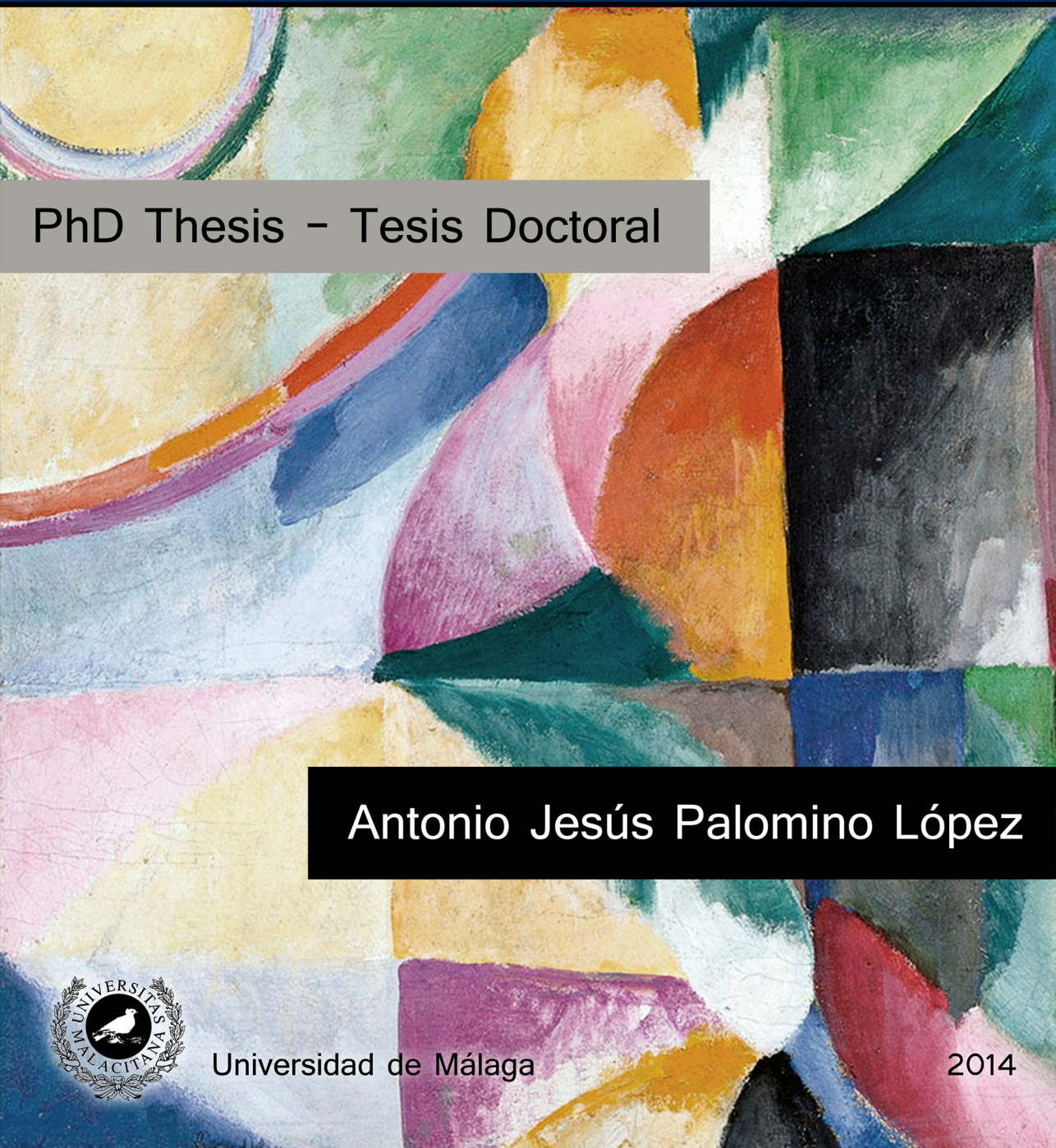
A Cognitive architecture for an Attention-based and BIdirectional Loop-closing DOmain

[CABILDO]



PhD Thesis - Tesis Doctoral

Antonio Jesús Palomino López

Universidad de Málaga

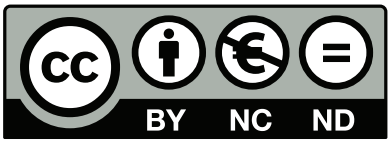
2014

UNIVERSIDAD DE MÁLAGA

Publicaciones y Divulgación Científica

AUTOR: Antonio Jesús Palomino López

EDITA: Publicaciones y Divulgación Científica. Universidad de Málaga



Esta obra está sujeta a una licencia Creative Commons:
Reconocimiento - No comercial - SinObraDerivada (cc-by-nc-nd):
Http://creativecommons.org/licences/by-nc-nd/3.0/es
Cualquier parte de esta obra se puede reproducir sin autorización
pero con el reconocimiento y atribución de los autores.
No se puede hacer uso comercial de la obra y no se puede alterar, transformar o hacer obras derivadas.

Esta Tesis Doctoral está depositada en el Repositorio Institucional de la Universidad de Málaga (RIUMA): riuma.uma.es

Publicaciones y Divulgación Científica

UNIVERSIDAD DE MÁLAGA





UNIVERSIDAD DE MÁLAGA

ESCUELA TÉCNICA SUPERIOR DE
INGENIERÍA DE TELECOMUNICACIÓN

Tesis Doctoral

A COGNITIVE ARCHITECTURE FOR AN ATTENTION-BASED AND BIDIRECTIONAL LOOP-CLOSING DOMAIN (CABILDO)

AUTOR: Antonio Jesús Palomino López
Ingeniero de Telecomunicación
2014

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

UNIVERSIDAD DE MÁLAGA
Publicaciones y Divulgación Científica

DRA. REBECA MARFIL ROBLES, INVESTIGADORA ASOCIADA AL DEPARTAMENTO DE TECNOLOGÍA ELECTRÓNICA DE LA UNIVERSIDAD DE MÁLAGA.

y

DR. JUAN PEDRO BANDERA RUBIO, PROFESOR DEL DEPARTAMENTO DE TECNOLOGÍA ELECTRÓNICA DE LA UNIVERSIDAD DE MÁLAGA.

CERTIFICAN:

Que D. Antonio Jesús Palomino López, Ingeniero de Telecomunicación, ha realizado en el Departamento de Tecnología Electrónica de la Universidad de Málaga, bajo nuestra dirección, el trabajo de investigación correspondiente a su Tesis Doctoral titulada:

"A COGNITIVE ARCHITECTURE FOR AN ATTENTION-BASED AND BIDIRECTIONAL LOOP-CLOSING DOMAIN (CABILDO)"

Revisado el presente trabajo, estimamos que puede ser presentado al Tribunal que ha de juzgarlo.

Y para que conste a efectos de lo establecido en el Real Decreto 1393/2007 regulador de los estudios de Tercer Ciclo–Doctorado, **AUTORIZAMOS la presentación de esta Tesis en la Universidad de Málaga**.

Málaga, a 24 de Mayo de 2014

Fdo. Rebeca Marfil Robles
Investigadora Dpto. Tecnología Electrónica

Fdo. Juan Pedro Bandera Rubio
Profesor Dpto. Tecnología Electrónica

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

UNIVERSIDAD DE MÁLAGA
Publicaciones y Divulgación Científica

Departamento de Tecnología Electrónica

E.T.S.I. Telecomunicación

Universidad de Málaga

Tesis Doctoral

A COGNITIVE ARCHITECTURE FOR AN ATTENTION-BASED AND BIDIRECTIONAL LOOP-CLOSING DOMAIN (CABILDO)

AUTOR Antonio Jesús Palomino López
Ingeniero de Telecomunicación

DIRECTORES Rebeca Marfil Robles
Dra. Ingeniera de Telecomunicación

Juan Pedro Bandera Rubio
Dr. Ingeniero de Telecomunicación

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

UNIVERSIDAD DE MÁLAGA
Publicaciones y Divulgación Científica

Abstract

This Ph. D. Thesis presents a novel attention-based cognitive architecture for social robots. The architecture aims to join perception and reasoning considering a double and simultaneous imbrication: the ongoing task biases the perceptual process to obtain only useful elements whereas perceived items determine the behaviours to be accomplished. Therefore, the proposed architecture represents a bidirectional solution to the perception-reasoning-action loop closing problem.

The basis of the architecture is an Object-Based Visual Attention model. This perception system draws attention over perceptual units of visual information, called proto-objects. In order to highlight relevant elements, not only several intrinsic basic features (such as colour, location or shape) but also the constraints provided by the ongoing behaviour and context are considered.

The proposed architecture is divided into two levels of performance. The lower level is concerned with quantitative models of execution, namely tasks that are suitable for the current work conditions, whereas a qualitative framework that describes and defines tasks relationships and coverages is placed at the top level. Perceived items determine the tasks that can be executed in each moment, following a need-based approach. Thereby, the tasks that better fit the perceived environment are more likely to be executed.

Finally, the cognitive architecture has been tested using a real and unrestricted scenario that involves a real robot, time-varying tasks and daily life situations, in order to demonstrate that the proposal is able to efficiently address time- and behaviour-varying environments, overcoming the main drawbacks of already existing models.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

UNIVERSIDAD DE MÁLAGA
Publicaciones y Divulgación Científica

Resumen

Esta Tesis Doctoral presenta una nueva arquitectura cognitiva basada en atención para robots sociales. El objetivo de la misma consiste en aunar percepción y razonamiento abstracto con el fin de que ambos puedan modularse simultáneamente: mientras la tarea en curso es capaz de guiar al sistema perceptivo para obtener objetos relevantes para la misma, los objetos percibidos repercuten en las acciones que pueden ser ejecutadas en cada momento. Por tanto, la arquitectura propuesta supone una solución bidireccional al problema del cierre del lazo percepción-acción.

La arquitectura emplea un mecanismo de atención basado en objetos como sistema perceptivo. Dicho modelo dirige la atención hacia unidades perceptivas de información visual, denominadas proto-objetos. Para conseguir detectar los elementos relevantes de una escena, se tienen en cuenta tanto algunas características básicas intrínsecas (como pueden ser su color, su forma o su localización) como las restricciones impuestas por el comportamiento y el contexto actuales.

La arquitectura propuesta se divide en dos niveles de desempeño. El nivel inferior engloba modelos cuantitativos de ejecución, es decir, conjuntos de tareas que son adecuadas para las condiciones de trabajo actuales. Por su parte, el nivel superior está ocupado por un modelo cualitativo capaz de describir y definir relaciones entre tareas y sus respectivos ámbitos de ejecución. Por tanto, los elementos percibidos determinan las acciones que se pueden llevar a cabo, siguiendo un modelo de necesidades satisfechas. De esta manera, las tareas que mejor se adaptan al entorno que se percibe son ejecutadas con una mayor probabilidad.

Finalmente, la arquitectura cognitiva completa se ha evaluado a través

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

de una serie de experimentos que analizan el comportamiento de un robot real, situado en un entorno sin restricciones, a través de situaciones cotidianas que conllevan tareas que van cambiando con el tiempo. Estos experimentos han demostrado que la arquitectura propuesta es capaz de abordar entornos con comportamientos variables en el tiempo, superando los principales inconvenientes que presentan los modelos ya existentes.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

Agradecimientos

Al llegar al final de una etapa tan importante como es la realización de una tesis doctoral, resulta complicado reunir en unas pocas líneas a todas aquellas personas que con su aliento, su ánimo, su compañía y su cariño han contribuido a que este trabajo se convierta, por fin, en una realidad.

Mis padres y mi hermana son, sin duda, los primeros a los que debo agradecer de corazón su apoyo incondicional y su comprensión incluso en lo incomprensible. No sólo ahora, sino siempre, a pesar de que la mayoría de las cosas que contaba sobre mi trabajo les resultaran normalmente de lo más extravagante. Llegar hasta aquí no es más que la consecuencia de lo que he recibido de ellos. Y a ti, abuela, que tanto insististe en que estudiara para que no me tuviera que "subir a los postes", te hablo desde estas líneas para que también llegue mi agradecimiento allí donde estás desde aquella mañana fría de febrero.

Durante estos años también he tenido la suerte de contar (y seguir contando) con esa otra familia repartida por distintas ciudades que son mis amigos. Quiero recordar especialmente aquí a Rafa, a Paulina, a Paqui y a Gloria. Vuestras palabras, vuestros silencios, vuestra complicidad, vuestras sonrisas, vuestros tirones de orejas, el tiempo robado a esa tesis sobre patrimonio y nuevas tecnologías... Sin todo eso no estaría ahora escribiendo estas líneas. Como tampoco hubieran sido lo mismo estos años en Málaga sin mis primos Loli y José, que hicieron de su tierra mi segunda casa.

Gracias también a mis directores de tesis, Rebeca y Juan Pedro. Su orientación, sus consejos y sus correcciones han contribuido a mejorar notablemente la calidad de este trabajo. A la hora de definir el cierre de lazo bidireccional que se propone en esta tesis, ha sido inestimable la

Publicaciones y Divulgación Científica
UMA UNIVERSIDAD DE MÁLAGA

ayuda de Fernando y Ángel, del grupo PLG de la Universidad Carlos III de Madrid. En todo momento me facilitaron la integración de sus sistemas de planificación deliberativa en la arquitectura propuesta. Tampoco puedo olvidarme de Antonio Bandera, que en la sombra ha estado siempre ayudando. Sin él, este trabajo ni siquiera hubiera podido empezar. Por último, gracias también a todos mis compañeros del Laboratorio 1.3.10 del Departamento de Tecnología Electrónica de la UMA y de mis "laboratorios de acogida", en el CUD de Zaragoza y el grupo PLG de Madrid. Esos buenos ratos han sido inolvidables.

et BVM

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

UNIVERSIDAD DE MÁLAGA
Publicaciones y Divulgación Científica

UNIVERSIDAD DE MÁLAGA
Publicaciones y Divulgación Científica

Contents

1 Introduction **1**

1.1 Motivation 2

1.2 Scope 4

1.3 Outline 6

2 Biological Fundamentals **9**

2.1 Visual Attention 11

2.1.1 The Human Model 11

2.1.2 The controversy about the locus of attention 13

2.1.3 Overt vs. Covert 15

2.1.4 The units of attention 16

2.1.5 Bottom-up vs. Top-down 16

2.2 Psychological Models of Attention 18

2.2.1 Feature Integration Theory 19

2.2.2 Guided Search Model 20

2.2.3 Inhibition of Return 21

2.2.4 Theory of Visual Attention 23

2.2.5 Basic units of attention: proto-objects 24

2.3 Connecting perception and reasoning 25

2.3.1 From cognition to perception: attentional sets 25

2.3.2 From perception to cognition: affordances and reference features 27

Publicaciones y Divulgación Científica

UNIVERSIDAD DE MÁLAGA

3 State of the Art **29**
3.1 Artificial Visual Attention Models 30
3.1.1 First models 31
3.1.2 Advanced models and top-down integration 33
3.1.3 Object-based models 37
3.1.4 Models using alternative paradigms 39
3.2 Linking perception and abstract reasoning 40
3.2.1 Cognitive architectures 41
3.2.2 Other systems involving perception and reasoning . 43
3.3 The bidirectional perception-reasoning loop problem 46

4 Object-based Visual Attention System **49**
4.1 Overview 50
4.2 Perceptual Segmentation 51
4.3 Multi-feature saliency computation 56
4.3.1 Colour contrast and Intensity contrast 58
4.3.2 Proximity 60
4.3.3 Roundness 61
4.3.4 Orientation 62
4.3.5 Symmetry 64
4.3.6 Similarity to basic colours 65
4.3.7 Attending humans: skin colour 66
4.3.8 Saliency Computation 68
4.4 Inhibition of Return 71
4.5 Top-down selection and management of the most relevant proto-objects 73

5 From action to perception and vice versa **81**
5.1 A need-based model 82
5.1.1 Basis: Rules and Knowledge in Human Performance Models 85
5.1.2 Rule-based level 87
5.1.3 Knowledge-based level 90
5.2 The Planning Framework 91
5.3 Closing the loop 94

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

6 Experiments and Results **101**
6.1 NOMADA: a social robot 101
6.1.1 Hardware architecture 102
6.1.2 Software architecture 103
6.2 Visual attention model 107
6.2.1 Qualitative evaluation 107
6.2.2 Quantitative evaluation 112
6.2.3 Discussion 115
6.3 Cognitive architecture 116
6.3.1 Experiment 1: Putting coloured cards in order . . . 117
6.3.2 Experiment 2: A social robot classifying objects . . 123
6.3.3 Discussion 134

7 Conclusions and Future Work **139**
7.1 Conclusions 139
7.2 Future Work 141
7.3 Publications 143

Ñ Resumen en español **145**

Appendix A The Bounded Irregular Pyramid (BIP) **167**

Appendix B Planning domains and problems in PDDL **171**

Bibliography **185**

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

UNIVERSIDAD DE MÁLAGA
Publicaciones y Divulgación Científica

1 Introduction

"He who has begun is half done: dare to know, dare to begin!"

Quintus Horatius Flaccus

In Disney's film *Wall-E* (fig. 1.1), a little robot is responsible for cleaning up a future, deserted and plenty of rubbish Earth. Our mechanical hero has to deal with huge mountains of waste that should be classified and put in order. Before receiving the visit of Eve, causing his monotonous life to turn upside down, Wall-E's main problem is the enormous amount of information that he has to deal with in order to find objects with specific features. How is he able to address such a problem? How should an autonomous and computationally restricted robot carry out that kind of mission? From the point of view of Computer Vision, is it possible to process the whole scenario each time? Does there exist a better strategy to follow?

Although *Wall-E* is only a nice story, some important issues involving the development of autonomous agents placed in a real open world are shown. The robot has not only to deal with a lot of visual information but also has to adapt itself to accomplish different behaviours. For example, Wall-E's principal task consists in packing some specific kind of rubbish but, sometimes, he picks up some interesting or curious objects, saving them in a warehouse. In other occasions, he must react to sand storms or look for pieces in order to repair himself. To sum up, the robot has to address different actions, different tasks that vary over the time and

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA



Figure 1.1: Frames from Disney's film *Wall-E*

need different objects to be perceived. Even more, the discovery of some particular things may produce a deep modification in the robot's current plan of actions. Therefore, additional questions come up: is it possible to modify the way a robotic agent perceives the world depending on its current responsibilities? And, vice versa, are new interesting objects able to modify the ongoing task? How can perception and reasoning interoperate simultaneously in an autonomous robot?

1.1 Motivation

Artificial vision systems can not process all the information that they receive from the world on-line, because it is highly expensive and inefficient in terms of computational cost. However, biological vision systems present an interesting set of features regarding adaptability and robustness, that allows them to analyse and process the visual information of a complex scene in a very efficient manner. Research in Psychology and Physiology demonstrates that the efficiency of natural vision has foundations in **visual attention**, which is a process that filters out irrelevant information and limits processing to items that are relevant to the present task [Duncan, 1984].

In the past few years, emphasis has increased in the development of robot vision systems that are inspired by the model of natural vision. An artificial attention system allows to optimize the required computational

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGA

resources due to they can be focused on the processing of a set of selected regions, which are important for the current task, instead of the whole image.

This ability is specially useful when developing a social robot, that is, an embodied agent which is part of a heterogeneous community of people and other robots [Dautenhahn and Billard, 1999]. In this case, added to the increased efficiency mentioned above, the agent is able to process the visual information in the same way that people do. Thus, the interaction between a human and a robot becomes easier because both of them share the representation of their surrounding world. Furthermore, mobile robots are often carelessly placed at complex environments where they have to apply their knowledge to figure out what needs to be attended to and when and what to do in correspondence with visual feedback [Paletta et al., 2005].

Analysing the problem from a deliberative point of view, there are several behaviours to be accomplished that depend on the perception of a specific set of objects. From that definition, the effects on deliberative planning can be deduced: there exists partial observability, since the attention model constrains the information that the robot perceives. Besides, there is uncertainty because perceived elements are not expected to remain as they were in the past (sometimes not even for a small period of time). In other words, there exists a very close relationship between an attention-driven perception system and a deliberative planner, typically included in the reasoning phase of the classical perception-reasoning-action loop.

The connection between perception, reasoning and action, specially when an artificial attention system is employed, is still an open question. Besides, the problem has been addressed from different points of view because it is a meeting point between Computer Vision and Automated Planning lines of research. Until now, this problem has been usually tackled in an unidirectional way. On the one hand, people working on developing attention models have faced the problem by including a task-dependant component in saliency computation (for example, the solutions proposed in Tünnermann et al. [2013], Navalpakkam and Itti [2006] and Torralba et al. [2006]). However, they do not specify a real link with a deliberative model that points out what elements in scene are relevant to the task. On the other hand, in models such as Sridharan et al. [2010], Laird et al. [1987]

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

or Gilet et al. [2011], the deliberative layer proposes a set of visual features to be found and the perception system only needs to look for them in order to increase the observability and reduce the uncertainty. These unidirectional models are often bounded to very specific applications and do not care about attention.

Nevertheless, physiological observations suggest that particular perceptual characteristics, such as location or shape, engage actions related to those characteristics. Thus, those activating action systems may prime the processing of stimuli defined by the perceptual characteristics related to these actions [Collins et al., 2010; Fagioli et al., 2007]. In consequence, a classical unidirectional assumption of the perception-reasoning loop is not enough. Hence, a complete solution for that problem must cover the double imbrication between perception, reasoning and action, overcoming the drawbacks of the prior models.

1.2 Scope

This thesis presents a **new cognitive architecture for an attention-based and bidirectional perception-reasoning-action loop closing**. The proposed architecture creates a symbiosis between the abstract reasoning and the attention modules of a robotic system, linking visual features with high level behaviours (see fig. 1.2). Concretely, the architecture is based on the interaction of a general purpose Planning Framework, that produces plans constrained by the information perceived from the vision system, and an object-based attention model, able to highlight elements that are suitable for the ongoing task. Through the exchange of *relevant elements* and *sets of perception parameters*, the cognitive system is able to decide what tasks are going to be executed, following a *need-based* approach. At last, the predominant task selects what kind of elements are going to be searched next and the system performs the loop again in a cyclic way.

The **object-based attention model** itself is the other main contribution of this work. The proposed model integrates task-independent bottom-up processing and task-dependent top-down selection. In a first stage, the

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGUNA

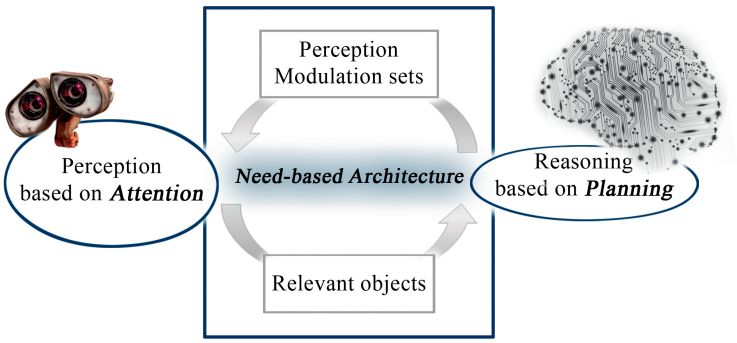


Figure 1.2: Overview of the proposed attention-based cognitive architecture

system is able to obtain the different visual entities in an image using a perceptual algorithm. Then, the most salient parts are computed taking into account different simple features (colour, location, shape...). The model uses templates of perception parameters, provided by a deliberative layer, to modify the influence of each feature in the global saliency computation. Thereby, attention is drawn to those objects that are relevant to accomplish the ongoing task. Mechanisms able to handle dynamic scenarios are also included in a second stage to avoid revisiting an already attended entity. Finally, the most relevant elements are provided to high level modules.

The cognitive architecture introduced in this monograph is clearly focused on a bidirectional consideration of the perception-reasoning connection. The relevance of an object changes depending on its utility for the ongoing task, whereas the perceived objects may constrain or provide new information that could suggest the modification of a current behaviour. Thus, the introduction of new objects in scene triggering new tasks (e.g. critical or higher priority tasks) is also taken into account. Consequently, a continuous adaptation of the plan depending on the perceived objects, and vice versa, is allowed.

In summary, using the development of an attention model as a basis, a cognitive architecture connecting active perception and abstract reasoning has been designed, providing a complete, reasonable and biologically founded solution to the perception-reasoning-action loop closing problem.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

1.3 Outline

The remainder of this thesis is organised as follows:

- In *chapter 2*, the biological foundations supporting this work are introduced. The chapter presents in detail the main concepts associated with attention in human vision. Besides, the most influential theories about visual attention and its relationship with abstract reasoning in humans are explained.

- *Chapter 3* is concerned with the state of the art of computational visual attention models and other systems linking perception, reasoning and action. The chapter analyses the solutions proposed by the two main lines of research working on this topic. Then, the convenience of a complete and bidirectional characterisation of the perception-reasoning loop closing problem is discussed.

- The first important contribution of this work, the object-based visual attention system, is fully described in *chapter 4*. This chapter shows how images are processed in order to divide them into visual units of perception and, then, how the relevance of those units is computed.

- *Chapter 5* deals with the other important contribution: the general purpose cognitive architecture closing the perception-reasoning-action loop in a bidirectional way. The different levels that compound the system are detailed and the employed planning framework is also introduced here.

- Experiments and results are shown in *chapter 6*. First, the hardware employed in the experiments is presented: NOMADA, a social robotic platform. Then, the attention model is independently evaluated in order to study its efficiency extracting and selecting relevant elements from a scene. At last, the whole architecture is tested through a pair of experiments bringing out the most representative characteristics of the proposal and its advantages, pointing out how the performance of a system using this architecture is effectively enhanced.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

- Finally, conclusions and future work are summed up in *chapter 7*, including a list of the publications related to the work presented in this monograph.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

UNIVERSIDAD DE MÁLAGA
Publicaciones y Divulgación Científica

2 Biological Fundamentals

"The first step in the acquisition of wisdom is silence, the second listening, the third memory, the fourth practice, the fifth teaching others"

Solomon Ibn Gabirol

Everyday, magicians surprise their audience with incredible tricks. They make things appear or disappear just in front of the eyes of the public. The spectators carefully observe while ignoring the fact that they are being fooled by their own visual system. Actually, people are not able to process in detail all the visual information that they receive at each moment. They need clever mechanisms to reduce that huge amount of heterogeneous visual data. Namely, that mechanism is the **visual attention** system.

Visual attention allows people to focus their efforts in those parts of the world that are relevant, filtering out all the unnecessary information. Fortunately for magicians, this means that people being concentrated in a specific task are usually "blind" to those objects that are not useful for the current action (this phenomenon is known as *inattentional blindness*; see fig. 2.1 for more detail). Therefore, the amount of information to be processed by the human brain becomes more useful and easier to handle.

For mobile autonomous robots, the ability to focus on the relevant data becomes crucial. Several modules share a set of limited resources in a robot and the computational load must be constrained. Besides, nowadays robots

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA



Figure 2.1: Inattentional Blindness. In the experiment, observers are told to count how many times a group of people passes a standard orange basketball. After about 30 seconds, a person in a gorilla suit walk across the scene, faced the camera, thumped his chest and walked away. Half the viewers did not notice the gorilla because they were focused on counting the passes [Simons and Chabris, 1999].

are not simple industrial tools. Current systems are supposed to deal with a complex, dynamic and previously unknown environment where they have to work autonomously even surrounded by people. Thus, limited resources requires an intelligent selection mechanism able to classify sensory input data from *very important and relevant* to *not useful at this moment.*

For many years, Computer Vision and Robotics scientists have tried to develop enhanced autonomous systems as practical applications of the idea of human selective attention. The applied concepts mainly come from Psychology and Physiology research about visual attention in humans. Consequently, this chapter firstly presents the fundamental theories supporting the attention system introduced in this monograph.

Finally, high-level cognitive and complex processes rely on data that have been transformed and constrained by the attention system. Therefore, it is also important to understand the different connections between perception and reasoning in humans. Psychologists have investigated about how abstract knowledge can modulate the way that people perceive the world and what mental processes are triggered when some specific features or objects are perceived. As a basis to develop the cognitive architecture presented in this thesis, the final part of the chapter copes with the theories studying the connection between perception and deliberative reasoning.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

2.1 Visual Attention

Throughout the years, the term **attention** has been used to refer to different psychological aspects such as the ability to keep alert for a period of time or the capacity to maintain a mental focus on a specific task or event. Nevertheless, this monograph lays emphasis on attention as a procedure of perceptual selectivity, that is, a mechanism to select from the incoming stimuli only those which are specially relevant. In these terms, a formal definition of attention is given in Corbetta [1998]: *"**Attention** defines the mental ability to select stimuli, responses, memories or thoughts that are behaviourally relevant among the many others that are behaviourally irrelevant"*. In other words, **attention** is the cognitive process of selectively concentrating on one thing while ignoring other things.

The concept of **saliency** is also closely related to attention. Saliency characterizes the elements or regions present in the scene that an observer could consider as standing out from the surrounding parts [Borji and Itti, 2013].

Although some authors like James [1890] had already emphasized the selective function of attention in the 19th century, psychologists have maintained an increased interest in studying this phenomenon since 1950. The first experiments, for example, were oriented to study the eye movements of a person observing a picture [Yarbus, 1967]. Hence, most of the theories about attention have been developed founded in both a background knowledge about human eye physiology and empirical studies with human subjects.

In order to correctly characterize the phenomenon of attention, this section covers the main concepts associated with visual input filtering and selection. After briefly describing the nature of the human vision system allowing an attention mechanism, questions about *when* the selection takes place, *where* attention is deployed and *how* the relevance of an element depends on both scene properties and observer's expectations are discussed.

2.1.1 The Human Model

The anatomy of human eyes (fig. 2.2) exhibits a set of characteristics which provide some interesting skills in visual data acquisition and filtering.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

The light travel through the cornea and the lens and arrives to the retina. This part of the eye is a light-sensitive layer of tissue plenty of photoreceptor cells. There exist mainly two different types of cells: rods and cones. Rods are more sensitive than cones so they are more suitable to dim light situations. However, rods only provide black and white information due to they are not sensitive to colour. On the contrary, cones are colour-sensitive but they need more intensive light in order to correctly capture the visual information. Finally, the perceived information is transmitted to the brain via the optic nerve.

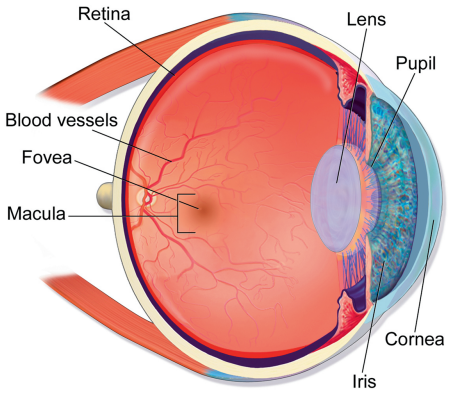


Figure 2.2: Eye Anatomy. Author: Bruce Blaus.

Another important issue refers to rod and cones distribution in the retina. As it can be seen in fig. 2.3, most of the cones are concentrated in the center of the retina. This zone is known as fovea. Because of cones are more accurate than rods, the part of the image meeting the fovea is analysed in more detail. Consequently, not all the regions in the image are processed with the same resolution.

The inability to process the whole image at full resolution seems to be a disadvantage. Quite the contrary, foveal vision allows a better management of the visual resources, reducing the amount of information provided to the brain. Thus, the only way to scan the whole scene consists in a continuous eye movement to sequentially select the different parts of the image. Hence, the human visual system needs a mechanism in order to prioritise which

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGUNA

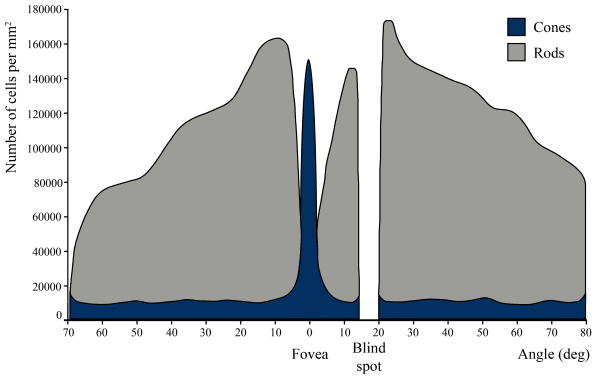


Figure 2.3: Distribution of rods and cones in the retina.

regions are going to be processed first. For that reason, the foveal nature of the human vision has motivated the development of the attention process. Thereby, the high-resolution area of the eye is focused on relevant parts of the image whereas not useful information is processed by the low-resolution peripheral cells.

In summary, human visual system provides a highly efficient solution in order to analyse and filter huge amounts of visual information employing a reduced set of resources.

2.1.2 The controversy about the locus of attention

Nowadays, an open debate still exits among the psychologists regarding *when* the attentional selection take place. Is this selection accomplished before or after an effective pattern recognition? In other words, do people deploy attention to recognized objects? Is only some basic knowledge about the elements in the image needed?

Based on the aforementioned criterion, attention theories are commonly classified by their assumption about the locus of attention into two categories: *early-selection* theories and *late-selection* ones.

In early-selection models, the region to be attended is selected before pattern recognition takes place. In consequence, the selection is based on a set of *physical* simple features (e.g. colour, location or shape) extracted by

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

pre-recognition processes. Therefore, the selected entities may have no full perceptive meaning, i.e., they could not correspond to real objects. The perception system is supposed to have a limited capacity and selective attention is considered as a strategy to distribute this capacity among several input channels. Models such as Treisman and Gelade [1980] or Broadbent [1958] are included in this category.

The late-selection theories, on the contrary, claim that pattern recognition precedes attentional selection. Then, the selection can be based on complex properties such as meaning or categories (e.g. alphanumerical classes) as well as simple features. Models included in this category are, for instance, Duncan [1980] and Posner et al. [1978].

In any case, the borderline between both categories is not clear. Even some theories like Bundesen's *Theory of Visual Attention (TVA)* [Bundesen and Habekost, 2008] establishes that selection and recognition occur simultaneously. That is, they are two aspects of the same process. Besides, there exists a question that is difficult to explain from the point of view of late-selection models: a recognizable form or object is not always needed to direct attention to. For example, in abstract paintings (fig. 2.4), the attention of the observer is drawn to those parts of the picture that have a set of elemental features differing from their surrounding. This process occurs regardless the presence of recognizable objects.

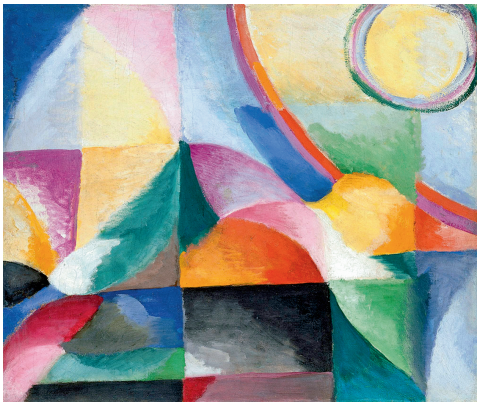


Figure 2.4: *Simultaneous Contrasts* (Sonia Delaunay-Terk, 1913). Even in the absence of recognisable forms or objects, attention is drawn to different parts of the painting which stand out from their neighbourhood.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

From the Computer Vision Engineering point of view, a full and effective object recognition is, at present, a quite complex task. Consequently, artificial attention models, such the one presented in this thesis, are commonly based on early-selection theories.

2.1.3 Overt vs. Covert

Attention can also be distinguished depending on *where* it is deployed. When people examine a scene, they usually attend to the different regions of interest by moving their eyes and head to correctly direct the fovea. This procedure is known as *overt* attention: the fovea points towards a stimulus. However, this is not the only manner to attend elements in the scene. *Covert* attention deals with the peripheral locations, so it is possible to attend certain elements without voluntarily moving the eyes (colloquially expressed as "looking at something out of the corner of one's eye"). This procedure can be seen as a mental focusing onto one or several external stimulus [Borji and Itti, 2013].

Car driving could be a good example for covert attention. The driver keeps the sight focused on the road while simultaneously attending the different traffic signals, lights or people coming across the road in a covert manner. Covert attention is usually explained as a mechanism for quickly scanning the field of view for an interesting location, deploying the computational resources needed for processing a specific part of the input sensory data. This covert shift produces a *saccade* (rapid and unconscious eye movement) to the attended location (overt attention) [Peterson et al., 2004]. Nevertheless, some studies demonstrate that, in some cases, an eye movement is not required to perform some tasks (e.g. object manipulation) [Johansson et al., 2001]. Hence, covert mechanisms of attention are not still well known. Thus, there is a lack of covert computational frameworks and covert attention is still difficult to measure [Borji and Itti, 2013].

Because of the intrinsic overlap and mutual relationship between covert and overt attention, saliency models are sometimes difficult to classify in these terms. Moreover, in non-foveal models and models deploying attention to several locations at the same time, as it occurs in the system introduced in this monograph (see chapter 4), the belonging to covert, overt or both mechanisms is even harder to determine.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

2.1.4 The units of attention

At present, there is no unique agreement about the units of attention, that is, about the scale of the target which attention is directed to. Most of the psychological and neurobiological studies assume that attention is related to spatial locations (e.g. Posner et al. [1978]). However, there are also evidences about the assumption that the boundaries of segmented objects, and not just spatial position, determine what is selected and how attention is drawn [Scholl, 2001; Duncan, 1984]. Therefore, these models reflect the fact that perception abilities must be optimized to interact with objects and not just with disembodied spatial locations. Furthermore, some authors claim for feature-based attention models [Treisman and Gelade, 1980].

This diversity about the units of attention has motivated attention models to be mainly classified into two categories: *space-based* and *object-based* systems. Although there exist more space-based models than object-based ones, it is also plausible to think that humans work and reason with objects and not with rough pixel values [Einhäuser et al., 2008]. Feature-based models adjust some specific features in order to stand out a target object from a distracting background. Because of the close relationship between object-based and feature-based models, both are typically included in the same group [Borji and Itti, 2013].

Finally, it should be pointed out that there is often more than a single unit of attention. It has been demonstrated that people are able to attend to multiple regions of interest simultaneously. Concretely, psychological studies and neurobiological experiments show that attention can be deployed up to four or five elements at the same time in covert mode [Pylyshyn and Storm, 1988; Awh and Pashler, 2000; Bundesen et al., 2011]. That is, covert attention entails saccades that alternatively move the fovea among a reduced set of elements in the scene. As a result, these four or five elements can be considered as being attended at the same time.

2.1.5 Bottom-up vs. Top-down

Visual attention is fundamentally driven by two major contributions: cognitive factors (**top-down**) such as knowledge, expectations and cur-

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

rent goals, and **bottom-up** factors that reflects the influence of sensory stimulation [Corbetta and Shulman, 2002].

On the one hand, stimulus-driven bottom-up attention is based on the basic features inherent to the objects in the scene. In other words, a region of interest draws attention, i.e. it is more *salient* or relevant, when it has one or several features sufficiently discriminative with respect to its surrounding elements. Bottom-up attention is fast, involuntary and, in principle, independent of the ongoing task. This attentional mechanism is also called exogenous, automatic, reflexive or peripherally cued [Egeth and Yantis, 1997].

On the other hand, top-down attention is consciously directed by the current behaviour of the observer. It reflects the fact that certain elements in the scene are more or less relevant depending on its utility to accomplish the action in progress. Top-own attention is slow, task-driven, voluntary and **closed-loop** [Itti and Koch, 2001]. Top-down attention is also referred as endogenous or centrally cued attention [Frintrop et al., 2010].

A person walking or driving a car along the street could be a good and intuitive example about top-down mechanisms. When driving a car, some elements such as traffic signals, semaphores or nearer cars are more likely to attract the driver's attention. On the contrary, when just walking along the street, the aforementioned elements are likely to remain ignored whereas holes or obstacles in the pavement, lampposts or nearby people and animals are more likely to be perceived by a pedestrian.

The idea of top-down guidance was empirically demonstrated by Yarbus [1967]. In his experiment, subjects were asked to observe a painting showing a room with a family and an unexpected visitor under different conditions. For example, Yarbus previously asked observers to figure out the material circumstances of the family, to guess the age of the people present in the scene or simply to freely examine the scene. Then, he recorded the eye movement of the subjects employing an eye-tracking system. The finding was that eye movements differed considerably for each case (fig. 2.5).

Neurophysiological experiments also conclude that bottom-up and top-down attention components are associated with two independent but interacting areas in the brain [Corbetta and Shulman, 2002]. Both mechanisms

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

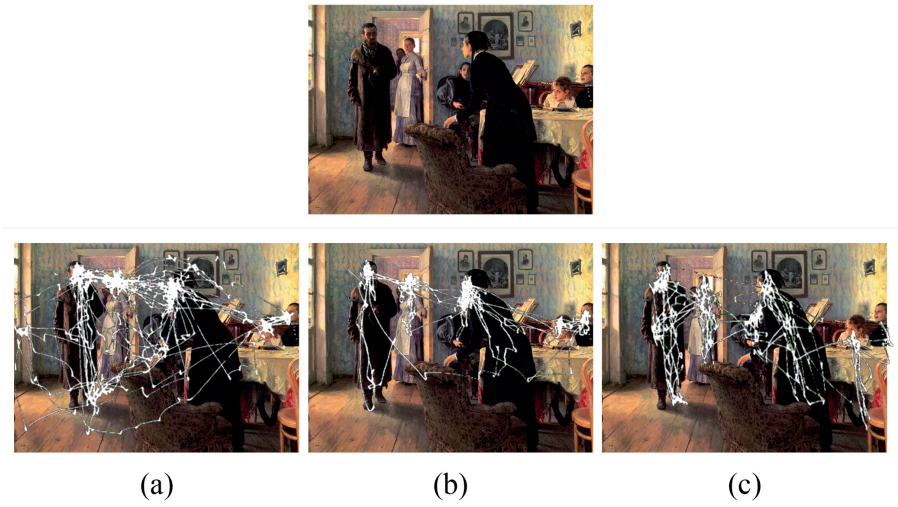


Figure 2.5: Some results of the eye-tracking study carried out by Yarbus [1967]. The painting *An Unexpected Visitor* (Ilya Repin, 1884) is presented to the observers in a number of different ways. Before presenting the picture, the subjects are asked to examine the scene under certain conditions (questions) in a task-dependent behaviour. (a) eye movement in free view; (b) eye movement when trying to figure out the age of the people; (c) eye movement when trying to remember the clothes worn by the people.

interact in the normal human perception. Although bottom-up influence is not voluntary suppressible [Theeuwes, 2004], there is evidence that it can be overridden in some cases by top-down processes [Bacon and Egeth, 1994]. Nevertheless, although the interaction between both contributions to attention has been widely empirically demonstrated, the fundamentals of such a relationship are not clear at present.

2.2 Psychological Models of Attention

A great variety of psychological theories and models about visual attention in humans have been developed during the last decades. Some of them try to fully explain the attentional process (Treisman and Gelade [1980], Wolfe [1994] or Bundesen and Habekost [2008]) while other lay emphasis on

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

specific attentional aspects such as the *Inhibition of Return* (for instance, Posner et al. [1985] and Klein [2000]) or the units of attention (Rensink [2000]).

In this section, the models and theories that support the computational attention system introduced in this thesis are detailed. A more detailed review of psychological models can be found in Bundesen and Habekost [2005].

2.2.1 Feature Integration Theory

Indubitably, the *Feature Integration Theory* elaborated by Treisman and Gelade [1980] is one of the most influential theories about visual attention. It has been taken as biological basis in most of the later artificial attention models. It was first introduced in 1980.

The Feature Integration Theory suggests that the human vision system detects separable features in parallel in an early step of the attention process (the *pre-attentive* stage). This procedure creates *feature maps* (also called *conspicuity maps*): maps that highlight locations according to the respective feature. Then, the different feature maps are combined in a *master map of saliency*. This map specifies the location of the most relevant elements in the image in descending order. Sequentially scanning through the master map focuses the attention and provides the data for further perception tasks. The process can be seen as a spotlight serially moving from one location to another in order to deploy attention individually to each relevant location.

Looking for a target is easier when more features differentiate the target from the possible distractors. When a target differs from the distractors only in how its features are combined, the search is more difficult and needs focused attention. This phenomenon is known as conjunctive search.

Finally, Treisman's findings suggest a convergence between parallel detection of visual targets and perceptual grouping. Both seems to depend on a distinction at the level of separable features so neither requires focal attention. Therefore, perceptual grouping and feature detection are addressed in the pre-attentive stage and both could be involved in the control of attention.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

2.2.2 Guided Search Model

Some years after Treisman's proposal, Wolfe introduced his *Guided Search* [Wolfe et al., 1989] as an answer to some questionable aspects of the Feature Integration Theory. Wolfe has improved his model over the years, assigning version numbers to each upgrade following the convention in software developing. Although there exist up to four updates of the model, versions 3.0 [Wolfe and Gancarz, 1997] and 4.0 [Wolfe, 2007] only introduce minor changes, being version 2.0 [Wolfe, 1994] the most elaborated one.

The Guided Search theory tries to explain and predict the results obtained in visual search experiments. Essentially, Wolfe proposed that low level visual processing may be guided depending on a high level task. That is, the model explains the influence of top-down information in saliency computation. Namely, humans only look for specific features to speed up the pre-attentive stage process. For instance, if the observer is searching for a "blue L" in a set of different coloured letters, not blue letters are directly discarded so low level parallel search is faster.

Wolfe's model is mainly an extension of the Feature Integration Theory. Therefore, both models share many concepts. For example, bottom-up saliency is also obtained as the summation of different feature maps (red, green, 45° orientation...). However, Wolfe considers only one map for each *feature dimension* (colour, orientation, etc.) although he also mentions that there is evidence for difference between features (e.g. there are multiple colour maps but only one orientation map). To introduce top-down information, an additional map is included for each feature. This map selects the feature type that allows the target to be better distinguished.

fig. 2.6 shows an example about the speeding up process in Guided Search. When presenting an image containing different coloured letters without any prior instruction (fig. 2.6.a), the observers usually deploy their attention to the blue L (or, less commonly, to the gray T) as Treisman's model explains. However, if an image containing more distractors is employed (fig. 2.6.b), it is difficult to find a salient letter. Nevertheless, if the observer is previously asked to look for the blue L in the image, the target is rapidly found as Wolfe's model established.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGUNA

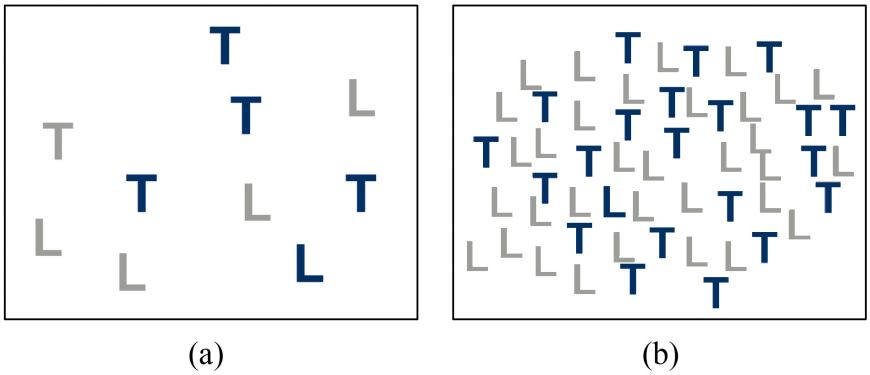


Figure 2.6: Guided Search. (a) image with few distractors. (b) image with a lot of distractors. A *blue L* can be easily found in (a) in a bottom-up manner. However, it is difficult to find it in (b). Top-down information speed up the search process of a *blue L* letter in (b) despite the amount of distractors.

Since Wolfe's model completes Treisman's theory, artificial attention systems are commonly based on both approaches. Nowadays, efforts in computational attention are being conducted to develop models which combine a task-independent bottom-up processing stage, according to Feature Integration Theory, with a top-down selection process, founded on Guided Search.

2.2.3 Inhibition of Return

The concept of **Inhibition of Return (IOR)** was firstly introduced by Posner et al. [1985] although the phenomenon had been observed and demonstrated one year earlier by the same authors [Posner and Cohen, 1984]. In their experiments, the authors found that, contrary to their expectations, following a peripheral cue did not predict the location of the next peripheral target [Klein, 2000]. Then, they realised that there exists a mechanism encouraging the orientation towards novel locations: *the inhibition of return*.

When attending an event at a peripheral location, the processing of near stimuli is facilitated for a short period of time. After attention is removed

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

from that previous location, the response to stimuli there is then delayed. Consequently, recently attended locations or elements remain *masked* during certain time. IOR can be easily observed in eye-tracking experiments (fig. 2.7). The focus of attention is directed to the relevant parts of the scene. However, each location draws attention only for a short time. Then, the focus is not re-oriented to the previous location until certain time.



Figure 2.7: Eye-tracking experiment. Three frames are shown (frame rate: 30 fps). The subject freely examines the scene. The sight of the observer is sequentially directed to the different relevant parts of the scene (in this case, mainly the faces). The focus of attention is maintained for a few milliseconds over the same location. After that, attention is deployed to another place and the previous location is inhibited for a time.

In summary, the cause of IOR is attributed to attention re-orienting towards a location and the subsequent removal of attention from that location. The effect allows attention to avoid coming back to the originally attended place. Although psychophysical experiments reveal an influence of IOR in saliency depending on time, computational models of attention have not satisfactorily addressed this problem at present. Most of the artificial attention systems employ a stationary implementation of IOR applying instantaneous inhibition masks (e.g. Itti et al. [1998], Backer et al. [2001], Frintrop [2006], Aziz and Mertsching [2008a] or Palomino et al. [2011]). Thereby, it is difficult to re-include inhibited locations again after some time being avoided. As it will be detailed in section 4.5, the computational attention system proposed in this thesis offers a more elaborated solution to compute time-dependent saliency values based on the dynamics of IOR.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGA

2.2.4 Theory of Visual Attention

Bundesen's *Theory of Visual Attention* (TVA) [Bundesen, 1990; Bundesen and Habekost, 2008; Bundesen et al., 2011] is an attempt to unify the several theories previously developed. In brief, visual recognition and attentional selection is based on the perceptual categorization of the elements in the visual field. The selection is determined by a processing race between possible perceptual categorizations towards a limited short-term memory.

TVA is based on the combination of choice models and race models of selection. Choice models predict only the final result of selection, so they are non-process models. Therefore, temporal aspect of attentional selection is reported by race models. Concreteley, TVA integrates a Biased Choice Model [Luce, 1963] into a Four-parameter Exponential Independent Race Model (FIRM) framework.

Bundesen postulates the existence of a **Limited-capacity short-term memory** (VSTM) containing targets. This memory is typically limited to three of four elements. The final selection is determined by a processing race among the possible input perceptual categorizations. Thus, those entities arriving first to the VSTM are attended (fig. 2.8).

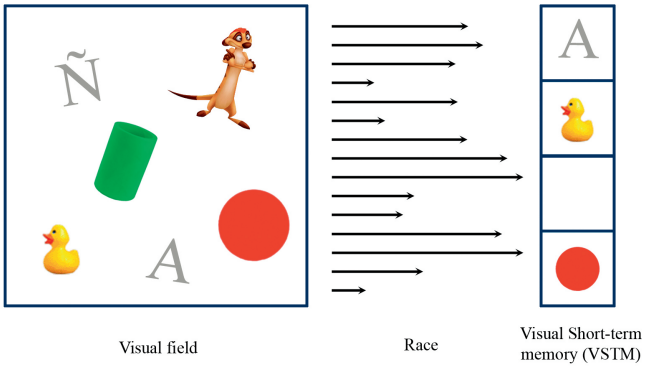


Figure 2.8: Theory of Visual Attention (TVA). The race model selects the perceptual categorizations to be attended as the first ones to arrive to the Visual Short-Term Memory (VSTM).

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

Finally, the top-down component of attention is addressed in TVA employing a combination of two mechanisms: *filtering* and *pigeonholing*. While the former is responsible for selecting target objects rather than distractors, the latter classifies objects with respect to specific categories that are useful for the current action. Given adequate top-down settings of pertinence, the system will perform the filtering and the pigeonholing processes employing elementary algebraic operations [Bundesen and Habekost, 2008].

Concerning the attention system and the cognitive architecture proposed in this monograph, TVA provides **two key ideas**: the small, fixed-size **visual short-term memory**, that allows to handle several attended elements at once, and the **categorization** process, that translates the perceived elements into abstract entities for reasoning and permits the top-down selection of particular useful elements.

2.2.5 Basic units of attention: proto-objects

In previous sections, the diversity about the units of attention has been discussed. Since the perception system presented in this thesis uses an object-based approach, it is important to correctly and unambiguously define the *entities* that draw attention.

Taking into account that, in early-selection models, the selection of relevant items precedes pattern recognition, an early processing to divide the scene into items is required. These items, or *pre-attentive* objects, group parts of the visual field as coherent wholes but they do not necessarily correspond to conceptual or recognizable objects [Orabona et al., 2007]. They can be considered as potential objects. Rensink gave the first definition to these prior entities in his *Coherence Theory of Attention* [Rensink, 2000]. He named them **proto-objects**.

Proto-objects are the resultant structures of a low-level (only geometric and photometric features are involved), rapid (it takes a few hundred milliseconds) and carried out in parallel pre-processing task. Proto-objects provide local descriptions of scene structure. They can be quite complex but they are coherent only over a small region of space. Another characteristic of proto-objects is their volatility, that is, they are constantly regenerated

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGA

and simply replaced by any new incoming stimuli to their location. As a consequence, proto-objects superposition is disallowed.

Finally, abstract reasoning is based on structural items rather than pure spatial and disembodied pixel-based locations. Consequently, a perception system directing attention to proto-objects provides a more convenient representation of the visual field to high-level deliberative layers. For example, it has been demonstrated that more object-based attention is present during grasping tasks [Fischer and Hoellen, 2004].

2.3 Connecting perception and reasoning

In the sections above, attention has been explained as a perception process able to take into account cognitive factors in a top-down manner. But, what are the underlying mechanisms that allow people to translate factors such as knowledge, expectations or goals into a variation in the relevance of the objects they perceive?

There are neurophysiological evidences to conclude that abstract planning can modulate the perceived information [Fagioli et al., 2007]. To achieve that, deliberative processes provide a group of perception parameters affecting the relevance of some specific features to pop out action-related elements: the **Attentional Sets** [Corbetta and Shulman, 2002].

Furthermore, not only cognitive information influences perception, the reverse direction is also possible. Some scientists have investigated how the presence of specific features or categories of objects involves the execution (or the possibility of execution) of certain tasks, triggering cognitive processes. **Affordances** [Gibson, 1979] and **Reference Features** [Pryor, 1994] are specially interesting approaches in this sense.

2.3.1 From cognition to perception: attentional sets

During the last two decades, work in Cognitive Psychology and Neuroscience has revealed that task demands and action requirements act on visual perception, the selection of visual objects and memory. This implies

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

a more dynamic and close relationship between cognitive systems and perception than the standard unidirectional assumption of human information processing [Ward, 2002].

Through a set of experiments, Fagioli et al. [2007] suggest that action can influence perception in at least two ways. On the one hand, actions bias the perception of objects and events at a *feature level*, giving preference to some concrete properties. On the other hand, at a *dimensional level*, actions facilitate the processing of the features defined in a specific perceptual dimension. As a consequence, features defined as task-relevant are more *heavily weighted* and, hence, more strongly represented than other being task-irrelevant.

Among the studies analysing the impact of action on perception and action-related feature biasing, the work by Corbetta and Shulman [2002] is specially thorough. They use functional magnetic resonance imaging (fMRI) experiments to analyse top-down control of attention. They noticed that the final saliency value of an object is the result of a sum of both bottom-up and top-down signals for different features.

Corbetta's findings evidence that human observers are faster detecting an object in a visual scene when they know in advance something about its features (colour, shape, location, etc.). This facilitation depends on human's ability to represent this advance information and to apply it to bias the processing of incoming visual data. They refer to the representations involved in the selection of task-relevant stimuli as **Attentional Sets**. In other words, an attentional set is a group of parameters that codes the top-down signals and *modulates* the influence of some properties in the computation of saliency maps.

Summarising, top-down and bottom-up contributions are integrated in saliency computation through weighting different features. In this process, the attentional sets translate cognitive processes, such as goals or expectations, into a modulation of perception by modifying the weights involved in the summation. Thereby, cognition assists vision by creating, maintaining and changing a representation of what is important while observing the scene.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGUNA

2.3.2 From perception to cognition: affordances and reference features

Analogously to the aforementioned theories, scientists have also tried to understand the relation between perceived objects and the actions and effects they can evoke in human brain.

One of the most renowned contributions to explain this phenomenon is the concept of **affordance**, introduced by Gibson [1979]. Gibson's theory is based on ecological observations. He defines affordances as what the environment offers and provides to animals. This implies that the *values* and *meanings* of things can be directly perceived. For example, a flat, horizontal and rigid surface affords *support* for an animal if the surface is strong enough in comparison with the weight of the animal (maybe is sufficient for a dog but not for an elephant); a tree could be only *climbable* by animals with specific capabilities, etc.

From a engineering point of view, affordances can be considered as *agent-dependent usages*, that is, the action possibilities given by an object to an agent with specific action capabilities [Montesano et al., 2008]. Thereby, for an autonomous system, it could be very interesting to know what actions, from its currently planned ones, can be executed while simply observing the elements in a scene.

In a similar way to Gibson's affordances, Pryor [1994] studied the influence of perceptual information in triggering cognitive actions. She addressed the problem from the Artificial Intelligence Automated Planning perspective and introduced the concept of **Reference Features**. When an agent has to face unexpected situations, it must be able to filter those goals that are likely to be achieved without performing large computational processes. One solution to perform this kind of filter consists in detecting features in objects that reveal their utility: the reference features. For example, a *sharp* object is frequently involved in performing *cutting* activities.

Reference features must be both easy to compute and highly predictive to be really efficient. Otherwise, they would be as inefficient as a detailed analysis. They also depend on the knowledge that people have about objects and other elements in their quotidian world. Finally, these features

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

also help focusing the reasoning that is performed. Thereby, an agent using them is able to concentrate its reasoning on the particular method or goals indicated by the reference features.

In conclusion, the problem of linking attention and abstract reasoning have been addressed from two points of view depending on the main lines of research of the authors. However, there is a lack of proposals providing integration solutions for such a challenge up to now in the literature. This deficit is even greater in approaches applied to artificial autonomous agents. Consequently, the novel cognitive architecture introduced in chapter 5 attempts to bridge this gap by assuming a symbiotic and interdependent combination of both philosophies. Thereby, cognitive factors can guide attention by modifying some perception parameters and, simultaneously, the presence of specific elements is able to suggest the execution of certain actions.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

3 State of the Art

"Ideas? My head is full of them, one after the other, but they serve no purpose there. They must be put down on paper, one after the other"

Camilo José Cela

Inspired by the biological fundamentals detailed in the previous chapter, researchers in the fields of Robotics and Computer Vision have developed artificial systems that take advantage of the inherent enhancements of such natural models. Concerning computational visual attention systems, there exists a wide range of different applications. Thereby, visual attention has been successfully applied to mobile robotics, computer vision and cognitive systems [Frintrop et al., 2010].

Although pure visual attention models have been deeply investigated, there are still open questions about an effective integration between artificial attention models and reasoning systems, as it was pointed out in the previous chapter. At present, some partial solutions to this problem have been proposed from different points of view, namely from Computer Vision and Automated Planning fields. However, computational attention proposals addressing the problem of task-driven attention do not explicitly define top-down maps generation. On the other hand, pure planning solutions and cognitive architectures typically deal with both perception and reasoning but in a classical unidirectional assumption. Therefore, attention-based

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

perception is not taken into account so perception can not modulate cognition. Besides, they are often quite restricted to specific applications and environments.

The first part of this chapter reviews the work about computational models of visual attention over the past few years. Then, the second part is concerned with cognitive architectures and other solutions that have faced the perception-reasoning-action loop closing problem. Finally, the competence of the reviewed models for addressing the perception-reasoning-action loop closing problem in a bidirectional manner is discussed.

3.1 Artificial Visual Attention Models

Since the first psychological models of attention, researchers have shown a growing interest in applying their advantages to computer vision and robotics systems. First artificial models of attention were mainly focused on implementing a computational version of Treisman's *Feature Integration Theory* [Treisman and Gelade, 1980]. These primitive models were pure bottom-up and static approaches.

Later, new findings in Psychology and Physiology research have entailed new improvements in already existing models as well as the development of new advanced proposals. Thereby, the top-down component of attention has been gradually introduced into prior and new models. Depending on the considered attention scale (recall section 2.1.4 on page 16), models has been also divided into *space-based* and *object-based* approaches. Finally, models derived from alternative paradigms to Treisman's model have also been developed, applying techniques from a variety of fields, such as Machine Learning, Information Theory or Bayesian Frameworks.

This section provides an overall review about the principal and more significant computational models of attention presented in the literature. Exhaustive surveys on this topic have been recently published by Borji and Itti [2013] and Frintrop et al. [2010]. Ferreira and Dias [2014] have also recently proposed a specific survey related to attention in socially interactive robots.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGUNA

3.1.1 First models

The first computational attention model was introduced by Koch and Ullman [1985]. Although it was a purely theoretical model (it had not been implemented when it was first published), it included all the necessary algorithmic reasoning. Hence, this model has been used as a foundation for many later approaches. Following the guidelines of the *Feature Integration Theory*, the model firstly computes a set of basic features in parallel. Then, the different feature maps are combined into a master *saliency map*. At last, a *Winner Takes All (WTA)* network is applied. A WTA network is a neural network that determines the most salient region in a topographical map. WTA shows how the computation of a maximum is possible in neural networks, that are compounded by single units being only locally connected.

After selection, the region of interest is routed into a *central representation*. Such representation contains only a single location in the scene at each time. However, this procedure is not detailed. Finally, the authors also suggest an Inhibition of Return mechanism in order to select the next most relevant region in the image. The proposed model is pure bottom-up, so the influence of task-driven cues are not considered. Besides, the saliency computation is also restricted by rules of *proximity* and *similarity*. That is, regions that are close or similar to the selected one are preferred. However, this proximity effect was refuted by later findings [Kröse and Julesz, 1989].

Milanese [1993] introduced one the first implementations based on Koch and Ullman's model. In his approach, feature maps are computed using filter operations. Milanese considered the following features to compute saliency: two opposite colour contrast (red-green and blue-yellow), 16 different orientations, local curvature and intensity (when no colour information is available). The saliency of each specific feature is computed using a *conspicuity operator*. This operator implements a *center-surround difference* which compares the local value of the feature map to their surround. As a result, a *conspicuity map* that highlights relevant parts of the image depending on the corresponding feature is obtained.

The different conspicuity maps are integrated into the final saliency map by applying a relaxation process. In this procedure, a small number of regions of interest are identified. However, a mechanism to sort the obtained

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

regions in terms of saliency is not provided. The main drawback of this model is the high computational cost associated with the many filter operations on different scales and the relaxation procedure. Even nowadays, the model is too computationally expensive for real-world applications [Frintrop et al., 2010]. Although the model was improved later, this first version is also a pure bottom-up approach.

The *Neuromorphic Vision Toolkit* (NVT) proposed by Itti et al. [1998] has become a reference in computational attention models. Both the theoretical model and its implementation have been used as a starting point for further developments due to its good documentation and the availability of the source code. It has been also employed as the standard benchmark for comparison. NVT applies the ideas of feature maps, saliency map, WTA procedure and Inhibition of Return from the Koch and Ullman's model. Features are computed using linear filters based on a center-surround mechanism in a similar way to Milanese's approach.

The main contributions of NVT refers to a detailed and concrete implementation of the system and its application to real-world environments, overcoming the implementation drawbacks of Milanese's proposal. Feature maps are computed using Gaussian image pyramids, that allow a faster multi-scale feature detection, ranging from 1:1 (scale 0) to 1:25 (scale 8). Each pyramid level is decomposed in channels for red, green, blue, yellow and local orientations (0°, 45°, 90° and 135°). Once the feature maps are obtained, they are normalised in order to promote those maps containing a small number of strong peaks of activity (conspicuous locations).

Feature maps are combined into three conspicuity maps, concerning with intensity, colour and orientation. Finally, the master saliency map is obtained as the arithmetic average of the three normalised conspicuity maps. The selection of the most relevant region and the inhibition of return are computed using a WTA network. Another important characteristic of NVT is that the model has been demonstrated to correlate with human eye movements in free-viewing tasks [Parkhurst et al., 2002; Itti, 2005].

Regarding the drawbacks of NVT, the employed center-surround mechanism, based on image pyramids, may lead to some inaccurate results because the method is fast but not very precise [Frintrop, 2006]. The system

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGA

also lacks in handling 2D image transformations such as translations, rotations and reflections [Draper and Lionelle, 2005]. Finally, similarly to other early artificial models of attention, the influence of top-down attention is not considered in this approach. Nevertheless, the NVT model has been kept up to date by the research group around Itti, producing enhanced models that will be detailed later.

3.1.2 Advanced models and top-down integration

A great variety of computational models of attention have been developed using the aforementioned approaches as reference. Some of them only introduce minor changes (for example, adding new features). However, other models have improved the primitive systems by overcoming their main drawbacks and introducing the top-down component of attention.

Milanese's model was one of the first systems to be enhanced by adding top-down information [Milanese et al., 1994]. Top-down cues are provided by an object recognition system, implemented using *Distributed Associative Memories* (DAM). The bottom-up part provides a reduced set of regions. Then, the object recognition process is applied to such regions. As a result, a map that highlights regions corresponding to recognised objects is obtained. In other words, regions corresponding to known objects are more salient. This procedure may lead to a visual search by providing the DAM with a single object. However, top-down influence is bounded to the conspicuity maps so it is not possible to strengthen single feature properties such as "round" or "green". Furthermore, the system is not able to independently learn the features of an object.

Backer and Mertsching present an innovative model of attention divided into two selection stages [Backer and Mertsching, 2000; Backer et al., 2001; Backer and Mertsching, 2003]. Furthermore, while the previous methods compute saliency pixel by pixel, the authors introduce a region-based model that performs a pixel clustering prior to the saliency computation. The first stage is similar to the previous proposals. Different conspicuity maps are computed based on basic features (symmetry, eccentricity, colour contrast and depth) and a small number of regions, typically 4, are selected. In the second stage, the *semi-attentive* stage, the selected regions are tracked and

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

top-down information is used in order to select a single element as the focus of attention. Therefore, whereas previous models mainly deal with static images, this approach is able to handle dynamic 3D environments.

As it was mentioned before, Itti's model has been improved over the years by the research group he leads. Navalpakkam and Itti [2005] add a multi-scale object representation in a long-term memory. The multi-scale object's features stored in this memory determine the relevance of the scene features depending on the current executed task. That is, a top-down behaviour is implemented in the system. The model learns the feature values from the different feature maps associated with the target and its close surrounding. This process is repeated for six scales, yielding a 42-dimensional feature vector. Then, the feature vector is used to bias the feature maps by applying a weighted sum. Thereby, top-down influence is considered by using a *feature maps fusion* procedure [Navalpakkam and Itti, 2006]. However, it is not possible to adapt the strength of the bottom-up and top-down respective influences. Thus, the system may be exclusively concentrated in the target's specific features. Additionally, top-down information is based on a prior supervised learning. Hence, it is difficult to search for unknown or generic targets.

Hamker [2006] has also developed an attention system that combines bottom-up and top-down contributions. The model mainly aims to imitate the neural processes presented in the human visual cortex. The bottom-up part is similar to Itti's NVT: several features are computed (intensity, orientation, red-green contrast, blue-yellow contrast and spatial resolution) and the corresponding conspicuity maps are obtained. Top-down influence is considered by learning the feature values of a presented target (usually on a black background). Such goal features are stored in a *Working Memory*. Then, the stored features influence the conspicuity of the computed features in a presented test scene. Thereby, the target information influences the conspicuity maps. Unlike Navalpakkam and Itti's approach, top-down information do not influence the early-stage processing of the feature maps.

Frintrop [2006] also proposes a model based on Itti's NVT, combining both top-down and bottom-up cues: Visual Object detection with a CompUtational attention System (VOCUS). In the bottom-up part, three feature dimensions are considered: intensity, orientation and colour. For each

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGA

dimension, conspicuities are computed on several scales and feature types (e.g., red, blue, green and yellow for the colour dimension) using image pyramids. Then, the maps are fused to a single saliency map. An Inhibition of Return procedure resets the values of the selected region when computing the next focus of attention. The top-down component of VOCUS is divided into two stages: learning and search. In the learning phase, VOCUS learns target-relevant features. A region of interest is manually provided in a test scene and its features are computed. In the search stage, the system considers the learned information in order to excite or inhibit specific features during saliency computing. As a result, a top-down saliency map is obtained. Finally, both top-down and bottom-up saliency maps are fused in a master saliency map. In comparison to Navalpakkam and Itti's approach, VOCUS yields a simplified and more user-friendly learning procedure, reducing the dimension of the feature vector employed (13 against 42). As drawbacks, Frintrop's model is mainly oriented to static scenes and the target still needs to be manually provided during the learning phase.

Kouchaki and Nasrabadi [2012] propose a modification for computing the top-down influence in a NVT-based attention model. They suggest a non-linear map fusion approach instead of the linear computation proposed by Navalpakkam and Itti [2006]. The bottom-up part of the model is identical to NVT. In order to compute the top-down contribution, they have designed a non-linear fusion kernel for combining the 42-dimensional feature vector obtained in the top-down extension of NVT. The 42 feature maps are applied as the inputs of a *Variadic Neural Network* [McGregor, 2007, 2008]. This supervised neural network can be trained using target information to purposefully weight the 42 feature maps. Although this method is faster in first hit detection than the original top-down extension of NVT, it has not been applied to dynamic scenarios and it still needs a previous training of the system with a specific target.

In contrast to the aforementioned *feature map fusion* approaches, the so-called*template-based models* work with abstract templates of low-level features (such as colour, shape, symmetry, etc.) of the target. These models work with a segmentation of the scene. Furthermore, no previous training is needed in order to search for specific objects and the system can manage

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

abstract information about the target ("look for green, rounded objects").

Aziz and Mertsching [2008a,b] suggest a template-based model for computing top-down cues. On the one hand, the bottom-up part is computed, as usual, as a sum of conspicuity maps. The system extends the idea of a region-based model proposed by Backer and Mertsching. Therefore, the first step consists in segmenting the input image into regions of homogeneous colour. Then, a set of low level features are computed using efficient algorithms, including colour contrast, orientation, size, symmetry and eccentricity. Finally, all the conspicuity maps are combined into a bottom-up saliency map applying a weighted sum. At first, the weights are initialised to specific and different values. For each subsequent step, the weights are adjusted to increase the contribution of the feature map offering the sharpest peak of saliency in the accumulated saliency map. On the other hand, the search target is defined as a set of top-down feature values to be sought among the early-segmented regions. The regions matching the template are marked as more relevant and a top-down saliency map is obtained. Finally, both bottom-up and top-down saliency maps are combined using a weighted sum. However, the procedure to calculate these weights depending on the current behaviour is not detailed.

Some years later, Tünnermann et al. [2013] modify the original Aziz and Mertsching's approach. This model focuses on enhancing the top-down contribution. Instead of a simple region target template, they propose the idea of multi-region templates. Thereby, the scene segmentation is searched for a specific configuration of template regions that are expected at their relative locations with certain spatial tolerance. An image of the target object is segmented, obtaining a set of regions. Then, one of such regions is selected as the *parent* of the template and compared with the segmented regions of the scene using the Aziz and Mertsching's procedure. When the parent is found, a *proto-object* is formed and the children of the target are used to adjust the overall top-down saliency. The multi-region target model improves the previous model by allowing heuristics to compensate perspective variations, occlusions and other alterations. Nevertheless, the model still lacks in explaining how the current behaviour selects the template to be searched, i.e., the variation of the top-down contribution over time.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGUNA

3.1.3 Object-based models

As it was explained in section 2.1.4 on page 16, computational attention models can be classified into *space-based* and *object-based* systems depending on the basic unit of attention employed. In contrast to space-based and region-based models, object-based approaches do not compute saliency pixel by pixel and they do not deploy attention to unstructured regions. Instead, they divide the scene into basic entities or pre-attentive objects, corresponding to coherent wholes.

One of the first object-based artificial attention model was proposed by Sun and Fisher [2003]. They present a grouping-based saliency method and a hierarchical selection of attention at different perceptual levels (points, regions and objects). In their proposal, attention is object-driven as well as feature-driven. First, a set of basic features (colour, intensity and orientation) are computed using multi-scale pyramid filters. Then, a perceptual grouping procedure is applied to obtain hierarchical structures of objects and space (they may be a point, a region, an object or a hierarchical structure of groupings). The salience of a grouping is a function of all saliency contributions coming from its components. Finally, groupings are fed to an attention competition pool where a WTA process selects the most relevant one. The main drawback of this model lies in the formation of the groupings, that must be manually drawn. Since the proposal is focused on the novelty of applying an object-based model, a detailed top-down process is not provided.

Another early approach computing saliency from cluttered scenes was proposed by Gao and Vasconcelos [2004]. They propose a *discriminant saliency* definition grounded on the recognition problem. In other words, the model is based on the idea that the salient features of a visual class are those that best distinguish it from all other visual classes of recognition. They define top-down attention as classification with minimal expected error. Difference-of-Gaussians and Gabor filters are used for bottom-up saliency, measuring the saliency as the Kullback-Leibler divergence between the histograms of the filters response in a center-surround processing. This model also requires a supervised learning process: a training database is needed in order to determine the salient features of each class.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

Orabona et al. [2005, 2007] propose a model of visual attention based on the concept of proto-objects (see section 2.2.5 on page 24) as units of visual information that can be bound into a coherent and stable object. They compute these proto-objects by employing the watershed transform to segment the input image using edge and colour features in a pre-attentive stage. As input, the model uses log-polar colour images. The bottom-up saliency is based on a center-surround computation of basic opponent colour channels (red-green, green-red and blue-yellow), resulting in a Difference-of-Gaussians response. The top-down component is computed as the euclidean distance in the colour opponent space between the average colour of perceived proto-objects and the average colour of the target. Therefore, the top-down component is bounded to the visual search of a given object. However, a supervised learning of the target is not necessary.

Another model based on the concept of proto-objects has been developed by Wischnewski et al. [2010]. The model is mainly based on Bundesen's Theory of Visual Attention (TVA) [Bundesen, 1990] (see section 2.2.4 on page 23). Their proposal draws attention depending on three levels: a low-level of basic features (such as colour or orientation), a medium level of more sophisticated features associated with ellipsoidal proto-objects (such as rough shape or size) and a task-dependent high-level (top-down). They employ a retinal model to compute static feature maps associated with orientation (five different angles), colour (red-green and blue-yellow contrasts) and intensity. Then, the ellipsoidal proto-objects are computed applying a multidimensional hierarchical cluster approximation algorithm. Not only static but also dynamic features and proto-objects are obtained using segmentation of motion energy features [Belardinelli et al., 2010]. Merging both static and dynamic proto-objects provides the bottom-up component of attention. Finally, a template-based top-down selection is applied. The top-down function assigns a weight to each proto-object depending on the pertinence of its features. That is, the proto-object matching the most of the searched features obtains the greatest weight.

Yu et al. [2010] propose a model of attention that segments the scene into proto-objects in a bottom-up strategy based on Gestalt theories. The top-down component is computed using models of objects that are relevant for a specific task. These models are stored in a long-term memory (LTM).

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGUNA

If a task-specific object is provided, the model recalls the corresponding representation from LTM and deduces the features to be highlighted. The models stored in the LTM are learned using a supervised algorithm. Top-down and bottom-up saliency maps are combined using a weighted sum, where the weights are manually predefined. In subsequent improvements of the model, the top-down biasing parameters are calculated from the task-relevant object using Bayesian inference [Yu et al., 2012] and the pre-attentive segmentation procedure has been refined using irregular image pyramids [Yu et al., 2013].

3.1.4 Models using alternative paradigms

Most of the computational attention models, such as the ones presented in the sections above, are directly inspired by psychological and neurophysiological findings. Nevertheless, other visual attention systems have been developed using alternative paradigms such as Bayesian approaches, Connectionist models, Information Theoretic models or Pattern Classification methodology [Borji and Itti, 2013]. In this section, some of these alternative models are presented.

Connectionist models process the input data with neural networks, mimicking brain neuronal processes. One of the most renowned *connectionist models* was proposed by Tsotsos et al. [1995]. The *Selective Tuning model of Visual Attention* consist of a pyramidal architecture. The selected item at the top of the hierarchy has a *pass zone*, the pathway selected for further processing, and an *inhibit zone*, locations that have been inhibited because they do not belong to the selected item. Top-down cues can be considered by inhibiting all regions with features that are different to the target's ones. The model has been implemented for different features such as luminance, orientation and colour opponency. Recent versions of the model also consider motion as feature [Tsotsos et al., 2005]. In these approaches, only one feature dimension is processed at once.

In Bayesian models, prior knowledge (such as gist or scene context) and sensory information are probabilistically combined applying the Bayes' rule in order to obtain and select objects of interest. Torralba [2003], Torralba et al. [2006] and Oliva et al. [2003] propose a Bayesian framework for visual

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

search tasks. They compute bottom-up saliency through a global feature that measures the probability density of presence of the target object in the scene. In order to apply Bayes' rule, gist top-down information is employed. For scene context modelling, they apply *Principal Component Analysis* (PCA) to a large number of natural images. It can be noted that, in Bayesian models, top-down and bottom-up contributions are not easily separable processes.

Machine Learning techniques have also been used to derive attention models. Judd et al. [2009] propose to learn a model of saliency directly from human eye movement data, acquired using an eye-tracking system. They trained a linear *Support Vector Machine* (SVM) using human fixation data. The feature vector is composed of a set of low (orientation, colour contrast, intensity, RGB channels...), mid (horizon line) and high level (faces) image features. They used a database of about 1000 images observed by 15 subjects. The most important disadvantage of these models lies in the subjective data employed to build the model. It is very difficult for people to disengage from their own expectations, motivations or even present concerns when observing a scene. Therefore, these models are data-dependent and a fair model comparison may not be easily addressed.

Finally, Information Theoretic models consider that localised saliency computation can be used to maximize the perceived information. These models are based on selecting the most informative parts of the scene while discarding the rest. Rosenholtz [1999] suggests a simple model for predicting ease of visual search based on a quantitative measure of target saliency. She considers that saliency computation is equivalent to a parametric test for outliers to a statistical distribution. From the distribution of features (an uniform space), mean and covariance of distractor features are computed. Then, the Mahalanobis distance between the target feature vector and the mean of distractor distribution is obtained. As a result, the obtained outliers represent the most relevant parts of the scene.

3.2 Linking perception and abstract reasoning

The connection between perception and abstract reasoning is a fundamental key when developing autonomous agents. Perceived entities need

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

to be attached to symbolic representations whereas the effects of abstract operations should imply a reallocation of perceptual capabilities, in order to focus on suitable real-world elements. Therefore, an autonomous agent should acquire sufficient internal representations to produce suitable responses to events occurring in its surrounding [Markman and Dietrich, 2000]. Without addressing the linking of perception, action and abstract reasoning, a system should be unable to autonomously analyse, understand and handle a real and continuous changing environment.

As it was explained above, top-down computational attention models try to bridge the gap between perception and deliberative reasoning by assuming that high-level information can bias the relevance of the elements in a scene. Nevertheless, since this problem is not exclusive of Computer Vision, scientists from other fields, such as Artificial Intelligence and Automated Planning, have also investigated solutions to successfully link perception and reasoning.

On the one hand, cognitive architectures aim to completely characterise an intelligent system, providing the necessary functional processes, representation of elements and relationships among them, including decision and environment observation. On the other hand, other models deal with the classical unidirectional assumption of the perception-action loop by applying hierarchical planning or Bayesian approaches. They specify the features to be searched in the scene in order to accomplish a specific task. Some example systems belonging to both categories are detailed in the next sections.

3.2.1 Cognitive architectures

Cognitives architectures try to explain the whole underlying structure for an intelligent system. They support the creation and understanding of synthetic agents, able to show the same capabilities as humans. Therefore, the problem of linking perception and reasoning is implicitly addressed. That is, the agent acquires information depending on its knowledge, beliefs and expectations. Indeed, a similar process to the top-down contribution addressed by computational attention models. This section summarises

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

some examples of cognitive architectures, specially focusing on those aspects related to closing the perception-reasoning-action loop.

The first versions of Soar [Laird et al., 1987; Laird, 2008] consists of a single long-term memory, which is encoded as production rules, and a symbolic single short-term memory. The latter holds the agent's assessment of the current situation derived from perception. The long-term memory maintains knowledge in form of production rules, organised in terms of operators associated with problem spaces. The decision procedure selects operators and detects *impasses*, i.e., situations where knowledge about operator selection is insufficient to determine the next operator to apply. Soar uses a processing cycle to deal with environmental changes. In this process, changes to perception are sent to the short-term memory. Then, production rules compute entailments of short-term memory, propose operators and establish preferences among them. After that, a fixed decision procedure selects the operator. If the preferences are insufficient for making a decision, an *impasse* arises and Soar recursively divide the goal into substates in order to resolve the *impasse*. Finally, the operator is applied. Later improvements of the model have added separate episodic and semantic memories.

ACT-R [Anderson and Lebiere, 1998; Anderson et al., 2004] is a family of cognitive architectures concerned with modelling human behaviour. The architecture is organised into a set of modules, processing a particular type of information, that form a short-term memory. There exist sensory modules, motor modules for action, an intentional module for goals and a deliberative one for long-term declarative knowledge. A long-term memory of production rules coordinates the processing of the different modules. The modules place chunks in buffers where they can be deleted by a production system that responds to patterns of information in the buffers. Although early versions of ACT-R mainly deal with higher level cognition, latest versions also aim to relate cognition to perception. The authors employ a quite simplified attention system as perception module. The production rules simply specify a set of feature constraints in the visual search (e.g. "colour: red"). However, the procedure to finally achieve the guidance is not detailed and it is suggested for future work.

ICARUS [Langley et al., 2004] is a hierarchical cognitive architecture that stores two different forms of knowledge. *Concepts* describe classes of

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LÁLAGA

environmental situations in terms of other concepts and percepts. *Skills* specify how to achieve goals by decomposing them into ordered subgoals. Concepts (grounded on perceptions) and skills (related to executable actions) define relations among objects. The architecture basically operates on a *recognise-act cycle*. At the beginning of this cycle, descriptions of visual objects are placed at a perceptual buffer. Then, the system compares primitive concepts to the perceived items and adds matched instances to a short-term memory as beliefs. This process is repeated until the system infers all deductively implied beliefs. Then, from a top-level goal, the architecture finds a downward path through the skill hierarchy in which each subskill has satisfied conditions but an unsatisfied goal. Once a primitive skill with executable actions is reached, the system applies such action to change the environment and the cycle begins again.

HERMES [Bischoff and Graefe, 2004] is a humanoid robot developed to study some key technologies that are important to service robots. It includes a system architecture that allows integration of multiple sensors and actuators as well as knowledge bases and a human-friendly interface. The architecture is based on Rasmussen's *Model of Human Performance* [Rasmussen, 1983] (cf. section 5.1.1 on page 85). A skill-based situation-oriented deliberative component is responsible for situation assessment and behaviour management in order to be able to select behaviours and to achieve long-term goals. The deliberative process analyses the perceived information to figure out the current environmental context (present objects and their relationships). Then, a set of adequate skills is selected and executed to change the situation in a desired way, closing the perception-action loop.

3.2.2 Other systems involving perception and reasoning

While cognitive architectures suggest a complete solution for intelligent agents, other approaches simply aim to characterise the perception-action loop closing problem. These models are usually less general than cognitive architectures and they are often quite restricted to specific applications.

Eidenberger et al. [2009] propose an active perception planning module embedded in a distributed cognitive system. The approach is based on the

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

wish list concept, enabling distributed planning, reasoning and decision actions. The architecture has been designed for a specific purpose: supporting a householder robot that mainly has to manipulate objects in a local environment. The perception system is based on the probabilistic modelling of multi-object scenes, using a Partially observable Markov Decision Process (POMDP). Since the proposal is a distributed architecture, all modules autonomously perform local planning and control actions. One module can not accomplish a task if the uncertainty in the belief of the actual scene is too high. Then, a wish list containing suggested actions, their benefits and costs is sent to the overall control system. After a general evaluation, a new request in form of a single action of the wish list is passed to the perception module. In this version of the architecture, the only available action for active perception modification consists in modifying the sensor position by mechanically changing the viewpoint.

Sridharan et al. [2010] has also developed a POMDP hierarchical planning framework whose levels match the cognitive requirements of visual processing on a robot. The framework links sensing and information processing to the current task. The perception system is based on the acquisition of regions of interest (ROI) from a image. These regions correspond to simple coloured shapes (squares, circles...) and they are computed as the difference between the actual image and a previously learned scene. Information is obtained from the scene using a set of visual operators, including colour, shape and pre-learned models of objects. Then, a hierarchical POMDP planning system plan a sequence of operators to answer queries about the scene (e.g., "which objects in the scene are blue?"). The resulting action consists in choosing a specific viewpoint to correctly observe the environment, reducing the uncertainty. It can be noted that no action modifying the state of the scene (such as moving an object) is executed.

The connection between perception and action is often studied when developing task-specific systems. For example, Gilet et al. [2011] propose a Bayesian Action-Perception (BAP) model to study the interaction of such processes in handwritten letter recognition and production. As main hypothesis, there exists an internal representation associated with each letter and each writer in the form of probability distributions for the representation of letters. These representations are considered as axes between

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LÁLAGA

perception and action. Therefore, perception and action are assumed to be independent. Letters are encoded as a sequence of via-points and the motor system influences perception via internal simulation of movements. The perception system is restricted to the simple task of extracting the aforementioned via-points from the trajectory. As a result, the model can be used to automatically solve cognitive tasks, such as reading with motor simulation, writer recognition or copying of letters and trajectories, using Bayesian inference. Thereby, visual information can be used to correct motor gestures, reaching the classical problem of modelling closed-loop control.

Ferreira et al. [2012, 2013] have also developed a Bayesian framework for multimodal active perception, the Bayesian Volumetric Map (BVM) [Ferreira et al., 2008], integrating vision, audition and vestibular sensing with no data association (i.e. object detection and localisation only, no recognition). This framework is based on an egocentric, log-spherical spatial configuration used as a short-term memory inspired on the dorsal pathways and the superior colliculus of the human brain. In addition to the commonly used colour, orientation and intensity features, the visual part of the perception system also uses optical flow and face detection for saliency computation, while auditory saliency relies on an intensity feature of sound sources localised in log-spherical space. Similarly to the Bayesian approaches mentioned above, perception is guided in form of gaze shifts to regions with a combined assessment of high uncertainty and high relevance, determined by both entropy of the occupancy grid and saliency measurements, respectively. A simple Inhibition of Return mechanism is also included based on the fixation point of the previous step. As future work, they plan to model goal-oriented active perception by applying the concept of *attentional sets* [Corbetta and Shulman, 2002]. Thereby an appropriate active perception behaviour would emerge from changes in observer goals.

Other authors have faced the problem of perception-action loop closing from the perspective of autonomous cognitive learning. The work developed by Shevchenko et al. [2009] and Windridge and Kittler [2010] aims to learn cognitive capabilities in a hierarchical open-ended manner. The system is based on a hierarchical *bootstrapping* model. The system modifies itself via recursive generalising of parametric spaces that links perceptual

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

and motor domains to represent environmental *affordances*. The learning process is performed as a continuous perception-action cycle and cognitive capabilities are represented by pairwise links between acquired behavioural skills and perceptual concepts. Their proposal presents a really interesting approach in terms of behavioural learning. However, they show a strong unidirectional assumption of the loop closing problem since they are mainly based on the concept of *affordances*. Thus, the effects of deliberative decisions on perceptual modulation are not conveniently considered.

3.3 The bidirectional perception-reasoning loop problem

As it has been detailed throughout the chapter, although top-down computational attention systems consider that perception might be influenced by high-level effects related to task motivation and abstract reasoning, they normally fail in explaining the mechanisms needed to derive such top-down influence. These models are very complete from the attention point of view. However, their integration into new or already existing general architectures of perception and reasoning is still an open challenge.

Regarding the models detailed in section 3.2, they principally lack in assuming both directions in the perception-reasoning loop closing problem. Thus, these solutions consider an unidirectional perspective of the problem. For instance, approaches such as ICARUS and Soar mainly focus on the perception to action connection. Perception is mostly employed to increase the knowledge about the environment and to select the actions that can be executed.

Otherwise, systems like ACT-R presents just the opposite lack of top-down models of attention. That is, the model is able to infer a set of visual features constraints but the connection with the perception system is roughly addressed.

Planning and Bayesian models (Sridharan et al. [2010]; Gilet et al. [2011]) are strongly unidirectional approaches. Naive vision systems are

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LA RIOJA

employed to simply increase the observability of the decision process, reducing the uncertainty. Besides, they are bounded to very specific environments and tasks. Although a more elaborated and general-purpose system is provided by Ferreira et al. [2012, 2013], their model still constrains the visual guidance to mechanical changes of the viewpoint, dismissing the advantages of visual attention modulation.

Perhaps HERMES skill-based model is the most interesting solution in terms of generality and bidirectionality. However, the system is quite monolithic, omitting Rasmussen's rule-based level. Therefore, the addition of new tasks and components is not trivial. Furthermore, the advantages provided by attention models are not considered.

Summarising, the problem of bidirectionally closing the perception-reasoning-action loop has not been successfully and completely addressed yet. On the one hand, attention models need to be explicitly connected to complex and abstract deliberative modules. On the other hand, a cognitive architecture able to adapt itself to perceived elements while guiding a visual attention system is also necessary.

Hence, this thesis aims to overcome these issues with a double contribution. First, **an object-based computational attention system** with a specific interface to be linked to deliberative layers (chapter 4). Then, **a general-purpose need-based cognitive architecture** that employs such attention system to bidirectionally connect perception and reasoning, while interacting with a real-world environment (chapter 5).

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

UNIVERSIDAD DE MÁLAGA
Publicaciones y Divulgación Científica

4 Object-based Visual Attention System

"Facilius per partes in cognitionem totius adducimur"

Lucius Annaeus Seneca

As the basis of the general purpose cognitive architecture introduced in this work, this chapter presents the perception part of the system, consisting in an **Object-Based Visual Attention System (OBVIAS)** for social robots. There exist two main advantages of using an attention model as perception system. First, the computational cost is reduced because only some parts of the whole image need to be processed. Therefore, it is an interesting solution to deal with huge amounts of information as occurs in a natural scenario. Second, the selected parts of the image are the most relevant for the ongoing task, so system performance is enhanced.

As it was mentioned in chapter 2, there exist two main contributions to attention. The system is explained meeting both top-down and bottom-up components of attention. Additionally, the perceptual algorithm that allows to divide the image into visual units of perception, called *proto-objects*, and the solution given to the phenomenon of the *inhibition of return* in dynamics scenarios are detailed. Finally, the interface between the perception module and other components in a complex system is explained.

Since the OBVIAS system is clearly oriented to Social Robotics, some additional considerations must be taken into account. First, the model has to be versatile enough to operate in an unrestricted environment. It should

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

be able to pop out a great variety of elements, including people. Besides, in a social robot, autonomy is imperative. Hence, it is very important to select computationally efficient algorithms and maintain a reasonable computational load.

4.1 Overview

Fig. 4.1, shows an overview of the perception system. The proposed attention system integrates task-independent bottom-up processing and task-dependent top-down selection. The units of attention are the so-called proto-objects [Rensink, 2000]. These proto-objects are defined as units of visual information that can be bounded into a coherent and stable object, which is consistent with Gestalt factors of closure and colinearity. On the one hand, the bottom-up component determines the set of proto-objects in the image, describing them by a set of low-level features that are considered relevant to determine their corresponding saliency values. On the other hand, the top-down component weights the low-level features that characterise each proto-object to obtain a single saliency value depending on the task to perform.

The general two-stages structure of the attention mechanism is related to a previous proposal [Backer and Mertsching, 2003]. In the *pre-attentive stage*, the different proto-objects in the image are extracted using a perceptual segmentation algorithm based on a hierarchical framework [Marfil and Bandera, 2009]. Then, the relevance of each proto-object is computed taking into account different low-level features weighted by an attentional set ($\vec{\lambda}$) [Corbetta and Shulman, 2002] stored in a *Perception-Modulation Memory* (PMM). Depending on the specific value of the attentional set, the system is able to modify the influence of each low-level feature in the global saliency computation, so attention can be guided in a top-down way. As the result of this stage, a set of proto-objects ordered by their saliency is obtained and saved in a *Working Memory*.

The next stage, the *semi-attentive* stage, deals with the management of the Working Memory (WM) and the *Inhibition of Return*(IOR). A tracker module keeps permanently updated the position of each element in the

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGA

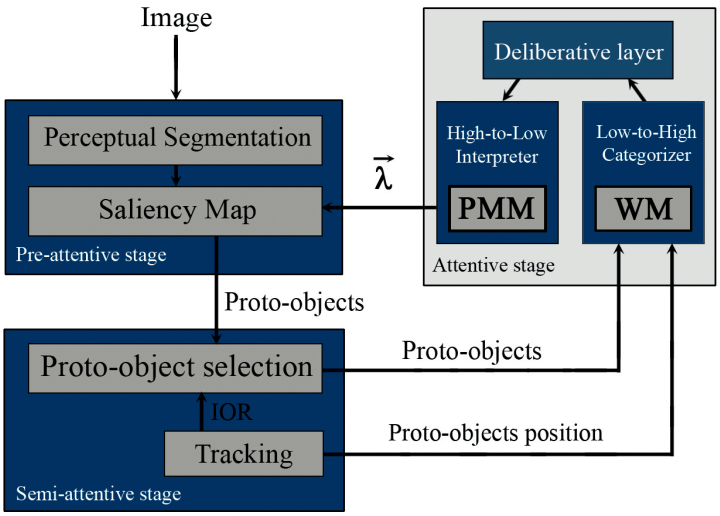


Figure 4.1: Overview of the Object-Based Visual Attention Model. The system is divided into two stages. In the *pre-attentive stage*, the *proto-objects* contained in the image are extracted and their saliency values are computed. The *semi-attentive stage* deals with the inhibition of return in dynamic scenarios. The *Working Memory (WM)* and the *Perception Modulation Memory (PMM)* are the interface with high level layers, storing the most relevant *proto-objects* and the *attentional sets*, respectively.

WM, allowing to manage not only moving objects but also camera and robot egocentric movements. Thereby, it is avoided to attend an already selected proto-object. In brief, the main steps of the attention model are described in Algorithm 1.

The following sections explain in detail the different algorithms and subsystems forming the object-based visual attention system.

4.2 Perceptual Segmentation

In OBVIAS model, attention is deployed to proto-objects instead of disembodied spatial locations of the image. In the system, proto-objects are defined as the blobs of uniform colour and disparity of the image that

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

Algorithm 1: Attention algorithm

repeat
- Acquire image (and depth data if available);
- Read attentional set from PMM;
- Divide image into proto-objects using a perceptual segmentation;
- Compute saliency of obtained proto-objects;
- Update position of older proto-objects in WM;
- Delete lost elements;
- Update saliency of older proto-objects in WM;
- Fill up WM with new proto-objects if they are more salient;

forever

are bounded by the edges obtained using a Canny edge detector [Canny, 1986]. Considering that an attention system is an early visual stage, it has to fulfil some requirements about speed and performance. Hence, in order to compute proto-obejcts, a computationally efficient hierarchical framework for image processing is employed: the *Bounded Irregular Pyramid* (BIP) [Marfil et al., 2006].

The BIP combines a $2 \times 2/4$ regular structure with a simple graph [Marfil and Bandera, 2009]. With respect to other irregular pyramids, the main advantage of the BIP is that it is able to obtain similar segmentation results but in a faster way [Marfil et al., 2006; Vázquez-Martín et al., 2009]. Both irregular and regular decimation processes are accomplished to build the BIP, using a specific similarity criterion. Then, the arcs between nodes in the same level are obtained applying a neighbourhood criterion. Not only simple distances (as colour difference) but also complex ones involving different image features can be used as similarity criteria. A more detailed description about the BIP is provided in Appendix A.

As the process to group image pixels into higher-level structures can be computationally complex, perceptual segmentation approaches typically combine a pre-segmentation step with a subsequent perceptual grouping step [Marfil and Bandera, 2009; Marfil et al., 2009]. The pre-segmentation step performs the low-level definition of segmentation. It groups pixels into homogeneous clusters. Thereby, pixels in input image are grouped into

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

blobs of uniform colour, replacing the pixel-based image representation. Besides, these regions preserve the geometric structure of the image because each significant feature contains at least one region. The perceptual grouping step conducts a domain-independent grouping which is mainly based on properties such as proximity, closure or continuity. Both steps are conducted using the aforementioned decimation process but employing different similarity criteria between nodes.

In order to compute the pre-segmentation stage, a basic colour segmentation is applied. In this case, a distance based on the HSV[1] colour space is used. Two nodes are similar (they share a similar colour) if their HSV values are less or equal than a similarity threshold τ_{colour}.

In the perceptual grouping step, the roots of the pre-segmented blobs are considered the first level of a new segmentation process. In this case, two constraints are taken into account for an efficient grouping process: first, although all groupings are tested, only the best groupings are locally retained; and second, all the groupings must be spread on the image so no part of the image takes advantage. As segmentation criterion, a more complex distance is employed instead of a simple colour threshold. This distance has three main components: the colour contrast between blobs, the edges of the original image, obtained using a Canny detector, and the depth information of the image blobs in form of disparity (if available). To avoid working at pixel resolution, which decreases the computational speed, a global contrast measurement is used instead of a local one. Then, the distance $\xi(n_i, n_j)$ between two nodes n_i and n_j is defined as:

$$\xi(n_i, n_j) = \sqrt{\omega_1 \left[\frac{d(n_i, n_j) \cdot p_i}{\alpha \cdot c_{ij} + \beta(b_{ij} - c_{ij})}\right]^2 + \omega_2 \left[\delta(n_i) - \delta(n_j)\right]^2} \qquad (4.1)$$

where $d(n_i, n_j)$ is the HSV colour distance between n_i and n_j, $\delta(x)$ is the mean disparity (if a sensor able to provide depth information is used) associated with the base image region represented by node x, p_i is the number of pixels in the perimeter of n_i, b_{ij} is the set of pixels in the common boundary between n_i and n_j and c_{ij} is the set of pixels in this common boundary which corresponds to pixels of the boundary obtained using the Canny detector. α and β are two constant values used to control the influence of

[1] HSV stands for **H**ue, **S**aturation and **V**alue

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

the Canny edges in the grouping process. ω_1 and ω_2 are two constants which weight the terms associated with colour and disparity. Their values depend on the accuracy of the depth information provided by the sensor. If no depth (or disparity) information is provided, ω_2 is set to 0. These parameters should be manually tuned depending on the application and the environment.

The new hierarchical levels in the BIP are obtained applying a new segmentation process. In this case, the algorithm takes the result of the pre-segmentation stage as input and the complex distance calculated in eq. (4.1) is employed. Therefore, two nodes are similar if the distance $\xi(n_i, n_j)$ between them is equal or less than a threshold τ_{percep}. The grouping process is iterated until the number of nodes remains constant among two consecutive levels. After the perceptual grouping, the nodes of the BIP with no parents are the roots of the proto-objects.

Some examples of computed proto-objects are shown in fig. 4.2. It can be noted that proto-objects usually fit real objects (see the four elements over the table in the second row image or the case, the green ball and the yellow ball in the third row example). Nevertheless, it should be mentioned that obtained proto-objects do not always correspond to natural objects. For instance, the cube in the first row image (a synthetic one) is split into three different proto-objects due to the great colour contrast between its faces. Illumination can also divide one natural element into several proto-objects as can be observed in the third row image, where the table and the red ball are separated into two proto-objects respectively. An object split into several proto-objects may cause that the same element could be attended twice. However, this is a minor issue from the attention point of view, where inhibition of return procedures are applied. Proto-objects can also be obtained even without depth (or disparity) information as it is shown in the first and the second row images.

In conclusion, the perceptual segmentation provides a reasonable, plausible and fast to compute explanation about the objects present even in complex scenes, despite the inherent errors associated with the segmentation process.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGUNA

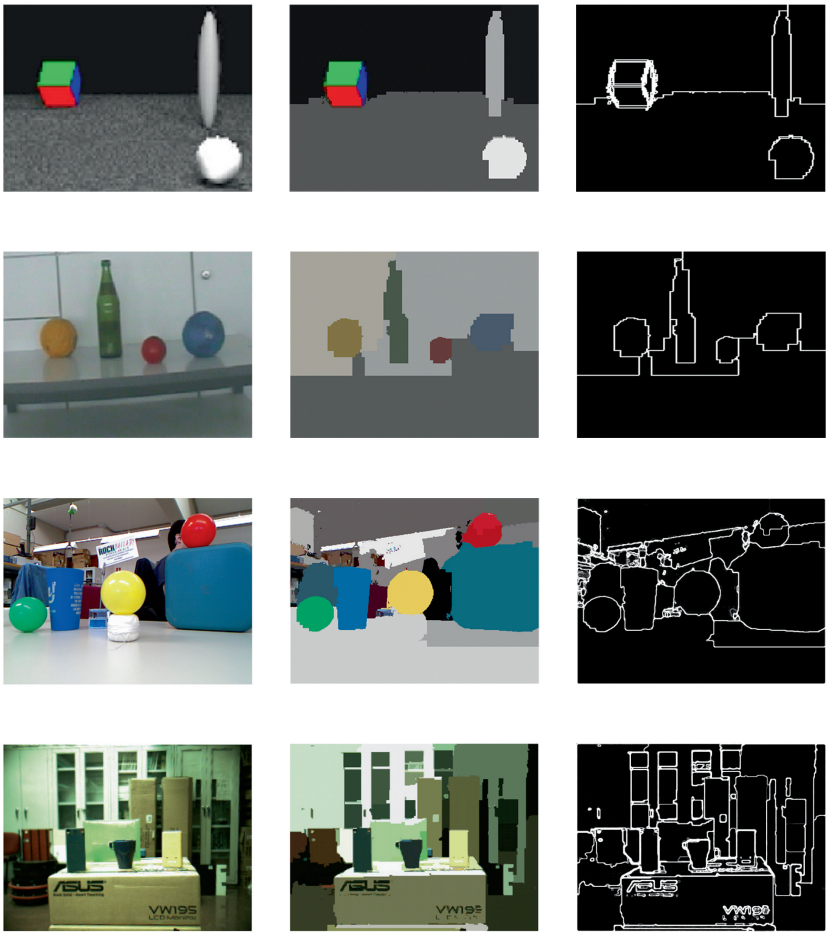


Figure 4.2: Some examples of Perceptual Segmentation and proto-objects computation. (a) original image; (b) perceptual segmentation (proto-objects); (c) boundaries of the obtained pro-objects. In the first row, a synthetic image is used. The second row corresponds to a standard webcam image. Image in the third row has been captured using a Kinect sensor. The image in the fourth row has been provided by a stereo pair of cameras.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

4.3 Multi-feature saliency computation

Once the scene is divided into proto-objects, the next step is the selection of the most relevant one. According to Treisman and Gelade [1980], this process is based on the computation of a set of low-level features. But, what features must be taken into consideration? What features really guide attention?

According to psychological studies, some features, such as colour [Treisman and Souther, 1985], motion [McLeod et al., 1988] or orientation [Wolfe et al., 1992], clearly influence in saliency computation. These three features, plus size, are catalogued in Wolfe and Horowitz [2004] as the only undoubted attributes that can guide attention. Wolfe also offers in his work a complete list of features that might guide the deployment of attention, grouped by their likelihood to be an effective source of attentional guiding. He differentiates among the aforementioned undoubted attributes, probable attributes, possible attributes, doubtful cases and probable non-attributes (see Table 4.1) [Wolfe and Horowitz, 2004].

Undoubted attributes	Probable attributes	Possible attributes	Doubtful cases	Probable non-attributes
Colour	Luminance onset	Lighting direction	Novelty	Intersection
Motion	Luminance polarity	Glossiness	Letter identify	Optic flow
Orientation	Vernier offset	Expansion	Alphanumeric category	Colour change
Size	Stereoscopic depth	Number		3-D volumes
	Pictorial depth cues	Aspect ratio		Faces
	Shape			Your name
	Line termination			Semantic category
	Closure			
	Topological status			
	Curvature			

Table 4.1: Attributes that might guide the deployment of attention. Source: Wolfe and Horowitz [2004]. Selected features for OBVIAS model are highlighted.

A recent survey on attention mechanisms for socially interactive robots [Ferreira and Dias, 2014] has reviewed the basic features employed in different models. They notice that, additionally to Wolfe's classification, some non-basic saliency metrics have had great success in predicting overt visual attention behaviours: *Bayesian surprise* (related to novelty) and *information maximization*. Moreover, they have observed that colour, motion-related and face-related basic features are used by 80%, 55% and 80% of the

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGA

surveyed work, respectively, while luminance and orientation are employed by 45%.

Another important issue when selecting features to develop an artificial attention system is concerned with computational cost. Computing a large number of features provides a richer description about elements in the scene. However, the associated computing time could be unacceptable. Hence, it is necessary a trade-off between computational efficiency and the number and type of the selected features.

Following the previous guidelines, eleven different features have been selected to compute saliency in OBVIAS model. From the undoubted attributes, orientation and colour have been chosen. Regarding colour, not only contrast is taken into account but also similarity with basic colours as blue, red, green and yellow. Because there is no background subtraction in the perceptual segmentation, larger proto-objects usually correspond to non-relevance parts of the image (e.g. walls, floor or empty tables). Therefore, size feature is not employed to avoid an erroneous highlighting of irrelevant elements. Motion is discarded due to computational cost restriction. Although intensity contrast is not considered an undoubted feature, it has also been included as a special case of colour contrast (intensity deals with gray, black and white elements).

From the remainder of available possible attributes, those describing shape and location have been considered as more suitable for a complete description of the objects in scene. Location is calculated in terms of proximity to the visual sensor. Regarding the shape, two features are taken into account: symmetry, which allows to discriminate between symmetric and non-symmetric elements, and roundness, a measure about the closure and the contour of an object.

Finally, in order to reach a social interaction with humans, it seems to be reasonable to include features able to pop out people from a scene. Although some works directly consider faces as a feature [Judd et al., 2009], experimental studies differ [Nothdurft, 1993; Suzuki and Cavanagh, 1995]. Faces themselves do not guide attention but they can be separated into basic features that really achieve the guidance [Wolfe and Horowitz, 2004]. In general, global properties are correlated with low-level features that explain search efficiency [Greene and Wolfe, 2011]. Consequently, OBVIAS

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

model uses similarity with skin colour as an undoubted feature to guide attention to human faces in combination with other features as roundness.

To summarise, saliency is computed in terms of the following features: colour contrast, intensity contrast, proximity, symmetry, roundness, orientation, similarity to basic colours (red, blue, green and yellow) and similarity to skin colour. All features values are normalized in the range $[0 \dots 255]$ in order to have an homogenized calculus.

In any case, the selected set of features should not be considered as an immovable and closed group. New features can be added and the existing ones can be replaced or removed. Nevertheless, these changes should always follow the aforementioned guidelines about attention guidance and computational cost.

4.3.1 Colour contrast and Intensity contrast

These features measure how different a proto-object is with respect to its surrounding in terms of colour and luminosity. Because they are part of Treisman's theory, these features have been used in artificial attention systems since the first models [Itti et al., 1998].

Since proto-objects are the result of a perceptual segmentation process, the colour contrast, $(ColCON)$, of a specific proto-object, $\mathcal{P}_i$, can be computed as the mean colour gradient along its boundary to the neighbours in the segmentation hierarchy:

$$ColCON_i = \frac{S_i}{b_i} \sum_{j \in N_i} b_{ij} \cdot d\left(< C_i >, < C_j >\right) \tag{4.2}$$

where b_i is the perimeter of $\mathcal{P}_i$, N_i is the set of proto-objects that are neighbours of $\mathcal{P}_i$, b_{ij} is the length of the perimeter of $\mathcal{P}_i$ in contact with proto-object $\mathcal{P}_j$, $d\left[< C_i >, < C_j >\right]$ is the HSV colour distance between the colour mean values $< C >$ of proto-objects $\mathcal{P}_i$ and $\mathcal{P}_j$ and S_i is the mean saturation value of proto-object $\mathcal{P}_i$.

Because of the use of S_i in the colour contrast equation, white, black and gray proto-objects are suppressed. Thus, a feature about intensity contrast is also introduced. The intensity contrast, $(IntCON)$, of a proto-object,

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

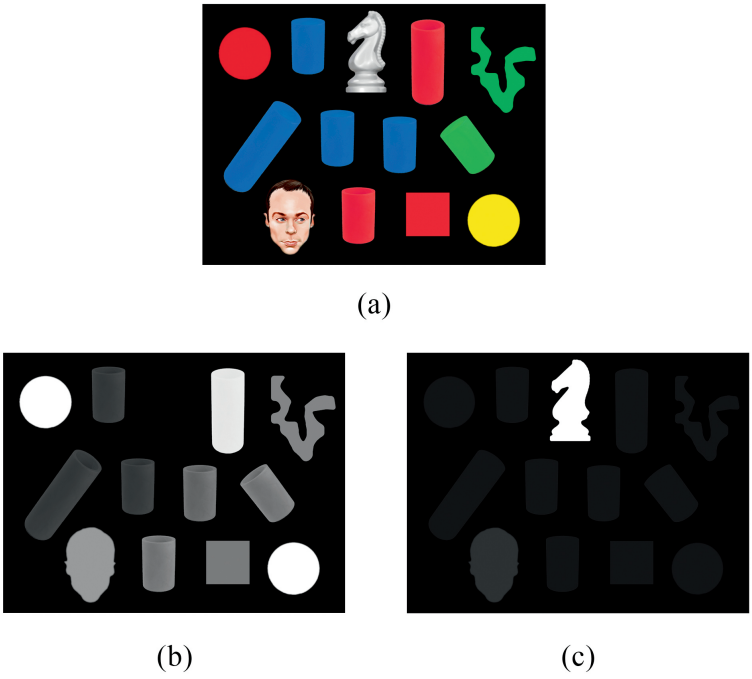


Figure 4.3: Example of colour contrast feature computation. (a) original image; (b) colour contrast conspicuity map; (c) intensity contrast conspicuity map. Relevant proto-objects are brighter.

$\mathcal{P}_i$, is computed as the mean luminosity gradient along its boundary to the neighbours:

$$IntCON_i = \frac{1}{b_i} \sum_{j \in N_i} b_{ij} \cdot d\left(< I_i >, < I_j >\right) \tag{4.3}$$

being $< I_i >$ the mean luminosity value of the proto-object $\mathcal{P}_i$.

As the result of the procedure, two contrast conspicuity maps are obtained, where proto-objects differing in colour (or intensity) from its surrounding are marked as more relevant. Fig. 4.3 shows an example of colour contrast and intensity contrast conspicuity maps. The colour contrast one (fig. 4.3.b), shows that the elements differing in colour from their neighbours are more salient (brighter in the conspicuity map). For instance, that is the case of the red circle (surrounded by blue cylinders), the yellow circle (next to red and green elements) or the red high cylinder. Elements

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

with similar colour neighbours obtain a lower saliency value. For example, the inclined blue cylinder is surrounded by other two blue elements, so it is almost no relevant. Note the special case of the gray chess knight. Due to the colour contrast computation process, gray parts are not taken into account. Therefore, there is no relevance assigned to this element despite the fact that it is the only one with that colour. This is the reason to compute an intensity contrast conspicuity map (fig. 4.3.c). In this case, elements with a saturated colour are omitted and the chess knight becomes the most relevant form.

4.3.2 Proximity

Another important parameter in order to characterise a proto-object is to determine its distance to the vision system. Nowadays, not only stereo pairs of cameras but also cheaper devices like Microsoft Kinect or ASUS Xtion provide accurate depth information of the captured image (fig. 4.4).



Figure 4.4: Some commercial RGBD sensors. From left to right: stereo pair of cameras (Videre Design), Xtion (ASUS), Time-of-Flight Camera (MESA Imaging) and Kinect (Microsoft)

When using a sensor able to directly provide depth information (e.g. a RGBD camera or similar), the proximity, ($PROX$), of a proto-object, $\mathcal{P}_i$, is directly obtained as the inverse of the depth value provided by the sensor:

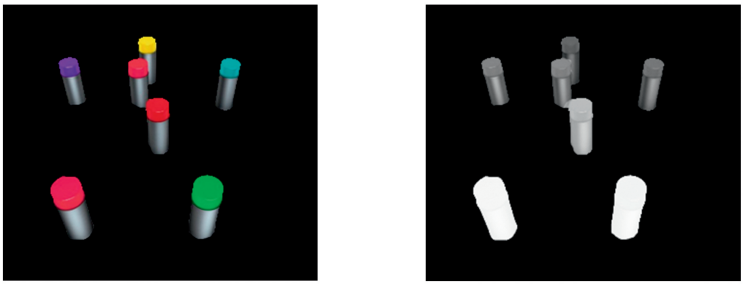
$$PROX_i = \frac{1}{depth_i} \tag{4.4}$$

In the case of using a stereo pair of cameras as depth sensor, the proximity can be obtained directly from disparity information. Note that, in order to maintain the computed values in the range $[0 \dots 255]$, the obtained

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGUNA

proximity does not correspond to a physical distance expressed in meters. It represents a normalised and relative measure with respect to the sensor.

An example of proximity conspicuity map is shown in fig. 4.5. Elements that are closer to the sensor are more relevant than those being far away.



(a) (b)

Figure 4.5: Example of proximity feature computation. (a) original image; (b) proximity conspicuity map. Relevant proto-objects are brighter.

4.3.3 Roundness

Roundness measurement reflects how similar to a circle a proto-object is. This feature provides information about convexity, closure and dispersion. Roundness is obtained employing a traditional technique based on image moments. Concretely, three different central moments are used:

$$\mu^i_{1,1} = \sum (x - \overline{x})(y - \overline{y}) \quad \forall (x, y) \in \mathcal{P}_i \tag{4.5}$$

$$\mu^i_{2,0} = \sum (x - \overline{x})^2 \quad \forall (x, y) \in \mathcal{P}_i \tag{4.6}$$

$$\mu^i_{0,2} = \sum (y - \overline{y})^2 \quad \forall (x, y) \in \mathcal{P}_i \tag{4.7}$$

being $(\overline{x}, \overline{y})$ the center of the proto-object $\mathcal{P}_i$.

From the combination of the equations above, it is possible to measure the difference between a region and a perfect circle. This measure is known as eccentricity and can be calculated as follows:

$$ecc_i = \frac{(\mu^i_{2,0} - \mu^i_{0,2})^2 + (2\mu^i_{1,1})^2}{(\mu^i_{2,0} + \mu^i_{0,2})^2} \tag{4.8}$$

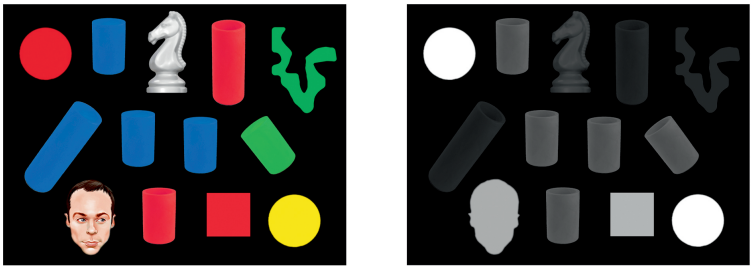
Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

being the result in the range $[0 \dots 1]$.

Finally, the roundness, $(ROUND_i)$, for a proto-object, $\mathcal{P}_i$, is obtained from the definition of eccentricity as:

$$ROUND_i = 1 - ecc_i \tag{4.9}$$

An example of a roundness conspicuity map can be seen in fig. 4.6. As it was expected, the two circles in the image obtain the largest saliency value. Something similar occurs to the red square. Due to squares (as all regular polygons) are easily circumscribed by a circle, they also present a large roundness value but less than a perfect circle. Faces are often quite round too. On the other hand, oblong forms, as cylinders or the chess piece, show low roundness values. The sparse and shapeless green element is, as expected, the least relevant in terms of roundness.



(a) (b)

Figure 4.6: Example of roundness feature computation. (a) original image; (b) roundness conspicuity map. Relevant proto-objects are brighter.

4.3.4 Orientation

As it was explained at the beginning of the chapter, orientation is one of the undoubted sources of attention guidance. The orientation of a region in a image can also be obtained from central moments computed in (4.5), (4.6) and (4.7):

$$\varphi_i = \frac{1}{2} \arctan\left(\frac{2\mu^i_{1,1}}{\mu^i_{2,0} - \mu^i_{0,2}}\right) \tag{4.10}$$

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGUNA

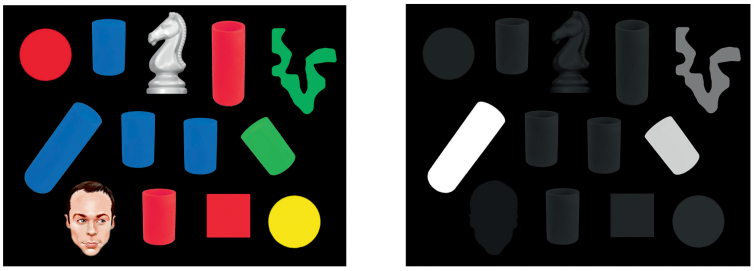
But the orientation of a proto-object, by itself, does not provide any useful information about its relevance. Only when comparing its orientation with the orientation of the rest of proto-objects in the image, a feasible measure of relevance is obtained. Thus, in fact, it is more interesting to compute saliency in terms of contrast with the surrounding elements. The orientation contrast, ($OriCON$), of a proto-object, $\mathcal{P}_i$, is obtained as:

$$OriCON_i = \sum_{j \in N_i} |\varphi_i - \varphi_j| \tag{4.11}$$

where N_i is the set of proto-objects that are neighbours of $\mathcal{P}_i$.

Although pure orientation information is not employed to calculate relevance, it is saved as a descriptor of the proto-object for further use (for example, to compute symmetry).

Fig. 4.7 presents an example about orientation contrast feature computation. Most of the elements are vertically oriented except the green and the blue cylinders diagonally placed. Because they are surrounded only by vertical forms, they are the most relevant. When computing the orientation of rounded forms, as circles, squares or regular polygons, where there is no dominant direction, the solution is equal to 0°.



(a) (b)

Figure 4.7: Example of orientation contrast feature computation. (a) original image; (b) orientation contrast conspicuity map. Relevant proto-objects are brighter.

Publicaciones y Divulgación Científica
UMA UNIVERSIDAD DE MÁLAGA

4.3.5 Symmetry

To compute the symmetry of a proto-object, an approach similar to Aziz and Mertsching [2008a] is followed. They propose a method to obtain symmetry using a scanning function $\psi(L, P_s)$ that counts the symmetric points around a point P_s along a line L. This procedure is repeated employing different lines of reference. For each line, the measure of symmetry is computed as:

$$S^{\theta} = \sum_{s=1}^{l} \frac{\psi(L, P_s)}{\alpha(R_i)} \tag{4.12}$$

where l and θ are the length and the angle of the line of reference and $\alpha(R_i)$ is the area of the region in order to normalise the result between 0 and 1.

Only an approximation of symmetry is needed in terms of attention systems. Thus, only 4 different angles for symmetry axes are considered: 0°, 45°, 90° and 135° respect to the orientation, φ_i, of the image (obtained in (4.10)) as it is shown in fig. 4.8.

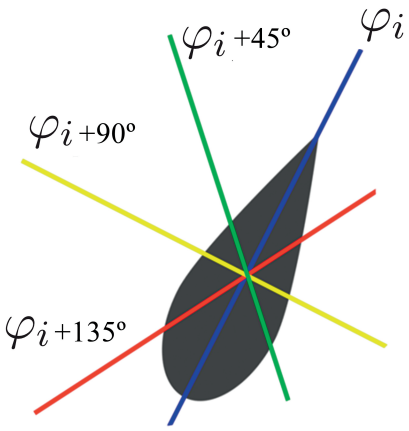


Figure 4.8: Symmetry is computed around 4 axes: from the major axis (orientation φ_i), three increments of 45° are considered.

In Aziz and Mertsching [2008a], the total measure of symmetry is computed as an average of the symmetry values in the different lines of reference. Nevertheless, such strategy can define a region with only one axis of symmetry as asymmetric, because non-symmetric axes cancel out the

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGA

contribution of the symmetric one. This is the case of the region in fig. 4.8. Around axis φ_i, the computed value of symmetry is high. However, a low value is obtained around the other axes. Applying an average, total symmetry remains small whereas the region is perfectly symmetric in the major axis.

As relevance is given to symmetry independently of the axis of symmetry, the maximum symmetry, $(SYMM)$, for a proto-object, $\mathcal{P}_i$, is computed as:

$$SYMM = max_\theta(S^\theta) \tag{4.13}$$

A symmetry conspicuity map is shown in fig. 4.9. As it can be noted, symmetric elements are relevant regardless of their orientation. For instance, both vertically and diagonally oriented cylinders are marked as equally relevant in terms of symmetry. On the other hand, elements with no axis of symmetry (the chess knight and the green shapeless form) are considered as no relevant at all.

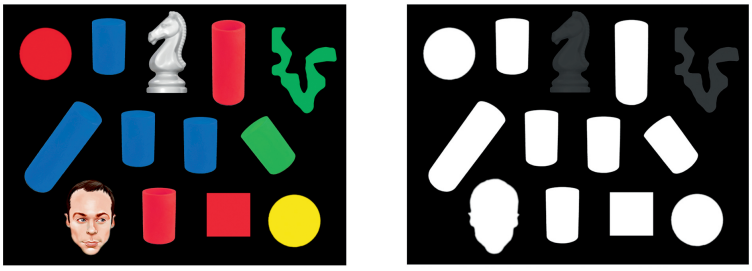


Figure 4.9: Example of symmetry feature computation. (a) original image; (b) symmetry conspicuity map. Relevant proto-objects are brighter.

4.3.6 Similarity to basic colours

Sometimes, an application requires objects of a specific colour to be more relevant than others regardless their colour contrast with the surrounding. For example, a fireman robot may look for red extinguishers or a refuse collector robot may look for specific containers identified by colours.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

For this reason, a set of dominant colour features have been added to the attention system. Concretely, OBVIAS model includes saliency computation for the four basic colours identified by Treisman as *canonical colour stimuli* [Treisman and Gormican, 1988]: red, blue, green and yellow.

To compute the similarity to basic colours, an HSV colour distance is employed. The hue and saturation values are compared with a reference. If the distance is less than a threshold Θ, the corresponding colour feature obtains a value of 255. Otherwise, the value is 0. The saturation value is used to avoid shadows being marked as colours. The red, (RED), blue, (BLU), green, (GRN), and yellow, (YLW), corresponding colours for a proto-object, $\mathcal{P}_i$, are computed as:

$$RED_i = \begin{cases} 255 & \text{if } d(< C_i >, < C_{red} >) \leq \Theta_{red} \\ 0 & \text{otherwise} \end{cases} \tag{4.14}$$

$$BLU_i = \begin{cases} 255 & \text{if } d(< C_i >, < C_{blue} >) \leq \Theta_{blue} \\ 0 & \text{otherwise} \end{cases} \tag{4.15}$$

$$GRN_i = \begin{cases} 255 & \text{if } d(< C_i >, < C_{green} >) \leq \Theta_{green} \\ 0 & \text{otherwise} \end{cases} \tag{4.16}$$

$$YLW_i = \begin{cases} 255 & \text{if } d(< C_i >, < C_{yellow} >) \leq \Theta_{yellow} \\ 0 & \text{otherwise} \end{cases} \tag{4.17}$$

By definition, a proto-object can only belong to one (or none) of the basic colours.

Examples of conspicuity maps associated with the four basic colours can be found in fig. 4.10. Note the special cases of the gray chess knight and the face. Because they have a colour that can not be included in any of the four categories, these elements are always marked as non relevant.

4.3.7 Attending humans: skin colour

As it was mentioned before, OBVIAS model is an attention system oriented to social robots. Hence, interaction with humans becomes an essential behaviour. One fast and simple way to detect humans and, therefore, deploy attention on them, consists in detecting their characteristic skin colour.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LA LAGUNA

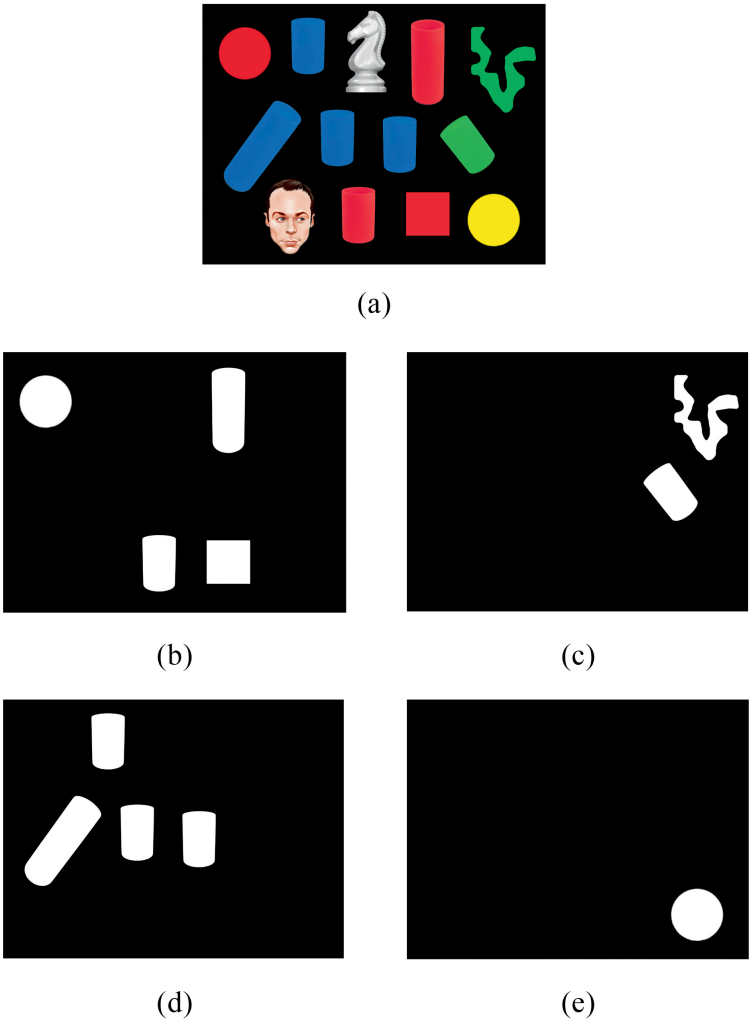


Figure 4.10: Example of similarity to basic colours feature computation. (a) original image; (b) red conspicuity map; (c) green conspicuity map; (d) blue conspicuity map; (e) yellow conspicuity map. Relevant proto-objects are brighter.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

In combination with other features (for example, roundness or symmetry), it is possible to guide attention to faces using only simple features as it is recommended in Wolfe and Horowitz [2004] and Greene and Wolfe [2011].

The computation is based on the skin colour chrominance model proposed by Terrillon and Akamatsu [Terrillon and Akamatsu, 1999]. First, the image is transformed into the TSL colour space. Then, the Mahalanobis distance between the colour of the proto-object and the mean vector of the skin chrominance model is computed. If this distance is less than a threshold Θ_{skin}, the skin colour feature is marked with a value of 255. Otherwise, it is set to 0.

$$SKN_i = \begin{cases} 255 & \text{if } d_M\left(< C_i^{TSL} >, < C_{yellow}^{TSL} >\right) \leq \Theta_{skin} \\ 0 & \text{otherwise} \end{cases} \tag{4.18}$$

Fig. 4.11 shows an example of skin colour detection. It can be observed that the face (excluding the hair and the eyes) is marked as relevant due to its characteristic colour.

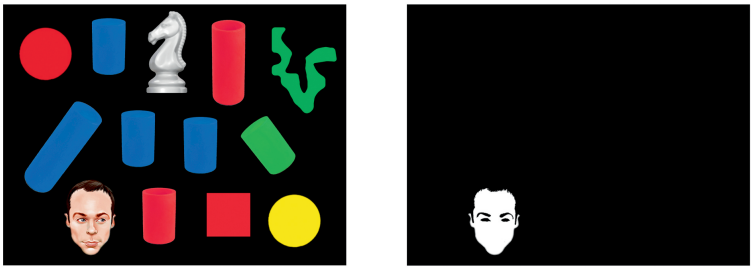


Figure 4.11: Example of similarity to skin colour feature computation. (a) original image; (b) skin colour conspicuity map. Relevant proto-objects are brighter.

4.3.8 Saliency Computation

As most of the artificial attention systems following Treisman's *Feature Integration Theory* [Treisman and Gelade, 1980], the total saliency of an

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGUNA

element in an image is the result of a linear combination of its low-level features.

In the proposed attention system, the final saliency value, SAL_i, for each proto-object, $\mathcal{P}_i$, is obtained as a weighted sum of all the previously described features, normalized in the range $[0 \dots 255]$:

$$SAL_i = \vec{\lambda} \cdot \vec{f} \tag{4.19}$$

where $\vec{\lambda}$ is the *attentional set* stored in the Perception Modulation Memory (PMM), verifying $\sum_i \lambda_i = 1$, and $\vec{f}$ is the feature vector, obtained as it was detailed in the sections above.

Expanding (4.19) in terms of (4.2), (4.3), (4.4), (4.9), (4.11), (4.13), (4.14), (4.15), (4.16), (4.17) and (4.18), the final saliency value can be expressed as follows:

$$\begin{aligned} SAL_i \quad &= \lambda_1 \cdot ColCON + \lambda_2 \cdot IntCON + \lambda_3 \cdot PROX + \lambda_4 \cdot ROUND \\ &+\lambda_5 \cdot OriCON + \lambda_6 \cdot SYMM + \lambda_7 \cdot RED + \lambda_8 \cdot BLU \\ &+\lambda_9 \cdot GRN + \lambda_{10} \cdot YLW + \lambda_{11} \cdot SKN \end{aligned} \tag{4.20}$$

Eq. (4.20) reflects the fact that the more different to other proto-objects in the image, the more salient the proto-object will be. Besides, both bottom-up and top-down contributions to attention are easily noticed. While the set of low-level features establishes the task-independent relevance of the proto-objects, the attentional set that weights the summation can bias the influence of each feature depending on the current high-level behaviour. At the end of the process, the most relevant proto-objects are saved in the Working Memory (WM). This top-down selection and management of WM are deeply described in section 4.5.

Finally, an example about whole saliency computation is shown in fig 4.12. Because saliency values are in the range $[0 \dots 255]$, a saliency map can be interpreted as a gray-scale image where brighter values correspond to larger values of relevance. In this case, all the weights are equal ($\lambda_i = \frac{1}{11} \forall i$), so no feature is advantaged in terms of attention guidance. In this case, the result is that the red ball is selected as the most salient proto-object in the scene.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

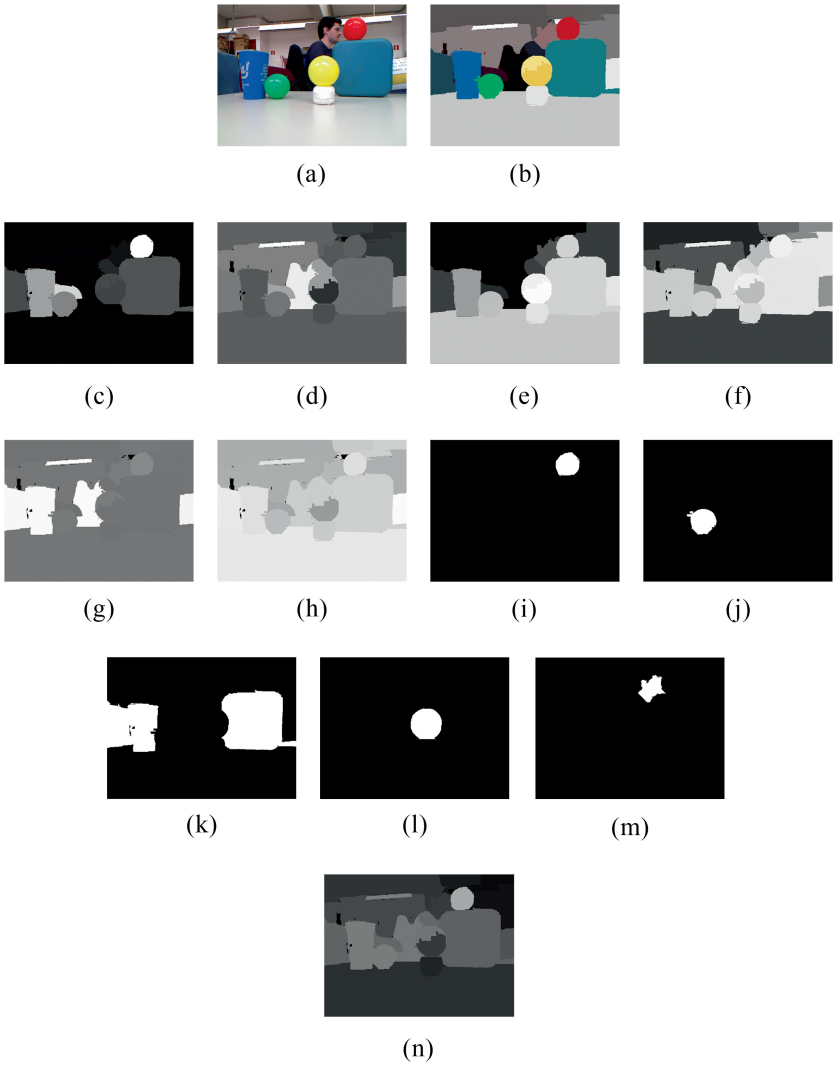


Figure 4.12: Features maps. (a) original image; (b) perceptual segmentation (extraction of proto-objects); (c) colour contrast; (d) intensity contrast; (e) proximity; (f) roundness; (g) orientation contrast; (h) symmetry; (i) dominant colour: red; (j) dominant colour: green; (k) dominant colour: blue; (l) dominant colour: yellow; (m) skin colour; (n) global saliency map.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

4.4 Inhibition of Return

Psychophysics studies about human visual attention have established that a local inhibition is activated in the saliency map when a region is already attended. This mechanism avoids directing focus of attention to a region immediately visited and it is normally called *inhibition of return* [Posner et al., 1985]. Posner's experiments demonstrate that, once an element in the scene is no longer task-dependent and attention has had time to disengage from it, such element is avoided in later fixations until certain time [Klein, 2000].

In the context of computational models of visual attention, an Inhibition of Return (IOR) mechanism becomes mandatory not only to achieve a more plausible and bio-related model, but also to avoid the system getting stuck in the most relevant element, ignoring the rest of the scene (as biological systems does). Therefore, the use of Inhibition of Return is due to both cognitive and pragmatic motivations. In fact, the latter is just a consequence of the former.

IOR is usually implemented using a 2D inhibition map which contains suppression factors for one or more recently used foci of attention. This approach is valid to manage static scenarios but it is not able to handle dynamic environments, where inhibited objects or the vision itself are in motion, or when minor illumination differences between consecutive frames cause shape changes in the objects. Hence, it is necessary to establish a correspondence between regions among consecutive frames.

In some models, such as Backer et al. [2001], this problem is solved relating the inhibition to features of activity clusters. However, the scope of dynamic inhibition becomes very limited because it is not related to objects. Thus, an object-based inhibition of return applying image tracking is employed instead. When the vision system moves, the proto-objects stored in the Working Memory are kept tracked. In the next frame, a new set of proto-objects is obtained from the image and the positions of the previously stored ones are updated. Then, from the new set of proto-objects, those occupying the same region as the already attended ones are suppressed. Discarded proto-objects are not taken into account in order to

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

fill up the WM. Thereby, it is highly unlikely to re-attend or duplicate an already existing proto-object in the WM. A visual example of inhibition of return is depicted in fig. 4.13. Using an image tracker is possible to identify the already attended proto-objects in the new frame (marked with orange bounding-boxes in the figure) and ignore them. The competition between new and existing proto-objects in the WM is explained in section 4.5.

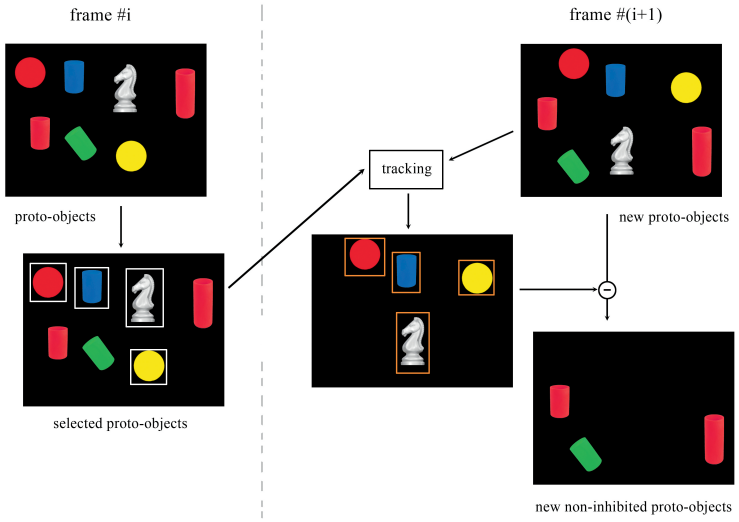


Figure 4.13: Inhibition of Return over two consecutive frames. Selected proto-objects in the first frame are marked with white bounding-boxes. In the second frame, the new positions of saved proto-objects are obtained using an image tracker (orange bounding-boxes). The proto-objects from the first frame are suppressed in the second frame as they already are in the WM

A tracker based on Dorin Comaniciu's *mean-shift* approach [Comaniciu et al., 2003] is employed to achieve the inhibition of return. Mean-shift algorithm is a non-parametric density estimator that optimizes a smooth similarity function to find the direction of movement of a target. A mean-shift based tracker is specially interesting because of its simplicity, efficiency, effectiveness, adaptability and robustness. Moreover, its low computational cost allows to track several objects in a scene maintaining a reasonable frame rate (real-time tracking of multiple objects). In the implementation used in OBVIAS, the target model is represented by a 16-bin colour his-

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

togram masked with an isotropic kernel in the spatial domain. Specifically, the Epanechnikov kernel is employed. The basics steps of the algorithm are briefly describe below:

1. Initialize the target model (in normalized RGB colour space).
2. Define the size of the target window.
3. Compute the model of the target on the current centre position.
4. Update the weights for each pixel in the target window.
5. Compute the target displacement and add it to the current position.
6. Repeats steps 3 to 5 over the current frame until convergence (zero displacement). If the Bhattacharyya distance to all targets is less than a threshold, the target is considered lost.
7. Load the next frame and repeat from step 3.

Finally, to speed up the process, the maximum number of iterations in step 6 is limited to 20 as is recommended in Comaniciu et al. [2003].

To sum up, the use of a mean-shift tracker to implement the inhibition of return provides several benefits. On the one hand, the system can deal with dynamic environments and attended elements are identified throughout the time regardless of their relative movement. On the other hand, the mean-shift approach is able to track non-homogeneous colour regions, so it is possible to track proto-objects with complex textures. Finally, the low computational cost allows to keep several proto-objects tracked at the same time.

4.5 Top-down selection and management of the most relevant proto-objects

Until now, it has been explained how OBVIAS system is able to extract proto-objects from a image, compute a set of low-level features related to saliency guidance, solve the inhibition of return problem and select the most relevant proto-objects depending on the task to accomplish. Nevertheless,

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

there still exist open questions concerning the number of proto-objects simultaneously attended and the top-down connection with high-level reasoning modules. These questions are mainly addressed in the Working Memory (WM) and the Perception Modulation Memory (PMM).

The integration between bottom-up and top-down contributions is an important issue in visual attention systems. As it was explained in section 3.1.2 on page 33, there are two main strategies to address this problem: *feature map fusion approaches* (e.g. Navalpakkam and Itti [2006] or Kouchaki and Nasrabadi [2012]) and *template-based approaches* (for instance, the models proposed by Aziz and Mertsching [2008b], Wischnewski et al. [2010] or Tünnermann et al. [2013]). While the first ones compute new feature maps based on high-level learned knowledge, template-based models work with abstract templates of low-level features.

In an autonomous system, top-down information is usually provided by a deliberative layer (typically, a planner). The planning system defines what to do. Therefore, it must suggest what type of information is relevant or not to the attention module. Thus, it is interesting that the latter can deal with abstract predicates. Consequently, a template-based approach, which is more suitable to manage that kind of information, is employed in OBVIAS.

As it was aforementioned, the ongoing task in the system can guide attention modifying the weight vector of saliency equation (see eq. (4.19) on page 69). Every task applies a different top-down template, i.e., a different attentional set. An attentional set is defined as a collection of different perception parameters able to highlight proto-objects with specific features [Corbetta and Shulman, 2002]. For example, a car-driving task looking for the nearest *stop signal* is translated to look for red and rounded proto-objects in the image which are close to the camera. Taking equation (4.20) as reference, the mentioned task implies large values for λ_3 (weighting proximity), λ_4 (weighting roundness) and λ_7 (weighting similarity to red colour). The remainder perception parameters obtain noticeably smaller values. This situation is depicted in fig. 4.14. When the task changes, a variation of perception parameters pops out different proto-objects in the scene. For example, the stop signal becomes the most relevant element in fig. 4.14.c when using a concrete attentional set. To establish a new

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

attentional set, only a new writing in the PMM is needed. It is possible to modify the content of the PMM every perception cycle.

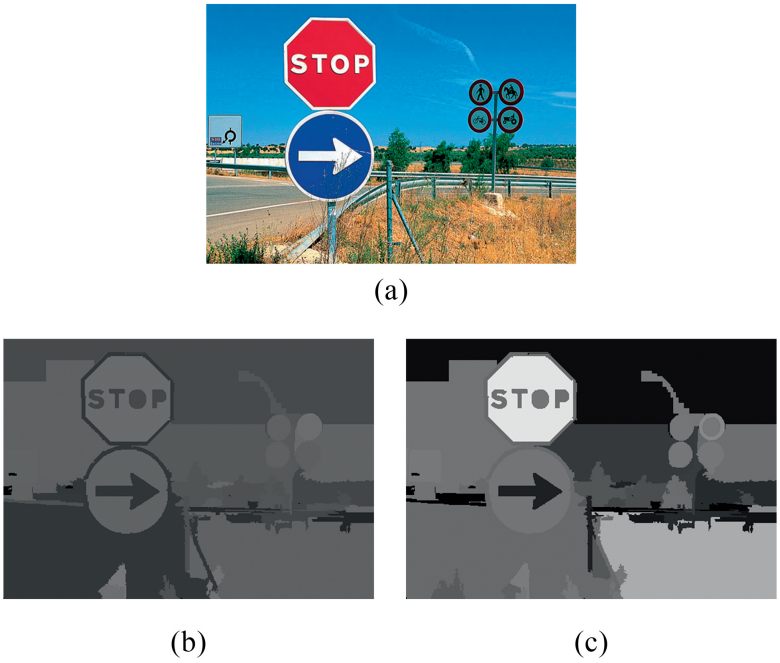


Figure 4.14: Top-down selection in OBVIAS model. (a) original image; (b) saliency map obtained when there is no guidance: all the perception parameters have the same value; (c) saliency map when attention is guided to near, red and round objects in order to look for *stop signals*. The stop signal becomes more relevant (brighter in saliency maps) when a correct guidance of attention is applied.

Guiding attention using attentional sets shows an important limitation: it is not possible to directly cope with "negative examples" or exclusion-related tasks (e.g. "find square objects that are not blue"). Nonetheless, the question can be indirectly addressed by assigning very small values to the corresponding parameters so that blue and round elements should not be relevant at all. Thereby, the system penalises round and blue object but it is not possible to explicitly highlight square and not blue elements. Consequently, this implies a partial solution that could not be enough for certain specific combinations of features.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

Psychophysics studies also refer to how many elements can be attended at the same time. Bundesen establishes in his *Theory of Visual Attention* [Bundesen and Habekost, 2008; Bundesen et al., 2011] that there exists a short-term memory where recently attended elements are stored. This memory has a fixed capacity usually reduced up to 3 or 5 elements. The function of the *Working Memory* in visual attention can be defined as *"the ability to maintain and manipulate information online in the absence of incoming sensory stimulation"* [Corbetta and Shulman, 2002]. Therefore, the WM establishes the maximum number of attended elements that can be maintained at once. In the case of OBVIAS system, the WM has a fixed capacity up to five elements. This value has been empirically obtained to maintain a reasonable computational time of the WM itself and the tracker following several regions at the same time.

Every proto-object in the WM is characterised by a set of descriptors:

- Saliency value.
- Depth.
- Orientation.
- Characteristic colour (in HSV space).
- Area.
- Position in image(updated by the image tracker presented in the previous section).
- The eleven computed low-level features values.
- A *time-to-live* value which establishes the maximum time that a proto-object can be stored in memory.
- A copy of the region of interest (ROI) occupied by the proto-object in the image.

ROI inclusion allows another modules in a complex system to use a more efficient representation about the real world: they received only the visual information they need for their tasks. As a consequence, less computational resources are required what is, indeed, one of the main objectives of an artificial attention system.

Regarding WM management, a new proto-object gets into the WM if and only if it is more salient than the currently stored elements. If the memory is full, the least salient element is dropped out. A proto-object

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

can also be removed from WM if it is lost by the image tracker. To keep a good performance, lost items are checked and dropped out before trying to add new elements. Total saliency of the elements in the WM is recomputed in every perception cycle in terms of the new attentional set. Thereby, the saliency of the stored proto-objects is always kept up-to-date.

The saliency of a proto-object also depends on the time-to-live parameter as the results of psychological studies about the dynamic of the IOR demonstrates (fig. 4.15) [Klein, 2000]. These studies conclude that the longer an element is kept in WM, the lower its saliency is. In other words, after some time, attention is no longer deployed on the same element of the scene and an inhibitory process begins. Thus, a proto-object already stored in WM has, at the beginning of its life, an increment in saliency due to a high time-to-live. On the contrary, older proto-objects receive a decrement of saliency inversely proportional to their time-to-live. In order to implement this behaviour in the WM, an exponential function is applied to time-to-live (see fig. 4.16). Then, the component of saliency depending on the time-to-live, Sal_{TTL}, of a stored proto-object in the WM can be approximated by the equation

$$Sal_{TTL} = \frac{2}{1 + e^{0.01(\epsilon - TTL)}} - 1 \tag{4.21}$$

where ϵ is the inhibitory threshold and TTL is the time-to-live. ϵ has been empirically set to 3-4 perception cycles. In summary, the WM operates as a cache memory. While a proto-object is relevant for the ongoing task, it is kept in memory. The content of the WM is only fully swapped when a new task comes out and no proto-object in the WM is suitable to accomplish the new behaviour.

An example of the final result of a perception cycle is shown in fig. 4.17. In the example, a search for round, blue and yellows objects is detailed. The five most salient proto-objects are shown in fig. 4.17.d. As it was expected, the object complying with most of the required features (the yellow ball) is selected as the most relevant. Although the blue case also verifies 2 features, the roundness of a square is a bit lower than the roundness of a ball. Thus, the blue case is selected as the second most relevant element. The remainder of the elements to be stored in the WM corresponds to blue or round objects.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

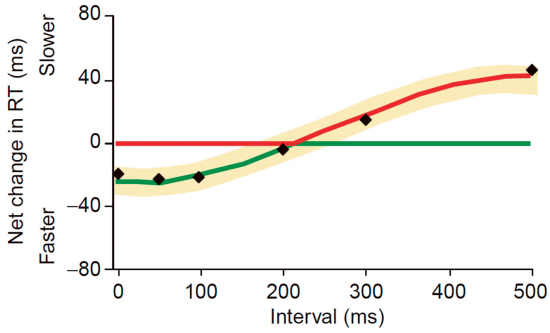


Figure 4.15: Results from Posner and Cohen's study. After some time attending an element in an image (200 ms), an inhibitory process begins and the saliency of the visual entity falls. Reference: Posner et al. [1985] via Klein [2000]

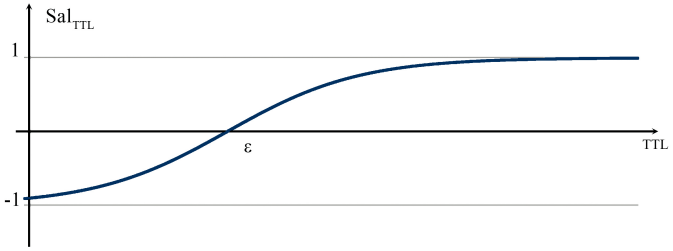


Figure 4.16: Saliency in terms of time-to-live. A new proto-object (high time-to-live), receives an increment of saliency. When time-to-live is less than ϵ time, the saliency of the proto-objects is penalised. Note that time-to-live is *decremented* throughout the execution time.

In conclusion, both Working Memory (WM) and Perception Modulation Memory (PMM) are the interface between early attention stages and the rest of the system, including the deliberative level. On the one hand, the WM provides the most relevant proto-objects to other applications, including a set of descriptors and the corresponding ROI. On the other hand, other modules are able to guide the attention system to specific and useful objects in the scene, just swapping the attentional set stored in the PMM. Assuming this interface as starting point, in the next chapter a whole cognitive architecture is developed in order to close the perception-reasoning-action loop in a bidirectional way.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

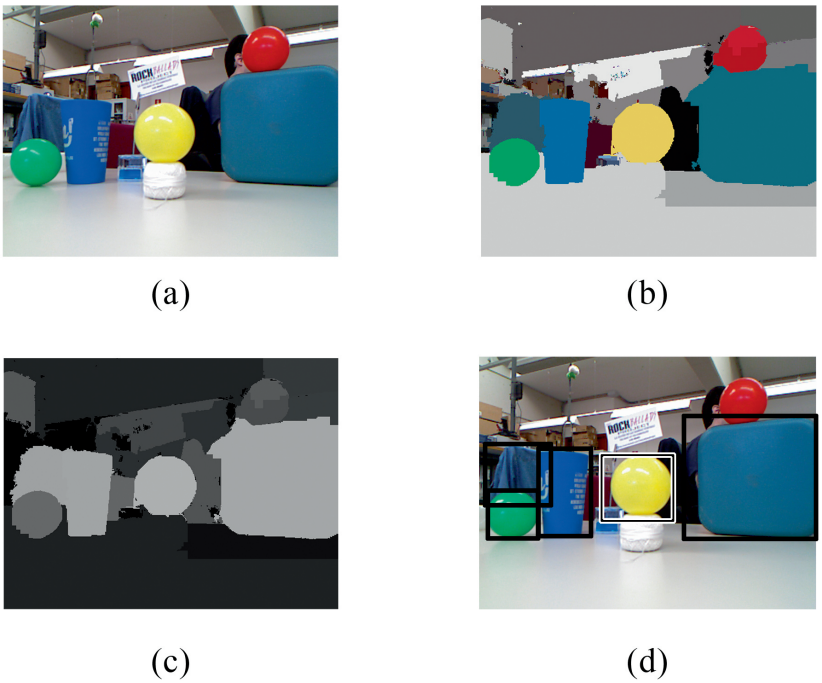


Figure 4.17: Saliency computation looking for round, blue and yellow proto-objects. (a) original image; (b) perceptual segmentation (extraction of proto-objects); (c) saliency map; (e) proto-objects selected to be stored in the WM (the most salient one is marked with a black and white bounding-box).

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

UNIVERSIDAD DE MÁLAGA
Publicaciones y Divulgación Científica

5 From action to perception and vice versa

"Thoughts are but dreams till their effects be tried"

William Shakespeare

As it has been pointed out in the previous chapters, most of the artificial attention systems lack connection with abstract reasoning layers. In their recent review about visual attention systems, Borji and Itti [2013] propose the development of models able to handle time varying task demands in interactive, complex and dynamic environments as a promising direction for future research. From the cognitive architectures development point of view, Langley et al. [2009] claim for expanding cognitive frameworks to manage agent's resources to selectively focus its perceptual attention and the tasks it can accomplish.

Although several solutions have been proposed from different points of view, they normally fall in an incomplete and unidirectional characterisation of the problem (fig. 5.1.a). These models reduce the problem to a visual search guided by the deliberative layer (recall section 3.2 on page 40). However, they do not consider that new perceived objects can modify the current set of tasks that can be executed. The architecture introduced in this thesis goes beyond, considering that attention-based perception, abstract reasoning and action present a close and symbiotic relationship: they mutually modulate each other simultaneously (fig. 5.1.b).

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

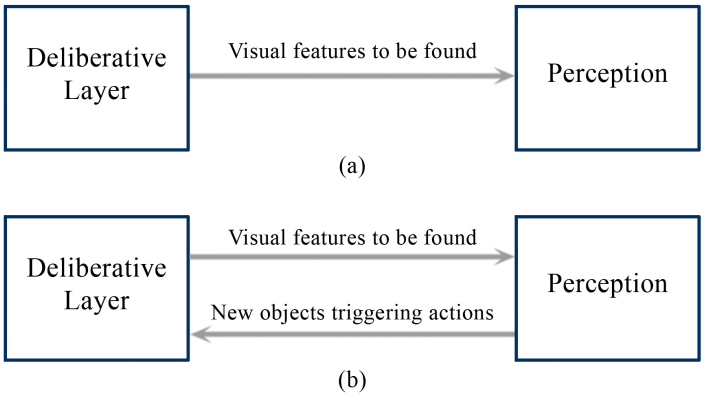


Figure 5.1: Perception-reasoning-action loop closing problem. (a) Classical and unidirectional assumption. (b) Bidirectional and complete characterisation.

This chapter copes with a **novel need-based cognitive architecture**, driven by the attention model presented in chapter 4 and a Planning Framework able to handle partial information. The chapter firstly introduces the cognitive architecture and its foundations. Then, the loop closing process is detailed, describing the planning framework involved and the procedures allowing an effective connection between attention and the deliberative level.

In terms of attention systems, the planner implements the top-down part of attention. Therefore, the planner influences the saliency computation of the elements in the scene depending on its requirements. Besides, the attention model only returns information from relevant areas and only such a partial information can be used to generate a plan. Hence, the perceived elements determine the actions to be achieved by the planner at each time.

5.1 A need-based model

When developing task-specific robotic agents, the classical solution usually entails the computation of quantitative models of the environment and the action to be accomplished. For example, designing a robotic arm for car

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGA

assembling commonly involves the application of Optimal Control Theory [Lewis et al., 2003]. The decision making procedure is based on perceived magnitudes (e.g. the distance to the piece or the size of the screw) and it is directly applied to the final effectors.

The aforementioned strategy is completely suitable for systems oriented to very specific behaviours. But a social and autonomous robot is a versatile agent usually placed in an unknown and changing environment. Social robots, in a similar manner to humans, can not be considered as simply deterministic input-output devices. On the contrary, they are likely to act as goal-oriented agents actively selecting their goals and looking for relevant information. As a consequence, a global quantitative model might not be needed. Instead, it is more interesting to employ a set of models, which are reliable for defined categories of work conditions, in conjunction with a qualitative framework describing and defining their relationships and coverages [Rasmussen, 1983].

The cognitive architecture presented in this thesis follows the previous guidelines. As it can be seen in fig. 5.2, the architecture is divided into two levels. The different tasks that can be performed at each moment are placed at the *Rule-based level*. Depending on the perceived elements, a task could be executed or not. This level is concerned with quantitative models of execution. Tasks are based on **needs**. In other words, the accomplishment of a task is closely linked to the presence in the scene of specific elements. The higher level, or *Knowledge-based level*, is related to the qualitative framework mentioned above. The aim of this abstract reasoning level is the selection of the tasks that will be active, by adding or removing them from the prior level, depending on context information. This level also establishes the priority of each task and it is informed about the state of achievement of the actions.

The proposed architecture is mainly inspired in the human performance model developed by Rasmussen [1983]. Rasmussen's work explains the different ways in which information is perceived at the different cognitive levels, distinguishing among skills, rules and knowledge. Although Rasmussen's approach was developed in the early 1980s, it is still used as basis in different disciplines, such as Social Robotics [Bischoff and Graefe, 2004], Human-Computer Interfaces [Rukšėnas et al., 2013] or human's behaviour

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

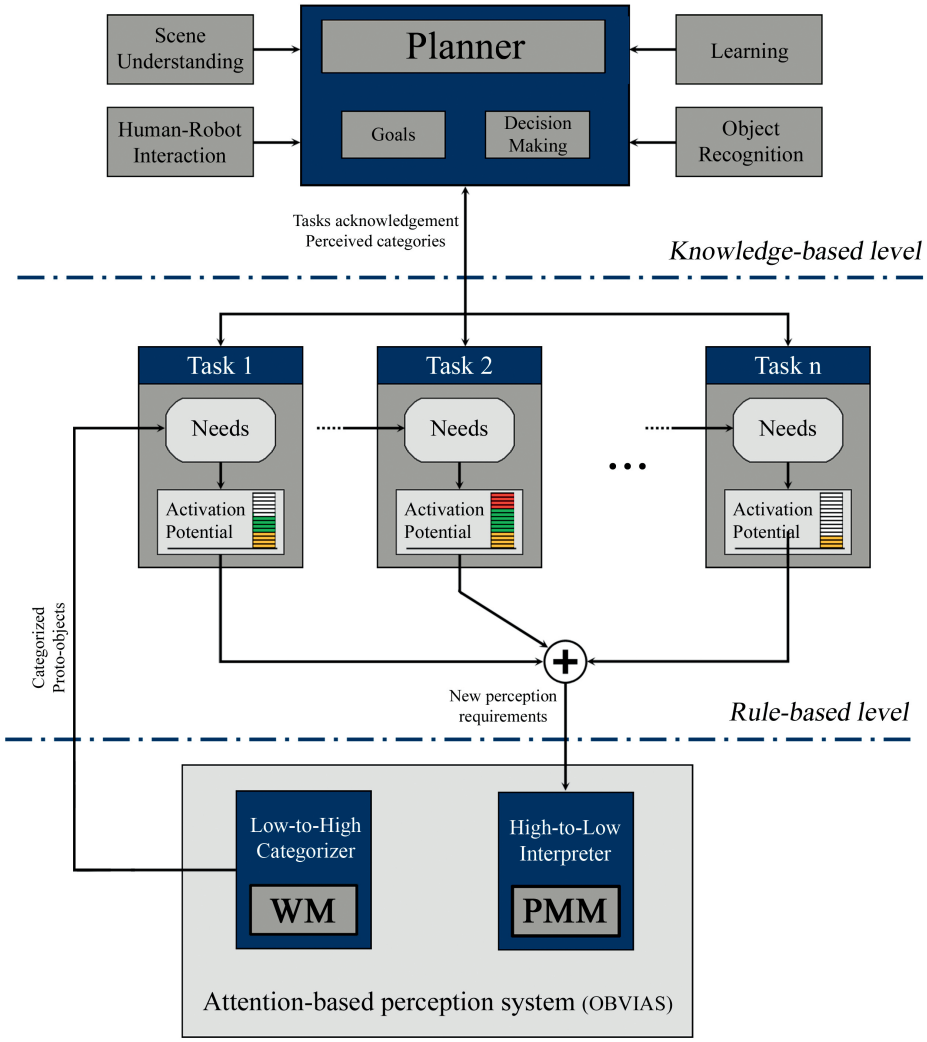


Figure 5.2: Need-based cognitive architecture. Tasks that can be accomplished at each time are placed at the *Rule-based level*. Depending on the perceived elements, some tasks can be activated. The activation potential associated with tasks measures their completeness, allowing a perception guidance based on the most dominant task. Modules placed at the *Knowledge-based level* obtain context information of the scene and deal with the selection and prioritisation of the tasks located at the Rule-based level.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

analysis in demanding environments [Lin et al., 2014]. The remainder of this section firstly introduces Rasmussen's biological model. Then, each level of the proposed architecture is deeply explained.

5.1.1 Basis: Rules and Knowledge in Human Performance Models

Due to the remarkable performance of natural cognitive systems, artificial models are often derived from biological solutions. In the same way that human attention provides several benefits when applying it to robotic agents, an artificial cognitive architecture can be inspired by natural models in order to obtain an enhanced information processing model.

While studying the design and evaluation of computer-based interface systems, Rasmussen [1983] developed and established an overall qualitative model of human performance. His model considers that the behaviour of humans is *teleological* by nature. Defined by Rosenblueth et al. [1943], a teleological behaviour is modified *during its course* by signals from the goal. In other words, the tasks involved in a complex behaviour can be modified depending on different factors such as experience in previous attempts or perception.

Rasmussen considers that human activity in a known environment is oriented towards the goal and controlled by a set of rules which has been successfully tested before. In unknown environments, where proven rules are not available, behaviour may be goal-controlled. Thus, different attempts are needed to reach the goal. This procedure corresponds with high level conscious planning in a classical sense. The efficiency of humans when dealing with complexity is due to their ability to generate *ad hoc* rules to control behaviour.

According to the diverse representations of behaviour constraints of a deterministic environment, Rasmussen distinguishes among three levels of performance: skill-, rule- and knowledge-based performance (fig. 5.3).

The *skill-based* behaviour represents sensory-motor performance during activities which take place without conscious control (e.g. handwriting).

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

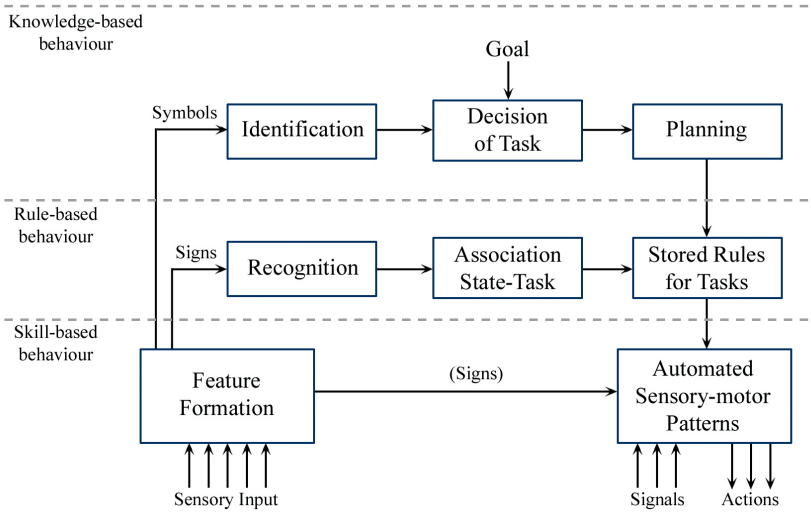


Figure 5.3: Simplified model of human information processing divided into three levels of performance [Rasmussen, 1983].

The body acts as a multivariable continuous control system, that synchronises movements. In general, human activities can be decomposed into a sequence of basic sensorimotor skills. For example, "grasp a bottle" can be divided into "locate object bottle + open hand + move hand towards object + close hand around object".

The composition of a familiar activity (e.g. the mentioned "grasp a bottle") selecting a sequence of subroutines from a repertoire of automated subroutines ("open hand", "locate object"...) is made at the **rule-based level**. At this level, the performance is goal-oriented and it is structured using rules. These rules, or procedures, can be derived from previous experience, communicated by other people (using a cookbook recipe, for instance) or instantaneously prepared by conscious problem solving and planning. The rule that better fits the environment constraints will control the activity, prevailing over the others. The information received in the rule-based level is considered as a *sign*. Thereby, it serves to activate or modify predetermined actions.

The boundary between skill-based and rule-based levels is quite subtle. It depends on the level of training and attention that a person requires

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGA

to perform a certain task. Whereas skill-based performance does not need focused attention, rule-based behaviour is generally based on explicit know-how and requires conscious focusing of attention.

When there are no applicable rules of control because of an unfamiliar environment, control moves to a higher level. At the **knowledge-based level**, performance is goal-controlled and explicitly formulated. In this case, a useful plan has to be derived based on an analysis of the environment and personal objectives. The effects of the plan can be physically tested by trial and error or they can be conceptually "simulated" by understanding the functional properties of the environment.

Following the same example as before, a person may has rules for "grasp a bottle" but not for "grasp a hot teapot". A *mental process* may lead people to analyse the possible threatening factors and actions (some parts of the teapot could be dangerous), looking for similar solutions in their knowledge ("grasp a bottle" may be similar to "grasp a hot teapot"). As a result, a new set of rules are obtained to solve the new problem (e.g. "grasp the teapot as a bottle but using only the handle to avoid injuries"). These new rules are added to the rule-based level.

5.1.2 Rule-based level

According to Rasmussen's work, the lower level of the cognitive architecture should consider a skill-based control of performance. However, in the model proposed in this thesis, that skill-based behaviour is not explicitly defined. When developing robotic systems, their skills are often addressed as isolated problems. For example, there exists a wide range of works facing problems such as grasping [Bicchi and Kumar, 2000], human-robot interaction [Bandera, 2010], navigation [Vázquez-Martín, 2009], etc. Conceptually, these abilities can be considered as a sequence of skilled routines that can be used by the rule-based level in order to accomplish certain tasks in terms of a conscious *executive program*. The work introduced here is intended to be a general solution for the perception-reasoning loop closing problem regardless the final application. Hence, the aforementioned skills are considered as black-boxes, i.e., as useful tools to execute the required tasks that do not have influence in the overall loop closing.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

There are two main novelties in the proposed approach. On the one hand, the execution of the tasks placed at the rule-based level is **need-based**. Each task has a set of *needs* in form of perceived object categories that must be covered to be plenty executed. For example, the task "look for a red card" should need elements belonging to the categories "red things" and "square things". On the other hand, the cognitive architecture has been specifically designed to take all the possible advantages from an attention-based perception system. Furthermore, the rule-based level is allowed to have influence over the attention system, guiding perception to task-relevant elements in the scene. This top-down consideration represents a considerable improvement of Rasmussen's model, permitting an effective symbiosis between perception and reasoning.

Multiple tasks can be placed at this level. For each one, a set of abstract object categories is specified. These **needs** must be fulfilled in order to execute the corresponding task. Each task has associated another important parameter: the **activation potential**. This factor measures the number of satisfied needs and, consequently, the probability of executing a task in the future. This information can be used later to guide the perception system. Thereby, those elements in the scene that can fully cover the needs of the task with the highest probability to be executed can be highlighted.

When designing a task, it is important to clearly define the categories involved in its activation. Some tasks may use several categories whereas others just need one (see some examples in fig. 5.4). The relationship between covered needs and the related activation potential strongly depends on the specific nature of the task. Although both of them are usually linear dependent (task 2 or task 3 in fig. 5.4), it is possible that some specific needs produce higher levels of activation. For example, a robotic arm employing the task "press the orange SOS button if a red light of danger is turned on". In this case, when the "red light" category is filled, it could be interesting a high activation potential to be able to rapidly satisfy the "orange button" condition. Thereby, the perception system is guided in order to execute that imperative task (e.g. task 1 in fig. 5.4). When several tasks get a similar activation potential, they are executed depending on their assigned priority.

Finally, the concept of perceived categories triggering different tasks

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LA LAGUNA

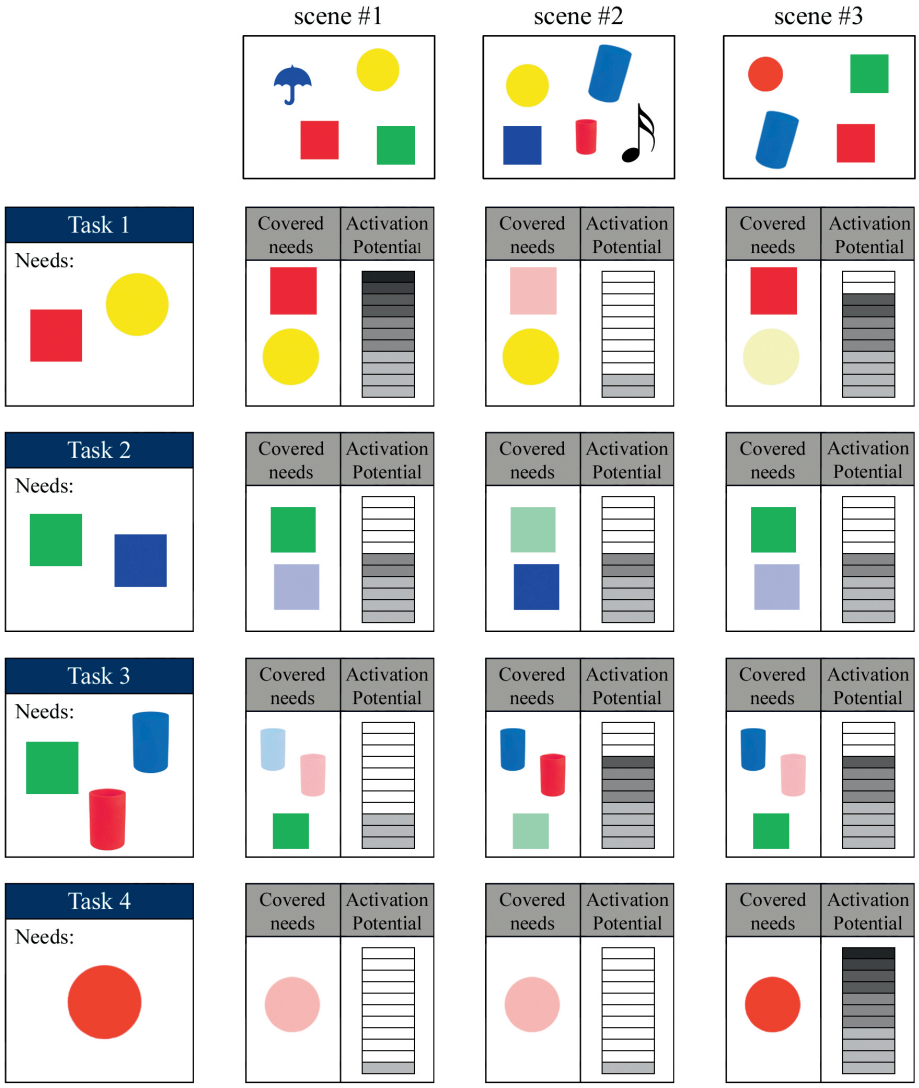


Figure 5.4: Example of tasks at the rule-based level. Depending on the observed items in the scene, different tasks may be executed. Each task has associated a set of needs that must be satisfied.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

also fits the biological theories about the influence of the observed environment in cognitive processes. As it was explained in section 2.3 (page 25), *Affordances* and *Reference Features* link perceived elements and tasks. The mere observation of an object affords an utility for the system if it belongs to a needed category. Thereby, an autonomous system is able to select the action to be executed among a variety of tasks by simple observation. Even a latent task could be instantaneously activated when observing a required item. Therefore, half of the loop closing problem is solved: the cognitive architecture is able to react to perception by activating those tasks that have fulfilled their needs.

5.1.3 Knowledge-based level

The **knowledge-based level** is responsible for the coordination and management of the tasks placed at the rule-based level. Maintaining lots of tasks at rule-based level to handle all the possible scenarios would be very inefficient. Moreover, certain behaviours would be likely to be wrongly accomplished due to ambiguity in action execution. For instance, two or more contradictory tasks could be ready for execution at the same time.

The overall planning, goals definition and decision making take place at the knowledge-based level. Therefore, a **General-purpose Automated Planner** is set as the core of this level. Unlike Rasmussen's approach, the planner receives information both from the sensory input and the tasks located at the rule-based level. The visual information is expressed in form of abstract predicates, derived from the perceived categories. This information can be used to elaborate new plans if needed. Since the visual data is supplied by an attention system, the Planning Framework must be able to deal with partial information. Thereby, a continuous adaptation of the plan to the perceived elements should be allowed.

Visual data can also be employed to figure out context information from the environment. The Planning Framework can be assisted by other high-level modules in order to determine the most suitable tasks to be applied in that context. This environment knowledge can be provided by different sources such as Machine Learning, Scene Understanding or effective Object Recognition. It can also be obtained from people through a Human-Robot

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGA

Interaction system. The Learning module could also allow the architecture to inductively infer new sets of rules to face unknown behavioural challenges. In other words, it would not be necessary to define all the possible actions a priori (*innate* rules and actions). Some models have been proposed to address this issue (e.g. Windridge and Kittler [2010]). However, the discussion about rules induction and learning exceeds the scope of this monograph.

Regarding the tasks placed at rule-based level, the Planning Framework can add, remove and prioritise them according to the context information and the perceived categories. In turn, the tasks inform the Planner about their state of execution (namely, their activation potential or an acknowledgement of successful execution). Depending on the specific environment and the current goals, the Planning Framework is able to load a completely new set of tasks at the rule-based level or change their priority in order to efficiently face a new scenario or role (fig. 5.5).

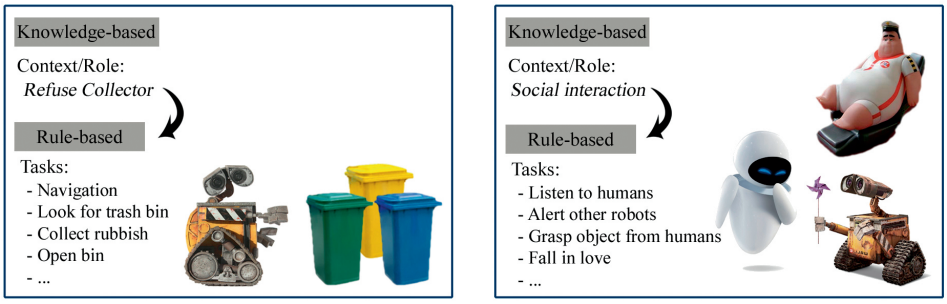


Figure 5.5: Depending on the specific scenario, the Knowledge-based level can modify the tasks present at rule-based level.

5.2 The Planning Framework

As it was aforementioned, a Planning Framework occupies the heart of the knowledge-based level, defining goals and managing the rule-based level. Since the attention model limits perception, the plan must be computed with only partial information about the initial state. Then, it can use *sensing actions* to increase the knowledge about the environment. Although this problem has been usually addressed using *Contingent Planning*

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

[Peot and Smith, 1992], **Oversubscription Planning** [Smith, 2004] provides a solution which fits better with the proposed cognitive architecture. Specifically, the solution proposed by García-Olaya et al. [2011] has been employed[1].

In general, the solution of a planning problem involves choosing a sequence of actions that transform the state of the world in order to likely satisfy a goal. The world is supposed to consist of atomic *facts* (states) and actions that make such states true or false. The procedure is accomplished step by step over the time. At a given step, the current state can be described using atomic formulas called *fluents* (e.g. `move(ob1,l1,l2,i)`, where `ob1` is an object, `l1` and `l2` are locations and `i` is a specific step).

A contingent planning problem is a tuple $P = \{F, A, I, G\}$, where F is the set of predicates and fluents, A is the set of actions, $I \subseteq F$ is the initial state and $G \subseteq F$ is the set of goals. Actions in contingent planning include conditional effects. Hence, the effects of actions depend on the real state to which the action is applied. Sensing actions discover the value of a certain previously unknown literal.

Contingent planning aims to find complete plans achieving all the goals by interspersing sensing actions whenever it is needed. This approach works well when the uncertainty about the initial state is small. However, it does not scale well in general. Recent approaches do not aim to produce complete plans. Instead, they only try to return at least a valid action to be executed [Shani and Brafman, 2011; Albore et al., 2009]. They mainly create a belief state by selecting a small subset of the possible initial states. Then, a plan according to this belief state is created. Plan is executed until an unexpected observation occurs or the preconditions of the next action to be applied do not hold. In this point, the belief state is updated and a new plan is generated.

In Oversubscription Planning (OSP), the planner is able to return a plan reaching just a subset of all the goals. Having a complete plan is not useful in non-deterministic domains, where external events can change

[1]The author would like to thank Dr. A. García-Olaya and Dr. F. Fernández their close collaboration in order to integrate the Oversubscription Planning Framework into the present work.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGA

the current world state. This is usually the case of autonomous and social robotics. OSP is a special case of planning with soft goals, where it is assumed that no plan achieving all soft goals exists. Usual causes making impossible to reach all the soft goals are limitation of resources or mutex goals.

In García-Olaya's approach [García-Olaya et al., 2011], they start only from the currently perceived state. Instead of a belief state, they use a single initial state containing the perception plus some static known facts. It is very likely that no plan achieving all goals exists given the limited available information. In this case, the lack of information makes some goals unachievable.

The oversubscription problem is solved following Algorithm 2. First, goals are selected and a new problem is constructed removing all the non-selected goals. Then, relaxed plans are constructed from the initial state to each goal and from every goal to each other. If a relaxed plan achieving a goal is found, the goal is added to the set of possibly achievable goals. Once goals are selected, the new problem is solved using any classical planner. If after a certain time no plan is found for the new problem, one of the goals is removed and a new problem is created. If a plan is found, it is executed and the environment is perceived again.

This approach has several advantages when compared against contingent planning. First, it results in a simpler model of actions because no conditional effects are needed. Second, it allows to overcome the problem of non-deterministic sensing actions by not reasoning about them; a sensing action is just applied after each planning cycle. Third, oversubscription planners tend to scale better than contingent ones. This is specially true in real environments with limited perception, where the number of objects for which the state is unknown is quite large and the results of the sensing actions are non predictable.

Finally, as the core of the knowledge-based level, the OSP Framework also deals with the allocation of suitable tasks at the rule-based level. In order to achieve that, the planner splits complex routines into a set of simple goals. The planner places all these tasks at the rule-based level and also handles what actions can be executed in parallel and what need the execution of prior tasks by exploring the related preconditions.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

Algorithm 2: Oversubscription problem solving algorithm

```
repeat
    Select goals;
    forall the goals do
        Look for relaxed plans achieving each goal;
        if solution is found then
            Add goal as achievable;
        else
            Remove the goal;
        end
    end
    Solve the new problem using a classical planner;
    if solution is found then
        Execute and perceive;
    else
        New perception cycle;
    end
until problem solved;
```

For example, the task "make a cup of tea" could be decomposed into the (simplified) goals *grasp the teapot*, *fill up the teapot with water*, *boil water*, *grasp a cup*, *put the tea bag in the cup* and *fill up the cup*. The tasks *boil water* and *put the tea bag in the cup*, for instance, could be achieved in parallel. On the contrary, the goal *fill up the cup* needs all the previous subtasks to be completed in order to be executed. This is possible since the tasks at the rule-based level inform to the OSP Framework about their state of execution. Thereby, all the parallel goals that can be accomplished are selected to elaborate the current plan (achievable goals in Algorithm 2).

5.3 Closing the loop

As it has been underlined throughout this chapter, the close relationship between perception and abstract reasoning becomes crucial in autonomous

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGA

robots placed in unknown and changeable environments. Despite the classical unidirectional assumption of such relationship, a thorough analysis reveals the **bidirectional** nature of the perception-reasoning-action connection. Perception and action modulates each other simultaneously. In fact, it has been demonstrated that certain perceptual features, such as shape or location, activate actions related to those features and, at the same time, those activated actions may prime the processing of stimuli defined by the perceptual features related to those actions [Collins et al., 2010; Fagioli et al., 2007].

The need-based cognitive architecture detailed in the previous sections extend the capabilities of OBVIAS model to effectively address the loop closing in a bidirectional way. The main actors in linking OBVIAS model and the cognitive architecture are the Working Memory (WM) and the Perception Modulation Memory (PMM), as it was mentioned in section 4.5 (page 73). In brief, depending on the number of satisfied needs due to the elements present in the WM, the influence of the task in the modulation of the perception parameters stored in the PMM will be greater (fully covered) or smaller (weakly satisfied).

The first step in order to obtain information that can be handle by a Planning Framework consists in translating the proto-objects stored in the WM into abstract predicates. Psychological studies support that high-level operations over visual entities are based on a *perceptual categorisation* of elements in the visual field [Bundesen, 1990]. Therefore, proto-objects in WM are classified into semantic categories according to their basic features.

The categorisation procedure could be tackle employing simple algorithms, based on the features provided as descriptors in the WM (see fig. 5.6), as well as more elaborated approaches based on face detection [Viola and Jones, 2004], Qualitative Image Descriptors [Falomir, 2013], an ontology [Johnston et al., 2008] or Fourier transforms [Oliva and Torralba, 2001] among others. However, the categorisation process must be fast and computationally efficient since it is part of the early-vision stage. That is, proto-objects do not need to be exactly recognised but classified into generic classes. An accurate object recognition is not necessary indeed. Nevertheless, slower but more precise approaches can be used at knowledge-based level to provide a enhanced scene context understanding. The category

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

assigned to each proto-object is also stored in the WM to make it available to other systems. Finally, a proto-object may be classified as not belonging to any currently known category.

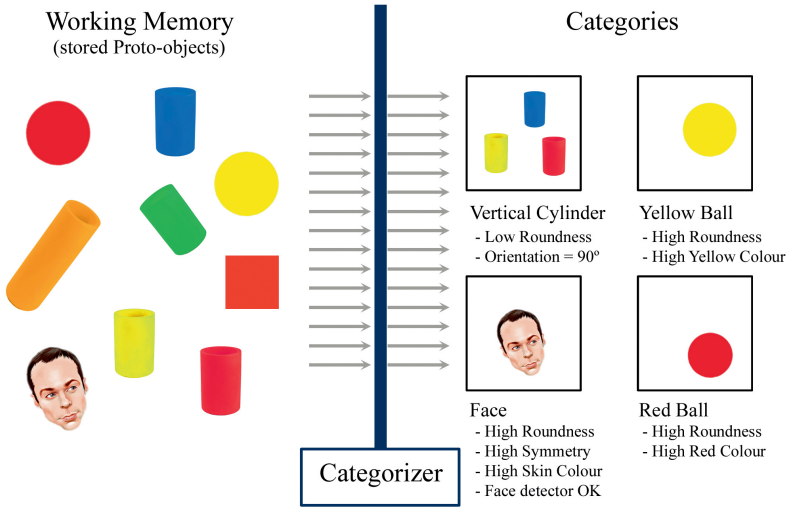


Figure 5.6: An example of basic categorisation process. A category is assigned to each proto-object depending on the presence of a specific set of features (one or several). A proto-object can also be considered as not belonging to any category (e.g. the red square or the orange cylinder).

Once the perceived categories are obtained, the tasks at rule-based level are activated depending on the number of covered needs as it was previously explained. Then, half of the problem is achieved: perceived categories are able to trigger latent tasks at rule-based level. Furthermore, the knowledge-based level can detect a modification in the scene context or the agent role. Thus, the Planning Framework is allowed to modify the tasks involved in the new activity.

The other half of the problem is concerned with the *reasoning to perception* part of the loop. In this case, the task with the higher activation potential biases the attention system. In terms of the Planning Framework, an action definition includes effects for attention guidance in the next step. This is specially important when no goal can be reached with the current information. These effects for visual biassing are also expressed as semantic

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGA

predicates (e.g. "look for green and rounded objects"). Hence, it is necessary an *interpreter* module in order to transform those abstract predicates into a new attentional set to be stored in the PMM (see fig. 5.7).

Although some approaches has tried to deduce the biasing parameters from a specific target object using Bayesian inference [Yu et al., 2012], there is no general solution for this kind of interpreters at present. The procedure strongly depends on the number and type of features employed in the attention model and the guidance to be achieved. So, at the moment, the solution should be manually developed, or the interpreter should be previously trained using Machine Learning techniques, for each concrete action and problem. For example, Frintrop [2006] uses supervised learning to accomplish a visual search.

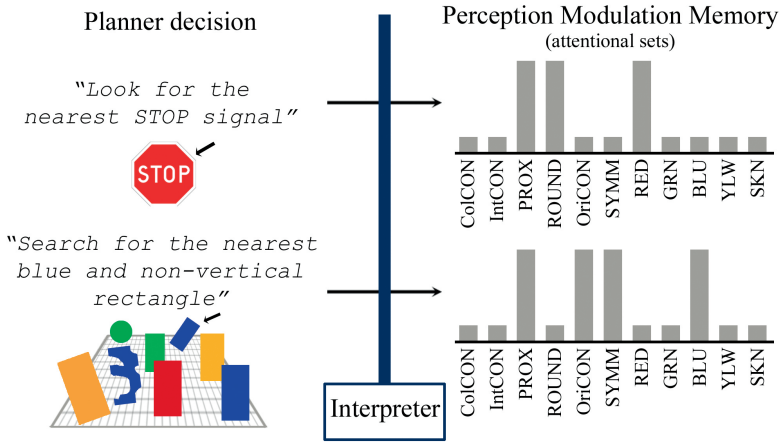


Figure 5.7: An example of a basic interpreter. Planning decisions are manually translated into features to be highlighted by the attention system.

From the Oversubscription Planner point of view, the process described above can be explained through an algorithm (see Algorithm 3). This algorithm receives the domain description, D, and the static part of the problem, P_s. The static part of the problem includes all the elements of the problem that do not change during the resolution: the header, the types and objects definition, the goals and all the static predicates that do not need to be perceived. It is assumed that the planner can generate partial

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

plans depending on the perceived information and the attentional set, that can be modified by the planning actions.

Algorithm 3: The perception-planning-action loop in terms of the Oversubscription Planner

given
The domain description, D ;
The static part of the problem description, P_s ;
An Oversubscription Planner, $OPlanner$;
The attentional set $\vec{\lambda}$;
repeat
$PerceiveState(\vec{\lambda}) \rightarrow s$;
$ComposeProblem(P_s, s, \vec{\lambda}) \rightarrow P$;
$Plan(OPlanner, D, P) \rightarrow Plan$;
$Execute(Plan)$;
$UpdateAttentionalSet(\vec{\lambda}, Plan)$;
until *a goal state is achieved*;

The algorithm repeats the loop until all the goals are achieved. In the first step of the loop, the current state, s, is perceived. With the static part of the problem, P_s, the perceived state, s, and the attentional set, $\vec{\lambda}$, the new planning problem, P, is composed. Next, the Oversubscription Planner is called, generating a plan able to reach the goals, that can be solved with the perceived information, if any. The generated plan is then executed and the attentional set $\vec{\lambda}$ is updated according to such plan.

To sum up, the proposed architecture represents a flexible, adaptable, easy-to-handle, biologically plausible and general-purpose cognitive model joining perception and reasoning together. Whereas other models fail in providing a complete characterisation of the problem, the proposed cognitive architecture, combining the OBVIAS attention system and a two-level need-based reasoning model, provides a complete and bidirectional loop-closing solution. The double imbrication between reasoning and perception is easily observable: depending on the categories perceived (covered needs), different tasks are triggered; and vice versa, depending on the dominant tasks, the perception system modifies its parameters so the most relevant

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGA

objects in the scene can change. This mutual modulation implies a dynamic balance of the system. Therefore, the whole system is not likely to remain locked in a specific task due to the combination of inhibition of return mechanisms in the attention model and knowledge-based planning, able to add, remove and prioritise tasks.

Finally, the proposed architecture can also be classified in relation to the three classical paradigms for organising robot information processing (reactive, deliberative and hybrid). From this point of view, the model clearly belongs to the hybrid paradigm. On the one hand, the knowledge-based level acts as a deliberative layer, dividing goals into a set of tasks. The activation of these behaviours placed at rule-based level only depends on the satisfaction of their requirements. On the other hand, tasks are executed according to the reactive paradigm and visual information is routed to each task. Visual data is also redirected to the Planning Framework to continuously update and manage the task-oriented world model.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

UNIVERSIDAD DE MÁLAGA
Publicaciones y Divulgación Científica

6 Experiments and Results

"Solvitur ambulando"

Saint Augustine

The performance of the visual attention system introduced in chapter 4 and the cognitive architecture allowing an attention-based bidirectional loop closing (chapter 5) has been tested using a real robot, placed in a natural environment with no specific restrictions about illumination or background elements. First, the OBVIAS model is separately evaluated in order to analyse its performance in selecting and extracting relevant proto-objects from a natural scene. Then, the whole cognitive architecture is tested. This final evaluation is carried out throughout two experiments. On the one hand, a simplified proof-of-concept highlights the basic concepts of the bidirectional loop-closing. On the other hand, a real social robot is employed to show the reliability of the proposal in a complex and dynamic environment with time-varying tasks.

6.1 NOMADA: a social robot

All the experiments detailed in this chapter have been performed using the same hardware platform: NOMADA, a social robot (fig. 6.1). NOMADA is an autonomous robot specially developed for human interaction. The robot mounts a variety of sensors and it moves using a wheeled holonomic platform. Regarding software performance, an open-source and component-based robotics framework is employed: *RoboComp*.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA



Figure 6.1: NOMADA, a social robot.

6.1.1 Hardware architecture

NOMADA is an experimental social robot developed at the University of Málaga by the scientists belonging to the ISIS research group[1]. The robot has been built as a set of several distributed subsystems and sensors, connected using standard communication protocols.

The mechanical part of the robot has been reused from a Nomadic's NOMAD-200 robot, including the original sensors (bumpers, infrared sensors and a sonar ring). This structure has been used as a starting point to mount new sensors and devices. A new electronic system has also been developed in order to interconnect all the existing subsystems using standard protocols. Fig. 6.2 shows the hardware architecture of NOMADA robot.

The robot is equipped with several devices that complement the aforementioned original sensors. The principal vision system is an ASUS Xtion device, able to compute depth information. The robot also includes a LIDAR sensor and a digital compass. All sensors and actuators are accessible

[1]ISIS stands for Ingeniería de Sistemas IntegradoS (Integrated Systems Engineering Group). The research group belongs to the Department of Electronics Technology at the University of Málaga (Spain).

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

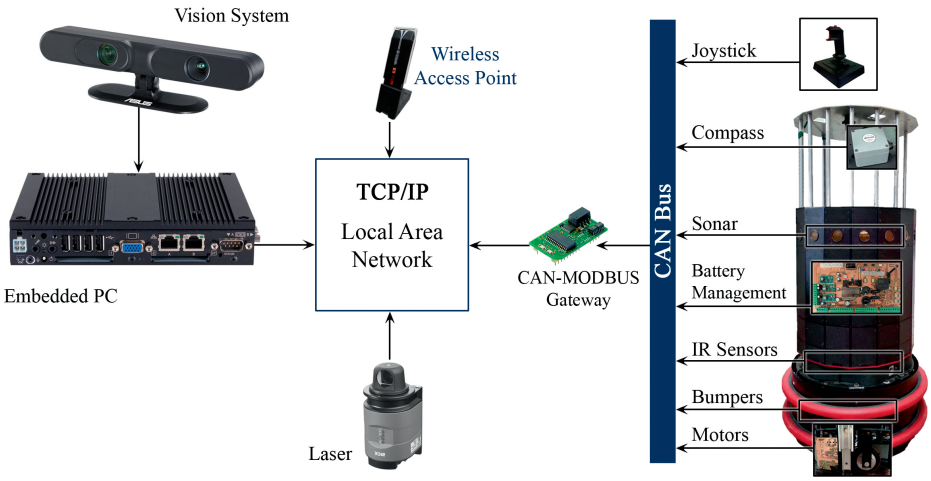


Figure 6.2: NOMADA: hardware architecture. The basic components (motors and simple sensors) are attached to a CAN (Controller Area Network) bus. They can be accessed via TCP/IP using the MODBUS application protocol. Complex sensors (laser or an ASUS Xtion device) are directly managed by an embedded PC.

via a TCP/IP network, using the MODBUS application protocol. An embedded PC manages all the capabilities of the robot. A wireless access point and a standard wired Ethernet port allow the embedded PC to cooperate with other external computers in a local network.

6.1.2 Software architecture

A social and autonomous robot is a very complex system. The hardware should be robust but the software must also be stable, efficient and reliable enough. Robotics software has to face important challenges such as code reuse, scalability, distribution or hardware independence. Software engineers have addressed these problems by developing different approaches that facilitate the development of new software to Robotics scientists. Concretely, NOMADA robot employs an open-source Robotics Framework: **RoboComp**[2].

[2]This framework is available at `www.robocomp.org`

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

RoboComp focuses on an ease of use and low development effort. It provides different software **components** for robotics and the necessary tools to use them [Manso et al., 2010]. The framework is model-driven and component-oriented. There are three key elements in RoboComp: a component model, a communications middleware and a set of tools that facilitates writing and maintaining code [Gutiérrez et al., 2013].

Running components can be distributed over different cores, CPU's or computers using software component technology. Communications are handled by the Ice framework. RoboComp uses in some of its components other tools or libraries such as CMake, Qt4, IPP, OpenSceneGraph and OpenGL (most of them are optional). Programming languages used in the project are mainly C++ and Python. RoboComp provides both client-server and publish-subscribe communication models. The components are connected using **interfaces**.

Latest version of the framework includes Domain Specific Language (DSL) tools so users only need to add the specific working code in their components. Thereby, RoboComp is allowed to support new communications frameworks such as *Nerve*, a lightweight C++ middleware for networked robotics, designed to guarantee the scalability and quality-of-service (QoS) of real-time and critical components [Martínez et al., 2012].

The whole cognitive architecture proposed in this monograph (fig. 5.2 on page 84), as well as the OBVIAS attention system (fig. 4.1 on page 51) have been implemented as RoboComp components in NOMADA robot (fig. 6.3). They can be divided into four groups: components belonging to OBVIAS model, components corresponding to the cognitive architecture and components handling hardware.

The OBVIAS model has been developed using the following components:

Attention: This component implements both the perceptual segmentation algorithm to obtain proto-objects and the saliency computation procedure. It also deals with the Working Memory (WM) management. The component receives a RGBD image from the Xtion device and an Attentional Set from the Perception Modulation Memory (PMM). It returns the list of the most relevant proto-objects.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGA

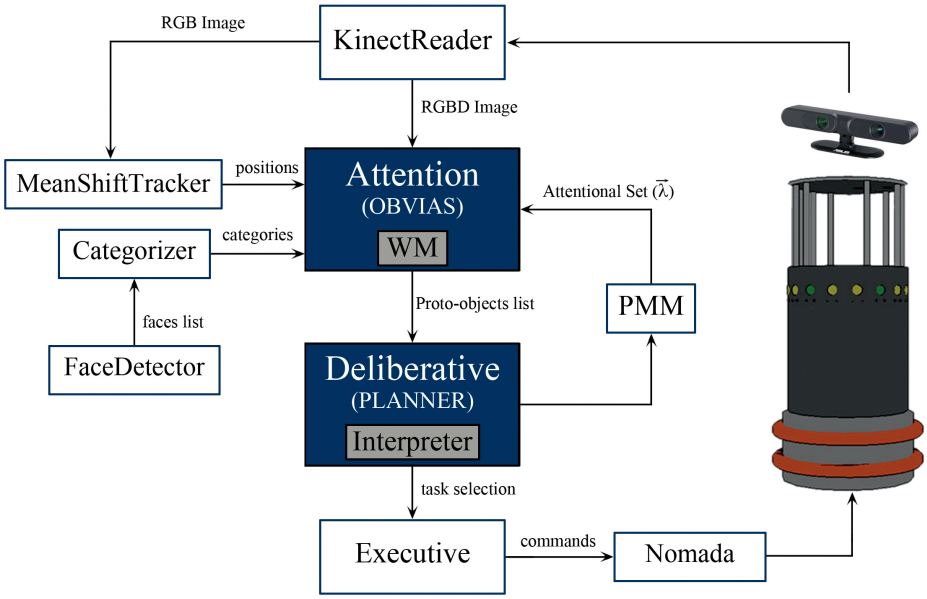


Figure 6.3: Robocomp components implementing the OBVIAS attention model and the need-based cognitive architecture for bidirectional loop-closing.

MeanShiftTracker: The component implements a mean-shift based image tracker. It is used by the attention model in order to carry out the Inhibition of Return (see section 4.4 on page 71). This component receives a RGB image from the sensor and keep all the initialised elements tracked. The *Attention* component controls the execution of the tracker.

PMM: This component corresponds to the Perception Modulation Memory (PMM). It receives the Attentional Set, $\vec{\lambda}$, (see section 4.5 on page 73) from the interpreter at the rule-based level and provides it to the *Attention* component.

Components belonging to the cognitive architecture:

Categorizer: This component reads the list of proto-objects in the Working Memory and assigns a category to them. The categorizer is able

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

to detect simple shapes and elements in the scene based on basic features (see descriptors saved in the WM in section 4.5). In order to detect people, a face detector component is also employed.

FaceDetector: It receives a region of interest from the categorizer and computes the number of available faces in it.

Executive: This component implements a set of basic movement routines. For example, "navigate to the point (x, y, z)", "get an object", "drop an object", "look at the point (x, y, z)", etc.

Deliberative: The Planning Framework from the knowledge-based level (section 5.2 on page 91) is placed here. This component receives the categorised list of proto-objects from the *Attention* component as input. Depending on the perceived elements and the current behaviour, this component selects the actions to be executed and the new Attentional Set, following the two-level reasoning methodology explained in chapter 5. The interpreter that transforms semantic predicates of visual guidance into Attentional Sets is also implemented in this component, in order to reduce the communication load.

Finally, two components directly operate over the hardware. They are the only ones designed for specific devices.

KinectReader: This component receives raw data from an ASUS Xtion or Microsft Kinect device as input. It can provide a RGB image, a RGBD image or a Depth image of the scene using RoboComp-compliant types.

Nomada: It translates the commands provided by the *Executive* component into NOMADA low-level instructions.

Since RoboComp allows components to be distributed over different computers, the embedded unit is aided by an external PC to enhance the overall performance. The external computer executes the *Deliberative*, *Categorizer*, *FaceDetector* and *PMM* components. Depending on the specific experimental set-up, some components can be disabled if they are not necessary.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

6.2 Visual attention model

The performance of the bottom-up part of an attention system is still difficult to quantify in an objective way. Despite the fact that computational attention models have been developed for about 30 years, there is no real consensus on any assessment method at present [Le Meur et al., 2006]. Some attempts have been made in order to establish standard metrics and measurement methods (e.g. Aziz and Mertsching [2010]). However, the different models are usually self-evaluated by their authors, applying their own metrics [Aziz and Mertsching, 2009; Borji and Itti, 2013]. Therefore, a full objective and quantitative comparison is not currently feasible.

The principal problem lies in defining a valid and objective ground-truth because it is difficult to obtain the "right" focus of attention on a natural scene. This problem is usually addressed by assuming the data obtained from eye-tracking experiments with humans as ground-truth. Nevertheless, it has been demonstrated that there is a great variety of available scan-paths, depending on each subject, their preferences and their motivations [Mannan et al., 1997]. For example, a person whose favourite colour is green is more likely to be attracted by this specific colour when observing a scene. Moreover, these experiments usually consider only static scenes.

Consequently, the bottom-up part of OBVIAS model has mainly been evaluated in terms of qualitative criteria, showing that the model obtains relevant objects in a reasonable manner from different environments. Nevertheless, in order to provide an objective performance measurement and allow a further comparison with other models, a quantitative evaluation has also been performed, computing the metrics suggested by Le Meur and Baccino [2013]. In this evaluation, all the parameters in the attentional set are adjusted to the same value (i.e., $\lambda_i = \frac{1}{11}$, $\forall i$) so there is no explicit attention guidance. Finally, the top-down influence that allows task-oriented behaviours will be evaluated as part of the loop-closing experiments involving the whole cognitive architecture (section 6.3).

6.2.1 Qualitative evaluation

As it was aforementioned, computational attention systems are often compared to human perception. Although this procedure is not accurate

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

and objective, it provides an intuitive and easy to perform method to evaluate how reasonable the computed foci of attention are. Therefore, the attention model needs to be applied to scenes where some elements are clearly recognised as relevant.

The History of Art is plenty of examples where the artists aim to draw the attention of the observer to specific parts of their artwork. Even Abstract Art attempts to evoke certain sensations by combining different colours and shapes. Does an artificial system direct attention to the same places than people do when observing a masterpiece? Fig. 6.4 shows the result of executing OBVIAS model over two different paintings.

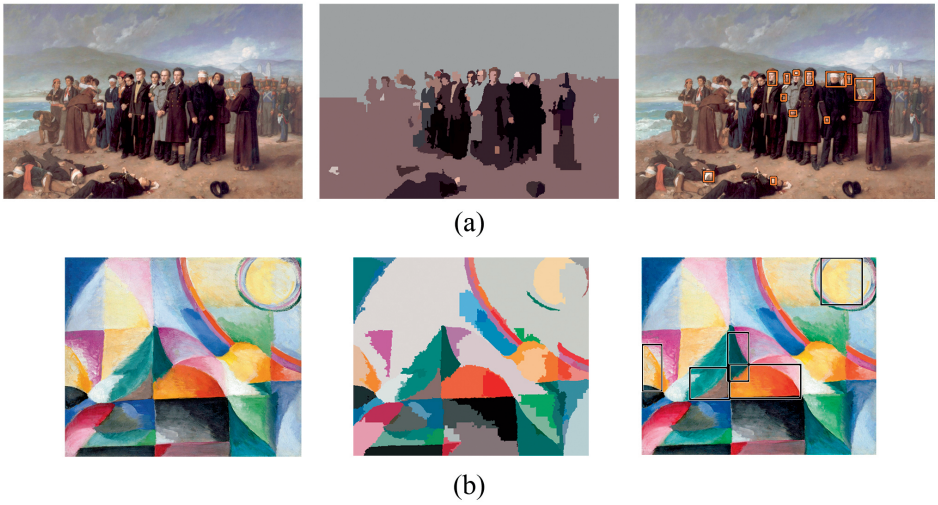


Figure 6.4: OBVIAS applied to Art Paintings. In (a), *The Execution by Firing Squad of Torrijos and his companions on the beach at Málaga* (Antonio Gisbert Pérez, 1888); the 12 first fixations are shown (orange bounding-boxes). In (b), *Simultaneous Contrasts* (Sonia Delaunay-Terk, 1913); the 5 first fixations are shown (black bounding-boxes). Center column corresponds to perceptual segmentation of the scene.

In the figurative painting (fig. 6.4.a), it is clear that the principal actors are Torrijos and his colleagues awaiting their execution. The artist depicts a dark and cloudy environment, contrasting with the pallid faces. Imagine yourself in front of this painting. You are probably looking at the condemned prisoners. When running OBVIAS over this painting, the

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

result highly fits the expected situation: faces and hands of people at the foreground are highlighted.

Obviously, abstract paintings are more difficult to analyse. The motivation and the intentions of the artist are usually hard to deduce. However, even in the absence of structured forms and objects, the picture is able to attract the attention of the observer. Fig. 6.4.b shows the first fixations of OBVIAS over an abstract painting. Due to the definition of saliency handled by OBVIAS, attention is drawn to elements with a high contrast with their surrounding, specially in terms of colour. It is worth mentioning that OBVIAS is able to address this kind of images despite the fact that the model is object-based. Hence, the model acts as all early-selection models (recall section 2.1.2 on page 13).

Subjective and qualitative analysis of computational attention systems usually also entails real-world images with an easy interpretation about relevance. The images are provided to the system and the evaluation of the results is left to the user [Frintrop, 2006]. Certain scenes and environments reveal intuitively clear salient regions. For example, traffic signals and vehicles lights are explicitly designed to stand out from their surrounding. Some sports also uses distinctive colours to designate the balls, the teams or the leader. Animals have characteristic colours and forms too. Fig. 6.5 shows some examples of such kind of images computed by OBVIAS.

For instance, the third and the sixth rows of fig. 6.5 show how attention is drawn to traffic signals and lights. The lights of the semaphore (note that all lights are turned on in the image) are clearly marked as the most relevant objects by the system. The emergency triangle rapidly pops out too. In this case, the rear light of the car is also highlighted due to their contrast with the colour of the car. Examples of animals are shown in the second and the fourth rows. Parrots and butterflies have lively colours that normally stand out the environment. Note that the perceptual segmentation algorithm divides the wings of the butterfly and the body of the parrots into several proto-objects due to their scale in the image. In the first row, the most relevant items coincide with the girl's face and dress, her hair and some object with garish colours. Finally, the fifth row provides a sports example. In the *Tour de France* bicycle race, the first classified rider

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

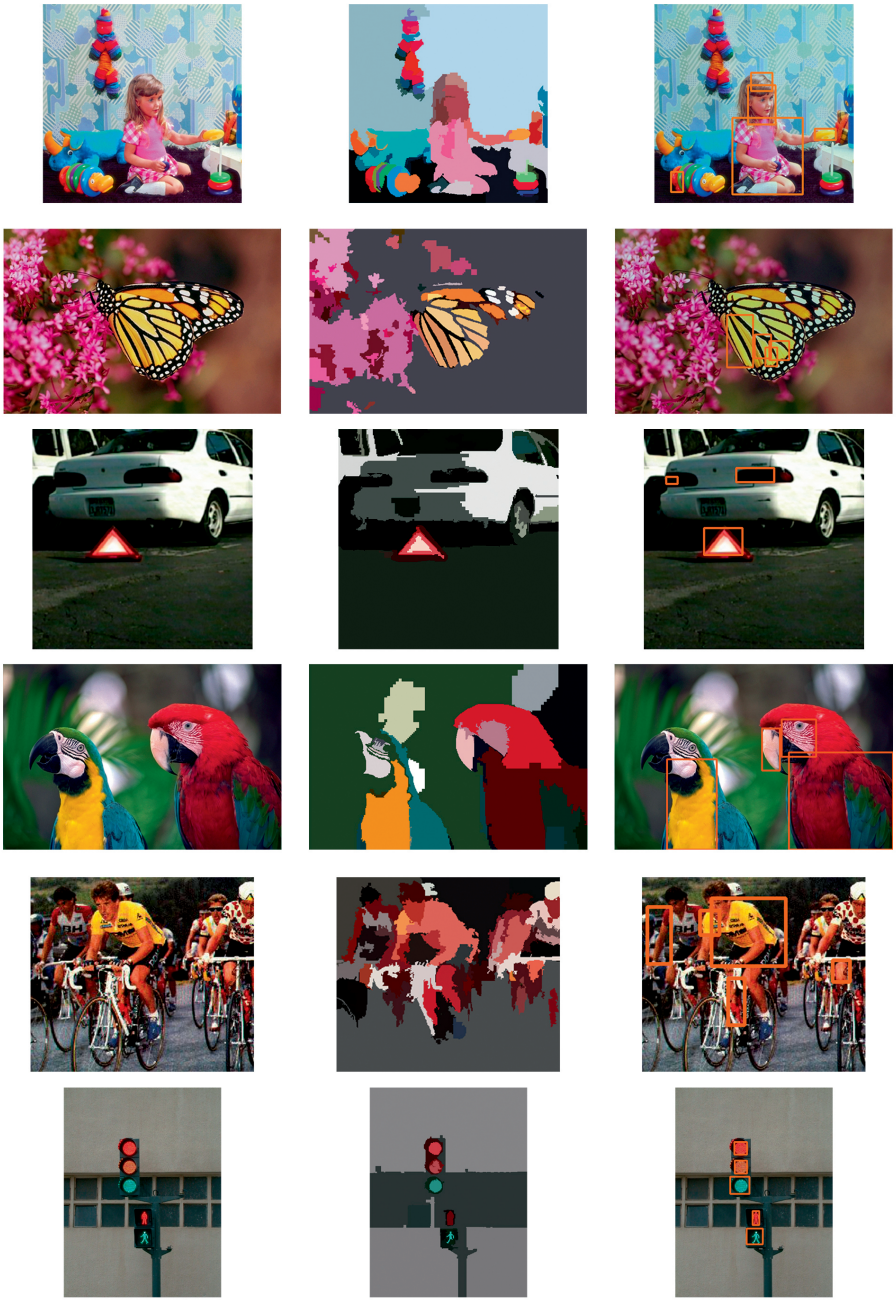


Figure 6.5: OBVIAS applied to real-world scenes. Up to 5 most relevant proto-objects are shown (orange bounding-boxes). Center column corresponds to perceptual segmentation of the scene.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

wears a specific yellow jersey in order to make him noticeable. Indeed, the results of OBVIAS computation really highlights the leader.

Another method to qualitatively evaluate a computational attention model consists in analysing *pop-out* scenes, that are often used in psychophysical experiments. Psychologists typically work with synthetic images composed by several and similar items. Only one or few elements differ in one feature from the remainder of the scene. Thereby, each target can be designed to specifically test the performance of the system to pop out a feature. Some examples of pop out scenes computed by OBVIAS model can be seen in fig. 6.6.

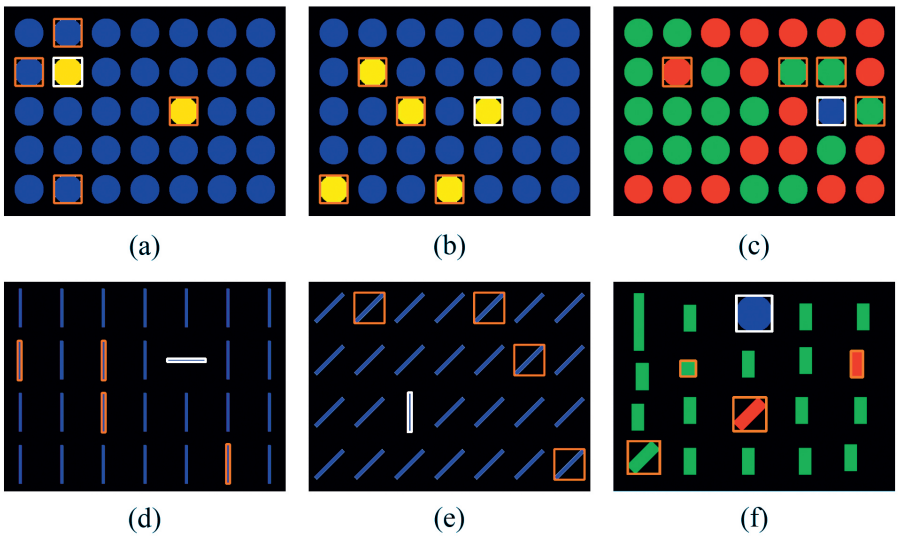


Figure 6.6: OBVIAS applied to psychological *pop-out* scenes. The most relevant proto-object (white bounding-box) and the remainder of the Working Memory (up to 5 elements; orange bounding-boxes) are shown. All popping out targets are successfully and immediately detected (one execution).

In fig. 6.6.a and fig. 6.6.b, the targets are the yellow circles among the blue ones. Recall that OBVIAS is able to compute up to 5 relevant objects at once. Therefore, when there exist only two targets (fig. 6.6.a), three blue circles are randomly selected to complete the Working Memory. In fig. 6.6.c, there are three colours involved. Again, the system rapidly focus on the

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

most differing target (in this case, the blue circle). The rest of proto-objects in the Working Memory are selected depending on the contrast with their neighbours. Fig. 6.6.d and fig. 6.6.e are similar to the first ones but using orientation as biasing feature. Finally, 6.6.f shows an image with targets distinguishing in different features. Note how the Working Memory is filled with items that are clearly different to the repeated green and vertical bars.

6.2.2 Quantitative evaluation

As it was previously pointed out, there exists a wide range of different metrics to quantitative measure the bottom-up performance of an attention model. Some of these methods normally consider saliency maps and eye movement data as probability distributions. Hence, their similarity can be measured by computing the distance between both distributions, using metrics such as the Kullback-Leibler Divergence or Percentile. Saliency maps can also be considered as binary classifiers. Then, signal detection theory analysis, such as the Receiver Operating Characteristic Analysis (ROC), can be applied to figure out the performance of the classifier. Other approaches identify saliency data with random variables. Therefore, their statistical relationship can be measured computing a Correlation Coefficient or a Normalised Scanpath Saliency (NSS). More evaluation measures can be found in Borji and Itti [2013], Aziz and Mertsching [2009] and Le Meur and Baccino [2013].

In this section, the quality of the saliency maps computed by OBVIAS is measured using two metrics: the NSS and the hit rate obtained from a ROC analysis. Both are hybrid metrics that involve a set of visual fixations, obtained from eye-tracking experiments with humans, and a saliency map. Unlike methods that previously compute saliency maps from eye movements data, hybrid metrics are non-parametric. Thus, the computed saliency map is directly compared against raw fixations data.

Nowadays, several eye movement datasets of still images are available in the internet. However, it should be pointed out that there is no consensus about some experimental settings, such as the stimuli presentation time. Concretely, OBVIAS model has been evaluated using the datasets provided

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGUNA

by Le Meur and Baccino [2013] and Bruce and Tsotsos [2006][3] (fig. 6.7). Le Meur's dataset contains 40 images, observed by 30 subjects in a free viewing task. Each image is shown for 14 seconds. Bruce's dataset contains 120 images, observed by 20 subjects in a free viewing task. Each image is shown for 4 seconds. Moreover, these datasets allow an objective comparison with the results obtained from other computational attention models, computed by Le Meur and Baccino [2013].

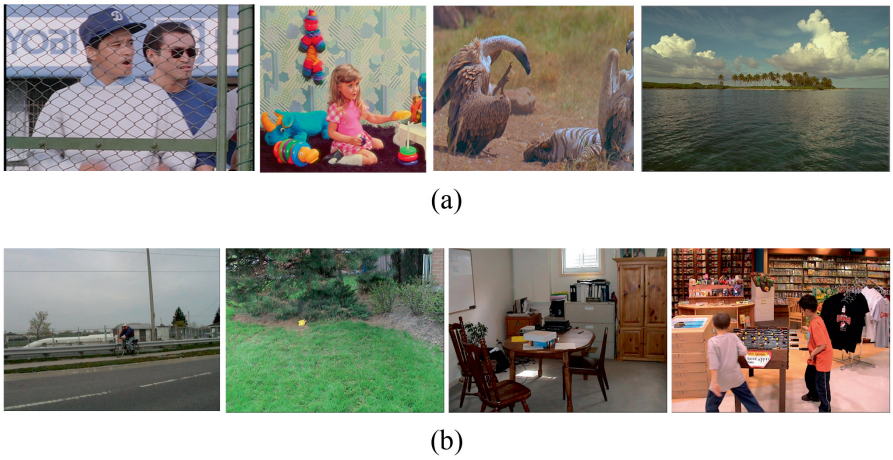


Figure 6.7: Some example images from datasets used for quantitative evaluation. (a) Le Meur's dataset [Le Meur and Baccino, 2013]. (b) Bruce's dataset [Bruce and Tsotsos, 2006].

Receiver Operating Characteristic Analysis (ROC)

ROC analysis [Green et al., 1966] is used for the evaluation of binary classifier systems with a variable threshold. The obtained saliency map is treated as a binary classifier in every pixel in the image. Pixels presenting larger saliency values than a threshold are classified as fixated whereas the rest of the pixels are classified as non-fixated. This procedure is repeated for different saliency thresholds, from 5% of saliency to 100%, using 5% intervals. Then, the *hit rate*, i.e. the amount of fixations that are correctly

[3]Both datasets have been downloaded from `http://people.irisa.fr/Olivier.Le_Meur/publi/2012_BRM/index2.html#suppMat`

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

predicted by the saliency map, can be represented as a function of the threshold. Thereby, a score of 1 corresponds to a perfect prediction.

In order to define a reasonable upper bound to ROC, the inter-observer variability is taken into account. This inter-observer congruency (IOC) can be measured by using a *one-against-all* approach. The procedure computes the similarity between the fixations of one observer against the other subjects. The final value is obtained as an average of similarity over all subjects [Torralba et al., 2006].

Fig. 6.8 shows the resulting curves of applying a ROC analysis to OBVIAS model. Curves representing an upper bound (due to IOC) and a lower bound (chance level) are also provided. It can be noted that the performance of the system is quite similar for both datasets.

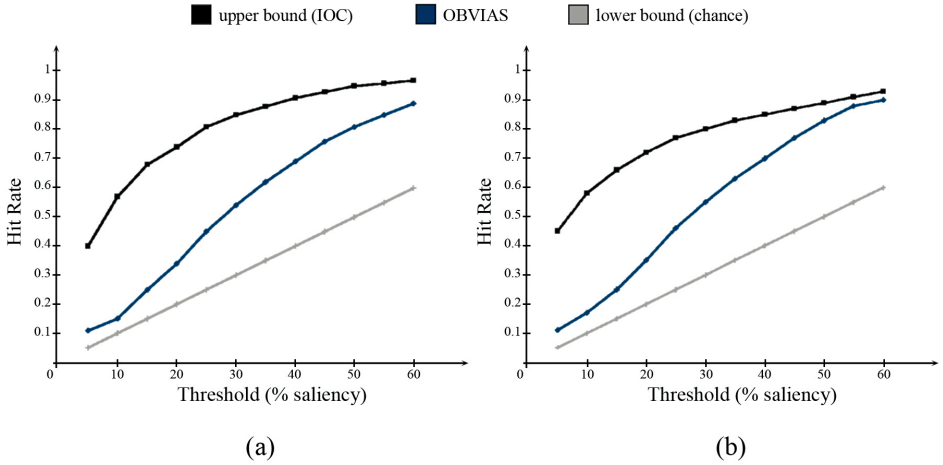


Figure 6.8: ROC analysis of OBVIAS. (a) Le Meur's dataset [Le Meur and Baccino, 2013]. (b) Bruce's dataset [Bruce and Tsotsos, 2006].

Normalised Scanpath Saliency (NSS)

The Normalised Scanpath Saliency (NSS) [Peters et al., 2005] is defined as the response value at the human eye position (x_n, y_n) in a saliency map that has been normalised to have zero mean and unit standard deviation. That is,

$$NSS = \frac{SM(x_i, y_i) - \mu_S}{\sigma_S} \tag{6.1}$$

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

being $SM(x, y)$ the computed saliency map and μ_S and σ_S the mean and the standard deviation of the computed map respectively.

NSS values greater than zero represents a greater correspondence than would be expected by chance between fixation locations and salient points predicted by the computational model. On the contrary, NSS values less than zero indicates that the model performs no better than randomly selecting a position on the computed saliency map.

Table 6.1 provides the averaged values of NSS for OBVIAS model, using the aforementioned datasets. The values computed by Le Meur and Baccino [2013] for other computational attention models are shown in table 6.2 for comparison purpose.

Dataset	NSS scores for OBVIAS	
	Average	SEM
Le Meur	0.68	± 0.18
Bruce	0.73	$\pm 0,16$

Table 6.1: NSS scores for OBVIAS model. Average and standard error of the mean (SEM) are shown.

Dataset	NSS scores for different models			
	Itti	Le Meur	Bruce	Judd
Le Meur	0.6 ± 0.1	0.77 ± 0.13	0.6 ± 0.09	0.82 ± 0.11
Bruce	0.99 ± 0.05	0.87 ± 0.03	0.72 ± 0.04	0.87 ± 0.05

Table 6.2: NSS (average±SEM) scores for Itti's model [Itti et al., 1998], Le Meur's model [Le Meur et al., 2006], Bruce's model [Bruce and Tsotsos, 2009] and Judd's model [Judd et al., 2009]. Reference: Le Meur and Baccino [2013].

6.2.3 Discussion

The qualitative evaluation of the bottom-up part of OBVIAS model reveals that attention is directed to relevant parts of the scene in a reasonable manner. Examples from diverse areas such as art paintings, daily life environments and psychological experiments data have been analysed. They

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

have demonstrated that the system is able to yield a good approximation of human behaviour.

Regarding the metrics employed in quantitative evaluation, the results obtained for NSS and ROC analysis are close to the compared attention systems average. In comparison to other models (see curves in Le Meur and Baccino [2013]), ROC response is slightly worst for the first thresholds. However, the performance is similar to other models when considering more than 30% of salient pixels. Since NSS and ROC are computed over pure saliency maps, they do not take into account that OBVIAS model is able to select up to 5 salient elements at once. Therefore, the selection of relevant items should be faster in real-world executions, improving the overall performance. Results for both datasets show that OBVIAS is highly consistent, yielding similar results.

Metrics proposed in the literature for quantitative evaluation still show some important lacks. The main one is concerned with dynamic considerations. All metrics compare saliency maps and eye fixations in a pure static manner. The amount of time that a part of the scene is fixated is not considered at all. Then, phenomena such as the Inhibition of Return are difficult to objectively measure. Furthermore, one of the main advantages of OBVIAS model is related to its continuous adaptability to time-varying environments.

In conclusion, evaluation has shown that bottom-up performance is reasonable and a wide range of environments can be successfully addressed. However, the objective metrics do not provide specially significant results, being the system close to the overall average. Nevertheless, the principal strength of the model lies in its effective connection with a complete deliberative layer, managing top-down cues in a effective manner. Thereby, the attention system is able to dynamically handle perceptual errors during execution, allowing to close the perception-reasoning loop as it will detailed in the next section.

6.3 Cognitive architecture

Architecture research takes place at *system level*. This implies that the evaluation of a cognitive architecture faces more challenges than the mere

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

evaluation of component knowledge structures and methods [Langley et al., 2009]. A cognitive architecture is a *whole*, where different modules interact. Hence, its evaluation can not be seen as a mere sum of the performance of its components. From an engineering point of view, an architecture should exhibit the same qualitative robustness that people do. Hence, the evaluation presented here mainly focuses on analysing the reliability of the need-based architecture when addressing complex environments and problems.

As it has been pointed out throughout this monograph, previous systems linking perception and reasoning assume an unidirectional and incomplete characterisation of the loop-closing problem. Therefore, from the evaluation point of view, it is difficult to achieve an objective comparison between the proposed architecture and the previously developed ones.

Domains and problems involved in the evaluation experiments are defined in PDDL (Planning Domain Definition Language) [McDermott, 1998; Gerevini et al., 2009]. Although the main parts of the different domains are explained in the subsequent sections, the complete PDDL code can be seen in Appendix B.

6.3.1 Experiment 1: Putting coloured cards in order

The first experiment domain involves a set of coloured cards labelled with letters. The goal of the problem consists in sorting all the cards into alphabetical order. The experiment is mainly a proof-of-concept study about the proposed solution to integrate attention and planning. The selected scenario allows to guide the perception system in order to look for a specific colour. Besides, the tasks that can be executed at once depend on the cards perceived at each moment. Thereby, the bidirectional loop-closing can be accomplished and studied, showing that perceptual guidance is able to enhance the overall performance of the system in comparison with the absence of visual guidance.

The PDDL static predicates of the planning domain contain the information about the correct ordering of cards `(order A B)` and also about the colour of each card `(colour A yellow)`. The initial state is completed

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

by the perception module by marking as visible the cards that are stored in the WM: `(visible ?x)`. A fourth predicate, `(stack ?x ?y)` stores information about the top and the bottom of the already created stacks. In the initial state, all the cards form a one-card stack. Deterministic non-sensing actions are assumed, so the information about already stacked cards can be rolled over next iterations.

There are only two actions in the planning domain. The first one stacks the visible cards in order if they can be stacked, namely, if two consecutive cards or stacks are visible. As a side effect, every time a card is stacked, the relevance of its colour is increased (perceptual guidance). Therefore, it is very likely to be perceived again in the next perception step. The second action is applied when no card can be stacked in an iteration. It randomly selects one of the visible cards and increases both the relevance of its colour and the colour of the next card. To avoid this action to be executed unless nothing else can be done, its cost is 300 times higher than the stack action's one. Since the aim of this first experiment is to show the relationship between perception and planning, the action proposed by the planner is executed by a human.

Fig. 6.9 shows the different configurations of the analysed domain. The actions in the domain are able to take all the possible advantages from the scenario described in fig. 6.9.a. In this case, there are 4 sets of letters with the same colour: A-F (yellow), G-M (blue), N-S (green) and T-Z (red). Consecutive letters have the same colour. On the contrary, another scenario is defined (fig. 6.9.b), where there are no consecutive letters with the same colour. Concretely, the four sets are:

- *Yellow* letters: A,E,I,M,Q,U,Y
- *Blue* letters: B,F,J,N,R,V,Z
- *Green* letters: C,G,K,O,S,W
- *Red* letters: D,H,L,P,T,X

Finally, a last configuration where all cards have the same colour is introduced (fig. 6.9.c) in order to cancel out the influence of the planner over the perception system (no perception guidance can be applied).

The attention system has also been simplified due to the restriction about the elements to be perceived (only rectangular and coloured cards).

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGA

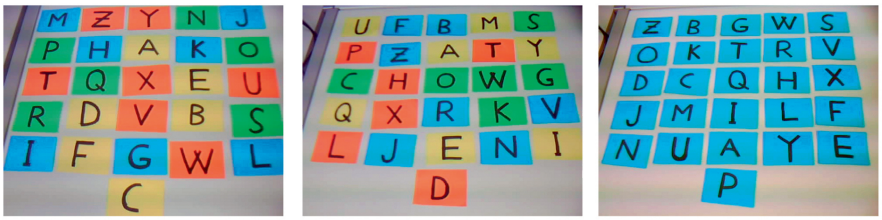


Figure 6.9: The three different scenarios for experiment 1. In (a), consecutive letters have the same colour. In (b), two consecutive letters never have the same colour. In (c), all letters are blue.

In order to clarify the interpretation of the results, the number of features computed in saliency equation (eq. (4.20) on page 69) is reduced to similarity to the four basic colours. Hence, the global saliency of each proto-object, (SAL_i), is computed as follows:

$$SAL_i = \lambda_7 \cdot RED + \lambda_8 \cdot GRN + \lambda_9 \cdot BLU + \lambda_{10} \cdot YLW$$

verifying $\sum_{i=7}^{10} \lambda_i = 1$. Other features also available in the attention model are not taken into account. Therefore, all the cards have the same *a priori* saliency, i.e., the influence of bottom-up attention is highly reduced.

Fig. 6.10 shows the image processing involved in obtaining the saliency map. Once the WM is filled up with the most relevant proto-objects, an OCR (Optical Character Recognition) algorithm is employed as a categorizer[4]. Consequently, the Planning Framework receives the corresponding letter as a predicate. If a letter is not recognised, the proto-object is not categorised. In fig. 6.10, the biggest relevance is given to blue and red color (λ_7 and λ_9 are greater than the others in this case) so, in the saliency map, blue and red proto-objects are brighter (more salient).

There exists a task for each pair of cards that can be sorted at the rule-based level. Thereby, there is a task for putting in order letters A-B (being the needs to be covered the perception of categories "A" and "B"), a task for putting in order letters B-C (being the needs to be covered the

[4]The OCR algorithm uses the Tesseract-OCR engine. Code available at `code.google.com/p/tesseract-ocr/`.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

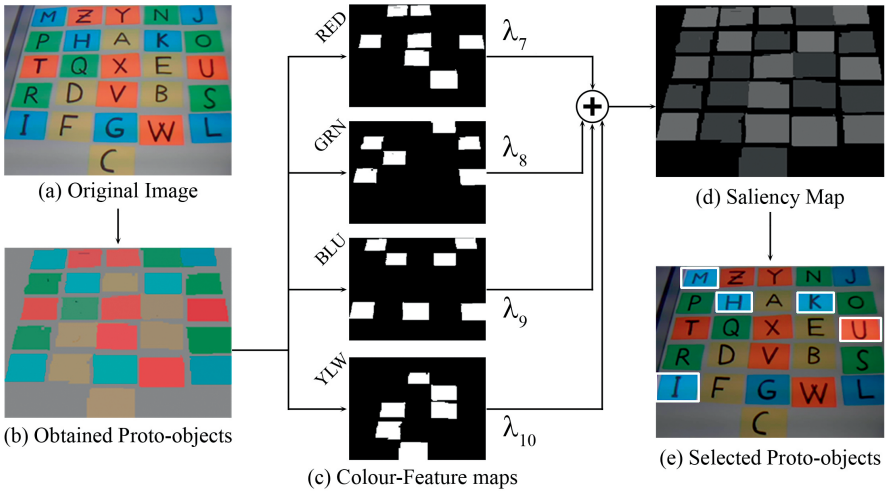


Figure 6.10: Visual Attention processing. The scene is divided into proto-objects (they normally match the real existing cards). Then, saliency is computed in terms of four features: similarity to red, green, blue and yellow colours. Finally, the most relevant cards are selected and stored in the Working Memory (these cards are highlighted with white bounding-boxes in the image).

perception of categories "B" and "C"), etc. When all the needs are satisfied (namely both letters are perceived), the action is executed. Fig. 6.11 shows additional examples.

The whole system is evaluated using 4 different approaches:

Case 1: The scenario from fig. 6.9.a is used. The planner provides the solution as aforementioned.

Case 2: The configuration from fig. 6.9.b is employed. The planner tries to follow the same strategy as in case 1 (a bad guiding strategy for this scene configuration).

Case 3: Same scenario as in case 2. However, the colour proposed by the planner is marked as **less relevant** than the others, instead of the strategy followed before.

Case 4: All the cards are blue (fig. 6.9.c). The planner is not able to highlight a specific colour as the most relevant. In this case, half of the bidirectional loop-closing is not available.

In fig. 6.12, two significant consecutive iterations of evaluation (case 1)

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

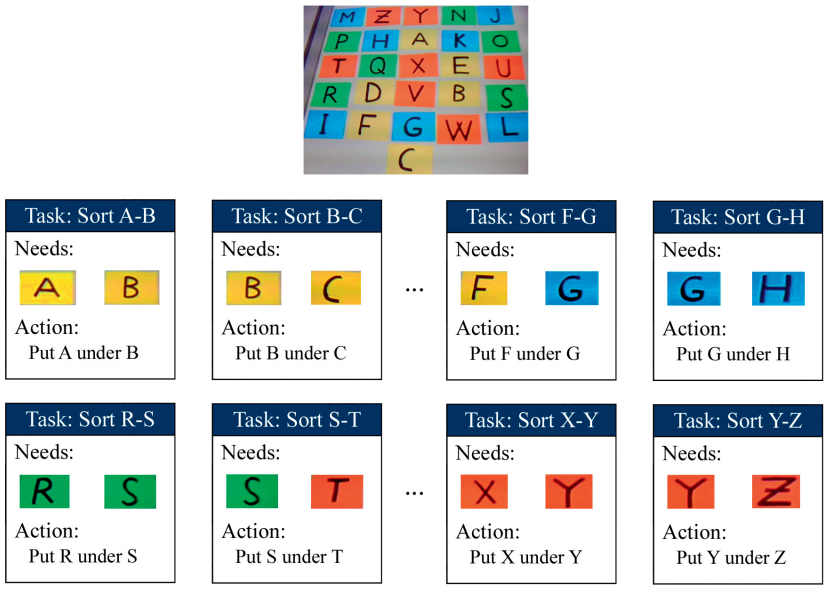


Figure 6.11: Tasks placed at rule-based level for experiment 1. Each task puts in order two letters, being the needs to be satisfied the perception of the categories corresponding to both letters.

are shown (the complete video sequence is available at `http://youtu.be/1fZWBJMnzXc`). In the first one (upper in the figure), only letters T and S can be sorted from the objects in the WM (marked with white bounding-boxes). Consequently, the solution plan consists in putting in order these letters and giving more relevance to the colour of the card in the top (in this case, red) varying the related λ_i. As a result, in the next iteration, the most salient objects are the red ones, allowing the planner to obtain more solutions at once.

The variation over time of the different perception parameters depending on the plan to execute is shown in fig. 6.13. As depicted in the figure, the system began putting in order the green cards, followed by the red ones, the blue ones and, finally, the remaining loose stacks. When all the letters of the same colour are stacked, that colour loses relevance because the planner guides the attention system to look for the rest of colours.

Finally, table 6.3 shows the results of the experiments and fig. 6.14 represents the number of iterations needed in each case to solve the problem.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

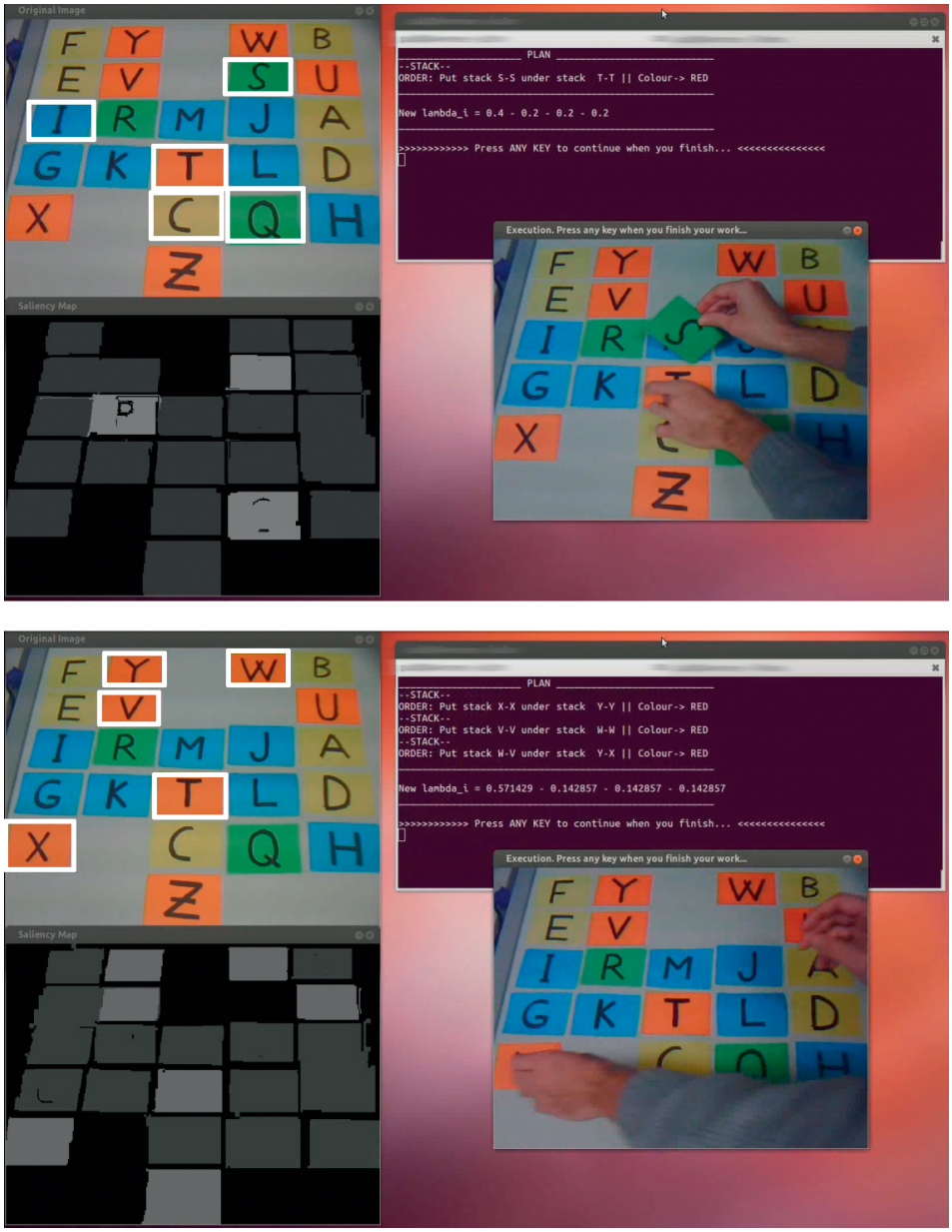


Figure 6.12: Two consecutive iterations in the evaluation task.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

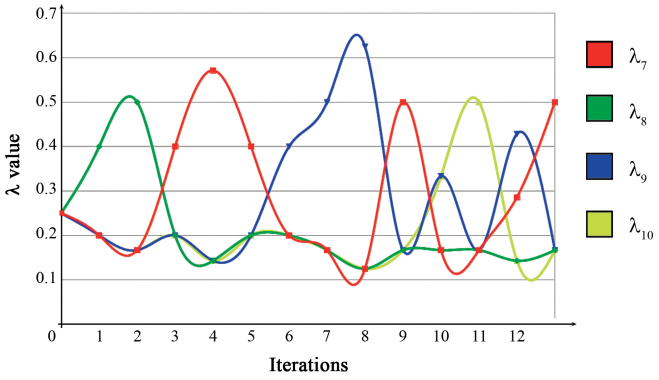


Figure 6.13: Evolution of attentional sets, λ_i, over time. The colour of the curves matches the basic colour feature weighted by each parameter.

The following results can be highlighted:

- As it was expected, the best results are obtained for *case 1*, where the planner is able to guide the perception system in an efficient way.
- *Case 2* produces the worst results due to the fact that the planner guides the vision system in a wrong manner. This solution is even worse than the one in *case 4*, where the planner can not guide the attention system at all. Thus, a bad guiding of the perception system is even worse than a no-guiding option.
- Comparing results for *case 3* and *case 4*, it can be noted that a subtle modification, looking for a more intelligent strategy of the planner, is enough to significantly increase the performance of the system. The iterations needed to solve the problem in *case 3* is reduced almost to the half.

6.3.2 Experiment 2: A social robot classifying objects

As it was mentioned in chapter 5, there is a lack of attention systems able to handle dynamic and interactive environments with time-varying behaviours. The experiment introduced here aims to evaluate the proposed cognitive architecture in terms of such an environment.

The robot has to classify different objects, given by a human, into the corresponding container. It can be seen as a refuse collector and recycling unit. The robot looks for people needing its services, receives objects

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

Case	Iterations		Total Time		Efficiency	
	μ_i	σ_i	μ_t	σ_t	μ_e	σ_e
1	11.9	2.03	2.87	0.56	2.15	0.34
2	55	12.94	13.92	3.16	0.48	0.18
3	27.3	7.68	7.13	2.12	0.98	0.3
4	43.7	13.16	11.18	3.53	0.63	0.22

Table 6.3: Experimental results for experiment 1. The first column shows the number of perception-planning-execution iterations needed to solve the whole problem (all cards are sorted). The second column shows the total execution time (in seconds). The third column expresses the efficiency of the system in terms of actions executed per iteration. Standard deviation is also included.

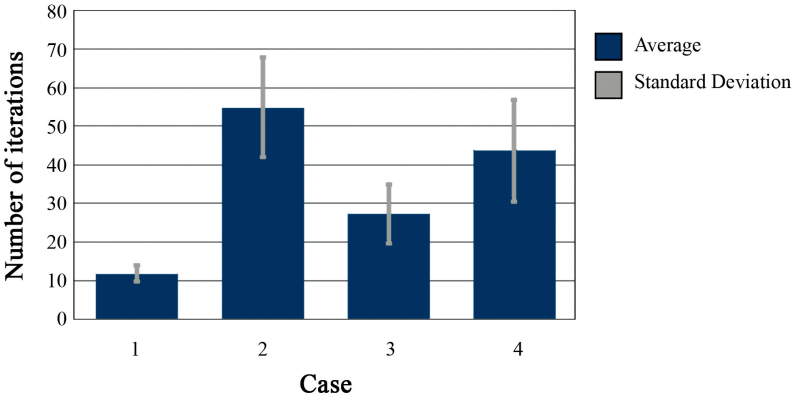


Figure 6.14: Number of perception-action iterations on average needed in each case to solve the whole problem. The standard deviation is also included in the graph. The less iterations needed, the better the performance.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

and, depending on their characteristics, takes them to a specific place (a green container for bottles, yellow ones for plastics, gray ones for organic waste...).

A real and complete implementation of such a recycling robot would require complex algorithms for both the *categorizer* and the *interpreter* (recall section 5.3 on page 94). Therefore, some simplifications have been assumed: (i) people always give balls to the robot; (ii) the containers are represented by coloured rectangles in a wall; (iii) the robot has to bring each ball to the rectangle that matches its colour.

Fig. 6.15 shows different views of the experimental set-up. There are no constraints about illumination, elements or number of people in scene. In order to validate the robustness of the system, some balls are placed as distractors in the middle of the environment.

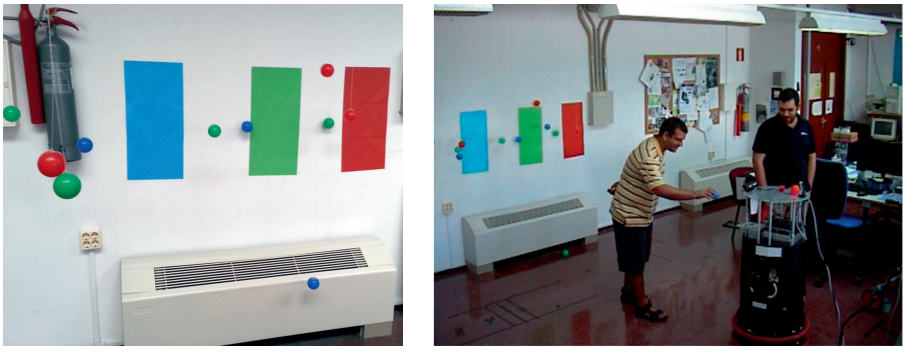


Figure 6.15: Scenario set-up for experiment 2. The robot has different tasks to accomplish in a complex environment, including human interaction.

The robot is supposed to achieve the following behaviours:

- By default, the robot is looking for humans to assist them. In this task, people are the most relevant elements in the scene. Concretely, the nearer person is always preferred. Therefore, these requirements are translated by the interpreter into increasing parameters related to skin colour, roundness and proximity in the associated attentional set.
- When the robot has found people, it moves towards them and detects if they have a ball. If a ball is perceived, the robot opens a compartment to receive the ball. In this case, the parameters to be increased

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

in the attentional set are roundness and proximity, maintaining the presence of humans as a need.

- Once the robot has obtained the ball, the next step consists in looking for the nearest corresponding rectangle in the wall and walking towards it. People and balls are not relevant now, so the attentional set is configured to increase the parameters related to symmetry, proximity and the specific ball colour. Finally, the robot drops the ball and returns to the default task.

An execution is considered successful when the robot performs all the possible actions, showing that:

- The architecture is able to swap among different behaviours, replacing the tasks placed at the rule-based level.
- The architecture is able to guide the attention system depending on the current task.
- The architecture allows to trigger specific tasks in terms of the perceived elements.
- The architecture is able to handle dynamic environments with time-varying tasks.

Fig. 6.16 summarises the tasks placed at the rule-based level, including their needs and their effects over attentional sets. Depending on the current behaviour of the robot, the knowledge-based level adds or removes certain tasks (as it was explained in section 5.1). Concretely, when the *default* behaviour is accomplished (a person is detected), the knowledge-based level removes the search-for-human task and the tasks corresponding to the *assisting human* activity (search for ball, move towards human and get ball) are added. In a similar way, these tasks are replaced by the ones belonging to the *recycling* behaviour when a ball is obtained. Thereby, the tasks that can be activated vary depending on the current context.

Regarding the attention system, there is no simplification at all. Hence, the saliency of each proto-object is computed taking into account all the available features in OBVIAS model (see eq. (4.20) on page 69). Besides, all the software components depicted in fig. 6.3 are employed.

The static predicates of the PDDL planning domain define the shape and number of the possible elements in the scene. For instance, `(colour`

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGA

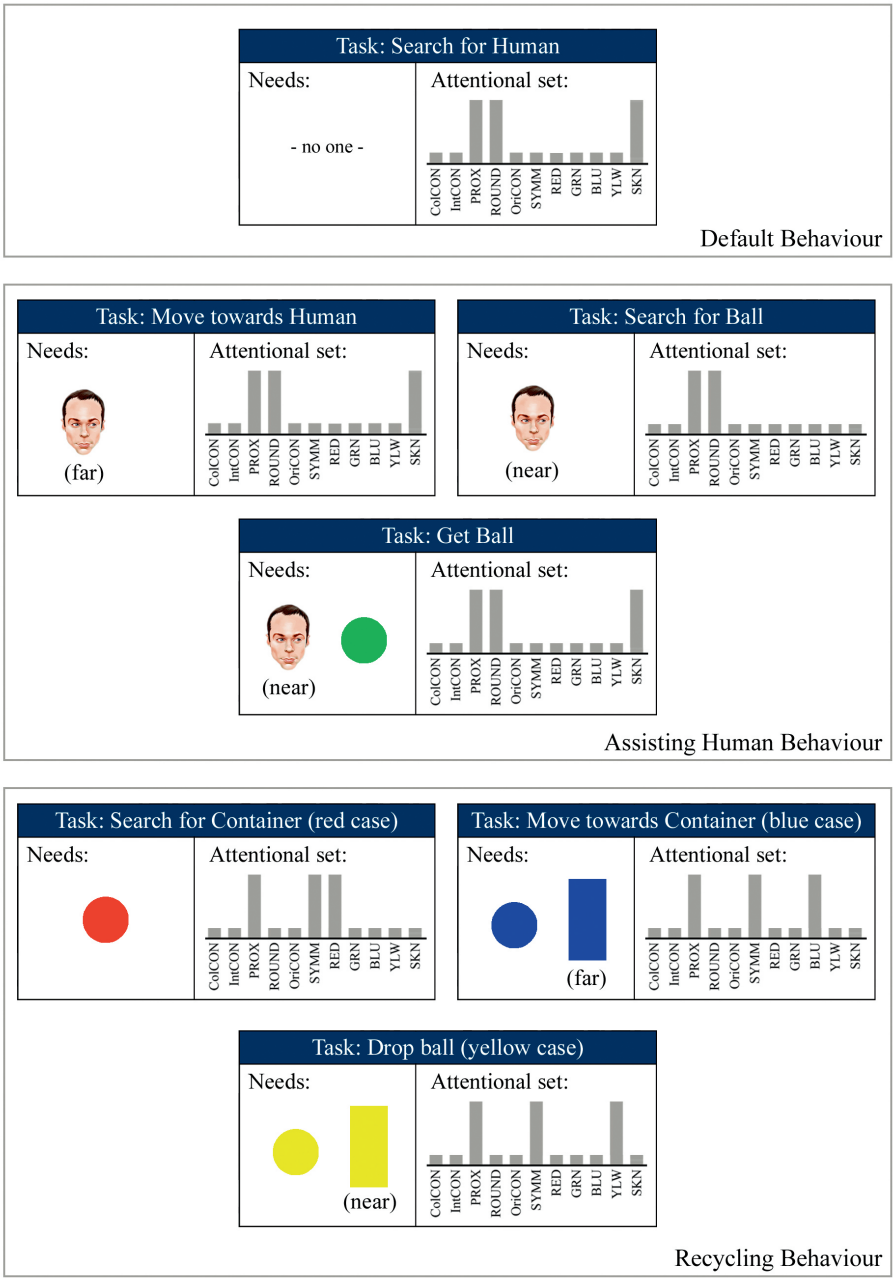


Figure 6.16: Tasks placed at rule-based level for experiment 2. Not all tasks are placed at the rule-based level at the same time. The knowledge-based level establishes the active tasks depending on the current behaviour.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

`human0 skin-color)`, `(shape ball5 round)`, `(shape pole5 vertical)`. Depending on the elements stored in the WM, the objects are marked as visible, `(visible ball3)` or `(visible human0)`, and their colours are obtained, `(colour pole2 yellow)`. Analysing depth information, the categorizer is also able to set some semantic relationships between objects: there is a person holding a ball `(has human2 ball1)`, a ball is at the waypoint 2 `(at ball1 wp2)`, robot is at the waypoint 5 `(at-robot robot1 wp5)`, etc. The complete PDDL domain definition is available in Appendix B.

Fig. 6.17 shows one successful execution of the experiment 2, including a SMPTE Timecode[5]. The categorised content of the Working Memory, the plan to be executed and the specific attentional set to be stored in the Perception Modulation Memory are also displayed. As it can be observed, the number of relevant proto-objects varies in each frame. In the first two frames (first row in figure 6.17), the robot is executing its default behaviour, looking for humans (note the high relevance given to skin colour in the attentional set). In the third frame (first in the second row), the knowledge-based level has swapped the tasks to those corresponding to the assisting behaviour. In this case, nothing but a person is visible. Since the robot is near the person, the *search for ball* task is executed. Otherwise, the moving task would have been executed to approach the far person.

When the robot finds a ball and a person, the *get ball* task is executed (fourth displayed frame). Immediately, another change of behaviour is performed and the tasks corresponding to the recycling behaviour are loaded on the rule-based level. While the "container" is not visible, the *search container* task is executed (frames 5 and 6). Notice that the colour corresponding to the obtained ball (in this case, blue) is highlighted in the attentional set in order to find the correct marker. In frames 7 and 8 (fourth row), the *moving* task is activated by a far visible marker. Finally, once the robot is near the "container", the ball is dropped (frames in the last row) and the behaviour is changed again to the default one. The robot employs about 55 seconds in executing all the tasks.

[5]SMPTE (Society of Motion Picture and Television Engineers) Timecode is a standard to label individual frames of video in the form *hours*:*minutes*:*seconds*:*frame*. The standard is defined in the SMPTE 12M specification.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

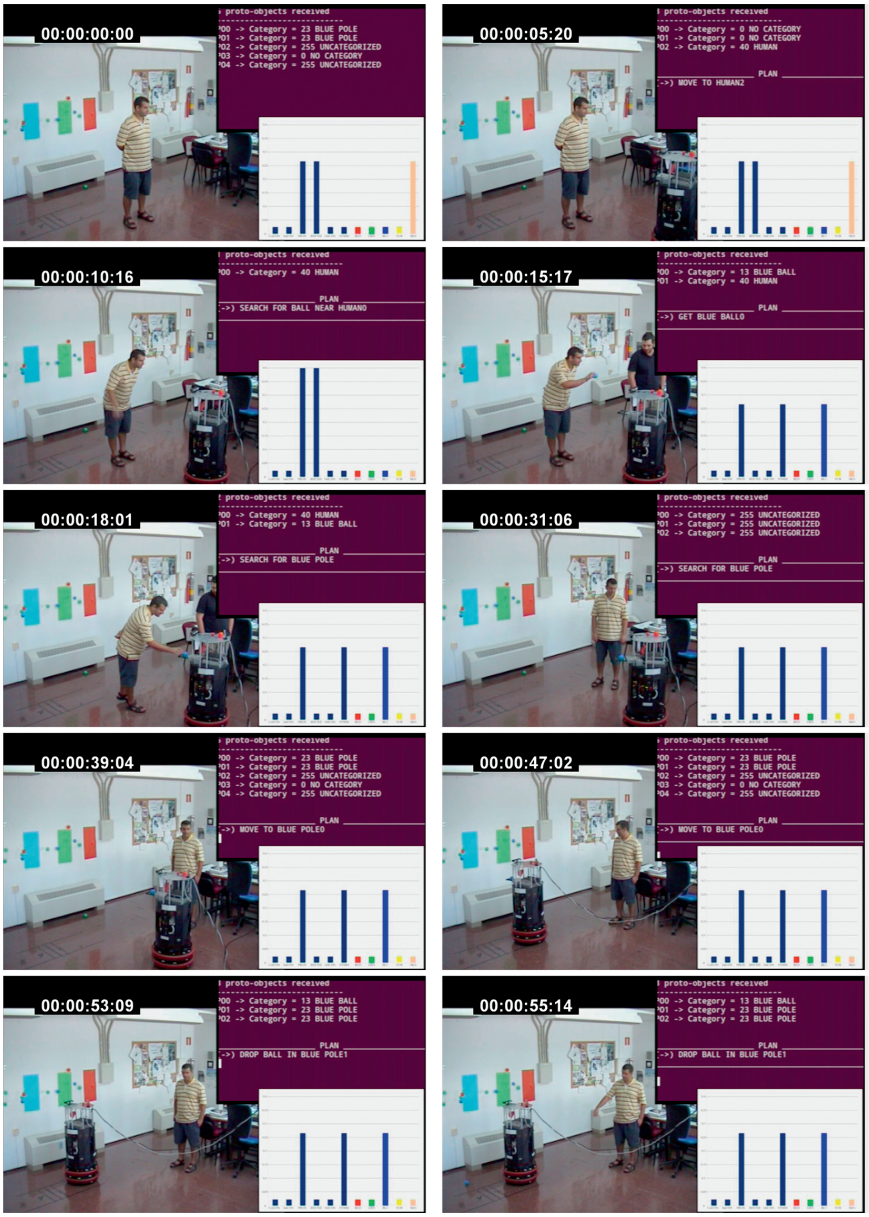


Figure 6.17: Some significant frames from an execution of experiment 2 (25 fps; SMPTE Timecode from video recorder is displayed). The output of the *Deliberative* component is also included, showing the categorised content of the WM, the plan provided by the Panning Framework and the attentional set to be stored in the PMM.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

This experiment has been repeated using different colours, with people located at different places. Some of these experiments are shown in fig. 6.18. Only the video stream provided by an external camcorder is displayed in order to simplify the figure. It can be noted that the robot is able to accomplish their tasks regardless the initial configuration or the external conditions.

Due to the continuous adaptability of the cognitive architecture, the robot is not likely to fall into inaction or to accomplish a wrong behaviour. Nevertheless, a successful execution is not always achieved. This is mainly caused by a wrong categorisation of the perceived proto-objects. For example, the face detector could not categorise a skin colour region as a face. This implies that the robot still continues looking for a person despite there is really someone in front of it. Another frequent errors are related to a bad colour assignation to the collected ball (often caused by people's clothes) or to a wrong identification of a goal marker (e.g., the "red container" category can be easily assigned to a red extinguisher). Some erroneous executions can be seen in fig. 6.19. These errors are mainly restricted to the categorizer module. Therefore, they could easily be overcome in future improvements of the system.

On the contrary, the inherent robustness of the architecture is remarkable. For instance, occlusions and lost items are handled in a natural way. If a target (e.g. a "container" marker) is lost or occluded during navigation, the needs of the *move towards container* task can not be completely fulfilled. As a consequence, the *search for container* task becomes executable and the target is sought again. Besides, attention is not drawn to elements belonging to other behaviours. If the robot is in the *recycling* behaviour, people walking or standing nearby and distracting balls are ignored by the perception system.

Some statistical results are shown in figures 6.20, 6.21 and 6.22. A total of **20 executions** of the experiment have been carried out in a similar scenario. The robot is supposed to accomplish the following tasks:

- Look for a person.
- Pick up the ball that the person is offering.
- Search for the corresponding marker of container.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGA



Figure 6.18: Several executions of experiment 2 (25 fps; SMPTE Timecode from video recorder is displayed).

Publicaciones y Divulgación Científica
UMA UNIVERSIDAD DE MÁLAGA



(a) (b)

Figure 6.19: Erroneous executions of experiment 2. (a) an element is wrongly identified as a blue "container"; (b) blue colour is erroneously assigned to a green ball.

- Drop the ball.

The execution finishes when the robot drops the ball. The initial position of the human and the colours involved vary throughout the different evaluations.

Fig. 6.20 shows the number of perception-reasoning-action iterations needed in average to perform the evaluation scenario. Concretely, the architecture needs **130 iterations** on average to completely solve the problem. Notice the high value obtained for the standard deviation due to the diversity of people's initial location and markers' position. The execution time has been also measured: the robot employs about **53 seconds** on average. This produces a ratio of about **2.4** $\frac{iterations}{second}$. Hence, the architecture is fast enough to react to behavioural changes.

The amount of activations of each task during a complete evaluation has been also measured (fig. 6.21). It can be noted that tasks involving search are activated considerably more times than the others. This behaviour is completely reasonable since these tasks have less needs to be satisfied and they are performed as default tasks when no other possible action is available. The case of *search for human* task is specially significant due to the absence of people in the scene at the beginning of some evaluations. On the contrary, tasks concerning ball getting and dropping are the less overall executed because these tasks require very specific conditions to be executed.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

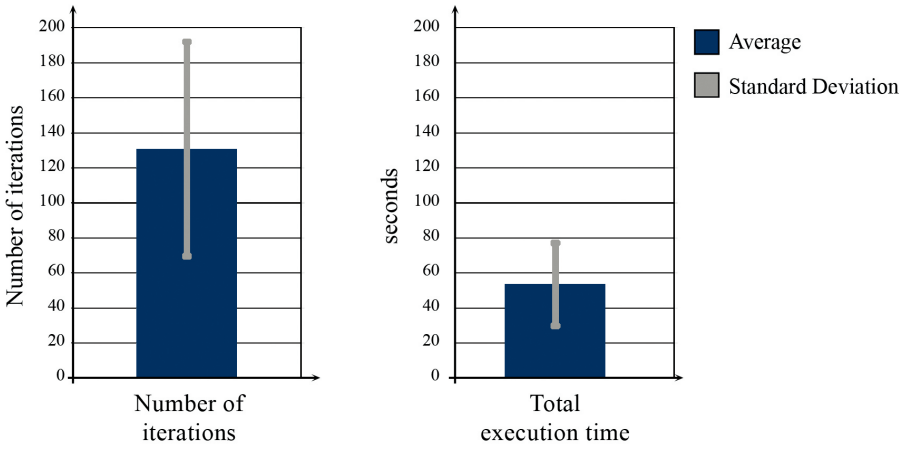


Figure 6.20: Number of iterations and execution time needed on average to solve the whole problem. The standard deviation is also included in the graph.

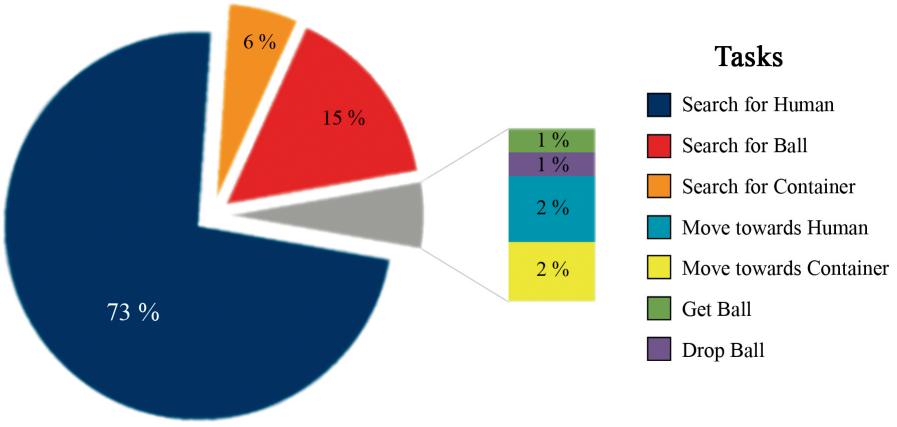


Figure 6.21: Activation distribution among the different tasks. The graph shows the averaged amount of times that a specific task is active during a whole execution.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

Finally, the success rate of the cognitive architecture for the current experiment is shown in fig. 6.22. As it was aforementioned, most errors are produced by an erroneous categorisation or by wrong target identification. Nevertheless, a **85%** of the evaluations were successfully accomplished.

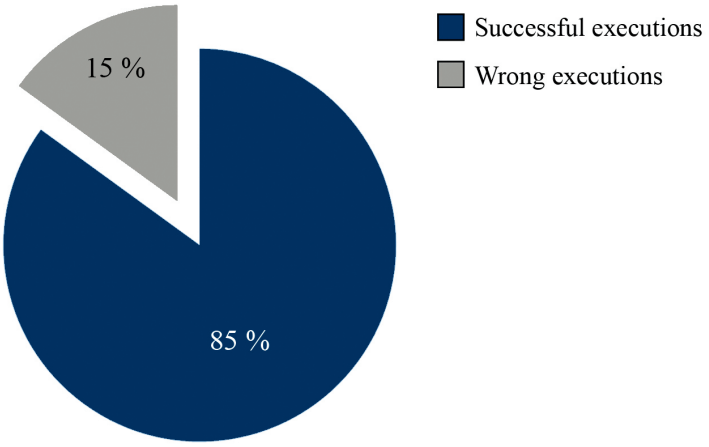


Figure 6.22: Success rate for the social robot classifying objects experiment.

6.3.3 Discussion

Based on qualitative aspects of human behaviour, Langley et al. [2009] propose a set of criteria in order to evaluate a cognitive architecture:

- *Generality*, *versatility* and *taskability*. These three criteria are related to the possibility of using the same architecture to address different tasks and problems without special development effort. The more environments in which the architecture supports intelligent behaviour, the greater its *generality*. The *versatility* is measured in terms of the difficulty encountered in constructing intelligent systems given a set of tasks and environments. Finally, *taskability* refers to the ability of the architecture to carry out different tasks in response to goals or other external commands. The more tasks an architecture can perform in response to such commands, the greater its *taskability*.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

- *Rationality* and *optimality*. *Rationality* refers to the ability of a system to select an action that accomplishes a current goal. *Optimality* measures the cost of a decision, in a classical Automated Planning sense (optimization of a cost function).
- *Efficiency* and *scalability*. *Efficiency* can be measured at the level of complete tasks in real-time or space-constrained environments. *Scalability* is related to the ability of an architecture to maintain a similar efficiency when addressing problems with a wide range of difficulty.
- *Reactivity* and *persistence*. The *reactivity* refers to the speed in responding to unexpected situations or events. On the other hand, *persistence* measures the capacity to remain focused on a long-term task despite small environmental changes.
- *Autonomy* and *extended operation*. These concepts measure the ability of an agent to avoid falling into inaction or requesting for external assistance.

Although this thesis is mainly focused on studying the bidirectional loop closing problem, not only the reliability of linking attention-based perception and abstract reasoning but also some of the aforementioned criteria are taken into account in order to obtain a qualitative evaluation of the proposed architecture. Specifically, the performance related to generality, versatility, rationality, efficiency, reactivity and persistence are discussed in this section. Optimality and scalability are not considered here because they are more closely associated with pure Planning Framework developments.

From the loop closing point of view, both experiments prove that the proposed cognitive architecture provides a **real and effective integration** between attention-based perception and abstract reasoning. In the first experiment (coloured cards), perceptual guidance is able to enhance the overall performance of the system. However, it is also shown that an erroneous guidance strategy may cause a considerable reduction of efficiency. The second experiment (recycling robot) shows the ability of the proposal to discriminate among different kind of information depending on time-varying tasks and behaviours. Thus, complex environments, where the computational cost associated with a full scene processing is unaffordable, can be addressed by highlighting simple, specific and suitable features, reducing the received amount of information.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

The **bidirectionality** of the proposal has also been demonstrated, specially in the second experiment. Not only the perception system is guided in terms of the ongoing task, but also the presence of significant objects triggers awaiting tasks or even changes the current behaviour of the robot (e.g., the detection of a person moves the robot to the assisting behaviour).

Concerning the *generality* of the proposed cognitive architecture, only two different environments can not be considered enough to perform a fair evaluation. Nevertheless, two quite different testing scenarios have been addressed using the same architecture, introducing only minor modifications. In fact, the system has just been simplified in the first experiment for a more understandable explanation of the results. Therefore, the same architecture has been successfully used for a diverse set of tasks and environments.

The modular configuration of the proposed need-based architecture provides a high degree of *versatility* by definition. The proposal can be adapted to different uses by defining the involved tasks (including their perceptual needs and effects) as well as providing the new domain definition in PDDL. In the particular case of the two experiments carried out in this chapter, all tasks have been individually defined. The domain information provided to the Planning Framework at the knowledge-based level is responsible for detecting the context information and selecting the corresponding tasks to be placed at the rule-based level (*stack G over F* in putting cards in order behaviour, *move towards human* in assisting behaviour or *drop ball* in recycling behaviour). In summary, task modularisation makes easier the development of new behaviours because interdependencies should not be taken into account at the task level. They are handle as abstract predicates at the knowledge-based level.

Rationality is also inherent to the proposed architecture definition. Only tasks that are supposed to accomplish behavioural useful goals are placed at the rule-based level. Besides, only those tasks getting all their needs fulfilled can be activated. Hence, only actions that accomplish goals can be selected.

Regarding *efficiency*, it has been measured in terms of perception-action iterations needed to solve a whole problem for both experiments (see Table

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

6.3 on page 124 and figures 6.14 and 6.20 on pages 124 and 133, respectively). In experiment 2 (assisting robot), the success rate is also given as an efficiency measure (fig. 6.22 on page 134).

The *persistence* of the architecture is easily observable in the assisting robot experiment. For instance, when the robot is approaching a "container", people are only taken into account as obstacles to be avoided instead of assisting them. That is, variations in the scene, that are not expected to change the ongoing behaviour, do not alter the current activity to be accomplished.

Finally, *reactivity* is intimately related to the rule-based level. Unexpected categories may be handled by a high priority task with only one need to be satisfied. Although there is no such a task among the ones presented in the evaluation experiments, its functionality may be similar to any standard task. Thereby, the system could instantaneously react to any situation.

In conclusion, the evaluation experiments reveal that the proposed cognitive architecture is able to correctly perform a bidirectional perception-reasoning-action loop closing integration. The architecture efficiently addresses time- and behaviour-varying environments while maintaining satisfactory qualities of persistence, reactivity and rationality. The architecture also promises to be general and versatile despite the insufficient range and variety of experiments. The proposal provides a modular model that facilitates a systematic and effortless development of new tasks and behaviours for further environments.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

UNIVERSIDAD DE MÁLAGA
Publicaciones y Divulgación Científica

7 Conclusions and Future Work

"There is joy in work. There is no happiness except in the realization that we have accomplished something"

Henry Ford

7.1 Conclusions

Attention-based perception and abstract reasoning have usually been investigated as separated problems. Computational attention models normally lack in explaining the mechanisms that derive task influence from a deliberative layer. On the contrary, solutions proposed from Cognitive Architectures and Automated Planning typically consider a quite simplified vision system, which is mainly used to increase observability. Therefore, the perception-reasoning-action loop is assumed to be unidirectional.

The work that has been proposed in this thesis bridges this gap, considering a complete and bidirectional characterisation of the loop closing problem. Both the attention model and the designed cognitive architecture pursue a double objective. On the one hand, the most relevant elements in the scene are selected, taking into consideration not only their intrinsic features but also the constraints provided by the ongoing behaviour and context. On the other hand, perceived items determine the tasks that

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

can be executed at each moment. Thereby, the tasks that better fit the perceived environment are more likely to be executed.

The first main contribution of this monograph is the attention-based early-vision system. The **Object-based Visual Attention System** (OBVIAS) is an attention model that integrates bottom-up processing of several features and a template-based top-down selection. In OBVIAS, attention is drawn to proto-objects, units of visual information that can be bound into coherent wholes. These proto-objects are the result of a perceptual grouping algorithm, performed using a hierarchical image structure.

The relevance of each proto-object is obtained as the weighted sum of different features describing colour, shape and location. The selection of these features has been proven to facilitate an effective attention guidance. The most relevant proto-objects are stored in a Working Memory able to keep their positions updated in a dynamic scenario. An Inhibition of Return (IOR) mechanism is responsible for avoiding selecting an already attended proto-object. Elapsed time is also considered when recomputing the saliency of the elements in the WM, providing a more plausible and realistic inhibition procedure. The top-down influence is tackled by modifying the weights that define the saliency computation. These perception parameters (attentional sets) are collected from a Perception Modulation Memory (PMM) and they skew the features considered in saliency computation.

The second main contribution of this thesis is a **need-based cognitive architecture** that extends the capabilities of OBVIAS model to effectively connect an attention-based perception system with an abstract reasoning layer. The architecture is divided into two levels, following a human performance model. On the one hand, multiple tasks are placed at the rule-based level. The execution of a task is subordinated to the presence of specific elements in the scene (needs to be covered). In addition, executed tasks also influence the attention system in order to guide perception to task-relevant elements.

On the other hand, the knowledge-based level is responsible for adding, removing and prioritising the tasks placed at the prior level, using a Planning Framework that can handle partial observation. An Oversubscription

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

Planner addresses the decision making and goal definition. This planner is able to build partial plans, constrained by the perceived information. Furthermore, other modules can be located at this level in order to figure out the scene context and select the most suitable set of tasks and their priorities.

The bidirectional loop closing is achieved through the WM and PMM. The stored proto-objects are translated into abstract predicates (categories) that can be handled by the Planning Framework. Then, perceived categories can trigger latent tasks at the rule-based level (perception-to-action direction). In order to perform the other direction of the loop (action-to-perception), an interpreter translates the effects for attention guidance associated with each task into a new attentional set.

In comparison with other existing models, it has been shown that the proposed architecture overcomes their current drawbacks. The top-down component of an attention system has definitely been connected with a functional deliberative module whereas the cognitive architecture is able to execute appropriate tasks depending on the perceived elements.

Finally, the architecture has successfully been implemented on a real robot, using RoboComp, an open-source and component-based robotics framework. The use of RoboComp increases the robustness and efficiency of the system and eases further developments and improvements. The reliability of the proposal has been evaluated through several experiments. As a result, not only the performance of the attention system has been demonstrated, but also the capacity of the cognitive architecture to tackle dynamic and behaviour-varying scenarios. The cognitive architecture promises to be general and versatile and presents satisfactory values of persistence, reactivity and rationality.

7.2 Future Work

The cognitive architecture presented in this thesis aims to be an open, general, versatile and scalable integration framework, not only for new developments but also for already existing models in different scientific fields.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

Therefore, the architecture could be easily improved by replacing some specific modules while maintaining the overall structure and interfaces.

Regarding the visual attention system (OBVIAS), the future work could be oriented to add new features in order to better characterise bottom-up saliency computation and attention guidance. We consider that features related to motion [Tünnermann and Mertsching, 2014; Belardinelli et al., 2010] are specially interesting since movement is an undoubted source of attention guidance (cf. table 4.1 on page 56). Nevertheless, it is also important to keep an acceptable computational load when adding or replacing features.

Although OBVIAS model roughly deals with time-varying saliency computation, a promising direction for future research would point towards automatically inferring the inhibition time. Thus, attention and inhibition times would depend on the features of an object and the ongoing tasks, as it can be observed in psychological studies [Klein, 2000; Posner et al., 1985].

As it was pointed out in section 2.1 on page 11, there exists a close relationship between attention and the foveal nature of the human vision system. Hence, a natural evolution of the model consists in replacing the current segmentation algorithm with a foveal one. Recently, the Bounded Irregular Pyramid (BIP), employed in the perceptual grouping procedure, has been adapted to deal with foveal information [Marfil et al., 2013].

Concerning other components of the cognitive architecture, we consider that one of the most promising lines of research is related to the *interpreter*. Some approaches have attempted to solve this problem by using Bayesian inference [Yu et al., 2012]. However, a general solution from Automated Planning is needed in order to automatically translate abstract planning actions into a new attentional set. For instance, Fuzzy Logic has played a similar role when applying abstract rules to control applications. A similar approach have been developed by Windridge et al. [2013].

Future work should also be directed to complete the *knowledge-based level* of the architecture. In this case, there exist many works on different disciplines (Machine Learning, Human-Robot Interaction, Ontologies, Emotional Affordances...) that could be placed at this level. Inductive

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

learning of new rules and behaviours also seems to be an interesting line of research, following approaches similar to the one proposed by Windridge and Kittler [2010]. Thus, the architecture also aims to facilitate to other research scientists the integration of their isolated works into a complex and multidisciplinary agent.

Another component at the *knowledge-based level* that should be addressed is a long-term memory. This memory should allow the robot to "keep in mind" recently observed elements regardless a momentarily disappearance from the scene. This would permit the robot to have an abstract model of the environment, enhancing visual search and scene context recognition. In this context, it is worth mentioning the solution provided by Ferreira et al. [2013], based on a probabilistic and egocentric framework. Although their work is currently focused on a short-term approach, we think that it could be a suitable starting point for a long-term extension.

In conclusion, the architecture presented in this thesis is intended to effortlessly support present and further developments about perception, action and cognition in social robotics from many different fields of research. Nowadays, the realisation of a real robot such as *Wall-E* is still a huge challenge. However, scientists are working day by day to finally develop completely autonomous robotic agents. And, perhaps, they will dream of electric sheep.

7.3 Publications

The main findings and continuous developments of the work introduced in this thesis have been published in different journals and communicated in international conferences. The related publications are referred below:

Journals:

- Palomino, A. J., Marfil, R., Bandera, J. P. and Bandera, A. (2011). A novel biologically inspired attention mechanism for a social robot. *EURASIP Journal on Advances in Signal Processing*. 10 pages. DOI: 1155/2011/841078.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

- Antúnez, E., Palomino, A. J., Marfil, R. and Bandera, J. P. (2013). Perceptual organization and artificial attention for visual landmarks detection. *Cognitive Processing*, 14(1): 13–18.

- Bandera, J. P. , Marfil, R., Palomino, A. J., Vázquez-Martín R. and Bandera, A. (2012). Visual attention mechanism for a social robot. *Applied Bionics and Biomechanics*, 9:409–425. DOI: 10.3233/ABB-2011-0030

Conference Proceedings:

- Palomino A. J., García-Olaya, A. , Fernández F. and Bandera J. P. (2013). From perception to action and vice versa: A new architecture showing how perception and action can modulate each other simultaneously. In *Proceedings of the 6th European Conference on Mobile Robots (ECMR 2013)*, pages 268–273, Barcelona (Spain).

- Palomino, A. J., Marfil, R., Bandera, J. P. and Bandera, A. (2013). Multi-feature bottom-up processing and top-down selection for an object-based visual attention model. In *Proceedings of the 2nd Workshop on Recognition and Action for Scene Understanding (REACTS 2013)*, pages 29–43, York (UK).

- Palomino A. J., Marfil, R., del Toro, J. C., Bandera, A. and Sandoval, F. (2009). An Object-based Attention Mechanism applied to a New Social Robot. In *Proceedings of X Workshop on Physical Agents (WAF 2009)*, pages 79–85, Cáceres (Spain).

- Cid, F., Palomino A. J. and Núñez, P. (2013). A New Paradigm for Learning Affective Behaviors: Emotional Affordances in Human Robot Interaction. In *Proceedings of the XIV Workshop on Physical Agents (WAF 2013)*, pages 47–52, Madrid (Spain).

- Bandera, J.P., Marfil, R., Palomino A. J., Bandera A. and Vázquez-Martín, R. (2010). Visual Perception System for a Social Robot. In *Proceedings of the 2010 IEEE Conference on Robotics, Automation and Mechatronics (RAM 2010)*, pages 243–249, Singapore.

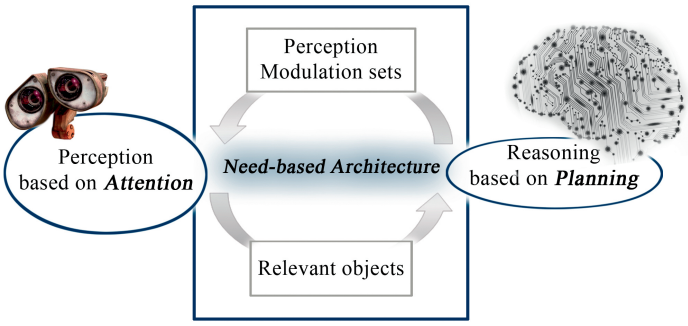
Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

Ñ

Resumen en español

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

UNIVERSIDAD DE MÁLAGA
Publicaciones y Divulgación Científica



Arquitectura Cognitiva para un cierre de lazo bidireccional basado en Atención (CABILDO)

Introducción

Cuando se sitúa a un robot autónomo en el mundo real, uno de los mayores inconvenientes con los que se encuentra radica en la enorme cantidad de información visual disponible. Además, el robot debe ser capaz de llevar a cabo diferentes acciones y tareas que se suceden en el tiempo, teniendo también que reaccionar ante situaciones inesperadas. Cuando se intenta desarrollar un sistema de percepción para un robot de tales características, surgen algunas preguntas fundamentales: ¿es posible modificar el modo en el que el robot percibe el mundo en función de sus responsabilidades? Y, viceversa, ¿la aparición de nuevos objetos en escena puede modificar la acción en curso? ¿Cómo pueden la percepción y el razonamiento abstracto interactuar simultáneamente en un robot autónomo?

En este sentido, el sistema de visión humano presenta una serie de características de adaptabilidad y robustez que le confieren una gran eficiencia a la hora de procesar la información visual presente en una escena. En gran medida, esta alta eficiencia se debe a la existencia de un *Mecanismo de Atención Visual*, gracias al cual la información irrelevante de la escena es automáticamente desechada, centrándose toda la

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

capacidad de proceso únicamente sobre aquellos elementos que realmente son importantes para la tarea que se está realizando [Duncan, 1984].

Desde hace algunos años, se ha incrementado el interés de los investigadores en aprovechar esta ventaja del sistema visual humano para aplicarla a los sistemas de visión artificial presentes en los robots. Esta estrategia resulta especialmente interesante cuando se desarrollan robots sociales, es decir, agentes autónomos capaces de desenvolverse en un ambiente heterogéneo, tanto de personas como de otros robots con los que puede interaccionar [Dautenhahn and Billard, 1999]. Es más, este tipo de robots suelen situarse de forma desatendida en entornos complejos donde deben establecer de forma autónoma cuándo y qué elementos deben ser atendidos y qué debe hacerse en respuesta a dicha información visual recibida [Paletta et al., 2005].

Analizando el problema desde un punto de vista deliberativo, las acciones y tareas que debe desarrollar un robot dependen en gran medida de la percepción de un conjunto específico de elementos. Debido a esto, un planificador deliberativo sólo podrá contar con una observación parcial del entorno, ya que ésta se encuentra limitada por el sistema de atención visual. Simultáneamente, la ejecución de una tarea determinada requiere que la atención recaiga sobre objetos específicos. Es decir, existe una relación muy estrecha entre un sistema de percepción basado en atención y la etapa de planificación que habitualmente se incluye en la fase de razonamiento del bucle percepción-razonamiento-acción.

Esta conexión entre percepción, razonamiento y acción, especialmente cuando se emplea un sistema de atención, supone una cuestión no resuelta en el ámbito científico. De hecho, ha sido abordada frecuentemente de una forma unidireccional, relegando las funciones del sistema de visión a una simple búsqueda guiada de objetos. No obstante, los resultados obtenidos de estudios fisiológicos sugieren que ciertas características relativas a la percepción visual, como la localización o la forma, también pueden provocar la activación de tareas y comportamientos relacionados con dichas características. Por consiguiente, el procesado de los estímulos relacionados con estas tareas debe ser prioritario [Collins et al., 2010; Fagioli et al., 2007]. De este modo, la clásica suposición unilateral del bucle percepción-razonamiento no es suficiente. Una solución completa

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGUNA

debe cubrir la doble imbricación entre percepción y razonamiento, salvando las desventajas de los modelos previos.

Por ello, esta tesis presenta una **nueva arquitectura cognitiva para un cierre de lazo bidireccional y basado en atención del bucle percepción-razonamiento-acción**. La solución propuesta genera una simbiosis efectiva entre los módulos de atención visual y razonamiento abstracto de un sistema robótico, conectando características visuales con comportamientos de alto nivel. A través del intercambio de *elementos relevantes* y *conjuntos de parámetros de percepción*, el sistema cognitivo es capaz de decidir qué tareas se ejecutarán, siguiendo una estrategia de *necesidades satisfechas*. De forma simultánea, la tarea predominante selecciona qué tipo de elementos serán relevantes en el siguiente ciclo de forma iterativa.

La otra contribución importante de este trabajo radica en el propio **mecanismo de atención basado en objetos** utilizado como sistema perceptivo. El modelo integra tanto un procesado *bottom-up* (de abajo arriba), independiente de la tarea, como una selección *top-down* (de arriba abajo), que depende de la tarea en curso. El sistema divide la imagen de entrada en entidades visuales básicas aplicando un algoritmo perceptivo. Para calcular la relevancia de cada entidad, el modelo usa tanto sus características básicas inherentes (color, localización, forma...) como plantillas de parámetros de modulación de la percepción proporcionadas por la capa deliberativa.

Fundamentos Biológicos y Trabajo Relacionado

El término **Atención** define la habilidad mental para seleccionar estímulos, respuestas, recuerdos o pensamientos que son relevantes para el comportamiento actual entre otros muchos que carecen completamente de importancia [Corbetta, 1998]. Debido a la naturaleza foveal de la visión humana, no todo el campo visual se procesa con la misma resolución. Esto, que a priori podría resultar una desventaja, ha permitido el desarrollo de un proceso de atención que permite obviar aquellas partes de la escena que aportan poca información y centrar el área de máxima

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

resolución del ojo sobre la región realmente relevante de la imagen. Ello permite un uso realmente eficiente de los recursos visuales.

El fenómeno de la atención visual ha sido ampliamente estudiado en los ámbitos de la Fisiología y la Psicología desde la década de 1950, proporcionando distintas teorías y modelos. Así pues, estos modelos tratan de explicar tanto el proceso de atención completo como otros aspectos específicos relacionados, como pueden ser la *Inhibición de Retorno* o las unidades básicas de atención. Posteriormente, este conocimiento ha sido trasladado a la disciplina de la Visión Artificial, con el fin de perfeccionar y mejorar los algoritmos existentes.

Sin duda, una de las teorías que más ha influenciado el trabajo en atención visual artificial ha sido la *Teoría de Integración de Características* [Treisman and Gelade, 1980]. De acuerdo con este modelo teórico, el proceso de atención se puede dividir en dos pasos. En concreto, existe un primer proceso de integración de características independiente de la tarea de tipo *bottom-up* (de abajo hacia arriba) y un proceso de selección posterior, denominado por Treisman búsqueda de conjunción de características, de tipo *top-down* (de arriba hacia abajo) que sí puede depender de un comportamiento determinado.

Algunos años después de la publicación del modelo de Treisman, Wolfe introdujo algunas modificaciones a la Teoría de Integración de Características en su *Búsqueda Guiada* [Wolfe et al., 1989; Wolfe, 1994]. Si bien Treisman se centraba en el componente *bottom-up* de la atención, Wolfe incide en que la parte *top-down* puede usarse para acelerar el proceso al identificar la presencia de un objetivo en la escena. Es decir, el proceso de extracción de características de bajo nivel puede ser guiado por una tarea de alto nivel, por lo que sólo es necesario buscar unas determinadas características.

Otro aspecto común a ambas teorías reside en la necesidad de un mecanismo que evite que el foco de atención vuelva a recaer sobre una zona que ha sido recientemente atendida. Este proceso se conoce como *Inhibición de Retorno* e impide que la atención recaiga siempre sobre el mismo objeto [Posner et al., 1985; Klein, 2000].

Por último, otros modelos establecen el número máximo de elementos que se pueden atender a la vez. Entre ellos, se encuentra la *Teoría de*

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGUNA

Atención Visual de Bundesen [Bundesen and Habekost, 2008]. Esta teoría postula la existencia de una memoria a corto plazo en la que se almacenan los objetos que han sido recientemente atendidos. Esta memoria tiene una capacidad limitada y fija, normalmente reducida a entre 3 y 5 elementos.

Tomando como base estos estudios, se han ido desarrollando distintos modelos computacionales de atención visual durante los últimos 30 años. Por ejemplo, los modelos desarrollados por Koch and Ullman [1985], Itti et al. [1998], Backer and Mertsching [2003], Navalpakkam and Itti [2006], Frintrop [2006] o Kouchaki and Nasrabadi [2012]. En contraposición a los modelos anteriores, basados en regiones, que computan la relevancia a nivel de píxel, los modelos *basados en objetos* ponen de manifiesto el hecho de que las habilidades perceptivas deben optimizarse para interaccionar con conjuntos coherentes de píxeles y no con meras regiones espaciales desestructuradas [Duncan, 1984]. Modelos como los de Sun and Fisher [2003], Wischnewski et al. [2010], Gao and Vasconcelos [2004] o Yu et al. [2013] siguen esta filosofía.

A pesar de que la mayoría de estos modelos considera en mayor o menor medida la influencia *top-down* a la hora de calcular la relevancia de una región o un objeto, normalmente no explican de forma explícita cómo se deriva dicha influencia a partir de una capa deliberativa de alto nivel. Por tanto, los modelos de atención se presentan como elementos aislados, careciendo de conexión real con niveles de razonamiento abstracto. No obstante, existen evidencias neurofisiológicas sobre cómo la planificación abstracta puede modular la información percibida. Para conseguirlo, los procesos deliberativos proporcionan un grupo de parámetros de percepción que afectan a la relevancia de algunas características específicas, con el fin de resaltar objetos relacionados con la tarea actual. Este grupo de parámetros recibe el nombre de *Conjunto Atencional* (*Attentional Set* [Corbetta and Shulman, 2002]).

Debido a que el problema del cierre de lazo no es exclusivo de la Visión Artificial, otras áreas de investigación, como la Inteligencia Artificial, también han propuesto diversas soluciones desde un punto de vista diferente. Por ejemplo, las *Arquitecturas Cognitivas* tratan de caracterizar completamente un sistema inteligente, incorporando los

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

procesos funcionales y representaciones de elementos necesarios, además de la observación del entorno.

En sistemas como ICARUS [Langley et al., 2004] y Soar [Laird et al., 1987], por ejemplo, el sistema de percepción se emplea principalmente para incrementar el conocimiento acerca de la escena y para determinar qué tareas se pueden llevan a cabo en cada momento. En cambio, sistemas como ACT-R [Anderson and Lebiere, 1998] son capaces de inferir una serie de restricciones visuales en forma de características a resaltar (o evitar). Sin embargo, su aplicación a un sistema de atención visual se detalla de manera muy somera. El sistema HERMES [Bischoff and Graefe, 2004] quizás es el que más se aproxima a la idea de bidireccionalidad planteado en este trabajo. No obstante, se trata de una solución bastante monolítica y, por tanto, difícil de generalizar, aparte de no aprovechar las ventajas que proporciona un sistema de atención.

Por otra parte, también existen modelos que proponen soluciones derivadas de la planificación jerárquica o la inferencia bayesiana. Modelos como los propuestos por Sridharan et al. [2010], Gilet et al. [2011] o Eidenberger et al. [2009] presentan una fuerte unidireccionalidad. En este caso, se emplean modelos de visión muy sencillos con el único objetivo de incrementar la observabilidad del proceso de decisión y control, reduciendo la incertidumbre. Por ello, el guiado visual queda normalmente restringido a un simple movimiento mecánico que cambie el punto de vista desde el que se observa la escena. Además, suelen ser sistemas limitados a actividades y escenarios muy específicos, aunque existen modelos más elaborados y de propósito general como el desarrollado por Ferreira et al. [2013]. Sin embargo, éste último también adolece de un guiado visual complejo que vaya más allá de un mero cambio de punto de vista mecánico.

En resumen, parece claro que el problema de la bidireccionalidad del cierre del bucle percepción-razonamiento no ha sido abordado de una forma satisfactoria hasta el momento. En consecuencia, resulta necesaria una solución que dote a un mecanismo de atención de una conexión real con módulos deliberativos complejos y defina una arquitectura cognitiva que sea capaz tanto de adaptarse a los objetos percibidos como de guiar al sistema de atención subyacente.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

Sistema de Atención Visual basado en objetos

En la arquitectura cognitiva propuesta en esta tesis, el **Sistema de Atención Visual Basado en Objetos (OBVIAS)** es el responsable del procesado visual de la escena, proporcionando las ventajas inherentes a los sistemas de atención visual: la reducción del coste computacional (al tener que procesar únicamente una parte de la escena) y la habilidad para seleccionar sólo los elementos que son realmente relevantes para la actividad en curso.

OBVIAS es un modelo que incluye tanto un procesado bottom-up como una selección top-down. La unidad básica de atención empleada es el *proto-objeto* [Rensink, 2000]. Los proto-objetos se definen como unidades de información visual que pueden corresponderse con objetos coherentes y estables. En otras palabras, se pueden definir como aquellas regiones de la imagen que pueden constituir potencialmente objetos reales.

El sistema está articulado en dos etapas. La etapa de *pre-atención* se encarga de obtener los proto-objetos, aplicando un algoritmo de segmentación perceptiva, y de calcular su relevancia en función de una serie de características de bajo nivel. Por su parte, la etapa de *semi-atención* es la responsable de la Inhibición de Retorno, la selección top-down y la gestión de los proto-objetos más relevantes.

Teniendo en cuenta que un sistema de atención actúa como una etapa de visión temprana, resulta imprescindible que cumpla unos requerimientos de inmediatez y eficiencia. Por ello, el algoritmo de segmentación perceptiva utiliza una estructura jerárquica de demostrada eficiencia computacional: la Pirámide Irregular Acotada (BIP, *Bounded Irregular Pyramid* [Marfil and Bandera, 2009; Marfil et al., 2009]. Dicho algoritmo combina un primer paso de pre-segmentación por color con un agrupamiento perceptivo posterior.

El siguiente paso dentro de la etapa de pre-atención corresponde al cálculo de las características básicas que describen a cada proto-objeto, con la finalidad de calcular su relevancia. De acuerdo con distintos estudios psicológicos, algunas características, como el color, el movimiento o la orientación, claramente determinan la relevancia *a priori* de un

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

objeto. Estas tres características, junto con el tamaño, son catalogadas por Wolfe and Horowitz [2004] como atributos que, indudablemente, guían la atención. Estos autores ofrecen en su trabajo una clasificación bastante exhaustiva de las características que podrían guiar la atención, dividiéndolas según su capacidad para actuar realmente como guías de la atención. Aunque la elección de un gran número de características proporcionaría una descripción más completa sobre los elementos en la escena, el coste computacional asociado podría llegar a ser inaceptable. Por consiguiente, resulta imprescindible alcanzar una solución de compromiso entre la eficiencia de cómputo y el número y tipo de las características seleccionadas para guiar la atención.

Siguiendo las directrices anteriores y teniendo en cuenta que el sistema está pensado para usarse en robots sociales, se ha seleccionado un conjunto compuesto por once características para calcular la relevancia de cada proto-objeto en el sistema OBVIAS. Dichas características son el contraste de color ($ColCON$), el contraste de intensidad ($IntCON$), la proximidad al sensor ($PROX$), la redondez del elemento ($ROUND$), el contraste de orientación entre proto-objetos ($OriCON$), la simetría ($SYMM$), similitud con el color piel (SKN) y similitud con colores básicos (rojo –RED–, verde –GRN–, azul –BLU– y amarillo –YLW–).

El valor final de relevancia, SAL_i, para cada proto-objeto, $\mathcal{P}_i$, se obtiene como la suma ponderada de todas sus características:

$$\begin{aligned} SAL_i \quad &= \lambda_1 \cdot ColCON + \lambda_2 \cdot IntCON + \lambda_3 \cdot PROX + \lambda_4 \cdot ROUND \\ &+\lambda_5 \cdot OriCON + \lambda_6 \cdot SYMM + \lambda_7 \cdot RED + \lambda_8 \cdot BLU \\ &+\lambda_9 \cdot GRN + \lambda_{10} \cdot YLW + \lambda_{11} \cdot SKN \end{aligned}$$

En forma compacta, la ecuación anterior se puede expresar como $SAL_i = \vec{\lambda} \cdot \vec{f}$, donde $\vec{\lambda}$ es el **Conjunto Atencional** almacenado en la Memoria de Modulación de Percepción (PMM), verificándose $\sum_i \lambda_i = 1$. Una vez obtenida la relevancia de cada proto-objeto, los cinco más relevantes se almacenan en una Memoria de Trabajo (WM), pasando el sistema a la etapa de semi-atención.

Una de las tareas fundamentales de la etapa de semi-atención se corresponde con la Inhibición de Retorno (IOR). Para poder aplicar la

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGUNA

IOR a escenarios dinámicos, resulta imprescindible establecer una correspondencia entre regiones a través de fotogramas sucesivos. En OBVIAS, esta correspondencia se obtiene desarrollando una IOR orientada a objetos y basada en seguimiento de regiones en secuencias de imágenes. Cuando se modifica el punto de vista, bien por desplazamientos de elementos que componen la escena o bien por movimientos del propio sistema de visión, los proto-objetos almacenados en la WM se mantienen en seguimiento. De esta forma, en el siguiente fotograma se obtiene un nuevo conjunto de proto-objetos y las posiciones de los anteriormente calculados se actualizan.

A continuación, del conjunto de nuevos proto-objetos, aquellos que ocupan las mismas posiciones que los que ya están siendo atendidos son descartados y no compiten para rellenar la WM. De esta manera, resulta poco probable que se duplique o se vuelvan a atender proto-objetos que ya existían en la memoria de trabajo. El proceso de seguimiento de imágenes está basado en la técnica *mean-shift*, propuesta por Comaniciu et al. [2003]. Este algoritmo es especialmente interesante debido a su simplicidad y su eficiencia, lo que permite mantener el seguimiento simultáneo de múltiples regiones dentro de una escena.

En cuanto a la integración de los componentes top-down y bottom-up, suelen usarse principalmente dos tipos de estrategia: la basada en *fusión de mapas de características* y la basada en *plantillas*. En la primera se calculan unos mapas de relevancia adicionales en función de un conocimiento de alto nivel previamente aprendido. Estos nuevos mapas top-down son después fusionados con los mapas bottom-up, resultando un mapa de relevancia global. Por su parte, los modelos basados en plantillas trabajan con plantillas abstractas de los objetos a buscar, compuestas por características de bajo nivel (color, forma...). Estos modelos no necesitan ningún entrenamiento supervisado previo y pueden manejar información abstracta acerca del objetivo (v.g. "buscar objetos verdes y redondos"). Ya que estos modelos presentan una mayor versatilidad y no necesitan entrenamiento previo, OBVIAS sigue una estrategia similar, modificando los parámetros de modulación de percepción almacenados en la PMM. En concreto, cada tarea aplica una plantilla diferente, es decir, un **conjunto atencional** diferente, que se define como una colección de parámetros de

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

percepción capaces de resaltar proto-objetos con unas características específicas.

Por último, la etapa de semi-atención también se ocupa de la gestión de la WM. Junto a cada proto-objeto, se almacena también una serie de descriptores: su valor de relevancia, profundidad (lejanía al sensor), orientación, color, área, posición en la imagen (para la IOR), sus características de bajo nivel, su tiempo de vida y una copia de la correspondiente región de interés ocupada por el proto-objeto en la escena. Un proto-objeto entra en la WM si y sólo si es más relevante que los elementos que ya están almacenados. Si la memoria estuviera llena, el elemento menos relevante de todos es desechado. Un proto-objeto también puede ser eliminado de la WM si desaparece de la imagen (o se pierde su seguimiento). Con el fin de mantener actualizada la relevancia de los proto-objetos almacenados, ésta se recalcula en cada ciclo de percepción según el nuevo conjunto atencional.

La relevancia de un proto-objeto también depende del tiempo de vida. Los estudios acerca de la IOR [Klein, 2000; Posner et al., 1985] demuestran que cuanto más tiempo se mantiene un elemento en la WM, menos relevante se vuelve. Dicho de otro modo, después de cierto tiempo un elemento deja de atraer el foco de atención y comienza un proceso inhibitorio. Debido a esto, los proto-objetos recientemente almacenados presentan un plus de relevancia debido a su alto tiempo de vida restante. Por el contrario, los elementos más antiguos (con poco tiempo de vida restante) ven penalizada su relevancia. Por tanto, la WM actúa como una memoria caché, de forma que mantiene elementos que son relevantes para la tarea actual y los reemplaza por unos nuevos cuando se origina un cambio de contexto. Este comportamiento se consigue expresando el aporte de relevancia como una función exponencial dependiente del tiempo de vida:

$$Sal_{TTL} = \frac{2}{1 + e^{0.01(\epsilon - TTL)}} - 1$$

En resumen, tanto la WM como la PMM actúan como interfaz con el resto del sistema. Mientras la WM proporciona proto-objetos relevantes a otros módulos o sistemas, éstos pueden guiar la atención hacia elementos específicos de la escena simplemente modificando el conjunto atencional almacenado en la PMM.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

Arquitectura cognitiva basada en necesidades

La arquitectura cognitiva introducida en esta tesis trata de ir más allá de los modelos desarrollados hasta la fecha, considerando que la percepción basada en atención, el razonamiento abstracto y la ejecución de acciones presentan una estrecha relación de simbiosis, modulándose mutuamente de forma simultánea.

En concreto, se ha propuesto una **arquitectura cognitiva basada en necesidades satisfechas** inspirada en el modelo humano de desempeño establecido por Rasmussen [1983]. En dicho modelo humano, la información se percibe en diferentes niveles cognitivos, distinguiéndose entre *habilidades*, *reglas* y *conocimiento*. Rasmussen considera que la actividad humana en un entorno conocido está orientada a una meta concreta y es controlada por un conjunto de reglas, cuya utilidad se ha demostrado con anterioridad. En entornos desconocidos, en cambio, donde no se dispone de reglas probadas, el comportamiento debe ser controlado directamente por el objetivo. Este proceso se corresponde con la planificación consciente de alto nivel en sentido clásico.

La arquitectura propuesta se sustenta principalmente sobre el Mecanismo de Atención previamente detallado y un Entorno de Planificación capaz de manejar información parcial. En términos de sistemas de atención, el planificador implementa la parte top-down del modelo, pudiendo modificar el nivel de relevancia de los elementos presentes en la escena dependiendo de sus requerimientos. En cambio, el modelo de atención devuelve únicamente información sobre áreas relevantes y sólo dicha información puede ser utilizada para generar un plan. Por tanto, los elementos percibidos determinan qué acciones puede ejecutar el planificador en cada momento.

La arquitectura cognitiva se divide en dos niveles: un *nivel basado en reglas* y un *nivel basado en conocimiento*. En el primero se sitúan las tareas que pueden ejecutarse en un instante determinado. La ejecución de estas tareas se basa en la satisfacción de unas **necesidades** determinadas: cada tarea tiene una serie de necesidades en forma de categorías de objetos percibidos, que deben cubrirse para que se pueda ejecutar

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

satisfactoriamente. Por ejemplo, la tarea "buscar una carta roja" necesitaría la percepción de elementos pertenecientes a las categorías "cosas rojas" y "cosas rectangulares" respectivamente. El número de categorías que necesita cada tarea depende estrechamente de su propia naturaleza. Algunas actividades necesitan unas cuantas categorías mientras otras, simplemente una. Además, la arquitectura se ha diseñado para aprovechar al máximo las ventajas ofrecidas por un sistema de atención, pudiendo guiar al sistema de visión hacia elementos relevantes para la actividad actual.

Otro parámetro importante asociado a las tareas ubicadas en el *nivel basado en reglas* es el *potencial de activación*. Este factor mide el número de necesidades satisfechas y, por consiguiente, la probabilidad de ejecutar una determinada tarea en un futuro cercano. En caso de que no sea posible ejecutar ninguna tarea en un instante determinado, esta información puede usarse también para guiar al sistema de percepción. De esta forma, pueden destacarse los elementos de la escena que cubren las necesidades pendientes de la tarea con mayor probabilidad de ser ejecutada a corto plazo.

Por su parte, el *nivel basado en conocimiento* es el responsable de coordinar y gestionar las tareas situadas en el nivel anterior. Sería muy ineficiente mantener un elevado número de tareas para poder afrontar todos los posibles escenarios. Es más, podrían ejecutarse ciertas acciones de manera errónea debido a ambigüedades. Por ejemplo, dos o más tareas contradictorias podrían estar listas para ejecutarse al mismo tiempo. Por ello, la planificación general, definición de metas y la toma de decisiones tiene lugar de forma centralizada en este *nivel basado en conocimiento*, siendo su núcleo un **Planificador automático de propósito general**.

Al contrario de lo que ocurre en el modelo de Rasmussen, el planificador recibe información tanto del sistema visual como de las tareas situadas en el *nivel basado en reglas*. La información visual se expresa en forma de predicados abstractos, derivados de las categorías percibidas. Dicha información puede usarse para elaborar nuevos planes si fuere necesario. Además, ya que es proporcionada por un sistema de atención, se trata de información parcial por lo que el Entorno de Planificación debe ser capaz de manejar este tipo de información. Es decir, se requiere una adaptación continua del plan a los elementos percibidos. Para solventar el problema de

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGUNA

la observabilidad parcial, se ha empleado un Entorno de Planificación con Sobresubscripción (Oversubscription Planning Framework) [García-Olaya et al., 2011]. Este planificador es capaz de elaborar planes parciales en función de los datos observados. De esta forma, aunque no se pueda cumplir una meta global, ésta se subdivide en metas parciales que sí puedan lograrse con los datos conocidos.

La información visual también puede usarse para deducir el contexto de la escena. El entorno de planificación, pues, puede servirse de otros módulos de alto nivel para determinar cuáles son las tareas más adecuadas para cada circunstancia. Esta información sobre el entorno puede obtenerse de diferentes fuentes, como módulos de reconocimiento de objetos, análisis de escenas o incluso de la interacción con humanos a través de interfaces hombre-máquina.

En relación a las tareas situadas en el *nivel basado en reglas*, el entorno de planificación puede añadir, quitar, reemplazar y priorizarlas en función de la información de contexto y las categorías percibidas. A cambio, las tareas informan al planificador acerca de su estado de ejecución (en concreto, informan sobre su potencial de activación y sobre si han terminado con éxito su ejecución).

Cerrando el lazo

La conexión de la arquitectura cognitiva y el modelo de atención OBVIAS posibilitan el cierre del lazo percepción-razonamiento-acción de forma efectiva y bidireccional. Este proceso se lleva a cabo fundamentalmente en la Memoria de Trabajo (WM) y la Memoria de Modulación de Percepción (PMM). De forma resumida, dependiendo del número de necesidades que satisfacen los proto-objetos almacenados en la WM, la influencia de las tareas en la modulación de los parámetros de percepción será mayor (muchas necesidades cubiertas) o menor (necesidades escasamente satisfechas).

El primer paso para cerrar el lazo consiste en transformar la información visual, almacenada en forma de proto-objetos, en predicados abstractos accesibles para el entorno de planificación. Existen estudios psicológicos

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

que sostienen que las operaciones de alto nivel sobre entidades visuales se basan en una categorización perceptiva previa de los elementos en el campo de visión. Por tanto, es necesario un **categorizador** que clasifique los proto-objetos de la WM en categorías semánticas.

Debido a que está integrado en una etapa de visión temprana, este proceso de categorización debe ser rápido y computacionalmente eficiente. Es decir, no se trata de hacer corresponder cada proto-objeto con un objeto real sino más bien de clasificarlo dentro de una clase genérica. No obstante, se pueden usar procedimientos y algoritmos más lentos pero más precisos en el*nivel basado en conocimiento* para proporcionar una mayor información de contexto acerca de la escena. La categoría asignada a cada proto-objeto también es almacenada en la WM para hacerla accesible a otros sistemas, existiendo la posibilidad de catalogar un proto-objeto como no perteneciente a ninguna categoría conocida.

Una vez que se obtienen las categorías percibidas, las tareas localizadas en el *nivel basado en reglas* se activan dependiendo del número de sus necesidades que sean satisfechas, tal como se ha explicado anteriormente. Por tanto, la mitad del problema se encuentra ya resuelto: las categorías percibidas son capaces de activar tareas latentes situadas en el *nivel basado en reglas*. Es más, el *nivel basado en conocimiento* es capaz de detectar modificaciones en el contexto de la escena o el rol del robot, de forma que se pueden modificar las tareas involucradas en la nueva actividad.

La otra parte del problema está relacionada con el sentido razonamiento-percepción del lazo. En este caso, la tarea con mayor potencial de activación es la que condiciona al sistema de atención para realizar una búsqueda determinada de elementos. En términos del planificador, la definición de una acción incluye también efectos sobre la búsqueda basada en atención, lo que resulta especialmente interesante cuando no se puede alcanzar ninguna meta con la información actual. Dicho condicionamiento del sistema de atención visual se expresa en forma de predicados semánticos del tipo “buscar objetos verdes y redondos”. Así, resulta necesario un módulo **intérprete** que transforme dichos predicados semánticos en un nuevo conjunto atencional que se almacenará en la PMM.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

Aunque algunos modelos han intentado deducir estos parámetros de modulación de percepción a partir de un objetivo específico usando inferencia bayesiana [Yu et al., 2012], actualmente no existe una solución general para este tipo de intérpretes. Este procedimiento depende en gran medida del número y tipo de características usadas en el modelo de atención. Por tanto, por el momento la solución pasa por definir manualmente la transformación para cada predicado. También es posible entrenar de forma supervisada el intérprete para cada acción y problema, empleando técnicas de Aprendizaje Automático (Machine Learning) (como en el método desarrollado, por ejemplo, por Frintrop [2006]).

Resultados experimentales

Tanto el sistema de atención como la arquitectura cognitiva propuestas en esta tesis se han evaluado usando un robot real (NOMADA) situado en un entorno dinámico y sin restricciones.

En primer lugar, se ha comprobado la eficacia del sistema de atención a la hora de extraer los proto-objetos más relevantes de una escena natural. Normalmente, los sistemas de atención se analizan tomando como referencia únicamente el componente bottom-up. Sin embargo, no existe hasta la fecha consenso acerca de la metodología y métricas que deben utilizarse para realizar dicha evaluación. De este modo, los modelos son evaluados a menudo por sus autores empleando sus propias métricas [Aziz and Mertsching, 2009; Borji and Itti, 2013].

Debido a esto, se ha evaluado el modelo OBVIAS considerando principalmente criterios cualitativos que demuestran que el modelo obtiene objetos relevantes de una forma razonable en diferentes escenarios. Entre estos escenarios de prueba se han incluido escenas cotidianas donde resulta fácil razonar qué elementos son los más relevantes (escenas con señales de tráfico, eventos deportivos o animales de colores vivos), obras de arte (donde el autor suele buscar un efecto determinado) e imágenes de experimentos psicológicos sobre la atención. Además, con el fin de proporcionar una serie de medidas objetivas que posibilite futuras comparaciones con otros sistemas, se han aplicado las métricas propuestas

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

por Le Meur and Baccino [2013]. Esta metodología está basada en la comparación estadística con bases de datos de movimiento de la vista al observar imágenes, obtenidas directamente de personas.

Como resultado de la evaluación cuantitativa, se han obtenido unos valores para ROC (*Receiver Operating Characteristic*, Característica Operativa del Receptor) y NSS (*Normalised Scanpath Saliency*, Relevancia Normalizada del recorrido visual) que sitúan a OBVIAS en la media de los sistemas de atención analizados con estas medidas. Sin embargo, el gran inconveniente de estas métricas radica en su carácter puramente estático. Por tanto, fenómenos como la Inhibición de Retorno no se pueden evaluar. Además, una de las mayores ventajas de OBVIAS está relacionada con su adaptabilidad continua a entornos variantes en el tiempo.

Por tanto, resulta mucho más interesante la evaluación del sistema completo en un escenario complejo con tareas que cambian en el tiempo. La fiabilidad de la arquitectura cognitiva se ha puesto a prueba en dos experimentos diferentes.

En primer lugar, se ha elaborado una prueba de concepto que pone de manifiesto los principales conceptos del cierre de lazo bidireccional. En este experimento, el escenario está compuesto por una serie de cartas de diferentes colores que contienen una letra del abecedario cada una. El objetivo consiste en dejar ordenadas todas las cartas alfabéticamente. En este caso, el guiado del sistema de atención se basa exclusivamente en el color de las cartas. Las cartas que se pueden ordenar en cada momento, es decir, las tareas que se pueden llevar a cabo, dependen de las cartas percibidas (hasta 5 como máximo, que es la capacidad de la memoria de trabajo visual). Los resultados de este primer experimento demuestran que, efectivamente, el guiado basado en atención contribuye a mejorar considerablemente la eficiencia general del sistema completo. No obstante, también queda demostrado que un guiado basado en una estrategia errónea puede resultar incluso más pernicioso que la imposibilidad de guiar al sistema de percepción.

El segundo experimento está enfocado a observar el rendimiento de la arquitectura cognitiva al enfrentarse a problemas complejos con diversas tareas y contextos. En concreto, se trata de un robot social real que tiene

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGA

que clasificar diferentes objetos, que le proporciona una persona, en su correspondiente contenedor. En cierto modo, puede verse como un robot basurero que separa y recicla los desechos que recoge de las personas. Debido a la complejidad que supondría un problema de este tipo absolutamente fiel a la realidad, se han asumida ciertas simplificaciones. Por ejemplo, los objetos entregados se corresponden siempre con bolas de diferentes colores y los contenedores son meras marcas rectangulares de colores en la pared.

El robot puede encontrarse en tres contextos diferentes. En el primero, denominado *comportamiento por defecto*, sólo existe una tarea: buscar a personas. Una vez encontrado alguien, el robot pasa al *comportamiento de asistencia*, en el que tiene que moverse hacia la persona y recoger el objeto a reciclar. Por último, el robot pasa al *comportamiento de reciclado*, en el que tiene que buscar el contenedor correspondiente y depositar el objeto. Por tanto, la configuración del experimento permite que, en función del contexto actual, el *nivel basado en conocimiento* pueda modificar las tareas situadas en el *nivel basado en reglas*. Además, para cada contexto, salvo para el comportamiento por defecto, existen distintas tareas, cuya activación total depende exclusivamente de la observación de las categorías expresadas en forma de necesidades. Por último, el guiado visual queda de manifiesto al observar cómo elementos que son muy relevantes para algunas tareas (por ejemplo, la de buscar personas), son completamente ignorados en otros instantes en los que, por ejemplo, lo que se busca es el contenedor.

El experimento se ha repetido para distintas condiciones, obteniéndose una tasa de éxito del 85%, ejecutándose el ciclo percepción-razonamiento-acción con una tasa de $2.4 \frac{iteraciones}{segundo}$. Por tanto, la arquitectura es lo suficientemente rápida como para reaccionar a cambios contextuales con solvencia.

Los experimentos llevados a cabo con la arquitectura completa demuestran la habilidad del sistema para discriminar diferentes tipos de información en función de comportamientos y tareas que varían en el tiempo. De esta forma, se pueden afrontar entornos complejos, donde el coste computacional asociado al procesado completo de la escena sería completamente inadmisible, simplemente destacando características específicas y simples que optimizan la cantidad de información recibida.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

Conclusiones

Hasta el momento, tanto la percepción basada en atención como los sistemas de razonamiento abstracto se han investigado como problemas separados, asumiendo una concepción unilateral del lazo percepción-razonamiento-acción. La arquitectura cognitiva presentada en esta tesis aúna ambas disciplinas, promoviendo una solución completa y bidireccional del problema del cierre de lazo.

Los experimentos realizados dejan patente dicha **bidireccionalidad**. No sólo se guía al sistema de percepción en relación a la actividad actual, sino que también la presencia de elementos significativos en la escena desencadena la ejecución de tareas latentes o incluso cambios completos de rol en un robot (por ejemplo, en las pruebas se aprecia cómo la detección de una persona mueve al robot a asistirla). La modulación mutua entre ambos implica un equilibrio dinámico del sistema. Por tanto, el sistema difícilmente quedará bloqueado en una tarea determinada debido a la combinación de un mecanismo de Inhibición de Retorno en el sistema de atención con una planificación basada en conocimiento abstracto.

La configuración modular de la arquitectura propuesta facilita también un alto grado de **versatilidad**. El modelo puede ser adaptado a diferentes usos simplemente definiendo las tareas involucradas (con sus necesidades) y proporcionando la definición del dominio. Esta concepción modular facilita el desarrollo de nuevos comportamientos ya que no es necesario tener en cuenta las dependencias en el *nivel basado en reglas*.

En definitiva, la arquitectura cognitiva propuesta en esta tesis representa un modelo flexible, adaptable, fácil de usar y ampliar, biológicamente plausible y de propósito general para interconectar percepción y razonamiento a través de una modulación mutua y simultánea.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

UNIVERSIDAD DE MÁLAGA
Publicaciones y Divulgación Científica

UNIVERSIDAD DE MÁLAGA
Publicaciones y Divulgación Científica

A The Bounded Irregular Pyramid (BIP)

Pyramids are hierarchical structures widely used in segmentation tasks. A pyramid segmentation algorithm describes an image employing multiple representations with decreasing resolution. Thus, local operations can adapt the pyramid hierarchy to the topology of the image, allowing the detection of global features and their representation at low resolution levels [Huart and Bartolino, 2005]. With respect to other irregular pyramids, the main advantage of the BIP is that it is able to obtain similar segmentation results but in a faster way [Marfil et al., 2006; Vázquez-Martín et al., 2009].

The BIP combines a $2 \times 2/4$ regular structure with an irregular simple graph with an union-find decimation process [Marfil et al., 2007]. The combination of both structures generates an irregular configuration which is described as a graph hierarchy in which each level $G_l = (N_l, E_l)$ is a set of nodes, N_l, linked by a set of intra-level edges, E_l. Each graph G_{l+1} is built from G_l by computing the nodes of N_{l+1} from the nodes of N_l and establishing the inter-level edges $E_{l,l+1}$. Therefore, each node n_i of G_{l+1} has associated a set of nodes of G_l, which is called the *reduction window* of n_i. This includes all nodes linked to n_i by an inter-level edge. The node n_i is called *parent* of the nodes in its reduction window, which are called *sons*. The successive levels of the hierarchy are built using a regular decimation process and a union-find strategy. Therefore, there are two types of nodes: nodes belonging to the $2 \times 2/4$ structure, named *regular nodes*, and *virtual nodes* or nodes belonging to the irregular structure. Fig. A.1 shows two levels of the BIP hierarchy.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

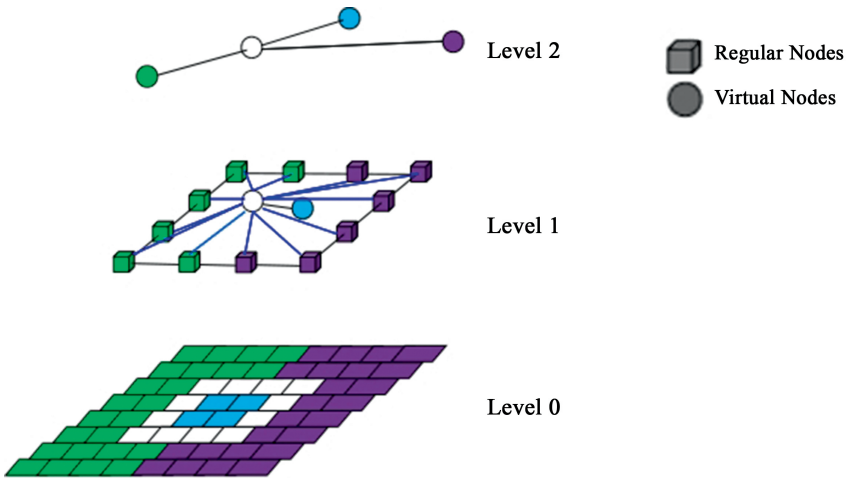


Figure A.1: Three levels of the BIP graph hierarchy.

To introduce colour information within the BIP, all nodes have associated two parameters: chromatic phasor $S_{\angle H}(n)$ and luminosity $V(n)$, where S, H and V are the saturation, hue and value of the HSV[1] colour space. The chromatic phasor and the luminosity of a node n at level l are equal to the average of the respective values of the nodes in its reduction window. As the employed similarity measurement is the HSV colour distance, two nodes are similar or have similar colour if their HSV values are less or equal than a similarity threshold T.

The graph $G_0 = (N_0, E_0)$ is a 8-connected graph where the nodes are the pixels of the original image. The chromatic phasors and the luminosity values of the nodes in $G_0 = (N_0, E_0)$ are equal to the chromatic phasors and luminosity values of their corresponding image pixels. Although the process to build the graph $G_{l+1} = (N_{l+1}, E_{l+1})$ from $G_l = (N_l, E_l)$ is closely explained in [Marfil and Bandera, 2009], a briefly description appears below:

1. *Regular decimation process.* If four regular neighbour nodes of the level l have similar colour, they are grouped together generating a regular node in $l + 1$.

2. *Irregular decimation process.* Each irregular or regular node $x \in N_l$ without parent at level $l+1$ chooses the closest neighbour y according

[1] HSV stands for **H**ue, **S**aturation and **V**alue

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

to its colour. If this condition is not satisfied by any node, then a new node x' is generated at level $l+1$. This node will be the parent node of x. Besides, it will constitute a root node and the set of nodes linked to it at base level will be an homogeneous set of pixels according to the defined criteria. On the other hand, if y exists and it has a parent z at level $l+1$, then x is also linked to z. If y exists but it does not have a parent at level $l+1$, a new irregular node z' is generated at level $l+1$. In this case, the nodes x and y are linked to z'. This process is sequentially performed and, when it finishes, each node of G_l is linked to its parent node in G_{l+1}. That is, a partition of N_l is defined.

3. *Definition of intra-level arcs.* The set of edges E_{l+1} is obtained by defining the neighbourhood relationships between the nodes N_{l+1}. Two nodes at level $l+1$ are neighbours if their reduction windows, i.e. the sets of nodes linked to them at level l, are connected at level l.

The construction of the BIP process ends when is not possible to link together any more nodes because of they are not similar. Experimental results show that the modified BIP decimation scheme is robust against slightly shifting of the input image.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

UNIVERSIDAD DE MÁLAGA
Publicaciones y Divulgación Científica

B Planning domains and problems in PDDL

The Planning Framework located at the knowledge-based level of the cognitive architecture proposed in this thesis (see section 5.2 on page 91) employs a PDDL (Planning Domain Definition Language) description of the different domains in order to solve problems. This appendix contains the PDDL definition for the domains and problem types corresponding to the experiments carried out in order to evaluate the proposed architecture (see section 6.3 on page 116).

Domain and problem for Experiment 1: Putting coloured cards in order

Domain PDDL definition for Experiment 1 (see section 6.3.1 on page 117):

```
(define (domain cards)
(:requirements :typing :fluents :preferences)

(:types card colour - object)

(:predicates
   (ordered ?c1 ?c2 - card)
   (order ?c1 ?c2 - card)
```

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

```
   (visible ?c - card)
   (colour ?ca - card ?co - colour)
   ;?c1 is the base of the stack, ?c2 is the top
   (stack ?c1 ?c2 - card)
   (_something-done)
)

(:functions
   (attentional ?c - colour)
   (total-cost)
)

(:action stack
:parameters
   ;?c1 ?c2 is the first stack, ?c3 ?c4 is the second
   (?c1 ?c2 ?c3 ?c4 - card
   ?co - colour)
:precondition
   (and (order ?c2 ?c3) (visible ?c2) (visible ?c4)
     (colour ?c4 ?co) (stack ?c1 ?c2) (stack ?c3 ?c4)
     (not (ordered ?c2 ?c3)))
:effect
   (and (ordered ?c2 ?c3)
     (not (visible ?c2))
     (stack ?c1 ?c4)
     (not (stack ?c3 ?c4))
     (not (stack ?c1 ?c2))
     (increase (attentional ?co) 1)
     (increase (total-cost) 1)
     (_something-done)
))

(:action search
:parameters
   (?c1 ?c2 ?c3 - card ?co1 ?co2 - colour)
:precondition
```

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

```
   (and (colour ?c1 ?co1) (visible ?c1) (not (visible ?c2))
       (not (visible ?c3))(colour ?c3 ?co2) (order ?c1 ?c2)
       (stack ?c2 ?c3) (not (_something-done)))
:effect
   (and
     (increase (attentional ?co1) 1)
     (increase (attentional ?co2) 1)
     (increase (total-cost) 300)
     (_something-done)
   )
)
)
```

Problem example for Experiment 1. The predicates below *Dynamic predicates* are filled up by the categorizer at the beginning of each iteration.

```
(define (problem p1)
  (:domain cartas)
(:objects  blue red yellow green - colour
 a b c d e f g h i j k l m n o p q r s t u v w x y z - card
)
  (:init

  ;; fixed part of the pddl problem (static predicates)
   (colour a blue)
   (colour b blue)
   (colour c blue)
   (colour d green)
   (colour e green)
   (colour f green)
   (colour g yellow)
   (colour h yellow)
   (colour i yellow)
   (colour j red)
   (colour k red)
   (colour l red)
```

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

```
(colour m blue)
(colour n blue)
(colour o green)
(colour p green)
(colour q yellow)
(colour r yellow)
(colour s red)
(colour t red)
(colour u blue)
(colour v blue)
(colour w blue)
(colour x blue)
(colour y blue)
(colour z blue)
(order a b)
(order b c)
(order c d)
(order d e)
(order e f)
(order f g)
(order g h)
(order h i)
(order i j)
(order j k)
(order k l)
(order l m)
(order m n)
(order n o)
(order o p)
(order p q)
(order q r)
(order r s)
(order s t)
(order t u)
(order u v)
(order v w)
```

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGA

```
 (order w x)
 (order x y)
 (order y z)

 ;begin dynamic predicates
 (visible a)
 (visible h)
 (visible c)
 (visible w)
 (visible t)
 ;end dynamic predicates

 (= (attentional blue) 5)
 (= (attentional yellow) 5)
 (= (attentional red) 5)
 (= (attentional green) 5)
 (= (total-cost) 0)
)
(:goal
 (and
  (ordered a b)
  (ordered b c)
  (ordered c d)
  (ordered d e)
  (ordered e f)
  (ordered f g)
  (ordered g h)
  (ordered h i)
  (ordered i j)
  (ordered j k)
  (ordered k l)
  (ordered l m)
  (ordered m n)
  (ordered n o)
  (ordered o p)
  (ordered p q)
```

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

```
    (ordered q r)
    (ordered r s)
    (ordered s t)
    (ordered t u)
    (ordered u v)
    (ordered v w)
    (ordered w x)
    (ordered x y)
    (ordered y z)
  ))
(:metric minimize (total-cost))
)
```

Domain and problem for Experiment 2: A social robot classifying objects

Domain PDDL definition for Experiment 2 (see section 6.3.2 on page 123):

```
(define (domain robot)
(:requirements :typing :fluents :preferences)

(:types thing robot color symmetry waypoint - object)

(:constants
   skin-color - color
   round vertical - symmetry)

(:predicates
  (at ?t - thing ?w - waypoint)
  (placed ?t - thing)
  (visible ?t - thing)
  (is-human ?t - thing)
  (is-pole ?p - thing)
  (color ?t - thing ?c - color)
```

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

```
  (shape ?t - thing ?s - symmetry)
  (at-robot ?r - robot ?w - waypoint)
  (has ?h - thing ?t - thing) ;human has a thing
  (has-robot ?r - robot ?t - thing)
  (conected ?w - waypoint ?w2 - waypoint)
  (something-done)
  (robot-has-something)
)

(:functions
  (attentional-c ?c - color)
  (attentional-s ?s - symmetry)
  (total-cost)
)

(:action move-to-ball
  :parameters (?b - thing ?h - thing
               ?r - robot ?w1 - waypoint ?w2 - waypoint)
  :precondition (and
          (not (is-pole ?b))
          (not (is-human ?b))
  (is-human ?h)
  (has ?h ?b)
   (visible ?b)
(visible ?h)
(at ?b ?w2)
(at-robot ?r ?w1)
(not(robot-has-something))
(conected ?w1 ?w2))
  :effect (and
      (at-robot ?r ?w2)
   (not (at-robot ?r ?w1))
   (increase (total-cost) 3)
   (something-done)))

(:action look-at
```

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

```
  :parameters (?t - thing ?r - robot)
  :precondition (and
        (not (is-pole ?t))
        (not (is-human ?t))
   (visible ?t)
(not(robot-has-something)))
  :effect (and
       (increase (attentional-c skin-color) 1)
   (increase (total-cost) 50)
   (something-done)))

(:action move-to-pole
  :parameters (?p - thing ?b - thing ?c - color
               ?r - robot ?w1 - waypoint ?w2 - waypoint)
  :precondition (and
       (is-pole ?p)
    (visible ?p)
 (at ?p ?w2)
 (at-robot ?r ?w1)
 (has-robot ?r ?b)
 (color ?b ?c)
 (color ?p ?c)
 (conected ?w1 ?w2))
  :effect (and
      (at-robot ?r ?w2)
   (not (at-robot ?r ?w1))
   (increase (total-cost) 3)
   (something-done)))

(:action move-to-human
  :parameters (?t - thing ?r - robot
               ?w1 - waypoint ?w2 - waypoint)
  :precondition (and
    (is-human ?t)
    (visible ?t)
 (at ?t ?w2)
```

Publicaciones y Divulgación Científica
UNIVERSIDAD DE LAGA

```
 (at-robot ?r ?w1)
 (not(robot-has-something))
 (conected ?w1 ?w2))
  :effect (and
      (at-robot ?r ?w2)
   (not (at-robot ?r ?w1))
   (something-done)
   (increase (total-cost) 1)))

(:action search-ball
  :parameters (?t - thing ?h - thing ?s - symmetry)
  :precondition (and
    (not (visible ?t))
    (is-human ?h)
 (visible ?h)

 (shape ?t ?s))
  :effect (and
      (increase (attentional-s ?s) 1)
   (increase (total-cost) 4)
   (something-done)))

(:action search-human
  :parameters (?h - thing)
  :precondition (and
    (not (visible ?h))
 (color ?h skin-color))
  :effect (and
      (increase (attentional-c skin-color) 1)
   (increase (total-cost) 5)
   (something-done)))

(:action search-pole
  :parameters (?p - thing ?b - thing
               ?r - robot ?c - color ?s - symmetry)
  :precondition (and
```

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

```
    (not (visible ?p))
 (is-pole ?p)
 (has-robot ?r ?b)

 (color ?b ?c)
 (shape ?p ?s))
  :effect (and
      (increase (attentional-c ?c) 1)
      (increase (attentional-s ?s) 1)
   (increase (total-cost) 4)
   (something-done)))

(:action get
  :parameters (?h - thing ?t - thing
               ?r - robot ?w - waypoint ?c - color)
  :precondition (and
    (is-human ?h)
 (visible ?h)
 (has ?h ?t)
 (visible ?t)
 (color ?t ?c)
 (at ?h ?w)
 (at ?t ?w)
 (not(robot-has-something))
 (at-robot ?r ?w))
  :effect (and
      (not (has ?h ?t))
   (has-robot ?r ?t)
   (not (at ?t ?w))
   (robot-has-something)
   (something-done)
   (increase (attentional-c ?c) 1) ;to search for the pole
   (increase (total-cost) 1)))

(:action drop
  :parameters (?p - thing ?t - thing
```

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

```
                ?r - robot ?w - waypoint ?c - color)
  :precondition (and
 (has-robot ?r ?t)
 (visible ?p)
 (is-pole ?p)
 (at-robot ?r ?w)
 (at ?p ?w)
 (color ?t ?c)
 (color ?p ?c)
 )
  :effect (and
      (not (has-robot ?r ?t))
   (placed ?t)
   (at ?t ?w)
   (something-done)
   (increase (total-cost) 1)))
)
```

Problem example for Experiment 2. The predicates below *Dynamic predicates* are filled up by the categorizer at the beginning of each iteration.

```
(define (problem p1)
  (:domain robot)
    (:objects red blue yellow green - color
  ball0 ball1 ball2 ball3 ball4 ball5 - thing
  human0 human1 human2 human3 human4 - thing
  robot1 - robot
  pole0 pole1 pole2 pole3 pole4 pole5 - thing
  wp1 wp2 wp3 wp4 wp5 wp6 wp7 - waypoint)
    (:init
(is-human human0)
(is-human human1)
(is-human human2)
(is-human human3)
(is-human human4)
(is-pole pole0)
```

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

```
(is-pole pole1)
(is-pole pole2)
(is-pole pole3)
(is-pole pole4)
(is-pole pole5)
(color human0 skin-color)
(color human1 skin-color)
(color human2 skin-color)
(color human3 skin-color)
(color human4 skin-color)
(shape ball0 round)
(shape ball1 round)
(shape ball2 round)
(shape ball3 round)
(shape ball4 round)
(shape ball5 round)
(shape pole0 vertical)
(shape pole1 vertical)
(shape pole2 vertical)
(shape pole3 vertical)
(shape pole4 vertical)
(shape pole5 vertical)
(conected wp1 wp2)
(conected wp1 wp3)
(conected wp1 wp4)
(conected wp1 wp5)
(conected wp1 wp6)
(conected wp1 wp7)
(conected wp2 wp1)
(conected wp2 wp3)
(conected wp2 wp4)
(conected wp2 wp5)
(conected wp2 wp6)
(conected wp2 wp7)
(conected wp3 wp1)
(conected wp3 wp2)
```

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

```
(conected wp3 wp4)
(conected wp3 wp5)
(conected wp3 wp6)
(conected wp3 wp7)
(conected wp4 wp1)
(conected wp4 wp2)
(conected wp4 wp3)
(conected wp4 wp5)
(conected wp4 wp6)
(conected wp4 wp7)
(conected wp5 wp1)
(conected wp5 wp2)
(conected wp5 wp3)
(conected wp5 wp4)
(conected wp5 wp6)
(conected wp5 wp7)
(conected wp6 wp1)
(conected wp6 wp2)
(conected wp6 wp3)
(conected wp6 wp4)
(conected wp6 wp5)
(conected wp6 wp7)
(conected wp7 wp1)
(conected wp7 wp2)
(conected wp7 wp3)
(conected wp7 wp4)
(conected wp7 wp5)
(conected wp7 wp6)

;begin dynamic predicates
    (at-robot robot1 wp1)
(at human2 wp1)
(at ball1 wp1)
(at ball2 wp1)
(at ball3 wp3)
(at ball4 wp4)
```

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

```
(at pole1 wp3)
(at pole2 wp4)
(at pole3 wp1)
(at pole4 wp6)

(color ball1 blue)
(color ball2 blue)
(color pole3 red)
(visible human2)
(has human2 ball1)
(visible pole3)
(visible ball2)
(visible ball1)
;end dynamic predicates

(= (attentional-c blue) 5)
    (= (attentional-c yellow) 5)
    (= (attentional-c red) 5)
    (= (attentional-c green) 5)
    (= (attentional-s round) 5)
    (= (attentional-s vertical) 5)
(= (attentional-c skin-color) 5)
    (= (total-cost) 0)
)
 (:goal
   (and (preference p01 (placed ball1))
    (preference p02 (placed ball2))
    (preference p03 (placed ball3))
    (preference p04 (placed ball4))
(preference p05 (something-done))
))
(:metric minimize (total-cost))
)
```

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

Bibliography

Albore, A., Palacios, H., and Geffner, H. (2009). A translation-based approach to contingent planning. In *IJCAI*, pages 1623–1628.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., and Qin, Y. (2004). An integrated theory of the mind. *Psychological review*, 111(4):1036–1060.

Anderson, J. R. and Lebiere, C. (1998). *The atomic components of thought*. Psychology Press.

Awh, E. and Pashler, H. (2000). Evidence for split attentional foci. *Journal of Experimental Psychology: Human Perception and Performance*, 26(2):834–846.

Aziz, M. Z. and Mertsching, B. (2010). Towards standardization of metrics for evaluation of artificial visual attention. In *Proceedings of the 10th Performance Metrics for Intelligent Systems Workshop*, pages 29–36. ACM.

Aziz, Z. and Mertsching, B. (2008a). Fast and robust generation of feature maps for region-based visual attention. *IEEE Transactions on Image Processing*, 17(5):633–644.

Aziz, Z. and Mertsching, B. (2008b). Visual search in static and dynamic scenes using fine-grain top-down visual attention. In *Computer Vision Systems*, pages 3–12. Springer.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

Aziz, Z. and Mertsching, B. (2009). Towards standardization of evaluation metrics and methods for visual attention models. In *Attention in Cognitive Systems*, volume 5395 of *LNAI*, pages 227–241.

Backer, G. and Mertsching, B. (2000). Integrating time and depth into the attentional control of an active vision system. In Baratoff, G. and Neumann, H., editors, *Proc. Dynamische Perzeption Workshop der GI-Fachruppe 1.0.4*, pages 69–74. Citeseer.

Backer, G. and Mertsching, B. (2003). Two selection stages provide efficient object-based attentional control for dynamic vision. In *WAPCV 2003*, pages 9–16, Heidelberg. Springer.

Backer, G., Mertsching, B., and Bollmann, M. (2001). Data- and model-driven gaze control for an active-vision system. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:1415–1429.

Bacon, W. F. and Egeth, H. E. (1994). Overriding stimulus-driven attentional capture. *Perception & psychophysics*, 55(5):485–496.

Bandera, J. P. (2010). *Vision-based gesture recognition in a robot learning by imitation framework*. PhD thesis, University of Málaga.

Belardinelli, A., Schneider, W., and Steil, J. J. (2010). Oop: object-oriented-priority for motion saliency maps. In *Procedings of Conference on Brain-Inspired Cognitive Systems (BICS 2010)*, pages 370–381.

Bicchi, A. and Kumar, V. (2000). Robotic grasping and contact: A review. In *ICRA*, pages 348–353. Citeseer.

Bischoff, R. and Graefe, V. (2004). Hermes-a versatile personal robotic assistant. *Proceedings of the IEEE*, 92(11):1759–1779.

Borji, A. and Itti, L. (2013). State-of-the-art in visual attention modeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):185–207.

Broadbent, D. E. (1958). *Perception and communication*. Pergamon, New York.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

Bruce, N. and Tsotsos, J. (2006). Saliency based on information maximization. In Weiss, Y., Schölkopf, B., and Platt, J., editors, *Advances in Neural Information Processing Systems 18*, pages 155–162, Cambridge, MA. MIT Press.

Bruce, N. D. and Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of vision*, 9(3):1–24.

Bundesen, C. (1990). A theory of visual attention. *Psychological review*, 97(4):523–547.

Bundesen, C. and Habekost, T. (2005). Attention. In Lamberts, K. and Goldstone, R., editors, *Handbook of Cognition*. Sage Publications, London.

Bundesen, C. and Habekost, T. (2008). *Principles of Visual Attention: Linking Mind and Brain*, chapter 3. Oxford University.

Bundesen, C., Habekost, T., and Kyllingsbaek, S. (2011). A neural theory of visual attention and short-term memory (ntva). *Neuropsychologia*, 49(6):1446–1457.

Canny, J. (1986). A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8:679–698.

Collins, T., Heed, T., and Röder, B. (2010). Visual target selection and motor planning define attentional enhancement at perceptual processing stages. *frontiers on Human Neuroscience*, 4:1–10.

Comaniciu, D., Ramesh, V., and Meer, P. (2003). Kernel-based object tracking. *IEEE Transanction on Pattern Analysis and Machine Intelligence*, 25(5):564–575.

Corbetta, M. (1998). Frontoparietal cortical networks for directing attention and the eye to visual locations: Identical, independent, or overlapping neural systems? *Proceedings of the National Academy of Sciences*, 95(3):831–838.

Corbetta, M. and Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, 3(3):201–215.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

Dautenhahn, K. and Billard, A. (1999). Bringing up robots or the psychology of socially intelligent robots: From theory to implementation. In *Proc. of the third annual Conf. on Autonomous Agents*, Seattle, Washington, United States.

Draper, B. A. and Lionelle, A. (2005). Evaluation of selective attention under similarity transformations. *Computer vision and image understanding*, 100(1):152–171.

Duncan, J. (1980). The locus of interference in the perception of simultaneous stimuli. *Psychological Review*, 87(3):272–300.

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology*, 113(4):501–517.

Egeth, H. E. and Yantis, S. (1997). Visual attention: Control, representation, and time course. *Annual review of psychology*, 48(1):269–297.

Eidenberger, R., Zoellner, R., and Scharinger, J. (2009). An integrated active perception module for a distributed cognitive architecture. In *Proceedings of the International Conference on Advanced Robotics (ICAR 2009)*, pages 1–7. IEEE.

Einhäuser, W., Spain, M., and Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, 8(14):1–26.

Fagioli, S., Hommel, B., and Schubotz, R. I. (2007). Intentional control of attention: action planning primes action-related stimulus dimensions. *Psychological Research*, 71:22–29.

Falomir, Z. (2013). Towards scene understanding using contextual knowledge and spatial logics. In *Procedings of the II Workshop on Recognition and Action for Scene Understanding (REACTS)*, pages 85–100.

Ferreira, J. and Dias, J. (2014). Attentional mechanisms for socially interactive robots – a survey. *IEEE Transactions on Autonomous Mental Development*, 6(2):110–125.

Ferreira, J. F., Bessière, P., Mekhnacha, K., Lobo, J., Dias, J., Laugier, C., et al. (2008). Bayesian models for multimodal perception of 3d structure

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

and motion. In *International Conference on Cognitive Systems (CogSys 2008)*, pages 103–108.

Ferreira, J. F., Castelo-Branco, M., and Dias, J. (2012). A hierarchical bayesian framework for multimodal active perception. *Adaptive Behavior*, 20(3):172–190.

Ferreira, J. F., Lobo, J., Bessière, P., Castelo-Branco, M., and Dias, J. (2013). A bayesian framework for active artificial perception. *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics*, 43(2):699–711.

Fischer, M. H. and Hoellen, N. (2004). Space-and object-based attention depend on motor intention. *Journal of General Psychology*, 131(4):365–378.

Frintrop, S. (2006). *VOCUS: A visual attention system for object detection and goal-directed search*, volume 3899. Springer.

Frintrop, S., Rome, E., and Christensen, H. I. (2010). Computational visual attention systems and their cognitive foundations: A survey. *ACM Transactions on Applied Perception (TAP)*, 7(1):1–39.

Gao, D. and Vasconcelos, N. (2004). Discriminant saliency for visual recognition from cluttered scenes. In *Proceedings of Advances in Neural Information Processing Systems*, pages 481–488. MIT Press.

García-Olaya, A., de la Rosa, T., and Borrajo, D. (2011). Using the relaxed plan heuristic to select goals in oversubscription planning problems. In Lozano, J., Gómez, J., and Moreno, J., editors, *Advances in Artificial Intelligence*, volume 7023 of *Lecture Notes in Computer Science*, pages 183–192. Springer Berlin- Heidelberg.

Gerevini, A., Haslum, P., Long, D., Saetti, A., and Dimopoulos, Y. (2009). Deterministic planning in the fifth international planning competition: Pddl3 and experimental evaluation of the planners. *Artificial Intelligence*, 173(5):619–668.

Gibson, J. (1979). *The ecological Approach to Visual Perception*. Houghton Mifflin (Boston).

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

Gilet, E., Diard, J., and Bessière, P. (2011). Bayesian action–perception computational model: Interaction of production and recognition of cursive letters. *PloS one*, 6(6):e20387.

Green, D. M., Swets, J. A., et al. (1966). *Signal detection theory and psychophysics*, volume 1974. Wiley New York.

Greene, M. and Wolfe, J. (2011). Global image properties do not guide visual search. *Journal of vision*, 11(6):1–9.

Gutiérrez, M., Romero-Garcés, A., Bustos, P., and Martínez, J. (2013). Progress in robocomp. *Journal of Physical Agents*, 7(1).

Hamker, F. H. (2006). Modeling feature-based attention as an active top-down inference process. *BioSystems*, 86(1):91–99.

Huart, J. and Bartolino, P. (2005). Similarity-based and perception-based image segmentation. In *Procedings of the IEEE International Conference on Image Processing*, volume 3, pages 1148–1151.

Itti, L. (2005). Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition*, 12(6):1093–1123.

Itti, L. and Koch, C. (2001). Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:1254–1259.

James, W. (1890). *The Principles of Psychology*. New York: Holt, Rinehart and Winston.

Johansson, R. S., Westling, G., Bäckström, A., and Flanagan, J. R. (2001). Eye–hand coordination in object manipulation. *the Journal of Neuroscience*, 21(17):6917–6932.

Johnston, B., Yang, F., Mendoza, R., Chen, X., and Williams, M.-A. (2008). Ontology based object categorization for robots. In *Practical Aspects of Knowledge Management*, pages 219–231. Springer.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

Judd, T., Ehinger, K., Durand, F., and Torralba, A. (2009). Learning to predict where humans look. In *IEEE 12th international Conference on Computer Vision*, pages 2106–2113. IEEE.

Klein, R. M. (2000). Inhibition of return. *Trends in cognitive sciences*, 4(4):138–147.

Koch, C. and Ullman, S. (1985). Shifts in selective visual attention: towards the underlaying neural circuitry. *Human Neurobiology*, 4(4):219–227.

Kouchaki, Z. and Nasrabadi, A. M. (2012). A nonlinear feature fusion by variadic neural network in saliency-based visual attention. In *VISAPP*, pages 457–461.

Kröse, B. and Julesz, B. (1989). The control and speed of shifts of attention. *Vision research*, 29(11):1607–1619.

Laird, J. E. (2008). Extending the soar cognitive architecture. In *Proceedings of the First Artificial General Intelligence Conference*, volume 171, pages 224–235. IOS Press.

Laird, J. E., Newell, A., and Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. *Artificial intelligence*, 33(1):1–64.

Langley, P., Cummings, K., and Shapiro, D. (2004). Hierarchical skills and cognitive architectures. In *Proceedings of the twenty-sixth annual conference of the cognitive science society*, pages 779–784. Citeseer.

Langley, P., Laird, J. E., and Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, 10(2):141–160.

Le Meur, O. and Baccino, T. (2013). Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior research methods*, 45(1):251–266.

Le Meur, O., Le Callet, P., Barba, D., and Thoreau, D. (2006). A coherent computational approach to model bottom-up visual attention. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(5):802–817.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

Lewis, F. L., Dawson, D. M., and Abdallah, C. T. (2003). *Robot manipulator control: theory and practice*. CRC Press.

Lin, C. J., Shiang, W.-J., Chuang, C.-Y., and Liou, J.-L. (2014). Applying the skill-rule-knowledge framework to understanding operators' behaviors and workload in advanced main control rooms. *Nuclear Engineering and Design*, 270:176–184.

Luce, R. D. (1963). Detection and recognition. *Handbook of mathematical psychology*, 1:103–189.

Mannan, S., Ruddock, K., and Wooding, D. (1997). Fixation sequences made during visual examination of briefly presented 2d images. *Spatial vision*, 11(2):157–178.

Manso, L., Bachiller, P., Bustos, P., Núñez, P., Cintas, R., and Calderita, L. (2010). Robocomp: a tool-based robotics framework. In *Proceedings of the International Conference on Simulation, Modeling, and Programming for Autonomous Robots (SIMPAR)*, pages 251–262. Springer.

Marfil, R., Antúnez, E., Arrebola, F., and Bandera, A. (2013). Towards active image segmentation: the foveal bounded irregular pyramid. In *Procedings of the II Workshop on Recognition and Action for Scene Understanding (REACTS)*, pages 15–27.

Marfil, R. and Bandera, A. (2009). Comparison of perceptual grouping criteria within an integrated hierarchical framework. In Escolano, F. and Torsello, A., editors, *Proceedings of the Graph-Based Representations in Pattern Recognition (GbRPR'09)*, volume 5534 of *Lecture Notes in Computer Science*, pages 366–375, Venice, Italy. Springer.

Marfil, R., Bandera, A., Antonio, R. J., and Sandoval, F. (2009). A novel hierarchical framework for object-based visual attention. *LNCS*, 5395:27–40.

Marfil, R., Molina-Tanco, L., Bandera, A., Rodríguez, J., and Sandoval, F. (2006). Pyramid segmentation algorithms revisited. *Pattern Recognition*, 39(8):1430–1451.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

Marfil, R., Molina-Tanco, L., Bandera, A., and Sandoval, F. (2007). The construction of bounded irregular pyramids with a union-find decimation process. In Escolano, F. and Vento, M., editors, *GbRPR 2007. LNCS*, volume 4538, pages 307–318, Heidelberg. Springer.

Markman, A. B. and Dietrich, E. (2000). Extending the classical view of representation. *Trends in Cognitive Sciences*, 4(12):470–475.

Martínez, J., Romero-Garcés, A., Bandera, J. P., Marfil, R., and Bandera, A. (2012). A dds-based middleware for quality-of-service and high-performance networked robotics. *Concurrency and Computation: Practice and Experience*, 24(16):1940–1952.

McDermott, D. (1998). PDDL – the planning domain definition language. Technical Report CVC TR-98-003/DCS TR-1165, Yale Center for Computational Vision and Control.

McGregor, S. (2007). Neural network processing for multiset data. In *Proceedings of the 17th Internation Conference on Artificial Neural Networks (ICANN 2007)*, pages 460–470. Springer.

McGregor, S. (2008). Further results in multiset processing with neural networks. *Neural Networks*, 21(6):830–837.

McLeod, P., Driver, J., and Crisp, J. (1988). Visual search for a conjunction of movement and form is parallel. *Nature*, (332):154–155.

Milanese, R. (1993). *Detecting salient regions in an image: from biological evidence to computer implementation*. PhD thesis, University of Geneva.

Milanese, R., Wechsler, H., Gill, S., Bost, J., and Pun, T. (1994). Integration of bottom-up and top-down cues for visual attention using non-linear relaxation. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*, pages 781–785. IEEE.

Montesano, L., Lopes, M., Bernardino, A., and Santos-Victor, J. (2008). Learning object affordances: from sensory–motor coordination to imitation. *Robotics, IEEE Transactions on*, 24(1):15–26.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

Navalpakkam, V. and Itti, L. (2005). Moelling the influence of task on attention. *Vision Research*, 45(2):205–231.

Navalpakkam, V. and Itti, L. (2006). An integrated model of top-down and bottom-up attention for optimizing detection speed. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2049–2056. IEEE.

Nothdurft, H.-C. (1993). Faces and facial expressions do not pop out. *Perception*, 22:1287–1298.

Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175.

Oliva, A., Torralba, A., Castelhano, M. S., and Henderson, J. M. (2003). Top-down control of visual attention in object detection. In *Proceedings of International Conference on Image Processing (ICIP 2003)*, volume 1, pages I–253. IEEE.

Orabona, F., Metta, G., and Sandini, G. (2005). Object-based visual attention: a model for a behaving robot. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops*, pages 89–96. IEEE.

Orabona, F., Metta, G., and Sandini, G. (2007). A proto-object based visual attention model. In Paletta, L. and Rome, E., editors, *WAPCV 2007. LNCS (LNAI)*, volume 4840, pages 198–215, Heidelberg. Springer.

Paletta, L., Rome, E., and Buxton, H. (2005). Attention architectures for machine vision and mobile robots. *Neurobiology of attention*, pages 642–648.

Palomino, A. J., Marfil, R., Bandera, J. P., and Bandera, A. (2011). A novel biologically inspired attention mechanism for a social robot. *EURASIP Journal on Advances in Signal Processing*.

Parkhurst, D., Law, K., and Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision research*, 42(1):107–123.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

Peot, M. and Smith, D. (1992). Conditional nonlinear planning. In *Proceedings of the first international conference on Artificial intelligence planning systems*, pages 189–197, San Mateo, California.

Peters, R. J., Iyer, A., Itti, L., and Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision research*, 45(18):2397–2416.

Peterson, M. S., Kramer, A. F., and Irwin, D. E. (2004). Covert shifts of attention precede involuntary eye movements. *Perception & Psychophysics*, 66(3):398–405.

Posner, M., Rafal, R., Choate, L., and Vaughan, J. (1985). Inhibition of return: Neural basis and function. *Cognitive neuropsychology*, 2(3):211–228.

Posner, M. I. and Cohen, Y. (1984). Components of visual orienting. *Attention and performance X: Control of language processes*, 32:531–556.

Posner, M. I., Nissen, M. J., and Ogden, W. C. (1978). Attended and unattended processing modes: The role of set for spatial location. *Modes of perceiving and processing information*, pages 137–157.

Pryor, L. (1994). Perceptual and functional representations: Bridging the gap. In *Notes of the AISB Workshop on Computational Models of Cognition and Cognitive Functions*, pages 1–10. Citeseer.

Pylyshyn, Z. W. and Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial vision*, 3(3):179–197.

Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *Systems, Man and Cybernetics, IEEE Transactions on*, SMC-13(3):257–266.

Rensink, R. (2000). Seeing, sensing, and scrutinizing. *Vision research*, 40(10-12):1469–1487.

Rosenblueth, A., Wiener, N., and Bigelow, J. (1943). Behavior, purpose and teleology. *Philosophy of science*, 10(1):18–24.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

Rosenholtz, R. (1999). A simple saliency model predicts a number of motion popout phenomena. *Vision research*, 39(19):3157–3163.

Rukšėnas, R., Curzon, P., Blandford, A., and Back, J. (2013). Combining human error verification and timing analysis: A case study on an infusion pump. *Formal Aspects of Computing*, pages 1–44.

Scholl, B. (2001). Objects and attention: the state of art. *Cognition*, 80(1-2):1–46.

Shani, G. and Brafman, R. I. (2011). Replanning in domains with partial information and sensing actions. In *IJCAI*, pages 2021–2026.

Shevchenko, M., Windridge, D., and Kittler, J. (2009). A linear-complexity reparameterisation strategy for the hierarchical bootstrapping of capabilities within perception–action architectures. *Image and Vision Computing*, 27(11):1702–1714.

Simons, D. J. and Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception-London*, 28(9):1059–1074.

Smith, D. (2004). Choosing objectives in over-subscription planning. In *Proceedings of the 14th Intl. Conf. on Automated Planning & Scheduling*, pages 393–401.

Sridharan, M., Wyatt, J., and Dearden, R. (2010). Planning to see: A hierarchical approach to planning visual actions on a robot using pomdps. *Artificial Intelligence*, 174(11):704–725.

Sun, Y. and Fisher, R. (2003). Object-based visual attention for computer vision. *Artificial Intelligence*, 146(1):77–123.

Suzuki, S. and Cavanagh, P. (1995). Facial organization blocks access to low-level features: An object inferiority effect. *Journal of Experimental Psychology: Human Perception and Performance*, 21(4):901–913.

Terrillon, J. and Akamatsu, S. (1999). Comparative performance of different chrominance spaces for color segmentation and detection of human faces in complex scene images. In *Proceedings of the 12th Conference on Vision Interface*, pages 180–187.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

Theeuwes, J. (2004). Top-down search strategies cannot override attentional capture. *Psychonomic Bulletin & Review*, 11(1):65–70.

Torralba, A. (2003). Modeling global scene factors in attention. *JOSA-A*, 20(7):1407–1418.

Torralba, A., Oliva, A., Castelhano, M. S., and Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766–786.

Treisman, A. and Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology*, 12(1):97–136.

Treisman, A. and Gormican, S. (1988). Feature analysis in early vision: evidence from search asymmetries. *Psychological review*, 95(1):15–48.

Treisman, A. and Souther, J. (1985). Search asymmetry: a diagnostic for preattentive processing of separable features. *Journal of Experimental Psychology: General*, 114(3):285–310.

Tsotsos, J. K., Culhane, S. M., Kei Wai, W. Y., Lai, Y., Davis, N., and Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial intelligence*, 78(1):507–545.

Tsotsos, J. K., Liu, Y., Martínez-Trujillo, J. C., Pomplun, M., Simine, E., and Zhou, K. (2005). Attending to visual motion. *Computer Vision and Image Understanding*, 100(1):3–40.

Tünnermann, J., Born, C., and Mertsching, B. (2013). Top-down visual attention with complex templates. In *International Conference on Computer Vision Theory and Applications*, pages 370–377.

Tünnermann, J. and Mertsching, B. (2014). Region-based artificial visual attention in space and time. *Cognitive Computation*, 6(1):125 – 143.

Vázquez-Martín, R. (2009). *onLine Environment Segmentation based on Spectral Mapping (LESS-Mapping)*. PhD thesis, University of Málaga.

Vázquez-Martín, R., Marfil, R., Núñez, P., Bandera, A., and Sandoval, F. (2009). A novel approach for salient image regions detection and description. *Pattern Recognition Letters*, 30(16):1464–1476.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, 57(2):137–154.

Ward, R. (2002). Independence and integration of perception and action: An introduction. *Visual Cognition*, 9(4-5):385–391.

Windridge, D. and Kittler, J. (2010). Perception-action learning as an epistemologically-consistent model for self-updating cognitive representation. *Advances in Experimental Medicine and Biology*, pages 95–134.

Windridge, D., Shaukat, A., and Hollnagel, E. (2013). Characterizing driver intention via hierarchical perception–action modeling. *IEEE Transactions on Human-Machine Systems*, 43(1):17–31.

Wischnewski, M., Belardinelli, A., Schneider, W. X., and Steil, J. J. (2010). Where to look next? combining static and dynamic proto-objects in a tva-based model of visual attention. *Cognitive computation*, 2(4):326–343.

Wolfe, J. (1994). Guided search 2.0: a revised model of visual search. *Psychonomic Bulletin and Review*, 1:202–238.

Wolfe, J., Cave, K., and Franzel, S. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 15(3):419–433.

Wolfe, J. and Horowitz, T. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, 5(6):495–501.

Wolfe, J. M. (2007). Guided search 4.0. *Integrated models of cognitive systems*, pages 99–119.

Wolfe, J. M., Friedman-Hill, S. R., Stewart, M. I., and O'Connell, K. M. (1992). The role of categorization in visual search for orientation. *Journal of Experimental Psychology: Human Perception and Performance*, 18(1):34–49.

Wolfe, J. M. and Gancarz, G. (1997). Guided search 3.0. In *Basic and clinical applications of vision science*, pages 189–192. Springer.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

Yarbus, A. L. (1967). *Eye movement and vision.* Plenum Press, New York.

Yu, Y., Gu, J., Mann, G. K., and Gosine, R. G. (2013). Development and evaluation of object-based visual attention for automatic perception of robots. *Automation Science and Engineering, IEEE Transactions on*, 10(2):365–379.

Yu, Y., Mann, G. K., and Gosine, R. G. (2012). A goal-directed visual perception system using object-based top–down attention. *Autonomous Mental Development, IEEE Transactions on*, 4(1):87–103.

Yu, Y., Mann, K., and Gosine, R. (2010). An object-based visual attention model for robotic applications. *IEEE Transactions on Systems, Man and Cybernetics B*, 40(3):1–15.

Publicaciones y Divulgación Científica
UNIVERSIDAD DE MÁLAGA

UNIVERSIDAD DE MÁLAGA
Publicaciones y Divulgación Científica

UNIVERSIDAD DE MÁLAGA
Publicaciones y Divulgación Científica

Publicaciones y
Divulgación Científica

UNIVERSIDAD
DE MÁLAGA