# Visual Attention for Robotic Cognition: A Biologically Inspired Probabilistic Architecture

by

Momotaz Begum

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2010

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

**Abstract**

The human being, the most magnificent autonomous entity in the universe, frequently takes the decision of 'what to look at' in their day-to-day life without even realizing the complexities of the underlying process. When it comes to the design of such an attention system for autonomous robots, all of a sudden this apparently simple task appears to be an extremely complex one with highly dynamic interaction among motor skills, knowledge and experience developed throughout the life-time, highly connected circuitry of the visual cortex, and super-fast timing. The most fascinating thing about visual attention system of the primates is that the underlying mechanism is not precisely known yet. Different influential theories and hypothesis regarding this mechanism, however, are being proposed in psychology and neuroscience. These theories and hypothesis have encouraged the research on synthetic modeling of visual attention in computer vision, computational neuroscience and, very recently, in AI robotics. The major motivation behind the computational modeling of visual attention is two-fold: understanding the mechanism underlying the cognition of the primates' and using the principle of focused attention in different real-world applications, e.g. in computer vision, surveillance, and robotics. Accordingly, we observe the rise of two different trends in the computational modeling of visual attention. The first one is mostly focused on developing mathematical models which mimic, as much as possible, the details of the primates' attention system: the structure, the connectivity among visual neurons and different regions of the visual cortex, the flow of information etc. Such models provide a way to test the theories of the primates' visual attention with minimal involvement from the live subjects. This is a magnificent way to use technological advancement for the understanding of human cognition. The second trend in computational modeling, on the other hand, uses the methodological sophistication of the biological processes (like visual attention) to advance the technology. These models are mostly concerned with developing a technical system of visual attention which can be used in real-world applications where the principle of focused attention might play a significant role for redundant information management. This thesis is focused on developing a computational model of visual attention for robotic cognition and, therefore, belongs to the second trend. The design of a visual attention model for robotic systems as a component of their cognition comes with a number of challenges which, generally, do not appear in the traditional computer vision applications of visual attention. The robotic models of visual attention, although heavily inspired by the rich literature of visual attention in computer vision, adopt different measures to cope with these challenges. This thesis proposes a Bayesian model of visual attention designed specifically for robotic systems and, therefore, tackles the challenges involved with robotic visual attention. The operation of the proposed model is guided by the theory of biased competition, a popular theory from cognitive neuroscience describing the mechanism of primates' visual attention. The proposed Bayesian attention model offers a robot-centric approach of visual attention where the head-pose of a robot in the 3D world

is estimated recursively such that the robot can focus on the most behaviorally relevant stimuli in its environment. The behavioral relevance of an object determined based on two criteria which are inspired by the postulates of the biased competitive hypothesis of visual attention in the primates. Accordingly, the proposed model encourages a robot to focus on novel stimuli or stimuli that have similarity with a 'sought for' object depending on the context. In order to address a number of robot-specific issues of visual attention, the proposed model is further extended to the multi-modal case where speech commands from the human are used to modulate the visual attention behavior of the robot. The Bayes model of visual attention, inherited from the Bayesian sensor fusion characteristic, naturally accommodates multi-modal information during attention selection. This enables the proposed model to be the core component of an attention oriented speech-based human-robot interaction framework. Extensive experiments are performed in the real-world to investigate different aspects of the proposed Bayesian visual attention model.

**Acknowledgements**

I would like to express my deep gratitude to my PhD supervisor Dr. Fakhri Karray for his strong support throughout my PhD program. His visionary thinking about robotic research had a great influence on me. I thankfully acknowledge the encouragements from Dr. George Mann and Dr. Raymond Gosine of Memorial University for my graduate research. They always have been a great source of inspiration to me. My special thanks to Dr. Jiping Sun of Vestec Inc. He always came up with a solution to every problem I had while talking to my robots in English language. I also like to thank the members of Pattern Analysis and Machine Intelligence Lab, specially Ahmad, Bahador, Ahmed, and Jamil, for their help, encouragement, and all the insightful discussions about the tricky problem of robotics and life. Finally, I would like to acknowledge the enormous support from my husband. I could not have done this without him.

# Contents

# List of Tables

# List of Figures

# List of Abbreviations

| | | |
|---|---|---|
| BC | : | Biased Competition |
| DM | : | Dorsal-stream Module |
| DOF | : | Degree Of Freedom |
| EM | : | Early visual Module |
| fMRI | : | functional Magnetic Resonance Imaging |
| GPU | : | Graphic Processing Unit |
| GART | : | Gaussian Adaptive Resonance Theory |
| HRI | : | Human Robot Interaction |
| HSV | : | Hue-Saturation-Value |
| IOR | : | Inhibition Of Return |
| IT | : | Inferior Temporal |
| KP | : | KeyPoints |
| LTM | : | Long Term Memory |
| LGN | : | Lateral Geniculate Nucleus |
| MST | : | Medial Superior Temporal |
| NVT | : | Neuromorphic Vision Toolkit |
| PTU | : | Pan-Tilt Unit |
| PPC | : | Posterior Parietal Cortex |
| PET | : | Positron Emission Tomography |
| PDA | : | Personal Digital Assistant |
| RF | : | Receptive Field |
| RGB | : | Red-Green-Blue |
| SIFT | : | Scale Invariant Feature Transform |
| SLAM | : | Simultaneous Localization And Mapping |
| VF | : | Visual Field |
| VOCUS | : | Visual Object detection with a CompUtational attention System |
| VM | : | Ventral-stream Module |
| WTA | : | Winner Take All |
| WM | : | Working Memory |

# List of Symbols

| | | |
|---|---|---|
| $\alpha$ | : | the pan angle |
| $\beta$ | : | the tilt angle |
| $\eta$ | : | normalizing constant |
| $\rho$ | : | the resolution of the head-pose space in terms of the number of image pixels |
| $\lambda$ | : | A parameter for dynamic IOR |
| $\chi$ | : | scale constraint for SIFT keypoints matching |
| $\psi$ | : | orientation constraint for SIFT keypoints matching |
| $\epsilon$ | : | the sub-image block size |
| $\bar{\mu}$ | : | the mean vector of a color cluster in the GART |
| $\sigma$ | : | the variance vector of a color cluster in the GART |
| $\gamma$ | : | default value of variance for a color cluster in the GART |
| $\zeta_1, \zeta_2$ | : | thresholds for image segmentation |
| $A_c$ | : | object location expressed in camera coordinate |
| $A_w$ | : | object location expressed in world coordinate |
| $\mathbf{b}^{WM}$ | : | the non-zero top-down bias from the working memory |
| $\mathbf{b}^{LTM}$ | : | the non-zero top-down bias from the long-term memory |
| $\mathbf{b}$ | : | the non-zero top-down bias |
| $\mathcal{C}_w(x_w, y_w, z_w)$ | : | the world coordinate system |
| $\mathcal{C}_b(x_b, y_b, z_b)$ | : | the base coordinate system |
| $\mathcal{C}_h(x_h, y_h, z_h)$ | : | the head coordinate system |
| $\mathcal{C}_c(x_c, y_c, z_c)$ | : | the camera coordinate system |
| $\mathcal{C}_i(x_i, y_i, z_i)$ | : | the image coordinate system |
| $\mathbf{F}$ | : | the sensor measurements |
| $\mathbf{F}_k^v$ | : | measurements from the camera (after processing) at time $k$ |
| $\mathbf{F}_k^a$ | : | measurements from the microphone (after processing) at time $k$ |
| $f_{match}$ | : | a set of keypoints that matches with an object |
| $I_k$ | : | the image frame capture at $k$ |
| $I_k^Y$ | : | the intensity image at time $k$ |

| | | |
|---|---|---|
| $I_k^r$ | : | the red plane at time $k$ |
| $I_k^g$ | : | the green plane at time $k$ |
| $I_k^b$ | : | the blue plane at time $k$ |
| $I_k^{r'}$ | : | the red plane at time $k$ after pyramid-based image segmentation |
| $I_k^{g'}$ | : | the green plane at time $k$ after pyramid-based image segmentation |
| $I_k^{b'}$ | : | the blue plane at time $k$ after pyramid-based image segmentation |
| $k$ | : | the discrete time index |
| $0:k$ | : | $\{0, 1, 2, \ldots, k\}$ |
| $L$ | : | the total number of particles |
| $l$ | : | particle index, $l = \{1, 2, \ldots, L\}$ |
| $\mathbf{M}_k$ | : | the long-term memory at $k$ |
| $\mathbf{m}_k$ | : | the working memory at $k$ |
| $N$ | : | total number of SIFT keypoints in the LTM |
| $n$ | : | number of time a color has been observed |
| $O$ | : | a hypothesis that an object has been observed before |
| $p_a$ | : | the probability that a WM has been created from the LTM |
| $p_v$ | : | the probability that a set of visual features is the |
| $p(.)$ | : | the probability |
| $P$ | : | total number of SIFT keypoints extracted from one image |
| $P_{xy}$ | : | position constraint for SIFT keypoints matching |
| $Q$ | : | the total number of matching SIFT keypoints |
| $q$ | : | index of the SIFT keypoints |
| ${}^hT_c$ | : | homogeneous transformation matrix between the camera and the PTU |
| ${}^bT_h$ | : | homogeneous transformation matrix between the PTU and the 'home' position of the PTU |
| ${}^wT_b$ | : | homogeneous transformation matrix between the 'home' position of the PTU and the world coordinate system |
| $(u, v)$ | : | the location of a pixel in the image |
| $\mathbf{U}_k^{(l)}$ | : | the set of image points focused at $\mathbf{x}_k^{(l)}$ |
| $w_k^{(l)}$ | : | the weight of the $l-th$ particle at time $k$ |
| $\mathbf{x}$ | : | the head-pose of the robot (system state) |
| $\mathbf{x}_k^{(l)}$ | : | the $l-th$ particle at time $k$ |
| $z$ | : | the measurements |

# Chapter 1

# Introduction

The research in cognitive robotics operates with the very focused goal of designing robots with human like abilities (albeit of reduced complexity) in perception, reasoning, decision making, and action execution. Integration of all these abilities entitle human as a cognitive entity. Reaching this goal, however, involves dispersed connectivity with several other research areas in different disciplines namely, AI robotics, cognitive psychology, developmental neuroscience, and linguistics. Based on the research conducted in these disciplines on different aspects of human cognitive development, it is now a well accepted fact that cognition is something that can not be fully hand-coded in the artificial agents (e.g., the robot), rather it emerges through a bi-directional interaction between the robot and its surroundings [1–4]. Modern robots, therefore, are equipped with a redundant number of sensors and actuators to perceive and perturb the surrounding as a way of developing their cognition. The increased number of sensors and actuators introduces the challenge of managing enormous amount of information steadily arriving through them. The first major challenge of developing cognition, therefore, lies at the perceptual level: information management. The primates master this information management skill through their custom-built attention mechanism. The underlying idea is simple yet robust: focus on the piece of information (in relative exclusion of the others) which is the most relevant to the current context. In case of humans the question of 'what is relevant?' is itself a 'discipline', but for robotic systems we optimize the definition of 'relevancy' in the context of some predefined tasks (e.g., entertainment, assistance, rescue operation). Mimicking the attention behavior of the primates in the design of robot's attention behavior has gained tremendous popularity in the recent years [5, 6]. The problem with redundant information management is the most severe in case of visual perception of the robots. Even a moderate size image of the natural scenes generally contains enough visual information to easily overload the real-time decision making process of an autonomous robot. A well accepted solution to tackle this problem is designing a human-like visual attention mechanism for

the robots where the robot will selectively (and autonomously) choose a 'behaviorally relevant' segment of visual information for further processing. The research on developing such a computational model of visual attention has experienced significant success during the last decade but we are still far away from having an artificial model of human-like visual attention which can serve as a component of robotic cognition. **The goal of this thesis is to contribute to the endeavor of cognitive robotics through developing a model of visual attention which will serve as a component of cognition of the autonomous robots**.

This thesis performs a comprehensive analysis of the existing computational models of visual attention to shed light on their strengths and shortcomings with respect to cognitive robotics and thereby defines a set of properties that are expected to be observed in a computational model of visual attention designed for cognitive robots. The thesis then proposes a probabilistic model of visual attention which accommodates all of these expected properties. Extensive experiments with a physical robot and sensors in the real-world are presented to validate the proposed visual attention model.

## 1.1 Background and Motivation

The endeavor of cognitive robotics to design a bio-inspired visual attention model for robots has strong connectivity with the research in cognitive psychology, computer vision, and computational neuroscience as these are the three disciplines which cultivated the basic research on the artificial modeling of human visual attention. The visual attention models developed for robotic cognition heavily rely on the computational models of visual attention proposed in computer vision and computational neuroscience while the inspiration of all these models is rooted in the theories of human visual attention proposed in cognitive psychology and neuroscience. The theories of primates' visual attention mechanism first took the form of a computational model by the research work reported in [7]. The major inspiration behind the development of computational models of attention was two fold.

1. Development of a computational tool to test the validity of the theories/hypothesis of attention proposed in psychology and neuroscience
2. The potential applications of the principle of focused attention in computer vision, video surveillance, and robotics.

Accordingly, we observe the rise of two distinct trends in the computational modeling of visual attention. The first one is mostly concerned about simulating the response of the visual cortex during attention related activities [8–14]. Majority of the models here are proposed in computational neuroscience. The second kind of computational models are

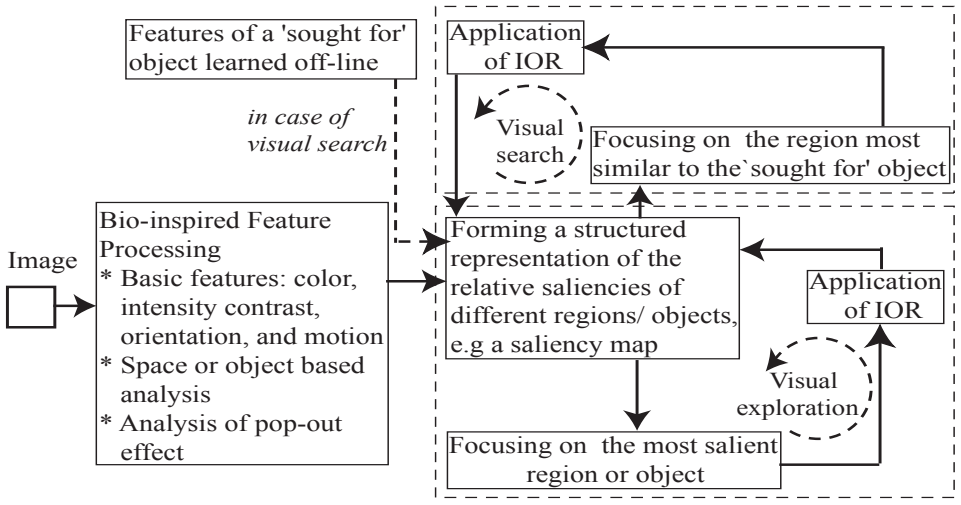
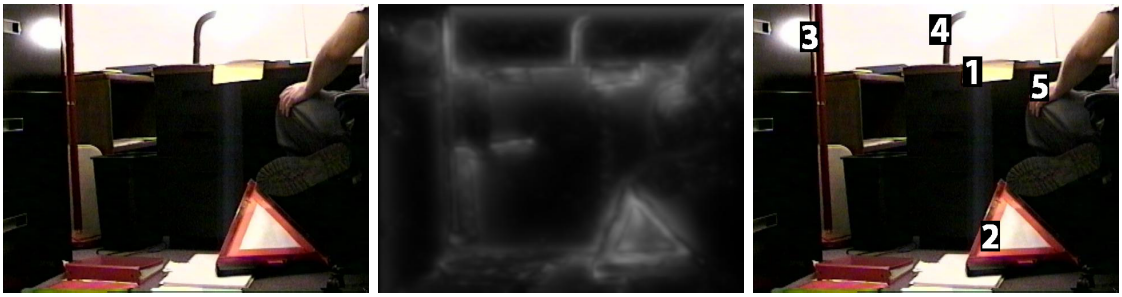



Figure 1.1: General architecture of the computer vision models of visual attention

concerned about developing a technical system of visual attention while utilizing the unique properties of the biological attention system [7, 11, 15–23]. The researchers in computer vision are the major developers of the technical models of visual attention and cognitive robotics, probably, is their most recent user. **This thesis is solely focused on the second group of computational models and is committed to propose a technical model of visual attention for the cognitive robots.**

The computational models of visual attention proposed in computer vision literature gained widespread popularity in many sectors of robotics research. This is mostly because of the fact that their strategies of analyzing visual features make them suitable to be applied on a real-time technical system like autonomous robots. But most of the existing computer vision models are characterized by a number of properties which impose some restrictions on their direct use in the robotic applications. They will be discussed later in this chapter.

The architecture generally followed by most of the computer vision models of visual attention is shown in Fig. 1.1. The key differences among different models occur in two sectors: 1) the methodology of implementing the overall architecture, e.g connectionist approach [20, 21], filter-based approach [7, 16–19, 22–24] and 2) the mechanism of constructing a saliency map. Despite these two sectors of mismatch, the computer vision models of visual attention share a set of common characteristics. They are summarized below.




(a) (b) (c)

Figure 1.2: The role of saliency operator in evaluating visual attention (a) A natural image from the standard image database [17] (b) The saliency map calculated using the visual attention model proposed in [23] (c) Five focus points in the image marked in the order of decreasing saliency

**Saliency operator**

A centralized 'saliency map', first introduced in the attention model reported in [7], plays a key role to guide attention toward different regions of an input image in almost all of the existing computer vision models of visual attention [18,19,21–24]. A saliency map, in simple words, is a two-dimensional image (of same size as the input image) in which the intensity value of a pixel represents the relative visual saliency of its corresponding pixel in the original input image. The higher the value is, the more salient the pixel is. A saliency map does not hold any information about the feature channel (e.g., color, intensity contrast, or orientation) that causes a particular pixel to appear more salient than the others. Besides, the relative feature strength of a pixel plays the main role to obtain higher value in the saliency map while the local absolute feature strength is of very little importance [25]. A saliency map-based attention model generally reports the most salient pixel in the saliency map as the current focus of attention. In order to prevent the attention from re-visiting the same location, the saliency of the current focus of attention is suppressed after being attended and thereby achieving the property of *inhibition of return* (IOR) [26]. Figure 1.2 shows a typical saliency map corresponding to a natural image (obtained from the freely available standard image database [17]) along with the five focuses of attention marked on the image in the order of decreasing saliency.

The most interesting fact about the saliency map is, it is a controversial notion in neurobiology. The hypothesis of attentional control based on a unique, centralized saliency map is, thus far, not supported by any of the existing neuro-physiological findings [7, 27–29]. Rather, majority of the functional magnetic resonance imaging (fMRI)- and positron emission tomography (PET)-based studies of attention in the primates advocates the idea that there are several areas in the primates' visual cortex (e.g., frontal eye field (FEF),

superior colliculus (SC), posterior parietal cortex (PPC)) which process different visual features in a distributed manner [30–32]. Even if the integrated operation of these brain areas helps to constitute some form of 'saliency map', the organization of that map is completely different than the one conventionally used in the existing computational models of visual attention [33]. As compared to the notion of 'saliency map', a more neurobiologically plausible way of explaining attentional shift is the concept of bias modulation as described in [34]. An early psychophysical theory of visual attention [35] also advocates on the role of bias modulation for attention selection. The concept of bias modulation, however, has gained very limited popularity among the researchers in computer vision for developing technical model of visual attention [11].

### Covert shift of attention

Majority of the computational models of visual attention in the existing literature are designed based on the assumption that neither the eye nor the head moves to perform attention. The attention mechanism in the primates which follows this assumption is called covert attention [36]. The absence of eye/head movement during the direction of attention has a number of consequences.

- The retinal input remains unchanged throughout the attentional task.
- The frame of reference remains unchanged in the subsequent directions of attention. That simplifies the implementation of the IOR.
- The scene saliency remains unchanged causing no further requirement to recalculate it after each attentional shift.

Most of the computational models of visual attention (e.g., [7,18–24]) enjoy the simplicity of computation arising from the above three consequences of the covert nature of attentional shift. The covert shift of attention, however, makes it difficult to compare the performance of the computational models with the ground truth, e.g., with the attention behavior of the human.

### Bottom-up and top-down analysis

The early computational models of visual attention (e.g., [7, 17, 18, 20, 24]) mostly dealt with bottom-up (or stimulus driven) influence in attention selection. To be consistent with the biological findings, the recent computational models started to invoke the effect of top-down influence [19,21,23,37,38]. These latter computational models [19,21,23,37,38], however, limit the influence of top-down information in attention selection only to the case

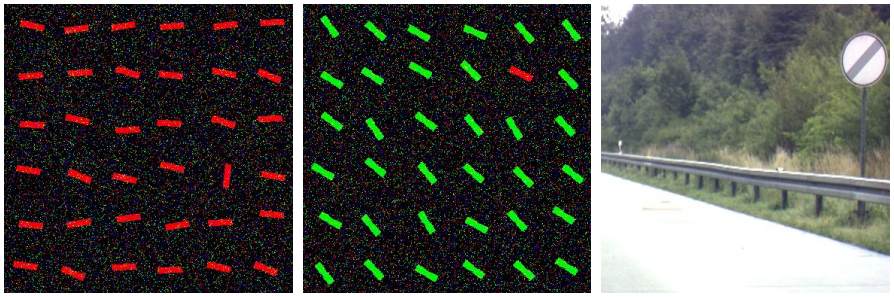


Figure 1.3: Standard images (available in [17]) commonly used to test the computational models of visual attention. One of the stimuli stands out of the scene in at least one feature dimension (e.g., orientation, color, intensity contrast). Some of the recent computational models can almost accurately identify such stimuli

of visual search. Thus the bottom-up cues guide the visual exploration (focusing on the most salient stimuli) while the top-down cues guide the visual search. In almost all of the existing models these two modes of attention (visual search and visual exploration) run in mutual exclusion of each other as shown in Fig. 1.1. In some of the models, even the process of generating the saliency map for visual exploration considerably differs from that for visual search. The desired mode of attention (visual exploration or visual search) is manually activated by the programmer depending on the task at hand.

### Off-line training for visual search

Almost all computer vision models of visual attention require an off-line training phase prior to performing visual search. The model learns the target-specific visual features during a training phase and the learned information is used to increase the saliency of the target-like features during the visual search. The success of visual search, therefore, strongly relies on the efficiency of the off-line training stage: type and quality of the training images, number of training images etc [23].

### Space- and object- based analysis

Inspired by the early psychophysical theories of attention [39, 40], the majority of the computer vision models hypothesizes 'space' as the elemental unit of attention selection [7, 17–21, 23, 24]. Accordingly, saliency and task-relevance are investigated at the pixel level without considering the concept of object. Increasing evidence, in the psychology and cognitive neuroscience, of 'object' being one of the elemental units of attention selection [41–46] has influenced the recent computer vision models of visual attention. Many of the

Figure 1.4: Natural images where the focus of attention of a human is never known precisely [18]

recent models perform object-based analysis for selective attention [22, 47, 48]. There are, however, only a few efforts which integrate space- and object-based analysis in the same framework [8, 10, 11, 49].

Although common in almost every computational model, the way the above mentioned characteristics are achieved differs in different attention models and hence the variation in performance. These characteristics along with their sophisticated implementation have enabled todays computational models to reach the stature where some of them can accurately mimic the 'what to focus on?' behavior of the human when presented with typical test images (e.g., the image shown in Fig. 1.3). The performance, however, is not that satisfactory in case of other natural images where the ground truth is never precisely known (e.g., the image shown in Fig 1.4).

An important fact about the computer vision models of visual attention is that their characteristics make them well suited for the applications where static images or images from a video stream are manually fed to the model in order to identify the most salient/task-relevant stimuli. In case of some real-time applications where the current visual input of the attention model has to be determined by the decision output of the model (i.e the focus of attention) at the immediate past, the traditional computer vision models of visual attention face severe limitations in a number of aspects. Using the visual attention models as a component of robotic cognition is an example of such applications. In this case the attention model should be able to locate the behaviorally-relevant stimuli in an ongoing stream of visual input and respond to it, perform learning in an on-line fashion and with minimal human supervision, and apply the learned knowledge for guiding the attention behavior in arbitrary environmental settings.

It is worth mentioning that a realization of the human-like attention system (a reduced complexity version) requires a complex interaction among attention, knowledge, emotion, and reasoning. Design of such an attention model might not be possible in the near future with our present understanding of human cognition and current technological sophistica-

tion, but there are some milestones in this journey that we can certainly achieve through adopting a slightly different design perspective.

## 1.2 Problem Statement

This thesis has identified a set of research issues that must be addressed in order to design an attention model which will serve as a component of robotic cognition. These issues has been identified based on an intensive investigation of the requirements of cognitive robots, the properties of the computer vision models of visual attention and the modifications required to fit them in cognitive robotic applications. A detailed analysis of these research issues is provided in this section.

**Issue 1. Overt shift of attention**

In robotic applications (e.g., social robots, assistive robots, entertainment robots) it is generally desired that visual attention will be accompanied by a saccadic movement of the camera head of the robot. Such movement is necessary to place the object of attention at the center of the camera frame and facilitates the learning of the focused object. A computational model of visual attention for the robots, therefore, requires an integration of the covert and overt modes in a common framework, much the same way the primates integrate covert and overt shift of attention. The overt shift of attention leads to the following issues that must be solved to design a model of attention for the robots.

**Issue 1.1 Change of reference frame:** In the simplest case, the visual attention hardware of a robot consists of a camera (generally, color camera) and a two degrees of freedom (DOF) pan-tilt unit (PTU) on to which the camera is mounted. There are at least five coordinate systems involved with such an arrangement for execution of overt attention.

- The world coordinate system, $\mathcal{C}_w : (x_w, y_w, z_w)$.
- The base coordinate system (the coordinate system attached with the home position of the PTU), $\mathcal{C}_b : (x_b, y_b, z_b)$
- The head coordinate system (the coordinate system attached with the current position of the PTU), $\mathcal{C}_h : (x_h, y_h, z_h)$.
- The camera coordinate system, $\mathcal{C}_c : (x_c, y_c, z_c)$.
- The image coordinate system, $\mathcal{C}_i : (x_i(u), y_i(v))$.

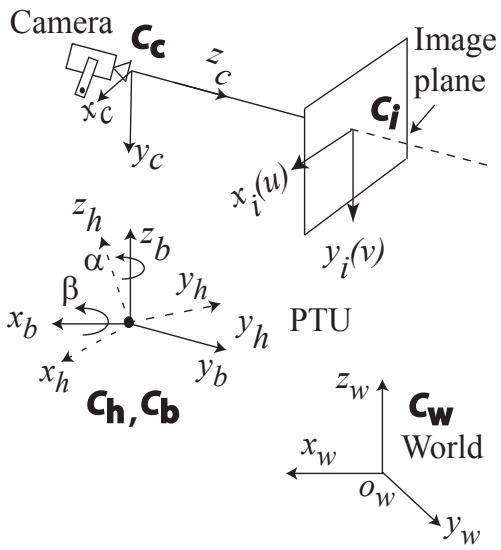
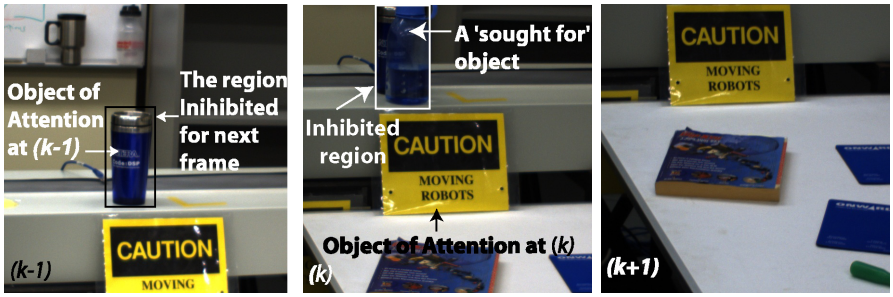



Figure 1.5: The coordinate systems involved with robotic overt visual attention for a simple camera-PTU arrangement (please see text for detail)

The coordinate systems are shown in Fig. 1.5 in case of a stationary robot. The world coordinate is fixed. For a given position of the robot in the 3D world, the base coordinate is also stationary. But the head, camera and the image coordinate systems are changing according to the movement of the PTU. A $(\alpha, \beta)$ amount of pan-tilt movements of the PTU cause the camera to perceive a different segment of the environment. Thus the content of the robot's visual field (VF) changes, although a considerable amount of overlap generally exists between two successive snap-shots of the environment. This makes the 'saliency map' calculated prior to the camera movement partially obsolete and demands either a fresh calculation of saliency or re-mapping of the previous saliency to the new image coordinate.

**Issue 1.2 Dynamic IOR:** The role of IOR in robotic attention is the same as that in the biological attention system: encouraging the shift of attention toward fresh stimuli/ location [26]. Failure to implement the IOR properly might cause a robot to oscillate between two stimuli. In overt attention camera movement causes the location of a stimulus to shift in the image coordinate. It is, therefore, required to design a dynamic IOR strategy where the location of the recently attended object will be mapped to the new image coordinate in order to inhibit its candidacy as the next focus of attention. The space-based dynamic IOR introduces the complexity that if, between two successive frame capture, a new object appears at the inhibited location, the robot completely ignores its presence. Fig. 1.6 demonstrates one instance of this problem. This might incur a longer time to identify a 'sought for' object during visual search which is undesirable in many




(a) (b) (c)

Figure 1.6: Difficulty in visual search with space-based dynamic IOR (a) The region of the attended object at time $(k-1)$ is made inhibited for time $k$, (b) The inhibited region is mapped to the new image coordinate system at time $k$. A 'sought for' object appears within the inhibited region and the robot ignores its presence, (c) A random head movement in search of the 'sought for' object causes it to go out of the VF. As a result, the robot requires longer time to find the 'sough for' object

robotic applications. This kind of problem can be avoided by implementing object-based IOR. The object-based IOR, however, introduces the problem of object correspondence. In order to inhibit a recently attended object from being attended again, the robot needs to identify it in the shifted image coordinate. This is generally a challenging task due to change in camera perspective, lighting, image blurring due to camera motion, and partial appearance of the objects.

**Issue 1.3 Partial appearance:** Due to head movement, it is highly likely that many objects will partially/ completely go out of the camera frame in course of time. The probability of this increases when the robot uses a narrow angle optics for the camera or when the objects are located either very close to the camera or near the periphery of the frame. Due to partial appearance the robot might fail to identify a recently attended object. This, in turn, results in a failure to apply the IOR on it and the robot might re-attend the same object. In the worst case, the attention of the robot will start oscillating among a set of objects. The same trapped situation might also occur if the robot always finds a set of objects 'attention worthy' (e.g., because of their novelty) as it can not match the partially perceived features of these objects with its memory database of previously attended and learned features. Application of space-based dynamic IOR on all of the previously attended locations might relax this problem at the expense of exacerbating the problem stated under *Issue 1.2*.

**Issue 2. Integrated space- and object- based analysis**

The space-based analysis, commonly used in the majority of the computational models of attention, does not practically fulfill the interest of most robotic applications. Instead of a single pixel reported as the focus of attention, the information about the object underlying that salient pixel is of greater interest in robotic applications. The space-based models of attention, which generally rely on a traditional saliency map, do not preserve the information of the underlying objects. Another serious problem with the space-based approaches is the common practice of using coarse scales of the input image for space-based saliency map construction [17–19, 23]. This causes the fading of many attention-worthy small regions which do not get a chance to be highlighted in the saliency map [50]. The problem of space-based dynamic IOR as stated under *Issue 1.2* is another consequence of space-based analysis. The object-based analysis works well in certain situations but it has its own shortcomings, e.g., the object correspondence problem as stated under *Issue 1.2*. Besides, extraction of meaningful objects from the scene is computationally more expensive and makes the object-based model of attention slower than its space-based counter part. Integration of the space- and object-based analysis in the same framework will have superior performance and is expected as the quality of a computational attention model for robotic systems.

**Issue 3. Optimal learning strategy**

This issue is particularly related to visual search. The robot needs to know the visual features of a target object prior to performing a search for it in the environment. Because of the extended number of sensors and actuators, modern-day robots are blessed with higher degrees of freedom in their visual perception. Even an static object in the environment can be perceived by the robot from arbitrary viewing angle. For a dynamic object the possibilities are even higher. To the best of our knowledge, there is no such image feature which is invariant to arbitrary affine transformation, change in viewing angle and lighting condition. Consequently, in order to identify an object in an arbitrary setting the robot requires to learn 'several' views of the object. The precise number to quantify the term 'several', however, is not known. The visual attention model of a robot, therefore, should have a reasonable strategy (with human supervision) to learn sufficient visual features of an object for identification in arbitrary setting.

**Issue 4. Generality**

As shown in Fig. 1.1, in majority of the computer vision models of visual attention, visual search and visual exploration run in mutual exclusion of each other. The desired loop of

attention (visual search or visual exploration) is manually activated by the programmer. Such a manual selection of the mode of visual attention significantly reduces the generality of an attention model and makes it unsuitable for robotic applications. A robotic visual attention model must be able to switch back-and-forth autonomously between the two modes of attention depending on the behavioral requirement.

**Issue 5. Prior training**

This issue is also related to visual search. The robotic applications can not afford to have a separate off-line training phase for visual search. A robot has a very little use as a task-assistant of human if it requires a precise training to learn every possible object prior to performing a search for it. Rather, it is generally expected in the cognitive robots that they will learn while working, much the same way we humans learn.

Many of the research issues stated above have strong mutual dependency on each other. For instance, a strategy to deal with the changing reference frame (*Issue 1.1*) will inherently provide a solution to implement the dynamic IOR (*Issue 1.2*). Again, for the sake of generality (*Issue 4*) if we integrate visual search and visual exploration in the same framework such that the model can switch back-and-forth between the two modes, there will be no room for prior training (*Issue 5*). In other words, the learning has to be performed on-line in an integrated framework of visual search and exploration. Again, if the target-learning is performed on-line, an intelligent learning strategy must be devised to ensure that the robot obtains enough information about the target for identification in arbitrary settings (*Issue 3*).

Addressing the *Issues 1 -5* is a crucial requirement to design a sound model of visual attention for cognitive robots. In response to this requirement we observe the rise of a separate group of visual attention models dedicated solely for robotic applications. There is no doubt that this new group of models are heavily inspired by the computer vision models of visual attention, specially when it comes to the detail of visual feature processing, but they attempt to address at least some of the research issues stated above. For instance, a popular choice to address *Issue 1* and *Issue 2* is robot-centric approach of visual attention as reported in [49, 51–55]. Imitation learning and scaffolding are obtaining increasing popularity to address the *Issues 3, 4,* and *5* [56–61]. Each of these approaches, however, has their pros and cons. A complete model of visual attention providing solution to all of these research issues is yet to be delivered.

## 1.3 Objectives

Two objectives are set for the research reported in this thesis.

**Objective I:** Development of a bio-inspired model of visual attention for cognitive robots which will

- permit the robot to execute overt attention with head-eye movements,
- be able to resolve the research issues arising from overt shift of attention, e.g., the change of camera and image coordinated systems, the implementation of dynamic IOR, the partial appearance of different objects,
- integrate space- and object-based analysis of visual attention in the same framework,
- run autonomously with minimum amount of human involvement, and
- be, as much as possible, independent of any prior training such that the success of the model does not depend on the robustness of a training algorithm/session.

**Objective II:** Implementation of the proposed model on a real robotic system.

## 1.4 Contributions

The thesis makes several contributions while meeting the objectives stated in section 1.3.

- The thesis proposes a novel Bayesian model of visual attention for cognitive robots.
- The proposed model makes the first attempt to exploit the theory of biased competition (BC) [34], a very famous neurodynamic theory of primates' visual attention, to design a model of visual attention for robotic cognition.
- The proposed Bayesian model of attention offers a robot-centric solution of visual attention to address the research *Issue 1*.
- An attention-oriented speech-based human robot interaction (HRI) framework is proposed to address *Issues 3, 4,* and *5*.

## 1.5 Organization

The rest of the thesis is organized in the following manner.

**Chapter 2** provides a review of literature on visual attention models developed for robotic systems. The major focus of this review is how the research issues stated in section 1.2 have been addressed in the current robotic literature and what are the existing open

challenges. For the sake of continuity the chapter will also provide a brief history of the computational modeling of visual attention.

**Chapter 3** will describe the proposed Bayesian model of visual attention for cognitive robots. The chapter provides necessary mathematical formulation to establish the model along with its proposed particle filter implementation.

**Chapter 4** presents a number of performance criteria which will be used to evaluate the performance of the proposed visual attention model. The chapter then presents a set of real-world experiments conducted on a robotic camera head for performance evaluation of the proposed model.

**Chapter 5** describes a multi-modal extension of the Bayesian model presented in chapter 3.

**Chapter 6** presents a set of experiments to validate the performance of the multi-modal Bayesian attention model described in chapter 5.

**Chapter 7** concludes this thesis with a summary of the works presented along with the future direction of research. This chapter also lists the publications originated from the research work presented in this thesis.

## 1.6 Conclusion

This chapter has discussed the motivation and specified the goals of this thesis. It has reported a brief background of visual attention modeling and described the motivation behind developing a visual attention model for cognitive robots. The chapter has also identified a set of research issues involved with the robotic visual attention and has set up two objectives of this thesis based on these research issues. Finally, the contribution of this thesis has been summarized.

The next chapter will provide a survey of the literature on visual attention with a special focus on how they have addressed the research issues involved with robotic visual attention discussed in this chapter.

# Chapter 2

# Literature Review

The major focus of this chapter is to shed light on the ongoing research of visual attention modeling for robotic cognition and to investigate on how the existing works deal with the research issues stated in section 1.2. As discussed in chapter 1, the robotic research on visual attention modeling are closely connected with the visual attention research in cognitive psychology, computational neuroscience, and computer vision. This chapter, therefore, also provides a brief history of evolution of the research on visual attention in psychology, neurobiology and computer vision.

## 2.1 Visual Attention: The Biological Basis

*The Principles of Psychology* [36] is probably the first effort to investigate on the visual attention mechanism in the primates. Since then numerous researchers in psychology and neuroscience investigated on the mechanism of visual attention in the primates although we are still far away from having a complete understanding of how attention works [62,63]. The recent imaging technologies (e.g fMRI, PET) have enabled cognitive neuroscience to perform non-invasive studies on the primates' brain which helps to improve our understanding about the operation of brain areas dedicated for visual processing. The current findings on attention allow us to state safely that the attention mechanism of the primates is carried out by a network of anatomical areas which frequently interact with other brain networks while maintaining its own identity. Each anatomical region in this attention network performs specified duty [64]. Several studies on monkey show that more than 30 separate brain areas are involved with the attention network [65]. The processing for visual features starts from lateral geniculate nucleus (LGN) and primary visual cortex (V1), and is believed to be carried out in two functionally specialized processing pathways, namely ventral stream or 'what' pathway and dorsal stream or 'where' pathway [62, 65–69]. The

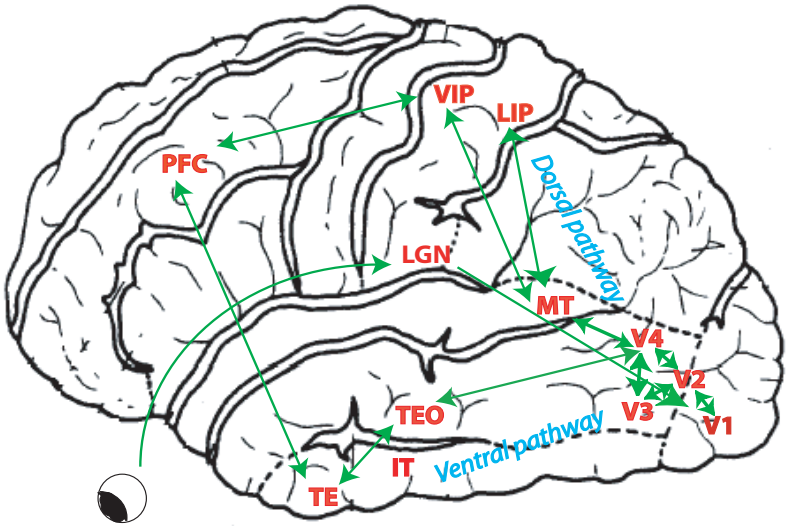



Figure 2.1: Visual pathways shown in a lateral view of the macaque brain

ventral stream runs from V1, to V2, V3, V4, and inferior temporal (IT) cortex areas TEO and TE. This pathway is responsible for object recognition. Along the ventral stream, the complexity of visual processing as well as the size of the receptive field (RF) increases from one step to another. Accordingly, the IT neurons deal with the most complex object feature (shape) and have larger RF. The dorsal stream runs from V1, to V2, V3, middle temporal (MT or V5) area, medial superior temporal (MST) areas, and finally on to the posterior parietal cortex (PPC) areas VIP and LIP. Fig. 2.1 shows the dorsal and ventral stream in the lateral view of the macaque brain. Generally, dorsal stream is responsible for spatial perception and visuomotor activities. Neurons in temporal areas are sensitive to the spatial distribution of the object's features. Similarly, the neurons in the PPC are more sensitive to the stimuli within the foveal region than to other spatial locations. As a whole, the spatial sensitivity of the neuronal RF can be summarized as: the foveal region of retina are mapped in a large region of the visual cortex while the peripheral region get smaller representation (retinotopic mapping) [29]. Consequently, neurons are more sensitive to stimulus in the foveal region than that on the peripheral region. Integrated operation of these brain areas is revealed in different attentional functions and plays a key role in the survival and normal operation of the biological entities.

There are three distinct activities related to the execution of visual attention in the primates [64, 65].

1. Visual orienting.

2. Feature processing.

3. Alertness or sustained attention.

These three subtasks of attention are briefly described below to provide the basic idea about some attentional terms which will be used frequently in the rest of this thesis.

### Visual orienting

Visual orienting refers to the process of shifting the attention to a particular spatial location/object. Orienting could be performed in pure **overt** fashion (where the eye moves for foveation of the stimulus/space of interest [66]), or **covert** fashion (where the object/space of interest undergoes enriched processing through attention without being foveated [70]) or a combination of both (where covert shift of attention acts as a guide to move the eye to the appropriate location [26]). The overt visual orienting could be of two types: 1) **smooth pursuit**, which is characterized as "continuous, slow, smooth, and automatic eye movements that can only be elicited by the tracking of a target moving slowly across the visual field" [63], and 2) **saccadic eye movement**, which is a kind of rapid movement of the eyes made to foveate a target abruptly appeared in the peripheral region of the visual field [71].

A topic of major debate is whether visual orienting directs the attention toward a 'spatial location' or an 'object' and has given birth to two distinct concepts: **object-based attention** and **space-based attention** [72]. A somewhat popular hypothesis on this dichotomy is that attention is directed toward some segmented region/blob (which could easily be a part of an object) in the space rather than to a spatial location [43, 73]. Another distinct phenomenon in visual orienting is the **pop-out effect** which refers to the process of automatically orienting toward any discrepant visual stimuli (e.g., a white car in the pool of red cars) [74–76]. An interesting component of visual orienting is **inhibition of return** (IOR) which suggests that the attention network in the brain resists the attentional shift to a previously attended location/stimuli [26,77]. The IOR manifests itself as a delayed response to a location or an object that has recently been cued.

The dorsal visual pathway in the visual cortex (also termed the posterior attention system) is believed to be involved with the attentional functions related to visual orienting.

### Feature processing

Feature processing refers to the process of foveal inspection of the attended object/location mediated by the ventral pathway. It involves object recognition and visual search process. The most popular hypothesis on feature processing is that the basic visual properties of the objects (e.g., color, intensity, orientation) are processed independent of one another

by different regions along the ventral pathway and re-integrated at a later stage to form the concept of the object for object recognition [78]. An interaction between the ventral and the dorsal pathway occurs to perform the space-feature binding for successful visual search [39, 79]

**Alertness**

Alertness is referred as being prepared to process high priority signals. Alertness or sustained attention is the property of the relatively mature brain and is mediated by the frontal brain areas [70]. Alertness assists the process of visual orienting

The attention of the primates is not yet a fully understood mechanism and, therefore, associated theories and hypotheses are being updated continuously. In spite of this lack of complete understanding, the research on visual attention in the last few decades has reached the status where we can derive a functional framework for attention related activities. This development inspired the researchers in biology, psychology, computational neuroscience, computer vision, and robotics to develop synthetic models of visual attention which have potential applications in their respective fields.

## 2.2 Evolution of Research on Visual Attention: From Biology to Synthetic Modeling

This section sheds light on the evolution of computational modeling of visual attention. The pioneering work on visual attention, the *feature integration theory* [39], was proposed in psychology. The generic purpose of the attention models in psychology is to use the behavioral data (of the primates) to explain human perception and cognition [39,40,42,80]. The *feature integration theory* went through several facets of development to accommodate the new findings on attention from psychological experiments. A comprehensive survey on this popular theory is available in [81]. One major drawback of the early *feature integration theory* is that it considers only the bottom-up effect in attentional selection. The *guided search* model [40] overcomes this limitation by invoking the effect of top-down selection. The *guided search* model is mostly focused on explaining the attentional functions related to visual search. Similar to the *feature integration theory*, continuous upgrading is observed in the *guided search* model [35,82,83]. Another influential model in psychology is the CODE theory of attention [42] which is an integration of the theory of visual attention in [80] with the theory of perceptual grouping by proximity [84]. A major difference of CODE theory as compared to the *feature integration theory* and *guided search* model is it considers both space and object during attentional selection. Besides these, there are many other models of visual attention available in psychology (please see [85] for a comprehensive survey).

The synthetic models of visual attention developed in neurobiology and computational neuroscience are based on the neurobiological findings from lesion study and brain imaging study (e.g., fMRI, PET). The goal of these models is to faithfully reproduce the results obtained from the study of different attentional networks in the primates' brain. Given the fact that study on the brain of live subjects is a very delicate matter and is subjected to ethical bindings, accurate models in computational neuroscience plays a critical role in understanding the operation of different brain networks. The models in computational neuroscience are generally not concerned about the technical applications. A survey on the neurobiological models of visual attention is available in [86].

One of the most popular theories of attention in neuroscience is the *biased competition hypothesis* (BC) (also known as *integrated competition hypothesis* [34, 45, 68, 87, 88]). A number of computational models have been proposed in computational neuroscience based on the postulates of the BC hypothesis [8–11, 89, 90]. These models invoke many new findings of attention, e.g., combination of object and space based analysis for attentional selection, integration of top-down and bottom-up bias, integration of covert and overt shift of attention.

Table 2.1: Synthetic models of visual attention.

| Model | Discipline | Synopsis |
|---|---|---|
| Feature integration theory [39, 81] | Psychology | • Different features (e.g., red, vertical) register their saliency in separate *feature maps* and *feature maps* are summed up to create a *master map* of saliency<br>• Primitive features: Color, intensity, orientation<br>• Visual search is fast and parallel for a target with at least one unique primitive feature but slow and serial if the target shares several primitive features with the surroundings |
| Guided search [40, 82] [35, 83] | Psychology | • Features (color and orientation) register their saliency in two bottom-up *feature maps*<br>• A top-down *feature map* for each feature is created using the unique feature<br>• *Feature maps* are summed up to create an *activation map* of saliency |
| CODE theory of attention [42, 80] | Psychology | • Integrates space- and object-based approach<br>• Proximity effect is used for perceptual grouping<br>• A group of stimuli is chosen based on their strength and the subject's bias in favor of them |
| Kochs's | Neurobiology | • Coins the term 'saliency map' which encodes the |

Continued on next page

**Table 2.1 – continued from previous page**

| Model | Discipline | Synopsis |
|---|---|---|
| model [7] | | conspicuousness of different image locations • Proximity effect is used during the computation of saliency • A *winner take all* (WTA) mechanism is proposed to identify the current focus of attention • A WTA-based IOR mechanism is proposed |
| Biased competition [34, 68] [45, 87] | Neuroscience | • Being excited by the visual stimuli, the visual neurons engage in a mutually suppressive interaction • The attention mechanism biases this competition through feedback bias mechanism • Feedback bias can favor neurons excited by a behaviorally relevant space or features (e.g., color, shape, texture) |
| Deco's model [9, 89] | Computational neuroscience | • Implements the visual attention mechanism in the framework of biased competition • A pool of neurons implements three modules: early visual module (EM), ventral-stream module (VM) and dorsal-stream module (DM) • The functions of the neurons in EM, VM, DM and the type of their connectivity are the same as that in the primates' visual cortex • The interaction between VM and DM through the EM enables translation-invariant object recognition and search |
| Linda's model [10, 90] | Computational neuroscience | • Implements the search behavior of a monkey in case of a conjunctive search for color and orientation in the frame work of biased competition • Pools of neurons with mean field population dynamics are used to represent the areas of visual cortex to model different stages of visual processing |
| Hamker's model [11, 91] | Computational neuroscience | • Introduces the concept of 'perceptual map' • Considers the top-down influence in attention • For visual search the target features are preserved in working memory which enhances the saliency of target-like features |

Continued on next page

**Table 2.1 – continued from previous page**

| Model | Discipline | Synopsis |
|---|---|---|
| | | • Considers both covert and overt attention |
| NVT & extensions [18, 19, 92] | Computer vision | • Inspired by Kochs's model in several aspects<br>• Primitive features: color, intensity, orientation<br>• Multi-scale analysis of image<br>• To include the top-down influence the feature maps are multiplying by some weights which are determined through off-line training |
| VOCUS [23, 93] | Computer vision | • A number of theoretical and implementation-related improvement over the NVT [18]<br>• A novel mechanism for feature maps fusion<br>• Uses background information during visual search<br>• Requires separate training phase to learn target features from several images |
| Selective tuning model [20, 38, 94] | Computer vision | • Luminance, orientation, color, and motion are analyzed<br>• Pyramid style processing of information where the stimuli of interest are located at the top and control an inhibitory beam which could inhibit or pass a zone for further processing<br>• The top-down influence is modeled through manipulation of the inhibitory beam |

The rise of the computer vision models of attention happened almost in parallel with the psychophysical models [7, 16, 18–24]. The goal of these models, however, is somewhat different from the psychophysical models of visual attention. The major focus of the computer vision models is to develop a technical system of attention which has potential application in pattern recognition, video surveillance, and AI robotics. The model proposed in [7] is the first computer vision model of visual attention and relies on the basic postulates of the *feature integration theory* [39]. Probably the most influential computational model in computer vision literature is the *neuromorphic vision toolkit* (NVT) [18] which has been extensively used in many other later models of visual attention. The early version of NVT [18] performed only bottom-up analysis of attention but a later modification in [19] invokes the effect of top-down selection. Some of the flaws of NVT have been alleviated in the NVT-based model *visual object detection with a computational attention system* (VOCUS) [23]. A very famous connectionist model of visual attention in the computer vision literature is the *selective tuning model* [20]. A unique characteristic of this model is, in spite of being a technical model of attention, the *selective tuning model* [20] and

all of its later variants [38, 94] are tightly coupled with biological principles. Table 2.1 summarizes the major properties of some of the most influential attention models/theories in psychology, neuroscience, and computer vision.

The computational models of visual attention proposed in computer vision gained widespread popularity in robotic research. The early works on attention in robotics mostly adopted the attention models in computer vision and modified them to meet the requirements of the robotic applications. The recent trend, however, is to design attention models which are specific to robots. The following section sheds light on the efforts in robotic research for visual attention modeling and investigates on how they address the research issues discussed in section 1.2.

## 2.3 Visual Attention for Robotics Systems: The Current Trend

A number of attempts are observed in robotic literature on the modeling of visual attention for cognitive robots. Many of these models propose general solution to tackle the research issues while some address them in task-specific manner. This thesis classifies the existing works on robotic visual attention into two groups based on their goal and motivation.

1. **Overt attention models:** The research works in this group focus on developing camera maneuvering technique based on the principle of overt visual attention. A considerable number of overt models are inspired by the covert attention models proposed in computer vision.

2. **Application-specific visual attention models:** The research works in this group develop robotic attention models which are tuned to specific task, e.g., localization, navigation, manipulation, HRI and joint attention. Many of these tasks consider the property of selectivity of the primates' visual attention as a mere technique to solve the desired task while others consider visual attention as a component of developing cognition in the robots. Most of the works related to HRI and joint attention fall under the second category while attention-based robot navigation, localization and manipulation are generally the members of the first category.

Analysis of each group with respect to the research issues stated in section 1.2 are described below.

### 2.3.1 Overt Attention Models

The attention mechanism in the primates integrates the overt and covert modes of attention in a highly efficient manner: the stimulus of interest is selected covertly and then placed at the foveal region through overt movement of the eyes [95]. Evidence is also available in favor of the independent occurrence of covert and overt attention [96,97]. In case of robotics applications, however, direction of attention mediated by eye/head movement is the most suitable choice. The major reason behind this is placing the object of interest at the center of visual field facilitates the learning process. Besides, head/eye movement of the robot provides a way for the user to understand the current gaze of the robot which is specially important in many applications (e.g., HRI). Inspired by these requirements, a number of efforts are observed in the robotic literature for modeling of overt visual attention. At the early stage of this research the principle of overt attention (to place an object of interest at the center of visual field) helped the concept of 'active vision' [98], 'active perception' [99], or 'animate vision' [100] to be established in computer vision. For instance, the theme of 'active vision' is to actively position a sensor (preferably a camera) for obtaining enriched information to solve the basic computer vision problems (e.g., shape from shading and depth computation, shape from contour, shape from texture, and structure from motion). A number of active vision models propose mechanism of positioning a camera based on the feedback from a visual attention model [101–106]. The major focus of most of these models is the control aspects of saccade generation and/or smooth pursuit tracking. A common practice among these works is the use of some well-known covert models of attention (e.g., [7, 18]) to identify the most interesting/salient region in a scene. These active vision models, therefore, are less concerned about the research issues stated in section 1.2 of this thesis.

The overt attention models described in [47, 51, 53, 54, 107–117] have been designed to be implement in the robots/robotic heads. Among them the models in [47, 107, 108, 110] adopted different variants of the covert model NVT [18] to identify the visually salient/task-relevant stimuli and introduced different measures to deal with the research issues involved with robotic overt attention. For instance, the model in [108] addresses the *Issue 1.1* by adopting the idea of shifting the entire content of the saliency map in the direction of head movement as stated in [12]. The object-based overt attention system proposed in [47] implements a simple form of integrated object- and space-based IOR to deal with *Issue 1.2* and *Issue 2*. The overt model described in [110] suggests to re-map the location of the recently attended object to the transformed image coordinate in order to implement a space-based IOR (*Issue 1.2*). The problem involved with the partial appearance of objects (*Issue 1.3*) is not noticeable in the experiments demonstrated in [110] due to the use of a wide angle camera. The model in [107] demonstrates few simple cases of overt attention and does not provide any effective solution to any of the research issues.

The neural network based overt model reported in [113] is tightly coupled with biology

(with respect to motor aspects of attention) and is focused on implementing visual exploration behavior guided by the novelty preference characteristics of primates' attention. The identification of novelty in [113], however, is achieved though the implementation of space-based IOR, i.e the robot moves to novel locations (through successive application of space-based IOR) and thereby attends to novel objects. The model [113] also relies on NVT [17] for visual saliency calculation. The issue of dynamic IOR (*Issue 1.2*) is addressed by remembering the locations of the previously visited stimuli. To comply with this strategy the model [113] assumes that all of the stimuli lie within the visual field of the robot at all times. This is a strong assumption which is valid in the experiments demonstrated in [113] but generally does not hold in most robotic applications. The *Feature Gate* model [21] based overt model in [114] claims to propose a general purpose model of visual attention for the humanoid robots but mostly focuses on mimicking the feature-processing attributes of the primates' attention system (e.g., log-polar retino-cortical mapping, banks of oriented filter).

All of the overt models discussed thus far follow an image-centric approach where the attention model operates absolutely in the image plane. Focus of attention is evaluated based on the content of a given image and necessary motion command is calculated based on the image dimension and the parameters of the camera optics. In contrast to this traditional image-centric approach, the recent models of overt attention adopt a robot-centric solution for attentional selection [51–54]. In case of robot-centric approach it is assumed that a robot is a human-like autonomous entity which decides 'what to look at?' based on its perception of surrounding with respect to an ego-centric frame of reference. For instance, the model in [51] considers an ego-sphere of infinite radius around a robotic head and the robot is able to project the perceptual information collected through different modality on the surface of the ego-sphere. The concept of the head-centric ego-sphere provides an elegant solution of the issues involved with overt shift of attention (*Issues 1.1, 1.2, 1.3*). The multi-modal attention model [51] considers both acoustic and visual information and combines them into a single head-centric saliency map by taking the maximum value between the two modes. This straightforward methodology of fusing multi-modal perception into a single saliency map has several shortcomings, e.g saliency maps from different modes have same influence on the aggregated saliency map. A detailed analysis of this problem is available in [23]. The model [51] operates in a purely bottom-up fashion and performs NVT [17] style space-based analysis for saliency calculation. The concept of an ego-sphere is also present in the attention model reported in [52]. The model [52], however, uses the principle of attention for updating a sensory ego-sphere with overlapping images perceived by the robot. The multi-modal attention model in [116] also uses sensory ego-sphere to focus, learn, and then track the salient stimuli (bright colored moving objects or human faces) in the visual field. The model [116] integrates the visual search and visual exploration in the same frame work and thereby eliminates the presence of a training phase during visual search (*Issue 4, 5*). The overt model in [54] uses the term 'scene space' instead of 'ego-sphere' to represent a

two dimensional surface which contains the information perceived by the robot with respect to the robot's head-centric coordinate system. The purpose of the model [54], however, is to track a set of predefined objects in the surrounding. To achieve this goal it uses only the color information of the target objects and performs object-based analysis to implement the IOR (*Issue 1.2, 2*). Although the model might have the potential to be extended for complex attention scenario, the current implementation in [54] is dealing with only few simple cases. The models in [56, 57] use scaffolding where the human operator heavily guides the robot to teach what to focus on through speech command and hand-gesture. This solves the problem of prior training (*Issue 5*) and optimal learning strategy (*Issue 3*) with the price of having a dedicated human operator throughout the attention process. Unfortunately, having such a dedicated human operator severs the generality problem (*Issue 4*). A reduced amount of human-dependency for learning of attention is observed in the multi-modal overt attention model described in [53]. The model [53] proposes the idea of an attention map, similar to 'probabilistic occupancy grid' widely used in robotic mapping [118], to encode the saliency of the robot's surrounding. The attention map can be modulated by the task-demand conveyed to the robot through speech command. The model, however, requires significant amount of prior training and manual work to create an useful attention map for any specific robotic application.

For quick reference, Table 2.2 shows a comparative analysis of the overt attention models discussed in this section. In the tables the issues are denoted by the letter *'I'*, e.g *'Issue 1.1'* is written as *'I 1.1'* etc.

Table 2.2: Overt attention models for the robots.

| References | Synopsis | Issue(s) addressed |
|---|---|---|
| [108] | • Relies on [7] for saliency<br>• Considers the bottom-up effect only | • *I1.1*:Re-maps the entire saliency map to new camera coordinate |
| [47, 109] | • Object-based attention model<br>• Integrates top-down and bottom-up effect<br>• Color-based bottom-up saliency | • *I1.2*: Memorizes the locations of the last visited objects along with the objects' features |
| [110] | • Attention model for stereo-vision<br>• Provision for top-down and bottom-up bias but no actual demonstration of top-down effect<br>• Saliency map is inspired by NVT | • *I1.2*: Locations of the attended object is mapped to the new camera frame |
| [113] | • Tightly coupled with biology<br>• Relies on NVT for saliency<br>• calculation | • *I1.2*: Remembers the location of the previous focus with the assumption that all objects |

Continued on next page

**Table 2.2 – continued from previous page**

| References | Synopsis | Issue(s) addressed |
|---|---|---|
| | | remain within visual field |
| [51] | • Robot-centric approach<br>• Operates only with bottom-up information<br>• Saliency calculation is inspired by NVT [18] | • *I1.1:* Projects sensor data to an ego-centric frame of reference<br>• *I1.2:* Performs space-based IOR<br>• *I5:* Fuses acoustic and visual information on a sensory ego-sphere through a 'Maximum' operator |
| [52] | • Robot-centric approach<br>• Uses visual attention to map overlapping images on a sensory ego-sphere<br>• Uses *Feature Gate* [21] model | - |
| [116] | • Learning of attention<br>• Object-based analysis<br>• Considers color, face, and sound as salient and tracks them in consecutive frames | • *I1.1:* Projects sensor data to an ego-centric frame of reference<br>• *I1.2:* Performs object-based IOR<br>• *I5:* Performs on-line learning of target and maintains a memory |
| [54, 119] | • Attention model for tracking a set of object<br>• Uses color information only to evaluated saliency | • *I1.1:* Projects sensor data to a 'scene space' which is expressed w.r.t the head-centric coordinate system<br>• *I1.2:* Performs object-based IOR |
| [53] | • Focuses only on the task-relevant objects<br>• Requires a lot of prior learning | • *I1.1:* Projects sensor data to an 'attention map'<br>• *I5:* Uses speech command to bias attention selection |

### 2.3.2 Application-specific Visual Attention Models

The application-specific visual attention models are tuned to the applications they are developed for. Visual attention mechanism has at least two properties which can be tuned in the application specific manner.

- Selectivity: The basic idea of attention is to focus on a relevant visual stimulus for further processing. The 'relevancy' of a stimulus can be defined in terms of its similarity with a set of predefined task-specific features. The irrelevant information

in the visual scene are not considered for further processing and thereby reducing the computational load on an artificial system.

- Visual Search: Visual search is an important property of the primates' visual attention mechanism which helps to focus on the target-related information in relative exclusion of the others. Thus the visual search is an special case of manifesting selectivity. The success of a visual search and the time requirement depends on the number of distractor stimuli present in the VF and the number of features they share with the target.

Exploitation of these two properties often causes visual attention to reduce to a tracking problem in many application-specific models of visual attention. In case of attention-based tracking, many of the research issues stated in section 1.2 do not arise. For instance, the object that is to be tracked is learned once and is tracked in the subsequent camera frames. Each incoming camera frame is searched for this specific object. Thus there is no need to implement the IOR and the change of coordinates does not have any significant effect on the tracking decision (hence, no need to address *(*Issues 1.1, 1.2)). An example of such attention-based tracking is demonstrated in [93]. Here a covert model of visual attention VOCUS [23] is used to perform simultaneous localization and mapping (SLAM) by a mobile robot [93]. The role of the attention model is to identify the most salient stimuli in the scene (the landmarks) and then keep on tracking that specific stimuli in the successive frames by adjusting the camera head. Similar strategy of attention-based tracking is also adopted in [120] for vision-based SLAM by mobile robots. To deal with the partial appearance of object (*Issue 1.3*) the model in [93] adopts the strategy that the landmarks that reside at the center of the visual field are given higher priority as it is likely that they can be tracked for an extended period of time.

The attention model in [115] exploits the principle of visual search for robot navigation and mapping. The robot learns the visual features of a set of objects during an off-line training phase. During the autonomous navigation the robot searches for the learned objects, which appear as landmarks, in natural indoor environment. A number of important parameters of the navigation model are chosen based on the off-line training phase. The objects location are projected in an ego-centric frame of reference in order to update a 3D occupancy grid which contains the information about the landmarks/obstacles in the robot's workspace (*Issues 1.1, 1.2, 1.3*). The model described in [117] is dedicated to design a Bayesian approach of fast visual search for human faces in a video stream. To achieve faster response the attention model sacrifices all other visual information except the intensity feature. Similar to [93], this model [117] also considers each incoming frame as an isolated static image and does not implement IOR. Similar kind of attention model (focusing on the visual search) is also proposed in [121]. Here the robot is provided with a predefined set of features to search for, e.g., a talking person, human face, human legs

located at the closest distance, etc. The robot then uses its multi-modal perception to search for these features and attend to them. The task-specific attention model proposed in [122] performs visual search for pre-specified object patterns (domino) and executes manipulative actions based on their 3D locations. The attention model in [123] is designed for social interaction with human. The model uses omni-directional camera and the nature of images obtained from such cameras enables the visual features to be registered directly in an ego-centric frame of reference. This inherently offers a solution to the problem of coordinate change (*Issue 1.1*), dynamic IOR (*Issue 1.2*), and partial appearance (*Issue 1.3*).

The visual attention model developed for HRI mostly considers attention as a step toward making the robots cognitive. Visual attention plays a significant role in HRI in order to establish joint attention [124] between the robot and the human. Establishing joint attention between a human and robot requires that a robot should be able to detect and manipulate the attention of the human, socially interact with the human, and finally see itself as well as the human as intentional agents. Joint attention, therefore, is an excellent tool to build a meaningful HRI system. A basic requirement of joint attention is that the robot should posses a human like attention model with the capacity to manipulate attention of other agent as well as of being manipulated by other agents. The visual attention models proposed in HRI literature, therefore, have strong emphasis on top-down modulation of attention. A number of approaches, inspired by the cognitive development of human child, are available to model the top-down influence in attention selection, e.g imitation learning, scaffolding. These methodologies have their own unique way of addressing the issues sated in section 1.2. In some cases, however, their way of addressing one issue worsens the consequence of the others.

In case of imitation based learning of attention, the robot imitates the movements (head/eye/hand) of a person (the user or the operator) to exhibit overt attention behavior [125]. Thus the top-down bias appears as the commands from the human operator conveyed through natural speech, hand gesture, gaze direction, etc. For instance, the models in [59–61] evaluate the gaze direction of the user to identify the object of interest to attend. Thus the model guides a robot to look at the objects to which its user is also looking and thereby establishing simultaneous looking behavior which is a major requirement of joint attention [124]. The work in [126] uses the head pose and eye-gaze direction of the user to identify the object to attend. To further enhance the quality of joint attention it uses pointing behavior by the robot once it attends to an object. The shared attention model in [127, 128] uses the gaze direction as a cue to decide which object to attend. An integration of imitation learning and visual search is observed in the connectionist model of joint attention reported in [129, 130] where the robot learns a set of motion patterns in an off-line training phase and reproduces them when it finds similar kind of motion pattern performed by the user. The model introduced in [131] performs overt attention

based on gaze direction of the user as well as spoken command. A major complexity of imitation learning is in order to be accurate it requires the robot to have an efficient learning strategy to conceptualize the underlying goal of the imitative actions and form knowledge from that [132]. In other words, the robot has to decide on its own about 'what to imitate?', which, by itself, is a type of skill that requires cognition.

A bit more relaxed approach (with respect to the amount of cognitive load on the robot) as compared to imitation learning is attention mediated by scaffolding [133]. Here the idea is to explicitly attract the attention of the robot to certain specific stimuli through different kind of actions, e.g., verbal command, hand-gesture, motionese. For instance, the attention model in [56] uses hand-gesture and verbal command to guide the attention of the robot toward novel objects. Similar approach of attention guiding has been used in [57] in order to perform grasping task by a robot manipulator. The attention model in [134] uses motionese in order to make certain stimuli to appear as extremely salient in the robot's perception. The model [134], however, relies on NVT [17] for calculation of saliency, does not implement any form of IOR, and operates in a pure off-line fashion. The attention model developed for HRI in [111] is based on the psychophysical model of visual search proposed in [40]. The model is sensitive to task-specific stimuli (e.g., human face, toys with specific color) and attends to them based on the task-context. This model also uses motionese to guide the robot's attention toward certain specific stimuli. The model performs the IOR and the habituation effect with moving camera but does not mention explicitly how the issues involved with camera movement have been addressed.

The imitation learning approach and scaffolding relieve a visual attention model from worrying about the issues such as change of image coordinates (*Issue 1.1*), implementation of IOR (*Issue 1.2*), partial appearance of the objects (*Issue 1.3*), and generality (*Issue 3*). The human operator takes care of these issues and the robot's attention model just mimics the operator. Such a huge benefit, however, comes with the heavy price that a human operator must be dedicated for a robot, which is often an unrealistic demand for autonomous robotic applications.

For quick reference, the Table 2.3 shows a comparative analysis of the application-specific attention models discussed in this section.

Table 2.3: Application-specific models of visual attention

| Application | Synopsis (of the attention model) | Issue(s) addressed |
|---|---|---|
| Vision-based SLAM [93] | • Extension of VOCUS [23]<br>• Identifies the most salient region in a frame and tracks it in the consecutive frames | • *I1.3:* Higher priority is given to the stimuli (landmark) located at the center of the frame |

Continued on next page

**Table 2.3 – continued from previous page**

| Application | Synopsis (of the attention model) | Issue(s) addressed |
|---|---|---|
| Navigation and mapping [115] | • Attention model for visual search<br>• Learns object in an off line training phase and detect them in the environment | • *I1.1:* Update a 3D occupancy grid with the locations of the objects in the 3D world |
| Search for human face [117] | • A Bayesian model of visual search<br>• Searches for face features in each incoming camera frame<br>• Only intensity feature is used | Visual attention reduces to target tracking problem which does not require to address the research issues |
| People tracking [121] | • Model of visual search to track multiple person<br>• Abstract level information about target are given to the robot and multi-modal data are analyzed to identify people | Visual attention reduces to target tracking problem which does not require to address the research issues |
| Social interaction [123] | • Integrates vision and audition to identify and focus on human<br>• Option for top-down and bottom-up influences is available but no actual demonstration of top-down effect | • *I1.1, 1.2, 1.3:* Omni-directional vision is used to maintain a $180^o$ wide attentional span |
| Joint attention in HRI [60, 61] [58, 59] | • Considers the caregiver's face as the most salient region<br>• Calculates the gaze direction of the caregiver from his/her face<br>• Attends overtly to the object(s) the caregiver is looking at | • *I1.1, 1.2, 1.3, 4, 5:* Use of imitation learning approach let the human operator to take care of the issues |
| HRI [126] | • Learns sensorimotor mapping through active interaction<br>• Uses head-pose and eye-gaze direction to identify the next focus | • *I 1.1, 1.2, 1.3, 4, 5:* Use of imitation learning approach let the human operator to take care of the issues |
| Joint attention [127, 128] | • Uses the gaze direction of the human to identify the next focus of attention | • *I 1.1, 1.2, 1.3, 4, 5:* Use of imitation learning approach let the human operator to take care of the issues |
| HRI [129, 130] | • Learns motion patterns off line and reproduces them when similar motion patterns are observed in the environment | • *I 1.1, 1.2, 1.3, 4, 5:* Use of imitation learning approach let the human operator to take care of the issues |

Continued on next page

**Table 2.3 – continued from previous page**

| Application | Synopsis (of the attention model) | Issue(s) addressed |
|---|---|---|
| HRI [56, 57] | • Hand-gesture and verbal command are used to guide the robot's attention to a novel object<br>• Object-based analysis | • *I 1.1, 1.2, 1.3, 3:* Uses scaffolding to draw attention of the robot to the target object<br>• *I4:* Switches back-and-forth between visual search and visual exploration based on speech command |
| Joint attention [134] | • Relies on NVT for saliency<br>• Use motionese to make some regions highly salient to the robot and thus drawing the attention toward that regions | • *I 1.1, 1.2, 1.3, 3:* Uses scaffolding to draw attention of the robot to the target objects |
| HRI [111] | • Implements Guided Search [40]<br>• Attends to task-specific stimuli, e.g., face, colored toy, etc.<br>• Motionese is used to make the task-specific stimuli to appear as the most salient | • *I 1.1, 1.2, 1.3, 3:* Uses scaffolding to draw attention of the robot to the target objects |

## 2.4 Conclusion

This chapter has presented a brief survey of research on synthetic modeling of visual attention. The chapter first introduces the biological basis of visual attention in a very concise manner. This is followed by a brief discussion on the rich literature of visual attention in psychology, neuroscience, and computer vision. Finally the existing works on visual attention in robotics literature and how they deal with the research issues identified by this thesis are discussed in detail.

The following chapter will introduce the proposed Bayesian model of visual attention.

# Chapter 3

# The Proposed Bayesian Model of Visual Attention

The Bayesian model proposed in this thesis provides a robot-centric solution of visual attention. The proposed model recursively estimates the next head-pose of a robot such that a behaviorally relevant object resides approximately at the center of its visual field (VF). The primates (more specifically the human) are able to perform overt and covert shift of attention in an integrated manner as well as independent of each other. For implementation of a technical system of overt visual attention, however, it is a requirement that every overt shift will be preceded by a covert shift of attention. In other words, a technical model of overt attention has to focus on the object of interest in a covert manner before physically orienting the camera toward that object. The proposed model also relies on this strategy. The transition between two successive head-poses is guided by a set of criteria which are inspired by the biased competition (BC) hypothesis of visual attention in the primates. The proposed Bayesian model is implemented using particle filter.

This chapter first describes the biological motivation of the model along with its functional overview for a clear understanding of the Bayesian formulation in the context of robotic visual attention. This will be followed by the detail of the particle filter implementation of the Bayesian attention model.

## 3.1 Biological Motivations of the Model

There are two major aspects where the proposed model mimics the primates visual attention system.

1. Transition of attention

2. Hierarchical processing of visual features

The biological motivation of the proposed model in these two aspects are discussed below.

### 3.1.1 Transition of Attention

The proposed model follows the postulates of BC hypothesis [34] to guide the switching of attentional focus from one stimulus to the other. For a better understanding of the biological inspiration of this behavior, a brief description of the BC hypothesis is provided here.

#### Biased competition hypothesis

Avoiding the biological details, which are available in [87,135], the BC hypothesis of visual attention in the primates can be summarized as follows.

- **Competition among visual stimuli:** When multiple stimuli appear in the VF of a subject, they activate populations of neurons in different areas of the visual cortex. These activated neurons engage themselves in a mutually suppressing interaction. The only objective of this competition is to win the limited processing power of the brain. If two stimuli in the VF excite the neurons of the same local region of the cortex then the competition is assumed to be stronger than the case when they excite two different regions of the visual cortex.

- **Biasing the competition:** The competition among visual neurons can be biased in favor of certain specific set of neurons. The result of this biasing effect manifests itself as the visual attention behavior of the subject. In other words, the subject focuses on the stimuli which have activated the set of neurons that received the biasing signal.

- **Criteria of feedback bias:** There are two well-investigated criteria of feedback bias. The first one is stimulus-driven. In this case the visual properties of a stimulus cause the set of neurons excited by this stimulus to win in the competition. For instance, one stimulus might have strong contrast as compared to the others and draws the attention quickly. The saliency of the stimuli, however, depends on a number of factors among which the spatial location of the stimuli and the contrast in color, orientation, and intensity are the most investigated. The stimulus-driven bias is commonly termed as the 'bottom-up bias' in the visual attention literature. The second criteria of feedback bias is independent of the visual strength of the stimulus and is commonly termed as the 'top-down bias'. In case of top-down bias a number of brain

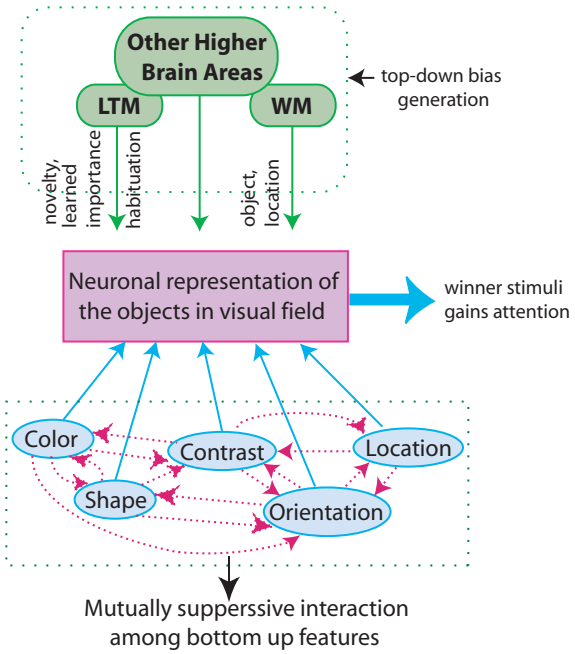



Figure 3.1: An abstract-level graphical representation of biased competition in the visual cortex

areas outside of visual cortex generate biasing signals which modulate the competition among neurons in favor of a specific set. This modulating effect can be revealed in this 'chosen set of neurons' in a number of different ways, e.g. enhancement of their neuronal response, increment of sensitivity to their target visual features, increase in their baseline activity. Irrespective of the way they reveal, the top-down biasing effect generally causes a 'chosen set of neurons' to win in the competition irrespective of the visual saliency of the stimuli that activated them in the first place. The specific regions of visual cortex which contribute in generating top-down biasing signals are not precisely known yet. Two possible sources of top-down bias, however, has been reported in [34]: the long term memory (LTM) and the working memory (WM).

- The attended stimulus achieves further access to the memory and motor system to control the behavior and action of the subject.

Fig. 3.1 demonstrates an abstract level graphical representation of biased competition in the visual cortex of the primates. The bottom-up bias is represented by the competitive interaction among the visual neurons excited by specific visual features, e.g. color, intensity, orientation.

The top-down bias can be classified into two categories: 1) bias in favor of object feature (known as object bias) and 2) bias in favor of spatial location (known as spatial selection). The object bias can be delivered either from the LTM or from the WM. When the bias is delivered from the LTM, it usually favors novel visual features [34, 136, 137]. An exactly opposite kind of bias can be delivered from the LTM where it chooses the most familiar stimuli for attention due to its long-term learned importance [138]. When the object bias is derived from the WM, the attentional process is generally termed as visual search. During visual search the WM holds the visual features of the target stimuli and the stimuli in the visual field which are a good match to these features receive strong top-down feedback bias. In case of spatial selection prior knowledge about the target's spatial location is stored in the WM and a top-down bias is delivered accordingly in favor of the stimuli at that specific location. These two forms of top-down biases seamlessly integrate the space- and object-based attention and thereby advocating the fact that the object- and space- based modes of visual attention are the mere manifestation of two different kind of top-down selection processes. This capacity to inherently integrate the space- and object- based modes of attention is a powerful characteristics of the BC hypothesis of visual attention.

The proposed Bayesian model takes the motivations from the BC hypothesis to guide the attentional switching of the robot among different stimuli. The head-pose space of a robot is continuous and, at a given time, each head-pose enables the robot to focus on one specific stimulus (i.e., placing that stimulus at the center of the camera frame). The head-poses in the pose space, therefore, are behaviorally analogous to the visual neurons with respect to the fact that they are competing with each other to select the stimulus associated with them for attending. The robot requires to resolve this competition among head poses through preferring the head-pose which satisfies its current behavioral requirement the most. Inspired by the tenets of BC hypothesis, the proposed model guides the robot to perform a biasing of head poses using the following criteria.

- A head-pose receives bottom-up bias proportional to the saliency of the stimulus it is focusing at.
- A head-pose receives top-down bias proportional to the behavioral relevance of the visual feature or spatial location it is focusing at.

Two types of behavioral relevance of the objects are considered during top-down biasing. The objects with novel visual features are considered as 'worthy to attend' and receives excitatory top-down bias from the LTM. The objects having similar visual features to a 'sought for' object is considered as 'worthy to attend' and receives excitatory top-down bias from the WM. The proposed Bayesian model, therefore, always chooses the novel objects and the 'sought for' objects as the focus of attention.

### 3.1.2 Hierarchical Processing of Visual Features

The computational models of visual attention (irrespective of overt or covert models) generally observe the property of the primates' visual cortex to hierarchically process the visual features. Accordingly, primitive image features like color and intensity contrast are processed at the first stage (similar to the processing in the V1 area [139]). High level features containing specific object information (e.g. SIFT keypoints, for this work) are processed at a later stage (similar to the processing in the areas V4, TE, MT [139]).

## 3.2 Functional Overview of the Model

The proposed Bayesian model guides the robot to choose a head-pose which has the highest probability of letting the robot to focus on the most behaviorally relevant stimuli in the environment. For evaluation of such probabilities the model uses current sensor measurements as well as the prior knowledge learned throughout the life-time of the robot. Fig. 3.2 shows a functional overview of the model. Here the LTM is a database of features of the attended objects and the WM is a database of features of a 'sought for' object. The prime modality for visual attention is vision, although there are evidences that other modalities (e.g audio, tactile) have modulatory effect on the visual attention behavior of the primates. The proposed Bayesian model, inherited from the Bayesian sensor fusion characteristics, is capable to accommodate the effect of multiple modalities while evaluating the focus of attention. This chapter, however, focuses only on the vision sensor (i.e., the camera). The multi-modal extension of the model to accommodate auditory measurements is discussed in chapter 5.

At each decision cycle the model guides the robot through the following stages.

1. The camera-head of the robot orients to the object identified as 'worthy to attend' during the immediate past cycle. A new frame is captured at this new head-pose.

2. The memory is updated with the high level features of the focused object (i.e., the object located at the center of the current frame).

3. From the current head-pose the robot can switch to a number of different head poses. The Bayesian model helps to identify the most probable head-pose for the robot. The probability of a head-pose, at this stage, depends on the saliency of the visual feature(s) it focuses on. The primitive visual features (e.g. color and intensity contrast) of the current frame are analyzed to identify a set of visually salient regions (a predefined priority order is given to the features as: intensity contrast> color). The head poses which focus on these regions are the potential candidates for the next

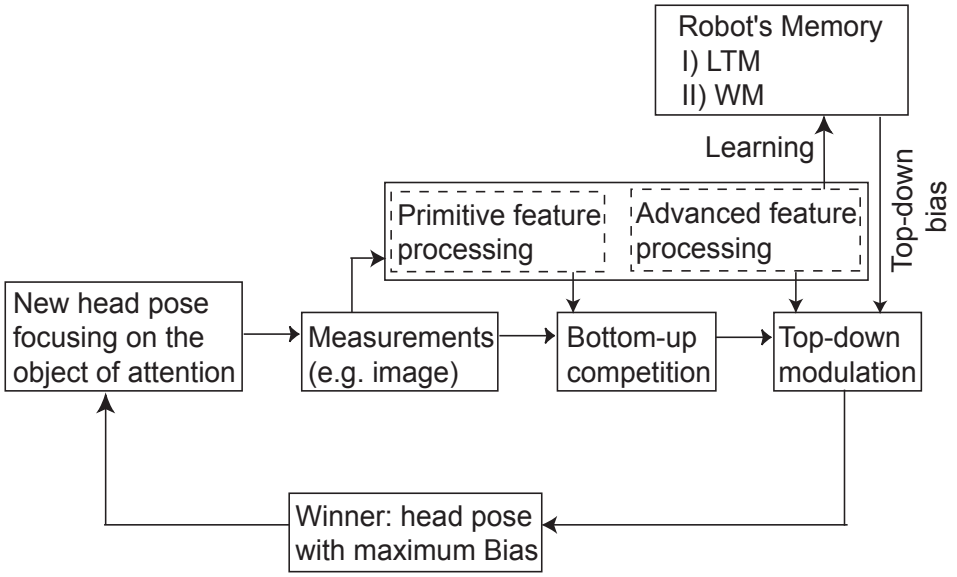



Figure 3.2: Functional description of the proposed Bayesian model

head-pose of the robot. Thus the head poses engage in a competitive interaction with each other to focus on certain visual features. In light of BC hypothesis this competition is termed as the 'bottom-up competition'.

4. The bottom-up competition among the head poses is then modulated by top-down biases from the LTM and the WM. Advanced visual features are used to evaluate top-down biases. The current implementation uses the SIFT keypoints [140] associated with different objects in this regard (complex feature like 'shape' can also be used at this stage to achieve further accuracy in attention decision). Based on the analysis of SIFT keypoints, the LTM generates non-zero bias for the head poses which focus on novel objects while the WM generates non-zero bias for the poses which focus on a 'sought for' object. A head-pose that focuses neither on a novel object nor on a 'sought for' object receives minimum top-down bias. Finally, the robot takes the head-pose which receives the maximum bias.

## 3.3 The Bayes Filter for Visual Attention

The goal of a Bayesian model is to recursively estimate the state of a dynamic system conditioned on the measurement data. In the context of robotic visual attention, the robot

and its surrounding environment constitute a dynamic system and the state $\mathbf{x}$ of this system is the head-pose of the robot. The head-pose of a (stationary) robot is expressed in terms of the pan ($\alpha$) and tilt ($\beta$) angles of the PTU where $\alpha$ is the rotation with respect to the $z_b$ axis and $\beta$ is the rotation with respect to the $x_b$ axis of the base coordinate system $\mathcal{C}_b$ (please see Fig. 1.5 for coordinate systems).

$$\mathbf{x} = \{\alpha, \beta\} \tag{3.1}$$

The center of the two-dimensional head-pose space of a robot, therefore, coincides with the center of the base coordinate system. In case of visual attention by a moving robot the location of the platform in the 3D world and its heading direction are also required to be considered as the state variables. At any time, the system state $\mathbf{x}$ defines a unique location for the camera coordinate system $\mathcal{C}_{c_k}$ and the image coordinate system $\mathcal{C}_{i_k}$. Here, $k$ is the discrete time index. The focus of attention of the robot in the image plane is denoted by $\mathbf{a}_k$ and is located at the center of the image $I_k$ which is situated along the $(x, y)$ plane of the $\mathcal{C}_{i_k}$. According to the requirement of the Bayes filter, the system state $\mathbf{x}$ is Markov, i.e, the past and the future data are independent of each other if the current state is known.

**Statement** The system state $\mathbf{x}$ is a Markov state.
*Justification:* The LTM of the robot holds the information of the objects attended throughout the life-time of the robot. Thus the robot has, at a given time, the knowledge of the objects attended at all other previous head poses including the current one. Prediction of the next head-pose, therefore, requires to analyze the current measurement (i.e., the camera image) and is independent of all other past measurements. This is a reasonable Markov assumption which renders $\mathbf{x}$ as a Markov state.

The symbols $\mathbf{M}_k$ and $\mathbf{m}_k$ are used to denote the LTM and the WM of the robot, respectively. The simplest way of realizing the LTM is to maintain a database of visual features attended throughout the life-time of the robot. In case of the WM, such a database contains the information of a 'sought for' object.

The proposed model recursively estimates the posterior probability $p(\mathbf{x}_k|\mathbf{z}_{0:k})$. The posterior probability is commonly termed as *belief* in the robotic literature. Accordingly,

$$Bel(\mathbf{x}_k) = p(\mathbf{x}_k|\mathbf{z}_{0:k}) \tag{3.2}$$

Here $\mathbf{z}_{0:k}$ denotes the measurements starting from time 0 up to time $k$. There are two types of measurement involved in robotic visual attention: sensor measurement $\mathbf{F}$ and top-down bias $\mathbf{b}$ derived from the robot's memory. The latter measurement does not require any physical sensor. Considering vision as the only modality used by the robot, the sensor measurement $\mathbf{F}$ consists of a set of primitive visual features and a set of advanced visual features.

$$z = \{\mathbf{F}, \mathbf{b}\} \tag{3.3}$$

Equation (3.2) can be written as follows.

$$Bel(\mathbf{x}_k) = p(\mathbf{x}_k|\mathbf{b}_k, \mathbf{F}_k, \mathbf{b}_{k-1}, \mathbf{F}_{k-1}, \ldots, \mathbf{b}_0, \mathbf{F}_0) \tag{3.4}$$

Applying Bayes rule we obtain the following expression.

$$\begin{aligned} Bel(\mathbf{x}_k) &= \eta p(\mathbf{b}_k|\mathbf{x}_k, \mathbf{F}_k, \mathbf{b}_{k-1}, \mathbf{F}_{k-1}, \ldots, \mathbf{b}_0, \mathbf{F}_0) \\ & \quad p(\mathbf{x}_k|\mathbf{F}_k, \mathbf{b}_{k-1}, \mathbf{F}_{k-1}, \ldots, \mathbf{b}_0, \mathbf{F}_0) \end{aligned} \tag{3.5}$$

where

$$\eta = \frac{1}{p(\mathbf{b}_k|\mathbf{F}_k, \mathbf{b}_{k-1}, \mathbf{F}_{k-1}, \ldots, \mathbf{b}_0, \mathbf{F}_0)}$$

As $\mathbf{x}_k$ is a Markov state, the past and future measurements will be independent of each other given the knowledge of $\mathbf{x}_k$. Therefore,

$$p(\mathbf{b}_k|\mathbf{x}_k, \mathbf{F}_k, \mathbf{b}_{k-1}, \mathbf{F}_{k-1}, \ldots, \mathbf{b}_0, \mathbf{F}_0) = p(\mathbf{b}_k|\mathbf{x}_k, \mathbf{F}_k) \tag{3.6}$$

This simplifies the belief as follows.

$$\begin{aligned} Bel(\mathbf{x}_k) &= \eta p(\mathbf{b}_k|\mathbf{x}_k, \mathbf{F}_k) p(\mathbf{x}_k|\mathbf{F}_k, \mathbf{b}_{k-1}, \mathbf{F}_{k-1}, \ldots, \mathbf{b}_0, \mathbf{F}_0) \\ &= \eta p(\mathbf{b}_k|\mathbf{x}_k, \mathbf{F}_k) p(\mathbf{x}_k|\mathbf{F}_k, \mathbf{b}_{0:k-1}, \mathbf{F}_{0:k-1}) \end{aligned} \tag{3.7}$$

Applying the Chapman-Kolmogorov equation to predict the transition density, we obtain the following expression.

$$\begin{aligned} Bel(\mathbf{x}_k) = \eta p(\mathbf{b}_k|\mathbf{x}_k, \mathbf{F}_k) \int p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{F}_k, \mathbf{b}_{0:k-1}, \mathbf{F}_{0:k-1}) \\ p(\mathbf{x}_{k-1}|\mathbf{F}_k, \mathbf{b}_{0:k-1}, \mathbf{F}_{0:k-1}) d\mathbf{x}_{k-1} \end{aligned} \tag{3.8}$$

The assumption of Markov state suggests that the knowledge of the immediate past state $\mathbf{x}_{k-1}$ and the current measurement $\mathbf{F}_k$ renders the current state $\mathbf{x}_k$ independent of all other previous measurements. This simplifies the belief as follows.

$$Bel(\mathbf{x}_k) = \eta p(\mathbf{b}_k|\mathbf{x}_k, \mathbf{F}_k) \int p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{F}_k) p(\mathbf{x}_{k-1}|\mathbf{F}_k, \mathbf{b}_{0:k-1}, \mathbf{F}_{0:k-1}) d\mathbf{x}_{k-1} \tag{3.9}$$

Again, the current measurement $\mathbf{F}_k$ does not alter our knowledge about the previous system state $\mathbf{x}_{k-1}$. Therefore

$$Bel(\mathbf{x}_k) = \eta p(\mathbf{b}_k|\mathbf{x}_k, \mathbf{F}_k) \int p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{F}_k) p(\mathbf{x}_{k-1}|\mathbf{b}_{0:k-1}, \mathbf{F}_{0:k-1}) d\mathbf{x}_{k-1} \tag{3.10}$$

The final expression for recursive state estimation is

$$Bel(\mathbf{x}_k) = \eta p(\mathbf{b}_k|\mathbf{x}_k, \mathbf{F}_k) \int p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{F}_k) Bel(\mathbf{x}_{k-1}) d\mathbf{x}_{k-1} \tag{3.11}$$

Equation (3.11) is a formal description of BC-based overt visual attention for a robot. Evaluation of (3.11) is computationally expensive due to the continuous nature of the head-pose space. Knowledge of the three following distributions are required for implementation of (3.11):

1. the initial belief $Bel(\mathbf{x}_0)$,
2. the transition probability $p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{F}_k)$, and
3. the measurement likelihood $p(\mathbf{b}_k|\mathbf{x}_k, \mathbf{F}_k)$.

The shape of these distributions are defined in the context of robotic overt attention.

**1.** $Bel(\mathbf{x}_0)$**:** The distribution $Bel(\mathbf{x}_{k-1})$ makes the state estimation recursive. Knowledge of this distribution at $k = 1$ is required to implement the (3.11). We can assume, without violating the generality of the model, a known orientation for the camera-head (e.g. $\alpha = \alpha_0$ and $\beta = \beta_0$) when the robot is first turned-on for an attention experiment. This makes $Bel(\mathbf{x}_0)$ a Dirac distribution located at $(\alpha_0, \beta_0)$ in the $(\alpha, \beta)$ space.

**2.** $p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{F}_k)$**:** This distribution describes the transition probability between two head-poses conditioned on the perceived visual features. In other words, if the visual features $\mathbf{F}_k$ are perceived at the head-pose $\mathbf{x}_{k-1}$, $p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{F}_k)$ evaluates the probability of the next head-pose of the robot such that the most salient feature in $\mathbf{F}_k$ is focused. Thus the distribution models the competition among head poses to focus on different visual features and hence termed as the **bottom-up competition model**.

**3.** $p(\mathbf{b}_k|\mathbf{x}_k, \mathbf{F}_k)$**:** This is the probability of receiving non-zero top-down bias from the memory if the robot takes the head-pose $\mathbf{x}_k$ and focuses on the stimuli with visual features $\mathbf{F}_k$. As the non-zero bias is assigned only for those head-poses which focus on the behaviorally relevant objects, the probability $p(\mathbf{b}_k|\mathbf{x}_k, \mathbf{F}_k)$, in other words, evaluates the degree of behavioral relevance of the object focused at $\mathbf{x}_k$. The probability $p(\mathbf{b}_k|\mathbf{x}_k, \mathbf{F}_k)$ is termed as the **top-down modulation model**.

The realization of the two models of Bayesian visual attention are discussed in the following section.

### 3.3.1 The Bottom-up Competition Model

The process of defining a shape for $p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{F}_k)$ is tricky as $\mathbf{F}_k$ and $\mathbf{x}_k$ are expressed in two different coordinate systems. To understand the process, let consider a hypothetical

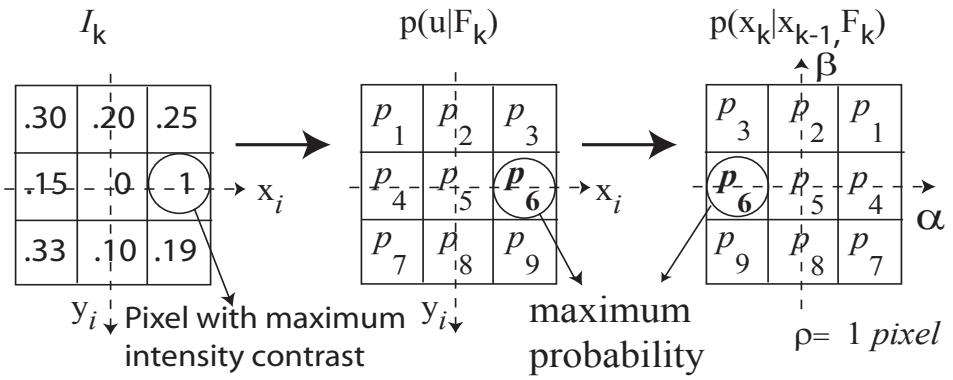
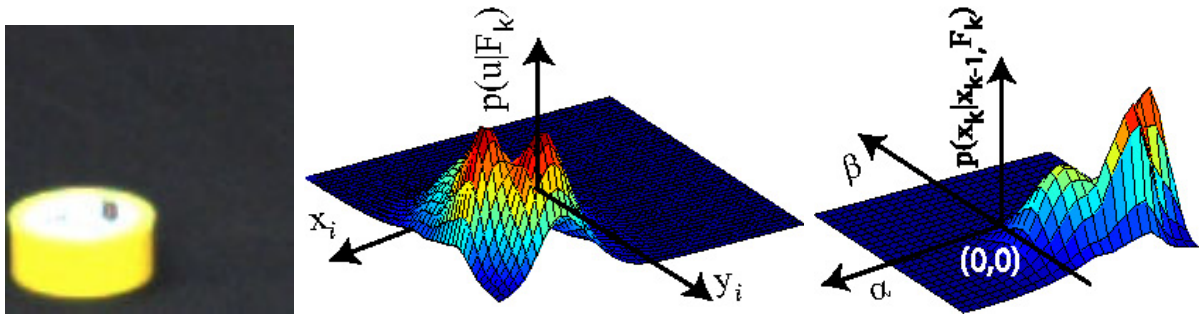



Figure 3.3: Evaluation of the bottom-up competition model for a hypothetical $3 \times 3$ camera image. The numbers in $I_k$ represent the intensity values of different pixels and $p_1, \ldots, p_9$ indicate their probabilities. Here, $\rho = 1$ (please see text for detail)

$3 \times 3$ image frame $I_k$ (as shown in Fig. 3.3) which is perceived at a head-pose $\mathbf{x}_{k-1}$. The numbers shown in $I_k$ indicate the intensity values of different pixels. The focus of attention is at the center of the frame, i.e the $(0, 0)$-$th$ pixel. Attending to any other point in the image means performing a head movement such that the point becomes the center of the frame. For a stationary robotic head as presented in this paper, there exists a one-to-one relation between an image point and a head-pose at which that point is focused. Such a one-to-one relationship does not hold good for a mobile camera-head or for a system with multiple camera-heads. In that case the relation between an image point and a head-pose can be retrieved by mapping the image point to the camera-heads' ego-centric frame of reference. For the current setting, focusing a point along the positive $x_i$ axis requires a negative (or clockwise) pan ($\alpha$) movement of the camera-head while focusing a point along the positive $y_i$ axis requires a positive (or anticlockwise) tilt ($\beta$) movement. A parameter $\rho$ defines the resolution of the head-pose space in terms of the number of image pixel. The value of $\rho$ depends on the 'angle of view' of the camera optics and the image size. The probability of a head-pose is proportional to the visual saliency of the image point it focuses on. For instance, the $(1, 0)$-$th$ pixel in $I_k$ (Fig. 3.3) has the highest intensity contrast and is calculated as the most probable point (with a probability value $p_6$) to focus on. Corresponding head-pose in the $(\alpha, \beta)$ space will have the most high probability among the possible nine head poses associated with the $3 \times 3$ frame.

For full scale images a mixture of Gaussian distribution $p(\mathbf{u}|\mathbf{F}_k)$ is used to represent the probability of different image points ($\mathbf{u} = \{u, v\}$) to be attended based on their visual features. In case of practical implementation the visual features are more meaningful when they are evaluated for a neighboring region than for a single isolated pixel. Each image frame, therefore, is divided into $\epsilon \times \epsilon$ sub-image blocks. Each sub-image block




(a) (b) (c)

Figure 3.4: (a) A $50 \times 50$ pixels image (b) The mixture of Gaussian $p(\mathbf{u}|\mathbf{F}_k)$ (c) The bottom-up competition model $p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{F}_k)$ with $\rho = 2$

holds one component Gaussian whose mean is located at the point of the highest intensity contrast and variance is calculated based on the color and intensity variance within that block. The distribution $p(\mathbf{u}|\mathbf{F}_k)$, therefore, acts as a scaled reflection of the distribution $p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{F}_k)$ with an scaling factor of $\rho$. Fig. 3.4 shows $p(\mathbf{u}|\mathbf{F}_k)$ and $p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{F}_k)$ for a $50 \times 50$ camera frame (with $\epsilon = 5$ and $\rho = 2$). The qualitative property of the distribution $p(\mathbf{u}_k|\mathbf{F}_k)$ indicates that it is a probabilistic generalization of the traditional 'saliency map' introduced in [7].

A notable characteristic of the bottom-up competition model is it performs space-based analysis of image features to identify a set of potential head poses. The contrast in color and intensity in different regions in the image are investigated without explicitly forming the notion of an object.

### 3.3.2 The Top-down Modulation Model

The top-down modulation model acts on the objects and evaluates their behavioral relevance with the current context of the robot. The top-down modulation model, therefore, operates with the assumption that a set of objects has been extracted from the potentially interesting regions identified by the bottom up competition model. The object segmentation process will be discussed in the next section along with the particle filter implementation of (3.11).

As discussed in section 3.1, two types of behavioral relevance of the objects are considered during top-down biasing. The objects with novel visual features are considered as 'worthy to attend' and receive excitatory top-down bias from the LTM. The objects having similar visual features as a 'sought for' object is considered as 'worthy to attend' and receive excitatory top-down bias from the WM. Establishing behavioral relevance for

attentional selection in these ways has strong biological evidence [34, 136]. Accordingly,

$$p(\mathbf{b}_k|\mathbf{x}_k, \mathbf{F}_k) = \lambda(\ p(\mathbf{b}_k^{LTM}|\mathbf{x}_k, \mathbf{F}_k) + p(\mathbf{b}_k^{WM}|\mathbf{x}_k, \mathbf{F}_k)) \tag{3.12}$$

Here $\lambda$ is a parameter which manipulates the importance of a head-poses to implement the dynamic IOR. The process of choosing a value for $\lambda$ will be discussed later in this chapter. The behavioral relevance of different objects are investigated through comparing the SIFT keypoints of the objects with that stored in the LTM and WM. The LTM $\mathbf{M}_k$ contains the SIFT keypoints of the objects attended throughout the life-time of the robot and is used when evaluating bias for the novel objects. The WM $\mathbf{m}_k$ contains the SIFT keypoints of an object which is being searched by the robot and is used when evaluating bias for the target-like objects. The top-down biases for an object which is focused at $\mathbf{x}_k$ and has the SIFT keypoints $\mathbf{F}_k$ is evaluated as follows.

$$p(\mathbf{b}_k^{LTM}|\mathbf{x}_k, \mathbf{F}_k) = 1 - p(O|f_{match}) \tag{3.13}$$

$$p(\mathbf{b}_k^{WM}|\mathbf{x}_k, \mathbf{F}_k) = p(O|f_{match}) \tag{3.14}$$

Here $O$ represents the hypothesis that the object has been observed before and $f_{match}$ is a set of $Q$ keypoints from $\mathbf{M}_k$ (in case of (3.13)) or $\mathbf{m}_k$ (in case of (3.14)) that matches with the $Q$ keypoints from $\mathbf{F}_k$. $p(O|f_{match})$, therefore, is the probability that the set of keypoints $f_{match}$ is a true match of the candidate object. For evaluation of $p(O|f_{match})$, a set of $\acute{Q}$ keypoints $\acute{f}_{match}$ is identified in the memory ($\mathbf{M}_k$ or $\mathbf{m}_k$, based on the context) as the nearest neighbors of the $\acute{Q}$ keypoints from $\mathbf{F}_k$ based on minimum Euclidean distance between keypoint descriptor vectors [140] ($\acute{Q} \leq Q$). For each pair of the matched keypoints the probability that the match is correct is evaluated based on the following three measures.

- Position constraint: Position constraint $P_{xy}$ determines how far the location of the candidate object's keypoint is from the location of its matched keypoint in the memory. Assuming that a 20% change in size in each direction is acceptable, $P_{xy}$ is the probability that the candidate object's keypoint satisfies this criteria. The probability is modeled using a Gaussian distribution with the mean located at the position corresponding to the matched keypoint in the memory and the standard deviation of 0.2.

- Scale constraint: The scale constraint $\chi$ is the probability that the scale of the candidate object's keypoint is within a close proximity of that of the matched database keyopoint. The probability is modeled using a Gaussian distribution with the mean equal to the scale of the matched keypoint and the standard deviation of 0.5 (the same value has been suggested for pattern matching in [141]).

- Orientation constraint: A 30% change in orientation between the object keypoint and the database match is considered acceptable. Accordingly, the orientation constraint

$\psi$ represents the probability that the orientation of the object keypoint satisfies this criteria. The probability is modeled using a Gaussian distribution with the mean located at the orientation of the database keypoint and the standard deviation is $30/360 = 0.085$.

The keypoints for which the probabilities $P_{xy}, \chi$, and $\psi$ are less than a certain threshold are discarded from the set $\acute{f}_{match}$. The remaining $Q$ keypoints constitute the set $f_{match}$ whose probability of being the true match of a candidate object is calculated as follows.

$$p(O|f_{match}) = \prod_{q=1}^{Q} P_{xy_q}\psi_q\chi_q \tag{3.15}$$

The next section describes the implementation of Bayesian model of attention using particle filter algorithm.

## 3.4 The Particle Filter Implementation

Evaluating the full distribution of (3.11) is a computationally expensive process. Particle filter [142] provides a time-efficient solution and has been used to implement (3.11). The particle filter represents the *belief* by a set of $L$ weighted samples who are distributed according to the original distribution. Accordingly, from (3.11)

$$Bel(\mathbf{x}_k) : \{\mathbf{x}_k^{(l)}, w_k^{(l)}, l = 1, \ldots, L\} \tag{3.16}$$

Here, each sample $\mathbf{x}_k^{(l)}$ represents a system state and $w_k^{(l)}$ is a non-negative number serving as the importance weight of the sample. The weights sum up to unity. In case of robotic visual attention where we can assume a full knowledge of the initial belief $Bel(\mathbf{x}_0)$, the $L$ samples (each with weight $\frac{1}{L}$) representing the $Bel(\mathbf{x}_0)$ are densely populated around the very first head-pose of the robot. At any time $k > 1$, the recursive attention equation (3.11) is realized through the following four steps.

**Step 1**. **Sampling from the prior belief:** A set of $L$ samples $\{\mathbf{x}_k^{\prime(l)}, l = 1, \ldots, L\}$ is collected from $Bel(\mathbf{x}_{k-1})$ according to the importance weights $\{w_{k-1}^{(l)}\}$, where $Bel(\mathbf{x}_{k-1}) : \{\mathbf{x}_{k-1}^{(l)}, w_{k-1}^{(l)}, l = 1, \ldots, L\}$.

**Step 2**. **Prediction through bottom-up competition model:** The bottom-up competition model $p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{F}_k)$ is used to generate a sample set $\{\mathbf{x}_k^{(l)}, l = 1, \ldots, L\}$ as a sample-based prediction of the current state.

$$\mathbf{x}_k^{(l)} \sim p(\mathbf{x}_k|\mathbf{x}_k^{\prime(l)}, \mathbf{F}_k) \tag{3.17}$$

**Step 3**. **Measurement update through top-down modulation model:** An importance weight $w_k^{(l)}$ is calculated for each sample in $\{\mathbf{x}_k^{(l)}\}$ according to (3.12)- (3.15).

$$w_k^{(l)} = p(\mathbf{b}_k|\mathbf{x}_k^{(l)}, \mathbf{F}_k) \tag{3.18}$$

The weights are normalized to form $\sum_l w_k^{(l)} = 1$ so that the samples constitute a probability distribution. This makes the weighted sample set $\{\mathbf{x}_k^{(l)}, w_k^{(l)}, l = 1, \ldots, L\}$ an approximate representation of the distribution in (3.11).

**Step 4**. **Reporting the current state** The sample with maximum weight is reported as the current head-pose of the robot. Taking this head-pose places the corresponding object at the center of the VF of the robot.

Further clarifications are required for **Steps 2** to **4** and are provided in the following sections. A set of experimental data are used to clarify these steps. The camera images used are of dimension $600 \times 600$ pixels and the value of the parameter $\rho$ is 0.06.

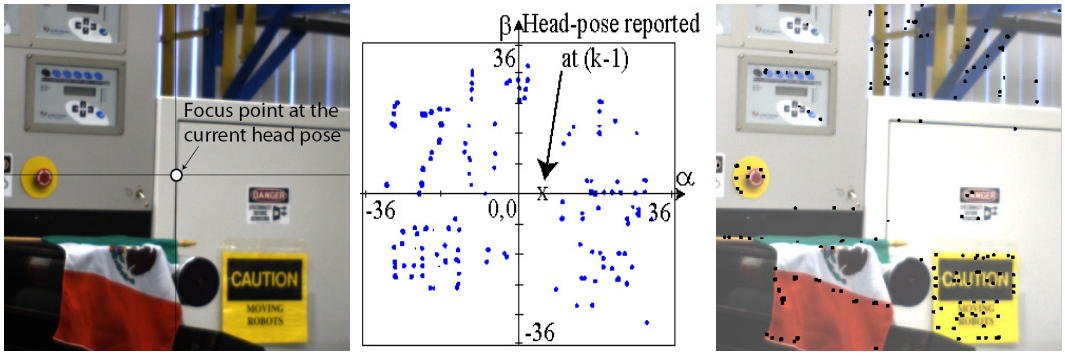
### 3.4.1 Prediction Through Bottom-up Competition Model

During the prediction stage a set of $L$ poses are predicted as the next head-pose of the robot. The sample set constructed in **step 1** along with the bottom-up competition model is used to make this prediction. For instance, the predicted head-pose $\mathbf{x}_k^{(l=1)}$ is a sample collected from the bottom-up competition model $p(\mathbf{x}_k|\mathbf{x}_k'^{(l=1)}, \mathbf{F}_k)$. To maintain diversity in the sample set, a sample head-pose which has been selected more than 10 times is prohibited from being selected again. This reduces the possibility of degeneracy, a common problem in particle filter [143].

Fig. 3.5 graphically demonstrates the prediction stage. Let us assume that the head-pose reported at $(k-1)$ focuses on the center of the frame shown in Fig. 3.5(a). Fig. 3.5(b) shows the set of samples drawn according to the *bottom-up competition model* using (3.17). The image points that will be focused at these predicted head poses are shown by black dots in Fig. 3.5(c). The points in the image plane corresponding to a predicted head-pose set are denoted by $\{\mathbf{U}_k^{(l)}\}$. These image points play a key role in object segmentation.

### 3.4.2 Measurement Update Through Top-down Modulation Model

During this stage each sample in $\{\mathbf{x}_k^{(l)}\}$ is assigned with an importance weight $w_k^{(l)}$. The weight of a sample head-pose is determined based on the behavioral relevance of the object to which the corresponding focus points in the image belongs. Object segmentation and SIFT keypoint extraction from the segmented objects, therefore, play a key role in this stage.




(a) (b) (c)

Figure 3.5: Graphical demonstration of the prediction stage in the particle filter for visual attention (a) The reported head-pose at time $(k-1)$ focuses on the center of the visual field (b) For prediction of $\mathbf{x}_k$ the head-pose samples drawn according to the bottom-up competition model (c) The points in the image that will be focused at the sample head-poses shown in (b) (please see text for detail)

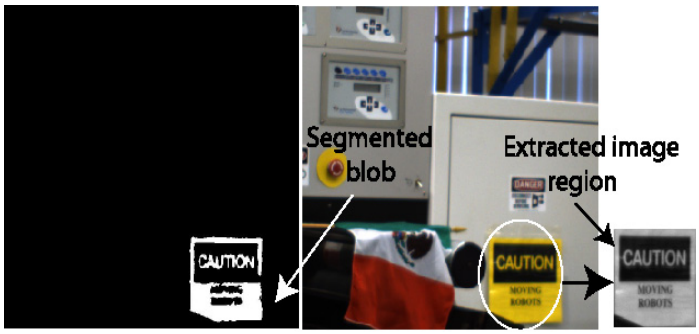
### Object segmentation and SIFT keypoints extraction

The set of image points $\{\mathbf{U}_k^{(l)}\}$ is used for object segmentation. Due to its process of construction (described in section 3.4.1) these points are located at the regions of the image which are visually salient due to contrast in color and/or intensity. In order to segment the objects underlying these regions the points in $\{\mathbf{U}_k^{(l)}\}$ are used as the 'seed' of a region growing algorithm. For better success in object segmentation the region growing algorithm is not directly applied on the raw camera image. Rather, the RGB image from the camera is preprocessed through the following steps prior to applying the region growing algorithm.

- The camera image $I_k$ is transformed to the *YCrCb* color space. Four images are created from the YCrCb image, one for the intensity ($I_k^{\mathrm{Y}}$) and three for the three distinct colors red ($I_k^{\mathrm{r}}$), green ($I_k^{\mathrm{g}}$), and blue ($I_k^{\mathrm{b}}$).

- A pyramid-based image segmentation technique is performed on $I_k^{\mathrm{r}}$, $I_k^{\mathrm{g}}$, and $I_k^{\mathrm{b}}$ (using the pyramid-based image segmentation algorithm implemented in the *open source computer vision library*). Three levels are used for the pyramid segmentation. There are two parameters involved with the pyramid segmentation process, $\zeta_1$ (the threshold value used to define connectivity among different pixels) and $\zeta_2$ (the threshold value used to merge the connected components into different clusters). The values used for these two parameters in the experiments described in this thesis are 12 and 10,


(a) (b) (c)




(d)

Figure 3.6: Object segmentation process corresponding to the image in Fig. 3.5(a). The images after the pyramid-based image segmentation (a) $I_k^{r'}$ (b) $I_k^{g'}$ (c) $I_k^{b'}$ (d) The segmented object blob corresponding to the head-pose ($\alpha = -9^o, \beta = -20^o$) and the image region considered for SIFT keypoint extraction based on the bounding rectangle of the segmented blob

respectively. These values are chosen on a trial and error basis after analyzing 20 indoor images taken with two different cameras having different angular resolutions.

- The images $I_k^{r'}$, $I_k^{g'}$, and $I_k^{b'}$, obtained after pyramid-based segmentation, are used to apply the region growing algorithm.
- The intensity image $I_k^{Y}$ is used for extracting SIFT keypoints.

The result of the pyramid-based segmentation for the image shown in Fig. 3.5(a) are shown in Fig. 3.6(a)- 3.6(c)

The segmentation of the object focused at a sample head-pose works as described below.

1. For each image point (which is focused at a sample head-pose) the mean values of the pixels within a $3 \times 3$ neighborhood is calculated for the $I_k^r$, $I_k^g$, and $I_k^b$. The image

which yields the maximum mean value is considered to be the dominant image and the corresponding segmented image is considered for further processing (for instance, if the mean value corresponding to $I_k^{\mathrm{r}}$ is the highest then $I_k^{\mathrm{r}'}$ is selected).

2. Region growing is performed on the segmented image considering the focused pixel as the 'seed'. All pixels that differs in value at most $\pm 5\%$ from the 'seed' pixel are merged together.

3. The validity of the blob obtained through region growing is analyzed based on its size. For instance, if the extracted blob occupies most of the scene (the dimension of the bounding rectangle of the blob is more than 50% of the dimension of the image), it is likely that the blob corresponds to the background. Similarly, too small blob is discarded as outliers (the dimension of the bounding rectangle is less than 4% of the dimension of the image).

4. A thumb rule is applied on an extracted blob to investigate if it is a part of an object which has been extracted earlier using another seed point: if more than 70% of an extracted blob lies within the bounding rectangle of another blob, they are merged together and is considered as a single object.

5. The image region within the bounding rectangle of the segmented object blob is considered to be a part of that object. An object blob might include several points from the point set $\{\mathbf{U}_k^{(l)}\}$. In other words, there might be more than one head poses which focus on the same object blob. In such case, the region growing algorithm will run once and the segmented blob will be used for assigning weights to all of the head poses that focus on it. Accordingly, the head poses that focus on the same object blob will receive similar weight.

The region growing from only one pixel generally can not fully segment an object if it has complex texture or have very large body with different colors. But in any case, it is likely that there will be some other head-poses that focus on other parts of the object. It will cause those parts to be segmented through region growing with different 'seed(s)'. In the worst case, different blobs from the same object might not be merged together. At the behavioral level, it is, therefore, possible that different parts of a large object will be attended separately. Fig. 3.6(d) demonstrates a case of object segmentation corresponding to five sample head-pose.

The SIFT filter is applied on the region of the image residing within the bounding rectangle of each extracted object blob. Fig. 3.7 shows the regions of the image (corresponding to different head-poses) selected for collecting SIFT keypoints.

Once an object corresponding to a sample head-pose is segmented, the SIFT keypoints associated with it are analyzed for evaluating a weight for the sample using equation (3.12) to (3.15). Fig. 3.8(a) shows the sample set of Fig. 3.5(b) after weight assignment.

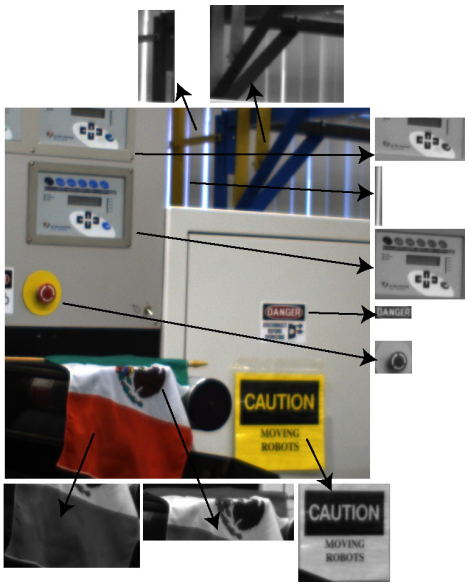
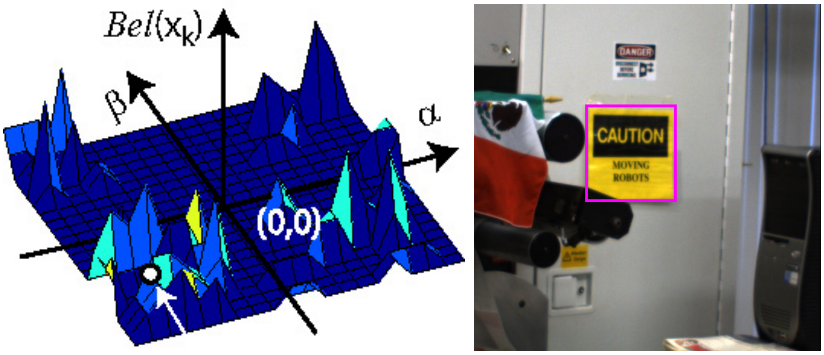



Figure 3.7: The regions of the image selected for SIFT keypoints extraction based on the predicted head-pose samples shown in Fig. 3.5(b)




(a) (b)

Figure 3.8: (a) The weighted samples representing the posterior for attention. The particle with highest weight (marked by the white circle) is reported as the next head-pose of the robot (b) The object (marked by a rectangle) focused at the new head-pose (please see text for detail)

### 3.4.3 Reporting the Current State

It is required to obtain an estimate of the current system state from the posterior distribution. For multi-modal posterior distribution the most commonly used measure of system state is the sample with maximum weight. The posterior for visual attention is generally multi-modal but might not have one unique maximum. In the simplest case of having one unique maximum, the head-pose sample that corresponds to that maximum is reported as the next head-pose $\mathbf{x}_k$ of the robot. Fig. 3.8(a) shows such a scenario where the sample with unique maximum weight is marked. When the robot takes this head-pose, the associated object resides at the center of the VF as shown in Fig. 3.8(b).

Reporting a system state is relatively complex when the posterior has several equal-amplitude peaks. This might occur when the robot observes more than one novel object (during a visual exploration) or several copies of a target (during a visual search) in its VF. Inspired by the characteristics of the primates' visual attention two different strategies are proposed here to resolve this confusion and choose a sample to report as the next head-pose of the robot.

- Relative proximity-based analysis: In this case, among the samples with equally high probabilities the one closest to the previous system state $\mathbf{x}_{k-1}$ is chosen as the next state of the system. This strategy of favoring the closest sample is inspired by the *proximity effect* in visual attention [7]. Some recent studies on visual attention, however, suggest that there is no *proximity effect* in the selective attention of the primates [144].

- Relative saliency-based analysis: In this case the objects associated with the samples with equally high probabilities are further analyzed to evaluate their relative visual saliency. The head-pose for which the associated object has the highest relative visual saliency is chosen as the next state of the system. Relative visual saliency is used as the major criteria (sometimes the only criteria) of visual attention in most of the existing computer vision models of visual attention [18, 19, 22, 23]. A number of methods, therefore, are available in the existing literature to evaluate the relative saliency of different objects in an image. Two very popular method in this category are NVT [18] and VOCUS [23]. Any of the existing algorithms can be used in this purpose.

When there is no novel stimulus in the VF (in case of visual exploration) or none of the objects have similarities with the target (in case of visual search), the posterior essentially becomes uniform. In such a case the robot remains at its current head-pose.

In case of visual search, all head-poses whose weights are within the 5% of the maximum weight are considered as the next head-pose of the robot. Relative proximity- or relative saliency-based analysis is performed to chose a sequence of attention.

## 3.5 Dealing with the Research Issues

The way the proposed model deals with the research issues involved with robotic visual attention are discussed in this section.

### 3.5.1 Integrated Space- and Object-based Analysis

The proposed model integrates the space- and object-based analysis while identifying the next head-pose of the robot (*Issue 2* in section 1.2). The bottom-up competition model operates on the space and identify the regions in the VF worthy to attend. This is done without forming any concept of object. The top-down modulation model, on the other hand, operates on the objects. As a result, only those regions of the VF are attended, among all of the 'attention worthy' regions selected by the bottom-up competition model, which are occupied by behaviorally relevant objects. This integrated analysis assists to resolve the problem of dynamic IOR and will be discussed later in this section.

### 3.5.2 Change of Reference Frame

Once the head-pose $\mathbf{x}_k$ is estimated from $Bel(\mathbf{x}_k)$, the robot executes the required pan-tilt movements in order to be at that pose. Accordingly, the content of the VF as well as the image coordinate system change. This changes a traditional saliency map and requires a re-mapping of the saliency information to the new image coordinate (discussed as *Issue 1.1* in section 1.2). As the particle filter of attention operates in the head-pose space, unlike an image-centric saliency map, the posterior for visual attention does not require any such re-mapping. The system state merely shift to a new location in the $(\alpha, \beta)$ space. This is a significant beauty of the robot-centric approach of visual attention. Some of the particles in the old posterior, however, might represent probabilities corresponding to objects which are not visible from the new system state. The prediction of next head-pose uses this old posterior as the prior information (**step 1** in section 3.4) and thereby rendering the visual attention of the robot as a process which has memory. This is also consistent with the attention system of the primates which is not believed to be a memoryless process [145].

### 3.5.3 Dynamic IOR and the Value of the Parameter $\lambda$

The robot-centric approach of the proposed model as well as the integration of space- and object-based analysis offer an interesting way of implementing dynamic IOR (*Issue 1.2* in section 1.2). The proposed model operates on the head poses and a head-pose always focuses on a space. As mentioned in section 3.5.2, the head-poses are invariant to the change

in image coordinate or the content of the VF. Implementing the space-based dynamic IOR, therefore, is as simple as inhibiting the previous head-pose (and its close neighborhood) in the $(\alpha, \beta)$ space from being the next state of the robot. This automatically prevents the robot to focus on the space that was attended in the immediate past, irrespective of the camera and image coordinates. But the space-based IOR, as mentioned in section 1.2, might introduce some unexpected delay in visual search. The proposed model, therefore, also includes the object-based analysis while implementing the IOR. The space which is focused at any head-pose might be occupied by an object or might simply belongs to the background. A head-pose, therefore, always associates a space in the 3D world but might or might not associate an object. Again, the object occupying a space might be familiar or novel, target or non-target. The proposed model selects a space to focus on based on the properties of the object lying in that space. This is mostly achieved by choosing a value for the parameter $\lambda$ in (3.12).

$$\lambda = \begin{cases} 0.01 & \text{if } s_1 \\ 2 & \text{if } s_2 \text{ or} s_1 \wedge s_2 \\ 1 & \text{if } s_3 \text{ or} s_1 \wedge s_3 \end{cases} \tag{3.19}$$

Here,

- $s_1$ : The head-pose is inhibited.
- $s_2$ : The head-pose focuses on a 'sought for' object.
- $s_3$ : The head-pose focuses on a novel object.
- $\wedge$ : The logical *and* operator.

The problem associated with the object-based IOR, as discussed in section 1.2, still exists in the proposed model. Because of that the robot might re-attend to an object when there is large change in the object's orientation, camera perspective, or illumination.

### 3.5.4 Partial Appearance of Features

In case of the SIFT keypoint-based object identification, the robot can successfully recognize an object as long as a certain number of matching keypoints can be extracted from the visible region of the object. The head-pose(s) at which a focused object is partially visible with dimension less than a certain threshold (the area is less than 2500 pixels) are assigned a predefined small weight during the top-down modulation.

The discussion in section 1.2 on the issues involved with robotic visual attention enlists three other research issues namely, optimal learning strategy, generality, and prior training.

The Bayesian model of visual attention proposed in this chapter is not capable to address these three research issues. A multi-modal extension of the model will be introduced in chapter 5 which offers solutions to deal with these three research issues.

## 3.6 Conclusion

This chapter has described the proposed Bayesian model of visual attention. At first the biological motivation of the proposed model is explained along with a brief summarization of the BC hypothesis of visual attention. The Bayesian formulation for the robotic visual attention is then discussed in light of the BC hypothesis. The chapter also provides a detailed description of the particle filter implementation of the proposed Bayesian model. Finally, it sheds light on how the proposed model deals with the research issues of robotic visual attention.

The next chapter will describe a set of experiments to evaluate the performance of the proposed model.

# Chapter 4

# Performance Evaluation of the Proposed Visual Attention Model

Analysis of performance in different real-world scenarios is a crucial requirement to evaluate any technical model. There is a lot of research on visual attention modeling but unfortunately no generic method has been proposed to evaluate the performance of a technical model of visual attention. The visual attention models in computational neuroscience generally use the response data of the primates's visual cortex (collected through single-cell experiments, e.g fMRI, PET) as the ground truth and compare the response of a model with them. In case of technical models of attention no such ground truth is available. This chapter focuses on the performance evaluation of the proposed model of visual attention.

## 4.1 Evaluation Criteria

Performance evaluation/comparison is not a very common practice in the computational modeling of visual attention. The handful of computational models which consider performance evaluation mostly use the following metrics.

1. **Accuracy of visual search:** A simple method for performance evaluation of a technical model of visual attention is to measure how accurately the model can identify a target during visual search. Measure of visual search accuracy, therefore, has been adopted by many visual attention models for performance evaluation. For instance, the covert model in [23] proposes a metric named *hit number* which indicates the required number of focus transitions before landing on the target object. A *hit number* close to unity is indicative to a good model of visual attention with respect to visual

search. The model in [146] uses *target detection speed* as a measure of goodness of the model in case of visual search.

The metrics like *hit number* and *target detection speed* are highly biased by different factors, e.g., type of training images, similarity of the test and training images etc. Besides, they are only applicable for visual search, but not for the overall attention behavior.

2. **Reaction time:** Reaction time is the time required to identify the focus of attention in a given image. This is the most commonly used evaluation criteria for both overt and covert models of visual attention. In case of robotic overt attention the reaction time also includes the time required for the camera to orient to the stimuli of attention. Reaction time in covert models are generally smaller than the overt models. Again, the reaction time of the models which deal only with bottom-up cues is less than those which deal with both the top-down and the bottom-up cues. The famous covert model NVT [17] reported a maximum of 10 seconds reaction time for visual search task in a $640 \times 480$ pixels static image. The NVT [17] also reports a linear relationship among the reaction time, complexity of the image (number of distractors), and the image dimension. The extension of NVT in [92] reports reaction time of approximately 1 minute to precisely detect novel events in the video sequence. In case of processing pure bottom-up cues the reaction time of VOCUS [23] is approximately around 1 second to 1.5 seconds. The reaction time of the attention model with VOCUS like bottom-up processing has been minimized to 10 milliseconds to 20 milliseconds in [147] through using the processing power of the graphic processing unit (GPU). Several research works on task-specific visual search in constrained environments report reaction time in millisecond range through being selective in feature processing, e.g., the work in [117] reports 11 milliseconds reaction time to identify human face in the indoor environment while using only intensity feature. Thus, the reaction time is a metric of attention which can be biased by three major factors: mode of attention (overt or covert), type of application, and complexity of the image.

3. **Human eye tracking:** Eye tracking of the human subject is becoming a popular method to evaluate the overt models of visual attention. In this case the output of an attention model (i.e., the attentional scan paths) for a set of images/ video sequence is compared with the result of an eye tracking experiment performed on a group of humans using the same images/ video sequence. Commercially available eye-tracking systems are used to record the gaze patterns of the human. The comparison is trivial for the standard psychophysical test images (e.g., the first two images in Fig. 1.4) but extremely non-trivial for the natural images. The overt model in [148] performs an analysis to compare the gaze pattern of a robotic head with that

of a human in a specially designed simplistic type experimental environment. The model [148] reports significant similarity between the two gaze patterns. The covert model in [92] uses the same method of comparison, except it is performed off line on a video sequence.

To further enhance the use of human eye tracking as a way of evaluating the visual attention models, a number of works have been reported in the recent literature focusing on different strategies to compare human gaze data with the synthetic models of attention. For instance, the models in [149,150] advocate a location-based strategy where a synthetic model's performance is considered as satisfactory if it focuses on the same spatial location as the human. The work in [151] further introduces the condition of feature similarity and suggests that a synthetic model and a human gaze-pattern will be 'close' if they focus on similar features irrespective of their spatio-temporal locations. All of these models [149–151] require a prior training phase with a considerably large amount of human gaze data. This prior training stage does not lend the method of comparison with human gaze pattern to be used as an unbiased metric of visual attention performance. This is mainly because the way the human gaze pattern data is collected in eye-tracking experiments is very likely to differ (with respect to eye-tracking hardware, experimental set-up, subjects' age, experience, and preference) from researcher to researcher. In order to use the comparison with the human gaze pattern as a statistically correct metric of visual attention performance, we have to have a general database of gaze patterns of considerably large numbers of people from different age groups, professions, and in a wide variety of environments.

4. **Qualitative analysis:** Qualitative analysis of attention behavior is performed by the majority of the attention model. There are some models which perform analysis of their performance with respect to their self-defined objectives (the majority of the models fall under this category, e.g., [51, 107, 117, 131]) while others compare their performance with similar existing models (e.g., [23, 50])

Based on the analysis of the commonly used performance metrics, this thesis focuses on three metrics for performance evaluation of the proposed attention model.

1. Self evaluation
2. Consistency of decision
3. Robustness against parameter variation

They are discussed below.

### 4.1.1 Self Evaluation

Self-evaluation is the process of evaluating the performance of an attention model with respect to the properties/ characteristics expected to be observed in that model. In other words, every attention model is designed to fulfill some specific goal(s) and is expected to have certain characteristics. Self evaluation evaluates how well the goals are fulfilled and how accurately the characteristics are manifested in the real-time implementation of the model. Self-evaluation is performed through a set of experiments specially designed to focus on the goals of a model and the characteristics expected to be observed in it. If other existing models of attention have similar goals or characteristics then it is strongly desirable to perform a direct comparison with such a model with respect to that common goal or characteristic.

The Bayesian model of visual attention proposed in this thesis is dedicated for autonomous robots. The autonomous robot which is being operated by the proposed model is expected to exhibit the following two attention behaviors.

- The novel stimuli in the VF of the robot will always be attended. Accordingly, the visual exploration of the robot is driven by the preference for novelty.
- While conducting visual search, the stimuli in the VF which have target-like features will be attended by the robot. Visual search has higher priority over the visual exploration.

While operating with these two properties, the major goal of the proposed model is to address the research issues reported in section 1.2. Self-evaluation of the proposed model, therefore, is concerned about how well the proposed model maintains these two characteristics while simultaneously tackling the issues of robotic visual attention.

### 4.1.2 Consistency of Decision

When it comes to visual attention, humans are generally never consistent in their decision. For the same scene the attentional scan-path of different human differs based on the individual's habit, background, personal preference, current state of mind, environmental condition (lighting) etc. Resulting from a complex interaction among all of these criteria it is not unusual that the same person will choose a different scan-path for the same scene at two different points of time. This apparent 'uncertainty' is one aspect of human cognition that makes us so special. As we are not able to mimic the whole cognitive system of the human, we are not interested in mimic this kind of uncertainty of human attention in a technical system. Rather, we want to develop a technical model of attention which is

consistent in its decision of 'what to focus on' as much as possible subject to reasonable change in illumination, viewing perspective (more accurately, moderate amount of affine transformation), and relative spatial position of the stimuli. It is worth mentioning that change of viewing perspective and varying illumination are very common phenomena in the robotic overt attention.

Consistency of decision is a very important criterion for a technical model of visual attention in order to demonstrate different concepts and to analyze and compare performance among different attentional scenario. Consistency of decision, therefore, will be used as a evaluation criteria for the experiments for self-evaluation of the proposed model.

### 4.1.3 Robustness Against Parameter Variation

Any mathematical model of the natural systems contains several parameters. It is extremely important to focus on the behavior of the model subjected to variation in the parameters' values. This helps to define a safe operating zone for the model. The proposed model has the following tunable parameters.

- The number of particles $L$.
- The block size $\epsilon$ (related to the bottom-up competition model)
- Image processing parameters $\zeta_1, \zeta_2$

The performance of the proposed model will be investigated subjected to the variation in values of these parameters. Aside these four parameters, the effect of the size of the LTM on the visual attention performance will also be investigated.

## 4.2 Experimental Hardware

A set of experiments is performed for self-evaluation of performance of the proposed Bayesian model of visual attention. The experiments are categorized into three groups.

1. Experiments related to visual exploration
2. Experiments related to visual search
3. Experiments related to parameter variation

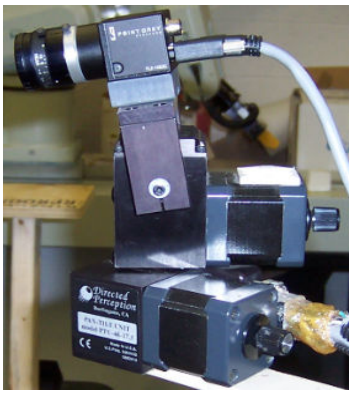
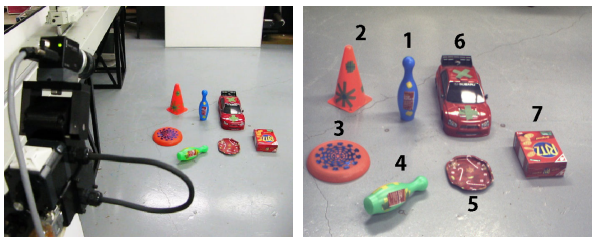


Figure 4.1: A *Point Grey Research* Flea2 color camera mounted on a *Directed Perception* PTU constitutes a robotic camera-head and is used during the experiments

Experimental hardware includes a *Point Grey Research* Flea2 color camera which is mounted on a *Directed Perception* PTU to serve as the camera-head of a robotic system (Fig. 4.1). The rest of this chapter will refer this camera-head as the 'robot'. The physical location of the robot in the 3D world does not change during the course of an experiment. Flea2 uses a narrow angle optics with the angle of view $38.47^o(H) \times 29.35^o(V)$. Image dimension is $640 \times 480$ pixels. The particle filter implementation of the Bayesian attention model uses 500 particles. Corresponding to one snap-shot of the environment the total time required for running the attention algorithm and executing the motion command by the PTU is considered as the time for one decision cycle. Each decision cycle begins with capturing an image at a new head-pose and ends with delivering the appropriate motion command to the PTU. A constant time delay of 1.5 seconds is introduced between two successive cycles in order to allow the mechanical movements of the PTU. The average time of a decision cycle in different experiments is found to be 5.6s with a 2GHz processor (including the constant time delay). No additional processing support (e.g., a GPU) is involved with this timing. It is worth to mention that currently the program code is not optimized and that code optimization will further reduce the time requirement of the algorithm.

## 4.3 Experiment 1: Visual Exploration

Two sets of experiments are performed to demonstrate the novelty-guided visual exploration characteristic of the proposed model. During the experiments the robot is exposed to different objects and it attends only to those objects which appear novel in its perception. The first set of experiments focuses on demonstrating the strategy of the proposed model to deal with multiple novel stimuli and the second set is focused on the demonstration of dynamic IOR.




(a) (b)

Figure 4.2: Novelty-guided visual exploration with relative proximity-based analysis (a) The experimental environment with seven novel objects (b) The VF of the robot at the first decision cycle. In the successive cycles the objects are attended in order of their proximity to the current focus of attention. The numbers denote the sequence of attention (please see text and Fig. 4.3 for detail)

### 4.3.1 Dealing with Competing Stimuli

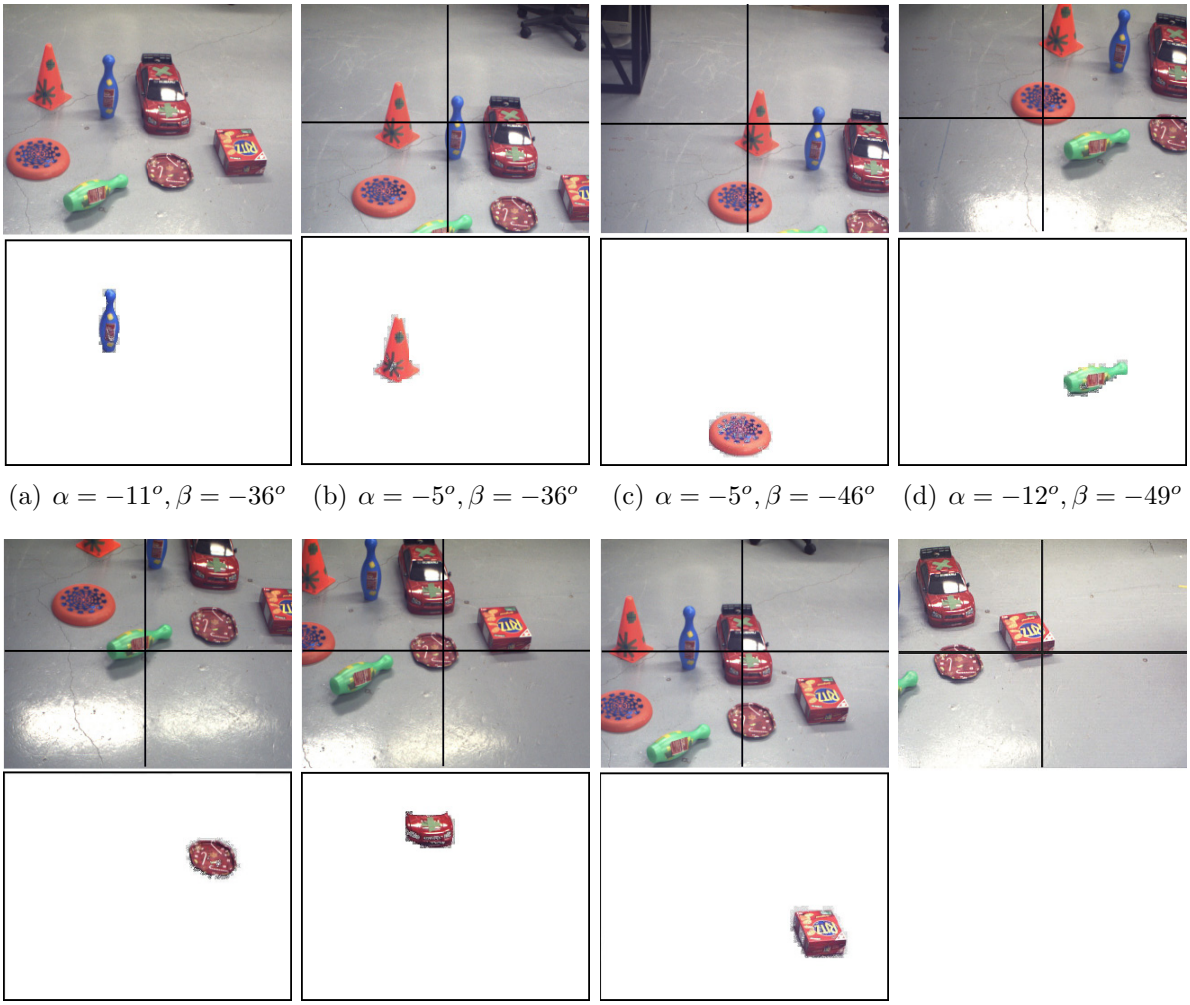
During this experiment the robot starts with an empty LTM. The robot is initially exposed to the environment shown in Fig. 4.2(a). There are seven objects in the VF (as shown in Fig. 4.2(b)), all of which are novel to the robot as it does not have any prior knowledge. This causes the posterior distribution to have multiple equal-amplitude peaks. To resolve the confusion arising from such competing novel stimuli the proposed model suggests two different strategies as discussed in section 3.4.3. The attention sequence resulting from each of these two strategies are discussed below.

#### Relative proximity-based analysis

The relative proximity-based analysis chooses the head pose, among a set of equally probable head-poses, which is the closest to the current head-pose as the next system state. Five trials of the same experiment are performed while using the proximity-based analysis to resolve the conflict among competing novel stimuli. During these five trials the relative positions of the objects with respect to the camera position are kept unchanged. In all of the trials the robot sequentially attends to the novel objects while following the sequence marked in Fig. 4.2(b). Fig. 4.3 shows different stages of the experiment.

Now the same experiment is performed with the same camera setting but some of the objects in the VF change their position as shown in Fig. 4.4(a). In all five trials of the experiment the attention sequence followed by the robot is shown in Fig. 4.4(b).

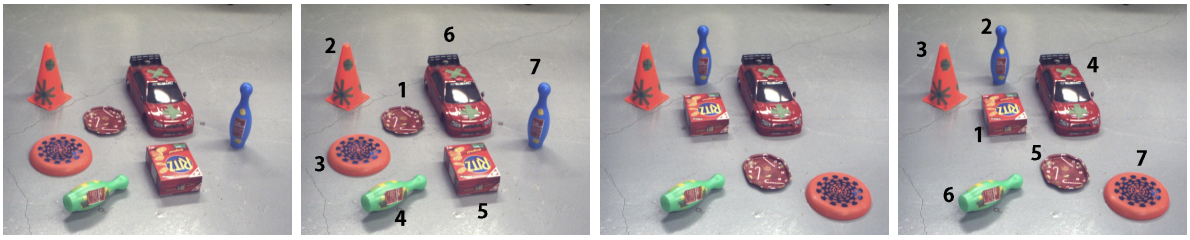
With further change in relative positions of the objects (as shown in Fig. 4.4(c)), the experiment is conducted again and the resulting attention sequence is shown in Fig. 4.4(d).



(a) $\alpha = -11^o, \beta = -36^o$ (b) $\alpha = -5^o, \beta = -36^o$ (c) $\alpha = -5^o, \beta = -46^o$ (d) $\alpha = -12^o, \beta = -49^o$

(e) $\alpha = -20^o, \beta = -46^o$ (f) $\alpha = -18^o, \beta = -39^o$ (g) $\alpha = -28^o, \beta = -44^o$ (h) $\alpha = -28^o, \beta = -44^o$

Figure 4.3: Different stages of the visual exploration experiment shown in Fig. 4.2. For each image the top row indicates the frame captured at the beginning of a decision cycle and the bottom row indicates the objects to be focused at the successive decision cycle based on the estimated head-pose $(\alpha, \beta)$ shown within parenthesis. After attending the object in (h), the robot can not identify any other novel stimulus in the VF and remains at the current state



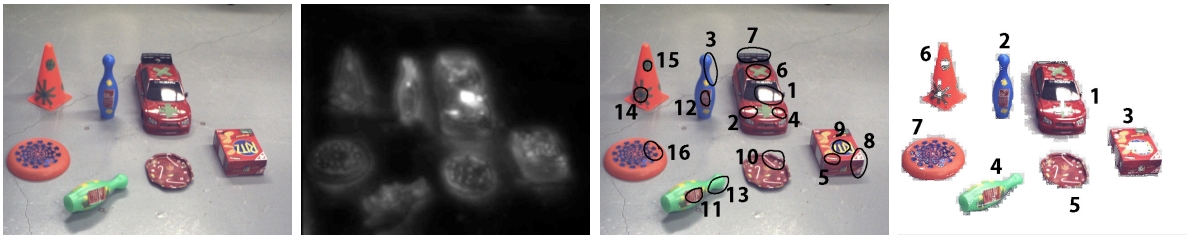

(a) (b) (c) (d)

Figure 4.4: Novelty guided visual exploration with relative proximity-based analysis: the novel objects change their relative positions as compared to the environment shown in Fig. 4.2(a). (a), (c) The VFs of the robot at the first decision cycle during the experiments with two different settings of the objects. The objects are attended sequentially in the subsequent decision cycles. The sequences of attention are shown in (b) and (d), respectively (please see text for detail)

Although the proximity-based analysis is applied on the head-poses in the pose-space, the distance between different objects visible in Fig. 4.2(a), 4.4(a), and 4.4(c) indicate that the sequences at which they are attended are highly consistent with the theme of the proximity-based analysis.

**Relative saliency-based analysis**

In case of relative saliency-based analysis, among a set of equally probable head-poses, the one focusing on the object with highest relative visual saliency is chosen as the next system state. The current experiments use VOCUS-like image processing [23] to evaluate the relative visual saliency of the objects in the VF that are associated with different equally probable head-poses. As the VOCUS algorithm performs space-based analysis of attention, a small modification of the original algorithm is performed to accommodate the object-based analysis of the current system. According to the original VOCUS algorithm several pixel-points from one object might be focused in order of their decreasing saliency, especially when the object is highly textured and can not be segmented through the simple color- or intensity-based region growing method [23]. Switching to a different focus point is achieved through application of space-based IOR. In the modified VOCUS algorithm, once a pixel-point is focused, the entire object to which the pixel belongs is inhibited for the next attention. The object blobs are extracted according to the object segmentation method described in section 3.4.2. An object, therefore, can be focused only once during a novelty exploration experiment. The most salient pixel of an object is the representative of the relative visual saliency of that object and, thereby, determines the position of the object in the attention sequence. For comparison purposes, the VOCUS-based saliency map of the



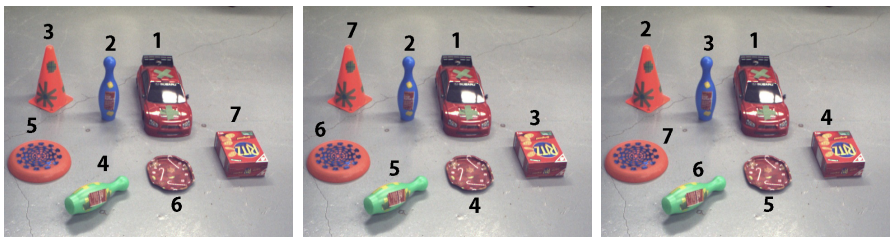

(a) (b) (c) (d)

Figure 4.5: Novelty guided visual exploration with relative saliency-based analysis (a) The visual field of the robot at the first decision cycle where all seven novel objects are visible. (b) The saliency map of the visual field generated using the VOCUS algorithm [23]. (c) The first sixteen focuses of attention suggested by the saliency map. (d) The first seven objects of attention evaluated according to the modified VOCUS algorithm (please see text for detail)

VF in Fig. 4.5(a) is shown in Fig. 4.5(b). The first sixteen focuses of attention, according to the original VOCUS algorithm, are marked in Fig. 4.5(c). The first seven focuses of attention, according to the modified VOCUS algorithm, are marked in Fig. 4.5(d).

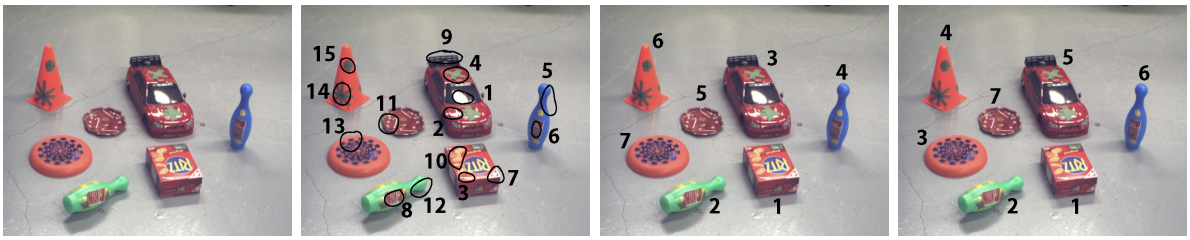
Five trials of the visual exploration experiment are performed while using the relative saliency-based analysis to resolve the conflict among competing novel stimuli. For the environment in Fig. 4.2(a) the robot executes three different attention sequences as shown in Fig. 4.6(a), 4.6(b), and 4.6(c). Completely different attention sequences are observed when the objects switch their positions. For instance, there are two different attention sequences for the environment in Fig. 4.4(a) and are shown in Figs. 4.7(c) and 4.7(d). For comparison purposes, the first sixteen focuses of attention for the same environment calculated according to the original VOCUS algorithm are marked in Fig. 4.7(b). Again, there are two different attention sequences for the environment in Fig. 4.4(c) and are shown in Figs. 4.8(c) and 4.8(d). For comparison purposes, the first sixteen focuses of attention for the same environment calculated according to the original VOCUS algorithm are marked in Fig. 4.8(b). The results from all of the relative proximity- and relative saliency- based analysis are summarized in Table 4.1.

Analysis of the results in Table 4.1 shows that although the relative saliency-based analysis is a more biologically-legitimate approach to deal with the competing stimuli, the resulting attention sequences are highly unstable. This is probably because of the fact that the saliency information never remains exactly the same in two images captured at two different points of time. The relative saliency has high degree of dependency on natural lighting variation, slight change in camera perspective, and even on the relative position of different objects. The relative proximity-based analysis, on the other hand, generates highly stable attention sequence in all conditions. This stability has significant importance



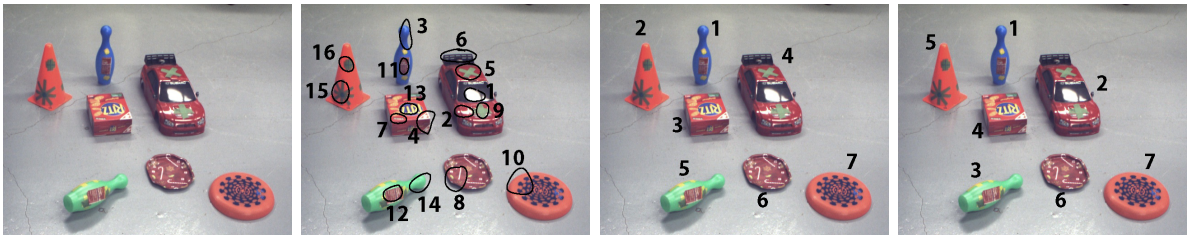

(a) (b) (c)

Figure 4.6: Novelty guided visual exploration with relative saliency-based analysis. The robot executes three different attention sequences for the environment in Fig. 4.2(a). The sequences are marked on the visual field of the robot at the first decision cycle (please see text for detail)




(a) (b) (c) (d)

Figure 4.7: Novelty guided visual exploration with relative saliency-based analysis: the novel objects change their relative positions as compared to the environment shown in Fig. 4.2(a). (a) The visual fields of the robot at the first decision cycle. (b) The first sixteen focuses of attention calculated according to the original VOCUS algorithm-based saliency map [23]. (c), (d) Two different attention sequences for the same environment resulting from using the modified VOCUS algorithm to resolve the conflict among the competing novel stimuli (please see text for detail)



(a) (b) (c) (d)

Figure 4.8: Novelty guided visual exploration with relative saliency-based analysis: the novel objects change their relative positions as compared to the environment shown in Fig. 4.2(a). (a) The visual fields of the robot at the first decision cycle. (b) The first sixteen focuses of attention calculated based on the VOCUS algorithm-based saliency map [23]. (c), (d) Two different attention sequences for the same environment resulting from using the modified VOCUS algorithm to resolve the conflict among the competing novel stimuli

Table 4.1: Novelty guided visual exploration: dealing with competing stimuli

| Environment | Relative-proximity | | Relative-saliency | |
|---|---|---|---|---|
| | Attention sequence | # of trial | Attention sequence | # of trial |
| Fig. 4.2(a) | Fig. 4.2(b) | 10 | Fig. 4.6(a) | 5 |
| | | | Fig. 4.6(b) | 6 |
| | | | Fig. 4.6(c) | 5 |
| Fig. 4.4(a) | Fig. 4.4(b) | 5 | Fig. 4.7(c) | 5 |
| | | | Fig. 4.7(d) | 3 |
| Fig. 4.4(c) | Fig. 4.4(d) | 5 | Fig. 4.8(c) | 3 |
| | | | Fig. 4.8(d) | 4 |

during demonstration of different attention-related phenomena as well as for performance analysis and comparison. The experiments described in the rest of this thesis, therefore, adopt the proximity-based analysis unless otherwise stated.

### 4.3.2 Demonstration of IOR

The goal of this experiment is to demonstrate the implementation of IOR in the proposed model. During the experiment the robot starts with an empty LTM and the camera is exposed to the scene of Fig. 4.9(a). There are seven novel objects in the VF. The robot focuses on each of them and updates the LTM with their associated SIFT keypoints. The sequence of attention is marked in Fig. 4.9(a). The VF lacks novel stimulus after the seventh object is attended (as shown in Fig. 4.9(b)). The robot never re-visits the locations

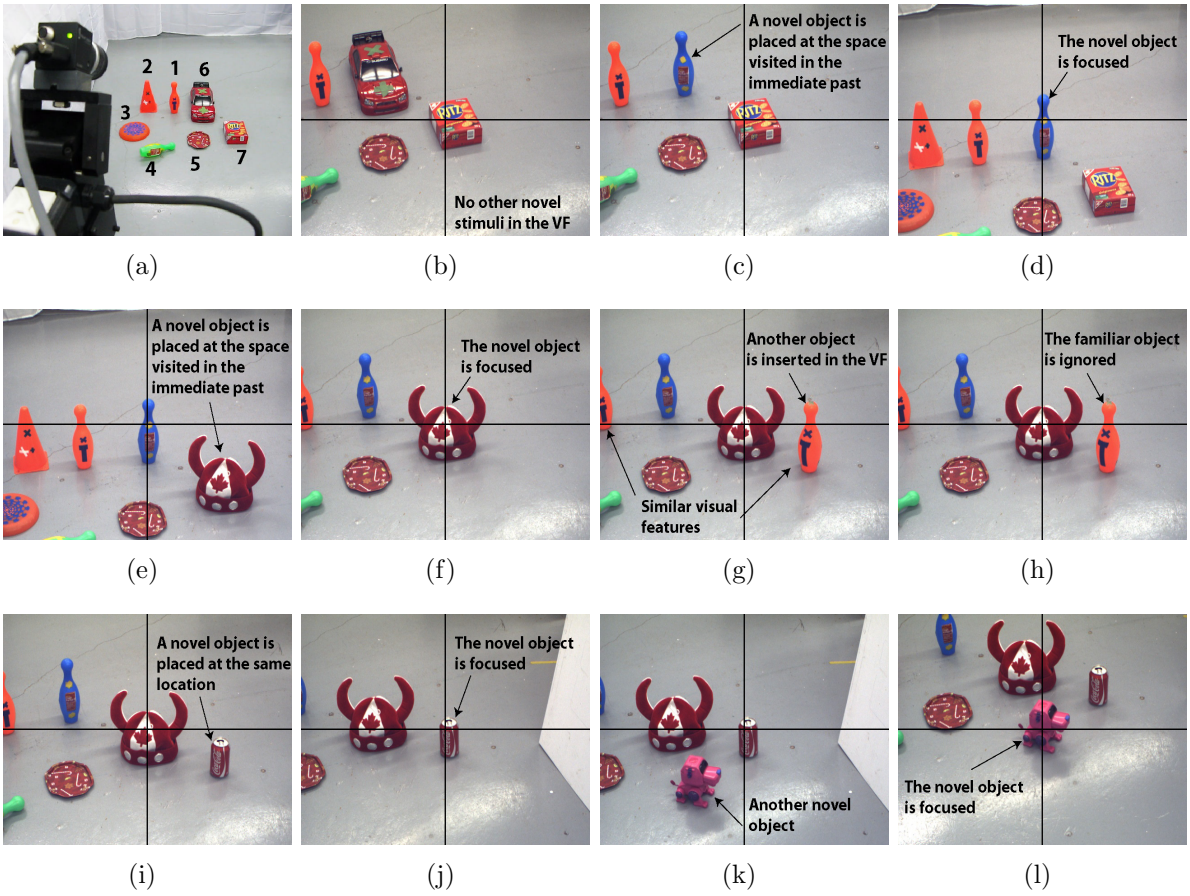



Figure 4.9: Demonstration of IOR: (a) The experimental environment and the sequence of attention. (b)-(j) Different stages of the experiment (please see the text and the multimedia file "Multimedia_IOR.wmv" for detail)

attended thus far as none of them is occupied by any novel stimulus. At this point a novel object is placed at the location which was visited at the immediate past (Fig. 4.9(c)). Although the last-visited head-pose is inhibited according to space-based IOR, the robot re-visits its neighborhood to focus on the novel object under the influence of object-based IOR (as shown in Fig. 4.9(d)). Similar reasoning applies for the attention behavior of the robot when another novel object is placed at the last visited location as shown in Fig. 4.9(e) and the robot focuses on it (Fig. 4.9(f)). At this point, another object is inserted in the VF as shown in Fig. 4.9(g). This newly inserted object has visual features similar to one of the previously attended objects (the first object in the attention sequence shown in Fig. 4.9(a)). Due to familiarity of its visual features the object does not get attention of the robot (as shown in Fig. 4.9(h)). When a novel object is inserted at the same place (Fig. 4.9(i)), the camera immediately identifies the novelty and focuses on it (Fig. 4.9(j)). The features of the object occupying a space make that space worthy to attend. Similar reasoning applies for the attention toward another novel object inserted in the VF as shown in Fig. 4.9(l). Thus, the property of the proposed model to prefer novel stimuli for attention inherently prevents the model from visiting previously visited locations, thereby implementing a form of space-based IOR. If, however, a novel object appears in any of the already visited locations (including the one visited in the immediate past), the attention model allows the robot to focus on it, thereby implementing a form of object-based IOR. A video of this experiment is available in the multimedia file "Multimedia_IOR.wmv" attached with this thesis.

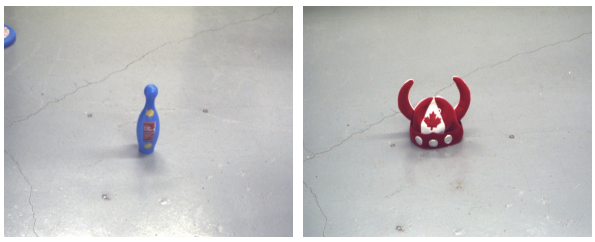
### 4.3.3 Analysis of Results

The results of the experiments reported in this section help to shed light on the following facts.

- The characteristic of the proposed model to attend the novel visual stimuli in the VF has been successfully implemented. The implementation of the proposed model on a robotic camera-head allowed the robot to identify and focus on the stimuli that appears as novel in its perception. This novelty preferring attribute guides the visual exploration behavior of the robot.

- The proposed model successfully tackles the research issues such as change of coordinates (*Issue 1.1*), dynamic IOR (*Issue 1.2*), partial appearance of features (*Issue 1.3*) to smoothly perform the visual exploration behavior.

- The integration of object- and space-based analysis in the proposed model helps to implement the dynamic IOR.



(a) (b)

Figure 4.10: Target objects for visual search experiments

- To resolve the conflict among competing stimuli, the relative proximity-based analysis provides more stable results than the relative saliency-based analysis.

## 4.4 Experiment 2: Visual Search

A set of experiments are performed to evaluate the performance of the proposed model in conducting a search for a specific target. Two target objects, a *red hat* and a *blue bowling pin* as shown in Fig. 4.10, are used throughout these experiments. The visual features of the targets are learned in a separate training phase. During the training phase the camera is exposed to the scenes of Figs. 4.10(a) and 4.10(b) separately. The model runs in visual exploration mode and the robot focuses on the only novel object present in each scene and learns the associated SIFT keypoints. The LTM developed in this way are used as the WM during the visual search experiments. The number of SIFT keypoints corresponding to the *blue bowling pin* and the *red hat* are 26 and 88, respectively. During the visual search the camera is exposed to a number of different environments where the targets share the VF with several other objects. A number of different environmental settings are chosen carefully in order to investigate the robustness of the visual search performed by the proposed model.

- Setting I: The target is rotated moderately with respect to the training case and placed with several other objects. Some of these objects share common feature(s) with the target. The background is the same as the training case.
- Setting II: The target is occluded partially and placed with other objects. The background is the same as the training case.
- Setting III: Multiple target-like objects are placed in the environment. The background is the same as the training case.

- Setting IV: The target is placed in backgrounds which are different than the training case.
- Setting V: The target is rotated significantly.
- Setting VI: The relative position of the camera and the target in the 3D world changes with respect to the training case.

For settings I- V the camera position in the 3D world remains the same as the training case. In each of these settings the robot tries to identify the target using the target information stored in the WM and focuses on it.

The results from the experiments are summarized in Table 4.2. In the majority of successful cases of visual search only one head movement, from the starting head-pose, is performed to focus on the target. In different trials of the same experiment the head-poses taken by the robot to focus on the target are averaged and is reported as the 'average pose sequence' in the Table 4.2. The robot successfully identifies the target with a moderate amount of occlusion and change in orientation. The visual search with two target-like objects in the VF creates the situation of competing target-like stimuli and the robot attends to both of the objects, one after another.

The most important aspect to notice in Table 4.2 is the failure of search in setting V and VI. In the case of setting V, the strong change in the target's orientation causes a failure in visual search for both of the target objects. In the case of setting VI, the targets are neither heavily rotated nor occluded as shown in Fig. 4.11(f) and 4.12(f). The model, however, fails to identify them using the knowledge of the targets extracted from the training images in Fig. 4.10(a) and 4.10(b). The reason behind this failure is the change in the camera perspective to which the SIFT keypoints are highly sensitive.

### 4.4.1 Analysis of Results

The results of the experiments reported in this section demonstrate the visual search characteristic of the proposed model. The visual search experiments are conducted with the prior-knowledge of the targets' visual features learned from one snap-shot of the target. With such a limited knowledge of the target, the proposed model demonstrates considerable robustness in visual search against change in orientation of the target, partial occlusion of target, and change in the background. In spite of this success, a notable shortcoming of the model lies in the fact that a separate training session was designed for target learning where a clean view of the target was manually arranged to provide the robot with enriched information about the target. This training stage significantly reduces the autonomy of the model. This issue has been discussed in section 1.2 of this thesis (*Issue 5*).

Table 4.2: Demonstration of visual search by the proposed model

| Setting | Target | | | | | |
|---|---|---|---|---|---|---|
| | Bowling pin | | | Red hat | | |
| | Result | # of trial | Average pose sequence | Result | # of trial | Average pose sequence |
| I | Success Fig. 4.11(a) | 5 | $(-12°, -42°) \rightarrow (2°, -50°)$ | Success Fig. 4.12(a) | 5 | $(-12°, -38°) \rightarrow (-24°, -34°)$ |
| II | Success Fig. 4.11(b) | 5 | $(-12°, -42°) \rightarrow (-30°, -38°)$ | Success Fig. 4.12(b) | 5 | $(-12°, -38°) \rightarrow (-7°, -34°)$ |
| III | Success Fig. 4.11(c) | 5 | $(-12°, -42°) \rightarrow (-17°, -34°) \rightarrow (-30°, -39°)$ | Success Fig. 4.12(c) | 5 | $(-12°, -38°) \rightarrow (0°, -46°) \rightarrow (-16°, -35°)$ |
| IV | Success Fig. 4.11(d) | 5 | $(-12°, -42°) \rightarrow (-28°, -38°)$ | Success Fig. 4.12(d) | 5 | $(-12°, -38°) \rightarrow (-30°, -44°)$ |
| V | Failure Fig. 4.11(e) | 5 | - | Failure Fig. 4.12(e) | 5 | - - |
| VI | Failure Fig. 4.11(f)-(top) | 3 | - | Failure Fig. 4.12(f)-(top) | 3 | - |
| | Failure Fig. 4.11(f) (middle) | 3 | - | Failure Fig. 4.12(f) (middle) | 3 | - |
| | Failure Fig. 4.11(f) (bottom) | 3 | - | Failure Fig. 4.12(f) (bottom) | 3 | - |

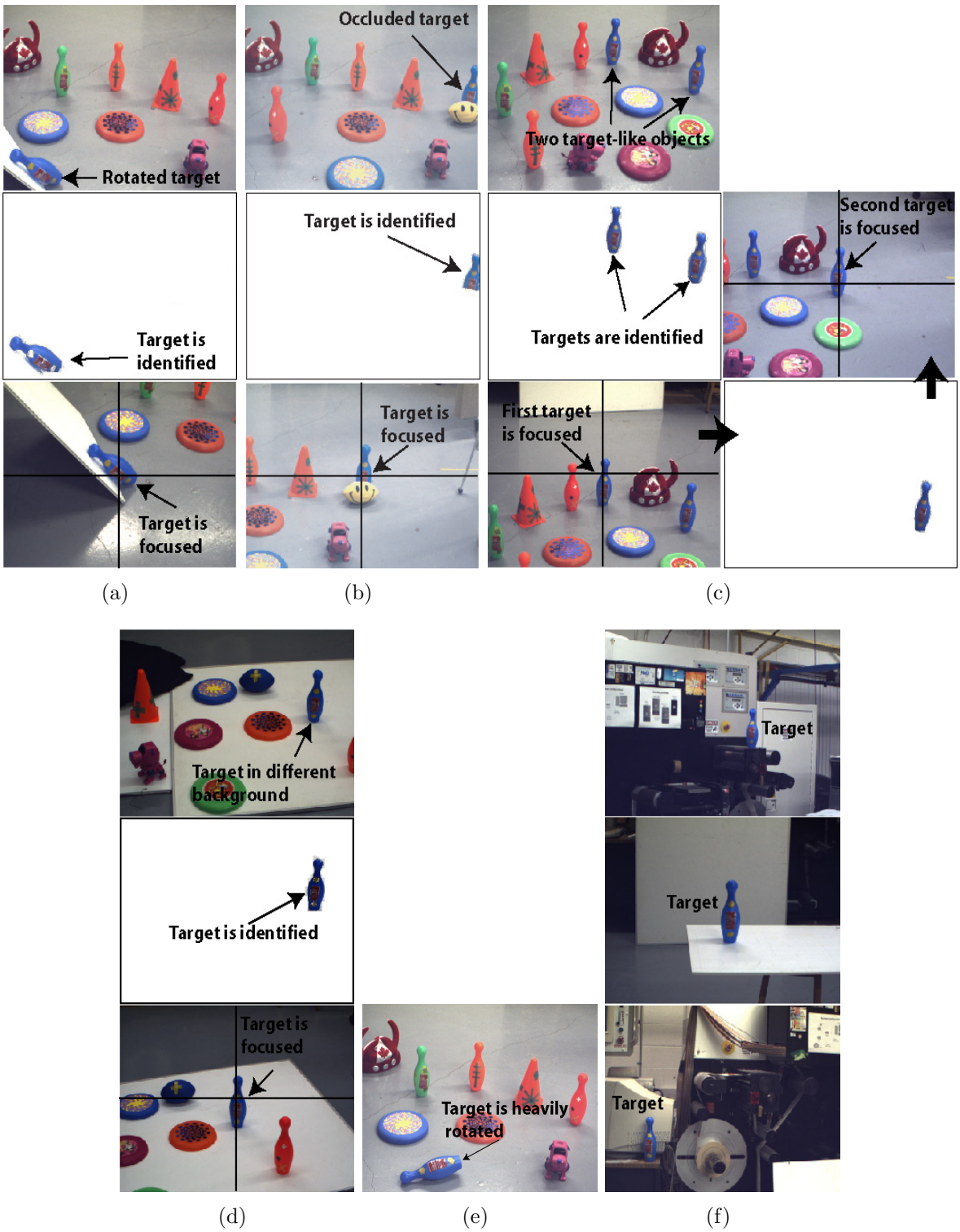



Figure 4.11: Visual search for the *blue bowling pin*. (a)-(d) Success. (e), (f) Failure (please see Table 4.2 for detail)

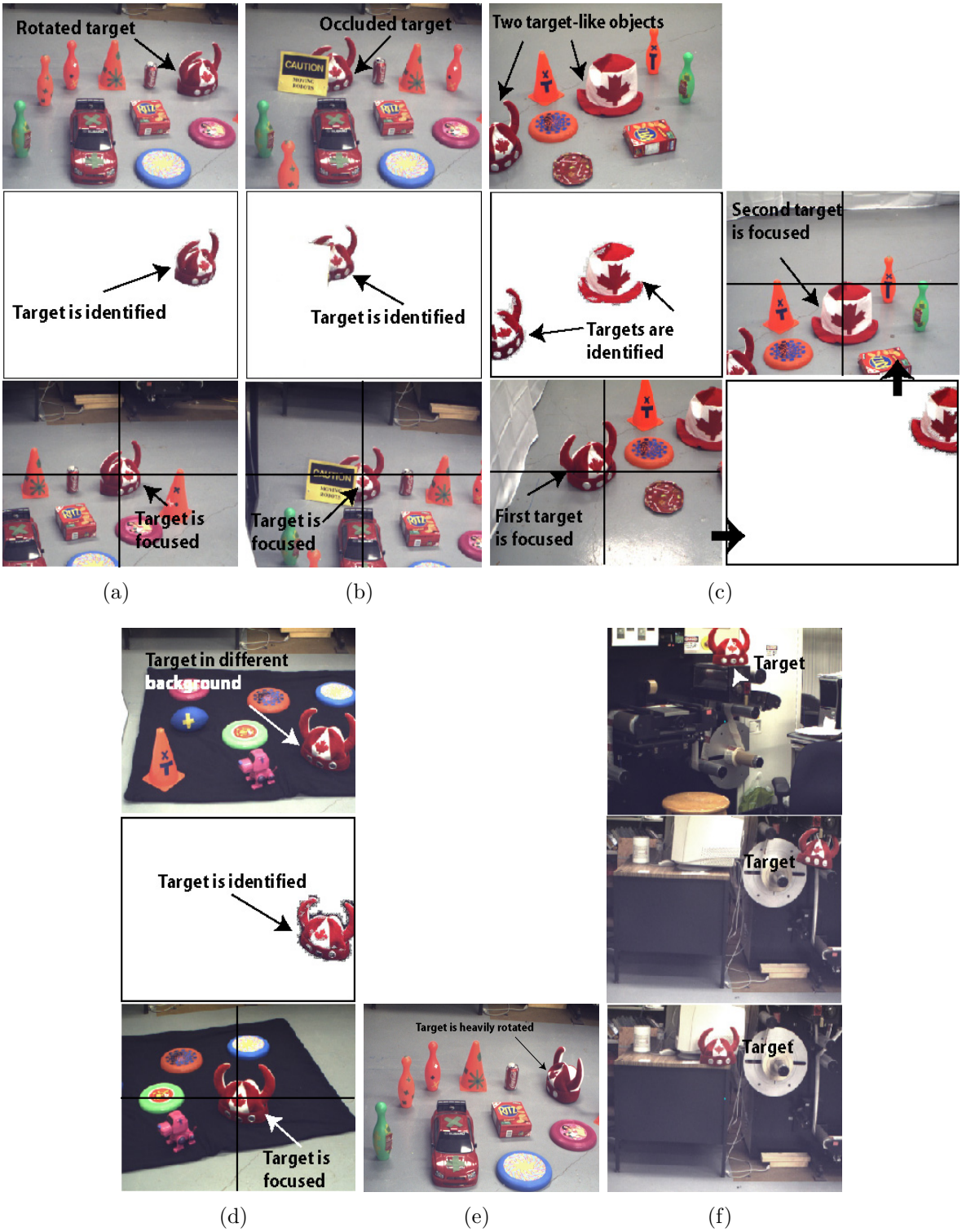



Figure 4.12: Visual search for the *red hat*. (a)-(d) Success. (e), (f) Failure (please see Table 4.2 for detail)

In addition to the training phase problem, the failure of the model to identify the targets in setting V and VI reveals a significant shortcoming of the proposed model in case of visual search. The proposed model is designed to serve as a component of robotic cognition. A cognitive robot is most likely to be a mobile platform with high degrees of freedom in perception and action. Such a robot is able to observe its surrounding from different perspectives. Besides, objects themselves often change their orientation (mediated by human) in the natural human environment. A visual attention model must be able to cope with the change in the camera perspective and large affine transformation of the object while performing visual search. Otherwise, an unrealistic amount of prior learning will be required to familiarize the robot with different objects in order to recognize them from arbitrary perspectives and/or orientations. A real-world robotic-application can not afford that sort of learning (*Issues 3, 4* in section 1.2).

## 4.5 Experiment 3: Performance with Parameter Variation

A set of experiments is conducted to investigate the performance of the proposed model subjected to variation in different parameters.

### 4.5.1 Number of Particle and Block Size

The success of a particle filter to approximate any posterior distribution highly depends on the number of particles. The higher the number of particles, the better the filter approximates the original distribution. But for a lower number of particles the estimated system state becomes a very crude approximation of the original system state. The way this incorrect state estimation affects the system behavior differs in different applications. In the case of the proposed Bayesian model of visual attention, the variation in the number of particles is manifested in the attention behavior of the robot in a very interesting and unique way.

To investigate the effect of particle size, four trials of a visual exploration experiment are conducted where each trail runs with a different number of particles. The number of particles used are $L = 100, 200, 500$, and 700. In each trial the robot starts with an empty LTM and is exposed to the VF shown in Fig. 4.13(a). The results of these experiments reveal two aspects of the proposed model which are affected by the number of particles: 1) the sequence of attention, and 2) the time required to detect a behaviorally relevant stimulus (in this case, a novel stimulus).

**The sequence of attention**

The attention sequences followed by the robot in different trials of the visual exploration experiment are shown in Figs. 4.13- 4.14. These sequences indicate that when the particle filter operates with fewer numbers of particles, the robot attends to the same object more than once. Attending to different parts of one object generally results in wiggly movements of the camera-head, especially when the object is small in size. The exact number of re-visit is generally proportional to the type and visual complexity of the objects. Highly textured and large objects (e.g., the red car in Fig. 4.13(b)) are re-visited more than the less-textured and relatively smaller objects (e.g., the orange toy in Fig. 4.13(b)). As the number of particles decreases, the number of re-visits increases. Besides, with too few particles the attention sequence sometimes may not even follow the theme of proximity-based analysis, i.e., the robot starts to attend the distant objects earlier than the closer objects. For around 500 particles and higher, the number of re-visit is optimized and the robot mostly likely attends to an object once. The reason behind this apparently strange attention behavior is the relation between the number of particles and the image segmentation process described in section 3.4.2. The image points corresponding to the head-poses (particles) are used as the seeds for the region growing algorithm which segments different objects. Fewer particles means fewer seeds for region growing. A region growing algorithm with a fewer number of seeds generally results in small object blobs which can not be merged together to form a complete view of the objects. The problem is severe for the objects with complex texture. The top-down modulation model considers different parts of a single object as isolated objects and assigns weights accordingly. This results in the re-visiting of a same object. When the particle number is high, several blobs are created from one object and the probabilities are higher that they will be merged together to form a complete view of the object. Consideration of the full object during the top-down biasing process reduces the probability of re-visiting it in the same trial. For better understanding of the fact, Fig. 4.14 shows the object blobs segmented with different numbers of particles from the visual field of the robot at the first decision cycle during the visual exploration experiments demonstrated in Fig. 4.13 (the blobs are shown in the RGB space). With higher number of particles the blobs are smoother and include more features from the underlying objects.

All of the experiments discussed in this section are performed with a value of $\epsilon = 20$. The parameter $\epsilon$ (the sub-image block size used for constructing the bottom-up competition model discussed in section 3.3.1) has an interesting effect on the relation between the number of particles and the image segmentation process. To investigate the effect of the parameter $\epsilon$, two more trials of the same experiment are conducted, one with $L = 100$ and $\epsilon = 10$ and the other with $L = 500$ and $\epsilon = 40$. The resulting attention sequences and segmented object blobs from the visual field of the robot at the first decision cycle are shown in Fig. 4.15. Comparison of the results shows that for a given number of particles,

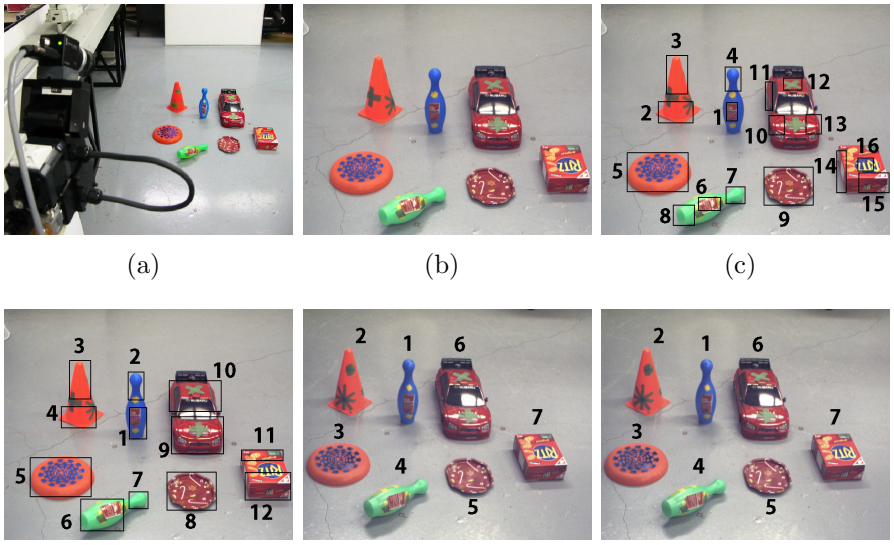
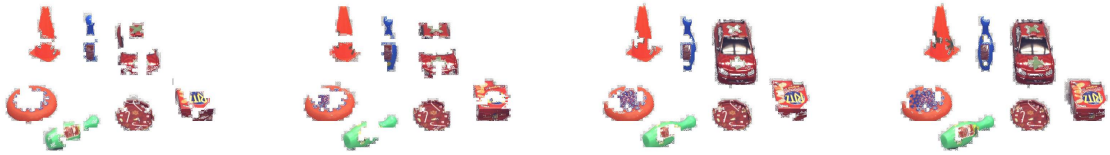
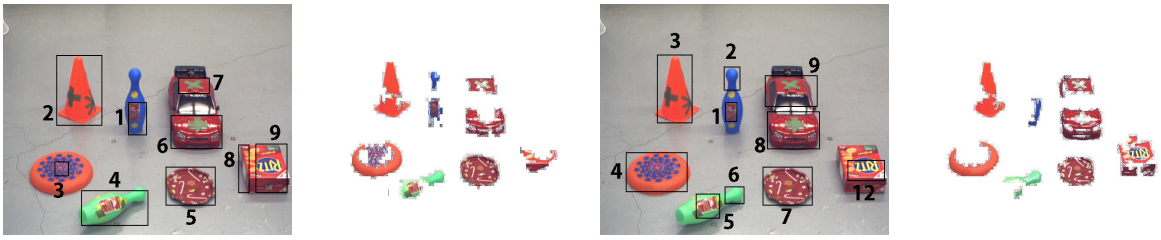



Figure 4.13: Effect of the number of particle $L$ on the visual attention behavior. The experimental environment is shown in (a) and the VF of the robot at the first decision cycle is shown in (b). The attention sequences resulted from using different numbers of particles are marked on the VF at the first decision cycle. (c) $L = 100$ (d) $L = 200$ (e) $L = 500$ (f) $L = 700$. As the number of particles decreases, the robot starts to attend to different parts of the objects rather than the entire object (please see text for detail)



(a) $L = 100$ (b) $L = 200$ (c) $L = 500$ (d) $L = 700$

Figure 4.14: The effect of the number of particles on visual attention behavior: the blobs segmented from the VF of the robot at the first decision cycle (during the visual exploration experiments demonstrated in Fig. 4.13) while using different number of particles (please see text for detail)



(a) $L = 100, \epsilon = 10$ (b) $L = 100, \epsilon = 10$ (c) $L = 500, \epsilon = 40$ (d) $L = 500, \epsilon = 40$

Figure 4.15: The effect of the block size on the visual attention behavior. The attention sequences resulted from using different combinations of values for $\epsilon$ and $L$ are marked on the first VF of the robot and shown in (a) and (c). The segmented object blobs are shown in (b) and (d) (please see text for detail)

small values of $\epsilon$ reduce the number of re-visits while large values of $\epsilon$ increase it. The underlying reason for this behavior is that smaller values of $\epsilon$ (i.e., small sub-image block) increase the effective resolution of the image plane by increasing the number of component distributions in the Gaussian mixture which represents the bottom-up competition model. The component distributions corresponding to image regions with bright color and/or with high contrast in color and intensity become narrower, while that corresponding to a plain background become wider. Sampling from such a bottom-up competition model increases the number of head-poses that focuses on meaningful objects rather than the background. Using the corresponding image points as the seeds for region-growing increases the chances of segmenting the full object. Accordingly, the chances of re-visiting the same object go down. On the other hand, for larger values of $\epsilon$ the effective resolution of the image plane decreases as the number of component Gaussian in the bottom-up competition model decreases. Due to a large area-coverage the majority of the component Gaussian become wider. Thus the probability of the samples collected according to such a bottom-up competition model to correspond to meaningful objects in the VF becomes lower.

### The time

The time for running the particle filter algorithm for visual attention varies with the number of particles as well as the block size. For the experiments demonstrated in Figs. 4.13- 4.15, the average time to detect a novel object is summarized in Table 4.3. Thus for a given environment, the attention filter needs more time to run with higher numbers of particles, and for a given number of particles the time requirement is inversely proportional to the block size.

Table 4.3: Effect of the parameters $L$ and $\epsilon$ on the time for novelty detection in a 2GHz processor

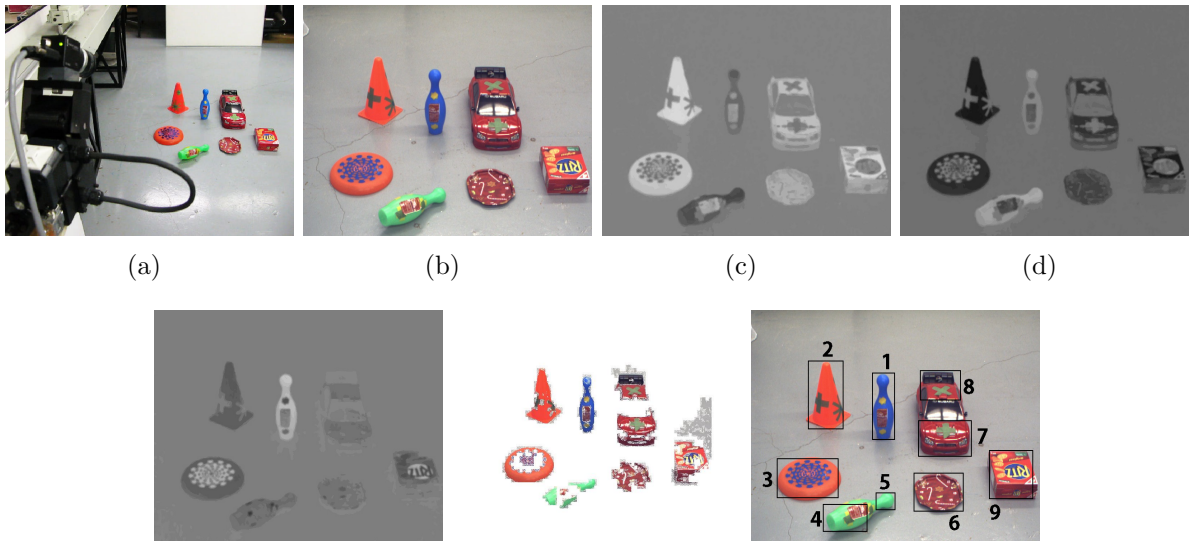
| L | $\epsilon$ | Time (ms) |
|---|---|---|
| 100 | 20 | 2850 |
| 200 | 20 | 3135 |
| 500 | 20 | 3984 |
| 700 | 20 | 4520 |
| 100 | 10 | 3138 |
| 100 | 40 | 3268 |
| 500 | 10 | 4781 |
| 500 | 40 | 3265 |

### 4.5.2 Image Processing Parameters $\zeta_1, \zeta_2$

The parameters $\zeta_1$ and $\zeta_2$ are involved with the pyramid-based image segmentation process discussed in section 3.4.2. The values for $\zeta_1$ and $\zeta_2$ are chosen to be 12 and 10, respectively on a a trial-and-error basis after analyzing 20 images of different indoor environments captured with a *Point Grey Research* Flea2 color camera and a Bumblebee stereo camera. The optics are different for these two test cameras. As the region growing algorithm (for segmentation of object blobs) is applied on the images resulted from the pyramid-based segmentation, the parameters $\zeta_1$ and $\zeta_2$ have some degree of influence on the proposed visual attention algorithm. To investigate this influence, two visual exploration experiments are conducted with different values for $\zeta_1$ and $\zeta_2$.

Similar to the other visual exploration experiments described in this section, the robot starts with an empty LTM and is initially exposed to the environment shown in Fig. 4.16(a). The values of the parameters for the first experiment are $\zeta_1 = 6$ and $\zeta_2 = 5$. The values of the other parameters involved are $L = 500$ and $\epsilon = 20$. The result of the pyramid segmentation for the first snap-shot of the environment (i.e., the image in Fig. 4.16(b)) is shown in Figs. 4.16(c)-4.16(e). While running the visual attention algorithm the object blobs segmented from the image in Fig. 4.16(b) are shown in Fig. 4.16(f). The attention sequence of the robot during the experiment is shown in Fig. 4.16(g).

Now the same experiment is conducted with $\zeta_1 = 24$ and $\zeta_2 = 20$. Corresponding results are presented in Fig. 4.17. Comparison of the results presented in Figs. 4.13, 4.16, and 4.17 indicate that for appropriately chosen values for $L$ and $\epsilon$, the attention sequence of the robot is not very sensitive to the variation of $\zeta_1$ and $\zeta_2$. The quality of the segmented object blobs, however, varies with the variation of $\zeta_1$ and $\zeta_2$. The smaller values of $\zeta_1$ and $\zeta_2$ make the pyramid segmentation process very restrictive and generate many small clusters in the segmented image. Application of region growing on this image results




(a) (b) (c) (d)

(e) (f) (g)

Figure 4.16: Effect of the image processing parameters: visual exploration experiment with $\zeta_1 = 6, \zeta_2 = 5$. (a) The experimental environment. (b) The VF of the robot at the first decision cycle. (c)-(e) The images $I^{r'}, I^{g'}$ and $I^{g'}$ corresponding to the VF. (f) The object blobs segmented from the image in (a) during the visual exploration experiment. (g) The attention sequence of the robot during the experiment (please see text for detail)

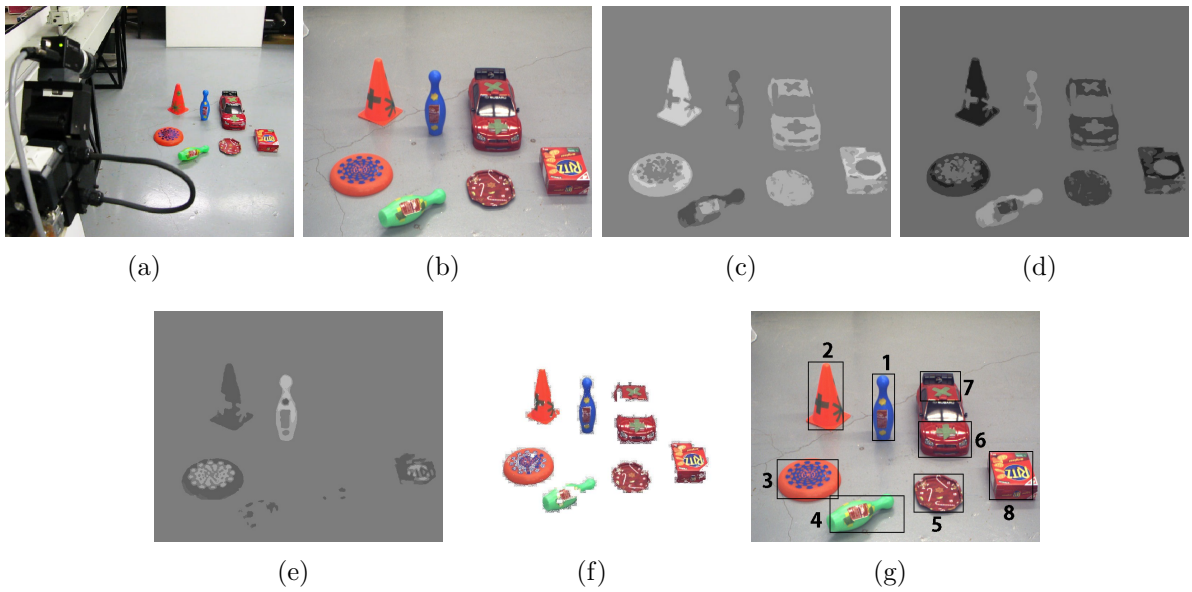
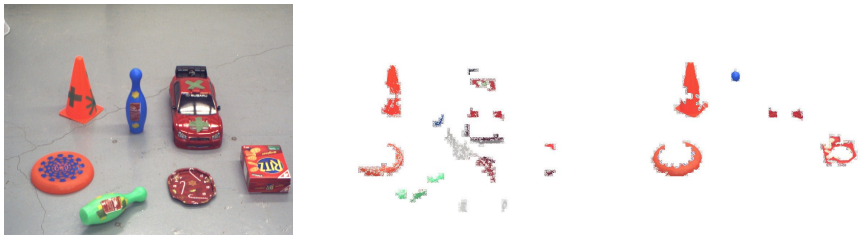



Figure 4.17: Effect of the image processing parameters: visual exploration experiment with $\zeta_1 = 24$ and $\zeta_2 = 20$. (a) The experimental environment (b) The visual field of the robot at the first decision cycle. (c)-(e) The images $I^{r'}, I^{g'}$ and $I^{g'}$ corresponding to the visual field (f) The object blobs segmented from the image in (a) during the visual exploration experiment. (g) The attention sequence of the robot during the experiment (please see text for detail)

in non-smooth object blobs as shown in Fig. 4.16(f). The blob merging technique used in this research, however, helps to merge such non-smooth blobs and selects a reasonable image region for SIFT keypoints extraction. For larger values of $\zeta_1$ and $\zeta_2$ the pyramid segmentation becomes relaxed and generates larger clusters. The region growing algorithm on such images results in relatively smoother object blobs as shown in Fig. 4.17(f). Too small or too large values of $\zeta_1$ and $\zeta_2$, however, may result in situations where the segmented object blobs will be too small or too big to qualify as an object (outliers or background). Thus there will be very few meaningful object blobs in the image and the majority of the head-pose samples will be assigned with zero weight. Fig. 4.18 shows such a scenario for two extreme values of $\zeta_1$ and $\zeta_2$. In such cases the attention behavior of the robot will change significantly.

The recommended operating regions for $\zeta_1, \zeta_2$ are the values around $\zeta_1 = 12$ and $\zeta_2 = 10$ but the region is not very restricted. It is, however, recommended to decide the values for these two parameters prior to working with a different camera (other than Flea2 or Bumblebee).



(a) (b) (c)

Figure 4.18: Effect of too large or too small values for $\zeta_1$ and $\zeta_2$. The object blobs segmented from the image in (a) during a visual exploration experiment using (b) $\zeta_1 = 2$ and $\zeta_2 = 1$ (c) $\zeta_1 = 72$ and $\zeta_2 = 60$. Too small values of $\zeta_1$ and $\zeta_2$ generate several small blobs corresponding to one object while too large values generate very large blobs which are discarded as background and their underlying objects can not compete for attention (please see text for detail)

### 4.5.3 The Size of the Memory

A series of experiments are conducted to investigate the effect of memory size on the time performance of visual attention. During the experiments the robot starts with an empty LTM and is initially exposed to the scene of Fig. 4.19(a). The model runs in visual exploration mode and attends to all of the seven novel objects. The time required to identify the first novel object is the time that the model requires to identify novelty with an empty LTM. The number of SIFT keypoints stored in the LTM after learning the seven novel objects is 801. With this LTM, the robot is then exposed to the scene of Fig. 4.19(b) which is the same scene as Fig. 4.19(a) except that one familiar object is replaced with a novel object. In this case, to identify the novel object the visual attention model has to perform a total of $NP$ keypoint-matching operations, where $N = 801$ and $P$ is the number of SIFT kyepoints associated with the objects segmented from Fig. 4.19(b). The robot attends to and learns the novel object. The time required to identify and focus on this object is the time that the model requires to identify novelty with a memory size of $N$. Learning the new object adds $P\prime$ more SIFT keypoints in the LTM where $P\prime$ is the number of keypoints associated with the novel object. In order to determine the time required for detection of novelty with a very large LTM, the number of SIFT keypoints in the LTM is doubled by manually copying the existing keypoints twice. The camera-head is then sequentially exposed to the scenes of Fig. 4.19(c)- 4.19(g). In each case the novel object is focused (and learned) and the LTM sized is doubled by copying the existing keypoints twice. The results are summarized in Fig. 4.20 where the time required to identify a novel object is plotted with respect to varying LTM sizes. The results of these experiments help to infer that a small increase in memory size does not have significant effect on the timing

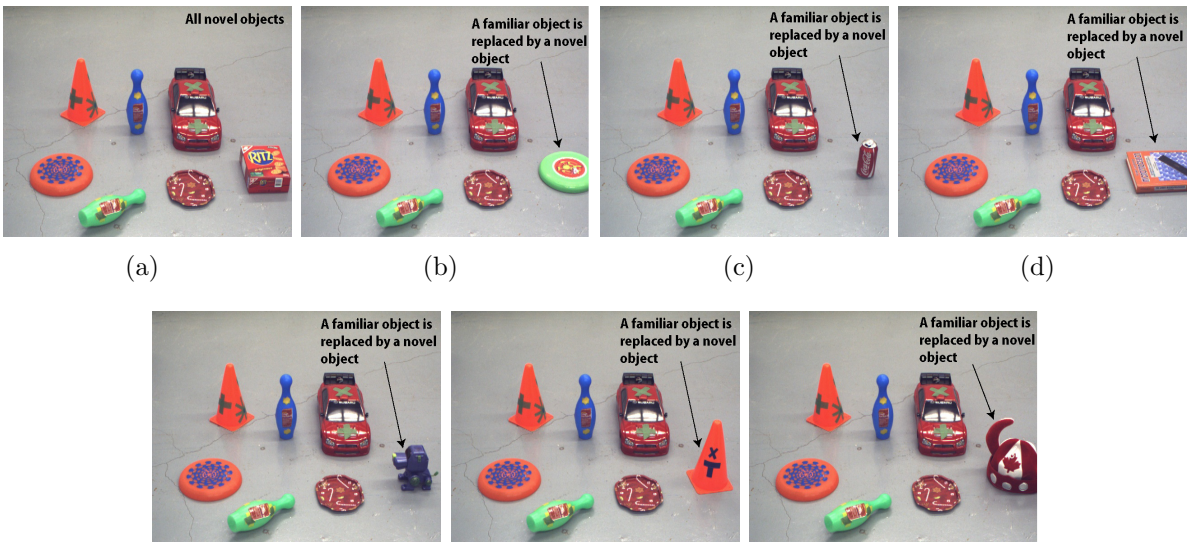



Figure 4.19: Effect of memory size on the time performance of visual attention. (a) A visual field with seven novel objects. (b)-(g) In a series of visual exploration experiments one familiar object from the scene is replaced by one novel object (please see text for detail)

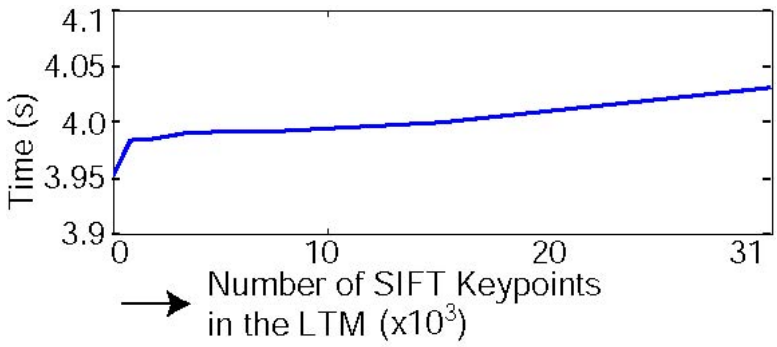



Figure 4.20: Time required to identify a novel object with varying memory (LTM) sizes (corresponding to the experiments demonstrated in Fig. 4.19)

of visual attention during visual exploration.

In case of visual search the number of keypoint-matching operation is also $NP$ where $N$ is the number of keypoints in the WM and $P$ is the number of SIFT keypoints associated with the objects segmented from a scene. As the time requirement is mostly related to the total number of keypoint matching operation, the effect of WM size on the time performance of visual search can also be expected as insignificant. Thus the memory size has small effect on the timing of visual attention.

### 4.5.4 Analysis of Results

The results of the experiments presented in this section shed light on the following facts about the parameters involved with the proposed Bayesian model of visual attention.

- The number of particles affects the sequence of attention as well as the time to identify a behaviorally relevant stimuli. The experimental investigation suggests that satisfactory performance is achieved when the number of particles is around 500.
- The sub-image block size involved with the computation of the bottom-up competition model also affects the sequence of attention as well as the time to identify a behaviorally relevant stimuli. The experimental investigation suggests that a satisfactory performance is achieved when operating with a $20 \times 20$ block size.
- The image processing parameters influence the top-down modulation model and moderately affect the sequence of attention. These parameters are mostly camera dependent and should be fixed prior to implementing the algorithm on a different camera.
- The increasing size of the LTM of the robot gradually makes the robot slow in responding to behaviorally relevant stimuli. The rate of this change, however, is very small.

## 4.6 Conclusion

This chapter has described a set of experiments for evaluating the performance of the Bayesian visual attention model proposed in chapter 3. The results of the experiments help to summarize the following characteristics of the Bayesian model.

1. The model successfully exhibits the visual exploration and visual search behavior in a natural environment.

2. The model efficiently addresses the research *Issues* 1 and 2 in different cases of visual exploration and visual search.

3. As demonstrated in the experiments, the visual search and visual exploration run in mutual exclusion of each other. This hampers the generality of the model and makes it unsuitable for most robotic applications (*Issue* 4)

4. The proposed model fails to address the *Issues* 3 and 5 in the case of visual search.

5. The parameters involved with the implementation technique of the proposed model have effect on the performance of the model. The performance variation with the change in parameters' values has been investigated in detail.

It is, therefore, evident that the Bayesian model proposed in chapter 3 does not completely fulfill the promises of this thesis made in section 1.2. In order to address the *Issues* $3, 4$, and 5 this thesis proposes a solution which relies on multi-modal interaction with the human during attentional selection. Accordingly, the proposed Bayesian model of visual attention is further extended to accommodate the multi-modal information and the human-robot interaction. The multi-modal attention algorithm will be discussed in the following chapter.

# Chapter 5

# A Multi-modal Extension of the Proposed Model

A multi-modal extension of the Bayesian visual attention model is proposed in this chapter. The major goal of this extension is to address a number of research issues namely, optimal learning strategy, generality, and prior training. The original Bayesian model proposed in chapters 3 and 4 fails to deal with these research issues. This thesis identifies the role of multi-modal information and occasional interaction with the human as a potential solution to deal with these research issues. Accordingly, it proposes an extension of the Bayesian attention model to design a visual attention-oriented speech-based HRI framework. The proposed model integrates the visual search and visual exploration. Visual exploration runs as the default mode of attention. Switching to the visual search mode occurs in response to the user request which is conveyed to the robot through speech command. Speech, being the most powerful modality of human communication, makes the attention-oriented bi-directional HRI natural. The robot maintains an audio-visual memory of the attended objects. This enables the robot to autonomously fetch the target information from the memory when a request for visual search is made. This eliminates the requirement of a training session for target learning. In case of failure in target identification, however, the speech-based HRI framework can assist the robot to recover from its fault through using human guidance. The novelty preferring characteristics of the Bayesian attention model along with the proposed attention-oriented speech-based HRI framework is used to design a strategy for optimal learning.

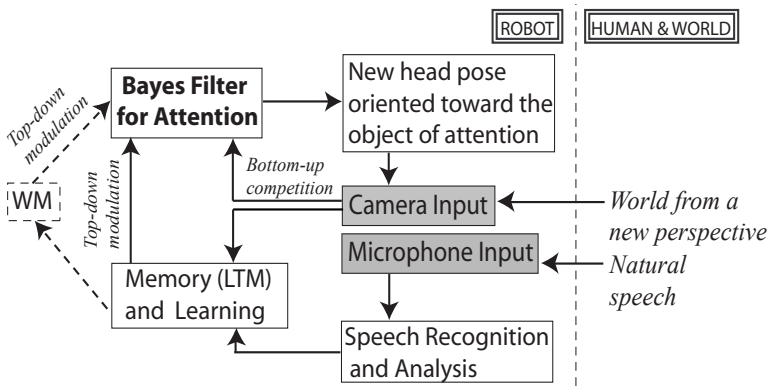



Figure 5.1: Functional description of the proposed model. Visual search and visual exploration is integrated through robot learning and speech-based interaction with the human

## 5.1 Functional Overview of the Model

The attention-oriented speech-based HRI framework of the proposed model binds the two modes of visual attention in the same framework. The default mode is the visual exploration behavior. The visual search mode is activated by the speech-based request from the user. The proposed attention-oriented HRI framework is bi-directional in the sense that the interaction can be triggered by both the human and the robot. Fig. 5.1 shows an overview of the operation of the model. At each decision cycle the model guides the robot through the following steps in order to identify the most behaviorally relevant object.

1. The robot visually orients itself to the object identified as 'worthy to attend' during the immediate past cycle. A camera image is captured at this new head-pose.

2. The robot triggers an interaction with the human. As per robot's perception, if the current object of attention is a 'novel object', then the goal of this interaction is to request the human to provide certain high-level knowledge about this object (e.g name of the object, name of its color). On the other hand, if the object is attended in response to a search request from the human (occurred during the immediate past cycle), the robot interacts with the human to confirm if the current object of attention is the 'sought for' object.

3. The speech feedback from the human is recognized and analyzed to determine its content. If the feedback is about high-level description of the current object of attention, the speech information along with the newly captured frame are used for learning. If the feedback is to affirm the success of a visual search then the WM is cleared and the robot switches back to the default mode of novelty exploration. If the

human negates the success of visual search, the robot retains the WM and continues the search.

4. In case of sustained visual search the Bayes filter for attention processes the newly arrived camera input along with the top-down modulating bias from the WM to identify the 'sough for' object. At this stage, if required, the human can provide the robot with speech-based assistive feedback to identify the 'sought for' object.

5. In case of switching back to novelty exploration mode, the Bayes filter for attention processes the newly arrived camera input along with the top-down modulating bias from the LTM to identify a novel object to focus at the next cycle.

6. At any time the human can request the robot to search for any specific object. In that case, the human triggers an interaction and provides the robot with high-level specifications of the 'sought for' object using speech command. Until a success of this visual search is confirmed by the human, steps 1 to 4 repeat themselves.

Prior to describing the multi-modal extension of the Bayes filter, the following section describes the audio-visual memory and the learning mechanism which play a key role in the operation of the model.

## 5.2 The Audio-Visual Memory

The LTM of the robot contains audio-visual information of the attended objects. The robot is able to automatically generate a WM from the LTM based on the contextual demand.

For the multi-modal extension of the model, two different kinds of visual features of the objects are considered during top-down bias generation: the SIFT keypoints and the color of the objects. The LTM stores the visual features in two separate linked lists: the first one is for the set of SIFT keypoints and the second one is for the color information of the attended objects. The process of extracting SIFT keypoints associated with each object is same as described in section 3.4.2. For each attended object blob, the mean values of the pixels in the YCrCb color space are used as color information. The color information, therefore, is represented by a three dimensional vector. After attending each new object, its name and color are introduced by the human. The name of an object is used as the 'object label' for its corresponding set of SIFT keypoint while the color name is used as the 'color label' for the corresponding color vector. A look-up table keeps record of the entries of linked list that corresponding to a specific 'object label'.

Other than the case of visual search, the attention of the robot is driven by its sense of novelty. Re-attending to a previously attended object, therefore, indicates that the visual

features of the object appear as novel to the robot, possibly due to change in camera perspective, lighting condition, or change in the object's own orientation in the 3D world. As the object was attended at some other time in the past, its name and color exist in the lists of 'object labels' and 'color labels' in the LTM of the robot. When such a match occurs, the robot stores the newly arrived SIFT keypoints under the already existing 'object label' of that object. Note that the entire set of keypoints are stored. This is because the current implementation considers an object 'novel' if the number of its SIFT keypoints that matches with any object in the LTM is less than 3. Therefore, re-identifying an object as 'novel' means that majority of the SIFT keypoints contain new information about that object. In case of color information, however, a Gaussian Adaptive Resonance Theory (GART)- style learning [152] is performed to update the LTM. Each color vector forms a cluster in the color feature space. A color cluster is characterized by a triplet $(\bar{\mu}, \bar{\sigma}, n)$, where $\bar{\mu}$ and $\bar{\sigma}$ are three-dimensional vectors representing the cluster mean and standard deviation in each feature dimension, respectively, and $n$ is a scalar denoting the number of times the color has been observed. For the first visual encounter of a color, the color feature vector itself serves as the mean of that color cluster and an arbitrary value $\gamma$ is used as the initial standard deviation in each feature dimension. For every subsequent encounter of the same color, which generates a color vector $\bar{y}$, the corresponding color cluster in the LTM is updated as follows.

$$
\begin{aligned}
n &= n + 1 && (5.1)\\
\bar{\mu} &= (1 - n^{-1})\bar{\mu} + n^{-1}\bar{y} && (5.2)\\
\bar{\sigma}_i &= \sqrt{(1 - n^{-1})(\bar{\sigma}_i)^2 + n^{-1}(\bar{\mu}_i - \bar{y}_i)^2} \text{ if } n > 1 \\
&= \gamma, \text{ otherwise } (i = 1, 2, 3) && (5.3)
\end{aligned}
$$

The proposed strategy to maintain and update the LTM has two significant consequences.

- The robot stores only those views of an object which, the robot 'feels', are required to identify it in a changed background or from a different perspective. This saves the LTM from being burdened with overlapping information.
- The high-level knowledge obtained from the human saves the system from running a complex pattern matching algorithm to decide if the two sets of SIFT keypoints observed at two different points of time and from two different perspectives belong to the same object or not. A major disadvantage of using such a pattern-matching algorithm is that the attention behavior of the robot will be heavily dependent on its success.

The WM is created during visual search. When the human makes a visual search request for an object, the robot places the object's specification (e.g., name, color) in the WM and

searches the LTM to find the most appropriate match for this specification. If a match occurs (several partial matches might also occur due to failure in speech recognition), the corresponding visual features (e.g SIFT keypoints and/or color information) are copied to the WM which then acts as a template of the 'sought for' object for the top-down bias generation. The WM is cleared after successfully processing every visual search request.

## 5.3 Bayesian Formulation for the Audio-Visual Attention

The Bayesian filter for robotic visual attention, as described in section 3.3, is expressed as follows.

$$Bel(\mathbf{x}_k) = \eta p(\mathbf{b}_k|\mathbf{x}_k, \mathbf{F}_k) \int p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{F}_k) Bel(\mathbf{x}_{k-1}) d\mathbf{x}_{k-1}$$

For the current audio-visual scenario there are two different types of sensor measurements: visual input from the camera ($\mathbf{F}_k^v$), and speech input from the microphone ($\mathbf{F}_k^a$), $\mathbf{F}_k = \{\mathbf{F}_k^v, \mathbf{F}_k^a\}$. The speech feature set $\mathbf{F}_k^a$ contains the name and/or color of the 'sought for' object spoken by the human and recognized by the speech recognition engine. The modification of the bottom-up and top-down models for the multi-modal case is discussed below.

### Bottom-up competition

The bottom-up competition model is generated based only on the visual features and is evaluated in the same way as the unimodal case described in section 3.3.1.

$$p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{F}_k) = p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{F}_k^v) \tag{5.4}$$

This distribution describes the transition probability between two head poses conditioned on the perceived visual features.

### Top-down modulation

The top-down modulation model considers both the visual and the speech information for bias generation and is modified as follows.

$$p(\mathbf{b}_k|\mathbf{x}_k, \mathbf{F}_k^v, \mathbf{F}_k^a) \quad = \quad p(\mathbf{b}_k^{LTM}|\mathbf{x}_k, \mathbf{F}_k^v) + p(\mathbf{b}_k^{WM}|\mathbf{x}_k, \mathbf{F}_k^v, \mathbf{F}_k^a) \tag{5.5}$$

The first term on the right hand side of (5.5) implements the default novelty preferring behavior of the robot. The probability $p(\mathbf{b}_k^{LTM}|\mathbf{x}_k, \mathbf{F}_k^v)$ is evaluated using equations (3.12)-(3.15). The second term in (5.5) performs a speech-based modulation of visual attention. The probability $p(\mathbf{b}_k^{WM}|\mathbf{x}_k, \mathbf{F}_k^v, \mathbf{F}_k^a)$ evaluates a non-zero bias for the head poses which focus on objects with visual features having the same 'object label' and/or 'color label' as $\mathbf{F}_k^a$. The probability is evaluated as follows.

$$p(\mathbf{b}_k^{WM}|\mathbf{x}_k, \mathbf{F}_k^v, \mathbf{F}_k^a) = p(Z|\mathbf{F}_k^v, \mathbf{F}_k^a) \tag{5.6}$$

Here $Z$ represents the hypothesis that the object focused at the head-pose $\mathbf{x}_k$ is the same object as the human 'asked for'.

$$p(Z|\mathbf{F}_k^v, \mathbf{F}_k^a) = p_a p_v \tag{5.7}$$

Here $p_a$ is the probability that a WM has been created from the LTM using the speech information and $p_v$ is the probability that the visual features of the object focused at $\mathbf{x}_k$ is the correct match of the current content of the WM. A value for $p_v$ is calculated using (3.15) but the WM, rather than the entire LTM, is used for matching. There are three possible cases in evaluating a value for $p_a$.

- **Case I** The audio input exactly matches with an 'object label' and/or a 'color label' in the LTM data base, $p_a = 1$
- **Case II** Due to failure in speech recognition the audio input partially matches with several 'object labels' and/or 'color labels' in the LTM database.

  $$p_a = \frac{c}{C} \tag{5.8}$$

  Here, $C$ : the total number of letters in the recognized 'label', and $c$ : the number of letters matched (from the left) with an LTM 'label' until the first mismatch is found.
- **Case III** The audio input does not match any of the 'labels' in the LTM database, $p_a = 0$.

In the third case, a failure in visual search is reported.

## 5.4 Speech Understanding

The open source speech recognition engine 'Julius' is used to understand the speech command. The vocabulary and the grammar developed for the current system enable the robot to understand speech input of four different categories.

- **Description:** This type of input consists of high level description of an attended object. Example grammar for a description type speech input is: *'That is a [color] [object]'*. Here *color*: red, green, blue, etc. and *object*: car, toy, book, etc.

- **Query:** The visual search request is made through this type of speech command. Example grammar for a query type command is: *'Find a [color] [object]'*.

- **Confirmation:** After executing a visual search task the robot checks with the human if the focused object is the 'sought for' object or not. The human performs this verification using confirmation type speech input. Example grammars are: *'Yes, thank you', 'that is correct','no, try again'*. An affirmative confirmation causes the robot to switch back to the visual exploration mode while a negative confirmation instructs to continue the search.

- **Assistance:** In case of visual search at least two confusing cases might arise: 1) There is more than one target-like object and the human wants a specific one of them, but the robot focuses on the others. 2) The robot has some knowledge of the target but can not recognize it in the current scene due to change in lighting or view angle. In both of these two cases the human can provide the robot with additional clues to narrow down the search and thereby the robot can identify the target successfully. The assistance type speech input are used to perform this. Example grammars are: *'Find something [color]', 'not this one', 'no, the other one',* etc. The first example grammar basically helps to narrow down the search to specific colored objects while the others cause the focus to travel through different target-like objects before landing on the expected one (which is confirmed using the *confirmation* type speech input).

Once a speech input is recognized, a speech understanding module, which is a simple state machine and operates like a parser, categorizes it into one of the above four categories and extracts the useful information, e.g., color and/or name of an object.

## 5.5 Dealing with the Research Issues

This section briefly summarizes the way the proposed model deals with the issues of *optimal learning strategy, generality,* and *prior training.*

### 5.5.1 Generality

The proposed multi-modal Bayesian framework of visual attention addresses the problem of generality by integrating visual search and visual exploration in the same framework.

The multi-modal attention model and the attention-oriented speech-based HRI framework enable the robot to switch back-and-forth between visual search and visual exploration modes depending on the context.

### 5.5.2 Optimal Learning Strategy

The proposed framework of visual attention suggests self-directed learning by the robot. The LTM of the robot helps it to develop its own sense of novelty while the Bayesian model causes the robot to prefer novel stimuli for attention (given that there is no request for visual search). Combination of these two phenomena helps to devise this self-directed learning strategy. The robot attends to whatever appears as 'novel' in its perception. If a previously attended object appears as novel at a later time due to change in lighting and view angle, the robot will re-attend (and learn) it. This makes the knowledge of the robot about that object more complete. This strategy of *"learn what you feel is required"* voids the requirement of teaching the robot with a number of carefully chosen views of different target objects. The high-level knowledge from the human operator still plays a key role to quickly re-organize the LTM every time an object is re-attended. This strategy does not guarantee the most optimal learning but provides a way to get rid of the huge manual work and uncertainty involved with target learning.

### 5.5.3 Prior Training

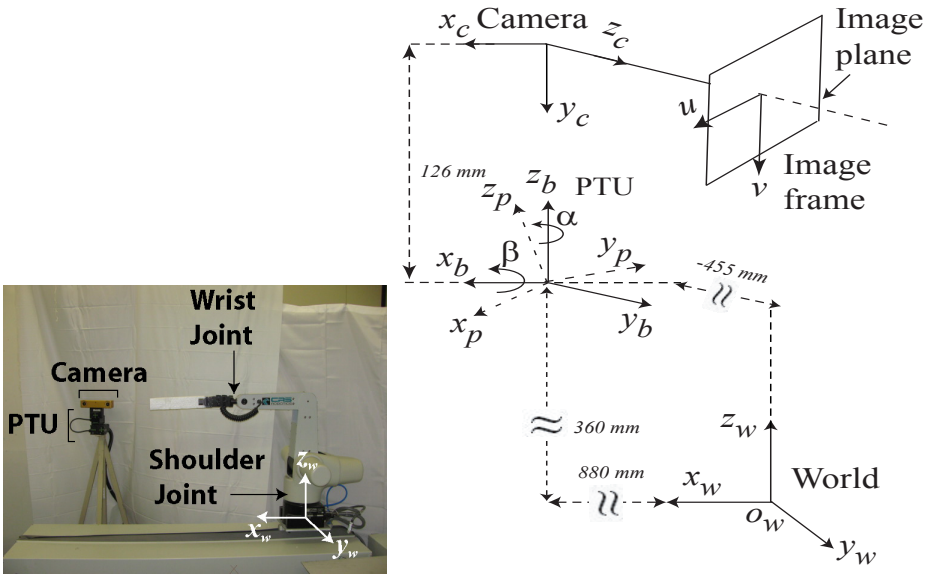
According to the proposed model, the robot maintains an LTM of the attended objects and is able to create the WM relevant to a visual search request. The more time the robot spends in an environment, the more it encounters different objects from different perspectives. This enriches the LTM and, in turn, enhances the probability of success in visual search in any arbitrary setting. The requirement of a separate training session, therefore, is essentially eliminated. If the robot, in spite of its long visual experience, fails to identify a target, an *on-spot learning* strategy is proposed to tackle such scenario. The *on-spot learning* strategy uses the *'assistance'* type speech input to conduct an interactive visual search in order to direct the attention of the robot toward the target object (similar to scaffolding). Depending on the complexity of the environment it may take several decision cycles before the robot focuses on the correct object. When the robot focuses on the correct object, the success of visual search affirmed by the human causes the robot to update its database with the new information of a familiar object. For instance, if the robot fails to identify an object due to huge change in its orientation or camera perspective (to which the SIFT keypoints are sensitive), the human can ask the robot to look for an object with a similar color as the target object (color is a more stable feature with respect to this kind of variation). If there ise more than one object with that specific color then

the human can guide the robot toward the correct object by negating the success of visual search every time the robot focuses on an undesired target object (e.g by saying *'try again'*, or *'not this object'*, or *'try the other one'* etc). A significant beauty of this strategy is that it does not expect the model to perform accurately at all times. Rather, if it fails, the interaction with the human allows the model to recover from its failure and learn from the experience.

## 5.6 The Proposed Model and the Operator Burden

The attention-oriented speech-based HRI framework proposed in this chapter plays a major role to address the *Issues* 3, 4 and 5 and thus enables the proposed Bayesian model of visual attention to address all of the research issues of robotic visual attention discussed in section 1.2 of this thesis. As the interaction with the human is speech-based, the human operator does not have to be an expert in handling the robots. Enriching the vocabulary and grammar will further enhance the situation in this regard. The proposed HRI framework, however, imposes one constraint on the human operator. That is, when the robot focuses on an object and interacts with the human for feedback, the human operator must be around the robot to identify the object toward which the camera is oriented. The situation becomes very confusing when the focused object is located closely with other objects. An obvious solution to this problem is that the image of the focused object can be made available to the human operator through WiFi devices e.g., PDA, iPhone. When the operator is located at a different place than the robot then the use of such WiFi devices is unavoidable. But when both the operator and the robot are located in the same room, it is possible to make the interaction independent of such devices. We propose to include a pointing behavior in the robot to point to the object it is currently focusing at using its available actuating resources (e.g., an arm-like robotic manipulator). During the developmental process of human, pointing behavior is considered as one of the strongest means of establishing joint attention with the caregivers [124]. Implementing pointing behavior in the robots, therefore, is becoming popular in robotic models of joint attention [126]. In the proposed work, a robotic manipulator is used to point to the attended objects. After identifying a focus of attention, camera calibration is performed to guide the manipulator to point to the focused object. Fig. 5.2 shows the relative position of the camera, PTU, and the manipulator used in the experiment as well as the coordinate systems involved with them. Here, a stereo camera is used to obtain accuracy in hand-eye coordination. The *Triclops* API from *Point Grey Research* is used to calculate the location of the focused object in the three-dimensional camera coordinate. The object location expressed in the camera coordinate is transformed to the world coordinate system as follows.

$$\mathbf{A}_w = {}^{w}T_b \, {}^{b}T_h \, {}^{h}T_c \mathbf{A}_c \tag{5.9}$$




(a) (b)

Figure 5.2: (a) The experimental set-up: the robotic manipulator used for pointing behavior, the PTU and the camera (b) The coordinate systems involved with the setup

Here,

- $\mathbf{A}_c$: Object location expressed in the camera coordinate system
- $\mathbf{A}_w$: Object location expressed in the world coordinate system
- ${}^hT_c$: transformation matrix between the camera and the current position of the PTU
- ${}^bT_h$: transformation matrix between the current position of the PTU and the 'home' position of the PTU
- ${}^wT_b$: transformation matrix between the 'home' position of the PTU and the world coordinate system

Among the transformation matrices, ${}^hT_C$ and ${}^wT_b$ are constant for a given setup and can be evaluated based on the dimension of the PTU and the physical separation between the robot and the PTU (as shown in Fig. 5.2(b)). The matrix ${}^bT_h$ depends on the current head-pose $\mathbf{x}_k$ as follows.

$$ {}^bT_h = \begin{bmatrix} cos\alpha & -sin\alpha cos\beta & -sin\alpha sin\beta & 0 \\ sin\alpha & cos\alpha cos\beta & cos\alpha sin\beta & 0 \\ 0 & -sin\beta & cos\beta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{5.10} $$

The current implementation is using the 'shoulder' and the 'wrist' joint of the manipulator for pointing to an object. The coordinate system of the 'shoulder' joint coincides with the world coordinate system. The location $\mathbf{A}_w$, therefore, can be used directly to deliver motion command to the 'shoulder' joint. The 'wrist' joint, however, operates with a coordinate system different than the world coordinate system. The location $\mathbf{A}_w$ is further transformed to the 'wrist' coordinate system for appropriate motion command to the wrist joint.

## 5.7 Conclusion

This chapter has presented a multi-modal extension of the Bayesian model for visual attention. The proposed extension includes a visual attention-oriented speech-based HRI framework which contributes to deal with the research *issues 3, 4, 5*. Through this multi-modal extension the proposed Bayesian model of visual attention addresses all of the research issues involved with robotic visual attention as discussed in section 1.2. The attention-oriented speech-based HRI integrates the two modes of visual attention (visual exploration and visual search) in the same framework. Thus the attention of the robot can switch back-and-forth between these two modes autonomously. The model maintains an audio-visual LTM of the attended objects which helps the robot to grow a sense of novelty. The novelty preferring characteristics of the model along with the robot's own sense of novelty are exploited to devise a learning strategy for the robot. The robot learns whatever it 'feels' is required to be learned in order to enrich its knowledge about an object. It is, therefore, not required to teach the robot with different views of the target for robust identification. The proposed model also allows the robot to autonomously create a WM (using the information from the LTM) corresponding to a visual search request. This eliminates the requirement of a separate training session prior to the visual search. This chapter has also described the implementation of pointing behavior in the robot where the robot will point to the focused object. This reduces the burden on a human operator to identify the focused object during the human-robot interaction. The next chapter will describe a set of experiments for performance evaluation of the model presented in this chapter.

# Chapter 6

# Performance Evaluation of the Multi-modal Attention Model

This chapter describes a set of experiments for self-evaluation of the performance of the multi-modal Bayesian visual attention model described in chapter 5. As the multi-modal extension of the Bayes model is proposed to address the research *issues 3,4,* and *5*, the experiments are specifically designed to investigate how well the proposed model addresses these research issues. Accordingly, the experiments are categorized into four groups:

1. Experiments related to optimal learning strategy
2. Experiments related to generality
3. Experiments related to prior training
4. Experiments related to pointing behavior

As the extension is made on the basic Bayesian model described in chapters 3 and 4, the extended model inherently tackles the research *issues 1, 2* whenever they arise, while performing these experiments.

## 6.1 Experimental Hardware

The majority of the experiments are conducted using the Flea2 color camera and a *Directed Perception* PTU as described in section 4.2. The experiments related to the pointing behavior, however, use the Bumblebee2 stereo camera from *Point Grey Research*. Bumblebee2 has a wide angle optics with angle of view of $97^o(H) \times 73^o(V)$. Stereo images are also

of dimension $640 \times 480$ *pixels*. A T265 CRS robot manipulator is used to implement the pointing behavior. The 'Active Robot' API supplied with the CRS manipulator is used to solve the inverse kinematics of the robot. A simple hand-held microphone is used to record the speech input. Similar to the experiments described in chapter 4 the particle filter implementation uses 500 particles. Corresponding to one snap-shot of the environment the total time required for running the attention algorithm, executing the motion command by the PTU (and the manipulator), and interacting with the human operator is considered as the time for one decision cycle. Thus the average length of the decision cycle of the experiments presented in this chapter is longer than that of the experiments presented in chapter 4.

Due to the nature of the experiments presented in this chapter the location of the camera in the 3D world will vary. The camera position in the 3D world is expressed with respect to the world coordinate system using the five-tuple $(r, \theta, \phi, \alpha, \beta)$, where $(r, \theta, \phi)$ represent the location of the PTU in the spherical coordinate system with respect to the center of the world coordinate system. Thus, $r$ is the radial distance of the PTU from the center of the world coordinate system, $\theta$ is the azimuth angle in the $(x_w, y_w)$ plane from the $x_w$ axis, and $\phi$ is the zenith angle from the positive $z_w$ axis. $\alpha$ and $\beta$ are the pan-tilt angles of the PTU (please see Fig. 1.5 for the definitions of different coordinate systems).

## 6.2 Experiment 1: Optimal Learning Strategy

A set of visual search experiments is performed first where target learning is conducted using the traditional method, i.e., in a separate training phase several views of the targets are manually arranged to learn target related visual features from different perspectives. The results from these visual search experiments are then compared with that from another set of experiments where target learning is conducted using the proposed self-directed learning strategy. The same two objects (a *red hat* and a *blue bowling pin*) as used in the visual search experiments described in chapter 4 are used as targets.

#### Experiment 1A

For this set of experiments the targets are learned in a separate training phase. Four different views of the *red hat* and seven views of the *blue pin* are chosen manually for the training purpose. The target views are shown in Fig. 6.2. The visual features of a target extracted from its different views are combined together to create a WM to be used during visual search for that target. Camera position during the training phase is approximately at $(r = 6ft, \theta = 85^o, \phi = 50^o, \alpha = 0^o, \beta = 0^o)$ while the targets are located roughly along the $x_w$ axis and approximately $2ft$ away from the center $o_w$.

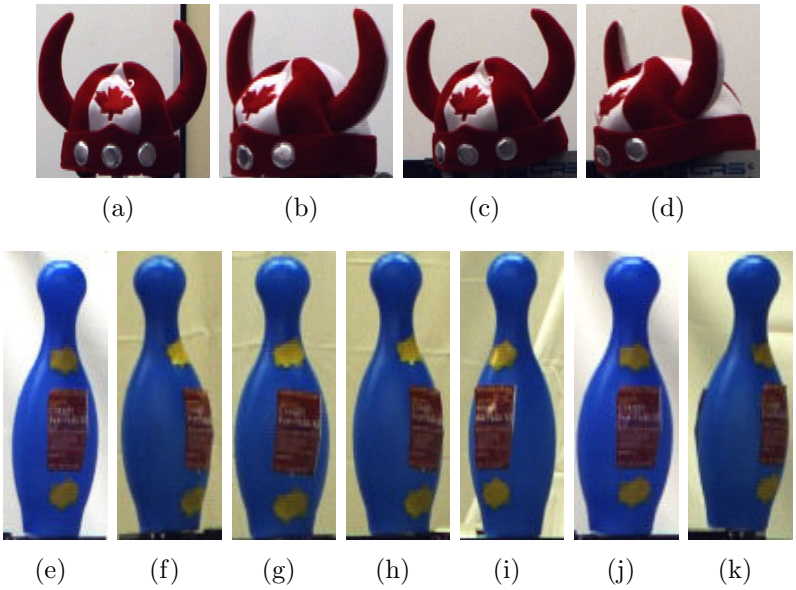



Figure 6.1: Experiment $1A$: views of the targets chosen manually to create the WM for visual search

During the test phase the search for targets are performed in four different environmental settings.

- Setting I: Camera position is same as the training case (i.e no change in camera perspective). The environment is same as the training case. The individual orientation of the target does not change significantly.
- Setting II: Camera position is same as the training case (i.e no change in camera perspective). The environment contains few other objects which share visual features with the target in one or more feature dimension(s). The individual orientation of the target does not change significantly
- Setting III: Camera position is different than the training case, approximately at $(7.6ft, 73^o, 57^o, 7^o, 0^o)$ (the $(\alpha, \beta)$ angles change during the course of the experiment). The individual orientation of the target changes significantly.
- Setting IV: Camera position is different than the training case, approximately at $(8.5ft, 65^o, 60^o, 13^o, 0^o)$ (the $(\alpha, \beta)$ angles change during the course of the experiment). The individual orientation of the target changes significantly.

Views of the targets in different orientations are obtained automatically by placing the target on a robotic manipulator and positioning the camera in front of it, as shown in Fig.

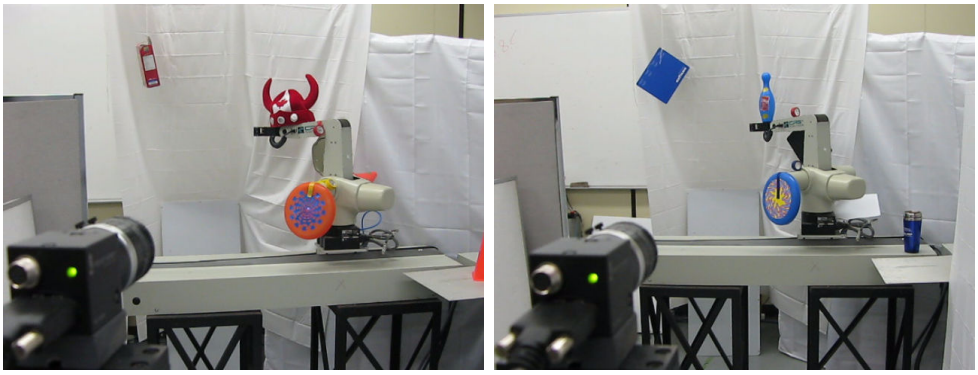


Figure 6.2: Experiment $1A$: Environmental setting during visual search for the targets

6.2. Note that the experiments here use the robotic manipulator only to rotate the targets in a controlled manner. The 'shoulder' joint of the manipulator performs a rotation of $200^o$ in $10^o$ steps causing different views of the targets to be presented to the camera gradually. During the experiments an occasional human intervention is made to control the amount of change in the target's individual orientation. In other words, when the orientation of the target changes heavily (due to movement of the manipulator), the target is rotated back (by a human) to one of the familiar orientations (approximately same as one of the training cases). Such human intervention evidently helps in target identification and leads to better success in visual search. The performance of visual search is quantified in terms of the number of views (out of $\frac{200^o}{10^o} = 20$ views) where the model successfully identifies the target. The quantity is expressed as a percentage and is termed as the success rate. To ensure the consistency of the results, 5 trials of each experiment are performed and the average of the success rates from the 5 trials is reported as the success rate of an experiment at a specific environmental setting. The results of these experiments are summarized in Table 6.1. A video of one trial of the second experiment in Table 6.1 (with the *blue pin* as target) is available in the multimedia file "Multimedia_ManualSearch.wmv" attached with this thesis.

The analysis of the results reported in Table 6.1 leads to the following conclusion.

1. Learning more views of the target enhances the search performance.

2. For a given number of learned target views, the search performance decreases when the camera perspective changes from the training case.

3. For a given number of learned target views, the search performance decreases if the target's orientation changes heavily from the training cases (the cases with no human intervention)

Table 6.1: Visual search with manually selected target views

| Experimental Specifications | | | Target | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Hat | | | Bowling pin | | |
| Setting | Human inter-vention | # of trial | # of training images | # of SIFT KP | Average success rate | # of training images | # of SIFT KP | Average Success rate |
| I | Yes | 5 | 1 Fig.6.1(a) | 120 | 50% | 1 Fig.6.1(e) | 27 | 22% |
| I | Yes | 5 | 4 Fig.6.1(a)-Fig.6.1(d) | 242 | 90% | 7 Fig.6.1(e)-Fig.6.1(k) | 163 | 80% |
| II | No | 5 | | | 50% | | | 60% |
| III | No | 5 | | | 45% | | | 55% |
| IV | No | 5 | | | 40% | | | 35% |

4. A considerable amount of manual works (starting from the training arrangement, up to human intervention during the test phase) is required to enhance the search performance.

As discussed in section 1.2, the required number of views to achieve maximum success rate in an arbitrary case of visual search is not known a priori. Besides, change of camera perspective is a fairly common event for any mobile robot. The objects in the real world might also appear with different orientations when the robot is in action. A better target learning strategy, therefore, is required to save a visual attention model from severe failure in visual search in case of real world applications.

**Experiment 1B**

This set of experiments are performed in the same environmental settings as that of the experiment $1A$ but here the target learning is performed using the proposed self-directed learning strategy. For the sake of performance comparison with experiment $1A$, these experiments do not use speech feedback from the human operator. A separate training phase, therefore, is used for target learning. The self-directed learning with speech feedback does not require a separate training phase and will be demonstrated later in this section. During the self-directed learning of the target the camera is exposed to the target which is placed on the manipulator. The 'shoulder' joint of the manipulator performs a rotation of $200^o$ in $10^o$ steps. The attention model runs in the novelty exploration mode and chooses only those views of the target which appear as 'novel' to learn and thereby developing a LTM of the target. The process runs autonomously without any human intervention. During this learning process the model chooses 6 views of the *red hat* and 10 views of the *bowling pin* to learn. The autonomously chosen views of the targets are shown in Fig. 6.3.

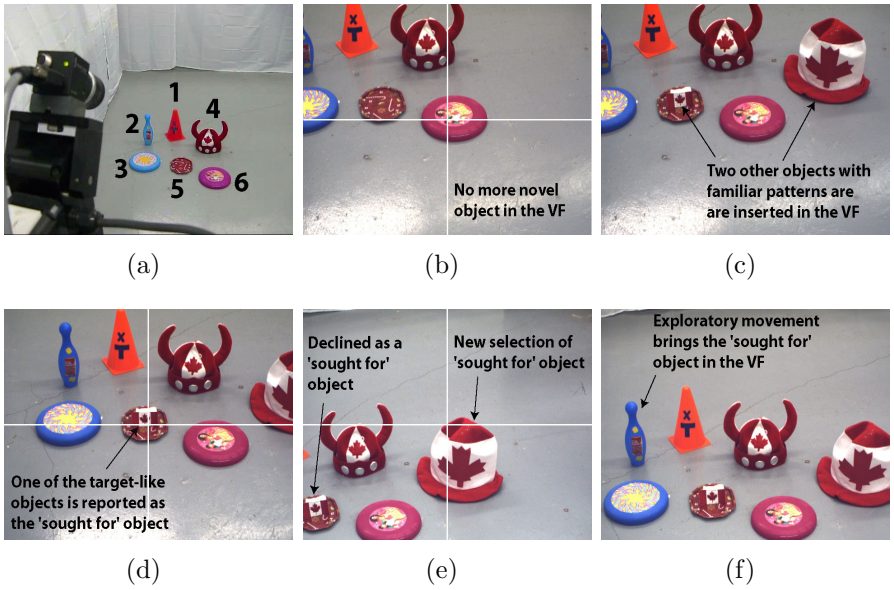
Figure 6.3: Experiment $1B$: views of the targets chosen autonomously by the robot using the proposed optimal learning strategy

Table 6.2: Visual search with the proposed self-directed target learning strategy

| Experimental Specifications | | | Target | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Hat | | | Bowling pin | | |
| Setting | Human inter-vention | # of trial | # of images learned | # of SIFT KP | Average success rate | # of images learned | # of SIFT KP | Average success rate |
| I | No | 5 | 6 (top) | 314 | 90% | 10 (bottom) | 253 | 95% |
| II | No | 5 | | | 90% | | | 88% |
| III | No | 5 | | | 85% | | | 75% |
| IV | No | 5 | | | 80% | | | 80% |

The LTM developed during the training stage is used as the WM for the visual search. The test phases run in the same way as that in experiment $1A$ but completely without human intervention. The results are summarized in Table 6.2. A video of one trial of the third experiment in Table 6.2 (with the *red hat* as the target) is available in the multimedia file "Multimedia_AutoSearch.wmv" attached with this thesis.

Comparison of the results reported in Tables 6.1 and 6.2 shows a significant average improvement of the visual search performance when the learning is performed using the proposed self-directed learning strategy. Besides, no manual training or human intervention are required here. The size of the WM is slightly larger in the case of self-directed learning (than that in manual learning). But according to the investigation reported in section 4.5.3 a small increase in memory size does not have any significant effect on the time




(a) (b) (c)

(d) (e) (f)

Figure 6.4: Experiment 2*A*: Integration of visual exploration and visual search (a) The experimental environment with several novel objects. The numbers denote the sequence at which different objects are attended (b)-(f) The visual fields of the robot at different stages of the experiment (please see text and the attached multimedia file "Multimedia_Generality.wmv" for detail)

performance of visual search.

## 6.3 Experiment 2: Generality

A set of experiments are performed to demonstrate the integration of visual exploration and visual search in the proposed multi-modal Bayesian model of attention. The experiments also demonstrate the effectiveness of the proposed attention-oriented speech-based HRI framework to resolve conflict in a visual search scenario. Such conflicts are very common in real-world applications of visual attention.

**Experiment** 2*A*

In this experiment the robot starts with an empty LTM and is exposed to the environment shown in Fig. 6.4(a). There are six objects and the robot attends to all of them because of their novelty. The sequence of attention is shown in Fig. 6.4(a).

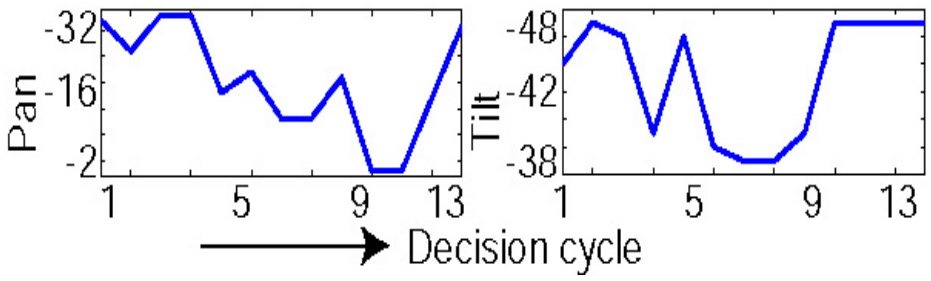
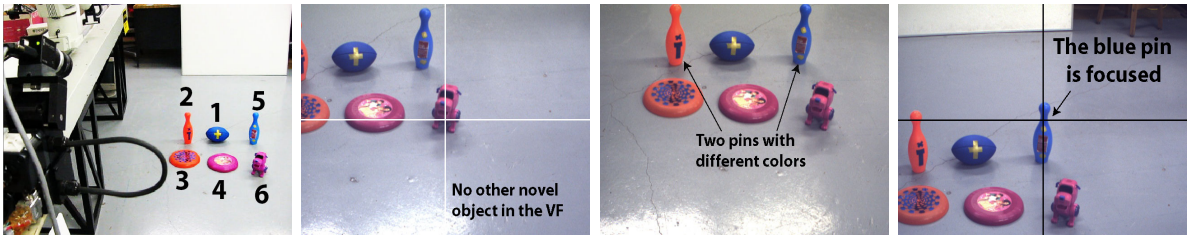



Figure 6.5: Experiment 2*A*: pan-tilt positions of the camera head

After focusing on each object the robot interacts with the human who provides the robot with the name and color of the focused object. In this way a LTM is developed which contains 524 SIFT keypoints under 5 different 'object labels' (*cone*, *pin*, *plate*, *hat*, and *toy*) and 4 color vectors under 4 different 'color labels' (*orange*, *blue*, *red*, and *purple*). After attending the *purple toy* (as shown in Fig. 6.4(b)) the robot can not identify any other novel object in the VF and the camera remains at its current position. At this point two other objects having very similar visual features as the *red hat* (attended fourth in the sequence) are inserted in the visual field of the robot (as shown in Fig. 6.4(c)). As both of these newly inserted objects have familiar patterns, the robot refrains itself from focusing on them. The human then triggers an interaction using 'Query' type speech input and asks the robot to look for a *red hat*. The robot extracts the color feature vector corresponding to the 'color label' *red* and 132 SIFT keypoints corresponding to the 'object label' *hat* from its LTM and creates a WM for the visual search task. The head-poses corresponding to all three of the *red hat*-like objects receive different amount of top-down modulating biases from the WM. The robot focuses on one of them as shown in Fig. 6.4(d) and interacts with the human for confirmation. The human uses 'Assistance' type speech input (e.g 'not this one') to negate the success of visual search. This nullifies the future candidacy of the head-pose corresponding to this object as a 'sought for' object for the ongoing request of visual search. The robot continues the search and focuses on another *red hat*-like object as shown in Fig. 6.4(e). This time the human affirms the success of the visual search. Consequently, the WM is cleared and the robot goes back to the visual exploration mode. But as there is no novel stimuli in the visual field, the camera remains at its current position. At this point the human further requests to look for a *blue pin*. Here the 'sought for' object is not within the visual field of the robot, as seen in Fig. 6.4(e). An exploratory movement in search of the 'sought for' object brings it within the visual field (as seen in Fig. 6.4(f)) and the robot focuses on the target (a default $(\pm 10^o, \pm 8^o)$ pan-tilt movement is predefined as the exploratory movement if a 'sought for' object can not be identified within the VF). Thus the visual search is performed along with visual exploration without having any separate training phase.




(a) (b) (c) (d)

Figure 6.6: Experiment $2B$: (a) The experimental environment with several novel objects. The numbers denote the sequence of attention (b)-(d) The VFs of the robot at different stages of the experiment (please see text for detail)

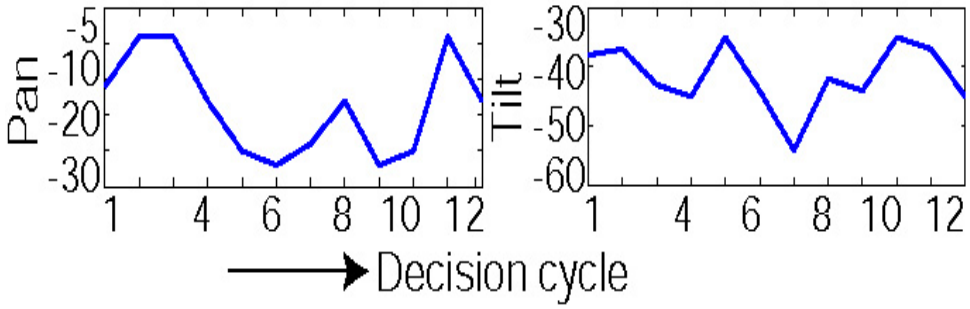



Figure 6.7: Experiment $2B$: pan-tilt positions of the camera head

A video of the experiment is available in the multimedia file "Multimedia_Generality.wmv" attached with this thesis. Fig. 6.5 shows the pan-tilt position of the camera head during the experiment. The camera angles are always expressed with respect to the base coordinate system (please see Fig. 5.2(b) for the coordinate systems).

### Experiment $2B$

This experiment shows an interesting case to demonstrate the effectiveness of the proposed model in real-world robotic applications. Here the robot learns a 'color label' at some point of time and uses this knowledge at a later time to differentiate between two objects who have, according to the robot's memory, similar 'object label' but no specific 'color label'.

During the experiment the robot starts with an empty LTM and is exposed to the environment shown in Fig. 6.6(a). In this case there are six objects with 4 distinct 'object labels' (*pin, ball, plate, toy*) and 3 distinct 'color labels' (*blue, orange, purple*). When the robot attends to the novel objects, the human operator mentions the 'color labels' only for the first four objects (i.e., *blue ball, orange pin, orange plate*, and *purple plate*).

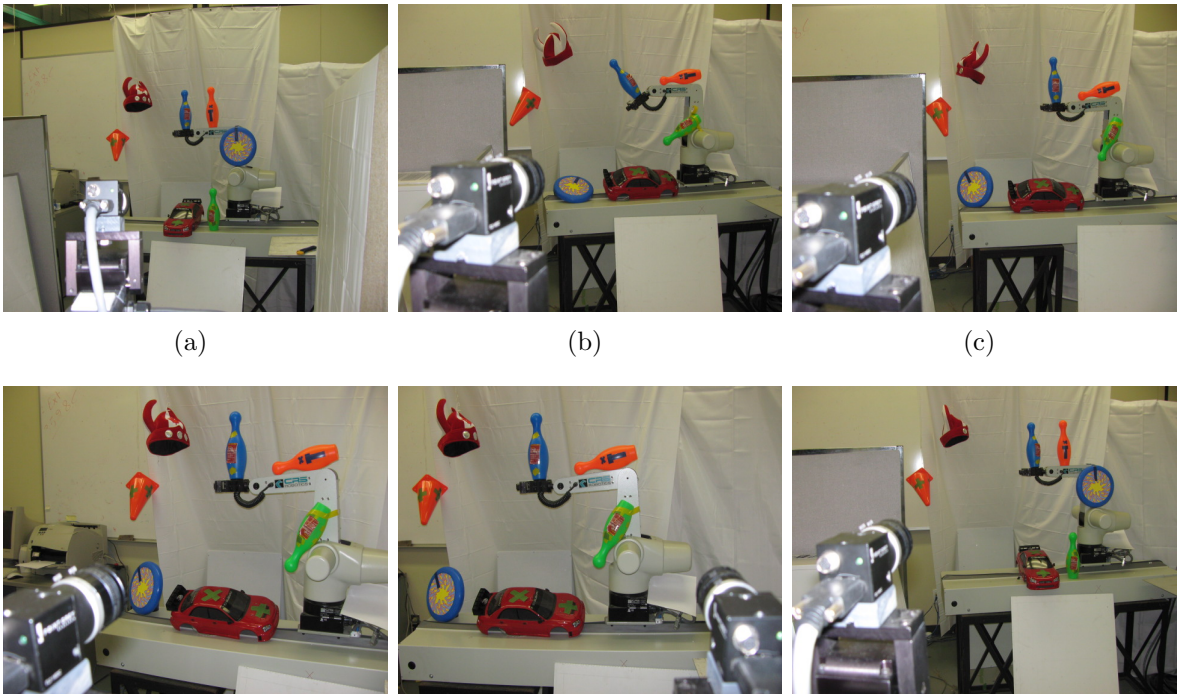
The fifth and sixth objects are introduced by their 'object labels' only (i.e., *pin* and *toy*, respectively). When there are no other novel object in the VF (as shown in Fig. 6.6(b)), the robot is asked sequentially to look for a *purple plate*, a *purple toy*, a *blue pin*, and an *orange pin*. The search for the *purple plate* and the *orange pin* was straightforward as the robot has knowledge of the colors and the SIFT keypoints that characterizes these two objects. The case, however, is a bit different for the *purple toy* and the *blue pin*. The robot had to apply its own sense of *purple* and *blue* color, developed from the color of the *purple plate* and the *blue ball*, to identify these two objects. Specially, the search for the *blue pin* is more interesting as the VF has two pins as shown in Fig. 6.6(c). The only way to prioritize the SIFT keypoints corresponding to the *blue pin* (in order to focus it instead of the *orange pin*) is through using the sense of color. The robot successfully performs that task and focuses on the *blue pin* (Fig. 6.6(d)). Fig. 6.7 shows the pan-tilt position of the camera head during the experiment.

Video of this experiment is available in multimedia file "Multimedia_Cross-modal.wmv" attached with this thesis.

## 6.4 Experiment 3: Prior Training

The goal of this experiment is to demonstrate the on-spot learning strategy discussed in section 5.5. At first a set of experiments are conducted to create the case of a mobile cognitive robot which is situated in a natural environment for an extended period of time and has gathered knowledge of different objects through the proposed visual attention model. This is done by exposing the robot to a set of objects from different locations in the world and allowing it to focus on and learn them while being guided by its sense of novelty. Fig. 6.8 shows different environmental settings at which the robot gathers information of seven different objects. A total of 1073 SIFT keypoints are stored in the LTM under 5 'object labels' (*car, pin, toy, cone, hat*) and 4 color vectors are stored under 4 'color labels' (*orange, blue, red, green*). Different orientations of the objects learned during this time are shown in Fig. 6.9

With such an enriched LTM the robot is then placed at the location $(4ft, 5^o, 39^o, 0^o, 0^o)$ where it is exposed to the environment shown in Fig. 6.10(a). There are eight objects among which one is novel. Among the seven familiar objects, five objects (*blue pin, green pin, red car, red hat, orange cone*) appear with orientations significantly different than that observed in Fig. 6.8. The robot first focuses on the novel object (the *purple dog*) and updates its LTM. The human then asks the robot to look for different familiar objects and it successfully identifies the *green pin*, the *blue toy*, the *red car*, the *orange pin*, and the *purple dog* after search requests were made for them (Figs. 6.11(a)- 6.10(f)). The robot, however, fails to immediately identify the *blue pin* (in Fig. 6.10(g)). During the



(a) (b) (c)

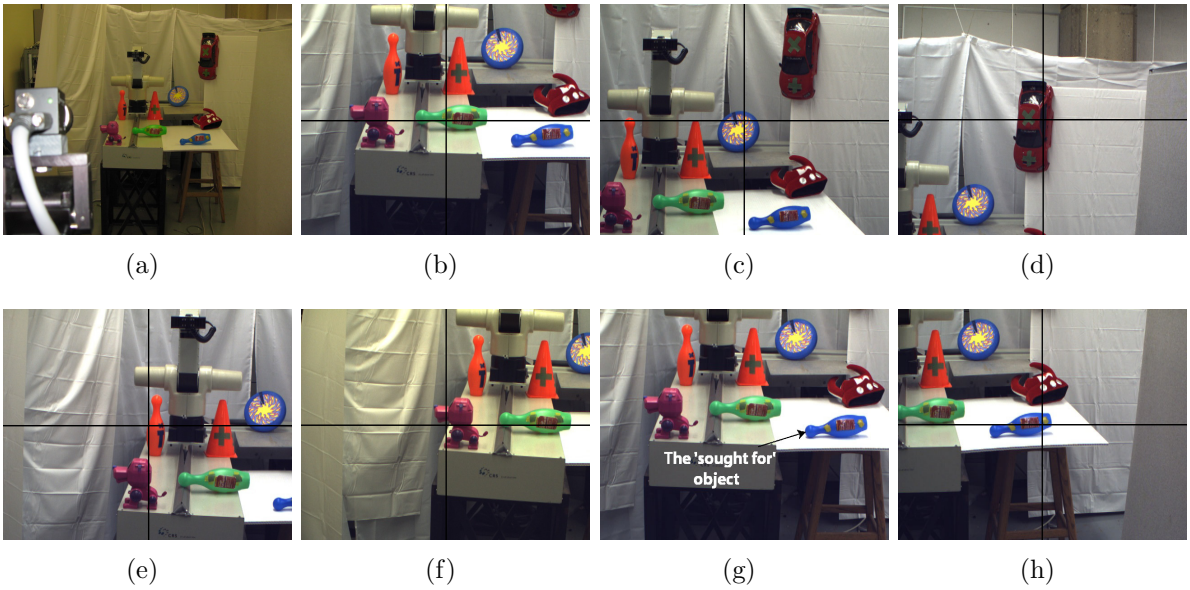
(d) (e) (f)

Figure 6.8: Experiment 3: The robot is exposed to a set of objects from different 3D locations in the world and learns about them while being guided by its continuously developing sense of novelty. The camera locations $(r, \theta, \phi, \alpha, \beta)$ are (a) $(6ft, 85^o, 50^o, -5^o, 0^o)$ (b) $(6.5ft, 68^o, 52^o, 15^o, 5^o)$ (c) $(6.5ft, 68^o, 52^o, -15^o, 5^o)$ (d) $(3ft, 75^o, 30^o, -5^o, 0^o)$ (e) $(3ft, 80^o, 30^o, 8^o, 5^o)$ (f) $(6.5ft, 68^o, 52^o, -15^o, 5^o)$

Figure 6.9: Experiment 3: On-spot learning. Different views of the objects chosen by the robot to learn. The SIFT keypoints corresponding to these views are stored in the LTM

search for the blue pin the robot first chooses the *blue toy* and then the *green pin* as the 'sought for' objects. The type of these two successive failures indicates that the robot has the knowledge of 'what is a *pin*' and 'what is *blue*' but it can not identify a *pin* which is *blue* in color. The human applies on-spot learning to teach the robot. First the robot is asked to look for any blue object (e.g by asking 'find something blue'). As the head-pose corresponding to the *blue toy* was already inhibited as a candidate for the next system state for the ongoing request of visual search, the robot focuses on the other blue object, i.e the *blue pin* (Fig. 6.10(h)). The human then affirms it as the correct choice. The robot updates its LTM with the new information about the *blue pin*. At a later time, this on-spot learning is checked by asking the robot to search for the *blue pin* again and the robot immediately identifies the target.

The visual search for the red hat was also failed initially but that was because of the object segmentation problem. Figs. 6.11(a) and 6.11(b) show the segmentation failure during the search for the *red hat*. Incorrect region growing causes the *red hat* to become a part of the large object blob which is discarded from further analysis of attention due to its large size. The robot selects the *red car* as the best match of the 'sought for' object and focuses on it. The human negates the success of visual search and allows the robot to continue the search. The segmentation problem no longer exists in the next decision cycle and the robot successfully identifies the red hat (Fig. 6.11(c)). Fig 6.12 shows the pan-tilt




(a) (b) (c) (d)

(e) (f) (g) (h)

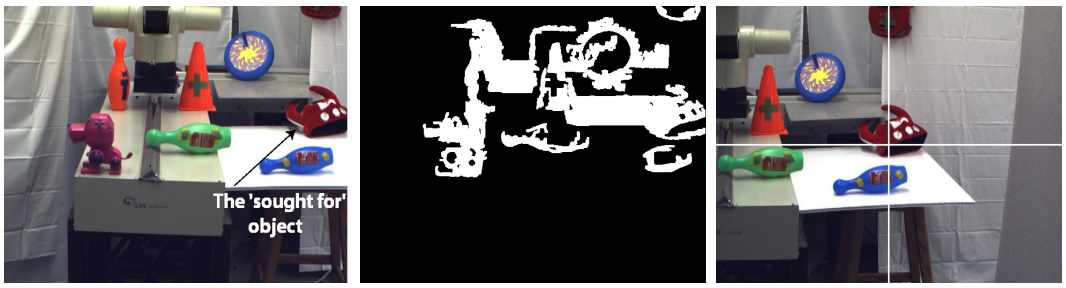
Figure 6.10: Experiment 3: on-spot learning. (a) The experimental environment. (b)-(f) Focusing on different 'sought for' objects: *green pin*, *blue toy*, *red car*, *orange pin*, and *purple dog*. (g) Failure to identify the *blue pin* due to large change in orientation. (h) The *blue pin* is focused through on-spot learning (please see text and the attached multimedia file "Multimedia_On-SpotLearning.wmv" for detail)

positions of the camera during the experiment.

Video of this experiment is available in the multimedia file "Multimedia_On-Spot Learning.wmv" attached with this thesis.

## 6.5 Experiment 4: Operator Burden

An experiment is performed where a robotic manipulator is used to point to the focused object. The Bumblebee stereo camera is used in this experiment for accuracy of hand-eye coordination. During this experiment, similar to experiment 3, both visual search and visual exploration are performed. But along with the visual orientation of the camera toward the focused object, a manipulator also points to the object of attention to make it easily identifiable to the human user (as shown in Fig. 6.13). This makes the task of providing high level information easier for the human user. Fig. 6.15 shows the camera position along with the position of the wrist and shoulder joints of the manipulator throughout the experiment. There are four novel objects in the VF (*orange cone, green pin, green toy, red hat*) and the robot attends to each of them due to their novelty as shown in




(a) (b) (c)

Figure 6.11: Experiment 3 (a) The visual field where the *red hat* is searched (b) Incorrect region growing causes the the *red hat* to become a part of the large object blob which is discarded from further analysis of attention due to its large size (c) The *red hat* is focused

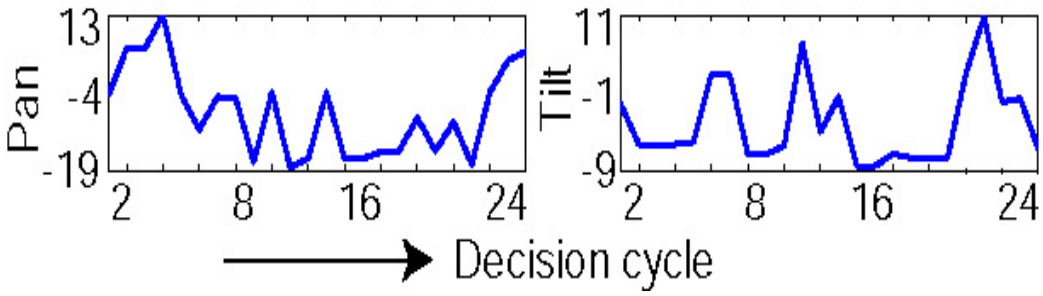



Figure 6.12: Experiment 3: pan-tilt positions of the camera-head

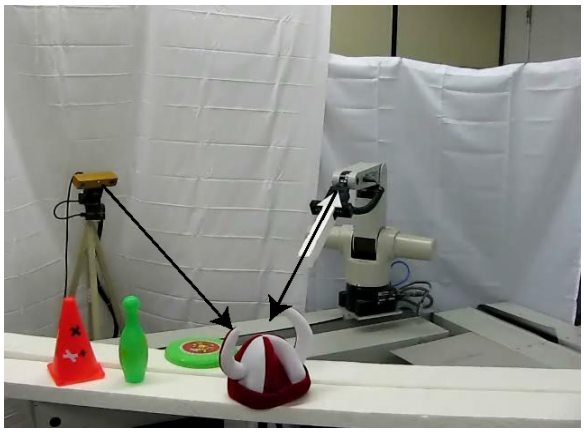
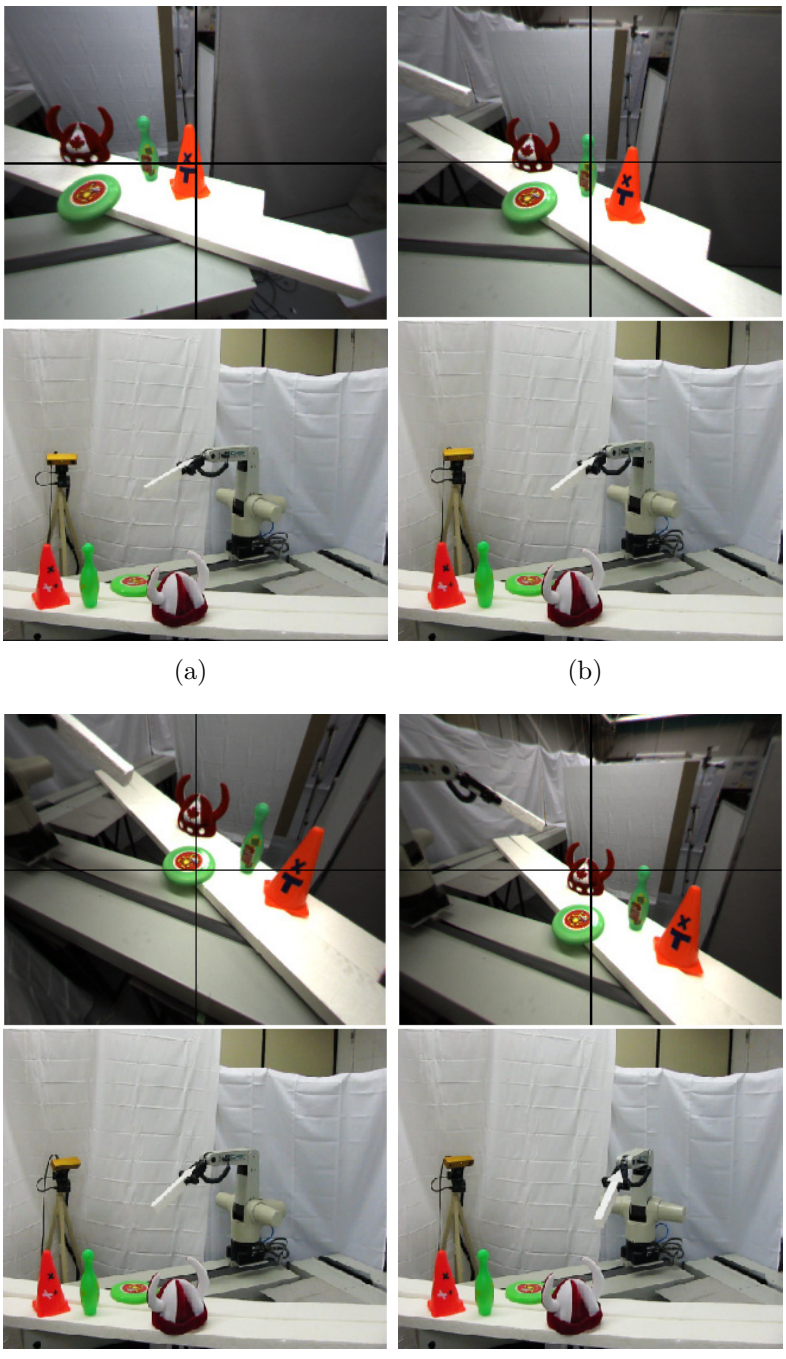


Figure 6.13: Experiment 4: The manipulator points to the object on which the camera focuses



(a) (b)

(c) (d)

Figure 6.14: Experiment 4: in each case the top image shows the camera focusing on an object (due to its novelty or search request for it) and the bottom image show the manipulator pointing to the focused object to make it more salient to the human operator (please see text and the attached multimedia file "Multimedia_AttentionAndPointing.wmv" for detail)

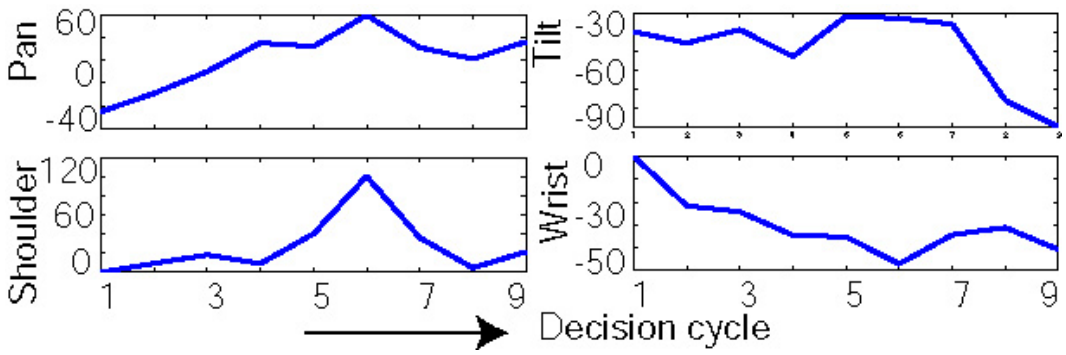



Figure 6.15: Experiment 4: Positions of the camera, shoulder and the wrist joint during the experiment. The shoulder and wrist angles are with respect to the world coordinate system

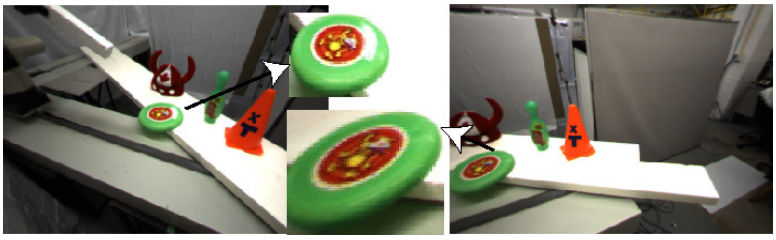


Figure 6.16: Experiment 4: Effect of lens distortion on the appearance of objects. The $(\alpha, \beta)$ angles of the camera are $(-25^o, -40^o)$ for the left image and $(35^o, -53^o)$ for the right image. The values of $(r, \theta, \phi)$ are the same for both cases. The non-affine stretching of the *green toy* in the right image makes it difficult to identify during the visual search

Fig. 6.14. The robot also fulfill the search request for two of the objects (*red hat, green toy*) and focuses on them. A video of this experiment is available in the multimedia file "Multimdedia_AttentionAndPointing.wmv" attached with this thesis.

The Bumblebee2, being a very wide angle camera, does not provide the detailed texture of the objects in its visual field. Fewer number of SIFT keypoints, therefore, are extracted from each object. Besides, the lens distortion is significant in wide angle optics. Fig. 6.16 demonstrates one example of the lens distortion effect. Fewer number of SIFT keypoints from each object as well as the strong changes in the object's appearance caused by lens distortion make visual search more challenging with wide angle cameras like Bumblebee2. On-spot learning, therefore, plays a significant role when dealing with lower resolution wide angle cameras.

## 6.6 Conclusion

This chapter has presented a set of experiments for self-evaluation of the proposed multimodal Bayesian model of visual attention. The experiments presented in this chapter show that the proposed model addresses all of the research issues involved with the robotic application of visual attention as discussed in section 1.2.

# Chapter 7

# Conclusion

Two objectives were set-up for the research presented in this thesis.

**Objective I:** Development of a bio-inspired model of visual attention for robotic cognition. Such a model is expected to have the following characteristics.

- The model should be able to execute overt attention with head-eye movement.
- The model should properly tackle the change in the content of the VF as well as the camera and image coordinate systems resulting from the execution of overt attention (*Issue 1* in section 1.2).
- The model should be able to run autonomously (if required, with minimum human supervision). Should it be required, the human supervision will be such that it does not interrupt the normal flow of operation of the model (*Issue 4* in section 1.2).
- The model should, as much as possible, be independent of any prior training such that the success of the model does not depend on the robustness of a training session (*Issue 3* and *Issue 5* in section 1.2).

**Objective II:** Implementation of the proposed model on a real robotic system and evaluation of its performance.

The research presented in this thesis fully observes these two objectives.

In relation to objective I, a comprehensive literature survey is first conducted to investigate the bio-inspiration in visual attention modeling, the ways of mimicking biological principles (related to visual attention) in a technical system, and the existing models of visual attention to identify their strengths and shortcomings. The analysis and the survey

has been presented in chapters 1 and 2 of this thesis. Based on this intense survey this thesis formally reports the following issues to be considered in the design of a robotic model of visual attention.

1. Consequences of the overt shift of focus: There are three consequences of overt shift of focus which are: 1) change of reference frame, 2)dynamic IOR, and 3) partial appearance of features. If not properly tackled, they might seriously hamper the visual attention behavior of a robot.

2. Integration of space- and object based analysis: To resolve conflicts in the implementation of many attention-related phenomena it is highly beneficial to consider both space and object as the elemental units of attentional selection.

3. Generality: The two modes of visual attention (visual exploration and visual search) should run in conjunction with each other. Separate processing for visual search and visual exploration is not desirable in robotic applications.

4. Prior training: The visual search of an object should not be preceded by an off-line training session for learning of the features of that object.

5. Optimal learning strategy: For successful visual search in the arbitrary settings (environment and camera perspective) there must be a learning strategy which ensures that the robot knows enough information about the target. Such learning must not violate the generality of the attention model and should not impose the burden of a prior training phase.

Based on this investigation this thesis proposes a Bayesian model of visual attention for robotic systems. The proposed model takes its inspiration from the theory of biased competition which is a widely accepted neurodynamic theory of visual attention in the primates. The Bayesian filter for attention proposes a robot-centric solution of visual attention where the robot is considered as an autonomous entity which attends to the stimulus of its choice. A set of possible choices of the robot for attention is derived from the tenets of biased competition. The choice of the stimuli to attend depends on the visual saliency of the stimuli as well as their relevance with the current behavior of the robot. The saliency and behavioral relevance are considered as the measurement data for the proposed visual attention model. The Bayes filter considers the head-pose of the robot as the state variable and recursively estimates the system state based on the current measurements. The Bayesian attention model has been implemented using a particle filter. The robot-centric approach of attention inherently addresses the issues resulting from the overt shift of focus mediated by head-eye movement. Besides, the particle-filter implementation integrates the space- and object- based analysis while predicting a head-pose as the next state of the visual

attention system. Details of the mathematics, operation, and implementation of the filter has been reported in chapter 3 of this thesis.

In relation to Objective II, the performance of the proposed Bayes filter for visual attention is evaluated in a number of real-world experiments conducted with a robotic camera-head. A set of criteria is defined for performance evaluation and the experiments are specifically designed to investigate the success of the proposed model with respect to its defined goal. The details of the experiments are reported in chapter 4 of this thesis. The analysis of the experimental results reveals that the proposed model integrates the space- and object-based analysis during attentional selection and successfully addresses the research issues involved with the overt shift of focus. The model, however, fails to address the issues of generality, prior training, and optimal learning strategy. In response to this failure, the Bayesian model described in chapters 3 and 4 is further extended to the multi-modal case where speech inputs from the human user are processed along with the visual information in order to develop an attention-oriented speech-based HRI framework. According to this framework the visual attention system of the robot maintains an occasional interaction with its user (or operator) to enhance its knowledge about the surrounding with multi-modal information. The interaction also enables the robot to switch back-and-forth between the two modes of attention. Besides, such an occasional interaction with the human assists to develop an optimal learning strategy for the robot. The detail of the multi-modal extension of the Bayes filter for attention is reported in chapter 5. A set of experiments are designed to evaluate the performance of the multi-modal Bayes filter for attention and are reported in chapter 6. Analysis of the experimental results shows that the proposed framework of visual attention successfully address all the research issues and fulfill the goals setup by this thesis.

## 7.1 The List of Publications

The research presented in this thesis has generated the following technical publications.

1. Momotaz Begum and Fakhri Karray, "Integrating Visual Exploration and Visual Search for Robotic Visual Attention: The Role of Human-Robot Interaction," Submitted to *IEEE International Conference on Intelligent Robots and Systems 2010*

2. Momotaz Begum and Fakhri Karray, "Visual Attention for Cognitive Robots: The Role of Multi-modality and Human Interaction," Submitted to *IEEE Transaction on System, Man, and Cybernetics. Part B*

3. Momotaz Begum and Fakhri Karray, "Visual Attention for Robotic Cognition: A Survey," Submitted to *IEEE Transaction on Autonomous Mental Development*

4. Momotaz Begum, Fakhri Karray, G. K. Mann, and R. Gosine, "A Probabilistic Model of Overt Visual Attention for Cognitive Robots,", Accepted for Publication In *IEEE Transaction on System, Man, and Cybernetics. Part B*,
DOI: 10.1109/TSMCB.2009.2037511

5. Momotaz Begum and Fakhi Karray, "Computational Intelligence Techniques in Bio-inspired Robotics," In *Computational Intelligence in Autonomous Robotic Systems, Springer 2008*, pp. 1-29.

6. Momotaz Begum, F. Karray, G. K. I. Mann, and R. G. Gosine, "A Probabilistic Approach for Attention-Based Multi- Modal Human-Robot Interaction," In *IEEE International Symposium on Robot and Human Interactive Communication 2009*, pp. 200-205.

7. Momotaz Begum, F. Karray, G. K. I. Mann, and R. G. Gosine, "Re-mapping of Visual Saliency in Overt Attention: A Particle Filter Approach for Robotic Systems," In *IEEE International Conference on Robotics and Bio-mimetic 2008*, pp. 425-430.

8. Momotaz Begum, George K. I. Mann, Raymond G. Gosine, and Fakhri Karray, "Object- and Space- based Visual Attention: An Integrated Framework for Autonomous Robots," In *IEEE/RSJ International Conference on Intelligent Robots and Systems 2008*, pp. 301-306.

9. Momotaz Begum, George K. I. Mann, and Raymond G. Gosine, "A Biologically Inspired Bayesian Model of Visual Attention for Humanoid Robots," In *IEEE-RAS International Conference on Humanoid Robots 2006*, pp. 587-592.

10. Rajibul Huq, Momotaz Begum, George K. I. Mann, and Raymond G. Gosine, "Biased Competitive Model of Humanoid Visual Attention Using Fuzzy Discrete Event System ," In *IEEE International Conference on Robotics and Biomimetics 2006*, pp. 1559-1564.

## 7.2 Future Works

### 7.2.1 The Model

There is a number of sectors where the Bayesian model proposed in this work can be further improved. They are summarize below.

- The top-down modulation model has the significant capacity to make any visually insignificant or less conspicuous stimulus as 'worthy to attend'. In order to receive

that favor from the top-down model, the stimuli, however, have to be selected by the bottom-up competition model at the first place. The bottom-up competition model in the current implementation considers only the color and intensity features to identify the potentially interesting regions in the snapshot of an environment. This causes the proposed model to consider only the colorful objects with noticeable texture and reasonable size (not too big or too small) for attention. Thus the stimuli that are less colorful or have very poor contrast with the surrounding have very small chance of getting selected by the bottom-up competition model. Use of more image features (e.g. motion, orientation) and intelligent analysis of them to define the parameters of the Gaussian mixture representing the bottom-up competition model will enable the model to identify regions in the visual field having small and less interesting stimuli to be identified as the potential candidate for attention. This is also a very legitimate conclusion with respect to the particle filter implementation. The bottom-up competition model serves as the proposal distribution for the Bayesian model of visual attention and, the better the proposal distribution is, the closer the weighted samples represent the true posterior.

- The image segmentation process has some influence on the operation of the top-down modulation model. Use of more robust segmentation techniques will make the algorithm more robust against the segmentation failure. For instance, the type of segmentation failure demonstrated in Fig. 6.11 of chapter 6 can be avoided through using improved segmentation techniques.

- The current implementation is operating with a very limited set of vocabulary and grammars for the speech-based HRI part. Improving the vocabulary and use of context-free grammar will make the visual attention model more general for any robotic application and will ease the interaction with the non-expert users. This will open up the possibilities of natural speech-based control of cognitive robots.

### 7.2.2 Research Direction

In the primates, visual attention is submerged in their perception, action, and in many of the cognitive functions. In addition to its trivial manifestation in the visual exploration and visual search, visual attention works underneath the action execution, planning, reasoning, and decision making process of the primates [153]. This ensures the survival and normal operation of the primates in their environment. Mimicking the visual attention of the primates in the robotic system will not be complete until we explore this hidden influence of attention in the overall cognition of the primates. That will also enable us to model the true nature of human cognition in a more realistic manner.

In robotics, the use of visual attention as a stand alone ability of the robot is far less appealing than the case where the visual attention operates in conjunction with the reasoning, decision making, and action planning of the robot. In all of these cases visual attention works at the perceptual level and passes only those information to the higher cognitive processes which are relevant to their respective requirements. The model of visual attention proposed in this thesis can easily blend in such a scenario. The proposed model maintains two different channels (bottom-up and top-down) to reflect the behavioral relevance of the stimuli on the attention behavior of the robot. Depending on the higher cognitive process it is cooperating with, the model can pass the appropriate visual stimuli to that process while blocking the others. For instance, when a robot is performing the task of manipulating an object, the relevant visual stimuli for the planning process is the current and the future position of the target object while the action execution process is only concerned about the visual features and locations of the graspable parts of the object. The role of the visual attention model, therefore, will be to deliver the position information of the object to the planning process and the visual feature-related information to the action execution process. A visual attention model can also invoke measurements from multiple sensors to further assist the operation of a cognitive process. The visual attention model, therefore, serves as the gate-keeper of information and thereby making the autonomous robots a bit more cognitive. Such robots have increasingly growing demand in service industries, assistive and health-care sectors, and entertainment industries.

# Bibliography


[1] R. A. Brooks, "Intelligence without representation," *Artificial Intelligence*, vol. 47, p. 139  160, 1991.

[2] M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishigurob, T. Inui, Y. Yoshikawa, M. Ogino, and C. Yoshida, "Cognitive developmental robotics: a survey," *IEEE Transaction on Autonomous Mental Development*, vol. 1, pp. 1 – 23, 2009.

[3] J. Weng, "Developmental robotics: Theory and experiments," *International Journal of Humanoid Robotics*, vol. 1, pp. 199 – 234, 2004.

[4] R. J. Brachman, "Systems that know what they're doing," *IEEE Intelligent Systems*, vol. 17, pp. 67 – 71, 2002.

[5] Y. Bar-Cohen and C. Breazeal, *Biologically Inspired Intelligent Robots*. SPIE Press, 2003.

[6] B. Webb and T. R. Consi, *Biorobotics*. The MIT press, 2001.

[7] C. Koch and S. Ullman, "Shifts in selective visual attention: toward the underlying neural circuitry," *Human Neurobiology*, vol. 4, pp. 219 – 227, 1985.

[8] G. Deco and T. S. Lee, "A unified model of spatial and object attention based on inter-cortical biased competition," *Neurocomputing*, vol. 44, pp. 775 – 781, 2002.

[9] G. Deco and E. T. Rolls, "A neurodynamical cortical model of visual attention and invariant object recognition," *Vision Research*, vol. 44, p. 621  642, 2004.

[10] L. J. Lanyon and S. L. Denham, "A model of active visual search with object-based attention guiding scan paths," *Neural Networks*, vol. 17, pp. 873 – 897, 2004.

[11] F. H. Hamker, *Distributed competition in directed attention*. Akademische Verlagsgesellschaft, 2000, pp. 39 – 44.

[12] P. F. Dominey and M. A. Arbib, "A cortico-subcortical model for generation of spatially accurate sequential saccades," *Cerebral Cortex*, vol. 2, pp. 153 – 175, 1992.

[13] A. Pouget and T. J. Sejnowski, "Spatial transformation in the parietal cortex using basis functions," *Journal of Cognitive Neuroscience*, vol. 9, pp. 222 – 237, 1997.

[14] R. P. N. Rao, "Bayesian inference and attentional modulation in the visual cortex," *Cognitive neuroscience and neuropsychology*, vol. 16, pp. 1843– 1848, 2005.

[15] G. Carpenter and S. Grossberg, "ART2: Self-organization of stable category recognition codes for analog input patterns," *Applied Optics*, vol. 26, no. 23, pp. 4919 – 4930, 1987.

[16] L. Itti and C. Koch, "Computational modeling of visual attention," *Nature Reviews: Neuroscience*, vol. 2, p. 194 203, 2001.

[17] ——, "A saliency based search mechanism for overt and covert shift of visual attention," *Vision Research*, vol. 40, pp. 1489 – 1506, 2000.

[18] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 1254 – 1259, 1998.

[19] V. Navalpakkam and L. Itti, "Top-down attention selection is fine-grained," *Journal of Vision*, vol. 6, pp. 1180 – 1193, 2006.

[20] J. K. Tsotsos, S. Culhane, Y. Winky, L. Yuzhong, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artificial Intelligence*, vol. 78, pp. 507 – 545, 1995.

[21] K. R. Cave, "The featuregate model of visual selection," *Psychological Research*, vol. 62, pp. 182 – 194, 1999.

[22] Y. Sun and R. Fisher, "Object-based visual attention for computer vision," *Artificial Intelligence*, vol. 146, pp. 77 – 123, 2003.

[23] S. Frintrop, *VOCUS: A Visual Attention System for Object Detection and Goal-directed Search*. Lecture Notes in Artificial Intelligence (LNAI), Vol. 3899, Springer Berlin/Heidelberg. ISBN: 3-540-32759-2, 2006.

[24] R. Milanese, "Detecting salient regions in an image: from biological evidence to computer implementation," Ph.D. dissertation, Univ. of Geneva, Switzerland, 1993.

[25] H. C. Nothdurft, "The role of features in pre-attentive vision: comparison of orientation, motion, and color cues," *Vision Research*, vol. 33, pp. 1937 – 1958, 1993.

[26] M. I. Posner and Y. Cohen, *Components of visual orienting.* Hillsdale, NJ, 1984, pp. 531 – 556.

[27] M. H. Johnson, *Developmental Cognitive Neuroscience.* Blackwell Publishing, 2005.

[28] C. A. Nelson, M. Hann, and K. M. Thomas, *Neuroscience of cognitive development: the role of experience and the developing brain.* John Wiley & Sons: New Jersey, 2006.

[29] E. T. Rolls and G. Deco, *Computational Neuroscience of Vision.* Oxford University Press, 2002.

[30] J. P. Gottlieb, M. Kusunoki, and M. E. Goldberg, "The representation of visual saliency in monkey parietal cortex," *Nature*, vol. 391, pp. 481 – 484, 1998.

[31] A. A. Kustov and D. L. Robinson, "Shared neural control of attentional shifts and eye movements," *Nature*, vol. 384, pp. 74 – 77, 1996.

[32] K. G. Thompson and J. D. Schall, "Antecedents and correlates of visual detection and awareness in macaque prefrontal cortex," *Vision Research*, vol. 40, pp. 1523 – 1538, 2000.

[33] S. Frintrop and P. Jensfelt, "Active gaze control for attentional visual slam," in *IEEE International Conference on Robotics and Automation*, 2008, pp. 3690–3697.

[34] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Annual Reviews of Neurscience*, vol. 18, pp. 193 – 222, 1995.

[35] J. M. Wolfe, "Guided search 4.0: A guided search model that does not require memory for rejected distractor," *Journal of Vision*, vol. 1, p. 349a, 2001.

[36] W. James, *Principles of Psychology, Vol I*, 1890.

[37] R. Milanese, H. Wechsler, S. Gil, J. Bost, and T. Pun, "Integration of top-down and bottom-up cues for visual attention using non-linear relaxation," in *IEEE International Conference on Computer vision and pattern recognition*, 1994, pp. 781 – 785.

[38] J. K. Tsotsos, Y. Liua, J. C. M. Trujilloc, M. Pomplund, E. Siminea, and K. Zhoua, "Attending to visual motion," *Computer Vision and Image Understanding*, vol. 100, pp. 2 – 40, 2005.

[39] A. M. Treisman and G. Gelade, "A feature integration theory of attention," *Cognitive Psychology*, vol. 12, pp. 97 – 136, 1980.

[40] J. M. Wolfe, K. Cave, and S. Franzel, "Guided search: an alternative to the feature integration model for visual search," *Journal of Experimental psychology: Human perception and performance*, vol. 15, pp. 419 – 433, 1989.

[41] R. A. Rensink, "The dynamic representation of scenes," *Visual Cognition*, vol. 7, pp. 17 – 42, 2000.

[42] G. D. Logan, "The CODE theory of visual attention: an integration of space-based and object-based attention," *Psychological review*, vol. 103, pp. 603 – 649, 1996.

[43] S. Yantis and J. T. Serences, "Cortical mechanisms of space-based and object-based attentional control," *Current Opinion in Neurobiology*, vol. 13, pp. 187 – 193, 2003.

[44] M. Goldsmith and M. Yeari, "Modulation of object-based attention by spatial focus under endogenous and exogenous orienting," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 29, pp. 897 – 918, 2003.

[45] J. Duncan, "Selective attention and the organization of visual information," *Journal of Experimental Psychology*, vol. 113, pp. 501 – 513, 1984.

[46] B. J. Schol, "Objects and attention: the state of the art," *Cognition*, vol. 80, pp. 1 – 46, 2001.

[47] F. Orabona, G. Metta, and G. Sandini, "Object-based visual attention: A model for a behaving robot," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[48] T. Wu, J. Gao, and Q. Zhao, "A computational model of object-based selective visual attention mechanism in visual information acquisition," in *IEEE Conference on Information Acquisition*, 2004, pp. 405 – 409.

[49] M. Begum, G. Mann, R. Gosine, and F. Karray, "Object- and space-based visual attention: An integrated framework for autonomous robots," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2008, pp. 301 – 306.

[50] M. Z. Aziz and B. Mertsching, "Fast and robust generation of feature maps for region-based visual attention," *IEEE Transaction on image processing*, vol. 17, pp. 633 – 644, 2008.

[51] J. Ruesch, M. Lopes, A. Bernardino, J. Hornstein, J. S. Victor, and R. Pfeifer, "Multi modal saliency-based bottom-up attention: A framework for the humanoid robot icub," in *IEEE International conference on Robotics and automation*, 2008, pp. 962 – 967.

[52] K. A. Fleming and R. E. B. R. A. Peter II, "Image mapping and visual attention on a sensory ego-sphere," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2006, pp. 241 – 246.

[53] J. L. Crespo, A. Faina, and R. J. Duro, "An adaptive detection/attention mechanism for real time robot operation," *Neurocomputing*, vol. 72, pp. 850 – 860, 2009.

[54] J. M. Canas, M. M. Casa, and T. Gonzalez, "An overt visual attention mechanism based on saliency dynamics," *International Journal of Intelligent Computing in Medical Sciences and Image Processing*, vol. 2, pp. 93 – 100, 2008.

[55] M. Begum, F. Karray, G. Mann, and R. Gosine, "Re-mapping of visual saliency in overt attention: A particle filter approach for robotic systems," in *IEEE International Conference on Robotics and Biomimetics*, 2008.

[56] A. Haasch, N. Hofemann, J. Fritsch, and G. Sagerer, "A multi-modal object attention system for a mobile robot," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2005, pp. 2712– 2717.

[57] P. Mcguire, J. Fritsch, J. J. Steil, F. Roethling, G. A. Fink, S. Wachsmuth, G. Sagerer, and H. Ritter, "Multi-modal human-machine communication for instructing robot grasping tasks," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2002, pp. 1082 – 1088.

[58] H. Sumioka, K. Hosoda, Y. Yoshikawa, and M. Asada, "Acquisition of joint attention through natural interaction utilizing motion cues," *Advanced Robotics*, vol. 21, pp. 983– 999, 2007.

[59] K. Hasoda, H. Sumioh, A. Morita, and M. Asada, "Acquisition of human-robot joint attention through real-time natural interaction," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2004, pp. 2867 – 2872.

[60] Y. Nagai, K. Hosoda, and M. Asada, "Joint attention emerges through bootstrap learning," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2003, pp. 168 – 173.

[61] Y. Nagai, M. Asada, and K. Hosoda, "A developmental approach accelerates learning of joint attention," in *Proceedings of IEEE International Conference on Development and Learning*, 2002, pp. 277 – 282.

[62] M. J. Webster and L. G. Ungerleider, *Neuroanatomoy of visual attention.* Cambridge, MA: MIT press, 1998, pp. 19 – 34.

[63] J. Colombo, "The development of visual attention in infancy," *Annual Reviews, Psychology*, vol. 52, pp. 337 – 367, 2001.

[64] M. I. Posner, "The attention system of the human brain," *Annual Reviews, Neuroscience*, vol. 13, pp. 25 – 42, 1990.

[65] S. Kastner and L. G. Ungerleider, "The mechanisms of visual attention in the human cortex," *Annual Review of Neuroscience*, vol. 23, pp. 315 – 341, 2000.

[66] R. Desimone and L. G. Ungerleider, *Neural mechanisms of visual processing in monkeys*. Elsevier: NY, 1989, vol. 2, pp. 267 – 299.

[67] R. Desimone, E. K. Miller, and L. Chelazzi, *Interactions of neural systems for attention and memory*. Cambridge, MA: MIT press, 1994, pp. 75 – 91.

[68] L. Chelazzi, J. Duncan, E. K. Miller, and R. Desimone, "Responses of neurons in inferior temporal cortex during memory guided visual search," *Journal of Neurophysiology*, vol. 80, no. 6, pp. 2918 – 2940, 1998.

[69] L. Chelazzi, E. K. Miller, J. Duncan, and R. Desimone, "A neural basis for visual search in inferior temporal cortex," *Nature*, vol. 363, pp. 345 – 347, 1993.

[70] M. I. Posner, *Structures and functions of slective attention*. Washington, DC: American Psychological Association, 1988, pp. 173 – 202.

[71] M. H. Jhonson, *The development of visual attention: a cognitive neuroscience perspective*. Cambridge, MA: MIT Press, 1995, pp. 735 – 747.

[72] R. Egly, J. Driver, and R. Rafal, "Shifting attention between objects and locations: evidence from normal and parietal lesion subjects," *Journal of experimental Psychology*, vol. 123, pp. 161 – 177, 1994.

[73] J. Driver and G. C. Baylis, *Attention and visual object segmentation*. Cambridge, MA: MIT press, 1998, pp. 299 – 326.

[74] K. Nakayama and J. Joseph, *Attention, pattern recognition, and pop-out effect in visual search*. Cambridge, MA: MIT press, 1998, pp. 279 – 326.

[75] S. A. Adler and J. Orprecio, "The eyes have it: visual pop-out in infants and adults," *Developmental Science*, vol. 9, pp. 189 – 206, 2006.

[76] J. L. Dannemiller, "Motion pop-out in selective visual orienting at 4.5 but not at 2 months in human infants," *Infancy*, vol. 8, pp. 201 – 216, 2005.

[77] M. I. Posner, C. R. R. Snyder, and B. J. Davidson, "Attention and the detection of signals," *Journal of Experimental Psychology: General*, vol. 109, pp. 160 – 174, 1980.

[78] L. G. Ungerleider, "Functional brain imaging studies of cortical mechanisms of memory," *Science*, vol. 270, pp. 769 – 775, 1995.

[79] J. Duncan and G. W. Humphreys, "Visual search and stimulus similarity," *Psychologycal review*, vol. 96, pp. 433 – 458, 1989.

[80] C. Bundense, "A theory of visual attention," *Psychological Review*, vol. 97, p. 523 527, 1990.

[81] P. T. Quinlan, "Visual feature integration theory: Past, present, and future," *Psychological Bulletin*, vol. 129, p. 643 673, 2003.

[82] J. M. Wolfe, "Guided search 2.0: A revised model of visual search," *Psychonomic Bulletin and Review*, vol. 1, pp. 202 – 238, 1994.

[83] J. M. Wolfe and G. Grancarz, *Guided Search 3.0: Basic and clinical applications of vision science*. Kluwer Academic: Netherlands, 1996, pp. 189 – 192.

[84] M. P. Oeffelen and P. G. Voss, "Configurational effects on the enumeration of dots: counting by groups," *Memory and Cognition*, vol. 10, pp. 396 – 404, 1982.

[85] D. Heinke and G. W. Humphreys, *Computational models of visual selective attention: a review*. Psychology press, pp. 273 – 312.

[86] J. K. Tsotsos, L. Itti, and G. Rees, *A Brief and Selective History of Attention*. Elsevier Academic Press, 2005.

[87] S. Kastner and L. G. Ungerleider, "The neural basis of biased competition in human visual cortex," *Neuropsychologia*, vol. 39, pp. 1263 – 1276, 2001.

[88] S. Luck, L. Chelazzi, S. A. Hillyard, and R. Desimone, "Neural mechanisms for directed visual attention," *Journal of Neurophysiology*, vol. 77, pp. 24 – 42, 1997.

[89] G. Deco, *Biased competitive mechanisms for visual attention in a multi-modular neurodyanmical system*. LNAI 2036: Springer-Verlag Berlin, 2001, pp. 114 – 126.

[90] L. J. Lanyon and S. L. Denham, "A biased competition computational model of spatial and object-based attention mediating active visual search," *Neurocomputing*, vol. 58-60, pp. 655 – 662, 2004.

[91] F. H. Hamker, *Modeling Attention: From Computational Neuroscience to Computer Vision*. Springer-Verlag Berlin, 2005, pp. 118 – 132.

[92] L. Itti and P. Baldi, "A principled approach to detecting surprising events in video," in *IEEE International Conference on Computer vision and pattern recognition*, 2005, pp. 631 – 637.

[93] S. Frintrop and P. Jensfelt, "Attentional landmarks and active gaze control for visual slam," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1054 – 1065, 2008.

[94] S. J. Dickinson, H. I. Christensen, J. K. Tsotsos, and G. Olofsson, "Active object recognition integrating attention and viewpoint control," *Computer Vision And Image Understanding*, vol. 67, pp. 239 – 260, 1997.

[95] H. Deubel and W. X. Schneider, "Saccade target selection and object recognition: Evidence for a common attentional mechanism," *Vision Research*, vol. 36, pp. 1827 – 1837, 1996.

[96] R. Johnason, G. Westling, A. Backstrom, and J. Flanagan, "Eye-hand coordination in object manipulation," *The Journal of Neuroscience*, vol. 21, pp. 6917 – 6932, 2001.

[97] J. M. Findlay and I. D. Gilcrist, *Active vision perspective*. Springer: Verlag, 2001, pp. 83 – 103.

[98] J. Aloimonos, I. Weiss, and A. Badyopadhyay, "Active vision," *International Journal of Computer Vision*, vol. 1, pp. 333–356, 1988.

[99] R. Bajscy, "Active perception," *Proceedings of the IEEE*, vol. 76, pp. 996–1005, 1988.

[100] D. Ballard, "Animate vision," *Artificial Intelligence*, vol. 48, pp. 57–86, 1991.

[101] J. J. Clark, "Spatial attention and saccadic camera motion," in *Proceedings of IEEE International Conference on Robotics and Automation*, 1998, pp. 3247 – 3252.

[102] J. J. Clark and N. J. Ferrier, "Modal control of an attentive vision system," in *Proceedings of IEEE International Conference on Computer Vision*, 1988, pp. 514 – 523.

[103] L. Manfredi, E. S. Maini, P. Dario, C. Laschi, B. Girard, N. Tabareau, and A. Berthoz, "Implementation of neurophysiological model of saccadic eye movements on an anthropomorphic robotic head," in *IEEE/RSJ International Conference on Humanoid Robotics*, 2006, pp. 438 – 443.

[104] D. Coombs and C. Brown, "Real-time smooth pursuit tracking for a moving binocular robot," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1992, pp. 23 – 28.

[105] C. Brown, "Prediction and cooperation in gaze control," *Biological Cybernetics*, vol. 63, pp. 61 – 70, 1990.

[106] J. A. Driscoll, R. A. Peters, and K. R. Cave, "A visual attention network for a humanoid robot," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1998, pp. 1968 – 1974.

[107] A. Ude, V. Wyart, L.-H. Lin, and G. Cheng, "Distributed visual attention on a humanoid robot," in *IEEE-RAS International Conference on Humanoid Robots*, 2005, pp. 381 – 386.

[108] S. Vijayakumar, J. Conrad, T. Shibata, and S. Schaal, "Overt visual attention for a humanoid robot," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2001, pp. 2332 – 2337.

[109] G. Metta, "An attentional system for humanoid robot exploiting space variant vision," in *IEEE-RAS International Conference on Humanoid Robots*, 2001.

[110] A. Dankers, N. Barnes, and A. Zelinsky, "A reactive vision system: Active-dynamic saliency," in *Proceedings of International Conference on Computer Vision Systems*, 2007.

[111] C. Breazeal, A. Edsinger, P. Fitzpatrick, and B. Scassellati, "Active vision for sociable robots," *IEEE Transaction on System, Man, and Cybernetics, Part A*, vol. 31, pp. 443 – 453, 2001.

[112] R. Fay, U. Kaufmann, and A. Knoblauch, *Combining Visual Attention, Object Recognition and Associative Information Processing in a Neurobotic System.* Springer-Verlag, 2005, pp. 117 – 142.

[113] J. Vitay, N. P. Rougier, and F. Alexandre, *A Distributed Model of Spatial Visual Attention.* Springer-Verlag, 2005, pp. 54 – 72.

[114] O. Stasse, Y. Kuniyoshi, and G. Cheng, "Development of a biologically inspired real-time visual attention system," in *IEEE International Workshop on Biologically Motivated Computer vision*, 2000, pp. 150 – 159.

[115] F. Saidi, O. Stasse, and K. Yokoi, "A visual attention framework for search behavior by a humanoid robot," in *IEEE International conference on Humanoid robots*, 2006, pp. 346 – 351.

[116] L. Aryananda, "Attending to learn and learning to attend for a social robot," in *IEEE International conference on Humanoid robots*, 2006, pp. 618 – 623.

[117] N. J. Butko, L. Zhang, G. W. Cottrell, and J. R. Movellan, "Visual saliency models for robot cameras," in *IEEE International conference on Robotics and automation*, 2008, pp. 2398 – 2403.

[118] A. Elfes, "Sonar-based real-world mapping and navigation," *IEEE Journal of Robotics and Automation*, vol. 3, pp. 249 – 265, 1987.

[119] J. M. Canas, M. M. Casa, P. Bustos, and P. Bachiller, "Overt visual attention inside jde control architecture," in *Portuguese conference on Artificial intelligenc*, 2005, pp. 226 – 229.

[120] A. Davison and D. Murray, "Simultaneous localization and map-building using active vision," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 7, pp. 865 – 880, 2002.

[121] S. Lang, M. Kleinehagenbrock, S. Hohenner, J. Fritsch, G. A. Fink, and G. Sagerer, "Providing the basis for human-robot-interaction: A multi-modal attention system for a mobile robot," in *International Conference on Multi modal Interfaces*, 2003, pp. 28 – 35.

[122] M. Bollmann, R. Hoischen, M. Jesikiewicz, C. Justkowski, and B. Mertsching, "Playing domino: A case study for an active vision system," in *First International Conference on Computer Vision Systems*, vol. LNCS 1542, 1999, pp. 392–411.

[123] O. Dniz, M. Castrilln, J. Lorenzo, M. Hernndez, and J. Mndez, "Multimodal attention system for an interactive robot," in *Lectures Notes in Computer Science, vol. 2652. First Iberian Conference on Pattern Recognition and Image Analysis*, 2003, pp. 212 – 220.

[124] F. Kaplan and V. V. Hafner, "The challenges of joint attention," *Interaction Studies*, vol. 7, pp. 135 – 169, 2006.

[125] M. Oginoa, H. Toichia, Y. Yoshikawaa, and M. Asadaa, "Interaction rule learning with a human partner based on an imitation faculty with a simple visuo-motor mapping," *Robotics and Autonomous Systems*, vol. 54, p. 414  418, 2006.

[126] M. W. Doniec, G. Sun, and B. Scassellati, "Active learning of joint attention," in *IEEE/RAS International Conference on Humanoid Robots*, 2006, pp. 34 – 39.

[127] A. P. Shon, J. J. Storz, and R. P. N. Rao, "Toward a real-time bayesian imitation system for a humanoid robot," in *IEEE International Conference on Robotics and Automation*, 2007, p. 2847  2852.

[128] M. W. Hoffman, D. B. Grimes, A. P. Shon, and R. P. N. Rao, "A probabilistic model of gaze imitation and shared attention," *Neural network*, vol. 19, pp. 299– 309, 2006.

[129] M. Ito and J. Tani, "On-line imitative interaction with a humanoid robot using a dynamic neural network model of a mirror system," *Adaptive Behavior*, vol. 12, pp. 93 – 115, 2004.

[130] ——, "Joint attention between a humanoid robot and users in imitation game," in *IEEE International Conference on Development and Learning*, 2004, pp. 277 – 282.

[131] Y. Yoshikawa, T. Nakano, M. Asada, and H. Ishiguro, "Multimodal joint attention through cross facilitative learning based on x principle," in *IEEE International Conference on Development and Learning*, 2008, pp. 226–231.

[132] M. Lopes and J. S. Victor, "A developmental roadmap for learning by imitation in robots," *IEEE Trans. Syst., Man, Cybern. B*, vol. 37, pp. 308 – 321, 2007.

[133] J. S. Bruner and L. Postman, "On the perception of incongruity: A paradigm," *Journal of Personality*, vol. 18, pp. 206–223, 1949.

[134] Y. Nagai and K. J. Rohlfing, "Computational analysis of motionese toward scaffolding based robot action learning," *IEEE Transaction on Autonomous Mental Development*, 2009.

[135] J. H. Reynolds, L. Chelazzi, and R. Desimone, "Competitive mechanisms subserve attention in macaque areas V2 and V4," *Journal of Neuroscience*, vol. 19, pp. 1736 – 1753, 1999.

[136] R. Fantz, "Visual experience in infants: Decreased attention to familiar patterns relative to novel ones," *Science*, vol. 146, pp. 364 – 370, 1964.

[137] L. E. Bahrick, M. H.-R.  , and J. N. Pickens, "The effect of retrieval cues on visual preferences and memory in infancy: Evidence for a four-phase attention function," *Journal of Experimental Child Psychology*, vol. 67, pp. 1 – 20, 1997.

[138] R. Shiffrin and W. Schneider, "Controlled and automatic human information processingII: Perceptual learning, automatic attending, and a general theory," *Psychological Review*, vol. 84, pp. 127 – 190, 1977.

[139] D. J. Felleman and D. C. V. Essen, "Distributed hierarchical processing in the primate cerebral cortex," *Cerebral cortex*, vol. 1, pp. 1 – 47, 1991.

[140] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91 – 110, 2004.

[141] ——, "Local feature view clustering for 3d object recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2001, pp. 682 – 688.

[142] M. A. Tanner, *Tools for statistical inference : methods for the exploration of posterior distributions and likelihood functions*. Springer: New York, 1996.

[143] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Artech House, 2004.

[144] B. Kroose and B. Julesz, "The control and speed of shifts of attention," *Vision Research*, vol. 29, pp. 1607 – 1619, 1989.

[145] M. S. Peterson, A. F. Kramer, R. F. Wang, D. E. Irwin, and J. S. McCarley, "Visual search has memory," *Psychological Science*, vol. 12, pp. 287 – 292, 2002.

[146] V. Navalpakkam, J. Robesco, and L. Itti, "Modeling the influence of task on attention," *Vision Research*, vol. 45, pp. 205 – 231, 2005.

[147] S. May, M. Klodt, E. Rome, and R. Breithaupt, "GPU-accelerated affordance cueing based on visual attention," in *IEEE International Conference on Intelligent Robots and Systems*, 2007, pp. 3385–3390.

[148] A. Dankers, "Real-time synthetic primate vision," Ph.D. dissertation, Australian National University, Canberra, 2007. [Online]. Available: http://users.rsise.anu.edu.au/~andrew/papers.html

[149] N. Ouerhani, R. V. Wartburg, and R. Muri, "Empirical validation of the saliency-based model of visual attention," *Electronic Letters on Computer Vision and Image Analysis*, vol. 3, pp. 13– 24, 2004.

[150] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vision Research*, vol. 42, pp. 107– 123, 2002.

[151] F. Shic and B. Scassellati, "A behavioral analysis of computational models of visual attention," *International journal of computer vision*, vol. 73, pp. 159–177, 2007.

[152] J. R. Williamson, "Gaussian ARTMAP: A neural network for fast incremental learning of noisy multidimensional maps," *Neural Networks*, vol. 9, pp. 881 – 897, 1996.

[153] D. Drubach, *The brain explained*. Prentice Hall Health: New Jersey, 2000.