4

Autonomous Knowledge Discovery Based on Artificial Curiosity-Driven Learning by Interaction

K. Madani, D. M. Ramik and C. Sabourin

Images, Signals & Intelligent Systems Lab. (LISSI / EA 3956) University PARIS-EST Créteil (UPEC) –Sénart-FB Institute of Technolog, Lieusaint, France
Corresponding author: K. Madani <madani@u-pec.fr>

Abstract

In this work, we investigate the development of a real-time intelligent system allowing a robot to discover its surrounding world and to learn autonomously new knowledge about it by semantically interacting with humans. The learning is performed by observation and by interaction with a human. We describe the system in a general manner, and then we apply it to autonomous learning of objects and their colors. We provide experimental results both using simulated environments and implementing the approach on a humanoid robot in a real-world environment including every-day objects. We show that our approach allows a humanoid robot to learn without negative input and from a small number of samples.

Keywords: Visual saliency, autonomous learning, intelligent system, artificial curiosity, automated interpretation, semantic robot-human interaction.

4.1 Introduction

In recent years, there has been a substantial progress in robotic systems able to robustly recognize objects in the real world using a large database of pre-collected knowledge (see [1] for a notable example). There has been, however, comparatively less advance in the autonomous acquisition of such

Advances in Intelligent Robotics and Collaborative Automation, 73–94.

knowledge: if contemporary robots are often fully automatic, they are rarely fully autonomous in their knowledge acquisition. If the aforementioned substantial progress is commonsensical regarding the last-decades' significant developments in methodological and algorithmic approaches relating visual information processing, pattern recognition and artificial intelligence, the languishing in the machine's autonomous knowledge acquisition is also obvious regarding the complexity of the additional necessary skills to achieve such "not algorithmic" but "cognitive" task.

Emergence of cognitive phenomena in machines have been and remain an active part of research efforts since the rise of Artificial Intelligence (AI) in the middle of the last century, but the fact that human-like machine-cognition is still beyond the reach of contemporary science only proves how difficult the problem is. In fact, nowadays there are many systems, such as sensors, computers or robotic bodies, that outperform human capacities; nonetheless, none of the existing robots can be called truly intelligent. In other words, robots sharing everyday life with humans are still far away. Somewhat, it is due to the fact that we are still far from fully understanding the human cognitive system. Partly, it is so because it is not easy to emulate human cognitive skills and complex mechanisms relating those skills. Nevertheless, the concepts of bio-inspired or human-like machine-cognition remain the foremost sources of inspiration for achieving intelligent systems (intelligent machines, intelligent robots, etc...). This is the way we have taken (e.g. through inspiration from biological and human knowledge acquisition mechanisms) to design the investigated human-like machine-cognition based system able to acquire high-level semantic knowledge from visual information (e.g. from observation). It is important to emphasize that the term "cognitive system" means here that characteristics of such a system tend to those of human cognitive systems. This means that a cognitive system, which is supposed to be able to comprehend the surrounding world on its own, but whose comprehension would be non-human, would afterward be incompetent of communicating about it with its human counterparts. In fact, human-inspired knowledge representation and human-like communication (namely semantic) about the acquired knowledge become key points expected from such a system. To achieve the aforementioned capabilities, such a cognitive system should thus be able to develop its own high-level representation of facts from low-level visual information (such as images). Accordingly to the expected autonomy, the processing from the "sensory level" (namely visual level) to the "semantic level" should be performed solely by the robot, without human supervision. However, this does not mean excluding interaction with humans, which is, on the contrary, vital for

any cognitive system, be it human or machine. Thus, the investigated system has to share its perceptual high-level knowledge of the world with the human by interacting with him. The human on his turn shares with the cognitive robot his knowledge about the world using natural speech (utterances) completing observations made by the robot.

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 real world. The other one comes from humans' richness concerning various ways they 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. This is shown in [2] regarding usage of color terms by different people around the world, or in [3] regarding cultural differences in description of spatial relations. A robot supposed to defeat 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, within the environment in which it evolves and by interaction with the people it encounters in that environment (see [4] for a survey on human-robot interaction and learning and [5] 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 [6]. This problem has been addressed to different degrees in various works. For example, in [7] a computational model of word-meaning, acquisition by interaction is presented. In [8], the authors present a computational model for the acquisition of a lexicon describing simple objects. In [9], a humanoid robot is taught to associate simple shapes to human lexicon. In [10], a humanoid robot is taught through a dialog with untrained users with the aim to learn different objects and to grasp them properly. More advanced works on robots' autonomous learning and dialog are given by [11, 12].

In this chapter, we describe an intelligent system, allowing robots (as for example humanoid robots) to learn and to interpret the world in which they evolve using appropriate terms from human language, while not making use of a priori knowledge. This is done by word-meaning anchoring based on learning by observation and by interaction with its human counterpart. Our model is closely inspired by human infants' early-ages learning behaviour (e.g.

see [13, 14]). The goal of this system is to allow a humanoid robot to anchor the heard terms to its sensory-motor 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.

The rest of this chapter is structured as follows. Section 4.2 describes the architecture of the proposed approach. In this section, we detail our approach by outlining its architecture and principles, we explain how beliefs about the world are generated and evaluated by the robot and we describe the role of human-robot interaction in the learning process. Validation of the presented system on colors learning and interpretation, using simulation facilities, is reported in Section 4.3. Section 4.4 focuses on the implementation and validation of the proposed approach on a real robot in a real-world environment. Finally, Section 4.5 discusses the achieved results and outlines future work.

4.2 Proposed System and Role of Curiosity

Curiosity is a key skill for human cognition and thus it appears as an appealing concept in conceiving artificial systems that gather knowledge, especially when they are supposed to gather knowledge autonomously. Accordingly to Berlyne's Theory of human curiosity [15], two kinds of curiosities stimulate the human's cognitive mechanism. The first one is the 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 relates reflexive or repetitive perceptual experiences. The other one is called "epistemic curiosity", which is more related to the "desire for knowledge that motivates individuals to learn new ideas, to eliminate information-gaps, and to solve intellectual problems.

According to [16] and [17], the general concept of the presented architecture could include one unconscious visual level which may contain a number of Unconscious Cognitive Functions (UCF) and one conscious visual level which may contain a number of Conscious Cognitive Functions (CCF). Conformably with the aforementioned concept of two kinds of curiosity,

an example of knowledge extraction from visual perception, involving both kinds of curiosity, is shown on Figure 4.1. The perceptual curiosity motivates or stimulates what we call the low-level knowledge acquisition and concerns "reflexive" (unconscious) processing level. It seeks "surprising" or "attention-drawing" information in given visual data. The task of the perceptual curiosity is realized by perceptual saliency detection mechanisms. This gives the basis for operation of high-level knowledge acquisition, which is stimulated by epistemic curiosity. Being previously defined as the process that motivates to "learn new ideas, eliminate information-gaps, and solve intellectual problems": as those relating the interpretation of visual information or the belief's generation concerning the observed objects.

The problem of learning brings an inherent problem of distinguishing the pertinent sensory information and the impertinent one. The solution to this task is not obvious even if we achieve joint attention in the robot. This is illustrated on Figure 4.2. If a human points to one object (e.g. an apple) among many others, and describes it as "red", the robot still has to distinguish which of the detected colors and shades of the object the human is referring to.

To achieve correct anchoring in spite of such an uncertainty, we adopt the following strategy. The robot extracts features from important objects found in the scene along with the words the tutor used to describe the objects. Then, the robot generates its beliefs about which word could describe which feature. The beliefs are used as organisms in a genetic algorithm. Here, the appropriate

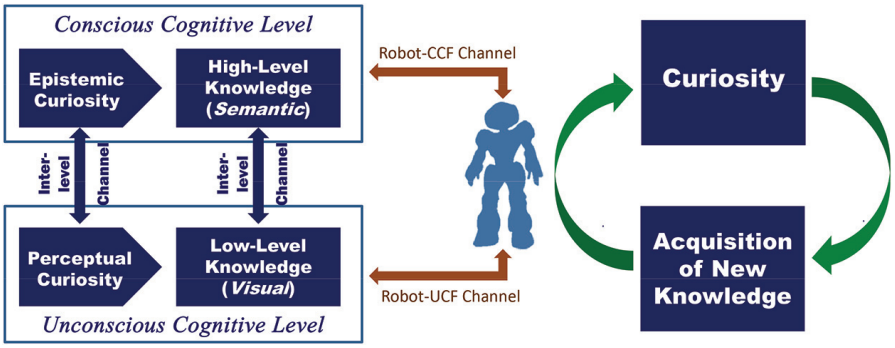

Figure 4.1 General Bloc-diagram of the proposed curiosity driven architecture (left) and principle of curiosity-based stimulation-satisfaction mechanism for knowledge acquisition (right).

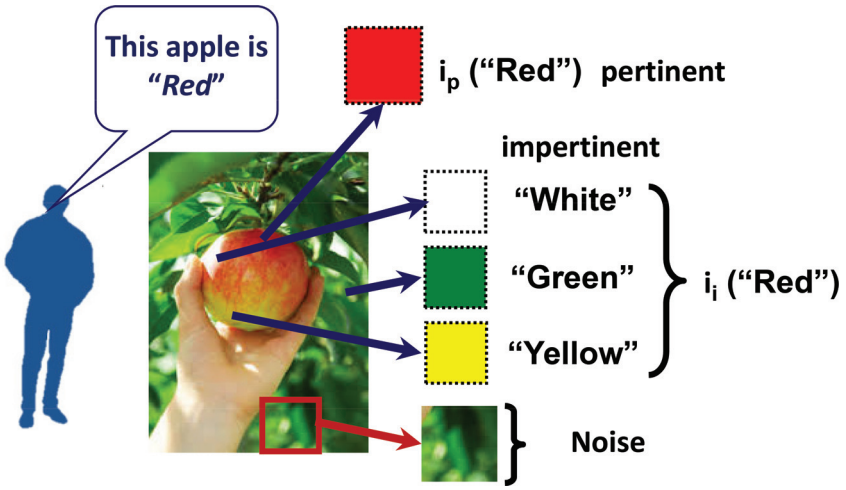

Figure 4.2 A Human would describe this Apple as "Red" in spite of the fact, that this is not the only visible color.

fitness function is of major importance. To calculate the fitness, we train a classifier based on each belief and using it, we try to interpret the objects the robot has already seen. We compare the utterances pronounced by the human tutor in the presence of each such an object 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. 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 Figure 4.3, important parts of the system proposed in this paper are depicted.

4.2.1 Interpretation from Observation

Let us suppose a robot equipped with a sensor observing the surrounding world. The world is represented as a set of features $I = \{i_1 , i_2 , \cdots , i_k\}$,which can be acquired by this sensor [18]. Each time the robot makes an observation *o,* a human tutor gives it a set of utterances U_m describing the important (e.g. salient) objects found. Let us denote the set of all utterances ever given about the world as *U*. The observation *o* is defined as an ordered pair $o = \{I_l , U_m\}$, where $I_l \subseteq I$, expressed by Equation (4.1), stands for the set of features obtained from observation and $U_m \subseteq U$is a set of utterances (describing *o*) given in the context of that observation. i_p denotes the pertinent information for a given *u* (i.e. features that can be described

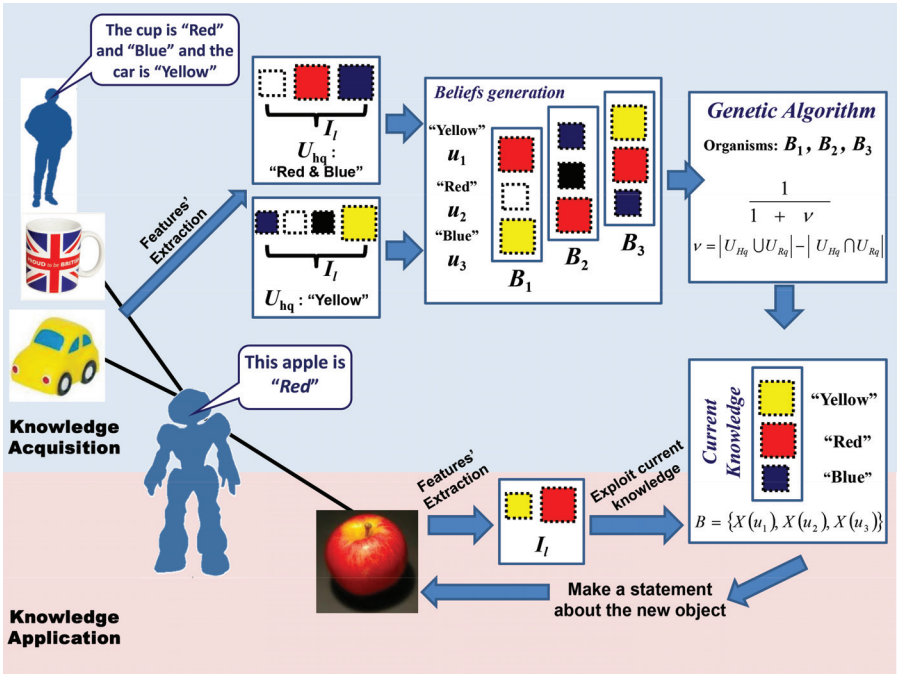

Figure 4.3 A Human would describe this Toy-frog as green in spite of the fact, that this is not the only visible color.

semantically as *u* in the language used for communication between the human and the robot), i_i 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.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 of the world. The robot is further allowed to interact with the human in order to clarify or verify its interpretations.

$$I_l = \bigcup_{U_m} i_p(u) + \bigcup_{U_m} i_i(u) + \varepsilon. \tag{4.1}$$

Let us define an interpretation $X(u) = \{u, I_j\}$of an utterance *u* an ordered pair where $I_j \subseteq I$ is a set of features from *I*. So, the belief *B* is defined according to Equation (4.2) as an ordered set of *X(u)* interpreting utterances *u* from *U*.

$$B = \{X\,(u_1),\ \cdots,\ X\,(u_n)\}\,. \tag{4.2}$$

According to the criterion expressed by (4.3), one can calculate the belief *B*, which interprets in the most coherent way the observations made so far: in other words, by looking for such a belief, which minimizes across all the observations $o_q \in O$ the difference between the utterances U_{Hq} made by the human, and those utterances U_{Bq}, made by the system by using the belief *B*. Thus, *B* is a mapping from the set U to I: 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*.

$$\arg\min_{B}\left(\sum_{q=1}^{|O|}|U_{Hq} - U_{Bq}|\right). \tag{4.3}$$

Figure 4.4 gives, through example, an alternative scheme of the defined notions and their relationship. It depicts a scenario in which two observations o_1 and o_2 are made corresponding to two description $U_1$and U_2 of those observations, respectively.

On first observation, features i_1 and i_2 were obtained along with utterances u_1 and u_2, respectively. Likewise for the second observation, features i_3, i_4 and i_5 were obtained along with utterance u_3. In this example, it is easily visible that the entire set of features $I = \{i_1\,,\ \cdots\,,\ i_5\}$ contains two sub-sets I_1 and I_2. Similarly the ensemble of whole utterances $\{u_1\,,\ u_2\,,\ u_3\}$ give the

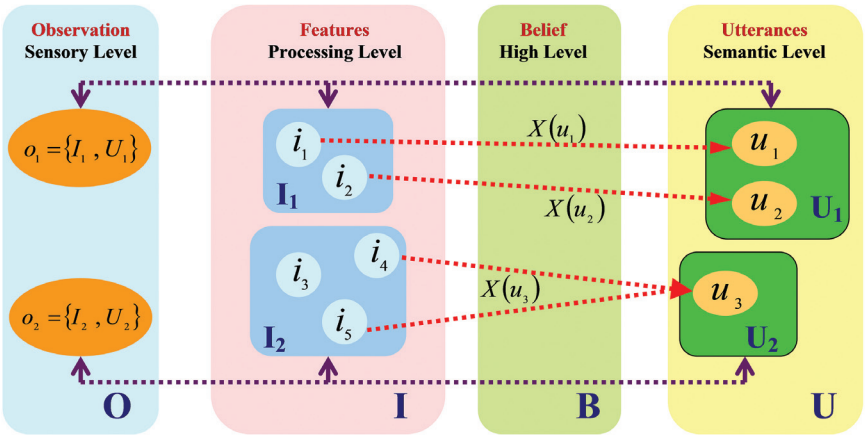

Figure 4.4 Bloc-diagram of relations between observations, features, beliefs and utterances in sense of terms defined in the text.

set U_H and their sub-sets U_1 and U_2 refer to the corresponding observations (e.g. $q \in \{1, 2\}$). In this view, an interpretation $X\,(u_1)$ is a relation of u_1 with a set of features from *I* (namely I_1). Then, a belief *B* is a mapping (relation) from the set *U* to *I*. All members of *U* map to one or more members of *I* and no two members of *U* are associated to the same member of *I*.

4.2.2 Search for the Most Coherent Interpretation

The system has to look for a belief *B,* which would make the robot describe a particular scene with utterances as close and as coherent as possible to those made by a human on the same scene. For this purpose, instead of performing the exhaustive search over all possible beliefs, we propose to search for a suboptimal belief by means of a genetic algorithm. For doing that, we assume that 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*. The task of coherent belief generation is to generate beliefs which are coherent with the observed reality.

In our genetic algorithm, the genomes' generation is a belief generation process generating genomes (e.g. beliefs) as follows. For each interpretation $X(u)$ the process explores the whole the set O. For each observation $o_q \in O$, if $u \in U_{Hq}$ then features $i_q \in I_j$ (with $I_j \subseteq I$) are extracted. As described in (1), the extracted set contains pertinent as well as impertinent features. The coherent belief generation is done by deciding, which features $i_q \in I_j$ may possibly be the pertinent ones. The decision is driven by two principles. The first one is the principle of "proximity", stating that any feature *i* is more likely to be selected as pertinent in the context of, if its distance to other already selected features is comparatively small. The second principle is the "coherence" with all the observations in *O*. This means that any observation $o_q \in O$, corresponding to $u \in U_{Hq}$, has to have at least one feature *i* assigned into I_j of the current $X\,(u) = \{u, I_j\}$ [19]. Thus, it is both the similarity of features and the combination of certain utterances, describing observations from *O* (characterized by certain features), that guide the belief generation process. These beliefs may be seen as "informed guesses" on the interpretation of the world as perceived by the robot.

To evaluate a given organism, a classifier is trained, whose classes are the utterances from *U* and the training data for each class $u \in U$ are those corresponding to $X\,(u) = \{u, I_j\}$, i.e. the features associated with the given *u* in the genome. This classifier is used through the whole set *O* of observations, classifying utterances $u \in U$describing

each $o_q \in O$according to its extracted features. Such a classification results in the set of utterances U_{Bq} (meaning that a belief B is tested regarding the q$^{\text{th}}$ observation). The fitness function evaluating the fitness of each above-mentioned organism is defined as "disparity" between U_{Bq} and U_{Hq} (defined in the previous subsection) which is computed according to the Equation (4.4), where v is given by Equation (4.5) representing the number of utterances that are not present in both sets U_{Bq} and U_{Hq}, which means that they are either missed or are superfluous utterances interpreting the given features.

$$D(\nu) = \frac{1}{1+\nu} \tag{4.4}$$

$$\nu = \left| U_{Hq} \bigcup U_{Bq} \right| - \left| U_{Hq} \bigcap U_{Bq} \right|. \tag{4.5}$$

At the end of the above-described genetic evolution process, the globally best fitting organism is chosen as the belief that best explains the observations O made (by the robot) so far about the surrounding world.

4.2.3 Human-Robot Interaction

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 previous subsections. The latter type of learning requires that the robot be able to communicate with its environment and is facilitated by learning by observation, which may serve as its bootstrap. In our approach, the learning by interaction is carried out in two kinds of interactions: human-to-robot and robot-to-human. The human-to-robot interaction is activated anytime the robot interprets the world wrongly. When the human receives a wrong response (from the robot), he provides the robot a new observation by uttering the desired interpretation. The robot takes this new corrective knowledge about the world into account and searches for a new interpretation according to this new observation. The robot-to-human interaction may be activated 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, then 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 epistemic curiosity, the robot asks its human counterpart to make an utterance about the uncertain observation. If the robot does not interpret according to the utterance given by the human (the robot's interpretation was

wrong), this observation is recorded as new knowledge and a search for the new interpretation is started.

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

4.3 Validation Results by Simulation

In the simulated environment, images of real-world objects were presented to the system alongside with textual tags describing colors present on each object. The images were taken from the Columbia Object Image Library database (COIL: it contains 1000 color images of different views of 100 objects). Five fluent English speakers were asked to describe each object in terms of colors. We restricted the choice of colors to "Black", "Gray", "White", "Red", "Green", "Blue" and "Yellow", based on the color opponent process theory [20]. The tagging of the entire set of images was highly coherent across the subjects. In each run of the experiment, we have randomly chosen a tagged set. The utterances were given in the form of text extracted from the descriptions. The object was accepted as correctly interpreted if the system's and the human's interpretations were equal.

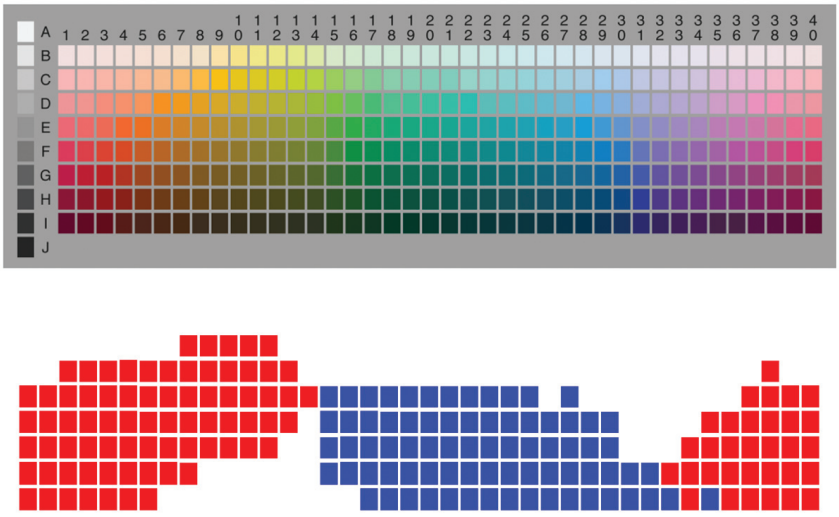

Figure 4.5 Upper: the WCS color table. lower: the WCS color table interpreted by robot taught to distinguish warm (marked by red), cool (blue) and neutral (white) colors.

The rate of correctly described objects from the test set was approximately 91% after the robot had fully learned. Figure 4.5 gives the result of interpretation by the system of the colors of the WCS table regarding "Warm" and "Cool" colors.

Figure 4.6 shows the learning rate versus the increasing number of exposures of each color. It is pertinent to emphasize the weak number of learned examples (required examples) leading to a correct recognition rate

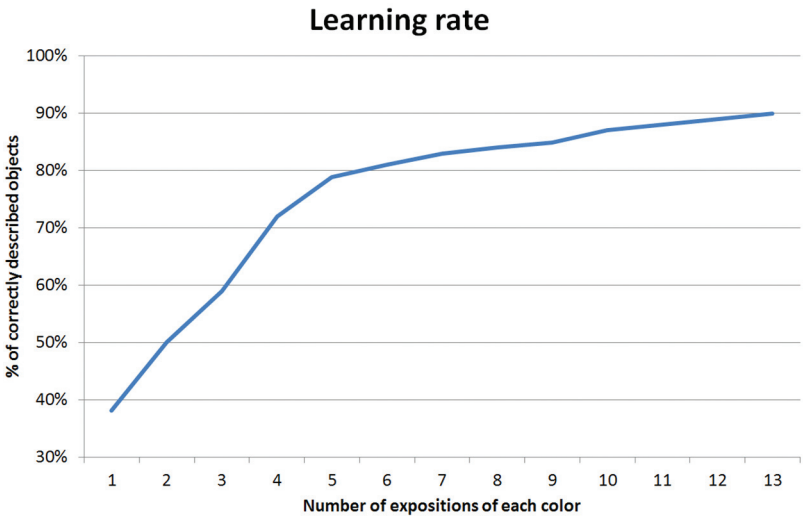

Figure 4.6 Evolution of number of correctly described objects with increasing number of exposures of each color to the simulated robot.

Figure 4.7 Examples of obtained visual colors' interpretations (lower images) and corresponding original images (upper images) for several testing objects from COIL database.

of 91%. Finally, Figure 4.7 gives an example of objects' colors interpretation by the system.

4.4 Implementation on Real Robot and Validation Results

The validation of the proposed system has been performed on the basis of both simulation of the designed system and by an implementation on a real humanoid robot[1]. As real robot we have considered the NAO robot (a small humanoid robot from Aldebaran Robotics) which provides a number of facilities such as onboard camera (vision), communication devices and onboard speech generator. The fact that the above-mentioned facilities were already available offers a huge save of time, even if those faculties remain quite basic in that kind of robot.

Although the usage of the presented system is not specifically bound to humanoid robots, it is pertinent to state two main reasons why a humanoid robot has been used for the system's validation. The first reason for this is that from the definition of the term "humanoid", a humanoid robot aspires to make its perception close to the human one, entailing a more human-like experience of the world. This is an important aspect to be considered in the context of sharing knowledge between a human and a robot. Some aspects of this problem are discussed in [21]. The second reason is that humanoid robots are specifically designed in order to interact with humans in a "natural" way by using a loudspeaker and microphone set. Thus, required facilities for bi-directional communication with humans through speech synthesis and speech recognition are already available on such kinds of robots. This is of major importance when speaking is a central item for natural human-robot interaction.

4.4.1 Implementation

The core of the implementation's architecture is split into five main units: Communication Unit (CU), Navigation Unit (NU), Low-level Knowledge Acquisition Unit (LKAU), High-level Knowledge Acquisition Unit (HLAU) and Behavior Control Unit (BCU). Figure 4.8 illustrates the bloc-diagram of the implementation's architecture. The aforementioned units control NAO robot (symbolized by its sensors, its actuators and its interfaces in Figure 4.8)

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

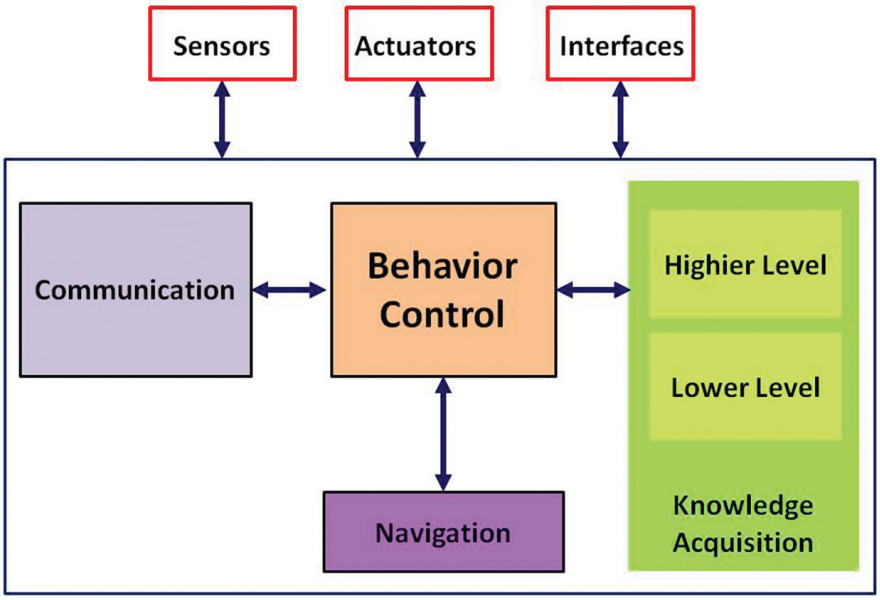

Figure 4.8 Block diagram the implementation's architecture.

through its already available hardware and software facilities. In other words, the above-mentioned architecture controls the whole robot's behavior.

The purpose of NU is to allow the robot to position itself in space with respect to objects around it and to use this knowledge to navigate within the surrounding environment. Capacities needed in this context are obstacle avoidance and determination of distance to objects. Its sub-unit handling spatial orientation receives its inputs from the camera and from the LKAU. To get to the bottom of the obstacle avoidance problem, we have adopted a technique based on ground color modeling. Inspired by the work presented in [22], color model of the ground helps the robot to distinguish free-space from obstacles. The 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 from monocular camera data. In [23], some aspects of distance estimation from a static monocular camera have been mentioned, proffering the robot the capacity to infer distances and sizes of surrounding objects.

The LKAU ensures gathering of visual knowledge, such as detection of salient objects and their learning (by the sub-unit in charge of salient object detection) and sub-recognition (see [18, 24]). Those activities are

carried out mostly in an "unconscious" manner, that is, 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 HKAU is the center where the intellectual behavior of the robot is constructed. Receiving its features from the LKAU (visual features) and from the CU (linguistic features), this unit processes the belief generation, the most coherent beliefs emergence and constructs the high-level semantic representation of acquired visual knowledge. Unlike the LKAU, this unit represents conscious and intentional cognitive activity. In some way, it operates as a baby who learns from observation and from verbal interaction with adults about what he observes developing in this way his own representation and his own opinion about the observed world [25].

The CU is in charge of robot communications. It includes 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 from the BCU. The input channel takes its input from a microphone and through an Automated Speech Recognition engine (available in NAO) the syntax and semantic analysis (designed and incorporated in BCU) it provides the BCU labeled chain of strings representing the heard speech. As it has been mentioned, the syntax analysis is not available on NAO. Thus it has been incorporated in BCU. To perform syntax analysis, the TreeTagger tool is used. Developed by the ICL at University of Stuttgart, the TreeTagger tool is a tool for annotating text with part-of-speech and lemma information. Figure 4.9 shows, through a simple example of an English phrase, the operational principle of syntactic analysis performed by this tool. "Part-of-speech" row gives tokens explanation 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

Phrase	Robots	are	our	friends
Tokens	NNS	VBS	PP$	NNS
Part-of-speech	Noun, plural	Verb, present	Possessive pron	Noun, plural
Lemma	robot	be	our	friend

Figure 4.9 Example of English phrase and the corresponding syntactic analysis output generated by treetagger.

declarative (for example: "This is a Box"), interrogative (for example: "What is the name of this object?") or imperative (for example: "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 make emerge the appropriate action by the robot. Figure 4.10 gives the flow diagram of communication between the robot and a human as it has been implemented in this work.

The BCU plays the role of a coordinator of robot's behavior. It handles data flows and issues command signals for other units, controlling the behavior of the robot and its suitable reactions to external events (including its interaction with humans). BCU received its inputs from all other units and returns its outputs to each concerned unit including robot's devices (e.g. sensors, actuators and interfaces) [25].

4.4.2 Validation Results

A total of 25 every-day objects was collected for experimental purposes of (Figure 4.11). They have been 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 from its surrounding environment containing objects from the above-mentioned set.

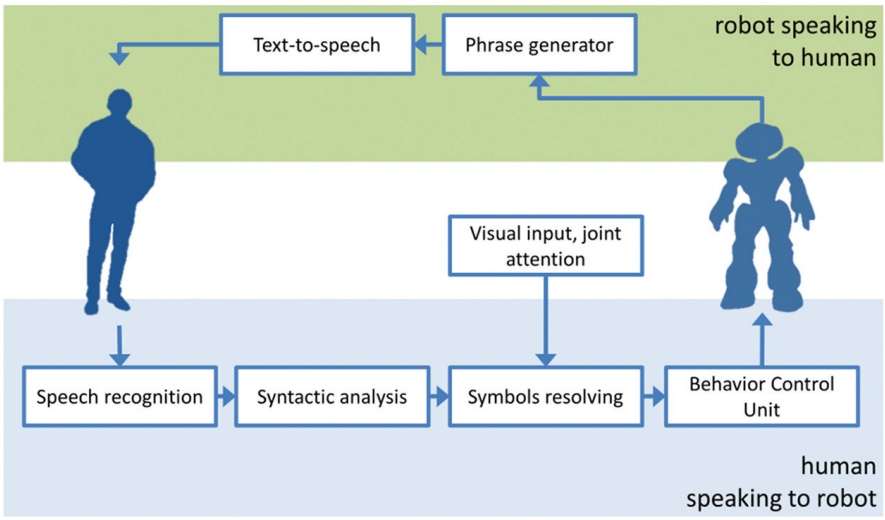

Figure 4.10 Flow diagram of communication between a robot and a human which is used in this work.

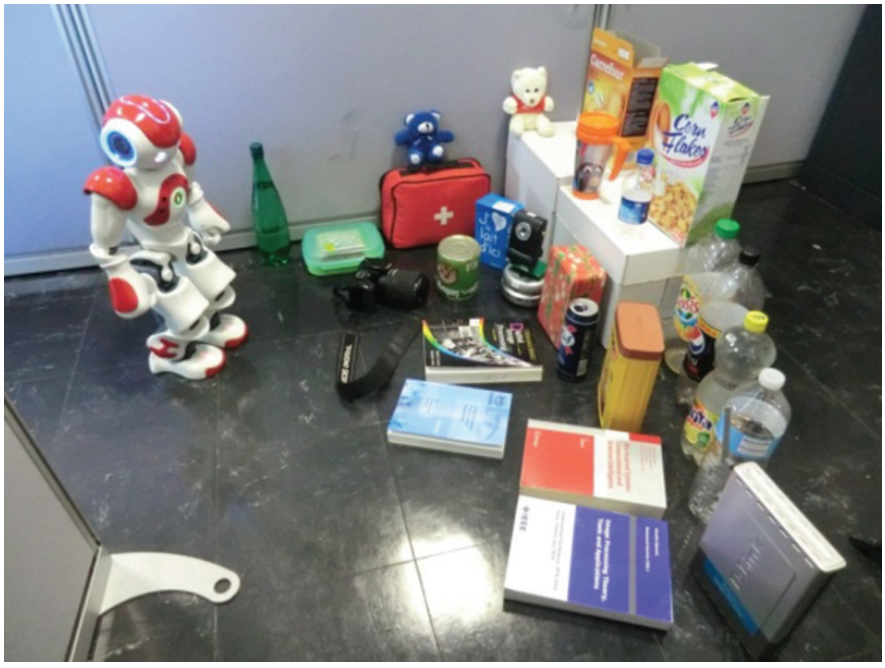

Figure 4.11 Everyday objects used in the experiments in this work.

The first validation involving the robot has aimed at verifying the leaning, color interpretation, interaction with human and description abilities of the proposed (e.g. investigated) system. To do this, the robot has been asked to learn a subset of the 25 objects: in terms of associating the name of each detected object to that object. At the same time, a second learning process has been performed involving the interaction with the tutor who has successively pointed the above-learned objects describing (e.g. telling) to the robot the color of each object. Here below, an example of the Human-Robot interactive learning is reported:

- **Human**: [pointing a red aid-kit] "This is a first-aid-kit!"
- **Robot**: "I will remember that this is a first-aid-kit."
- **Human**: "It is red and white."
- **Robot**: "OK, the first-aid-kit is red and the white."

After learning the names and colors of the observed objects, the robot is asked to describe a number of objects including also some of the already learned objects but in a different posture (for example the yellow chocolate

box presented in reverse posture, etc.) and a number of still unseen objects (as for example a red apple or a white teddy-bear). The robot has successfully described, in coherent linguistics, the presented seen and unseen objects. Here below is an example of Human-Robot interaction during the recognition phase:

- **Human**: [pointing the unseen white teddy-bear]: "Describe this!"
- **Robot**: "It is white!"
- **Human**: [pointing the already seen, but reversed, yellow chocolate box]: "Describe this!"
- **Robot**: "It is yellow!"
- **Human**: [pointing the unseen apple]: "Describe this!"
- **Robot**: "It is red!"

Figure 4.12 shows two photographs of the above-reported experimental validation, where the robot completes its knowledge by interacting with a human and learning from him. Figure 4.13 shows another two photographs where the robot describes different objects it could detect within its surrounding environment.

In another experiment, taking into account the previously learned objects, among which the robot has detected and has learned a black book (the only available book in the learning set), the robot has been asked to search for the book in a new (e.g. unseen) environment where a number of different objects, including the above-mentioned black book, have been placed with two new books (unseen) as well as other objects looking similar to a book. The robot's behavior and Human-Robot interaction during this experiment is given here below:

- **Human**: "Go find the book!"
- **Robot**: "OK, I am searching for the book!"

Figure 4.12 Tutor pointing an aid-kit detected by robot describing its name and color to the robot (left-side picture). Pointing, in the same way, another visible objects detected by robot, tutor describes them to the robot (Right-Side picture).

Figure 4.13 Tutor pointing a yellow chocolate box which has been seen, interpreted and learned (by the robot) in terms of colors then asking the robot to describe the chosen object (left-side picture). Tutor pointing an unseen white teddy-bear asking the robot to describe the chosen object (right-side picture).

- **Robot**: [after visually exploring the new surrounding environment]: "I can see three books!"
- **Human**: "Get the red one!"
- **Robot**: [after turning toward the "red and grey" book and after walking and reaching the aforementioned book's location] "I am near the red book!"

Figure 4.14 shows two photographs of the robot during its searching for the required "red" book. Besides the correct recognition of the desired object (asked by the tutor), what is pertinent and interesting to note is that the robot also found two other unseen books. What is also very pertinent to emphasize, and very interesting, is that even if there is no "red" book in that environment, the robot has correctly interpreted the fact that the red book required by the human was the "red and grey" book: the only book that may coherently

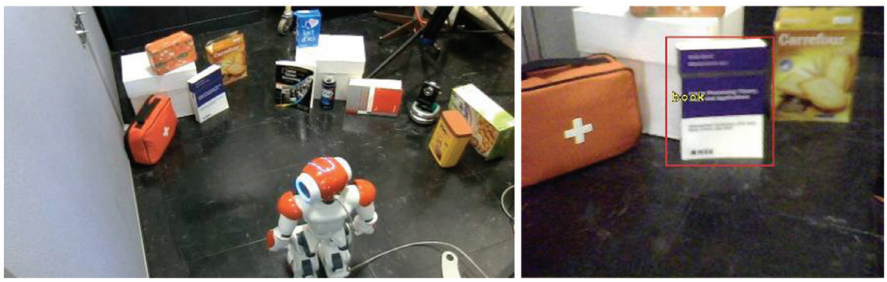

Figure 4.14 Images from a video sequence showing the robot searching for the book (left-side picture) and robot's camera view and visualization of color interpretation of the searched object (right-side picture).

be considered as "red" by the human. A video showing the experimental validation may be found on http://youtu.be/W5FD6zXihOo. More details of the presented work with complementary results can be found in [19, 25].

4.5 Conclusions

This chapter has presented, discussed and validated a cognitive system for high-level knowledge acquisition from visual perception based on the notion of artificial curiosity. Driving as well the lower as the higher levels of the presented cognitive system, the emergent artificial curiosity allows such a system to learn in an autonomous manner new knowledge about the unknown surrounding world and to complete (enrich or correct) its knowledge by interacting with a human. Experimental results, performed as well on a simulation platform as using the NAO robot, show the pertinence of the investigated concepts as well as the effectiveness of the designed system. Although it is difficult to make a precise comparison due to different experimental protocols, the results we obtained show that our system is able to learn faster and from significantly fewer examples than most of more-or-less similar implementations.

Based on the results obtained, it is thus justified to say that a robot endowed with such artificial curiosity-based intelligence will necessarily include autonomous cognitive capabilities. With respect to this, the further perspectives regarding the autonomous cognitive robot presented in this chapter will focus on integration of the investigated concepts in other kinds of robots, 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.

References

[1] D. Meger, P. E. Forssén, K. Lai, S. Helmer, S. McCann, T. Southey, M. Baumann, J. J. Little and D. G. Lowe, 'Curious George: An attentive semantic robot', Robot. Auton. Syst., vol. 56, no. 6, pp. 503–511, 2008.

[2] P. Kay, B. Berlin and W. Merrifield, 'Biocultural Implications of Systems of Color Naming', Journal of Linguistic Anthropology, vol. 1, no. 1, pp. 12–25, 1991.

[3] M. Bowerman, '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, 1983.
[4] M. A. Goodrich and A. C. Schultz, 'Human-robot interaction: a survey', Found. Trends Hum.-Comput. Interact., vol. 1, no. 3, pp. 203–275, 2007.
[5] S. Coradeschi and A. Saffiotti, 'An introduction to the anchoring problem', Robotics & Autonomous Sys., vol. 43, pp. 85–96, 2003.
[6] T. Regier, 'A Model of the Human Capacity for Categorizing Spatial Relations', Cognitive Linguistics, vol. 6, no. 1, pp. 63–88, 1995.
[7] J. de Greeff, F. Delaunay and T. Belpaeme, 'Human-robot interaction in concept acquisition: a computational model', Proc. of Int. Conf. on Development and Learning, vol. 0, pp. 1–6, 2009.
[8] P. Wellens, M. Loetzsch and L. Steels, 'Flexible word meaning in embodied agents', Connection Science, vol. 20, no. 2–3, pp. 173–191, 2008.
[9] J. Saunders, C. L. Nehaniv and C. Lyon, 'Robot learning of lexical semantics from sensorimotor interaction and the unrestricted speech of human tutors', Proc. of 2nd International Symposium on New Frontiers in Human-Robot Interaction, Leicester, pp. 95–102, 2010.
[10] Lütkebohle, J. Peltason, L. Schillingmann, B. Wrede, S. Wachsmuth, C. Elbrechter and R. Haschke, 'The curious robot - structuring interactive robot learning', Proc. of the 2009 IEEE international conference on Robotics and Automation, Kobe, pp. 2154–2160, 2009.
[11] T. Araki, T. Nakamura, T. Nagai, K. Funakoshi, M. Nakano and N. Iwahashi, 'Autonomous acquisition of multimodal information for online object concept formation by robots', Proc. of IEEE/ IROS, pp. 1540–1547, 2011.
[12] D. Skocaj, M. Kristan, A. Vrecko, M. Mahnic, M. Janicek, G.-J. M. Kruijff, M. Hanheide, N. Hawes, T. Keller, M. Zillich and K. Zhou, 'A system for interactive learning in dialogue with a tutor', Proc.of IEEE/ IROS, pp. 3387–3394, 2011.
[13] C. Yu, 'The emergence of links between lexical acquisition and object categorization: a computational study', Connection Science, vol. 17, 3–4, pp. 381–397, 2005.
[14] S. R. Waxman and S. A. Gelman, 'Early word-learning entails reference, not merely associations', Trends in cognitive science, 2009.
[15] D. E. Berlyne, 'A theory of human curiosity', British Journal of Psychology, vol. 45, no. 3, August, pp. 180–191, 1954.

[16] K. Madani, C. Sabourin, 'Multi-level cognitive machine-learning based concept for human-like artificial walking: Application to autonomous stroll of humanoid robots', Neurocomputing, S.I. on Linking of phenomenological data and cognition, pp. 1213–1228, 2011.

[17] K. Madani, D. Ramik, C. Sabourin, 'Multi-level cognitive machine-learning based concept for Artificial Awareness: application to humanoid robot's awareness using visual saliency', J. of Applied Computational Intelligence and Soft Computing,. DOI: 10.1155/2012/354785, 2012. (available on: http://dx.doi.org/10.1155/2012/354785).

[18] D. M. Ramik, C. Sabourin, K. Madani, 'A Machine Learning based Intelligent Vision System for Autonomous Object Detection and Recognition', J. of Applied Intelligence, Springer, Vol. 40, Issue 2, pp. 358–374, 2014.

[19] D-M. Ramik, C. Sabourin. K. Madani, 'From Visual Patterns to Semantic Description: a Cognitive Approach Using Artificial Curiosity as the Foundation', Pattern Rgognition Letters, Elsevier, vol. 34, no. 14, pp. 1577–1588, 2013.

[20] M. Schindler and J. W. v. Goethe, 'Goethe's theory of colour applied by Maria Schindler', New Knowledge Books, East Grinstead, Eng., 1964.

[21] V. Klingspor, J. Demiris, M. Kaiser, 'Human-Robot-Communication and Machine Learning', Applied Artificial Intelligence, pp. 719–746, 1997.

[22] J. Hofmann, M. Jngel, M. Ltzsch, 'A vision based system for goal-directed obstacle avoidance used in the rc'03 obstacle avoidance challenge', Lecture Notes in Artificial Intelligence, Proc. of 8th International Workshop on RoboCup, pp. 418–425, 2004.

[23] D. M. Ramik, C. Sabourin, K. Madani, 'On human inspired semantic slam's feasibility', Proc. of the 6th International Workshop on Artificial Neural Networks and Intelligent Information Processing (ANNIIP 2010), ICINCO 2010, INSTICC Press, Funchal, pp. 99–108, 2010.

[24] R. Moreno, D. M. Ramik, M. Graña, K. Madani, 'Image Segmentation on the Spherical Coordinate Representation of the RGB Color Space', IET Image Processing, vol. 6, no. 9, pp. 1275–1283, 2012.

[25] D. M. Ramik, C. Sabourin, K. Madani, 'Autonomous Knowledge Acquisition based on Artificial Curiosity: Application to Mobile Robots in Indoor Environment', J. of Robotics and Autonomous Systems, Elsevier, Vol. 61, no. 12, pp. 1680–1695, 2013.