A Hierarchical Active Binocular Robot Vision Architecture for Scene Exploration and Object Appearance Learning.

Gerardo Aragón-Camarasa

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy
School of Computing Science
College of Science and Engineering
University of Glasgow

October, 2012

Abstract

This thesis presents an investigation of a computational model of hierarchical visual behaviours within an active binocular robot vision architecture. The robot vision system is able to localise multiple instances of the same object class, while simultaneously maintaining vergence and directing its gaze to attend and recognise objects within cluttered, complex scenes. This is achieved by implementing all image analysis in an egocentric symbolic space without creating explicit pixel-space maps and without the need for calibration or other knowledge of the camera geometry. One of the important aspects of the active binocular vision paradigm requires that visual features in both camera eyes must be bound together in order to drive visual search to saccade, locate and recognise putative objects or salient locations in the robot's field of view. The system structure is based on the "attentional spotlight" metaphor of biological systems and a collection of abstract and reactive visual behaviours arranged in a hierarchical structure.

Several studies have shown that the human brain represents and learns objects for recognition by snapshots of 2-dimensional views of the imaged scene that happens to contain the object of interest during active interaction (exploration) of the environment. Likewise, psychophysical findings specify that the primate's visual cortex represents common everyday objects by a hierarchical structure of their parts or sub-features and, consequently, recognise by simple but imperfect 2D view object part approximations. This thesis incorporates the above observations into an active visual learning behaviour in the hierarchical active binocular robot vision architecture. By actively exploring the object viewing sphere (as higher mammals do), the robot vision system automatically synthesises and creates its own part-based object representation from multiple observations while a human teacher indicates the object and supplies a classification name. Its is proposed to adopt the computational concepts of a visual learning exploration mechanism that controls the accumulation of visual evidence and directs attention towards the spatial salient object parts.

The behavioural structure of the binocular robot vision architecture is loosely modelled by a WHAT and WHERE visual streams. The WHERE stream maintains and binds spatial attention on the object part coordinates that egocentrically characterises the location of the object of interest and extracts spatio-temporal properties of feature coordinates and descriptors. The WHAT stream either determines the identity of an object or triggers a learning behaviour that stores view-invariant feature descriptions of the object part. Therefore, the robot vision is capable to perform a collection of different specific visual tasks such as vergence, detection, discrimination, recognition localisation and multiple same-instance identification. This classification of tasks enables the robot vision system to execute and fulfil specified high-level tasks, e.g. autonomous scene exploration and active object appearance learning.

Contents

List of Figures

List of Tables

Acknowledgements

I am deeply thankful to my PhD supervisor, Dr Paul Siebert. It is difficult to not state that his guidance, patience, friendship and help over the last years were fundamental in this thesis.

Advice given by my second supervisor, Prof Roderick Murray-Smith, was an invaluable help at the earlier stages of my research. My deepest gratitude is also due to Prof Emanuele Trucco and Dr Bernd Porr for their useful and constructive recommendations that improved greatly this research work.

I also thank the support during this research project to the Programme Alβan, the European Union Programme of High Level Scholarships for Latin America (no. E07D400872MX) and CONACYT-Mexico.

Special thanks to my friends Maria Carla, Asma, Konstanze, Shazad, Juan Carlos and the *Mexicanos* who made Glasgow like my second home. Thanks to the members of the Computer Vision and Graphics Lab. They made the office and, specially, research fun and exciting.

Lastly, and most importantly, I wish to thank my family, parents (*Jefa* y *Jefe*), *lil-sis* (carnalita) and Steph. Although I was far from home, they gave me the strength, support and love I needed to complete my research. To them I dedicate this thesis.

Author's Declaration

This thesis presented a work that was carried at the University of Glasgow under the supervision of Dr. J. Paul Siebert, School of Computing Science, during the period between October 2007 to September 2011. I declare that this thesis is entirely my own work and it has not been previously submitted for any other degree or qualification in any university.

Gerardo Aragon-Camarasa

Glasgow, October 2012

Notation

Standard

- (x, y, σ, θ): x and y locations and scale and orientation of SIFT features, respectively.
- $\mathbf{A}$: attentional shroud.
- $\mathbf{B}$: bounding box.
- c_x, c_y: x and y image coordinate centres.
- C: covariance matrix.
- D_M^2: Mahalanobis distances.
- D_E: Euclidean distances.
- $\mathbf{F}$: SIFT features that are being learned.
- $\mathcal{G}$: cluster assignments.
- $\mathbf{H}$: projection of SIFT features into continuous Hough space.
- H_0: statistical null hypothesis.
- $\mathcal{H}^A$: set of attended/identified objects.
- $\mathcal{H}^P$: set of putative objects.
- $\mathcal{H}^S$: set of salient features.
- $H(x, y, S, \varphi) + +$: Hough Transform histogram.
- $\mathcal{I}$: numeric index of the the object pose class in the database that produces the recorded object hypothesis.
- $\mathbf{L}$: Sift features of the left camera.
- $\mathbf{M}$: Sift features of a model image.
- MQ: match quality values of the multiple same-class object instance detector.
- $\mathcal{M}^{(\mathbb{L},\mathbb{V})}$: list of logical indexes between SIFT feature matches of $\mathbf{L}$ and $\mathbf{V}$.

- η_i: object detection confidence score.
- PS_{ratio}: number of motor steps per pixel required to translate$x$ into actuator space.
- $\mathbf{R}$: Sift features of the right camera.
- std: standard deviation.
- s_t: quality view score where clusters are grouped with high confidence.
- $\mathcal{S}$: Saliency test.
- S: Silhouette numbers.
- $\bar{S}$: average Silhouette number.
- $\mathbf{T}$: Sift features of a test image.
- $\mathcal{T}_{SIFT}$: threshold value for the log likelihood test for the SIFT feature matching framework.
- $\mathcal{T}_{fixation}$: maximum threshold value of the two dimensional standard deviation of the coordinates in $\mathbf{B}$.
- $\mathbf{V}$: Sift features from the object database.
- $\vec{x}$: horizontal velocity vector.
- $\mathcal{X}$: fixation point in retinotopic coordinates.
- $\vec{y}$: vertical velocity vector.
- $\mathcal{Y}$: translated retinotopic coordinates with respect to the image centre.
- $\mathbf{Z}_{t-1}$ and $\mathbf{Z}_t$: compound matrices of SIFT feature, velocity vectors and angle poses.
- $\mathcal{E}_k$: stores the test and model SIFT features coordinates of the object instance.
- $\mathcal{E}^U$: contains the attended angular positions of the object's viewing sphere.

Greek Letters

- α: degree of overlap threshold for the multiple same-class object instance detector.
- β : discrete angle step.
- δ: pairwise difference of the x and y SIFT feature components of the left and right matched sets.
- ϵ_k: score of the kth attentional shroud.
- κ: saliency score.

- $\mu_{1/2}$: statistical median value.
- ϱ: radial polar coordinate.
- υ: statistical variance of a sample.
- ϕ: random generated angle.
- φ: direction polar coordinate.
- Φ: angular position of the object at the given observation.
- Φ_{sacc}: rotational saccade.
- Φ_t: current attended angular pose.
- Φ_{t-1}: previous attended angle pose.

Chapter 1

Introduction

The objective of this thesis is to investigate and design an integrated active binocular vision architecture built on 2D spatial sensing controlled by an autonomous attention control behaviour that interprets images in terms of local visual features and descriptions and directs the active vision system to achieve the task at hand. Specifically, the active binocular robot system is able to maintain vergence, direct its gaze, recognise multiple same objects, autonomously explore the environment and semi-automatically learn object representations in a unified and parsimonious hierarchy of visual behaviours.

1.1 Introduction

How does a robot "see" what is contained in the environment? How does it become an observer of the world and act upon the observed surrounding? This thesis attempts to address the above general questions by means of an active binocular robot vision system that integrates a hierarchy of visual behaviours in a structured, functional, and practical robotic architecture. The envisaged hierarchical architecture is inspired by the Sensor Fusion Effects (SFX, originally proposed by Murphy and Mali (1997)) which loosely resembles biological behaviours of a neurophysiological model of sensing in animals (McHaffie et al., 1989) without being an attempt to model the mammal brain itself. Therefore, this robot vision architecture is able to:

- carry out the specific tasks of *autonomous scene exploration and object appearance learning*.
- actively execute the visual tasks of *discrimination, detection, recognition, identification, and same-different identification* (as defined by Tsotsos (2008)), and,

- to address ultimately the *"lost and found"* problem (Forssen et al., 2008), i.e., the action of finding a set of objects present in cluttered, complex scenes.

The purpose of this thesis is therefore to explore and investigate state-of-the-art computer vision and image processing techniques in order to advance the current knowledge on the robot vision and the active vision fields.

1.1.1 Visual Tasks

Defined visual search tasks within the scope of this thesis and in accordance with (Tsotsos, 2008) are briefed as follows:

- *Discrimination tasks* embrace the ability of the robot to discern from multiple visual stimulus inputs. This visual task specifically concerns the distinction between stimulus classes in which the robot must determine the identity of the perceived stimulus. As pointed by (Tsotsos, 2008), this is basically how modern object recognition models work, e.g. an object is classified according to its extracted visual features in an image (as it will be described in Section 2.3).
- *Detection tasks* comprise the capability of the robot to determine whether the observed stimulus is noise or a known object class and, in consequence, the robot must act according to the perceived stimuli. This task enables a robot to localise visual cues that need further investigation (i.e. to attend the location of the observed stimuli).
- *Recognition tasks* only include the verification of a single stimulus. That is, the robot directs its gaze towards the stimuli and determines the identity of the observed stimulus. Whereas *Identification tasks* consist in assigning responses to each perceived stimulus in the environment (i.e. active exploration of a scene).
- *Same-Different Identification tasks* cover the case where the robot should decide if two or more elements are of the same class or different.

1.2 Motivation

At the beginning of the author's research project, an initial design of an active binocular robot vision system (Fattah, 2007) was provided. This robot vision system was structured as a collection of ad-hoc functions which were capable of autonomously exploring the environment

but only able to recognise, verge on and saccade to just one single instance per known object class for each invoked visual task. That is, this system, in its original form, was able to perform *detection and recognition* visual tasks of single object classes. This system is further studied in detail in Chapter 3. Similarly, current state-of-the-art robot vision systems have been proposed under the same trend, although few are reported in the literature. For example, in (Kragic et al., 2005; Björkman and Eklundh, 2005a; Wallraven and Bülthoff, 2007a; Rasolzadeh et al., 2010; Meger et al., 2008; Kootstra, 2010), robot vision systems have been enabled to somehow address the *"lost and found"* problem, as coined by Meger et al. (2008) (i.e. robot, go and find me my brown socks).

In the above robotic systems, the following key limitations are identified:

- Vergence and object recognition operating in conjunction with a visual attention model and visual search strategies by means of a single visual representation have not been entirely enabled for autonomous scene exploration such that the robot operates in complex, dynamic, and unstructured environments containing multiple instances of the same object class.

- The aforementioned robot vision systems are only capable of recognising and locating individual objects in a controlled environment (i.e. stable lighting conditions and isolated objects with some overlap between known and unknown objects). In consequence, *Same-Different Identification* visual tasks are not considered.

- Manually pre-trained and segmented object databases are used for the correct characterisation of the object appearance over the viewing sphere. Thus, such systems are able to execute *detection, recognition and identification* visual tasks and to properly operate in the specified application which achieves high recognition rates at the cost of a large object knowledge database and within only slightly cluttered environments.

- The active vision paradigm is partially adapted since some of these systems rely on different sensory inputs (e.g. range laser finder sensors), as in (Meger et al., 2008) in order to support the next location to be visited. Similarly, Björkman and Eklundh (2005a) have adopted a human's input to weakly supervise the selection of the next location for the correct recognition and localisation of objects.

- In the same trend, these robot vision systems have been proposed as a collection of ad-hoc visual mechanisms which perform a particular visual task. To overcome this limitation, Kragic et al. (2005) (and in an extended version reported in (Rasolzadeh et al., 2010)) have proposed a behavioural structure; however, its components are designed as small ad-hoc functions for the specific purpose these functions are designed.

This reduces the applicability and portability of a general purpose robot application into different context domains as further discussed in Chapter 2.

- Kootstra (2010) and Wallraven and Bülthoff (2007a) have devised semi-autonomous active exploration-learning strategies integrated into robot vision systems in order to learn an object's appearance. However, their strategies are conceived as a fixed set of sampling positions over the object's viewing sphere. This constraint therefore biases the representation of the object internally, and, in consequence, a trade-off in the recognition rate is obtained (as discussed in Section 2.7.2).

Hence, the purpose of this research is to integrate visual competences loosely inspired by the WHERE and WHAT streams into a hierarchy of visual behaviours and, consequently, integrated as a robot vision architecture for the robustness and applicability in a wider range of robotic applications. In other words, the more general objective of this research is to advance towards a general purpose robot vision architecture that can be applied to *flexible robot work-cell automation, space exploration, telemedicine, humanoid robots for elderly care, Unmanned Aerial Vehicles (UAV) and so forth.*

The following specific objectives are established in order to address current limitations of state-of-the-art robot vision technology and, in consequence, advance the active vision field.

- To integrate visual competences in order to improve visual search capabilities and therefore to explore cluttered, complex scenes
- To develop a multiple same-class object instances detector to improve perceptual capabilities on a visual object search exploration task
- To develop a hierarchy of visual behaviours which in turn provides an "open-ended" architecture in terms of the task specification
- To semi-automatically learn an object in order to synthesise its visual feature representation based on canonical views observed over the viewing-sphere
- To evaluate the performance and quantitatively measure the robustness of the described robot vision system with respect to current the state-of-the-art robot systems

1.2.1 Scientific Questions

This thesis concretely attempts to answer the following scientific questions.

- How does a robot explore autonomously a cluttered, complex scene?
- How can a robot be capable of detecting and localising multiple same-class object instances with occlusion and self-occlusion scenarios (i.e. Same-Different visual task)?
- Can the adoption of an hybrid reactive/deliberative robot architecture potentially provide the means of modelling visual attention and visual streams for robotic applications?
- Which type of robot architecture would provide an *"open-ended"* architecture in terms of the task specification (i.e. visual competences can be incorporated while preserving its underlying basic structure)?
- How could an active exploration-learning strategy of the object's appearance across its viewing sphere characterise and synthesise robustly robot's object knowledge?
  - What are the required visual streams involved in order to learn the object's appearance?
- To which extent the active interaction with the object facilitates learning feature descriptions?
- Validation of the hierarchical robot vision architecture:
  - To which extent a learned or manually segmented object database improves the ability of the robot vision system to carry out *Discrimination, Detection, Recognition, Identification, and Same-Different Identification* visual tasks?
  - Is the recognition rate achieved by a learned object representation above current state of the art robot vision systems?
  - What are the benefits of a learned object representation as opposed to a manually segmented and fixed sample object knowledge in visual tasks?
  - Is the robot more robust in terms of its visual search capabilities and the detection of multiple same object instance classes than state-of-the-art robot vision systems?

1.3 Background

Probably the first robot vision project dates back to 1966 when Marvin Minsky (founder of the Artificial Intelligence Lab, MIT) asked his undergraduate student, Gerald Sussman, to connect a television camera to a machine during a summer project such that *a machine could visually recognise objects in real-world settings* (Crevier, 1993). The project was fairly complex and

ambitious for the time frame and, moreover, for the computing resources and computer vision techniques available at that time. This initial attempt, however, provided the required insights to design capable seeing machines. In that regard, successful robot vision systems were designed over the next two decades (from mid 1960 to 1980). Early recorded attempts of developing robot vision machines included Freddy (Fikes and Nilsson, 1971) and Shakey (Rosen et al., 1965) robots; however, they were constrained to specific objects and environments and simple object recognition techniques, as further described in Section 2.3.

Current robots are no longer constrained to specific objects or environments. They can carry out more complex tasks on evolving and dynamic scenarios and are more aware of their interactions with the world. The continual advancement in imaging hardware and image processing and analysis techniques has enabled scientists to devise and to design robust real-world robotic vision-based systems (Meger et al., 2008; Rasolzadeh et al., 2010). Active robot vision systems are generally required to explore the environment, track objects, identify interesting objects, and so forth (Aloimonos et al., 1988; Ballard, 1991). The development of these systems has found its design principles in the biological vision machinery.

On the one hand, the attentional mechanisms which demand to carry out visual tasks (as those defined in Section 1.1.1) are equally divided into two types of modalities: the so-called *pre-attentive* and *attentive* modes of attention (Chun and Wolfe, 2004; Aziz et al., 2006; Berman and Colby, 2009). On the other hand, the active vision paradigm asserts that vision is a behaviour structured as a *perception-action cycle* paradigm and the perceptual visual behaviours act according to the physical visual *stimuli* (Aloimonos et al., 1988). Behaviour is believed to be the building block of animal intelligence (Murphy and Mali, 1997) and, in consequence, it is defined in this thesis as a process that maps sensory inputs to a pattern of motor actions that contributes to achieve a visual task. Thus, active vision research seeks to build computational seeing machines (or robot vision systems) based on a broad variety of approaches in order to attempt to answer those questions posed at the beginning of this introductory chapter (Section 1.1). The goal is thus *to enable a robot to "see" what is happening in the environment and, consequently, to "act" upon the changing surrounding*. Binocular robotic vision, in this thesis, is adopted to calculate relative depths of the observed environment, to provide more visual information than monocular vision and, consequently, to improve decision-making for visual operations while executing the above described visual tasks.

Vision in mammals is an active process that closely follows the perception-action cycle paradigm. That is, mammals direct their eyes successively towards a visual stimuli in order to produce single precept of the environment and, therefore, act upon the environment (Girard and Berthoz, 2005; Berman and Colby, 2009; Wurtz et al., 2011). In the literature, it has been proposed several cognitive models of eye movements (i.e. saccadic movements). These models

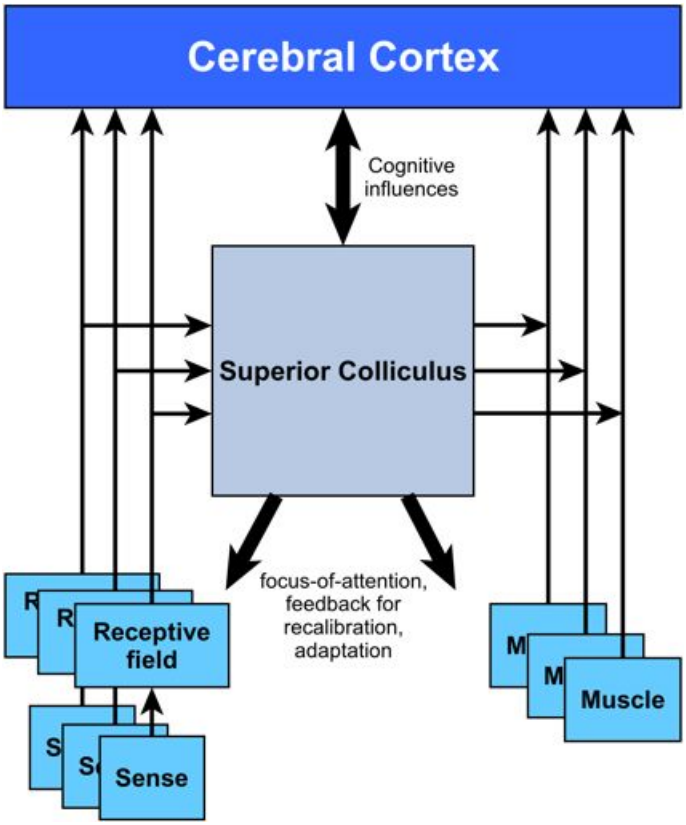


Figure 1.1: Neurophysiological model of sensing in animals. (McHaffie et al., 1989; Murphy and Mali, 1997)

attempt to reproduce the building blocks of the mammal's brain in order to understand how it works while carrying out similar visual tasks as those described in Section 1.1.1. In (Girard and Berthoz, 2005), it is reviewed computational models of eye movements according to the activated brain areas while emphasising in their psychophysical and philosophical implications. Overall, the main purpose of the the models described in (Girard and Berthoz, 2005) has been to create biological plausible computational models in order to simulate the behaviours of the mammal brain. It must be noted, however, that this thesis does not attempt to develop a biological plausible robotic system but to design a functional and parsimonious system that adopts concepts found in biology and is able to solve similar visual behaviours. Therefore, these models are only employed as an inspiration to develop a robotic working system.

In that regard, the link between perception and action, in the robot community, has been achieved by means of robotic architectures (as reviewed in Chapter 2). Robotic architectures are classified according to their purpose: pro-reactive, hierarchical and hybrid deliberative/reactive architectures. In this thesis, a reactive/deliberative architecture is adopted as discussed in the following subsection. These architectures have been inspired by biological models found in nature, e.g. the neurophysiological model of sensing in animals (McHaffie

et al., 1989) depicted in Figure 1.1. Specifically, this model associates action-motor commands with perceptual inputs that are relevant for a particular behaviour of an animal (Murphy and Mali, 1997). That is, an animal only perceives what they need to, following closely the perception-action cycle describe above (Aloimonos et al., 1988; Murphy and Mali, 1997; Berman and Colby, 2009).

This configuration provides the required abilities to establish high-level task-goals according to the state of the environment. Whilst reactive/simple behaviours either serve the ultimate goal, or inhibit the high-level layer which allows the robot to reflexively act in relation to events in the environment. A more in-depth discussion can be found in Chapter 2.

1.4 The Hierarchical Binocular Robot Vision Architecture

Figure 1.2 shows a conceptual illustration of the hierarchical binocular robot vision architecture. The devised architecture is inspired by the *Sensor Fusion Effects* architecture (Murphy and Arkin, 1992) (SFX, further discussed in Section 2.6.3). The design rationale of this architecture closely follows the neurophysiological model of sensing in animals depicted in Figure 1.1 and the perception-action paradigm.

As observed in Figure 1.2, visual and motor behaviours in the robot vision architecture are distributed over different layers accordingly, each subserving specific goals according to the level they are arranged. In this case, low-level layers are concerned to simple, innate behaviours to either process perceptual inputs or execute motor commands that come from higher levels. In this thesis, sensor information is represented in terms of iconic visual features which areeither (a) passed to high-level cognitive behaviours or (b) processed in the mid-level visual behaviours(Fazl et al., 2009; Shen et al., 2011).Mid-level behaviours are loosely modelled according to the dorsal (WHERE) and ventral (WHAT) streams. The WHAT stream consists of visual feature matching operations, object recognition and memory management of feature descriptions stored in an object knowledge database (i.e. the analogue of the IT cortex in the human brain). The WHERE stream, on the other hand, carries out coordinate-based operations such as the detection of objects, tracking of the spatial locations of either putative objects or attended objects, the control of the robot's gaze towards the indicated locations, and, finally, the management of feature locations in the object knowledge database. Both streams are connected to the action-motor stream (the analogous of the superior colliculus as depicted in Figure 1.1) that moves the cameras in order to enable the robot to explore actively the environment.

The above streams include visual behaviours that essentially execute the visual tasks described

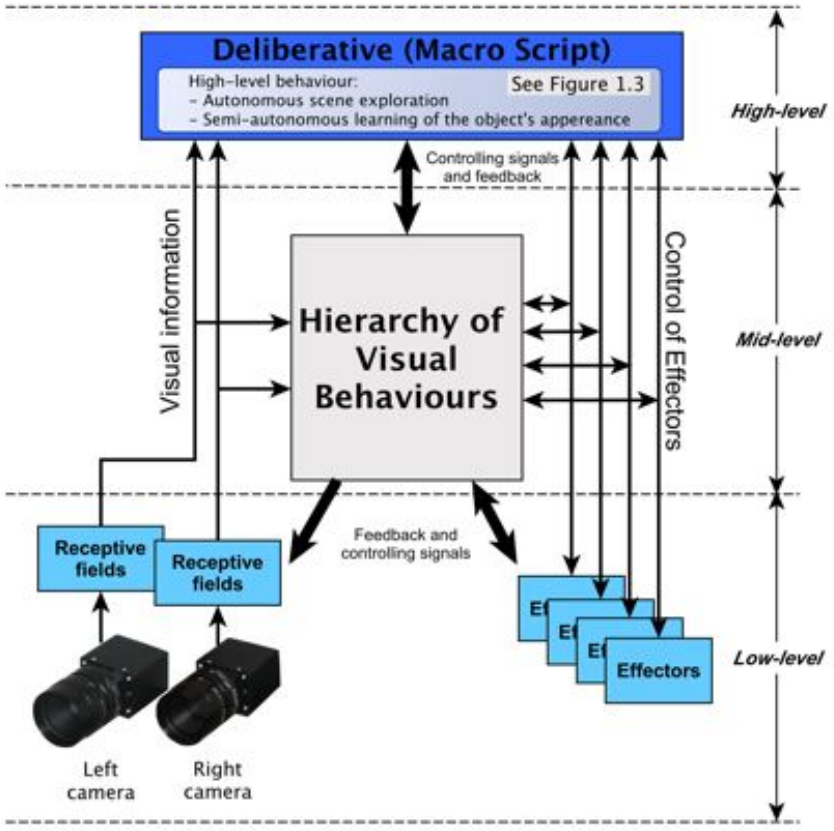


Figure 1.2: Concept design of the hierarchical active binocular robot vision architecture.

in Section 1.1.1. These streams are linked to the pre-attentive and attentive mode of attention and they are, in turn, divided into purpose orientated behaviours that can be triggered by a high-level signal in order to fulfil a high-level task-goal of the robot vision system. These oriented behaviours closely follows the philosophical foundations of paradigm of the SFX architecture. Specifically, the design of mid-level behaviours is related to a hierarchical structure where visual behaviours only serve a specific purpose (e.g. object recognition, binocular vergence, to name but a few), whilst more complex behaviours determine the attentional mode of operation for the current goals of the system (e.g. pre-attentive and attentive behaviours). This configuration is specifically represented by means of a *hierarchical robot architecture*. The details for each visual behaviour within the hierarchical architecture are described in the forthcoming chapters.

Finally, the high-level/deliberative layer is defined as a macro script where a user indicates the sequential activation of visual behaviours in order to accomplish the goal. In overall, this macro script can be devised by following the flow diagram described in Figure 1.3. As stated in previous sections, this thesis implements two types of high-level task-goal behaviours: *autonomous active scene exploration* and *semi-autonomous learning of the appearance of an object* (as depicted in the deliberative/high-level layer in Figure 1.2). Both high-level

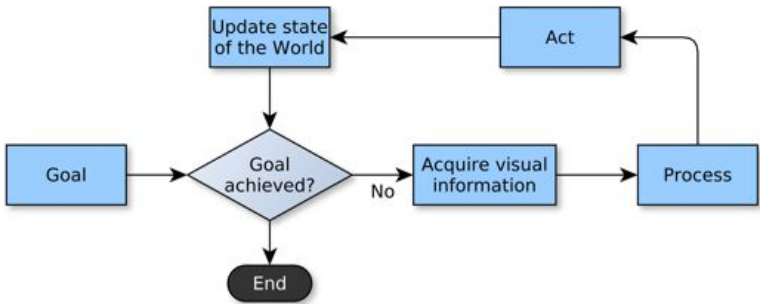


Figure 1.3: Overall flow diagram for the task-goal specification in the macro script.

behaviours therefore activate the required mid-level behaviours according to the task-goal specification. The subsequent chapters detail the design of the high-level macro script.

1.5 Contributions

This thesis advances current literature in the robot and active vision fields with the following key contributions (ordered as they are presented in this thesis):

- A complete investigation and characterisation of the initial binocular robot system (Fattah, 2007).
- Functional demonstration of a novel multiple same-class object instance detector that allows a robot vision system to pre-attentively detect and localise up to 6 objects with high confidence.
- Design and development of an integrated and novel hierarchical architecture for the active binocular robot vision system which is capable of performing *discrimination, detection, recognition, identification, and same-different identification* visual tasks. This architecture features binocular vergence, attention control, identification of multiple same-class object instances, and inhibition of return behaviours.
- Novel demonstration in the literature of the full active vision paradigm applied to learn the appearance of an object. This semi-autonomous active visual learning is integrated within the hierarchical robot vision architecture in order to automatically synthesise and create a part-based object representation knowledge by means of the dynamic interaction with the object.
- Functional demonstration and validation of the hierarchical architecture in cluttered and complex scenes featuring multiple same-class object instances with occlusion and self-occlusion.

The research reported herein has appeared in the following publications:

- Fattah, Haitham and Aragon-Camarasa, Gerardo and Siebert, J Paul, "Towards Binocular Active Vision in a Robot Head System", in Ramamoorthy, Subramanian and Hayes, Gillian M., ed., Towards Autonomous Robotic Systems, TAROS 2009 (University of Edinburgh, 2008), pp. 25–32.
- Aragon-Camarasa, Gerardo and Siebert, J Paul, "A Hierarchy of Visual Behaviours in an Active Binocular Robot", in Kyriacou, Theocharis and Nehmzow, Ulrich and Melhuish, Chris and Witkowski, Mark, ed., Towards Autonomous Robotic Systems, TAROS 2009 (University of Ulster, 2009), pp. 88–95.
- Aragon-Camarasa, Gerardo and Fattah, Haitham and Siebert, J Paul, "Towards a unified visual framework in a binocular active robot vision system", Robotics and Autonomous Systems 58, 3 (2010), pp. 276–286.
- Aragon-Camarasa, Gerardo and Siebert, J Paul, "Unsupervised clustering in Hough space for recognition of multiple instances of the same object in a cluttered scene", Pattern Recognition Letters 31, 11 (2010), pp. 1274–1284.

This thesis has also been presented in the following venues:

- Poster presentation of the initial development of a multiple same objects class detector method on a British Machine Vision Association technical meeting held on the 14th of May, 2008 in London, UK.
- Oral presentation of an improve hierarchy of visual behaviours in the Third EUCogII Members Conference, Palma de Mallorca held on 8th-9th October, 2011.
- Poster presentation of the Semi-supervised learning behaviour in the Third EUCogII Members Conference, Palma de Mallorca held on 8th-9th October, 2011.

1.6 Thesis Outline

The structure of this thesis is organised as follows:

- Chapter 2 provides an overview of different concepts, theories, methodologies, approaches, and applications required to develop and design a robot vision system. This chapter also surveys and discusses related approaches and methods.

- Chapter 3 reviews the initial active binocular robot vision system and presents a full validation of the system.
- The multiple same-class object instance detector is introduced and developed in Chapter 4.
- Chapter 5 presents the formulation and derivation of a hierarchical active binocular robot vision architecture.
- Afterwards, Chapter 6 details the semi-autonomous learning of the object's appearance behaviour.
- Chapter 7 presents the validation of the overall hierarchical architecture.
- Finally, conclusions and future work are presented in Chapter 8.

Chapter 2

Literature Review

The purpose of this chapter is to provide an overview of the different concepts, theories, methodologies, approaches and applications that are required to design a robotic system that can see and explore its environment. The design of robot vision systems is not constrained to specific areas of study but to a wide range of areas as presented. Thus, the study of different research fields is required in order to gain the necessary foundation knowledge. This chapter therefore covers an in-depth study of current robot vision systems, feature extraction techniques, robot architectures, computational models of human visual attention and the psychology and physiology of the human visual system. The chapter concludes by identifying the advantages and shortfalls of existing computer and robot vision technology and suggests the required methods and technologies to advance current state-of-the-art robot vision systems and scientific knowledge.

2.1 Robot Vision

Robots are machines that perform a predefined task either autonomously or guided. Such machines are equipped with different sensors in order to achieve the desired goals and tasks. Among current sensor technologies available (i.e. haptic, positional, ultrasonic range sensors and so forth), the most powerful sensor that can capture a wide variety of information is vision. Therefore, vision is the process of acquiring and extracting visual information from digital images of the observed three dimensional world such that the robot perceives, acts and interacts within a dynamic world.

Robot vision systems base their foundations on two different vision paradigms; the *passive* and *active vision* paradigms. On the one hand, the passive approach is mainly inspired by Marr

et al. (1983) seminal publication in which it is outlined a series of visual processing steps, the so-called "*Primal Sketch*". This theoretical framework has been used to develop computer vision techniques for object recognition, 3-D reconstruction and so forth. On the other hand, the active vision[1] paradigm (Aloimonos et al., 1988) is formulated based on the dynamic nature of the human visual architecture, that is, the human vision system is capable of moving the eyes in order to act based on the recovered information from dynamic observations. Both paradigms are reviewed in the forthcoming subsection.

2.1.1 Vision Paradigms

The relevance of using the passive approach in computer vision lies in the parallel processing of the imaged environments and the delivery of immediate information about the scene. In general, the computer vision community has mainly focused on the static analysis of passive observations of the environment (Aloimonos, 1993; Findlay and Gilchrist, 2003) which are the main building blocks of current research in computer vision and active vision itself.

Passive vision in a robotic context provides a pictorial view of the world. This view is captured by means of a single or a sequence of frames where all the visual information is analysed and processed without orientating the sensor to capture a different portion of the environment. Therefore, passive systems rely on complicated reasoning and computations in order to create useful percepts that contribute to the task at hand (Aloimonos et al., 1988; Aloimonos, 1993). For example, Ferrari et al. (2006) devise a system that simultaneously classifies objects based on contour information. Likewise, Mikolajczyk et al. (2006) develop an image retrieval engine based on the classification of visual features in accordance with the texture and geometric invariant measures of feature locations. Despite passive systems have produced successful commercial products such as *Google Images* (being the most notable example in the image information retrieval context), passive systems are not capable of selecting how to observe the scene and they are therefore limited to the current view to extract all meaningful information to fulfil the task requested. In addition, some problems in passive vision might become ill-posed such as locating multiple objects shape from shading, shape from contour, shape from texture, structure from motion and optic flow as pointed out by Aloimonos et al. (1988).

In the robotic context, Walther et al. (2005) has exploited the passive vision paradigm in order to learn visual features of salient regions for robot localisation. Their approach consisted of capturing image data over a fixed time interval using a static camera setting while the robot navigated a pre-defined path. Their system was capable of recognising previously observed

[1]It has been argued in the early nineties that "*active vision*" might be confused with "*active sensing*" (Ballard, 1991); however, robot vision scientists have used the term "*active vision*" extensively to define the control of cameras mounted in a robot. Therefore, the term "active vision" in this thesis will be employed.

locations within the navigation path; however, it suffered from a combinatorial explosion as the robot maintained an accurate representation of every salient region in the sampled imaginary without deciding if the visual information was relevant to the task at hand. Despite the limited capacity of passive systems within the robotics context, passive vision approaches has solved underlying techniques such as feature extraction, object recognition and so forth that are now used in the active vision field.

The environment is dynamic and observers should therefore be capable of engaging with it (Aloimonos et al., 1988; Findlay and Gilchrist, 2001). This, active vision systems are able to modify its intrinsic and external parameters to perceive the environment in a controlled manner while solving tasks and, consequently, to extract useful percepts about the environment that favourably contributes to the observer's goal. These visual paradigm therefore allows an observer to obtain information when it is needed (a requirement that has been discussed in Section 1.3). The underlying active vision premise states that vision is a behaviour geared by a *perception-action cycle* and the perceptual visual behaviours act according to the physical *stimuli* (Aloimonos et al., 1988). Active vision research has therefore sought to build robot vision systems based on a broad variety of approaches to achieve the definitive goal: *to enable a robot to "see" what is occurring in the environment and, consequently, to "act" upon the changing surrounding*. Indeed, active robot vision systems are dynamic observers that exploit recovered information from the imaged scene and perform actions and fulfil tasks (Ballard, 1991). The active approach features a potentially unlimited amount of information due to the range of different viewpoints. The environment can therefore be perceived from different view locations when the visual tasks become ill-posed and, consequently, more possible solutions can be obtained to solve them.

Aloimonos et al. (1988) and Ballard's (1991) seminal publications in the early 90s inspired several scientists to build and design active vision systems. Overall, active vision system considered anthropomorphic features and consisted of steerable monocular or stereo cameras settings mounted over programmable mechanical structures: the so-called "*robot heads*" (further described in section 2.1.2). These early systems were the founding experimental test beds to devise attention mechanisms such as attention/gaze control, saccadic movements, vergence (stereo case), smooth pursuit and so forth. Moreover, video-metric active systems began to appear in order to automate gaze point selection and vergence of the cameras such that it is possible to build three-dimensional model of a human faces for telemedicine (Urquhart, 1997), telepresence applications (Heng, 1995) and autonomous 3D scene reconstruction (Boyling, 2002).

Within active robot vision research, Westelius (1995) produced the most notable research in the 90s by devising attention mechanisms closely related to the human visual system; pre-

attentive, attentive and habituation functions that allow the system to explore and manipulate objects in a computer simulated environment. By using a virtual world, he did not consider the dynamically changing world (i.e. camera sensor noise, lighting variations and so forth); nevertheless, later active robot vision systems have been inspired by his robot architectural design. The most notable examples are described in the following Subsection.

The active paradigm has also found applications for navigation and robot localisation. In this context, Spatial Localisation and Mapping techniques (SLAM) (Davison and Murray, 2002) have allowed a robot to create landmarks according to the dynamic acquisition of the environment. In other words, SLAM facilitates the selection of visual features based on competition and assigning relevant visual attributes. Hence the robot is aware of the three-dimensional spatial configuration of the observed environment and is capable of planning a path in order to navigate without colliding into objects. These type of applications include navigation over closed-world environments such as office and industrial settings (Davison and Murray, 2002) and urban environments (Newman et al., 2009) to name but a few.

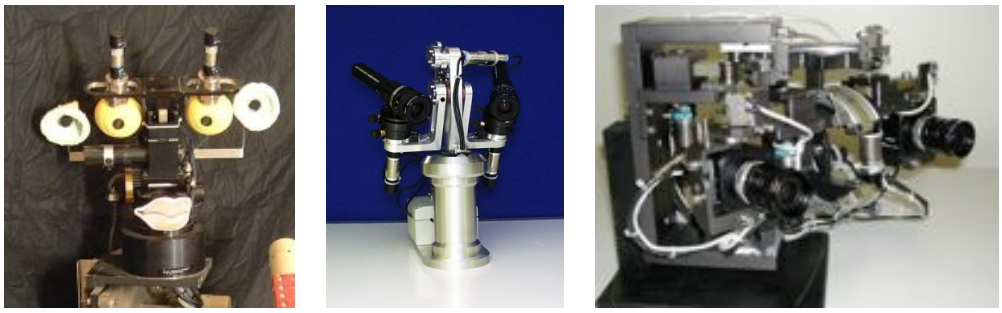
The ultimate goal in this thesis is to produce an active binocular system that explores scenes in order to solve *'lost and found'* tasks and to learn object appearances in a semi-supervised manner (as described in section 1.2). Therefore the underlying theoretic paradigm in which this thesis is built on is *active vision*.

2.1.2 Robot Heads Review

As a corollary of active vision research in the early eighties, scientists started designing embodied robotic machines that were capable of moving its cameras according to the physical stimuli of the environment. These type of robots were thus defined as "*robot heads*" and they generally simulate to some extent the behaviours of the human visual system. They also provide, in the engineering perspective, a solid point of reference in which to develop programmable visual behaviours, visual tasks and complex low-level and high-level operations. Robot heads are commonly used as a sensor component within more sensing and motor capabilities such as humanoid robots (Welke et al., 2010) and mobile robot platforms (Meger et al., 2010).

Several attempts have been made to build such heads. One of the earliest robot heads reported was "*Richard the First*" from the *Turing Institute, Glasgow* (Mowforth et al., 1990) that was capable of controlling the azimuth, elevation and neck of the dynamic head. Stereo microphones were also integrated to extend its application to telepresence and video conference.

Further examples in the same period are the "*Harvard Head*" (Ferrier and Clark, 1993)



(a) (b) (c)

Figure 2.1: (a) LIRA-robot head; (b) Yorick robot head; (d) Medusa robot head.

that could perform tracking, attentive servoing and depth computations form controlled ego-motions; the "*KTH*" robot head (Li et al., 1994) from the *Royal Institute of Technology* in Stockholm was considered one of the best heads available in the nineties (Boyling, 2002). Among its abilities, it could replicate human head and oculomotor motion speeds and move independently through all its 13 degrees of freedom in parallel. Likewise, the anthropomorphic sensor platform designed at the Turing Institute/University of Strathclyde (Heng, 1995), developed at the same time as the *Active Stereo Probe* (ASP) (described in section 2.1.3), was a low-cost robot head with high anthropomorphic and anthropometric accuracy. It could perform long-range telepresence and was built to operate with a low-cost personal computer. The Strathclyde anthropomorphic robot head received media popularity due to British television appearance and articles on newspapers[2].

Recent robot heads include the LIRA-head (Natale et al., 2002) (Fig. 2.1(a)), where acoustic and visual stimuli based on a space-variant sensor was exploited to drive the head gaze. The head was part of a humanoid robot which was used in the design of intelligent robotic systems with the intention of learning object representations by grasping actions on a fixation based task. This head later became part of the state-of-the-art cognitive robot platform, *iCub* (Orabona et al., 2005).

The Yorick head (Björkman and Eklundh, 2005a,b) (Fig. 7.1(b)), which integrates two cameras for foveal and other two for peripheral vision, is able to localise and detect objects using multiple 3D cues and object appearance. The same robotic platform has been used over different projects; for instance, Kragic et al. (2005) designed a system to recognise objects by matching through a predefined database using local feature descriptors (as discussed subsequently in 2.2). The authors estimated an object's pose by combining object appearance, geometric models and depth recovery based on different visual feature descriptions and search strategies. Another modern robot head is the Medusa head (Bernardino, 2004) (Fig. 7.1(c)) where high-accuracy calibration, gaze control, control of vergence and real-time speed track-

[2] `http://www.telepresence.strath.ac.uk/television.htm` (verified on the 30th, July 2012)

ing at 50Hz with a space-variant sensor were successfully demonstrated.

Current state-of-the-art robot head systems include the research by Rasolzadeh et al. (2010), where an embodied robot is capable of grasping and manipulating isolated objects in the robot's workspace, thus being the immediate descendant of the Yorick head. Similarly, (Meger et al., 2008; Forssen et al., 2008; Meger et al., 2010) have developed a mobile active robot platform that is capable of finding objects over controlled office settings. This system consists in a trinocular camera arrangement: two wide-field cameras for depth perception and attention control; and one high-resolution foveated camera for object recognition. This robot obtained the first place in the 2007, 2008, and 2009 "Semantic Robot Vision Challenge"[3] editions.

2.1.3 The Active Stereo Probe - Origins of the Active Binocular Robot Head

The robot head that this project is built on, found its origins in the *Active Stereo Probe* (ASP) (Urquhart, 1997) (descendent of *"Richard the First*", section 2.1.2) prototype form the *University of Glasgow* and *Turing Institute*. The hardware configuration consisted of two steerable cameras (8-bits grey-scale images at 768 by 576 pixels of resolution interfaced with a video frame grabber) and a steerable mirror mounted in dedicated *Physik Instrumente* high-accuracy pan and tilt platforms. Each camera and mirror had independent actuation which was driven by two rotatory platforms and brushed servo-motors respectively. A projector was also mounted at the bottom of the mirror such that light was beamed and actively directed towards the scene; likewise, the baseline between cameras could be manually adjusted as required. This work demonstrated active vision, vergence, stereo matching, camera calibration and 3D perception in a complete video-metric system. Software components developed during this research work were then incorporated in the so-called photogrammetry analysis tool *C3D* (Siebert and Marshall, 2000).

Thereafter, the ASP was upgraded in the actuation control (Boyling, 2002), as depicted in Fig. 2.2, which consisted of a *Physik Instrumente C-842.40* 4-channels card for the actuation of the cameras and a *Physik Instrumente C-842.20* 2-channels card to control the optical lens. The previous projector and some mechanical components were also updated as the system was in storage for a number of years. The ultimate objective of this research project (Boyling, 2002) was to investigate a dynamic foveated active vision system for dynamic photogrammetry analysis and autonomous 3D reconstruction of scenes. By adopting a foveation scheme in stereo matching and 3D reconstruction, it was required to develop a new dynamic calibration approach not previously reported in the literature. This system was therefore capable

[3]http://www.semantic-robot-vision-challenge.org/ (verified on 30th, July 2011)

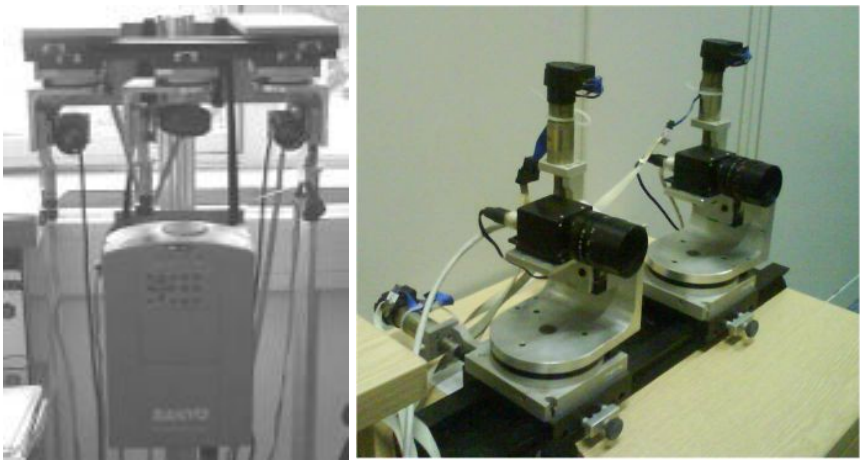


Figure 2.2: Left: The second version of the ASP as in Boyling (2002); right: Current robot head hardware.

of reconstructing 3D scenes by selecting portions of the image to be observed at the highest spatial resolution possible in a multi-resolution strategy for improved performance.

The second version of the ASP became unused for a couple of years until McDougall's MSc project (McDougall, 2004). The ASP was thus heavily transformed to become the binocular hardware system at which this study began with (as depicted in Fig. 2.2). Therefore, the modified robot head comprised in *Physik Instrumente* pan and tilt mechanisms mounted such that the baseline can be adjusted as required by the user. This head also featured a colour and a monochrome Sony cameras (DFW-X700 and XCD-X700 respectively) at 1024 by 768 pixels of resolution. The projector and the optical bench were not longer considered in the development. The ultimate goal of the project was to control the actuators and camera capturing from direct calls using the MATLAB API and C/C++ instructions. A MATLAB GUI provided the software interface to control and adjust actuator and camera parameters.

This thesis is developed from the founding robot vision software of Fattah's MSc project (Fattah, 2007) also reported in (Fattah et al., 2008). In this project, Fattah designed the initial gaze control, vergence, and object recognition engines. A detailed description of this project is given in chapter 3.

2.2 Image Feature Extraction

Image feature extraction is a dimensionality reduction technique such that the resultant information provides a low dimensional description of the image contents. Examples comprise image/template indexing, matching, tracking, camera calibration for 3D reconstruction, pose estimation, image alignment and so forth (Tuytelaars and Mikolajczyk, 2007). In general, two

types of feature extraction techniques can be found over the literature:

- *Global features* describe properties/contents of an image by taking into consideration all pixels in the image. Approaches include colour histograms, principal component analysis and statistical measures for textures, to name a few. Application domains include scene and object classification, image retrieval and video mining. Global features present drawbacks in terms of dealing with occlusions, shape variability, image clutter and they are not entirely invariant to geometric deformations (Tuytelaars and Mikolajczyk, 2007).

  In the context of robot vision, global features have proven to be useful mainly for robot self localisation (Zhou et al., 2003). One of the main aims of this thesis is to allow a binocular robot to localise objects despite occlusions, image clutter and view pose transformations where global features fail to produce a repeatable, accurate, distinctive and efficient visual representation of the observed scene. Therefore, these type of features is no longer reviewed in the forthcoming sections in this thesis; a complete review of these techniques can be found in (Tuytelaars and Mikolajczyk, 2007).

- *Local features* are either points, edges or small patches within an image that subserve as anchors from which descriptions of the analysed region can be computed (Hubel and Wiesel, 1962; Tuytelaars and Mikolajczyk, 2007). Generally, a two-step procedure follows the computation of local features; firstly, it finds local *interest points,* and secondly it computes a descriptor vector of the region at which the interest point is localised. As opposed to global features, local feature extraction is the most used low-level visual representation in a wide range of applications due to its informativeness, locality, repeatability, accuracy and efficiency. These type of features are the visual representation on which the binocular vision system of this study relies on. Further details on local features are reviewed on the forthcoming subsections.

2.2.1 Local Feature Extraction Overview

The development of local feature extraction techniques has been evolving for over 50 years. The pioneering work of Hubel and Wiesel (1962) have inspired computer vision scientists to develop local feature extraction techniques in order to attempt to solve the general problem of perception. Initial efforts focused on extracting edges (Roberts, 1963) and later on interest points (Tuytelaars and Mikolajczyk, 2007) around corners and junctions as they were clear indications of structures recognised by humans. Earlier techniques consisted in extracting interest points along contour intersections and junctions, image intensities and colour histograms. These approaches were limited to the scope of the problem: they were not invariant

to geometric transformations and they usually performed well with very limited imaginary configurations.

Local feature extraction techniques have evolved according to the development of new hardware technology and the discovery of new insights about the human visual system. Over the last two decades, efforts on investigating more robust local features have been focused on finding interest points that are stable and distinctive. According to Tuytelaars and Mikolajczyk (2007) and Itti and Koch (2001), it is assumed now that corners, junctions or contours do not have to be necessarily extracted but anything that is an extremum of evaluated functions on digital edges, shapes or contours. For example, such techniques include corner detectors as the *Harris detector* (Lowe, 2004), blob detectors as the *Hessian detector* (Bay et al., 2008) or region detectors as the *Maximally Stable External Regions* (MSER) (Matas et al., 2002). Current robust local feature techniques, predominantly Lowe's *Scale Invariant Feature Transform* (SIFT) (Lowe, 2004), are inspired in Crawley's receptive field histograms (Schiele and Crowley, 2000) and Mikolajczik and Schmid's scale invariant interest point detector (Mikolajczyk and Schmid, 2002). Such visual representation has transformed the perceptual capabilities and tasks a robot can perform.

In the context of this thesis, the different visual tasks a robot can perform rely on the representational richness of the visual description capable of extracting in observed environments. The idea of using local features over different visual scales has shown itself to be the most effective and mature visual representation. Probably the most robust image descriptor for object recognition/detection are based on the computation of *histogram of edge orientations*. This type of descriptor has been employed in successful approaches such as the *Histogram of Oriented Gradients* (HOG) (Dalal and Triggs, 2005) and SIFT. The HOG share similarities to Lowe's SIFT descriptors, the main difference is where and how the histogram of edge gradient orientations is computed and their invariance properties. That is, HOG descriptors are extracted from a dense overlapped grid of uniformly spaced cells while SIFT descriptors are computed from localised interest points. Both approaches have been demonstrated in diverse recognition systems such as pedestrian detection (Dalal and Triggs, 2005), object recognition (Björkman and Eklundh, 2004; Dong et al., 2010), free path navigation (Johnson and Matthies, 2000) and spatial localisation and mapping (i.e. SLAM) (Davison and Murray, 2002; Newman et al., 2009).

One of the limitation of HOG like descriptors, in the context of this thesis, is that they are only invariant to local geometric and photometric transformations while SIFT descriptors are invariant to scale and rotation and some geometric transformations. In that regard, the object orientation (rotation invariance) in the image is a key aspect in the robot vision system described in this thesis while inspecting a scene and, consequently, HOG descriptors are not

suitable for the purpose of this thesis. Hence, the underlying visual representation adopted in this thesis is SIFT descriptors. The following subsection presents an overview of this technique and its object recognition pipeline.

2.2.2 Scale Invariant Feature Transform (SIFT)

The Scale Invariant Feature Transform (SIFT) is firstly described in (Lowe, 1999) and later in its extended paper version in (Lowe, 2004). SIFT have been extensively exploited; for example, in object recognition, image restoration, image similarity, and so forth. Moreover, Mikolajczyk and Schmid (2005) have demonstrated that SIFT overcomes, in terms of reliability, popular local feature extraction algorithms. Applications in the robot vision context can be found in (Kragic et al., 2005; Fattah et al., 2008). (A brief description of this algorithm is given below.)

The rationale behind SIFT is to transform a digital image into a collection of local feature coordinates and descriptors with invariant properties over scale. SIFT features are therefore invariant to scale and rotation, and reasonably invariant to illumination changes, image noise, deformations and pose changes within a range of ± 20 degrees for off-plane rotations of planar surfaces. SIFT features consist in four key stages summarised as follows:

1. *Scale space extrema detection*. To detect interest points that are invariant to scale and orientation, a scale-space detection scheme is used to find stable features for each level in the image pyramid. Each level in the pyramid is firstly convolved with a Gaussian filter operator. A Difference of Gaussians (DoG) is applied for each scale level from adjacent blurred images (as depicted in Figure 2.3) in order to find local extrema (i.e. interest points).

2. *Keypoint localisation*. In order to accurately locate interest points, each candidate keypoint found is interpolated with nearby pixels. This localisation is achieved by drawing an analogy between the located pixel and its neighbouring pixels within the extrema found in scale-space. For each extrema, the spatial coordinates are located by fitting a 3D quadratic function on the nearby data to select an extremum above a threshold. The rejection of responses along edges and low contrast interest points are removed following a ratio of principle curvature calculation and a given threshold. The principal curvature ratio is defined as $\frac{(r+1)^2}{r}$.

3. *Orientation assignment*. Rotation invariance is accomplished by computing histograms of local gradient directions obtained at each specific scale in the neighbourhood of the

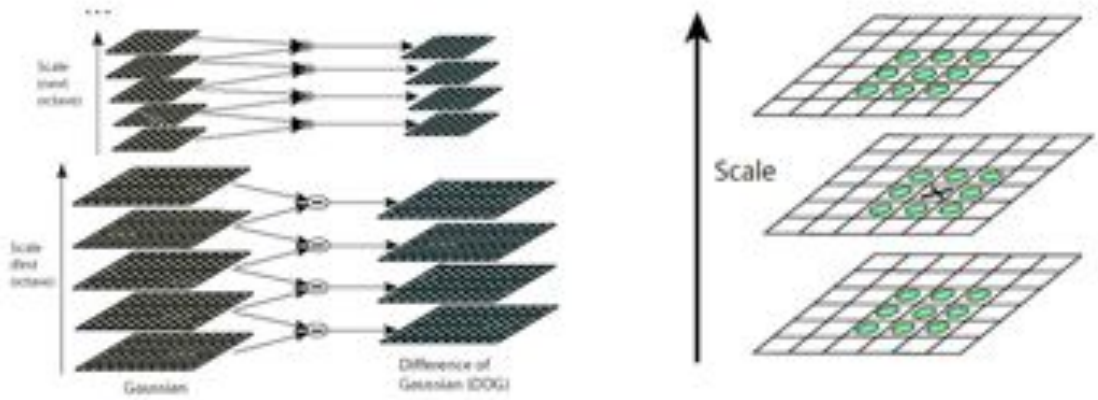


Figure 2.3: Scale space representation used and localisation through the difference of Gaussians (Lowe, 2004).

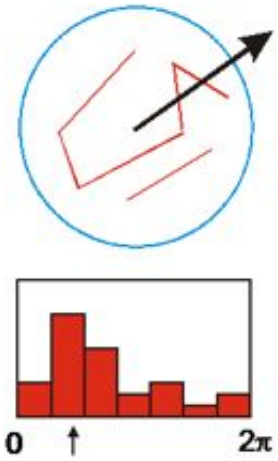


Figure 2.4: Orientation assignment by local gradient direction (Lowe, 2004).

keypoint (as showed in Figure 2.4). The canonical orientations are assigned to each keypoint. The peak of the gradient direction found corresponds to the dominant orientation. If multiple orientations are found, they are assigned to the same keypoint location.

4. *Keypoint descriptor*. Computing a gradient histogram sampled over a 16-by-16 pixel area in the scale space generates feature descriptors. The magnitude of gradients in eight directions is evaluated in a 4-by-4 pixel histogram (Figure 2.5 shows a toy example of the descriptor computation). The vector containing values of all orientation histograms entries results in the 128 dimensional SIFT descriptor ($4 \times 4 \times 8 = 128$). Finally, the feature vector is normalised in order to reduce effects of illumination changes. For each feature descriptor vector there is a corresponding keypoint that specifies 2D location, scale, and orientation.

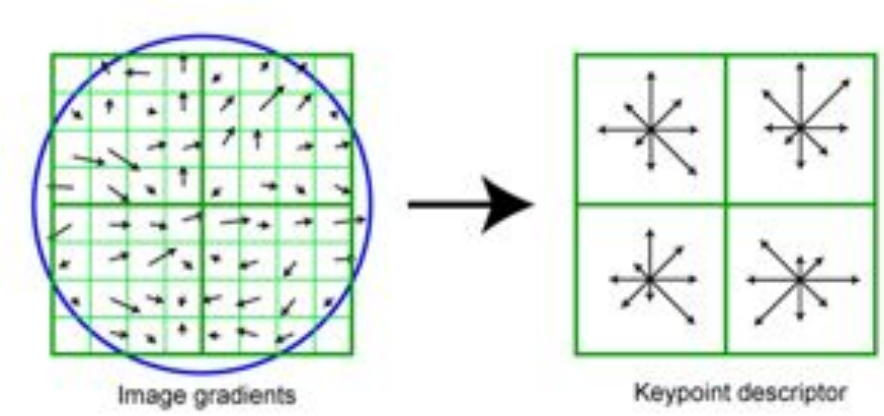


Figure 2.5: 2-by-2 pixel histogram over a 8-by-8 pixel area for the descriptor computation; although the actual descriptor defined in Lowe (Lowe, 2004) is obtained from a 4-by-4 pixel histogram over a 16-by-16 pixel area .

2.3 Object Recognition

In general, two types of object recognition models exist in the computer vision community; *model-based* (model-dependent) and *exemplar-based* (view-dependent) models.

M*odel-based systems* relate the intrinsic 3D properties of objects (i.e. simple geometrical primitives) and assume '*a 3D mental rotation paradigm*' in order to align and project the stored 3D model into the input image (Marr et al., 1983; Styles, 2005; Findlay and Gilchrist, 2003). It has been proven, however, that the human visual system recognises objects by simple but imperfect 2D view object part approximations (Bülthoff and Edelman, 1992) rather than perfect 3D models. Furthermore, the model-based paradigm, in a robot vision context, results in a system that increases its visual knowledge complexity over time as 3D models for each object observed have to be stored in memory. Therefore, this paradigm becomes computationally inefficient and complex as observed by Bülthoff et al. (2008).

E*xemplar-based models* have been the most successful resemblance of human visual operations according to neuroscience and psychophysical findings (Blanz et al., 1996; Styles, 2005; Bülthoff et al., 2008; Fazl et al., 2009). These models consist of view-based object representations that are later compared to simple transformations between input and stored views (Lowe, 2001; Kragic et al., 2005; Rasolzadeh et al., 2010). The most successful exemplar-based model reported is the so-called SIFT framework (Lowe, 2004; Mikolajczyk and Schmid, 2005); the following section overviews such technique.

2.3.1 SIFT Object recognition pipeline

A vision system capable of recognising objects in its field of view is a challenging task; however, the advancement of local feature extraction technology has proven to be the most useful visual primitive representation for object recognition. SIFT features, as described in section 2.2.2, are extremely useful to find patterns in images within some affine deformations. Lowe (2004) has described an exemplar-based object recognition pipeline (as depicted on Figure 2.6[4]) that consists in:

1. extracting SIFT features from a test image,
2. comparing features from the test image against a previously defined SIFT feature database (this database consist of a collection of images with their corresponding SIFT features),
3. finding geometric consistency across different poses stored in database, and, finally,
4. verifying object pose hypotheses based on an affine pose estimator that relates the location, scale and orientation of database and test features.

Matching SIFT features between a test and model images (top section in Figure 2.6) consists of finding the first and second nearest neighbours within minimum Euclidean distance of a test image descriptors with respect to a database of SIFT features of the model images. In order to increase robustness in the match (i.e. features that are diagnostic), the distance ratio between the first and second nearest neighbour is computed and compared with a predefined log-likelihood threshold not greater than 0.8. This threshold has been experimentally determined in (Lowe, 2004).

To find geometric consistency between input and database images, the *Hough transform* is used to detect any arbitrary shape by verifying that all edges of a detected object match a consistent centroid location. The Hough transform, first developed by Hough and Paul (1962), is a technique to locate curves in some parametric form of a given shape, for example; lines, circles, ellipses, and, also, to detect features (Yanez-Suarez and Azimi-Sadjadi, 1999) from any given image. The Hough transform has the property of being robust to discontinuities and being somewhat tolerant to image noise; however, the parametric function used to describe the shape is dependent on the parameter numbers used and, consequently, the computational cost increases in proportion to the parameter dimensions.

[4]MATLAB functions that illustrate the voting scheme and the Hough space accumulator were kindly provided by Mr. Euan Strachan.

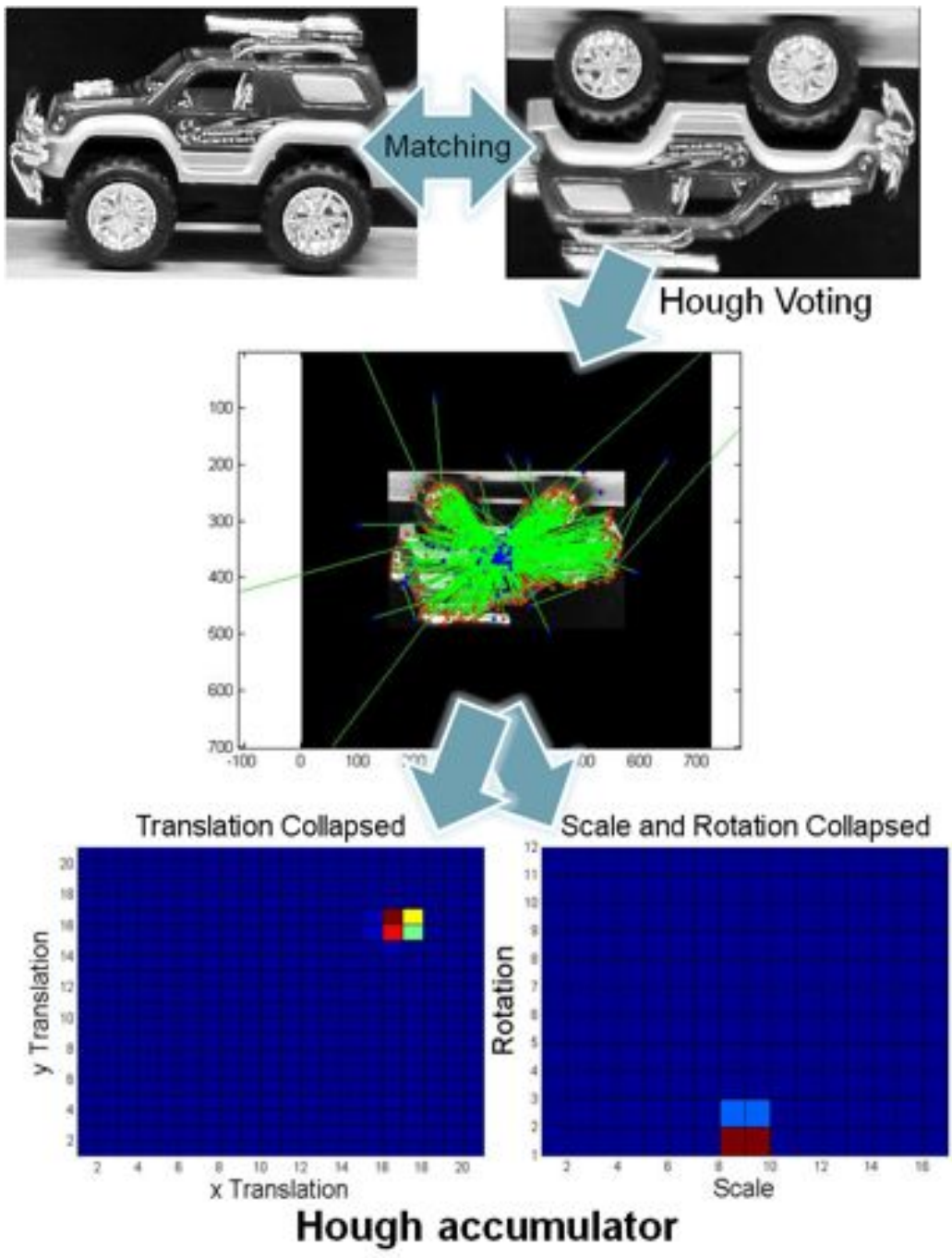


Figure 2.6: SIFT object recognition pipeline. Top: test and model images. Middle: The Generalised Hough Transform voting scheme. Bottom: Hough Space accumulator.

In the feature extraction context, the Hough transform can be generalised by means of a look-up table (R-table) instead of a parametric function as described in (Ballard, 1981). Lowe (2004) took advantage of this generalisation concept to identify feature groups by voting for all object poses stored in an image database that is uniform with the test features. The *Generalised Hough Transform* (GHT) in Lowe's paper (Lowe, 2001) is therefore employed to identify clusters of matches that vote for the same object pose and to reinforce a detection hypothesis between a particular object and a database of objects. In other words, each matched SIFT feature is assigned into a four dimensional Hough-space histogram such that the highest peak denotes the geometrical consistent interpretation for a group of features. This process is described as follows. a) A set of SIFT features of the database/model image (denoted as m): defined as $(x, y, \sigma, \theta)^m$, where x and y represent the location, and σ and θ denote the scale and orientation of the features. b) The corresponding matched set of the test image (denoted as t), $(x, y, \sigma, \theta)^t$, the geometric reference points are then calculated by:

$$\Upsilon^m = x_i^m + \sigma_i^m \cdot \left[\begin{array}{cc} \cos\theta_i^m & \sin\theta_i^m \end{array} \right] \tag{2.1}$$

$$\Upsilon^t = x_i^t + \sigma_i^t \cdot \left[\begin{array}{cc} \cos\theta_i^t & \sin\theta_i^t \end{array} \right] \tag{2.2}$$

where Υ^m and Υ^t denote the tuple of x and y locations; and i, each feature matched in the set. The middle section in Figure 2.6 shows the graphical representation when finding geometric reference points. Following this, the feature projections into Hough space are computed by finding the similarity transformation between model and test features. Thus, the linear equation that defines the relation between features is $Ab = c$, where:

$$A = \left[\begin{array}{cccc} x_i^t & -y_i^t & 1 & 0 \\ y_i^t & x_i^t & 0 & 1 \\ \Upsilon_x^t & -\Upsilon_y^t & 1 & 0 \\ \Upsilon_y^t & \Upsilon_x^t & 0 & 1 \end{array} \right] \tag{2.3}$$

$$c = \left[\begin{array}{cccc} x_i^m & y_i^m & \Upsilon_x^m & \Upsilon_y^m \end{array} \right]^T \tag{2.4}$$

$$b = \left[\begin{array}{cccc} x & y & S & \varphi \end{array} \right] \tag{2.5}$$

Hence, the least-square solution for b is the result of the x and y location, scale and rotation tuple as in Equation 2.5, respectively. Each type is then stored in a 4-dimensional Hough histogram accumulator. Projections are distributed into the four nearest neighbours for location and the two nearest neighbours for the scale and rotation. The highest bin in the 4D Hough

histogram represents the most geometrically consistent representation of all object poses (as represented at the bottom portion in Figure 2.6).

An affine pose estimator can be computed based on the transformation solution which yields on the location points and the weighted score of the Hough transform from model and test objects. The solution is found by using RANSAC and weighted least squares. Lowe (2004) stated that only three matched SIFT descriptors are needed to perform an object recognition task, providing some robustness to illumination changes, overlap between different objects and cluttered environments.

The GHT is used in this thesis to determine centroids of a hypothesised object in a robotic visual search task and to find natural keypoint clusters among objects for multiple instances of the same object detection and localisation.

2.4 Visual Attention

The development of computational visual attention models have been influenced by the need to understand and model the biological foundations of human vision. These models mimic operating modes of the human visual system and they are therefore employed to acquire relevant visual information over space and time. As Berman and Colby (2009) point out: *"attention acts upon sensory signals at many levels to construct a selective representation of the imaged space"*. Visual attention mechanisms within the scope of this thesis include *target selection*, *gaze control*, *eye convergence* and *inhibition of return mechanisms*; these concepts are structured below.

Attention, in the biological context, facilitates the selection of a subset of the observed world in accordance with the goals of the perceiver (Chun and Wolfe, 2004). It also enables to perceive at different levels the contents of the environment, namely *local* and *global processing* (Styles, 2005) (e.g. in a domestic setting, a cereal box is perceived at the local level, whereas, at the global level, a cupboard is perceived). Hence, attention is a behavioural system which assists the acquisition of information so that an observer can interact with the environment (Ballard, 1991).

Overall, attention mechanisms follow a "*perception-action cycle*" in which the observer acts and interacts with the environment (Neisser, 1976). The perception-action cycle, in the visual attention context, allows performing cognitive activities, for instance:

- *the perception phase* creates visual hypotheses of the perceived stimuli that need to be

inspected, and

- *the action phase* selects perceptual hypotheses in accordance with the task and/or the events in the environment and shifts attention to the reported locations.

In accordance with the perception-action cycle, an observer can direct its eyes towards interesting locations of the perceived environment (Chun and Wolfe, 2004; Arnow and Bovik, 2008). The reason is that human vision has about 1.25 millions photo-receptors cells densely packed in the foveal region in the human retina (1% of the total number of photo-receptor cells) (Grosso et al., 1995; Balasuriya, 2006) such that approximately 2^o of the field of view is perceived with high visual acuity. The purpose of such constrained visual representation in the human eye is that full field of view processing would require a massive amount of neural resources. The foveal region thereby reduces the information processing capabilities needed in the brain and allows casting aside large parts of the field of view that are not relevant to the goals and/or tasks of the observer. It must be noted that attention is directed to the location before the actual eye-shifts and, in consequence, the movements of the eyes are not the selection process itself, but only the outcome of attentional processes (Theeuwes, 1993). In that respect, attentional processes can be captured by the *attentional spotlight metaphor* (Posner and Cohen, 1984). This metaphor suggests that attention is directed as a spotlight over the visual field, and this spotlight improves the detection of visual events within its beam (Posner and Cohen, 1984). Thus, by having a reduced field of view and enabling to direct the eyes towards imaged targets, the attentional spotlight metaphor is an efficient and effective form to capture the nature of human information processing mechanisms and to allocate vision resources accordingly (Posner and Cohen, 1984; Posner and Petersen, 1990).

As human vision is *binocular*, , the ability to focus on the same real-world fixation point with both eyes after each eyes movement is performed by a *vergence* mechanism. This is responsible for minimising the induced disparities between eyes such that the inward movement of the eyes provide perceptual depth cues of the observed target in the environment (Bernardino and Santos-Victor, 1996). Furthermore, by verging both eyes, it enables an observer to maximise the visual information that can be extracted and perceived within the high-acuity region.

In addition to the attentional spotlight metaphor, it has been suggested that the attentional human visual attention is divided into two independent, hierarchical modalities of analysis running simultaneously, with the output of one feeding into the other. These modalities are categorised into *pre-attentive* and *attentive* stages (Chun and Wolfe, 2004; Styles, 2005).

The pre-attentive stage (or *covert attention*) operates in parallel across the entire visual field without changing the gaze of the observer (Findlay and Gilchrist, 2003; Chun and Wolfe, 2004). This mode is governed by a bottom-up (exogenous, stimulus-driven) and a top-down

(endogenous, conceptual-goal driven) cueing mechanisms. On the one hand, bottom-up mechanisms are captured by external stimulus and select salient regions over a physical input stimuli; neither recognition nor reasoning process are involved (Chun and Wolfe, 2004). On the other hand, top-down mechanisms are found when further understanding of the visual information must be done as part of a long term cognitive process. They are driven either by goals, memories or knowledge. In this thesis, these cueing mechanisms have an important role on the rationale of how the visual information is processed and arranged on the robot's attentional architecture.

The attentive stage (or *overt attention*) is limited in capacity in the sense that it is capable of focusing one item each time (Chun and Wolfe, 2004). This mode is responsible for the selection of pre-attentive cues that are informative with respect the goal or events and to modify and control the gaze of the observer in order to obtain high-resolution foveal vision (Findlay and Gilchrist, 2003). The gaze control is based on the movements of the eyes which are normally refereed in the literature as *saccadic movements*. In that regard, saccades are rapid motions such that the direction and destination cannot be altered once they have been initiated. Saccades occur after covertly selecting the next location to direct the gaze. Therefore, the attentive mode, as Yarbus (1967) observed in his study, governs a sequence of saccades to investigate and explore regions of the perceived environment in the quest of the best coherent interpretation of the available visual data.

Each time the eyes saccade to a new location in the visual space, new visual information is acquired. This continuous operation of the perception-action cycle is only possible by an inhibition of the observed locations of the visual space. Thus, *Inhibition of Return* (IOR), as termed in the literature (Itti, 2000; Chun and Wolfe, 2004; Styles, 2005; Aziz and Mertsching, 2008), enables the attentional system to explore the visual field of view without endless backtracking into the same location. IOR follows overt shifts of attention where the location of the cue is inhibited such that the attentional mechanisms are oriented towards novel items and environmental events not explored yet.

Hence, the pre-attentive and attentive modes control the analysis of information which happens to be in the the *ventral stream* and the *dorsal stream* (Fazl et al., 2009). The ventral stream analyses *WHAT* an object is; while the dorsal stream analyses *WHERE* an object is located in the perceived stimuli. Specifically, the WHERE stream maintains and binds spatial attention to the object coordinates that characterise the location of the object of interest and create the topological relations of feature coordinates with respect to a reference frame (Wallraven and Bülthoff, 2007a; Fazl et al., 2009). The WHAT stream either determines the identity of an object or triggers visual learning mechanisms (Wallraven and Bülthoff, 2007a; Fazl et al., 2009).

The behavioural structure of the binocular robot vision system in this thesis is specifically modelled by the pre-attentive, attentive, and inhibition behaviours in addition to the WHAT and WHERE information processing streams as described in Chapter 3. The study of vergence mechanisms is out of the scope in this thesis since the behavioural design proposed in (Fattah, 2007; Fattah et al., 2008) is implemented. For completeness, a brief description of the design rationale of the vergence behaviour is given in section 3.4.

2.4.1 Visual Search

The selection of where to deploy attention is carried out by attending locations or objects. That is, *location-based* selection states that attention is allocated based on the stimuli produced by spatial locations, whereas *object-based* selection defines that attention is deployed into object or groups of visual features (Findlay and Gilchrist, 2001, 2003). Therefore, attention mechanisms have different operational modes: to explore the visual field, to sense possible dangers in the environment, to inspect novel visual stimuli and so forth.

Visual search studies have been employed to discover insights about attentional mechanisms in human visual behaviour. The most influential visual search model in the literature is the *Feature-Integration Theory* (FIT) (Treisman and Gelade, 1980).

FIT defines that a variety of basic visual features are grouped automatically and in parallel across the field of view whereas objects are perceived and localised by serially deploying attention into them (Treisman and Gelade, 1980). Basic visual features comprise: *colour*, *intensity,* and *orientation,* which are pre-attentively coded into different feature maps. Further visual features have been integrated in recent models, for example, *motion* (Milanese, 1992), *depth* (Kragic et al., 2005; Frintrop et al., 2005) and *size* (Björkman and Eklundh, 2004) to name but a few. Feature maps are thereafter aggregated to produce a master map where top-down object knowledge bias the selection of attention. The final map is then perceptually grouped in order to produce potential loci hypotheses based on a Winner-Take-All (WTA) neural network approach such that every location is serially attended. Object recognition is then carried out by matching those group features in the observed location with respect to descriptions in long-term memory. Finally, each observed location is then inhibited by decreasing the strength of the signals in the master feature map.

Several other theories have been proposed, the most notable being the *Guided-Search* (GS) model (Wolfe, 2007). This model is a refined extension of FIT that takes into account visual search asymmetries of loci points in the feature maps and variability in search times.

The underlying visual search strategy developed in this thesis shares similarities with the

models described above in terms of dividing attention on several attention spotlights. However, attention control relies on the features produced by the visual representation, i.e. SIFT features (section 2.2.2) which are intensity, scale, and orientation.

2.4.2 Related Literature in Robot Vision

The most recent and successful work, inspired directly on FIT, is the computational attention model of Itti (2000). In this paper, the author proposes a *contrast-saliency map* which characterises the above metaphor in a passive setting. A *saliency map* is created by aggregating basic visual features that share similarities with the SIFT detector's difference-of-Gaussian computations (section 2.2.2). Specifically, three basic features are utilised in this model: *intensity*, *colour*, and *orientation*. The main idea is to depict how human vision distinguishes an object by its features and from its surroundings. In the first stage, visual input is represented as an iconic structure over topographic feature maps (salient features). Thereafter, a combination of these feature maps produces a global saliency map with respect to its neighbourhoods. The maximum salient region in the map drives and shifts attention towards that specific location. The next location is selected and the saliency map values decrease in order to inhibit visited locations (i.e. inhibition of return). Finally, this model has shown high correlation with actual human eye saccades under similar visual search tasks (Ouerhani et al., 2004).

Visual attention models have recently integrated top-down cues into the computation of saliency maps. For example, the passive robot vision system developed in (Walther et al., 2005) integrates bottom-up feature maps (i.e. the contrast-saliency model in (Itti, 2000)) in order to learn visual features from a collection of fixation points while recognising object classes in a top-down manner based on SIFT descriptors. Similarly, the *VOCUS* (Visual Object detection with a Computational attention System) attentional model (Frintrop, 2006) includes the bottom-up contrast-saliency model in (Itti, 2000) and top-down cues incorporating a goal-directed map. This model also features range sensing data from a laser scanner to aggregate depth features into the saliency map. This systems shows that a robot behavioural system is improved by fusing bottom-up and top-down cues with different sensing capabilities for object classifications and SLAM tasks.

Passive visual attention models have successfully demonstrated basic functions of the human visual system, even in some constrained robotic frameworks; however, these models have been tested with static images, without addressing the dynamic nature of the active vision in visual search applications. These models thus require the development of further gaze and attentional controls. Feature locations in space subserve the guidance of the attentional spotlight of the above attentional models; however, recent works have stated that the metaphorical

spotlight selects objects rather than locations in space (Chun and Wolfe, 2004; Fazl et al., 2009; Wallraven and Bülthoff, 2007b).

In that respect, Ballard (1991) has argued that different attentional gaze controls modes in accordance with the spotlight metaphor are classified as *holding* and *changing the gaze* behaviours. On the one hand, a robot has two different modes while *holding the gaze*; fixating or pursuing the moving target. Both modes, in a binocular case, include a vergence system that maintains the foveal region of both cameras over the same target by minimising local disparities. On the other hand, *changing the gaze* is carried out by high speed motions, which modify the position of cameras over a different target. This type of movement is commonly known as s*accadic movements* followed by verging behaviour.

Westelius (1995) has introduced the above gaze behaviours into a virtual robot application context. The aim is to drive the attention of his hierarchical gaze control based on three different components: *pre-attentive*, *attentive*, and *habitation*. The first two components work in a close loop and are based on an attention object based framework. The last one works as a crude inhibition of return map by reducing the significance of visited locations in the pre-attentive model, being one of the earliest attempts of integrating inhibition of return in the robotic context.

Recent reported developments in robot gaze control include the systems implemented by Forssen (2007), which performs automatic saccadic gaze control in a mobile robot unit with active binocular cameras based on SIFT features of object locations. However, in an effort to model the nature of visual search scan paths, robotic scientists have devised heuristic visual search mechanisms according to the task in hand. These heuristics schemas are mainly driven by the outputs of available feature extraction techniques. For example, Kragic et al. (2005) use depth recovery to segment the scene according to the distance between the targeted object and the robot as part of a visual object-search strategy. Likewise, Meger et al. (2008) implement a saliency map including intensity, colour and depth features to drive attention biased by top-down features based on the MSER feature detector (Matas et al., 2002) for object recognition and navigation.

2.4.3 Foveated Vision

Over the last decades, the concept of foveated imaging has become popular among computer vision scientists. This visual representation is cast as a non-uniform spatial allocation of visual information on the input stimulus for visual processing and reasoning in the early primate's visual pathway (Balasuriya, 2006). Researchers have studied the implement-

ation of this concept on computer vision approaches (Schwartz, 1995; Balasuriya, 2006), and in the robotics field (Panerai et al., 2000; Orabona et al., 2005; Bernardino, 2004), in both Charge-Coupled Devices (CCD) cameras or space-variant software implementations (Balasuriya, 2006). Nevertheless, small objects in the periphery within a cluttered scene might be invisible to a foveated system as noted by Ballard (1991), foveation provides the ability to have high spatial resolution over small areas and a low spatial resolution in the periphery in order to reduce the amount of information that needs to be processed in high resolution while still capturing in low detail events that might be of interest to the observer..

Software based foveation can also be approximated without resorting to complex mapping processes. That is the case of using *symbolic foveation* of visual features from the image, *foveated image pyramids* and *focused image segmentation* of a small portion of the image. Each software based foveation is described below.

- *Symbolic segmentation* refers to the extraction of computed visual features in accordance with a region of interest, such that focus of attention is only concentrated on that portion of the observed image. This case has been used extensively in robot vision while attending objects in the field of view (Fattah et al., 2008; Meger et al., 2008).
- *Focused image segmentation*, or *active segmentation* as in (Mishra et al., 2009) is accomplished either by extracting portions of the image around salient regions (Björkman and Eklundh, 2004; Walther et al., 2005; Rasolzadeh et al., 2010) or by indicating fixation points based on salient features from the image (Mishra et al., 2009).

The use of the above types of foveation in this thesis is threefold. Firstly, foveation is accomplished by symbolically segmenting visual features from the image, while invoking a visual exploratory task for object recognition and identification. Secondly, segmenting target objects around fixations based on motion and edge cues creates foveated spatial regions in order to bind visual features into objects while learning objects. And finally, the integration of foveation in the scale-space computation of the SIFT framework could potentially enable the study of the efficiency and the capacity of adaptation to the behavioural architecture with different visual representations.

2.5 Detection and Localisation of Multiple Same Objects

There have been several attempts to tackle the problem of simultaneously detecting multiple objects in the computer vision community. Such approaches comprise the multiple recognition of same class objects and/or instances of the same object class by relying on either local

feature-based or *template-based* matching procedures in combination with unsupervised or supervised learning techniques or geometrical frameworks. In the biological context, the multiple detection of objects while pre-attentively analysing the field of view is a task-demanding process. By assuming that objects are the units for dividing the attentional spotlight (section 2.4.1), object-based theories suggest that the analysis of the perceived environment is constrained to a restricted number of objects (Chun and Wolfe, 2004).

The above hypothesis has been demonstrated to hold true under different psychophysical experiments. These experimental frameworks have mainly consisted of asking subjects to keep track of multiple moving objects while fixating a target point in a screen (Chun and Wolfe, 2004). The studies have demonstrated that humans own a limited capacity to covertly divide the attentional spotlight among 5-6 independent objects at the same time without making eye movements (Chun and Wolfe, 2004), and detection is limited by the overlap degree among objects (Franconeri et al., 2010). The simultaneous localisation of many objects therefore restricts the attentional resources as each object in the observed field of view competes for the peripheral resources (Ballard, 1991). A survey of the most notable techniques is given in the following subsection.

2.5.1 Multiple-Instance Detection Literature Review

In the passive vision context, one of the first attempts to detect many objects is the work proposed by Yanez-Suarez and Azimi-Sadjadi (1999). They employed self-organising neural network maps in order to group collinear and parallel projected edges in Hough space. That is, their approach finds specific geometric shapes in scanned images based on an object shape defined in terms of simple geometric primitives. In their method, it is only possible to detect specific shapes defined by their geometric primitives, in this case irregular parallelograms produced by airborne fibreglass particles. Furthermore, they claim a template-free point of view for the detection of partially occluded objects but their approach is only capable of detecting instances of a single object class.

Template and/or area based matching approaches are reviewed in (Davies, 2005). Specifically, an application of object localisation is provided based on a simple hole detection in conjunction with a graph-theoretic approach on a pair of biscuits. The computational load is one of the biggest problems of this method, and, in real matching tasks, it can be unmanageable. El-Sonbaty and Ismail (2003) describes a similar approach, where they propose an algorithm using area-based matching to recognise occluded objects. By searching for three connected lines in the test and model objects, these lines are marked as they share the same distance ratio and angle relations to the last matched template. This approach achieves robustness in

partially occluded objects. It is evident that progress has been made on template/area-based approaches; however, Davies (2005) argued that the "*search for objects via their feature descriptions is far more efficient than template matching*".

Hence, with the advancement of feature extraction technologies, it is now possible to detect many objects of different classes based on local descriptions that are invariant to geometric transformations. For instance, the *PASCAL Visual Object Classes challenge*[5] is intended as a platform to develop object class recognition algorithms based on local feature extraction techniques. This challenge has been running each year from to 2005 to date. The challenge consists of twenty object classes that are grouped into an annotated database consisting of persons, animals, vehicles and indoor realistic images. Such database is therefore employed to train unsupervised or supervised learning algorithms to retrieve objects within test images. The evaluation criteria is based on *Receiver Operating Characteristic* (ROC) curves and *Precision and Recall* (PR) curves; these graphical tools are better described in (Fawcett, 2006). The challenge has two main competitions, which are the core of the contest, named classification and detection, where learning and classification algorithms in combination with a defined visual understanding model are used to compete in one of these contests.

Within this challenge, the proposed approaches differ in the use of the learning technique or in the selection of the visual understanding model. That is, the SIFT derived approaches are the ones with more reliable results in simultaneously detecting object classes. For example, (Mikolajczyk et al., 2006) reports a generative model to recognise and localise multiple object classes simultaneously. Their method consists of building hierarchical code-book representations (i.e. bag-of-words model) of several object classes in order to model the joint appearance-geometry probability distributions of local scale and orientation invariant feature descriptors. Their probabilistic recognition approach allows them to deal with possible overlaps between objects as well as ambiguities occurring between similar object classes. Although their approach achieves a degree of robustness in an image retrieval context, they demonstrate the detection of only five different object classes and this requires several thousand training images. Consequently, the computational cost of this approach appears to exceed by far that viable in a robotics scenario.

In the robot vision context, the detection of multiple same object class instances methods have served to covertly divide attention over many objects within the observed environment. In that regard, Zickler and Veloso (2006) proposed a method using PCA-SIFT descriptors for multiple object localisation, and, in a similar approach, (Zickler and Efros, 2007) are able to cope with a degree of shape variability exhibited by deformable objects. In this paper, SIFT keypoint matches with respect to a trained database are grouped within a voting space based on

[5] `http://pascallin.ecs.soton.ac.uk/challenges/VOC/` (verified on the 3th, July 2011)

hierarchical clustering techniques and then used in a humanoid robot scenario. They control the sensitivity of multiple instance detection by defining a distance threshold for merging groups such that if the Euclidean distance does not meet the threshold criterion, two clusters are merged into one until there are no more possible groups. While this approach achieves multiple instance same-class recognition, they restrict their method to two object classes. The objects are assumed to appear in isolation under different scenarios and the acquisition of visual information is through a passive camera setting, i.e. object hypotheses are not further validated.

As discussed above, the detection and localisation of multiple instances is carried out by grouping image features that observe certain characteristics that describe an instance. In the following section, grouping techniques available in the literature are reviewed.

2.5.2 Clustering Algorithms

Several clustering algorithms exist over the literature, each one is designed for a particular condition, application or input data. These grouping techniques are commonly applied to separate data in accordance with some properties in order to infer and describe (and sometimes learn) what the intrinsic properties of the data are.

Clustering algorithms are classified as: *hard partitioning* and *hierarchical* (Kaufman and Rousseeuw, 2005). Algorithms for each category are explained below, as well as some recent approaches.

2.5.2.1 Hard Partitioning Techniques

The most popular clustering algorithm in the hard partitioning category is *k-means*. This algorithm attempts to iteratively minimise a metric distance with respect to prototypical centroids (i.e. "*cluster means*") until an objective function is minimised. That is, a collection of observations, $X = \{x_1, x_2, \ldots, x_n\}$, are partitioned into c cluster means such that an objective function, J, (equation 2.6) is minimised.

$$J = \sum_{i=1}^{n} \sum_{j=1}^{c} \|x_i - c_j\|^2 \tag{2.6}$$

where x_i is the ith of $d - dimensional$ measured data, c_j is the $d - dimensional$ centre of the cluster, and $\|x_i - c_i\|$ is the norm expressing the similarity between measured data and the centre.

K-means is a simple clustering algorithm; however, it suffers from being outlier sensitive. In that regard, *K-medoids* or *Partitioning Around Medoids* (PAM) share similarities with k-means but the underlying difference prevails in the selection of cluster means, of which actual data points are employed in order to find representative groups in the data set (Kaufman and Rousseeuw, 2005).

A different hard partitioning is the *fuzzy C-means* (FCM) clustering algorithm, proposed by Bezdek (1981). It attempts to partition a finite collection of elements $X = \{x_1, x_2, \ldots, x_n\}$ into a collection of c fuzzy clusters with respect to an objective function defined as equation 2.7.

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{c} u_{ij}^m \left\| x_i - c_j \right\|^2 \tag{2.7}$$

where m is the degree of partial members of a cluster that affects the clustering result, u_{ij}^m is the degree of membership of x_i in the cluster j, x_i is the ith of $d - dimensional$ measured data, c_j is the $d - dimension$ centre of the cluster, and $\|x_i - c_i\|$ is any norm expressing the similarity between any measured data and the centre.

This clustering algorithm is powerful due to the use of an objective function (equation 2.7) that assigns probabilities to points that belong to a certain cluster by means of a membership matrix. That is not the case of the previous partitional clustering methods; *k-means* and *partition around medoids* which do not avoid the hard partitioning, and, in consequence, they eventually find solutions with empty clusters.

2.5.2.2 Hierarchical Clustering Techniques

Hierarchical clustering algorithms generally fall into two categories: *divisive* and *agglomerative*. Divisive clustering approaches produce a top-down hierarchy (Kaufman and Rousseeuw, 2005) by assuming that all data points form a single cluster. Subsequent clusters are split based on a flat clustering algorithm by means of a hard partitioning clustering technique (i.e. k-means) until each data point is its own member cluster. Divisive approaches are computationally demanding as a partitioning technique is evaluated at each level and each branch of the hierarchy. In computer vision, these approaches are generally employed to create codebooks of visual features as in (Muja and Lowe, 2009) where the authors propose a fast approximate nearest neighbour solution for matching large databases of visual features such as SIFT (section 2.2.2).

Agglomerative techniques assume that all data points are cluster centres from which sub-

sequent clusters merged based on a bottom-up analysis between-cluster dissimilarity metric (Kaufman and Rousseeuw, 2005) (equations 2.8 and 2.9). There are several agglomerative techniques, where the underlying difference one from the other is the dissimilarity distance between clusters. The most representative agglomerative techniques are described in what follows.

- *Single linkage* is the most simple agglomerative algorithm. Groups are formed in accordance with the minimum distance between closest observations with respect to the minimum inter-group pairwise distances. Clusters are merged into/with the furthest pair of observations until there is only one group (equation 2.8). This algorithm performs better with an elongated shape data distribution; however, this grouping technique is known to provide the least useful results in this category (Kaufman and Rousseeuw, 2005).

- *Complete linkage* is best summarised by Johnson (1967) where he proposed to merge groups with the furthest distances between the closest pair of observations until there is only one group. This clustering method provides acceptable results with a compact but not well separated data shape. Thus, given a $m \times n$ data matrix to be clustered, its ultra metric inequality is expressed in equation 2.9):

$$D \leq \min \{d(x, y), d(y, z)\} \tag{2.8}$$

$$D \leq \max \{d(x, y), d(y, z)\} \tag{2.9}$$

In equation 2.8 and 2.9, x, y and z are objects of a certain group and $d(x, y)$ and $d(x, z)$ might be any given metric defined. The metric employed in this thesis is thereby the standard euclidean distance (equation 2.10) for all the above clustering techniques described:

$$d_{a,b} = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \tag{2.10}$$

2.5.2.3 Recent Clustering Techniques

The selection of a clustering technique relies on the data to be processed, and the aim of the grouping process. Several variations of the above clustering algorithms might robustly perform for specific conditions or data types; however, it does not guarantee that the chosen algorithm suffice to all possible cases.

In that regard, modern clustering algorithms are moving towards the combination of different grouping models, towards convergence into an optimal solution or reduction of spatial data in order to develop more sophisticated and general approaches. For example, Frigui and Krishnapuram (1999) used a robust competitive clustering algorithm to take advantage of the hierarchical and hard partition clustering approaches. They also determined the optimum number of clusters via a process of competitive agglomeration. Rosenberger and Chehdi (2000) established that the optimum number of clusters in the data is based on a modified LGB clustering algorithm (simple k-means) and the stabilisation among partitions. An image segmentation application is given in order to show the robustness of the method.

A more recent approach is described by Sanguinetti (2008) where he proposed the reduction of the observed data dimension using a probabilistic latent variable model; the model is evaluated in synthetic and real data. The disadvantage lies when overlapped clusters have a condition which, as the author denoted (Sanguinetti, 2008), "*may be difficult to detect*". Nevertheless improvements on data clustering have been made, a truly general model for any kind of data type is yet to be devised.

In this thesis, clustering algorithms are employed to find natural groups of projected SIFT features in Hough Space for multiple instance of the same object class localisation and detection. Furthermore, clustering algorithms are also considered to group potential object parts based on depth and motion cues (as described in chapter 6). Specifically, the implemented algorithms in this thesis are the *fuzzy C-mean* and the *complete agglomerative hierarchical* algorithms.

2.6 Robot Architectures

In the human visual system, scientists have asserted that attention is part of a behavioural system (Ballard, 1991; Chun and Wolfe, 2004) that tightly couples sensing and acting to perform, in conjunction with other sensing capabilities, intelligent actions in the real-world. In that regard, behavioural systems in insects and mammals are believed to be a modular arrangement of simple behaviours that improve their survival likelihood within the environment (Styles, 2005). A group of behaviours thereby conforms a whole system capable of performing abstract computations, complex tasks and actions that modify the environment.

The study of animal behaviours has served as a biological foundation to design robots that exhibit intelligence in real world situations (Brooks, 1991; Murphy and Mali, 1997). It is agreed that behaviours are the fundamental building blocks of intelligence on which robot architectures have been designed (Murphy and Mali, 1997). Behaviours balance, interact,

cooperate, dominate, or cancel each other according to the goals and state of the environment. Robots are thus equipped with behaviours that allow the interaction with the world in order to make changes, perform actions, and/or sense according to the perceived world, i.e. perception-action cycle described in section 5.4.

In the biological context, a behaviour is *"a mapping of sensory inputs to a pattern of motor actions that are used to achieve a task"* (Brooks, 1991; Murphy and Mali, 1997). Behaviours can be categorised in terms of how they were acquired. These are (according to (Murphy and Arkin, 1992)) *"innate behaviours*, *inborn with a sequence of innate behaviours*, *innate behaviours that need an initialisation sequence* and *learned behaviours"* (responses are learned from experience). There are also three different categories based on the purpose of a behaviour:

- *Reflexive behaviours* that are innate and simple stimulus-response mappings in accordance with the environment.
- *Pro-reactive behaviours* that are learned stimulus-response mappings but without any conscious deliberation.
- *Deliberate behaviours* that perform decisions on the stimulus presented in order to output a response.

Robot architectures are thus programmed by coupling the perception-action cycle into arrangements of independent behaviours such that they subserve complex actions and tasks. Robots therefore become more intelligent by having more behaviours (Brooks, 1991; Bonasso et al., 1997; Murphy, 2001). Robot architectures should be designed according to a set of principles (as described by Murphy (2001)). That is, architectures should:

- preserve modular engineering principles,
- exhibit a viable degree of precision and reliability while executing the task,
- be portable to other robotic domains, and
- be sufficiently robust for their purpose.

In the literature, there are mainly three robot architectures: the *reactive*, *hierarchical* and *hybrid-deliberative/reactive* architectures (Murphy, 2001). These type of architectures are outlined on what follows.

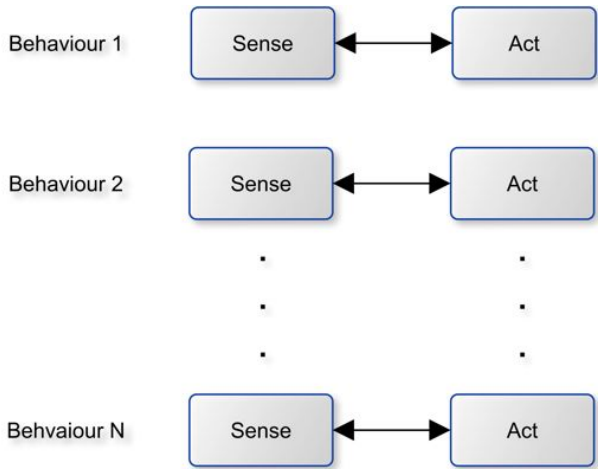


Figure 2.7: Composition of the reactive architecture.

2.6.1 Reactive Architectures

Reactive architectures find its origins in Brook's seminal works (Brooks, 1986, 1991): the *Subsumption Architecture*. This architecture operates in accordance with a "*sense then act*" paradigm (as depicted in Figure 2.7); that is, robots under this paradigm mimic insect behaviours. They gained popularity due to its simplicity, low-cost operation and their biological resemblance with animal behavioural analysis (i.e. the *ethological approach* as termed in the literature). They are commonly defined as a set of behaviours that instinctively react to events in the world without reasoning about the actions, i.e. sense then act (Murphy, 2001). Therefore, local sensing of the environment is reflexive and directly mapped to motor responses in order to achieve a task.

Under the *Subsumption Architecture*, behaviours do not own memory and are vertically decomposed in layers, each with a predefined goal (Brooks, 1991). Behaviours are thereby grouped into competence layers, and they start with general and primitive abilities that are controlled, overridden, and/or subsumed by more specific goal-directed behaviours in higher levels. For instance, if a level of the hierarchy is active, lower levels are subsumed in their defined behaviour. Hence, this architecture enables an incremental and sequential, bottom-up operation of the system. The overall behaviour of the robot is thus a consequence of the responses within the environment and behaviours rely on the state of the world without maintaining a global internal representation (Brooks, 1991).

Pro-reactive architectures are by definition limited on the tasks. That is, new behaviours must be reprogrammed in other to increase the robot's competences. Such architectures also present portability issues as they are underpinned by the the changes of the task and the environment which limits the functionality to the targeted environment. However, they maintain modular programming and exhibit robust reliability within the defined task. Furthermore, if a higher

Figure 2.8: Composition of the hierarchical architecture.

level of the hierarchy fails, it does not imply that lower levels will not be active; therefore, the robot is capable of surviving within the environment.

2.6.2 Hierarchical Architectures

Early attempts to provide robots with some degree of awareness arose out of the hierarchical arrangement of behaviours (a review of this type of robots is given by (Murphy and Mali, 1997)). The hierarchical architecture follows, in accordance with the perception-action cycle, a control theoretic, top-down sense-plan-act loop (as depicted in Figure 2.8). In other words, the robot senses the environment, plans based on the perceived stimuli, and, latterly, carries out the actions.

According to the design principles in Section 2.6, early architectures were not robust at performing the goals they were designed for and they required vast amounts of computational resources to reason, plan and maintain a detailed world map of the environment. Brooks (Brooks, 1986, 1991) also argued that behaviours should be designed in a vertical decomposition (in accordance with the ethological approach) as opposed to a horizontal one, e.g. the *Subsumption Architecture*. However, the hierarchical architecture presents superior modularity and portability as opposed to the reactive architecture. Furthermore, modern hierarchical architectures have been proposed with improved and robust performance at fulfilling tasks, e.g. mobile robots in a robot football league (Behnke and Rojas, 2001) and autonomous operation in unsupervised urban driving (Urmson et al., 2007). Some scientists have agreed that hierarchical arrangements support the modularity of the mind and the evolution of intelligence (Ullman, 2007; Arbib et al., 2008).

In that regard, second-generation architectures have been based on Brook's *Subsumption Architecture* while preserving its horizontal decomposition. Thus, the underlying function relies on the time interval at which the level of planning and execution complexity is required. In other words, low-level behaviours are simple stimulus-response mappings that perform basic functions which are thus controlled by higher layers. Low-level behaviours subsume some other low-level behaviours, whereas higher levels do not inhibit low-level functions.

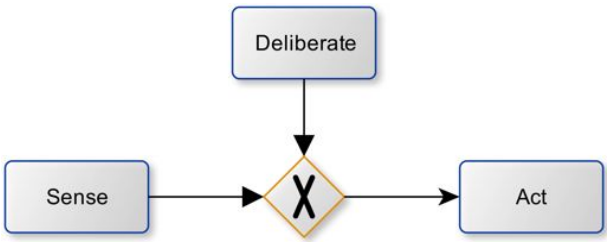


Figure 2.9: Composition of the hybrid deliberate/reactive architecture.

2.6.3 Hybrid Deliberative/Reactive Architectures

Hybrid deliberative/reactive architectures embrace two different architectures that they closely operate together: *deliberative* and *reactive* architectures. The reactive component, as previously defined, fulfils the requirements of reflexive actions without reasoning about the state of the world, while the deliberative component refers to the layer of intelligence that is conscious of the actions of the robot (Murphy and Arkin, 1992). Therefore, sensing the state of the world is directly mapped to reflexive actions while the planning stage is decoupled from the reactive behaviours (as depicted in Figure 2.9): the robot senses then acts while plans. The deliberative layer is therefore composed of different distributed and parallel modules that enable the deliberation and reasoning of the actions and interactions of the robot within the environment. The most successful hybrid deliberate/reactive architecture is the *Sensor Fusion Effects* (SFX) (Figure 2.10). The SFX mimics the neurophysiological arrangement of the human visual system as depicted in Figure 1.1. The SFX's deliberative modules, as described by (Murphy and Mali, 1997), are:

- a *sequencer* that controls a set of behaviours to accomplish a tasks,
- a *resource manager* that allocates computational resources to the behaviours,
- a *cartographer* that maintains and updates the model of the world and the knowledge representation,
- a *mission planner* that bridges the interaction with the environment and a human, and creates the corresponding command to execute the mission plan, and, finally,
- a *performance monitoring and problem solving* module that keeps track of the progress of the robot towards the goal, and allows a degree of consciousness while performing the task.

These types of architectures have been employed to developed cognitive functions. For example, Fay et al. (2005) have developed a robot vision system that integrates visual attention,

object recognition with language and auditory capabilities. Their robot is capable of recognising objects by either indicating the object or verbally commanding the requested object. Likewise, Welke et al. (2010) describe a humanoid robot performing object recognition and manipulation tasks. However, both systems only consider simple shapes and non-complex scenes (i.e. toy environments).

2.7 Visual Object Appearance Learning

Humans are capable of recognising any sort of objects despite imperfect 2D view representations of them. Moreover, the human visual system is robust and efficient regardless challenging environments, different poses, occlusions, and lighting conditions. This remarkable ability is thus enabled by a the interaction with the environment and a learning behaviour that allows creating and building a visual knowledge about objects and structures of the environment.

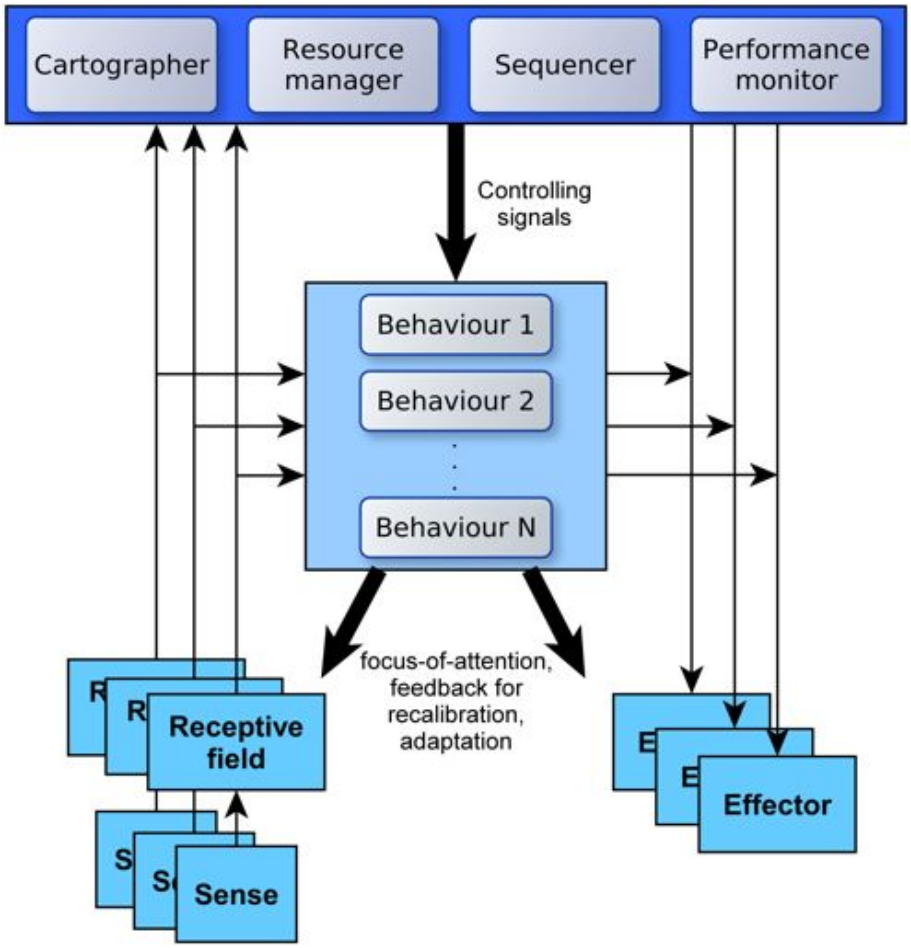


Figure 2.10: SFX architecture. (Murphy, 2001)

As the ultimate goal (as outlined in Section 1.1) is to enable a binocular vision robot head to actively and semi-automatically synthesise its own object knowledge, the robot must be capable of creating and building its own appearance representations by means of the active visual interaction with the object (i.e. an actuated turn-table). To that end, psychophysical findings on how humans represent objects for recognition have inspired the computer vision and robot vision communities to devise learning models capable of capturing the intrinsic properties of objects. The following subsections outlines human studies and related approaches in computer vision and robot vision.

2.7.1 Psychophysics evidence

Psychophysical studies have shown that the human brain represents and learns objects for recognition by snapshots of two-dimensional views of the imaged scene (Bülthoff and Edelman, 1992; Edelman and Weinshall, 1991), that happen to contain the object of interest and the active interaction, i.e, *exploration,* with the objects. Such psychophysical findings argue that the primate's visual cortex represents common everyday objects by a hierarchical structure of their parts or sub-features (Edelman and Weinshall, 1991). For example, attention, while observing an object, is thought to occur at different levels of processing: at a local level objects are imaged for what they are independently of the general configuration; whereas, at a global level, a label is learned or defined to the set of sub-parts to form a concept which recognises the object's identity (i.e. the local and global attentional beam as previously discussed in section 2.4). The recognition of objects is thereby carried out by attending the spatial locations of the pre-attentive localised object.

As determined by image-based theories of object recognition and visual attention, objects are learned by collections of view-specific representations (Fazl et al., 2009). These internal representations are employed to attend objects while they are recognised by a close match previously stored in memory, despite occlusions, fragmentations, and/or deformations (Chun and Wolfe, 2004). Therefore, recognition is a function of imperfect object representations previously observed (Bülthoff and Edelman, 1992; Chun and Wolfe, 2004; Styles, 2005; Fazl et al., 2009) that supports the discrimination between similar object classes and categories.

Specific observation poses have been called *canonical views* in the literature Peters et al. (2002); Palmeri and Gauthier (2004). A canonical view is thus defined (for the purpose of this thesis and according to Blanz et al. (1996) and Peters et al. (2002)) as a "*stable view that provides highly distinctive visual features over the object's viewing sphere and enables a viewpoint invariant representation of an object*".

The selection of canonical view representations has been supported by psychophysical evidence. For instance, **?** allowed participants to freely manipulate 3D household articles over two experimental settings. Firstly, participants were allowed to rotate the object to the view they would normally take photograph. Secondly, they were asked to imagine an article, such that, latterly, they could adjust it in accordance with their imagined view with an interactive tool. Results showed that while exploring/recognising a "known" object, the view-point selection was biased towards a "three-quarters" view of the object, whereas, during the imagination task, participants preferred the most frontal or lateral views of the object. The authors asserted that the selection of the canonical view depended on the task. For instance, recognising an object is biased towards the view where more visual information is observed. But, while the participants were imagining the object, views were selected in terms of how objects were originally learned and how they interacted. Therefore, this suggests that the visual properties play a crucial role on how representations are learned for different tasks.

During acquisition of view-specific information, objects are represented in terms of relevant parts that are the most informative (Palmeri and Gauthier, 2004; Fazl et al., 2009). The foveal regions are thus directed towards those parts in order to observe their visual features with high resolution (Ballard, 1991; Fazl et al., 2009). Hence, attention selection, as described in previous sections, is contained within two quasi-parallel interacting stages: pre-attentive and attentive. Both stages control the information flow that occurs in the WHAT and WHERE streams, and, latterly, in the *object-coding* stream (the *IT cortex* in the human visual system) (Bulthoff et al., 2002; Fazl et al., 2009; Op de Beeck and Baker, 2010).

While learning, the WHERE stream maintains and binds spatial attention on the object part coordinates that characterise the location of the object-part of interest, and creates the spatial relationships (i.e. topological relations) of feature coordinates. Whereas the WHAT stream either determines the identity of an object or triggers a learning behaviour that stores the episodic view-invariant feature descriptions of the object part. Therefore, the object-coding stream is responsible of storing and consolidating visual features that present some form of consistency over a set of viewing poses and saccades (Ullman et al., 2002; Op de Beeck and Baker, 2010). That is, neurophysiological and neuroimaging evidence demonstrates that neurons in the IT cortex are optimally tuned in accordance with observed visual features since they respond more strongly to those trained views than other observed views (Ullman et al., 2002; Op de Beeck and Baker, 2010). Ullman et al. (2002) have shown that individual neurons in the IT cortex are selectively tuned by maximising the visual information with respect to the perceived view-changes.

Within attention selection, Fazl et al. (2009) have argued that attention deployment is driven by an *attentional shroud* that fits a salient object surface, the foveal region is directed towards

object boundaries and, finally, attention selects prototype visual features that characterise a learned object inside the attentional shroud. Thus, attentional shrouds allow exploring and learning corners, intersections and other feature types that are the most informative visual cues in an object.

2.7.2 Related Literature

Despite learning being an active process, some robot vision researchers have failed to consider this assumption. For example, the robot vision system proposed in (Forssen et al., 2008; Meger et al., 2008, 2010) employs static images gathered from the Internet to train a bag-of-words classification scheme in order to allow robust recognition over different view poses. Thus, visual object learning relies on training machine learning algorithms with massive amounts of data, e.g. (Mikolajczyk et al., 2006; Tuytelaars and Mikolajczyk, 2007). (Newman et al., 2009) proposed a Simultaneous Localisation And Mapping (SLAM) mobile robot application based on appearance information and a generative probabilistic model for online learning. The authors used the concept of visual words to build the image vocabulary by acquiring high detection on loop closures in outdoor environments.

Recent research literature regarding active robot vision systems performing visual active learning of an object appearance is scarce. The most notable works are the robot system by (Kootstra, 2010) and Wallraven and Bülthoff (2007b).

On the one hand, (Kootstra, 2010) propose a mobile robot system that actively explore objects over the viewing-sphere. Their approach consists of allowing the robot to move around in a similar trend of having an object in a robotic hand. Local visual features, in this case SIFT features, are classified according to their displacements in order to achieve figure-ground segmentation. At the same time, those features that are stable over different viewpoints are filtered in order to remove features that do not present invariant properties. Active viewpoint selection follows a probabilistic framework where the most informative view is chosen based on the amount of observed information. Features from all informative views around the viewing-sphere are then clustered in order to group visual information that shares similar attributes over different poses. By grouping similar features, a compression ratio of $\sim 37\%$ with respect to simply storing SIFT features at $30\,degrees$ of interval is achieved. The authors also show that using an active learning exploration scheme achieves better recognition results than its passive vision counterpart; however, using a learned feature database represents a trade-off of $\sim 85\%$ of success recognition rate with respect to $\sim 99\%$ of using raw SIFT-features passively observed at fixed intervals.

On the other hand, the work of Wallraven and Bülthoff (2007b) finds its origins in the so-called exemplar-based *key-frame framework* (Wallraven and Bülthoff, 2001) which is proposed as an attempt to model the human visual streams computationally. The initial framework combines recognition and visual object learning appearance in order to exploit and learn local spatio-temporal consistencies of visual features from a fixed-viewpoint image sequence (i.e. passive vision system). This model also provides dynamic information on local texture changes across the object's viewing sphere. The authors also claim that their framework is scalable (knowledge is increased over time) and automatic (it decides whether it should learn or recognise the incoming video stream). An extension of the framework (Wallraven et al., 2003) has relied on support vector classification schemes over the learned views in order to achieve state-of-the-art recognition results on controlled and real image databases. However, the link between perception and action is not entirely established as the authors considered a passive computer vision system and a pre-recorded video stream, i.e. the original key-frame framework does not control its interaction with the environment.

To overcome the above limitations, the authors then extended the key-frame framework to a robotic application (Wallraven and Bülthoff, 2007b) that integrates haptic and proprioceptive information with visual information. Motor control sequences and image acquisition are previously defined giving less importance to the active interaction with the object while learning its appearance; although the subsequent recognition phase is controlled according to the active affordances with the object that shares similarities with the behaviour herein described.

Some other closely related and recent research is found in (Hyundo et al., 2006). The authors present a binocular robot head that semi-automatically learns individual object representations by actively tracking and segmenting while a user shows objects in a cluttered and real environment. However, the authors do not consider stable canonical representations and, in consequence, employs a wide visual feature vocabulary (trained off-line) to obtain acceptable recognition rates. This makes the system intractable to match in a robot visual exploration task because of the large size of the visual knowledge database. Therefore, further extensions (e.g. a scalable and continuous learning of objects) are not considered as they depend on training a supervised classification algorithm each time the robot observes a new object.

Finally, the robot vision system in (Modayil and Kuipers, 2008) relies on a mobile robot application and a planar laser range-finder to construct and learn its own object representations contained in the environment; however, the authors do not consider active sensing and visual information is set aside as future extensions of their model. Their proposed model has the limitation of using a laser range-finder as part of the tracking and perceptual system while vision would only provide further information of the environment where, in this thesis, it is argued that only stereo visual information is sufficient for the visual learning of the appearance

of an object.

2.8 Summary and Discussion

This chapter presented an overview of the current state-of-the-art in robot vision systems. Their characteristics and limitations in terms of perceptual capabilities, are identified as follows.

Robot vision technology have used a combination of several low-level feature extraction techniques; for example, saliency maps, SIFT, SURF and/or colour histograms to name a few. In that respect, this chapter surveyed popular local feature extraction techniques in order to understand the underlying properties of their representation as attentional cues for gaze control, object recognition and so forth. Among current feature extraction techniques, SIFT was found to be the most reliable local feature extraction technique that has been employed in a wide range of computer vision and robotic applications. A brief review was also provided.

Several robot vision systems share the same basic visual principles underlying bottom-up and top-down attentional cueing mechanisms. Firstly, saliency maps for bottom-up, low-level cueing are designed to invoke the application of a more computationally costly investigation of specific visual object locations. Secondly, an object-based, top-down biasing is adopted in order to tune the cueing process according to the needs of a high-level visual task. Both cueing systems thence subserve the deployment of attention of robot systems in order to explore the contents of the perceived environment. This, in turn, allows developing visual search strategies that carries out and executes visual tasks. In overall, these strategies have been modelled by means of the close interaction between the attentive and pre-attentive modes of attention. However, developed visual search strategies in the literature have been essentially "hard wired" and difficult to extend or generalise.

In that regard, this chapter analyses the properties and characteristics of visual attention models reported in the literature. Scientists have agreed that attention is governed by the "perception-action" cycle principle. In the robotic context, most of the reviewed robot vision systems in this chapter devised visual operations as a collection of "ad-hoc" functions. These functions have been "hard wired" by means of the described visual search strategies and constrained for the specific purpose the system was designed. Nevertheless, there are few reported robot vision systems that have cast attention as a structure, parsimonious robotic architecture, these architectures have only considered simple shapes and non-complex scenes. Thus, in order to develop a functional architecture, this chapter examined prominent robot architectures that have been successfully developed for different purposes and tasks than vision.

Object recognition modules developed in robot vision systems are normally able to covertly recognise only one object instance in an image and, moreover, in the environment. This limitation is inherent in the design of the feature extraction technique, commonly the SIFT object recognition pipeline (Section 2.3.1). It is therefore required to extend the perceptual capabilities in robots as common scenes and environments contain multiple instances of the same-class object. As surveyed in Section 2.5, multiple detection has been initially enabled based on manually-tuned thresholds and bag-of-words models over large image databases. However current methods have assumed that object instances are isolated entities in the scenes and their detection is cast in terms of thresholds which must be adjusted according to the current object class being detected.

Similarly, typical robotic visual search strategies have relied on manually pre-trained image databases or bag-of-words models as in (Kragic et al., 2005; Björkman and Eklundh, 2006; Rasolzadeh et al., 2010) or (Meger et al., 2010), respectively. The former is employed to correctly characterise the object's appearance over the viewing sphere and, consequently, to enable a robot to properly operate in the specified application. While the latter uses images retrieved from Internet in order to create a view-point category invariant classification engine. Both approaches thus consists of a large image database in order to achieve high recognition rates which implies that a large memory size is used. Nevertheless, it is found in the literature that the active visual exploration of an object while learning its appearance (ref. Section 2.7) overcomes the above limitations. That is, a robot is enabled to actively interact with an object in order to capture spatio-temporal properties that reliably characterise the object's appearance over the viewing-sphere. However, these approaches have reduced the size of the object database at the cost of average recognition rates. Therefore, the need of a robust and reliable computational model has yet to be devised.

The underlying design of robotic systems, as reviewed in this chapter, comprises an arrangement of more than two cameras for foveated and wide-field vision for recognition and inspection. This configuration implies that the extrinsic camera geometry and a global-metric world representation must be computed and updated for each camera movement invoked and for visual competences and the control system (as in (Björkman and Eklundh, 2005a, 2006; Rasolzadeh et al., 2010)). In consequence, visual operations need to be synchronised geometrically and globally in order to fulfil a high-level task-goal specification. Thus, computational complexity and evaluation times increases as the system progresses in its task.

In order to advance the current state-of-the-art research and address shortfalls of robot vision research, this thesis proposes to firstly investigate and characterise an active binocular robot head which integrates object recognition, binocular vergence and gaze control by means of a single visual feature representation (i.e. SIFT). Since the methods developed on the above

binocular robot vision system serve as the founding robot vision software, a complete study and characterisation of the system's performance is needed in order to identify and acquire the required foundation knowledge to advance this particular active binocular robot head and, consequently, the state-of-the-art of robot vision. These are presented in Chapter 3. Shortfalls are thus identified such as the detection of single instances, "ad-hoc" arrangements of visual functions and the use of manually pre-trained image databases or bag-of-words models for objects' characterisation. It is thus proposed to address the foregoing limitations as follows.

- Chapter 4 describes a novel covert multiple same-class instance detector which avoids the need for manually-tuned thresholds and large databases in order to obtain higher recognition rates over a wide collection of trained objects in different settings and domains.
- Chapter 5 develops a fully integrated hierarchical robot vision architecture in accordance with the cognitive model of the mammal's brain (Figure 1.1) and the design robot architecture principles in Section 2.6. This architecture features visual behaviours that are not constrained to the hierarchy but can be adapted and applied to different contexts.
- And, finally, Chapter 6 proposes a semi-supervised visual learning behaviour that actively acquires visual knowledge in terms of the interaction with the object under different environmental conditions (i.e. different view-points, illumination changes and so on) and creates canonical representations by means of an active clustering and visual knowledge consolidation.

In must be pointed out that the described active robot vision system adopted stereo vision. This assumption afforded the means of adopting a symbolic map representation of the world for object recognition and saliency detection is used (as discussed in Chapters 3 and 5) since the camera geometry does not need to be known. Furthermore, the attention and inhibition of return rely on a motor-ocular egocentric space as described in Chapter 5.

The next chapter will thus describe the founding robot vision software and the complete validation of this system.

Chapter 3

The Active Binocular Robot Head

This chapter[1] *presents the pilot investigation and initial development of the active binocular robot head. This system as devised integrates vergence, object recognition and gaze control in an integrated robotic application capable of performing simple visual search tasks. The underlying idea is to unify the above visual competences with a single feature extraction. For completeness, the results of the initial investigation carried out by* Fattah (2007) *and later reported in* (Fattah et al., 2008) *are presented here. Likewise, design deficiencies of the hardware protocols of the robotic system are identified after analysing the aforementioned initial results. The last sections discuss the rationale and methodology of a complete and extended characterisation of the system's performance in order to gain the required scientific insight to advance the active binocular robot head and the state-of-the-art in the field of robot vision.*

3.1 Introduction

This chapter thus presents the initial software framework of the active binocular head (Fattah, 2007; Fattah et al., 2008; Aragon-Camarasa et al., 2010) which serves as the founding concepts of the forthcoming methods presented in this thesis. In these papers, the authors de-

[1]Parts of this chapter have appeared in the following peer-reviewed papers:

- Fattah, Haitham and Aragon-Camarasa, Gerardo and Siebert, J Paul, "Towards Binocular Active Vision in a Robot Head System", in Ramamoorthy, Subramanian and Hayes, Gillian M., ed., Towards Autonomous Robotic Systems, TAROS 2008 (University of Edinburgh, 2008), pp. 25–32.
- Aragon-Camarasa, Gerardo and Fattah, Haitham and Siebert, J Paul, "Towards a unified visual framework in a binocular active robot vision system", Robotics and Autonomous Systems 58, 3 (2010), pp. 276–286.

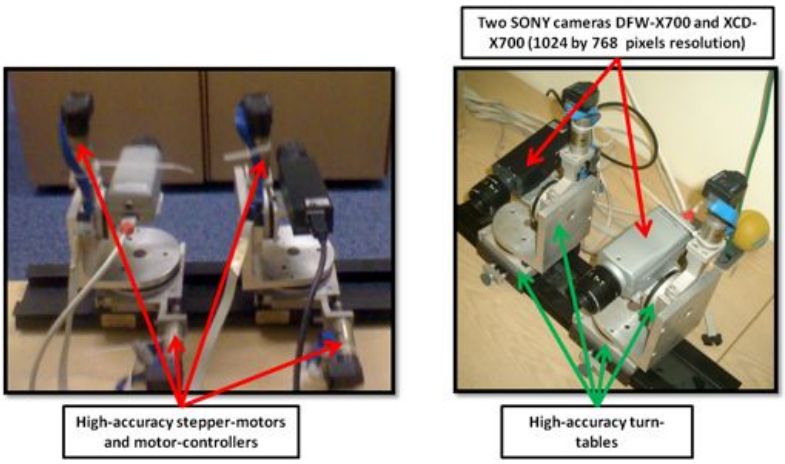


Figure 3.1: Specifications of the active binocular robot head

scribe the development and a pilot study of a system that integrates object recognition (Section 3.3), vergence (Section 3.4) and gaze control (Section 3.5) based on point matches extracted by means of the Scale Invariant Feature Transform Lowe (2004) (SIFT). As demonstrated in Section 3.6, the system as devised provides efficient means of controlling an active binocular robot head (depicted in Figure 3.1) by integrating low-level and high-level visual components in a uncomplicated and unified framework. This chapter also addresses the identification of design deficiencies in the robot's hardware interface and a complete and extended investigation of the initial research of Fattah (2007).

3.2 Motivation and Objectives

Few binocular robot head systems integrate vergence, gaze control and object recognition that explore cluttered visual scenes (ref. Section 2.1.2). Binocular robot vision, as opposed to monocular vision, enables a machine to perceive instantaneously depth of visible surfaces. In the object recognition context, binocular vision allows gathering more information in order to generate stronger object identity hypotheses and project the observed environment into detailed machine representations. Thus, the design of an active binocular robot system that features basic vision functions (e.g. object recognition, visual attention, binocular convergence and so on) not only serves as a founding benchmark platform to assist the understanding of human vision, but has also the potential to be applied over different robotic applications.

As described above, each visual functionality in the active binocular robot head is structured according to a unified and robust visual representation (i.e. SIFT features). The idea of using SIFT in robot vision has already been explored. These include: the Yorick head (Björkman and Eklundh (2005a)) and the robot vision system devised by Kragic et al. (2005) (Section

Figure 3.2: (a) Robot head exploring objects (as appeared in Aragon-Camarasa et al. (2010)) (b) A complex, cluttered scene.

2.1.2). However, the aforementioned systems utilise the SIFT framework as a general purpose object recognition system; vergence and visual attention mechanism rely on different visual representations. In that regard, the described robot system in this chapter only adopts SIFT features as the underlying visual representation for all image processing and control operations. This system demonstrates the application of several novel design principles in a basic functional integrated robot vision system.

A typical scenario that this robot vision technology finds applications, are, for example, *flexible robot work-cell automation* or CIM (Computer-Integrated-Manufacturing) production lines. However, the scene settings of the devised robot vision system herein reported consists of typical office settings and such contains everyday highly textured objects. Figure 3.2 illustrates such office settings employed during validation experiments in this thesis.

3.2.1 Mechanical Specifications

The physical robot head McDougall (2004) employed in this chapter comprised the following (Figure 3.1). Two SONY cameras: DFW-X700 and XCD-X700 (colour and mono respectively at1024×768 pixels resolution); fitted with IEEE Firewire interfaces and four high-accuracy stepper-motors and motor-controllers (Physik Instrumente GmbH & Co.). The hardware was interfaced to a Pentium 4 computer with a CPU clock speed of 2 GHz, with 2 GB in RAM running under Windows XP and MATLAB.

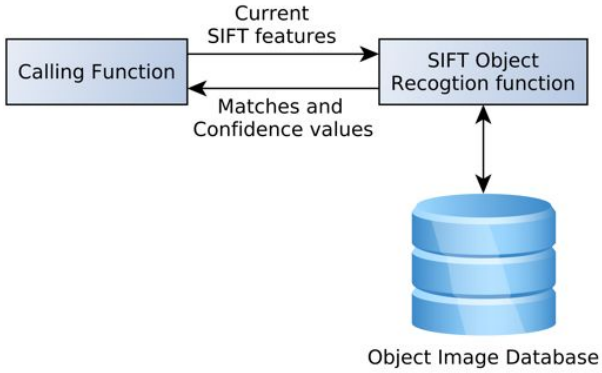


Figure 3.3: The object recognition adaptation within the overall framework. (Fattah et al., 2008; Aragon-Camarasa et al., 2010)

3.3 Object Recognition in the Robot Head

The design of the object recognition system is an adaptation of the exemplar-based model described by Lowe (2004) (Section 2.3.1). In the robot vision context, Björkman and Eklundh (2005a) has initially described the algorithmic steps of using SIFT for their object recognition module in the Yorick Head (Section 2.1.2). This module is composed of a database of *'known' objects* where it is stored a number of image poses of several objects and, consequently, SIFT features are extracted for each image. The object recognition engine consists of comparing extracted SIFT features from captured images of the robot cameras to all SIFT features object samples stored in the database.SIFT features also provide the means of processing captured images into an encoded low-dimensional data representation which are, in turn, passed onto other visual competences. Therefore, the importance of the design resides in the integration of the object recognition in the robot vision software framework. In Section 2.3.1, the key steps of an object recognition pipeline have been described based on SIFT. Figure 3.3 depicts the adaptation of the object recognition engine into the active robot head software framework.

The object recognition module (as reported in (Fattah et al., 2008; Aragon-Camarasa et al., 2010)) is a two-fold process. On the one hand, this system provides the required matching operations of SIFT features while verging the cameras (Section 3.4) and to verify the object's identity in the gaze control system (Section 3.5). On the other hand, it also enables the capability of locating and identifying putative objects such that the robot is capable of actively exploring the contents of the imaged scene (as described in Section 3.5).

3.4 Vergence

In the human visual system, *binocular vergence* refers to the ability of simultaneously adjusting, in opposite directions, the relative angles of both eyes' foveal regions in other to target the same region in both eyes. Thus, the converging angles of both eyes are directly proportional to the measured disparities of the observed region of interest such that both eyes are centred and focused on the same portion of the environment. Vergence is a two-fold visual competence: it minimises imaged dissimilarities between the foveal region of both eyes and maximises the visual information that can be perceived from the environment. Vergence is thus a reflexive behaviour (Section (2.6), page 43) since it is a consequence of directing the attentional spotlight to either salient regions or object targets. That is, the visual stimuli of the newly invoked fixation triggers the behaviour without any conscious deliberation of the observer.

In a binocular vision context, there is a broad range of circumstances and tasks which the verging mechanism must satisfy. That is, both cameras are driven by means of inter-camera disparities and thus they target the same real-world position. Fattah (2007) determines that the design of the vergence component based on point feature matching techniques is regulated according to the following principles:

- All matched point features or only a subset of them are considered such that both cameras target the same real-world position (i.e. this subset is segmented from the global set with respect to depth, camera geometry and/or so forth) .
- Selection of the verging region is determined by the analysis of either the observed environment or the current observed visual information.
- Vergence is driven by top-down cues from an object recognition function or by a reflexive action on the available visual input.

Each modality subserves a specific task in line with the current state of the robot (i.e. task-based defined). As vergence is a reflexive behavioural system, Fattah et al. (2008) suggested to structure the above vergence modalities as a *hierarchy of behaviours* consistent with *Brooks' Subsumption Architecture* (Brooks, 1991) (Section 2.6.2 on page 46). Fattah's behavioural hierarchy was thereby divided in layers of increasingly complexity and act reflexively according to the current objective. These layers were defined as follows (Fattah, 2007):

- 0th layer - global, non-selective vergence
- 1st layer - image-independent selective, or foveated vergence

- 2nd layer - image-dependent selective vergence
- 3rd layer - attended, selective vergence

According to the *Subsumption Architecture*, each layer satisfies objectives based on different rules. The first layers are subsumed by more complex higher layers and are in harmony with the top-level objective, i.e. the main task of the robot. However, in Fattah's behavioural hierarchy, layers are independent in their operation and they are activated according to the available information and commands received form a high-level control system. This type of configuration is therefore in accordance with the second generation of the hierarchical architectures described in Section 2.6.2. The modes of behaviour of the high-level control are discussed in Section 3.5.

Specifically, only two layers of the behavioural hierarchy are considered in the design of active robot head. Originally, Fattah (2007) specified that the *0th* and *3rd layers* were deemed relevant to the design as the *1st* layer and *2nd layer* were somehow contained or are similar in their functional operation. The vergence behavioural hierarchy was thus initially design for the *0th layer* (non-selective vergence) as the *3rd layer* was developed during the design of the gaze control system (Section 3.5) and shared the same algorithmic principles of the non-selective vergence layer.

Thus, the vergence behavioural architecture can be designed under different visual understanding models such as stereo feature matching (Boyling, 2002; Fattah et al., 2008) and area-based matching (Björkman and Eklundh, 2005a; Boyling and Siebert, 2000). Fattah et al. (2008) pointed out that *"feature based matching offers advantages over area based techniques [...] when the surfaces are jagged or "spiked" or the local disparity gradient is near to an occlusion"*. Therefore, SIFT features provide the required visual representation to bring the cameras into convergence.

The ability to verge the cameras is thus based on a global set of SIFT feature matches between the two camera eyes. As SIFT features are invariant to some geometric transformations, it is highly probable that several matches are found. Vergence is implemented as a closed-loop system behaviour based on extracting and matching SIFT features from each of the images of a stereo-pair captured by the robot vision system (Figure 3.4 depicts the overall algorithm of the vergence system). That is, binocular disparities are calculated between matching SIFT keypoints. These disparities are then histogrammed in order to determine the highest peak of the disparity function in terms of depth and the most dense/compact cluster in the visual field. The resulting disparity associated to this highest peak is then used to estimate the actuator movement required to rotate the gaze angle of each camera until the vergence adjustment loop halts when the highest disparity peak has been shifted close to zero disparity (i.e. both

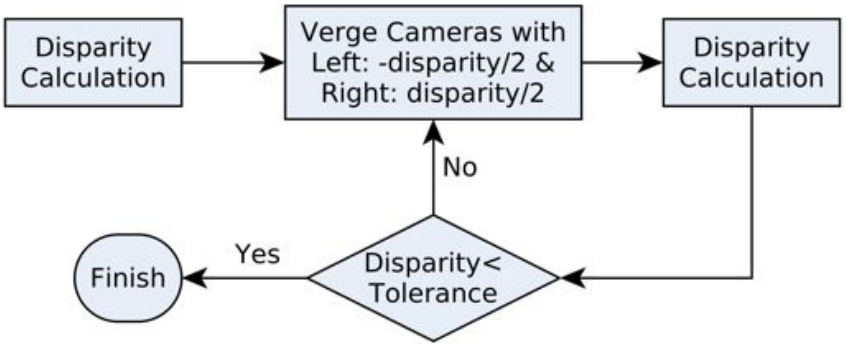


Figure 3.4: Flow chart of the vergence algorithm.

cameras target the same real-world position). Figure 3.5 depicts an example of a completed vergence loop where the anaglyph qualitatively depicts the amount of registration (i.e. vergence) between the stereo pair. In this thesis, an anaglyph encodes a 3-dimensional image in a single colour image by superimposing a pair of pictures separated by a baseline in the red, blue and green channels.

Fattah (2007) (and consequently in Fattah et al. (2008); Aragon-Camarasa et al. (2010)) assumed that the cameras were always in vertical alignment as the cameras were maintained in a relatively fronto-parallel arrangement over all possible pan and tilt camera positions. Therefore, the vertical axis generally registered disparities around zero when situated in their initial positions (as observed in Figure 3.5; bottom row) but, as the cameras rotated away from their home position, the highest peak of the vertical disparity histogram started deviating from zero (this limitation of the system is further discussed in Chapter 5 and 6). In order to reduce false positives of match features from both cameras, Fattah et al. (2008) proposed to constrain the disparity computation by selecting those SIFT features that were consistent in terms of their *scale* and *orientation* components. Such constrains were defined heuristically as:

$$\left|\theta^L - \theta^R\right| \leq 20^\circ \tag{3.1}$$

$$\frac{\sigma^L}{\sigma^R} \leq 0.45 \tag{3.2}$$

where θ is the SIFT canonical orientation value of the corresponding feature matches on the left and right cameras; and σ, the SIFT scale component of both camera images. Both constrains (Equations 3.1 and 3.2) filter feature matches such that the computation of disparity histogram enables the active robot head to robustly direct its cameras onto objects in complex and cluttered scenes. Furthermore, the vergence system is a critical component as the visual input description is obtained and passed over to other visual functions.

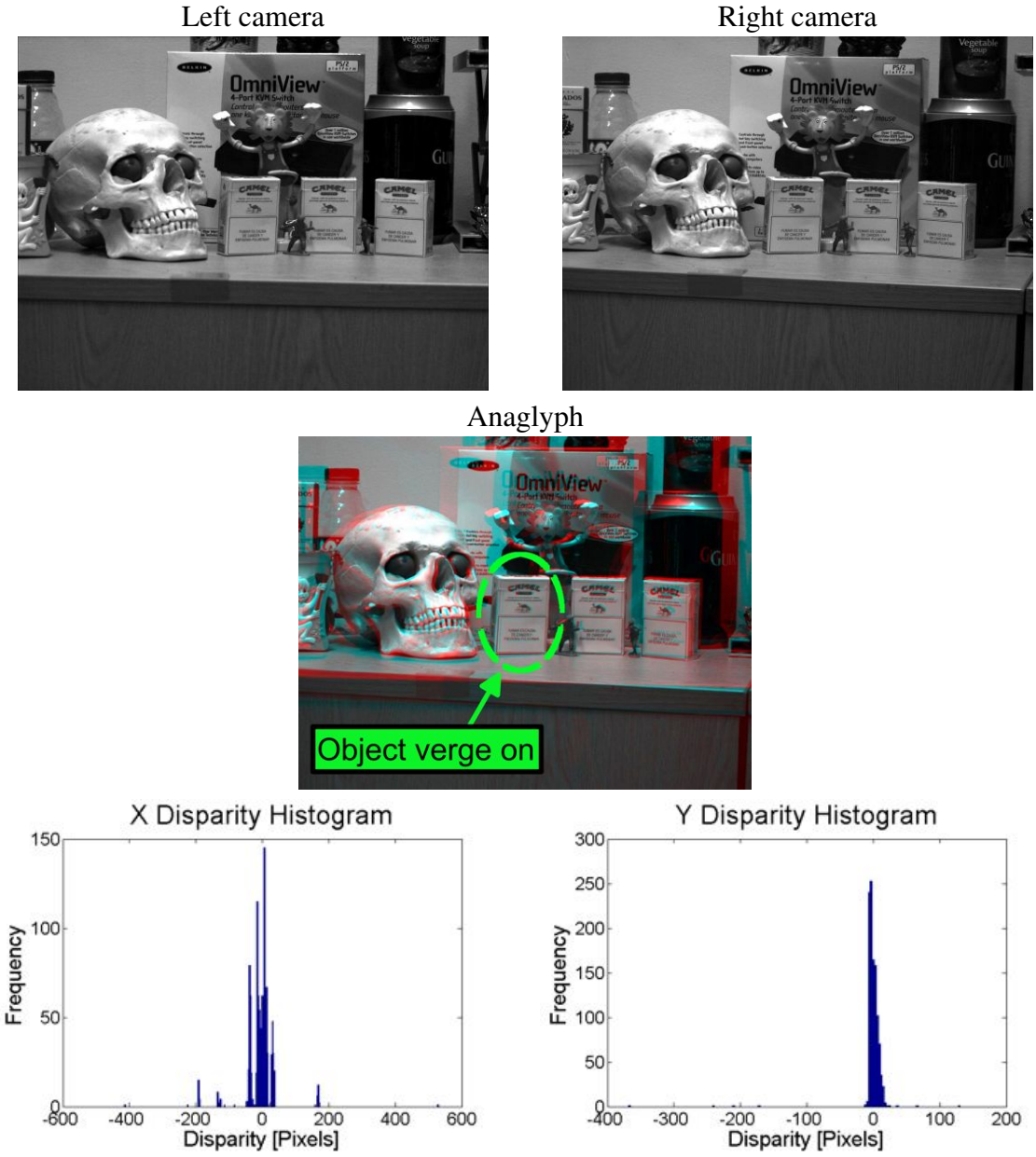


Figure 3.5: Top: captured stereo pair of a cluttered scene. Middle: anaglyph of the stereo pairs. Bottom: Disparity histograms of the x and y axes.

3.5 Gaze Control

The design of the gaze control system is devised as a high-level "*ad-hoc*" set of control functions that governs the execution of the object recognition and vergence systems. Fattah et al. (2008) designed their object recognition and vergence systems based on the assumption of a symbolic visual representation of SIFT features. The gaze control system therefore employed the same visual representation to control the ability to shift the *attentional beam* (Section 2.4) and to use the same visual information on each processing stage for the active investigation of observed scene.

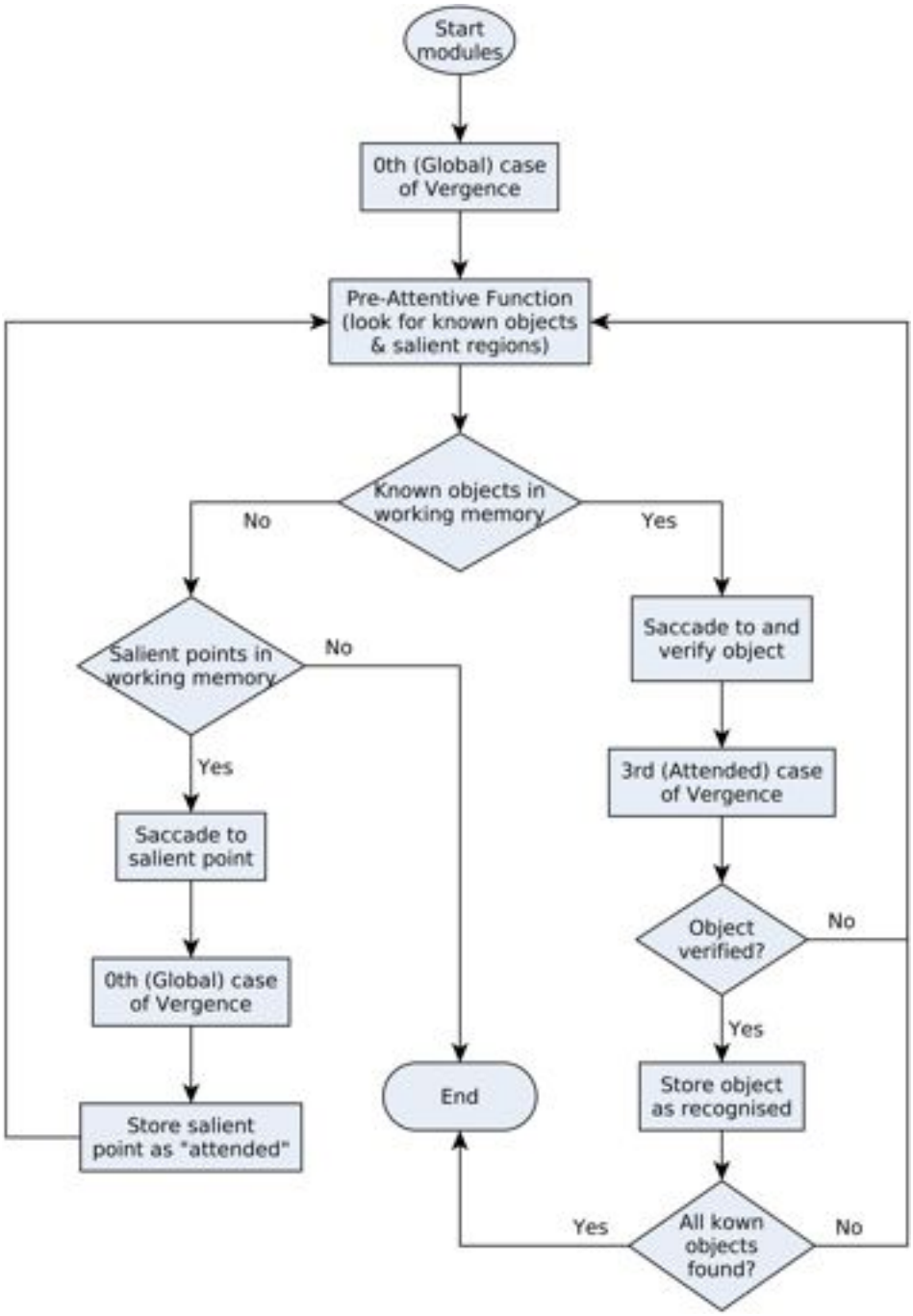


Figure 3.6: Flow chart of the the gaze control system. Fattah et al. (2008)

The design was inspired on the *pre-attentive* and *attentive* behavioural approaches that operate in a quasi-parallel manner: *the output of one feeds the input of the other and vice-versa*. In other words, the system was designed in accordance with the *"perception-action cycle"* principle. A flow chart of the gaze control system configuration can be depicted in Figure 3.6.

The pre-attentive stage was concerned with the detection of objects and salient features that are of interest to the search strategy and current objectives and state of the robot. This function analysed the current field of view before invoking saccadic camera movements. Each

highlighted object and salient feature were then passed to the attentive function and in order to either verify the object's identity, or decide where to look next. Objects and salient features, in that regard, were only localised and detected on the predefined dominant eye (the left camera in Fattah's software design) by following a *"stepping stone"* search pattern. That is, an object is only noticed when it appears in the field of view of the dominant camera eye or it is reached *by "a bridge of other objects or salient features"* (Fattah et al., 2008; Aragon-Camarasa et al., 2010). Hence, the pre-attentive was divided in two processing stages: a) detection and localisation of *"known"* objects and b) detection of salient features as discussed in Sections 3.5.1 and 3.5.2, respectively.

The attentive stage, on the contrary, is responsible for guiding and controlling the attentional beam and making recognition decisions based on the reported locations. That is, every potential and interesting object/salient location detected in the pre-attentive stage is serially examined in detail by targeting the cameras on the found object/salient locations. In Fattah's design, putative objects were attended to first whereas salient features were only visited when there were not any object hypotheses in working memory as depicted in Figure 3.6. The robotic system directed the cameras towards an object/salient feature with the highest confidence/score computed in the pre-attentive stage (as described in Section 3.5.1 and 3.5.2). Hence, the visual search was guided by two different visual processes: *attending to object*s and *attending to salient features*. Both are described on what follows and, as well, illustrated in Figure 3.6.

- While *attending to objects*, the camera-pair was verged before making recognition decisions such that the object of interest was centred in each camera's field of view by means of the 3rd layer of the behavioural vergence hierarchy (Section 3.4). The 3rd vergence layer consisted of firstly directing the dominant camera (left camera) to the approximate location passed from the pre-attentive stage (as described in Section 3.5.1). While verging the cameras, the target object was symbolically segmented by considering only those features that matched among the attended object class in the database, the target object and the SIFT features from the left camera. Those segmented features were then used in the disparity computation (as in the algorithm of Figure 3.4). As both cameras are kept in vertical alignment, the non-dominant camera (right camera) was vertically rotated to the the same y actuator steps of the left camera. Thereafter, the right camera was horizontally rotated in function of the disparity induced until the disparity error was minimised (Figure 3.4).

  By matching between those features that survived the vergence cycle and the database object features which remained from the pre-attentive stage, an object was verified if the ratio of the first and second nearest neighbours were below a predefined threshold value of 0.4 *"AND"* at least 3 features were matched in either left or right cameras. For each

identified object , the corresponding object class in the database was not longer considered in subsequent pre-attentive cycles. This thereby acted as a symbolic inhibition of return mechanism.

- While *attending to salient features*, the cameras were directed towards those unknown but interesting feature points. The cameras were verged according to the $3rd$ layer of vergence (Section 3.4). To avoid endless backtracking of attended salient features, a list was maintained of already observed salient items.

  While pre-attentively analysing the scene, each unknown feature was compared to all attended salient features in the aforementioned list by matching their corresponding vectors an exhibiting a nearest neighbour ratio value above 0.5. This operation therefore inhibited salient locations in a purely symbolic space. This allowed a semi-guided visual search method to explore unrecognised image structures and, therefore, served as a bridge to discover other possible *"known"* objects in the scene that would not be possible to reach if the attentional beam followed recognised objects (Fattah, 2007; Aragon-Camarasa et al., 2010).

3.5.1 Pre-attentive Object Detection and Localisation

One of the functions of the pre-attentive stage was the detection and localisation of objects. This consisted of generating feature correspondences for each input image with respect to all pre-stored object examples contained in a manually trained database. In such database, there were several images of each object class (i.e. different poses) where SIFT features were indexed and labelled in accordance with the object class. Matching SIFT features of the left camera image (dominant eye) with the database features therefore returned all possible object class sets identified.

The Generalised Hough Transform (Section 2.3.1 on page 28) was employed to geometrically evaluate and strengthen the localisation and detection of the feature coordinates of the guessed object classes: Equations 2.1, 2.2 and **??**. Furthermore, the value of the highest peak in the Hough accumulator was used as a confidence of the detection and, subsequently, employed to select the next object saccade in the attentive stage.

The affine pose estimator was thereafter applied to each winning set of feature clusters of the identified object classes (i.e. the highest peak in the Hough accumulator). That is, the projection of any feature points in the database into the scene was defined as:

$$y = Ax \cdot PS_{ratio} \tag{3.3}$$

where A is the resulting affine transformation; x, the centre coordinate of the database image; PS_{ratio}, the number of motor steps per pixel required to translate x into actuator space, and y is the spatial location of the object in the scene in motor steps units. The actuators could then be driven to fixate the cameras using Equation 3.3. This projected point was then used to support the selection of the next saccade in the attentive stage as described in Section 3.5.

Fattah et al. (2008) experimentally determined the motor steps per pixel relationship by tracking a feature location over a serie of camera angular motor positions. It was assumed that the camera principal points was nearly close to the rotational axis of each camera and, consequently, the step-pixel relationship behaved as a linear system. Hence, the relationship was thus computed by applying a linear regression of the sampled data and the final value is defined as the inverse of the slope in the linear regression model.

3.5.2 Salient Feature Detection

Salient items are those features in the dominant camera (left camera) image that do not match to the features in database image and their saliency score are above a threshold. The saliency score (κ) was defined as follows:

$$\kappa = x_{offset} \cdot y_{offset} \cdot \sigma \tag{3.4}$$

Note that x_{offset} and y_{offset} denote the horizontal and vertical distance from the left image centre coordinate, respectively, and σ the corresponding scale of the input features. The saliency score was then filtered by computing their mean and standard deviation. Thus, those features that exhibited a saliency score above three standard deviations were stored in working memory and, therefore, employed to guide attention.

3.6 Pilot Experiments Overview

Fattah et al. (2008) perform initial experiments in order to validate the designed systems while the robot operates within real-world settings. Specifically, these experiments include the validation of the 0th and 3rd layers of the vergence and the gaze control systems (Sections 3.6.1 and 3.6.2).

3.6.1 Vergence

The first set of experiments concerns the verification of the vergence behavioural hierarchy. In these experiments, the 0th layer is only considered (global, non selective) vergence since the correct function of the 3rd layer is verified as part of the gaze control system. Thus, the main aim is to measure the statistical accuracy and reliability of the 0th layer of vergence when presented with a number of different and isolated scenarios. Specifically, Fattah et al. (2008) describe two different experiments as briefed on what follows:

1. **Experiment:** All visual features come from a single depth plane in the field of view.

   (a) **Set-up.** A printed image was mounted on a board and placed at an initial distance from the cameras baseline. The cameras were directed towards the target board. The scene only contained the target image in a planar background without clutter.

   (b) **Methodology.** The vergence algorithm (Figure 3.4) was invoked and executed until the disparity error was below a threshold of 10 pixels. This process was repeated six times over six different depth distances from the cameras.

   (c) **Results.** From Figure 3.7, it is observed that the RMS final verge errors were sufficiently small to enable a foveated robot vision system to target 3D structures in a scene accurately around the high-resolution fovea such that the degree of overlap achieved is sufficient for 3D reconstruction. The RMS errors are the final disparity value where the vergence loop halted with respect to the mean value of all the experiments. The following was obtained:

      i. Algorithm stabilisation in 2 vergence cycles in all experiments.

      ii. Maximum error: ~ 5.3 pixels.

      iii. Overall error: ~ 1.4 pixels.

2. **Experiment:** Visual features appear over two juxtaposed planes at different depths distances.

   (a) **Set-up.** A second printed image was mounted adjacent to the previous first printed image. The cameras and scenes exhibited the same configuration as above.

   (b) **Methodology.** The above methodology was used in these experiments. One image was maintained static while the second was translated over six different depth distances from the camera baseline; that is, two different objects with six different depth distances.

(c) **Results.** From Figure 3.8, it was inferred that there was not a clear correlation between the depth distance and accuracy of the vergence. In average, the vergence cycle of all recorded experiments stabilised after 78.8 seconds. The following was obtained:

i. A maximum of 6 vergence cycles took to stabilise the algorithm.

ii. Maximum error: ~ 6.5 pixels

iii. Overall error: ~ 3.8 pixels.

3.6.2 Gaze Control and Scene Exploration

After verification the vergence behaviour, the second set of experiments consisted in the validation of the overall gaze control system. Fattah et al. (2008) divided the operational modalities of the gaze control in three isolated functions in order to objectively validate the system. These modalities were defined as (according to Fattah et al. (2008); Aragon-Camarasa et al. (2010)):

1. The system should pre-attentively detect "known" objects when they appear on the field of view of the dominant camera and their location is recorded in terms of motor-steps units. If the pre-attentive stage finds several putative objects, the gaze control must select the object with the highest confidence value in order to saccade the cameras to the reported location.

2. The system is capable of actively exploring and identifying all objects in a scene presented to the robot by employing recognised objects and/or salient features in a "stepping-stone" search manner.

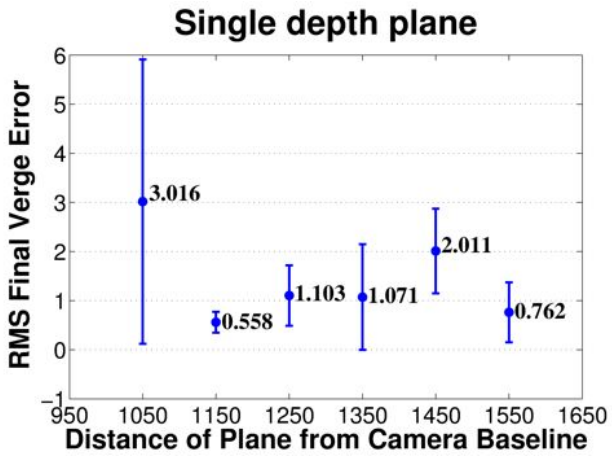


Figure 3.7: Root-Mean-Square (RMS) vergence errors while verging the cameras on a single depth plane.(Fattah et al., 2008; Aragon-Camarasa et al., 2010)

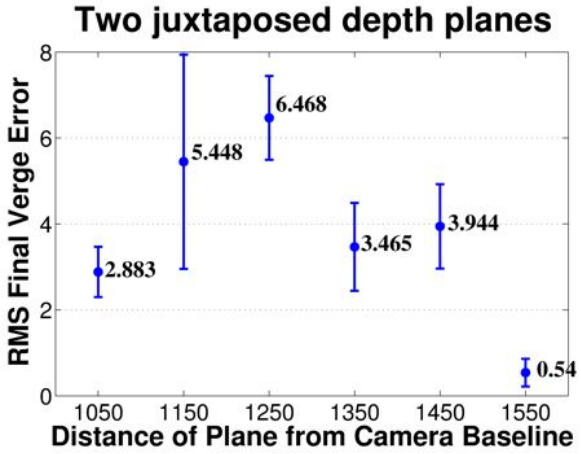


Figure 3.8: RMS vergence errors while verging the cameras on two juxtaposed depth planes.(Fattah et al., 2008; Aragon-Camarasa et al., 2010)

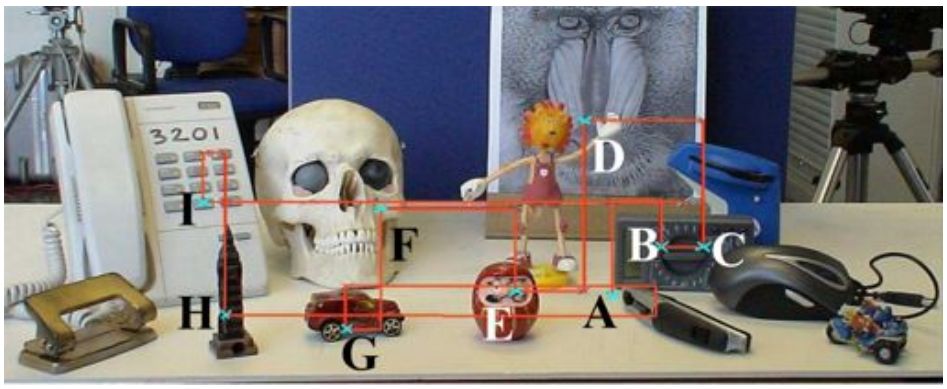
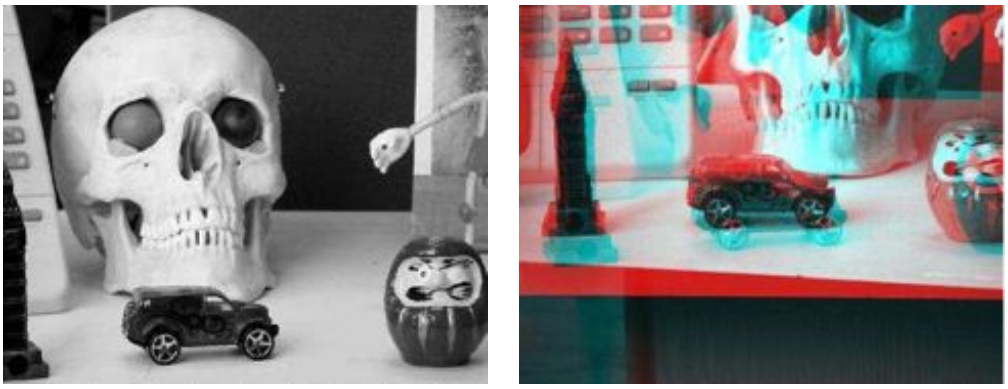


Figure 3.9: Cluttered scene used in the validation of the gaze control system. Overlaid camera traces depict the fixations of the dominant camera over the scene.(Fattah et al., 2008)

3. The system will not attempt to direct the cameras towards an object if this object has already been identified. Likewise, if there are not object hypotheses in working memory, the system will select a previously unseen salient feature as the next saccadic movement.

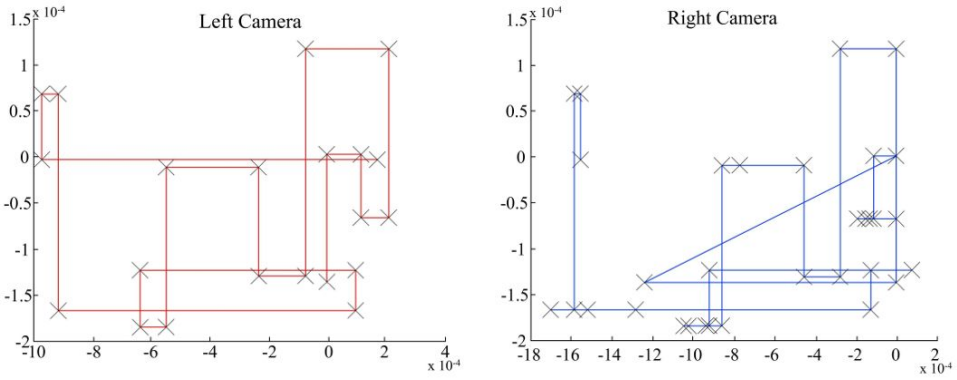
Hence, the correct operation of the above modalities was verified by presenting a cluttered scene to the system that contains nine *"known"* objects, as depicted in Figure 3.9. Fattah et al. (2008) employed a single visual search task to validate the three different functions of the gaze control system. The system was allowed to execute a maximum of 20 saccades. Fattah's experimental results are outlined for each modality in the following sections.

The verification of the first modality consisted of presenting to the system a scene that contained two "known" objects with clutter (being the objects named, *"skull"* and *"car"* in Figures 3.9 and 3.10(a)). Fattah et al. (2008) described that the pre-attentive stage successfully detected and localised both objects; the *"car"* was the one that exhibited a highest confidence score. By visually inspecting Figure 3.10(b), it was observed that the car was first attended



(a) (b)

Figure 3.10: (a) Field of view of the predominant camera prior while pre-attentively looking for "known" objects, (b) Anaglyph of the camera images after saccading to the "car" and prior the 3rd layer of vergence.(Fattah et al., 2008)



(a) (b)

Figure 3.11: Camera traces of: (a) the left (dominant) camera and (b) the right (subordinate) camera.(Fattah et al., 2008)

and, consequently, the saccadic movement was near the centre of the image. The first gaze control functions is therefore validated.

The last two remaining modalities were corroborated by inspecting the overall result of the visual search task. Figure 3.11, (a) and (b), shows the actuator camera trace after completing the search task. The "stepping-stone" search pattern can be depicted in Figure 3.9 where each saccade is denoted with a capital letter. By close examination of Figure 3.9, it can be seen that two objects were not identified/localised correctly. That is, the *"bike toy"* did not have a registered saccade and the saccade denoted with the capital letter "C", which corresponds to the *"mouse"* object, did not correctly centre the object.

3.6.3 Discussion

From the above pilot results, Fattah et al. (2008) initially demonstrated the ability of the binocular robot vision system to actively search and recognise objects in a complex and cluttered scene. The "stepping-stone" search pattern was also demonstrated to be a basic but robust visual search strategy that did not require knowledge of the camera geometry. Specifically, by couching the visual representation in terms of SIFT features, it enabled to combine binocular vergence, object recognition and gaze control into an operational and integrated robot vision software that fulfilled the requirements and objectives of the initial investigation on the active binocular robot head (Section 3.2).

The design of the behavioural hierarchy of the binocular vergence is robust in different operational contexts as the single worst vergence error obtained, ~ 6.5 pixels, was within practical limits for stereo reconstruction and depth recovery (Section 3.6.1). Although Fattah's original vergence algorithm was extended and modified to suit the requirements of the robotic architecture as described in the following chapters of this thesis, the computation of the disparity histogram and the scale and orientation constrains (Equations 3.1 and 3.2) remained in their original form throughout this thesis.

Fattah et al. (2008) argued that his SIFT-based *"attentional spotlight"* implementation allowed the robot vision system to explore the contents of the scene in a structured and conceptually uncomplicated manner. The system was able to direct and centre its gaze towards objects with a recognition rate of $\sim 77\%$. Therefore, the novel integration of the SIFT object recognition system as the founding principle for visual processing operations subserved the ability to carry out visual search tasks in fairly complex and cluttered settings.

However, the experiments in Fattah et al. (2008) showed a capable system with some degree of visual maturity that had not been reported at the start of this project, an extended experimental evaluation was required to characterise the performance of the system. This allows to devise novel visual competences that fulfil the requirements and the desired goals in Chapter 1 and in the design of robot vision systems. In that regard, it is initially observed from this pilot experiment that the operation of the robotic system presents slow execution times while searching a given scene. This is mainly due to the deficient design of the software driver connectivity with the hardware. In the forthcoming sections, the extended validation and shortfalls are addressed.

3.7 Improved Hardware Interfaces

McDougall (2004) integrated the current hardware components of the robot head and devised a general purpose robot vision benchmarking tool. He developed the actuator control and image acquisition modules that were latterly used in Fattah's MSc project (Fattah (2007)). By examining Fattah's results and his robot vision software, it was identified that the actuator control that drives the robot head cameras and the image capturing modules were inefficient. That is, the system took nearly 4 minutes to verge the cameras in the 0th (i.e. Global, non-selective) and 3rd (i.e. Attended, selective) layers of vergence. Similarly, the execution time of a simple visual search task necessitated two hours to search over a scene as in Figure 3.9.

Hence, both modules have been improved in order to reduce execution time from one hour to 20 minutes for any of the defined visual tasks in previous sections. Additionally, during a familiarisation period with Fattah's software, some programming errors were corrected which affected some of his initial results. These include the object recognition system and the saccade selection of the system.

It is considered, however, that these enhancements do not contribute to the founding research questions and objectives (ref. Sections 1.2.1 and 1.1) of the active binocular robot vision head and they are therefore included in Appendix A.

3.8 Extended Experiments

To fully verify the correct function, recognition performance and robustness of the system, it is required to test the system under different operational settings. The following sections report the validation results obtained.

3.8.1 Materials and Methodology

This experimental design adopts the operational gaze control functions and hypotheses described in Fattah et al. (2008) and Aragon-Camarasa et al. (2010) which, in turn, have been already defined in Section 3.6.2. Additionally, these experiments include the described enhancements in Section 3.7.

As the vergence system has been validated for the 0th layer of vergence, only a case-study is presented in Section 3.8.2. The 3rd layer, however, is included and validated as part of the gaze control system. The objective of these experiments is to obtain a full characterisation

of the system's performance. Thus, the materials used in these experiments are described as follows:

- Five different scenes are prepared with a combination of ten known and unknown objects in random positions. Objects objects have been selected arbitrarily from everyday objects of different texture compositions and geometrical shapes.
- Each scene contains between five and six known objects while the rest is clutter.
- These known objects classes are thus stored in a database as described in Section 3.3, i.e. the system is trained with hand-tuned images and a range of different viewing poses. Figure 3.12 shows the six object models used in these experiments.

The experimental methodology consists of generating five random initial fixations for each presented scene. The experiments therefore constitute 25 visual search tasks which, in turn, produce 145 object observations. Opposed to Fattah's pilot experiments where the visual search task was allowed to perform a maximum of 20 saccades, these experiments consider a maximum of 45 saccades per each visual search task invoked. These constrains were determined heuristically in order to capture the behaviour of the system and characterised false positive detections while exploring the scene. Furthermore, if there are not new known objects to be attended (i.e. each object class is inhibited after being attended and recognised as described in Section 3.5), the visual search task is interrupted and reports all the objects found in the scene.

Bear

Car

Cigarette Box

Iceman

Orange Juice

Skull

Figure 3.12: (a) The six object models used in the experiments.

While actively exploring a scene, there are three possible outcomes in the detection and recognition of objects:

- *False positives* are those errors produced when the system is not able to centre the object in the field of view of both cameras during the attentive cycle.
- *Not found* comprises the system's failures of not noticing pre-attentively an object in the visual search task.
- *False hypotheses* are those incorrect detections of objects that cannot be verified while attending to them. That is, the object class hypothesis found during the pre-attentive cycle does not correspond to the object class while attending to the object hypotheses.

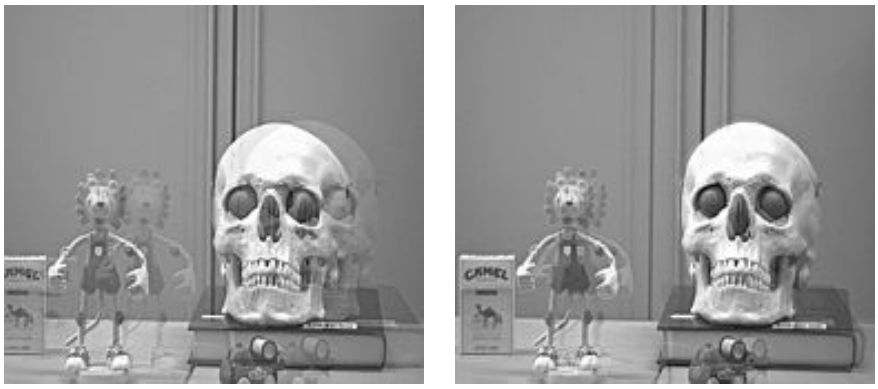
These possible outcomes are summarised in a table and in a cumulative frequency graph of identified objects for each saccade.

Conversely, for each fixated object in the attentive cycle, the motor coordinate-space location is recorded in order to quantitatively measure the system's performance. Specifically, the possible outcomes consist of the ability of the system to centre the hypothesised object in the field of view and to fixate at the same camera-space location in each of different initial random fixations per scene. A plot of the Root-Mean-Square (RMS) fixation errors of the projected camera coordinate-space of the x and y pixel coordinates, respectively is employed to depict the system's accuracy and repeatability.

Finally, to verify the ability of the system to use unrecognised elements to guide the exploration (i.e. the third operational function of the gaze control system), a scene, which contains two known objects widely separated from each other, is created. This scene contains background clutter and unknown objects. This experiment consists of allowing the system to exploit salient structures in the scene such that both objects are attentively recognised, i.e. by means of a bridge that interconnects them. The scene is thus arranged such that one known object is not present in the field of view when the system is targeting a known object.

3.8.2 Vergence

In a real world scenario, the 0th layer of vergence is verified qualitatively by observing Figure 3.13. In this case, the 3rd layer of vergence can only be verified qualitative as there is not a clear verging point in a real-world scenario and, in consequence, ground truth measurements. For example, Figure 3.13(a) shows an anaglyph of the camera views after verge. The *"Lion toy"* object consists of fewer SIFT features than the *"Skull"* object and, in consequence, the



(a) (b)

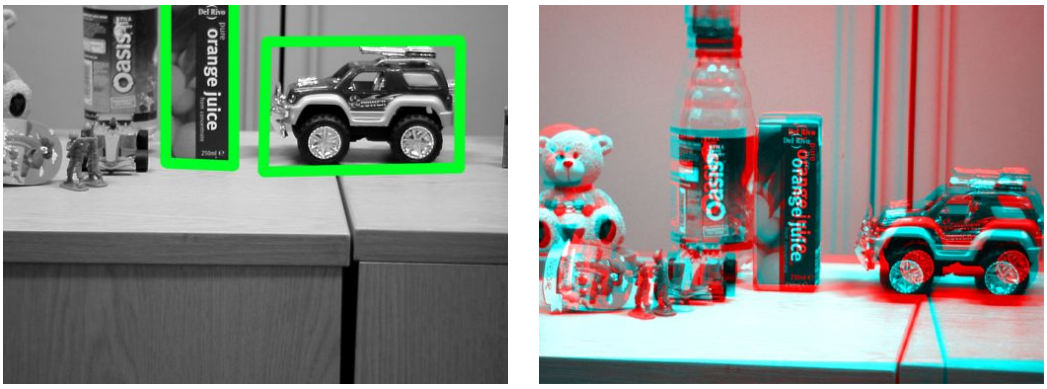
Figure 3.13: (a) Anaglyph of both cameras before verging; (b) Anaglyph of both cameras verging on the "Skull" (Aragon-Camarasa et al., 2010).

cameras converges on the most structured object (in terms of SIFT features) which, in this particular example, is the "Skull" object. In both cameras, the correct alignment of the "Skull" object is depicted in Figure 3.13(b). Notably, regions of the "Skull" where SIFT features are densely distributed, both cameras are more aligned than regions that are further away. That is, these regions depict less overlap in Figure 3.13(b). It is therefore argued that the degree of overlap of the "Skull" achieved is sufficient for 3D reconstruction purposes. The overall execution time while verging the cameras is ~ 74 seconds.

3.8.3 Gaze Control

The first gaze control modality is validated by inspecting Figure 3.14. In this figure, the system pre-attentively analyses the current field-of-view for interesting putative objects and salient features. Each localised object is denoted in Figure 3.14(a) by the rectangular shape. Following the saccade selection criterion in Section 3.5.1, the highest confidence value designates the next object in working memory to be attentively verified. That is, the "Orange Juice" object exhibits the highest confidence value in Figure 3.14(a) and, consequently, the system saccades to the reported location as depicted in Figure 3.14(b). The ability of the system to centre the spatial location and correctly identify the object in the scene validates the first gaze control function and, therefore, the 3rd layer of vergence.

The second gaze control function is a generalisation of the above as each recorded object location (i.e. fixation) is stored and examined in order to fully validate the performance of the system in a common visual search task. To that end, the five scenes created in these experiments are depicted in Figure 3.15 (each scene corresponds to a single initial random fixation trial). In these figures, each fixation can be object targets/identifications (represented with a black circle), salient items (depicted with a black filled square) and, finally, initial



(a) (b)

Figure 3.14: (a) Pre-attentive detection of known objects (rectangles depict the found objects); (b) Anaglyph of both cameras verging on the *"Orange juice"* (i.e. the object with the highest confidence score).

and final fixation points (denoted as black upward and white downward pointing triangles). The "stepping-stone" search pattern thus emerges by overlaying the camera traces on the five stitched scenes of the dominant camera. In Figure 3.15, the selection of the next attended object is determined from previous fixations while either targeting an object or attending a salient item. For instance, in the initial fixations of Figures 3.15(b)(c)(d) and (e), no putative objects are discovered in the robot's field of view; therefore, the gaze control system selects, as the next saccade, a salient feature. It can be observed that the detection of object hypotheses in the pre-attentive cycle occurs after the second or third saccades, that is, the next object attended is not necessarily selected from the current fixation but from previous pre-attentive cycles on each attended locations (as depicted in Figure 3.15(a) where objects are identified after fixating salient features).

As the visual search task actively explores the scene, new visual evidence is gathered. This allows to reject outliers (incorrect feature matches) that do not contribute to the Hough voting scheme, and, in consequence, in the affine estimation processes of the pre-attentive detection of object hypotheses (ref. Section 3.3).

Table 3.1 summarises the system's detection and recognition performance for all visual search tasks. The most difficult objects to locate are the "Bear," "Iceman," and "Skull" objects. False localisations are due to the sudden change of the object's 3D structure. That is, the SIFT algorithm does not produce enough invariant descriptions in order to characterise the object. It is also noted that for the "Bear" and "Skull" objects, their constituent micro-structures share similar texture details over the object. These hypotheses are further supported given the fact that the "Iceman" and the "Skull" are also in the *Not Found* object column. Therefore, the 2D appearance of the trained object poses does not contain stable feature locations and descriptions that confound the SIFT matching process. It is also possible that SIFT features

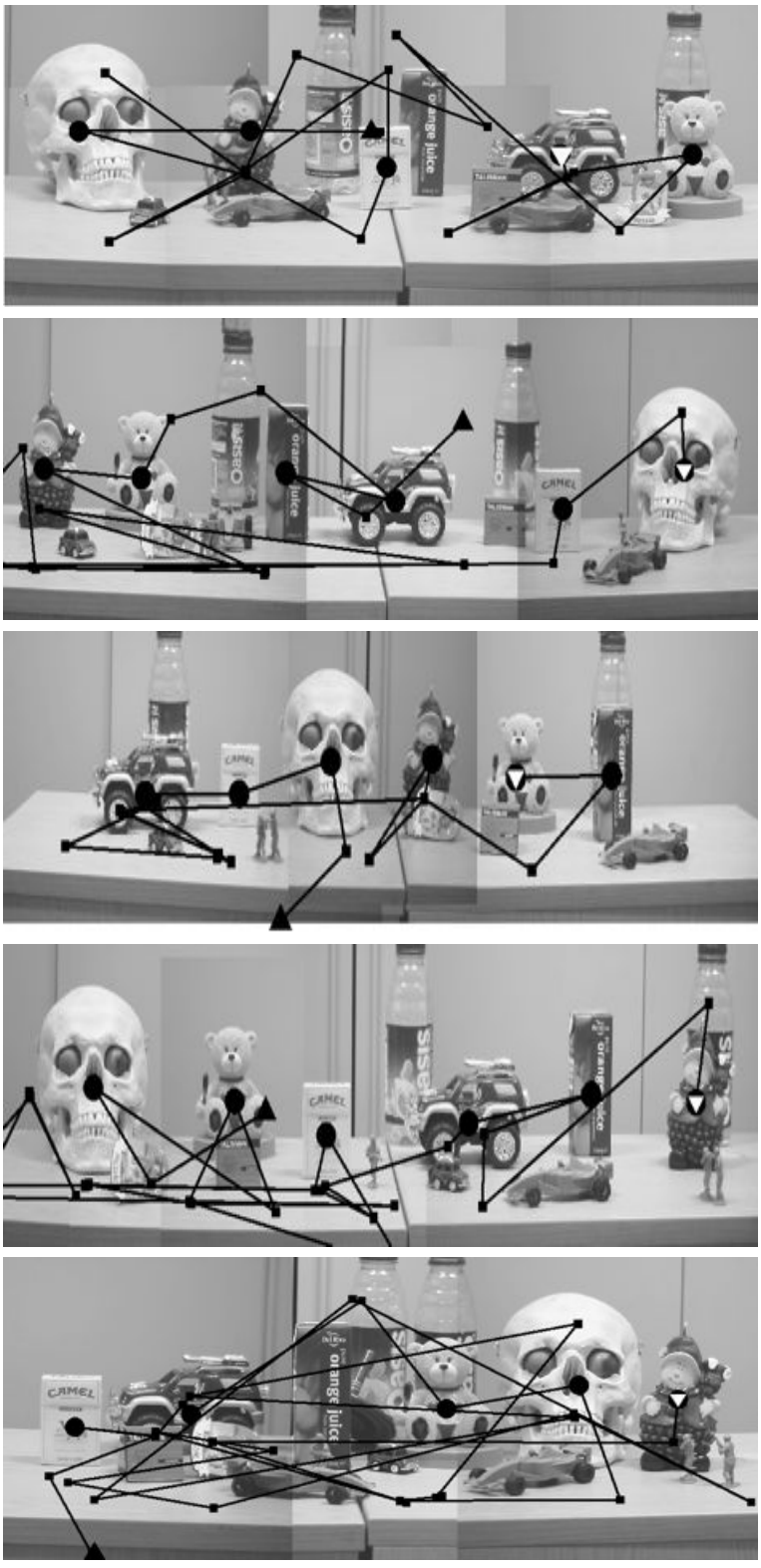


Figure 3.15: (a)(b)(c)(d)(e) The five different scenes with approximated overlaid camera traces of the dominant camera (in pixels).(Aragon-Camarasa et al., 2010)

match with clutter and unknown objects. Specifically, the object "Car" scores more *Not Found* failures mainly because its feature descriptions are classified as outliers and, in consequence,

Table 3.1: Outcomes failures for all visual search tasks (Aragon-Camarasa et al., 2010)

Object	False Localisation	Not Found	False Hypothesis	Total no. of failures
Bear	3	0	1	4
Car	0	3	0	3
Cigarette Box	1	1	1	3
Iceman	3	2	0	5
Orange Juice	1	0	0	1
Skull	3	2	0	5
Total	11	8	2	21

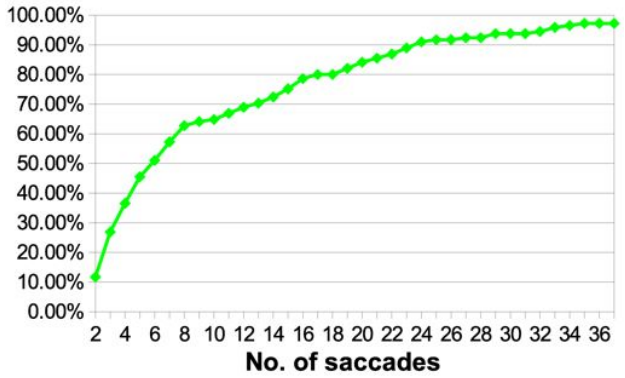


Figure 3.16: Cumulative frequency of identified objects.(Aragon-Camarasa et al., 2010)

these are not geometrically consistent with respect to the trained object's reference centre.

The 45 saccade halt condition is activated eight times as the system does not notice the eight objects listed in the *Not Found* failures column. Finally, the system rejects only two object hypotheses while verifying their identities due to the structure of the scene of the object's pose in it (Table 3.1: False Hypothesis column). This thereby denotes consistent behaviour while performing a visual search task. The overall identification performance while exploring all 25 presented scene is 85.5% over a total of 145 objects in all visual search trials.

The average execution time incurred in the exploration of the presented scenes is ~ 32 minutes with an average of 37 saccades per trial in order to recognise all the requested "known" objects. Notably, the system is capable of starting to recognise objects after performing the second saccade. This is further demonstrated in Figure 3.16 where it is depicted the identification rate per the number of saccades performed. Notably, the system's ability to target and identify correctly 80% of objects is within the first 17 saccades. These results therefore validate the second and third gaze functions.

To additionally support the listed gaze control function, the system's performance is measured to correctly fixate and centre the object of interest in the field of view. Figures 3.17(a) and (b) depict the RMS fixation errors of the projected camera-motor space into pixels of the x

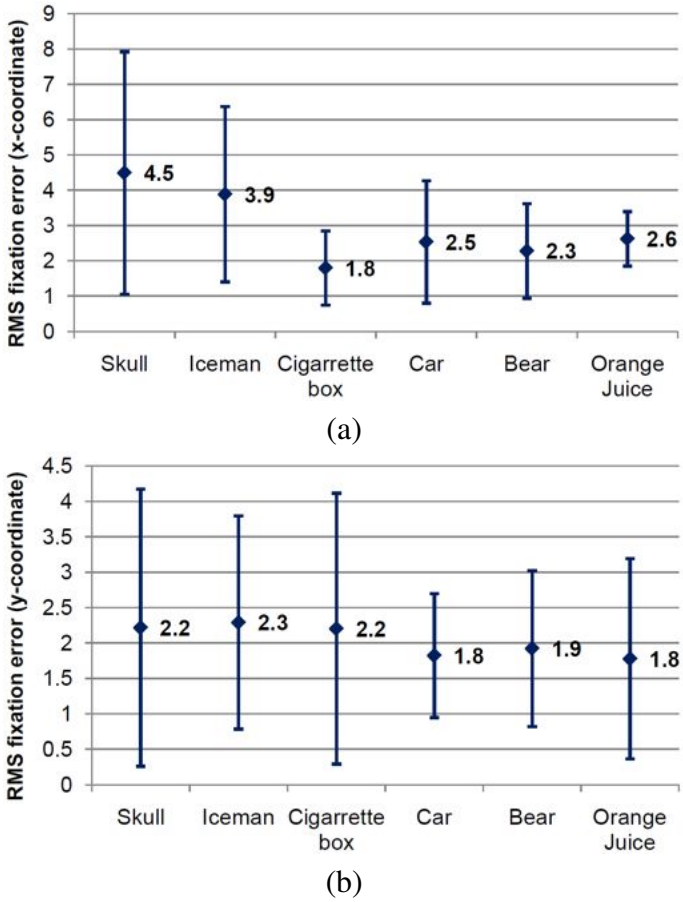
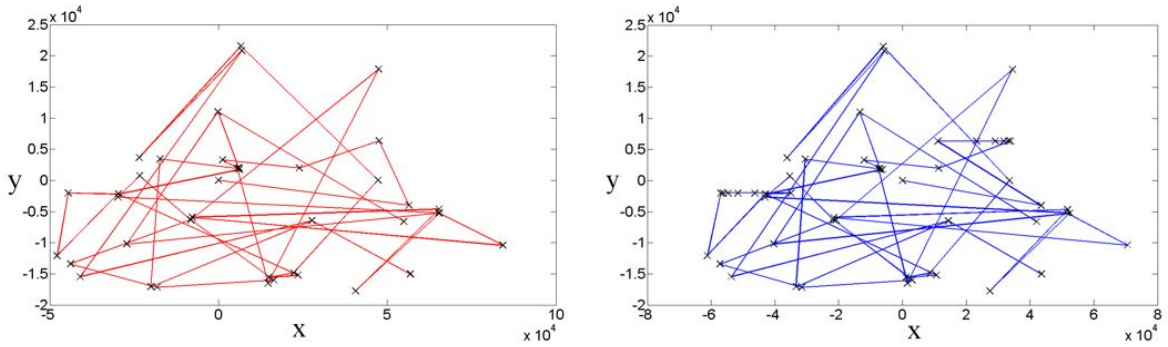


Figure 3.17: The (a) x- and (b) y- axes fixation errors for each of the six objects over all visual search trials reported. The RMS error is in pixels, and the error bars are at 1 standard deviation.(Aragon-Camarasa et al., 2010)

and y coordinates, respectively. The RMS error is measured in terms of the deviation between the centre coordinate of the ground truth (manually determined) and the attended fixation. In that regard, the "Skull" object produces the highest deviation error of 4.5 in the x axis and 2.2 pixels in the y axis. It is therefore argued that these errors do not represent a spatially significant deviation measures that impact the system's performance considerably. Hence, the reported fixation errors are within viable limits for robotic vision applications. Specifically, for the "Skull" object, its fixation deviations correspond to a viewing error of only $\sim 1\%$ of resolution occupied in the image (e.g. 500 by 200 pixels). The fixation errors on the y axis exhibit more consistent measures with an average error of 2 pixels. This is due to the fact that both cameras are aligned vertically in all visual search trials; therefore, vergence errors do not impact the overall fixation accuracy.

Figure 3.18: Defined scene to verify the "stepping-stone" visual search strategy and the "known" object: the "Skull" and the "Lion toy" objects (both bounded by black boxes). (Aragon-Camarasa et al., 2010)



(a) (b)

Figure 3.19: Resulting camera traces of (a) the left and (b) right cameras in the "stepping-stone" search strategy experiment.(Aragon-Camarasa et al., 2010)

3.8.4 Stepping-Stone Search Strategy

In order to characterise the ability of the system to explore a scene in a "stepping-stone" manner, a scene, that contains two widely separated known objects, is created. For the purpose of this experiments and in accordance with Section 3.8.1, a separation of $30\,cm$ between known objects is employed as depicted in Figure 3.18. A visual search task is therefore invoked until both objects are correctly identified.

In this particular experiment, the system performs 40 saccades to target and find both objects. Figure 3.19(a) and (b) show the left and right camera motion traces. It is observed that the "stepping-stone" visual search strategy adopted allows the gaze control system to attend both unknown and salient scene portions in order to identify both known objects. 3.19(a) and (b) also suggest that the gaze control system exhibits a consistent behaviour across different scene arrangements and requested tasks. Therefore, these results support the visual search results presented in Figures 3.15(a)-(e) discussed in the previous section and the third gaze control function.

3.9 Conclusions

This chapter firstly describes the research reported in (Fattah, 2007) where a novel active binocular robot vision system is devised. This system is capable of autonomously exploring a presented scene by finding and localising known objects while maintaining in convergence the binocular arrangement. Several novel principles have been demonstrated to fulfil the initial objectives defined in Section 3.2 and in (Fattah, 2007). That is, the adoption of SIFT features as the unique visual representation enables the system to employ a single visual mechanism in order to operate all defined visual competences, i.e. vergence, object recognition and gaze control.

The pilot experiments presented verify the operational function in each of the described visual competences. It is demonstrated that the gaze control mechanism, specifically the *"attentional spotlight"* metaphor, in conjunction with the "stepping-stone" visual search strategy provide the ability to localise and identify automatically known objects in different and challenging scenarios.

By carrying out a deep analysis of the initial results described in Section 3.6, it is identified that the hardware interface protocols of Fattah's implementation are incorrectly developed and, likewise, the characterisation of the system's performance is deemed incomplete. Particularly, the enhancement of the robot's hardware interface does not only permits a smooth operation of the overall function of system but also the execution time is reduced by almost half of the time (e.g. scenes in Figure 3.15).

The extended characterisation of the system's performance exhibits a classification rate of 85% and a single worst fixation error of 8 pixels from optimal. Likewise, the system required between 2 and 37 saccades to successfully find all known objects presented in 25 different scenes.

It is also noticeable that, from the results of the extended experimentation, the gaze control system tends to direct the cameras towards an object that resembles the trained examples in the database during the pre-attentive localisation. That is, objects in the periphery that exhibit in-plane rotations from their original trained examples produce a lower confidence detection score than those towards the centre of the scene. Consequently, they are attended and identified after several saccade cycles. This behaviour in conjunction with the classification rate obtained (i.e. 85% of correct identifications) suggests that a learning mechanism must be devised in order to produce object representations that are invariant to the object pose. Furthermore, Chapter 5 verifies this assumption but it is not until Chapter 6 when it is addressed.

Similarly, the detection of salient features as formulated in Equation 3.4, expresses saliency

as the rectangular area from the SIFT point to the image centre coordinate weighted by the corresponding SIFT scale component. This equation however does not enable the robotic system to attend salient locations over the periphery for the current saccade and causes that the adopted "stepping-stone" search strategy exhibits a random behaviour (as depicted in Figure 3.19, while validating the stepping-stone search strategy). Hence, saliency detection is reformulated in Chapter 5 in order to enable the search strategy to direct attention towards the periphery of each saccade if putative objects are not in working memory.

It is therefore concluded that the outcomes of this extended investigation are positively comparable to even current state-of-the-art systems at the time of writing this thesis (e.g. Rasolzadeh et al. (2010); Meger et al. (2010)). Nevertheless, there exists shortfalls in the robot vision system herein presented and, accordingly, in the current literature. That is, these systems are, in general, capable of recognising only one object class per visual search task. In a natural environment, however, several instances of the same object class might be present in the environment. In order to address this visual deficiency, the next chapter presents a method, which allows the active binocular robot head to accurately localise and detect multiple same-class object instances under significant occlusion, including self-occlusion.

Chapter 4

Detecting Multiple Same-Object Class Instances

Following the experimental validation in the previous chapter, it is deduced that the active binocular robot head and current robot vision technology is constrained to recognise only one instance per object class per visual search invoked. To that end, this chapter[1] *therefore addresses this limitation by devising an algorithm to covertly detect and localise multiple instances of the same object class. This algorithm implements a pre-attentive detection of objects prior to performing camera movements. The crucial objective of this chapter is to improve object detection and localisation tasks in active robot vision by providing an algorithm to pre-attentively find multiple instances of the same object class. The design rationale consists of clustering transformed SIFT features into a non-quantised, continuous Hough space. The proposed Hough space representation allows to group multiple peaks of such projected SIFT features. This algorithm is capable of simultaneously detecting up to* 6 *same object class instances while perceiving up to* 66% *of occlusion between same-class object instances. Validation experiments are conducted over* ~ 2900 *synthetic composited and real-world images in order to measure its performance and robustness under different scene settings.*

[1] *This chapter is based on the following peer-reviewed paper:*

- Aragon-Camarasa, Gerardo and Siebert, J Paul, "Unsupervised clustering in Hough space for recognition of multiple instances of the same object in a cluttered scene", Pattern Recognition Letters 31, 11 (2010), pp. 1274–1284.

4.1 Motivation

The active binocular robot vision system of the previous chapter exploits the *pre-attentive* and *attentive* modes of attention based on the standard Lowe's SIFT object recognition pipeline (Lowe, 2004) integrated with a high level visual search strategy. As a result, this robot vision system is capable to automatically search and recognise single-class objects within cluttered and different scene settings. However, the pipeline described in Section 2.3.1 implicitly limits the detection of just one instance for each object model class in any given digital image.

Reported robot vision systems included this limitation and as such were not able to disambiguate observed same-object classes. Specifically, current systems (Meger et al., 2008; Forssen et al., 2008; Meger et al., 2010; Rasolzadeh et al., 2010) and the robot system in the previous chapter established that an object was unique in any defined arrangement of objects presented in a scene. This limitation emerges since the multidimensional representation of the Hough Transform space is coarsely quantised to produce a single peak of all geometrically consistent SIFT features matches between observed and trained images. Hence, only those features that resemble the trained object the most, results in the highest peak in Hough space.

However, this limitation needs to be further addressed: (Aloimonos et al., 1988) states that the localisation of multiple objects is an ill-posed problem in a passive vision setting (ref. Section 2.1.1). The idea of using the pre-attentive mode of vision is supported by the ability of the human visual system to covertly attend multiple targets in parallel across the entire visual field without performing any saccadic movements (as discussed in Section 2.4). Each detected target can be serially attended and, therefore, the problem becomes well-posed (the ability to attentively verify multiple same-class objects is further discussed in Chapter 5).

Current literature related to the problem of simultaneously detecting multiple objects has already been reviewed in Section 2.5. A more general and robust approach is nevertheless required in order to overcome the limitations discussed in Section 2.5. Consequently, the devised algorithm must obtain higher recognition rates with different trained objects in different settings and extend visual capabilities in robot vision systems. It must be remarked that the robot vision in this thesis is based on SIFT features, and, consequently, the object detection and recognition paradigm in this chapter consists of iconic appearance modelling. Hence, in order to extend the robot's visual capabilities to specifically localise and detect multiple same-object class instances within a cluttered scene under occlusion and self-occlusion, it is required to avoid the coarse quantisation of the Hough Transform space in Lowe's object recognition pipeline.

This solution thereby consists of dividing a finely quantised or even *continuous Hough space*

(i.e. the space is not quantised) into groups and, consequently, detecting and localising pre-attentively multiple peaks in such space corresponding to potential object hypotheses. As above, each group is geometrically validated by means of an affine transformation relating both input and trained matched feature sets.

Overall, the classification of the contents of an unknown input image involves extracting SIFT features from the input image and then comparing them to the SIFT features of the trained image set for each object class. Each matched feature is projected into the continuous Hough space. Each stored object class has its own Hough space representation. The determination of the number of same object-class instances presented in the input image consists of applying an unsupervised clustering technique to each object class found. This obviates the need of manually tuned thresholds while remaining invariant to different scene settings. The proposed solution therefore depends on the reliability of the SIFT feature extraction as same object-class instances share same feature descriptions and, in consequence, the robustness of the matching process and the Hough space mapping.

For the purposes of this chapter, a *continuous Hough space* is defined as the point to point mapping of the observed image and the database image into an essentially analogous (i.e. continuous) space that is not discretised. That is, this modified version of the Hough Transform does not follow the multidimensional voting scheme (Equation **??**), but it stores each mapping in Hough space as a list data structure. These points are therefore grouped by means of an unsupervised clustering technique in order to identify multiple object instances.

The remainder of this chapter is organised as follows: Section 4.2 details the algorithmic steps of the proposed approach. The design rationale and methodology of the projection of SIFT features into the continuous Hough space and the determination when only one instance is presented, are described in Sections 4.3 and 4.4. The unsupervised clustering of the Hough space and the algorithm stabilisation are discussed in Sections 4.5 and 4.6. The last Sections (4.7 and 4.8) present the validation experiments by testing the algorithm over state-of-the-art images databases (specifically, Nene et al. (1996); Geusebroek et al. (2005); Burghouts and Geusebroek (2009a)) and real-world images captured from the active binocular robot head.

4.2 Algorithm Overview

As discussed, the objective is to find natural SIFT feature groups from the resultant matches between an input and a model image and project them into continuous Hough Space in order to find instances of the same-class object. Thus, the SIFT feature matching process is carried out exactly as in Section 2.3.1 on page 28. Hence, the proposed algorithm is composed of the

following six steps (as well depicted in Figure 4.1):

1. All surviving SIFT feature matches are projected into continuous Hough space (Section 4.3).
2. Determine in image space if there is a single or multiple clusters of the same object class in the input image, as described in Section 4.4. If one single instance is detected, the algorithm goes to step 5 (as depicted in Figure 4.1).
3. In case multiple instances of the same object class are detected, an iterative unsupervised clustering technique is employed in order to find multiple distinct peaks in Hough space (Section 4.5). The algorithm is initialised with $k =$ two initial cluster seeds, and $K =$ user defined maximum number of clusters.
4. The average Silhouette numbers are computed for each k iteration in order to determine how well the grouped features are classified with respect to a specific k (as discussed in Section 4.5).
5. The localisation and detection is rectified by removing outliers and evaluating each cluster with a set of conditional rules (as explained in Section 4.6). If the conditional rules do not hold true, then the outliers are affecting the grouping result, and, in consequence, the process must be evaluated again after filtering out such outliers. Therefore, the algorithm returns to step 3. Otherwise, the clustering process iterates until the maximum number of clusters, K, is reached.
6. Each potential group hypotheses is evaluated at the end of the iterative clustering process by a defined match quality criterion in conjunction with the computed average Silhouette numbers (Section 4.6). The result of this final evaluation denotes the solution, and, hence, the total number of instances hypotheses object class analysed.

4.3 Continuous Hough Space

The Generalised Hough Transform (GHT), specifically in Lowe's object recognition pipeline, has already been discussed in Section 2.3.1. In this context, the GHT provides a four-dimensional space (i.e. translation, scale, and rotation degrees of freedom) that relates the transformation and, in consequence, determines the geometric configuration between the model and input SIFT features.

In this chapter, it is therefore proposed to project all surviving SIFT feature matches between model and input images into a continuous Hough space representation. Several groups are

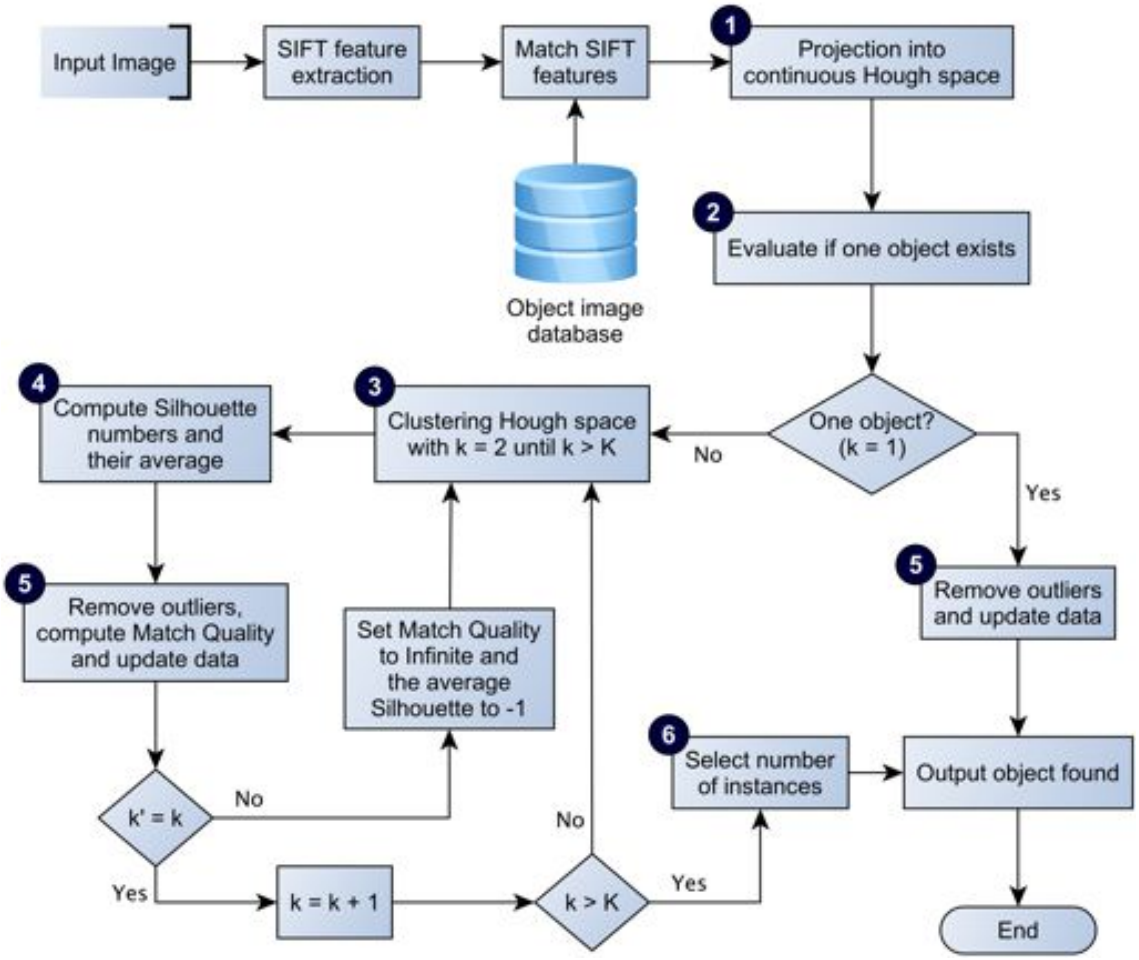


Figure 4.1: Flow diagram of the multiple same-class object instance detector.

created, each associated with one of the object instances present in the input image. The accumulation of geometrical evidence in the four-dimensional histogram is not performed but, instead, grouped by means of an unsupervised clustering algorithm (as described in Section 4.5). The projection into the continuous Hough space exactly follows the derivation presented in Section 2.3.1; however, for completeness, its deduction in this context is given below. The matching ordered set between the SIFT features of the model and test images are defined as, respectively:

$$\mathbf{M} = [m_{ij}]_{i=1,j=1}^{a,4} \tag{4.1}$$

$$\mathbf{T} = [t_{ij}]_{i=1,j=1}^{a,4} \tag{4.2}$$

where $m_{ij} = (x, y, \sigma, \theta)_{ij}^{M}$, $i = 1...a$, $j = 1, ..., 4$, and $t_{ij} = (x, y, \sigma, \theta)_{ij}^{T}$, $i = 1, \ldots, a$, $j = 1, \ldots, 4$, such that x, y and σ, and θ represent location and scale and orientation for each entry $1, 2, ..., N$ in both feature sets; and M and T denote the model and test images,

respectively. Likewise, a and b are the cardinalities of the SIFT features in set $\mathbf{T}$ and $\mathbf{M}$, respectively. Therefore, the projection into the Hough space consists of finding the reference locations of each associated feature in the model, Υ^m, and test, Υ^t, image spaces:

$$\Upsilon^m = x_i^m + \sigma_i^m \cdot \left[\begin{array}{cc} \cos\theta_i^m & \sin\theta_i^m \end{array} \right] \tag{4.3}$$

$$\Upsilon^t = x_i^t + \sigma_i^t \cdot \left[\begin{array}{cc} \cos\theta_i^t & \sin\theta_i^t \end{array} \right] \tag{4.4}$$

where Υ^m and Υ^t represent a tuple of x and y locations, and $i = 1, \ldots, n$ corresponds to the logical index that associates SIFT features matches between $\mathbf{T}$ and $\mathbf{M}$. These features can then be projected into a continuous Hough space by solving the linear equation (4.5) for each corresponding match i, where:

$$b = P^{-1}c \tag{4.5}$$

$$P = \left[\begin{array}{cccc} x_i^t & -y_i^t & 1 & 0 \\ y_i^t & x_i^t & 0 & 1 \\ \Upsilon_x^t & -\Upsilon_y^t & 1 & 0 \\ \Upsilon_y^t & \Upsilon_x^t & 0 & 1 \end{array} \right] \tag{4.6}$$

$$c = \left[\begin{array}{cccc} x_i^m & y_i^m & \Upsilon_x^m & \Upsilon_y^m \end{array} \right]^T \tag{4.7}$$

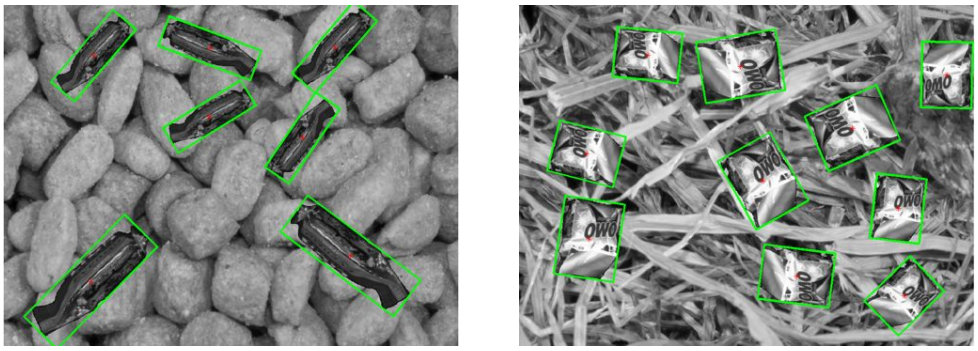
Thus, the least square solution of Equation 4.5 yields the (x, y) location, scale, and rotation tuple denoted as:

$$\mathbf{H} = [h_{ij}]_{i=1,j=1}^{n,4} \tag{4.8}$$

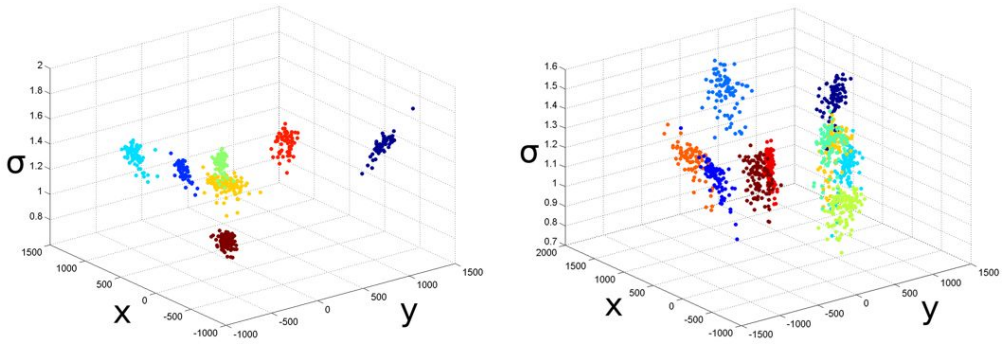
where $h_{ij} = (x, y, \sigma, \theta)_{ij}^{Hough}$, $i = 1, ..., n$, $j = 1, ..., 4$. Hence, $\mathbf{H}$ denotes the final projection into continuous Hough space. Figure 4.2 exemplifies the continuous Hough space mapping of two input test images where objects are positioned relatively close[2] to each other (Figure 4.2: top).

At the end of this step (step 1 in Figure 4.1), three $4D$ arrays are stored in correspondence to the position, scale, and orientation tuples, (x, y, σ, θ) for the sets of: the model, $\mathbf{M}$, and test, $\mathbf{T}$, images; and the projected points into continuous Hough space, $\mathbf{H}$.

[2]"close", as defined in Aragon-Camarasa and Siebert (2010), is *the separation between the objects being less than the smallest spatial dimension in pixels of the smaller object in the image*.



Input images



Continuous Hough space

Figure 4.2: Top: Identified instances (denoted as bounding boxes) in the input images; Bottom: 3D continuous Hough space representation.

4.4 Single-Object Threshold Test

In a non-trivial real-world case, a scene might be composed of either a single or several instances of the same object class. In that regard, Zickler and Efros (2007) approached this problem by defining a distance threshold within a bottom-up agglomerative clustering technique. However, this approach included the detection of single and/or multiple instances in one simple step, this threshold definition is a highly sensitive parameter that must be manually adjusted for different object classes and scenes. Therefore, this chapter proposes to decouple the distance threshold parameter from the grouping process and determine in advance if a single group instance is present in the input image.

By assuming, as defined by Kaufman and Rousseeuw (2005), that clustering algorithms partition data containing more than one cluster, i.e. $k \geq 2$, it is required to determine if only a single object instance is present in the test image. Therefore, the devised solution consists of evaluating the Mahalanobis distances, in order to measure the multidimensional variability in image space between SIFT feature matches sets $\mathbf{M}$ and $\mathbf{T}$. The existence of a single instance is therefore expressed as an ellipsoidal confidence interval test around the statistical mean of the SIFT features of model and test images, i.e. $\overline{\mathbf{M}} = [m_{ij} - \mu(\mathbf{M})]_{i=1,\,j=1}^{n,4}$

and $\overline{\mathbf{T}} = [t_{ij} - \mu(\mathbf{T})]_{i=1,\,j=1}^{n,4}$, respectively. The Mahalanobis distances, D_M^2 and $D_M^2(\mathbf{T})$, are therefore evaluated for each set as:

$$D_M^2(\mathbf{M}) = \overline{\mathbf{M}} C_m^{-1} \overline{\mathbf{M}}^{\mathrm{T}} \tag{4.9}$$

$$D_M^2(\mathbf{T}) = \overline{\mathbf{T}} C_m^{-1} \overline{\mathbf{T}}^{\mathrm{T}} \tag{4.10}$$

where C^{-1} is the inverse of the covariance matrix between the SIFT feature sets (either $\mathbf{M}$ or $\mathbf{T}$).

It must be noted that, at this point, data might contain outliers. Thereby, the Grubbs' statistical test (Grubbs, 1969) is a simple yet effective univariate, iterative outlier detection technique and, therefore, it is performed to only remove outliers in $D_M^2(\mathbf{T})$ since it is assumed that the SIFT features of the model image do not contain outliers. That is, training images have been previously segmented such that SIFT features are specifically extracted from the trained object. The Grubbs' test statistic is defined (the two-sided version) as the largest absolute deviation from the sample mean in units of the standard deviation. This technique is employed to detect outliers in univariate normal distributions. Since the Mahalanobis distances, as an outlier removal technique, converge to a univariate normal distribution (Manly, 2004), the Grubbs' test is formulated as follows;

$$G > \frac{(N-1)}{\sqrt{N}} \left(\sqrt{\frac{t^2_{(a/2N,N-2)}}{N-2+t^2_{(a/2N,N-2)}}} \right) \tag{4.11}$$

where N indicates the degrees of freedom, $t^2_{(a/2N,N-2)}$ denotes the critical value of the *Student's t-distribution*, $a/2N$, the significance level and G is the Grubbs' statistic in which outliers exist for either a minimum,

$$G = \frac{(\mu\left(D_M^2(\mathbf{T})\right) - \min\left(D_M^2(\mathbf{T})\right))}{\upsilon} \tag{4.12}$$

or a maximum value,

$$G = \frac{(\mu\left(D_M^2(\mathbf{T})\right) - \max\left(D_M^2(\mathbf{T})\right))}{\upsilon} \tag{4.13}$$

where υ denotes the variance of the sample.

Hence, the single object threshold test is formulated in terms of the median value, $\mu_{1/2}$, and the standard deviation, std, of $D_M^2(\mathbf{M})$ and $D_M^2(\mathbf{T})$ as follows:

$$\mu_{1/2}\left(D_M^2(\mathbf{T})\right) \leq \left[\mu_{1/2}\left(D_M^2(\mathbf{M})\right) + \alpha \cdot \mathrm{std}\left(D_M^2(\mathbf{M})\right)\right] \cdot s_f \tag{4.14}$$

where α is a user defined *overlap* threshold that defines the perception threshold of multiple same-class object instances and, s_f, a scale factor. This scale factor is introduced as the intrinsic scales between the imaged objects in model and test images (and, in consequence, in sets $\mathbf{M}$ and $\mathbf{T}$) respectively. Thus, the scale factor is defined as the ratio of the mean values of the input scale in set $\mathbf{T}$ and the model scale in set $\mathbf{M}$, such that;

$$s_f = \frac{\mu\left([t_{i3}]_{i=1}^n\right)}{\mu\left([m_{i3}]_{i=1}^n\right)} \tag{4.15}$$

4.5 Clustering Process

As the ultimate goal of the proposed algorithm is to automatically detect multiple distinct peaks of SIFT features projected into continuous Hough space, it is proposed to employ a grouping (i.e. clustering) mechanism. Several clustering algorithms have been devised, as reviewed in Section 5.5.2. They are usually designed for a particular application context and specific input data with defined characteristics and shapes. A technique that meets the requirements of all possible non-trivial cases is yet to be devised.

Kaufman and Rousseeuw (2005) state that the best clustering algorithm must be chosen according to their specific algorithmic properties and performance characteristics. Accordingly, it is decided to apply a clustering technique of each of the two categories described in Section 5.5.2; specifically, the *hard partitioning* and *hierarchical* clustering algorithms.

Among the hard partitioning techniques, the most robust and data shape independent is the *fuzzy C-means* approach. This algorithm consists of assigning probabilities for cluster membership such that it does not require SIFT feature groups to adopt any particular shape. On the contrary, the *hierarchical complete clustering* approach is also chosen since peaks, or projected SIFT features, in the continuous Hough space has a similar shape form as the examples shown in Figure 4.2 and, likewise, defined in (Kaufman and Rousseeuw, 2005). In both cases, the standardised Euclidean distances are selected as the default *objective* and *pairwise metric* functions for the *fuzzy C-means* and *hierarchical complete* algorithms, respectively (Equation 2.10).

Clustering algorithms therefore partition the data in accordance with the defined number of seeds. In order to design an unsupervised process, the clustering algorithm has to iteratively

Algorithm 4.1 Algorithm prototype for the unsupervised clustering approach.

Inputs: $\mathbf{X}$, data matrix of size $\mathrm{n} \times \mathrm{m}$ where n denotes samples and, m, variables; K: maximum number of cluster seeds.

Outputs: $\mathcal{G}$, cluster assignments in each iteration of size $\mathrm{n} \times \mathrm{K}$; kOut, found optimal number of clusters

1: FOR k = 2 to K
2: $\mathcal{G}(\mathrm{k}) \leftarrow$Clustering Algorithm($\mathbf{X}$, k)
3: $S \leftarrow$Silhouette numbers($\mathcal{G}(\mathrm{k})$, $\mathbf{X}$)
4: $\bar{\mathrm{S}}_{\mathrm{k}} \leftarrow$Statistical mean of S
5: END FOR
6: kOut $\leftarrow$find the index where there is a maximum value in $\bar{\mathrm{S}}$

evaluate a given number of cluster seeds in the input data. That is, each iteration consists of initialising a cluster seed k to 2 and then partition the data (i.e. projected SIFT features in $\mathbf{H}$) until a maximum user defined number of seeds, K, is reached (Algorithm 4.1). As part of the evaluation of the optimal number of cluster seeds (i.e. same-class object instances), the quality of the grouping process is carried out by means of their average *Silhouette numbers* (Kaufman and Rousseeuw, 2005) (step 4 on Figure 4.1).

By definition, Silhouette numbers are employed to evaluate and to obtain the number of potential groups in unsupervised clustering techniques by determining which clustered points lay well within their clusters centres, and which do not. A Silhouette number is bounded within the limits $[1, -1]$ such that a Silhouette value near to 1 indicates that the data point is clustered to the appropriate group whilst a value close to -1 determines how dissimilar the data point is with respect to the allocated cluster. The algorithmic steps of a general unsupervised clustering technique in this context are described in Algorithm 4.1. The average of such numbers at each iteration step, k, therefore indicates the overall quality of the point membership to each found cluster.

Silhouette numbers are thus computed as follows (Equation 4.16):

$$S_i = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{4.16}$$

where a_i is the average similarity distance of each i data point with respect to all other points in the cluster, and b_i is the average dissimilarity distance of neighbouring clusters where the i data point is not a member. The *average Silhouette number*, $\bar{S}_k$ (where $k = 2, 3, ..., K$), of the entire measures set indicates how tightly grouped is the point data in all k clusters found. Therefore, the maximum number in $\bar{S}$ determines the optimal result of the clustering process which, within the scope of this chapter, denotes the putative number of objects found in the

input image.

At this point, the set $\mathbf{H}$ contains possible outliers that might not belong to the found cluster and, in consequence, the clustering process is not able to correctly classify them. Such outliers can thereby result in different clustering outputs at each iteration. Hence, a mechanism that deals with outlier data within the unsupervised algorithm is required. The following section describes this mechanism and, likewise, the algorithmic steps needed to improve the detection of multiple instances and, therefore, the final solution of the approach (steps 5 and 6 of the flowchart in Figure 4.1).

4.6 Improving Detection and Localisation

As discussed in the previous section, outliers have an effect on the performance of clustering algorithms and, as a consequence, impact on the final result. In order to refine and deduce the final result, it is required to reevaluate the clustering process without the existence of outliers. (It must be pointed out that the outlier removal technique proposed in Section 4.4 is only applied to $D_M^2(\mathbf{T})$). This process is carried out by means of an affine pose estimator, as described in (Lowe, 2004) and Section 2.3.1, in conjunction with a match quality evaluation criterion and simple conditional constrains (Steps 5 and 6, as shown in Figure 4.1).

Specifically, the affine pose estimator determines the localisation of the object instance(s) and acts as a mechanism to remove outliers. On the one hand, object localisation is carried out by projecting the model's image frame where all SIFT features in set $\mathbf{M}$ are contained. The model image frame (i.e. bounding box, $\mathbf{B}$ which is a 4×2 matrix) is defined by hand during a training phase (as described in Section 2.3.1 and 3.3) and is consequently projected into the input test image to specify the localisation of each of the object instances found (as depicted in Figures 4.2, 4.3 and 4.5). On the other hand, RANSAC[3] and reweighted least squares operate as the underlying core of the affine pose estimator which, in turn,serves as a mechanism to remove outliers. This is carried out by projecting the coordinates of the matched SIFT features in set $\mathbf{M}$ into the input test image in the form: $w' = \mathbf{T}w$ (where $\mathbf{T}$ corresponds to the found affine transformation in the $2D$ Euclidean space; $w_{ij} = [m_{ij}|1]_{i=1,j=1}^{n,2}$ and w', the transformed coordinates) and computing the Euclidean distances, $D_E(\mathbf{M}', \mathbf{T})$, between each corresponding coordinates of the projected set $\mathbf{M}'$ and $\mathbf{T}$ sets. An outlier is detected if it exceeds 2.5 standard deviations from the statistical median of $D_E(\mathbf{M}', \mathbf{T})$ (Equation 4.17).

$$D_E(\mathbf{M}', \mathbf{T}) < \mu_{1/2}\left(D_E(\mathbf{M}', \mathbf{T})\right) + 2.5 \cdot \mathrm{std}\left(D_E(\mathbf{M}', \mathbf{T})\right) \tag{4.17}$$

[3]RANdom SAmple Consensus is a non-deterministic and iterative process that estimates the parameters of a mathematical model from a data set which may contain potential outliers.

Likewise, the recently computed Euclidean distances of each data point are employed to compute the Mean Square Error (MSE) (in pixels). The MSE resulting value is therefore established as the *match quality* criterion in order to evaluate the back projection error of the matched coordinates in $\mathbf{M}$ and $\mathbf{T}$ sets.

These mechanisms are evaluated for each cluster found of the current seed in the unsupervised clustering cycle. It is assumed that a cluster symbolises an object instance; therefore, if a cluster contains less than 3 SIFT feature points, this cluster is removed and MQ and $\overline{S}$ are set to infinity and -1, respectively. Under this assumption, the current cluster seed, k, is decremented by one (denoted as a temporary variable k'), and the clustering process is re-evaluated with the current cluster seed, k, until $k = k'$; i.e. no more outliers groups are detected (as depicted in Figure 4.1). When $k' = 0$ (i.e. all groups are treated as outliers), MQ and $\overline{S}$ are set to infinity and -1, respectively, and the described approach is stopped for the current cluster seed cycle. Consequently, the algorithm resumes with the next iteration, $k+1$, until K is reached. Therefore, the final algorithm including the unsupervised clustering process is depicted in Algorithm 4.2.

The final step (step 6 in Section 4.2 and in Figure 4.1) and, thus, the final result consists of inspecting the average Silhouette numbers and the statistical mean of the match quality values for each evaluated cluster seed. That is, this process searches the value of the optimal cluster seed (i.e. final number of object instances) when $\overline{S}$ is a maximum and the mean of the MQ values is a minimum, as summarised in Algorithm 4.3. If any of these conditional rules fail (e.g. the case when the Algorithm 4.2 finds that all cluster seeds are outlier groups, $k' = 0$), the proposed approach terminates and assumes that one object instance is present in the input image and, consequently, proceeds to step 5 (as depicted in Figure 4.1).

4.7 Pilot Experiments

A database of 620 synthetically composite images, which comprises varying numbers of objects segmented from real images and set against a plain background, is generated in order to evaluate the overall performance of the proposed algorithm. Composite images provide known ground truth in terms of the numbers instances present for different objects and also their pose, scale and degree of occlusion. Seven objects are taken from the *Columbia University Image Library* (Nene et al., 1996) and images of three more objects are captured by means of the binocular robot head described in Chapter 3. Each training image is segmented to isolate each object (in total, 5 image poses for each object class with ± 5 degrees of offset of in-plane rotation).

Algorithm 4.2 Pseudo-code for steps 3, 4 and 5 in Figure 4.1.

Inputs: $\mathbf{H}$, projected SIFT features into continuous Hough space of size $\mathrm{n} \times 4$; $\mathbf{M}$ and $\mathbf{T}$ model and input matched SIFT features of size $\mathrm{n} \times 4$ ($\mathbf{H}$, $\mathbf{M}$ and $\mathbf{T}$ are ordered sets of equal size); K, maximum number of cluster seeds.

Outputs: $\left(\left([B_{ij}]_{i=1,j=1}^{4,2}\right)_k\right)_{k=1}^{K}$, a $\mathrm{K} \times \mathrm{K}$ cell array of the bounding boxes of all object instances; MSE a $\mathrm{K} \times \mathrm{K}$ cell array of the Mean Square Error of SIFT projections; MQ, a $\mathrm{K} \times 1$ cell array of the average match quality values; $\overline{\mathtt{S}}$, a $\mathrm{K} \times 1$ cell array of the average Silhouette numbers; $\mathcal{G}$, cluster assignments in each iteration of size $\mathrm{n} \times \mathrm{K}$.

1: FOR $\mathrm{k} = 2$ to K
2: $\mathrm{k}' = 0$
3: $\mathbf{H}'$, $\mathbf{M}'$ and $\mathbf{T}'$←copy $\mathbf{H}$, $\mathbf{M}$ and $\mathbf{T}$ respectively
4: WHILE $\mathrm{k}' \neq \mathrm{k}$
5: $\mathcal{G}_\mathrm{k}$←Clustering Algorithm($\mathbf{H}'$,k)
6: $\mathtt{S}$←Silhouette numbers($\mathcal{G}_\mathrm{k}$)
7: $\overline{\mathtt{S}}_\mathrm{k}$←Average Silhouette numbers($\mathtt{S}$)
8: $\mathrm{k}' \leftarrow \mathrm{k}$
9: tempMSE ←a $\mathrm{k} \times 1$ matrix with zeros
10: FOR $\mathrm{i} = 1$ to k
11: IF LENGTH($\mathcal{G}(\mathrm{k})) \geq$ 3 SIFT feature points
12: [Outliers, $(\mathrm{tempMSE})_i$, $(B)_i$] ← ... affine pose estimator($\mathcal{G}(\mathrm{k})$, $\mathbf{M}'$, $\mathbf{T}'$)
13: UPDATE $\mathbf{H}1$, $\mathbf{M}1$ and $\mathbf{T}1$ without Outliers
14: ELSE
15: $\mathrm{k}' \leftarrow \mathrm{k}' - 1$
16: $\mathcal{V}_\mathrm{i} \leftarrow \infty$
17: $\overline{\mathtt{S}}(\mathtt{k}) \leftarrow -1$
18: END IF
19: END FOR
20: MQ_k = Statistical mean of $(\mathrm{tempMSE})_{i=1}^{k}$
21: $\mathrm{MSE}_\mathrm{k} = \mathrm{tempMSE}$
21: IF $\mathrm{k}' < \mathrm{k}$ && $\mathrm{k}' > 0$
22: $\mathrm{MQ}_\mathrm{k} \leftarrow \infty$
23: $\overline{\mathtt{S}}_\mathrm{k} \leftarrow -1$
24: ELSE
25: STOP WHILE
26: END IF
27: END WHILE
28: END FOR
29: CALL Algorithm 4.3

Algorithm 4.3 Pseudo-code for step 6 in Figure 4.1.

Inputs: $\overline{\mathrm{S}}(\mathrm{k})$, a $\mathrm{K} \times 1$ cell array of the average Silhouette numbers; MQ, a $\mathrm{K} \times 1$ cell array of the average match quality values.

Outputs: $\mathrm{K}_{\mathrm{final}}$, final number of object instances found.

$\mathrm{C} \leftarrow \max(\overline{\mathrm{S}})$
WHILE $\mathrm{MQ_C} \neq \min\left([\mathrm{MQ_i}]_{i=1}^{K}\right)$
  IF $\overline{\mathrm{S}}_{\mathrm{C}} \leq 0$
    $\mathrm{MQ_C} \leftarrow \infty$
    $\overline{\mathrm{S}}_{\mathrm{C}} \leftarrow -1$
  ELSE
    $\mathrm{C} \leftarrow 1$
    STOP WHILE
  END IF
  $\mathrm{C} \leftarrow \max\left([\overline{\mathrm{S}}_{\mathrm{i}}]_{i=1}^{K}\right)$
END WHILE
$\mathrm{K}_{\mathrm{final}} \leftarrow \mathrm{C}$

The test image data set is therefore generated by a composition of the segmented objects of the trained database images, transformed with random position, scale and orientation. Two image data set categories are generated;

1. From 1 to 15 separated and multiple overlapping same object instances (as depicted in Figures 4.3(a) and (b), respectively).
2. Two overlapping same object instances (as shown in Figure 4.3(c)).

The validation methodology comprises the verification of how many instances of the same object the proposed algorithm accurately detects over different positions, orientations, and scales and, also, partially occluded by other instances of the same objects in synthetic composited imaginary. *Receiver Operating Characteristic curves* (ROC-curves) (Fawcett, 2006) are employed to measure the system performance. That is, the closer the curve to the upper-left hand of the plot, the better an algorithm performs. Thus, the classification of the obtained results are briefed as follows:

- The *true positive rate* includes all correct localisations over all object instances in the input test image such that the match quality value (described in Section 4.6) is less than ± 1 pixel of average back projection error.
- The *false positive rate* encompasses all incorrect localisations over the number of detections (correct and incorrect) and, also, denotes incorrect localisations when the match quality does not meet the above match quality threshold.

Table 4.1: Maximum percentage perception overlap between same-class object instances.

Database object class	1	2	3	4
Fuzzy C-means [%]	68.42	60.93	59.37	46.87
hierarchical complete [%]	68.42	61.10	59.37	46.87

As the only parameter threshold that can be manually set is the log-likelihood threshold test of the SIFT feature matching process, it is therefore varied in the range of 0.4 to 0.6 with a step of 0.025 in order to test the performance of the method with different match sensitivities. The overlap threshold α used for all the experiments is 0.7 (70% of overlapped between instances).

The algorithm performance evaluation consists of measuring the correct number of objects in the image versus the localised objects and the match quality value, MQ, obtained in Section 4.6. Consequently, in order to summarise the ROC-curves, threshold average ROC-curves are employed for all different runs over a range of log-likelihood thresholds of the matching process. The error bars in the thresholded ROC curves are at 95% confidence interval.

Finally, the perceptual degree of overlap of the algorithm that can successfully tolerate is experimentally measured as follows:

$$A_{ov} = \frac{\text{area}(B_1 \cap B_2)}{\max\left(\text{area}(B_1), \text{area}(B_2)\right)} \tag{4.18}$$

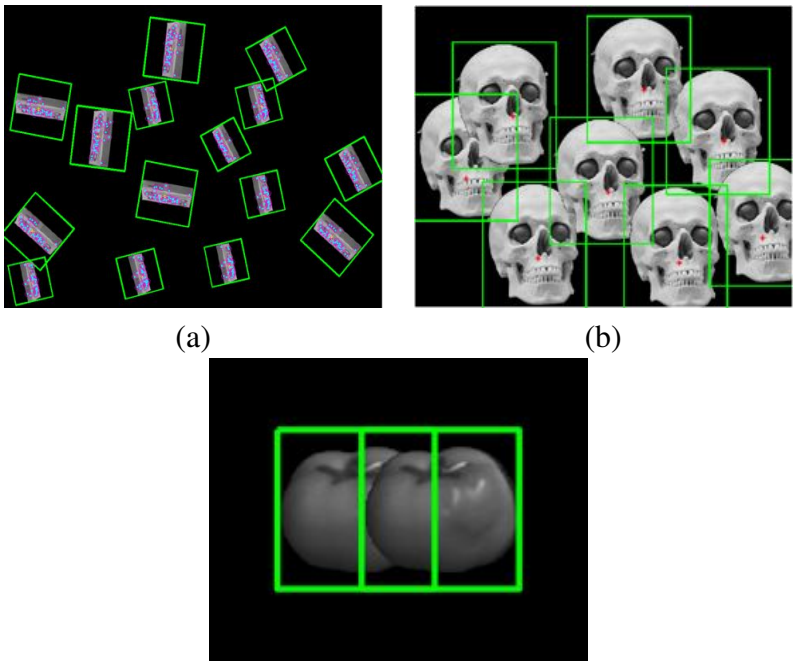
where A_{ov} is the percentage area of occlusion between two objects, and B_1 and B_2, the bounding boxes of each object (bounding boxes are previously defined in Section 4.6 and depicted in Figures 4.2, 4.3 and 4.5).

4.7.1 Synthetic Composite Image Experiments with a Plain Background

This section presents the overall performance of the proposed algorithm applied to images synthetically composited against a plain black background. Figure 4.3 depicts example images and Figure 4.4 illustrates threshold average ROC curves for each clustering algorithm. Table 4.1 summarises the percentage of overlap perception degree tolerated by clustering the Hough space with each of the proposed clustering approaches.

4.7.2 Discussion

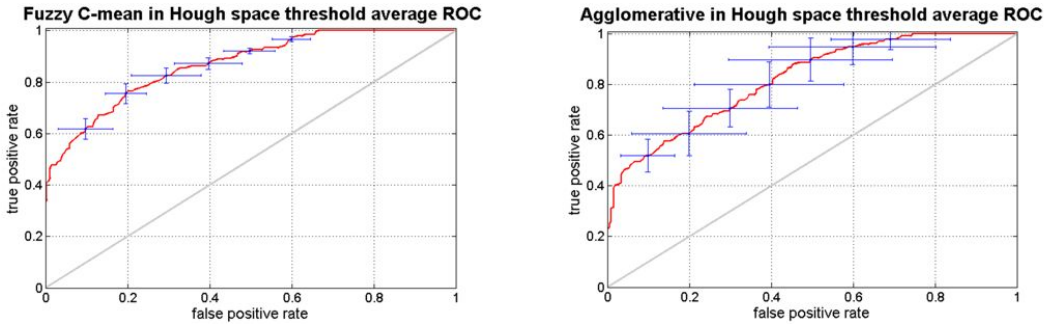
The fuzzy C-means clustering outperforms the hierarchical complete clustering as depicted in Figure 4.4(a) and (b), respectively. The maximum percentage of recall using fuzzy C-means



(a) (b)

(c)

Figure 4.3: Examples of synthetically composited images; (a) multiple instances separated (with SIFT features of the $\mathbf{M}'$ and $\mathbf{T}$ sets overlaid), (b) multiple instances overlapped, and (c) two overlapping instances.



(a) (b)

Figure 4.4: Threshold average ROC curves of (a) fuzzy C-means and (b) hierarchical complete clustering.

clustering exceeds that achieved by its hierarchical counterpart; 40.4% and 36.8% respectively. As it can be observed, the Hough space provides high-confidence localisations. This is due to the compactness offered by the Hough space as it maximises the probability of finding the right number of clusters and, additionally, outliers in the data easily.

By inspecting Figure 4.4(a), the fuzzy C-means clustering exhibits a steady performance without decreasing the average recall rate for different log-likelihood thresholds. Whilst its hierarchical counterpart reveals high variance over the ROC curve (Figure 4.4(b)) because the performance is affected directly by the log likelihood threshold (similar results are obtained in (Zickler and Efros, 2007)).

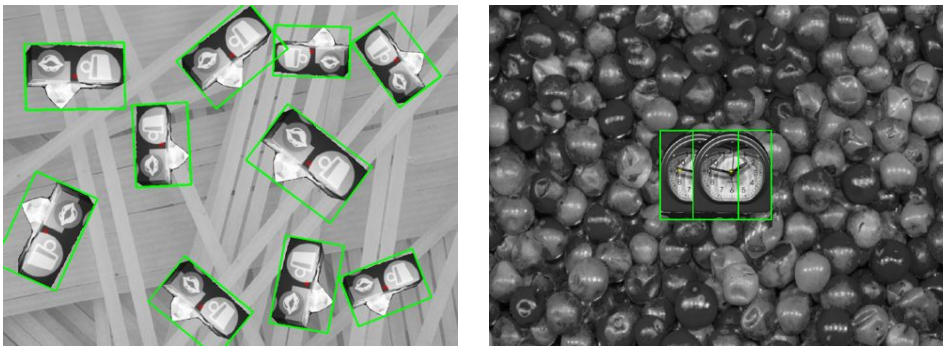
The probability of occurring correct localisations using the fuzzy C-means clustering, the so-called Area Under the Curve (AUC), is 86.8% as opposed to 84% achieved by the hierarchical analogue. The AUC is defined as the probability (i.e. performance) of correctly detecting same-class object instances on the image with respect to detect incorrect ones, a closer value to 1 denotes the best performance. Notably, the fuzzy C-means clustering achieves a maximum AUC of $\sim 87\%$ with a log likelihood ratio of 0.525 to correctly localise multiple instances of the same object in these experiments.

Both clustering approaches are capable of correctly detecting up to 15 objects as illustrated in Figure 4.3(a). Nevertheless, up to 6 objects can be detected and localised with reasonably high confidence by the fuzzy C-means clustering whilst the hierarchical complete agglomerative clustering can operate with up to 4 objects. Above these values, incorrect detections and localisations increase considerably.

The maximum overlap between same-object class instances (as summarised in Table 4.1) obtained with the proposed algorithm is 68.4% of occlusion in Hough space with both clustering methods over 4 different object classes. These observations demonstrate the perceptual robustness of the proposed continuous Hough space representation. That is, the Hough space allows to separate two different peaks in accordance with their geometric configuration of the visual features in the image.

4.8 Final Experiments and Discussion

Following the previous section, the adoption of an unsupervised clustering approach as the means of grouping projected features in Hough space provides a general method to localise same object instances without adjusting initial parameters for each observed synthetic composite images. However, the full characterisation of the proposed algorithm requires extensive



(a) (b)

Figure 4.5: Examples of synthetic composite image datasets employed: (a) multiple same-object instances (object class: 815) and (b) two overlapping same-object instances (object class: 208).

validation experiments, which consider the presence of distractors and real-world settings.

For this purpose, a database of 2000 synthetically composited images are generated and set against a real-world textured background. Each composited image comprises varying numbers of objects in different positions, scales, and orientations as the previous experiments. Two test image datasets are considered in these experiments; from 1 to 10 separated and multiple overlapping same object instances (as depicted in Figure 4.5(a)) and two overlapping same object instances (as shown in Figure 4.5(b)).

By adopting a textured background, these images provide a considerable source of false positive SIFT feature matches; and it also allows validating the performance of the proposed algorithm under more controlled and natural image settings than those results presented in Section 4.7. Hence, one hundred objects were selected at random from the *ALOI image database* (Geusebroek et al., 2005). (Table 4.2 lists the object class numbers used in these experiments). This state-of-the-art image object database consists of a collection of 1000 images of objects captured under controlled lighting conditions and camera settings). For each random selected object , five different image poses of ± 5 degrees of in-plane rotation offset are employed to train an image database as in Section 3.3 (i.e. 350, 355, 0, 5 and 10 degrees as specified in the *ALOI image database*). Likewise, fifty textured images are randomly selected from (Burghouts and Geusebroek, 2009a); this state-of-the-art image texture database consists of a collection of 250 different textures.

The main objective of this algorithm is however to improve object detection tasks, in the context of active robot vision, and provide a tool for automatically determining the number of instances for each object class detected in a visual search task. It is therefore required to evaluate the accuracy of the proposed algorithm as a covert and endogenous visual behaviour mechanism in real world settings. To that end, 342 stereo image pairs are captured using

Table 4.2: Object class numbers of the ALOI image database.

5	104	244	315	445	570	703	758	814	893
9	110	260	317	451	575	707	760	815	894
48	114	269	319	454	578	722	772	820	896
50	127	271	323	471	593	727	773	825	918
54	134	275	348	482	612	728	782	830	933
74	170	289	377	497	616	729	789	832	939
80	208	294	393	505	629	734	793	843	951
86	228	297	398	521	681	741	799	853	962
88	235	307	409	556	686	749	804	854	963
97	243	308	427	561	701	754	808	863	978

the active binocular robot head described in Chapter 3. Each image pair captured consists of different viewpoints while the cameras are verging on an unknown scene point whilst objects are still within the field of view of both cameras as depicted in FIgure 4.9.

The robot head experiments therefore consist of five object classes. The trained database is constituted under the same specifications as the synthetically composited experimental test-bed. However, the maximum number of objects considered is 5 objects per image since the proposed algorithm is capable of accurately detecting and localising multiple instances when fewer than 6 objects are present in the image (ref. Section 4.7.2). The captured test image dataset is comprised of different in-plane rotations without exceeding 20 degrees of in-plane rotation, a constraint experimentally determined by (Lowe, 2004). Similarly, the best clustering algorithm determined from the results of the synthetic images is thus selected and measured its performance against the captured real-world images. It must be noted that the results presented are restricted to the pre-attentive localisation and detection of multiple instances of the same object class in both camera eyes; therefore, attentive functions in the active binocular robot head are disabled during these experiments.

The validation methodology for both experimental test-beds follows the same formulation described in Section 4.7. That is, ROC curves are employed to measure the robustness and performance of the method. However, the proportion of examples detected as correct that are indeed correct is not reflected in ROC curves since these curves only summarise the trade-off between correct versus incorrect multiple same-object class detections. For that reason, Precision and Recall curves (PR-curves) (Davis and Goadrich, 2006) are employed to identify subtle differences between clustering approaches. Davis and Goadrich (2006) define that *Recall*, in PR-curves, is exactly the true positive rate on ROC-curves whereas *Precision* captures the rate of correct detections over the total number of correct and incorrect detections (i.e. the probability of producing a correct detection when the match quality error is below one pixel). Therefore, a correct performance is denoted by observing that the PR-curve approx-

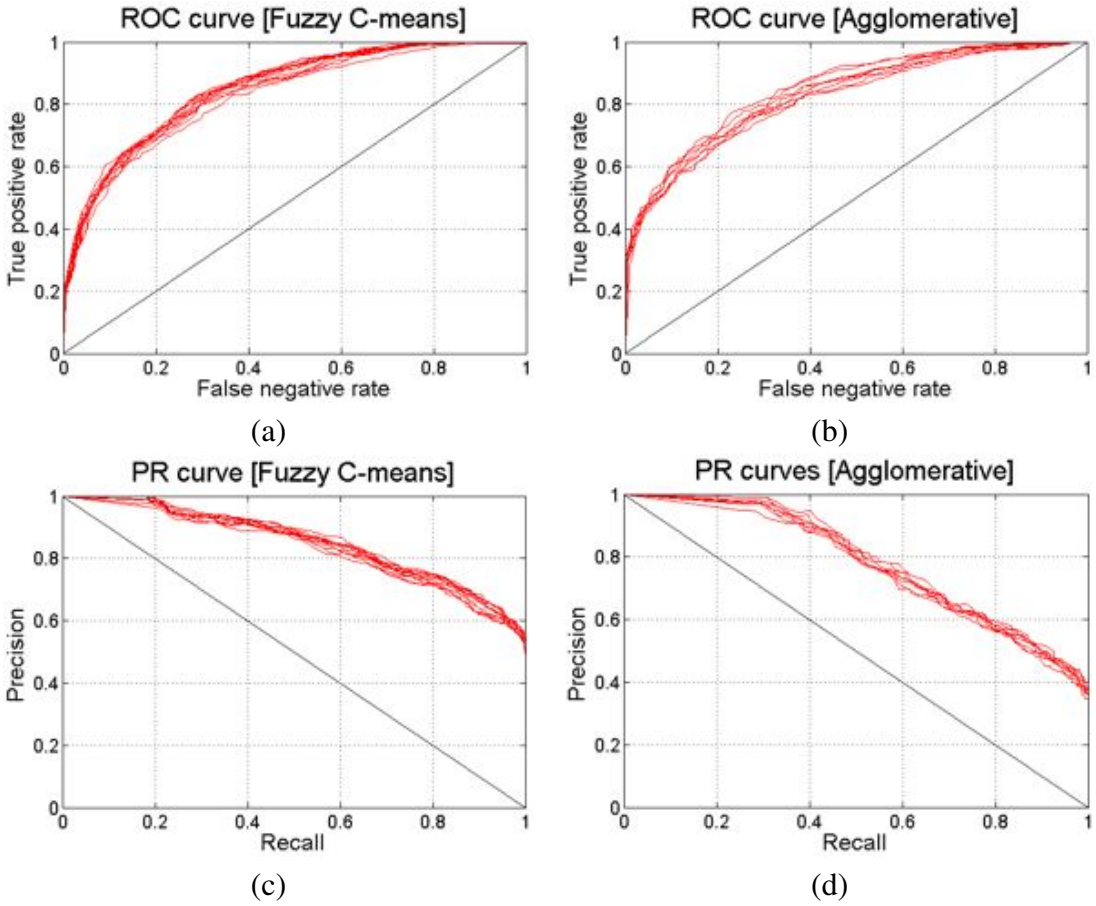


Figure 4.6: (a) and (b) ROC-curves, and (c) and (d) PR-curves of fuzzy C-means and hierarchical complete clustering, respectively (Aragon-Camarasa and Siebert, 2010).

imates to the upper-right hand corner of the graph. Therefore, the performance of the proposed algorithm is measure by means of ROC- and PR-curves over a range of log-likelihood thresholds.

4.8.1 Synthetic Composite Image Experiments

This section presents the characterisation and performance of the proposed algorithm tested against real texture composited images. These experiments are validated with a range of log-likelihood thresholds from 0.5 to 0.7 in steps of 0.01 for each described clustering algorithm. The overlapped perceptual threshold, α, employed in the experiments, is set to 0.7 (increasing this value results in a greater number of incorrect detections).

Figures 4.5(a) and (b) show experimental samples of the employed image dataset, and Figure 4.6 presents the performance characterisation by means of ROC-curves and PR-curves for each clustering algorithm.

By visual examination of Figures 4.6(a) and (b), the trade-off between correct (true positive rate) and incorrect (false positive rate) detections shows signs of comparable performance for both clustering algorithms. In that respect, a comparison of the AUC (Area Under the Curve as defined in Section 4.7.2) corresponds to ~ 0.82 and ~ 0.83 for the fuzzy C-means and the hierarchical agglomerative clustering algorithms, respectively. In this experimental setup, both ROC-curves (Figures 4.6(a) and (b)) present similar behaviour for both clustering algorithms since the proposed image database observes a high skew in the class distribution. In on the words, the confidence in detecting two object instances of the same object class is not analogous to the obtained confidence of detecting ten object instances; therefore, similar performance results are expected for both clustering algorithms, as discussed by Fawcett (2006) and Davis and Goadrich (2006).

Hence, Figures 4.6(c) and (d) capture different classification performance measurements for both implemented clustering algorithms. The agglomerative case presents inferior precision over the recall rate as observed in Figure 4.6(d) while the fuzzy C-means clustering is close to the upper-right hand corner of the graph. The precision at which instances can be correctly recalled using fuzzy C-means clustering at 0.5 is ~ 0.89 whilst its agglomerative counterpart achieves ~ 0.83. This means that the fuzzy C-means returns high correct detections with a favourable confidence almost half of the times the proposed algorithm is invoked. As the number of object instances increases on the image, a greater number of incorrect detections are depicted. That is, at greater recall rates, the precision rate declines since the ability of both clustering algorithms to produce correct detections decreases (as depicted on Figures 4.6(c) and (d)). However, the precision rate of the fuzzy C-means never falls below chance as opposed to the agglomerative method.

The overlap perception degree achieved by the proposed algorithm for both clustering algorithms is depicted in Figure 4.7. This overlap perception threshold is defined as the percentage of mutual occlusion between same-object class instances that the proposed algorithm is able to perceive. By inspecting Figure 4.7, it is observed that the object class number "228" achieves the greatest perception percentage of occlusion in these experiments with a value of 65.9%. However, by observing the mean and deviation error (at 1 standard deviation) values for each clustering algorithm; the fuzzy C-means (mean: 51.7%, deviation error: 11.9%) achieves the best rate of perceptual overlap as opposed to its agglomerative equivalent (mean: 44.2%, deviation error of 16.3%).

Similarly, Figures 4.8(a)-(b) and (c)-(d) illustrate the SIFT feature samplings of the extreme object classes that observe the lowest and greatest overlap perception degree in Figure 4.7, respectively. From these results, it is inferred that the co-existence of more than one object class instance is proportional to the SIFT feature density of the object class. That is, a highly

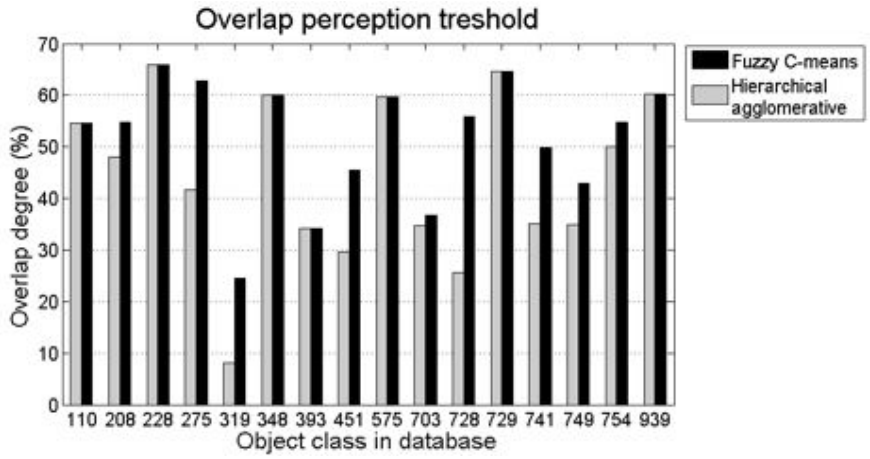


Figure 4.7: Overlap perception threshold in percentage. Numbers of the object classes correspond to those found in (Geusebroek et al., 2005).(Aragon-Camarasa and Siebert, 2010)

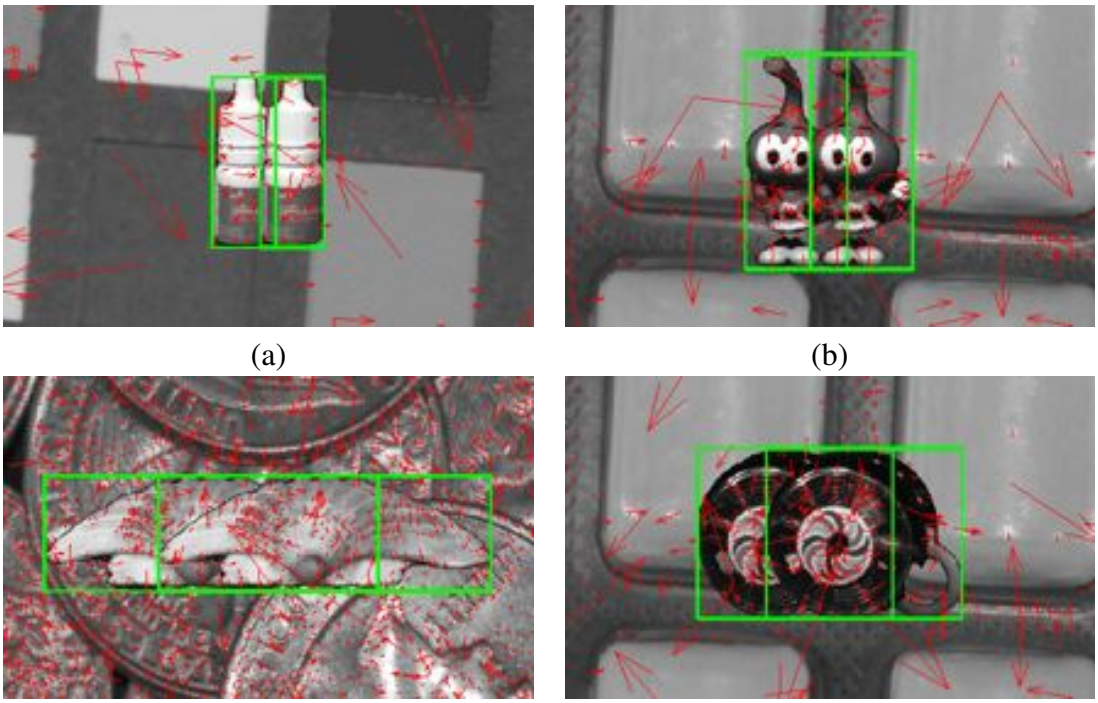
texture object results in a greater number of SIFT keypoints (Figure 4.8(c)); therefore the degree of overlap perceived is generally greater than an object class with lack of texture, e.g. Figure 4.8(a).

Therefore, the proposed algorithm in combination with the fuzzy C-means clustering is capable of detecting and localising up to 6 objects (AUC: ~ 0.89 at 0.5 recall rate) with a reasonable high confidence (as already discussed in Section 4.7.2) while the hierarchical complete agglomerative clustering, up to 4 objects (AUC: ~ 0.83 at 0.5 recall rate). Above these values, incorrect detections and localisations increase considerably. The results obtained in these experiments are supported with those found in the pre-attentive function of human visual system (ref. Section 2.5 on page 37). That is, the devised algorithm is capable of pre-attentively keeping track to up to 5 objects at the same time without invoking eye movements (as discussed in Section 2.5, e.g. (Chun and Wolfe, 2004; Styles, 2005; Franconeri et al., 2010)).

Following a series of validation experiments of the devised algorithm, it is thereby argued that the selected fuzzy C-means clustering algorithm configuration allows detecting and localising same object instances that clearly outperforms state-of-the-art approaches (Zickler and Efros, 2007).

4.8.2 Robot Head Image Experiments

As discussed in Sections 4.7.2 and 4.8.1, the fuzzy C-means clustering algorithm has been demonstrated to localise, with a high detection probability (AUC: ~ 0.86), multiple instances of the same object . Therefore, this clustering technique is selected for the purpose of a covert and endogenous visual behaviour mechanism in the investigated active binocular robot head.



(a) (b)

(c) (d)

Figure 4.8: SIFT feature samplings of (a)(b) low degree of overlap (object classes: 319 and 703) and (c)(d) significant degree of overlap (object classes: 228 and 729).(Aragon-Camarasa and Siebert, 2010)

Figure 4.9 shows stereo image pairs of real-world settings where the active binocular robot head operates. It can be seen that the different scenes contain different unknown objects that are a typical source of potential outliers that affect the overall performance of the devised algorithm. Figure 4.9 shows the ROC-curves obtained using fuzzy C-means clustering over a range of likelihood thresholds from 0.5 to 0.6, in steps of 0.025 as the best observed performance in previous experiments is within this range.

The ROC- and PR-curves in Figures 4.10(a) and (b) demonstrates that the system has good performance over a set of real-world settings where noise and illumination changes decrease the performance of any machine vision algorithm. Also, consistent performance, i.e. low statistical variance, over different likelihood thresholds is observed. The probability of correct detections (i.e. AUC) measured is 0.92 and the true positive rate is 75% versus 10% of false positive rates. Figures 4.10(a) and (b), depict that the overall performance is relatively invariant and insensitive to the log-likelihood threshold. It is therefore concluded that the devised algorithm demonstrates the ability to detect and localise same object instances, within viable levels of performance under real-world settings.

Figure 4.11 shows examples of occluded instances of the same object. The overlap perception degree in these experiments achieves a similar overlap percentage as in previous experiments (as in Table 4.2 and Figure 4.7); an average of $\sim 55\%$. However, the overall performance

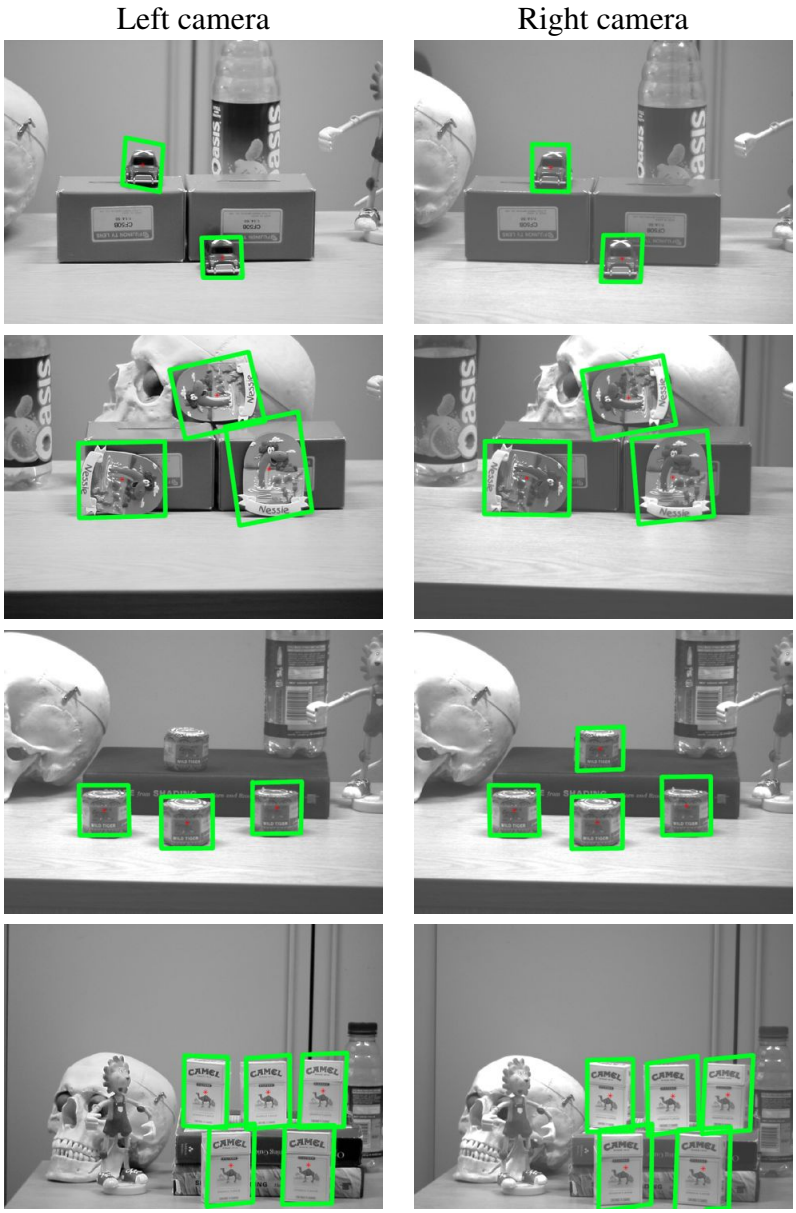
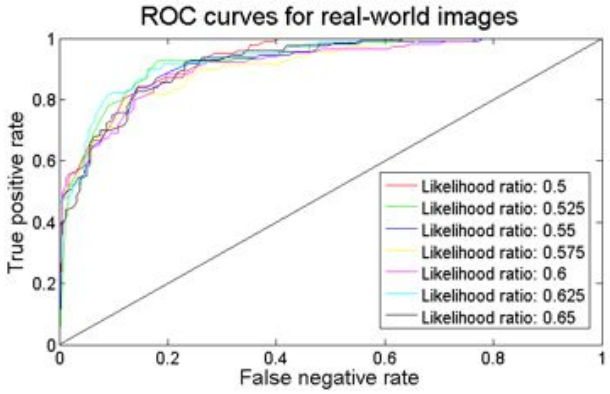
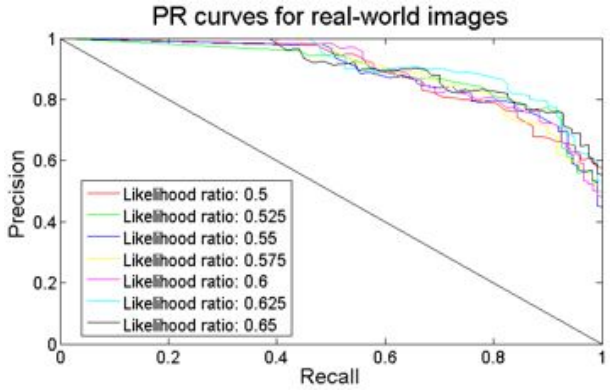


Figure 4.9: Robot head output examples of the left camera from 2 to 5 objects and the right camera from 2 to 5 objects.(Aragon-Camarasa and Siebert, 2010)



(a)



(b)

Figure 4.10: Robot head output examples; (top row) left camera images from 2 to 5 objects, (bottom row) right camera from 2 to 5 objects.(Aragon-Camarasa and Siebert, 2010)

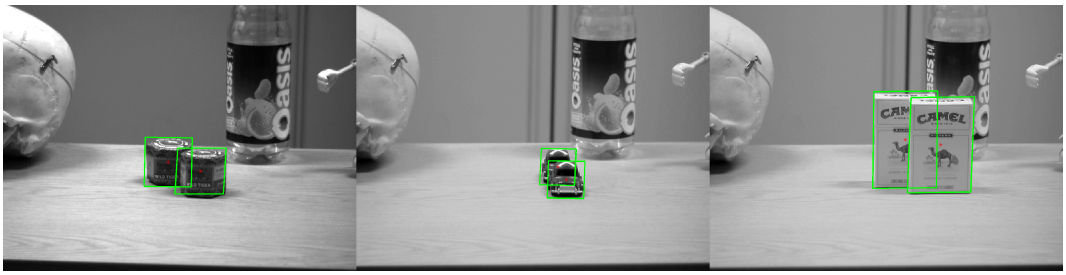


Figure 4.11: Overlap object samples from the real-world images.(Aragon-Camarasa and Siebert, 2010)

of the algorithm decreases considerably (more incorrect detections) if an overlap perception degree threshold is set to 0.7. Therefore, by setting the overlap degree threshold to 0.6, it is observed that the false positive rate decreases considerably. Thereby, the optimal α of approximately 0.6 indicates the best performance that must be considered in the forthcoming Chapters of this thesis.

4.9 Conclusions

This chapter presents an algorithm to localise and detect same-class object instances in a covert and endogenous visual behaviour competence. This visual behaviour serves as a pre-attentive perceptual core for a robot vision system, investigated in this thesis, which is able to explore the environment autonomously with a viable level of visual competence. The devised algorithm exploits the intrinsic properties of a continuous Hough space to find projected SIFT feature groups represented as multiple distinct peaks in the said space. Similarly, this algorithm adopts well-known pattern recognition techniques in order to group projected SIFT features into a continuous Hough space representation without requiring to tune clustering distance thresholds in accordance with specific object classes and image contents (as in Zickler and Efros (2007)).

Validation experiments show that the fuzzy C-means clustering algorithm achieves a successful precision performance between ~ 0.83 and ~ 0.98 with a recall of 0.5; and, an average accuracy above chance (as depicted in Figures 4.6(a) and 4.10(c)) over synthetically composited and real-world images datasets. It is also demonstrated that the perceptual capabilities of the devised algorithm achieve an overall overlap percentage of $\sim 66\%$. The processing time incurred in the validation experiments is subject to the complexity of the image contents and the number of matched features. Overall, the computational processing cost per image (with a resolution of 1024×768 pixels) is between 0.9 and 5 seconds. 40% of the consumed time belongs to the SIFT descriptor generation, 25%, to the matching process; and the remaining

time stands for the projection into Hough space, the clustering algorithm, and the stabilisation steps of the proposed algorithm.

It must be noted that in Figure 4.9 (second top image from right to left which corresponds to the left camera), 4 objects are in the field view but only 3 are correctly localised. This is due to the fact that the different viewing perspectives of the left camera and the trained object pose in the database do not allow the system to detect and localise all the objects present in the camera's field of view. However, in the right camera image (bottom image in the same column), the proposed algorithm detects and localises accurately all 4 objects. As the robot vision investigated adopts binocular vision, it is thereby possible to merge each camera solution into a single representation of the current observed portion of the scene. *Binocular rivalry* is found in biological systems, in accordance with (Styles, 2005), as the scene is imaged as a single panoramic view captured from both converging eyes. This binocular rivalry should therefore be included in the development of any robotic visual search strategy implementation and, specifically, the active robot head investigated in this work. Hence, the next chapter addresses the implementation of this multiple same object-class instance detector into the overall framework of the active binocular robot head.

The proposed multiple same-class instance detector herein described is demonstrated to subserve as a pre-attentive, covert detection mechanism for the perception of multiple instances in a scene. however, there are limitations that must be addressed and investigated, for example: the localisation of multiple instances in the three-dimensional domain for robotic object manipulation and the integration of colour or 2.5D descriptors (as reported in Burghouts and Geusebroek (2009b) and Lo and Siebert (2009), respectively). These limitations are discussed in Chapter 8.

Finally, this detector, as already discussed, will be part of the pre-attentive cueing system of the visual search strategy. However, the active robot vision system presented in Chapter 3 develops visual behaviours as a collection of "ad-hoc" functions that do not comply with the design principles of robotic systems described in Section 2.6. Thus, the integration of the developed detector in this chapter and the outlined shortfall in the system's design are addressed in the forthcoming chapter.

Chapter 5

The Hierarchical Active Binocular Robot Vision Architecture

The ability of the prior state of the active binocular robot head (as discussed in Chapter 3) is only capable of detecting a single object class present in the scene. That is, only the most confident object class observed in the pre-attentive module can be attended whilst other instances of the same object class are inhibited. In that respect, the previous chapter proposes a novel method to covertly detect multiple same-class object instances to extend the visual capabilities of the active binocular robot vision system in Chapter 3. However, the described visual modules are integrated as hard-wired, ad-hoc functions that are constrained to a specific visual purpose, as outlined above. Therefore, this chapter[1] *presents a binocular, active robot vision architecture that synthesise visual behaviours in a parsimonious and uncomplicated manner and, in consequence, permit to integrate the multiple same-class object instance detector. The devised architecture features visual behaviours that are not constrained to the hierarchy but can be adapted and applied to different contexts as demonstrated in the forthcoming chapters. This chapter concludes with a pilot investigation of the visual search stability and repeatability in a complex and cluttered scene where known and unknown multiple same-class object instances are present.*

[1] *This chapter is based on and extends the following peer-reviewed paper:*

- Aragon-Camarasa, Gerardo and Siebert, J Paul, "A Hierarchy of Visual Behaviours in an Active Binocular Robot", in Kyriacou, Theocharis and Nehmzow, Ulrich and Melhuish, Chris and Witkowski, Mark, ed., Towards Autonomous Robotic Systems, TAROS 2009 (University of Ulster, 2009), pp. 88–95.

5.1 Introduction

A central goal in the field of computer vision is the development of active binocular vision systems which have the ability to interact with dynamic and complex environments. By coupling sophisticated computer vision algorithms with high-level reasoning mechanisms, it is envisaged that these systems will eventually lead to machines aware of their surroundings and mimic the motion of human heads undertaking visual search tasks (as outlined in Chapter 1). In that regard, computer vision algorithms, such as SIFT, now provide basic mechanisms for supporting visual understanding of the environment (as demonstrated in Chapter 3) and can be used to extract highly discriminative visual features that can then be transformed into actions/events which, in turn, command a robot, i.e. (Kragic et al., 2005).

Chapter 3 described the initial development of the active binocular vision system investigated in this thesis. This system employs a standard implementation of the SIFT algorithm as a basic visual representation, within a gaze control system that exploits biological motivated concepts such as: pre-attentive search, attentive saccadic targeting, binocular vergence and the "attentional spotlight" metaphor for visual search tasks (Styles, 2005). In its original form, this robot system is conceived as a unified framework of hard-wired, ad-hoc functions and is capable of detecting a single instance, within each observation of a scene, of each object class contained in a database of pre-trained object examples.

To that end, in the previous chapter, a novel detector is devised and validated in order to localise multiple instances of the same object class. The integration of this detector as part of the visual capabilities repertoire of the robot vision system in Chapter 3, provides the means of enabling the system to detect and localise multiple same-class object instances in cluttered and complex settings, which contain both same-class known and unknown objects.

This chapter therefore details the specific functional improvements (i.e. multiple same-class instance detector) in the investigated robot vision system. However, during the system's integration, it was observed that the implementation must redefine the overall structure of the robot vision system. This is due to the fact that visual abilities are defined as ad-hoc functions that only subserve a high-level task-goal specification. To that end, this chapter also proposes and introduces a novel hierarchical robot vision architecture which is designed under the design paradigms described in Section 2.6, and develops attentional mechanisms to suit the requirements of multiple object detection in the overall robot vision system.

The remainder of this chapter is therefore organised as follows: Section 5.2 outlines the objectives and motivation of designing the active binocular robot vision architecture, while Section 5.3 presents its design rationale and overall specification. Thence, Section 5.4 introduces the

mathematical framework on which the devised architecture is based and, as a consequence, Sections 5.5, 5.6 and 5.7 describe the visual behaviours included in the robot vision architecture. Section 5.8 defines the task-goal specification of the hierarchy and finally, in Section 5.9, the hierarchical architecture is initially validated by measuring the viability and stability of a visual search task over cluttered, complex settings. The reported results provide the required scientific insights in terms of the system's behaviour and shortfalls in the intrinsic design of the architecture.

5.2 Motivation

Despite the fact that the initial robot vision software presented in Chapter 3 observes robust performance in its defined visual competences, it is limited in its capacity of detecting and recognising multiple instances of the same object class. Furthermore, the described visual abilities are integrated as an "ad-hoc" function in the system that constrains the devised framework to the specified application (i.e. exploration of single object class present in the scene). Thus, the integration of the multiple same-class instance detector in the overall robot vision framework is required to define a collection of visual behaviours in accordance with the robot architecture design principles discussed in Section 2.6.

Chapter 3 concludes that autonomous exploration of a scene is accomplished by operating solely with SIFT features as the underlying basic visual representation of the imaged environment. As a corollary, the behaviour of the system can be further abstracted according to how the operational mode of the active binocular head acts in accordance with the *attention for perception* principle (Styles, 2005). That is, the defined visual search strategy assumes an object-based visual attention cueing approach in order to guide the camera-pair to fulfil visual exploration tasks.

In that regard, the above principle is consistent with the "attentional spotlight" metaphor and, in consequence, with the devised "stepping-stone" visual search pattern. In other words, this attentional principle guides attention towards *objects rather than locations in space* (Posner and Petersen, 1990; Chun and Wolfe, 2004; Fazl et al., 2009; Wallraven and Bülthoff, 2007b). Psychophysical research literature has proposed that perception is contained under two interacting but different attentional systems (related to the *perception-action cycle* and according to Posner and Petersen (1990)) as follows:

- *Covert attentional orientating* (pre-attentive) allows guiding attention towards stimulus locations and is responsible for finding putative known objects and discovering salient regions as they emerge in the visual field of view.

- *Overt attentional orientating* (attentive) drives the "attentional spotlight" in order to focus an object hypothesis or salient item in the field of view of both cameras and such serves to categorise identified objects in the scene.

The pre-attentive attentional function is further divided into two operational modalities: *endogenous, goal-driven* (for perception) and *exogenous, stimulus-driven* (for action). Scientists have therefore agreed that the above different attentional modalities and systems are structured in a hierarchical arrangement of behaviours (Chun and Wolfe, 2004; Styles, 2005) which is consistent to the cognitive model depicted in Figure 1.1.

Hence, the proposed behavioural architecture in this chapter is based on the hybrid deliberative/reactive *Sensor Fusion Effector* architecture (SFX, ref. Section 2.6.3) and the *hierarchical* paradigm (Section 2.6.2). The former is said to resemble the configuration of the neurophysiological model of sensing in animals (Murphy and Mali, 1997); whilst, the latter is considered among robot scientists as the paradigm which resembles primate's behaviours in the early stages of human evolution.

Specifically, the SFX architecture, as implemented in the investigated robot vision system, relates how deliberative and reactive modules are interconnected with sensor and actuator functions; whereas the hierarchical paradigm conforms the arrangement of visual behaviours of visual pathway streams in the reactive layer of the SFX architecture. Thus, a hierarchical visual behaviour arrangement exploits sensed visual information in order to explore the observed environment without further reasoning or understanding of the environment (e.g. senses, plans and acts accordingly to what it is specified (ref. Figure 2.8)) while the deliberative layer provides the required set of commands to carry out a visual task or goal. Thus, the active binocular robot vision system is designed such that:

- It preserves modular engineering principles.
- It shows a precision and a reliable performance above the level of acceptance.
- It can be easily implemented into other robotic domains (as described in Section 2.1).
- It is robust enough for the intended application.

As the overall design of the robot vision architecture is devised in accordance with the above design principles, there most exist a cognitive, deliberative layer. Deliberation, for the objectives of this thesis, is therefore cast as macro scripts where an user indicates the sequential activation of visual behaviour in order to fulfil the specific high level task-goal. This layer is conceptualised and included to enable the development of cognitive functions in future augmentations of the robot system herein described (as discussed in Chapter 8). .

For the ability of the robot head to discern which same-class object instances have been already attended while exploring a scene, it is required to devise an inhibitory mechanism into the active binocular robot head such that the visual search strategy is able to disambiguate between attended and putative objects observed. That is, the functional operation of the initial inhibition of return mechanism in Chapter 3 assumes that objects are unique in the observed scene (i.e. attention is directed to one instance per stored object class in the database). These object classes are therefore not further considered if they have been successfully attended. Specifically, this symbolic inhibition of return mechanism is not longer valid under the new system visual requirements since there might be more object instances in the given scene.

To that end, psychological and psychophysical studies suggest that inhibitory mechanisms can reduce ambiguity of subsequent observations by suppressing previously investigated information and thereby assisting the deployment of the "attentional spotlight" towards potentially interesting regions (Posner and Cohen, 1984). Posner and Cohen (1984) further claim that the inhibitory process is performed on the spatial location of the cue observed in the scene, rather than on the 2D retinal location. Additionally, this process subserves the orientation of the gaze towards novel regions/objects present in the environment. Therefore, the robot's inhibitory mechanism must be implemented in order to allow autonomous exploration such that an object reference frame serves as a mechanism for suppressing visual information during the visual exploration task. The above assumption affords a means of adopting a symbolic map representation of the world for object recognition, since the camera geometry does not need to be known to achieve the visual tasks and goals described in this thesis. In other words, the object's reference frame (expressed in retinotopic coordinates as found in biology (Girard and Berthoz, 2005)) is projected into an egocentric spatio-temporal map.

The described egocentric spatio-temporal map is defined in this robotic architecture as a relative pixel coordinate system where the frame of reference is established with respect to *a "home"* position, relative to the robot head state. Thus, for each saccadic camera motion, the state of the robot head is updated accordingly by means of this map. Furthermore, this pixel map allows the system to maintain an egocentric coordinate system that is decoupled from the actuation behaviours and such is only transformed into motor steps units when it is required (i.e. the number of motor steps per pixel units required to transform these coordinates into actuator space are described in Section 3.5.1). The latter is biologically supported from the above studies (Girard and Berthoz, 2005) and, additionally, complies with the requirements of the proposed architecture.

5.3 The Active Binocular Robot Vision Architecture

As discussed in the previous section, the robot vision architecture herein proposed is based on the *SFX architecture*. Figure 5.1 illustrates such configuration for the overall robot vision architecture investigated. In the devised architecture, the processing levels are classified in terms of their behavioural function (i.e. low-, mid- and high-levels as depicted in Figure 5.1). The corresponding low- and mid-level functions consists of simple yet effective visual operations that subserve upper-level goals, whilst the high-level functions relate to the intelligence layer of a cognitive robot.

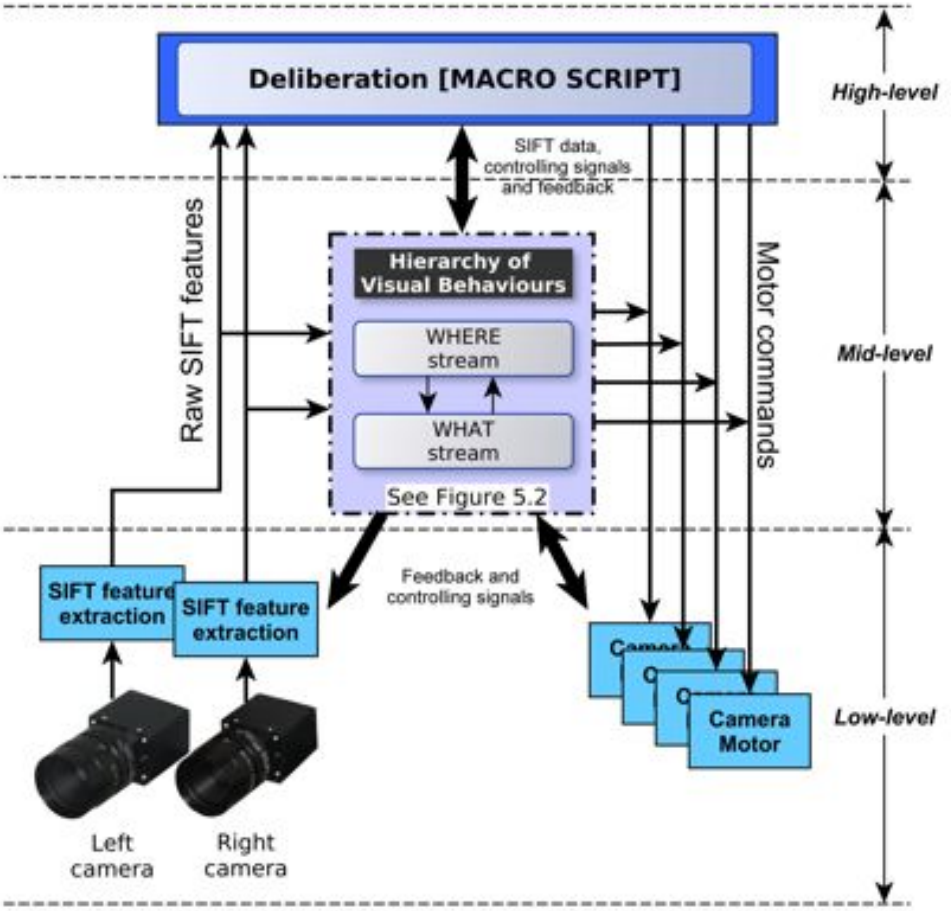


Figure 5.1: The active binocular robot vision architecture. This figure is an specialisation/abstraction of Figure 1.1.

The deliberative layer (or cognitive layer) in Figure 5.1 is inspired by the *cerebral cortex* (and related brain modules such as the basal ganglia (Girard and Berthoz, 2005)) as illustrated in Figure 1.1 which, in this thesis, specifies visual tasks and goals (i.e. *autonomous scene exploration and recognition of multiple same class objects* as defined in Section 1.2). It is also employed to define other types of tasks in order to achieve complex goals (e.g. an object

appearance learning behaviour devised in Chapter 6). As previously stated, this layer is out of the scope in this thesis and it is defined as a *macro script* where a human operator specifies the activation sequences of behaviours in lower layers required by the defined visual search strategy (Section 5.8). However, this layer is considered and integrated in order to design a general purpose robot vision architecture that fulfils the design requirements described in Section 5.2.

Low- and mid- processing levels incorporate an arrangement of *reflexive* and *reactive* behaviours. Such behaviours have already been reviewed in Section 2.6. These are further classified with respect to their configuration and functional operation in *primitive* and *abstract behaviours*. On the one hand, primitive behaviours are monolithic methods that only serve a single (i.e. mono) purpose. That is, they are simple stimulus-response mappings that transform a collection of sensed information into data structures (Murphy and Mali, 1997). On the other hand, abstract behaviours are composed of a collection of primitive or other abstract behaviours that serve high-level task-goal specifications. Therefore, these behaviours transform data structures and can also invoke sensor and motor related functions as required.

It can be observed from Figure 5.1 that sensor and motor related primitive behaviours (i.e. *low-level* functions) are decoupled from the deliberative layer and the *"hierarchy of visual behaviours"*. This arrangement therefore allows to maintain an egocentric coordinate system (ref. the egocentric spatio-temporal map discussed in Section 5.2) in higher-levels which are not related to the real-world units of the observed environment. Hence, low-level functions link visual behaviours and cognitive processes with the environment and, in turn, only deal with (stereo pair) image acquisition, SIFT feature extraction and motor actuation commands. In regard with the SIFT visual representation, Section 5.4 describes the theoretical framework in which the proposed architecture is specifically based on.

In spite of the fact that SIFT features are employed in this thesis, the overall architecture is not constrained to such feature extraction technique to the extent that the related behaviours which process SIFT features are modified accordingly to the new feature description approach. Moreover, image acquisition and motor control are implemented in accordance with specific hardware configurations and as such are thereby adapted to the required data structures of the architecture. In that regard, the robot head hardware (described in Sections 2.1.3 and 3.2, Figure 3.1 and Appendix A) are modified to suit the needs of this architecture accordingly. Therefore, the design of low-level behaviours is not considered in this chapter as the underlying objective of the described robot vision architecture is specifically to be general in its purpose and not hardware related.

The mid-processing level is composed of visual behaviours structured in terms of WHERE (dorsal) and WHAT (ventral) visual processing streams (as discussed in Sections 1.4 and 2.4).

Both streams, as depicted in Figure 5.1, are composed of hierarchical configurations of abstract and primitive behaviours that model the categorisation of attentional behaviours (Fazl et al., 2009). Therefore, the *hierarchy of visual behaviours* is modelled in accordance with the WHAT and WHERE stream. Specifically, the WHAT stream, in the context of this thesis, consists of SIFT matching operations and memory management of feature descriptions stored in an object knowledge database. Whilst the WHERE stream carries out coordinate based operations such as the detection of objects, tracking of the spatial locations of either putative objects or attended objects, the control of the attentional spotlight towards the indicated object/feature locations and, finally, the management of feature locations in the object knowledge database.

Both streams behave in a closely parallel operation according to the current visual task-goal specification and such are therefore devised in terms of the *hierarchical paradigm*, as discussed in Section 5.2. On the following subsection, the devised hierarchy of visual behaviours is described.

5.3.1 Hierarchy of Visual Behaviours

Figure 5.2 illustrates the overall structure of the hierarchy of visual behaviours within the overall robot vision architecture (i.e. this hierarchy corresponds to the mid-level behaviours in Figure 5.1). As depicted in Figure 5.2, abstract behaviours are the *pre-attentive*, *attentive*, *inhibition of return* and *binocular vergence*. Each is composed of other primitive and abstract behaviours.

Thus, in order to extend the capabilities of the active binocular robot head to enable it to detect and localise multiple instances of same-class objects, three principal modifications had to be specifically effected in the original robot vision framework in Chapter 3:

- the detection of feature space clusters corresponding to each object instance present,
- the gaze control strategy to handle multiple instance hypotheses, and,
- the inhibitory process to enable the system to search multiple object instances.

To that end, each of the above-stated abstract behaviours is cast in accordance with the new visual requirements. Each of these behaviours and their particular visual objectives are briefly described as follows.

The *pre-attentive behaviour* is solely responsible for detecting SIFT features in the current field of view that may be of interest to the visual search strategy and overall task. The specified visual objectives of this behaviour are thereby summarised as follows:

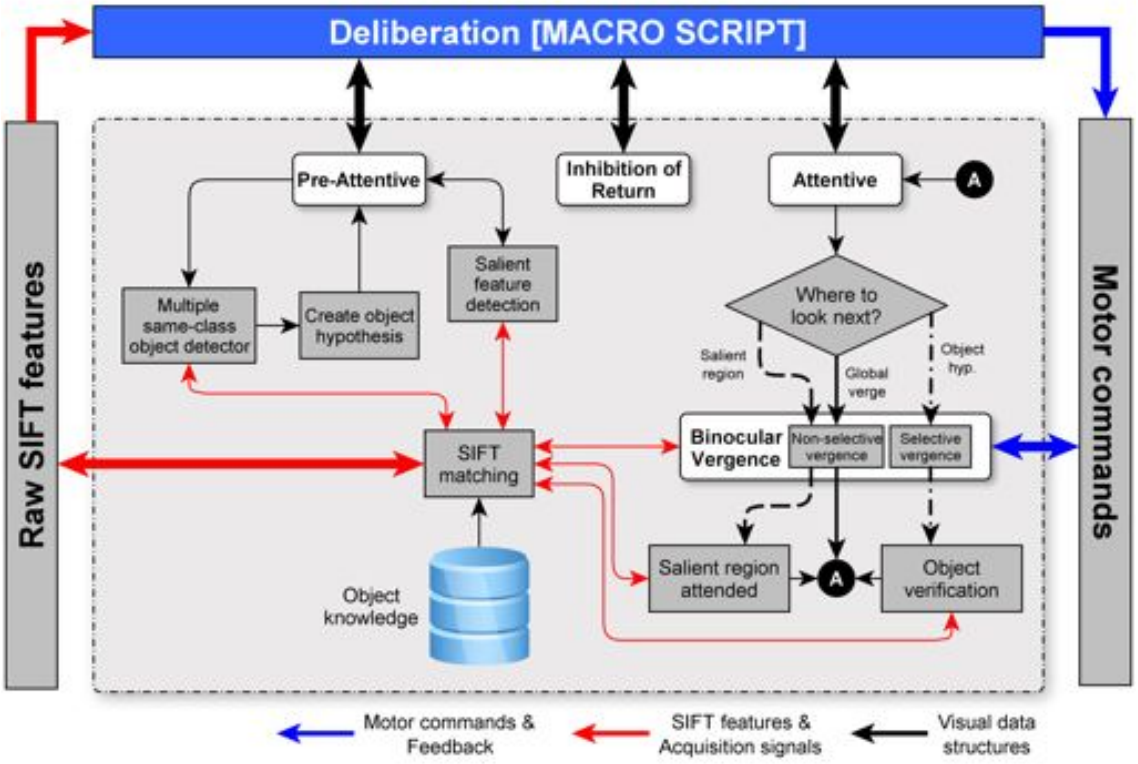


Figure 5.2: Hierarchy of visual behaviours within the devised robot vision architecture. White boxes denote abstract behaviours, whereas grey boxes represent primitive behaviours.

- Covert, top-down attention of multiple same-class object instance detection (Section 5.5.1)
- Covert, bottom-up attention of salient feature detection (Section 5.5.3).

On the contrary, the inhibition of return behaviour is required to suppress object and salient items of either attended or putative object locations labelled in the spatio-temporal map (as described in Section 5.6). The specific visual task of this behaviour is thus:

- Inhibition of objects and salient items such that they stop being repeatedly detected and attended on successive pre-attentive and attentive cycles respectively.

Hence, the attentive behaviour employs generated data structures of the pre-attentive and inhibition of return behaviours to target the stereo-pair cameras towards found objects/salient items that are of interest to the system. This behaviour also performs recognition decisions in order to verify that the pre-attentively found object is indeed the recognised one. Similarly, the visual tasks conveyed by this behaviour are listed below:

- Saccadic camera movements and vergence towards objects or salient features and recognition.
- Global, non-selective vergence (0th layer as in Section 2.4).

It must be noted that the vergence behaviour is mainly an implementation of the hierarchy detailed in Section 2.4. Therefore, both vergence modalities (*global non-selective vergence* (0th layer) and *attended, selective vergence* (3rd layer)) are integrated within the attentive behaviour as described in Section 5.7. In this chapter, however, the *selective vergence* mode is extended in order to enable the system to verge on multiple same class objects present in the scene. This is further described while designing the attentive behaviour in Section 5.7.1. Additionally, as the camera actuation is only required while targeting object hypotheses, the communication with motors and their respective controlling signals are embedded in the binocular vergence behaviour in this hierarchy and, in consequence, the attentive behaviour.

5.4 Visual Representation Framework

The design of the hierarchy of behaviours is specifically inherent to the representation of the sensed visual information (i.e. SIFT features) in order to define and prototype the required visual behaviours. Hence, the visual representation in the hierarchical architecture context is formulated as follows.

Let d be the 128-dimensional SIFT descriptor vectors (expressed as $d = (d_1, d_2, \ldots, d_{128})$) after applying the SIFT algorithm (Lowe, 2004) on a captured image while executing a visual task. Similarly, there exists a corresponding SIFT feature location, scale and orientation values, (x, y, ρ, θ), that are logically indexed to a descriptor vector. Thus, the set of SIFT features for the left, L, and right, R, cameras are defined as:

$$\mathbf{L} = \begin{bmatrix} x_1^L & y_1^L & \sigma_1^L & \theta_1^L & (d_1)_1^L & (d_2)_1^L & \ldots & (d_{128})_1^L \\ x_2^L & y_2^L & \sigma_2^L & \theta_2^L & (d_1)_2^L & (d_2)_2^L & \ldots & (d_{128})_2^L \\ \vdots & & & & & & \ddots & \vdots \\ x_a^L & y_a^L & \sigma_a^L & \theta_a^L & (d_1)_a^L & (d_2)_a^L & \ldots & (d_{128})_a^L \end{bmatrix} \tag{5.1}$$

$$\mathbf{R} = \begin{bmatrix} x_1^R & y_1^R & \sigma_1^R & \theta_1^R & (d_1)_1^R & (d_2)_1^R & \ldots & (d_{128})_1^R \\ x_2^R & y_2^R & \sigma_2^R & \theta_2^R & (d_1)_2^R & (d_2)_2^R & \ldots & (d_{128})_2^R \\ \vdots & & & & & & \ddots & \vdots \\ x_b^R & y_b^R & \sigma_b^R & \theta_b^R & (d_1)_b^R & (d_2)_b^R & \ldots & (d_{128})_b^R \end{bmatrix} \tag{5.2}$$

where a and b ($a \neq b$) are the cardinalities of each set of SIFT descriptor vectors found in both left, L, and right, R, camera images respectively. Likewise, Equations 5.1 and 5.2 can be expressed in matrix notation as $\mathbf{L} = [l_{ij}]_{i=1,j=1}^{a,132}$ and $\mathbf{R} = [r_{ij}]_{i=1,j=1}^{a,132}$, respectively. The latter definition is thereby employed in the remainder of this thesis.

On the contrary, the set of SIFT features for all object images stored in a database is:

$$\mathbf{V} = [v_{ij}]_{i=1,j=1}^{c,132} \tag{5.3}$$

where v_{ij} has a similar definition as in Equations 5.1 and 5.2 and each v_i vector is logically indexed to an object image canonical view-point, and c ($a \neq b \neq c$) is the cardinality of SIFT descriptor vectors found for all the object view-point images. It must be noted that an object entry in the database is manually annotated and segmented in order to contain only the object of interest; thus, the resulting image size denotes the region of interest (i.e. bounding box) where the object's detail is contained.

Hence, the SIFT matched pair relation is denoted by $\sim$ (for simplicity, it is only considered the SIFT feature matched pairs between $\mathbf{L}$ and $\mathbf{V}$) :

$$d_p^L \sim d_q^V \Leftrightarrow \frac{D_E(d_p^L, d_q^V)}{D_E(d_p^L, d_r^V)} < \mathcal{T}_{SIFT} \tag{5.4}$$

where $D_E(d_p^L, d_q^V)$ and $D_E(d_p^L, d_r^V)$ are the Euclidean distances between the descriptor $d_p^L = [l_{pj}]_{j=5}^{132}$ and the first, $d_q^V = [v_{qj}]_{j=5}^{132}$, and second, $d_r^V = [v_{rj}]_{j=5}^{132}$, nearest neighbour, respectively; and $\mathcal{T}_{SIFT}$ is a threshold value for the log likelihood test. Lowe (2004) defines that $\mathcal{T}_{SIFT}$ must be within the range of 0.4 and 0.8 in order to obtain true positive matches. Finally, if the log likelihood ratio is below a given threshold value, a distinctive enough SIFT match is found; thereby, SIFT matches are denoted as:

$$\mathcal{M}_p^{(\mathbf{L},\mathbf{V})} = \begin{cases} \mathrm{q} & \text{If } d_p^L \sim d_q^V, \forall\, d_p^L \in \mathbf{L},\, d_q^V \in \mathbf{V}, \text{ holds true} \\ 0 & \text{Otherwise} \end{cases} \tag{5.5}$$

where $\mathcal{M}_p^{(\mathbf{L},\mathbf{V})}$ is a one dimensional array of the indexes between SIFT feature matches of $\mathbf{L}$ and $\mathbf{V}$ such that their cardinalities are $a = c = n$ and $p = 1, \ldots, n$ (n denotes the total number of correspondences between sets). This notation is employed in the remainder of this thesis to express similar relations for the other match sets, e.g. $\mathcal{M}^{(\mathbf{R},\mathbf{V})} : \ b = c = n$), $\mathcal{M}^{(\mathbf{L},\mathbf{R})} : \ a = b = n$, and so forth.

5.5 Pre-Attentive Behaviour

The pre-attentive behaviour is concerned with analysing the current field of view to detect and localise object instances and individual salient features not associated with any object grouping. The design rationale for the pre-attentive behaviour follows the *"stepping-stone" search pattern* discussed in Section 3.5.1 since an object or salient SIFT feature will only reach the attention of the system if it appears close enough to the current fixation and is of sufficiently high contrast. In that regard, this strategy therefore biases the robot vision system to discover large scale salient items in the periphery of the field of view of each observed fixation.

The pre-attentive search mechanism is therefore responsible for detecting SIFT features in the current field of view of each camera that are of interest to the visual search strategy and to *"notice"* objects and salient SIFT features only when they appear in the field of view of both cameras (as summarised in Figure 5.2). Similarly, this behaviour also binds a single object and salient item data structure of each corresponding found putative object instance and detected salient features respectively in both camera images. Hence, the pre-attentive behaviour is further divided in the hierarchy of visual behaviours into three different visual processing stream behaviours (as depicted in Figure 5.2). The operational function of each constituent behaviours are described in what follows.

5.5.1 Overt Attention - Multiple Same-Class Object Instance Detection

The *"Multiple Same-Class Object Instance"* (MSCIO) detector (as described in Chapter 4) affords the ability to covertly localise multiple same-class object instances in order to generate object hypotheses within the observed field of view. The working assumption is to reduce the granularity of the generalised Hough transform quantisation (Lowe, 2001) in order to allow the formation of multiple (same class) distinct peaks of projected SIFT features in a continuous Hough space representation. By applying an unsupervised clustering technique, it is possible to find natural groups in such space for each object instance present.

The implementation of the MSCIO detector in the pre-attentive behaviour is summarised as follows. SIFT features correspondences, the sets $\mathcal{M}^{(\mathbf{L},\mathbf{V})}$ and $\mathcal{M}^{(\mathbf{R},\mathbf{V})}$, are the input of this algorithm (as described in Section 4.3). For simplicity, the algorithmic steps of this behaviour are only described between right, $\mathbf{R}$, and database, $\mathbf{V}$, sets; however, the same process applies for the left camera (i.e. between $\mathbf{L}$ and $\mathbf{V}$ sets).

Thus, for each object matched in the database found after applying $\mathcal{M}^{(\mathbf{R},\mathbf{V})}$, the MSCIO

detector is invoked with $\mathbf{T} = [r_{bj}]_{j=1}^{4}$ and $\mathbf{M} = [v_{cj}]_{j=1}^{4}$ (where a and b contain the index entries in which a match exists such that $\mathbf{T} \subseteq \mathbf{R}$ and $\mathbf{V} \subseteq \mathbf{M}$, and their cardinalities are $b = c = n$ as discussed in Section 5.4), maximum number of instances, K (it is demonstrated in Sections 4.7 and 4.8 that a maximum of 6 object instances with high confidence score can be detected per each saccade in order to avoid false positive localisations), and a maximum perceptual overlap factor, α (it is set to 0.60 as determined experimentally in Section 4.8.2). The MSCIO detector then returns for each matched object in the database:

- the number of instances, K_{final}, detected,
- the Mean Square Error, MSE, for each instance $k = 1, \ldots, K_{\text{final}}$, of the projection error of SIFT features (as defined in Section 4.6), and,
- the object localisation in terms of the bounding box matrix, $\left(\mathbf{B}^R\right)_k$ for each instance $k = 1, \ldots, K_{\text{final}}$ (as defined in Section 4.6).

Thereafter, the fixation point in retinotopic coordinates (in pixels), $\mathcal{X}_k$, is the centre coordinate of each instance's bounding box (i.e. the mean value of the coordinate vertices of the bounding box) as follows:

$$\mathcal{X}_k^R = \left(\mu\left(\left(\left[B_{i1}^R\right]_{i=1}^{4}\right)_k\right), \mu\left(\left(\left[B_{i2}^R\right]_{i=1}^{4}\right)_k\right)\right), \text{ for } k = 1, \ldots, K_{\text{final}}. \tag{5.6}$$

Finally, each object instance found is stored into an object hypothesis set for the right camera, $\mathcal{H}^R$, with the following structure:

$$\mathcal{H}^R = \left\{\mathcal{I}_k^R, \mathcal{X}_k^R, \epsilon_k^R, \left(B^R\right)_k, \mathbf{F}_k^R, \mathcal{E}_k^R\right\}_{k=1}^{K_{\text{final}}} \tag{5.7}$$

where $\mathcal{I}_k^R$ denotes the numeric index of the object class in the database that produces the recorded object hypothesis, ϵ_k^R denotes the detection confidence such that $\epsilon_k^R = \left(\text{MSE}_k^R\right)^{-1}$, $\mathbf{F}_k^R = \mathbf{V}\left(\mathcal{M}_{\mathcal{I}_k}^{(\mathbf{R},\mathbf{V})}\right)$ (i.e. all matched SIFT features in the database that are indexed to the $\mathcal{I}_k$ object class); and, $\mathcal{E}_k^R$ stores the test and model SIFT features coordinates of the object instance (Equations (4.2) and (4.1), respectively) and their corresponding projected coordinates into Hough space (Equation (4.4)). It must be noted that $\mathcal{E}_k$ is employed for validation purposes and such is not used in the visual search task.

After MSCOI has found object hypotheses in both right and left camera images ($\mathcal{H}^L$ and $\mathcal{H}^R$) as shown in Figure 5.3, this behaviour then verifies that all found object instances are inside the current image field of view. That is, false object hypotheses can be detected when the object centres and the bounding boxes, given by $\mathcal{X}$ and B respectively for each camera, are located outside the current image field of view. The current image field of view is established

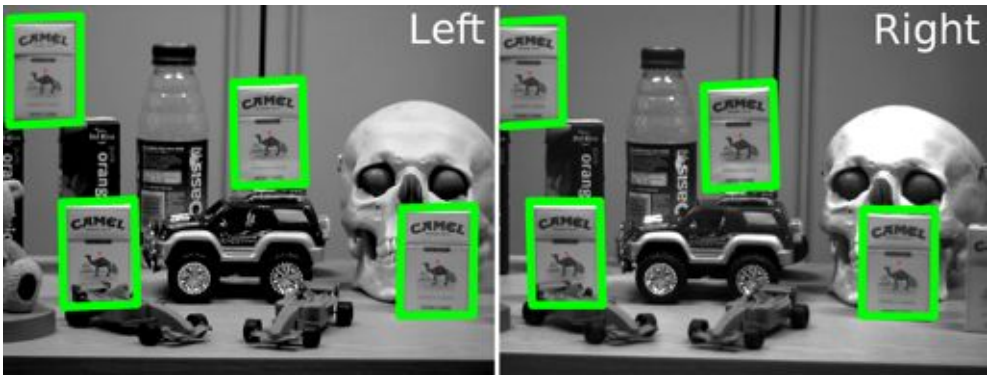


Figure 5.3: Multiple same-class object instances detection of one object class in the left and right camera images, respectively.

as 120% of the width and height of the image (in pixels) of each camera. This constrain sets the degree of occlusion of objects in terms of the limits of the field of view. Therefore, the following conditions must be satisfied: if the coordinates in B and $\mathcal{X}$ are smaller or greater than image boundaries, the object hypothesis is rejected; otherwise, it is accepted.

5.5.2 Hypotheses Generation

In order to provide a single set of new putative object hypotheses ($\mathcal{H}^{P}_{new}$), the pre-attentive behaviour must be able to locate correspondences between the found object instances in $\mathcal{H}^{L}$ and $\mathcal{H}^{R}$. The working hypothesis is that both cameras are currently verged and both cameras roughly observe the same portion of the scene (within a disparity tolerance). This assumption is independent of the layer of vergence currently active (Section 3.4). Hence, an object hypothesis is approximately the same in terms of its spatial location in both observed images if it belongs to the same object class and the Euclidean distance between the retinotopic fixation coordinates is minimal. That is, hypotheses, $H^{\mathcal{I}}_{0}$ and $H^{D_E}_{0}$, are formulated as:

$$H^{\mathcal{I}}_{0} = \begin{cases} 1 & \mathcal{I}^{L}_{i} = \mathcal{I}^{R}_{j}, \forall \mathcal{I}^{L}_{i} \in \mathcal{H}^{L}, \mathcal{I}^{R}_{j} \in \mathcal{H}^{R} \\ 0 & \text{Otherwise} \end{cases} \tag{5.8}$$

$$H^{D_E}_{0} = \begin{cases} 1 & \min\left(D_E\left(\mathcal{X}^{L}_{i}, \mathcal{X}^{R}_{j}\right)\right) < \mathcal{T}_{fixation}, \forall \mathcal{X}^{L}_{i} \in \mathcal{H}^{L}, \mathcal{X}^{R}_{j} \in \mathcal{H}^{R} \\ 0 & \text{Otherwise} \end{cases} \tag{5.9}$$

where $\mathcal{T}_{fixation}$ is defined as the maximum value of the two-dimensional standard deviation of the bounding box multiplied by the perception threshold defined in Section 4.4, i.e. $\mathcal{T}_{fixation} = \max\left(\left(\text{std}\left(\left(\left[B^{L}_{i1}\right]^{4}_{i=1,j=1}\right)_{k}\right), \text{std}\left(\left(\left[B^{L}_{i2}\right]^{4}_{i=1}\right)_{k}\right)\right) \cdot \alpha\right)$ for the surveyed kth object element

in set $\mathcal{H}^L$. Hence, an object in both cameras is considered to be the same if the following hypothesis rule holds true:

$$H_0 = \begin{cases} 1 & \text{If } H_0^{\mathcal{I}} \wedge H_0^{D_E} \text{ is true} \\ 0 & \text{Otherwise} \end{cases} \tag{5.10}$$

The last null hypothesis statement is valid if and only if the population size of H^L and H^R are identical. However, there are cases when they are unequal because MSCOI fails in the detection of objects (as discussed in Chapter 4). It is therefore denoted that $\mathcal{H}^P_{new}$ corresponds to the set with the maximum number of object hypothesis in the set. Thereby, the set with the missing object hypotheses is used to corroborate the putative objects of the latter; whilst the missing object instances hypothesis of the previous set are duplicated into the other camera. After having both sets equalised, the above algorithmic steps are carried out. If $\mathcal{H}^L$ and $\mathcal{H}^R$ are both empty, no object hypothesis is generated.

Thus, a single set of new putative objects is represented by the following structure:

$$\mathcal{H}^P_{new} = \left\{\mathcal{I}_i, \eta_i, \mathcal{Y}^L_i, \mathcal{Y}^R_j, \mathbf{F}, \mathbf{B}^L_i, \mathbf{B}^R_j, \mathcal{E}\right\} \tag{5.11}$$

where i and j denote the object elements when Equation 5.10 holds true, $\mathbf{F} = \left[\mathbf{F}^L_i \cup \mathbf{F}^R_j\right]$; $\mathcal{E} = \left\{\mathcal{E}^L_i, \mathcal{E}^R_j\right\}$ (see Section (5.5.1) and Equation (5.7)); η_i, the confidence score such that $\left(\max\left(\epsilon^L_i, \epsilon^R_j\right)\right)^{-1}$ (ref. Equation 5.7); and,

$$\mathbf{B}^L_i = \left[\left(\left[B^L_{p1}\right]^4_{p=1} - \left[c_x\right]^4_{p=1}\right)_i, \left(\left[B^L_{p2}\right]^4_{p=1} - \left[c_y\right]^4_{p=1}\right)_i\right] \tag{5.12}$$

and

$$\mathbf{B}^R_j = \left[\left(\left[B^R_{p1}\right]^4_{p=1} - \left[c_x\right]^4_{p=1}\right)_j, \left(\left[B^R_{p2}\right]^4_{p=1} - \left[c_y\right]^4_{p=1}\right)_j\right] \tag{5.13}$$

the translated bounding boxes of the ith and jth object elements in sets $\mathcal{H}^L$ and $\mathcal{H}^R$ with respect to the corresponding image coordinate centres, (c_x, c_y), of the stereo-pair (both cameras are assumed to have the same image resolution).

Similarly, $\mathcal{Y}^L_i$ and $\mathcal{Y}^R_j$ contain the translated retinotopic coordinates, $\mathcal{X}^L_i$ and $\mathcal{X}^R_j$ (see Equation 5.6), of the left and right cameras with respect to the coordinate centre of the image and the current camera position in an egocentric spatio-temporal pixel map. This map is contained within the variables $\mathcal{Y}^L_i$ and $\mathcal{Y}^R_j$ as it is stored the absolute position of the observed instance fixation point with respect to the recorded home position of the robotic system. Therefore, it is not required to explicitly create and store this map in working memory. Hence, $\mathcal{Y}^L_i$ and $\mathcal{Y}^R_j$

are defined (in accordance with Equation 5.6) as follows:

$$\mathcal{Y}_i^L = \left(\mathcal{X}_{i1}^L - c_x, \mathcal{X}_{i2}^L - c_y\right) \tag{5.14}$$

$$\mathcal{Y}_j^R = \left(\mathcal{X}_{j1}^R - c_x, \mathcal{X}_{j2}^R - c_y\right) \tag{5.15}$$

This egocentric spatio-temporal map further maintains a record of all visited locations of detected object instances in the scene that is later employed in the "inhibition of return" behaviour as described in Section 5.6.

Hence, $\mathcal{H}_{new}^P$ denotes the set of all putative objects found that is then passed to the pre-attentive behaviour.

5.5.3 Saliency Detection

As the cameras are only driven to look at object/salient locations, salient items are only registered if they appear in the field of view of the dominant eye (the left camera) when the cameras are targeting an object or salient item. That is, salient items are those SIFT feature locations in the dominant eye that $\neg\mathcal{M}_i^{(\mathbf{L},\mathbf{V})}$ and exhibit a saliency score above a threshold value (Equation 3.4).

The saliency score as computed in Equation 3.4 does not bias the system to attend salient locations over the periphery for the current saccade. In consequence, the "stepping-stone" search strategy presents a random behaviour as observed in the camera traces of Figure 3.19 in Section 3.8.4 (while validating the stepping-stone search strategy).

The proposed formulation of the saliency score herein presented (Equation 5.16) now includes the relative distance at which the observed salient item is located such that only those in the periphery (filtered by Equation 5.17) are employed within the visual search strategy. The saliency score is therefore the Euclidean distance from the image centre coordinate to the location of the salient SIFT feature weighted by the SIFT scale component. The saliency score, κ, is thus defined as:

$$\kappa_i = l_{i3}\left((l_{i1} - c_x)^2 + (l_{i2} - c_y)^2\right)^{1/2},\ i = 1, \ldots, n \tag{5.16}$$

where n is the population size of $\mathcal{M}_i^{(\mathbf{L},\mathbf{V})}$. Thereafter, the mean and standard deviation is evaluated of each ith element in κ and a SIFT feature that depicts a score above the following condition rule is retained:

$$\mathcal{S}_j = \begin{cases} 1 & If\ \kappa_j > \mu\left([\kappa_i]_{i=1}^{n}\right) + 3 \cdot \mathrm{std}\left([\kappa_i]_{i=1}^{n}\right),\ j = 1, \ldots, n \\ 0 & \text{Otherwise} \end{cases} \tag{5.17}$$

This conditional rule thus enables this behaviour to select those features that are located on the periphery and are sufficiently salient in terms of their SIFT scale component (i.e. $\sim 99\%$ of saliency scores are removed). Thus, if $\mathcal{S}_j$ holds true, the salient feature is appended to the new salient hypotheses set, $\mathcal{H}_{new}^{S}$, as:

$$\mathcal{H}_{new}^{S} = \left\{ f, \kappa_j, [l_{jk}]_{k=1}^{132}, \mathcal{Y}_j^{S_L}, \mathcal{Y}_j^{S_R} \right\} \tag{5.18}$$

where f is a variable that denotes if a salient entry has been attended: $f = 1$, otherwise $f = 0$; and, $[l_{ik}]_{k=1}^{132}$, the corresponding SIFT features of the current jth entry in $\mathcal{S}_j$. $\mathcal{Y}_j^{S_L}$ and $\mathcal{Y}_j^{S_R}$ are the projected retinotopic SIFT feature coordinate (for the left and right camera, respectively) of the jth salient item into the spatio-temporal pixel map defined in Section 5.5.2. If previous salient features have been detected in previous pre-attentive cycles, they are consecutively appended to $\mathcal{H}_{new}^{S}$. Thus, $\mathcal{H}_{new}^{S}$ denotes the output of the "*salient feature detection*" behaviour as depicted in Figure 5.2.

5.6 Inhibition of Return

During the progression of the visual exploration task, the binocular robot head might have detected other object hypotheses during previous pre-attentive cycles. In Chapter 3, a symbolic inhibition of return is devised where object classes are directly inhibited from the database of images by labelling the attended object class in the database as *"visited"* and therefore ignored in subsequent pre-attentive cycles. The new visual requirements of the robot head system, however, demands to inhibit same-class object instances of detected instances while exploring the scene. To that end, the egocentric spatio-temporal map (as discussed in Section 5.2) maintains a continuous record of the visited object/salient locations and their corresponding spatial region (i.e. denoted by bounding boxes) occupied in the observed scene.

Hence, the ability to inhibit multiple instances of each object class following each saccade of the visual search task, the Inhibition of Return (IOR) behaviour must determine whether detected instances have been either:

- identified objects that have been attended and verified, $\mathcal{H}^{A}$, as described in Section 5.7.1; or,

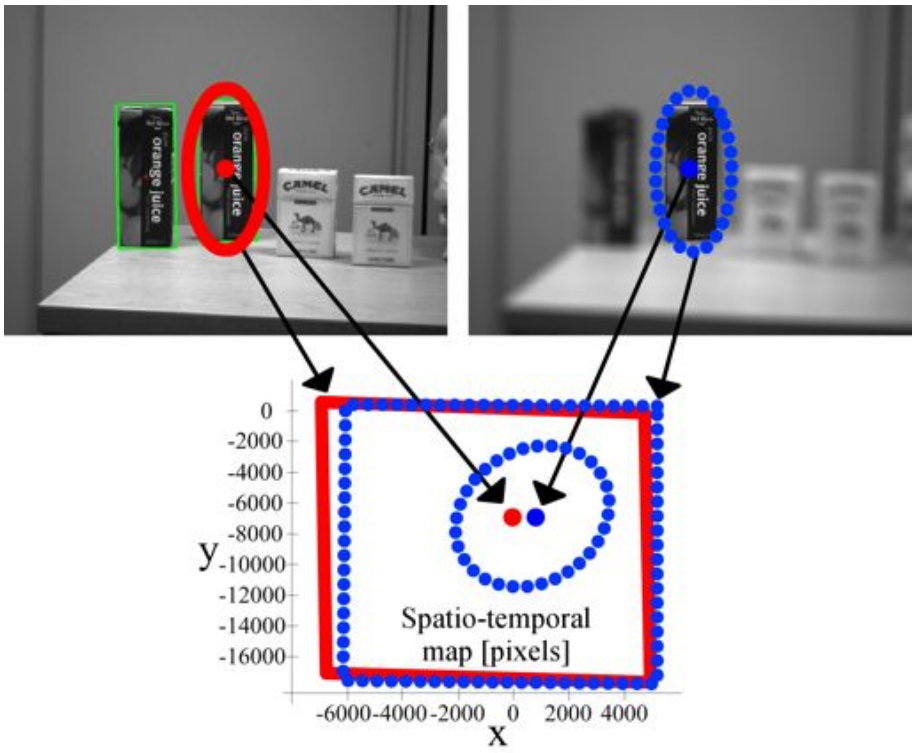


Figure 5.4: Example of the inhibition of return behaviour applied to the sets $\mathcal{H}^P_{new}$ and $\mathcal{H}^A$. X and Y axes depict the internal map of the system described and these are expressed in pixel units with respect to the home position of the cameras as described in Section 5.2.

- pre-attentively localised objects (the set of putative objects of previous saccades, $\mathcal{H}^P$, that have not been attended yet).

This behaviour is defined by means of a confidence ellipsoid test, where the critical assumption hypothesis is that an object in $\mathcal{H}^P_{new}$ is inhibited if and only if its fixation coordinate, $\mathcal{Y}^L$ (represented by points inside the drawn ellipse in Figure 5.4), falls inside the region confidence interval of either an attended object ($\mathcal{H}^A$) or an object hypothesis ($\mathcal{H}^P$) in the dominant eye (left camera). Figure 5.4 depicts an example of the inhibition process between $\mathcal{H}^P_{new}$ and $\mathcal{H}^A$.

For simplicity, the implementation of this behaviour is described for the inhibition process between $\mathcal{H}^P_{new}$ and $\mathcal{H}^P$ (current and previous putative objects). However, the exact steps are also valid while inhibiting between $\mathcal{H}^P$ and $\mathcal{H}^A$.

Thus, the algorithmic steps of this behaviour is explicitly modelled in the spatio-temporal pixel map domain by means of the statistical confidence ellipse test as follows, see Equations 5.14 and 5.15 (this test is only applied when the object belongs to the same object class):

$$y = \left(\mathcal{Y}^L_i\right)^P_{new} - \left(\mathcal{Y}^L_j\right)^P \tag{5.19}$$

$$\mathcal{K}_{ij} = yC^{-1}y^T : \mathcal{I}_i = \mathcal{I}_j, \forall \mathcal{I}_i^L \in \mathcal{H}_{new}^P, \mathcal{I}_j^R \in \mathcal{H}^P \tag{5.20}$$

where $\mathcal{K}_{ij}$ is a confidence factor determined by the χ^2 distribution of the ith and jth object elements in $\mathcal{H}_{new}^P$ and $\mathcal{H}^P$; and, C, the covariance matrix between bounding boxes ($\mathbf{B}^L$) of ith and jth object elements in $\mathcal{H}_{new}^P$ and $\mathcal{H}^P$; p and q denote the population size of $\mathcal{H}_{new}^P$ and $\mathcal{H}^P$, respectively. The null hypothesis is defined as the probability that the ith object in $\mathcal{H}_{new}^P$ appears in the interior ellipsoid of jth attended object (defined by the Equation 5.20) is equal to $P_{\chi^2}(\mathcal{K}, d)$. Each result of the null hypothesis is thereby stored in an array, $G = [g_{ij}]_{i=1,j=1}^{pq}$, as follows:

$$G = \begin{cases} 1 & 1 - P_{\chi^2}\left([\mathcal{K}_{ij}]_{i=1,j=1}^{pq}, d\right) > 0.1 \\ 0 & \text{Otherwise} \end{cases} \tag{5.21}$$

where P_{χ^2} is the probability of the χ^2 distribution; 0.1, the 90% of significance level of being the null hypothesis true; and d, the degrees of freedom which in this case is 2 as the visual coordinates are in the two dimensional image plane. Thereafter, G is then reduced to a column vector in order to determine the inhibited object of the ith element in $\mathcal{H}_{new}^P$, hence;

$$G' = [g_{i1}]_{i=1}^{p} \vee [g_{i2}]_{i=1}^{p} \vee \ldots \vee [g_{iq}]_{i=1}^{p} : g = \{0, 1\} \tag{5.22}$$

A pre-attentively observed object in the current saccade might be detected with a better confidence value than the one from previous pre-attentive cycles; therefore, the new object hypothesis presenting the highest confidence replaces the previous hypothesis. This special case is only valid while inhibiting objects in $\mathcal{H}_{new}^P$ and $\mathcal{H}^P$. Specifically, this case allows the system to correct detections and to suppress visual object information that might not contribute to the overall visual search task.

Finally, the symbolic inhibition of return mechanism described in Chapter 3 is adopted in this behaviour for the inhibition of salient features. The final result of inhibited sets thus consist of appending to $\mathcal{H}_{new}^P$ and $\mathcal{H}_{new}^S$ in $\mathcal{H}^P$ and $\mathcal{H}^S$, respectively. Therefore, $\mathcal{H}^P$ and $\mathcal{H}^S$ are the output of this behaviour.

5.7 Attentive Behaviour

As the attentional "spotlight" in the human visual system appears to select objects rather than locations (as discussed in Section 5.2), the critical assumption adopted for the visual search

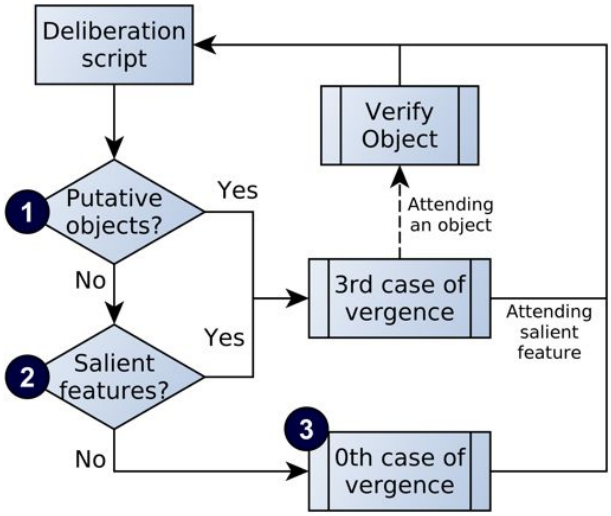


Figure 5.5: Flow diagram of the implemented attentive behaviour while searching a scene. The three described operational bases are marked accordingly (part of this diagram has appeared in (Aragon-Camarasa and Siebert, 2009).

strategy is to attend only those putative object locations that appear in the field of view of the stereo-pair camera configuration. As illustrated in Figure 5.2, the attentive behaviour is defined by a collection of primitive and abstract behaviours. An overview of the algorithmic implementation of the attentive behaviour is given in Figure 5.5.

Hence, the attentive behaviour employs the information provided by the pre-attentive and inhibition of return behaviours to target the cameras, to perform recognition decisions and, specifically, to bring the cameras into convergence in the 3rd layer of vergence (as illustrated in Figure 5.2). This behaviour selects *"where to look (attend) next"* (as well depicted in Figure 5.2) and the cameras are verged on the reported location. The design principles adopted in this behaviour follow those devised in the original robot vision framework (Section 3.5). The most confident object hypothesis or salient feature (whichever the case) is selected to drive attention based on the confidence score of $\mathcal{H}^P$ (Equation 5.11) or the saliency score of $\mathcal{H}^S$ (Equation 5.18).

As in Figure 5.5, the attentive mechanism operates under the following three bases (each basis is numerically labelled in Figure 5.3):

1. If $\mathcal{H}^P$ is not empty, the object hypothesis with the highest confidence score is then targeted, verged on (3rd layer of vergence) and verified.

2. If $\mathcal{H}^P$ is empty but $\mathcal{H}^S$ is not, the cameras are targeted and verged on (3rd layer of vergence) the salient feature with the highest score.

3. If $\mathcal{H}^P$ and $\mathcal{H}^S$ are empty, the cameras are maintained in convergence under the 0th layer

of vergence.

The third point is a special case; this occurs: a) when the formulated hypotheses and salient features have been attended and they are inhibited (Section 5.6), but the visual search task has not meet its halting criterion (as defined in Section 5.8); or b) when the robot vision system is initialising.

5.7.1 Saccadic Targeting

The saccadic targeting within the attentive behaviour (denoted as *"Where to look next?"* in Figure 5.2) consists of two functional modes:

- *top-down object based attention* ($\mathcal{H}^P$);
- *bottom-up salient based attention* ($\mathcal{H}^S$)

As discussed in the previous section, the attentive behaviour prioritises $\mathcal{H}^P$ as the most important visual information to be observed. Therefore, the second case only occurs when any object hypotheses are found in the pre-attentive cycle (as depicted in Figure 5.5).

Hence, when targeting to the most confident object hypothesis, it is crucial to verge the camera on the object of interest such that it is centred in the field of view of each camera (3rd layer of vergence, as illustrated in Figure 5.5). The location of the object hypothesis is known, since the object fixation coordinates in the spatio-temporal pixel map, $\mathcal{Y}^L$ and $\mathcal{Y}^R$, are obtained form $\mathcal{H}^P$ (Equation 5.11). Thus to ensure that the vergence system only verges on the desired object, those database SIFT features, $\mathbf{F}$ in Equation 5.11, that satisfy $\mathcal{M}^{(\mathbf{F},\mathbf{L})}$ and $\mathcal{M}^{(\mathbf{F},\mathbf{R})}$, are used in the disparity calculation, and such are only considered during the vergence cycle (Figure 3.4). Thereafter, the selected object hypothesis is verified as described in Section 5.7.2. If the system fails to target and verge on the current object hypothesis, this behaviour then removes the hypothesis form $\mathcal{H}^P$ and targets a salient feature as described below.

While attending to the most salient feature (based on the saliency score, s_j; Equation 5.18), it is only required to saccade to a single specific point of location in the scene; thereby, the 3rd layer of vergence operates without a defined target. Thus, all extracted SIFT features are considered in the disparity calculation and the dominant camera remains static while the other reduces the induced disparities. That is, both cameras fixates on the reported salient location and, in consequence, the dominant camera is not allowed to move during the vergence cycle. This allows the non-dominant camera to target the same location by means of the induced

Left Right

Figure 5.6: Segmented region of interest while verifying an object.

disparity errors. Afterwards, the targeted salient feature is thus marked as "attended" ($f = 1$ of Equation 5.18) and, in consequence, the behaviour returns the updated $\mathcal{H}^S$ set.

5.7.2 Object Verification

This behaviour is solely concerned with corroborating the selected ith object hypothesis in $\mathcal{H}^P$. As the system is enabled to localise multiple same class object instances, the object hypothesis attended can be close to or overlapped by other same-class instances. The bounding box coordinates are thus employed ($\mathbf{B}^L$ and $\mathbf{B}^R$ from Equation 5.11) to *"segment"* the region of interest in the current camera image pair and, in consequence, only those SIFT features inside both segmented regions are matched in order to verify the identity of the object. The system is thereby biased to focus only on the object of interest while possible false positives matches of same-class instances in the observed scene are ignored. Figure 5.6 illustrates a segmented region of interest while the robot investigates a "cigarette box" object contained in a scene.

Hence, an object hypothesis is verified if and only if the population size of either $\mathcal{M}^{(\mathbf{F},\mathbf{L})}$ or $\mathcal{M}^{(\mathbf{F},\mathbf{R})}$ (see Equation 5.11) with $\mathcal{T}_{SIFT} = 0.4$ are greater than 3 SIFT features matches. If an object is verified, the object hypothesis data is stored in the set of attended and identified objects, $\mathcal{H}^A$; otherwise, the object hypothesis is rejected and removed from $\mathcal{H}^P$.

5.8 Visual Search Strategy Definition - Macro Script

The deliberative layer (as illustrated in Figure 5.1) enables the system to perform cognitive operations in proportion to the designed behaviours in lower layers. As discussed in Section 5.2, deliberation is not considered in this thesis but this layer is included to enable the

Algorithm 5.1 Pseudo-code of macro script of the deliberative layer in Figure 5.1.

Inputs: None

Outputs: $\mathcal{H}^A$ with the same structure as in Equation 5.11 and as defined in Section 5.7.2.

$\mathbf{V} \leftarrow$Generate database in accordance with Equation 5.3
$mode \leftarrow$"explore" (as character string)
$[\mathbf{L}, \mathbf{R}] \leftarrow$Verge cameras under the 0th layer
$[\mathcal{H}^P, \mathcal{H}^S] \leftarrow \emptyset$
$[\mathcal{H}^P, \mathcal{H}^S] \leftarrow$Pre-attentive$(mode, \mathbf{L}, \mathbf{R}, \mathbf{V})$
$saccades \leftarrow 1$
WHILE ($\mathcal{H}^P$ or $\mathcal{H}^S$ are not empty;...
    or $saccades <$user-defined number)
  $[\mathbf{L}, \mathbf{R}, \mathcal{H}^P, \mathcal{H}^S, \mathcal{H}^A] \leftarrow$Attentive$(mode, \mathcal{H}^P, \mathcal{H}^S, \mathbf{V})$
  IF $\mathbf{L}$ and $\mathbf{R}$ are not empty
    $[\mathcal{H}^P_{new}, \mathcal{H}^S_{new}] \leftarrow$Pre-attentive$(mode, \mathbf{L}, \mathbf{R}, \mathbf{V})$
    $[\mathcal{H}^P, \mathcal{H}^S] \leftarrow$IOR$(mode, \mathcal{H}^P_{new}, \mathcal{H}^S_{new}, \mathcal{H}^P, \mathcal{H}^S, \mathcal{H}^A)$
  END IF
  $saccades \leftarrow saccades + 1$
END WHILE
Report objects stored in $\mathcal{H}^A$

development of reasoning and in future augmentations of the robot system herein described. Therefore, deliberation in this thesis is manually defined as a macro script that specifies the visual search task, controls and schedules behavioural resources, and monitors the progress of the task. Algorithm 5.1 depicts the macro script for the specific task of *active autonomous scene exploration of multiple same-class object instances.* This macro script can take several forms in order to allow the robot vision system to perform different visual tasks, as demonstrated in Chapter 6.

5.9 Experiments

The operation of the active binocular robot vision architecture is predicated on the design principles described in Section 5.2 which embeds the following critical requirement: *the hierarchy of visual behaviours linked with the deliberative layer in the overall architecture shall be capable of localising, identifying and reporting known object instances present within a scene containing unknown and known multiple same class object instances. In addition, the scene configuration may contain clutter and also overlapped and occluded objects*. The proposed architecture is however not fully demonstrated until its utility is applied in a different task specification or application domain. That is, the design of a different high level behaviour, which validates its applicability and robustness of the architecture, is described in Chapter6.

Hence, the devised architecture is thus validated within the specific application domain to ini-

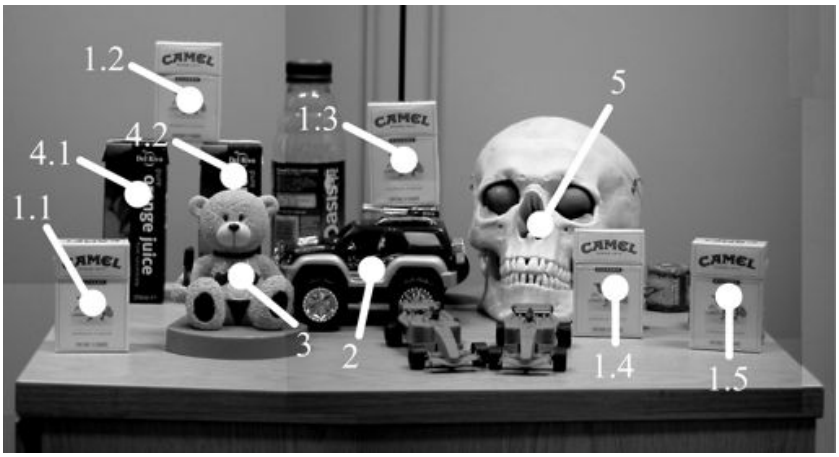


Figure 5.7: Reconstructed scene used in the experiments and objects labels. (Aragon-Camarasa and Siebert, 2009)

tially characterise its performance and, specifically, the visual search stability while using the designed hierarchy of visual behaviours. To that end, a scene is configured such that all identifiable object instances are mixed with unknown objects in a complex cluttered setting. Ten objects in total featuring five different object classes (box cigarette, skull, car, juice box and bear; numerically labelled as observed in Figure 5.7) are arranged in the scene. From a given initial home actuator position, the active binocular robot head is thus allowed to search for known objects previously defined in the database. It must be noted that the task is contextually defined within the visual search strategy designed in Algorithm 5.1: *the active binocular robot head must find the objects requested in database using the visual search strategy*.

As discussed in Section 5.3.1, the vergence behaviour within the architecture is a direct implementation of the hierarchy detailed in Section 2.4 and as such it is therefore not considered in these experiments because it has been extensively validated in Chapter 3. Nevertheless, the correct operation of the 3rd layer (*attended, selective*) of vergence is demonstrated, as the visual search task in this chapter now features multiple same-class object instances detection and, therefore, an attended object must be in the centre of the field of view in both camera eyes when the robot is verifying it. The reliability of the designed inhibition of return behaviour to suppress incoming visual information during the visual search strategy adopted is tested similarly.

The experimental methodology thus consists of invoking five visual searches in order to measure:

- the RMS fixation error to characterise the system's repeatability, and,
- the correlation between all combinations of pairs of searches in terms of the attended locations.

On the one hand, the former shares similarities with the experimental validation in Section 3.8.3; however, it is now verified under the task of attending multiple same-class objects where clutter might affect the ability of the search strategy to attend same-class instances within the scene. The RMS fixation error is measured with respect to the mean fixation value of each instance in the scene for all the experiments (see Section 3.8.3). On the other hand, the latter is based on the validation framework proposed in (Ouerhani et al., 2004) that objectively measures the correlation among different fixations maps and subjectively compares these maps. A fixation map is created for each visual search experiment. This map is a transformed image space from the recorded locations of the egocentric spatio-temporal map described in this chapter (as illustrated in Figure 5.8). Therefore, this experiment tests the following hypothesis: *if a hierarchy of visual behaviours is adopted in order to search a cluttered scene with multiple same object class instances, the designed robot vision architecture will attend locations whose positions correlate spatially above chance across different trials, as experimentally observed in previous research (Ouerhani et al., 2004).*

To construct each fixation map, a 2D Gaussian weighting function is thus computed for each fixation point as follows

$$Map(x, y) = \lambda e^{-\left(\frac{(x-\mathcal{Y}_x)^2+(y-\mathcal{Y}_y)^2}{\sigma^2}\right)} \tag{5.23}$$

where $\mathcal{Y}_x^{\boldsymbol{u}}$ and $\mathcal{Y}_y^{\boldsymbol{u}}$ are the x and y fixation coordinates such that $\boldsymbol{u}$ is either the left, L, or right, R, camera fixation point in set $\mathcal{H}^A$; x and y, given in pixels, define the coordinate space of the fixation point; λ specifies the importance of each fixation point (this is set to $\lambda = 1$, since all fixations share the same importance in the visual search task in this chapter). In order to objectively measure the spatial correlation between attended and verified object fixations with respect to the spatial localisation on the image space, an acceptable fixed standard deviation (σ) of $\sim$ 20 (pixels) is adopted. This corresponds to 2% of the image dimensions in each saccade.

5.9.1 Exploration of Multiple Same-Class Object Instance

Figure 5.8 depicts the object positions (bounding boxes), fixation points, and the camera saccades/traces for both cameras in image space. A square denotes a fixation point, circular dot represents a salient item and an upward and downward pointing triangles indicate the initial and final position of the search process.

By inspecting the five visual searches (50 object instance observations) in both cameras as shown in Figure 5.8, it is stated that the ability to recognise same class object instances in a

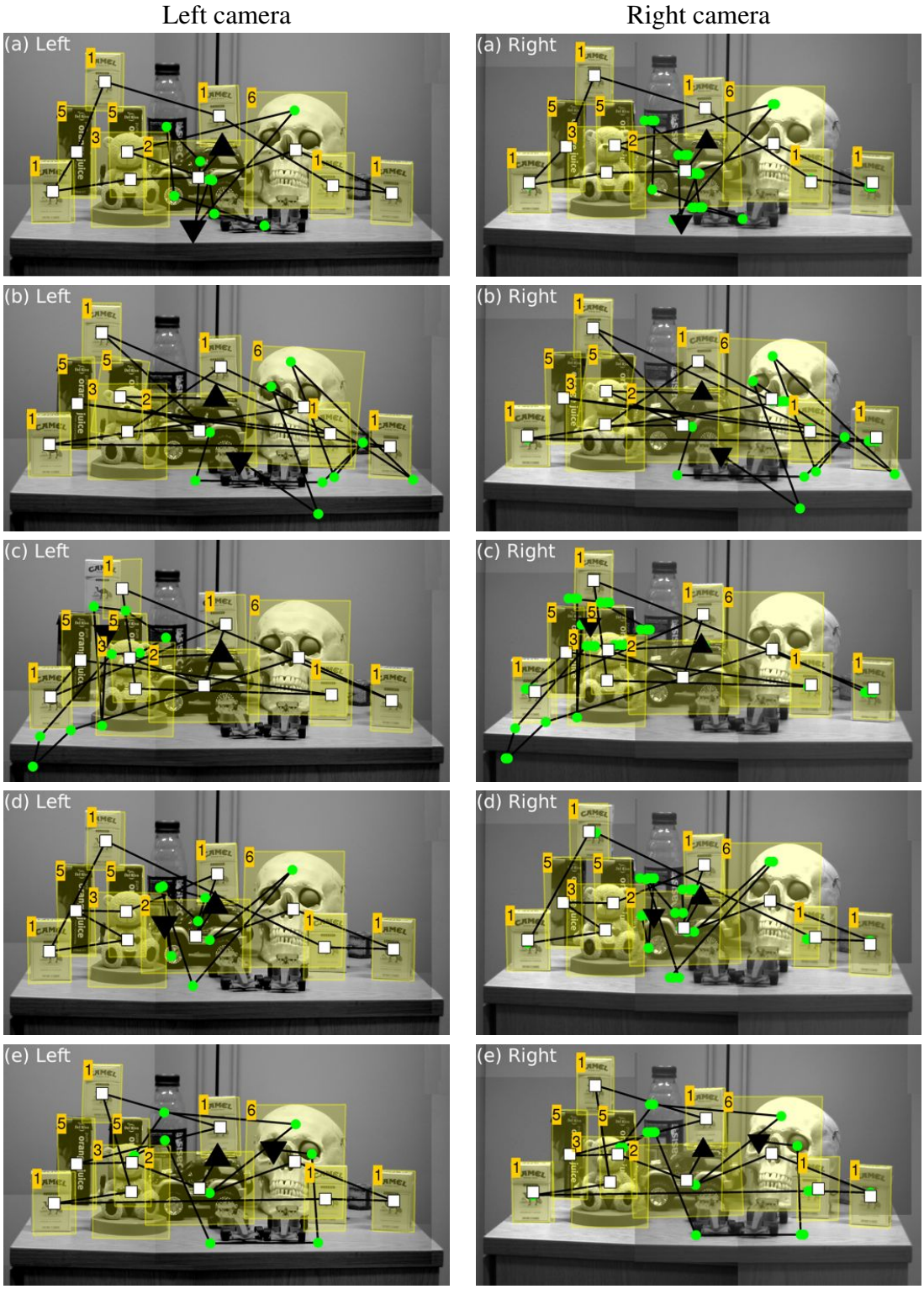


Figure 5.8: Camera traces of the left and right camera approximately overlaid in the scene employed for: (a) experiment 1, (b) experiment 2, (c) experiment 3, (d) experiment 4, and (e) experiment 5.

cluttered background has been initially demonstrated, since all identified objects are successfully localised within the first eleven saccades performed in each trial. Overall, the investigated robot vision system performed less than 15 fixations to identify all requested objects.

Figure 5.8 demonstrates the ability of the system to perform autonomous scene exploration. The average computational cost to carry out the visual search was ~ 26 min. Similarly, the inhibition of return mechanism behaviour is demonstrated since the system avoids revisiting the same object or salient region that has already been attended, as in Figure 5.8 (circular dots in the figure).

Table 5.1: RMS error incurred on the left camera on the X and Y axes given in pixels.

	Object	1.1	1.2	1.3	1.4	1.5	2	3	4.1	4.2	5
Left	X axis	10.5	26.5	3.7	10.4	13.5	9.4	8.4	11.8	9.2	7.2
Camera	Y axis	4.0	1.9	5.4	2.6	5.5	3.0	4.5	2.2	2.9	2.1
Right	X axis	8.0	4.8	12.8	10.5	8.1	3.6	2.3	13.9	3.8	3.4
Camera	Y axis	5.6	2.2	2.0	1.4	6.2	2.6	4.5	10.1	2.7	4.0

The *attended, selective vergence* mechanism is able to maintain vergence while the gaze is simultaneously directed towards the object being attended. Since correct object localisations are evident in both image pairs, Figure 5.8 confirms that the system successfully verges on the putative objects where close enough object instances could distract the attention of the system (i.e. the orange box objects). The RMS errors of attended object localisations that this system incurred is given in Table 5.1. The worst observed error is ~ 26.5. In this case, it is inferred that the determined fixation point in both cameras is distant from the inaccurately located instance, and the object is therefore not accurately localised (object 1.2 on Figure 5.8); however, this error only represents 2.5% of the employed image size (1024×768 pixels).

5.9.2 Visual Search Stability

The above RMS errors indicate satisfactory detection and localisation of known objects; however, it is required to obtain a deeper insight into the visual search behaviour exhibited by the hierarchy of behaviours and, specifically, determine its stability.

Hence, from Figure 5.8 and Table 5.1, it can be depicted that objects 1.2 and 5 exhibit the largest fixation deviations of all observed object instances. Specifically, object 1.2 is the most difficult object to localise since the fixation point (as illustrated in Figure 5.8(c): Left camera) is on the object edge despite its visual features being visible (as delimited by the bounding box). Furthermore, Figure 5.8(c) (left and right camera) shows that the 3rd layer of vergence fails to centre correctly the object. Conversely, object 4.2 is correctly identified even though

Figure 5.9: Attentional map of fixated object instances for the five visual searches in this experiment. Aragon-Camarasa and Siebert (2009)

it is approximately 60% occluded. That is, this object's fixation point is exactly located on the head of the toy bear (object 3). The latter demonstrates that the hierarchy of behaviours is robust in its behaviour capabilities but it is constrained the degree of invariance of SIFT features and, in consequence, to the canonical view-point stored in the database (this is further discussed in Section 5.10).

By carefully inspecting the visual scan paths, it is observed that the camera traces do not share corresponding fixation point locations. To investigate this phenomenon, only those fixation points of attended and verified objects are merged in order to produce a graphical representation of the object localisation deviation incurred by our binocular robot vision architecture, Figure 5.9 (black circles indicate the attended and verified object centre locations).

The localisation repeatability incurred is thus measured by the mean $(\bar{x},\, \bar{y})$ and standard deviations $(\sigma_x,\, \sigma_y)$ of the fixation coordinates for each object instance (Table 5.2). Similarly, the mean and standard deviation ($\bar{\rho}$ and σ_ρ, respectively) of the correlation coefficients for each fixation point belonging to the object instance are computed. This indicates the spatial correlation and the error induced, respectively. Hence, 100 correlation coefficients are computed for all possible combinations between fixation points ($10(C_2^5)$ since 10 "known" objects are present in the scene).

Hence, Table 5.2 further reveals that object 1.2 exhibits the highest deviation error over the horizontal axis ($\sigma_x = 26.8$) and, in consequence its correlation coefficient is the lowest ($\bar{\rho} = 0.70$) for the attended objects. This result is consistent with the one found in Table 5.1 of Section 5.9.1. It is thus deduced that the object view-point representation stored in the database does not provide sufficient SIFT features matches to locate and characterise it correctly. Nevertheless, this figure represents $\sim 9.3\%$ of error with respect to the object area

Table 5.2: Visual search stability statistics.

Object	$\bar{x}$	$\bar{y}$	σ_x	σ_y	$\bar{\rho}$	σ_ρ
1.1	188.5	583.4	11.3	4.3	0.91	0.07
1.2	393.8	181.1	26.8	2.1	0.70	0.27
1.3	808.6	305.9	11.4	2.6	0.92	0.08
1.4	1204.3	563.6	6.7	10.4	0.91	0.07
1.5	1435.9	585.3	13.7	7.5	0.83	0.10
2	730.7	532.6	10.8	3.5	0.92	0.07
3	478.5	542	8.4	4.7	0.94	0.06
4.1	284.8	438.5	11.7	2.8	0.91	0.07
4.2	466.4	431.6	14.1	10.2	0.83	0.12
5	1088.9	435.2	18.2	8.8	0.80	0.21
Overall	$\cdots$	$\cdots$	13.35	5.72	0.87	0.11

Figure 5.10: Attended salient locations illustrating that the visual search strategy inspects the entire scene over all visual search experiments conducted (Aragon-Camarasa and Siebert, 2009).

in the image (stored in the database with a size of 288×432 pixels).

All fixation points associated with salient items of all visual searches invoked are finally merged, as depicted in Figure 5.10. This fixation map indicates that there is not a visible spatial correlation among such fixations since there is no indication of a noticeable pattern in the camera traces of each trial (Figure 5.8). However, the defined visual search strategy inspects the entire scene across each of the visual search trials as illustrated. That is, salient saccades attend different visual regions of the scene, and such contribute to the overall progress of visual search task.

5.10 Summary and Discussion

This chapter presented the development of an active binocular robot head architecture that integrates reactive behaviours in a hierarchical, parsimonious manner. The designed robot

Figure 5.11: Verified object after attending a salient item (circles denote object locations while squares indicate salient keypoints) (cfr. Figure 5.8(a)) (Aragon-Camarasa and Siebert, 2009).

vision architecture was capable of autonomous scene exploration, with the specific task of identifying and localising same-class object instances while maintaining vergence and directing the system's gaze towards scene regions and objects in cluttered, complex settings. Likewise, the new algorithms and the devised architecture had considerably enhanced the capabilities of the investigated robot vision system, which now outperforms favourably state-of-the-art systems (Björkman and Eklundh, 2004, 2005a; Kragic et al., 2005; Das et al., 2008; Aragon-Camarasa et al., 2010) (as reviewed in Sections 2.1, 2.5 and 2.6).

Section 5.9.1 demonstrated the correct operation of the system when exploring a scene containing unknown and known multiple same-class object instances (Figure 5.8). Close inspection of Figure 5.8 confirmed that the hierarchical robot vision architecture was capable of localising, identifying, and reporting all objects despite occlusion. Results showed that the robot head system identified and localised all requested objects in a cluttered and occluded scene. The mean and standard deviation of all RMS errors of fixations performed towards known objects were ~ 6.4 and ~ 2.02 pixels, respectively (Table 5.1), over 5 experimental trials locating a total of 50 object instances. Similarly, an analysis of this visual search process revealed that object instances could be reliably localised with worst observed errors of: ~ 13.35 pixels standard deviation in the spatial image frame (Table 5.2). In terms of search stability, object instance fixations were demonstrated to be stable, producing an average spatial correlation repeatability value of ~ 0.87 with a maximum of 0.27 spatial correlation variation being observed over the 5 trials.

Close inspection of Figure 5.10 revealed that object fixations were repeatable, whereas, saccadic fixations towards isolated salient keypoints were not. This system's behaviour substantiated the the active vision paradigm: for each fixation performed, the system had the opportunity to notice something that previous pre-attentive cycles had missed. For instance, in Figure 5.11, the system detected and localised object 2 after fixating a salient feature. Accordingly, the visual search properties observed above loosely resembled to those found in biological systems (e.g. Styles (2005) and Yarbus (1967)).

Although the experimental validation in this chapter depicted favourable results, further characterisation of the devised behaviours was required while performing several trials over a variety of different types of scene (and observed objects) in order to characterise the hierarchical architecture reliably. This extended validation will be addressed in Chapter 7.

Hence, it is envisaged that the adoption of the devised hierarchical architecture, a cognitive active robot vision system could be designed in order to generate (by reasoning) the sequence of behaviours that must be triggered in order to achieve a specified goal. Therefore, the deliberative macro script described in Algorithm 5.1 might be replaced by a possible high-level cognitive machine which:

- translates user input into robot commands (e.g. natural language programming, for example as in Veres (2008)),
- allocates available resources to fulfil the task-goal specification and
- evaluate the performance while executing the task and recover from system failures.

Finally, when the cameras target objects closed to the centre of the scene or the object's canonical corresponded to the one stored in the object knowledge database, its was higher than that object which observed a different different canonical view-point (e.g. objects 1.2, 1.5 and 5, Figure 5.9). This was due to the fact that captured canonical poses of such objects in database did not adequately represent their appearance over the viewing sphere. It would be required to capture a greater range of canonical poses in order to provide the pre-attentive behaviour with enough visual knowledge to successfully detect and locate objects over different poses. The above, however, implied that the object database increased in size and, in consequence, SIFT matching related operations became computationally expensive. In that respect, the forthcoming chapter describes a visual object appearance learning behaviour that further improves object knowledge database by enabling the investigated robot vision architecture to create its own visual knowledge.

Chapter 6

Learning the Appearance of Objects

As observed in the previous chapters, the objects' canonical representations in manually segmented databases give rise to false positive detection while exploring a scene. This is because captured canonical poses of these objects do not adequately characterise their appearance over the view-sphere; therefore, a greater range of canonical poses must be acquired in order to enable the pre-attentive and attentive behaviours to successfully locate and recognise the object. To that end, this chapter introduces an active visual learning behaviour that allows the investigated robot vision system to automatically synthesise, create and build its own part-based object representation knowledge from multiple observations while a human teacher indicates the object and supplies a classification name. The devised behaviour permits to actively explore an object view-sphere by loosely simulating what higher mammals do. The computational concepts of a visual learning exploration mechanism are adopted to control the accumulation of spatio-temporal visual evidence and direct attention towards representative object parts. Therefore, the devised visual learning behaviour introduces a structured and parsimonious computational model that has not been yet reported in the robot vision community.

6.1 Introduction

Scientific studies about the human brain have established that recognition of common 3-dimensional objects is carried out by highly tuned two-dimensional views of the imaged object. That is, object learning for recognition (in accordance with the exemplar-based object recognition paradigm described in Section 2.3) consists of two-dimensional snapshots of the

most representative views across the object's view-sphere[1]. Specifically, this visual learning behaviour is loosely based on the WHAT and WHERE streams as depicted in Figure 5.1.

The defined visual behaviours are therefore structured in terms of the hierarchical architecture described in Chapter 5. Such visual behaviours are extended/designed in accordance with a new high-level task-goal specification: *semi-autonomous object appearance learning*. Additionally, this new visual behaviours extend the application domain of the devised robot vision architecture. As in the hierarchical architecture, SIFT features are the adopted visual representation for all visual abilities and thus enable the robotic head to learn the objects' appearance.

Hence, the purpose of an object appearance learning behaviour[2] is twofold:

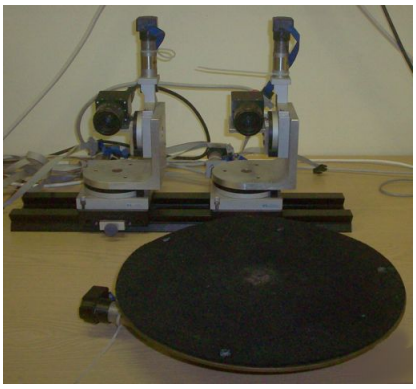
- To validate the applicability and robustness of the devised robot vision architecture in a different visual task.
- To create and build object representations by means of the active visual interaction with the object (i.e. an actuated turn-table) in order to improve recognition tasks.

The remainder of this chapter is organised as follows: Section 6.3 introduces the founding concepts of the visual learning behaviour and, in consequence, Section 6.2 outlines hypotheses and objectives of such behaviour. The design rationale and methodology is presented in Section 6.4. Whilst Sections 6.5, 6.6 and 6.8 discuss the modifications and extensions carried out on the proposed active binocular robot vision hierarchical architecture. Section 6.9 presents the conclusions of this chapter.

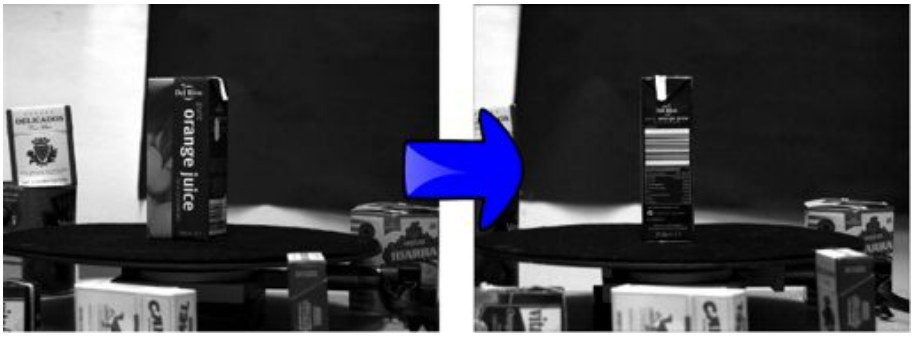
6.1.1 Binocular Robot Head Adjustments

In order to enable the robot vision system to synthesise the acquired visual information by means of the active interaction with objects and the environment, the system in this chapter features an actuated turn-table (as depicted in Figure 6.1(a)). The actuated turn-table enables the robot vision system to manipulate an object over different spatial configurations. With this turn-table, however, the object is only revolved with respect to the rotation axis parallel to the image image plane (as illustrated in Figure 6.1(b)). Therefore, this angular transformation resembles similar movements if the object has been placed in a human's palm.

[1] In the context of this thesis, the *object's view-sphere* relates those possible spatial positions and view points the robot vision system can observe in an object.
[2] In the remainder of this thesis, this behaviour is termed as: *visual learning behaviour*.



(a)



(b)

Figure 6.1: (a) Robot head featuring the actuated turn-table, (b) The turn-table employed to explore an object across the view-sphere.

During the design of the visual learning behaviour, the stereo-pair cameras of the robot head employed thus far failed and, in consequence, where replaced by: two Prosilica cameras: GC2450C and GC2450 (colour and mono respectively at 5 mega pixels of resolution) fitted with Gigabit Ethernet interfaces. Due to the new hardware requirements, low-level components (Figure 5.1) were interfaced to the Pentium 4 computer described in Section 3.2.1. Mid- and high-level components were interfaced to a 4-core Intel Xeon (model E5502) with a CPU clock speed of 2 GHz, with 24 GB in RAM running under Windows XP and MATLAB R2009. As both computers were located in different places, they were interconnected through the local network by means of a collection of network socket functions for use in MATLAB[3].

6.2 Motivation

As the key objective is to enable a binocular vision robot head to actively and automatically synthesise its own object knowledge, the robotic system must be capable of creating and building its own appearance representations by means of the active visual interaction with the object (i.e. an actuated turn-table). To that end, visual learning as a behaviour can take the

[3] `http://code.google.com/p/msocket/` (verified on the 30th September, 2011)

Figure 6.2: An infant exploring an object. The dynamic interaction and exploration with the object enables the infant to actively investigate what the object looks like from different viewpoints.

following modalities (as similarly defined in Hyundo et al. (2006)):

- *Semi-autonomous learning*: This is closely related on how biological organisms learn to recognise objects by "exploring" them (as illustrated in Figure 6.2), or by a "teacher" who defines what the observed object is (i.e. teacher-student protocol).
- *Fully autonomous learning*: This mode relates how biological organisms learn by means of situations, experiences or interactions with the environment in an unsupervised way. Robot applications in this context can be found in Modayil and Kuipers (2008).

It must be pointed out that humans bind different sensory features (i.e. haptic, visual, auditory and so forth) to learn an object concept. In this thesis, visual features are only considered. Thus, the underlying task of the robotic system is to find and locate an indicated object (by a human teacher) and build an egocentric internal map of the relative position of object parts for topological abstractions of the object concept. The proposed visual learning behaviour is therefore modelled as an *exemplar-based, semi-autonomous learning robot vision system*. That is, by actively exploring the object view-sphere (as high-order mammals do, see Figure 6.2), the robot vision system automatically synthesises, creates, and builds its own part-based object appearance knowledge (i.e. it automatically finds and selects attentional shrouds (Fazl et al., 2009)) from multiple observations while a human teacher indicates the object and supplies a classification name. The robot thus interacts closely with the object and the environment by means of an actuated turn-table (e.g. perception-action cycle). Accordingly, the robot does not rely on an off-line classifier to train the system but employs an active unsupervised clustering approach to group visual features that maintain spatio-temporal relationships and topological connections across the view-sphere (i.e. object parts are synthesised as a unique canonical 2D snapshot for a particular viewing angle).

The working assumption is defined as: *a robot is capable of directing and holding its gaze towards an unknown but interesting attentional shrouds (i.e. object parts in this chapter) while an exploratory, learning, behaviour is triggered to capture distinctive features of object-parts over a set of sampling angular positions, and, eventually, those visual features constitute canonical, object-based, representations of the imaged object.*

Thus, *how could a binocular robot vision system create its own canonical object representations from different viewpoints while inspecting an object with camera saccadic movements in a semi-supervised manner?* And, *what are the underlying principles of attention and camera saccadic movements, and computer vision techniques to enable object learn*ing *in a robot vision context?*

To characterise the above questions, a set of learning design principles must be defined (inspired by Weng et al. (1993)):

- The robot should be able to automatically learn rules that a designer might not manually define. Learning is not constrained to a fixed set of parameters or/and rules such that the robot knowledge is scalable to solve complex problems (e.g. to recognise object instances in the environment, to perform active selection of objects of interest in an exploration task, to name but a few).
- Consequently, the structure of the knowledge should be automatic, object pattern structures have to be automatically identified and their decomposition and projection through a hierarchy of pattern structures must be unsupervised (the designer may only offer a label for the overall concept of the object).
- The object of interest must be automatically segmented from background.
- Object views are minimised and decomposed into simpler representations (i.e. feature description and canonical 2D representations).
- The subsequent recognition rate achieves a specific level of representation for any intermediate view without relying on a mental rotation approach (i.e. model-based object recognition as described in Section 2.3).
- Different viewpoints on the object are sampled automatically (it saccades to the most structural representative object part that contributes to the learning process).
- Learned features are classified as new input features without forgetting the stored ones. It is incremental and scalable.
- Exploration and discovery of novel object views is performed automatically and actively.

The overall visual learning behaviour is outlined in the following section.

6.3 Visual Learning Concepts/Principles

As discussed in previous chapters, the research community has agreed that attention-selection mechanisms are contained into two different but interacting stages: a *pre-attentive* and *attentive stage*. They operate closely one after the other in a closed-loop cycle. Both stages control the information flow that happens to be in the *ventral* and *dorsal stream* of the human brain (ref. Figure 5.1). Therefore, the behavioural learning structure is inspired by the WHAT and WHERE visual streams. This implements the devised visual learning behaviour into the context of the hierarchical robot vision architecture described in Chapter 5.

Specifically, the WHERE stream maintains and binds spatial attention on an object coordinates that egocentrically characterises the location of the object-part of interest and creates the spatio-temporal relationships (i.e. topological relations) of feature coordinates. Whereas the WHAT stream either determines the identity of an object or triggers a learning behaviour that stores the episodic view-invariant feature descriptions of the object.

Episodic view-invariant descriptions are commonly termed in the literature as canonical views. Therefore, a canonical view is thus defined (for the purpose of this chapter and according to Blanz et al. (1996); Peters et al. (2002)) as a "*stable view that provides highly distinctive SIFT features over the view-sphere and enables a viewpoint invariant representation of an object for recognition*". In an active vision context, the dynamic exploration of an object (i.e. camera-eye movements) improves the robustness of finding canonical object parts. That is, the selection of the two-dimensional canonical view is based on the most informative view for recognition purposes. Thus, canonical representations become stable by gaining visual information while saccading to structural representative object parts (i.e. active exploration of the object) and storing and clustering those features that present spatio-temporal continuity over a set of viewing poses and saccades.

Attention, according to the '*spotlight of attention*' metaphor, is deployed into objects rather than locations in space (Posner and Petersen, 1990; Chun and Wolfe, 2004; Fazl et al., 2009; Wallraven and Bülthoff, 2007b). This leads to assume that object-based attention is controlled by factors such as perceptual grouping and completion operations (i.e. Gestalt's Laws (Yu et al., 2010)) which, latterly, reconstruct the imaged scene structure (as opposed to the model-based system described in section 2.3). Thus, in this chapter, the distribution of spatial attention over objects comes in a form-fitting procedure according to the object surface/shape termed as *attentional shrouds* (Fazl et al., 2009). Therefore, the *attentional shrouds* concept

plays an important role on the proposed behaviour as it is the driver to direct and control attention towards the object parts across the view-sphere. The system thus pre-attentively fits attentional shrouds into potential object parts and, subsequently, attend those putative object components.

Learning object parts, in this chapter, therefore consists of two-dimensional canonical representations that compress the object appearance in terms of SIFT descriptors. Object parts are defined as broken up snapshots that conserve the distinctiveness of an object observed in a particular period of time (Lehrer and Bianco, 2000; Styles, 2005) (i.e. spatio-temporal continuity). In order to pre-attentively segment, select and create putative attentional shrouds, the learning behaviour employs SIFT features to compute depth and motion information about the structure of the scene. This enables the system to segment an object from background and control the gaze towards the attentional shroud of interest. The latter has biological foundations; for instance, bees employ depth and motion to create landmarks of the object spatial configuration in a "turn-back-and-look" manner (Lehrer and Bianco, 2000) (in this paper, the authors present a biologically motivated mobile robot that exploits motion and depth visual cues for robot navigation and exploration).

The canonical view selection has biological foundations. For instance, Ullman et al. (2002) have showed that individual neurones in the human brain (specifically in the *IT cortex*) are selectively tuned by measuring the *mutual information* between trained and input visual features of object part fragments over a set of sample views. It must be noted, that, in the investigated robot vision system context, IT neurones are specifically represented by means of SIFT feature descriptions.

On the contrary, binocular gaze holding (i.e. smooth pursuit) enables a binocular vision system to focus on a target while clutter is ignored. This assertion has been successfully demonstrated over the literature in robotic systems (Coombs (1992); Bernardino and Santos-Victor (1998)). To date, there exist several approaches that tracks a moving object within dynamic and complex worlds. The most widely used algorithm in the robotic context is the *pyramidal Lucas-Kanade tracking method* devised by Bouguet (2002). This algorithm consists of a point-based computation (i.e. commonly Harris features, Section 2.2.1) across a pyramidal image representation between two images in order to calculate the induced velocity components. Additionally, this pyramidal image representation allows to find feature points that observe large displacements between images.

In the context of the hierarchical architecture and the visual learning behaviour described herein, binocular gaze holding is casted by means of the pre-attentive and attentive behaviours (i.e. perception-action cycle). Therefore, there must exist a behaviour that pre-attentively computes the induced optical flow and binds together the spatial locations of the moving fea-

tures, and, in consequence, a behaviour that targets the object in motion. The interaction of both abstract behaviours loosely resembles a smooth pursuit operation in the human visual system; however, the operation of the system is not intended to be in real-time and, in consequence, the study of a smooth pursuit behaviour is out of the scope of this thesis.

6.3.1 Properties of Canonical Views

Three different underlying properties characterise a canonical view (according to Blanz et al. (1996) and Ullman et al. (2002)):

1. *Goodness of recognition:* This depicts that salient and significance of visual features and stable views are preserved over small linear transformations. In consequence, an unstable view is one that for small rotations produces considerable changes on the appearance of the visual feature, and, thereby, recognition becomes difficult.

2. *Familiarity:* Object representations and visual sampling strategies applied for recognition are determined by the most frequently views and the ones used for initial learning. This suggests that visual features should be learned if a visual pattern is observed over different saccades/rotations over the object. That is, visual information of previously observed features against input features are compared such that the object visual description is maximised towards a highly-tuned object representation stored in the object database knowledge. Therefore, it is deduced that the WHAT stream measures the mutual information of acquired knowledge at different periods of time.

   At the visual feature description level, information is defined according to the spareness and clustering of neurones found in the *IT cortex* of the human brain (as investigated in Op de Beeck and Baker (2010)). That is, spareness involves the geometric distribution across the visual feature population, whilst, clustering selectively tunes feature descriptions by encoding the spatio-temporal relations across the sampled view-sphere.

3. *Functionality:* This defines that the object representation is linked to the common poses of the object observed. Therefore, recognition is interconnected by means of the action-perception cycle in terms of the most relevant views during training, after learning the object appearance and on the usefulness of the object. In the context of the robot head, the object's usefulness is not relevant as it is not considered a grasping behaviour; however, the relevance on how an object is seen while looking at/for it would allow adapting the overall concept of the object over time.

Thus, the above mentioned canonical view properties are considered in order to design and implement a novel active visual learning behaviour. This behaviour is specifically devised

from the founding principles of the devised hierarchy of visual behaviours architecture in Chapter 5. Moreover, the visual learning behaviour further validates the hierarchical architecture as discussed in Sections 5.9 and 5.10 ("a different high-level behaviour which validates its applicability and robustness of the architecture", ref. 5.9). The following section outlines the working hypotheses and defines visual objectives for this behaviour.

6.4 Visual Learning Behaviour Overview

The automatic synthesis of the object's appearance consists in the interaction between exploration and learning visual behaviours. That is, visual behaviours are iterated in accordance with the task-goal specified in the deliberative layer. Thus, the learning related behaviours allow to learn the object's appearance in terms of SIFT features whilst the exploration related behaviours enable the system to verify and, subsequently, determine if a new canonical view is required when the learned object pose is not longer valid. This interaction between behavioural modes is termed as a *exploration-learning* strategy and this strategy is employed in the remainder of this chapter.

A canonical representation is determined by clustering spatio-temporal properties of the imaged object in order to extract and find local representative feature groups. The centre coordinate of such groups are considered as attentional seeds in order to produce putative object parts that can be attended (attentional shrouds). Thus, attending to those salient informative object parts, the system is expected to learn and represent potentially the most informative, viewpoint invariant canonical view of an object. An exploration active paradigm (i.e. camera-eye movements) improves the robustness of finding canonical object parts. Thus, canonical representations become stable by gaining more visual information while saccading to structural representative object parts (i.e. active interaction with the object).

Figure 6.3 depicts the macro script for the task of *active visual learning of the appearance of the object's view-sphere*. The behaviours required in order to carry out the specified task are described in the following sections. Hence, the devised visual learning behaviour is composed of the following algorithmic steps:

1. Attentional shrouds are computed by covertly clustering depth and motion cues of extracted SIFT features of the imaged "unknown"object over a period of time. These consist of three consecutive frames of a portion of the object's view-sphere without saccading or verging the cameras. The object's view-sphere is discretised every 1 degree (i.e. frames are captured each 1 degree). This point iterates until attentional shrouds are found. Accordingly, sections 6.6.1.1, 6.6.2 and 6.6.3 describe the behaviours involved.

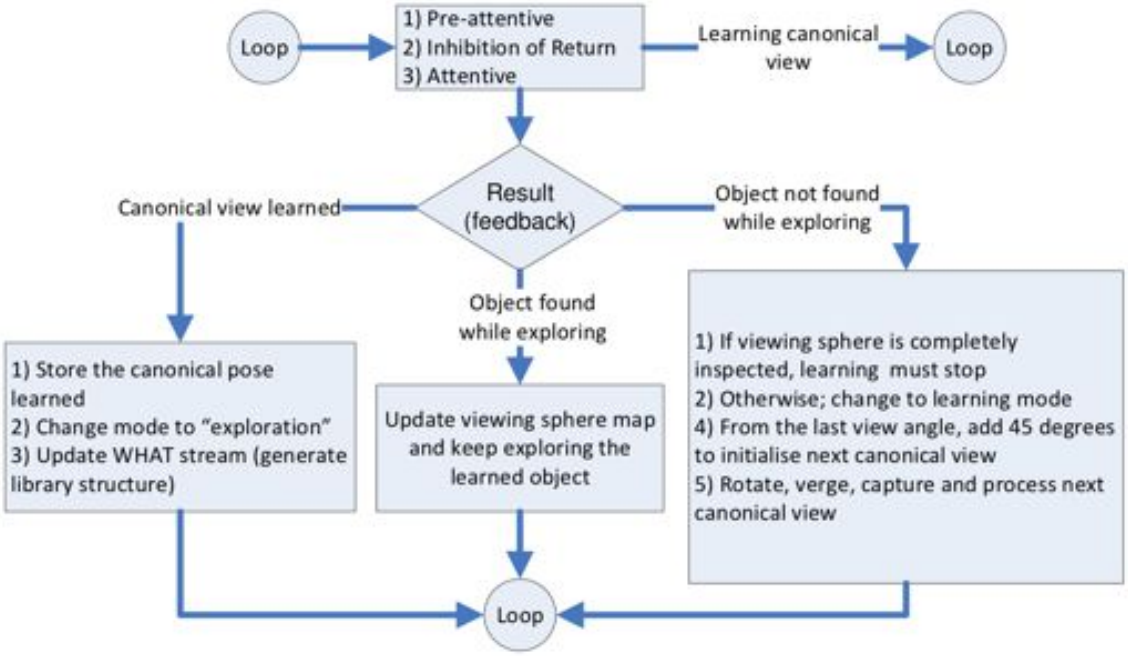


Figure 6.3: Flow diagram of the macro script for the visual learning behaviour.

2. Found attentional shrouds are cast as object hypotheses (as defined in Section 5.5.2) and passed to the attentive behaviour in order to start investigating an attentional shroud.

3. The attentive function then saccades to the reported location on the selected viewing angle (Section 6.2) verges on the targeted attentional shroud (Section 6.7.1) and verifies the fixated shroud.

4. While targeting the attentional shroud for the first time, an active clustering process is initialised (Section 6.7.3) (i.e. loosely based on the bag-of-words model as discussed in Section 2.5.1).

5. The object is rotated at a discrete angular step ($\beta = 1$ degree) until the next rotational saccade[4] (Section 6.6.4) is reached. Focus on the fixated shroud is maintained (i.e. binocular gaze holding). Therefore, the pre-attentive behaviour tracks and estimates the updated location of the attentional shroud (Section 6.6.1.2) and updates where to saccade next (Section 6.6.4).

6. While rotating the object at discrete steps in order to maintain spatio-temporal continuity, SIFT descriptors continue being accumulated and clustered. If the rotational saccade is reached, the attentive behaviour clusters new visual evidence and evaluates whether enough information has been acquired and the process must stop or it continues learning the attentional shroud. (Section 6.7.3).

[4]A rotational saccade is defined in this chapter as the angular positions where the learning behaviour determines to continue learning features. This allows establishing a halting criterion based on the mutual information between rotational saccades (as described in Section 6.7.3).

7. If there are attentional shrouds to be learned, the processes from points 2 to 6 are repeated (i.e. *"Learning canonical view"* case as depicted Figure 6.3). Otherwise, a canonical view is consolidated and the learned view is stored in working memory. Subsequently, the attentive and pre-attentive behavioural functions are commanded to operate in an exploration mode (i.e. *"Canonical view learned"* case as depicted Figure 6.3).

8. When the currently learned canonical view is not detected (pre-attentive) or recognised (attentive), the high-level macro script modifies the operational mode to "learning" while selecting a new object view (i.e. *"Object not found while exploring*" case as depicted Figure 6.3). A new view (a clock wise rotation in the turn-table) is selected by rotating the turn-table 45 degrees[5] with respect to the last recognised object's pose angle.

9. If the system has not investigated the object's view-sphere (i.e. a complete revolution of the object's view-sphere), the algorithm returns to point 1; otherwise, learned canonical views stored in working memory are saved to a database (as Equation 5.3) and the visual learning behaviour terminates.

It must be pointed out that the implemented deliberative visual behaviour while learning specifically consists of inspecting the object's view-sphere in a counter- and clock-wise direction, respectively. That is, the system initially rotates the "unknown" object in one direction with respect to the selected starting view. Related learning behaviours acquire visual information until enough features have been gathered. Thereafter, the system rotates the unknown object in the other direction with respect to the selected starting view such that the system samples and learns different view points (i.e. the object's appearance of the object might be slightly different) of the the object's view-sphere.

6.5 Hierarchical Architecture - Visual Learning Case

The object learning behaviour, as discussed in Section 6.3, is as well inspired by the WHERE and WHAT streams as the active binocular robot vision hierarchical architecture of the previous chapter. In that respect, this architecture serves as a founding design principle to devise and develop the said visual object learning behaviour of this chapter. To that end, the hierarchical architecture (Figure 5.1) is adopted and such is extended to suit the new visual requirements of the high-level task-goal specification.

[5]From empirical investigation, this angular threshold is found to work well in practice.

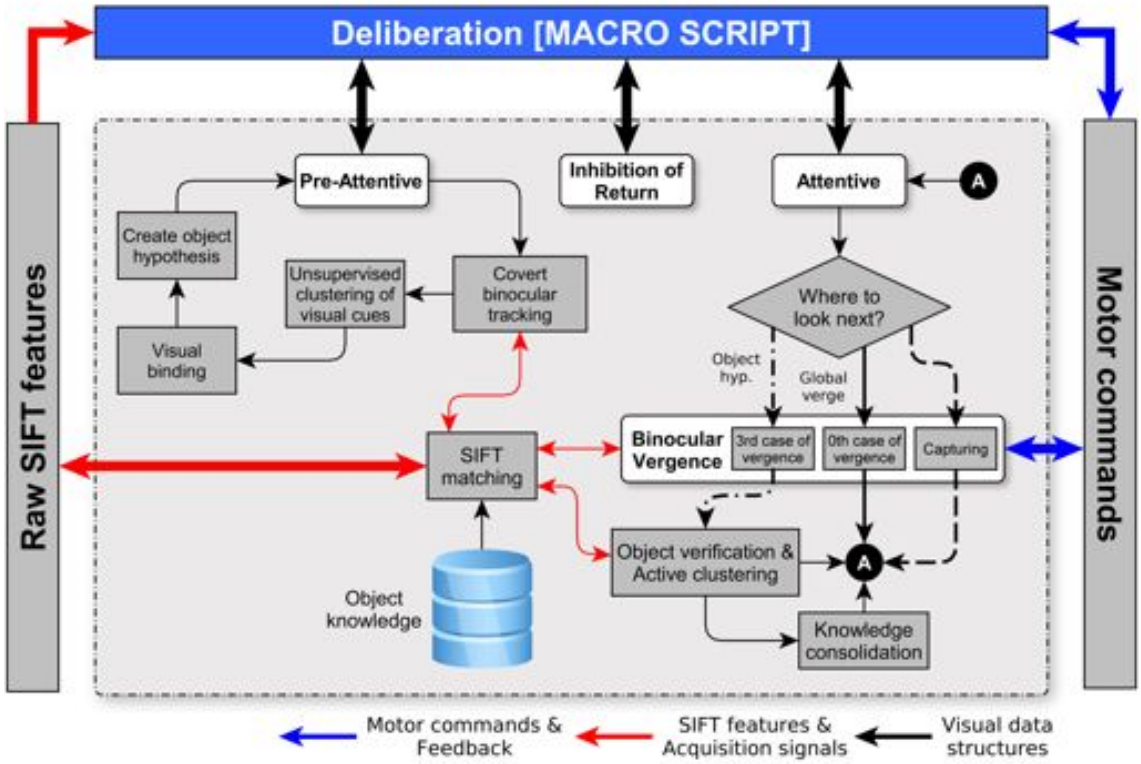


Figure 6.4: Hierarchy of visual behaviours for the visual learning behaviour within the devised robot vision architecture. White boxes denote abstract behaviours, whereas grey boxes represent primitive behaviours.

Figure 6.4 depicts the overall structure of the hierarchy of visual behaviours for the object appearance learning behaviour. While Figure 6.4 only depicts those visual behaviours related to learning case, explorative related behaviours are still contained within their corresponding abstract behaviours (see Figure 5.2).

As observed in Figure 6.4, this hierarchy shares the same abstract visual behaviours depicted in Figure 5.2 and defined in Section 5.3.1. Such behaviours are the *pre-attentive*, *attentive* and *inhibition of return*. However, their intrinsic primitive and abstract behaviours are extended with respect to those defined since the high-level task-goal specification is modified to enable the robot vision system to explore and learn the appearance of an object. The behavioural capabilities of the active binocular robot head are thus cast in accordance with the new visual requirements.

Accordingly, the ability to learn the appearance of an object requires that the pre-attentive behaviour processes SIFT features into meaningful data structures that are of interest to the visual learning behaviour and, similarly, tracks and maintains attention focused on an attentional-shroud/object-part while investigating it. The extended visual tasks of this behaviour are thereby briefed as follows:

- Point-based correspondence tracking of SIFT features in order to enable the system to pursuit a target from both cameras over a period of time (i.e. binocular gaze holding).
- Automatic and unsupervised segmentation of the indicated object to be learned from background in terms of depth and motion visual cues.
- Formulation of attentional shrouds hypothesis in order to bind together SIFT features with similar depth and motion properties.

On the contrary, the attentive behaviour employs generated data structures of the pre-attentive behaviour to either rotate and verge on the object in the turn-table, saccade both cameras towards found attentional shrouds or generate the object's knowledge structure. This abstract behaviour therefore carries out the following visual tasks:

- Saccadic camera movements that enables the robot vision system to attend each of the found attentional shrouds.
- Active clustering of visual features while learning the object in order to create and build the object appearance knowledge in terms of canonical views.
- Knowledge consolidation of the canonical view after all attentional shrouds have been investigated

Finally, the inhibition of return behaviour is only responsible for suppressing observed angular positions while learning the object. Thus, the defined visual task consists of:

- Inhibition of angular positions in order to enable the robot vision system to sample the object's view-sphere without returning to previously visited angular positions.

6.5.1 Actuation of the Turn-table

The overall function of the binocular vergence behaviour defined in the hierarchical architecture is not modified and such is executed as described in Section 3.4. However, this behaviour is extended in order to cope with the actuation commands of the turn-table employed in this chapter; the latter is described as follows.

The motors' control related behaviours are couched as primitive behaviours as described in Section 5.3. Therefore, actuation of the turn-table is an extra primitive behaviour in the low-level layer of the devised robot vision architecture (camera motor primitive behaviour as depicted in Figure 5.1). As one key objective of the hierarchical architecture is to decouple

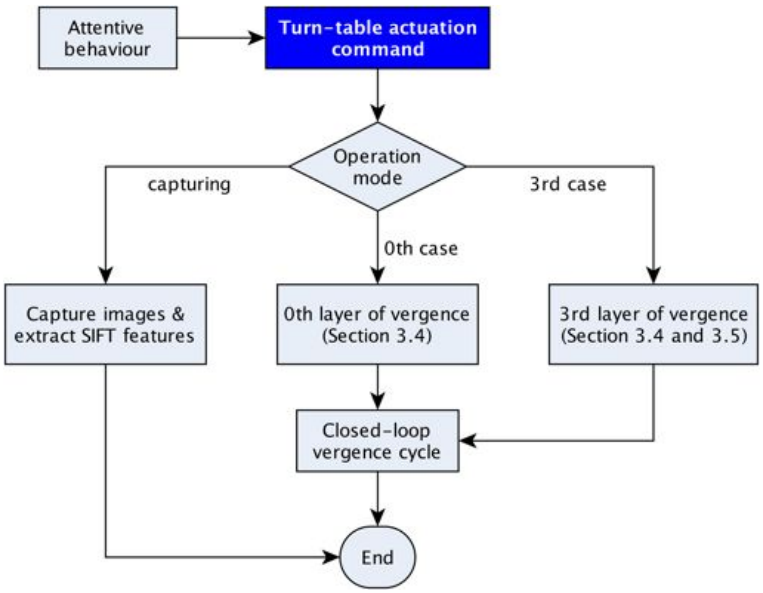


Figure 6.5: Binocular vergence behaviour adapted with the turn-table actuation control (highlighted) and featuring a passive acquisition primitive behaviour.

sensors and actuators from the mid- and high-level layers (ref. Section 5.3), this extra primitive behaviour is implemented in accordance with the turn-table's low-level hardware functions and such is thereby adapted to the required data structures of the architecture.

The mid-level abstract behaviour that specifically commands actuation signals to the camera's motors is the binocular vergence (as illustrated in Figure 5.2). This abstract behaviour is thereby adapted to control the actuated turn-table and, specifically, command the angular position of the turn-table as required while verging the cameras. That is, the robot head must verge into the object each time the object rotates. Therefore, its implementation consists in simply invoking the actuation of the turn-table before the vergence closed loop cycle for either the $0th$ or $3rd$ layer of vergence is executed (as illustrated in Figure 6.7). This allows keeping the cameras in convergence.

Figure 6.5 also depicts a primitive behaviour that is responsible for solely actuating the turn-table, capturing images and extracting SIFT features without invoking any of the above layers of vergence. This primitive behaviour is therefore defined for the specific purpose of preserving the perception-action cycle while tracking visual features. That is, the system is able to maintain spatio-temporal relations of the revolved object in the turn-table and, consequently, to allow tracking visual features as described in Section 6.6. Likewise, this also allows the system to rotate the object while exploring learned canonical views across the object's view-sphere (as discussed in Section 6.7).

6.5.2 Depth Perception

Depth perception is the visual ability to perceive the environment in terms of a three dimensional coordinate reference frame (Styles, 2005). This ability allows the design of robotic systems that can navigate (Svedman et al., 2005), drive a car (Newman et al., 2009), manipulate objects (Rasolzadeh et al., 2010) and so forth. There exist several visual cues that allow to compute either the relative or absolute depth distances from both retinal images (i.e. stereo vision). Such visual cues in stereo vision comprise *binocular disparities* (or stereopsis) and o*bject shadows*. Depth perception research, in the computer vision field, is vast, and there exist well established methods to compute and reconstruct the three-dimensional world from a pair of images. Specifically, Hartley and Zisserman (2004) and Cyganek and Siebert (2009) have condensed the most successful approaches for depth reconstruction and perception and, also, demonstrated them under different application contexts.

In this thesis, however, full depth perception is not required as the devised visual learning behaviour must the pre-attentive behaviour (in the visual learning behaviour context) must perceive the overall three-dimensional structure. This 3-dimensional visual cues are thus used to perceptually group features and, consequently, formulate attentional shrouds. Therefore, the 3-dimensional structure of the scene can be extracted from the stereopsis induced while verging the cameras (i.e. disparity maps). Figure 6.6 depicts binocular disparities of SIFT features from the left and right image after the cameras have been verged on a targeted object. As observed, the overall depth structure of the object is captured by the binocular disparities.

As described in Section 3.4, binocular disparities are computed from a set of SIFT features matches between the left and right camera, $\mathcal{M}^{(\mathbf{L},\mathbf{R})}$. For completeness, its definition within the overall hierarchical architecture is described as follows.

Binocular disparities, δ, is the pairwise difference of the x and y SIFT feature components of the left and right matched sets. These are expressed as (see Equation 5.1, 5.2 and 5.5 for definition of $\mathbf{L}$ and $\mathbf{R}$):

$$\delta_i = [l_{ij}]_{j=1}^{2} - [r_{ij}]_{j=1}^{2} \;\; \forall \, l_{ij} \in \mathbf{L}, r_{ij} \in \mathbf{R} \tag{6.1}$$

such that $i = 1, \ldots, n$ where n is the total number of correspondences between matched sets that satisfy: $a = b = n$ (ref. Section 5.4). Hence, these values create a sparse disparity points (attached to the SIFT features in the left camera) that encodes the approximate relative depth distances from the robot head to the points observed.

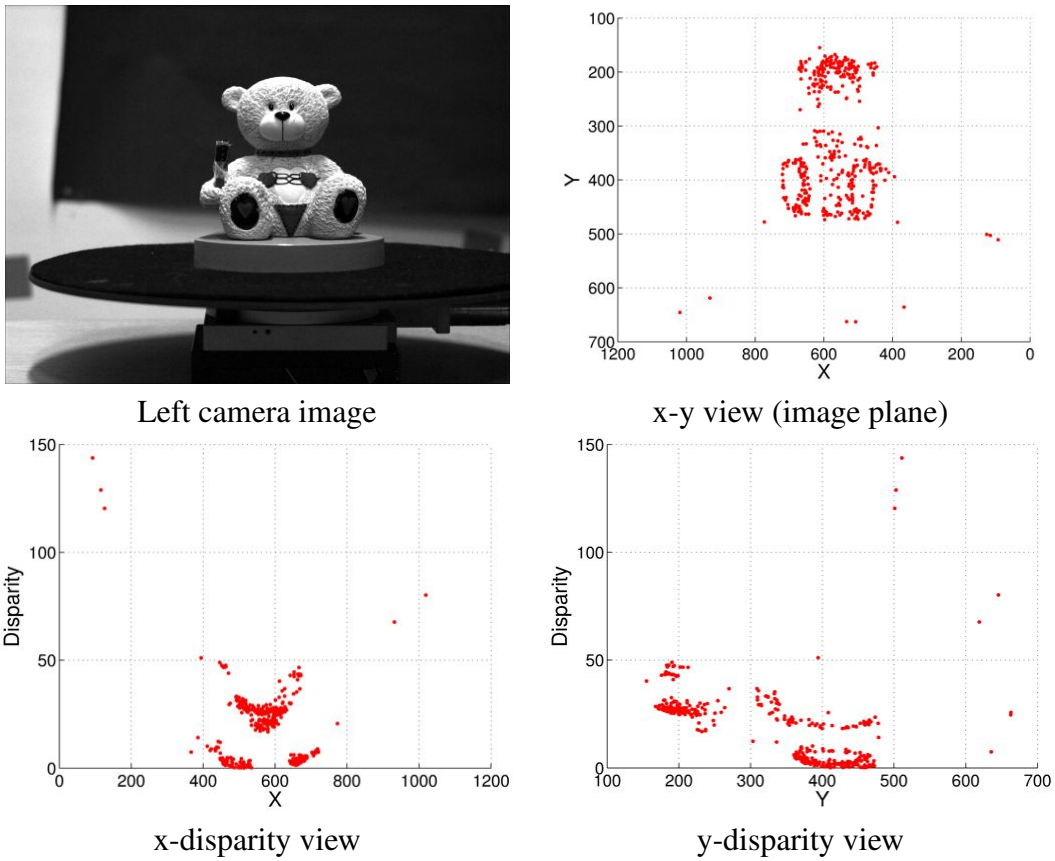


Figure 6.6: Binocular disparities of the bear object after verging the cameras.

6.6 Pre-attentive - Learning Case

The operational function of the pre-attentive behaviour described in the previous chapter does not provide the required visual capabilities for tracking down visual features (i.e. SIFT features are traced over the object's viewing sphere) and attentional shrouds formulation. The pre-attentive behaviour is therefore extended in accordance with the visual tasks described in Section 6.5.

Figure 6.7 depicts the overall hierarchical configuration of the pre-attentive behaviour. As observed, the devised architecture allows a new branch in the behaviour to be created while preserving the visual object exploration capabilities previously defined. Thus, by invoking the pre-attentive behaviour in the *"learning"* mode, this behaviour encompasses two further different operational modes. These are:

1. Finding and formulating object parts (step 1 in Section 6.4).
2. When an object part is being attended (step 5 in Section 6.4).

The first operational mode consists of tracking visual features over a set of a fixed angular step

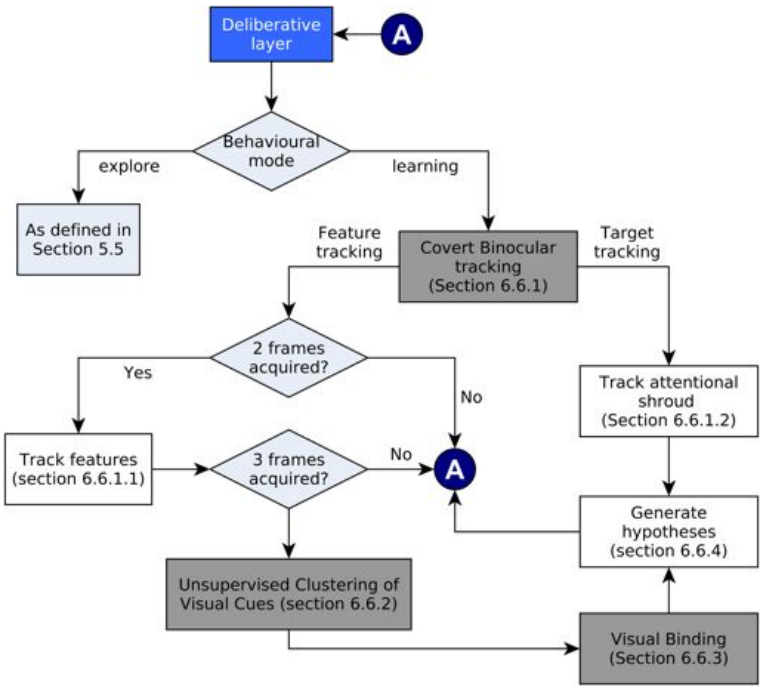


Figure 6.7: Flow diagram of the overall pre-attentive behaviour including the visual learning case. Grey boxes denote abstract behaviours whereas white boxes represent primitive behaviours.

samples. Thereafter, tracked visual features that are consistent across the angular sampling steps are grouped in terms of the three-dimensional spatial coordinates, and motion vectors. Each clustered group therefore defines an object part, and such clusters are employed to create putative attentional shrouds. The second mode iteratively employs those primitive behaviours of the first mode in order to track a currently fixated attentional shroud and determine where to saccade next. That is, the object is rotated during the attentive cycle (as described in Section 6.7) and, in consequence, the attentional shroud must be located.

The above operational modes are invoked in terms of the current behavioural state of the robot. Specifically, $\mathcal{H}^P$ (see Equation 5.11) provides the means of dictating the operational mode of the pre-attentive behaviour. That is, this behaviour is controlled by querying whether the first entry of the numeric index of the object pose class, $\mathcal{I}_i$ for $i = 1$, is infinite (i.e. an object is been tracked and learned) or empty (i.e. there are not attentional shrouds hypotheses).

The abstract and primitive behaviours that constitute the pre-attentive abstract behaviour for the visual learning case are described in the following subsections.

6.6.1 Covert Binocular Feature Tracking

The ability of the system to hold its gaze on a single moving object relies on the close-loop interaction of the pre-attentive and attentive abstract behaviours as discussed in Section 6.2. In this thesis, the working assumption (as discussed in Section 6.3) is that a human operator provides an unknown object to the robot by placing it on the turn-table and, in consequence, the robot dynamically controls the angular position of the object's view-sphere. This allows to collect visual evidence by revolving the object over a set of angular positions. In this section, however, it is only considered the pre-attentive analysis of moving features between two acquired frames while attentive behaviours are discussed in Section 6.7.

Hence, this behaviour is twofold:

1. It tracks SIFT features and process them to a data structure that consists of velocity vectors, the angular object position where the optical flow is computed and binocular disparities; logically indexed to only those SIFT features in motion.
2. It tracks attentional shrouds while attentively inspecting and learning the ith object in set $\mathcal{H}^P$.

The implementation of each thus follows a hierarchical arrangement and such is divided according to the above described operational modes. Each mode is described as follows.

6.6.1.1 Feature Correspondence

Feature tracking techniques are categorised into two main groups: correspondence and texture correlation techniques. The pre-attentive feature tracking/correspondence behaviour is therefore based on feature correspondence based approaches. The idea of using features points date back to the *Lucas-Kanade feature tracker algorithm* (LK tracker) *(Lucas and Kanade, 1981)*. In this tracker, the second-moment matrix (i.e. also known as structure tensor) around a point have a significant role in the computation of "good features" which, in turn, can be used for correspondence purposes. In recent years, this type of feature detection has been substituted for more sophisticated techniques, e.g. corner/blob extractors, which, in essence, have been on the structure tensor rationale. For instance, the pyramidal version of the LK tracker (Bouguet, 2002) employs Harris features (Harris and Stephens, 1988) in order to compute a set of robust feature points that might potentially be of interest to the tracker. In that respect, SIFT features are an improved and extended version of the Harris detector and shares similarities to the pyramidal image representation algorithmic steps of the method presented

in (Bouguet, 2002). As SIFT features are the underlying visual representation for all visual behaviours described in this thesis, the feature correspondence behaviour is thus approximated in terms of the SIFT feature matching operations (see Equation 5.5). A description of the feature correspondence based on SIFT features within the hierarchical structure of the robot vision system is presented below.

It is assumed that two different frames from each camera have been acquired. Between observed frames, it is established that1 degree of rotation in the turntable is small enough in order to densely sample the object's view-sphere and, in consequence, to loosely simulate a smooth pursuit behaviour while trying to localise the object in motion. The time steps at given angular positions between frames is denoted as: t and $t+1$. Thus, $\mathbf{L}_t$ and $\mathbf{L}_{t+1}$ are the extracted SIFT features for the left camera of each frame respectively; and, conversely, $\mathbf{R}_t$ and $\mathbf{R}_{t+1}$, for the right camera. For simplicity, the algorithmic description is described only for the left camera; although the same algorithmic steps apply for the right camera.

Hence, SIFT features are matched, i.e. $\mathcal{M}^{(\mathbf{L}_t,\mathbf{L}_{t+1})}$ and $\mathcal{M}^{(\mathbf{R}_t,\mathbf{R}_{t+1})}$, for the left and right camera, respectively. The algorithmic steps are described specifically for the left camera; however, the same holds true for the right camera. Thus the induced horizontal and vertical optical flow, $\vec{x}_i^L$ and is computed by taking the difference of the x and y SIFT feature coordinates. These velocities are then transformed to a polar representation:

$$\varrho_i^L = \sqrt{\vec{x}_i^2 + \vec{y}_i^2} \tag{6.2}$$

$$\varphi_i^L = \arctan\left(\frac{\vec{y}_i}{\vec{x}_i}\right) \tag{6.3}$$

where ϱ_i^L and φ_i^L are the radial and direction components. Thus, as the object in the turn-table can only present horizontal motion, the direction component is accumulated in a 12 bin histogram (with a bin width of $\frac{\pi}{6}$) (this number of bins is found to work well in practice) in order to determine the dominant orientation of the induced optical flow of the object in motion (i.e. the biggest bin in the histogram determines the dominant direction of rotation). Likewise, the rotated object might present velocity vectors in both directions as the object roughly revolves near the rotational axis of the turn-table. The velocity vectors are thus selected according to the positive and negative dominant direction of the object with a tolerance equal to the size of the specified bin width. This captures motion in opposite directions, approximately segment SIFT features based on the observed motion and, likewise, remove false positive matches.

Thereafter, an affine pose estimator (as in Section 4.6) is employed to obtain a registration error of the moving visual features. That is, only those SIFT features that survive the above dominant direction filtering, are used to initialise the affine pose estimation described in Sec-

tion 4.6. Therefore, a registration error less than 10 pixels[6] for each velocity component denotes that those tracked features between frames are stable.

The final result is therefore expressed as two matrices according to the hierarchical architecture. The matrix for the left camera (similarly for the right camera) is:

$$\mathbf{Z}_t^L = \begin{bmatrix} 1 & \left([l_{1j}]_{j=1}^{4}\right)_t & (\delta_1)_t & \left(\varrho_1^L\right)_t & \left(\varphi_1^L\right)_t & \Phi_t & \left([l_{1j}]_{j=5}^{132}\right)_t \\ 2 & \left([l_{2j}]_{j=1}^{4}\right)_t & (\delta_2)_t & \left(\varrho_2^L\right)_t & \left(\varphi_2^L\right)_t & \Phi_t & \left([l_{2j}]_{j=5}^{132}\right)_t \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ n & \left([l_{nj}]_{j=1}^{4}\right)_t & (\delta_n)_t & \left(\varrho_n^L\right)_t & \left(\varphi_n^L\right)_t & \Phi_t & \left([l_{nj}]_{j=5}^{132}\right)_t \end{bmatrix} \tag{6.4}$$

$$\mathbf{Z}_t^L = \left([z_{ij}]_{i=1,j=1}^{n,137}\right)_t^L \tag{6.5}$$

where t is the recorded time step, $[l_{ij}]_{j=1}^{132}$ $(i = 1, \ldots, n)$, SIFT features as in Equation 5.1, and Φ, the angular position of the object at the given observation. The first column in Equation (6.5) denotes classification labels of the SIFT features tracked and such labels are updated each time this behaviour is invoked. These labels are subsequently used to actively cluster visual evidence in the attentive behaviour.

Hence, the learning behaviour (as described in Section 6.4) must acquire three consecutive frames for each camera without performing eye movements in order to obtain stable tracked features. This operational function allows to formulate meaningful object parts in terms of depth and motion visual cues.

In that respect, this behaviour has been defined thus far to cope with two frames; however, it might be invoked with previous processed frames, i.e. $\mathbf{Z}_{t-1}^L$ and $\mathbf{Z}_{t-1}^R$. That is, this behaviour employs previous tracked information in order to maintain spatio-temporal continuity across the object's view-sphere. Therefore, for those SIFT feature points in $\mathbf{Z}_{t-1}^L$ and $\mathbf{Z}_{t-1}^R$ that are still tracked for the the current time step, t; they are logically indexed in terms of their classification number label of the previous frame. Those features that are "lost" or "newly discovered" are labelled as such and stored in working memory for future frames; however, for the purpose of the learning behaviour, these are not further considered.

Hence, $\mathbf{Z}_t^L$ and $\mathbf{Z}_t^R$ (as in Equation 6.5) are therefore passed to the unsupervised clustering behaviour in Section 6.6.2 in order to group visual cues in terms of their motion and depth and, in consequence, to formulate attentional shroud hypotheses.

[6]From empirical investigation, a global error less than 10 pixels is found to work well in practice.

6.6.1.2 Attentional Shroud Tracking/Correspondence

An attentional shroud spatially characterises a surface or shape of an object (as defined in Section 6.3). This shrouds is defined as a convex hull that contain SIFT features that are currently learned (expressed as $\mathbf{F}$ as defined by the Equation 6.25). Thus, an attentional shroud is expressed as $\mathbf{A}^U = \left[a_{ij}^U\right]_{i=1,j=1}^{p,2}$ where p is an integer that denotes the number of nodes of the convex hull and U denotes either for the left, L, or right,R, camera.

Hence, the target tracking behaviour only consists of finding those SIFT features that match the targeted attentional shroud and, in consequence, find the projection and the fixation point, $\mathcal{X}^U$ (see Equation 5.6), where this attentional shroud maps to the current observed frame. The algorithmic steps are summarised as follows.

For those SIFT features in $\mathbf{F}$, the tracked position is thus found by $\mathcal{M}^{(\mathbf{F},\mathbf{L}_t)}$ and $\mathcal{M}^{(\mathbf{F},\mathbf{R}_t)}$ for the current time step, t. In order to find the transformation that maps $\left[a_{ij}^U\right]_{i=1,j=1}^{p,2}$ into the current image frame (for either the left or right camera), the affine pose estimator described in Section 4.6 is invoked with the ordered matched sets previously mentioned for each camera. After projecting the attentional shroud into the current frame, the current attentional shroud hypothesis is updated accordingly (see Equation 5.7):

$$\mathcal{H}^U = \left\{\mathcal{I}^U, \mathcal{X}^U, \epsilon^U, \mathbf{A}^U, \mathbf{F}, \mathcal{E}^U\right\} \tag{6.6}$$

where $\mathcal{I}^U = \epsilon^U = \infty$ (which denotes that it is the current targeted object in the learning behaviour as defined in Section 6.6);

$$\mathcal{X}^U = \left(\mu\left(\left[a_{i1}^U\right]_{i=1}^{p}\right), \mu\left(\left[a_{i2}^U\right]_{i=1}^{p}\right)\right), \tag{6.7}$$

$\mathbf{F}$ contains those features that are being learned, and $\mathcal{E}^U$ contains the attended angular positions of the object's view-sphere (ref. Section 6.6.4). Figure 6.8 depicts an example of a tracked attentional shroud in both cameras. Hence, $\mathcal{H}^L$ and $\mathcal{H}^R$ are passed to the hypotheses generation behaviour described in Section 6.6.4.

6.6.2 Unsupervised Clustering of Visual Cues

Several types of visual cues might create an object part such as colour, shape and texture, to name but a few. In this thesis and for the specific purpose of the visual learning behaviour, an object part is however defined as a group that is within a similar depth plane and preserves spatio-temporal continuity properties. To that end, this behaviour enables the system to group

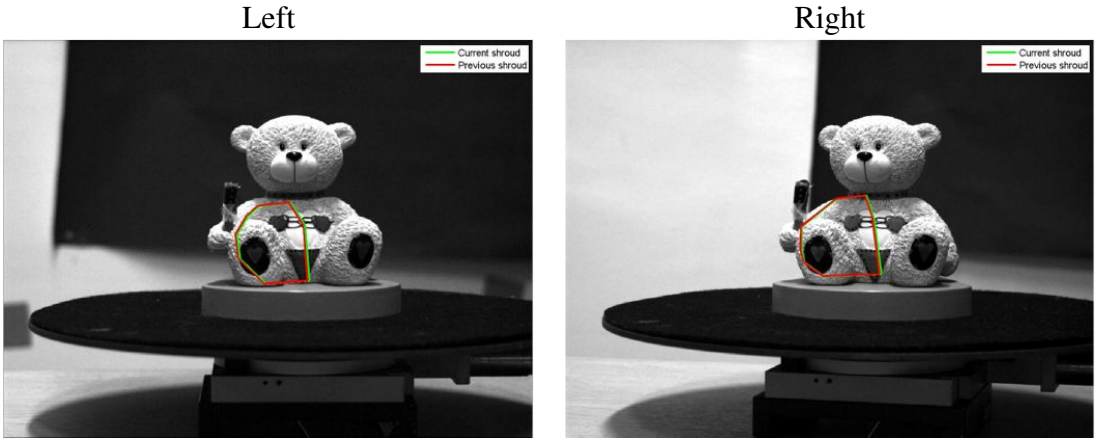


Figure 6.8: Tracked attentional shroud in both cameras.

such visual cues into initial object parts that subsequently subserve the definition of attentional shroud hypotheses.

This behaviour therefore employs Algorithm 4.1 in order to cluster SIFT feature locations indexed with corresponding disparity and polar velocity components; i.e. $(x, y, \delta, \varrho, \varphi)$, which are consistent over all processed frames in each camera. That is, two data matrices for each camera are considered: $\mathbf{X}^U_{t-1}$ and $\mathbf{X}^U_t$; and such are defined as follows (see Equation (6.5)):

$$\mathbf{X}^U_t = \left[\begin{array}{ccccc} \left(z^U_t\right)_{i2} & \left(z^U_t\right)_{i3} & \left(z^U_t\right)_{i6} & \left(z^U_t\right)_{i7} & \left(z^U_t\right)_{i8} \end{array} \right] \tag{6.8}$$

where i is the ith entry that observes spatio-temporal continuity properties for each processed time step ($t-1$ and t) and U denotes either the left, L, or right, R cameras. Thus, Algorithm 4.1 is invoked with $K = 6$ and, $\mathbf{X}^U_{t-1}$ and $\mathbf{X}^U_t$ accordingly. The selected clustering algorithm is the fuzzy C-means as it does not require that $\mathbf{X}$ takes a particular shape. Likewise, a maximum number of 6 clusters is established as the attentional spotlight in biological systems is not capable of localising more than 6 objects/blobs without making eye movements (as discussed in Section 2.5). In addition, those features that are classified as "lost" or "newly discovered" in Section 6.6.1.1 are not considered as they might be outliers and, consequently, affect the performance of the unsupervised clustering.

Thus, the group, $\mathcal{G}$, that holds the optimal number of clusters is selected, and $\mathbf{Z}^L_{t-1}$ and $\mathbf{Z}^L_t$ are sorted in accordance with their cluster assignments. The output of this behaviour is therefore the average silhouette score of each processed frame:

$$\overline{\mathrm{S}}^U = \left(\overline{\mathrm{S}}^U_{t-1}, \overline{\mathrm{S}}^U_t\right) \tag{6.9}$$

(ref. Algorithm 4.1 and Equation (4.16)); and, likewise, the ordered matrices, $\mathbf{Z}^U_{t-1}$ and $\mathbf{Z}^U_t$:

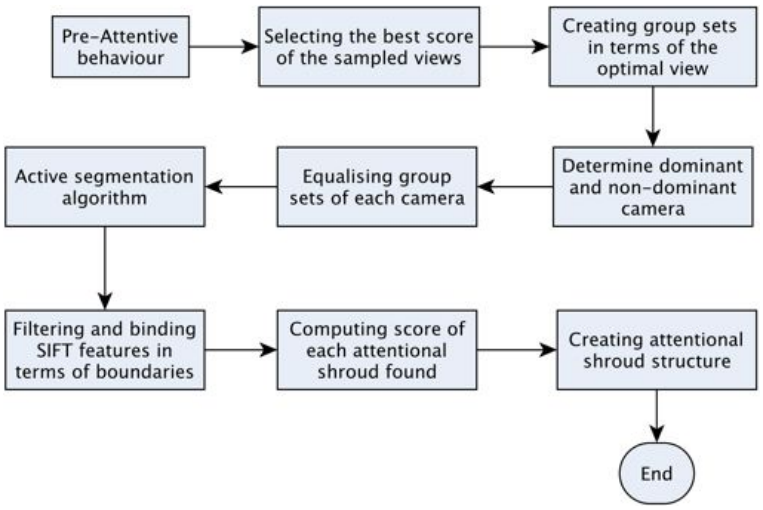


Figure 6.9: Pipeline of the visual binding behaviour.

$$\mathbf{Z}_{t-1}^{U} = \mathbf{Z}_{t-1}^{U}\left(\mathcal{G}_{t-1}^{U}\right) \tag{6.10}$$

$$\mathbf{Z}_{t}^{U} = \mathbf{Z}_{t}^{U}\left(\mathcal{G}_{t}^{U}\right) \tag{6.11}$$

where U is either the left, L, or right, R cameras and $\mathcal{G}$ contains cluster assignments. These are subsequently passed to the visual binding behaviour as described in the following section.

6.6.3 Visual Binding

The visual learning behaviour requires a behaviour that correctly "glues" or puts together those visual properties (e.g. binocular disparities, sparse optical flow and SIFT features) into meaningful object data structures (e.g. Equation 5.11). This behaviour thus determines which visual featres belong to the object of interest to the learning behavior. In the literature, scientists have discovered that this visual behaviour is governed by a *visual binding paradigm* (Styles, 2005; Tsotsos et al., 2008). In this chapter, the visual binding behaviour allows to put together perceptual properties found in the WHERE stream in order to create attentional shrouds that capture the object's perceptual intrinsic properties and, consequently, to drive attention, while learning the appearance of the object (as defined in Section 6.3).

Hence, the visual binding behaviour is summarised in Figure 6.9 and described as follows. As feature groups of the left and right camera have been clustered for each processed frames (Equations (6.10) and (6.11)), the optimal observed view is selected in terms of the computed silhouettes numbers of Equation (6.9) as follows:

$$s_t = \left| 1 - \frac{\min\left(\overline{\mathrm{S}}_t^L, \overline{\mathrm{S}}_t^R\right)}{\max\left(\overline{\mathrm{S}}_t^L, \overline{\mathrm{S}}_t^R\right)} \right| \tag{6.12}$$

where s_t denotes the quality view score where clusters are grouped with high confidence for the specified time step, t (i.e. the spatial configuration of clusters does not undergo a significant statistical variance). This quality score is therefore evaluated for all processed frames (i.e. $t-1$ and t) and the overall best observed $view$ is determined as follows:

$$view = \begin{cases} t-1 & \text{If } s_{t-1} > s_t \\ t & \text{If } s_{t-1} \leq s_t \end{cases} \tag{6.13}$$

The features on the selected $view$ are thus employed, $\mathbf{Z}_{view}^U$, and group sets are created for each camera respectively. There are cases where the clustering algorithm finds clusters with less or equal than 3 features. Such groups are not considered and their SIFT feature members are thus reclassified to the closest group in terms of their minimum Euclidean distance between their SIFT feature coordinates (see Equations (5.1) and (5.2)) and the group coordinate centres.

Thereafter, the camera holding the fewest number of groups is selected as the dominant one. That is, an object that is split into several and small parts does not symbolise a significant perceptual representation of the image object and this camera becomes the non-dominant view. This process of selective covert attention in the human visual system is termed as the *"binocular rivalry"* phenomenon (Styles, 2005). It controls perceptual awareness, or conscious experience of the observed stereo retinal images. Competition of the perceived stimuli is driven in terms of the visual properties (e.g. brightness, high contrasts, to name a few). In this chapter, binocular rivalry specifies how grouped features compete for the resources of the system. Hence, the attention, in the context of the learning behaviour, is thereby biased to focus on large attentional shrouds in order to capture and learned observed visual information (i.e. greedy algorithm).

SIFT features of the non-dominant camera eye are thus grouped according to those features in the dominant camera by finding their respective correspondences between the stereo pair, i.e. $\mathcal{M}_p^{(\mathbf{Z}_{view}^L, \mathbf{Z}_{view}^R)}$. For this matching operation, $\mathcal{T}_{SIFT} = 0.8$ as SIFT features between cameras might be slightly different. Hence, groups are formed in accordance with the feature matches in the dominant camera. Similarly, classification labels of the suppressed camera are renamed in relation to their corresponding match label of the dominant camera. There are cases that features that do not match between cameras as they might be occluded or not tracked. In that respect, unmatched features in each camera are thus grouped to the closest cluster of the corresponding camera. This is carried out by computing the spatial Euclidean

distance between each unmatched SIFT feature coordinate and the formed cluster centres. The minimum distance to a cluster centre therefore depicts the closest group, and the unmatched feature is assigned to selected cluster.

In order to finally *bind* SIFT features to the object in motion, an *active segmentation algorithm*[7] proposed by Mishra et al. (2009) (ref. Section 2.4.3) is implemented in order to delineate the bounding contour of the object in motion. The *active segmentation algorithm* exploits motion (i.e. an optical flow map[8]), colour and texture visual cues in order to generate a probabilistic boundary edge map that highlights the boundaries of the object in motion. The image segmentation is thus carried out by transforming the image to a log-polar representation. Thereafter, it is applied a standard *"graph-cut*" segmentation algorithm in order to find the shortest path of the edge with high probability. Therefore, the active segmentation algorithm is only employed in this thesis as an extra and extended visual ability of the investigated robot vision system. It must be also noted that the visual binding behaviour employs a different understanding model as SIFT features are limited to texture description and such do not do not suffice to extract bounding contours of the imaged object (Chapter 8 further discusses this limitation).

Thus, the active segmentation algorithm is only applied to the colour image of the robot head since it employs colour features to find bounding contours as described above (the robot head only has one colour camera as described in Section 6.1.1). Nevertheless, the feature groups from the dominant and non-dominant cameras are roughly equalised by means of the *binocular rivalry* greedy algorithm described above.

Hence, Figure 6.10 depicts an example of applying the *active segmentation algorithm* (Mishra et al., 2009). The dense optical flow map is computed for the selected $view$ (Equation (6.13)) to the next frame, $t + 1$. Similarly, fixation points used to initialise the active segmentation algorithm are the cluster centres of the feature groups of the left (colour) camera. The output of the implementation of this algorithm is illustrated at the top-left corner in Figure 6.10.

As observed in this example, sometimes the active segmentation algorithm does not produce a single, well-defined bounding contour of the object, but creates several broken up contours around the fixation points. This is because this algorithm is highly sensitive to image noise and to the quality of the optical flow map (the reported experiments in (Mishra et al., 2009) employ a laser range finder to compute such optical flow map, whereas, in this thesis, only vision is used). To that end, the initial segmentation is further refined by merging, dividing or eliminating segmentation hypotheses giving preference to bigger regions. That is, this greedy

[7]Code can be retrieved from: `http://www.umiacs.umd.edu/~mishraka/code.html` (verified on the 30th September, 2011)

[8]The code to compute the dense optical flow map is retrieved from: `http://people.csail.mit.edu/celiu/OpticalFlow/` (verified on the 30th September, 2011)

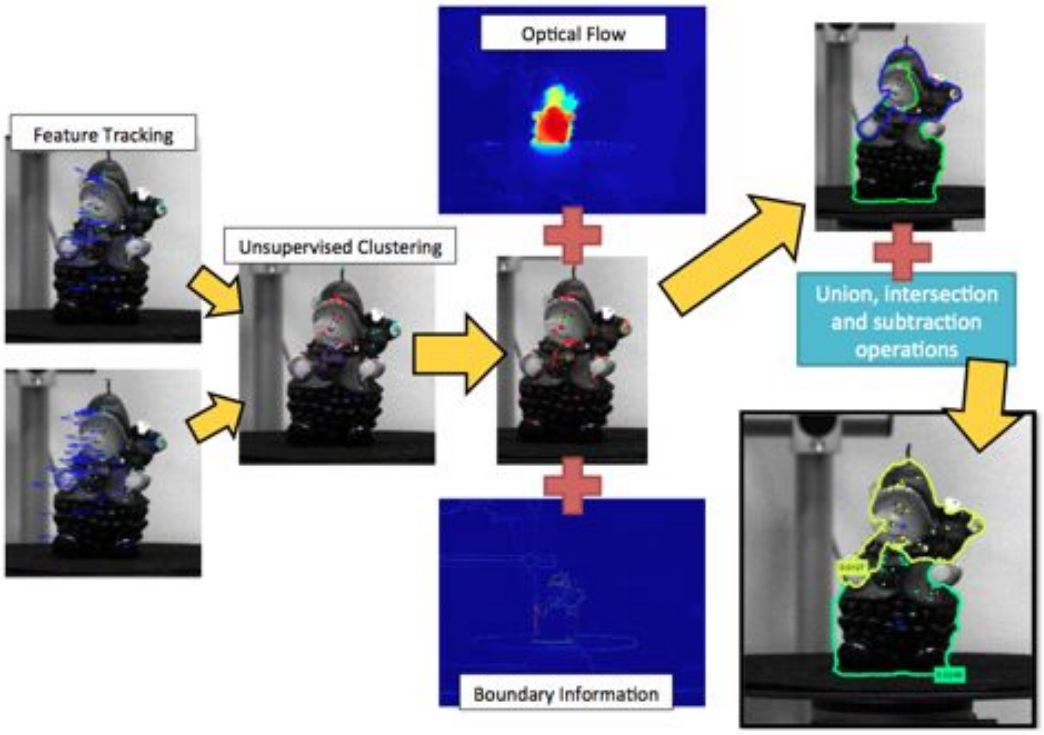


Figure 6.10: Active segmentation.

algorithm consists of computing the area (in pixels) of a segmentation hypothesis in order to compare it with the areas of other hypotheses. Thereafter, the algorithm decides if this segmented contour is merged, divided or eliminated such that a bigger region is maintained. This algorithm is halted when segmentation hypotheses do not overlap between each other. Algorithm 6.1 summarises this process.

The final bounding contours are thus projected into the right camera (the affine pose estimator described in Section 4.6 computes the required transformation in terms of the clustered SIFT features). Such bounding contours in both cameras specify the spatial image limits where SIFT feature coordinates must be located. SIFT features that are found outside these bounding contours are thus removed from their corresponding cluster. Hence, the convex hull of the SIFT feature coordinates set of each group denotes the attentional shroud, A (as described in Section 6.6.1.2) for each camera.

As described in Section 6.4, each attentional shroud is serially attended in order to gather and learn the sensed visual information. In order to enable the system to attend shroud locations, the attentive behaviour must know which shroud has to be attended first. To that end, a confidence score is defined for each as follows (Equations are expressed for the left camera but the same holds for the right):

$$r_i^L = D_E\left(\left(l_{i1}, l_{i2}\right), \mu\left(\left[l_{jk}\right]_{j=1,k=1}^{n,2}\right)\right),\ i = 1, \ldots, n \tag{6.14}$$

Algorithm 6.1 Set operations for the active segmentation.

Inputs: $\mathbf{A} = \left([A_{ij}]_{i=1,j=1}^{k,2}\right)_{p=1}^{q}$, initial attentional shrouds where k is the number of shrouds and q is the total number of found initial shrouds.

Outputs: $\mathbf{A}^{out}$, refined attentional shrouds with a similar structure as above.

1: $a \leftarrow$Compute areas for each shroud in $\mathbf{A}$
2: $\mathbf{A}^{out} \leftarrow$Initialise with the same entries as $\mathbf{A}$
3: WHILE size of a > 1
4: [$\sim$, idx]$\leftarrow$Minimum area and logical index on $area$
5: For $i = 1,\ldots,$size of a
6: IF $i \neq idx$
7: IF $\mathbf{A}_{idx} \cap \mathbf{A}_i$ is not empty
8: $a1 =$area$(\mathbf{A}_{idx} \cap \mathbf{A}_i)$
9: $a2 =$area$(\mathbf{A}_{idx} \cup \mathbf{A}_i)$
10: END IF
11: IF $\mathbf{A}_{idx} \cap \mathbf{A}_i$ is not empty and $a2 > 0$
12: $test1 = \frac{a1}{a2}$
13: IF $test1 > 0.9$
14: $\mathbf{A}_i$ is inside $\mathbf{A}_{idx}$; thus remove it
15: break
16: ELSE
17: $test2 = \frac{a1}{a_i}$
18: IF $test2 < 0.1$
19: IF $test2 < 0.1$
20: $C = \begin{cases} 1 & \text{If } \mathbf{A}_{idx} \text{ is minimum} \\ 2 & \text{If } \mathbf{A}_i \text{ is minimum} \end{cases}$
21: ELSE
22: $C \begin{cases} 1 & \text{If } \mathbf{A}_{idx} \text{ is maximum} \\ 2 & \text{If } \mathbf{A}_i \text{ is maximum} \end{cases}$
23: END IF
24: IF $C = 1$
25: $\mathbf{A}_i = \mathbf{A}_i - \mathbf{A}_{idx}$ (set operation)
26: ELSE
26: $\mathbf{A}_{idx} = \mathbf{A}_{idx} - \mathbf{A}_i$ (set operation)
28: END IF
29: break
30: END IF
31: END IF
32: IF $i =$size of $area$
33: $\mathbf{A}_i^{out} = \mathbf{A}_{idx}$
34: $\mathbf{A}_{idx} = \emptyset$
35: break
36: END IF
37: END IF
38: END FOR
39: $a \leftarrow$Compute areas for each shroud in $\mathbf{A}$
40: END WHILE

$$\epsilon_k^L = \max\left(\left(\frac{r_i^L \cdot l_{i3}}{\text{std}\left(\left[A_{pq}^L\right]_{p=1,q=1}^{p,2}\right)}\right)_{i=1}^{n}\right) \tag{6.15}$$

where $l_{i3} \in \mathbf{Z}$ are the corresponding SIFT features of the group in the left camera with population size n, r_i^L is the Euclidean distance from the ith feature coordinate to the cluster centre coordinate of the left camera, and, ϵ_k^L, the score of the kth attentional shroud.

Finally, an attentional shroud (i.e. object part) is stored for each k group (where $k = 1, \ldots, K$) with a similar structure as in Equation (5.7):

$$\mathcal{H}^L = \left\{\mathcal{I}_k^L, \mathcal{X}_k^L, \epsilon_k^L, \left(A^L\right)_k, \mathbf{F}_k^L, \mathcal{E}_k^L\right\}_{k=1}^{K} \tag{6.16}$$

with $\mathcal{I}_k = 0$ to indicate that the observed object does not exist in the database knowledge, $\mathcal{X}_k$ as defined in Equation (6.7), and $\mathcal{E}_k^L = \{\Phi_{view}\}$ (i.e. the angular position of the turn table for the optimal view). Similarly, $\mathbf{F}_k$ stores those visual features associated to the kth group as follows (see Equation 6.5):

$$\mathbf{F}_k = \begin{bmatrix} \left[\left(z_t^L\right)_{i1}\right]_{i=1}^{a} & \left[\left(z_t^L\right)_{i9}\right]_{i=1}^{a} & [0]_{i=1}^{a} & \left[\left(z_t^L\right)_{ij}\right]_{i=1,j=2}^{a,5} & \left[\left(z_t^L\right)_{ij}\right]_{i=1,j=10}^{a,137} \\ \left[\left(z_t^R\right)_{i1}\right]_{i=1}^{b} & \left[\left(z_t^R\right)_{i9}\right]_{i=1}^{b} & [1]_{i=1}^{b} & \left[\left(z_t^R\right)_{ij}\right]_{i=1,j=2}^{b,5} & \left[\left(z_t^R\right)_{ij}\right]_{i=1,j=10}^{b,137} \end{bmatrix} \tag{6.17}$$

where a and b denote the population size of the kth group for the left or right camera, respectively, and $\mathbf{F}_k = [f_{ij}]_{i=1,j=1}^{n,135}$ (n is the population size of $\mathbf{F}_k$). Similarly, the first entry in $\mathbf{F}_k$ corresponds to the classification labels, the second is the angle pose where features are captured (i.e. Φ), the third determines where the SIFT feature are observed (either left, 0 or right, 1 camera), the fourth stored SIFT feature coordinates, and, finally, the last column, SIFT descriptors. It must be noted that disparity and motion vectors are only employed within the pre-attentive behaviour while formulating attentional shrouds and such are not stored in $\mathbf{F}_k$ as they are not considered in the overall learning behaviour. Figure 6.11 illustrates the output of this behaviour.

6.6.4 Hypotheses Generation - Visual Learning Case

The hypotheses generation behaviour defined in Section 5.5.2 is therefore extended according to the passed data structures of the defined behaviours in Sections 6.6.1.2 and 6.6.2. Specifically, this behaviour includes the following operational functions:

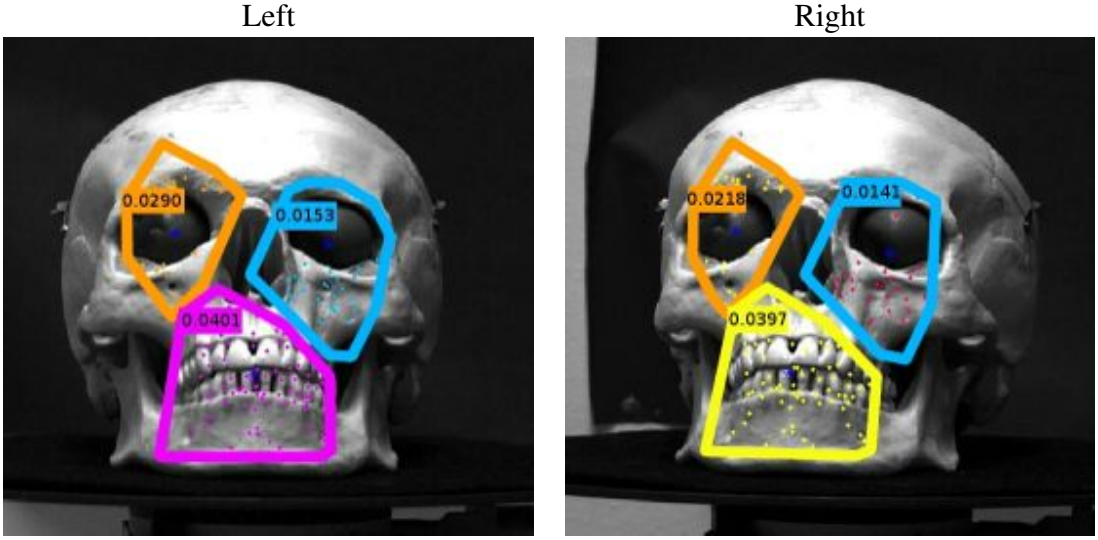


Figure 6.11: Attentional shrouds bound onto the object with their corresponding confidence score.

1. Visual exploration of a scene.

2. Formulation of attentional shrouds or tracking an object part (step 2 in Section 6.4).

The visual exploration mode holds the same behavioural properties defined in Section 5.5.2. When the learning behaviour is formulating attentional shroud hypotheses (Section 6.6.3), the created data structures for each camera (see Equation (6.16)) are compatible with the visual exploration mode inputs. However, the overall behavioural function is employed as defined in Section 5.5.2, it is slightly extended in order to generate random sample viewing angles for the investigation and learning of the object appearance.

To that end, an initial rotational saccade angle (defined in Section 6.7.3), Φ^j_{sacc} (where j indicates the number of the saccade which, in this case, is $j = 1$), is generated for all attentional shrouds found in the current pre-attentive cycle. Thus:

$$\Phi^j_{sacc} = \Phi_{view} + \phi \tag{6.18}$$

where $\phi = \{\phi \in \mathbb{R} | 4 \leq \phi \leq 8\}$ (these limits are determined empirically during the design of the learning behaviour). The sign of Φ^j_{sacc} is as well selected randomly. In the context of this chapter, Φ_{sacc} is defined as a rotational saccade where the learning strategy determines if the acquired visual information fulfils a learning halting criterion (as described in Section 6.7.3). A rotational saccade specifically captures visual information at a determined angular position and such is stochastically determined as described in Section 6.7.4. Thus, $\mathcal{H}^P_{new}$ (see Equation 5.11 and according to Section 5.5.2) is created with (cfr. Equations 5.12 and 5.13):

$$\mathbf{B}_i^L = \left[\left(\left[A_{p1}^L \right]_{p=1}^4 - \left[c_x \right]_{p=1}^4 \right)_i, \left(\left[A_{p2}^L \right]_{p=1}^4 - \left[c_y \right]_{p=1}^4 \right)_i \right] \tag{6.19}$$

and

$$\mathbf{B}_j^R = \left[\left(\left[A_{p1}^R \right]_{p=1}^4 - \left[c_x \right]_{p=1}^4 \right)_j, \left(\left[A_{p2}^R \right]_{p=1}^4 - \left[c_y \right]_{p=1}^4 \right)_j \right], \tag{6.20}$$

the translated attentional shroud convex hulls of the ith and jth object elements in sets $\mathcal{H}^L$ and $\mathcal{H}^R$. Similarly,

$$\mathcal{E} = \left\{ \Phi_{view}, \Phi_{view} + \beta, \Phi_{sacc}^j \right\} \tag{6.21}$$

where β is a discrete angle step ($\beta = 1$ degree as defined in Section 6.4). $\mathcal{E}$ is now defined as a list that stores information regarding the angular positions at which the object must be sampled for the kth attentional shroud. The remaining variables that are not mentioned herein are exactly defined as in Section 5.5.2. It must be pointed out that the visual learning behaviour does not consider salient features. Therefore, $\mathcal{H}^S$ is set as an empty set while learning the object's appearance.

6.6.4.1 Tracking an Attentional Shroud - Hypothesis Generation case

While learning an object part, the system is required to update only the fixation locations and the attentional shroud of the rotated target in the field of view of both cameras as follows. It is assumed that an object hypothesis exists and such hypothesis must be updated. Thus, $\mathcal{H}^L$ and $\mathcal{H}^R$ from Section 6.6.1.2 are employed to updated the ith tracked object hypotheses as follows (see Equations 6.6 and 5.11):

$$\left\{ \mathcal{H}_{new}^P \right\}_i = \left\{ \mathcal{I}_i, \eta_i, \mathcal{Y}^L, \mathcal{Y}^R, \mathbf{F}, \mathbf{B}^L, \mathbf{B}^R, \mathcal{E} \right\} \tag{6.22}$$

where $\mathbf{F} = \left[\mathbf{F}^L \cup \mathbf{F}^R \right]$ (expressed as in Equation 6.17), $\mathcal{Y}^L$ and $\mathcal{Y}^R$ as defined in Equations 5.14 and 5.15, $\mathcal{I}_i$ and η_i are set to infinity (as described in Section 6.7.2) and,

$$\mathbf{B}^U = \left[\left(\left[A_{p1}^U \right]_{p=1}^k - \left[c_x \right]_{p=1}^k \right)_i, \left(\left[A_{p2}^U \right]_{p=1}^k - \left[c_y \right]_{p=1}^k \right)_i \right], \tag{6.23}$$

(Section 6.6.1.2) where U denotes either the left, L, or right,R, camera. Finally, if there are "non-attended" objects in $\mathcal{H}^P$, the ith tracked object hypothesis is then added at the beginning of this set.

6.7 Attentive - Learning Case

While learning the object's appearance, the attentive abstract behaviour manipulates WHERE and WHAT behaviours to fulfil its specified visual tasks. That is, this behaviour selects the attentional shrouds that must be investigated and, in consequence, accumulates visual evidence across the object's view-sphere in each specified direction (as described in Section 6.4). The attentive behaviour is thus divided into two behaviours as follows:

1. Visual exploration of either the scene or object's view-sphere.
2. Visual learning of attentional shrouds.

The former carries out exploration tasks as designed in Section 5.7. In the visual learning behaviour, the exploration of learned canonical views is required in order to verify if such canonical views actually characterise the object's view-sphere. Therefore, each view of the object's view-sphere that is successfully targeted and verified is appended to $\mathcal{H}^A$ (Section 5.7.2). While actively exploring the object's view-sphere, the visual learning behaviour requires to rotate the object after each successful pre-attentive and attentive cycle since the current attended view pose has already been recognised. The defined *exploration-learning* strategy is enabled by setting $\mathbf{L} = \emptyset$ and $\mathbf{R} = \emptyset$ (as depicted in Figure 6.3). The exploration behaviour mode is extended in order to allow the passive acquisition of visual information while the turn-table is rotated if and only if $\mathcal{H}^P$ and $\mathcal{H}^S$ are empty. Section 6.5.1 outlines this extension.

The second operational mode allows the system to focus and verge on an attentional shroud (step 3 in Section 6.4) and, subsequently, to actively cluster (steps 4 and 6 in Section 6.4) gathered visual information within this shroud (Sections 6.7.1 and 6.7.3). While clustering SIFT features, it is measured the mutual information between previous and current rotational saccades. By computing the mutual information between rotational saccades, a halting criterion is establish in order to determine whether visual information characterise the appearance of the object or the learning behaviour must investigate further the current attentional shroud (Section 6.7.3). After all possible attentional shroud hypotheses have been investigated, the attentive behaviour invokes a knowledge consolidation behaviour (Section 6.7.5) in order to unite attentional shrouds into a single canonical view of the overall object's concept.

Similarly, Figure 6.12 depicts the structure of the attentive visual learning mode (second behavioural mode as described above). As observed, there are four possible feedback signals which are employed by the deliberative layer to determine the operation of the hierarchy of behaviours (Figure 6.3). In addition, feedback signals determine the output of the attentive behaviour as summarised in Figure 6.12.

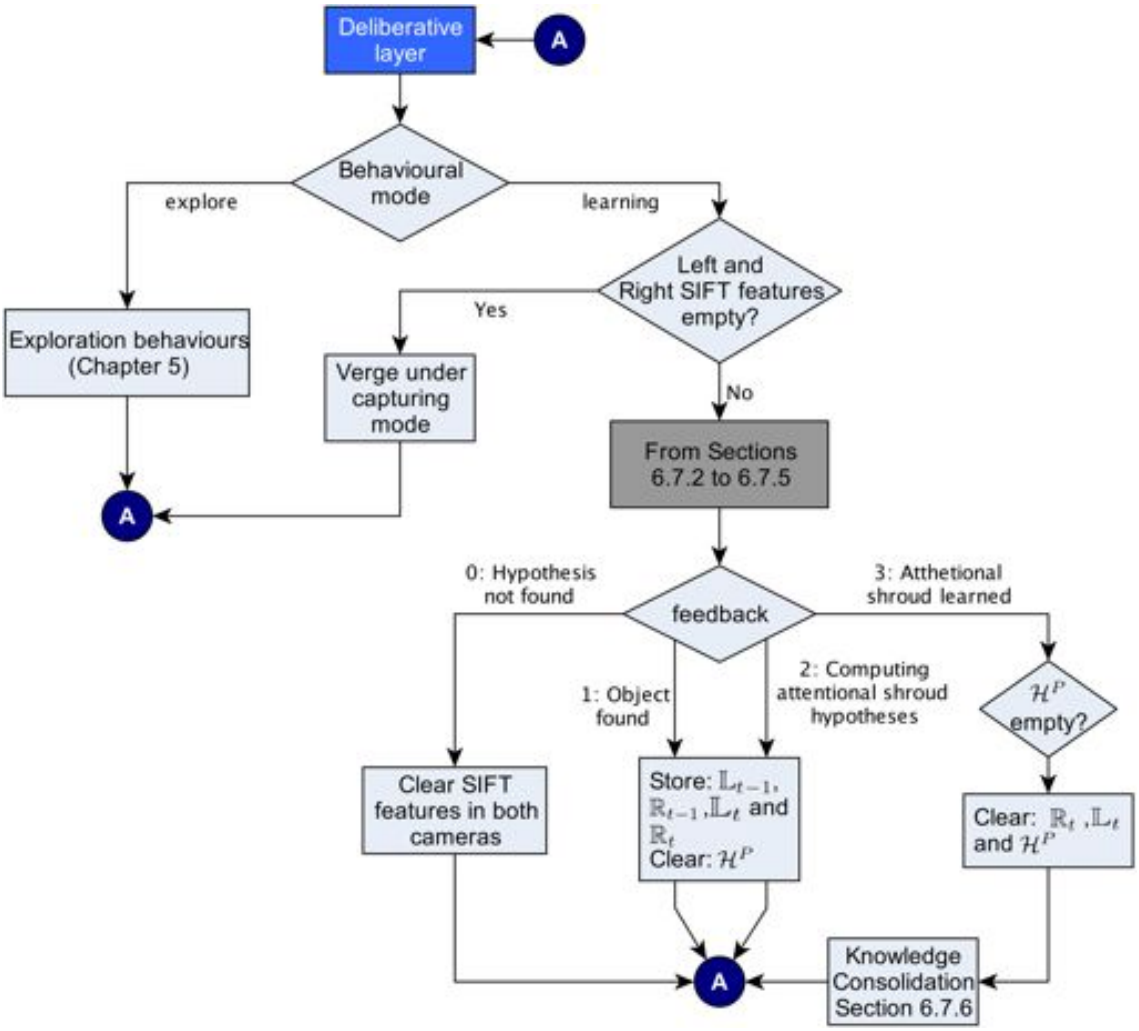


Figure 6.12: Attentive behaviour while learning.

6.7.1 Saccadic Targeting - Learning Case

The saccadic targeting behaviour while learning is further divided into two operational cases:

1. While the *learning* mode in the high-level layer (as described in Section 6.4) is selected and not enough visual information has been acquired to learn an object (i.e. $\mathcal{H}^P$ is empty), the system passively captures images and extract SIFT features (Sections 6.5.1 and 6.7). Thereafter, this operational case sets that the pre-attentive behaviour is gathering visual information in order to formulate attentional shroud hypotheses ("2: *Computing attentional shroud hypotheses*" case in Figure 6.12).

2. Attentional shroud hypotheses have been formulated, and the system must target them and learn SIFT features within the selected attentional shroud.

The second case initially follows the behaviour defined in Section 5.7.1. That is, the most confident attentional shroud hypothesis in $\mathcal{H}^P$ is selected and such is attended and verged on. This second case also contains object verification routines in order to confirm the identity of the fixated attentional shroud and, in consequence, to determine which SIFT features are geometrically stable (Sections 6.7.2). Similarly, this case governs the active clustering of visual evidence and controls the object's view pose being learned (Section 6.7.3). However, if the attentional shroud is successfully targeted but it cannot be geometrically verified (Section 6.7.2), this behaviour determines whether to invert the direction of angular movement or the selected attentional shroud is removed from $\mathcal{H}^P$ and the the feedback signal sent to the attentive behaviour is set to "3: *Attentional shroud learned*" case (as depicted in Figure 6.12).

There might be cases while targeting the cameras in which the attentional shroud does not contain enough SIFT features to verge on a target as the shroud becomes degenerate (i.e. SIFT features within the shroud are not sufficiently distinctive in order to be matched between cameras or the pre-attentive behaviour did not successfully track the attentional shroud). Therefore, the selected attentional shroud is removed from $\mathcal{H}^P$ and, subsequently, the feedback signal sent to the attentive behaviour is set to "3: *Attentional shroud learned*" case (as depicted in Figure 6.12). The latter allows the system to continue learning the object's view-sphere despite false positive shrouds.

6.7.2 Object Verification - Learning Case

This behaviour is responsible for geometrically verifying SIFT features within ith selected attentional shroud in $\mathcal{H}^P$ (i.e. *Goodness of recognition* learning principle in Section 6.3.1). The shroud's bounding contour is thus employed to selectively bias attention to a specific region of interest. To that end, $\mathbf{B}^L$ and $\mathbf{B}^R$ (Equations 6.19 and 6.20, respectively) delineate the region of interest on each camera and only those extracted SIFT features (either $\mathbf{L}$ or $\mathbf{R}$, respectively) inside the segmented regions are employed.

Hence, the candidate attentional shroud is thus verified as described in Section 5.7.2. Thereafter, $\mathcal{M}^{(\mathbf{F},\mathbf{L})}$ or $\mathcal{M}^{(\mathbf{F},\mathbf{R})}$ are employed to assess the spatial configuration between SIFT features of the current object's pose (either $\mathbf{L}$ or $\mathbf{R}$) and the learned visual features $\mathbf{F}$ (which will be defined by the Equation 6.25). This process is carried out by means of the standard object recognition pipeline described in Section 2.3.1. That is, the Generalised Hough Transform finds those SIFT features that are geometrically consistent with the attentional shroud reference centroid and, in consequence, the affine pose estimator computes the required transformation to project each ith feature coordinate in $\mathbf{F}$ into either the left or right camera respectively. A MSE (Section 4.6) less than 20 pixels, determines that an observed SIFT feature in either

the left or right camera is geometrically consistent with the learned representation, and this feature is classified as stable and, thence, learned. Accordingly, an empirical investigation establishes that the above global MSE threshold performs well in practice.

If the candidate attentional shroud is geometrically verified and the total population size of SIFT feature matches is greater than 10 features in either the left or right camera, matched SIFT features in each camera are thus clustered as described in the following section ("1: *Object found*" case in Figure 6.12). Otherwise, the attentional shroud is removed and the saccadic targeting behaviour thus returns to the attentive behaviour that "0: *Hypotheses not found*" (as depicted in Figure 6.12).

On the contrary, if the attentional shroud is verified but it is not geometrically consistent, the SIFT features of the current object's pose are not considered in the active clustering process and, thence, the object is rotated in order to obtain a new pose view (as described in the following section). In other words, the extracted features for the current object's pose are inadequate to characterise this particular view as they might be potential outliers.

Finally, the object pose class, $\mathcal{I}$, and the confidence score, η in the ith selected attentional shroud in $\mathcal{H}^P$ are set to infinity in order to indicate the attentive behaviour that this attentional shroud is currently investigated and learned. Accordingly, $\mathcal{M}^{(\mathbf{F},\mathbf{L})}$, $\mathcal{M}^{(\mathbf{F},\mathbf{R})}$ and $\mathcal{H}^P$ are thus returned to the attentive behaviour.

6.7.3 Active Clustering

The system's ability to learn canonical representations of the observed "unknown" object relies on the accuracy of the intrinsic properties that characterise the object over a period of time. The devised visual learning behaviour enables the system to find stable visual features (i.e. SIFT) that describe the object's appearance. In that regard, the active clustering behaviour specifically groups features that are geometrically stable and observe spatio-temporal continuity over observed sequence of views (i.e *Familiarity* learning principle in Section 6.3.1). The specific purpose of this behaviour is to group SIFT features that present the described properties.

To that end, the active clustering behaviour consists of grouping stable SIFT features (Section 6.7.2) with respect to their classification labels (see Equation 6.5 in Section 6.6.1.1). That is, the objective function of this active clustering process is symbolically defined in terms of the matching operations: $\mathcal{M}^{(\mathbf{F},\mathbf{L})}$ or $\mathcal{M}^{(\mathbf{F},\mathbf{R})}$ of Section 6.7.2. Each matched feature in $\mathbf{L}$ and $\mathbf{R}$ is thereby assigned to the corresponding match class labels given in $\mathbf{F}$; i.e. $[f_{i1}]_{i=1}^{n} : \ f_{i1} \in \mathbf{F}$, where n is the population size of $\mathbf{F}$.

Knowledge scalability while learning the object's appearance (Section 6.2) is achieved by allocating new classification labels to those SIFT features where $\neg\mathcal{M}^{(\mathbf{F},\mathbf{L})}$ and $\neg\mathcal{M}^{(\mathbf{F},\mathbf{R})}$ (unmatched features with cardinalities a and b, respectively) hold true. Thus,

$$\mathbf{X}^L = \mathbf{L}\left(\neg\mathcal{M}^{(\mathbf{F},\mathbf{L})}\right) \text{ and } \mathbf{X}^R = \mathbf{R}\left(\neg\mathcal{M}^{(\mathbf{F},\mathbf{R})}\right) \tag{6.24}$$

where $\mathbf{X}^U = \left[x_{ij}^U\right]_{i=1,j=1}^{a,132}$ (see Equations 5.1 and 5.2 and U denotes either the left, L, or right, R SIFT features). These unmatched features are thereby appended to $\mathbf{F}$ as follows (ref. Equation 6.17):

$$\mathbf{F} = \begin{bmatrix} [f_{i1}]_{i=1}^{n} & [f_{i2}]_{i=1}^{n} & [f_{i3}]_{i=1}^{n} & [f_{ij}]_{i=1,j=4}^{n,7} & [f_{ij}]_{i=1,j=8}^{n,135} \\ (n+i)_{i=1}^{a} & [\Phi_t]_{i=1}^{a} & [0]_{i=1}^{a} & \left[\left(x^L\right)_{ij}\right]_{i=1,j=1}^{a,4} & \left[\left(x^L\right)_{ij}\right]_{i=1,j=5}^{a,132} \\ (n+a+i)_{i=1}^{b} & [\Phi_t]_{i=1}^{b} & [1]_{i=1}^{b} & \left[\left(x^R\right)_{ij}\right]_{i=1,j=1}^{n,4} & \left[\left(x^R\right)_{ij}\right]_{i=1,j=5}^{132} \end{bmatrix} \tag{6.25}$$

where Φ_t is the current attended angular pose. Similarly, the active clustering behaviour maintains in working memory a list logically indexed to $\mathbf{F}$ that contains classification label occurrences. This enables the system to register if a SIFT feature is matched across the investigated portion of the object's view-sphere. In consequence, the learning behaviour is not overloaded with false positive SIFT features and, moreover, it only stores stable spatio-temporal relationships. Therefore, visual information that has been appended to $\mathbf{F}$ is removed from $\mathbf{F}$ if it is not matched in subsequent views and such is not further considered in the clustering process.

While investigating and learning an attentional shroud, $\mathbf{F}$ is stored in working memory at specific episodic angular positions. These episodic positions have beed termed as *rotational saccades* and they are expressed as Φ_{sacc} (as defined in Section 6.6.4).. In addition, rotational saccades serve as reference points where the active clustering behaviour determines whether the learning behaviour:

- continues gathering new visual information,
- investigates a different side of the "unknown" object (as described in Section 6.4), or
- stops learning the current attentional shroud.

Therefore, for each rotational saccade, $\mathbf{F}_{prev}$ and $\mathbf{F}$ denotes learned features for the previous and current rotational saccade, respectively. $\mathbf{F}_{prev}$ is only stored in working memory if the current attentive cycle is indeed investigating and learning a rotational saccade.

The active clustering behaviour thus measure the mutual information between classification labels in $\mathbf{F}_{prev}$ and $\mathbf{F}$ (i.e. *Familiarity* learning principle in Section 6.3.1). As described in Section 6.2, the evaluation of the mutual information between trained and current views observes biologically motivated principles. Specifically, Pluim et al. (2003) describe the use of mutual information in the context of medical image registration. That is, an image is similar if and only if the measured mutual information between images is approximately equal to their measured joint entropy distribution. In this chapter, the computation of mutual information between two saccades follows the algorithmic steps defined in (Pluim et al., 2003). Specifically, the mutual information between rotational saccades is the amount that the uncertainty in $\mathbf{F}$ is reduced when visual information in $\mathbf{F}_{prev}$ is known. Hence, classification labels in $\mathbf{F}$ and $\mathbf{F}_{prev}$ are quantised in frequency histograms; g and g_{prev}, respectively (as illustrated in Figure 6.13(a)). The mutual information is thus expressed as follows (Pluim et al., 2003):

$$I\left(g_{prev}, g\right) = H\left(g_{prev}\right) + H\left(g\right) - H\left(g_{prev}, g\right) \tag{6.26}$$

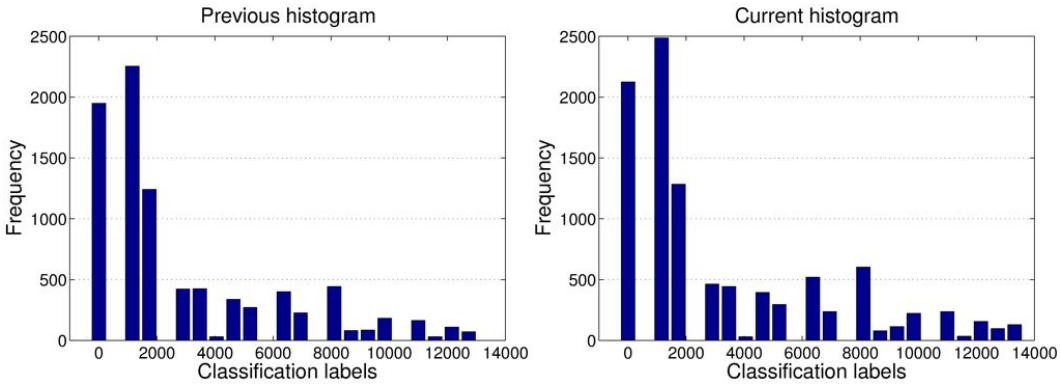
where $I\left(g_{prev}, g\right)$ is the measured mutual information, $H\left(g_{prev}\right)$ and $H\left(g\right)$ are the univariate entropy distribution; and $H\left(g_{prev}, g\right)$ is the joint entropy distribution. The halting criterion is thereby cast in terms of a null hypothesis definition expressed as follows:

$$H_0^{MI} = \begin{cases} \text{true} & \text{If } \left|I\left(g_{prev}, h\right) - H\left(g_{prev}, h\right)\right| < \mathcal{T}_{MI} \\ \text{false} & \text{Otherwise} \end{cases} \tag{6.27}$$

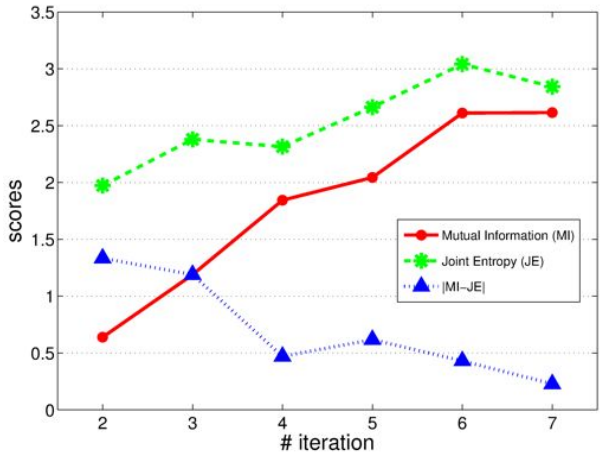
where $\mathcal{T}_{MI} = 0.3$ for the initial rotation direction, whereas $\mathcal{T}_{MI} = 0.2$, when the rotation direction is modified[9]. Figure 6.13(b) depicts the minimisation of the mutual information with respect to the joint entropy distribution while inspecting an attentional shroud across a portion of a unknown object's view-sphere. By inspecting Figure 6.13(b), it can be observed that the object has been inspected in both directions. That is, the inflection point at the fourth rotational saccade denotes the change of direction of movement, i.e. $H_0^{MI} < \mathcal{T}_{MI} = 0.3$. Therefore, the next rotational saccade (the fifth rotational) presents increments on the difference between mutual information and joint entropy; this denotes that the direction change allows the system to gather new visual information. The system tracks the attentional shroud until the second threshold definition is achieved, i.e. $H_0^{MI} < \mathcal{T}_{MI} = 0.2$, and, in consequence, the system selects the next putative attentional shroud or consolidates acquired knowledge (Section 6.7.5). A second threshold is set lower than the initial threshold in order to enable the system to tune learned SIFT descriptors over a different object's view poses.

Similarly, classification labels that are subsequently assigned do not observe large amounts of occurrences (labels greater than 2000, as observed in Figure 6.13(a)). It is therefore inferred

[9]From empirical validation, these thresholds are found to work well in practice.



(a)



(b)

Figure 6.13: (a) Previous and current histograms, and (b) mutual information and joint entropy over different rotational saccades.

that initial learned features (labels less than 2000) present stable spatio-temporal descriptions of the object and, in consequence, the learning behaviour is biased to the initial angle view. This bias, however, does not affect the robustness of the system while performing exploration tasks (as demonstrated in Chapter 7).

Hence, $\mathbf{F}_{prev}$, $\mathbf{F}$ and H_0^{MI} denote the output of this behaviour and are then passed to the attentive behaviour. Similarly, this behaviour returns the message: "3: *Attentional shroud learned*" (as depicted in Figure 6.12) in order to indicate the behaviour above in the hierarchy the current state of the system.

6.7.4 Sampling the view-sphere of an Object

This primitive behaviour allows the system to rotate the object on the turn-table and, in consequence, acquire stereo-pairs and extract SIFT features (Section 6.5.1). Similarly, this behaviour also determines the direction of movement of the object while investigating and learning

attentional shroud hypotheses (as described in Section 6.4). Hence, this behaviour consists of three operational modes, described as follows.

The first mode involves rotating the object at discrete angular steps while learning the object's view-sphere. To that end, Equation 6.21 is updated accordingly as follows (see Equations 6.18 and 6.21):

$$\mathcal{E} = \left\{\Phi_{view}, \Phi^1_{sacc}, \ldots, \Phi_{t-1}, \Phi_t + \beta, \Phi^{i+1}_{sacc}\right\} \tag{6.28}$$

where i is the number of attended rotational saccades thus far; Φ_{t-1} and Φ_t, the previous and current attended angle pose at t time steps respectively; and β, the discrete angular step ($\beta = 1$ degree, ref. Section 6.4). $\mathcal{E}$ is thereby updated in the currently attended attentional shroud, $\mathcal{H}^P$. As observed in Equation 6.28, all attended angles poses are stored in order to maintain a record of visited angular positions of the object's view-sphere for the inhibition of return and knowledge consolidation behaviours. Thereafter, the system rotates the object, acquires images and extracts SIFT features (capturing primitive behaviour in Section 6.5.1).

The second operation mode consists of determining the rotational saccades each time one is reached. In consequence, the system has to rotate the object by β discrete angular steps towards the defined rotational saccade. Therefore, the new rotational saccade is determined as follows:

$$\Phi^{i+1}_{sacc} = \Phi^{i}_{sacc} + \phi \tag{6.29}$$

where ϕ is a random angle expressed as: $\phi = \{\phi \in \mathbf{R} | \mathrm{s} \cdot 4 \leq \phi \leq \mathrm{s} \cdot 8\}$ where s gives the sign which depends on the current direction of movement of the object. The direction sign is randomly determined as described in Section 6.4). Thereafter,

$$\mathcal{E} = \left\{\Phi_{view}, \Phi^1_{sacc}, \ldots, \Phi_{t-1}, \Phi^{i}_{sacc} + \beta, \Phi^{i+1}_{sacc}\right\} \tag{6.30}$$

and $\mathcal{E}$ is accordingly updated in the currently attentional shroud,$\mathcal{H}^P$. The system then rotates the object and, subsequently, acquires images and extracts SIFT features (capturing primitive behaviour in Section 6.5.1).

Finally, the third mode controls the direction of the object's movement. That is, if and only if H^{MI}_0 (see Equation 6.27) holds true, the direction sign is modified (Equation 6.29). Thus,

$$\mathcal{E} = \left\{\Phi_{view}, \Phi^1_{sacc}, \ldots, \Phi_{t-2}, \Phi^{i}_{sacc}, \Phi_{view} + \phi\right\} \tag{6.31}$$

where $\phi = \{\phi \in \mathbf{R} | \mathrm{s} \cdot 4 \leq \phi \leq \mathrm{s} \cdot 8\}$ with $\mathrm{s} = -1\mathrm{s}$. Thence, the robot vision system targets

and verges on the attentional shroud of the initial view, Φ_{view} (Section 6.6.4).

6.7.5 Knowledge Consolidation

Knowledge consolidation in the human brain (in accordance with (Styles, 2005)) refers to the process that allows to store in memory an enduring event. This behaviour in the hierarchical architecture exactly allows this process.

Specifically, after the system has investigated all putative attentional shrouds that characterise a portion of the object's view-sphere (point 9 in Section 6.4), it is required to create a canonical view (as defined in Section 6.3) that captures the object's appearance for a given pose (i.e *Familiarity* and *Functionality* learning principles in Section 6.3.1). To that end, this behaviour selects the most representative view and, subsequently, consolidates clustered features.

Hence, the choice of the most representative view consists in quantising the population size of learned features in each attentional shroud (i.e. $(\mathbf{F})_i^n$ where n is the number of attentional shrouds in $\mathcal{H}^P$) with respect to the recoded angle poses (see Equations 6.28, 6.30 and 6.31). The highest peak in the computed histogram of frequencies of learned features is selected and its corresponding recorded angle pose therefore determines the canonical view.

The canonical view, Φ_c, is thus employed as the reference angle pose to create a code-book of clustered features. Thus, for each cluster found (Section 6.7.3), it is queried if a feature in the surveyed group (denoted as $\mathbf{X}_i = \Big((\mathbf{F})_j\Big)_i$, where j denotes the surveyed cluster) has been observed in the canonical view (i.e. a canonical feature, $[x_{cq}]_{q=1}^{135}$, where c is the canonical feature entry in the group). In that regard, there are three possible cases as briefed below:

1. *More than one canonical feature exists in the group*. This can be caused by the presence of outliers and, therefore, they have to be identified and removed. To that end, the Euclidean distances are computed between all non-canonical features descriptors and each canonical feature descriptor in the group. The minimum Euclidean distance therefore specifies the optimal canonical feature whilst the other canonical features are removed (as these are outliers). Thereafter,

$$\mathbf{X}_i = \Big[\; [x_{p1}]_{p=1}^{r} \quad [\theta_c]_{p=1}^{r} \quad [x_{p3}]_{p=1}^{r} \quad \Big[\; x_c \quad y_c \quad \sigma_c \quad \theta_c \; \Big]_{p=1}^{r} \quad [x_{pq}]_{p=1,q=1}^{r,135} \; \Big] \tag{6.32}$$

   where $r = 1, \ldots,$ population size of the jth surveyed group. This specifies that the canonical angle, the camera where the canonical feature is observed and the corresponding SIFT feature components, (x, y, σ, θ), are assigned to all members in the group.

2. *One canonical feature exists*. This case only consists in assigning the canonical locations into all the members of the group as in Equation 6.32. Finally, each updated group is returned into an array.

3. *No canonical feature exists*. This depicts the case when SIFT features are not registered in the canonical angle pose. As a possible extension of the visual learning behaviour, this case might include non-canonical features since these depict regions that might be occluded and, in consequence, cannot be observed from the canonical view. The affine pose estimator in Section 4.6 can be used in order to approximate the required transformation to project these features into the canonical view.

Finally, the consolidation of clustered learned features consists of computing the statistical mean value of the features (i.e. keypoint and descriptor) ofeach cluster in $\mathbf{X}$. In other words, the mean value for each feature found depicts a codebook for the current canonical view. $\mathbf{F}$ is therefore stored in $\mathcal{H}^A$ as follows (see Equation 5.11):

$$\mathcal{H}^A = \{\mathcal{I}_i, \eta_i, \mathcal{Y}_i^L, \mathcal{Y}_j^R, \mathbf{F}, \mathbf{B}_i^L, \mathbf{B}_j^R, \mathcal{E}\} \tag{6.33}$$

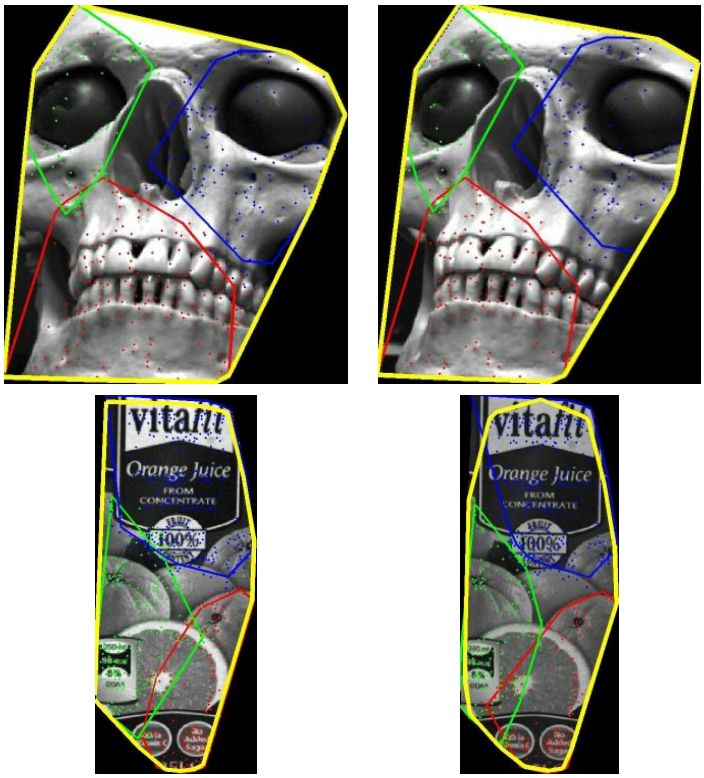
where $\mathcal{I}_i = \eta_i = \mathcal{Y}_i^L = \mathcal{Y}_j^R = \mathbf{B}_i^L = \mathbf{B}_j^R = \emptyset$, $\mathbf{F} = (\mathbf{X}_i)_{i=1}^n$ (n is the number of attentional shrouds in $\mathcal{H}^P$) and $\mathcal{E}$ is the recorded the view-sphere samples while learning the canonical view (ref. Equations 6.28, 6.30 and 6.31). Figure 6.14 illustrates how the system internally represents a learned canonical view. Thus, $\mathcal{H}^A$ is the output of this behaviour that is then passed to the attentive behaviour.

6.8 Inhibition of Return

In the context of the visual learning behaviour, inhibition of return is only required in order to suppress attended view poses. In that respect, this abstract behaviour is divided into two behavioural functions: a) while exploring a scene as described in Section 5.6 and b) while learning the object's appearance. The second function is described as follows.

In the learning case, the system might be either:

- formulating attentional shrouds or tracking an attentional shroud, or,
- actively exploring a learned canonical view (as described in Section 6.7.1).



Left Right

Figure 6.14: Canonical views of two object for the left and right camera.

The former consists in specifying: $\mathcal{H}^P = \mathcal{H}^P_{new}$, while the latter must inhibit those angle poses where the canonical view has been verified. Thus, for each verified object pose while exploring the learned canonical view, Φ is contained in $\mathcal{E}$ (such that $\mathcal{E} \in \mathcal{H}^A$) is registered in a view-sphere map, Ψ, in order to register the angles where the object has been observed while learning or exploring the view-sphere.

6.9 Conclusions

This chapter presents a semi-autonomous visual object appearance learning behaviour that allows the investigated active binocular robot vision system to synthesise and characterise automatically its own part-based object representation knowledge from multiple observations while a human teacher indicates the object and supplies a classification name. Specifically, the hierarchy of visual behaviours is extended in order to enable the robotic vision system to find those objects' poses that are relevant for visual exploration of a scene. Therefore, this behaviour avoids the need to build manually segmented databases which, as demonstrated in Chapter 5, produces incorrect localisations while searching a scene.

The visual learning behaviour adopts biological motivated principles to drive the learning

strategy. That is, the active binocular robot vision system builds and synthesises canonical pose representations that are preserved despite linear transformations within the object's view-sphere (i.e. *goodness of recognition*). Likewise, knowledge is acquired by the active interaction with the object (i.e. *familiarity* and *functionality*), such that the system is trained according to the observed view poses. The primitive representations are finally consolidated by aggregating observed visual features across the object's view-sphere.

This chapter also serves as a validation of the hierarchical architecture proposed in Chapter 5. The definition of a visual learning of the object's appearance behaviour qualitatively demonstrates the robustness of the hierarchical architecture. That is, visual behaviours are integrated within the architecture and visual behaviours devised in this chapter acquires the properties of the hierarchical architecture (as described in Sections 5.3 and 5.10). Hence, the hierarchical architecture now features visual behaviours for object and feature tracking, visual binding, active clustering of acquired knowledge that operates parsimoniously with previously defined behaviours. Furthermore, they can potentially enable the system to accomplish a number of different complex tasks; for example, *smooth pursuit*, location learning for navigation to name a few.

The following chapter is thus concerned with the overall validation of the visual learning behaviour in terms of the repeatability of finding similar canonical views regardless the initial viewing pose. Similarly, this devised behaviour is further demonstrated in an active exploration of a scene task. These experiments specifically consist of evaluating the robustness of the learned representations against manually segmented databases in a visual search task.

Chapter 7

Visual Learning Behaviour and Visual Search Experiments

This chapter presents a study on the hierarchical active binocular robot vision system architecture in two different contexts: while semi-automatically learning object appearances and while actively exploring the contents of a scene. The aim of this chapter is twofold. On the one hand, the visual learning behaviour is validated in terms of the stability, repeatability and the underlying properties of canonical views (Section 6.3.1). Likewise, the performance of the learned object's canonical views is compared to a manually annotated ground truth in a passive pre-attentive mode while the object is rotated over the viewing sphere. On the other hand, the active visual exploration of a scene is validated with learned and manually segmented object databases in order to compare the overall performance of the robot vision system with different knowledge representations.

7.1 Introduction

A manually segmented database has been demonstrated in Section 3.8.3 to produce false positive identifications in a visual search task. Specifically, object models stored in such database bias the adopted attentional spotlight metaphor in conjunction with the "stepping-stone" search pattern and, in consequence, incorrect identifications appear as soon as the the observed object deviates from the canonical representation. Furthermore, such views might contain outliers that affect the detection of object instances. This chapter therefore presents a complete investigation of the semi-autonomous binocular object appearance learning behaviour that enables the robot vision system to create its object representations. Similarly, the learned

Task-goal Behaviour	Experiment	Section
Visual object appearance learning	Training objects	7.2.1
	Pre-attentive localisation	7.2.2
Autonomous scene exploration	Visual search task	7.3.1
	Binocular vergence	7.3.2

Table 7.1: Table of the experiments conducted in this chapter.

object database are validated in an autonomous visual exploration task and compared against manually pre-trained and segmented object databases. Table 7.1 shows the list of experiments conducted in this chapter.

7.2 Visual Learning Behaviour Experiments

The previous chapter describes the design of a set of visual behaviours within the hierarchical architecture in order to actively learn the object's appearance while a human teacher indicates the object and supplies a classification name. The operation of the active visual learning behaviour is inspired by the learning principles described in Section 6.3, and this behaviour equips the robot vision system with the required visual competences and abilities to explore an object over different viewpoints. The aim of this behaviour is to create a condensed but descriptive object knowledge database that captures the intrinsic properties of the observed objects. Thus, the system actively acquires robust object part representations that are then consolidated as canonical views over the viewing sphere. The working assumption is therefore: *object parts defined as meaningful visual representations of the object surrounded by attentional shrouds provide a better representation using smaller parts than considering one big attentional shroud (i.e. manual segmentation) of the object from the scene*.

This section therefore presents the validation of the visual learning behaviour in two different contexts:

1. While learning the object appearance by means of the devised visual learning behaviour (Chapter 6).
2. While pre-attentively localising single, isolated instances of the recently learned object representations across the objects' viewing-sphere.

The following sections describe the methodology and results of the above described experiments.


1: Bear

2: Cigarette box 1


4: Vase

5: Cigarette box 2

6: Chocolate

7: Juice box 1

8: Juice box 2


9: Skull

3: Soup can

10: Juice box 3

Figure 7.1: The ten objects employed in the experiments of this chapter and their corresponding object class numbers.

7.2.1 Learning the Object Appearance

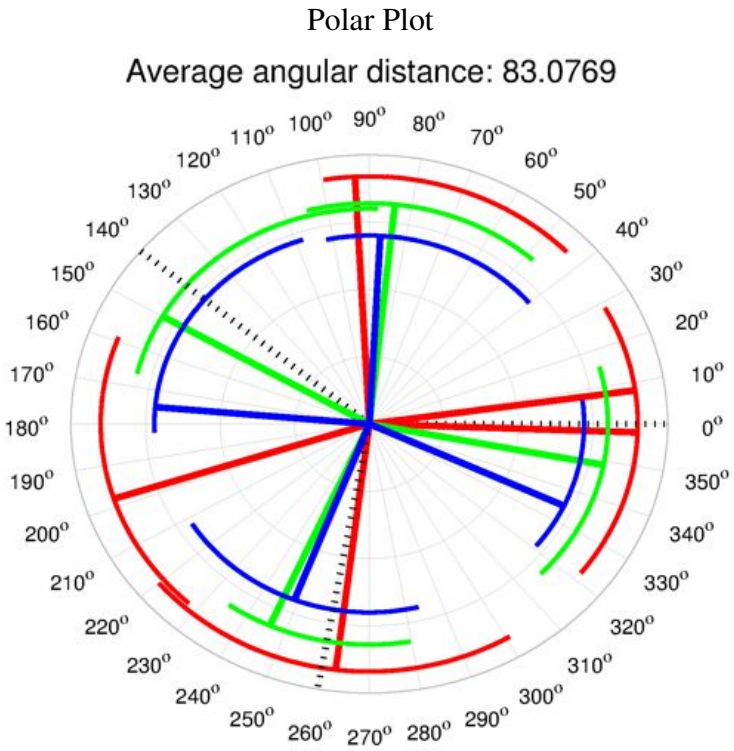
To validate the integration of the overall visual behaviour, it is proposed to allow the robot vision system to learn 10 objects by means of the visual learning behaviour described in the previous chapter. These 10 objects have been selected arbitrarily from everyday objects of different texture compositions and geometrical shapes. Similarly, the *goodness of recognition, functionality* and *familiarity* of the object properties (as defined in Section 6.3.1) are also tested. The hypotheses considered in these experiments are described as follows.

A canonical representation is determined by clustering spatio-temporal properties of the imaged object such that it extracts and finds local representative feature groups. The centre coordinates of each feature group are considered as endogenous attentional seeds in order to produce putative object parts that can be attended (attentional shrouds). Thus, *attending those salient informative object parts, the system is expected to learn and represent potentially the most informative, view-point invariant canonical view of an object*. Similarly, an "explorative" active paradigm (i.e. camera-eye movements) improves the robustness of finding canonical object parts. Thus, *canonical representations become stable by gaining more visual informa-*

tion while saccading to structural representative object parts (i.e. active interaction with the object).

To test the validity of the above assumptions, the system formulates attentional shrouds (Section 6.6.3), attends to them, and, finally, learns canonical representations of the unknown object (Section 6.7.5). To that end, ten objects are considered in this experiment (as depicted in Figure 7.1); each object presents different visual properties such as shape, texture and so forth.

The experimental methodology thus consists of invoking the visual learning behaviour three times on the same object over different initial view-points. Polar plots of each object are employed to qualitatively measure the degree of similarity and repeatability between observed views. Thus, this polar representations objectively illustrate the range of observations required to determine the most informative canonical views of the explored object. On the contrary, the repeatability of the behaviour to select similar canonical views over the viewing sphere regardless the initial pose is estimated in terms of the statistical mean and confidence intervals of the found canonical views for each visual learning behaviour invoked. These results are summarised for each object depicted in Figure 7.1 from Figure 7.2 to Figure 7.11.



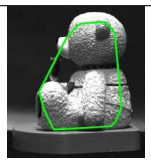
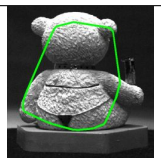
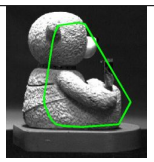
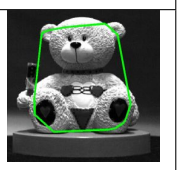
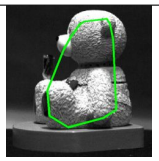
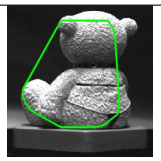
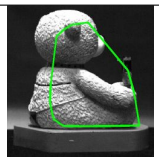
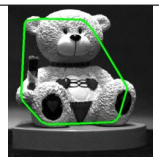
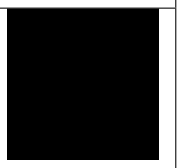
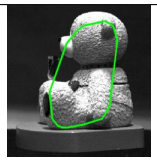
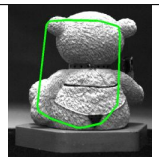
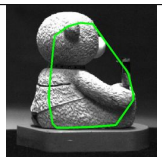
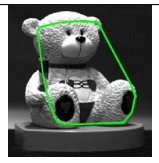
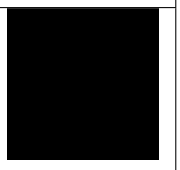
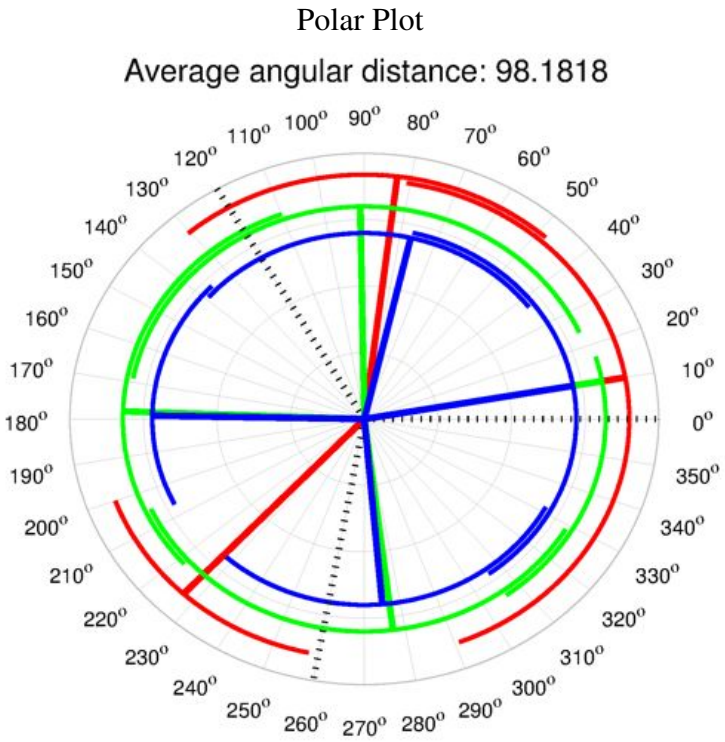
Experiment	**Canonical Views**				
#1: red	93^o	198^o	263^o	358^o	368^o
					
#2: green	84^o	150^o	246^o	349^o	–
					
#3: blue	87^o	175^o	249^o	335^o	–
					
Confidence Interval (@95%)	88 ± 5.1	174.3 ± 27.1	252.6 ± 10.2	352.5 ± 13.7	–

Figure 7.2: Polar plot and canonical views of the *"Bear"* object. Black dotted lines in the polar plot denote the initial angular view point where visual learning behaviour is invoked. Red, Green and Blue lines denote the number of experiment described in the bottom table.



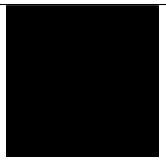
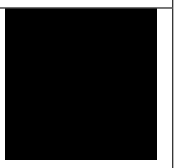
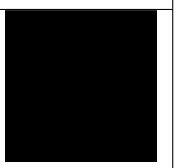
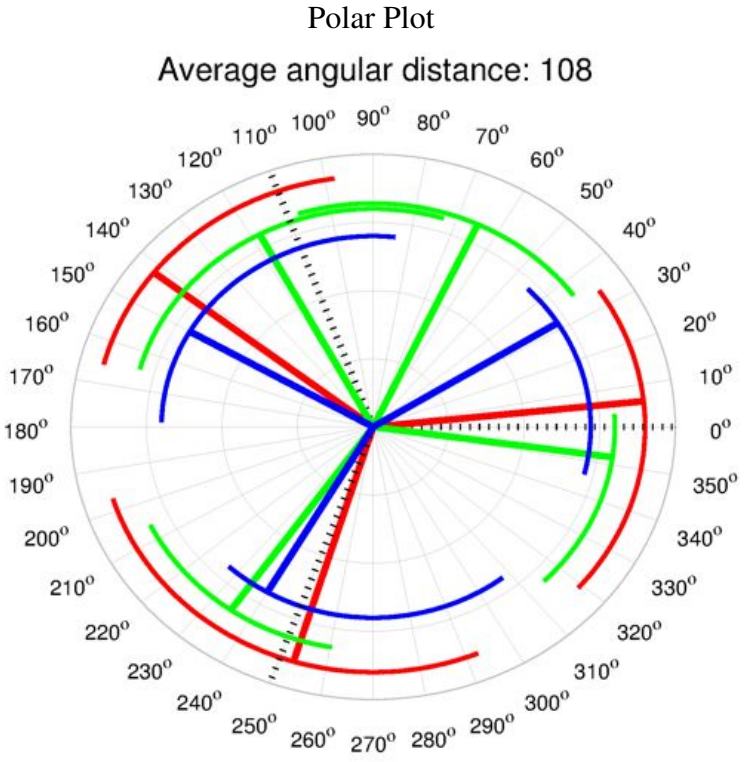
Experiment	Canonical Views				
#1: red	10^o	83^o	–	–	227^o
					
#2: green	10^o	91^o	178^o	277^o	–
					
#3: blue	10^o	77^o	179^o	275^o	–
					
Confidence Interval (@95%)	10 ± 0.0	86.6 ± 7.9	178.5 ± 0.9	276 ± 1.9	227 ± 0.0

Figure 7.3: Polar plot and canonical views of the *"Cigarette box 1"* object. Black dotted lines in the polar plot denote the initial angular view point where visual learning behaviour is invoked. Red, Green and Blue lines denote the number of experiment described in the bottom table.



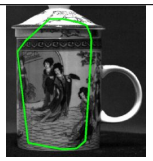
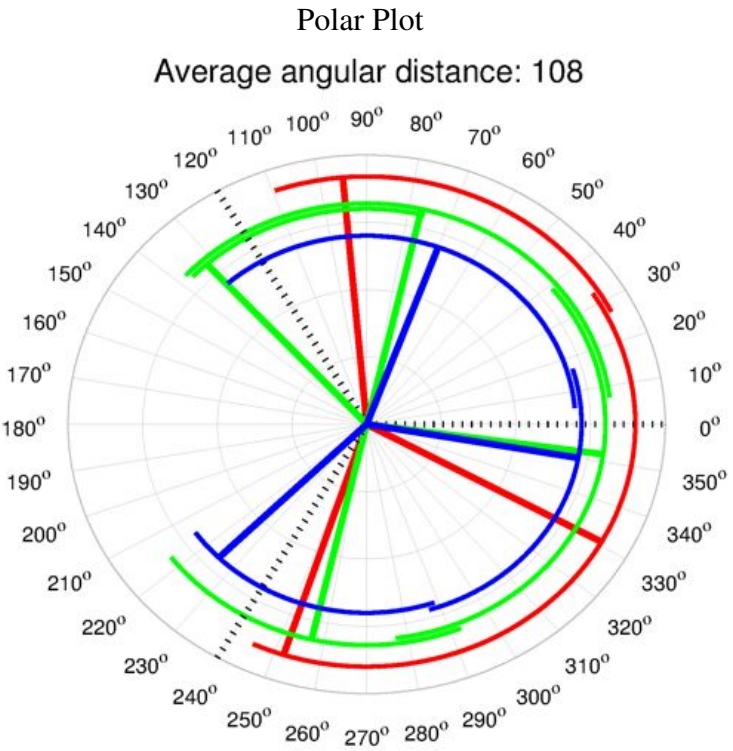
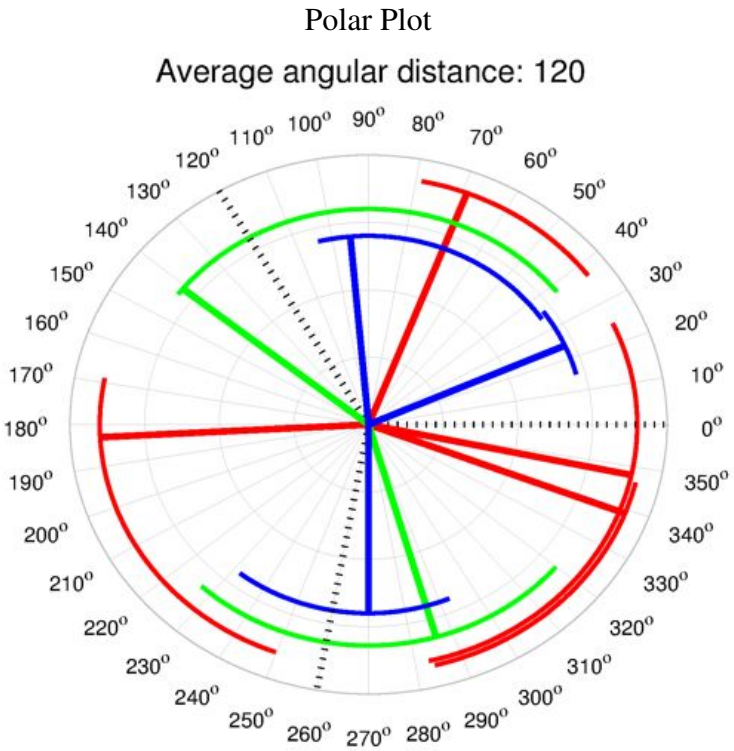
Experiment	**Canonical Views**			
#1: red	6^o	–	142^o	253^o
				
#2: green	-8^o	65^o	118^o	235^o
				
#3: blue	32^o	–	150^o	240^o
				
Confidence Interval (@95%)	10 ± 22.96	65 ± 0.0	136 ± 18.8	242 ± 10.5

Figure 7.4: Polar plot and canonical views of the *"Vase"* object. Black dotted lines in the polar plot denote the initial angular view point where visual learning behaviour is invoked. Red, Green and Blue lines denote the number of experiment described in the bottom table.



Experiment	Canonical Views			
#1: red	95^o	–	252^o	331^o
				
#2: green	77^o	132^o	257^o	352^o
				
#3: blue	70^o	–	225^o	350^o
				
Confidence Interval (@95%)	80.6 ± 14.6	132 ± 0.0	244.6 ± 19.4	344 ± 13.1

Figure 7.5: Polar plot and canonical views of the *"Cigarette box 2"* object. Black dotted lines in the polar plot denote the initial angular view point where visual learning behaviour is invoked. Red, Green and Blue lines denote the number of experiment described in the bottom table.



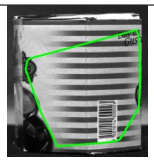
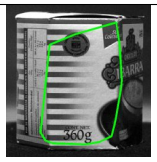
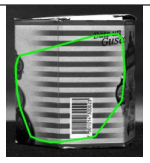
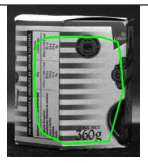
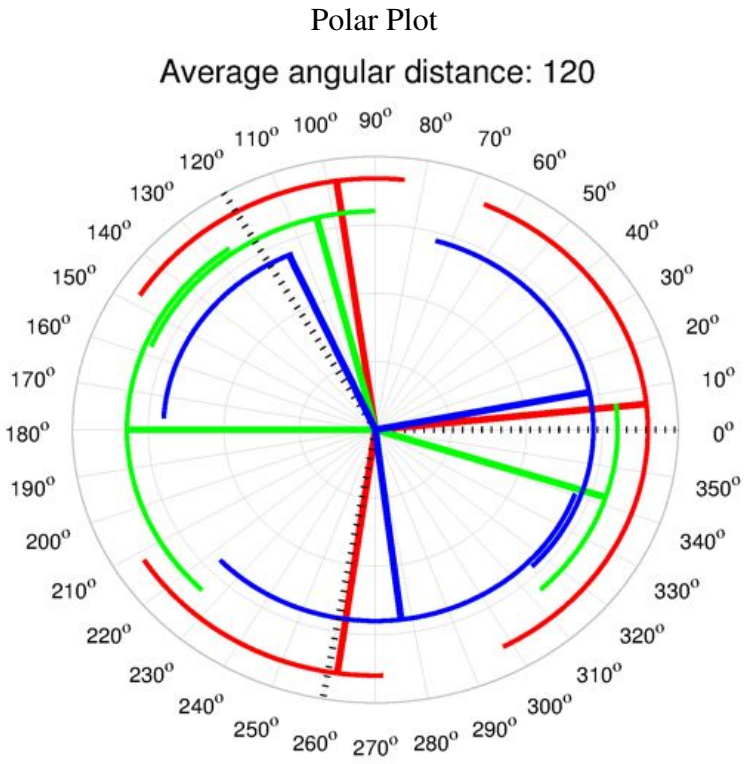
Experiment	**Canonical Views**				
#1: red	-12^o	-21^o	69^o	183^o	–
					
#2: green	–	–	–	141^o	286^o
					
#3: blue	24^o	–	95^o	–	270^o
					
Confidence Interval (@95%)	-3 ± 26.9	–	82 ± 25.4	162 ± 41.5	278 ± 15.6

Figure 7.6: Polar plot and canonical views of the *"Chocolate"* object. Black dotted lines in the polar plot denote the initial angular view point where visual learning behaviour is invoked. Red, Green and Blue lines denote the number of experiment described in the bottom table.



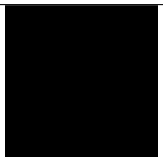
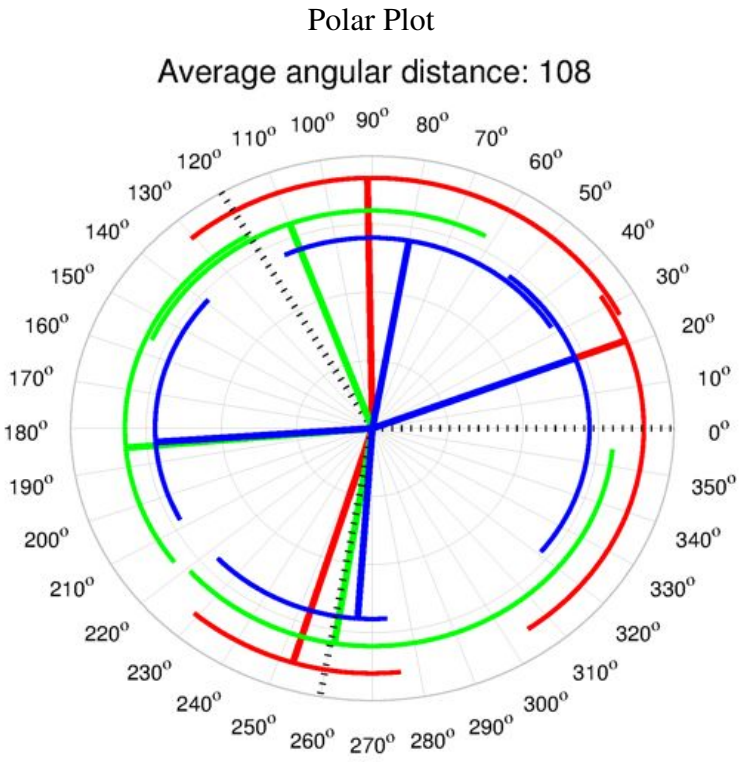
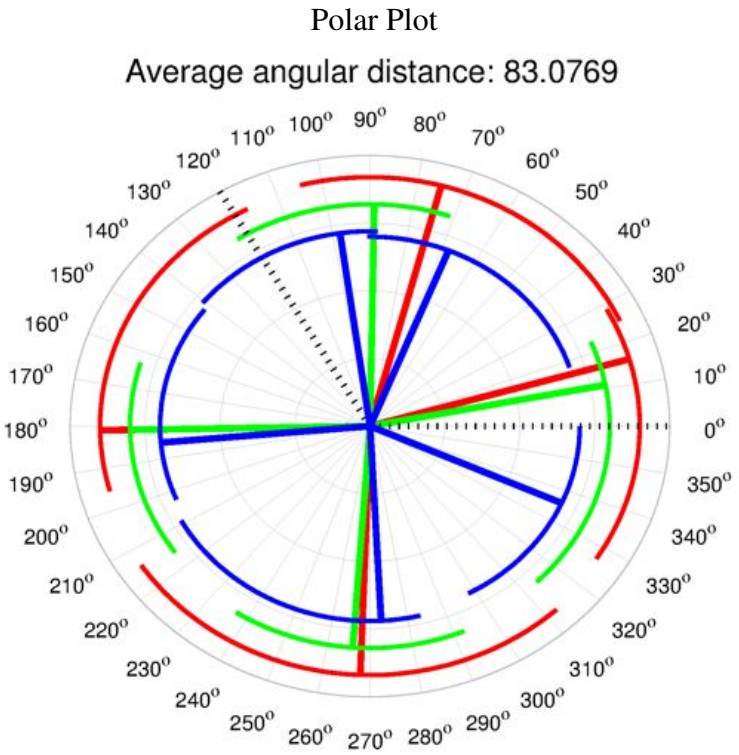
Experiment	**Canonical Views**			
#1: red	6^o	98^o	–	262^o
				
#2: green	-18^o	104^o	180^o	–
				
#3: blue	11^o	114^o	–	277^o
				
Confidence Interval (@95%)	-0.3 ± 17.5	105 ± 9.1	180 ± 0.0	269.5 ± 14.6

Figure 7.7: Polar plot and canonical views of the *"Juice box 1"* object. Black dotted lines in the polar plot denote the initial angular view point where visual learning behaviour is invoked. Red, Green and Blue lines denote the number of experiment described in the bottom table.



Experiment	**Canonical Views**			
#1: red	21^o	91^o	–	253^o
				
#2: green	–	110^o	185^o	261^o
				
#3: blue	21^o	80^o	184^o	266^o
				
Confidence Interval (@95%)	21 ± 0.0	93.6 ± 17.1	184.5 ± 0.9	260 ± 7.4

Figure 7.8: Polar plot and canonical views of the *"Juice box 2"* object. Black dotted lines in the polar plot denote the initial angular view point where visual learning behaviour is invoked. Red, Green and Blue lines denote the number of experiment described in the bottom table.



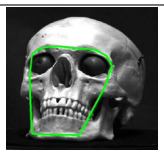
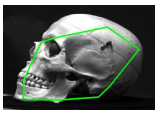
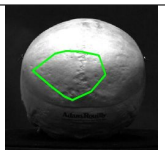
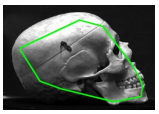
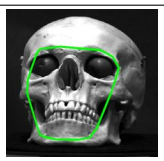
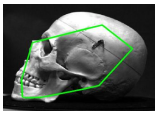
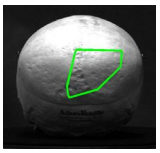
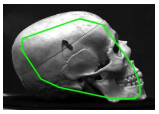
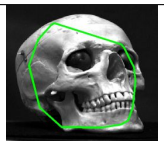
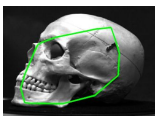
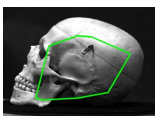
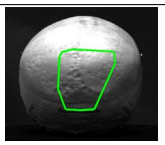
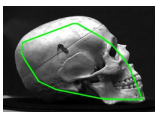
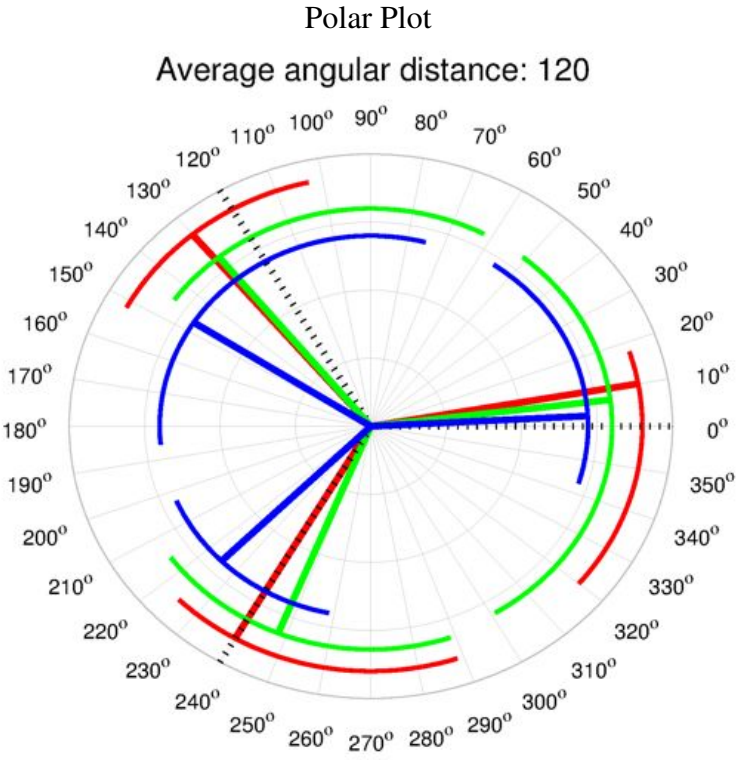
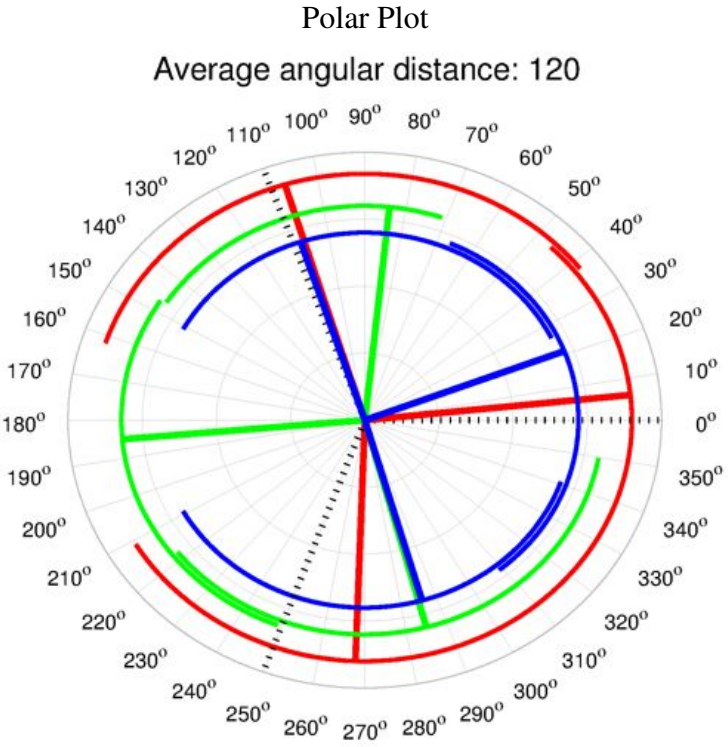
Experiment	**Canonical Views**				
#1: red	16^o	75^o	–	181^o	268^o
					
#2: green	11^o	89^o	–	181^o	266^o
					
#3: blue	-24^o	68^o	98^o	185^o	273^o
					
Confidence Interval (@95%)	1 ± 24.6	82.5 ± 10.4	–	181 ± 2	269 ± 4

Figure 7.9: Polar plot and canonical views of the *"Skull"* object. Black dotted lines in the polar plot denote the initial angular view point where visual learning behaviour is invoked. Red, Green and Blue lines denote the number of experiment described in the bottom table.



Experiment	Canonical Views		
#1: red	10^o	130^o	240^o
			
#2: green	7^o	129^o	248^o
			
#3: blue	3^o	147^o	225^o
			
Confidence Interval (@95%)	6.6 ± 3.9	135.3 ± 11.4	237.6 ± 13.2

Figure 7.10: Polar plot and canonical views of the *"Soup can"* object. Black dotted lines in the polar plot denote the initial angular view point where visual learning behaviour is invoked. Red, Green and Blue lines denote the number of experiment described in the bottom table..



Experiment	**Canonical Views**		
#1: red	6^o	107^o	268^o
			
#2: green	$185^o - 180^o = 5^o$	84^o	285^o
			
#3: blue	21^o	108^o	286^o
			
Confidence Interval (@95%)	10.6 ± 10.1	99.6 ± 15.3	279.6 ± 11.4

Figure 7.11: Polar plot and canonical views of the *"Juice box 3"* object. Black dotted lines in the polar plot denote the initial angular view point where visual learning behaviour is invoked. Red, Green and Blue lines denote the number of experiment described in the bottom table..

7.2.1.1 Results and Discussion

Learned canonical views for all the objects considered in this chapter (Figure 7.1) are illustrated from Figures 7.2 to 7.11. These figures comprise polar plots and canonical view tables for each of the initial view points. Polar plots also depict the explored portion of the object's viewing range before consolidating a canonical view.

Overall, the visual learning behaviour finds between 3 and 4 canonical views per object. It is inferred that the canonical view selection correlates to the intrinsic geometry of the observed object. In other words, the objects employed (Figure 7.1) have clear geometrical symmetries over the viewing sphere , and, in turn, the visual learning behaviour is biased towards the most planar view. For example, the *"Cigarette box 1"* (Figure 7.3), *"Cigarette box 2"* (Figure 7.5), *"Juice box 1"* (Figure 7.7), *"Juice box 1"* (Figure 7.7), and *"Juice box 3"* (Figure 7.11) objects observe the most symmetrical structures of the trained objects, and, therefore, the visual learning behaviour establishes that the most representative and stable views are most planar surface of the object.

From the presented results, the canonical views found can be visually correlated despite the initial angular position of the turn-table and, furthermore, the observed view-sphere range. In that regard, the worst recorded confidence interval (at 95% of significance level) is 41.6 from the statistical mean (the "Chocolate" object, Figure 7.6). Additionally, canonical views of the *"Chocolate"* object are not consistent in any of the three experiments. That is, visible texture information appears to be similar over the object's viewing sphere and, in consequence, the visual learning behaviour incorrectly groups spatio-temporal properties. However, by close inspection of Figure 7.6, it can be deduced that the learning behaviour is still able to capture similar object appearances but at different viewing poses (e.g. canonical views at -21^o and 24^o, or 69^o and 95^o for experiments 1 and 2, respectively).

On the contrary, closer examination of the polar plot of the *"Bear"* object (Figure 7.2), there does not exist a high degree of variation among canonical views. Therefore, this satisfies and demonstrates the familiarity and functionality properties of canonical views. This characteristic is repeatedly observed over the presented results. Therefore, this property corroborates that the visual learning behaviour characterise the object's viewing sphere in accordance with the underlying learning properties described in Section 6.3.1. Furthermore, it also demonstrates that the active exploration of the salient, informative attentional shrouds influence the selection of the the most informative, view-point invariant canonical view of an object.

It is worth noting that the maximum value of the average angular distance recorded is 120 degrees between canonical views. This suggests that the visual learning behaviour while using this consolidated object representation, can potentially achieve a recognition rate above

the 85% of recognition success for anyintermediate view (this limit is set in accordance to the overall performance of the learning approach reported in (Kootstra, 2010)). This assertion will be further demonstrated in the forthcoming sections.

7.2.2 Pre-attentive Localisation

This section aims to further test the *goodness of recognition, familiarity* and *functionality* properties (as defined in Section 6.3.1) in isolation for each object. That is, visual knowledge is robustly consolidated from the active interaction with objects (as humans do, Figure 6.2) and semantic small attentional shrouds, as opposed to considering a manually segmented database (i.e. bounding boxes as defined in Chapters 3 and 4). Therefore, these experiments test the following assumption:

- *Learning discriminant canonical views from an object presents a trade-off in terms of localisation accuracy and recognition with respect to object database size as opposed to using a manually segmented object representation over the view-sphere of the same object.*

It is worth noting that few approaches related to the devised visual learning behaviour have been reported in the literature. The most related work with respect to the reported visual learning behaviour in this thesis is the *active learning exploration* approach presented by (Kootstra, 2010) (Section 2.7.2). In that regard, the experiments in this section are to some extent similar to those reported in (Kootstra, 2010). The author's experimental set-up consists in allowing a robot to actively move around the object's viewing sphere every 10 degrees and measure the recognition rate achieved with a densely sample (intervals of 30 degrees) and learned object databases. It must be noted that the mobile robot does not perform camera saccadic movements while exploring/learning an object. The experimental methodology used in this section is thus described as follows.

The best canonical views of from experiments described in Section 7.2.1.1 are selected by the author of this thesis and these views are therefore stored in an object database as defined in equation 5.3 (Section 5.4). Likewise, the same objects employed while validating the visual learning behaviour (Figure 7.1) are sampled at fix intervals of 30, 45 and 60 degrees between snapshots. SIFT features (Section 7.2.2) are extracted for each captured image and, therefore, three manually segmented object databases are created which corresponds to each of the above sampling intervals.

Figure 7.12: Experimental setup for the pre-attentive localisation experiments.

The robot vision system is thus allowed to pre-attentively (i.e. passively) localise and detect a single instance of an object class (the designed pre-attentive behaviour in Section 5.5) against a slightly cluttered background, as depicted in Figure 7.12. That is, the object is embedded in a scene comprising this object on a turn-table beside distracting unknown objects (Figure 7.12). The visual task consists in passively inspecting the presented scene while the object is rotated across the viewing sphere at intervals of 5 degrees (in a similar fashion as the mobile robot described in (Kootstra et al., 2008)). The pre-attentive behaviour is invoked 50 times for each object using one of the above 4 object databases.

From the acquired experimental data (i.e. 3600 observations for all objects and databases), the back projection error (Mean Squared Error or MSE as defined in Section 4.6) while pre-attentively localising and detecting objects is stored in order to validate and compare quantitatively the *goodness of recognition, functionality* and *familiarity* of the learned canonical views (ref. Section 6.3.1) with respect to the manually segmented databases. False positives are thus determined by measuring the degree overlap between ground truth and the localised object hypothesis. In that regard, the 3600 observations are manually annotated by means of bounding boxes. The confidence ellipse test (Section 5.6) is employed to measure the accuracy of the object's localisation. Therefore, if the localised object's bounding box is within $\sim 68\%$ of significance level (i.e. 1 standard deviation from the centre point of the object's bounding box) with respect to the annotated ground truth, such localisation is marked as true positive; otherwise it is considered as a false positive. A table will summarise the recognition rate (i.e. *goodness of recognition*) of using the learned object canonical representations and the sampled databases.

Functionality and *familiarity* are finally demonstrated in terms of the degree of dispersion of the recorded MSE for each experiment. To that end, box plots will illustrate the spread of the MSE for each database employed which illustrate the degree of dispersion with respect the annotated centroid location.

	DB 30	DB 45	DB 60	Learn
Bear	100%	100%	100%	100%
Cigarette Box 1	100%	100%	*99.3%*	*86.1%*
Vase	100%	100%	100%	*94.4%*
Cigarette Box 2	*90.2%*	*68%*	*67.3%*	*96.5%*
Chocolate	100%	100%	*95.1%*	100%
Juice box 1	100%	100%	*97.9%*	100%
Juice box 2	100%	100%	100%	100%
Skull	100%	100%	100%	*94.4%*
Soup can	100%	100%	100%	*97.2%*
Juice box 3	*97.2%*	*98.6%*	*97.2%*	100%
AVERAGE	**98.7%**	**96.1%**	**95.7%**	**96.8%**

Table 7.2: Recognition rates while pre-attentively locating an object while using each of the four databases considered in a slightly cluttered scene (Figure 7.12).

7.2.2.1 Results and Discussion

Table 7.2 summarises the recognition rate percentages achieved for each databased and trained objects. Accordingly, *"DB 30"*, *"DB 45"* and *"DB 60"*, in the table, denote the object databases sampled at fixed intervals of 30, 45, and 60 degrees, respectively. Whilst *"Learn"* indicates the database acquired by means of the visual learning behaviour. Likewise, recognition rates below 100% are highlighted.

Hence, Table 7.2 exhibits that the success recognition rate for the manually segmented object databases is in accordance with the sampling intervals. That is, the overall recognition rate of the system declines as the sampled interval increases. The worst recorded rate is 67.36% for the *"Cigarette Box 2"* while employing *DB 60*. The learned object database, on the contrary, observes a recognition rate within viable limits of acceptance (as asserted in Section 7.2.1.1). Therefore, this observation proposes that the learned representations introduce a trade-off as the object deviates form the registered canonical view; however, in this case, the worst recognition rate never falls below 85% in all recorded observations.

The localisation errors incurred by the system are depicted in Figures 7.13 and 7.14. By carefully inspecting these results, it is observed that there indeed exists a trade-off while employing the learned canonical views. However, the worst observed error (5.4 of back projection error) only represents $\sim 2.86\%$ of the object's bounding box region in the observed image (this back projection error corresponds to the "Cigarette Box 1" object which occupies 191×244 pixels in the image, Figure 7.13).

Although the pre-attentive behaviour successfully detects all the object posses, it fails to correctly localise the object using the learned database. Therefore, the observed recognition error

rate only depicts detections that do not satisfy the statistical confidence ellipse test. In that regard, close inspection of Figures 7.13 and 7.14 and Table 7.2 demonstrates that the degree of dispersion of the back projection errors correlate to the loss of recognition performance, e.g. the *"Cigarette Box 1"* object.

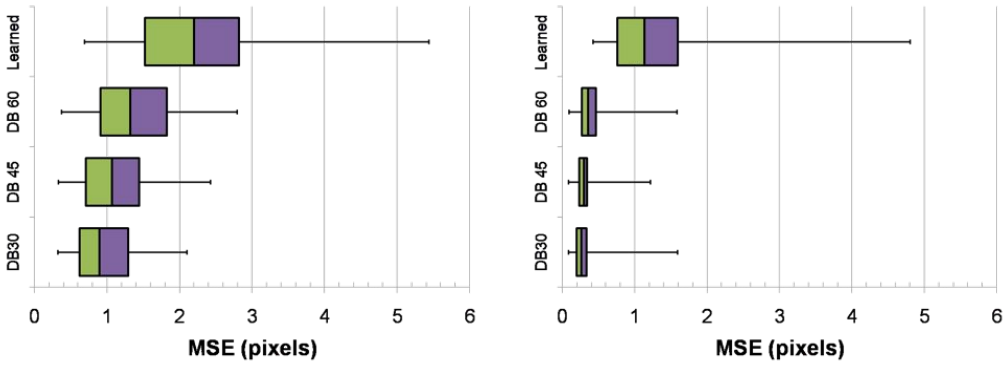
As discussed in the previous section, these experiments only show the recognition rates and back projection of errors while passively detecting and localising an object across the viewing sphere and, thereby, employing the corresponding object database of the observed instance (i.e. the case of multiple same-class object instances is not considered). The presented results therefore outperform the *active learning exploration* approach (Kootstra, 2010) as the learned object database with the visual learning behaviour represents an average trade-off of only 96.8% of recognition rate with respect to 98.7% of an object database sampled at intervals of 30 degrees (Table 7.2).

The following section concerns with the study of the learned object database while exploring complex and cluttered scenes and, in consequence, it further extends the reported experiments described in this section.

7.3 Active Visual Exploration Experiments

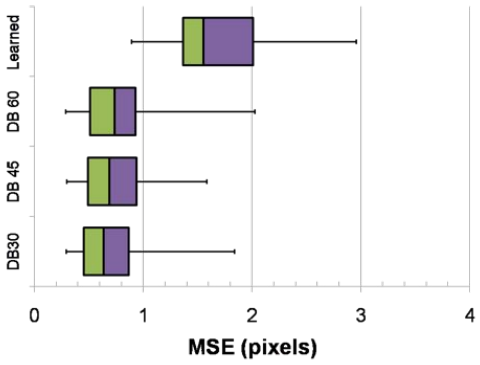
In previous chapters, the active visual exploration of a scene has been validated with manually segmented object databases. The hierarchical architecture has been demonstrated for slightly cluttered and complex visual scenes. However, it is essential to fully corroborate the robustness of the system while carrying out a visual task over challenging, complex, and cluttered scenes that contain several object instances observing different poses and with similar visual features. In addition, these experiments also describe the performance of the learned appearance representation of the 10 objects with respect to manually segmented databases comprising these same objects.

Therefore, these experiments aim *to validate the learned object representations in previous Section 7.2 in visual discrimination, recognition, identification, and categorisation visual tasks*. Likewise, *it also compares the performance of the learned representations with respect to manually segmented object databases*. These experiments thus confirm that *by using stereo vision into a robotic head and adopting the active vision paradigm structured as a hierarchy of visual behaviours (which integrates low- and high-level visual competences), it is possible to enable a robot vision system to find requested objects in autonomous mode (i.e. 'lost and found' problem) by either using a view-point invariant learned object database or a manually segmented object database.*

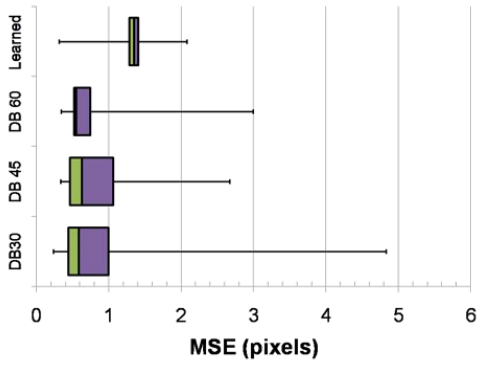


Bear

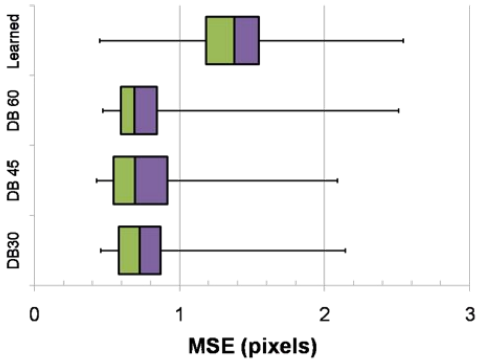
Cigarette Box 1



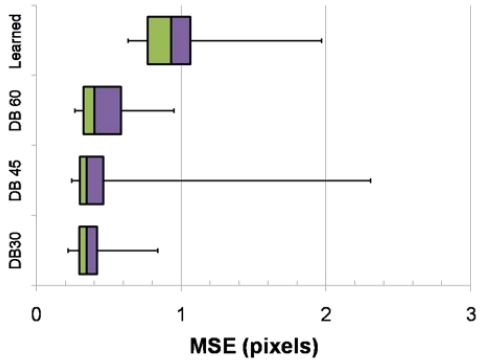
Vase



Cigarette Box 2

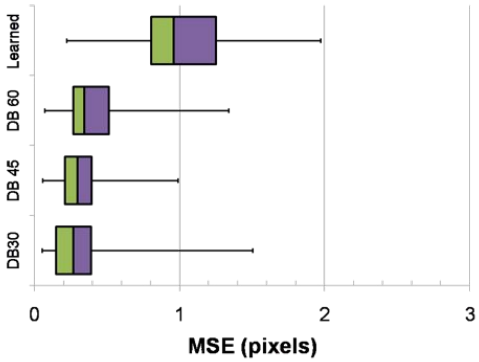


Chocolate

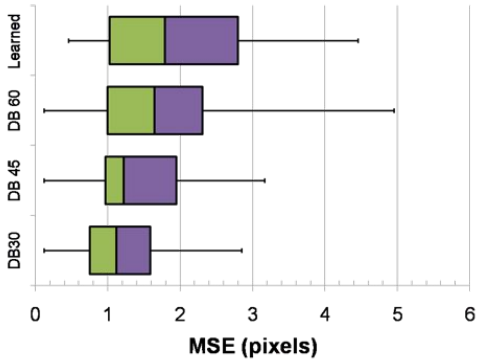


Juice box 1

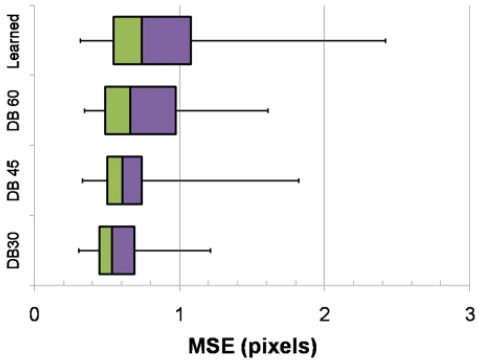
Figure 7.13: Box plots illustrating the degree of dispersion while pre-attentively localising an object.



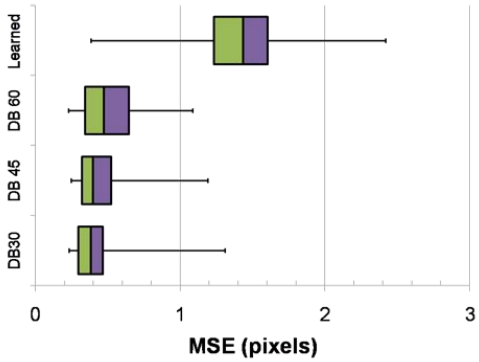
Juice box 2



Skull



Soup can



Juice box 3

Figure 7.14: Continuation of Figure 7.13.

The validation rationale of these experiments follows the methodology framework previously reported in Sections 3.8 and 5.9. That is, scenes contain cluttered backgrounds comprising known and unknown objects (as depicted in Figure 7.15). In this chapter, however, a total of 7 different scenes are arranged in increasing complexity such that 20 known instances of the ten object classes are disposed in arbitrary poses and locations. Scene complexity is defined in the context of this chapter as the quantity of similar unknown objects in the scene (i.e. a typical source of potential outliers) and background clutter. It must be noted that the last and most difficult scene comprises 21 object instances due to an author's error while counting the objects in the scene.

Trials of this validation consist of three visual search tasks: for each of the manually segmented object databases sampled at intervals of a) 45 and b) 60 degrees, respectively; and c) for the learned object database. In that regard, the object databases created in Section 7.2.2 are employed in these experiments.

Each of the above visual exploration tasks also includes 3 random initial fixations points which, in turn, produce a total of 63 visual search tasks. Opposed to previous experiments reported in Sections 3.8 and 5.9, where the visual search task is allowed to perform a maximum of 20 and 45 saccades, respectively, these experiments allow the system to halt the visual search task if and only if the robot does not attend an object within 5 consecutive saccades. That is, if there are no new object hypotheses to be attended (i.e. the system is only targeting salient features) because all possible objects within the scene have been successfully attended and identified and, in consequence, they are inhibited after each pre-attentive cycle, as described in Chapter 5, the visual search task is interrupted and reports the objects found in the scene. The above is considered in order to shorten the visual search time while conducting these experiments.

Hence, 1269 object observations are recorded in all visual search tasks (i.e. 6 scenes comprising 20 objects each explored by the 3 object databases times over 3 initial fixations plus 1 scene comprising 21 objects with the same object databases and initial fixations). While actively exploring a scene, there are three possible outcomes in the detection and recognition of objects (as it has been employed in previous validations of this thesis, e.g. Chapters 3):

- *False positives* include when the system localises an object hypothesis but without being able to centre the object in the field of view of both cameras during the attentive cycle or, similarly, an attended object hypothesis does not correspond to the object class in the scene.
- *Not found* comprises the system's failures of not pre-attentively noticing an object in the visual search task.

- *False hypotheses* are those objects that do not correspond to the object class identity in the pre-attentive cycle while attentively verifying its identity.

Similarly, the egocentric spatio-temporal attended location (ref. Sections 5.2 and 5.5.2) is recorded for each fixated object in order to quantitatively measure the system's performance. The possible results consist in the ability of the system to centre the hypothesised object in the field of view and to fixate at the same camera-space location in each of the different initial random fixations per scene. A plot of the average Root-Mean-Square (RMS) fixation errors (as defined in Section 5.9) of the egocentric x and y fixation coordinates locations is employed to depict the system's accuracy, robustness and repeatability. Finally, correct and incorrect recognition of objects (Section 7.3.1) for each of the three object databases are measured in terms of ROC- and PR- curves (ref. Section 4.8) in order to characterise and compare the overall performance of the hierarchical architecture between manually segmented object databases.

Conversely, while conducting pilot experiments with the hierarchical architecture, the fixation accuracy towards objects on the periphery of the scene was observed to decrease considerably. This is because the hierarchy of behaviours designed to verge the cameras have assumed in previous experiments that the cameras are in a close fronto-parallel configuration. This constrain holds true while objects are near to the homing position of both cameras. However, objects located towards the periphery of the scene are not correctly verged since the residual vertical disparity increases considerably. Hence, the vergence behaviour is extended to calculate vertical disparities and, in consequence, enable the cameras to adjust the vertical error in the vergence close-loop cycle. The extended algorithm consists of only considering the residual vertical disparity in the overall algorithm in Figure 3.4.

In that regard, the performance of the extended vergence behaviour is analysed in Section 7.3.2. The residual x and y disparities for the 0th layer and the 3rd layer of vergence (Section 3.4, and Section 5.7) are collected in all executed visual searches in this experiments to statistically describe the robustness and performance of the vergence system.

7.3.1 Visual Exploration Results

Figure 7.15 depicts the scenes employed in these experiments; likewise, the fixation points and camera saccades/traces for the left camera in image space are illustrated (a square denotes a fixation point, circular dot represents a salient item, and the downward and upward pointing triangles indicate the initial and final position of the search process). Inspection of 7.15 confirms the ability of the system to perform autonomous scene exploration. The average

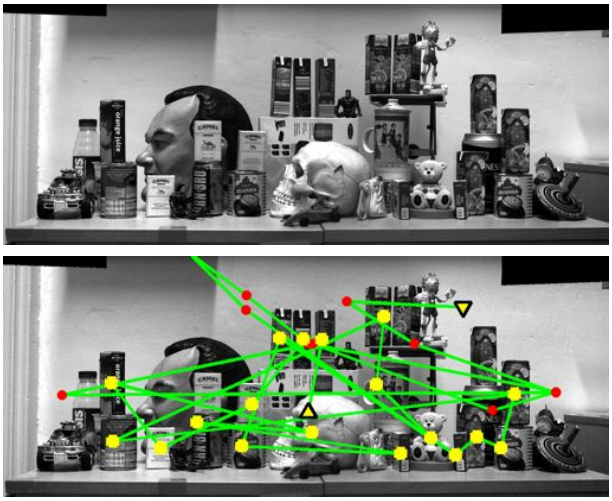
time of the visual searches considered is ~ 90 minutes in order to potentially find all possible objects within the observed scene. It must be noted that the image resolution of these experiments is set to 1024×768 pixels since the average execution while employing the maximum image resolution of the cameras (5 mega pixels) corresponds to almost three times the above described reported time.

The average RMS errors of all attended objects in the observed scenes are summarised in Figure 7.18. Specifically, the worst measured error (~ 27.7 pixels on the x-axis) is observed in the scene 1 (Figure 7.15) and, thereby, the average RMS error is ~ 12.8. This error corresponds to the *"Cigarette box 1"* object stored in the database sampled at 60 degrees (i.e. DB 60). In this case, it is inferred that the found fixation point is influenced by the distracting unknown objects in the scene. It must be noted that, this RMS error is registered on the most complex scene ("scene 1", Figure 7.15); however, this error only represents $\sim 6.7\%$ with respect to the total size of the "Cigarette Box 1" in image (191×244 pixels in the image). Notably, fixation errors in the x and $y-$axes while using a learned database are smaller than those observed for the manually segmented databases. This suggests that the learned canonical views are highly tuned in accordance with the spatio-temporal feature properties of the object's appearance. Whilst, manually segmented databases are thereby prone to include unstable visual features that, in the context of a active visual exploration, become potential outliers in the presence of roughly similar and same-class objects.

Figure 7.19(a) illustrates the system's recognition rate for all visual search tasks. As described in Section 7.3, Scene 1 (Figure 7.15) emerges as the most difficult scene due to image clutter and similar unknown similar textures; however, the lowest recorded recognition rate is $\sim 85\%$ (i.e. 19 objects are recognised of a total of 21) while using a database sample every 60 degrees. In that regard, the learned object database exhibits in Figure 7.19(a) the best recognition rate for scene 1; that is, the success rate with this database is above 95% in all invoked visual search tasks. Therefore, the average successes recognition rate is $\sim 98.8\%$ which, in turn, outperforms its manually segmented object databases counterparts and, moreover, current state-of-the-art visual learning approaches (Kootstra, 2010). The latter is further demonstrated (Figure 7.19(b)) where the learned database presents few visual failures. Specifically, the robot vision system did not find only 5 objects of a total of 423 object observations while employing the learned object database.

Finally, the overall performance of the system is further confirmed by means of ROC and PR curves (as depicted in Figures 7.20(a) and (b) respectively). Accordingly, the active binocular robot vision system equipped with a learned database exhibits a recall rate of ~ 0.84 of correct object identifications (Figure 7.20(a)). Similarly, this particular configuration maintains a satisfactory precision rate above a specific level of acceptance of 85% of recognition success

Scene 1



Scene 2

Scene 3

Figure 7.15: Scenes created for these experiments whereas just right at the bottom of each scene it is depicted camera traces of a selected visual search task for only the left camera; these traces are *approximately overlaid*. Scene 1 is regarded as the most complex whereas Scene 7 the least.

Scene 4

Scene 5

Scene 6:

Figure 7.16: Continuation of Figure 7.15.

Scene 7

Figure 7.17: Continuation of Figure 7.15.

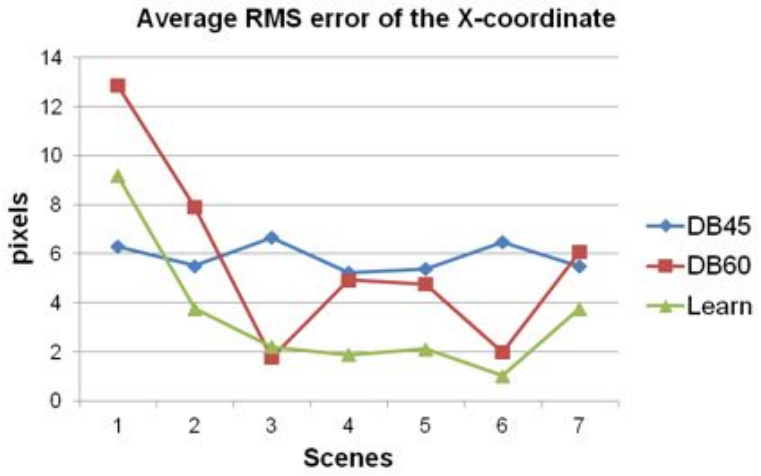
for any a intermediate view (this limit is set in accordance to the overall performance of the learning approach reported in (Kootstra, 2010)) while the system actively explores a scene (as depicted in Figure 7.20(b)).

7.3.2 Vergence Behaviour

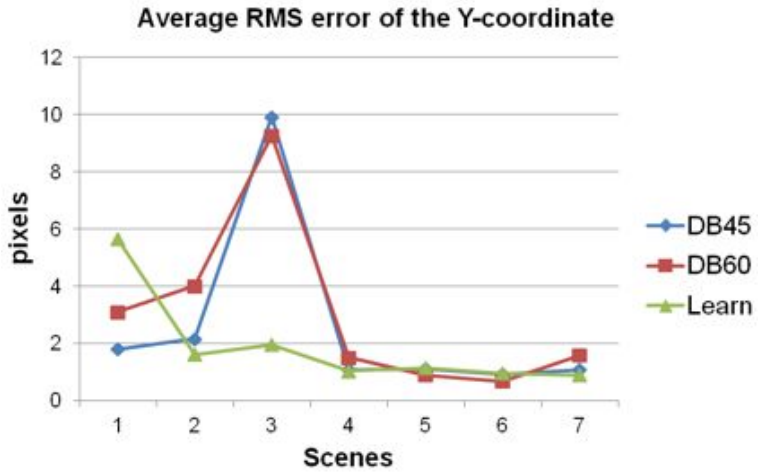
The validation of the extended version of the vergence behaviour featuring convergence of the $y - axis$ is presented in this section. The vergence behaviour has already been validated in previous chapters; therefore, this section only comprises the residual disparity incurred by the system while executing the above described visual search tasks.

Hence, Figures 7.21(a) and (b) illustrate the performance of the vergence accuracy while converging the cameras in the 0th and 3rd layers of vergence while carrying out the above reported visual exploration tasks. The 3rd layer of vergence, as described in Section 5.7 consists of two modes of operation: while targeting either an object or a salient feature.

As there is not a defined vergence point, it is only record the residual disparity (i.e. the final highest peak in the histogram of disparities close to zero) in both axes and the number of vergence iterations required to stabilise the algorithm. On the one hand, the non-selective vergence behaviour is only activated at the beginning of the visual search task; therefore, the number of sample disparity points corresponds to the total number of visual search experiments, i.e. 63. On the other hand, the selective vergence is activated when the system is either attending an object (target) or a salient feature (no target); therefore, the sample disparity data points are 852 (i.e. total number of identified objects, including false positives) and 815 data

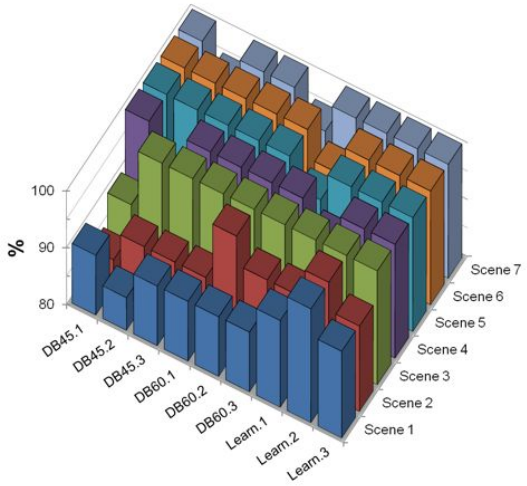


(a)

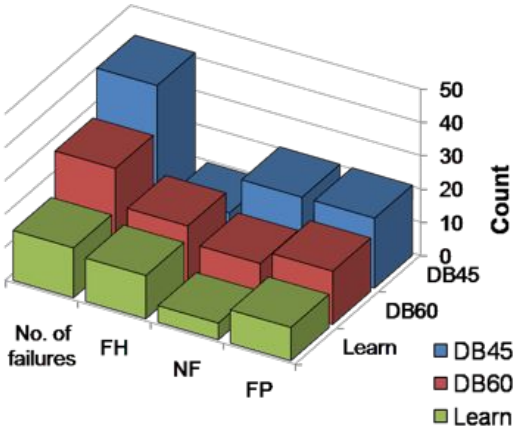


(b)

Figure 7.18: Overall RMS error observed for each scene and each object database in the (a) x- and (b) y-axes. Scenes' numbers correspond tho those depicted in Figure 7.15. Object database abbreviations follow the naming convention described in Section 7.2.2.1. The RMS error is defined in Section 5.9.

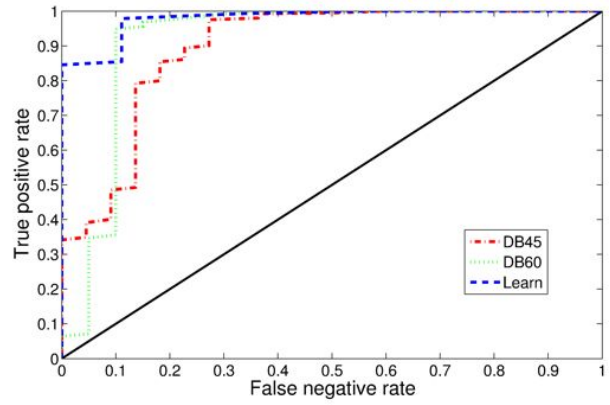


(a)

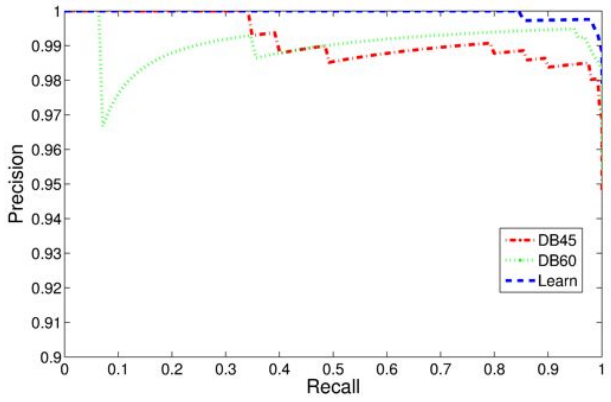


(b)

Figure 7.19: (a) Recognition rate for all the 63 visual tasks. Numbers after the object database name corresponds to the random fixation experiment. (b) Visual failures of each object database where FP, NF and FH stand for *False Positive*, *Not Found* and *False Hypotheses* as described in Section 7.3.

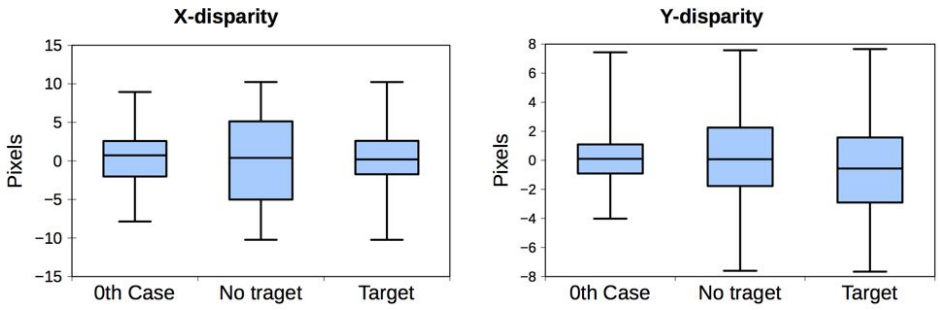


(a)



(b)

Figure 7.20: (a) ROC and (b) PR-curves for the manually segmented object databases sampled at 45 and 60; and the learned object database.



(a) (b)

Figure 7.21: Measured residual disparities in the (a) x- and (b) y-axis while verging the cameras in the 0th and 3rd layers of vergence.

points while attending salient features. As depicted in Figure 7.21, it is possible to observe that the residual disparity are within ± 10 pixels from the statistical zero mean. Therefore, the degree of dispersion achieved is still within practical limits for stereo reconstruction and depth recovery as pointed out in Section 3.6.1. Finally, the vergence algorithm takes an average of 1 iteration in order to reduce the level of residual disparity below the terminating threshold.

7.4 Conclusions

The validation process and results of the visual learning behaviour and the integrated hierarchical robot vision architecture are presented in this chapter. The aim of the presented validation is to investigate and study the performance of the devised algorithms over challenging settings and scenarios.

The ultimate objective of this thesis (as outlined in Chapter 1) is to investigate and design a binocular active vision platform which maintains vergence, directs its gaze, recognises multiple same objects, autonomously explores the environment, semi-automatically learns object representations over the viewing-sphere and, finally, increases the visual understanding, richness and representation of what it is observing in a unified hierarchy of visual behaviours architecture. The purpose is to integrate visual competences into a hierarchy of visual behaviours for the robustness and applicability in a wider range of robotic applications.

As reviewed in Chapter 2, ill-posed problems in passive vision systems become objectively solvable by adopting the active vision paradigm. That is, the environment can be observed from different viewpoints when the visual tasks become ill-posed, and, consequently, more possible solutions can be obtained to solve them. In the context of this thesis, the active and dynamic interaction with the environment allows the described robot vision system to create robust, view-point invariant object representations and, furthermore, to provide the means of

autonomous exploration of complex and cluttered office-like environments in order to solve the *"lost and found"* problem as outlined in Section 1.2. Segmentations of the found canonical views, as observed from Figures 7.2 to 7.11, do not describe the object's shape. This is due to the fact that SIFT features (the adopted visual representation in this thesis) close to the object's contour edges are not as stable as those found on surfaces within the overall shape. This suggests that the SIFT framework must be extended in order to include boundary description. This is further discussed in the future work section of this chapter (Section 8.2).

The semi-automatic learning behaviour is demonstrated in terms of the overall performance of the investigated robot vision system validated under different settings and experimental frameworks. Therefore, the robot vision architecture exhibits an average success recognition rate of $\sim 98.8\%$, and a recall and precision rate of ~ 0.84 and ~ 0.99, respectively, while actively exploring, maintaining camera vergence, and recognising multiple same-class object instance with high confidence. It is therefore concluded that the devised hierarchical robot vision architecture favourably compares and outperforms state-of-the-art robot vision systems, i.e. (Rasolzadeh et al., 2010; Meger et al., 2010; Kootstra, 2010) (0.92<0.1 FPR, 90% categories and 72.2%, respectively).

Similarly, it is also confirmed that the visual learning behaviour creates robust, view-point invariant canonical views that can be used in visual search tasks. It is worth noting that approaches related to the devised visual learning behaviour scarcely appear in the literature. The most related research in regard to the reported visual learning behaviour in this thesis is the *active learning exploration* approach presented by (Kootstra, 2010). The author demonstrates, as previously discussed in Section 2.7.2, that by allowing a mobile robot to actively explore an object, it is possible to build a condensed SIFT feature database with a compression ratio of $\sim 37\%$ with respect to a manually segmented object database sampled at $30\,degrees$ of interval. As previously discussed, the active learning exploration outperforms its passive vision counterpart; however, such compressed feature database exhibits a trade-off of $\sim 85\%$ of recognition rate against $\sim 99\%$ of using SIFT features passively observed at fixed intervals. Hence, it is concluded that the results presented in this chapter outperform those reported in the current literature. Furthermore, the overall compression ratios of the learned database are 14.2%, 21.2% and 28.5% with reference to the manually segmented object databases sample at fix intervals of 30, 45 and 60 degrees, respectively. Figure 7.22 illustrates the number of keypoints for each each object class (Figure 7.2) stored in the above described object databases.

The exploration of an object while learning its appearance (i.e. camera-eye saccadic movements) improves the robustness of characterising canonical object parts and, in consequence, creating a condensed but robust description of an observed object. That is, these canonical

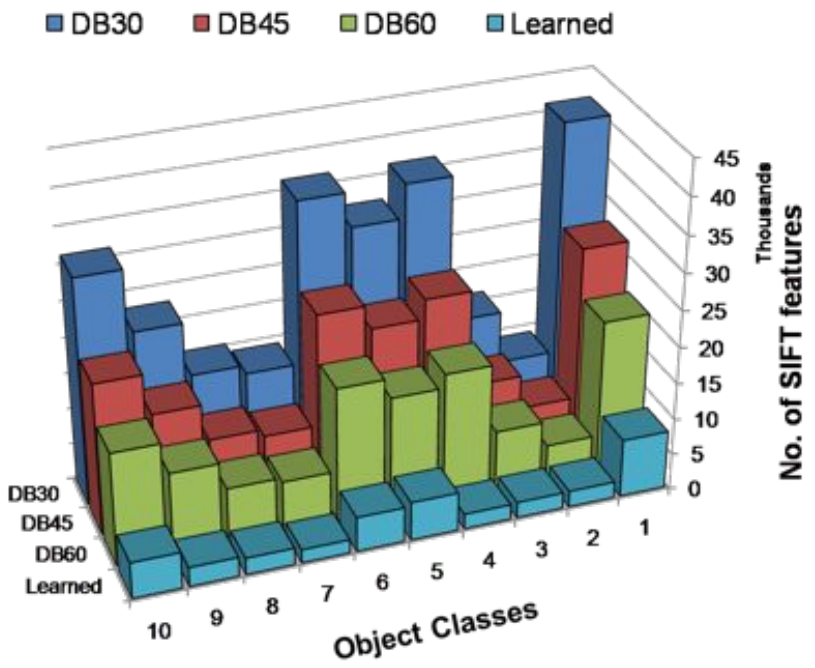


Figure 7.22: Relation of number of keypoints found for each object class over the object databases employed in this chapter.

representations become stable when the robot actively interacts with the object in order to gain more visual information, while saccading to structural representative object parts or attentional shrouds (i.e. active exploration of the object) and storing those features that are spatio-temporal consistent over a set of angular poses, dynamic changes and saccades. Therefore, object canonical representations are characterised by three different underlying properties: *goodness of recognition*, *familiarity,* and *functionality* (as defined in Section 6.3.1) which, as demonstrated in this chapter, enables the devised hierarchical architecture to detect, locate, and verify objects despite their pose and viewpoint in cluttered scenarios.

It must be pointed out, however, that the hierarchical robot vision architecture is only validated according to the feature description adopted (i.e. SIFT) and, in consequence, objects considered in this study are selected with visibly enough texture detail; otherwise the robot vision system would fail in their recognition. However, the devised visual behaviours are designed such that their operation is not constrained to the adopted visual understanding model. Thus, the forthcoming and last chapter addresses the above limitation, draws conclusions from the outcomes of the PhD research project, and provides future work directions in which this active robot vision system can be further improved, extended, and applied.

Chapter 8

Conclusions

This chapter presents a summary of the research conducted in this thesis. The significance and novelty of the reported research is compared with the current literature. The achievements and limitations of this thesis are described. The chapter concludes with future directions where this thesis can be extended.

8.1 Contributions

This thesis advances current literature in terms of the following key major contributions:

- Development and demonstration of a novel covert multiple same-class object instance detector using a continuous Hough space representation and unsupervised clustering techniques.
- Conceptualisation and development of an integrated and novel hierarchical architecture for the active binocular robot vision system being capable of performing the described visual tasks (Section 1.1.1) in real-world environments.
- First reported development and demonstration of the full active vision paradigm applied to learn the appearance of an object.
- First functional demonstration of the integrated hierarchical robot vision architecture operating in cluttered and complex real-world environments, featuring multiple same-class object instances detection with occlusion and self-occlusion while using a self-learned object knowledge database.

Minor contribution comprises:

- Analysis and functional demonstration of the initial binocular robot system reported in (Fattah, 2007) and (Fattah et al., 2008).

The validity of the objectives and scientific questions described in Sections 1.2 and 1.2.1, respectively, have been demonstrated through out this thesis. These are summarised in the following sections.

8.1.1 Extended Validation of the Active Binocular System

The analysis and study of the initial and founding active binocular robot head system is given in Chapter 3. Several novel principles were integrated in this system for the purpose of carrying out *detection and recognition* visual tasks of single object classes. That is, this system adopted SIFT features as the underlying visual representation to execute autonomous scene exploration of a real-world scene and such features subserved as attentional cues for the "attentional spotlight" and "stepping-stone" visual search strategies. As reviewed, this system featured visual competencies for binocular vergence, object recognition and gaze control.

As presented in 3, a pilot investigation conducted by Fattah (2007) demonstrated the functional operation of the system in a moderate complex and cluttered real-world scene. In this first validation, the system was capable of directing and centring its gaze towards objects with a successful recognition rate within viable limits of acceptance (i.e. $\sim 80\%$ of successful recognition rate). However, this pilot characterisation of the system was deemed incomplete because a single scene and and visual search did not entirely characterise the system's performance.

It was thus proposed to fully verify the correct function, recognition performance and robustness of the system by testing it under different operational and challenging settings in order to fully characterise and identify shortfalls in its design and, in consequence advance the literature. In that respect, five different scenes were created featuring a combination of ten known and unknown objects in random poses. The experimental methodology consisted of invoking the autonomous scene exploration over 25 trials. That is, it was proposed to randomly select five initial fixation points for each scene in order to measure the repeatability of the system while targeting the object presented in the scene. Results demonstrated a recognition rate of 85% while observing fixation deviation of 8 pixels from optimal. Similarly, the system took 37 saccades to successfully find all known objects presented in 25 different scenes.

From these extended validation, it was noticed that the ability of the system conveyed successfully the specified visual tasks; however, the system was only capable of attending one single instance of an object class for each visual search invoked. This shortfall on the design of current robot vision systems is mainly produced by the object recognition engine. In the active binocular robot, specifically, the SIFT object recognition pipeline is adopted which, in turn, features the generalised Hough transform. In the feature extraction context, the underlying algorithmic design of the Hough transform limits the detection of multiple same-class object instances since it is coarsely quantised to produce a single peak for all geometrically SIFT features that match between observed and trained images.

Similarly, the incurred recognition error was due to the deficiency of pre-attentively localising an object because there was not a trained example in the object database that resembled the observed object's pose (an inherent limitation of the SIFT feature which presents an out-of-plane invariance of up to 30 degrees with respect to the trained example (Lowe, 2004)). It was therefore observed that objects with poses not trained in database were detected with a lower detection score than those that resembled the trained examples.

Therefore, the scientific insights gained from the analysis of the above results motivated the design of the covert multiple same-class object detector to allow multiple detections in a single image and the semi-autonomous object's appearance learning behaviour to allow training visual data to be gathered over each object's viewsphere in order to eliminate potential bias in the exploration task (as reported in Chapters 4 and 6, respectively). Specifically, this research work fulfils and answers the research question: "*How does a robot explore autonomously a cluttered, complex scene?*"

8.1.2 Multiple Same Object Class Instance Detection and Localisation

In order to endow the binocular robot head with the ability to detect and localise multiple same object class instances, it was proposed in Chapter 4 to divide a continuous (non discretised) Hough transform space into groups and, in consequence, apply an unsupervised clustering technique to detect multiple peaks in this space. The devised algorithm exploited the intrinsic properties of the continuous Hough space to localise SIFT feature groups represented by multiple distinct peaks in this space. This algorithm therefore employed well-known pattern recognition algorithms to group multiple peaks in the continuous Hough space in order to detect and localise multiple same-class object instances.

A complete investigation of this algorithm was also provided in Chapter 4. It was demonstrated that this method, operating with a fuzzy C-means approach, was capable of detecting

up to 5 objects with a recall rate of $\sim 98\%$ as a pre-attentive module of the active binocular robot vision system. This algorithm thus overcame current limitations in the literature on the detection of multiple objects, e.g. the need to tune clustering distance thresholds according to the object-class being detected and the use of large image databases (as described in Chapter 4). Thus, the formation of multiple peaks was performed in an unsupervised manner which, in turn, permitted the detection of self-occlusion between instances. In that regard, the perceptual capabilities of the devised algorithm achieved an overall overlap percentage of $\sim 66\%$.

The detection and localisation of up to 5 objects was found in accordance with the capabilities of the human visual system and, furthermore, with the active vision paradigm (Sections 2.1.1 and 2.5). This was due to the fact that active vision, according to Aloimonos et al. (1988) and Ballard (1991), divide attention in order to focus on portions of the environment and to manage visual resources efficiently to achieve a high-level task-goal specification. Active vision granted the ability to dynamically sense the environment and therefore the detection of more than 5 objects in the scene was not required. That is, the "stepping-stone" visual search strategy could be devised such that if objects were not noticed in a pre-attentive cycle, the active exploration of the environment enabled the robot to observe a better viewpoint and, thereby, detect those unnoticed objects of previous pre-attentive cycles (as demonstrated in Chapter 7). This observation also enabled the investigated robot vision system to achieve multiple instance perception tasks. Therefore, active visual search would enable a robotic vision system to locate and detect object instances reliably, despite being there the strong possibility that not every instance in the field of view was detected in each fixation performed.

The multiple same-class object instance detector is believed to be the first successful attempt in the literature to tackle multiple detection and localisation of same-class object instances as a covert, endogenous visual competence. Its functional validation in a robot vision context was further demonstrated to successfully carry out *detection, identification, and same-different identification* visual tasks (as presented in Chapters 4, 5 and 7). Therefore, this detector addresses the second research question specified in this thesis, namely, "*How can a robot be capable of detecting and localising multiple same-class object instances of the same object class with occlusion and self-occlusion settings (i.e. Same-Different visual task)?*"

8.1.3 Visual Behaviours in the Hierarchical Robot Architecture

Given the availability of multiple same-class object instances detection in the active binocular robot vision system, the integration of this detector within the robot vision software was complex as this system had been developed as a set of "ad-hoc" functions to carry out the

specific visual scene exploration task. This approach contrasts with typical robot vision applications reported in the literature that were limited to the scope of their application and, in consequence, generalisation or even adaptations of further visual competences would be cumbersome to implement. The above provided the motivation to develop an integrated and novel hierarchical architecture for the active binocular robot vision system, as presented in Chapter 5.

The system was therefore conceptualised as an arrangement of behaviours abstracted in terms of how the robot vision system acted according to the attention for perception principle and object-based attention cueing approach. That is, the behaviour of the system was loosely based on two operational modalities of the human visual system: *pre-attentive* (for perception) and *attentive* (for action). Similarly, the system's behavioural architecture was inspired on the hybrid deliberative/reactive *Sensor Fusion Effector* architecture (SFX, Section 2.6.3) and the *hierarchical* paradigm (Section 2.6.2). In current literature, both architectures have been conceptualised as the paradigms that mimic the configuration of cognitive processes in the mammalian brain and resemble the arrangement of visual behaviours in the early stages of human evolution, respectively. The adoption of the SFX architecture in the active binocular robot vision therefore allowed the system to relate reasoning/deliberative components with motor and reactive visual behaviours. Whilst the hierarchical visual arrangement allowed reactive visual behaviours to be based on models of human visual streams (i.e. WHERE and WHAT streams).

By adopting the deliberative-hierarchy interconnection, it became possible to establish high-level task-goal specifications as macro scripts that are schedule/executed according to the state of the environment. By decoupling the high-level task from the specific robot's visual competences and sensing and motor related functions, visual behaviours were divided according to their visual objectives and functional operation and classified into a hierarchy according to their function and complexity. In consequence, they executed simple visual tasks, but when working in conjunction, as specified in the high-level script, the robot vision system was able to operate into two different contexts, namely, autonomous scene exploration and semi-autonomous object's appearance learning high-level tasks (Chapter 6).

In addition, the above separation of visual behaviours eliminated the need to integrate motor-related functions within the overall operation of the visual competences. Specifically, this thesis proposed an egocentric spatio-temporal map. This map was defined as a relative coordinate system where the frame of reference was deliberately located with respect to an internal *"homming"* position of the robot head state. Therefore, visual operations (e.g. inhibition of return, pre-attentive and attentive behaviours) were evaluated in terms of the retinotopic coordinates rather than in a global actuator space. This avoided the specification of a

global world map and, in consequence, actuation behaviours were entirely decoupled from the overall architecture. The above is biologically plausible (as pointed out by (Styles, 2005; Chun and Wolfe, 2004)) and specifically addresses the objective stating: *"How can a hierarchy of visual behaviours provide an open-ended architecture in terms of the task specification?"*

Particularly, the reported hierarchy of behaviour could not be compared with respect to state-of-the-art robot vision systems (Kragic et al., 2005; Björkman and Eklundh, 2005a; Wallraven and Bülthoff, 2007a; Rasolzadeh et al., 2010; Meger et al., 2008; Kootstra, 2010) since reported systems in the literature has only analysed the recognition rate achieved and, in consequence, the overall fixation performance of such systems has been qualitatively analysed (i.e a user assessed visually whether the object has been detected correctly or incorrectly). In addition, these systems have only been demonstrated in moderately cluttered scenes. In that regard, the investigation presented in Chapter 5 reported 100% of identifications. Furthermore, by carrying out a qualitative comparison with respect to these systems, it was argued therefore that the proposed hierarchy clearly outperforms such systems in terms of the system's innate visual capabilities, e.g. multiple same-class object identification and recognition.

The above results compared favourable with respect to the extended validation carried out in Chapter 3; however similar fixation errors and low correlation repetition agreed with the problem of locating an object due to the lack of a trained canonical pose in the database. This shortfall was initially identified in the extended validation of Chapter 3 and subsequently discussed in Section 8.1.1. Similarly, it was concluded that further characterisation of the devised behaviours was required in order to reliably characterise the hierarchical architecture while performing several trials over a variety of different types of scene (and observed objects). This extended validation was addressed in Chapter 7.

The hierarchical active binocular robot vision architecture is believed to be the first reported functional system in the literature. At present time, there is not a reported robot vision system capable of performing *discrimination, detection, recognition, identification, and same-different identification* visual tasks in complex and cluttered real-world environments. In addition, there is no other such robot architecture able to maintain vergence, direct its gaze, recognise multiple same objects, autonomously explore the environment in a unified and parsimonious hierarchy of visual behaviours architecture. Hence, the above addresses the third research question specified in this thesis, which is; *"Can the adoption of the cognitive model of the mammalian's brain provide potentially the means of modelling visual attention and visual streams for robotic applications?"*

8.1.4 Semi-automatic Object Appearance Learning Behaviour

The penultimate question specified in this research, namely, *"How could an active exploration-learning strategy of the object's appearance across its viewing sphere characterise and synthesise robustly robot's object knowledge?"* is addressed in Chapter 6.

To overcome the limitation of the system to correctly detect the object's pose when it was deviated from the canonical pose of the trained example, Chapter 6 presented a high-level visual learning behaviour that exploited the active vision paradigm and spatio-temporal properties of observed features in order to characterise and consolidate viewpoint invariant representation of the object's appearance. By adapting a motorised turntable into the actuation layer of the hierarchical architecture, this behaviour enabled the robot vision system to create its own visual knowledge by means of the dynamic interaction with the observed object (closely related to the behaviour observed in humans). Therefore, this behaviour avoided the need to build manually segmented databases which, as demonstrated in Chapter 5, produced incorrect localisations while searching a scene.

The underlying idea behind this behaviour was that objects are learned by snapshots of highly discriminative 2-dimensional views of the imaged object (Bülthoff and Edelman, 1992; Edelman and Weinshall, 1991). Its development therefore integrated and synthesised several novel learning and biological motivated principles reported in the literature. Similarly, it was established and defined in terms of underlying properties from which learned canonical representations must be composed of, namely, *goodness of recognition, familiarity, and functionality* (according to Blanz et al. (1996) and Ullman et al. (2002)). The primitive representations were finally consolidated in order to sample visual features across the object's viewing sphere.

In addition, such behaviour was integrated with the hierarchical architecture and employed a subset of previously defined visual behaviours to execute the task. Hence, the development of this behaviour demonstrated qualitatively the utility and applicability of the hierarchical architecture. In that respect, the hierarchical architecture now features visual behaviours for object and feature tracking, visual binding, active clustering of acquired knowledge that operates parsimoniously with previously defined behaviours. Thus, learning related visual behaviours inherited the properties of the hierarchical binocular robot architecture.

The literature contains no previous records of a visual learning behaviour that demonstrates the full active vision paradigm in a object appearance learning task.

8.1.5 Final Validation Outcomes

The last part of this thesis, as presented in Chapter 7, was twofold: the validation of the visual learning behaviour and the demonstration of the integrated hierarchical active binocular robot vision system featuring a self-learned or manually pre-trained object knowledge databases.

Hence, the validity of visual learning behaviour performance was tested in Section 7.2. These experiments consisted of learning the appearance of a collection of object examples by means of the devised learning behaviour. The experimental methodology firstly comprised on testing the overall integration of such visual learning behaviour. The hierarchical system is thus allowed to learn canonical representations of a set of presented unknown objects. Ten objects were considered and each consisted of different visual properties such as shape, texture, and so forth. The visual learning high-level task was invoked three times, each at different initial angular positions. These different initial positions characterised quantitatively the ability of the system to find similar canonical object poses despite the initial observed pose of the object. As demonstrated in Chapter 7, the system robustly found canonical representations within an average confidence interval (at 95% of significance) of ~ 15.4 degrees observing the same pose over the three trials. The visual learning behaviour required in average 4 canonical poses in order to characterise and, consequently, learn the object's viewsphere.

After the object's appearance of the proposed set was learned , the *goodness of recognition, familiarity* and *functionality* properties were then tested in isolation for each object in a pre-attentive localisation visual task. The experimental set-up in (Kootstra, 2010) was adopted in order to objectively compare both visual learning approaches. From a ground truth annotated database comprising 3600 images, it was therefore found that the reported approach in Chapter 6 outperformed the state-of-the-art learning algorithm reported in (Kootstra, 2010). Notably, the devised behaviour achieved an average of 96.8% of recognition rate as opposed to 85% of its counterpart. This experiment also measured the localisation error in order to test the *familiarity* and *functionality* properties. It was found that the worst MSE percentage observed error was $\sim 2.86\%$ with respect to the annotated region of interest of the detected object in the image.

The final demonstration and extended validation of the integrated hierarchical system was also presented in Chapter 7. These experiments consisted of 7 complex scenes which comprised 20 multiple same-class object instances embedded with clutter. Objects were positioned such that they observed occlusion with similar unknown objects and, also, self-occlusion with one or more identical instances. The hierarchical robot vision architecture successfully explored each scene while using: manually segmented object databases sampled at intervals of 45 and 60 degrees, respectively; and, also, the learned object database of previous experiments.

Therefore, it was observed that the learned database, in an active exploration of the scene task, outperformed its manually pre-trained object databases counterparts; that is, an average of $\sim 98.8\%$ for the learned database as opposed to 95% with a pre-trained database sampled at 45%. Furthermore, an analysis of the incurred fixation errors while attending objects demonstrated that the learned object database observed lower fixation deviations than the manually pre-trained databases. Finally, the maximum overall compression ratio observed was 14.2% with respect to a pre-trained object database sample at 30 degrees.

Therefore, the above results validated the last objective described in Chapter 1, which is: *"To evaluate the performance and quantitatively measure the robustness of the described robot vision system with respect to current the state-of-the-art robot systems"*. In consequence, the functional operation of the hierarchical active binocular robot vision architecture confirms the validity of the underlying hypothesis of this thesis.

8.2 Future Work

This section provides possible research directions and future extensions of the research presented in this thesis.

8.2.1 Visual Tasks

The main objective the robot vision architecture herein reported is to execute *discrimination, detection, recognition, identification, and same-different identification*, visual tasks for autonomous scene exploration and object appearance learning. Nevertheless, (Tsotsos et al., 2008) has defined some extra tasks that were not considered in this thesis such as categorisation and within-category visual tasks. These could therefore be potentially integrated into the hierarchical robot architecture.

- *Categorisation tasks* would dictate the robot to link observed stimuli to a class exemplar of similar visual stimuli rather than specific identities of each stimuli. For instance, the robot must determine whether dogs belong to the class of animals or mammals.
- *Within-Category* tasks would require the robot to have a stimulus associated with a particular sub-category from a class (e.g., bungalows, flats, house tree and so forth).

These visual tasks, therefore, required to develop new visual behaviours within the hierarchy. These could be based on the popular bag-of-words models (Mikolajczyk et al., 2006) or a

hierarchical representation of extracted features sub components (Fidler and Leonardis, 2007). To that end, SIFT features might be limited and further basic features need to be considered, these are briefed in the following section.

8.2.2 Visual Feature Representation

In order to tackle visual understanding in dynamic and complex environments and, in turn, have better visual knowledge for object-based attention models, it is proposed to investigate additional means of enriching the underlying visual representations, Accordingly, it might be possible to implement state-of-the-art feature extractors.

For example, the colour-SIFT feature detector, reported in (Burghouts and Geusebroek, 2009b), which can drive attention based on the chromatic properties of objects, would underpin further visual competences and, in consequence, would make a direct impact on the classification of same object instances with same/different colours. Similarly, a 2.5D feature descriptor (Lo and Siebert, 2008) applied to range images which can directly improve object identification by encoding stable features based on range surface topology.

The visual representation in the hierarchical architecture can potentially include the core architecture on the Gaussian and Laplacian multi-resolution image pyramids. These structures serve the needs of the above 2D intrinsic image property extraction, local feature extraction, optical flow-field recovery and dense range map acquisition. Recovered range maps are also computed as pyramids to provide multi-resolution 2.5D intrinsic property maps such as shape index and range surface normals.

A key aspect of the proposed vision architecture would be the use of foveated pyramids, where a constant "core" set of image fields might be extracted from top to bottom of the image pyramid and serves all image functions from feature extraction to depth recovery. At the centre of the systems Field Of View (FOV) full resolution is available while the lowest pyramid level (the same size as the core) provides coverage of the full FOV, and intermediate pyramid levels provide varying degrees of coverage accordingly. Two benefits accrue from this fovea visual representation, namely execution speed since only a small area of the input image needs to be processed at full resolution which results of the suppression of unwanted high frequency features, i.e. attention control.

The current object learning system is based on extending and integrating a number of standard modules for SIFT feature extraction boundary detection and flow-field recovery. The potential exists to improve and adapt these modules, for example boundary segmentation could be integrated with anisotropic diffusion to provide simplified boundaries or boundary extraction

over multi-modal image properties (Mabaar M.;Siebert, 2008). Likewise the potential for characterising and matching boundaries would be supported by the proposed division system architecture.

8.2.3 Extending the Multiple Same-Class Detector

Despite being the multiple same-class object instances detector robust to carry out the task required in this thesis, the localisation of multiple instances was constrained to the two-dimensional domain. Thus, the full three-dimensional pose registration of object instances for object manipulation was not considered. Consequently, in order to extend the system is required to either:

- adopt a 2.5D visual understanding model (as reported in Lo and Siebert (2009)) such that depth information is encoded into the overall visual processing capabilities of the system, or,
- include pose and orientation information in the object class database such that the projective transformation in the image plane captures the perspective orientation of the objects.

In the one hand, the continuous Hough space adopted in the multiple same-class object detection algorithm requires to be extended into 2.5D domain by considering a 6-dimenasional space (x and y position, scale, orientation, tilt, and pan). Thereafter, the devised algorithm could be evaluated as described. On the other hand, a database of an unordered image dataset could be created as reported by Brown M. and Lowe (2005) (this paper presents an algorithm to recover the three-dimensional structure of objects based on a point-based feature extraction technique, e.g. SIFT features) such that pose information would be logically indexed to the object class model. Therefore, the affine pose estimator (Section 4.6) step must be upgraded such that a projective transformation matrix would be estimated from the model object class samples in the database. In this case, the features projected into the continuous Hough space and the unsupervised clustering algorithm would be evaluated as described in Chapter 4.

Similarly, in order to improve the detection and localisation performance, it would be necessary to investigate more advanced clustering methods (for e.g. Frigui and Krishnapuram (1999); Rosenberger and Chehdi (2000); Sanguinetti (2008)). Notably, soft-assignment techniques, such as the work reported in van Gemert et al. (2010) for visual code book modelling, have demonstrated to avoid the hard partitioning of fuzzy C-means and enhance the performance of the proposed method in this thesis.

8.2.4 Visual Learning Behaviour Extension

Based on the underlying learning principles outlined in this thesis, it might be possible to extend the learning behaviour to cope with categorisation learning and create abstract representations of visual knowledge (as pointed out by Palmeri and Gauthier (2004)). This would link the gap between robotics and the computationally expensive bag of words models into a parsimonious integrated architecture. That is, if an object is "similar" (in terms of its class category) to a learned object, a new category (or class) would be created, and a hierarchical structure can be adopted.

The devised learning behaviour could also be extended to allow the system to learn from visual experience. That is, it might be possible to start learning and understanding objects, then categorise similar objects and create classes for those objects (a human teacher might supervise these processes) until a hierarchy of abstract concepts is created in terms of a collection of several basic features.

8.2.5 Deliberative/Reasoning Layer

The macro script employed to defined the high-level task (as described in Chapters 1, 5, 6) specified the sequential activation of behaviours in order to accomplish a defined goal (sense-plan-act convention). However, the hierarchical architecture followed what it is specified without being "*conscious*" of the undertaken task. In this regard, a *deliberative/reactive architecture* (as conceptualised in this thesis) could potentially enable the robot vision system to have some degree of intelligence to interact and/or solve complex goals over challenging situations (examples can be found in Bonasso et al. (1997) and Murphy and Mali (1997)). That is, the *deliberative* component would refer to the layer of intelligence that is conscious of the actions of the robot (sense-act while planning convention) (Murphy and Mali, 1997) and this layer would potentially afford operational flexibility. By providing a reasoning mechanism, visual search strategies might be no longer "hard wired", but constructed in response to specific goals.

Hence, it is envisaged that the adoption of the devised hierarchical architecture, a cognitive active robot vision system could be designed in order to generate (by reasoning) the sequence of behaviours that must be triggered in order to achieve a specified goal. Therefore, the deliberative macro script described in Algorithm 5.1 might be replaced by a possible high-level cognitive machine which:

- translates user input into robot commands (e.g. natural language programming, for

example as in Veres (2008)),

- allocates available resources to fulfil the task-goal specification and
- evaluate the performance while executing the task and recover from system failures.

8.2.6 Foveated Vision

As the hierarchical architecture is decoupled from the actuation and visual sensing functions, a promising and immediate extension of the reported research in this thesis could comprise the implementation of foveated visual representation. That is, *foveated image pyramids* based on Burt's model Burt (1988); Boyling and Siebert (2000) could be employed to limit the processing field of view of the robot vision system. This model consists of iteratively reducing a portion of the periphery, such that the high resolution part is used for verging the cameras of a binocular robot (Boyling, 2002). Specifically, this model would create a stack of sub-sampled images like the scale-space pyramid employed in SIFT (section 2.2.2), and would crop the pyramid by employing a window patch of the same size of the smallest image in the said pyramid. Therefore, the high acuity detail would be concentrated in a small area in the image whereas the smallest, low acuity detail covers the full field of view. The above could be implemented within the SIFT framework and, in consequence, it would explore the paradigms that foveated vision embraces while exploring a scene.

8.2.7 Cognitive Robot Vision Machines

An exciting but rather ambitious extension of this research is to develop a fully cognitive, reasoning machine that can carry out human-like tasks. For example, manipulation of domestic objects, scene/object visual understanding, social interactions and so forth. Notably, in these target applications, cognition might rely upon the recursion of basic visual and interaction capabilities in order to start with an evolution process towards problem-solving, robot localisation, action learning and world/environment interaction, reasoning and understanding (i.e. the system would develop its own environment representations, actions and language structure grammar with either human-supervision or unsupervised).

Appendix A

Enhancements to the Robot Head

A.1 Actuator Control Module

The actuator control module was designed by McDougall (2004) and latterly employed in its original form by Fattah (2007) in order to operate and control the four motors of the active binocular robot head. As MATLAB is the programming language of choice in (McDougall, 2004; Fattah, 2007) and, in consequence, in this thesis, an interface to access the actuator cards is designed to allow the communication with the hardware and native commands in accordance with the hardware specifications.

Physik Instrumente GmbH & Co.,[1] the manufacturer of the actuator cards, provides a standalone software and a complete set of commands to be accessed in different programming languages such as; *C*, *Basic* and *FORTRAN*. McDougall (2004) implements this interface by means of MATLAB API calls which allows to be accessed by means of external routines written in C-language. Specifically, these routines are written in C-code and compiled into MATLAB format (known as *MEX files*) in the form of a *Windows Dynamic-Link Library* (DLL). A short tutorial on how to compile *MEX files* in MATLAB and the integration to the *API* is described in McDougall's MSc project (McDougall, 2004).

The ultimate aim of designing this interface consists of controlling the motors as part of an integrated software vision framework and it must be invoked through MATLAB command line instructions. The designed steps to drive the actuators are as follows (according toMcDougall (2004));

1. Activate actuator cards.

[1] `http://www.physikinstrumente.co.uk/`

2. Set the default base address for the card which is going to be used (0x210H or 0x214H).
3. Initialise board.
4. Initialise the axis to be used.
5. Set the actuator maximum velocity.
6. Move the desired actuator for the required number of steps using the translate command.
7. Deactivate actuator cards.

It can be noticed that each routine call only drives one axis on the actuator card at a time. This results in an inefficient design as the robot head comprises four actuators and, in a common visual search task, all of them have to be moved simultaneously. Additionally, the routine call waits until the desired angular displacement of the camera is reached. This is consequently an extremely slow process. It can be further notice that the aforesaid displacement produce a linear translation on the camera traces illustrated in Figure 3.11.

Hence, the integration of 4 motors can be implemented by a single routine. That is, this routine should call and access those actuators that are required. In order to fully implement this capability, a detailed search on the manufacturer manuals reveals that a native command function (named *"VectorA"*) might operate two actuators at the same time in just one procedure call. That is, the camera angular displacement, velocity and acceleration are treated as directional vectors and, therefore, the movement is synchronised in the perpendicular and parallel components. The capital letter *"A"* in the instruction denotes an absolute movement. This operation is commonly known as a linear interpolation towards a desired coordinate point. The syntax of this function is;

$$VectorA(x, y, d_1, d_2, speed, acceleration) \tag{A.1}$$

where x and y denote the desired axis to be moved and; d_1 and d_2 the corresponding distances.

Since the described instruction can only drive two actuators at the same time, the implementation of the 4 actuator case is enabled by integrating in the same C-routine two linear interpolation instructions. Thus, the routine is modified accordingly at the 4 at 7 steps (described above) by initialising the desired axes to be moved and the '*VectorA*' instruction. The pseudo-code for this implementation can be depicted on what follows.

This code is therefore compiled in a DLL-MEX file. Additionally, the routine in Algorithm A.1 returns the relative distances travelled.

Algorithm A.1 Pseudo-code for controlling the Robot Head motors

```
Defining variables in the MEX file();
Assigning velocity and acceleration constrains();

Initialise and Activate actuator cards();

VectorA(1, 2, amount1, amount2, speed, acceleration);
VectorA(3, 4, amount3, amount4, speed, acceleration);

WHILE axis 1, axis 2, axis 3 or axis 4 is moving
  Wait until they have reached the desired angular positions;
ENDWHILE

Print results();
Deactivate actuator cards();
```

In regard with the integration on the robot head system, it is necessary to modify the previous commands with the command in A.2. Furthermore, the gaze control module is also adapted to operate with the new hardware interface.

$$moveACT(speed, acceleration, amount1, amount2, amount3, amount4); \qquad \text{(A.2)}$$

The incurred execution time to move the actuators is successfully reduced by three quarters of the overall time. Furthermore, the cameras are now capable of following a smooth angular displacement while executing visual tasks. This enables to develop visual competences where a continuous and smooth translation is required, such as: smooth pursuit Coombs (1992).

A.2 Image Capturing Module

The *Image Acquisition Toolbox* in MATLAB provides a complete set of functions to operate cameras of almost any model existing in the market. Two main model types are defined in the toolbox; "*WinVideo*" driver for almost all webcams and "*CMU*" driver which is a generic driver developed by the *Carnegie Mellon University* for the IEEE-FireWire interface. The latter is therefore used in McDougall (2004) and Fattah (2007); however, the implemented function is extremely slow when acquiring a single image with a resolution of 1024×768 pixels.

Originally, the capturing function performs a 3 step process: *Initiate cameras-Acquire-Close cameras*, each time it is invoked. For instance, in the vergence behaviour of Section 3.4, it is necessary to capture an image each vergence cycle, the camera object, as MATLAB required,

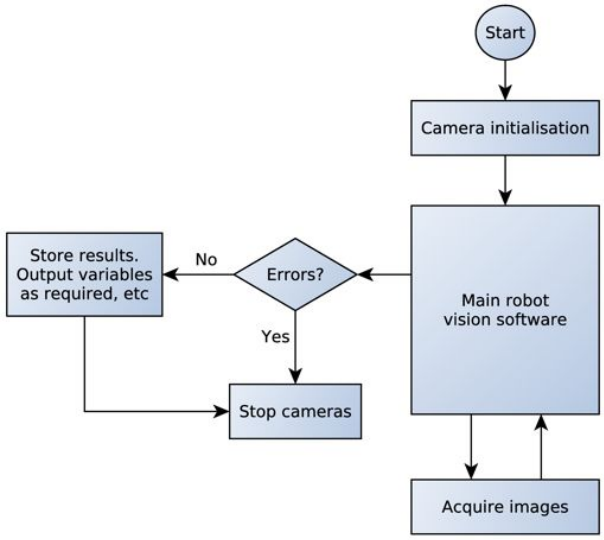


Figure A.1: Proposed image capturing module.

is initiated; then images are captured and, finally, the camera object is closed. This process is appropriate if only one image have to be captured; however, in this example two images are captured and, as described in the pilot experiments of Section 3.6, the vergence cycle takes up to 6 iterations to stabilise. Thus, this process must be initiated twelve times and, furthermore, the camera object initialisation under MATLAB requires an average of 30 seconds. This incurred time is therefore accumulated on a normal visual search task that increasingly delays the system.

Thus, it is clear that the problem lies in the camera object initialisation step and the repeated creation of the camera object. To that end, the proposed solution consists of three different functions, not sequential steps as it was conceived initially. The first function is related to the object cameras initialisation which is invoked for both cameras when a visual tasks is required. The second function is only dedicated for the image acquisition of both cameras. Finally, the object is closed either when the visual task is finished or when an error event occurs. This algorithm can be depicted in Figure A.1.

With this improvement, the overall time is reduced by a half of the incurred time. Moreover, both improvements implemented on the active binocular robot head, a visual search task as in Figure 3.9 is reduced from one hour to $20 - 25$ minutes.

It must be pointed out that the image capturing hardware and, in consequence, the software implementation of the active binocular robot head were modified during the course of the research project. However, the design principle of the proposed solution is the same in future hardware integrations.

Bibliography

Aloimonos, J., Weiss, I., Bandopadhay, A., Jan. 1988. Active vision. International Journal of Computer Vision 1 (4), 333–356. 1.3, 1.3, 2.1, 2.1.1, 4.1, 8.1.2

Aloimonos, Y., 1993. Introduction: Active vision revisited. Active perception Hillsdale, NJ: Lawrence Erlbaum Associates., Ch. Chapter 0, pp. 1–18. 2.1.1

Aragon-Camarasa, G., Fattah, H., Siebert, J. P., Mar. 2010. Towards a unified visual framework in a binocular active robot vision system. Robotics and Autonomous Systems 58 (3), 276–286. (document), 3.1, 3.2, 3.3, 3.4, 3.5, 3.6.2, 3.7, 3.8, 3.8.1, 3.13, 3.15, 3.1, 3.16, 3.17, 3.18, 3.19, 5.10

Aragon-Camarasa, G., Siebert, J. P., 2009. A Hierarchy of Visual Behaviours in an Active Binocular Robot. In: Kyriacou, T., Nehmzow, U., Melhuish, C., Witkowski, M. (Eds.), Towards Autonomous Robotic Systems, TAROS 2009. University of Ulster, pp. 88–95. (document), 5.5, 5.7, 5.9, 5.10, 5.11

Aragon-Camarasa, G., Siebert, J. P., Aug. 2010. Unsupervised clustering in Hough space for recognition of multiple instances of the same object in a cluttered scene. Pattern Recognition Letters 31 (11), 1274–1284. (document), 2, 4.6, 4.7, 4.8, 4.9, 4.10, 4.11

Arbib, M., Metta, G., der Smagt, P., 2008. Neurorobotics: From Vision to Action. In: Siciliano, B., Khatib, O. (Eds.), Springer Handbook of Robotics. Springer Berlin Heidelberg, pp. 1453–1480. 2.6.2

Arnow, T., Bovik, A. C., Mar. 2008. Foveated Object Recognition Using Corners. IEEE. 2.4

Aziz, M., Mertsching, B., Salah, M., Shafik, E.-N., Stemmer, R., 2006. Evaluation of Visual Attention Models for Robots. IEEE, Washington, DC, USA. 1.3

Aziz, M. Z., Mertsching, B., 2008. Visual search in static and dynamic scenes using fine-grain top-down visual attention. In: Proceedings of the 6th international conference on Computer vision systems. ICVS'08. Springer-Verlag, Berlin, Heidelberg, pp. 3–12. 2.4

Balasuriya, L. S., 2006. A Computational Model of Space-Variant Vision Based on a Self-Organised Artificial Retina Tessellation. Ph.D. thesis, Department of Computing Science, University of Glasgow. 2.4, 2.4.3

Ballard, D., 1981. Generalizing the Hough transform to detect arbitrary shapes. Pattern Recognition 13 (2), 111–122. 2.3.1

Ballard, D. H., Feb. 1991. Animate vision. Artificial Intelligence 48 (1), 57–86. 1.3, 1, 2.1.1, 2.4, 2.4.2, 2.4.3, 2.5, 2.6, 2.7.1, 8.1.2

Bay, H., Ess, A., Tuytelaars, T., VanGool, L., Jun. 2008. Speeded-Up Robust Features (SURF). Computer Vision and Image Understanding 110 (3), 346–359. 2.2.1

Behnke, S., Rojas, R., 2001. A Hierarchy of Reactive Behaviors Handles Complexity. In: Balancing Reactivity and Social Deliberation in Multi-Agent Systems, From RoboCup to Real-World Applications (selected papers from the ECAI 2000 Workshop and additional contributions). Springer-Verlag, London, UK, pp. 125–126. 2.6.2

Berman, R., Colby, C., Jun. 2009. Attention and active vision. Vision Research 49 (10), 1233–1248. 1.3, 1.3, 2.4

Bernardino, A., 2004. Binocular Head Control with Foveal Vision: Methods and Applications. Phd, Instituto Superior Técnico, Lisbon. 2.1.2, 2.4.3

Bernardino, A., Santos-Victor, J., 1996. Sensor Geometry for Dynamic Vergence: Characterisation and Performance Evaluation. In: Workshop on Performance Characteristics of Vision Algorithms. pp. 55–59. 2.4

Bernardino, A., Santos-Victor, J., Nov. 1998. Visual behaviours for binocular tracking. Robotics and Autonomous Systems 25 (3-4), 137–146. 6.3

Bezdek, J. C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms (Advanced Applications in Pattern Recognition). Springer. 2.5.2.1

Björkman, M., Eklundh, J. O., 2004. Attending, foveating and recognising objects in real world scenes. In: Proceedings of British Machine Vision Conference. 2.2.1, 2.4.1, 2.4.3, 5.10

Björkman, M., Eklundh, J. O., 2005a. Recognition of objects in the real world from a systems perspective. Kuenstliche Intelligenz 19 (2), 12–17. 1.2, 2.1.2, 2.8, 3.2, 3.3, 3.4, 5.10, 8.1.3

Björkman, M. r., Eklundh, J. O., 2005b. Foveated Figure-Ground Segmentation and Its Role in Recognition. In: British Machine Vision Conference. 2.1.2

Björkman, M. r., Eklundh, J.-O., 2006. Vision in the real world: Finding, attending and recognizing objects. International Journal of Imaging Systems and Technology 16 (5), 189–208. 2.8

Blanz, V., Tarr, M. J., Bülthoff, H. H., 1996. What Object Attributes Determine Canonical Views? Tech. rep., Max-Planck-Institut. 2.3, 2.7.1, 6.3, 6.3.1, 8.1.4

Bonasso, R. P., Firby, J., Gat, E., Kortenkamp, D., Miller, D., Slack, M., 1997. A Proven Three-tiered Architecture for Programming Autonomous Robots. Journal of Experimental and Theoretical Artificial Intelligence 9 (2), N/A. 2.6, 8.2.5

Bouguet, J., 2002. Pyramidal implementation of the lucas kanade feature tracker: Description of the algorithm. Tech. rep., Microprocessor Research Labs, Intel Corporation.
URL `robots.stanford.edu/cs223b04/algo_tracking.pdf` 6.3, 6.6.1.1

Boyling, T. A., 2002. Active Vision for Autonomous 3D Scene Reconstruction. Ph.D. thesis, University of Glasgow, Department of Computing Science. (document), 2.1.1, 2.1.2, 2.1.3, 2.2, 3.4, 8.2.6

Boyling, T. A., Siebert, J. P., Jun. 2000. A Fast Foveated Stereo Matcher. In: Conference on Imaging Science Systems and Technology (CISST 2000). AAAI Press, Las Vegas, USA, pp. 417–423. 3.4, 8.2.6

Brooks, R., 1986. A robust layered control system for a mobile robot. Robotics and Automation, IEEE Journal of 2 (1), 14–23. 2.6.1, 2.6.2

Brooks, R. A., 1991. How to build complete creatures rather than isolated cognitive simulators. In: Architectures for Intelligence. Erlbaum, pp. 225–239. 2.6, 2.6.1, 2.6.1, 2.6.2, 3.4

Brown M., Lowe, D., 2005. Unsupervised 3D object recognition and reconstruction in unordered datasets. In: Fifth International Conference on 3-D Digital Imaging and Modeling, 2005, 3DIM 2005. pp. 56–63. 8.2.3

Bülthoff, H., Wallraven, C., Giese, M. A., 2008. Perceptual Robotics. In: Siciliano, B., Khatib, O. (Eds.), Springer Handbook of Robotics. Springer, Ch. Part G, pp. 1481–1498. 2.3

Bülthoff, H. H., Edelman, S., Jan. 1992. Psychophysical support for a two-dimensional view interpolation theory of object recognition. Proceedings of the National Academy of Sciences of the United States of America 89 (1), 60–4. 2.3, 2.7.1, 8.1.4

Bulthoff, H. H., Wallraven, C., Graf, A., 2002. View-based dynamic object recognition based on human perception. In: Pattern Recognition, 2002. Proceedings. 16th International Conference on. 2.7.1

Burghouts, G. J., Geusebroek, J.-M., Feb. 2009a. Material-specific adaptation of color invariant features. Pattern Recognition Letters 30 (3), 306–313. 4.1, 4.8

Burghouts, G. J., Geusebroek, J. M., Jan. 2009b. Performance evaluation of local colour invariants. Computer Vision and Image Understanding 113 (1), 48–62. 4.9, 8.2.2

Burt, P. J., 1988. Attention mechanisms for vision in a dynamic world. In: 9th International Conference on Pattern Recognition, 1988. pp. 977–987 vol.2. 8.2.6

Chun, M. M., Wolfe, J. M., 2004. Visual Attention. In: Goldstein, E. B. (Ed.), The Blackwell Handbook of Perception. Blackwell Publishers Ltd., Ch. 9, pp. 272–310. 1.3, 2.4, 2.4.2, 2.5, 2.6, 2.7.1, 4.8.1, 5.2, 6.3, 8.1.3

Coombs, D. J., Jan. 1992. Real-time gaze holding in binocular robot vision. Ph.D. thesis, University of Rochester, Department of Computer Science. 6.3, A.1

Crevier, D., 1993. AI: the tumultuous history of the search for artificial intelligence. Basic Books. 1.3

Cyganek, B., Siebert, J. P., Jan. 2009. An Introduction to 3D Computer Vision Techniques and Algorithms. John Wiley & Sons, Ltd, Chichester, UK. 6.5.2

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005 (CVPR 2005). Vol. 1. pp. 886–893. 2.2.1

Das, D., Mansur, A., Kobayashi, Y., Kuno, Y., Dec. 2008. An Integrated Method for Multiple Object Detection and Localization. In: Bebis, G., Boyle, et. al. (Eds.), Advances in Visual Computing. Vol. 5359 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 133–144. 5.10

Davies, E. R., 2005. Machine Vision: Theory, Algorithms, Practicalities (Signal Processing and its Applications), 3rd Edition. Morgan Kaufmann. 2.5.1

Davis, J., Goadrich, M., 2006. The relationship between Precision-Recall and ROC curves. ACM Press, New York, New York, USA. 4.8, 4.8.1

Davison, A. J., Murray, D. W., Jul. 2002. Simultaneous localization and map-building using active vision. Pattern Analysis and Machine Intelligence, IEEE Transactions on 24 (7), 865–880. 2.1.1, 2.2.1

Dong, L., Yu, X., Li, L., Hoe, J., Dec. 2010. Hog based multi-stage object detection and pose recognition for service robot. In: Control Automation Robotics Vision (ICARCV), 2010 11th International Conference on. pp. 2495–2500. 2.2.1

Edelman, S., Weinshall, D., Jan. 1991. A self-organizing multiple-view representation of 3D objects. Biological Cybernetics 64 (3), 209–219. 2.7.1, 8.1.4

El-Sonbaty, Y., Ismail, M. A., 2003. Matching Occluded Objects Invariant to Rotations, Translations, Reflections, and Scale Changes. Lecture Notes in Computer Science 2749/2003, 836–843. 2.5.1

Fattah, H., 2007. Combining Binocular Vergence, Gaze Control and Object Recognition in Robotic Vision. Master's thesis, University of Glasgow, Department of Computing Science. 1.2, 1.5, 2.1.3, 2.4, 3, 3.1, 3.4, 3.4, 3.5, 3.7, 3.9, 8.1, 8.1.1, A.1, A.2

Fattah, H., Aragon-Camarasa, G., Siebert, J. P., 2008. Towards Binocular Active Vision in a Robot Head System. In: Ramamoorthy, S., Hayes, G. M. (Eds.), Towards Autonomous Robotic Systems, TAROS 2008. University of Edinburgh, pp. 25–32. (document), 2.1.3, 2.2.2, 2.4, 2.4.3, 3, 3.1, 3.3, 3.4, 3.4, 3.5, 3.6, 3.5, 3.5.1, 3.6, 3.6.1, 3.6.2, 3.7, 3.8, 3.9, 3.6.2, 3.6.2, 3.10, 3.11, 3.6.3, 3.8.1, 8.1

Fawcett, T., Jun. 2006. An introduction to ROC analysis. Pattern Recognition Letters 27 (8), 861–874. 2.5.1, 4.7, 4.8.1

Fay, R., Kaufmann, U., Knoblauch, A., Markert, H., Palm, G., 2005. Combining Visual Attention, Object Recognition and Associative Information Processing in a NeuroBotic System. In: Biomimetic Neural Learning for Intelligent Robots. Vol. 3575. Lecture Notes in Computer Science, Ch. Part I: Bi, pp. 118–143. 2.6.3

Fazl, A., Grossberg, S., Mingolla, E., Feb. 2009. View-invariant object category learning, recognition, and search: how spatial and object attention are coordinated using surface-based attentional shrouds. Cognitive psychology 58 (1), 1–48. 1.4, 2.3, 2.4, 2.4.2, 2.7.1, 5.2, 5.3, 6.2, 6.3

Ferrari, V., Tuytelaars, T., Gool, L., Jan. 2006. Simultaneous Object Recognition and Segmentation from Single or Multiple Model Views. International Journal of Computer Vision 67 (2), 159–188. 2.1.1

Ferrier, N. J., Clark, J. J., 1993. The Harvard Binocular Head. International Journal on Pattern Recognition and Aritifial Intelligence 7 (1), 9–31. 2.1.2

Fidler, S., Leonardis, A., 2007. Learning hierarchical representations of object categories for robot vision. In: International Symposium of Robotics Research. 8.2.1

Fikes, R. E., Nilsson, N. J., 1971. Strips: A new approach to the application of theorem proving to problem solving. Artificial Intelligence 2 (3-4), 189–208. 1.3

Findlay, J. M., Gilchrist, I. D., 2001. Visual attention: the active vision perspective. Springer Verlag, Ch. Vision and Attention, pp. 85–106. 2.1.1, 2.4.1

Findlay, J. M., Gilchrist, I. D., 2003. Active Vision: The Psychology of Looking and Seeing. Vol. 1 of Oxford Psychology Series. Oxford University Press. 2.1.1, 2.3, 2.4, 2.4.1

Forssen, P.-E., May 2007. Learning Saccadic Gaze Control via Motion Prediciton. In: Fourth Canadian Conference on Computer and Robot Vision (CRV '07). IEEE, Montreal, Quebec, Canada, pp. 44–54. 2.4.2

Forssen, P.-E., Meger, D., Lai, K., Helmer, S., Little, J. J., Lowe, D. G., May 2008. Informed visual search: Combining attention and object recognition. In: 2008 IEEE International Conference on Robotics and Automation. IEEE, pp. 935–942. 1.1, 2.1.2, 2.7.2, 4.1

Franconeri, S. L., Jonathan, S. V., Scimeca, J. M., 2010. Tracking Multiple Objects Is Limited Only by Object Spacing, Not by Speed, Time, or Capacity. Psychological Science 21 (7), 920–925. 2.5, 4.8.1

Frigui, H., Krishnapuram, R., May 1999. A robust competitive clustering algorithm with applications in computer vision. IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (5), 450–465. 2.5.2.3, 8.2.3

Frintrop, S., 2006. VOCUS: a visual attention system for object detection and goal-directed search. Vol. 3899 of Lecture notes in artificial intelligence. Springer. 2.4.2

Frintrop, S., Nüchter, A., Surmann, H., 2005. Visual Attention for Object Recognition in Spatial 3D Data. In: Paletta, L., Tsotsos, J. K., Rome, E., Humphreys, G. (Eds.), Attention and Performance in Computational Vision. Vol. 3368 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 168–182. 2.4.1

Geusebroek, J.-M., Burghouts, G. J., Smeulders, A. W., Jan. 2005. The Amsterdam Library of Object Images. International Journal of Computer Vision 61 (1), 103–112. (document), 4.1, 4.8, 4.7

Girard, B., Berthoz, A., 2005. From brainstem to cortex: Computational models of saccade generation circuitry. Progress in Neurobiology 77 (4), 215–251. 1.3, 5.2, 5.3

Grosso, E., Manzotti, R., Tiso, R., Sandini, G., 1995. A space-variant approach to oculomotor control. In: Proceedings of International Symposium on Computer Vision - ISCV. IEEE Comput. Soc. Press, pp. 509–514. 2.4

Grubbs, F. E., 1969. Procedures for Detecting Outlying Observations in Samples. Technometrics 11 (1), 1 – 21. 4.4

Harris, C., Stephens, M., 1988. A Combined Corner and Edge Detector. In: Proceedings of the Fourth Alvey Vision Conference. No. 15. pp. 147–151. 6.6.1.1

Hartley, R., Zisserman, A., 2004. Multiple View Geometry in Computer Vision. Cambridge University Press. 6.5.2

Heng, J., 1995. The design and application of a unique anthropomorphic sensor platform. Ph.D. thesis, University of Strathclyde, Department of Design, Manufacture, and Engineering Management. 2.1.1, 2.1.2

Hough, V., Paul, C., 1962. Methods and means for recognizing complex patterns. 2.3.1

Hubel, D. H., Wiesel, T. N., 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology 160 (1), 106–154. 2.2, 2.2.1

Hyundo, K., Murphy-Chutorian, E., Triesch, J., 2006. Semi-autonomous Learning of Objects. In: 2006 IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06). pp. 145–145. 2.7.2, 6.2

Itti, L., Jun. 2000. A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Research 40 (10-12), 1489–1506. 2.4, 2.4.2

Itti, L., Koch, C., Mar. 2001. Computational modelling of visual attention. Nature reviews. Neuroscience 2 (3), 194–203. 2.2.1

Johnson, A., Matthies, L., 2000. Machine vision for autonomous small body navigation. In: 2000 IEEE Aerospace Conference. Proceedings (Cat. No.00TH8484). pp. 661–671. 2.2.1

Johnson, S. C., Sep. 1967. Hierarchical clustering schemes. Psychometrika 32 (3), 241–254. 2.5.2.2

Kaufman, L., Rousseeuw, P. J., 2005. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley-Blackwell. 2.5.2, 2.5.2.1, 2.5.2.2, 4.4, 4.5

Kootstra, G., 2010. Visual Attention and Active Vision: From natural to artificial systems. Phd thesis, University of Groningen, Faculty of Mathematics and Natural Sciences. 1.2, 2.7.2, 7.2.1.1, 7.2.2, 7.2.2.1, 7.3.1, 7.3.1, 7.4, 8.1.3, 8.1.5

Kootstra, G., Ypma, J., de Boer, B., May 2008. Active exploration and keypoint clustering for object recognition. In: Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on. pp. 1005–1010. 7.2.2

Kragic, D., Björkman, M., Christensen, H. I., Eklundh, J. O., Jul. 2005. Vision for robotic object manipulation in domestic settings. Robotics and Autonomous Systems 52 (1), 85–100. 1.2, 2.1.2, 2.2.2, 2.3, 2.4.1, 2.4.2, 2.8, 3.2, 5.1, 5.10, 8.1.3

Lehrer, M., Bianco, G., 2000. The turn-back-and-look behaviour: bee versus robot. Biological Cybernetics 83 (3), 211–229. 6.3

Li, M., Betsis, D., Lavest, J., 1994. Kinematic calibration of the KTH head-eye system. Tech. Rep. CVAP171, Royal Institute of Technology (KTH), Stockman, Sweden. 2.1.2

Lo, T. R., Siebert, J. P., 2008. SIFT keypoint descriptors for range image analysis. Annals of the BMVA 3, 1–18. 8.2.2

Lo, T.-W. R., Siebert, J. P., Dec. 2009. Local feature extraction and matching on range images: 2.5D SIFT. Computer Vision and Image Understanding 113 (12), 1235–1250. 4.9, 8.2.3

Lowe, D., 1999. Object recognition from local scale-invariant features. In: Proceedings of the Seventh IEEE International Conference on Computer Vision. Vol. 2. IEEE, pp. 1150–1157 vol.2. 2.2.2

Lowe, D., 2001. Local feature view clustering for 3D object recognition. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001. IEEE Comput. Soc, pp. I–682–I–688. 2.3, 2.3.1, 5.5.1

Lowe, D. G., Nov. 2004. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60 (2), 91–110. (document), 2.2.1, 2.2.2, 2.3, 2.4, 2.5, 2.3, 2.3.1, 2.3.1, 2.3.1, 3.1, 3.3, 4.1, 4.6, 4.8, 5.4, 5.4, 8.1.1

Lucas, B., Kanade, T., 1981. An iterative image registration technique with an application to stereo vision. In: Proceedings of the 7th international joint conference on Artificial intelligence. Vol. 2. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 674–679. 6.6.1.1

Mabaar M.;Siebert, J. P., 2008. Smoothing Disparity Maps Using Intensity-Edge Guided Anisotropic Diffusion. In: Medical Image Understanding and Analysis 2008, 2nd-3rd July 2008, University of Dundee, Dundee, Scotland. 8.2.2

Manly, B. F., 2004. Multivariate Statistical Methods: A Primer. Chapman and Hall/CRC. 4.4

Marr, D., Ullman, S., Poggio, T., 1983. Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information. W. H. Freeman & Company, San Francisco. 2.1, 2.3

Matas, J., Chum, O., Urban, M., Pajdla, T., 2002. Robust wide baseline stereo from maximally stable extremal regions. In: In British Machine Vision Conference. Vol. 1. pp. 384–393. 2.2.1, 2.4.2

McDougall, A. A., 2004. Interfacing a Robot Head in MATLAB. Master's thesis, University of Glasgow, Department of Computing Science. 2.1.3, 3.2.1, 3.7, A.1, A.2

McHaffie, J. G., Kao, C. Q., Stein, B. E., 1989. Nociceptive neurons in rat superior colliculus: response properties, topography, and functional implications. Journal of Neurophysiology 62 (2), 510–525.
URL `http://jn.physiology.org/content/62/2/510.abstract` (document), 1.1, 1.1, 1.3

Meger, D., Forssen, P., Lai, K., Helmer, S., McCann, S., Southey, T., Baumann, M., Little, J. J., Lowe, D. G., Dow, D., Jun. 2008. Curious George: An attentive semantic robot. Robotics and Autonomous Systems 56 (6), 503–511. 1.2, 1.3, 2.1.2, 2.4.2, 2.4.3, 2.7.2, 4.1, 8.1.3

Meger, D., Gupta, A., Little, J. J., May 2010. Viewpoint detection models for sequential embodied object category recognition. In: 2010 IEEE International Conference on Robotics and Automation (ICRA). pp. 5055–5061. 2.1.2, 2.1.2, 2.7.2, 2.8, 3.9, 4.1, 7.4

Mikolajczyk, K., Leibe, B., Schiele, B., 2006. Multiple Object Class Detection with a Generative Model. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). Vol. 1. pp. 26–36. 2.1.1, 2.5.1, 2.7.2, 8.2.1

Mikolajczyk, K., Schmid, C., 2002. An Affine Invariant Interest Point Detector. In: Proceedings of the 7th European Conference on Computer Vision-Part I. ECCV '02. Springer-Verlag, London, UK, pp. 128–142. 2.2.1

Mikolajczyk, K., Schmid, C., Oct. 2005. Performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (10), 1615–30. 2.2.2, 2.3

Milanese, R., Oct. 1992. Detection of salient features for focus of attention. In: Proceedings 3rd SGAICO Meeting (Swiss Group for Artificial Intelligence and Cognitive Science), Biel-Bienne, Switzerland. pp. 87–101. 2.4.1

Mishra, A., Aloimonos, Y., Fah, C. L., 2009. Active segmentation with fixation. In: 2009 IEEE 12th International Conference on Computer Vision. pp. 468–475. 2.4.3, 6.6.3

Modayil, J., Kuipers, B., Nov. 2008. The Initial Development of Object Knowledge by a Learning Robot. Robotics and autonomous systems 56 (11), 879–890. 2.7.2, 6.2

Mowforth, P., Siebert, J. P., Jin, Z., Urquhart, C., 1990. A head called Richard. In: Proceedings of the British Machine Vision Conference. pp. 361–366. 2.1.2

Muja, M., Lowe, D. G., 2009. Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration. In: International Conference on Computer Vision Theory and Application (VISSAPP'09). INSTICC Press, pp. 331–340. 2.5.2.2

Murphy, R., Arkin, R., jul 1992. Sfx: An architecture for action-oriented sensor fusion. In: Proceedings of the 1992 lEEE/RSJ International Conference on Intelligent Robots and Systems. Vol. 2. pp. 1079 –1086. 1.4, 2.6, 2.6.3

Murphy, R., Mali, A., Apr. 1997. Lessons learned in integrating sensing into autonomous mobile robot architectures. Journal of Experimental & Theoretical Artificial Intelligence 9 (2), 191–209. (document), 1.1, 1.3, 1.1, 1.3, 2.6, 2.6.2, 2.6.3, 5.2, 5.3, 8.2.5

Murphy, R. R., 2001. Introduction to AI Robotics (Intelligent Robotics & Autonomous Agents). MIT Press. (document), 2.6, 2.6.1, 2.10

Natale, L., Metta, G., Sandini, G., 2002. Development of auditory-evoked reflexes: Visuo-acoustic cues integration in a binocular head. Robotics and Autonomous Systems 33, 87–106. 2.1.2

Neisser, U., 1976. Cognition and reality: principles and implications of cognitive psychology. Books in psychology. W. H. Freeman. 2.4

Nene, S. A., K., N. S., Murase, H., 1996. Columbia Object Image Library (COIL-100). Tech. rep., Columbia University, Computer Science. 4.1, 4.7

Newman, P., Sibley, G., Smith, M., Cummins, M., Harrison, A., Mei, C., Posner, I., Shade, R., Schroeter, D., Murphy, L., Churchill, W., Cole, D., Reid, I., Jul. 2009. Navigating, Recognizing and Describing Urban Spaces With Vision and Lasers. The International Journal of Robotics Research 28 (11-12), 1406–1433. 2.1.1, 2.2.1, 2.7.2, 6.5.2

Op de Beeck, H. P., Baker, C. I., Jan. 2010. The neural basis of visual object learning. Trends in Cognitive Sciences 14 (1), 22–30. 2.7.1, 2

Orabona, F., Metta, G., Sandini, G., 2005. Object-based Visual Attention: a Model for a Behaving Robot. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops. IEEE, Los Alamitos, CA, USA, pp. 89–89. 2.1.2, 2.4.3

Ouerhani, N., von Wartburg, R., Hugli, H., Rene, M., 2004. Empirical Validation of the Saliency-based Model of Visual Attention. Letters on Computer Vision and Image Analysis 3 (1), 13–24. 2.4.2, 5.9

Palmeri, T. J., Gauthier, I., Apr. 2004. Visual object understanding. Nat Rev Neurosci 5 (4), 291–303. 2.7.1, 8.2.4

Panerai, F., Metta, G., Sandini, G., Jan. 2000. Visuo-inertial stabilization in space-variant binocular systems. Robotics and Autonomous Systems 30 (1-2), 195–214. 2.4.3

Peters, G., Zitova, B., von der Malsburg, C., Apr. 2002. How to measure the pose robustness of object views. Image and Vision Computing 20 (4), 249–256. 2.7.1, 6.3

Pluim, J. P. W., Maintz, J. B. A., Viergever, M. A., Jul. 2003. Mutual-information-based registration of medical images: a survey. IEEE Transactions on Medical Imaging 22 (8), 986–1004. 6.7.3

Posner, M. I., Cohen, Y., 1984. Components of Visual Orienting. Vol. 10 of Attention and Performance. Lawrence Erlbaum Associates Ltd, Hove, U.K., Ch. 10, pp. 531–556. 2.4, 5.2

Posner, M. I., Petersen, S. E., Jan. 1990. The attention system of the human brain. Annual review of neuroscience 13, 25–42. 2.4, 5.2, 6.3

Rasolzadeh, B., Bjorkman, M., Huebner, K., Kragic, D., Aug. 2010. An Active Vision System for Detecting, Fixating and Manipulating Objects in the Real World. The International Journal of Robotics Research 29 (2-3), 133–154. 1.2, 1.3, 2.1.2, 2.3, 2.4.3, 2.8, 3.9, 4.1, 6.5.2, 7.4, 8.1.3

Roberts, L. G., 1963. Machine Perception of Three-Dimensional Solids. Outstanding Dissertations in the Computer Sciences. Garland Publishing, New York. 2.2.1

Rosen, C. A., J, N. N., Adams, M. B., Jan. 1965. A Research and Development Program in Applications of Intelligent Automata to Reconnaissance-Phase I, proposal ESU 65-1 Costing (ESU 65-117), January 1965.
URL `http://www.ai.sri.com/pubs/files/rosen65-esu65-1tech.pdf`
1.3

Rosenberger, C., Chehdi, K., 2000. Unsupervised clustering method with optimal estimation of the number of clusters: application to image segmentation. In: Proceedings 15th International Conference on Pattern Recognition. ICPR-2000. Vol. 1. IEEE Comput. Soc, pp. 656–659. 2.5.2.3, 8.2.3

Sanguinetti, G., Mar. 2008. Dimensionality reduction of clustered data sets. IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (3), 535–40. 2.5.2.3, 8.2.3

Schiele, B., Crowley, J. L., 2000. Recognition without Correspondence using Multidimensional Receptive Field Histograms. International Journal of Computer Vision 36 (1), 31–50. 2.2.1

Schwartz, E., 1995. Space-variant active vision: Definition, overview and examples. Neural Networks 8 (7-8), 1297–1308. 2.4.3

Shen, K., Valero, J., Day, G. S., Pare, M., 2011. Investigating the role of the superior colliculus in active vision with the visual search paradigm. European Journal of Neuroscience 33 (11), 2003–2016. 1.4

Siebert, J. P., Marshall, S. J., 2000. Human body 3D imaging by speckle texture projection photogrammetry. Sensor Review 20 (3), 218–226. 2.1.3

Styles, E., 2005. Attention, Perception and Memory: An Integrated Introduction (Psychology Focus). Psychology Press. 2.3, 2.4, 2.6, 2.7.1, 4.8.1, 4.9, 5.1, 5.2, 5.10, 6.3, 6.5.2, 6.6.3, 6.6.3, 6.7.5, 8.1.3

Svedman, M., Goncalves, L., Karlsson, N., Munich, M., Pirjanian, P., 2005. Structure from Stereo Vision using Unsynchronized Cameras for Simultaneous Localization and Mapping. In: 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 2993–2998. 6.5.2

Theeuwes, J., 1993. Visual selective attention: A theoretical analysis. Acta Psychologica 83 (2), 93 – 154. 2.4

Treisman, A. M., Gelade, G., 1980. A feature-integration theory of attention. Cognitive Psychology 12 (1), 97–136. 2.4.1

Tsotsos, J. K., Aug. 2008. What roles can attention play in recognition? In: 2008 7th IEEE International Conference on Development and Learning. IEEE, pp. 55–60. 1.1, 1.1.1

Tsotsos, J. K., Rodriguez-Sanchez, A. J., Rothenstein, A. L., Simine, E., Aug. 2008. The different stages of visual recognition need different attentional binding strategies. Brain Research 1225, 119–32. 6.6.3, 8.2.1

Tuytelaars, T., Mikolajczyk, K., 2007. Local Invariant Feature Detectors: A Survey. Foundations and Trends in Computer Graphics and Vision 3 (3), 177–280. 2.2, 2.2.1, 2.7.2

Ullman, S., Feb. 2007. Object recognition and segmentation by a fragment-based hierarchy. Trends in Cognitive Sciences 11 (2), 58–64. 2.6.2

Ullman, S., Vidal-Naquet, M., Sali, E., Jul. 2002. Visual features of intermediate complexity and their use in classification. Nature Neuroscience 5 (7), 682–687. 2.7.1, 6.3, 6.3.1, 8.1.4

Urmson, C., Anhalt, J., Clark, M., Galatali, T., Gonzalez, J., Gowdy, J., Gutierrez, A., Harbaugh, S., Johnson-Roberson, M., Kato, H., Koon, P., Peterson, K., Smith, B., Spiker, S., Tryzelaar, E., W.L., W., June 2007. High Speed Navigation of Unrehearsed Terrain: Red Team Technology for Grand Challenge. Tech. Rep. TR-04-37, Carnegie Mellon University. 2.6.2

Urquhart, C., 1997. The Active Stereo Probe: The design and implementation of an active videometrics system. Phd thesis, University of Glasgow, Department of Computing Science. 2.1.1, 2.1.3

van Gemert, J. C., Veenman, C. J., Smeulders, A. W. M., Geusebroek, J.-M., 2010. Visual Word Ambiguity. IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (7), 1271–1283. 8.2.3

Veres, S. M., 2008. Natural Language Programming of Agents and Robotic Devices. No. ISBN 978-0-9558417-0-5. SysBrain Ltd. 5.10, 8.2.5

Wallraven, C., Bülthoff, H., 2001. Automatic acquisition of exemplar-based representations for recognition from image sequences. In: CVPR 2001 - Workshop on Models versus Exemplars. IEEE. 2.7.2

Wallraven, C., Bülthoff, H., 2007a. Object recognition in Man and Machine. In: Object Recognition, Attention and Action. Springer, Ch. Part I, pp. 89–104. 1.2, 2.4, 8.1.3

Wallraven, C., Bülthoff, H. H., 2007b. Object Recognition in Humans and Machines. Springer, pp. 89–104. 2.4.2, 2.7.2, 5.2, 6.3

Wallraven, C., Caputo, B., Graf, A. B. A., 2003. Recognition with local features: the kernel recipe. In: Press, I. (Ed.), Proceedings of the Ninth IEEE International Conference on Computer Vision. Vol. 2. pp. 257–264. 2.7.2

Walther, D., Rutishauser, U., Koch, C., Perona, P., Oct. 2005. Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. Computer Vision and Image Understanding 100 (1-2), 41–63. 2.1.1, 2.4.2, 2.4.3

Welke, K., Issac, J., Schiebener, D., Asfour, T., Dillmann, R., May 2010. Autonomous acquisition of visual multi-view object representations for object recognition on a humanoid robot. In: Robotics and Automation (ICRA), 2010 IEEE International Conference on. pp. 2012–2019. 2.1.2, 2.6.3

Weng, J., Huang, J. S., Ahuja, N., 1993. Learning Recognition and Segmentation of 3D objects from 2D images. In: Procedings IEEE International Conference in Computer Vision. pp. 121–128. 6.2

Westelius, C. J., 1995. Focus of Attention and Gaze Control for Robot Vision. Phd thesis, Linkoping University, Department of Electrical Engineering. 2.1.1, 2.4.2

Wolfe, J. M., 2007. Guided Search 4.0: Current Progress with a model of visual search. In: Gray, W. (Ed.), Integrated Models of Cognitive Systems. Oxford, New York, pp. 99–119. 2.4.1

Wurtz, R. H., Joiner, W. M., Berman, R. A., 2011. Neuronal mechanisms for visual stability: progress and problems. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 366 (1564), 492–503. 1.3

Yanez-Suarez, O., Azimi-Sadjadi, M., 1999. Unsupervised clustering in Hough space for identification of partially occluded objects. IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (9), 946–950. 2.3.1, 2.5.1

Yarbus, A. L., 1967. Eye Movements and Vision. Plenum Press. 2.4, 5.10

Yu, Y., Mann, G. K. I., Gosine, R. G., 2010. An Object-Based Visual Attention Model for Robotic Applications. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 40 (5), 1398–1412. 6.3

Zhou, C., Wei, Y., Tan, T., 2003. Mobile robot self-localization based on global visual appearance features. In: Robotics and Automation, 2003. Proceedings. ICRA '03. IEEE International Conference on. Vol. 1. pp. 1271–1276 vol.1. 2.2

Zickler, S., Efros, A. A., 2007. Detection of Multiple Deformable Objects using PCA-SIFT. In: Proceedings of the 22nd national conference on Artificial intelligence - Volume 2. AAAI Press, Vancouver, British Columbia, Canada, pp. 1127–1132. 2.5.1, 4.4, 4.7.2, 4.8.1, 4.9

Zickler, S., Veloso, M., Dec. 2006. Detection and Localization of Multiple Objects. In: 2006 6th IEEE-RAS International Conference on Humanoid Robots. IEEE, Genova, pp. 20–25. 2.5.1