Contribution to complex visual information processing and autonomous knowledge extraction : application to autonomous robotics

Dominik Maximilián Ramik

UNIVERSITÉ PARIS-EST

Thesis

Presented to obtain the title of

DOCTOR OF UNIVERSITY PARIS-EST

Specialization: Signal & Images Processing

By **Dominik Maximilián RAMÍK**

Contribution to Complex Visual Information Processing and Autonomous Knowledge Extraction: Application to Autonomous Robotics

Defended on December the 10th 2012 in presence of commission composed by

Prof.	Serge	MIGUET	Rapporteur / Université Lyon 2 – LISIR UMR 5205 CNRS
Prof.	Samia	BOUCHAFA	Rapporteur / Université EVE – Laboratoire IBISC
Prof.	Patrick	GARDA	Examinateur / Université Pierre et Marie Curie – LIP6
Dr. Hab.	Eva	VOLNA	University of Ostrava
Dr.	Christophe	SABOURIN	Examinateur / Université PARIS-EST Créteil – LISSI
Prof.	Kurosh	MADANI	Directeur de thèse / Université PARIS-EST Créteil – LISSI

UNIVERSITÉ PARIS-EST

Thèse

Présentée pour l'obtention du titre de

DOCTEUR DE L'UNIVERSITÉ PARIS-EST

Spécialité: Traitement du Signal et des Images

Par **Dominik Maximilián RAMÍK**

Contribution au Traitement d'Informations Visuelles Complexes et à l'Extraction Autonome des Connaissances : Application à la Robotique Autonome

Soutenue publiquement le 10 décembre 2012 devant la commission d'examen composée de

Prof.	Serge	MIGUET	Rapporteur / Université Lyon 2 – LISIR UMR 5205 CNRS
Prof.	Samia	BOUCHAFA	Rapporteur / Université EVE – Laboratoire IBISC
Prof.	Patrick	GARDA	Examinateur / Université Pierre et Marie Curie – LIP6
Dr. Hab.	Eva	VOLNA	University of Ostrava
Dr.	Christophe	SABOURIN	Examinateur / Université PARIS-EST Créteil – LISSI
Prof.	Kurosh	MADANI	Directeur de thèse / Université PARIS-EST Créteil – LISSI

Acknowledgement

At the first place I would like to voice my deepest gratitude to Prof. Kurosh Madani for his sustained and invaluable support and patience while guiding and supervising my PhD studies. Not only I am grateful to him for his scientific experience, which he was so gladly sharing with me, but in a special way I want to recognize here his human qualities, thanks to whose our team is working in conditions motivating us to give the best of each of us.

I would also like to express my thanks to my tutor, Dr. Christophe Sabourin. His support, help and his very pragmatic and positive attitude has encouraged me a lot through my studies. The "learning of how to be a scientist" would be certainly much more painful for me without his guidance. I thank him very much.

I want to thank Prof. Serge Miguet from Université Lyon 2 – LISIR UMR 5205 CNRS and Prof. Samia Bouchafa from Université EVE – Laboratoire IBISC for having accepted to be my PhD thesis referees. I am equally grateful to Prof. Patrick Garda from Université Pierre et Marie Curie – LIP6 for being my thesis committee member. I also remember with much gratitude Dr. Hab. Eva Volná, my Master degree supervisor and also my thesis committee member. It was she who influenced me to endeavor to continue my studies.

I would like to thank the faculty members of the IUT of Senart including Dr. Amine Chohra, Dr. Veronique Amarger, Dr. Abdennasser Chebira and Dr. Aurélien Hazan.

My thanks belong to the members and ex-members of LISSI including: Dr. Ivan Budnyk, Dr. Arash Bahrammirzaee, Dr. Ting Wang, Dr. Ramón Moreno, Jingyu Wang, Ouerdia Megherbi and Assia Aziez. I am truly grateful for their friendship. The same is true for my friends Marie Anne Vakalepu, Colette Nana, RAMBELOHARINIRINA Marie Claire, René Prieur, Bernadette and Jean-Louis Van Kelst and many others who encouraged me and supported me in various ways.

At last, but far from least I would like to express on this place my most sincere thanks and recognition to my parents, my little sister, my brother-in-law, my little brother and in a very special way to my beloved fiancée, for without their unconditional love and their support I would certainly not withstand the very demanding course of my doctoral studies.

I dedicate this work
to my most beloved mother Maria
and to all my family
for their love

Table of Contents

List of Figures

List of Tables

List of Symbols

Ω	Image
θ	Zenithal angle in siRGB
ϕ	Azimuthal angle in siRGB
l	Intensity in siRGB
Ψ	Chromaticity
R	Red component in RGB, absolute
G	Green component in RGB, absolute
B	Blue component in RGB, absolute
r	Red component in RGB, normalized
g	Green component in RGB, normalized
b	Blue component in RGB, normalized
Π_Ψ	Chromatic plane
L_c	Equichromatic line
x	Pixel coordinates in an image
M_l	Intensity saliency map
$M_{\phi\theta}$	Chromatic saliency map
C	Color saturation
M	Composite saliency map
P	Sliding window
p	Sliding window size, visual attention scale
H_C	Center histogram
H_S	Surround histogram
d	Center-surround feature
D	Center-surround saliency
M_{final}	Final saliency map
W	Homogeneity predicate
S	Image segment
α	Chromatic activation function
a, b, c	Chromatic activation function parameters

$d_h(p,q)$	Hybrid distance between pixels
$d_\Psi(p,q)$	Chromaticity distance between pixels
δ	Segmentation algorithm threshold
$L(x)$	Label of a pixel in the segmentation algorithm
t_{Var}	Threshold of variance
$t_{\bar{S}}$	Threshold of average saliency
F	Fragment
G	Fragment group
t_{area}	Threshold of area
t_{aspect}	Threshold of aspect
$t_{\theta\phi}$	Threshold of chromaticity distribution
$t_{uniformity}$	Threshold of uniformity
w	Weak classifier
z	Texture uniformity
H_{SA}	Histogram of segment sizes
H_{SR}	Relative histogram of segment sizes
I	Set of features representing the world
i	A feature
U_m	Set of utterances, a phrase
U	A set of all heard utterances
u	Utterance
o	An observation
O	Set of all observations made
ϵ	Noise
$X(u)$	Interpretation of u
B	Belief
$o_q[U_m]$	A set of utterances made by a human on the observation o_q
U_{Bq}	Utterances made by the robot on observation o_q with belief B
v	Disparity of $o_q[U_m]$ and U_{Bq}
pop_{total}	Total size of population in genetic algorithm
pop_u	Number of unique genomes in the population
ϱ	Saturation of population
P_i	Probability of i-th organism to be selected as parent
fit_i	Fitness of the i-th organism

X_S	Interpretation regarding sensor S
B_S	Belief regarding sensor S
h_{cam}	Camera height above the ground
d_{obj}	Distance of object
h_{obj}	Height of object
τ	Tilt of camera
φ_{fov}	Camera field of view
h_y	Height of object on image
Δ_y	Distance of the bottom of object from the middle of image
h_Ω	Height of image
γ_1, γ_2	Angle between the vertical and the line of sight to lower, resp. upper boundary of object
$\mathcal{A}$	Set of ambiguous utterances
$\mathcal{D}$	Set of unambiguous utterances
$\mathcal{U}$	Set of all sets of utterances given on all observations

General Introduction

Foreword

Since the dawn of mankind, through sciences and philosophy, we have developed a staggering body of knowledge both about the universe that surrounds us and about ourselves. This gave us the role of a prominent element in our ecosystem on global scale. This is of course due to our intelligence, which is incomparable to any other known living kind. This intelligence, however, does not come from nowhere, but reposes on the powerful cognitive system that we, humans, have in possession.

Human cognition is indeed a huge source of inspiration and since the time of classical philosophers it remains one of the key research domains in human sciences. Being too difficult to be grasped in a single theory or model, it has been studied from many perspectives. Some scientists explore its links to human culture (Tomasello, 1999), others investigate its connection to brain structures and functions (Koechlin et al., 1999), yet others focus on human cognition as a social function (Wyer and Srull, 1986) and the list could go on and on.

A dictionary definition[1] describes cognition as a "mental action or process of acquiring knowledge", the term originating through late Middle English from Latin cognitio, from cognoscere: "get to know". While the definition implies that cognition is proper to humans, cognitive phenomena in machines are making part of research efforts since the rise of artificial intelligence in the middle of the last century. The fact that human-like (and even animal-like) machine cognition is still beyond the reach of contemporary science only proves how difficult the problem is. Partly this is certainly caused by the fact that we still do not fully understand the human cognitive system. Yet what we know about it is a valuable inspiration for designing machine cognitive systems. This is precisely the way I have taken in this thesis, whose goal is a contribution to development of human-like machine cognition through inspiration from biological and human systems.

[1] Oxford Dictionary of English, (Pearsall, 2010)

Motivation and Objectives

The objective of this thesis, which has been previously informally stated as "a contribution to development of human-like machine cognition through inspiration from biological and human systems" will be described more specifically in following lines. However, why the problem of developing autonomous and human-like machine cognition is so important? For me, the motivation comes from the state of the art in robotics, intelligent systems and technology in general. Nowadays there exist many systems, such as sensors or robots that outperform human capacities. Yet none of existing machines can be called truly intelligent and humanoid robots sharing everyday life with humans are still far away. It is so because contemporary machines are often automatic, but rarely fully autonomous in their knowledge acquisition. This is why the conception of bio-inspired human-like machine cognition is so important for future systems and intelligent robots. It is because this is the way of a major contribution to a true autonomy of future intelligent systems. For example, but not being limited to, the development of humanoid robots capable of autonomous operation in real world conditions.

The term "cognitive system" means here specifically that characteristics of such a system are similar to those of human cognitive system. It refers the reader to the fact that a cognitive system, which would be able to comprehend the world on its own, but whose comprehension would be non-human, would subsequently be incapable of communicating about it with its human counterparts. It is obvious that such a cognitive system would be useless.

Machine cognition implies an autonomous machine knowledge acquisition. Therefore the work presented in this thesis falls to the general domain of development of autonomous knowledge acquisition system. In order to be called "autonomous", a knowledge acquiring system should satisfy the condition that it develops its own high-level (semantic) representation of facts from low level sensory data, such as image, sound, tactile perception etc. All this way of information processing from the "sensory level" to the "semantic level" should be performed solely by the machine, without human supervision. This however does not exclude human interaction, which is, on the contrary, vital for any cognitive system, be it human or machine. This may be seen on the example the troublesome effects, which the condition of deprivation of the presence of other persons has on so called "feral children" (McNeil, Polloway and Smith, 1984).

In accordance with the requirement of autonomy in context of the machine cognitive system and knowledge acquisition capacities, I set up the following objectives for the work that is developed throughout the present thesis:

- Explore and realize the state of the art of autonomous knowledge acquisition (cognition) performed by an embodied agent in real world environment.
- Understand basic principles of human cognition and draw an inspiration from them to conception of a machine cognitive system enabling the machine to be itself the initiator and the actor of the knowledge acquisition.
- Contribute to a conception of an intelligent system capable of autonomous knowledge acquisition from low level data by a) observation and b) interaction with its environment including humans; of a system capable of high-level representation of such a knowledge, which would be practicable in real world conditions (i.e. robust, real-time, …)

Especially the third objective is very challenging. To reflect the fact, that a doctoral study has a firmly given 3-years-long time-frame, I have focused on fulfilling the objective in a way, that, while still keeping its generality, tends to its application in the area of humanoid robotics. This is also the scope in which the final system has been tested.

Contribution

The work accomplished in this thesis has allowed for bringing several contributions to the conception of artificial autonomous cognitive systems:

- First, the state-of-the-art on autonomous acquisition of knowledge has been realized in various research domains and it has demonstrated the complexity of the problems being dealt with in this thesis. Giving an overview of existing solutions, the bibliographical study has notably pointed out problems related to previously proposed solutions. It has allowed for an objective evaluation of achievements concerning autonomous acquisition of knowledge by machines as well as an outline of open problems currently existing in this domain.
- The second contribution is the study, the conception and the realization of a low-level cognitive system based on the principle of perceptive curiosity. The approach is based on salient object detection realized by creation of a fast, real-

world robust algorithm for salient object detection and learning. This algorithm has several important properties that make it stand out from similar existing algorithms and which make it particularly suitable for use in the cognitive system developed in this thesis, such as real-time speed, self-tuned visual attention scale and robustness to difficult illumination conditions.

- The third major contribution of this thesis is the conception of a high-level cognitive system, based on a generic approach, which allows for acquisition of knowledge from observation and from interaction with the environment (including humans). Based on the epistemic curiosity, thesis high-level cognitive system allows a machine (e.g. a robot) to become itself the actor of its own learning. One important consequence of this system is the possibility of conferring multimodal cognitive capacities to robots in order to increase their autonomy in a real environment (human environment). Comparing with existing high-level cognitive systems, the present one bring important functionalities such as significantly less exposures needed for successful learning. It is also possible to generalize this approach in order to process, on high-level, as well absolute visual features (e.g. color, shape) as immaterial relative features (e.g. motion or position).
- The last major contribution of this work is the realization of the strategy proposed in the context of autonomous robotics. The studies and experimental validations done had confirmed notably that our approach allows increasing the autonomy of robots in real-world environment.

Thesis Organization

This thesis is constituted of five chapters, leading the reader from the state of the art and theoretical basis of my research, through conception of different parts of the proposed cognitive system up to its concrete application in real world environment.

Chapter 1 introduces the reader into the state of the art in different domains which are concerned by autonomous acquisition of knowledge. It discusses different works which inspired me and influenced me in my research. Among others it discusses works from the field of semantic simultaneous localization and mapping (SLAM), visual saliency detection and works accounting on human-robot communication and interaction in knowledge sharing. This chapter should give the reader a general overview on the state of the art, existing

techniques and terminology on which I further develop my own research described in chapters that follow.

In Chapter 2 theoretical foundations of my work are presented. A cognitive architecture is devised in order to cope with problems and objectives that have been discussed earlier in "Motivation and Objectives". This chapter provides a theoretical framework defining constitutive parts of my research work, which are subsequently concretized in chapters 3 and 4. It shows how the concept of curiosity is used within the presented cognitive system to motivate its actions and its seeking for new knowledge both on the low, sensory level and on the high, semantic level.

Chapter 3 is dedicated to the realization of what will be further explained as the "perceptual curiosity", i.e. a lower cognitive lever of the system. I propose a technique of doing this based on visual saliency and on detection of salient objects. The first part of the chapter describes a novel approach to salient object detection and extraction. Then a learning method for the extracted objects is presented. The salient object extraction method is further refined my automatic estimation of the visual attention scale. Finally several experiments are described validating the proposed salient object extraction and learning approach.

If the previous chapter described in some way "unconscious" or "low-level" cognitive mechanism of acquiring information about surrounding objects, Chapter 4 presents a comparatively higher-level cognitive function of the entire system, realizing the "epistemic curiosity" (explained in sub-section 2.2.2.1). It uses partly results of salient object learning and partly it relies on other mechanisms in order to present a general and flexible framework for knowledge acquisition and knowledge sharing between human and robot. It allows a robot to build up its own representation of the world from observation and from interaction with human beings found in the environment.

Concrete approaches for realization of different parts of the cognitive system have been developed in previous two chapters. It is the purpose of Chapter 5 to presents how they work together constituting a single system. The chapter reports on experiments made in real world indoor environment and it shows how the system allows us to make a step towards a fully autonomous acquisition of knowledge in machines. For didactical purposes the chapter also briefly remembers the reader basis of the most important algorithms and techniques used in realization of the system.

The closing chapter of this thesis is the General Conclusion. That is where the reader is given a summary conclusion and an evaluation of the research presented here. Finally, perspectives of possible future directions of the work are provided.

Chapter 1. State of the Art

1.1. Introduction

This chapter will help us, first, to get a better understanding of the main concepts with which I am dealing in this thesis. In addition, it will provide us with the terminology and existing approaches to which I am referring in next chapters. It will also enlighten achievements made so far on several different research fields that are implicated in the research presented here as well as the unsolved problems and difficulties that each of the domains is encountering in present days.

One of the most prominent keywords of this thesis is "knowledge acquisition". In the Oxford Dictionary of English[2], the word "knowledge" is defined as:

- Facts, information, and skills acquired through experience or education; the theoretical or practical understanding of a subject
- The sum of what is known
- Information held on a computer system.
- In philosophy: true, justified belief; certain understanding, as opposed to opinion.
- Awareness or familiarity gained by experience of a fact or situation

In this thesis the word "knowledge" is used predominantly in sense of the first and the last of the definitions given. Otherwise said, it is perceived as a sum of facts contributing to familiarity with the subject, which is acquired through proper experience (cf. Chapter 3 and Chapter 4) and education (cf. Chapter 4, especially section 4.4). The double aspect of knowledge acquisition, which is the "proper experience" on one side and the "education" on the other one are further investigated in the following Chapter 2.

Knowledge acquisition is not a research field *per se*. Rather it is an objective which different research domains are trying to attain by using different methods and starting from different theoretical foundations. In this work I am building on the work done in several domains including cognitive science, machine learning, machine perception and especially machine vision, linguistic and developmental psychology. What I propose in following

[2] Adopted from (Pearsall, 2010)

chapters is a fruit of an inspiration gained while studying these domains. Each of them deals, from its own point of view, with knowledge acquisition in machines, in humans and in mixed human-robot groups. It is thus well-founded if I conceive following sections of this chapter as a cross-section of the fields and of works that inspired me and that directed my attention and my efforts of accomplishing the research presented this thesis. Although certain of them have later proved to only be a dead-end, I include them too as I wish to provide the reader the same point of departure that I had, before diving in succeeding chapters.

As it has been stated previously, the problem of cognition – be it human or machine – is an extremely vast one and it is not the purpose of this thesis, nor it is permitted by its scope, to address it fully in its completeness. Instead of this, in the following section I will focus on research efforts that were in some way influential for my work or that are closely related to its subject. It therefore focuses on cognitive systems and the appearance of curiosity in existing works concerning specifically cognition in robots. The next section accounts on perception in context of perceptual curiosity and notably visual saliency. As autonomous learning requires the capacity of distinguishing the pertinent information from the impertinent one, visual saliency is one of the key techniques used here to extract important information in context of visual stimuli. Further, works concerning knowledge acquisition in human infants and works relating the biologically inspired knowledge acquisition techniques to artificial agents, i.e. mobile robots, are presented.

1.2. Cognitive Systems and Curiosity

As a departure point it is worth of mentioning the work of (Langley, Laird and Rogers, 2009). It brings an in-depth review on a number of existing cognitive architectures such as ACT-R, which adheres to the symbolic theory and reposes on the assumption that human knowledge can be divided to two kinds: declarative and procedural. Another discussed architecture is Soar, which is based on a system of if-then production rules or ICARUS, based again on two kinds of knowledge: concepts and skills … only to mention a few examples and to give the reader an idea about the heterogeneity in the field. The work also discusses challenges for their future development. Further in the work, a thorough discussion is provided on necessary capabilities of cognitive systems as well as on the properties that make such a system viable. The work written by (Vernon, Metta and Sandini, 2007) provides a survey on cognitive systems from another point of view. It accounts on

different paradigms of cognition in artificial agents notably on the contrast of emergent vs. cognitivist paradigms and on their hybrid combinations. In contrast to previously mentioned works, which accounts on cognitive architectures in a wide manner, the work of (Levesque and Lakemeyer, 2010), while still useful for its broad perspective, focuses on the area of research on cognition, which is much closer to the subject of this thesis, i.e. the cognitive robotics. The work discusses questions like knowledge representation in cognitive robots, sensing, reasoning and several other areas. However, there is no cognition without perception (a cognitive system without the capacity to perceive would miss the link to the real world and so it would be impaired) and thus autonomous acquisition of knowledge from perception is a problem that should not be skipped when dealing with cognitive systems. More importantly, what is the drive or the motivation for a cognitive system to acquire new knowledge? For human cognitive system (Berlyne, 1954) states, that it is the curiosity that is the motor of seeking for new knowledge. Consequently a number of works have been since there dedicated to incorporation of artificial curiosity into a variety of artificial systems including embodied agents or robots. However the number of works using some kind of curiosity motivated knowledge acquisition with implementation to real agents (robots) is still relatively small. Often authors view curiosity only as an auxiliary mechanism in robot's exploration behavior.

One of early implementations of artificial curiosity may be found in (Schmidhuber, 1991). In this work a model-building control system is extended by a form of artificial curiosity in order to learn to distinguish situations, in which the system has previously learned new information. This further helps the system to actively seek similar situations in order to learn more. On the field of developmental and cognitive robotics a partially similar approach may be found in the work of (Oudeyer, Kaplan and Hafner, 2007), which presents an approach including artificial curiosity mechanism (called "Intelligent Adaptive Curiosity" in the work). This mechanism drives the robot into situations where finding new information to learn is more likely. Two experiments with AIBO robot are presented showing that the curiosity mechanism successfully stimulates the learning progress and that it drives the robot out of situations where no useful or new information could be learned. Authors of (Macedo and Cardoso, 2012) implement the psychological construct of surprise-curiosity into the process of decision making while their agent is exploring an unknown environment. Authors conclude that the surprise-curiosity driven strategy outperformed classical exploration strategy regarding the time/energy consumed in exploring the entirety of the environment. Also the self-organizing multi-robot system accounted in (Kernbach et al., 2009) takes curiosity in consideration, but rather as a general bias towards explorative behavior without a

strict connection with robots cognition. The concept of surprise, which is closely related to curiosity, is exploited in (Maier and Steinbach, 2011) where a cognitive robot uses the surprise in order to discover new objects and acquire their visual representations. It is worth mentioning that the concept of curiosity is not bound only to cognitive systems and it has been successfully used in robotics e.g. for learning affordances in traversability task for a mobile robot in (Ugur et al., 2007).

The mentioned works (and it is also the case of this thesis) consider autonomy of the agent (robot) in performing cognitive tasks as the desired state. In this context, the work of (Vernon, 2011) is of interest as it addresses the problem of robot autonomy in contrast to the external control. Otherwise said, the author is attempting to respond the question: "how can an autonomous cognitive system be designed so that it can exhibit the behaviours and functionality that its users require of it" (Vernon, 2011: p. 1). The conclusion is that a cognitive system is trainable and will respond to instructions if the exploration and the social motivation are properly balanced in it.

1.3. Perception in Context of Perceptual Curiosity

In this work perceptual saliency is used in order to realize the "perceptual curiosity" mechanism. This section will deal with existing techniques on the field of visual saliency, which is itself a specialized form of perceptual saliency dealing with visual inputs. As such visual saliency is acting as a tool for extracting important visual information from the background and it allows for autonomous learning from observation. The following text will discuss on the basic principles and state of the art in the domain of visual saliency detection.

1.3.1. Basic Principles

Visual saliency (also referred in the literature as visual attention, unpredictability or surprise) is described as a perceptual quality that makes a region of image stand out relative to its surroundings and to capture attention of observer (from (Achanta et al., 2009)). The inspiration for the concept of visual saliency comes from the functioning of early processing stages of human vision system and is roughly based on previous clinical research.

In early stages of the visual stimulus processing, human vision system first focuses in an unconscious, bottom-up manner, on visually attractive regions of the perceived image. The visual attractiveness may encompass features like intensity, contrast and motion. Although

solely biologically based approaches to visual saliency computation do exist, most of the existing works do not claim to be biologically plausible. Instead, they use purely computational techniques to achieve their goal.

In this work visual saliency is used build a low level cognitive system. As such, visual saliency is acting as a tool for extracting important visual information from the background, a realization of "curiosity" on the sensory level.

1.3.2. Existing Techniques Overview

One of the first works to use visual saliency in image processing has been published by (Itti, Koch and Niebur, 1998). Authors there use a biologically plausible approach based on a center-surround contrast calculation using Difference of Gaussians. Other common techniques of visual saliency calculation published more recently include graph-based random walk (Harel, Koch and Perona, 2007), center-surround feature distances (Achanta et al., 2008), multi-scale contrast, center-surround histogram and color spatial distribution (Liu et al., 2011) or features of color and luminance (Achanta et al., 2009).

Figure 1: Top row: original images. Middle row: saliency maps for each photo. Bottom row: extracted salient object masks. Adopted from **(Cheng et al., 2011)**

In image processing, identification of visually salient regions of an image is used in numerous areas including smart image resizing (Avidan and Shamir, 2007), adaptive image display on small devices screens (Chen et al., 2003), amelioration of object detection and recognition (Navalpakkam and Itti, 2006), web image search intelligent thumbnail generation

(Wang et al., 2012), real-time pedestrian detection (Montabone and Soto, 2010), content based image retrieval and adaptive image compression or image collection browsing to mention only a few.

Depending on the particular technique, many approaches like (Achanta et al., 2009), (Achanta et al., 2008) or (Liu et al., 2011) output a saliency map, which is an image whose pixel intensities correlate with the saliency of the corresponding pixels of the original image. An example of this is shown on Fig. 1. Selection of the most salient regions from saliency map by application of a threshold or a segmentation algorithm is subsequently performed. It results into extraction of visually important object or a patch of objects rather than just of a semantically incoherent fragment of the image. This property is exploited by several authors. In (Borba et al., 2006) a biologically-motivated saliency detector is used along with an unsupervised grouping algorithm to group together images containing visually similar objects. Notably in the work (Rutishauser et al., 2004) a purely bottom-up system based on visual attention is presented, investigating the feasibility of unsupervised learning of objects from unlabeled images. Experiments are successfully conducted by its authors on real world high-resolution still images and on a camera-equipped mobile robot, where the capacity to learn landmark objects during its navigation in an indoor environment is shown. The main difference between this approach and the presented by us (cf. section 3.4) is that (Rutishauser et al., 2004) use visual saliency rather to indicate interesting parts of the input image, while in the present work it is used use explicitly for extraction of individual visually important objects. More recently (Frintrop and Kessel, 2009) has used a real-time method for salient object tracking on a mobile robotic platform. However, objects are learned here in a supervised manner with assistance of the operator.

A singular approach is presented in (Hou and Zhang, 2007), where the saliency of image regions is calculated based on spectral residues found on image when log-spectrum is calculated. The technique is very fast and authors show correct responses for natural images and psychological patterns. The method, however, is scale dependent and its authors provide only results at one scale claiming it yields best results for normal visual conditions. Another uncommon approach is described in (Liang et al., 2012). It uses content-sensitive hypergraph representation and partitioning instead of using more traditional fixed features and parameters for all the images. In this regard it exhibits the same per-image auto-tuning properties as my approach presented in section 3.6. However it works on completely different basis. From the description it results that it is incapable of extracting multiple salient objects at once without re-calculating the potential region of interest.

By contrast to somehow "exotic" mathematical approaches used in the last mentioned works, a recent publication by (Cheng et al., 2011) uses very simple and straightforward features. Nonetheless, its authors show it outputs full-resolution saliency maps with precise salient object masks and they claim the method outperforms consistently existing state of the art methods. The method is based on global contrast calculated using simple histograms. In addition spatial information is included using contrast between regions by sparse histogram comparison. The regions are generated using a graph-based image segmentation method. The approach has low computational complexity and saliency map is calculated very fast (approximately only two times slower than in (Achanta et al., 2009)). It is, however, unclear, what is the subsequent speed of the saliency cut, i.e. the method authors use to get saliency mask from the saliency map.

1.4. Autonomous Systems for High-level Knowledge Acquisition

1.4.1. General Observations

In recent years, there has been a substantial progress in robotic systems able to robustly recognize objects in real world using a large database of pre-collected knowledge (see (Meger et al., 2008) for a notable example). There has been, however, comparatively less advance in autonomous acquisition of such knowledge. In fact, if a humanoid robot is required to learn to share the living space with its human counterparts and to reason about it in "human terms", it has to face at least two important challenges. One, coming from the world itself, is the vast number of objects and situations, the robot may encounter in the world. The other one comes from humans: it is the richness of the ways that we use to address those objects or situations using natural language. Moreover, the way we perceive the world and speak about it is strongly culturally dependent. It is shown e.g. in (Kay, Berlin and Merrifield, 1991) regarding usage of color terms by different people around the world, or in (Bowerman, 1983) regarding cultural differences in description of spatial relations.

A robot, that is supposed to respond correctly to those challenges, cannot rely solely on a priori knowledge that has been given to it by a human expert. On the contrary, it should be able to learn on-line, in the place where it is used and by interaction with the people it

encounters there. On this subject, the reader may refer to (Kuhn et al., 1995) for a monograph on knowledge acquisition strategies, to (Goodrich and Schultz, 2007) for a survey on human-robot interaction and learning and to (Coradeschi and Saffiotti, 2003) for an overview of the problem of anchoring. This learning should be completely autonomous, but still able to benefit from interaction with humans in order to acquire their way of describing the world. This will inherently require that the robot has the ability of learning without an explicit negative evidence or "negative training set" and from a relatively small number of samples. This important capacity is observed in children learning the language (see e.g. (Regier, 1995)).

This section discusses first so-called Semantic Simultaneous Localization and Mapping (semantic SLAM). This technique is interesting as it allows autonomous acquisition of high-level knowledge from environment by a mobile robot. Further in this section state of the art in techniques concerning autonomous learning and human-robot interaction in the context of autonomous knowledge acquisition are presented.

1.4.2. Semantic SLAM

In this research domain techniques from the classical SLAM are combined with a higher-level (semantic) notion of the places and objects encountered, resulting in a form of autonomous high-level knowledge acquisition. This gives the robot performing semantic SLAM better capacities in understanding its environment and the human way of referring to it. However, as it will be shown, it relies often on rigid pre-programmed structures, while the aim of the present work is to propose a very flexible framework with as less of a-priori knowledge as possible.

1.4.2.1. Basic Principles

One of the latest research directions on the field of SLAM, the so-called *semantic SLAM*, is discussed here. While being so recent – virtually all the works linking semantics to SLAM belong to the current decade, many of them to the past five years – the concept itself may be perceived as a very important and pertinent one for future mobile robots, especially those who will interact directly with humans and perform tasks in human-made environment. In fact, it is the human-robot interaction, which is probably one of the main motives for passing "from zeros and ones to words and meanings" in robotic SLAM.

Semantics may be clearly incorporated into the concept of robotic localization and mapping in many different ways to achieve different goals. One aspect of this may be the introduction of human spatial concepts into maps. In fact, humans usually do not use metrics to locate themselves but rather object-centric concepts and use them for purposes of navigation ("I am in *the kitchen* near *the sink*" and not "I am on *coordinates [12, 59]*"). Moreover, the presence of certain objects is often the most important clue for human place recognition. An interesting work addressing the mentioned problems has been published in (Vasudevan et al., 2007), in which the world is represented topologically with a hierarchy of objects and place recognition is performed based on probability of presence of typical objects in an indoor environment. A part of this work shows a study based on questioning about fifty people with the aim to understand human concepts of self-localization and place recognition. It suggests that humans generally tend to understand places in terms of significant objects present in them and in terms of their function. A similar way (i.e. place classification by presence of objects) has been taken by (Galindo et al., 2005) where low-level spatial information (grid maps) is linked to high-level semantics via anchoring. During experiments, the robot was interfaced with humans and performed tasks based on high-level orders (e.g. "go to the bedroom") involving robots "understanding" of the concept of bedroom and usage of low-level metric map for path planning. However, in this work, object recognition is black-boxed and the robot is not facing any real objects in the experiments but only boxes and cylinders of different colors representing different real-world objects.

1.4.2.2. Object Recognition and the Semantic SLAM

An approach considering this "gap" between object recognition and semantic SLAM is presented in (Persson et al., 2007). Here, a system based on a mobile robotic platform with an omnidirectional camera is developed to map an outdoor area. The outcome is a semantic map of surroundings with buildings and non-buildings marked on it. In (Nüchter and Hertzberg, 2008), a more general system is presented, using a wheeled robot equipped with a laser 3D scanner. Authors show the ability of their robot to evolve in an indoor environment constructing a 3D semantic map with objects like walls, doors, floor and ceiling labeled. The process is based on Prolog clauses enveloping common knowledge about such an environment (i.e. the doors are always a part of a wall and never a part of the floor), which enable the robot to reason about the environment. Further in the paper, an object detection method using the laser range data is shown with a classifier able to distinguish and tag objects surrounding the robot like humans and other robotic platforms. In (Ekvall, Jensfelt and

Kragic, 2006) active object recognition is performed by a mobile robot equipped by a laser range finder and a camera with zoom. A semantic structure is extracted from the environment and integrated to robots map, allowing it to search objects in an indoor environment. Another object recognition technique is shown in (Meger et al., 2008) including an attention system. Based on recognized objects a spatial-semantic map is built. An inverse approach is presented in (Hertzberg et al., 2010), where a concept of anchored knowledge base is presented. By contrast to more common bottom-up approaches in semantic mapping, which build a geometry map with a set of tags, the proposed technique is based on instantiation of a knowledge base. It does so by providing sensor data and spatial information concerning instances of object and further it aggregates categories in the knowledge base, which finally results in anchoring of the perceived objects in the knowledge base. A more recent contribution from (Civera et al., 2012) presents yet another unconventional technique merging monocular SLAM and a structure-from-motion technique and object recognition techniques allowing insertion of pre-computed known objects into a standard point-based monocular SLAM map.

The left part of Fig. 2 is adopted from (Galindo et al., 2005) and it describes the way in which spatial and semantic information is commonly anchored together in semantic SLAM. On the right sub-image, originating from (Nüchter and Hertzberg, 2008), a 3D laser scan is presented with tags associated to different "meaningful" objects perceived on the scene.

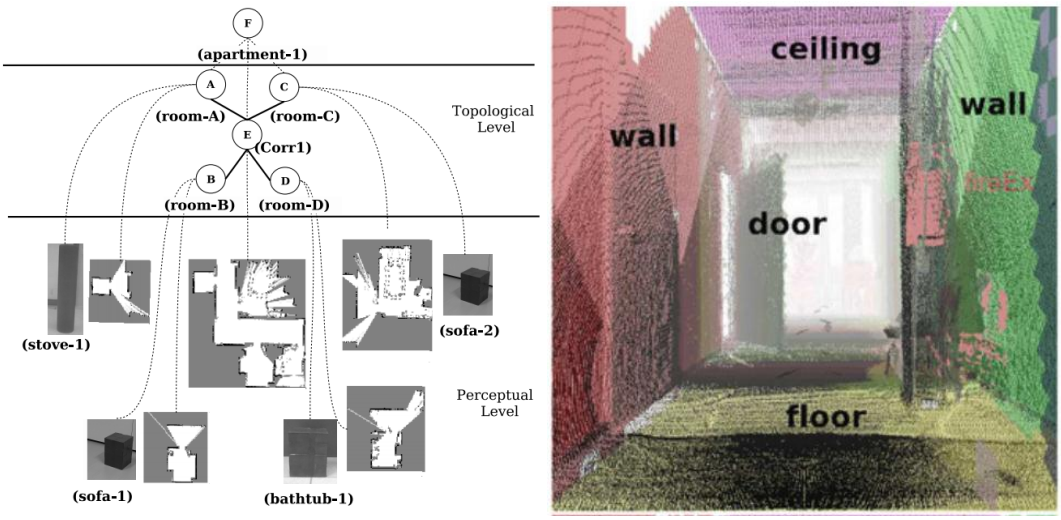

Figure 2: Examples of different representation of acquired sematic knowledge about the space in cases of two different Semantic SLAM approaches.

1.4.3. Autonomous Learning and Human-robot Interaction in Knowledge Acquisition

The problem of autonomous learning has been addressed on different degrees in previous works. For example, in (Greeff, Delaunay and Belpaeme, 2009) a computational model of word-meaning acquisition by interaction is presented. The work also discusses the problem of word-meaning acquisition in young children. In (Wellens, Loetzsch and Steels, 2008) authors present a computational model for acquisition of a lexicon describing simple objects. The model is verified in a population of humanoid robots. While it does not directly aim to learning by interaction with humans, it shows an interesting approach to autonomous forming of concepts in robots. On the other hand the work is interesting because it shows how the knowledge propagates in the population of robots in a way much resembling to what is happening in similar populations of humans. It would be very appealing to use the concepts proposed in the mentioned work to apply them on a mixed human-robot community using human language and to observe sharing of human knowledge with robots.

In (Saunders, Nehaniv and Lyon, 2010), a humanoid robot is taught by a human tutor to associate simple shapes to human lexicon in an interactive way. The interactive learning is explored also in (Griffith et al., 2009), where a robot is required to learn to distinguish two classes of objects. In (Lütkebohle et al., 2009), a humanoid robot is taught through a dialog with untrained user with the aim to learn different objects and to grasp them properly.

When robot learning is mediated through human-robot interaction, identification (verbal or nonverbal) of the referred-to objects is very important. See (Schauerte and Fink, 2010) for a recent contribution on joint attention in human-robot dialogs.

In the work of (Ogino, Kikuchi and Asada, 2006), a lexical acquisition model is presented combining more traditional approaches with the concept of curiosity to alternate the attention of the learning robot. A more advanced work on autonomous robot learning using a weak form of interaction with the tutor has been recently presented in (Araki et al., 2011). Its authors propose an online algorithm allowing a robot to perform multimodal categorization of objects with limited verbal input from human. Another interesting approach to autonomous learning of visual concepts in robots has been published in (Skocaj et al., 2011). Authors show capacity of their robotic platform to engage in different kinds of learning in interaction with a human tutor. The latter two mentioned works are to date perhaps the most advanced examples of autonomous acquisition of knowledge by observation and interaction in embodied agents, i.e. humanoid robots. Both approaches bear some

resemblance with what I describe of my approach in Chapter 4 and it is also in this chapter that some observations are given comparing my work to theirs.

1.5. Conclusion

This chapter has given the reader an overview of the fields and of the techniques that had an influence on my work and on which my work is based. It has familiarized the reader with important works and approaches that have been published in A) domains of low level knowledge acquisition concerning notably visual saliency, and of B) high-level knowledge acquisition techniques such as semantic SLAM and works that deal with autonomous learning and learning by human-robot interaction. Some critical observations have been made regarding the mentioned works. The aim was to discuss certain aspects occurring in state-of-the art works that are linked to problems that are dealt in the present thesis.

I would like to emphasize two observations regarding existing visual saliency techniques. The first is concerning psychological patterns. On Fig. 3 (adopted from (Hou and Zhang, 2007)) several forms of so-called psychological patterns are shown. These are synthetic images devised specially to test how the attention of a human (or a machine) vision is driven towards specific kinds of irregularities on the image. It should be further investigated which of the psychological pattern images, and to what extent, are processed in purely in the bottom-up manner in human vision system and which of them require on the contrary the top-down attention. Most of the existing approaches to visual saliency calculation focus on the bottom-up saliency without considering psychological or task-specific top-down feedback. It is therefore arguable whether correct response to psychological patterns should be considered as one of criteria in evaluation of quality of a given approach. Techniques like (Hou and Zhang, 2007), (Li et al., 2011) or (Sun, Yao and Ji, 2012) show correct responses to at least some of psychological patterns. On the other hand approaches like (Achanta et al., 2009) or (Cheng et al., 2011), which sometimes outperform the previously mentioned on natural image benchmarks, do not show psychological patterns response. The second observation is that many methods (e.g. like (Liu et al., 2011) or (Cheng et al., 2011)) work in a manner that directly presumes the existence of only one salient object on the image. While this presumption may be plausible for a number of applications, it does not hold for general natural images, especially in mobile robotic applications. On a scene the attention of the spectator is rarely attracted only by one object, but commonly several fixation points exist and the spectator fixes his attention successively on several objects in order of

their visual attractiveness (or saliency). It would be therefore reasonable to expect salient object detectors to output natively not only one, but several salient objects, if they are present on the examined scene.

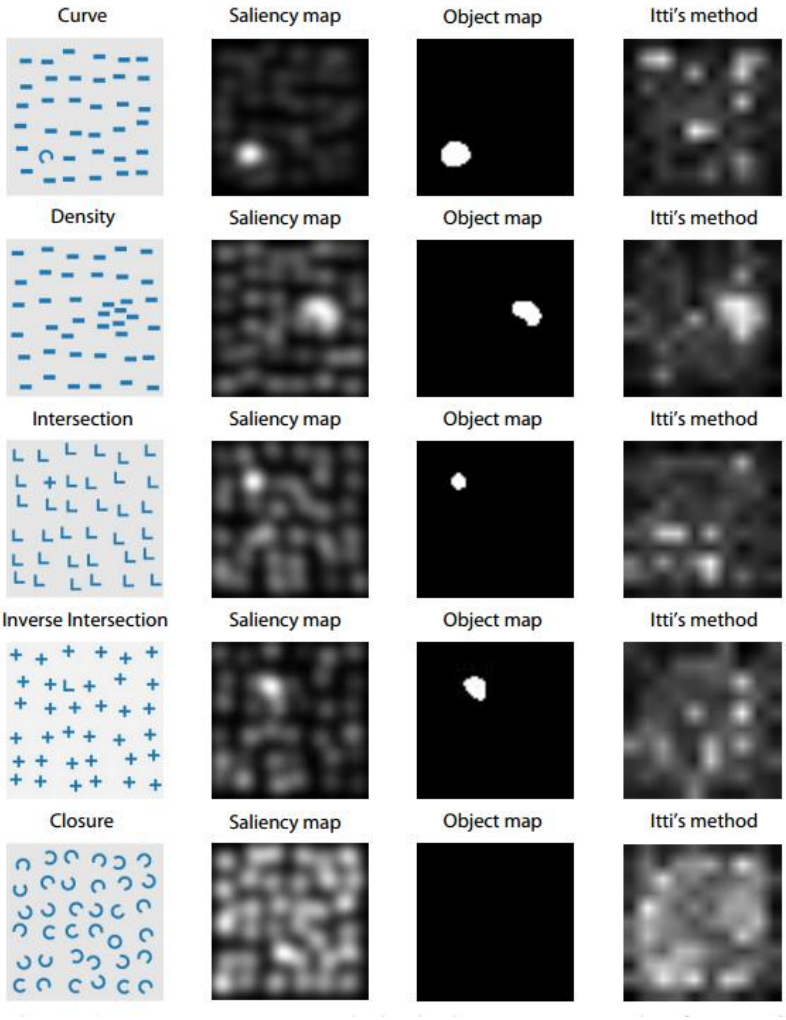

Figure 3: Various so-called psychological patterns (first column) with saliency maps and objects detected by **(Hou and Zhang, 2007)** (2nd and 3rd column). Finally the 4th column shows saliency map obtained by **(Itti, Koch and Niebur, 1998)**.

Concerning knowledge acquisition, in context of semantic SLAM, it should be stressed that in the predominant approach, seen in works like (Galindo et al., 2005) and (Vasudevan et al., 2007), to mention only a few, is to anchor the detected objects to a form of a hand-made taxonomy or ontology hierarchy (see Fig. 2). For example a room is called "kitchen" when a sink and a cooker is found inside. This approach by itself is a very pertinent one as it is coherent with the way that humans use to call places a name. However the

awareness of the fact that places e.g. with a sink and a cooker are usually called "kitchen" is most usually hard-coded by an expert and as such it is inflexible and incapable to accommodate to special properties of a particular environment. As a consequence the robot will produce consistently wrong behavior in situations like e.g. a dining hall that by accident contains also a sink and a cooker for re-heating of served meal. From the point of view of autonomous knowledge acquisition used to improve autonomy of mobile robots, an approach allowing learning of the "semantic" relation of the "kitchen" and the "sink" and "cooker", instead of having them hard-coded, would be much more flexible and also more natural. It is in this direction that my research is aimed; see Chapter 2 and Chapter 4.

Concerning the use of curiosity in machine cognitive systems, by observing the state of the art it may be concluded that the curiosity is usually used as an auxiliary, single-purpose mechanism, instead of being the fundamental basis of the knowledge acquisition. To my best knowledge there is no work to date which considers curiosity in context of machine cognition as a drive for knowledge acquisition on both low (perceptual) level and high ("semantic") level of the system, as it is done in this thesis.

The next chapter will be consecrated to the theoretical background of my work and to conception of the two-level cognitive system which has been outlined in General Introduction.

Chapter 2. Machine Cognition and Knowledge Acquisition as Foundations of the Presented System

2.1. Introduction

The previous chapter has given a frame of research initiatives in various fields contributing to the conception of systems with the capacity of turning low-level sensory data into high-level semantic information or knowledge. In the following chapter the main strategy to the solution of the problem of autonomous knowledge acquisition in context of the present thesis is presented. Step by step a theoretical scheme of a cognitive system is devised while respecting the objectives drawn in the General Introduction.

As already discussed in the Motivation, the conception of a human-like machine cognitive system represents an important step towards the autonomy of systems such as mobile robots. Indeed, over the last roughly five decades, autonomous robotics has been and continues to be subject to an ever increasing interest and has been the origin of numerous works and realizations. However, the performances of existing robots are still far from those of humans and in the 21st century robots will be supposed to share with humans' their living space (and vice-versa) and they won't be any more operated by skilled technicians. In fact, they will have to be self-sufficient enough to perform tasks in co-operation with their human users, who may have no a-priori technical skills. In order to achieve this, future works in the robotic field should focus on the increasing of robots' autonomy and indeed one of the major ways of contributing to this autonomy is research on machine cognition and, in relation to this, on autonomous knowledge acquisition. This is also the aim of the system described in this chapter and further developed in the rest of this thesis. The approach that is presented in this work is general. However, for the sake of feasibility in the course of three-years-long PhD studies I have focused particularly to its robotic application.

This chapter first discusses two main sources of inspiration for this work. First, it is a specific machine learning approach, in which the machine itself is an independent and active actor of its learning. Second, it is the concept of two distinct types of curiosity, the

"perceptual curiosity" and the "epistemic curiosity", and its implications for a conception of a two-level cognitive system.

2.2. Sources of Inspiration for Designing of the Envisaged Cognitive System

There are two main sources of inspiration that have played an important role in conception of the presented system. Each of them is explained in following sub-sections.

The first one is concerning different machine learning paradigms. I first discuss the traditional supervised and unsupervised approach, and then I describe the approach to machine learning taken in this thesis, which I call "Human-like learning".

The second source of inspiration concerns curiosity as a fundamental motivation for knowledge acquisition in cognitive systems. It should be stressed here that the use I make of curiosity in the present cognitive system is merely inspired by its biological function in human, but in no way I am attempting to model it in a biologically plausible way.

2.2.1. "Human-like learning"

In order to clarify the position of the machine learning approach used my work in context of other machine learning approaches, I first discuss the supervised and unsupervised learning. The focus is put here on roles of humans and machines it the process of knowledge acquisition and use of acquired knowledge.

2.2.1.1. Supervised Learning based Intelligent Systems

This kind of intelligent systems needs a human to provide features e.g. in form of labeled data, hand-segmented images or other. From this data the system is able to learn the intended function.

To illustrate it, let us take an example from the Viola-Jones detection framework. The framework is capable of learning to detect a class of objects (e.g. human faces). However, before the detector is capable to detect faces on an image, it has to be learned on a database of human face specimens extracted by hand by a human expert. This means that the features

have to be provided by the human and the machine is not able to collect them in an autonomous manner.

2.2.1.2. Unsupervised Learning based Intelligent Systems

In this kind of systems, the necessity of human in the feature extraction or data acquisition step is generally suppressed. They can learn from unlabeled data, thus features are acquired solely by the machine. The use of them is, too, performed by the machine.

For a concrete example, let us take a self-organizing map from (Kohonen, 1982). It takes features directly from the raw input. After the adaptation it can be used e.g. to cluster input patterns. An important point here is that the acquired knowledge is not directly intelligible to humans. It is so for two reasons. The first and the most obvious is that it is encoded into the system structures. In our example it would be the weights of neurons in the map. The second reason is that the system has, in fact, evolved without interference with humans and thus it does not share any substantial common ground with them. This makes the question of human intelligibility of such systems somehow obscure. All this does not mean that such systems are black-boxes with no connection with the surrounding world. It simply means that "nobody knows what happens inside" of such a system. To explain this point better, let us mention the work of (Wellens, Loetzsch and Steels, 2008), in which a community of humanoid robots is developing its proper language to describe objects they perceive. They do so using unsupervised learning techniques without any human interfering. The robots are thus fully autonomous in terms of cognition, meaning that they do not rely on humans, however as there is no interaction with humans, the terms they use to describe objects and their qualities is not immediately intelligible to humans.

2.2.1.3. "Human-like Learning" based Approach

By contrast to the previously mentioned approaches, the approach to machine learning taken in this thesis is characterized by the following principles:

- **Autonomous acquisition of features:** features, or, in general, information about the world, are, by contrast to the supervised learning, gathered autonomously. This on the other hand does not exclude their acquisition by means of communication with human (which contrasts to the paradigm of unsupervised

learning) as far as this is done in a seamless, natural way, i.e. as far as the intelligent system would pass the so-called Turing test.

- **Autonomous use of features:** the system handles the knowledge in an autonomous way, there is no need for human to supervise the course of their processing.
- **The knowledge is intelligible to both human and machine:** Internally it may be encoded in a "machine way", which is comparable to knowledge encoding in neural synapses in human brain. However, the intelligent system should be able to communicate his encoded knowledge to humans in human way and should in turn comprehend the knowledge communicated by humans in natural language, gestures etc.
- **The machine is itself an active actor of learning**: it does not passively rely on knowledge inserted by human, but it actively seeks new knowledge about the world A) by its observation and B) by knowledge sharing with other entities (including humans) and it decides independently what knowledge should be learned.

Concerning the last point: in the domain of supervised learning, an approach called Active learning exists (see (Settles, 2010) for a recent survey on the topic). While this approach bears some resemblance with the "Human-like learning" approach, notably by the fact that the machine actively queries a human in order to obtain labels for new features (data points), it does not fall completely into this domain. The ability of filling knowledge gaps or to precise uncertain knowledge is indeed a part of the proposed learning approach and it is explained more in detail in the following sub-section accounting for the curiosity. However, the notion of learning in "Human-like learning" is much broader than in the case of active learning. Notably it is the fact that learning is not performed only by querying new data from human, but also by observation. The observation here means bot observation of the state of the world and observation of human speech. The latter mentioned kind of observation is often found in children, who learn basic language concepts not only from direct discourse with adults, but even by a mere listening to discussions between adults that are not directed towards the child (cf. (Saffran et al., 1997)). Provided all this, the learning approach I use is probably closer to the active learning as a human educational technique (see e.g. (Bonwell and Eison, 1991) for a reference), than to the notion of this term in machine learning community.

Intelligent systems based on completely implemented "Human-like learning" clearly are, and for years will still remain, beyond the reach of state-of-the-art in the domain of artificial intelligence. However, it is precisely this direction, that is the most appealing and to which this thesis is making an unpretentious contribution. Classical and currently commercially used systems rely often on principles of supervised learning[3] or unsupervised learning[4]. By setting this work in context of what I call "Human-like learning", my aim is to contribute to conception of a system, where human being is no more an internal element, without which the system cannot work, but where it is an external element with which the machine cognitive system collaborates as an equal-to-equal.

2.2.2. Curiosity

2.2.2.1. Role of Curiosity

Curiosity is indeed an important factor both for human cognition and in conceiving an artificial system that gathers knowledge autonomously. To explain this affirmation, let us focus on curiosity in more depth.

In the introduction to his "Theory of human curiosity", Berlyne says, that *"Few phenomena have been the subject of more protracted discussion than human knowledge. Yet this discussion has usually paid little attention to the motivation underlying the quest for knowledge, with the result that two important questions still confront us. The first question is why human beings devote so much time and effort to the acquisition of knowledge"* (Berlyne, 1954: 180). The question is posed: why are humans keen to acquire knowledge? Why do people always seek new information, a more comprehensive knowledge? In the same work, the author proposes splitting up the curiosity into two kinds.

The first is so-called "perceptual curiosity", which leads to increased perception of stimuli. It is a lower level function, more related to perception of new, surprising or unusual sensory input. It contrasts to repetitive or monotonous perceptual experience.

The other one is called "epistemic curiosity", which is more related to the "desire for knowledge that motivates individuals to learn new ideas, eliminate information-gaps, and

[3] Take for example digital camera systems, which are capable of detection of human faces and of focusing on them.

[4] See e.g. Google Translator, whose operation is mostly given by unsupervised learning on extremely large corpuses of multi-language texts.

solve intellectual problems" (Litman, 2008). It also seems that it acts to stimulate long-term memory in remembering new or surprising (e.g. which is contradictory to what has been previously learned) information (Kang et al., 2009).

Without striving for biological plausibility, what has been previously said about the curiosity gives an important biological motivation for building of our system. On Fig. 4, a diagram is shown, which depicts the way in which curiosity stimulates acquisition of new knowledge and in turn the newly learned knowledge appeases or inhibits the curiosity. I adopt this as the basic working scheme for the envisaged system. By consequence, it is the curiosity, which motivates any action of the system.

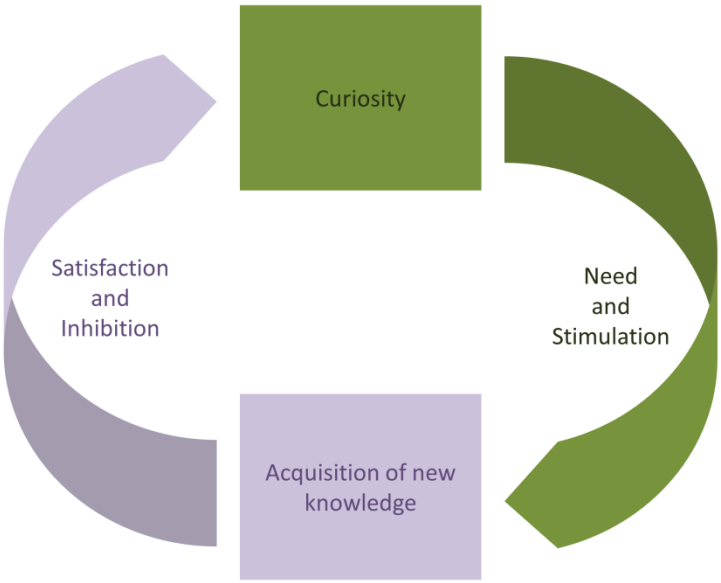

Figure 4: Diagram of the role of curiosity in stimulation of new knowledge acquisition in a cognitive system.

In conformity with the aforementioned concept of two kinds of curiosity, i.e. the "perceptual curiosity" and the "epistemic curiosity", I break down the process that has been shown on Fig. 4 to capture the role of both kinds of curiosity. This is shown on Fig. 5. On the left hand side of the figure, sample sensory data (an image) is shown. On this data the perceptual curiosity motivates or stimulates what I call the low level knowledge acquisition. It seeks "surprising" or "attention-drawing" information in given sensory data and thus devises of it a low-level knowledge. The task of the perceptual curiosity is realized by perceptual saliency detection mechanisms (see further in Chapter 3). This gives the basis for operation of high level knowledge acquisition, which is stimulated by the epistemic curiosity. Being previously defined as the process, that motivate to "learn new ideas, eliminate

information-gaps, and solve intellectual problems", the epistemic curiosity is here the motor of a) learning new concepts based on what has been gathered on the lower-level and b) eliminating information gaps by encouraging an active search for the missing information (see further in Chapter 4). Finally, this high-level (semantic) knowledge is stored and used when needed.

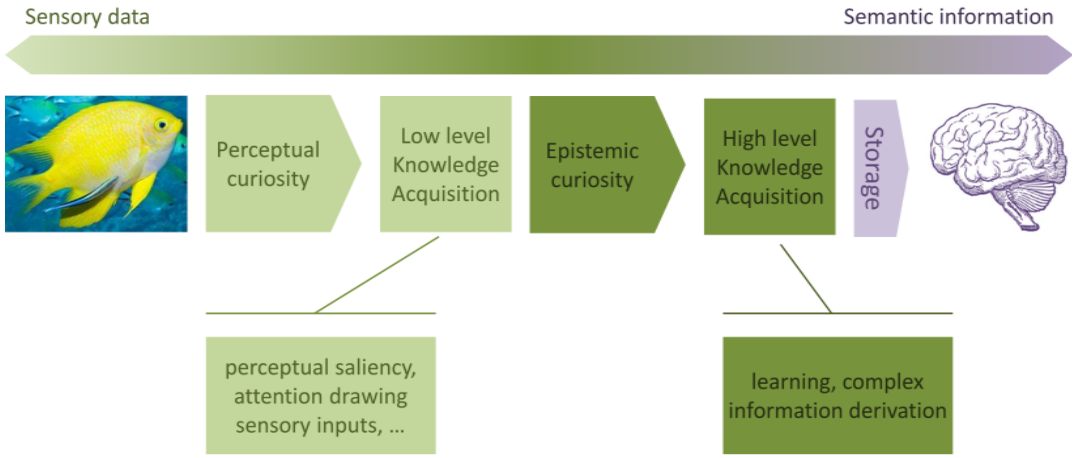

Figure 5: The place of the perceptual and the epistemic curiosity in learning of complex knowledge from raw sensory data.

2.2.2.2. Perceptual Curiosity Realization through Perceptual Saliency

In their perception, humans rely strikingly much on vision. It is then only pertinent to consider chiefly the visual information and learning processes connected to it. This sub-section focuses on this fundamental skill, a learning process based on visual perception, in relation to humanoid robots. Following the scheme from Fig. 5, this is the place, where the perceptual curiosity is realized through a perceptual saliency detection approach.

The design of perceptual functions is a major problem in robotics. Fully autonomous robots need perception to navigate in space and recognize objects and environment in which they evolve. However the question of how humans learn, represent, and recognize objects under a wide variety of viewing conditions presents a great challenge to both neurophysiology and cognitive research (Bülthoff, Wallraven and Giese, 2008). If we want an intelligent system to learn an unknown object from an unlabeled image, a clear need is the ability to select from the overwhelming flow of sensory information only the pertinent one.

And it appears appropriate to draw inspiration from studies on human infants and robots learning by demonstration. Experiments in (Brand, Baldwin and Ashburn, 2002) show that it is the explicitness or exaggeration of an action that helps a child to understand, what is important in the actual context of learning. It may be generalized, that it is the saliency (in terms of motion, colors, etc.) that lets the pertinent information "stand-out" from the context (Wolfe and Horowitz, 2004). This is supported by a number of existing works. For example (Zukow-Goldring and Arbib, 2007) are convinced that important variations in the input sensory signal make the child distinguish the pertinent information from the informational background. Similarly, experiments conducted in (Brand, Baldwin and Ashburn, 2002) show, that it is the explicitness or certain exaggeration of an action or presented information (in terms of voice, movement, color, etc.) that helps a child to understand, what is significant in the actual context of learning and what is unimportant. I argue that in this context the visual saliency may be helpful to enable unsupervised extraction and subsequent learning of a previously unknown object by a machine in a way that realizes the perceptual curiosity.

Further in this thesis, an approach is presented, enabling unsupervised real-time learning of objects from unlabeled images, and recognition of those objects when seen again. The cognitive visual architecture is given on Fig. 6.

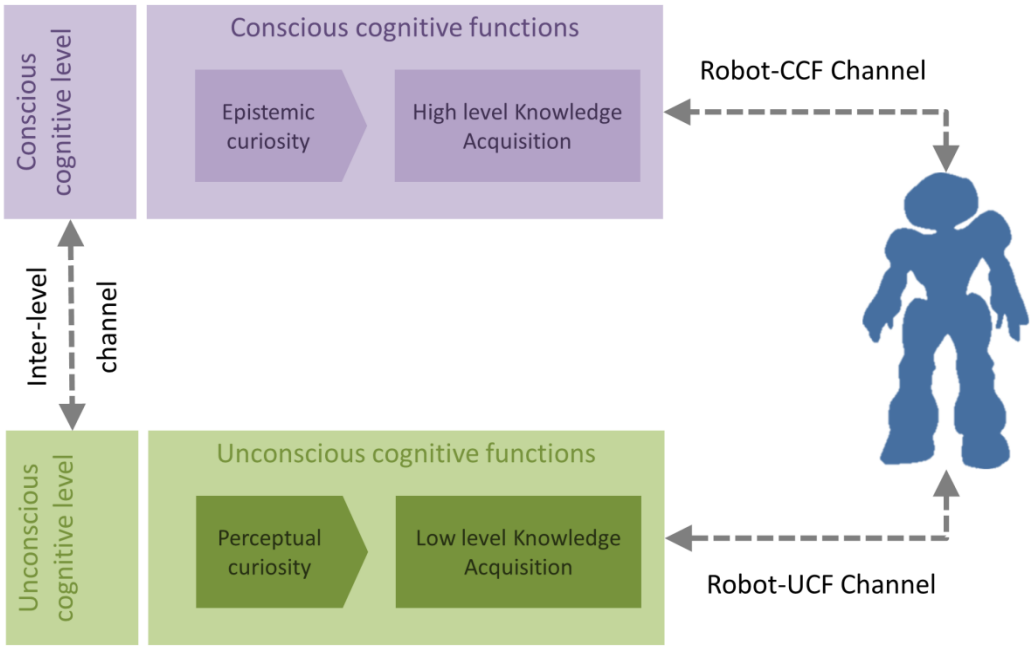

Figure 6: Block diagram of robot's cognitive architecture.

In the present work the term "cognition" is considered as human-like functionality (behavior) of humanoid machines (robots) and their autonomy. In (Madani and Sabourin,

2011), a multi-level cognitive machine-learning based concept for human-like "artificial" walking is proposed. This paper defines two kinds of cognitive functions: the "unconscious cognitive functions" (UCF: that is identified as "instinctive" cognition level handling reflexive abilities) and "conscious cognitive functions" (CCF: that is distinguished as "intentional" cognition level handling thought-out abilities).

The present approach is inspired by human vision system and by existing research on juvenile human (infants) learning process. The proposed approach extracts, first, objects of interest by means of visual saliency and secondly categorizes those objects using such acquired data for learning, which is identified as an unconscious cognitive function (UCF). Then a conscious cognitive function (CCF) is realized, on a higher level, by an intentional acquisition of new knowledge and seeking to fill informational gaps. This is discussed in the following sub-section.

2.2.2.3. Epistemic Curiosity Realization through Learning by Observation and by Interaction with Humans

In this sub-section the high level knowledge acquisition mechanism (cf. Fig. 5) is briefly described. It is this mechanism, which is stimulated by the epistemic curiosity in order to produce new semantic knowledge and to fill the gaps of missing knowledge. Contrary to the previously described perceptual curiosity apparatus, which is performed in an unconscious manner, the realization of the epistemic curiosity is inherently a conscious cognitive function, as it requires an intentional search and interaction with the environment. The mechanism allows an embodied agent (e.g. a humanoid robot) to learn to interpret the world, in which it evolves, using appropriate terms from human language. It is important to stress that this is done without making use of a priori knowledge. The task is realized by word-meaning anchoring based on learning by observation (see section 4.3) and by interaction with its human tutor (cf. sub-section 4.4).

The model is closely inspired by learning behavior of human infants (see e.g. (Yu, 2005) or (Waxman and Gelman, 2009)). The robot shares the world with a human tutor and interacts with him. The tutor on his turn shares with the robot his knowledge about the world in the form of natural speech (utterances), which accompany observations made by the robot. The goal of this system is to allow a humanoid robot to anchor the heard terms to its sensorimotor experience and to flexibly shape this anchoring according to its growing knowledge about the world.

The described system can play a key role in linking existing object extraction and learning techniques (e.g. SIFT matching or salient object extraction techniques) on one side, and ontologies on the other side. The former ones are closely related to perceptual reality, but are unaware of the meaning of objects they are treated, while the latter ones are able to represent complex semantic knowledge about the world, but, they are unaware of the perceptual reality of concepts, which they are handling.

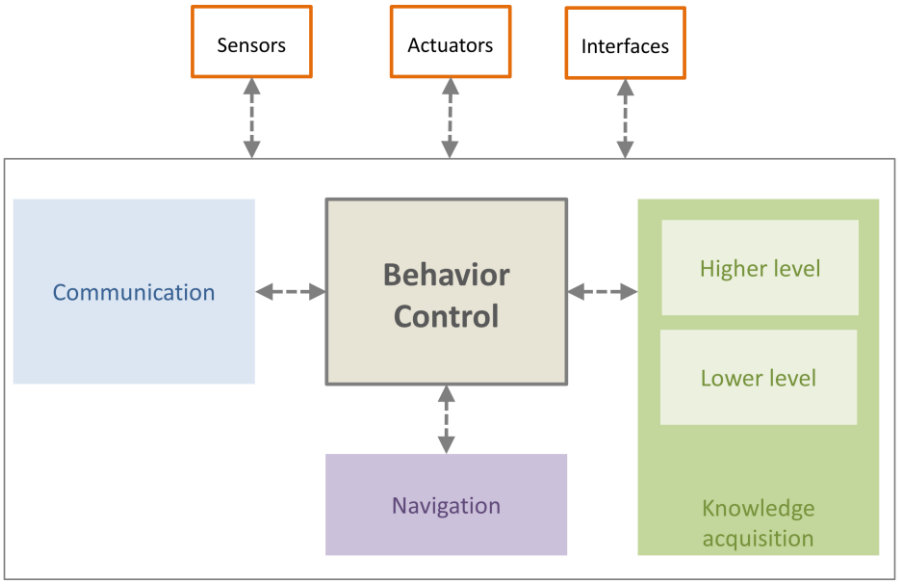

Figure 7: Block diagram of constitutive units of the system.

2.3. General Constitution of the System

Provided what has been told about the inspiration for conception of the cognitive system, more precision on its development are presented here. On Fig. 7, four main units of the system are identified. Their function is derived from the needs outlined earlier in sub-section 2.2.2 and from what has been previously said about the role of curiosity. The Knowledge acquisition unit's function is knowledge gathering and handling, derivation of high-level representation from low-level sensory data. Its higher level and lower level correspond to what is shown on Fig. 5.

The task of the Communication unit is to allow communicating this knowledge to the outer world and to handle inputs from humans and transfer them into a machine readable

form. It enables the system to communicate in two ways with other actors, be it similar intelligent machines or human beings.

An intelligent system that is not omnipresent and that is intended to evolve in the real world should be able to move freely in its environment in order to discover and interact with it. This is the purpose of the Navigation unit. Finally the Behavior Control Unit is acting as a controller. It supervises the interaction of other units and defines the system behavior logic, i.e. its reactions to external stimuli coming from the environment and its reactions to internal stimuli, coming from the "compulsions" of the machine itself (e.g. the "curiosity"). All units are connected to the sensors, interfaces and actuators that they use.

2.4. Conclusion

In this chapter the problem of autonomous knowledge acquisition has been defined. A cognitive system for autonomous knowledge acquisition has been proposed and put in context of an informal classification of existing intelligent systems.

The notion of curiosity has been inspected and it has been identified as the key motor of the mentioned cognitive system. This notion of curiosity helped to drawing some interesting inspiration for the system design. Especially the concept of the perceptual and the epistemic curiosity were related to two key mechanisms of knowledge acquisition. This has helped us to prepare the ground for following chapters 3 and 4, where both of the mentioned mechanisms are separately concretized and developed.

Chapter 3. Autonomous Detection and Learning Objects by means of Visual Saliency

3.1. Introduction

In accordance with what has been said in the previous chapter, in this chapter a system for lower level knowledge acquisition is presented. As the realization of the perceptual curiosity has been identified with the perceptual (especially visual) saliency, the chapter focuses on a visual saliency based autonomous technique for object detection and learning.

In the past decade, the scientific community has witnessed great advance on the field of techniques for object detection and recognition, such as SIFT (Lowe, 1999), SURF (Bay et al., 2008), Viola-Jones detection framework (Viola and Jones, 2004), color co-occurrence histograms (Chang and Krumm, 1999), to mention only a few. Many of them were so successful, that we are already meeting them in commercial applications like cameras focusing automatically on human faces or product logo recognition in mobile applications. While these methods show often high rates of recognition and are able to operate in real time, they all rely on human made databases of manually segmented or labeled images containing the object of interest without extensive spurious information and background. Some of the techniques use such a database as learning samples to learn e.g. a set of classifiers (Viola and Jones, 2004) others use it as a bank of templates for matching process (e.g. (Bay et al., 2008)). The mentioned database is sine qua non for a successful recognition process, but its manual creation often requires a considerable time and a skilled human expert. This impedes design of a fully autonomous machine vision system, which would learn to recognize new objects on its own.

Motivated by the mentioned shortcoming regarding existing object recognition methods, in this chapter an intelligent machine vision system is presented. It is able to learn autonomously individual objects present in real environment. Its key capacities are the following ones:

- Autonomous extraction of multiple objects from raw unlabeled camera images,
- Learning of those objects autonomously without human intervention

- Recognition of the learned objects in different conditions or visual contexts.

The goal for this system is to allow an embodied agent, e.g. a humanoid robot to learn to recognize objects encountered in its environment in a completely autonomous manner. The system itself is however not limited to mobile platforms and it can be very well used in context of sensor networks, intelligent houses etc. With respect to this envisaged goal, the system is designed with emphasis on on-line and real-time operation and we have validated it on a color camera equipped mobile robot in an explore-and-learn task performed in a real-world office environment.

In design of such a system, the approach was inspired partly by existing clinical investigations describing human vision system and partly by the way human infants learn objects in pre-lingual age. The extraction of objects of interest from raw images is driven by visual saliency. Building on existing work on the field of visual saliency, a novel salient object detection algorithm is proposed. It works in a spherical interpretation of RGB color space (cf. e.g. (Mileva, Bruhn and Weickert, 2007) and (Moreno, Graña and d'Anjou, 2011)), thus making use of photometric invariants. This, along with a fast image segmentation algorithm, which is robust to real-world illumination conditions, serve to extract image fragments containing objects. These image fragments are further used for learning. This way the perceptual (visual) saliency enables learning of objects from raw, i.e. unlabeled images without human supervision.

Resulting extracted objects can be exploited by most of the up to date object recognition methods. Here it is demonstrated how the present system performs when employing two fast recognition methods. It is the Speeded-up Robust Features (SURF) introduced in (Bay et al., 2008) and the Viola-Jones object detection framework, presented in (Viola and Jones, 2004). The machine learning aspects of this work have been specifically detailed in (Ramík, Sabourin and Madani, 2011).

The reminder of the chapter is organized as follows. Section 3.2 present constitutive parts of our system and explains their interaction. In section 3.3 the saliency detection approach is presented. In section 3.4, I detail on the salient object extraction technique and compare the quality of object extraction of this algorithm with other state-of-the-art approaches. Learning procedures are detailed in section 3.5. In section 3.6 details are given about our approach to estimation of visual attention scale. Section 3.7 reports and discusses results of validation of the proposed concepts. Conclusion and further work perspectives are presented in section 3.8.

3.2. General Overview of the Approach

The system I propose here consists of several units which collaborate together to accomplish the goal, which have been fixed in the previous section. On Fig. 8 a block-diagram of the system is depicted showing the individual units and their relations. Two main parts may be identified. The first one, labeled "Acquisition of new objects for learning" takes a raw image from the camera, detects visually important objects on it and extracts them so that they can be used as prospective samples for learning. These samples are then used in the section 3.5, where learning of the extracted objects is done and thus further recognition of those objects is made possible.

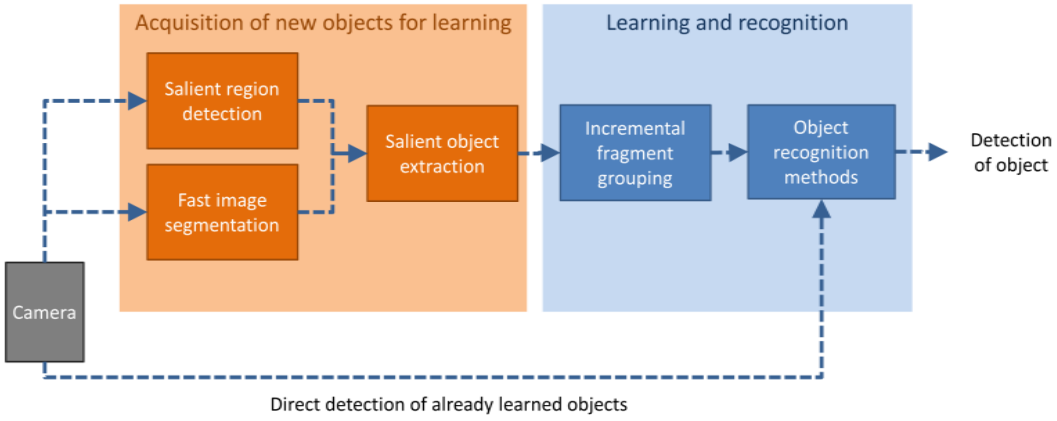

Figure 8: Block diagram of the system with the salient object detection unit and the learning unit.

Each of the two mentioned parts contains several processing units. In the first unit, as a new image is acquired by the camera, it is processed by the "Salient region detection" unit (described in section 3.3). Here, using features of chromaticity and luminosity along with local features of center-surround histogram calculation, a saliency map is constructed. It highlights regions of the image that are visually important, i.e. that are visually more salient with respect to the rest of the image. In parallel the input image is processed in the "Fast image segmentation" unit (sub-section 3.4.1), which splits the image into a set of segments according to the chromatic surface properties. The algorithm is shown to be robust to common illumination effects like shadows and reflections, which helps our system to cope with real illumination conditions. Finally the "Salient object extraction" unit (sub-section 3.4.2) combines results of the two previous, extracting the segments found on regions that

exhibit significant saliency and forming them together to present at the end salient objects extracted from the input image.

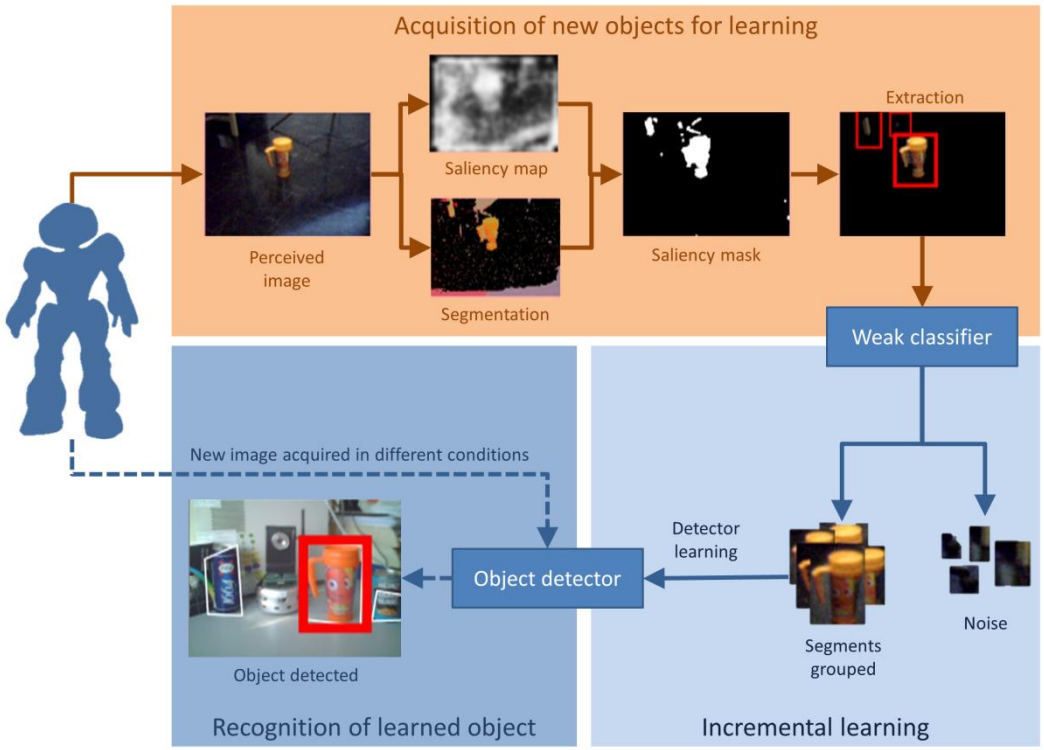

Figure 9: Overview of the entire proposed system's work-flow. An unknown object is incrementally learned by extracting it and providing it as learning samples to the object detector (solid arrows). This enables recognition of the object when encountered again (the dotted arrow).

As images are taken consecutively by the camera, salient objects extracted from each image are fed into the "Incremental fragment grouping" unit (sub-section 3.5.1). Here, an on-line classification is performed on each object by a set of weak classifiers and incrementally groups containing the same object extracted from different images are formed. These groups can be then used as a kind of visual memory of visual database describing each of the extracted objects. This alone could be enough for recognition of each of the objects, if it was ensured that each particular object will be found in the same visual context (i.e. in the context where the object is salient with respect to its surroundings) next time it is encountered by our system. This is clearly too restrictive for a system with a goal to recognize the once learned objects in any conditions. That is why the last unit of the system, tagged "Object recognition methods", is added, (sub-section 3.5). Its role is, by employing existing object recognition algorithms, to learn from the visual database built by "incremental fragment grouping" unit

and to recognize those objects regardless to their saliency in new settings. Thus for once learned objects, they can be recognized directly on the input image, which is denoted by the very bottom arrow on the Fig. 9 labeled "Direct detection of already learned objects". A different view on our system is presented on Fig. 8, where its work-flow is visualized.

3.3. Visual Saliency Calculation

3.3.1. A Spherical Interpretation of RGB Color Space

In the proposed saliency computation algorithm, colors are represented using a spherical interpretation of RGB color space (siRGB further on). This allows us to work with photometric invariants instead of pure RGB information. There are several works that explain and deal with the siRGB color space and photometric invariants, see (Mileva, Bruhn and Weickert, 2007) and (van de Weijer and Gevers, 2004). We are particularly interested in the correspondence between the angular parameters (θ, ϕ) and the chromaticity Ψ. The choice of siRGB over a more common HSV has been motivated by roughly 20% increase of performance of salient object extraction when using siRGB over HSV. Precise analysis of reasons of this increase are, however, beyond the scope of this thesis.

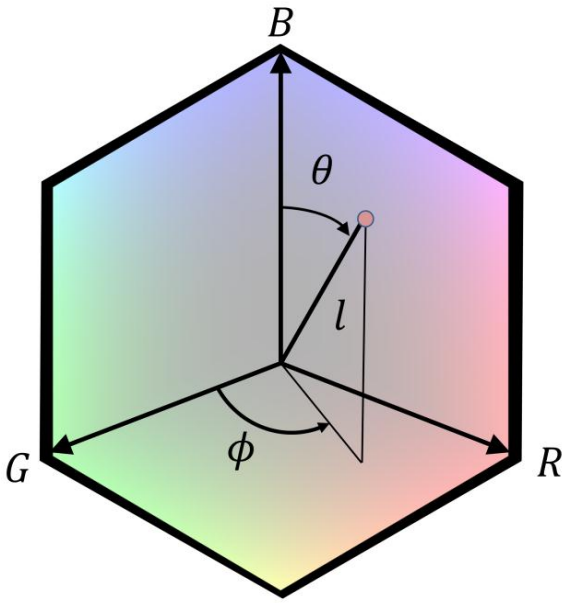

Figure 10: Diagram of relation between RGB and its spherical representation.

Any image pixel's color corresponds to a point in the RGB color space $c = \{R_c, G_c, B_c\}$. The vector going from the origin up to this point can be represented using spherical coordinates $= \{\theta_c, \phi_c, l_c\}$ (see Fig. 10), where θ, is zenithal angle, ϕ is azimuthal angle and l is the vector's magnitude (intensity). In RGB color space, chromaticity Ψ_c of a color point is represented by its normalized coordinates $r_c = \frac{R_c}{R_c+G_c+B_c}$, $g_c = \frac{G_c}{R_c+G_c+B_c}$, $b_c = \frac{B_c}{R_c+G_c+B_c}$, such that $r_c + g_c + b_c = 1$. That is, chromaticity corresponds to the projection on the chromatic plane Π_Ψ, defined by the collection of vertices of RGB cube $\{(1,0,0), (0,1,0), (0,0,1)\}$, along the line defined as $L_c = \{y = k \cdot \Psi_c; k \in \mathbb{R}\}$. In other words, all the points in line L_c have the same chromaticity Ψ_c, which is a 2D representation equivalent to one provided by the zenithal and azimuthal angle components of the spherical coordinate representation of the color point.

Given an image $\Omega(x) = \{(R, G, B)_x; x \in \mathbb{N}^2\}$, where x refers to the pixel coordinates in the image grid domain, he corresponding spherical representation is denoted as $\Omega(x) = \{(\theta, \phi, l)_x; x \in \mathbb{N}^2\}$, which allows us to use $(\theta, \phi)_x$ as the chromaticity representation of the pixel's color. For computational purposes, further the angle θ and ϕ and the value l is normalized into a range from 0 to 255.

3.3.2. Saliency Calculation in siRGB Color Space

In the present system, object of interest extraction is driven by the perceptual curiosity realized as the visual saliency. Therefore accurate and fast salient region detection is crucial for our system. Although there exist numerous approaches described in the literature, not all of them are suitable for this purpose. Often they lack precision or good resolution in frequency domain, are only able to extract the one most salient object from the image, or are computationally too heavy to be used in real-time. A comparison of some state-of-the-art algorithms in these terms may be found in (Achanta et al., 2009).

I propose a novel visual saliency detector composed of two independent parts, which can be computed in parallel. The first part captures saliency in terms of hybrid distribution of colors (i.e. a global saliency characteristic, sub-section 3.3.2.1). The second part calculates local characteristics of the image using a center-surround operation (sub-section 3.3.2.2). Their resulting saliency maps are eventually merged together using a translation function, resulting in the final saliency map. Images used in for evaluation throughout this work come

from MSRA Salient Object Database from (Liu et al., 2011) and the authors own data-set acquired by camera of Aldebaran Nao humanoid robot.

3.3.2.1. Global Saliency Features

For the first part, calculation of color saliency is done using two features: the intensity saliency (defined by Eq. (1)) and the chromatic saliency (defined by Eq. (2)). Here the saliency is defined as Euclidean distance of intensity l (or azimuth ϕ and zenith θ respectively) of each pixel to the mean of the entire image. Index l stands for intensity channel of the image, $\Omega_{\mu l}$ is the average intensity of the channel, similarly for azimuth ϕ and zenith θ in Eq. (2). Term (x) denotes coordinates of a given pixel on the image.

$$M_l(x) = \left\| \Omega_{\mu l} - \Omega_{\mathrm{l}}(x) \right\| \quad \textbf{(1)}$$

$$M_{\phi\theta}(x) = \sqrt{\left(\Omega_{\mu\phi} - \Omega_{\phi}(x)\right)^2 + \left(\Omega_{\mu\theta} - \Omega_{\theta}(x)\right)^2} \quad \textbf{(2)}$$

The composite color saliency map M is a hybrid result of combination of maps resulted from Eq. (1) and Eq. (2). Blending of the two saliency maps together is driven by a function of color saturation of each pixel. For this purpose, the color saturation C_c is defined. It is calculated from RGB color model for each pixel as pseudo-norm given by $C_c = \max[R,G,B] - \min[R,G,B]$ normalized to $0-1$ range. When C_c is low (too dull, unsaturated colors), more importance is given to intensity saliency (Eq. (1)). When C_c is high (vivid colors), chromatic saliency (Eq. (2)) is emphasized. As blending function we use the logistic sigmoid, so that the composite saliency map M is calculated following Eq. (3), where $C = 10(C_c - 0{,}5)$ in order to fit the logistic sigmoid.

$$M(x) = \left(\frac{1}{1+e^{-C}}\right) M_{\phi\theta}(x) + \left(1 - \frac{1}{1+e^{-C}}\right) M_l(x) \quad \textbf{(3)}$$

A similar feature as the one computed in Eq. (1) is used by (Achanta et al., 2009). However its authors use there only a single distance for all three channels, mixing chromaticity and intensity value of pixels together, while my approach respects the color saturation, which allows treating separately chromatic and achromatic regions. This is particularly helpful in cases where both chromatic and achromatic objects are present on the image.

3.3.2.2. Local Saliency Features

On Fig. 11 some resulting saliency maps of the presented algorithm are shown. Note that for the second image (leopard) the saliency map M (i.e. the global features, column b) does not highlight entirely the leopard's body. This image was selected to illustrate cases, where saliency consists in shape or texture of an object, which is distinct to its surroundings, rather than simply in its color. To capture this aspect of saliency, I compute the second (local) feature over the image: a center-surround difference of histograms (feature originally inspired by (Liu et al., 2011)). The idea is to go through the entire image and to compare the content of a sliding window with its surroundings to determine, how similar the two are. If the similarity is low, it may be a sign of a salient region. Let us have a sliding window P of size p, centered over pixel (x). Define a (center) histogram H_C of pixel intensities inside it. Then let us define a (surround) histogram H_S as histogram of intensities in a window Q surrounding P in a manner that the area of $(Q - P) = p^2$. The center-surround feature d is then given as

$$d(x) = \sum_i \left| \frac{H_C(i)}{|H_C|} - \frac{H_S(i)}{|H_S|} \right| \quad \textbf{(4)}$$

over all histogram bins (i). The $|H_C|$ and $|H_S|$ are pixel counts for each histogram allowing it to be normalized although a part of the windows is out of the image frame. In this case only pixels inside the image are counted. Calculating the $d(x)$ throughout all the L, ϕ and θ channels, we can compute the resulting center-surround saliency D on a given position (x) as follows in Eq. (5). To improve the performance of this feature on images with mixed achromatic and chromatic content, a similar approach of hybrid combination of chromaticity and intensity is used as the one described by Eq. (3). However, here the color saturation C refers to average saturation over the sliding window P.

$$D(x) = \left(\frac{1}{1 + e^{-C}}\right) d_l(x) + \left(1 - \frac{1}{1 + e^{-C}}\right) \max\left(d_\phi(x), d_\theta(x)\right) \quad \textbf{(5)}$$

In the column c of Fig. 11, sample center-surround saliency maps are presented. By using integral histograms described in (Porikli, 2005), all the mentioned histogram operations can be done very efficiently in constant time with respect to parameter p. This parameter permits moreover a top-down control of the attention and of the sensitivity of the feature in scale space. High p value with respect to the image size will make the feature more sensitive to large objects; low values will allow focusing to smaller objects and details. Note however,

that the experiments described further on were carried out with a constant value of p fixed on 0.4, unless stated otherwise.

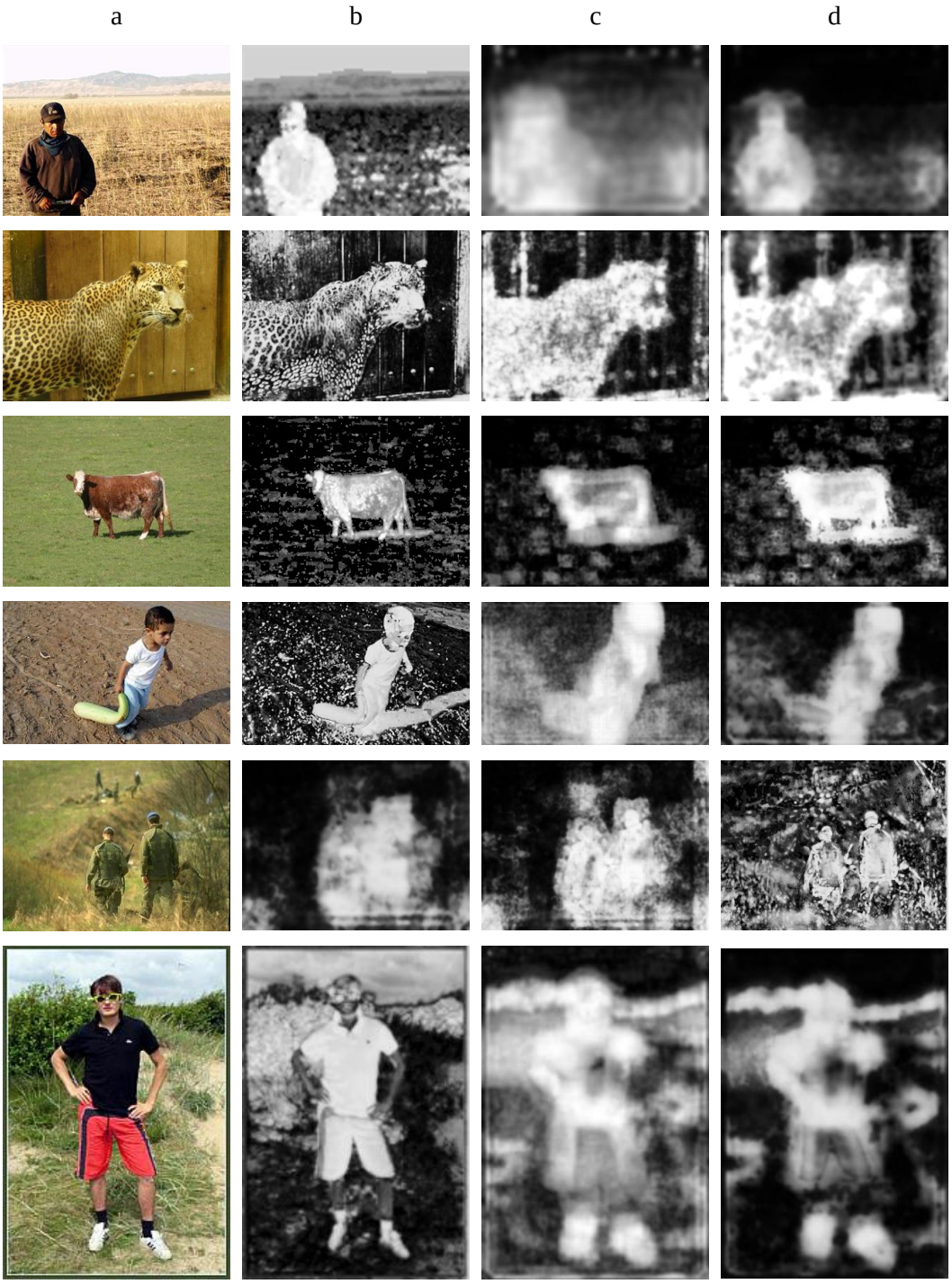

Figure 11: Sample saliency maps for different features. Column a: original images, b: composite saliency map $\boldsymbol{M}$, c: center-surround saliency $\boldsymbol{D}$, d: the final saliency map $\boldsymbol{M_{final}}$.

$$M_{final}(x) = \begin{cases} D(x) & if\ M(x) < D(x) \\ \sqrt{M(x)D(x)} & else \end{cases} \quad \textbf{(6)}$$

As the last step, both the global color saliency $M(x)$ from Eq. (3) and the local center-surround feature $D(x)$ from Eq. (4) are combined together by application of Eq. (6), resulting in the final saliency map M_{final}, which is then smoothed by Gaussian filter of size 3x3 pixels (for 320x240 pixels images). The upper part of the condition in Eq. (6) describes a particular case, where a part of image consists of a color, that is not considered salient (i.e. pixels with low $M(x)$) measure), but which is distinct to the surroundings by virtue of its texture. Several final saliency map samples are shown on the very right column of Fig. 11.

Regarding the features used for saliency map calculation, our algorithm belongs to the group of saliency detection approaches, which are not able to cope with psychological patterns like "curve", "intersection", "closure" etc. However, we do not perceive this as a shortcoming as our algorithm is primarily aimed for processing natural images and not to mimic precisely human psychological or vision system. This issue has been already discussed in sub-section 1.5.

3.4. Salient Object Extraction

Having the saliency map of the input image computed, we can proceed to extraction of visually salient objects themselves. A manual fixed-value thresholding on the final saliency map and automatic thresholding using the Otsu's method from (Otsu, 1979) have proven themselves as impracticable as well as other statistics based methods that have been applied on the saliency map. The problem is that all these methods work only over the saliency map and do not take into account the original image. Given this observation, I have decided to first apply a segmentation algorithm on the original image to obtain coherent parts of it and then to extract only those segments that are salient enough.

3.4.1. Fast Image Segmentation in siRGB Color Space

There are sophisticated techniques for image segmentation like growing-neural-gas approaches applied in real time, such as (García-Rodríguez and García-Chamizo, 2011) or

(Angelopoulou et al., 2008), however here the focus will be given to reflectance physics properties of the image for the following segmentation process.

3.4.1.1. Main Segmentation Problems

Image segmentation can be defined as a process which divides an image into different regions such that each region is homogenous, but the union of any two adjacent regions is not homogeneous. A formal definition of image segmentation is given in (Fu and Mui, 1981). According to this work: If W is a homogeneity predicate defined on groups of connected pixels, then segmentation is a partition of the set F into connected subsets or regions $(S_1, S_2, \dots, S_n)$ such that with $\bigcup_{i=1}^{n} S_i = F$ and $\forall i \neq j, S_i \cap S_j = \emptyset \; and \; \forall x, y \in S_i; W(x) = W(y)$.

There are four main problems in image segmentation: these are problems derived of a) the illumination, b) noise effects, c) edge ambiguity and d) the computational cost. The first three problems are closely related.

In segmentation processes the use of a suitable distance measure is very important. Therefore a hybrid distance is introduced, which works with intensity and chromaticity. On one hand, this hybrid distance allows parameterization of noise tolerance and on the other hand, we can adapt this distance for optimal edge detection. Furthermore, this distance is grounded in the dichromatic reflection model from (Shafer, 1985) by a spherical interpretation of the RGB color space from (Moreno, Graña and d'Anjou, 2011). So, this approach helps to avoid the first mentioned problem as well. Finally, in this method only 4 neighboring pixels will be used, instead of the full 8 pixel neighboring. This helps to decrease the computing time. The presented segmentation algorithm has thus the following properties: a good behavior in shadows and shines, avoids effect of noise and finally it is cheap in terms of computing time.

3.4.1.2. Distance

A distance based in the spherical interpretation of the RGB color space is proposed here. Given an image $\Omega(x) = \{(R, G, B)_x; x \in \mathbb{N}^2\}$ where x refers to the pixel coordinates in the image grid domain, the corresponding spherical representation is denoted as $\Omega(x) = \{(\phi, \theta, l)_x; x \in \mathbb{N}^2\}$, which allows us to use $(\phi, \theta)_x$ as the chromaticity representation of the pixel's color.

Empirical experiments tell us that intensity is the most important clue in overly dark regions, and that on the other hand it is better to use the chromaticity component when the illumination is good. Like in previous works of (Moreno, Graña and Zulueta, 2010) a hybrid distance is proposed. Fig. 12 shows the chromatic activation function. For values less than a, the chromatic component is inactive, for values that belong to the interval $[a, b]$, we take into account the chromatic component from its minimum energy to its maximum energy c by following a sinusoidal shape. Finally for values bigger than b its energy is always c. The three parameters a, b, c are in the range $[0,1]$. The region under the green line is the chromatic importance and its complementary, the region over this line is the intensity importance.

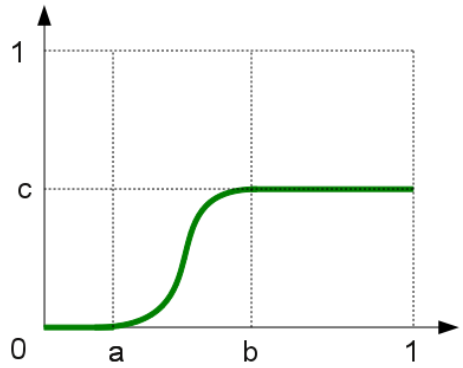

Figure 12: Chromatic activation function $\boldsymbol{\alpha(x)}$.

The function $\alpha(x)$ depends of the image intensity. Its complementary function $\bar{\alpha}(x)$ is the intensity activation function where $\bar{\alpha}(x) = 1 - \alpha(x)$ and hence $\bar{\alpha}(x) + \alpha(x) = 1$. The below equation is the mathematical expression of $\alpha(x)$.

$$\alpha(x) = \begin{cases} 0 & x \leq a \\ \dfrac{c}{2} - \dfrac{c}{2}\sin\left(\dfrac{(x-a)\pi}{b-a}\right) & a < x < b \\ c & x \geq b \end{cases} \quad \textbf{(7)}$$

Using this expression we can formulate a hybrid distance between any two pixels p, q on an image as follows:

$$d_h(p,q) = \bar{\alpha}(p,q) \cdot d_l(p,q) + \alpha(p,q) \cdot d_\Psi(p,q) \quad \textbf{(8)}$$

In Eq. (8), the relationship between $\alpha(x)$ and $\alpha(p,q)$ is given by $x = \frac{l_p + l_q}{2}$, where l_p and l_q are the intensities l in spherical coordinates. The distance d_l is an intensity distance defined by Eq. (9) and the chromaticity distance d_Ψ is finally defined by Eq. (10).

$$d_l(p,q) = |l_p - l_q| \tag{9}$$

$$d_\Psi(p,q) = \sqrt{(\phi_q - \phi_p)^2 + (\theta_q - \theta_p)^2} \tag{10}$$

3.4.1.3. Segmentation Method

All of the previously described techniques are joined here covering the four desired goals. On one hand we are going to use the spherical interpretation of the RGB image, and on other hand we are going to use the aforementioned hybrid distance expressed in the Eq. (8).

For edge detection a formal gradient is not necessary because it can be calculated "ad-hoc" using the hybrid distance and a threshold. In fact this method is focused to detection of homogeneous regions. When the distance between some pixels is less than this threshold we are going to admit that these pixels are homogeneous and then they belong to the same region. Homogeneous connected regions are easily identified because all of them have the same label. The method is explained by the following algorithm.

3.4.1.4. Algorithm

This algorithm returns a bi-dimensional integer matrix of labels. For computation of this algorithm a structure is also needed that relates each label with a chromaticity and the number of pixels labeled with it. That is necessary because each time a new pixel is assigned to a label the chromaticity of this label has to be actualized. This is given by the mean chromaticity of all pixels labeled with it.

The most important parameter for this algorithm is the threshold δ. Both the granularity and the noise tolerance depend on this parameter. For a very small value we will obtain a lot of small regions, and on the contrary, by using a high value we obtain several large and visually more important regions. On the other hand the parameters a, b and c used

in Eq. (8) allows to adjust the distance type. If $b = 0$ and $c = 1$ it is a purely chromatic distance. If $a = 1$, it is a purely intensity distance. In other cases it is a hybrid distance.

The Algorithm 3 gives details of our method. For the sake of fluidity of the text, in its formal version it is placed in Appendix A. In this algorithm $L(x)$ denotes the label of pixel x and $L_4(x)$ denotes the set of labels of the neighbors of pixel x. This can be expressed as $L_4(x) = \bigcup_{x' \in N_4(x)} L(x')$, where $N_4(x)$ are the 4 neighboring pixels of pixel x. The algorithm may be applied to any color image $\Omega(x)$. There is a need for specification of the distance $d_H(x, y)$, which provides a measure of the similarity between pixel colors $\Omega(x)$ and $\Omega(y)$. To label the regions we keep a counter R, and we build up a map Ψ_R assigning to each region label a chromatic value. We also define a counter C_R counting the number of pixels in the image region of a particular label R.

On Fig. 13, several sample results of the described segmentation algorithm are shown. First two scenes have been arranged to contain important highlights and reflections on reflective surfaces. The second two cases are natural images containing shadows. The way how the algorithm reacts to difficult illumination conditions is presented. Correct segmentation results are obtained even in presence of strong directional light, reflections and shadows.

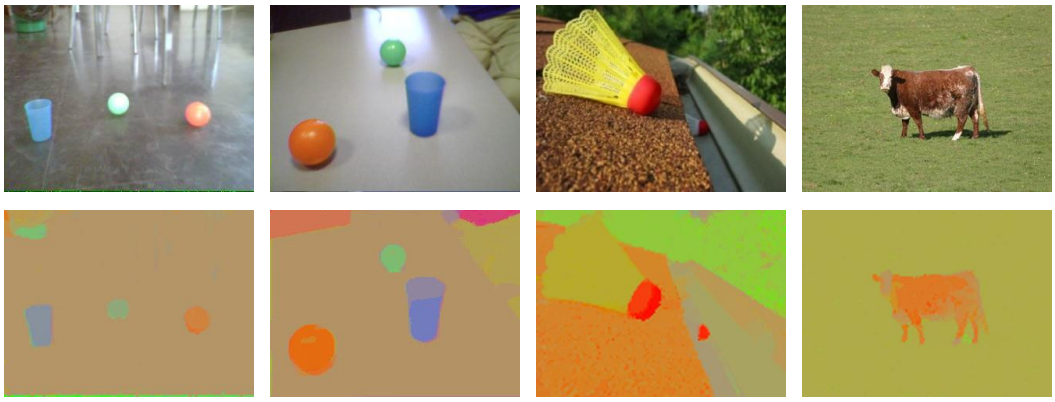

Figure 13: Sample results of the present segmentation algorithm. Top row: original images, bottom row: the resulting segmentation.

3.4.2. Extraction of Salient Objects using Segmented Image

The segmentation described by Algorithm 3 (in Appendix A) splits an image into a set of chromatically coherent regions. Objects present on the scene are composed of one or

multiple such segments. For objects that conform to conditions of "explicitness" discussed in 2.2.2.2, the segments forming them should cover areas of saliency map with high overall saliency. On the other hand visually unimportant objects and background should have this measure comparatively low.

$$\forall S_i \in \{S_1, S_2, \dots S_n\}; \forall \Omega(x) \in S_i;$$
$$\Omega(x) = \begin{cases} 1 & if\ \overline{S_i} > t_{\bar{S}}\ and\ Var(S_i) > t_{Var} \\ 0 & otherwise \end{cases} \quad \textbf{(11)}$$

The input image is thus segmented into connected subsets of pixels or segments $(S_1, S_2, \dots S_n)$. For each one of the found segments $S_i \in \{S_1, S_2, \dots, S_n\}$ its average saliency $\overline{S_i}$ is computed over the saliency map M_{final} as well as the variance of saliency values $Var(S_i)$. All the pixel values $\Omega(x) \in S_i$ of the segment are then set following Eq. (11), where $t_{\bar{S}}$ and t_{Var} are thresholds for average salience and its variance respectively. The result is a binary map containing a set of connected components $C = \{C_1, C_2, \dots, C_n\}$ formed by adjacent segments S_i evaluated by Eq. (11) as 1. To get rid of noise, a membership condition is imposed that any $C_i \in C$ has its area larger than a given threshold. Finally, the binary map is projected on the original image, which gives as a result parts of the original image containing its salient objects.

For our experiments, we set $t_{\bar{S}}$ to 50% of the maximal possible saliency, t_{Var} to 20 and the minimal area to 1% of total image area.

3.4.3. Salient Object Extraction Results

On Fig. 14, sample results of my salient object detection algorithm are compared with ground truth and two others state of the art algorithms. I have chosen to compare with the work presented in (Achanta et al., 2009) and (Liu et al., 2011) as the first one presents a fast algorithm potentially suitable for real-time application in machine vision, while the latter one shows high performance in terms of precision and correctness. No claims are made by authors of the latter one about its speed, but with respect to the description of algorithm provided in (Liu et al., 2011) it may be assumed that it is not suitable for a real-time application.

Shown results illustrate the typical performance of presented algorithms. Although (Achanta et al., 2009) is computationally very cheap (saliency map calculation takes about 45ms on a 320x240px image), its results vary largely in quality depending on the nature of

salient objects on the image. Algorithm of (Liu et al., 2011) produces results much close to human perception and more precise in terms of resolution (sample results are published online[5]). However, it suffers from two major drawbacks in context of the learning system presented here. It does not claim to be applicable in real time, and more importantly it outputs only one salient object (i.e. the most salient one) at time, although authors suggest for future work a workaround to this using inhibition-of-return technique. However this would come with even more increased expenses in terms of time.

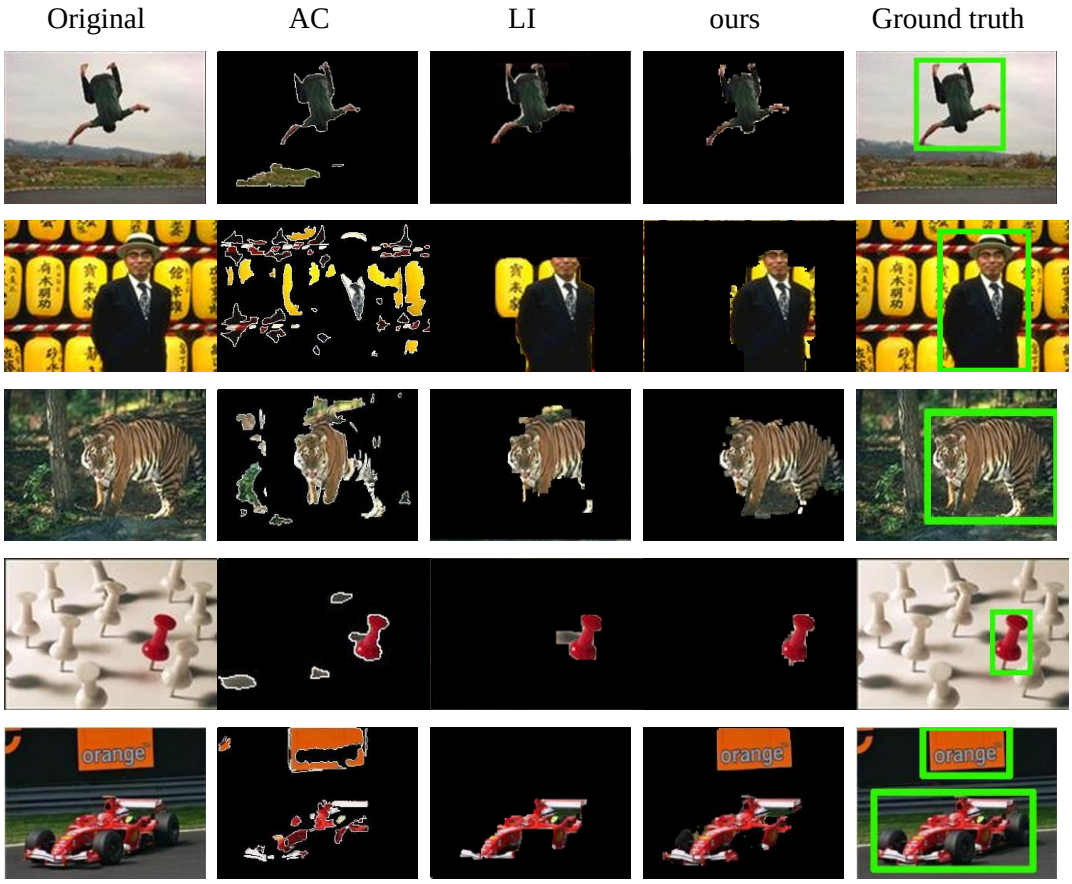

Figure 14: Comparison of different salient object detection algorithms. First column: original image. Second column: results of the approach of **(Achanta et al., 2009)**. Third column: results of the approach of **(Liu et al., 2011)**. Fourth column: results of our approach. Last column: ground truth (taking into account multiple objects in the scene).

On the other hand our approach outputs natively multiple salient objects if they are present on the image. An illustrative example may be found on Fig. 14 the last row, where

[5] accessible on Microsoft Research website: http://research.microsoft.com/en-us/um/people/jiansun/salientobject/submitted_1303/index.htm

two visually attractive objects are found on the same image: the F1 racing car and the "orange" logo. As they are both highly salient and clearly distinct in terms of their position on the image, our algorithm marks them both as visually salient. This property appears to be crucial while extracting unknown objects for learning as there is no reason why only the most salient object should be considered. This is especially true in real conditions with highly structured environment and many objects present in the field of view.

The present algorithm has been tested against the benchmark on MSRA Salient Object Database[6]. The scores that have been obtained are resumed on Table 1. While these results are close to the results obtained by Liu et al. (2011) (the F-measure differs from the Liu et al. (2011) only by about 0,05), our algorithm brings the benefit of high-speed processing and native output of multiple salient regions, if they are present on the image.

	Precision	**Recall**	**F-measure**
Data-set A	0,73	0,75	0,74
Data-set B	0,75	0,76	0,75

Table 1: Scores obtained by our salient object detection algorithm on the MSRA dataset.

Values of precision, recall and the F-measure were calculated following the appropriate equations given on Eq. (12) (adapted from (Liu et al., 2011)) and with parameter α set to 0,5. In the equation, g_x stands for the ground truth for the x-th pixel and a_x stand for the label (salient/non-salient) given by the algorithm for the x-th pixel of the image.

$$\begin{aligned} Precision &= \frac{\sum_x g_x a_x}{\sum_x a_x} \\ Recall &= \frac{\sum_x g_x a_x}{\sum_x g_x} \\ F-measure &= \frac{(1+\alpha)\cdot Precision \cdot Recall}{\alpha \cdot Precision + Recall} \end{aligned} \quad \textbf{(12)}$$

In terms of average speed, on 320x240px the method of Achanta calculated the saliency map in 45ms, but takes another 2900ms per image to extract salient segments using

[6] available on: http://research.microsoft.com/en-us/um/people/jiansun/salientobject/salient_object.htm

mean-shift segmentation[7]. Our algorithm in its non-optimized version takes in average 100ms per image (saliency map and image segmentation are calculated in parallel as they are two independent processes), which allows us to run it on speed about 10 frames per second. All algorithms were run on an Intel i5 CPU at 2,25Ghz machine.

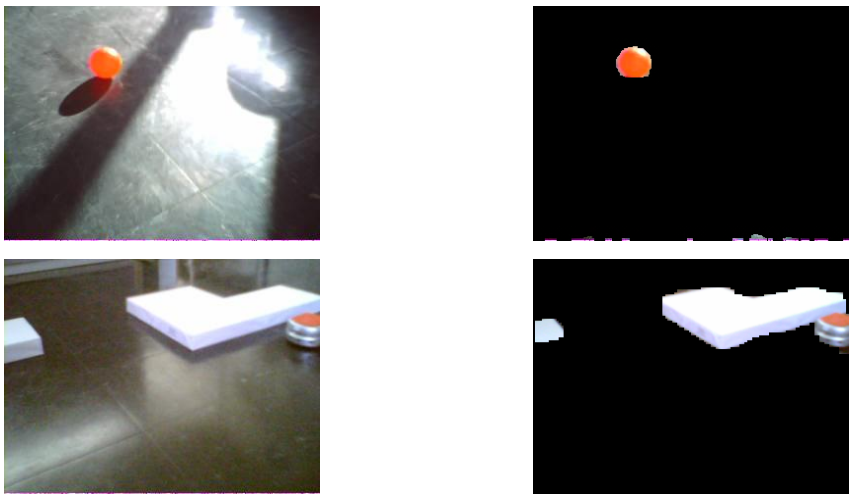

Figure 15: Particular cases of our algorithm in conditions of strong illumination which is causing reflections and shadows. Left: original images. Right: extracted salient objects.

The described algorithm has been run on some specifically selected images in order to illustrate how it is able to of coping with certain particular illumination conditions or cases. On Fig. 15 it is first shown how the algorithm copes with difficult illumination conditions in presence of strong directional light. The red ball is extracted correctly and the strong reflection is not marked as a salient object as the shine does not change the chromatic property of the surface. A similar case can be noticed on the following image, where a Khepera robot is shown on a reflective floor.

In case of Fig. 16, we can observe changes in parameter p from Eq. (11). In the middle column, the saliency is focused towards larger objects (high p), extracting mainly the human head and the entire American flag. On the other hand in the second case the emphasis is put on smaller details (low p), which allows extraction of eyes, mouth, hair and other small details on the face image, or star, stripes and windows on the flag image.

[7] Based on executable available online: http://ivrg.epfl.ch/supplementary_material/RK_CVPR09/index.html

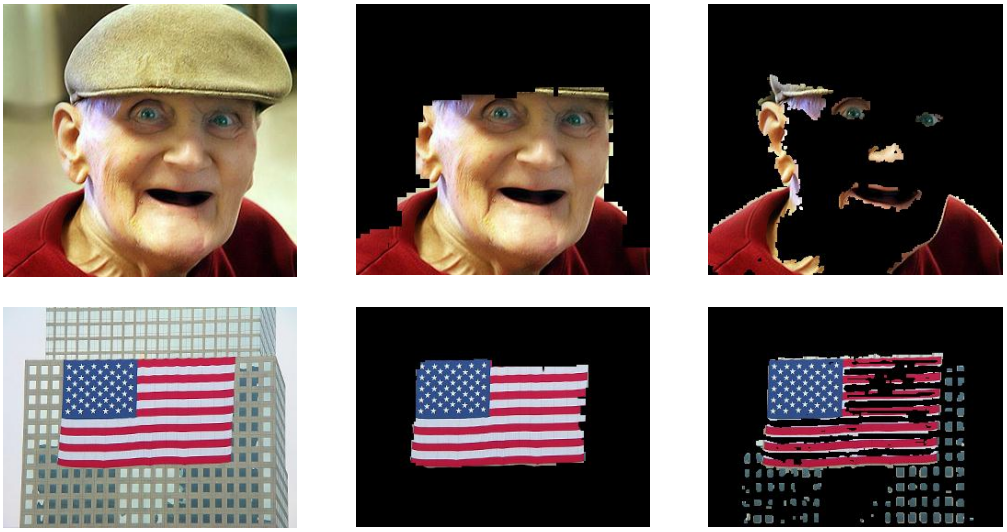

Figure 16: Particular cases of our algorithm: adjusting sensitivity to large or smaller objects using parameter ***p***. Left: original images. Middle: Extracted objects with focus on large objects. Right: extracted objects with focus on smaller details.

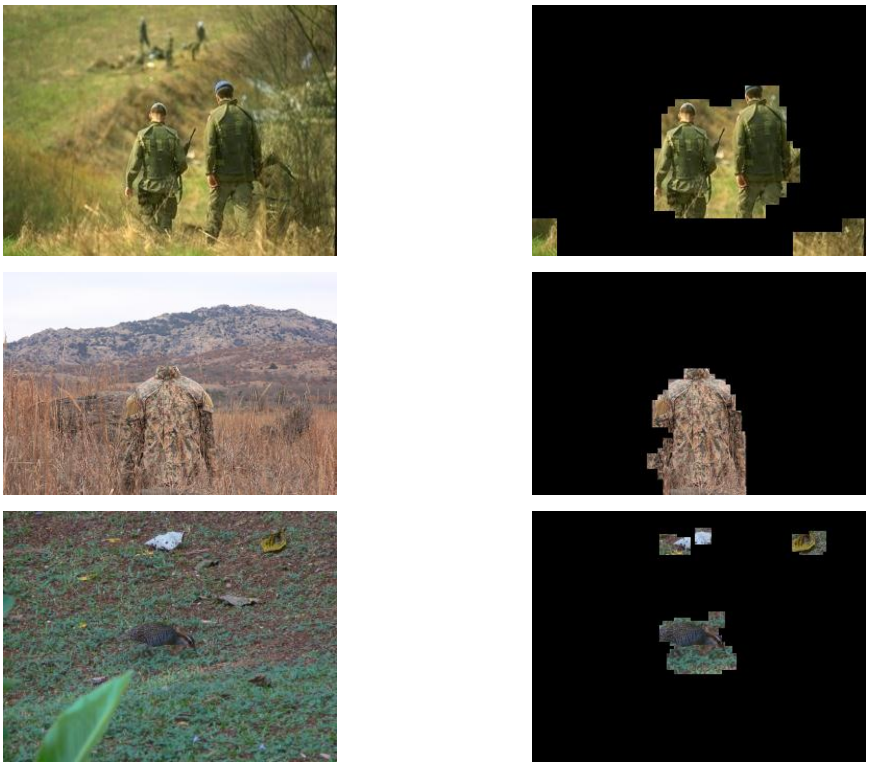

Figure 17: Particular cases of our algorithm: objects with camouflage pattern or colors similar to background close to the background such as military uniforms or animal camouflage. Left: original images. Right: extracted salient objects.

Finally on Fig. 17 we can observe extraction of objects with color and texture close to image background, such as military disruptive patterns found on combat uniforms in the first two images, or a natural animal camouflage in the case of veka bird captured on the last image.

3.5. Learning and Recognition of Salient Objects

The approach described in sections 3.3 and 3.4 allows us to split an image into a set of fragments, each containing a visually salient object. In this section it is explained how we can use this to enable a machine vision system to learn an object from unlabeled images. For experiments in real environment described further a mobile humanoid robot has been used, which was equipped with a color CMOS camera as a source of images.

```
acquire image
extract fragments by salient object detector
for each fragment F
   if(F is classified into one of groups)
      populate the group by F
   if(F is classified into multiple groups)
      populate by F the closest one by Euclidian dist. of features
   if(F is not classified to any group)
      create a new group and place F inside
select the most populated group G
   use fragments from G as learning
```

Algorithm 1: On-line salient object learning.

When acquired, images are processed to extract fragments containing salient objects; those fragments are grouped online using approach presented in sub-section 3.5.1. Only those groups with a significant number of members are used as samples database for object recognition methods. This has several effects. It enables recognition of previously seen objects in different visual context or environment and moreover it allows for learning of multiple objects in the same time.

Algorithm 1 describes the learning work-flow. At first time, the algorithm classifies each found fragment, and in the second step, the learning process is updated (on-line learning).

3.5.1. Incremental Fragment Grouping

To preserve the on-line and real time nature of learning, image fragments have to be grouped incrementally as they come from salient object detector with comparatively low calculation efforts. For this task a combination of weak classifiers $\{w_1, w_2, \dots, w_n\}$ is employed, each one classifying a fragment as belonging (result 1) or not belonging (result 0) to a certain class. Each classifier has a high level of false positives but a very low level of false negatives. In this case four weak classifiers ($n = 4$) are employed, covering several chief properties of object on the fragment.

A fragment belongs to a class if and only if $\prod_{i=1}^{n} w_i = 1$. A class is allowed to be populated only once by one fragment per image to prevent overpopulation by repeating patterns on the same image. If a fragment is not put into any class by classifiers, a new class is created for it. If a fragment satisfies this equation for multiple classes, it is assigned to the one whose Euclidian distance is smaller in terms of features measured by each classifier (i.e. c_{wn}). Features taken into account by weak classifiers are as follows. In all equations, F denotes the currently processed fragment, whereas G denotes an instance of the group in question. All other symbols are explained further on in the text.

Area: the w_1 in Eq. (13) classifier separates fragments, whose difference of areas is too large. In experiments, t_{area} is set to 10.

$$w_1 = \begin{cases} 1 & if\ c_{w1} < t_{area} \\ 0 & otherwise \end{cases}$$
$$where\ c_{w1} = \frac{\max(G_{area}, F_{area})}{\min(G_{area}, F_{area})} \quad \textbf{(13)}$$

Aspect: the w_2 in Eq. (14) classifier separates fragments, whose aspect ratios are too different to belong to the same object. In experiments, t_{aspect} is set to 0.3.

$$w_2 = \begin{cases} 1 & if\ c_{w2} < t_{aspect} \\ 0 & otherwise \end{cases}$$
$$where\ c_{w2} = \left| \log\left(\frac{G_{width}}{G_{height}}\right) - \log\left(\frac{G_{width}}{G_{height}}\right) \right| \quad \textbf{(14)}$$

Chromaticity distribution: the w_3 in Eq. (15) classifier separates fragments with clearly different chromaticity. It works over 2D normalized histograms of ϕ and θ component of fragment denoted by $G_{\phi\theta}$ and $F_{\phi\theta}$ respectively with N histogram bins,

calculating their intersection. N equal to 32 is used to avoid too sparse histogram and $t_{\phi\theta}$ equal to 0.35.

$$w_3 = \begin{cases} 1 & if\ c_{w3} < t_{\phi\theta} \\ 0 & otherwise \end{cases}$$
$$where\ c_{w3} = \frac{\sum_{j=1}^{N}\sum_{k=1}^{N} \min\left(G_{\phi\theta}(j,k) - F_{\phi\theta}(j,k)\right)}{L^2} \quad \textbf{(15)}$$

Texture uniformity: the w_4 in Eq. (16) classifier separates fragments, whose texture is too different. The measure of texture uniformity is used, calculated over the l channel of fragment. In Eq. (16), $\Omega(z_i); i = 1,2, \dots L$ is a normalized histogram of l channel of the given fragment and N is the number of histogram bins. In experiments, 32 histogram bins is used to avoid too sparse histogram and value $t_{uniformity}$ of 0.02.

$$w_4 = \begin{cases} 1 & if\ c_{w4} < t_{uniformity} \\ 0 & otherwise \end{cases}$$
$$where\ c_{w4} = \left| \sum_{j=1}^{N} \Omega_G^2(z_j) - \sum_{k=1}^{N} \Omega_F^2(z_k) \right| \quad \textbf{(16)}$$

3.5.2. Object Detection Algorithms

We are able to extract individual objects by means of their visual saliency (c.f. section 3.4). However, this ability alone cannot be used for their further re-detection in different conditions e.g. by a mere comparison of the extracted object with the ones already acquired. It is because there is no guarantee that next time we encounter the object it will be distinct to its surroundings (i.e. salient) and it won't be cluttered or partially occluded by other objects. To cope with this, we use existing object recognition approaches to detect in new conditions the objects we already acquired.

In any time of learning the fragment grouping algorithm provides us a set of groups, each one populated by fragments of images containing the same objects, seen from different viewpoints or different distances. We can choose any of those groups and use fragments contained in it as a database of samples for an object recognition algorithm.

For detection of objects in context of the present system, any suitable real time recognition algorithm can be used. To demonstrate how different kinds of recognition

algorithms can be employed within the system, I have employed two widely used object recognition algorithms. In following sub-sections I will explain basics of their function and how they make use of data about objects acquired in form of groups of fragments in order to learn the representation of each object and to enable its detection.

3.5.2.1. Speed-up Robust Features

The first object recognition technique, we use, Speed-up Robust Features, or SURF, described in (Bay et al., 2008) is a well-established technique based on matching interest points on the source image with interest points coming from the template. It describes a scale- and rotation-invariant interest point detector and descriptor. It allows detection robust to partial occlusions and perspective deformations.

The detector is based on the Hessian matrix. For a given pixel $x = (x, y)$ in image Ω, the Hessian matrix $\mathcal{H}(x, \sigma)$ in pixel x and scale σ is defined by (Bay et al., 2008) following Eq. (17). Symbol $L_{xx}(x, \sigma)$ stands for the convolution of the Gaussian second order derivate described as $\frac{\partial^2}{\partial x^2} g(\sigma)$ on image Ω. For $L_{xy}(x, \sigma)$ and $L_{yy}(x, \sigma)$ the expression under the line of fraction is ∂xy and ∂y^2 respectively.

$$\mathcal{H}(x, \sigma) = \begin{bmatrix} L_{xx}(x, \sigma) & L_{xy}(x, \sigma) \\ L_{xy}(x, \sigma) & L_{yy}(x, \sigma) \end{bmatrix} \quad \textbf{(17)}$$

In order to enable object recognition, the SURF method needs first a template. A template is an image showing the target object with very little background. Some background, however, is needed to be present on the image so that the edges of the object (the transition between the "object" and the "background") are clearly visible. From this template, key-points are extracted and stored. When a new image is presented, key-points are extracted from it and are matched against the set of template key-points. The matching follows the nearest-neighbor scheme, working with the Euclidian distance of feature vectors of each key-point.

To increase the processing speed, only key-points with similar contrast are matched. On the top of this, additional geometrical constraints are imposed on matching key-points in order to exclude false positive matches. When enough matching key-points between the template set and the image set are found, the area occupied by them is said to contain the searched object. Using the matching key-points the projection of the original template to the current image is calculated, which allows for drawing an appropriate bounding box around

the found object. An example of the described process is presented on Fig. 18[8]. On the right side of the image, the template (a food box) is shown. On the left, an entire scene containing the searched object is shown. Circles mark positions of key-points; matching key-points are linked by white lines. The found object on the scene is marked by a red rectangle, that has undergone an affine transformation to represent the most probable position of the searched box in space.

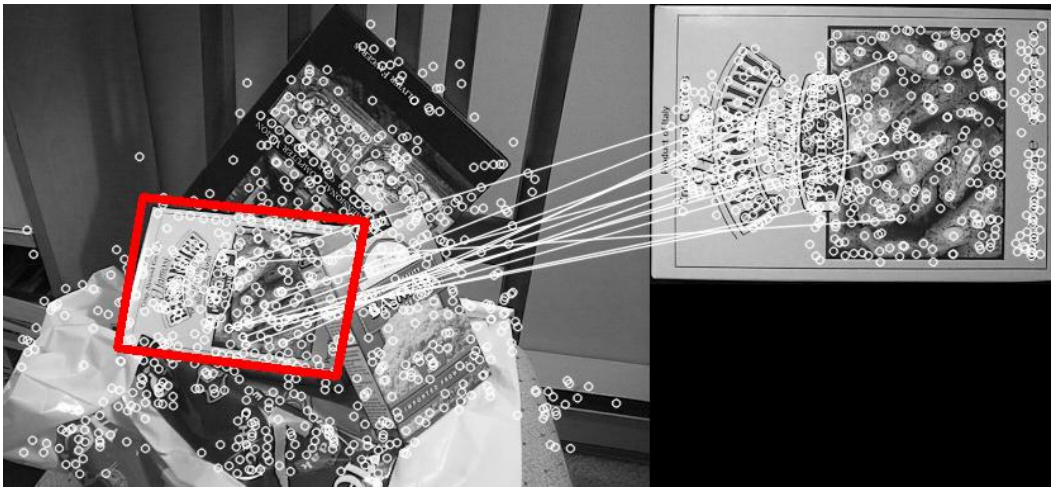

Figure 18: An example of SURF matching in object recognition. On the right there is the template. On the left, a scene containing the searched object.

In our case we use the fragments acquired as matching templates. To preserve the real-time operation of detection even with high numbers of templates, we pre-extract key points from each template in advance. In detection stage, we match first in several parallel threads templates with the greatest number of key-points (i.e. containing more visual information) and stop this process when another image from camera arrives. This allows testing up to few tens template matches per frame. Further important speed-up can be achieved using parallel computation power of CUDA-like architectures on modern GPUs.

3.5.2.2.Viola-Jones Detection Framework

The second object recognition method, used for applications in the present work, is the Viola-Jones detection framework. The framework has been published in the work of

[8] Adopted from the EmguCV code sample for SURF matching, online on: http://www.emgu.com/wiki/index.php/SURF_feature_detector_in_CSharp

(Viola and Jones, 2004) with a notable application to real-time human face detection. Its principle relies on browsing sub-windows over the target image and on a cascade of classifiers. This cascade determines, whether the processed part of image does, or does not belongs to a class of objects on which the classifier was trained.

The classifier relies on rectangular features, which are calculated as sums of pixels in adjacent rectangular areas of different kinds. In order to compute those features sufficiently rapidly, the notion of integral images is developed. An integral image is an image, on which each pixel contains the sum of all the pixels above and on the left of it. This is captured by Eq. (18). Using two simple recurrences from Eq. (19) an integral image can be computed from the original one in constant time. Using integral images, features at any given scale can be computed in constant time.

$$\Omega_{integral}(x,y) = \sum_{x' \le x; y' \le y} \Omega(x,y) \tag{18}$$

$$\Omega_{integral}(x,y) = \Omega_{integral}(x-1,y) + s(x,y)$$
$$where\ s(x,y) = s(x,y-1) + \Omega(x,y) \tag{19}$$

As the total number of possible features is overwhelming, the AdaBoost learning algorithm from (Freund and Schapire, 1995) is first used to select best features to train the classifier. In order to create a classifier which can evaluate the input in real time, a cascade of classifiers is created, in which each successive classifier is trained only on those selected samples which pass through the classifiers that precede it. The classifiers in the cascade have relatively low false negative rate and a very high false positive rate. By cascading them, most of non-perspective areas of the image are rejected in early stages and only promising regions are further processed.

In case of this work, we use acquired fragments of an object as positive samples to learn the cascade of classifiers (the learning here is carried out offline due to the nature of this method). As this method requires negative samples as well, we use the original images with the learned object replaced by a black rectangle. To be precise enough, the method needs up to several thousands of samples for learning. We achieved this number by applying random contrast changes, perspective deformations and rotation of learning fragments. Although Viola-Jones framework was originally designed to recognize a class of objects (i.e. human faces), rather than single instances, in our case we use it in the way that it recognizes a class of only one object (i.e. the one found on learning fragments). It must be noticed that having

the object detector learned, known objects can be detected directly from the input image when seen again, without passing by the salient object detection.

3.6. Focusing Visual Attention

Local features from sub-section 3.3.2.2 are scale-dependent. This gives us two options. First, a "universal" fixed value of the visual attention parameter p (the sliding window size) can be used, which was the case of results presented on Fig. 14. This approach as the advantage of simplicity (the p can be set once for all and its value may be estimated by a heuristic). However, the downside of this approach is that it does not allow reflecting the nature of any particular image. It is obvious that if the salient object extraction algorithm is run on images containing objects of different sizes and/or different texture properties, it will yield best (i.e. closes to the reality) extraction results on different values of p.

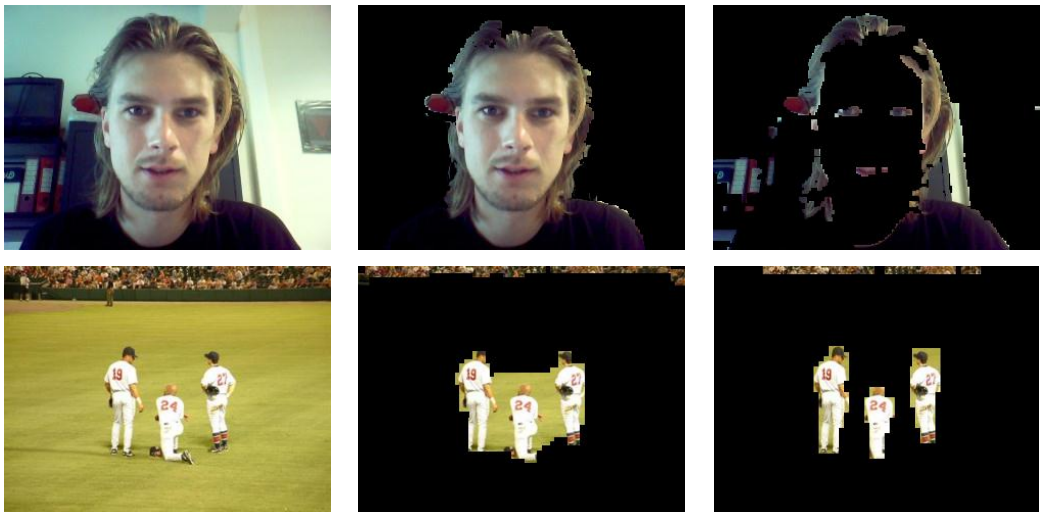

Figure 19: Impact of different values of visual attention $\boldsymbol{p}$ setting. Left: original images. Middle: salient objects extracted with high $\boldsymbol{p}$ value. Right: salient objects extracted with low $\boldsymbol{p}$ value.

As it has been explained previously, the visual attention parameter p permits for a top-down control of the attention and of the sensitivity of the feature in scale space. High p value (resulting in a large sliding window size) with respect to the image size will make the local saliency feature more sensitive to large coherent parts of the image. On the other hand, low values of p will allow focusing to smaller details. The situation is described on Fig. 19, where large p allows for extraction of the entire face, while low p focuses on smaller features

like eyes and lips and patches of hair. A similar situation may be observed on the image of sportsmen, where the entire group of three persons may be perceived as a salient object, or each person can be considered individually, depending on the level of granularity desired. However, as shown in the labeling consistency analysis in (Liu et al., 2011), for most images there is very little doubt about what object is (or what objects are) salient and what are not. This permits for each image to determinate the appropriate granularity level and consequently the most appropriate visual attention parameter p.

3.6.1. Examining Possible Correlation between the p and the Salient Object Size

The common sense would suggest that high values of p would give better results of salient object extraction for images with large salient objects and small p values would be more suitable for small salient objects. While this observation may be plausible in particular cases (see Fig. 16), I have decided to examine its generality. The number of 525 images has been selected from the MSRA dataset. Those images contain only one salient object per image, for the sake of transparency of results. The salient object sizes are ranging from approximately 10% of the image area up to approximately 75%.

For each image, an exhaustive search for the best p value (between 0.2 and 0.5) has been performed. The best p value has been determined as the one, which would yield the best F-measure of the extracted region. Results are plotted on Fig. 20. Each point of the plot represents an image taken from the MSRA selection. On the x-axis the proportional size of the salient object contained on the image is shown. On the y-axis, the corresponding best p value for the given image is captures.

There are two obvious conclusions we may make based on the results shown on Fig. 20. The first one is, that fixed values of p between approximately 0.35 and 0.45 will provide "good enough" results of extraction for most of the images, regardless to the size of the salient object. This justifies the heuristic choice of the p parameter that we have made in sub-section 3.3.2.2.

The second conclusion is, that there is no important correlation between the proportional size of salient objects and the value of p. The R^2 value (coefficient of determination) for the obtained data was approximately 0.015, which means no significant correlation. This fact is not particularly surprising if we consider, that the hybrid center-

surround feature as defined in sub-section 3.3.2.2 is influenced not only by the size of the object, but also by the texture of the surface of the object and of the background.

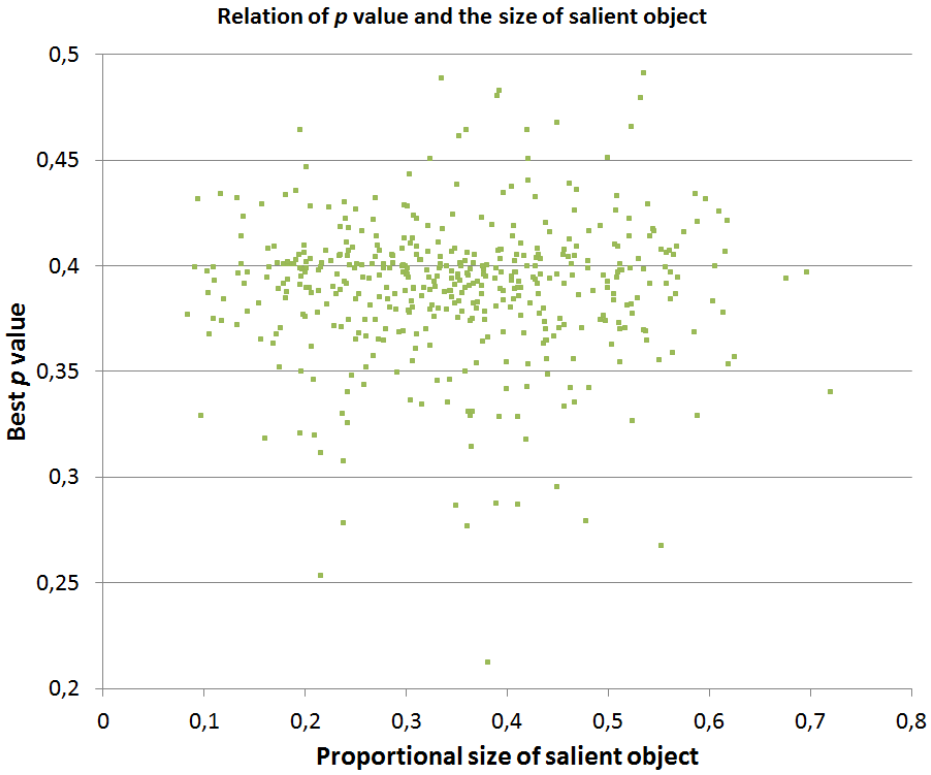

Figure 20: Graph of relation between the best performing value of visual attention parameter $\boldsymbol{p}$ and the proportional size of the salient object on image.

3.6.2. Visual Attention Parameter Estimation

Given observations I have made in the previous sub-section, it is pertinent to say, that an approach enabling correct estimation of the visual attention parameter on a per-image basis could improve salient object extraction results. This estimation must be, however, fully automatic, if we want to preserve the autonomous nature of the entire process of salient object extraction.

As we have seen, the value of p is dependent on the size of coherent regions (i.e. segments). I propose an estimation method based on calculation of a histogram of segment sizes from the input image. This gives us a feature vector, which is provided as an input of an artificial neural network trained to output the sliding window value. The weights of the neural network are adapted in training stage using a genetic algorithm. The process of automatic

estimation of the visual attention parameter, that I propose, is depicted on Fig. 21 (by orange boxes) in context of the entire salient object extraction (cf. Fig. 8).

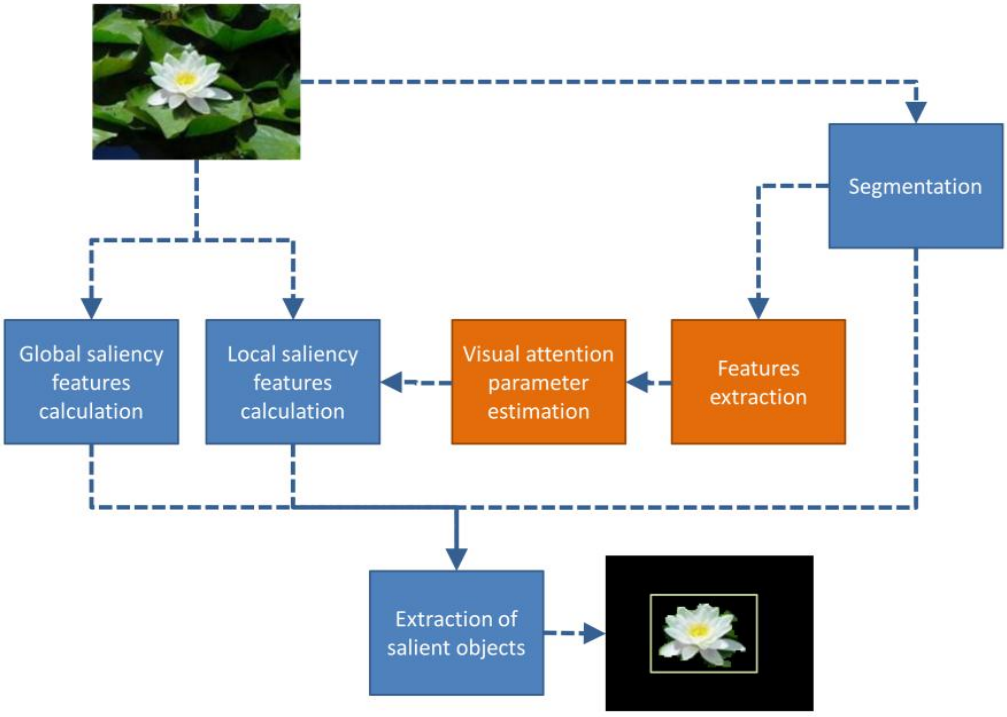

Figure 21: Graphical depiction of the work-flow of our salient object extraction system. The two orange boxes represent visual attention estimation process, which are discussed in this section.

3.6.3. Features Extraction

In sub-section 3.4.1, an algorithm for image segmentation is presented, having some interesting properties like robustness to difficult real-world illumination conditions such as shadows and shines and its relatively high speed. Results of this algorithm are exploited here for both feature extraction for visual attention parameter estimation and for the eventual extraction of salient objects (in sub-section 3.4.2).

To obtain the feature vector, the input image is segmented into segments$(S_1, S_2, \dots, S_n)$. For each one of the found segments $S_i \in \{S_1, S_2, \dots, S_n\}$ its size in pixels $|S_i|$ is divided by the overall image size $|\Omega|$. An (absolute) histogram H_{SA} of segment sizes is then constructed. To avoid a too sparse histogram, the i-th bin of histogram H_{SA} is populated by a sum of similarly large segments as described by Eq. (20). This ensures that the

first histogram bin contains the number of segments with area larger that 1/10 of the image size, the second contains segments from 1/10 to 1/100 of the image size etc. For practical reasons a 4-bin histogram is used, as fragments counted in the 5th and succeeding bins would be insignificantly small.

$$H_{SA}(i) = \sum_{j=i}^{n} \begin{cases} 1 & if\ 10^{i-1} \leq \left(\frac{|S_j|}{|\Omega|}\right) < 10^{i} \\ 0 & otherwise \end{cases} \quad \textbf{(20)}$$

To obtain relative values instead of absolute counts, we calculate a (relative) histogram H_{SR}, where each bin is assigned a number following Eq. (21).

$$H_{SR}(i) = \frac{H_{SA}(i)}{\sum_{j=1}^{n} H_{SA}(j)} \quad \textbf{(21)}$$

3.6.4. Construction and Learning of Visual Attention Parameter Estimator

The core of the proposed visual attention parameter estimator is an artificial neural network, a multilayer perceptron (see (Minsky and Papert, 1988)) with a sigmoidal activation function. A fully connected three-layer feed-forward network (MLP) is used. Its structure is: four input nodes, three hidden neurons and one output neuron. The number of hidden neurons has been determined comparing trials with different neuron numbers. The four input nodes are connected each to its respective bin from the H_{SR} histogram. The value of the output node ranging from 0 to 1 is interpreted as the ratio of the estimated sliding window size p and the long side of the image. Before being used, the MLP needs to have its weights adjusted. This is done in the learning loop, which is described by Algorithm 2. The learning makes use of a genetic algorithm (cf. (Holland, 1992)). As learning data-set, 10% of the MSRA-B data-set images are used. Measures proposed in the same work are then used to evaluate quantitatively the salient object extraction. The remaining 90% of the data-set images are left out for the purpose of validation.

Each organism in the population consists of a genome - an array of floating point numbers whose length corresponds with the number of weights in MLP. To calculate the fitness of each organism, the MLP weights are set according to its genome.

```
set popSize /population size
set genSize //length of genome
set population //a set of organisms
set mlp //multilayer perceptronacquire image

//random initialization of population
for(i = 1 to popSize){
   set newOrganism
   for(j = 1 to genSize)
      newOrganism[j]=rnd //initialization by a small random number
   population.Add(newOrganism)
}

//evolution starts
do{

   set fitnesses //array of fitnesses of members of the population
   for(i = 1 to popSize){
      set organism = population[i]
      initialize mlp weights with organism
      calculate H_SR
      input H_SR into mlp and do feed-forward
      set p according to mlp output
      foreach(image in learning set){
         extract salient objects with p parameter
         compare results with ground truth
         calculate F-ratio
      }
      fitnesses[i] = average F-ratio
   }
   set newPopulation //new generation
   //elitist selection
   newPopulation.Add(organism with max(fitnesses))
   for(i = 1 to popSize - 1)
      newPopulation.Add(crossover rnd organisms with high fitness)
   mutate randomly newPopulation
   population = newPopulation
   //stopping condition with an arbitrary threshold

}until(max(fitnesses) > threshold)
```

Algorithm 2: Adjusting neural network weights by genetic algorithm.

The input image from training set is segmented, H_{SR} calculated and used as input to the MLP. Once visual attention parameter p is calculated according to the MLP output,

saliency is computed over the image and salient objects are extracted. The result is compared with ground truth and the precision, recall and the F-measure are calculated (according to (Liu et al., 2011)). The F-measure, representing the overall quality of the extraction, is then used as the measure of fitness for the given organism. In each generation, the elitism rule is used in order to explicitly preserve the best solution found so far. Organisms are mutated with 5% of probability.

Once the MLP is learned, its weights are saved and the estimator is ready to be used. Input image processing follows then the work-flow depicted on Fig. 21.

3.7. Experiments

To verify the performance of our system, a number of experiments have been performed with learning objects present in a common office environment. For the sake of repeatability and convenience in evaluation of results, ten common house or office objects have been collected in order to be explicitly learned (although the system naturally learns any salient objects in its surroundings without any specific preference). A sample of scene images containing those objects is presented of Fig. 26. Illustrative photos of the objects that have been used in this experiment are shown on Fig. 25 in order to give the reader a better idea about their nature.

In order to approach to the real conditions as much as possible objects with different surface properties (chromatic, achromatic, textured, smooth, reflective …) have been chosen and they were put in a wide variety of light conditions and visual contexts. The number of images acquired for scenes containing each object varied between 100 and 600 for learning image sequences and between 50 and 300 for testing sequences, always with multiple objects occurring on the same scene. Note that the high number of learning images was taken primarily in order to test sufficiently our saliency detection and segment grouping algorithm. The learning process itself would generally require significantly less samples acquired in order to perform sufficiently well, depending on the actual object detection algorithm employed.

On Fig. 26 random images from the learning sequence are presented for each learned object along with fragments extracted from them. These fragments, containing salient objects found on each image are subsequently processed by the incremental fragment grouping (sub-section 3.5.1) and the selected ones are used for learning of the object detector.

3.7.1. Validation of Visual Attention Parameter Estimation

To evaluate the presented approach of automatic visual attention parameterization, its results have been compared to those achieved by the same approach using a fixed p value and to results of other state of the art algorithms. For quantitative comparison, the previously mentioned MSRA-B data-set was used. It contains 5000 images with hand-labeled salient objects as well as measures to quantify the correctness of salient object extraction. For evaluation, 90% of this set was used, excluding the 10% used for training.

Regarding the learning process, it usually converged in about 30 generations. A sample fitness evolution curve is presented on Fig. 22. The fitness curve is monotonically increasing because the elitism rule is used in the genetic algorithm. This ensures that in any new generation the previous best solution is automatically included and thus the best fitness of the population never decreases.

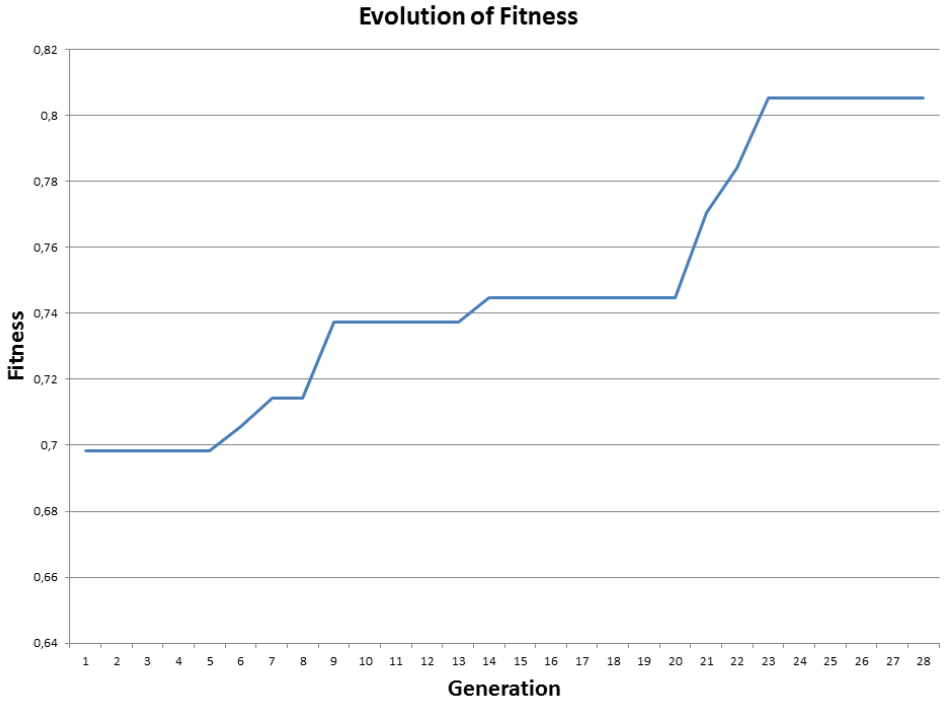

Figure 22: An instance of fitness convergence during the evolution.

When being run over the testing data-set, the p algorithm was resulting in average F-measure of 0,75 (compare to Table 1 in sub-section 3.4.3). After the visual attention estimator has been fully learned, the performance of salient object extraction over the testing data-set in terms of F-measure was 0,84 in average. This gives us about 11% increase in performance. It

may be also observed that the approach with automatically estimated p tends to be more precise in extraction (the measure of precision increased from 0,75 to 0,87). The results of the fixed p approach are compared to those that were obtained using automatic estimation of p in Table 2.

	Precision	**Recall**	**F-measure**
Best fixed p value	0,75	0,76	0,75
Automatic estimation	0,87	0,83	0,84

Table 2: Comparison of scores obtained in MSRA-B dataset using a fixed $\boldsymbol{p}$ approach and using the automatic estimation of $\boldsymbol{p}$.

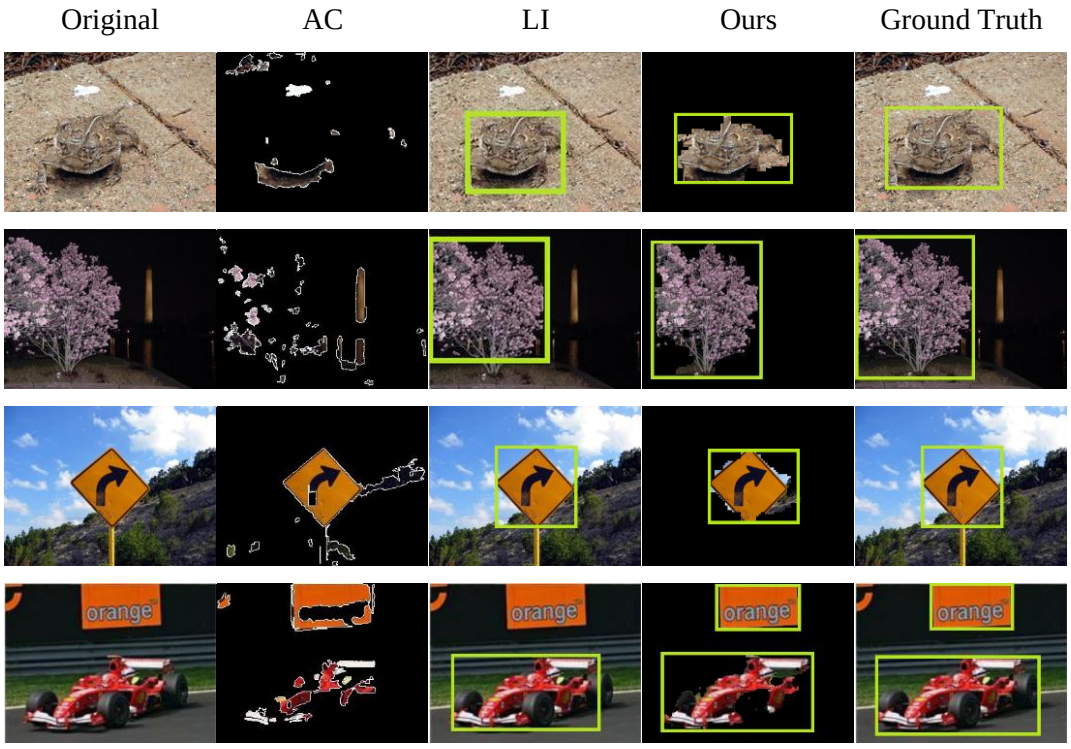

Figure 23: Comparison of different salient object detection algorithms. First column: original image. Second column: results of (AC) **(Achanta et al., 2009)**. Third column: results of (LI) **(Liu et al., 2011)**. Fourth column: results of the present approach with automatic estimation of $\boldsymbol{p}$. Last column contains ground truth, considering multiple objects in the scene.

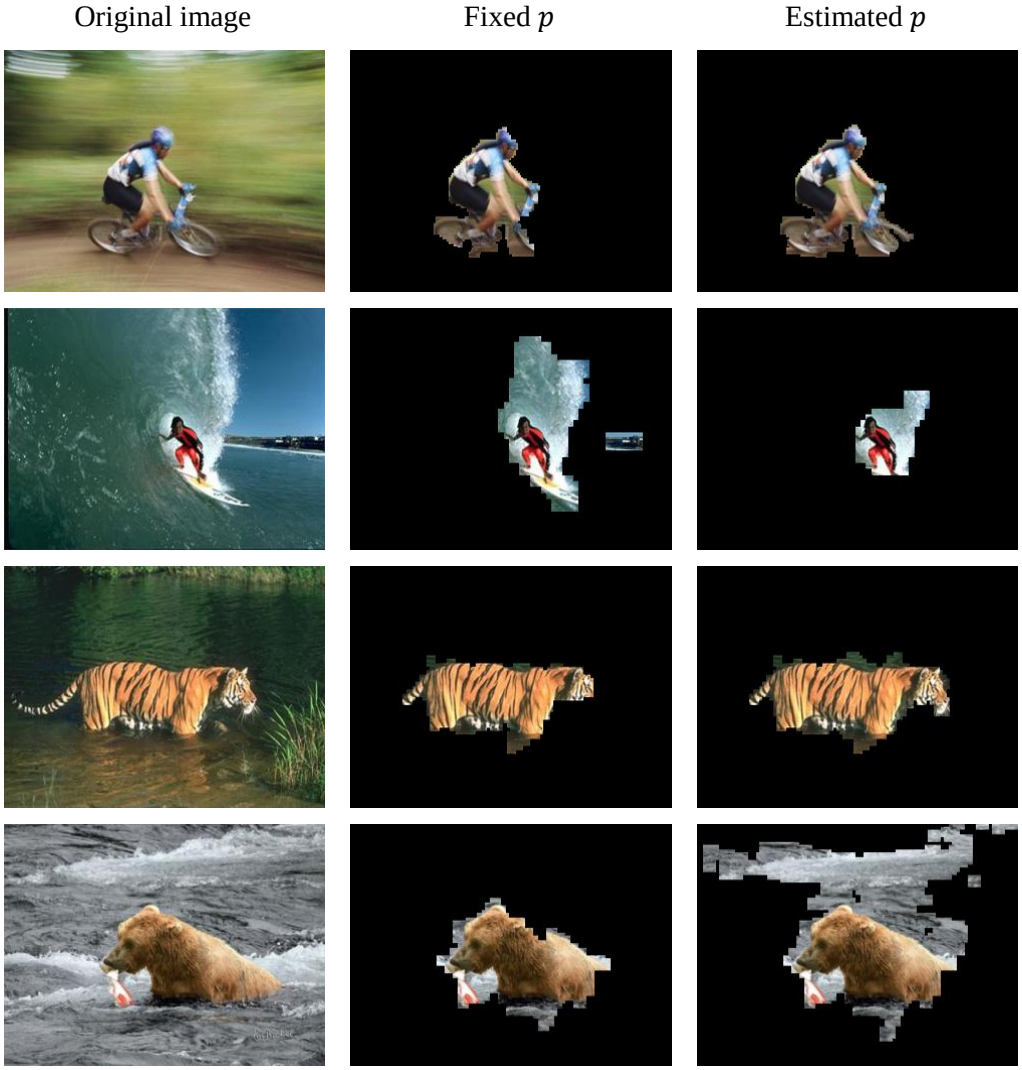

Figure 24: Random images from the MSRA data-set processed by the fixed $\boldsymbol{p}$ salient object extraction and compared to results achieved with automatic $\boldsymbol{p}$ estimator. The last row illustrates a case when the estimator failed to find the appropriate $\boldsymbol{p}$.

On Fig. 23, sample results of our algorithm with automatic estimation of the visual attention scale are compared with the ground truth and two others state of the art algorithms. As it was the case of Fig. 14, we have chosen to compare with the work presented in (Achanta et al., 2009) (AC) as it presents a computationally very fast approach and (Liu et al., 2011) (LI) as a source of the used benchmark. After programming code optimization, no major increase in computing time of our algorithm has been observed, with respect to the fixed p version. This means that the algorithm is still capable of real time operation.

Although (Achanta et al., 2009) is computationally very cheap, its results vary largely in quality depending on the nature of salient objects on the image. Algorithm of (Liu et al.,

2011) produces results much close to human perception and more precise in terms of resolution. Our approach outperforms considerably the one of (Liu et al., 2011), while maintaining both important benefits, that have been discussed in sub-section 3.4.2, that is its high processing speed and the fact, that it outputs natively multiple salient objects if they are present on the image.

On Fig. 24 some random images from the MSRA data-set are presented, showing processing results of the fixed p approach and those of the automatically estimated p approach. The difference is particularly visible in the second row, where the size of the salient object (the surfer) is proportionally very small with respect to the image size.

3.7.2. Validation of Salient Object Extraction and Learning

First, results of salient object extraction and fragment grouping are presented. To investigate the effectiveness of the salient object extraction, the percentage of learning set images, on which the learned object have been correctly detected and extracted by salient object detector, has been counted. Correct extraction means here that the object has been extracted entire and without any other objects co-occurring on the fragment.

Figure 25: Images of objects used throughout the described experiments. Note that the white background is here merely for the sake of clarity. During experiments those objects were presented in various situations, visual contexts and backgrounds.

Illustrative photos of the objects we used in this validation are shown on Fig. 25 in order to give the reader a better idea about their nature. Objects with different surface

properties (chromatic, achromatic, textured, smooth, reflective …) have been chosen and put in a wide variety of light conditions and visual contexts.

On Fig. 26 some images from the training set are shown along with salient objects extracted from them. Usually multiple salient objects were extracted from each scene. Successful extraction has been achieved approximately on 82% of images in the set. The subsequent grouping of fragments has achieved on the same data-set success rate of 96%, i.e. only 4% of fragments, which were usually bearing visual resemblance to other objects on the scene, were placed into a wrong group.

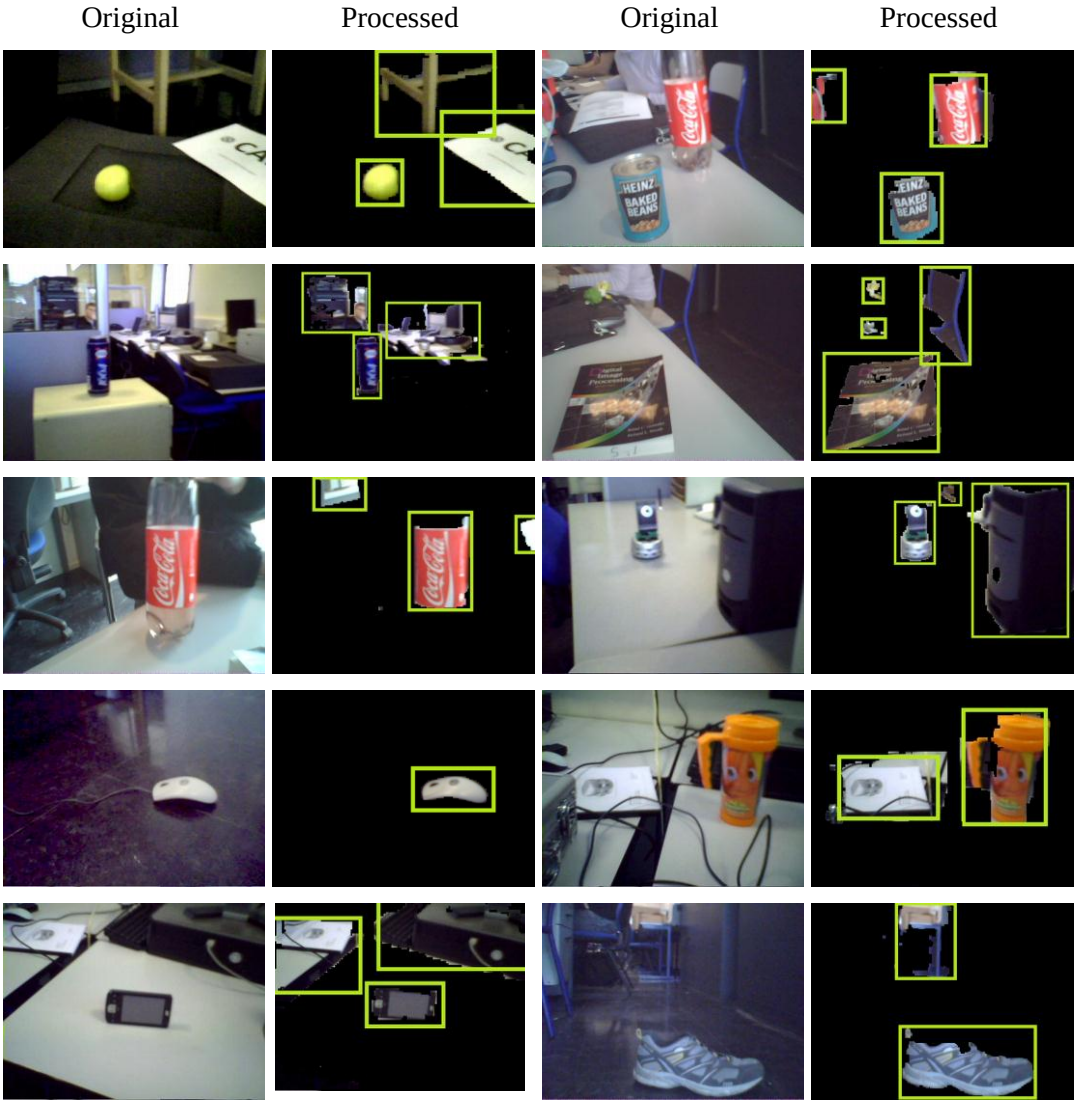

Figure 26: Sample images from the training sequence for each of the used objects. Fragments containing salient objects detected by our algorithm are marked by green rectangles.

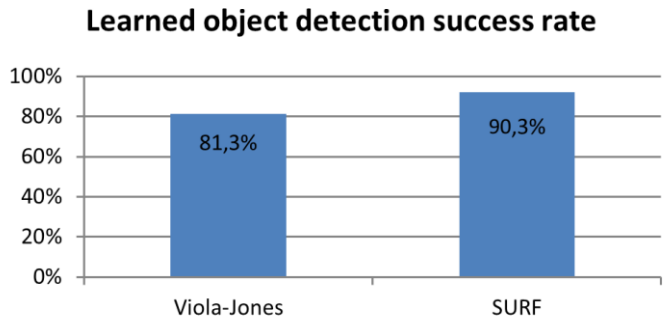

Figure 27: Percentage of correct detections of learned objects over testing image set using Viola-Jones algorithm and SURF algorithm.

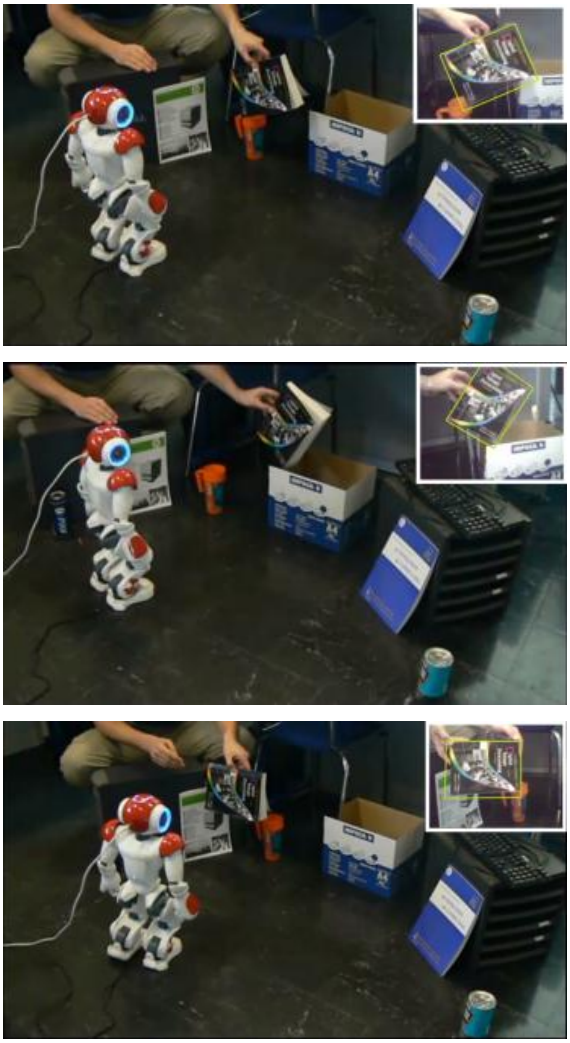

Figure 28: Images from tracking a previously learned moving object. Robot camera picture is shown in upper right corner of each image.

On Fig. 27 detection rates over testing data-set using a trained Viola-Jones detection framework are provided along with performance of SURF algorithm on the same data-set. In average over all the objects in testing set, the detection rate for Viola-Jones was about 81.3%. In case of SURF, the average detection rate was, higher about 90.3%. The numbers reflect only true positive detections. Average rate of false positive detections was around 0.5% for both methods.

Figure 29: Camera pictures from single (first column) object detection and multiple (second column) previously learned object detection. Successfully detected objects are marked by green lines.

To demonstrate real-time abilities of our system, several experiments were successfully run, where a mobile robot equipped with color camera was required to learn a presented object. When learned, the robot was required to find the object in its environment

and to track and follow it. Images from a video[9] acquired during those experiments are shown on Fig. 28. On Fig. 29, sample detection results are shown with a system having learned several objects. Boundary lines determine objects previously encountered by the robot and successfully recognized on the new scene.

3.7.3. Discussion

In sub-section 3.4.3, some quantitative results of our salient object extraction technique were given and compared with existing approaches. We have seen that the quality of our approach is comparable with the existing ones (considering the fixed p version) or superior to them (when using the automatic estimation of p). At the same time it brings the advantage of real-time processing, native extraction of multiple salient objects and robustness to certain difficult illumination conditions. This confirms that our salient object extraction technique is performing well enough to play its part in the proposed system for autonomous acquisition of knowledge.

On Fig. 26, some qualitative results of our salient object extraction in real environment are given. The figure shows several typical views acquired during learning of the system. Salient objects extracted from them are marked by green rectangles. On all images (except of the "mouse" one, where only one object is present) multiple visually important objects were extracted apart of the one we placed intentionally to the scene. This indeed is the desired behavior as the system is expected to extract (and learn) autonomously the encountered objects without any a-priory preference. On the other hand, as illustrated on the "mouse" and "shoe" images, the algorithm does not extract the "false objects" created by reflections found on the floor.

The percentage of successfully extracted samples of a same object usable to learn this object is 82% of its occurrences throughout a sequence of images. This means that more than 4 of 5 acquired images of that object contribute in fact to correct learning of this object. As our learning system is incremental, the needed number of sample images for each object can be achieved accurately and fast enough.

Two fundamentally different object recognition algorithms have been employed in this system, each yielding different rate of recognition. The SURF detector has shown superior performance 90.3% of average detection rate by contrast to Viola-Jones detection

[9] This video can be found to the following address:
http://www.youtube.com/watch?v=xxz3wm3L1pE

framework, which performed by about 9% worse. This shows that Viola-Jones framework may not be best suited for this kind of task. It is presumes that it is so mainly because of the fact that in order to achieve high recognition rates it needs typically thousands of learning samples, however the number of unique samples acquired for each learned object was in order of hundreds. Also its long learning time makes it impractical in the strict sense of on-line learning.

On the other hand results achieved with SURF are encouraging both for the relatively high percentage of correct recognitions and for the fact that it allows recognition of the learned object even with only several samples acquired. Some camera views of already learned system are shown on Fig. 29 with objects recognized marked by bounding shapes. On the first row individual objects are correctly detected. On the second row, three views on similar scenes containing multiple objects are shown. Between the scenes the system was progressively learning new objects so that e.g. the orange mug is recognized only on the last scene as prior to this it was not learned. These images show the flexibility of recognition of the learned objects that are recognized in different orientation and perspective (the book), different illumination conditions (the shoe) or different distance and orientation (the coke bottle).

Regarding experiments with a robot searching or tracking a previously learned object (see Fig. 28), the present system was successfully validated. It has enabled the robot to fulfill the required cognitive tasks, correctly responding to the input. Because of limited computing capacity of the robot used, it has been chosen to run the system on a remote computer. In this experimental context, despite of the specific communication protocol implemented by the constructor on the robot, the system itself has been capable of real-time processing performance. However, one may observe observed a slow-down in robots reactions due to the limited bandwidth. This is the consequence of inadequacy of the aforementioned protocol regarding image transfer.

Certain shortcomings have been also identified in learning chain naturally bound to the method of object extraction that was used. In fact, our system shows worsening performance in learning objects that are not enough visually distinct with respect to their background. The same happens in cases where two visually important objects are seen one behind another and thus are wrongly extracted as one by our current system. By consequence, in order to respond correctly to this complex situation, it would be necessary to extend the present salient object extraction system by an additional level of machine intelligence. However, it is pertinent to emphasize, that in the actual system, once an object is correctly

learned, its further detection (thanks to the object detectors employed) is practically independent from its visual context.

3.8. Conclusion

In this chapter a low level cognitive system is proposed, benefiting from the perceptual (visual) saliency as an implementation of the concept of perceptual curiosity. It has the capacity of autonomous learning of objects present in real environment. It has been inspired by early processing stages of human visual system and by existing work studying the way human infants learn. In this context a novel algorithm for visually salient object detection is suggested, taking advantage of using photometric invariants. The algorithm has low complexity and can be run in real-time on contemporary processors. Moreover it exhibits robustness to difficult real-world light conditions.

Further a machine learning approach is developed using an artificial neural network and genetic algorithm, which allows us to estimate automatically the visual attention parameter for each image based on its features. Observations have been made supporting the fact, that the learning process converges and once fully learned, the MLP allows for a consistent estimation of the p parameter. This has been verified by a quantitative evaluation on the MSRA benchmark. The results show an increase in overall quality and precision of salient object extraction, when compared to our previous approach with a fixed visual attention parameter.

The presented algorithm is the first key part of the proposed lower level knowledge acquisition unit. It is demonstrated that the detected salient objects can be efficiently used for training the second key part of this unit, which ensures a machine learning-based object detection and recognition. Encouraging results were obtained especially when SURF detector was employed as an object detector.

In future this approach could evolve in several ways. As there does not exist a universal recognition algorithm that suits any existing class of objects, other object detection algorithms, like GLOH in (Mikolajczyk and Schmid, 2005) or receptive field co-occurrence histograms in (Ekvall and Kragic, 2005), could be adopted along with surface descriptors for each learned object. Objects of different characteristics could be then learned by algorithms that best suit the nature of the object in a "mixture of experts" manner.

As to the visual saliency detector, the center-surround feature detector could be supplied by or replaced by an interesting approach of spectral residua detection published in

(Hou and Zhang, 2007). A top-down feedback based on already acquired and grouped fragments could also greatly improve the saliency detector. The results presented here have been achieved with a monocular camera. However, there are valid reasons to believe that the performance of the entire system could be enhanced by use of a stereo camera. In this case the depth-separation of objects would serve side-by-side with the segmentation algorithm to cope with the mentioned cases, where two visually important objects are one behind another.

An open question is, whether the presented technique, instead of learning solely individual objects, could be used as well for place learning and recognition, extracting visually important objects from the entire place like room or office or for visual navigation of a mobile robot. It would also be interesting to investigate, how the saliency-based method could have an overlap outside the image processing domain, to be applied for learning of other than visual data (e.g. audio).

Chapter 4. Learning by Interaction

4.1. Introduction

In this section, I detail on my approach to autonomous knowledge acquisition by interaction. This represents the high level, epistemic curiosity driven knowledge acquisition process discussed in Chapter 2. At first I outline its general principles concerning learning of a single type of features (originating from one sensor only) at one time. Then I explicate how beliefs about the world are generated by the robot based on its autonomous observations of the world and on its interaction with a human tutor. Then I detail on how the robot uses those beliefs to interpret the environment in which it evolves. Next, the role of interaction between the human tutor and the robot during different stages of the learning process is described. Finally the outlined learning principles are generalized in order to allow learning of multiple types of features from multiple sensors at the same time.

The problem of learning brings an inherent problem of distinguishing the pertinent sensory information (the one to which the tutor is referring) and the impertinent one. It indeed is a paradox, but in contrary to what one may believe, sensors provide generally too much data input, a lot more than effectively needed. It is the task of higher structures (e.g. an attention system or in general a machine learning system adapted to this task) to draw the attention to particular features of the data, which are pertinent in context of a particular task. This problem has been addressed by researchers on different fields (for a reference, see e.g. (Blum and Langley, 1997) or (Soderland, 1999)). The solution to this task is not obvious even if we achieve joint attention in the robot. This is illustrated on Fig. 30. Consider a robot learning a single type of features, e.g. colors. If a tutor points to one object (e.g. a yellow fish) among many others, and describes it by saying "The fish is yellow!" the robot still has to distinguish, which of the several colors and shades found on the object the tutor is referring to. This step is an inevitable one before we can proceed to the learning itself. In traditional learning systems, this task-relevant (i.e. pertinent) information is extracted by hand by a human expert. In a system capable of autonomous learning, however, this has to be done in an autonomous way and without recourse to human-extracted features.

4.2. General Overview of the System

As it has been depicted on Fig. 30, sensor data bring inherently both pertinent and impertinent information mixed up. To achieve correct detection of pertinent information in spite of such an uncertainty, we adopt the following strategy. The robot extracts features from important objects found in the scene along with words the tutor used to describe the presented objects. Then, the robot generates its beliefs about which word could describe which feature (see subsection 4.3.2). The beliefs are used as organisms in a genetic algorithm. Here, the appropriate fitness function is of major importance. To calculate the fitness, a classifier is trained based on each belief about the world. Using it, the cognitive system tries to interpret the objects the robot has already seen. The utterances pronounced by the human tutor in presence of each such object are compared with the utterances the robot would use to describe it based on the current belief. The closer the robot's description is to the one given by the human, the higher the fitness is.

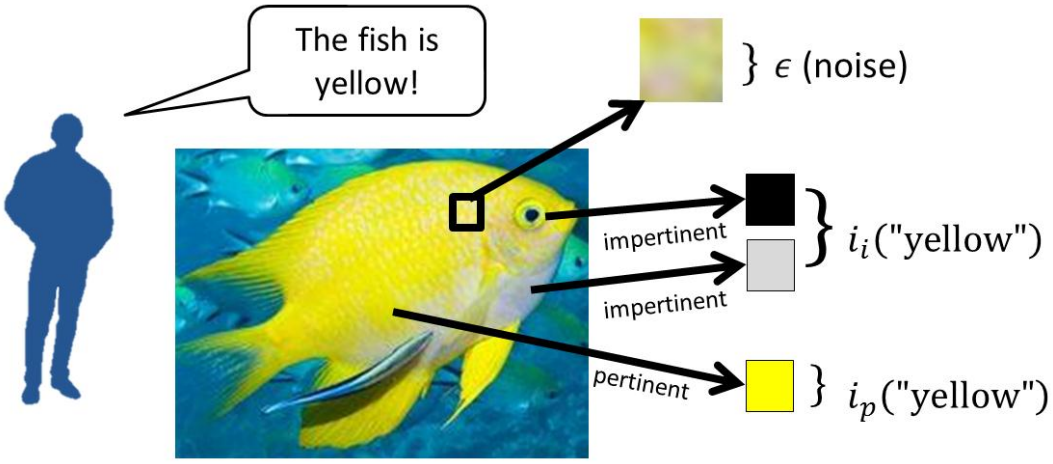

Figure 30: A human would describe this fish as being yellow in spite of the fact, that this is not by far its only color. Symbols $\boldsymbol{i_i}$, $\boldsymbol{i_p}$ and $\boldsymbol{\epsilon}$ refer to Eq. **(22)**.

Once the evolution has been finished, the belief with the highest fitness is adopted by the robot and is used to interpret occurrences of new (unseen) objects. On Fig. 31, important parts of the system proposed in this chapter are depicted. The "Genetic algorithm" box is further expanded and explained in sub-section 4.3.3. Further in the chapter the Fig. 39 shows the humanoid robot along with different objects that have been used to evaluate the present learning system in real-world conditions.

Although one could argue the system presented here is not specifically bound to humanoid robots, it is pertinent to state two main reasons why a humanoid robot is used for the system's validation. The first reason for this is that a humanoid robot by definition possesses a number of sensors (such as camera and microphones) that make its perception close to the human perception, entailing a more human-like experience of the world. This is an important aspect to consider in context of sharing knowledge between a human and a robot endowed with a cognitive system. Some aspects of this problem are discussed e.g. in (Klingspor, Demiris and Kaiser, 1997)).

The second reason is that humanoid robots are specifically designed to interact with humans in a "natural" way by using e.g. a loudspeaker and microphone set in order to allow for a bi-directional communication with human by speech synthesis and speech analysis and recognition. This is of importance when speaking about a natural human-robot interaction during learning.

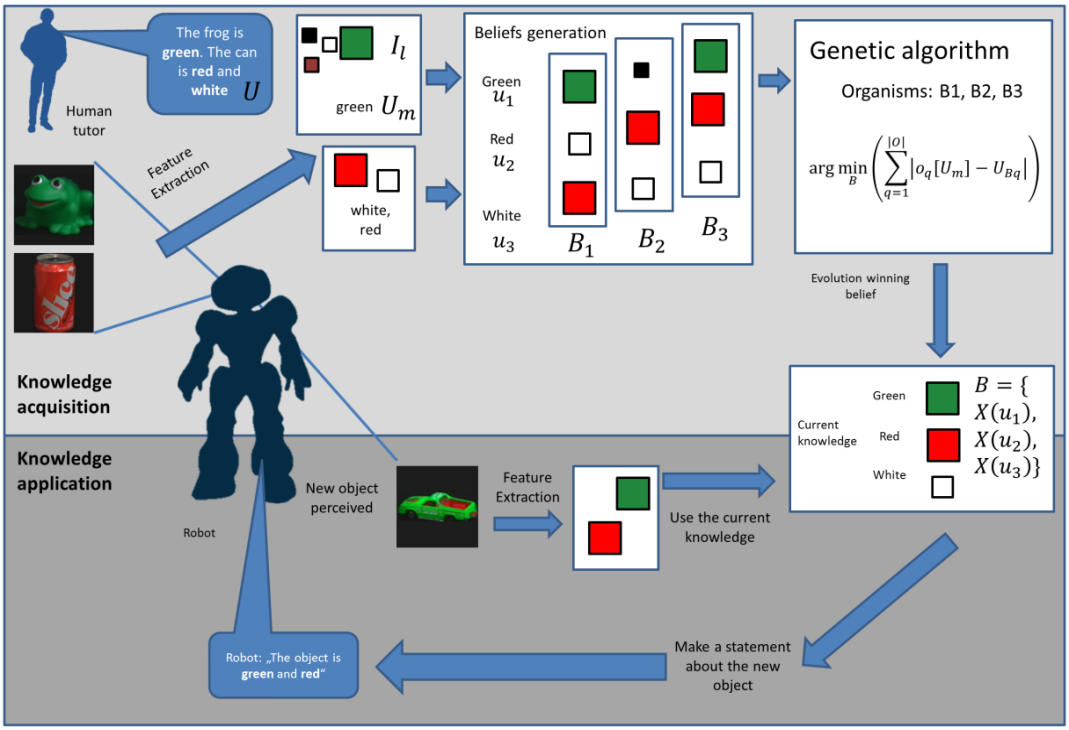

Figure 31: Graphical depiction of the proposed system for learning a single type of features. For the sake of comprehensibility it is shown in context of a particular learning task, i.e. color learning, instead of a purely symbolic description. Symbols used refer to those defined in sub-section 4.3.1 and following.

4.3. Learning by Interaction from One Sensor

4.3.1. Observation and Interpretation

Let us have a robot endowed with by a sensor, which makes it able to observe the world around it. The world is represented as a set of features $I = \{i_1, i_2, \dots i_k\}$, which can be acquired by this sensor. Each time the robot makes an observation o, a human tutor gives it a set of utterances U_m describing important objects found currently in the world. Let us denote the set of all utterances ever given about the world as U. The goal for the robot is to distinguish the pertinent information present in the observation from the impertinent one and to correctly map the utterances to appropriate perceived stimuli (features). In other words, the robot is required to establish a word-meaning relationship between the uttered words and its own perception. The robot is further allowed to interact with the human in order to clarify and verify its interpretations, following the stimulation of curiosity.

For this purpose, let us define an observation o as an ordered pair $o = \{I_l, U_m\}$, where $I_l \subseteq I$ stands for the set of features obtained by observing the world and $U_m \subseteq U$ is a set of utterances given in the context of the observation. Following Eq. (22), I_l is a union of all the pertinent information i_p for a given u (i.e. features that can be described as u in the language used for communication between the human and the robot), all the impertinent information i_i (i.e. features that are not described by the given u, but might be described by another $u_i \in U$) and sensor noise ϵ

$$I_l = \bigcup i_p(u) + \bigcup i_i(u) + \epsilon \quad \textbf{(22)}$$

Let us define an interpretation $X(u)$ of an utterance u as an ordered pair $X(u) = \{u, I_j \subseteq I\}$, which denotes that a set of features I_j from all the features I of the world is interpreted as u. Then a belief is defined following Eq. (23) as an ordered set of $X(u)$ interpreting all utterances u from U.

$$B = \{X(u_1), \dots X(u_n); n = |U|\} \quad \textbf{(23)}$$

Now, according to Eq. (24), we can calculate the belief B, which interprets in the most coherent way the observations made so far. It is done by looking for such a belief,

which minimizes across all the observations $o_q \in O$ the difference between the utterances $o_q[U_m]$[10] made on each particular observation by human, and those utterances U_{Bq}, which would make the robot by using the belief B. In other words, we are looking for a belief B, which would make the robot describe a particular scene with utterances as much close as possible to those, that would make a human on the same scene.

$$\arg\min_{B}\left(\sum_{q=1}^{|O|}\left|o_q[U_m] - U_{Bq}\right|\right) \tag{24}$$

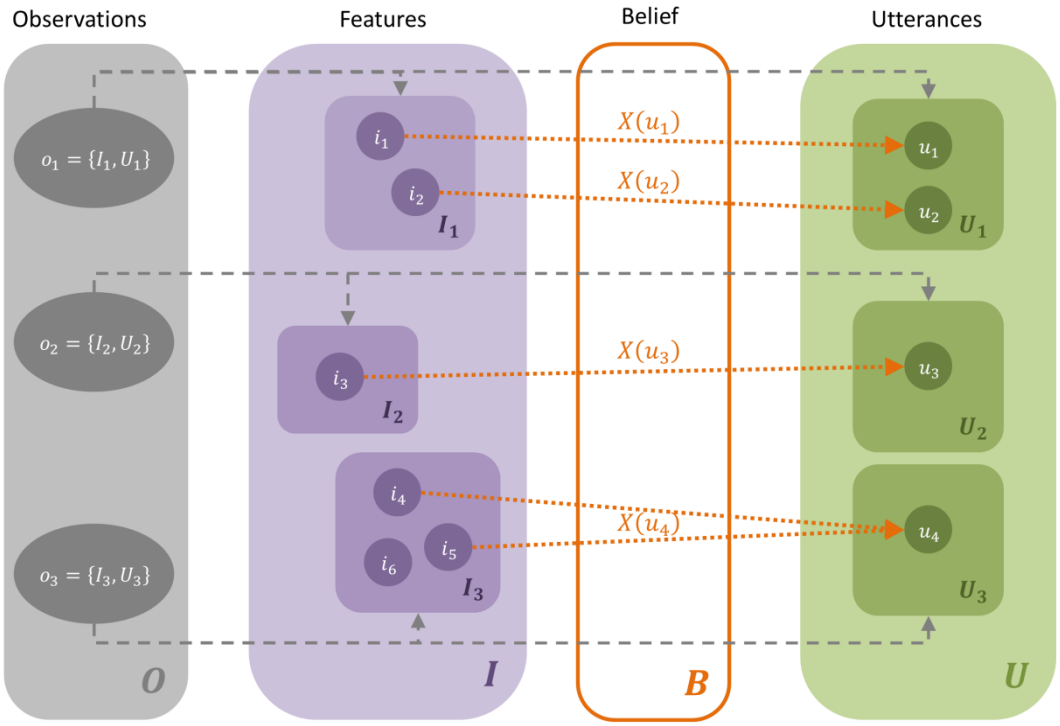

Figure 32: Graphical depiction of relations between observations, features, beliefs and utterances in sense of terms defined in the text.

On Fig. 32 an alternative view on the previously defined terms and their relations is presented. It depicts a state when three observations o_1, o_2, o_3 were made. On first observation, features i_1, i_2 were observed along with utterances u_1, u_2 and likewise for the

[10] To simplify the text and to avoid too much repetition, henceforth I will adopt the following notation. I use square brackets as e.g. in $o_q[U_m]$ to denote the set of utterances U_m from the observation o_q where $o = \{I_l, U_m\}$. Similarly $o_q[I_l]$ would be the set of features I_l belonging to o_q and so forth …

second and the third observation. It is visible that the entire set of features $\{i_1, \dots, i_6\}$ gives together the set I of all features ever observed, while sub-sets I_1, I_2, I_3 refer to features observed on particular corresponding observations. Similarly the utterances $\{u_1, \dots, u_4\}$ give the set U of all utterances and their sub-sets U_1, U_2, U_3 refer to corresponding observations. In this view an interpretation $X(u_1)$ is a relation of u_1 with a set of features from I. Then a belief B is a relation between the set of U to I following Eq. (25). All members of U map to one or more members of I and no two members of U map to the same member of I.

$$B: I \rightarrow U \tag{25}$$

4.3.2. Most Coherent Interpretation Search

To find a solution to Eq. (24), an exhaustive search over all possible beliefs will ensure that we eventually find the best fitting belief to explain the perceived world. However, this approach is impractical in real world conditions due to its high complexity, entailing an enormously large search space. Instead, we propose to search for a (sub)optimal belief B according to Eq. (24) by means of a genetic algorithm. Each organism within it has its genome constituted by a belief, which, results into genomes of equal size $|U|$ containing interpretations $X(u)$ of all utterances from U.

Let us have a belief generation process to generate genomes of organisms for the genetic algorithm as follows. For each interpretation $X(u)$ let us go through the entire set O of all observations made so far. On each observation $o_q \in O$, if $u \in o_q[U_m]$, features $o_q[I_l]$ are extracted. This set of features, as described in Eq. (22), contains pertinent and impertinent features (with respect to current u) and noise. The task of coherent belief generation is to generate beliefs, which are coherent with the observed reality. This is done by deciding, which features $i \in i_q[I_l]$ may possibly the pertinent ones. The decision is driven by two principles. The first one is the principle of proximity. As it is well known, similar things are more likely to be called the same name, than those less similar. As an application of it, any feature i is more likely to be selected to be pertinent in the context of u, if its distance to other features already selected is comparatively small. If a feature is too dissimilar to features interpreting a particular utterance, the feature is more likely to be considered impertinent in context of that particular utterance. The second factor is the coherence with all the observations in O. This means, that any observation $o_q \in O$, where $u \in o_q[U_m]$, has to have

at least one feature from $o_q[I_l]$ assigned into I_j of the current $X(u) = \{u, I_j\}$. Thus, it is both the similarity of features and the combination of certain utterances with certain features in observations from O, that guide the belief generation process. These beliefs may be perceived as "informed guesses" on the interpretation of the world made by the robot. The coherent belief generation procedure is outlined on Algorithm 5, which is for space reasons placed in Appendix C.

Before generating beliefs about the world, under some circumstances it might be appropriate to determine, whether there is enough information for a successful solution of Eq. (24). Take into account the following example. A robot is learning to name different shapes. It observes at one time a ball and a box while hearing "round" and "rectangular". Next time it observes a soup plate and a book, again hearing utterances "round" and "rectangular". Given those two observations, there is no way (without introducing a supplementary information e.g. by joint attention) to find a unique solution for Eq. (24). It is impossible to distinguish, whether the word "round" is applied to circular or to rectangular things (and likewise for the word "rectangular") as both interpretations have the same probability. Obviously this ambiguity may occur even with more than only two utterances. If such ambiguous situation happens, it may be identified by means of Algorithm 4 described under Appendix B. In this case the robot will attempt to make additional observations or will ask the human tutor to introduce new information which would allow for an unambiguous interpretation. In the given example, the robot could interact with the tutor in the following way: "I am unable to distinguish what is 'round' and what 'rectangular'. Please, show me some other round objects."

4.3.3. Evolution

4.3.3.1. Genetic Algorithms

Before explaining the further specific use of genetic algorithms in search of interpretation of observations, let us remind briefly some of their general principles. In the domain of machine learning, the idea of genetic algorithms was first exploited by (Holland, 1992) in his book "Adaptation in Natural and Artificial Systems". A genetic algorithm is a search heuristic through the solution space. In its function it is inspired by the process of natural selection, roughly based on the Darwinian theory of evolution from (Darwin, 1859) (in contrast to the Lamarckian theory of evolution, cf. (Ross, 1999)). It is used prominently,

but not solely, in cases, where the solution space is too large or in process of optimization, scheduling or engineering and design.

The main terminology of genetic algorithms is borrowed from biology and will be explained further:

- **Organism**: represents a single solution to the given problem, is constituted of a series of genes, which encodes the solution
- **Gene**: a constitutive part of each organism, in the basic generic algorithm, it may have binary values of 0 and 1, alternatively it may be arbitrary numbers, strings or complex objects
- **Population**: a group of organisms evolving together, each organism is evaluated by the fitness function, which determines its chances to take part on reproduction in creation of next generation's population
- **Fitness function**: is a function evaluating how "good" is the solution encoded by the given organism with respect to the desired state, when an organism receives high enough value of fitness, the evolution is usually stopped and the organism is presented as the final solution
- **Crossover**: best fitting organisms have better chances to transfer their genes to the next generation's population, it takes parts of genomes of parenting organisms and recombines them creating offspring; the aim is to transfer genes that contribute to the desired solution, into the next generation
- **Mutation**: in order to maintain genetic diversity, mutation is applied, changing with a small probability certain genes in a random fashion
- **Elitist selection**: best solutions (organisms) may be lost due to mutation or bad crossover, hence one or several best fitting organisms are automatically transferred to the next generation without any alternation

The operation cycle of a genetic algorithm is resumed on Fig. 33. The first generation population is generated randomly. Then fitness of each organism is calculated. If the best fitness is above the given threshold, evolution stops, otherwise reproduction is performed, generating new organisms based on best fitting parents. Mutation is performed as well, the resulting population is transferred to the next generation and the cycle continues.

In this work, I am mostly using an extended notion of genetic algorithm. Contrary to (Holland, 1992), in my work genomes are not composed of a series of bits (0 or 1), but are rather represented by chains of real numbers (as in 3.6.4) or complex objects (as in 4.3.3). Nonetheless, the scheme of operation of the genetic algorithm remains the same.

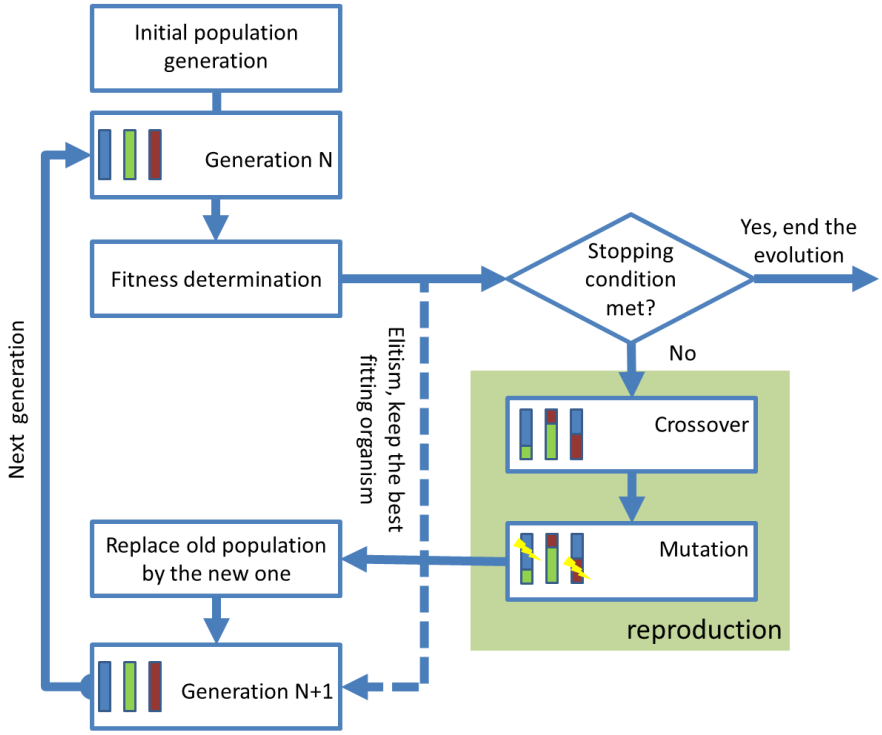

Figure 33: A general schema of genetic algorithm operation.

4.3.3.2. Fitness Evaluation

In the previous section, an approach has been defined for generation of coherent beliefs about the world, which are coherent with existing observations. Each of these beliefs makes one organism, which is used inside a genetic algorithm. To evaluate the fitness of each organism, a good fitness function is crucial. Here, the fitness function is defined as an inverted value of the sum in Eq. (24). To evaluate a given organism, a classifier is trained, whose classes are the utterances from U and training data for each class u are given by $X(u)[I_j]$, i.e. the features associated with the given u in the genome. Then, we can use this classifier through the entire set of observations O, and for each $o_q \in O$ its observed features $o_q[I_l]$ are classified, which results in a set of utterances U_{Bq} (provided that a belief B is tested on the q-th observation from O).

Having previously calculated the set U_{Bq}, which are the robot's utterances interpreting features observed in the q-th observation, this set can be compared with $o_q[U_m]$, i.e. the utterances made on the same features by a human. Here, we are in the core of the sum from

Eq. (24) and the distance $\left|o_q[U_m] - U_{Bq}\right|$ can be finally calculated as the disparity between sets $o_q[U_m]$ and U_{Bq} for each q respectively. The disparity is calculated as follows: let $\frac{1}{1+v}$ be the fitness. The value of v is given as in Eq. (26), i.e. it is the number of elements that are not present in both sets, which means missed and superfluous utterances interpreting the given features.

$$v = \left|o_q[U_m] \cup U_{Bq}\right| - \left|o_q[U_m] \cap U_{Bq}\right| \quad \textbf{(26)}$$

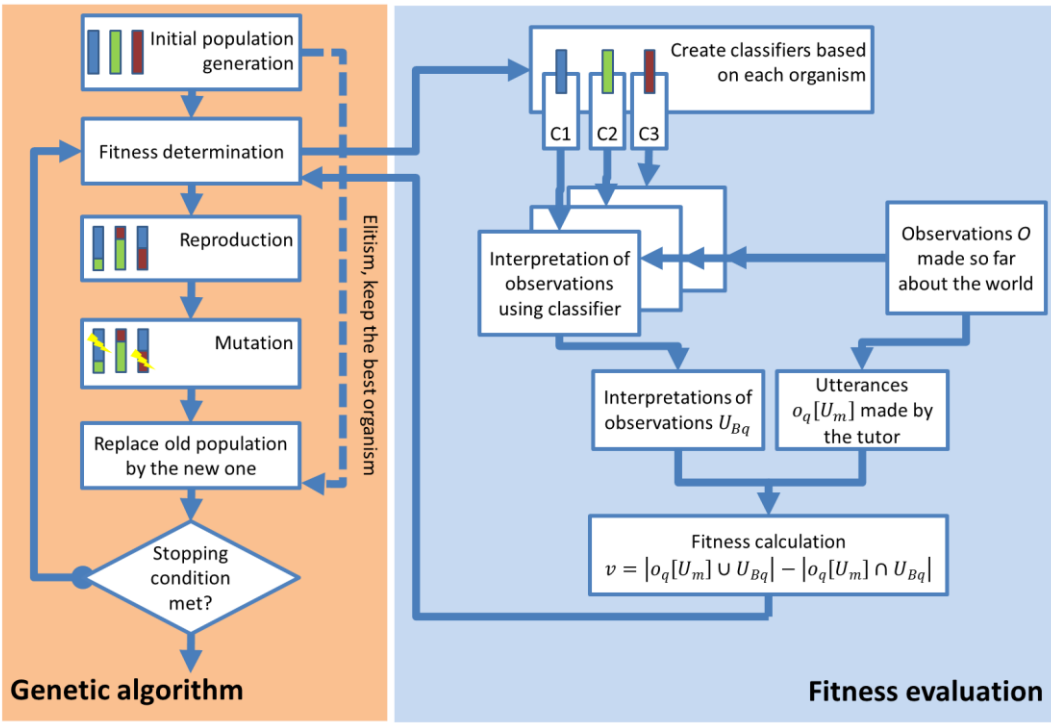

Figure 34: Graphical depiction of genetic algorithm workflow described in sub-section 4.3.3. The left part describes the genetic algorithm itself, while the right part focuses on the fitness evaluation workflow.

At the end of the evolution, the globally best fitting organism is chosen as the belief that best explains observations O made so far about the world.

On Fig. 34 a functional diagram of the genetic algorithm described in this sub-section is depicted. On the right-hand side of the diagram the workflow of fitness evaluation is depicted identifying the key parts of the fitness calculation process. On the left hand side (the

genetic algorithm), the step of reproduction and the mutation are further detailed in sub-section 4.3.3.4.

4.3.3.3. Population Control

To avoid possible loss of the best performing organism in each generation, I apply the elitism rule, which automatically preserves the best fitting organism into the next generation. Thus the evolution of fitness is monotonically increasing.

To avoid stagnation of evolution due to too high homogeneity of the population, the saturation of population ϱ is calculated according to (27), where pop_u is the number of unique genomes in the population and pop_{total} is the total size of the population. If in any generation the ϱ is higher than 0,5 (i.e. less than half of the population is constituted of unique genomes), all the organisms are mutated with increased probability, thus introducing genetic diversity into the population. This is done with the exception of the best fitting organism, which is preserved from mutation.

$$\varrho = 1 - \frac{pop_u}{pop_{total}} \quad \textbf{(27)}$$

4.3.3.4. Crossover Operator

The crossover operator serves to produce offspring of the actual population in order to propagate potentially benefit genes into the new generation. The standard one-point crossover approach is used here, where the genome of two organisms (parents) is split at random point. All genes beyond that point in either organism string are swapped so that both offspring receive the first part of the genome from their first parent and the second one from their second parent. Parents are selected by "Roulette wheel selection" method from (Holland, 1992), where the probability of an organism being selected is proportionate to its fitness according to Eq. (28), where P_i is the probability of the i-th organism to be selected as a parent. fit_i denotes the fitness of the i-th organism.

$$P_i = \frac{fit_i}{\sum_{j=1}^{pop_{total}} fit_j} \quad \textbf{(28)}$$

4.3.3.5.Mutation Operator

The primary reason for the mutation operator is to maintain genetic diversity in the population in order to avoid local minima. The most common way of realization of mutation is to pick up a random gene of an organism and change its value randomly to another value from the range of possible values.

In the population, beliefs are playing the role of organisms and interpretations are playing the role of genes. Application of the mutation operator would thus require a particular interpretation $X(u) = \{u, I_j \subseteq I\}$ of utterance u to be changed randomly. However this is not acceptable, because this way the entire belief could lose its coherency. As a consequence, such organism could achieve only low fitness. This would be due to its lack of coherency with the real world and inherently incorrect interpretation of the world. Remember that in sub-section 4.3.2 beliefs are not produced randomly, but are carefully generated in a way that makes them coherent with the observed reality.

Instead of making random changes in the genome, which would mostly lead to lose of fitness, I re-define the mutation in the following way. Let us mutate a particular $X_i(u) \in B$, where B is the belief constituting the currently processed organism. As the interpretation $X_i(u)$ is defined as $X_i(u) = \{u, I_j \subseteq I\}$, the mutation will induce changes of the set of features I_j associated to interpretation $X_i(u)$ in the way that does not harm the coherency of the entire belief B. It is achieved by at first generating an auxiliary set of features I_{aux}. This is a set of all the features that have been observed in presence of utterance u from $X_i(u)$ but which are not contained in any other interpretation of the current B. In other words, I_{aux} is a set composed of "free" features and those that were previously rejected as being possibly impertinent. All of them, of course, are connected to u by observation. Constructing the set I_j of $X_i(u) = \{u, I_j \subseteq I\}$ from members of I_{aux} ensures the entire belief B remains still coherent with reality.

Having the feature set I_{aux} ready, its random member is picked and placed into the set I_j. For all the other features in I_{aux} their average distance to members of I_j is measured. Based on this distance they are either included into the I_j or they are discarded as impertinent in a similar way features are distributed in Algorithm 5. Eventually the interpretation $X_i(u) = \{u, I_j \subseteq I\}$ contains a new set I_j, which is different to the original one and thus the genome has been mutated while its coherency has been preserved.

4.4. Human-robot Interaction during Learning

As showed in works from domains of linguistics and psychology (see e.g. (Waxman and Gelman, 2009)), human language is not a mere static set of "tags", that we give to entities of the world around us, but it is a dynamic system, which influences the perception and which is at the same time influenced by the perception. An example of this flexibility has been reported in (Tomasello, 2003), p. 73: "For example, many young children overextend words such as dog to cover all four-legged furry animals. One way they home in on the adult extension of this word is by hearing many four-legged furry animals called by other names such as horse and cow".

Another important remark is that human beings learn both by observation and by interaction with the world and with other human beings. The former is captured in our system in the "best interpretation search" outlined in sub-section 4.3.2. It is a state resembling to human infants in pre-lingual age. The latter type of learning requires that the robot is able to communicate with its environment (as it is a case for a child with developed speech capabilities) and is facilitated by previous learning by observation, which may serve as its bootstrap. In the present approach, this learning by interaction is carried out in two manners implying two directions: human-to-robot and robot-to-human.

Let us have a robot co-operating with its human counterpart. The first manner (human-to-robot) is employed anytime the robot interprets wrongly the world (due to incomplete knowledge about it, e.g. by bringing a "purple" mug when asked for a "red" one, provided that it has never encountered a "purple" thing before and thus it interprets it as a "red" one). If the human sees this wrong response, he provides the robot a new observation by uttering the desired interpretation ("purple" in our example) in presence of the wrongly interpreted features (i.e. the purple mug). The robot takes this new, corrective, knowledge about the world into account and searches for a new interpretation of the world (sub-section 4.3.2), which would be conform to this new observation.

The second manner (robot-to-human) may be employed when the robot attempts to interpret a particular feature. If the classifier trained with the current belief classifies the given feature with a very low confidence, this may be a sign, that this feature is a borderline example. In this case, it may be beneficial to clarify its true nature. Thus, led by the epistemic curiosity, the robot asks its human counterpart to make an utterance about the observation in question. If the robot's interpretation is not conforming to the utterance given by the human (robot's interpretation was wrong), this observation is recorded as a new knowledge and a search for the new interpretation of the world is started as in the previous case.

Using these two ways of interactive learning, the robot's interpretation of the world evolves both in quantity, covering increasingly more phenomena as they are encountered, and in quality, shaping the meaning of words (utterances) to conform with the perceived world.

4.5. Generalization for Multimodal Data

So far we have been interested in establishing the word-meaning relationship in an intelligent agent using one sensor only. In this section, this approach is going to be extended to multiple sensors. It should be noted here, that these sensors do not necessarily need to be the classical ones (camera, laser or ultrasound device) and we can work with soft-sensor. Thus, an agent equipped with a camera can work with several soft-sensors based on the camera image, such as sensors of color, shape, motion, egocentric position etc. It is in this context, that terms like "sensor" or "multimodal data" will be used henceforth.

The situation, with which we are dealing in the case of multiple sensors, can be described by the following example. We have an agent perceiving through its camera a big yellow book on a shelf (along with other "impertinent" visual information) and a ping-pong ball. This view (observation) is accompanied by an utterance "big yellow rectangle" and "white small sphere" given by a human.

Let us have this agent equipped by sensors of color, shape and size. The task of the agent now is not only to associate the appropriate feature (e.g. the RGB[242, 251, 0] value) to the correct utterance, which in this case would be "yellow". Moreover he needs to decide, which utterances belong to which sensor to prevent a cross-talk situation in the genetic algorithm, i.e. trying to interpret a particular feature by an utterance that can never be associated with it. In our example this would be the situation of trying to interpret the square-like shape feature by utterances like "yellow" or the round shape by "small", which were in fact uttered in the presence of the mentioned features, but were not addressing them.

4.5.1. Observing Features from Multiple Sensors at the Same Time

Let us have an ordered set of sensors $S = \{s_1, \dots, s_r\}$. To reflect the presence of multiple sensors in observations, I extend the definition of an observation from sub-section 4.3.1 following Eq. (29):

$$o = \left\{\{I_{l1}, \dots, I_{lr}\}, \{U_{m1}, \dots, U_{mx}\}\right\} \quad \textbf{(29)}$$

It means that the human provides on each observation a set of sets of utterances (i.e. a set of phrases) $\{U_{m1}, \dots, U_{mx}\}$ to accompany the set of sets of features $\{I_{l1}, \dots, I_{lr}\}$, each set from I_{l1} to I_{lr} retrieved by one of the sensors from $S = \{s_1, \dots, s_r\}$. Each set of utterances (each phrase) from U_{m1} to U_{mx} is describing a particular subject on the scene. Following the example given earlier, we could denote seeing a ping-pong ball and a big yellow book on the same scene as an observation denoted by the following: $o = \left\{\{I_{l1}, \dots, I_{lr}\}, \left\{\{big, yellow, rectangle\}, \{white, sphere, small\}\right\}\right\}$, provided that $\{I_{l1}, \dots, I_{lr}\}$ would represent the set of all features extracted from the scene by each of the sensors. To address the cross-talk problem and to establish the word-meaning relation using multimodal data, the previously described approach is extended by a new layer of processing. The previously (sub-section 4.3.1) introduced concept of interpretation of an utterance defined an interpretation as $X(u) = \{u, I_j\}$. In a similar way as let us have an interpretation of a particular sensor $s_t \in S$ defined as an ordered pair $X_S(s_t) = \{s_t, U_m\}$. This means that utterances from U_m are all associated to a particular sensor s_t and they describe linguistically different features, which can be found by this sensor. For example the set U_m belonging to a shape-detecting sensor may contain utterances like "round", "rectangular", "triangular" and so forth. The union of all utterances associated to all sensors gives exactly the set U. On the other hand in order to function correctly, there is absolutely no need for the cognitive system to have each object fully qualified with respect to all its sensors. For example, if a robot has the capacity to perceive the color, sound and the tactile information, there is no need to require the tutor to qualify an object as "red, furry and clicking". The qualification can be (and in most cases is) only partial. The tutor could e.g. say only that the object is "red". The robot would then add all the missing perceptual categories to the object from its previous experience with other "furry" or "clicking" objects.

4.5.2. Belief Generation and Co-evolution with Multiple Sensors

Again similarly to the definition of a belief $B = \{X(u_1), \dots, X(u_n); n = |U|\}$ from sub-section 4.3.1, let us define a belief about multimodal data as an ordered set of interpretations of sensors, following Eq. (30).

$$B_S = \{X_S(s_1), \dots, X_S(s_r); r = |S|\} \quad \textbf{(30)}$$

Beliefs generated about appurtenance of each of the utterances from U to a particular sensor can be generated in the manner similar to coherent belief generation procedure from sec. 4.3.2.

We are thus searching for a belief B_S, which divides the set of U among all existing sensors S in the most coherent manner, i.e. in the manner with the lowest possible crosstalk. As lower crosstalk will generally lead to more accurate interpretations, the fitness of a particular belief B_S can be safely defined as the best average fitness f_a achieved in evolution for each single sensor using interpretation from B_S. Therefore I propose to run a co-evolution of several such beliefs.

Each belief gives an environment for a genetic algorithm (sub-section 4.3.3), which is then run in the context of a particular sensor. The description of the mentioned genetic algorithm is modified so that given a $X_S(s_t) = \{s_t, U_m\}$, where each $X_S \in B_S$, we try to interpret all features observed by sensor s_t by using only utterances from U_m. As a result of this process, we obtain a belief B_S showing which utterances are commonly associated to which sensor, and at the same time a theory B for each sensor interpreting the observations made on it. Thus the word-meaning anchoring is achieved in in multimodal data. A schematic depiction of such co-evolution is shown on Fig. 35. Each box from the "co-evolution" field represents one complete genetic algorithm described in sub-section 4.3.3.

It is possible that the approach of multilevel evolutionary algorithm presented recently in (Akbari and Ziarati, 2011) could be adopted or modified to be used in place of the aforementioned co-evolution process. This is a matter of further investigation.

4.6. System Validation

4.6.1. Simulation

To evaluate the behavior of the described system, it has been implemented first in a simulated environment with a virtual robot. Then, it has been have deployed on a real humanoid robot and tested using complex real-world objects. Both experimental settings are described below. The approach is applied on the color names learning problem. In everyday

dialogs, people tend to describe objects, which they see, with only a few color terms (usually only one or two), although the objects in itself contains many more colors. Also different people can have slightly different preferences on what names to use for which color and all those aspects are highly varying across the cultures. Due to this, learning color names is a difficult task and it is a relevant sample problem to be used to test the present system.

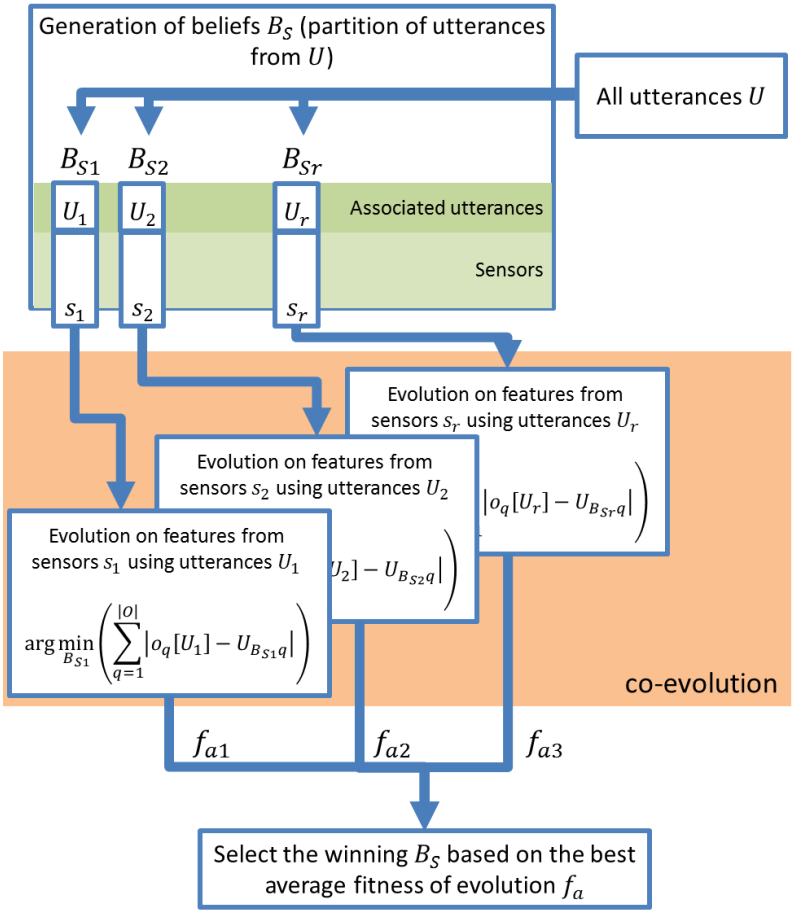

Figure 35: Schema of co-evolution during search of interpretation in case where multiple sensors are used. Symbols used are explained in text.

In the simulated environment, images of real-world objects were presented to the robot alongside with textual tags describing colors present on each object. The images were taken from the Columbia Object Image Library database (Nene, Nayar and Murase, 1996), which contains color images of 100 every-day objects taken from different perspectives, in total 1000 images. Five subjects (all fluent English speakers) were asked to describe each object in terms of colors as if they were describing it in a normal conversation. For coherency the choice of colors was restricted to "black, gray, white, red, green, blue and yellow", based

on the color opponent process theory (cf. (Schindler and Goethe, 1964)). The average number of colors per object given by the subjects was two. The tagging of the entire set of images was highly coherent across the subjects. In each run of the experiment, one tagged set coming from a random subject was chosen.

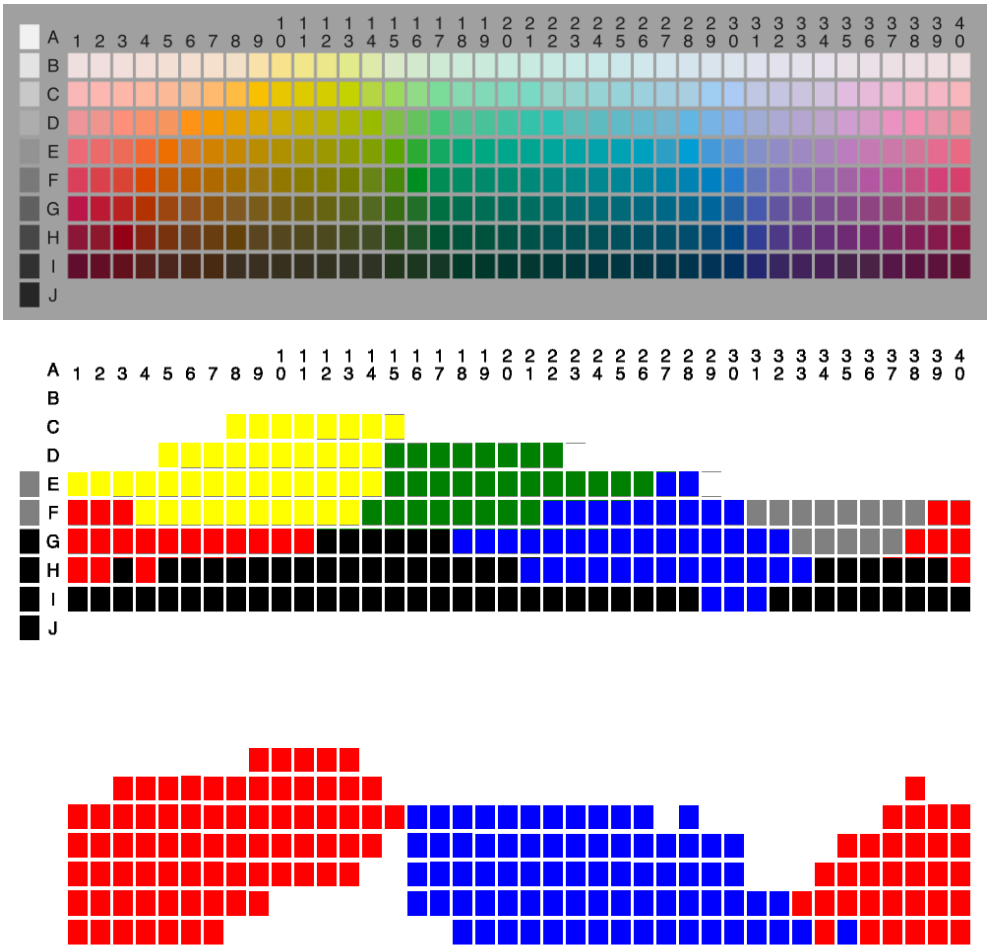

Figure 36: Upper: the WCS color table. Middle: interpretation made by the robot regarding each color present. Lower: the WCS color table interpreted by robot taught to distinguish warm (marked by red), cool (blue) and neutral (white) colors.

In this experiment, the utterances (see sub-section 4.3.1) were given in the form of text strings as color terms extracted from the descriptions. Images were segmented by algorithm described in 3.4.1 and the average color of each segment was used as a feature. As for the classifier required in sub-section 4.3.3, a classical Bayesian classifier has been used.

First, the available images of objects were divided randomly into two equally large sets, one for learning and the other for testing. After the learning took place, objects from the

testing set were presented to the simulated robot. Each object was interpreted by a set of color names and the robot's interpretation was compared to the ground truth given by the human subject (see. Eq. (24)). The object was accepted as correctly interpreted if the robot's and the human's interpretations were equal. The rate of correctly described objects from the test set was approximately 91% in average with the robot fully learned. To allow for visual evaluation, a color table[11] from World Color Survey (Kay et al., 2003) has been used. On Fig. 36 (middle) a sample interpretation of the colors of the WCS table are given after the system was learned. Fig. 37 shows visualization of interpretations of several testing objects from COIL database made by our system after learning took place.

Figure 37: Several objects from the COIL database. The second row shows visualizations of interpretation of those objects by our system fully learned.

To investigate the speed of learning, a series of 13 tests have been run. In the first test, the number of the objects in the training set was selected in the manner that it contained only one occurrence of each color. The second test contained two occurrences of each color and so forth. One half of the COIL objects (containing unused objects) formed the testing set. Each test was run ten times and the average values were plotted on the graph on Fig. 38. With one exposure, the system was capable to describe using correct terms only about 38% of the testing set. This number rises fast to approximately 80% after 5 or 6 exposures and then continues to rise slowly towards 90%.

[11] Available online on: http://www.icsi.berkeley.edu/wcs/data.html

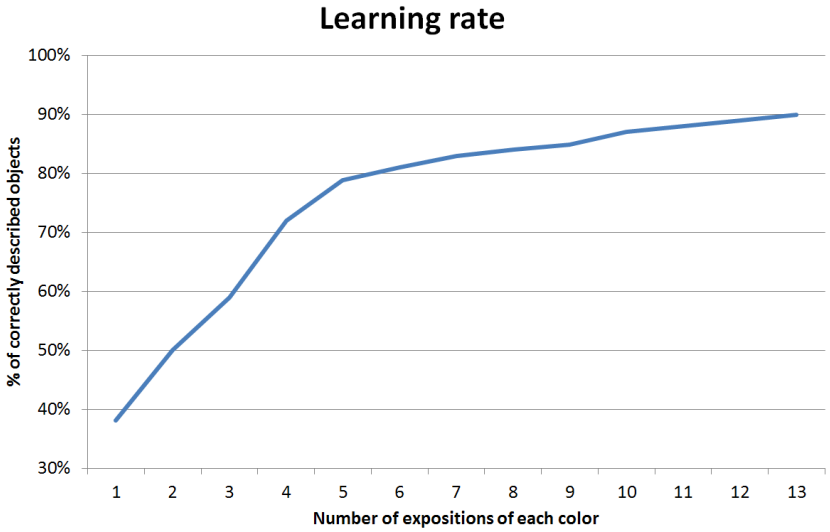

Figure 38: Evolution of number of correctly described objects with increasing number of exposures of each color to the simulated robot.

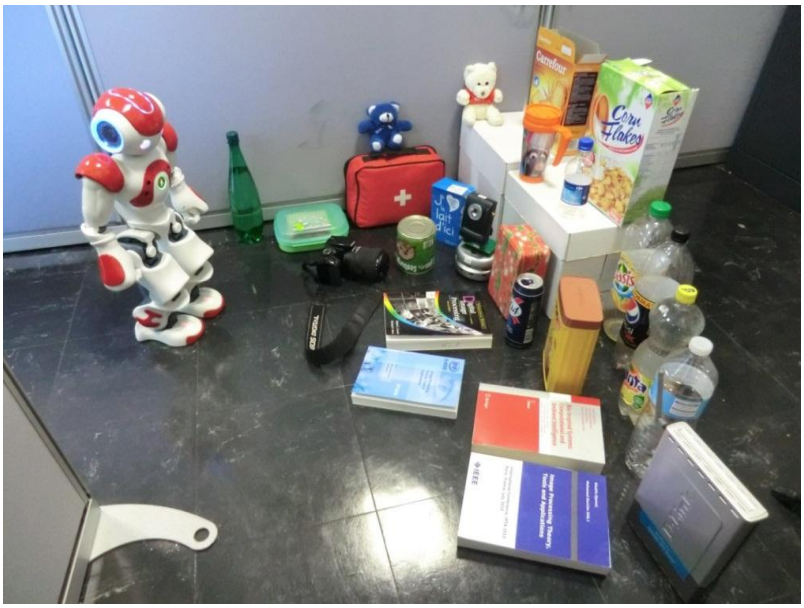

Figure 39: The humanoid robot alongside the objects used in experiments described in sub-section 4.6.2.

Finally, the same five subjects were asked to go again through the COIL image set, this time determining on each object, whether it contained warm colors (red, yellow, orange, ...), cool colors (blue, green, ...) or neutral colors (white, gray, black, ...) or any combination of them. Selecting again randomly one half from the whole image set for training, the virtual

robot was required to learn to distinguish colors based on their temperature. In average 96% of objects were correctly described once the learning was finished. On Fig. 36 (lower) is presented an example of interpretation of the WCS color table done by the robot after learning took place.

4.6.2. Experiment in Real Environment with a Humanoid Robot

After verifying the approach in a simulated environment, it has been tested using a real humanoid robot[12]. Several experimental scenarios have been created to verify both the acquisition of knowledge and its use by the robot. The total of 25 every-day objects was collected for purposes of the experiment (see Fig. 39). They were randomly divided into two sets for training and for testing.

The learning set objects were placed around the robot and then a human tutor pointed to each of them calling it by its name. Using its 640x480 monocular color camera, the robot discovered and learned the objects around it by the salient object detection approach I have described earlier in sub-section 3.4. Here, this approach has been extended by detecting the movement of the tutor's hand to achieve joint attention. This way the robot was able to determine what object the tutor is referring to and to learn its name. The tutor addressed to the robot in natural speech. The TreeTagger tool[13] was used in combination with robot's speech-recognition system to obtain the part-of-speech information from situated dialogs. Standard English grammar rules were used to determine whether the phrase is demonstrative (e.g. “This is an apple.”), descriptive (e.g. “The apple is red.”) or an order (e.g. “Describe this thing!”). To communicate with the tutor, the robot used its text-to-speech engine.

After learning the names of objects, those were presented randomly to the robot alongside of other objects. Then the tutor described each object in the view by saying e.g. “The handbook is black”. The robot was required to localize the object (e.g. “handbook”) among the presented objects based on the previous learning and to extract its color features along with the uttered information that the book was “black”. Then learning of the colors took place.

[12] A video capturing different parts of the experiment may be found online on: http://youtu.be/W5FD6zXihOo

[13] Developed by the ICL at University of Stuttgart, available online at: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger

To verify the results of learning, objects from the testing set were presented to the robot by the tutor, who then asked the robot to describe objects he was pointing towards (see Fig. 40). Using the same joint attention scheme, as described before, the robot extracted the object in question, interpreted its appearance (see Fig. 41 for visualization) and spoke aloud the colors it believes the object contains. When the robot described the object by wrong terms, the tutor corrected it by uttering the correct properties. In this case, the robot would add this information to the already gathered knowledge and use it in further learning. The tutor would latter verify whether the robot has assimilated correctly the new information by presenting the same object in new conditions. The last experimental scenario we used involved distinguishing objects of the same class based on their color. First, several objects of the same class (e.g. a "book") were presented to the robot. Then, the robot was asked to locate a particular object among all others given its color (see Fig. 42).

Figure 40: Left: the experimental setup. The tutor is asking the robot to describe the box he is pointing on. Left: the robot's view, with the object in question detected. The robots response was "It is yellow".

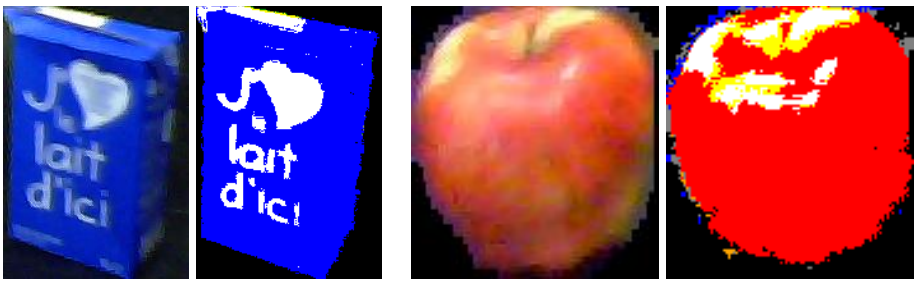

Figure 41: Two objects extracted from robot's surroundings. Right: the original image, left: features interpreted. For the "apple", the robot's given description was "the object is red". For the box, the description was "the object is blue and white".

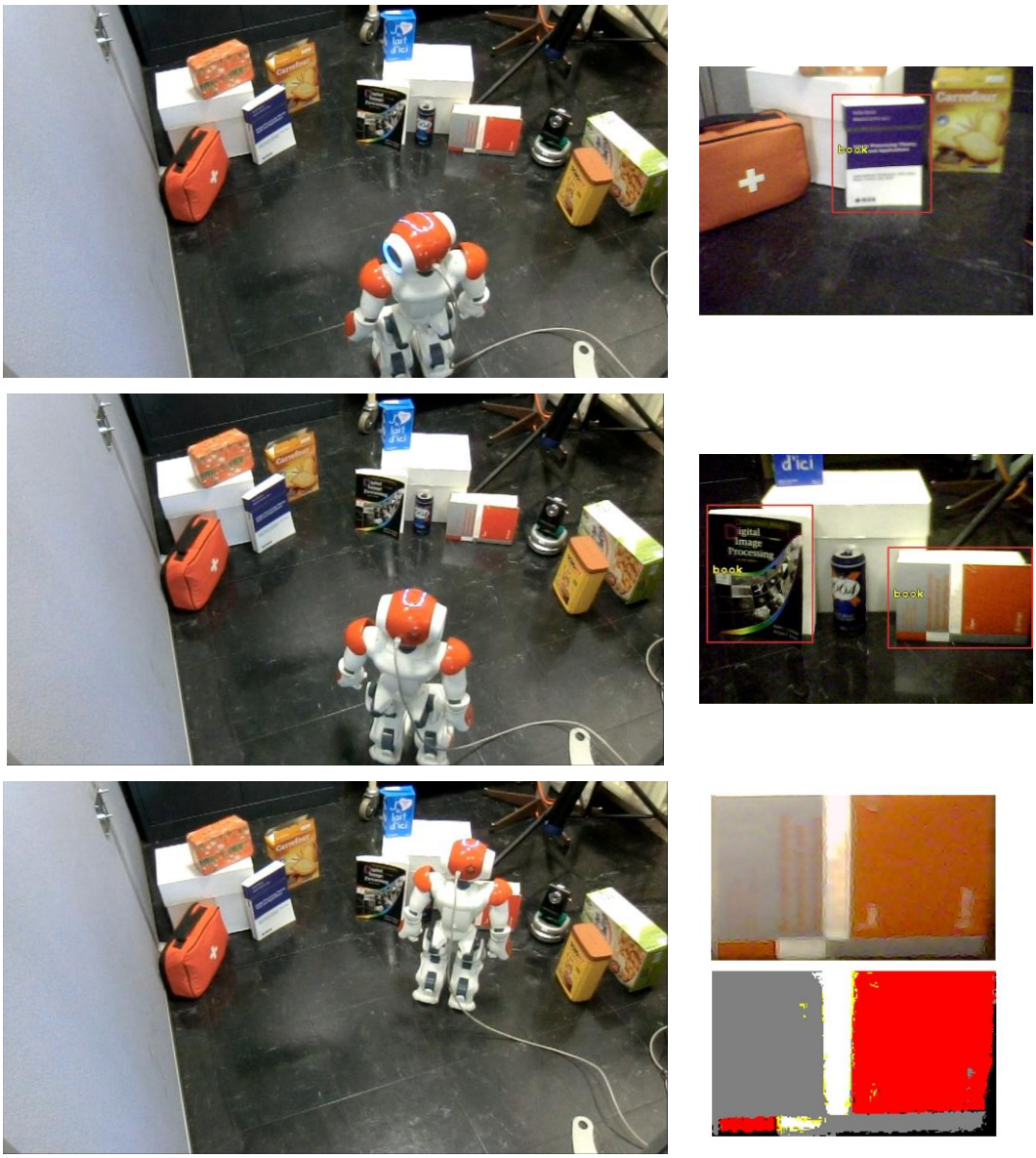

Figure 42: Images from a video sequence showing the robot searching for the book (1st row). Localizing several of them, it receives an order to fetch the one which is "red" (2nd row). The robot interprets colors of all the detected books and finally reaches the desired one (3rd row). The right column shows robot's camera view and visualization of color interpretation of the searched object.

4.7. Conclusion

This chapter discusses details of realization of an epistemic curiosity driven high level cognitive system allowing a humanoid robot to learn in an autonomous manner new knowledge about the surrounding world. The system has been first described in a general manner. Then, in order to validate it, it has been applied on a concrete problem, i.e. autonomous learning of names of colors. This learning has been done based on independent observation of colored objects and from further interaction with humans in non-trivial conditions using real-world objects.

Experimental results in simulated environment are provided and then the approach is verified on a humanoid robot in a real-world environment using every-day objects. However, it is pertinent to underline, that no constitutive part of the described approach is task-specific. Although in system validation the results obtained on the problem of learning of colors are presented, the approach is not tied to any particular type of features. Rather it provides a generic framework which allows learning of any kind of features … as far as there exists an appropriate classifier for them. As a consequence a robot may benefit from this versatility to acquire knowledge about physical qualities such as color, shape, size as well as about "non-material" features such as spatial position relations (on the left, on the top, …) or names of types of motion (falling, rolling left, …).

We have observed that using the described learning approach, the robot endowed with the present cognitive system was able to successfully anchor the meaning of uttered words to its perception. This was verified by learning basic color terms and by learning to distinguish color temperature. In both tasks, the system was able to achieve high success rate (91% and 96% respectively) on the testing set. The robot's interpretations of colors from the WCS table were also close to human (see Fig. 36).

The present approach exhibits two interesting properties found in human children learning. The first one is the ability to learn without a negative input, just by a mere observation. The second one is the capacity to achieve high percentage of correct interpretations after only a few exposures to the stimulus (see Fig. 38, the learning curve is steepest during first few exposures). Experiments with the real humanoid robot have confirmed the practicability of our approach in real-world environment. The robot was able to interact with the tutor in real time.

At the current state I have provided validation results showing the robot is able to learn a single category or property at a time, (e.g. the color in utterances like "it is red"). In section 4.5 I lay down a basis for an extension of this system in order to learn multiple

categories at the same time and to distinguish, which of the used words are related to which category. This will be done by means of co-evolution of several instances of the system described in this chapter, each of them dedicated to a different category. The validation of the proposed extension will be a part of future work on this learning system.

Chapter 5. Towards Autonomous Knowledge Acquisition in Real World Conditions

5.1. Introduction

In Chapter 2 a theoretical basis of my approach to autonomous acquisition of knowledge has been outlined and its concept has been presented. Through Chapter 3 and Chapter 4 solutions have been proposed to different aspects of this concept. Each of them has been validated separately in order to verify if they are capable to play their role it he envisaged cognitive system. In this chapter I am going to bring all the previously mentioned parts together in one functioning system and I will describe its application. Through the chapter, the design choices made in practical realization of the system, will be explained as well as results obtained by its deployment in real world conditions. Thus, this chapter is meant to provide an "all-in-one" realization description and validation of the entire system and to show how it contributes to research towards enabling fully autonomous knowledge acquisition capabilities in a robot.

In the first part of this chapter, section 5.2, the NAO robot is presented. It is a humanoid mobile robot platform used in experiments throughout this work. It is pertinent to describe here its general properties and its particularities as they inherently influence the practical application of my work in real environment. In section 5.3 I describe algorithms and techniques that have been adopted in realization of the described system. The section 5.4 is dedicated to description of performance of the entire system in real world conditions. The chapter is concluded with section 5.5.

5.2. NAO, the Humanoid Robotic Platform

NAO is a humanoid robot manufactured and merchandised since 2004 by a French company Aldebaran Robotics[14]. This sub-section familiarizes the reader with this robot.

[14] Company website: http://www.aldebaran-robotics.com

5.2.1. Overall Description of the Platform

Compared to other humanoid platforms like the HPR or Honda's Asimo, the robot is relatively small: about 58cm in height with weight slightly exceeding 4kg. The hardware version of the robot used by our laboratory is V3 and the following technical details refer to this version. It has 25 degrees of freedom, resumed on Table 3. They enable the robot to walk, grasp small objects and turn its head in exploring the environment with sufficient freedom.

For simulation purposes, a virtual version of Nao is available for the Webots simulation program, which was developed by Cyberbotics[15]. Basic simulation capabilities are offered also by Choreographe, an application shipped with the robot. Those were, however, used only in early stages of development and practically all the presented work has been done using the real robot.

	Head	**Arms**	**Grasping**	**Pelvis**	**Legs**
DOF #	2	5 (2x)	1 (2x)	1	5 (2x)

Table 3: Degrees of freedom of NAO, the humanoid robot

5.2.1.1. Sensors and Communication Devices

Concerning the available sensors, the robot is equipped with two color CMOS cameras with resolution up to 640x480 pixels each in a non-stereo arrangement. One camera is on the front of the head and ensures the forward vision. The other one is mounted lower on the head and is covering the space around the feet of the robot that is unreachable by the main camera. Two-channel sonar is mounted into the robot's chest and two infrared distance meters are mounted where the robot would normally have eyes. The robot also possesses a tactile sensor, bumpers and inertial sensors. To interact with humans, robot is equipped with two loudspeakers and a voice synthesizer capable of "text-to-speech" in English and French. This is completed with four microphones a speech recognition unit. The robot is depicted on Fig. 43 with important parts being annotated.

[15] Company website: http://www.cyberbotics.com/

The NAO can operate in fully autonomous mode using its onboard AMD Geode 500MHz processing unit to run programs and behaviors stored in its memory. Alternatively, it can be operated remotely from another computer via a network (WiFi or Ethernet) connection.

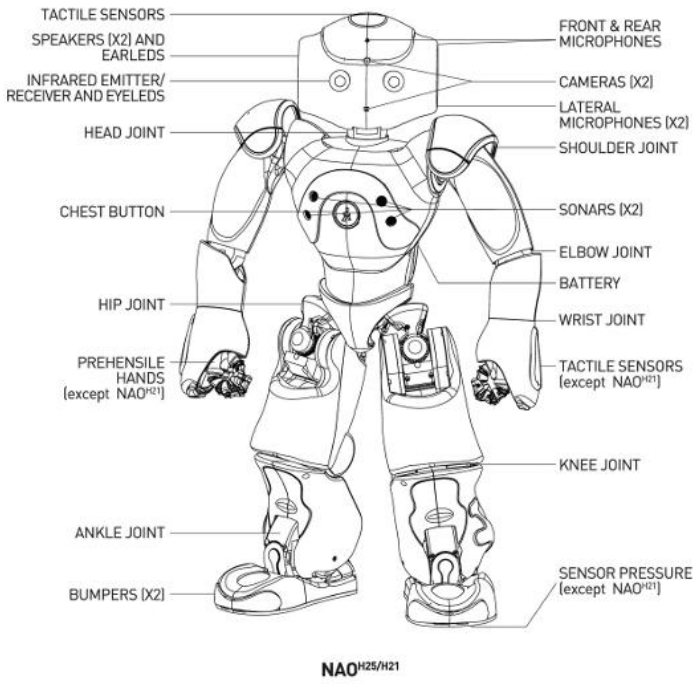

Figure 43: NAO robot V3. Scheme adopted from Technical documentation provided by Aldebaran Robotics.

5.2.1.2. Motion Control

When discovering the environment, the robot cannot stay motionless, but it walks around to take a closer look to objects, search for an object or simply to follow a human tutor. For this reason the main principles of control of the robot's walking will be very briefly presented here.

NAO is capable of omnidirectional walk. The walk is stabilized using feedback from sensors placed in its joints and inertial sensors. This makes the walk robust against small disturbances and absorbs torso oscillations in the frontal and lateral planes. The robot is capable of walking on different floor surfaces such as carpet, tiles and wooden floors;

however the assumption is that the floor is flat. As a consequence the robot can have difficulties with walking on uneven surface.

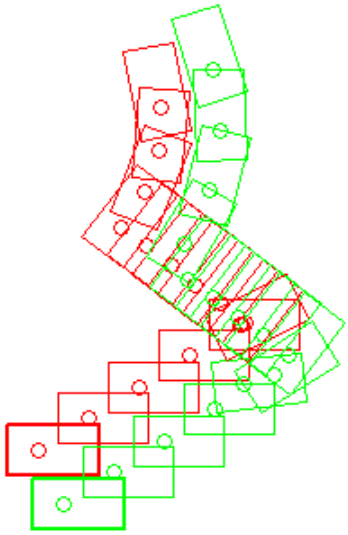

Figure 44: Omnidirectional walk of NAO showing different walking patterns. Red is the left leg, green is the right leg. Adopted from Tech. doc. from Aldebaran.

The robot's walk control is using a simple dynamic model called Linear Inverse Pendulum. This model is inspired by the classical work of (Kajita and Tanie, 1995) and detailed e.g. in (Kajita et al., 2009). The model is solved using quadratic programming method presented in (Wieber, 2006). Each step is composed two phases: a double leg support and a single leg support phase. The double leg support time uses one third of the step time. The preview controller length is 0.8s. The walk is initialized and ended with a 0.6s phase of double support. More recently in (Gouaillier, Collette and Kilner, 2010) authors from Aldebaran research have presented an updated closed loop walk algorithm improving both the stability and speed of the walk. On Fig. 44 a graph is shown presenting three different walking patterns generated by omnidirectional walk control unit of NAO.

5.2.2. Software and Hardware Architecture Employed

5.2.2.1. Remote Processing Architecture

As it has been said in sub-section 5.2.1.1, NAO possesses an on-board CPU AMD Geode 500MHz. However its processing speed and its limited memory do not allow real-time

execution of computationally heavy tasks such as many image processing and machine learning algorithms. This has determined the choice of remote processing architecture. As a consequence all computationally demanding tasks are executed on a remote full-featured PC and the input and the output is communicated via network. An appealing option would be to use a wireless connection. However the NAOqi[16] API the Aldebaran is currently providing uses a transmission protocol which has limited efficiency on some data structures, such as images. Such transfers consequently require a large bandwidth and this demand cannot be satisfied by standard WiFi transfer. For this reason during most of the experiments with NAO it had to be connected by an Ethernet cable to the remote processing station.

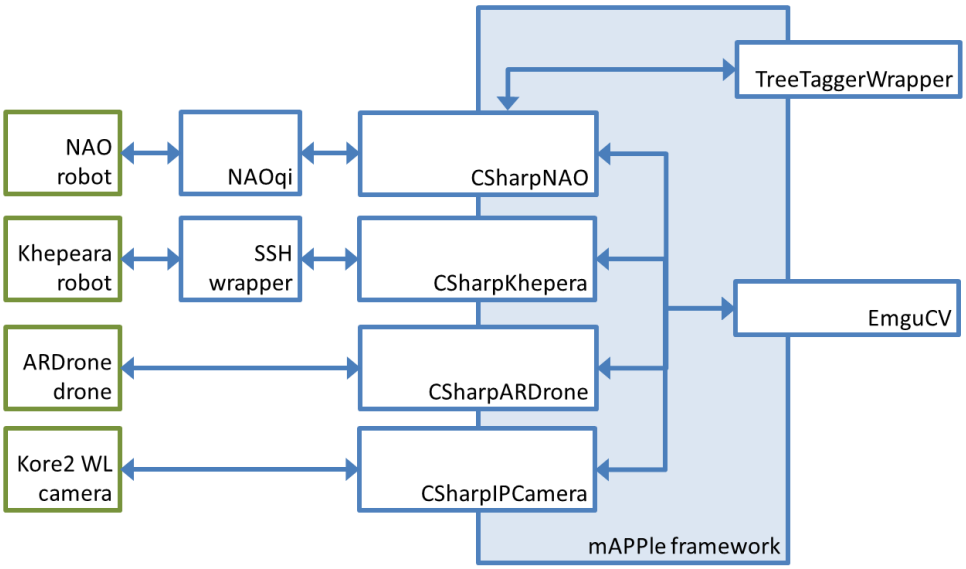

Figure 45: Software architecture used for implementation of the work presented in this thesis.

5.2.2.2. Wrapper Class and Framework for Robot Programming

The entire development has been done in C# on .NET platform. To facilitate implementation of the previously proposed algorithms on the NAO robot and to enable an efficient use of 3rd party data structures and programs, a wrapper class "CSharpNAO" has been developed over the standard NAOqi API. This wrapper class hides some lower-level

[16] The application programming interface of NAO, the documentation can be found online on: http://developer.aldebaran-robotics.com/doc/1-12/naoqi/index.html

aspects of the NAOqi and facilitates tasks like speech recognition, robot walk command or image processing (done using EmguCV[17], a C# wrapper of OpenCV).

Although there exist software frameworks for robot software development such as ROS[18], they could not be used in context of this work due to existing incompatible dependency software. Instead I have developed a framework "mAPPle" based on generic programming techniques and webserver architecture. Wrapper classes for other robots used by our laboratory have been developed too. The framework allows interfacing with various applications including Matlab via HTTP protocol. It has been used for other research projects in our laboratory as well, for example for the one presented in (Wang et al., 2011). The entire software architecture is depicted on Fig. 45.

5.3. Techniques and Algorithms Used

In the following, I am going to remind the reader some of the most prominent techniques and algorithms, I have been using, and which were not mentioned in the previous text along with basic principles of their operation. It is also pertinent to make such clarification on this place, because it is in this chapter where I report on practical implementation of my research work and the nature of the techniques and algorithms used has obviously an influence on the behavior of the system as a whole.

5.3.1. Multilayer Perceptron

A multilayer perceptron (MLP) is a class of artificial neural networks (cf. (Haykin, 2008)) consisting of several interconnected layers of perceptrons (Rosenblatt, 1957). An MLP is able to distinguish data, which are not linearly separable (Barron, 1994) and is inherently capable of approximation of any continuous function from an interval of the real numbers into the interval of [-1,1] with arbitrary precision, as shown in (Auer, Burgsteiner and Maass, 2008).

Each layer of MLP consists of artificial neurons (cf. (McCulloch and Pitts, 1988)), which have inputs, outputs and an activation function much like their biological counterparts. A mathematical model of a single neuron with a sigmoidal activation is described by Eq.

[17] Available on: http://www.emgu.com
[18] Available on: http://www.ros.org/wiki/

(31). y_k stands for the output of the k-th neuron, function $\mathcal{S}$ is the activation function, N is the number of inputs of the neuron, w_i and x_i are the weight of the i-th input and its value respectively.

$$y_k = \mathcal{S}\left(\sum_{i=1}^{N}(w_i x_i)\right) \qquad \textbf{(31)}$$
$$where\ \mathcal{S}(x) = \frac{1}{1+e^{-x}}$$

Neurons in an MLP are organized into layers. Neurons of N-th layer are not interconnected, but they take all their inputs from the (N-1)-th layer and produce their outputs to the (N+1)-th layer. Inputs of the first layer are connected to inputs from the environment; outputs of the last layer are considered as output of the entire neural network and can be further post-processed. In operation mode, inputs are provided to the first layer and for each of the input neurons their y is computed following Eq. (31). This y serves in turn as the input of the second layer and so forth until the output of the last layer is not calculated. This process is called feed-forward as there are no recurrent or backward connections in the network.

In this work I use neuroevolution for adjusting neural network weights, while keeping a fixed network topology. The technique is roughly inspired by the work of (Stanley, D'Ambrosio and Gauci, 2009) and the one of (Stanley and Miikkulainen, 2002). The choice of this approach has been made due to the nature of the problems treated here. Regarding problems discussed in sub-section 3.6.4, it would be very difficult to create a well-balanced training set of input-output pairs. Instead of it the performance of the neural network is measured, which can be done easily, and then this performance is converted into fitness value for the genetic algorithm.

5.3.2. Orientation of Robot in Environment

5.3.2.1. Obstacle Avoidance

When exploring the environment and gathering new knowledge about it, the robot is often required to move around. A basic capacity of orientation in the environment is thus necessary. Capacities needed in this context are obstacle avoidance and determination of

distance to objects. This has to be done using only the NAO's monocular vision as the sonar and infrared sensors of the robot are not completely suitable for this task.

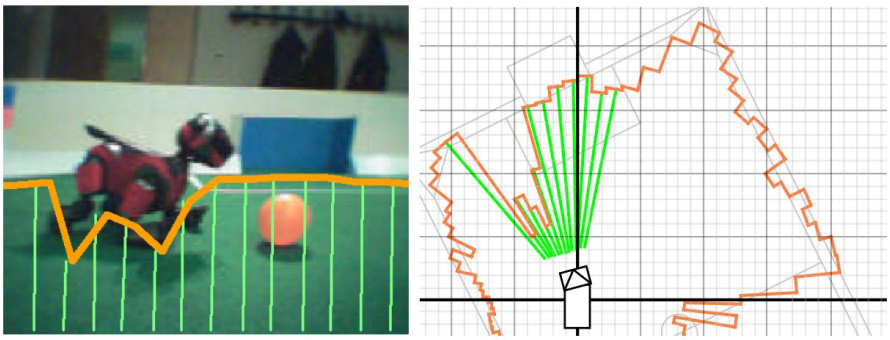

Figure 46: Left: robot's vision with obstacle and ground detection superimposed. Right: map of obstacle around the robot.

To resolve the obstacle avoidance problem, I have adopted a technique based on ground color modeling. Color model of the ground helps the robot to distinguish free-space from obstacles. The approach is loosely inspired by the work presented in (Hoffmann, Jüngel and Lötzsch, 2004). An assumption is made that obstacles repose on ground (i.e. overhanging and floating objects are not taken into account). With this assumption the distance of obstacles can be inferred even from monocular camera data. The approach of distance inference will be detailed in the next sub-section. On Fig. 46 (adopted from (Hoffmann, Jüngel and Lötzsch, 2004)) an example of obstacle detection results using ground color model and monocular vision is shown.

5.3.2.2. Inference of Distance and Size using Monocular Vision

When vision is used to estimate the 3-D model of the scene, techniques like stereo-imaging or structure-from-motion are the usual choice (for an overview on 3-D vision see (Daniilidis and Eklundh, 2008)). With a static monocular camera, on the other hand, it is impossible to recover the distance and the size of objects seen on the scene. To enable this capacity, an additional knowledge is needed.

In (Ramík, Sabourin and Madani, 2010) some aspects of distance estimation from a static monocular camera have been mentioned. I have developed the approach presented there in order to give the robot the capacity to infer distances and sizes of surrounding objects. This is achieved by making the following assumptions:

- **Objects repose on the ground.** However inference of distance and size of objects superposed on another object is also feasible after the height of the bottom object has been estimated.
- **Height and tilt of the camera is known.** This is a reasonable assumption as this information can be gathered from robot's technical documentation and from its internal state sensors.
- **Camera parameters are known.** The angle of camera's field of view and resulting image size in pixels can be obtained from robot's documentation.

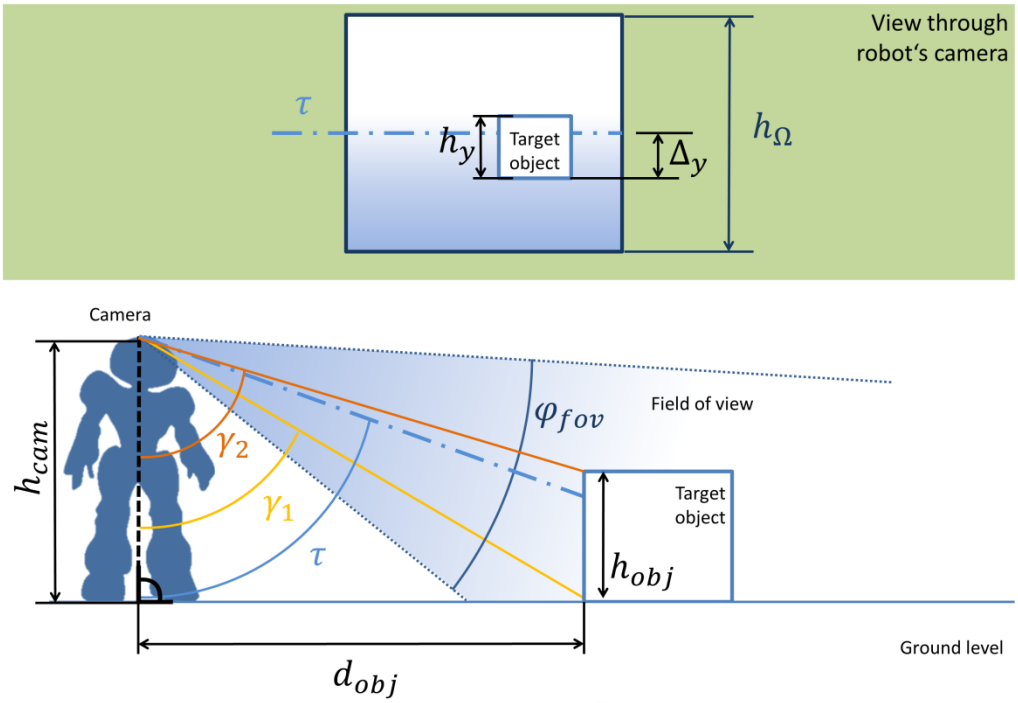

Figure 47: Distance inference from known camera spatial position using monocular camera. Symbols used are explained in text.

Knowing the position of the camera in space, it is possible to estimate the distance and the size of objects on the image in the way depicted on Fig. 47. The height of the camera is denoted by h_{cam}, d_{obj} and h_{obj} (all in meters) stand for the distance of the object and the height of the object respectively. τ (radians) is the tilt of the camera, φ_{fov} (radians) is the field of view. h_y (pixels) is the height of the object on camera image, Δ_y (pixels) is the distance of the bottom of the object from the middle of the image, which is at the same time

projection of the tilt angle of the camera to the image. h_Ω (pixels) denote the height of the camera image Ω.

Calculation of d_{obj} and h_{obj} from given camera height h_{cam}, field of view φ_{fov} and camera tilt τ and from known image height h_Ω and object apparent height h_y and position Δ_y is done as shown on Eq. (32), (33) and (34). The expression $\frac{\varphi_{fov}}{h_\Omega}$ in Eq. (32) is the number of radians per pixel of the image. γ_1 and γ_2 then stand respectively for the angle between the vertical and the line of sight to the lower and upper boundary of the object.

$$\gamma_1 = \tau - \frac{\varphi_{fov}}{h_\Omega}(\Delta_y)$$
$$\gamma_2 = \tau - \frac{\varphi_{fov}}{h_\Omega}(\Delta_y + h_y) \quad \textbf{(32)}$$

$$d_{obj} = \frac{\sin\gamma_1 \, h_{cam}}{\sin\left(\frac{\pi}{2} - \gamma_1\right)} \quad \textbf{(33)}$$

$$h_{obj} = \frac{\sin(\gamma_2 - \gamma_1)}{\sin(\pi - \gamma_1)} \sqrt{{d_{obj}}^2 + {h_{cam}}^2} \quad \textbf{(34)}$$

Calculation of width of the object is analogous to the calculation of the height or it can be derived from h_{obj} and proportions of the object on image in pixels. The calculations presented above are simplified as they do not take into account picture distortion due to projection through lenses (for a discussion on this problem see (Fisher and Konolige, 2008)). However, this does not cause any major detriment of estimation precision as the used camera does not use wide-angle lenses and the image distortion is only minor.

A similar approach to the one described here has been independently presented in (Hoiem, Efros and Hebert, 2008), but with some additional techniques like viewpoint discovery.

5.3.3. Human-robot Verbal Communication

For knowledge transfer between a human and a robot a suitable communication channel is necessary. The choice of verbal communication is obvious as it is arguably the most natural way for humans to share their thoughts and their knowledge. In order to enable

verbal communication capabilities in a robot, there must be at least two requirements fulfilled: a hardware support present in the robot (i.e. microphones and loudspeakers) and a software support (a text-to-speech generator and automated speech recognition). In NAO robot both hardware components are present, as described in sub-section 5.2.1.1. For the software part, it is equipped by text-to-speech module and automated speech recognition developed by Nuance[19]. They are both exposed on NAOqi API.

Another part of the problem of human-robot verbal communication is the understanding of what the human is saying. The product of speech recognition is a string containing words heard by the robot. To obtain the important information from the string, e.g. the subject and object of the phrase or the verb, a syntactic analysis is necessary. To perform syntax analysis, TreeTagger[20] tool is used. TreeTagger is a tool for annotating text with part-of-speech and lemma information. On Table 4 a simple English phrase is shown along with syntactic analysis output of TreeTagger in form of tokens. The "Part-of-speech" row gives tokens explanation[21] and the "Lemma" row shows lemmas output, which is the neutral form of each word in the phrase. This information along with known grammatical rules for creation of English phrases may further serve to determine the nature of the phrase as declarative ("This is a box."), interrogative ("What is the name of this object?") or imperative ("Go to the office!"). It can be also used to extract the subject, the verb and other parts of speech, which are further processed in order to motivate the appropriate response action in the robot.

Algorithms used by the TreeTagger tool are based on the work published in (Schmid, 1994) and (Schmid, 1995).

Phrase	Robots	are	our	friends
Tokens	NNS	VBP	PP$	NNS
Part-of-speech	noun, plural	verb, pres. t.	possessive pron.	noun, plural
Lemma	robot	be	our	friend

Table 4: A sample English phrase and its corresponding syntactic analysis output generated by TreeTagger.

[19] Company website: http://www.nuance.com
[20] Available online at: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
[21] Documentation for English tokens is available online on: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/Penn-Treebank-Tagset.pdf

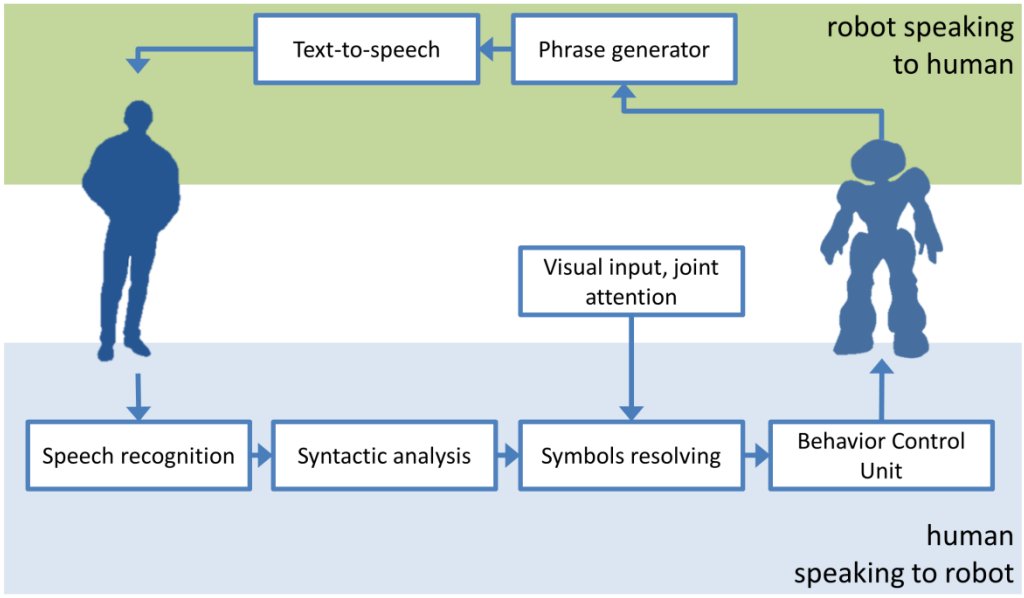

Figure 48: Flow diagram of communication between a robot and a human which is used in this work.

In spite of its usefulness, syntactic analysis only cannot provide sufficient information in a number of cases. Think about a situation, where a human says to the robot "go to the office". Syntactic analysis may result into an order ("go") and a place ("office"). This may be related by the robot to an entity on its map and the robot can plan its path to the office and finally get there. Consider on the other hand this counter-example. A human is pointing towards the office and says "go there". With syntactic analysis done, the meaning of "there" is still difficult to resolve as it is unrelated to any physical entity of the world. This task of symbol resolving requires additional information. In this case it is the visual information about in which direction the speaker was pointing while speaking. This problem is addressed in works covering joint attention and situated speech, c.f. (Roy and Mukherjee, 2005). As it has been mentioned in sub-section 4.6.2, I use a hand detection algorithm and couple it with the object detection algorithm in order to provide this supplementary information. When the robot is unable to resolve a particular phrase, it recourses to this joint attention mechanism as shown of the flow diagram on Fig. 48.

Phrases the robot uses to communicate with the tutor are generated from templates that take keywords from the context of the communication as parameters. The approach is in some of its aspects inspired by ELIZA, the influential artificial agent processing natural language as a chatterbot described in (Weizenbaum, 1966).

5.4. Autonomous Knowledge Acquisition by Robot in Real Environment

Previous sub-sections have discussed all the techniques implicated in application of our work to the real world. Having accomplished this, we can finally proceed to description the environment used and of the performance and the behavior of our system.

5.4.1. Environment Description

Although the NAO humanoid robot is far smaller than an average human, the decision has been taken to employ it in a human-sized environment for the sake of realism. For practical reasons we have chosen to deploy it in an office environment on the site of our university, which bring us several benefits. Most importantly it is the fact that we have control over the conditions. It allows us to use different light sources such as direct and diffused daylight or artificial illumination and observe how robust the behavior of our system is in changing conditions. Besides it is a readily available common kind of environment, in which future operation of companion (humanoid) robots is very likely. It also allows us to use the existing networking infrastructure for communication between the robot and the remote processing platform.

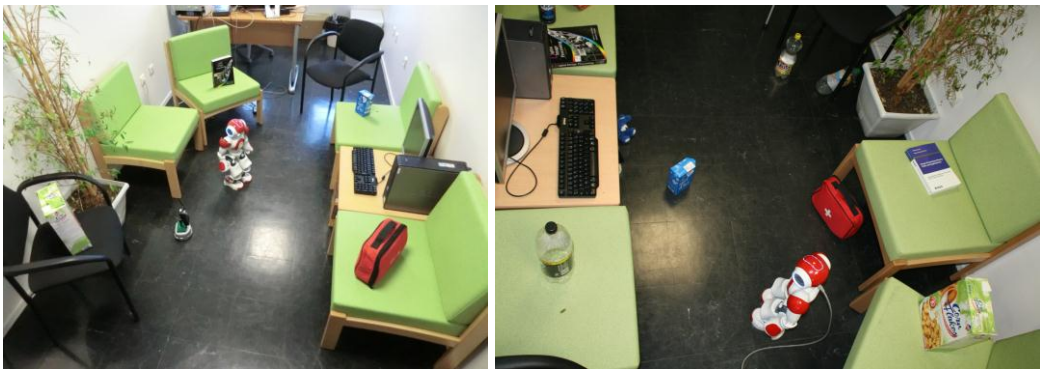

Figure 49: Two photos showing the robot in a real office environment. Some every-day objects like books, bottles and product boxes were added in order to enrich the environment by new visual stimuli.

On Fig. 49 two pictures are shown capturing the office room in which we have deployed NAO, the humanoid robot with our system. Different everyday objects have been

distributed in the environment in different manners. The reason for this was to increase the number of possibilities of the robot to learn and to interact with the environment. Most of the objects that have been used are depicted on Fig. 39 in section 4.6.2. Varying intensity of illumination in different parts of the room as well as slightly glossy floor causing reflections made the tasks involving camera image processing challenging.

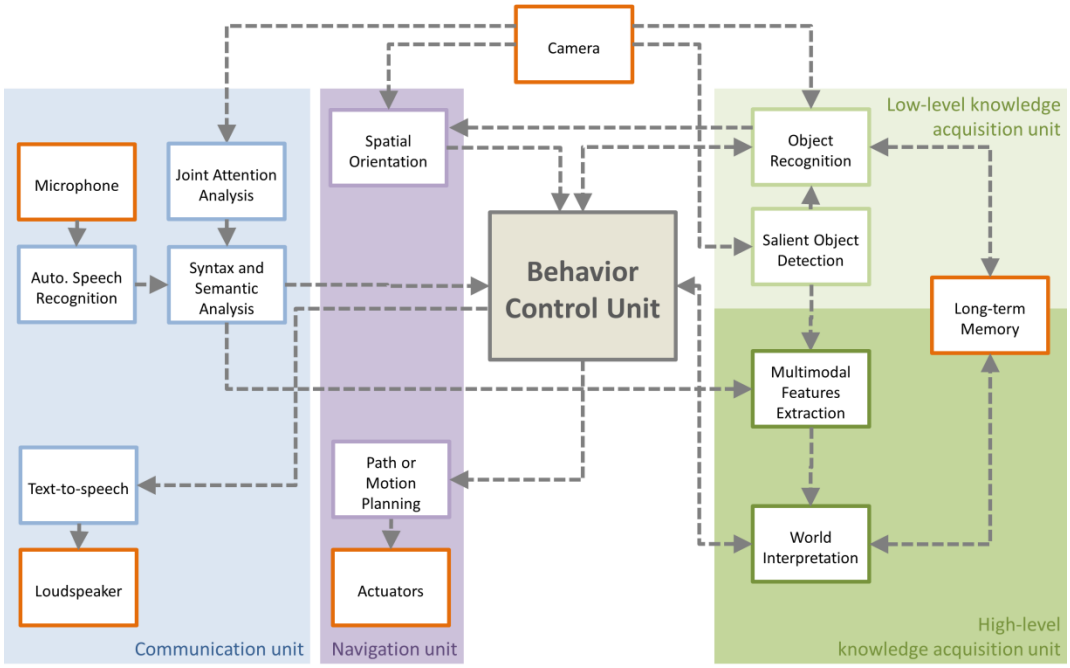

Figure 50: Composition of the entire system in deployment. Each box corresponds to a processing unit described in text.

5.4.2. Scheme of System Operation

This sub-section aims to explain the composition of the entire system and the way in which its different parts interact. On Fig. 50 the composition of the system is depicted. Please note that this is a concretized version of the general diagram provided on Fig. 7 in Chapter 2. It is split into five main units (Communication Unit, Navigation Unit, Low-level Knowledge Acquisition Unit, High-level Knowledge Acquisition Unit, Behavior Control Unit), which are explained in following sub-sections. Most of the main units are composed of smaller processing units, which were mostly subject of description in previous chapters and for this reason their nature is just briefly reminded while making reference to their corresponding part

of text. Fig. 50 contains also parts marked by dark-orange rectangles. Those parts represent hardware units with whose units of the system are interfaces.

5.4.2.1. Communication Unit

The Communication Unit has been outlined in sub-section 5.3.3. It has an output communication channel and an input communication channel. The output channel is composed of a Text-to-speech engine which generates human voice through Loudspeakers. It receives the text to say from the Behavior Control Unit. The input channel takes its input from a Microphone and through an Automated Speech Recognition engine and the Syntax and Semantic Analysis it provides the Behavior Control Unit labeled chain of strings representing the heard speech. It does so with the aid of visual cues from the Joint Attention Analysis. It also provides its outputs to the High-level Knowledge Acquisition Unit, where it is used in process of generation of beliefs about the world (see sub-section 4.3.1).

5.4.2.2. Navigation Unit

The purpose of this unit is to allow the robot to position itself in space with respect to objects around it and to use this knowledge to navigate in the environment. Its sub-unit for Spatial Orientation takes its input from the camera and from the Object Recognition sub-unit from the Low-level Knowledge Acquisition Unit. It operates using the approach described in sub-section 5.3.2 and feeds the Behavior Control Unit with data about navigable space and distances to different objects. The Path and Motion Planning sub-unit supplies Actuators with command signals. Those are used to accomplish what is required by the Behavior Control Unit in order to go towards or to avoid certain objects.

5.4.2.3. Low-level Knowledge Acquisition Unit

This unit ensures gathering of knowledge on lower levels of semantics, such as detection of salient objects (by the Salient Object Detection sub-unit) and their learning and subsequent recognition (in the Object Recognition sub-unit). Those activities are carried out mostly in an "unconscious" manner, i.e. they are run as an automatism in "background" while collecting salient objects and learning them. The learned knowledge is stored in Long-term Memory for further use.

The "lower level of semantics" means that the knowledge extracted on this level bears somewhat less semantic information if compared to outputs of the High-level Knowledge Acquisition Unit. The knowledge here is bound mostly to visual features without any link to linguistic terms describing it. In other words using this unit the robot acquires an animal-like capacity of learning objects and e.g. finding them again in different conditions, which means it has built certain inner representations of them. However it does not give the robot the capacity to communicate this knowledge to humans. This lower level is important for two reasons: it enables lower cognitive level actions like object recognition and it provides valuable features for the High-level Knowledge Acquisition Unit (see the connection to the Multimodal Features Extraction sub-unit).

As this unit is one of the key parts of the entire system, most of the Chapter 3 has been dedicated to description of its principles and its operation.

5.4.2.4. High-level Knowledge Acquisition Unit

The High-level Knowledge Acquisition Unit is the place where outputs from other units (prominently the Low-level Knowledge Acquisition Unit for its features output and the Communication Unit for its linguistic output) are combined together and where high-level semantic representation is derived from them. Unlike the Low-level Knowledge Acquisition Unit, this unit represents conscious and intentional cognitive activity much like a baby which learns from observation and from verbal interaction with adults about the world and develops in this way its own representation of the world.

The complete structure of the unit is presented in Chapter 4, where other important details about functioning of the unit are provided. For the sake of readability this unit is depicted on Fig. 50 only by two sub-units. The first one is the Multimodal Feature Extraction sub-unit. It is responsible for extraction of useful features from the linguistic input from the Syntax and Semantic Analysis sub-unit and of features derived from the vision input. It should be emphasized that on the place of the vision input there could be any sensor or a set of different sensors as discussed in section 4.7.

The second sub-unit is called World Interpretation. It is presented in detail in section 4.3 and following. Its main role is to develop a high-level (semantic) representation of the world based on past sensory experience and on the linguistic input and interaction with human tutor (see section 4.4). The sub-unit is connected to long-term memory, where the gathered sensory experience and the acquired knowledge are stored for further use.

5.4.2.5. Behavior Control Unit

This unit plays the key role of a coordinator among other units of the system. It directs data flows and issues command signals for other units. Also, as its name suggests, it controls the behavior of the robot and alternates it in order to respond properly to external events.

On Fig. 51 the Behavior Control Unit (BCU) operation is depicted in form of a flowchart. Some of the boxes of the flowchart are filled with colors corresponding to one of the four main units. This is to indicate that the particular unit is involved in the operation described by the box.

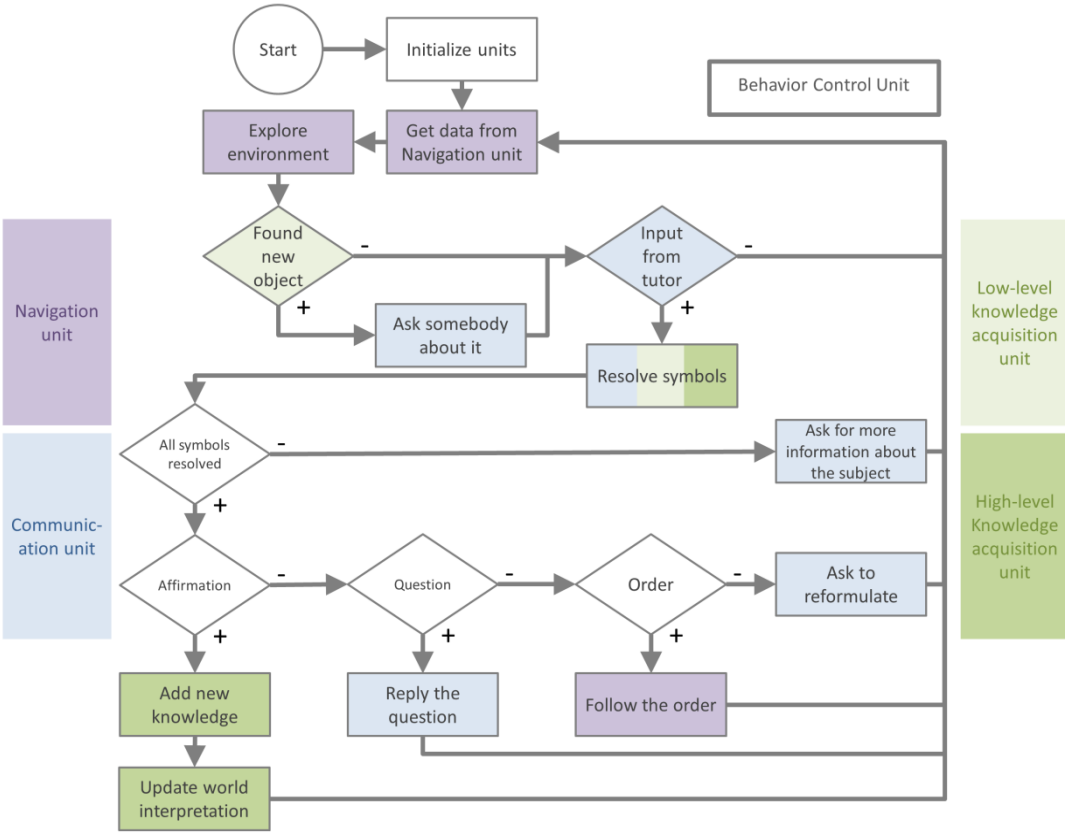

Figure 51: A flowchart representing the **Behavior Control Unit** operation. Color of boxes indicates which of the four main units is participating in each particular step. The processing logic is explained in text.

The BCU operation starts with initialization of all the units and with starting of their proper operation cycle. Then, data about the environment from the Navigation unit is gathered and the robot starts moving through the environment in order to explore it. This

could be considered “idling around” or “free exploration” and satisfying its own curiosity in moments when the robot has got no particular order. If a new object is found, the robot attempts to ask somebody about the object in order to learn more knowledge about it. This behavior cycle breaks when an input from the tutor is received, be it a response to a robot’s enquiry or the tutor’s proper intention to interact with the robot. Remember that the tutor could be any human willing to interact with the robot while possibly sharing his own knowledge.

If a (vocal) input is received, the robot tries to resolve symbols used in the input. Consider the input stating “Can you tell me, what the red thing over there is?” The symbols used in the phrase would be “red”, “thing over there”. The BCU would receive an input from the High-level Knowledge Acquisition Unit saying the concept of “red” is already known, it would receive an input from the Communication Unit resolving the joint attention problem of “over there” and finally the Low-level Knowledge Acquisition Unit would say that the object is not yet in its long-term memory and thus it is unknown to it. At this stage, the symbol cannot be resolved and the BCU will initiate a behavior directed to get information about the unknown symbol (i.e. about “the red thing over there”) possibly by asking somebody about it by saying “Please, tell me what is the name of this object”.

This “curious” behavior would also be initiated if all the symbols were resolved, but some of them without a sufficient certainty. Imagine the robot had seen very few red things so far and the concept of “red” was anchored with high uncertainty. In this case, the robot could engage in a behavior trying to refine its knowledge by asking “Sorry, I am unsure about what is ‘red’, could you show me some more objects that are red”. Both behaviors described here are intended to drive the robot towards enriching and enhancing its knowledge about the world, which has been discussed in section 4.4.

If all symbols are known with a sufficient certainty, the input from the tutor is further processed. If it is an affirmative phrase, i.e. it contains a knowledge explicitly expressed like in the phrase “The book is heavy”, it is extracted and the robot’s inner representation of the world is updated. If the phrase is a question, it is replied using symbols resolved earlier. If it is an order, it is executed. In case the type of phrase could not be decided (possibly a grammatically incorrect or incoherent phrase), the robot asks for a reformulation of the input. The same happens if the phrase is beyond the robot’s understanding, which would be most likely due to limited vocabulary or comprehension of the robot.

When interaction is finished the BCU falls back to the “free exploration” behavior until new interaction is initiated from the side of the robot or of the tutor.

5.4.3. System Behavior and Outcomes

In the following sub-section different aspects of the behavior of the present cognitive system in various stages of its operation will be described.

5.4.3.1. Free Exploration

As mentioned in sub-section 5.4.2.5, the system starts with a free exploration behavior. The robot freely navigates through the environment and tries to learn by observation about it. This behavior is depicted on Fig. 52. On its left part the robot is shown in its environment while walking around. The yellow arrow indicates approximately its motion in the free space between sofas during exploration. As described in sub-section 5.4.2.3, the Low-level Knowledge Acquisition Unit is operating and it extracts visually important objects from the environment.

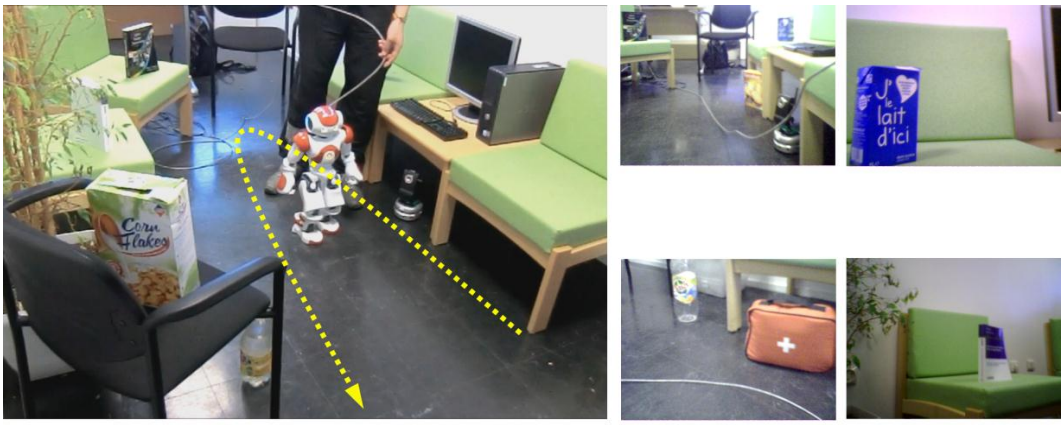

Figure 52: NAO the humanoid robot during free exploration behavior (on the left). The red arrow indicates robots trajectory. Behind the robot the operator is holding robot's cable for security reasons. On the right four sample views from the robot's camera taken during the exploration are presented.

On Fig. 53 some of objects discovered during exploration are shown. Note that varying quality of the samples is caused by different distance from which they were captured resulting in different resolution of the extracted object on the image. Color tone and contrast of images varies due to the fact that they were captured from different viewpoints as the robot

was walking and looking around. As a result, automatic white balance and sensitivity was applied causing slight differences in the appearance of objects.

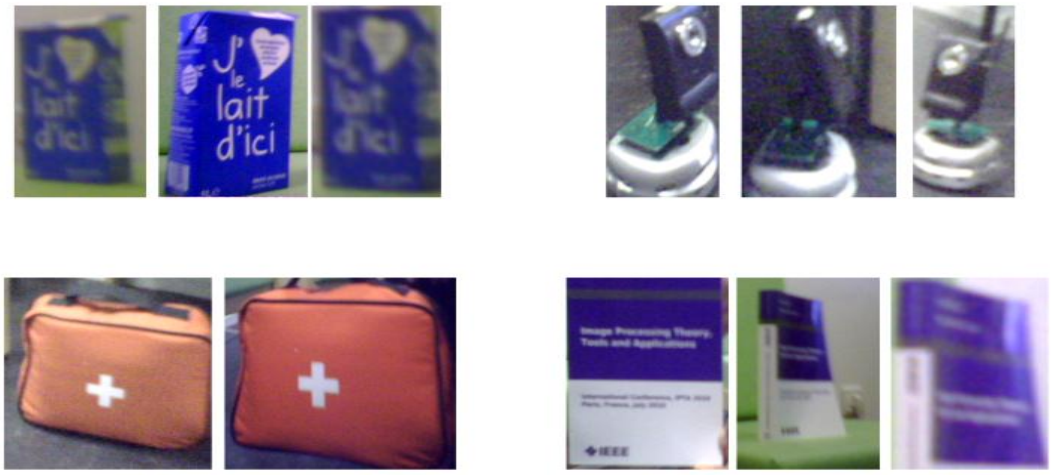

Figure 53: Various objects captured and extracted from different points of view during free exploration of the environment.

5.4.3.2. Interaction with Tutor

At one moment the "passive" learning-by-observation behavior is interrupted and the robot engages in a dialog with the tutor in order to learn more about objects it has seen during its exploration of the environment. The robot shares with the tutor its knowledge about the objects, which is merely limited to whether it had or had not seen them before during the exploration. The tutor in turn shares some of his much complete knowledge about the object with the robot. A sample of such a dialog follows, accompanied by Fig. 54 where two snapshots from two stages of the dialog are shown:

- **Robot**: I have finished the exploration and I have seen some interesting objects here. Could you tell me about them?
- **Human**: [takes a first-aid-kit and points to it, Fig. 54 left] Have you seen this?
- **Robot**: Yes, I have seen this, tell me what is it?
- **Human**: This is a first-aid-kit!
- **Robot**: OK, I will remember that this is a first-aid-kit. [… and learns this new information]
- **Human**: It is red and white.

- **Robot**: I am learning the red and the white. [… takes the features and learns/improves the concepts of "red" and "white" as in sub-section 4.6.2]
- **Tutor**: [draws from his pocket a teddy-bear, which the robot had not seen previously, Fig. 54 right] Have you seen this?
- **Robot**: No, I haven't seen this yet. Tell me, what is it? [… and learns the visual appearance of the new object, the teddy-bear]
- **Tutor**: This is a teddy-bear and it's color is blue.
- **Robot**: OK, I will remember it. [… links the appearance to the name "teddy-bear" and learns its color as in the previous case]
- **Tutor**: [places the first-aid-kit somewhere in the environment] Search for the first-aid-kit!
- **Robot**: I am searching the first-aid-kit. [… resolves that it knows what is the first-aid-kit and starts exploration behavior while actively searching for it using the knowledge it has learned about it before]

This dialog sample shows how the robot and the tutor communicate on different subjects. First the robot learns the name of an object it has previously seen, but did not know the name. Then it anchors two terms for colors, "white" and "red" using features from the visual appearance of the first-aid-kit. Then it encounter a new object, that the tutor shows, a teddy-bear, and the robot learns both its visual appearance and its name and ultimately it anchors a new color term, "blue". Finally thanks to the newly learned knowledge the robot is able to respond properly to the order "Search the first-aid-kit!".

The communication can be, of course, much trickier as in the following example, when the robot has learned to associate the term "book" with several different books. The tutor asks the robot to fetch a book and further specifies that he is interested in the red one. The robot interprets features of all the books it sees around and evaluates which one is fitting the concept of "red" which the robot has learned previously. Based on this interpretation the robot fetches the book it believes to be red.

- **Tutor**: Fetch a book.
- **Robot**: [looks around] I can see two books, which one should I fetch?
- **Tutor**: Fetch the red one.
- **Robot**: [resolves the visual appearance of both books based on its current understanding of the term "red" and goes towards it] OK, I have found the red book and I am fetching it.

Figure 54: Interaction of the tutor and the robot on subject of things the robot had or had not seen during its exploration. Left: learning the name of a previously seen object (a first-aid-kit). Right: learning the visual appearance and the name of a completely new object (a teddy-bear).

Figure 55: Two different illumination conditions applied while the robot was searching for the same object. Top: external view on the environment. Bottom: robot's proper view through the camera. Left: direct artificial illumination (causing reflections). Right: Natural ambient light illuminating the room through the window. A cloudy day.

5.4.3.3. Different Illumination Conditions

Robustness to changing conditions is one of the key parameters of any system deployed in real environment. Here it is described how the system performs in changing light conditions. Fig. 55 shows two different settings of illumination. One is the natural ambient light (right); the other is artificial illumination from ceiling (left). It is clearly visible that the later one is causing reflections on floor and glossy objects like the one the robot is searching for (a blue box of milk).

Apart of this, the robot's camera is obviously having difficulties with white balance for this particular color temperature. This is alternating the color balance, rendering the entire image yellowish. Both effects combined make this illumination particularly challenging. On the other hand the left part of Fig. 55 shows conditions of natural ambient illumination of the environment. Due to cloudy weather the amount of light coming to the room was insufficient and the robot's camera was producing images with significantly more noise and with a bluish tint.

Although the system was tested several times in such greatly varying conditions of illumination, no visible impact on the behavior of the system itself has been observed and the robot was fully able to pursue its normal cycle of operation.

5.5. Conclusion

The purpose of this chapter was to "close the loop" of design of the system presented here and to show practical aspects of its deployment in real conditions and its behavior. The reader was first familiarized with NAO, the humanoid robotic platform used in the deployment. Then, in the next section, foundations of some of the most important techniques used were reminded to the reader. The aim of this was to further enlighten the design choices that have been made and the influence that the nature of those methods have to the behavior of the system as a whole. Then the scheme of the system with regard to each of its unit has been explained. Some concrete examples from the deployment have finally been given in the last section.

In this chapter all the constitutive parts of the system were put together and operated as one structure. This has been successful and all units of the system were effectively collaborating in acquiring high-level semantic knowledge from unstructured data from the

environment. The system in its complete state has shown that it is capable of fulfilling tasks for which it was designed.

A video showing different aspects of what has been described in this chapter is available online[22]. It captures the system's operation in a real office environment and it completes videos shown in sub-section 3.7.2[23] and in sub-section 4.6.2[24].

[22] http://youtu.be/Y_JM0KfJb8Q
[23] http://www.youtube.com/watch?v=xxz3wm3L1pE
[24] http://youtu.be/W5FD6zXihOo

General Conclusion

Conclusion

Autonomous machine cognition is an important, yet extremely difficult problem. Its importance for conception of a true autonomy for future intelligent systems, including humanoid robots, has been discussed in the Introduction to this thesis. The difficulty of this problem comes from multiple sources. First, it is due to the complexity and incessant change present in the real world, which is thus difficult to capture or model. Secondly, the nature of machines and their function is fundamentally different from the nature and the function of human cognitive apparatus. It is thus a major challenge to reunite these two entities so different in their essence in a way which would allow a seamless knowledge transfer from one to the other.

In order to address the problem of machine cognition, different methods have been developed, approaching it from different points of view or addressing its particular aspects. A representative sample of existing methods, that in some way cope with this problem, have been provided in Chapter 1.

In this thesis, in order to contribute to development of autonomous machine cognition, the way I have taken reposes on the assumption that it is the curiosity which motivates a cognitive system to acquire new knowledge. Further two distinct kinds of curiosity are identified in conformity to human cognitive system. On this I build a two level cognitive architecture. I identify its lower level with the perceptual saliency mechanism (cf. Chapter 3), while the higher level performs knowledge acquisition from observation and interaction with the environment (cf. Chapter 4). The interaction also includes interaction with humans. This interaction is crucial for forming of human-like concepts in the mentioned bio-inspired cognitive system; however it is the robot who is the principal actor of the learning.

Constitutive parts of this system, i.e. both the low-level and the high-level cognitive units, have been separately tested. This has been done both in a virtual environment via simulations and through experiments in real world. Results of these partial validations have shown that each of the parts separately can be efficiently applied to the problem areas they were designed for. In particular, the algorithm for salient object detection, which has been

developed in order to realize the low level (perceptual) curiosity mechanism, is shown to be superior to other comparable contemporary approaches in terms of processing speed and correctness of its output. Equally, on the higher cognitive level of the system, which realizes the mechanism of epistemic curiosity, the presented approach has shown the capacity of acquisition of high lever knowledge from low-level features. Notably it has shown its capacity to acquire the desired knowledge on higher rate, i.e. needing fewer exposures to the subject, than other comparable approach recently published.

Finally the entire described system has been implemented on an embodied agent, a humanoid robot, and tested in a series of experiments in challenging real world conditions of an indoor (human) environment. This has been done not only with the aim to validate the ensemble of all sub-units of the system, but also to verify the concept of "human-like learning" introduced earlier in this work. Apart of the primary problem of autonomous machine cognition, many secondary problems had to be addressed in implementation of this system in order to make it capable of operation in physical world. These problems include e.g. navigation in space and realization of communication with humans and they are developed in Chapter 5. Results of the experiments conducted with this entire system implemented to a humanoid robot show an augmentation of the robot's overall autonomy. They notably show that, through embodiment to a humanoid robot, the present cognitive system is capable of a completely autonomous knowledge acquisition from observation and of interaction with humans in real time.

In first chapters of this thesis we have seen, that the lack of autonomy in current machine systems (e.g. robots) is due in particular to the lack of an appropriate autonomous machine cognitive system, through which such a robot could gradually acquire knowledge about the world and apprehend it in an autonomous manner. The work accomplished through this thesis is a step towards such autonomous machine cognition applicable in mobile robotics in conditions of real world. Its notable contribution lies both in theoretical development and in practical realization of an autonomous knowledge acquisition system and its embodiment to a mobile robot.

Knowledge acquisition is motivated here by the curiosity, a biologically inspired concept. Although usage of this concept in cognitive sciences and cognitive robotics is not new, in this thesis it is viewed from a new point of view, different from its usual use in existing works. In the present thesis, curiosity acts in its different forms (perceptual vs. epistemic) on both cognitive levels of the system. Through the two-tier architecture with a lower level (close to sensory data) and a higher level (close to semantic information) the cognitive system achieves the capacity of linking the perceptual experience with its high level

(semantic) representation. This link is acquired through a process called "human-like learning" – a human-inspired learning approach. It enables a humanoid robot endowed with this system to acquire knowledge in a manner that is inspired from young children learning and which makes the robot an active, enterprising partner in human-robot knowledge sharing, rather than a passive receptor. This represents a step from human-supervised learning towards a fully autonomous and self-motivated knowledge acquisition in robots and ultimately, on a broader view, to the full autonomy of robots.

Perspectives

The cognitive system that has been presented in this thesis has met the objectives given in the General Introduction and has proven itself to be practicable in real environment. This being said, as a matter of fact, a number of perspectives remain still open and some aspects of the work could not be reasonably addressed in the limited timeframe of the thesis. These are subject for future development of the work accomplished here.

On the lower cognitive level, the problems of perceptual saliency that have been addressed cover specifically visual saliency. However, image can be interpreted simply as a general 2D signal. It is thus reasonable to expect that algorithms developed in this thesis will be in general applicable to other signal data, e.g. the sound. As an interesting future direction this could enrich the lower level cognition by adding other (non-visual) sources of information in a multimodal fusion. This general saliency detecting algorithm can be expected to be of interest for another of research interests of our laboratory, which is intelligent fault detection. The fault, being an anomalous state, would be detected as it stands out form the normal pattern of the monitored system's operation.

The functioning of the higher cognitive level is motivated by epistemic curiosity. Consequently the knowledge acquisition on this level relies on a) evolution of beliefs about the world, which are coherent with the currently known state of the world and b) identification of "knowledge gaps", i.e. missing knowledge which should be filled. Validation results in Chapter 4 show the robot endowed with the present system is able of learning of a single category semantic at one time. Meanwhile, in the same chapter an extension of this is proposed in order to learn multiple categories at the same time. This extension inherently enables distinguishing of which of the used words are related to which category. As for the perspective of the future work, the validation of this extension will take

place. A natural evolution following this will tend towards a seamless integration of the cognitive experience coming from multiple sensors.

In humans, the general intelligence reposes on human cognitive system as a basis (c.f. (Geary, 2005)). It is thus justified to say, that a machine system endowed with human-like intelligence will necessarily include an autonomous cognitive system as its basis, over which this artificial intelligence will operate as a higher functional level. With respect to this, the long-term perspectives regarding the autonomous cognitive system presented in this thesis will focus on its integration to a system of larger scale realizing artificial-intelligence in machines, such as mobile robots. There, it will play the role of an underlying system for machine cognition and knowledge acquisition. This knowledge will be subsequently available as the basis for tasks proper for (machine) intelligence such as reasoning, decision making and an overall autonomy.

Appendices

Appendix A. Image Segmentation in siRGB

Input: $\Omega(x)$ //the color image in spherical coordinates
$\delta,\ a,\ b,\ c$ // threshold and three distance parameters values
Output: bi-directional matrix L containing pixel labels l and an array of region chromaticity representations $\Psi_l \forall l$

$x_0 = \Omega(0,0): L(x_0) = newlabel,\ \Psi_{l(x_0)} = \Psi(x_0)$
for each x **do**
  if $L_4(x) = \{l\}$ //there is only one label in $N_4(x)$
    evaluate_neighbor(x)
  else
    $D \leftarrow \{d_h(L(y), L(z)) = d_{y,z} \,|y, z \in N_4(x)\ \&\ L(y) \neq L(z)\}$
    for all $d_{y,z} \in D$ **s.t.** $d_{y,z} < \delta$ //region merging
      $L(y) = merge(y, z)$ //merge both regions into $L(y)$
      $\Psi_{L(y)} = avg\ \Psi(\mathrm{w})\ \ \forall w \in \Omega$ where $L(w) = L(y)$ //update reg. chroma.
    if $L_4(x) = \{l\}$ //there is only one label in $N_4(x)$
      evaluate_neighbor(x)
    else
      $d \leftarrow \min\{d_h(x, y) | y \in N_4(x)\}$
      if $d < \delta$ //assign to region with lower distance
        $L(x) = L(y)\ \boldsymbol{s.t.}\ d_h(x, y) = d$
        $\Psi_{L(y)} = avg\ \Psi(\mathrm{w})\ \ \forall w \in \Omega$ where $L(w) = L(y)$ //update reg.chroma.
      else //current pixel cannot be assigned to any region
        create_new_label(x)
end for

function: create_new_label(x)
  $L(x)$ = newlabel //create a new region label
  $\Psi_{L(x)} = \Psi(x)$ //init region label chromaticity

function: evaluate_neighbor(x)
  $d \leftarrow \min\{d_h(x, y) | y \in N_4(x)\}$
  if $d < \delta$ //neighbor colors are similar
    $L(x) = l;\ \Psi_{L(y)} = avg\ \Psi(\mathrm{w})\ \ \forall w \in \Omega$ where $L(w) = L(y)$ //upd.reg.chroma.
  else create_new_label(x)

Algorithm 3: Hybrid segmentation method in siRGB color space.

Appendix B. Detection of Ambiguities in a Set of Utterances

Under some conditions, a set of utterances on observations (see sub-section 4.3.1) may be ambiguous, which would lead to multiple plausible interpretations of the world by the cognitive system although indeed only one would be conforming to the reality. In order to detect such ambiguities and to initiate the search for completing the missing information (i.e. to disambiguate the set), the following algorithm is used.

$\mathcal{U}$ is the set of all sets of utterances given on all observations. $\mathcal{D}$ is the set of disambiguated utterances and $\mathcal{A}$ is the set of ambiguous utterances. We go through all combinations of any two members of $\mathcal{A}$ and test if a) their intersection produces one and only one utterance (let us call it u_i), or b) their symmetric difference produce one and only one utterance u_i. In case a) it means that "the only similarity in observations belonging to sets of utterances a and b is called u_i". In case b) it means that "the only difference in observations belonging to sets of utterances a and b is called u_i". Thus we have disambiguated the utterance u_i and we include it into $\mathcal{D}$. When no further additions into $\mathcal{D}$ are possible, the set of ambiguous utterances $\mathcal{A}$ is given as the relative component of all utterances U and disambiguated utterances $\mathcal{D}$.

Initialization:
$\mathcal{U} = \{U_1, \dots, U_m\}$ // set to all sets of utterances
$\mathcal{D} = \emptyset$

while ($\mathcal{D}$ changes){
  //find any combination of sets of utterance from $\mathcal{U}$
  foreach (a from $\mathcal{A}$)
    foreach (b from $(\mathcal{A} - a)$){
      $a = (a - \mathcal{D})$ //remove already disambiguated utterances
      $b = (b - \mathcal{D})$
      if($|a \cap b| = 1$) $a \cap b \rightarrow \mathcal{D}$ //intersection
      else if($|a \Delta b| = 1$) $a \cup b \rightarrow \mathcal{D}$ //symmetric difference
    }
}

$$\mathcal{A} = (U - \mathcal{D})$$

if($|\mathcal{A}| = 0$) no ambiguity
else ambiguity detected

Algorithm 4: Detection of ambiguities in a set of utterances.

Appendix C. Coherent belief generation procedure

```
Initialization:
int = new array of length |U| //interpretations of all utterances
rem_f = I //set remaining features to all observed features
do{
  rem_u = U //set remaining utterances to all heard utterances
  do{
    //pick up a random utterance and construct its interpretation
    r = random utterance from rem_u

    if(int[r] has no features assigned yet){
      //get all features observed in presence of the r-th utterance
      possible_f = features from rem_f observed in presence of r
      //get the feature the most dissimilar to any feature already
      //assigned to any of the interpretations
      f = feature from possible_f where |f − f_p| = max∀f_p ∈ int
      add f to int[r]
      remove f from rem_f
    }
    else{
      //get all features observed in presence of the r-th utterance
      possible_f = features from rem_f observed in presence of r
      //get the feature the most similar to features from int[r]
      f = feature from from possible_f where |f − f_p| = min∀f_p ∈ int[r]
      //get average distance of f normalized to 0..1 scale
      dst = average distance |f − f_p| ∀f_p ∈ int[r] from features in int[r]
      //probability of the feature being pertinent is proportional to
      //its distance to other features in int[r]
      if(random number from 0 to 1 > dst){
        //consider the feature as pertinent in context of r
        add f to int[r]
      }else{//do nothing … consider the feature as impertinent}
      remove f from rem_f}
    remove r from rem_u
  }while(|rem_u| > 0) //repeat while there are unassigned utterances
}while(|rem_f| > 0) //repeat while there are still unassigned features
```

Algorithm 5: Coherent belief generation procedure.

Publications

As the First Author

D. M. Ramík, C. Sabourin, K. Madani, “On Human Inspired Semantic SLAM’s Feasibility”, Artificial Neural Networks and Intelligent Information Processing, INSTICC Press, ISBN: 978-989-8425-03-4, pp. 99-108, 2010

D. M. Ramík, C. Sabourin, K. Madani, "Toward Human Inspired Semantic SLAM", Proceedings of International Conference on Informatics in Control Automation and Robotics (IFAC/IEEE ICINCO 2010), Vol. 2, Funchal, Madeira, Portugal, June 15 - 18, ISBN: 978-989-8425-01-0, pp. 360-363, 2010.

D. M. Ramík, C. Sabourin, K. Madani, "Hybrid Artificial Vision System Combining Salient Object Extraction and Machine-Learning Skills", Proceedings of International Conference on Pattern Recognition and Information Processing (PRIP 2011), Minsk, Belarus, May 18 - 20, ISBN: 978-985-488-722-7, pp. 366-369, 2011.

D. M. Ramík, C. Sabourin, K. Madani, “A Cognitive Approach for Robots’ Vision Using Unsupervised Learning and Visual Saliency”, Advances in Computational Intelligence,, LNCS series, Vol. 6691, Springer Verlag, ISBN: 978-3-642-21500-1, International Work-conference on Artificial Neural Networks (IWANN 2011), pp. 65-72, 2011

D. M. Ramík, C. Sabourin, K. Madani, "A Real-time Robot Vision Approach Combining Visual Saliency and Unsupervised Learning", Field Robotics, Ed. P. Bidaud, M. O. Tokhi, C. Grand, G. S. Virk, 14th International Conference on Climbing and Walking Robots and the Support Technologies for Mobile Machines (CLAWAR2011), Paris, France, , September 6 - 8, ISBN: 978-981-4374-27-9, pp. 241-248, 2011.

D. M. Ramík, C. Sabourin, K. Madani, "Hybrid Salient Object Extraction Approach with Automatic Estimation of Visual Attention Scale", Proc. of Seventh International Conference on Signal Image Technology & Internet-Based Systems (IEEE – SITIS 2011), Dijon, France, November 28 - December 1, pp. 438-445, 2011.

Co-authored

V. Amarger, D. M. Ramík, R. Moreno, L. Rossi, K Madani, M. Graña, " Wildland Fires' Outlines Extraction: a Spherical Coordinates Framed RGB Color Space Dichromatic Reflection Model Based Image Segmentation Approach", Proceedings of International Conference on Pattern Recognition and Information Processing (PRIP 2011), Minsk, Belarus, May 18 - 20, ISBN:978-985-488-722-7, pp. 451-454, 2011.

T. Wang, D. M. Ramík, C. Sabourin, K. Madani, "Machine Learning for Heterogeneous Multi-Robots Systems in Logistic Application Frame", Field Robotics, Ed. P. Bidaud, M. O. Tokhi, C. Grand, G. S. Virk, 14th International Conference on Climbing and Walking Robots and the Support Technologies for Mobile Machines (CLAWAR2011), Paris, France, , September 6 - 8, ISBN: 978-981-4374-27-9, pp. 207-214, 2011. (obtained the "Highly Recommended Award" / "Industrial Robot Journal's" Award).

R. Moreno, M. Graña, D. M. Ramík, K Madani, " Image Segmentation by Spherical Coordinates", Proceedings of International Conference on Pattern Recognition and Information Processing (PRIP 2011), Minsk, Belarus, May 18 - 20, ISBN: 978-985-488-722-7, pp. 112-115, 2011.

K Madani, C. Sabourin, D. M. Ramík, " Hybrid Artificial Vision System combining Salient Object Extraction and Machine-Learning Skills", Proceedings of International Conference on Pattern Recognition and Information Processing (PRIP 2011), Minsk, Belarus, May 18 - 20, ISBN: 978-985-488-722-7, pp. 112-115, 2011.

Journal Articles

Ting Wang, Dominik M. Ramík, Christophe Sabourin, Kurosh Madani, "Intelligent systems for industrial robotics: application in logistic field", Industrial Robot: An International Journal, Vol. 39, Issue 3, pp.251 – 259, 2012

K. Madani, D. M. Ramík, C. Sabourin, "Multilevel Cognitive Machine-Learning-Based Concept for Artificial Awareness: Application to Humanoid Robot Awareness Using Visual

Saliency," Applied Computational Intelligence and Soft Computing, Vol. 2012, Article ID 354785, 11 pages, 2012 – in press

R. Moreno, D. M. Ramík, M. Graña, K. Madani, "Image Segmentation on the Spherical Coordinate Representation of the RGB Color Space", IET Image Procession, ISSN: 1751-9659, 2012. Accepted.

Bibliography

[**Achanta, *et al.* 2008**] Achanta, R., Estrada, F., Wils, P. and Süsstrunk, S. (2008) 'Salient Region Detection and Segmentation', Computer Vision Systems, 6th International Conference, Santorini, 66-75.

[**Achanta, *et al.* 2009**] Achanta, R., Hemami, S., Estrada, F. and Süsstrunk, S. (2009) 'Frequency-tuned Salient Region Detection', IEEE Computer Society Conference on Computer Vision, Miami, 1597-1604.

[**Akbari, *et al.* 2011**] Akbari, R. and Ziarati, K. (2011) 'A multilevel evolutionary algorithm for optimizing numerical functions', *International Journal of Industrial Engineering Computations* , vol. 2, may, pp. 419-430, Available: 1923-2926.

[**Angelopoulou, *et al.* 2008**] Angelopoulou, A., Psarrou, A., García Rodríguez, J. and Gupta, G. (2008) 'Active-GNG: model acquisition and tracking in cluttered backgrounds', Proceedings of the 1st ACM workshop on Vision networks for behavior analysis, New York, 17-22.

[**Araki, *et al.* 2011**] Araki, T., Nakamura, T., Nagai, T., Funakoshi, K., Nakano, M. and Iwahashi, N. (2011) 'Autonomous acquisition of multimodal information for online object concept formation by robots', IEEE/RSJ International Conference on Intelligent Robots and Systems, San Francisco, 1540-1547.

[**Auer, *et al.* 2008**] Auer, P., Burgsteiner, H. and Maass, W. (2008) 'A learning rule for very simple universal approximators consisting of a single layer of perceptrons', *Neural Networks*, vol. 21, june, pp. 786-795, Available: 0893-6080.

[**Avidan, *et al.* 2007**] Avidan, S. and Shamir, A. (2007) 'Seam carving for content-aware image resizing', *ACM Transactions on Graphics*, vol. 26, no. 3, July, Available: 0730-0301.

[**Barron 1994**] Barron, A.R. (1994) 'Approximation and Estimation Bounds for Artificial Neural Networks', *Machine Learning*, vol. 14, pp. 115-133.

[**Bay, *et al.* 2008**] Bay, H., Ess, A., Tuytelaars, T. and Van Gool, L. (2008) 'Speeded-Up Robust Features (SURF)', *Computer Vision and Image Understanding*, vol. 110, june, pp. 346-359, Available: 1077-3142.

[**Berlyne 1954**] Berlyne, D.E. (1954) 'A theory of human curiosity', *British Journal of Psychology*, vol. 45, no. 3, August, pp. 180-191.

[**Blum, *et al.* 1997**] Blum, A.L. and Langley, P. (1997) 'Selection of relevant features and examples in machine learning', *Artificial Intelligence*, 1-2 dec, pp. 245-271.

[**Bonwell, *et al.* 1991**] Bonwell, C.C. and Eison, J.A. (1991) *Active Learning: Creating Excitement in the Classroom (J-B ASHE Higher Education Report Series (AEHE))*, School of Education and Human Development, George Washington University.

[**Borba, *et al.* 2006**] Borba, G.B., Gamba, H.R., Marques, O. and Mayron, L.M. (2006) 'An unsupervised method for clustering images based on their salient regions of interest', Proceedings of the 14th ACM International Conference on Multimedia, Santa Barbara, 145-148.

[**Bowerman 1983**] Bowerman, M. (1983) 'How Do Children Avoid Constructing an Overly General Grammar in the Absence of Feedback about What is Not a Sentence?', *Papers and Reports on Child Language Development*.

[**Brand, *et al.* 2002**] Brand, R.J., Baldwin, D.A. and Ashburn, L.A. (2002) 'Evidence for 'motionese': modifications in mothers infant-directed action', Developmental Science, 72-83.

[**Bülthoff, *et al.* 2008**] Bülthoff, H.H., Wallraven, C. and Giese, M.A. (2008) 'Perceptual Robotics', Springer Handbook of Robotics, 1481-1498.

[**Civera, *et al.* 2012**] Civera, J., Gálvez-López, D., Riazuelo, L., Tardós, J.D. and Montiel, J.M.M. (2012) 'Towards semantic SLAM using a monocular camera', International Conference on Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ, 1277-1284.

[**Coradeschi, *et al.* 2003**] Coradeschi, S. and Saffiotti, A. (2003) 'An introduction to the anchoring problem', *Robotics and Autonomous Systems*, vol. 43, pp. 85-96.

[**Daniilidis, *et al.* 2008**] Daniilidis, K. and Eklundh, J.-O. (2008) '3-D Vision and Recognition', in Siciliano, B. and Khatib, O. (ed.) *Springer Handbook of Robotics*, Springer.

[**Darwin 1859**] Darwin, C. (1859) *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*, John Murray.

[**Ekvall, *et al.* 2006**] Ekvall, S., Jensfelt, P. and Kragic, D. (2006) 'Integrating Active Mobile Robot Object Recognition and SLAM in Natural Environments', in *International Conference on Intelligent Robots and Systems, 2006 IEEE/RSJ*, Beijing: IEEE.

[**Ekvall, *et al.* 2005**] Ekvall, S. and Kragic, D. (2005) 'Receptive field cooccurrence histograms for object detection', *IEEE/RSJ International Conference on Intelligent Robots and Systems, 2005*, aug, pp. 84-89.

[**Fisher, *et al.* 2008**] Fisher, R.B. and Konolige, K. (2008) 'Range Sensors', in Siciliano, B. and Khatib, O. (ed.) *Springer Handbook of Robotics*, Springer.

[**Freund, *et al.* 1995**] Freund, Y. and Schapire, R.E. (1995) 'A decision-theoretic generalization of on-line learning and an application to boosting', Proceedings of the Second European Conference on Computational Learning Theory, London, 23-37.

[**Frintrop, *et al.* 2009**] Frintrop, S. and Kessel, M. (2009) 'Most salient region tracking', IEEE International Conference on Robotics and Automation, Kobe, 1869-1874.

[**Fu, *et al.* 1981**] Fu, K.S. and Mui, J.K. (1981) 'A survey on image segmentation', *Pattern Recognition*, vol. 13, pp. 3-16, Available: 0031-3203.

[**Galindo, *et al.* 2005**] Galindo, C., Saffiotti, A., Coradeschi, S., Buschka, P. and Fernandez-Madrigal, J.A. (2005) 'Multi-Hierarchical Semantic Maps for Mobile Robotics', Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Edmonton, 2278-2283.

[**García-Rodríguez, *et al.* 2011**] García-Rodríguez, J. and García-Chamizo, J.M. (2011) 'Surveillance and human-computer interaction applications of self-growing models', *Journal of Applied Soft Computing*, vol. 11, no. 7, pp. 4413-4431.

[**Geary 2005**] Geary, D.C. (2005) *The Origin of Mind: Evolution of Brain, Cognition, and General Intelligence*, American Psychological Association.

[**Goodrich, *et al.* 2007**] Goodrich, M.A. and Schultz, A.C. (2007) 'Human-robot interaction: a survey', *Foundations and trends in human computer interaction*, vol. 1, jan, pp. 203-275, Available: 1551-3955.

[**Gouaillier, *et al.* 2010**] Gouaillier, D., Collette, C. and Kilner, C. (2010) 'Omni-directional closed-loop walk for NAO', Proceedings of the 10th IEEE-RAS International Conference on Humanoid Robots, Nashville, 448-454.

[**Greeff, *et al.* 2009**] Greeff, J.d., Delaunay, F. and Belpaeme, T. (2009) 'Human-Robot Interaction in Concept Acquisition: a computational model', Proceedings of the 2009 IEEE 8th International Conference on Development and Learning, Washington, 1-6.

[**Griffith, *et al.* 2009**] Griffith, S., Sinapov, J., Miller, M. and Stoytchev, A. (2009) 'Toward interactive learning of object categories by a robot: A case study with container and non-container objects', Proceedings of the 2009 IEEE 8th International Conference on Development and Learning, Shanghai, 1-6.

[**Harel, *et al.* 2007**] Harel, J., Koch, C. and Perona, P. (2007) 'Graph-based visual saliency', Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, 545-552.

[**Haykin 2008**] Haykin, S. (2008) *Neural Networks and Learning Machines (3rd Edition)*, 3rd edition, Prentice Hall.

[**Hertzberg, *et al.* 2010**] Hertzberg, J., Albrecht, S., Günther, M., Lingemann, K., Sprickerhof, J. and Thomas, W. (2010) 'From Semantic Mapping to Anchored Knowledge Bases', Proceedings of the 10th Biannual Meeting of German Socciety

of Cognitive Science, Symposium Adaptivity of Hybrid Cognitive Systems, Potsdam, 33-37.

[**Hoffmann, *et al.* 2004**] Hoffmann, J., Jüngel, M. and Lötzsch, M. (2004) 'A Vision Based System for Goal-Directed Obstacle Avoidance Used in the RC'03 Obstacle Avoidance Challenge', Lecture Notes in Artificial Intelligence, In 8th International Workshop on RoboCup 2004 (Robot World Cup Soccer Games and Conferences), 418-425.

[**Hoiem, *et al.* 2008**] Hoiem, D., Efros, A.A. and Hebert, M. (2008) 'Putting Objects in Perspective', *International Journal of Computer Vision*, vol. 80, November, pp. 3-15, Available: 0920-5691.

[**Holland 1992**] Holland, J.H. (1992) *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*, Cambridge: MIT Press.

[**Hou, *et al.* 2007**] Hou, X. and Zhang, L. (2007) 'Saliency Detection: A Spectral Residual Approach', IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, 1-8.

[**Chang, *et al.* 1999**] Chang, P. and Krumm, J. (1999) 'Object Recognition with Color Cooccurrence Histograms', Proceedings of Conference on Computer Vision and Pattern Recognition, Fort Collins, 2498-2504.

[**Cheng, *et al.* 2011**] Cheng, M.-M., Zhang, G.-X., Mitra, N.J., Huang, X. and Hu, S.-M. (2011) 'Global contrast based salient region detection', The 24th IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, 409-416.

[**Chen, *et al.* 2003**] Chen, L.-Q., Xie, X., Fan, X., Ma, W.-Y., Zhang, H.-J. and Zhou, H.-Q. (2003) 'A Visual Attention Model for Adapting Images on Small Displays', *Multimedia Systems*, vol. 9, oct, pp. 353-364.

[**Itti, *et al.* 1998**] 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*, vol. 20, pp. 1254-1259, Available: 0162-8828.

[**Kajita, *et al.* 2009**] Kajita, S., Hirukawa, H., Harada, K. and Yokoi, K. (2009) *Introduction à la commande des robots humanoïdes: De la modélisation à la génération du mouvement*, Springer-Verlag France.

[**Kajita, *et al.* 1995**] Kajita, S. and Tanie, K. (1995) 'Experimental Study of Biped Dynamic Walking in the Linear Inverted Pendulum Mode', Proceedings of IEEE International Conference on Robotics and Automation, Nagoya, 2885-2891.

[**Kang, *et al.* 2009**] Kang, M.J.J., Hsu, M., Krajbich, I.M., Loewenstein, G., McClure, S.M., Wang, J.T.T. and Camerer, C.F. (2009) 'The wick in the candle of learning: epistemic curiosity activates reward circuitry and enhances memory', *Psychological science*, vol. 20, no. 8, August, pp. 963-973, Available: 1467-9280.

[**Kay, *et al.* 2003**] Kay, P., Berlin, B., Maffi, L. and Merrifield, W.R. (2003) *World Color Survey*, Center for the Study of Language and Information.

[**Kay, *et al.* 1991**] Kay, P., Berlin, B. and Merrifield, W. (1991) 'Biocultural Implications of Systems of Color Naming', *Journal of Linguistic Anthropology*, vol. 1, pp. 12-25.

[**Kernbach, *et al.* 2009**] Kernbach, S., Hamann, H., Stradner, J., Thenius, R., Schmickl, T., Crailsheim, K., Rossum, A.C.v., Sebag, M., Bredeche, N., Yao, Y., Baele, G., Peer, Y.V.d., Timmis, J., Mohktar, M., Tyrrell, A., Eiben, A.E. and McKibbin, S. (2009) 'On Adaptive Self-Organization in Artificial Robot Organisms', Proceedings of the 2009 Computation World: Future Computing, Service Computation, Cognitive, Adaptive, Content, Patterns, 33-43.

[**Klingspor, *et al.* 1997**] Klingspor, V., Demiris, J. and Kaiser, M. (1997) 'Human-Robot-Communication and Machine Learning', *Applied Artificial Intelligence*, pp. 719-746.

[**Koechlin, *et al.* 1999**] Koechlin, E., Basso, G., Pietrini, P., Panzer, S. and Grafman, J. (1999) 'The role of the anterior prefrontal cortex in human cognition', *Nature*, vol. 399 (6732), May, pp. 148-151.

[**Kohonen 1982**] Kohonen, T. (1982) 'Self-organized formation of topologically correct feature maps', *Biological Cybernetics*, vol. 43, no. 1, pp. 59-69, Available: 0340-1200.

[**Kuhn, *et al.* 1995**] Kuhn, D., Garcia-Mila, M., Zohar, A. and Andersen, C. (1995) 'Strategies of knowledge acquisition', *Society for Research in Child Development Monographs*, vol. 60 (4), no. 245.

[**Langley, *et al.* 2009**] Langley, P., Laird, J.E. and Rogers, S. (2009) 'Cognitive architectures: Research issues and challenges', *Cognitive Systems Research*, vol. 10, no. 2, June, pp. 141-160.

[**Levesque, *et al.* 2010**] Levesque, H.J. and Lakemeyer, G. (2010) 'Cognitive robotics', in Lakemeyer, G., Levesque, H.J. and Pirri, F. *Handbook of Knowledge Representation*, Dagstuhl: Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik.

[**Liang, *et al.* 2012**] Liang, Z., Chi, Z., Fu, H. and Feng, D. (2012) 'Salient object detection using content-sensitive hypergraph representation and partitioning', *Pattern Recognition*, vol. 45, November, pp. 3886-3901, Available: 0031-3203.

[**Li, *et al.* 2011**] Li, C., Qu, Z., Lu, K. and Gao, Y. (2011) 'Salient Region Detection by Tuning Spectrum Based on Lateral Inhibition', Proceedings of the 2011 First International Conference on Instrumentation, Measurement, Computer, Communication and Control, Washington, 289-292.

[**Litman 2008**] Litman, J.A. (2008) 'Interest and deprivation factors of epistemic curiosity', *Personality and Individual Differences*, vol. 44, no. 7, pp. 1585-1595.

[**Liu, *et al.* 2011**] Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X. and Shum, H.-Y. (2011) 'Learning to Detect a Salient Object', Computer Vision and Pattern Recognition, Los Alamitos, 353-367.

[**Lowe 1999**] Lowe, D.G. (1999) 'Object Recognition from Local Scale-Invariant Features', Proceedings of the International Conference on Computer Vision, Washington, 1150-1157.

[**Lütkebohle, *et al.* 2009**] Lütkebohle, I., Peltason, J., Schillingmann, L., Wrede, B., Wachsmuth, S., Elbrechter, C. and Haschke, R. (2009) 'The curious robot-structuring interactive robot learning', Proceedings of the 2009 IEEE international conference on Robotics and Automation, Kobe, 2154-2160.

[**Macedo, *et al.* 2012**] Macedo, L. and Cardoso, A. (2012) 'The exploration of unknown environments populated with entities by a surprise-curiosity-based agent', *Cognitive Systems Research*, vol. 19-20, September - October, pp. 62-87, Available: 1389-0417.

[**Madani, *et al.* 2011**] Madani, K. and Sabourin, C. (2011) 'Multi-level cognitive machine-learning based concept for human-like "artificial" walking: Application to autonomous stroll of humanoid robots', Neurocomputing, Amsterdam, 1213-1228.

[**Maier, *et al.* 2011**] Maier, W. and Steinbach, E.G. (2011) 'Surprise-driven acquisition of visual object representations for cognitive mobile robots', IEEE International Conference on Robotics and Automation, Shanghai, 1621-1626.

[**McCulloch, *et al.* 1988**] McCulloch, W.S. and Pitts, W. (1988) 'A logical calculus of the ideas immanent in nervous activity', in Anderson, J.A. and Rosenfeld, E. (ed.) *Neurocomputing: foundations of research*, Cambridge: MIT Press.

[**McNeil, *et al.* 1984**] McNeil, M.C., Polloway, E.A. and Smith, J.D. (1984) 'Feral and isolated children: Historical review and analysis', *Education & Training of the Mentally Retarded*, vol. 19, no. 1, February, pp. 70-79.

[**Meger, *et al.* 2008**] Meger, D., Forssén, P.-E., Lai, K., Helmer, S., McCann, S., Southey, T., Baumann, M., Little, J.J. and Lowe, D.G. (2008) 'Curious George: An attentive semantic robot', *Robotics and Autonomous Systems*, vol. 56, no. 6, June, pp. 503-511.

[**Mikolajczyk, *et al.* 2005**] Mikolajczyk, K. and Schmid, C. (2005) 'A performance evaluation of local descriptors', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, oct, pp. 1615-1630, Available: 0162-8828.

[**Mileva, *et al.* 2007**] Mileva, Y., Bruhn, A. and Weickert, J. (2007) 'Illumination-Robust Variational Optical Flow with Photometric Invariants', Pattern Recognition, 29th DAGM Symposium, Heidelberg, 152-162.

[**Minsky, *et al.* 1988**] Minsky, M. and Papert, S. (1988) *Perceptrons: An Introduction to Computational Geometry*, Expanded ed. edition, MIT Press.

[**Montabone, *et al.* 2010**] Montabone, S. and Soto, A. (2010) 'Human detection using a mobile platform and novel features derived from a visual saliency mechanism', *Image and Vision Computing*, vol. 28, no. 3, March, pp. 391-402, Available: 0262-8856.

[**Moreno, *et al.* 2011**] Moreno, R., Graña, M. and d'Anjou, A. (2011) 'Illumination source chromaticity estimation based on spherical coordinates in RGB', *Electronics Letters*, vol. 47, pp. 28-30.

[**Moreno, *et al.* 2010**] Moreno, R., Graña, M. and Zulueta, E. (2010) 'RGB colour gradient following colour constancy preservation', *Electronics Letters*, vol. 46, jun., pp. 908-910, Available: 0013-5194.

[**Navalpakkam, *et al.* 2006**] Navalpakkam, V. and Itti, L. (2006) 'An Integrated Model of Top-Down and Bottom-Up Attention for Optimizing Detection Speed', IEEE Conference on Computer Vision and Pattern Recognition, New York, 2049-2056.

[**Nene, *et al.* 1996**] Nene, S., Nayar, S. and Murase, H. (1996) *Columbia Object Image Library (COIL-100)*, Columbia University.

[**Nüchter, *et al.* 2008**] Nüchter, A. and Hertzberg, J. (2008) 'Towards semantic maps for mobile robots', Robotics and Autonomous Systems, Amsterdam, 915-926.

[**Ogino, *et al.* 2006**] Ogino, M., Kikuchi, M. and Asada, M. (2006) 'How can humanoid acquire lexicon? active approach by attention and learning biases based on curiosity', IEEE/RSJ International Conference on Intelligent Robots and Systems, Beijing, 3480-3485.

[**Otsu 1979**] Otsu, N. (1979) 'A threshold selection method from gray-level histograms', *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 1, January, pp. 62-66.

[**Oudeyer, *et al.* 2007**] Oudeyer, P.-Y., Kaplan, F. and Hafner, V.V. (2007) 'Intrinsic Motivation Systems for Autonomous Mental Development', *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 2, April, pp. 265-286.

[**Pearsall 2010**] Pearsall, J. (2010) *Oxford Dictionary of English*, 3rd edition, Oxford: Oxford University Press.

[**Persson, *et al.* 2007**] Persson, M., Duckett, T., Valgren, C. and Lilenthal, A. (2007) 'Probabilistic Semantic Mapping with a Virtual Sensor for Building/Nature detection', Proceedings of the 7th IEEE International Symposium on Computational Intelligence in Robotics and Automation, Jacksonville, 236-242.

[**Porikli 2005**] Porikli, F. (2005) 'Integral histogram: a fast way to extract histograms in Cartesian spaces', IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 829-836.

[**Ramík, *et al.* 2010**] Ramík, D.M., Sabourin, C. and Madani, K. (2010) 'On Human Inspired Semantic SLAM's Feasibility', Proceedings of the 6th International Workshop on Artificial Neural Networks and Intelligent Information Processing, ANNIIP 2010, In conjunction with ICINCO 2010, Funchal, 99-108.

[**Ramík, *et al.* 2011**] Ramík, D.M., Sabourin, C. and Madani, K. (2011) 'A Cognitive Approach for Robots Vision Using Unsupervised Learning and Visual Saliency', Proceedings of the 11th international conference on Artificial neural networks conference on Advances in computational intelligence, Torremolinos, 81-88.

[**Regier 1995**] Regier, T. (1995) 'A Model of the Human Capacity for Categorizing Spatial Relations', *Cognitive Linguistics*, vol. 6, pp. 63-88.

[**Rosenblatt 1957**] Rosenblatt, F. (1957) *The Perceptron, a perceiving and recognizing automaton*, Cornell Aeronautical Laboratory.

[**Ross 1999**] Ross, B.J. (1999) 'A Lamarckian Evolution Strategy for Genetic Algorithms', in Chambers, L.D. (ed.) *Practical Handbook of Genetic Algorithms: Complex Coding Systems*, 3rd edition, Florida: CRC Press.

[**Roy, *et al.* 2005**] Roy, D. and Mukherjee, N. (2005) 'Towards situated speech understanding: visual context priming of language models', *Computer Speech and Language*, vol. 19, pp. 227-248.

[**Rutishauser, *et al.* 2004**] Rutishauser, U., Walther, D., Koch, C. and Perona, P. (2004) 'Is bottom-up attention useful for object recognition?', IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Los Alamitos, 37-44.

[**Saffran, *et al.* 1997**] Saffran, J.R., Newport, E.L., Aslin, R.N., Tunick, R.A. and Barrueco, S. (1997) 'Incidental Language Learning: Listening (and Learning) Out of The Corner of Your Ear', *Psychological Science*, vol. 8, pp. 101-105, Available: 0956-7976.

[**Saunders, *et al.* 2010**] Saunders, J., Nehaniv, C.L. and Lyon, C. (2010) 'Robot learning of lexical semantics from sensorimotor interaction and the unrestricted speech of human tutors', Second International Symposium on New Frontiers in Human-Robot Interaction, Leicester, 95-102.

[**Settles 2010**] Settles, B. (2010) 'Active Learning Literature Survey', in *Computer Sciences Technical Report 1648*, University of Wisconsin - Madison.

[**Shafer 1985**] Shafer, S.A. (1985) 'Using color to separate reflection components', *Color Research and Application*, pp. 210-218.

[**Schauerte, *et al.* 2010**] Schauerte, B. and Fink, G.A. (2010) 'Focusing computational visual attention in multi-modal human-robot interaction', International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction, Beijing, 6:1-6:8.

[**Schindler, *et al.* 1964**] Schindler, M. and Goethe, J.W. (1964) *Goethes theory of colour applied by Maria Schindler*, New Knowledge Books, East Grinstead, Eng.

[**Schmid 1994**] Schmid, H. (1994) 'Probabilistic Part-of-Speech Tagging Using Decision Trees', Proceedings of the International Conference on New Methods in Language Processing, Manchester, 44-49.

[**Schmid 1995**] Schmid, H. (1995) 'Improvements In Part-of-Speech Tagging With an Application To German', Proceedings of the European Chapter of the Association for Computational Lingusitic, SIGDAT-Workshop, Dublin, 47-50.

[**Schmidhuber 1991**] Schmidhuber, J. (1991) 'Curious model-building control systems', *Proceedings of International Joint Conference on Neural Networks*, vol. 2, pp. 1458-1463.

[**Skocaj, *et al.* 2011**] Skocaj, D., Kristan, M., Vrecko, A., Mahnic, M., Janicek, M., Kruijff, G.-J.M., Hanheide, M., Hawes, N., Keller, T., Zillich, M. and Zhou, K. (2011) 'A system for interactive learning in dialogue with a tutor', IEEE/RSJ International Conference on Intelligent Robots and Systems IROS 2011, San Francisco, 3387-3394.

[**Soderland 1999**] Soderland, S. (1999) 'Learning Information Extraction Rules for Semi-Structured and Free Text', *Machine Learning - Special issue on natural language learning*, vol. 34, feb, pp. 233-272, Available: 0885-6125.

[**Stanley, *et al.* 2009**] Stanley, K.O., D'Ambrosio, D.B. and Gauci, J. (2009) 'A hypercube-based encoding for evolving large-scale neural networks', *Artificial Life*, April, pp. 185-212, Available: 1064-5462.

[**Stanley, *et al.* 2002**] Stanley, K.O. and Miikkulainen, R. (2002) 'Evolving neural networks through augmenting topologies', *Evolutionary Computation*, vol. 15, june, pp. 99-127, Available: 1063-6560.

[**Sun, *et al.* 2012**] Sun, X., Yao, H. and Ji, R. (2012) 'Visual attention modeling based on short-term environmental adaption', *Journal of Visual Communication and Image Representation*, in press, Available: 1047-3203.

[**Tomasello 1999**] Tomasello, M. (1999) *The Cultural Origins of Human Cognition*, Harvard University Press.

[**Tomasello 2003**] Tomasello, M. (2003) *Constructing a Language: A Usage-Based Theory of Language Acquisition*, Harvard University Press.

[**Ugur, *et al.* 2007**] Ugur, E., Dogar, M.R., Cakmak, M. and Sahin, E. (2007) 'Curiosity-driven learning of traversability affordance on a mobile robot', IEEE 6th International Conference on Development and Learning (ICDL' 07), 13-18.

[**van de Weijer, *et al.* 2004**] van de Weijer, J. and Gevers, T. (2004) 'Robust optical flow from photometric invariants', Proceedings of the 29th DAGM conference on Pattern recognition, Heidelberg, 1835-1838.

[**Vasudevan, *et al.* 2007**] Vasudevan, S., Gachter, S., Nguyen, V. and Siegwart, R. (2007) 'Cognitive maps for mobile robots-an object based approach', *Robotics and Autonomous Systems*, 5 May, pp. 359-371.

[**Vernon 2011**] Vernon, D. (2011) 'Reconciling Autonomy with Utility: A Roadmap and Architecture for Cognitive Development', Biologically Inspired Cognitive Architectures 2011 - Proceedings of the Second Annual Meeting of the BICA Society, Arlington, 412-418.

[**Vernon, *et al.* 2007**] Vernon, D., Metta, G. and Sandini, G. (2007) 'A Survey of Artificial Cognitive Systems: Implications for the Autonomous Development of Mental Capabilities in Computational Agents', *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 2, April, pp. 151-180, Available: 1089-778X.

[**Viola, *et al.* 2004**] Viola, P.A. and Jones, M.J. (2004) 'Robust Real-Time Face Detection', *International Journal of Computer Vision*, vol. 57, no. 2, May, pp. 137-154, Available: 0920-5691.

[**Wang, *et al.* 2011**] Wang, T., Ramík, D.M., Sabourin, C. and Madani, K. (2011) 'Machine learning for heterogeneous multi-robots systems in logistic application frame', in Bidaud, P., Tokhi, M.O., Grand, C. and Virk, G.S. (ed.) *Proceedings of the 14th International Conference on Climbing and Walking Robots and the Support Technologies for Mobile Machines*, Paris: University Pierre et Marie Curie.

[**Wang, *et al.* 2012**] Wang, P., Wang, J., Zeng, G., Feng, J. and Zha, H. (2012) 'Salient object detection for searched web images via global saliency', In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Rhode Island, 3194-3201.

[**Waxman, *et al.* 2009**] Waxman, S.R. and Gelman, S.A. (2009) 'Early word-learning entails reference, not merely associations', *Trends in cognitive science*, vol. 13, jun, pp. 258-263, Available: 13646613.

[**Weizenbaum 1966**] Weizenbaum, J. (1966) 'ELIZA - a computer program for the study of natural language communication between man and machine', *Communications of the ACM*, vol. 9, January, pp. 36-45, Available: 0001-0782.

[**Wellens, *et al.* 2008**] Wellens, P., Loetzsch, M. and Steels, L. (2008) 'Flexible word meaning in embodied agents', *Connection Science*, vol. 20, June, pp. 173-191.

[**Wieber 2006**] Wieber, P.-B. (2006) 'Trajectory Free Linear Model Predictive Control for Stable Walking in the Presence of Strong Perturbations', Proceedings of the 6th IEEE-RAS International Conference on Humanoid Robots, Genova, 137-142.

[**Wolfe, *et al.* 2004**] Wolfe, J.M. and Horowitz, T.S. (2004) 'What attributes guide the deployment of visual attention and how do they do it?', Nature Reviews Neuroscience, 495-501.

[**Wyer, *et al.* 1986**] Wyer, R.S. and Srull, T.K. (1986) 'Human cognition in its social context', *Psychological Review*, vol. 93, no. 3, pp. 322-359.

[**Yu 2005**] Yu, C. (2005) 'The emergence of links between lexical acquisition and object categorization: a computational study', *Connection Science*, vol. 17, December, pp. 381-397.

[**Zukow-Goldring, *et al.* 2007**] Zukow-Goldring, P. and Arbib, M.A. (2007) 'Affordances, effectivities, and assisted imitation: Caregivers and the directing of attention', *Neurocomputing*, vol. 70, no. 13-15, August, pp. 2181-2193, Available: 0925-2312.

Abstract

The work accomplished in this thesis concerns development of an autonomous machine cognition system. The proposed solution reposes on the assumption that it is the curiosity which motivates a cognitive system to acquire new knowledge. Further, two distinct kinds of curiosity are identified in conformity to human cognitive system. On this I build a two level cognitive architecture. I identify its lower level with the perceptual saliency mechanism, while the higher level performs knowledge acquisition from observation and interaction with the environment. This thesis brings the following contribution: A) Investigation of the state of the art in autonomous knowledge acquisition. B) Realization of a lower cognitive level in the ensemble of the mentioned system, which is realizing the perceptual curiosity mechanism through a novel fast, real-world robust algorithm for salient object detection and learning. C) Realization of a higher cognitive level through a general framework for knowledge acquisition from observation and interaction with the environment including humans. Based on the epistemic curiosity, the high-level cognitive system enables a machine (e.g. a robot) to be itself the actor of its learning. An important consequence of this system is the possibility to confer high level multimodal cognitive capabilities to robots to increase their autonomy in real-world environment (human environment). D) Realization of the strategy proposed in the context of autonomous robotics. The studies and experimental validations done had confirmed notably that our approach allows increasing the autonomy of robots in real-world environment.

Resumé

Le travail effectué lors de cette thèse concerne le développement d'un système cognitif artificiel autonome. La solution proposée repose sur l'hypothèse que la curiosité est une source de motivation d'un système cognitif dans le processus d'acquisition des nouvelles connaissances. En outre, deux types distincts de curiosité ont été identifiés conformément au système cognitif humain. Sur ce principe, une architecture cognitive à deux niveaux a été proposée. Le bas-niveau repose sur le principe de la saillance perceptive, tandis que le haut-niveau réalise l'acquisition des connaissances par l'observation et l'interaction avec l'environnement. Cette thèse apporte les contributions suivantes : A) Un état de l'art sur l'acquisition autonome de connaissance. B) L'étude, la conception et la réalisation d'un système cognitif bas-niveau basé sur le principe de la curiosité perceptive. L'approche proposée repose sur la saillance visuelle réalisée grâce au développement d'un algorithme rapide et robuste permettant la détection et l'apprentissage d'objets saillants. C) La conception d'un système cognitif haut-niveau, basé sur une approche générique, permettant l'acquisition de connaissance à partir de l'observation et de l'interaction avec son environnent (y compris avec les êtres humains). Basé sur la curiosité épistémique, le système cognitif haut-niveau développé permet à une machine (par exemple un robot) de devenir l'acteur de son propre apprentissage. Une conséquence substantielle d'un tel système est la possibilité de conférer des capacités cognitives haut-niveau multimodales à des robots pour accroitre leur autonomie dans un environnement réel (environnement humain). D) La mise en œuvre de la stratégie proposée dans le cadre de la robotique autonome. Les études et les validations expérimentales réalisées ont notamment confirmé que notre approche permet d'accroitre l'autonomie des robots dans un environnement réel.