To appear in the *Journal of Experimental & Theoretical Artificial Intelligence*
Vol. 00, No. 00, Month 20XX, 1–26

Long-term Knowledge Acquisition using Contextual Information in a Memory-inspired Robot Architecture

Ferdian Pratama[a*], Fulvio Mastrogiovanni[b], Soon Geul Lee[c], and Nak Young Chong[a,c]

[a] *Japan Advanced Institute of Science and Technology, Japan*;
[b] *University of Genoa, Italy*; [c] *Kyung Hee University, South Korea*

(*Received 00 Month 20XX; final version received 00 Month 20XX*)

In this paper, we present a novel cognitive framework allowing a robot to form *memories* of relevant traits of its perceptions and to recall them when necessary. The framework is based on two main principles: on the one hand, we propose an architecture inspired by current knowledge in human memory organization; on the other hand, we integrate such an architecture with the notion of *context*, which is used to modulate the knowledge acquisition process when consolidating memories and forming new ones, as well as with the notion of *familiarity*, which is employed to retrieve proper memories given relevant cues. Although much research has been carried out, which exploits Machine Learning approaches to provide robots with internal models of their environment (including objects and occurring events therein), we argue that such approaches may not be the right direction to follow if a long-term, continuous knowledge acquisition is to be achieved.

As a case study scenario, we focus on **both robot-environment and human-robot interaction processes. In case of robot-environment interaction, a robot performs pick and place movements using the objects in the workspace, at the same time observing their displacement on a table in front of it,** and progressively *forms* memories defined as relevant cues (e.g., color, shape or relative position) in a context-aware fashion. **As far as human-robot interaction is concerned,** the robot can recall specific snapshots **representing** past events using both sensory information and contextual cues **upon request** by humans.

Keywords: robot cognitive architectures; developmental learning; long-term knowledge acquisition; context-based memory retrieval.

1. Introduction

The role of *natural context* in human and animal behavior proves to be fundamental at various levels. We refer to a natural context as **those elements involved or associated with familiar environments** (including their physical laws and social rules) where a human or animal lives and operates (Allen & Bekoff, 1999). **Since contextual information originates from the relationship between a human and its environment, multiple contexts can be certainly pertinent to human behaviour at the same time.** As a matter of fact, both human-human and human-environment interaction processes are believed to be greatly affected by their natural context (Mehl & Conner, 2012). **The presence of multiple contexts can affect the way humans interact with each other and the environment. As the experience of someone develops, so do the contexts involving that individual, therefore modifying**

*Corresponding author. Email: ferdian.adi@gmail.com

their semantics even in sensible ways.

In humans, experimental evidence suggests that context-aware mechanisms are mostly represented in the hippocampus (Smith & Mizumori, 2006). Such mechanisms contribute to the identification of specific contexts with respect to other situations so that the most appropriate behavioral response can be executed or the most relevant mnemonic output can be retrieved. The contextual information elicited by the surrounding environment, the objects therein (as well as their properties), and the flow of events we perceive, play a significant role in our daily behavior, eventually influencing the way we create mental models of what we perceive and remember (Godden & Baddeley, 1975; Smith & Kosslyn, 2009).

In order to design robots able to pro-actively and sensibly understand their environment and to engage humans in long-term interaction processes, a similar concept of *robot natural context* must be envisaged and integrated into their cognitive architecture. Given the role of natural context in the formation of human mental models and memories, it is necessary to investigate what are the implications of the equivalent concept of robot natural context on a robot cognitive architecture that must be necessarily inspired by memory-related mechanisms and processes.

The goal of this paper is to investigate the role of robot natural context in a robot able to progressively acquire, consolidate and recall knowledge **during the execution of goal-oriented behavior (for both human-robot and robot-environment interaction tasks), and during human-robot verbal interaction for memory recollection.** In order to achieve such a goal, two research strands must be considered.

(1) No integrated and context-based robot architecture (possibly inspired by the developmental paradigm) is currently available, which is aimed at long-term human-robot interaction processes using a precise characterization of memory components, i.e., taking into account their interconnectivity.
(2) In spite of recent research activities on memory-inspired architectures (Bellas, Faina, Varela, & Duro, 2010; Morse, de Greeff, Belpaeme, & Cangelosi, 2010; Nuxoll & Laird, 2004), which are based on individual memory components, no holistic approach has been devised to provide robot architectures with the necessary flexibility to efficiently deal with contextual information.

An analysis of the literature shows that research in memory-inspired architectures is mainly focused on modeling memory components on an individual basis, such as the Working Memory (WM) (Phillips & Noelle, 2005), the Episodic Memory (EM) (Dodd & Gutierrez, 2005; Jockel, Weser, Westhoff, & Zhang, 2008; Jockel, Westhoff, & Zhang, 2007; Kasap & Magnenat-Thalmann, 2010; Kuppuswamy, Cho, & Kim, 2006; Nuxoll, 2007; Nuxoll & Laird, 2004, 2012; Stachowicz & Kruijff, 2012; Tecuci & Porter, 2007) or the Procedural Memory (PM) (Salgado, Bellas, Caamano, Santos-Diez, & Duro, 2012).

A precise (i.e., as formal as possible) characterization of architectural components and the associated information flow must be defined. On the one hand, Stachowicz and Kruijff (2012) provide an thorough discussion of both design requirements and formal concepts needed to characterize EM and its storage structure. However, the focus of their work is on the notion of *event*, its properties, and its use in such cognitive processes as event recognition and recursion of events. Despite their claim of having designed an EM-like memory structure, it is noteworthy that they do not exploit the notion of *context*, which is considered of the utmost importance for EM by Godden and Baddeley (1975) and Smith and Kosslyn (2009). On the other hand, when an attempt is made to design a more comprehensive memory-inspired robot architecture (Bellas et al., 2010; Morse et al., 2010; Nuxoll & Laird, 2004), the goal is limited to finding a solution to very specific problems, instead of providing the robot with the capability of developing its own

knowledge on the long-term. Furthermore, in such approaches, neither the relationships between different memory components is explicitly addressed, nor the mutual influence between components is considered. In any case, no clear use of the notion of context is provided.

In this paper, we present a memory-inspired robot cognitive architecture that allows a robot to progressively acquire knowledge using context-based information. **To demonstrate the features of the proposed architecture, we focus on a scenario where robot-environment and human-robot interactions are involved. During robot-environment interactions, the robot performs pick and place movements involving objects located on a table in front of it, and is expected to progressively form memory items by observing visual changes in the environments affected by external influences (i.e., human actions, or in this case, robot movements). In particular, pick and place actions deal with objects characterized by different physical properties (i.e., colors and shapes), and displace one object at a time to a different position within the workspace, hence resulting in a different workspace configuration. In case of the human-robot interactions, after the robot forms memory items by the case of robot-environment interactions, the robot interacts with a human, where the latter inquire verbal questions to the former involving contextual information, specifically regarding the robot past experience. Examples of inquiries include: *"What orange objects do you know?", "What color was the leftmost object, when three objects were presented?", or "How many objects were presented when the orange lamp was the rightmost?"*. We will also discuss how human-robot interaction can occur in parallel with robot-environment interaction.**

The contribution of the paper is two-fold: (i) we demonstrate the robot ability to store and **recall** memory items as a result of acquiring personal experience, on the basis of specific cues provided by a human; (ii) we show that contextual information (which may be *familiar* to the robot), is fundamental for the retrieval process.

The proposed architecture is culturally based on the two following arguments. On the one hand, avoiding the currently widespread mindset that robot developmental approaches are to be identified with Machine Learning frameworks, we argue that continuous knowledge acquisition allows for a progressive evolution of the stored knowledge and its representation, which is based on a continuous interaction with the robot natural environment. On the other hand, inspired by state of the art studies in Developmental Psychology by Baddeley (2000); Eichenbaum and Howard (2001); Godden and Baddeley (1975); Smith and Mizumori (2006); Smith and Kosslyn (2009); Tulving (2001, 2002), we argue that an explicit addressing of the role of memory in human-robot interaction processes is crucial in robot knowledge development.

The paper is organized as follows. Section 2 discusses relevant literature. Section 3 introduces the main concepts of the approach, as well as the system architecture. Section 4 elaborates on the conducted experiments with a specific, real-world scenario for the application domain. Conclusions follow.

2. Related Work

2.1. *Memory Models and Terminology*

Albeit there is no widespread consensus about a general framework, memory models typically assume a multi-storage organization. Two models constitute fundamental milestones in the literature, namely the *multi store model* by Atkinson and Shiffrin (1968) and the *working memory model* by Baddeley and Hitch (1974).

Adopting a computational approach, the multi store model describes how information **is formed into memory items and organized into different memory models.** Three stores are usually identified, namely the **Sensory Memory, the Short-Term Memory (STM) and the Long-Term Memory (LTM).** Different processes are involved in the management of such an information flow. After being perceived and properly conveyed to the brain through relevant neural pathways, sensory information is represented inside Sensory Memory (available for less than 1 *sec*). If sensory information is *attended*, the relevant part of it is transferred to STM, where it is processed for immediate use (occurring between 0 and 18 *sec*). Then, if such a representation is *rehearsed* (an elaborative process further developed by Raaijmakers and Shiffrin (2003)), it is transferred to LTM (in principle, therein available forever). Otherwise, it is lost from STM according to a memory-trace decay process.

The evidence for a distinction between STM and LTM is given in various studies related to amnesia since the well-known case of patient H. M. (Squire & Kandel, 2000), who still exhibits capabilities in retaining memories in STM but he is not **able anymore** to consolidate any new memory items in LTM.

However, this model suffers from a number of limitations, namely a quite simplified structure related to both STM and LTM, as well as a biased focus on attention and rehearsal, which turned out not to be essential, as described in later studies.

Baddeley and Hitch (1974) proposed a model for STM (which they call Working Memory - WM) that aims at better characterizing its subcomponents, each one devoted to represent and process different types of information. Specifically, WM consists of the Central Executive that orchestrates the behaviors of two subcomponents, namely the Visuo-Spatial Sketchpad and the Phonological Loop. **The Central Executive** is believed to deal with cognitive tasks related to logic and to make an on-demand use of subcomponents. **The Visuo-Spatial Sketchpad** processes visual and spatial based information, e.g., related to any motion in the environment. **The Phonological Loop** deals with symbol-mediated information (i.e., which can be written or spoken), and can be further divided in two parts, namely the Phonological Store (linked to speech perception) and the Articulatory Control Process (linked to speech production), see Jones, Macken, and Nicholls (2004); Shaw and Tiggemann (2004).

As a consequence of follow-up experiments, the original model has been updated by Baddeley (2000) to include a third subcomponent managed by **the Central Executive**, namely the Episodic Buffer. The role of **the Episodic Buffer** is to mediate between LTM and other components of WM: when **WM** is capable of identifying an observable relevant event (as a result of **Visuo-Spatial Sketchpad** and **Phonological Loop** processing), **the Episodic Buffer** appropriately manages its storage in LTM. Nowadays, there is no shortage of reasons to believe that STM is made-up of a number of subcomponents. The WM model accounts for a number of real-world functional behaviors, such as task and verbal-level reasoning, reading and comprehension, problem solving, as well as visual and spatial information processing.

With respect to LTM, as proposed by Atkinson and Shiffrin (1968), two parts can be identified, i.e., explicit and implicit memory (Wood, Baxter, & Belpaeme, 2011). Explicit memory (also referred to as Declarative Memory) refers to consciously available memory items. It can be further divided in **two** subcomponents, namely the Episodic Memory (EM) and the Semantic Memory (SM). EM is related to the encoding of generic events localized in time. An example of EM is the set of specific event occurred during the interaction with someone or with the environment. Knowledge about facts and their meaning is stored in SM. Differently from the content of EM, SM is not believed to depend on contextual information (Spaniol, Madden, & Voss, 2006). Finally, implicit memory (also known as Procedural Memory) refers to motor action, specifically actions

involved in the use of objects (including grasping, manipulation and tool use), as well as body motions (Bullemer, Nissen, & Willingham, 1989).

2.2. *Models of Memory Components*

Although the systematic study of human memory traces back to 1960s (refer to the book by Squire and Kandel (2000) for a historical account and the references therein), only in the past few years a number of approaches have been presented, which aim at modeling different aspects of human cognition and reasoning, as well as at developing computational paradigms to encode them in robot cognitive architectures.

In the following paragraphs, we limit our attention to literature explicitly taking memory components modeling into account, possibly grounded in a robot implementation.

In the past few years, two approaches have been presented, which attempt at modeling architectural aspects of memory *as a whole*, namely the work by Bellas et al. (2010) and Morse et al. (2010). Both the approaches put a great emphasis on memory components and their interconnections.

Bellas et al. (2010) employ the concept of Multilevel Darwinist Brain proposed by Bellas and Duro (2004) to develop an evolutionary behavior-based robot architecture. The framework is based on an Artificial Neural Network (ANN) neuro-evolutionary approach. Experiments are conducted on a Sony AIBO robot, which learns the basic sensorimotor behaviors associated with a ball catching task. Both STM and LTM memory components are modeled. Specifically, STM is further organized as a WM component (limited to vision processing) and an **Episodic Buffer** component, later elaborated by Salgado et al. (2012). The framework allows for concurrent behavior execution and ANN-based continuous knowledge evolution. A clear description of the advantages of applying evolutionary approaches to a robot cognitive architecture is not adequately motivated. Furthermore, the proposed framework lacks much **detail** about the actual organization of both STM and LTM, as well as their mutual relationships.

A number of approaches are devoted to model specific memory components. With respect to SM, two approaches are particularly interesting in our case, namely those put forward by Dodd (2005) and Dayoub, Duckett, and Cielniak (2010).

The objective of the SM component designed by Dodd (2005) is to maintain information about objects located in the environment. This is achieved using a novel architecture combining the so-called Sensory EgoSphere later refined by Peters II, Hambuchen, and Bodenheimer (2009), as well as SM, WM and **the Central Executive**. Although interesting, the framework is characterized by a number of drawbacks, as follows: (i) *a priori* knowledge about objects and the associated symbol grounding (Harnad, 1990) is required; (ii) since SM is designed to model and recognize objects in a very specific application domain, SM lacks the ability to represent anything that is not related to objects. In spite of these flaws, the framework has nonetheless the advantage of exhibiting a partial interconnectivity between the memory items pertaining to EM and SM.

Dayoub et al. (2010) propose a SM component based on the multi store model of human memory advocated by Atkinson and Shiffrin (1968), specifically in the context of semantic mapping tasks carried out by a mobile robot. The robot is able to track the displacement of several objects using omni-directional vision and it is able to provide humans with the most likely suggestion about the location of any tracked objects within the map. The overall behavior is managed using finite state machines. The advantages of the framework include: (i) a strong interconnection between the representation of objects, their locations and the capability of updating the internal model of the environment (i.e., the map); (ii) SM is tightly connected with the object tracking module, and it provides humans with comprehensive information about the map as a result of a human-robot

interaction process. Specifically, humans may pose questions such as *Where was object* x *the last time you have seen it?* or *What are the most likely locations to find object* x *in the map?* Since robot knowledge is only limited to object properties and locations, the scope of questions that can be posed by humans is limited. However, the possibility of posing questions inspired us to implement a **query**-based knowledge information retrieval process.

As far as PM is concerned, the approaches by Salgado et al. (2012) and Dodd (2005) have been considered.

The PM component by Salgado et al. (2012) stores basic skills and behaviors as a library to ground robot learning. Specifically, a Sony AIBO robot is expected to learn a ball catching behavior. Whilst the architecture has been designed to implement adaptive learning techniques, it features a model of PM that turns out not to be consistent with state of the art psychological studies. Furthermore, the information that can be obtained as a result of human-robot interaction processes is limited due to the inability of the system to store any information other than the learned associated behavior.

In the PM design proposed by Dodd (2005), robot motions are represented as nodes in a graph-like structure labeled as behavior nodes, motion primitive nodes, and example nodes. The architecture is designed to select PM nodes and to properly sequence them (Mastrogiovanni & Sgorbissa, 2013; Ratanaswasd, Gordon, & Dodd, 2005). Even though motions generated by sequencing robot behaviors are claimed to be fairly smooth, the PM design is highly dependent on the employed modular controller, as it as been pointed out by Ratanaswasd et al. (2005), the used behavior interpolator, and the trajectory error-reduction algorithm. Furthermore, since the structure of PM nodes only contains information about the associated behavior and the corresponding 3D trajectory, no comprehensive memory component interconnection is actually possible.

Finally, three approaches to model EM have been considered in our analysis, namely the work by Jockel et al. (2007); Stachowicz and Kruijff (2012) and, again, by Dodd (2005).

Consistently with the notion of EM, the approach proposed by Stachowicz and Kruijff (2012) is focused on a formal framework used to represent and relate events occurring in both space **and** time into *spatio-temporal contexts*. In particular, a hierarchy of events is envisaged, where an *event* can be either *atomic* or *complex*. Atomic events can be combined in different ways to form so-called *sub*events and *super*events. Unfortunately, no formal account is provided about the adopted notion of context and – above all – its influence on the other components of the architecture. The proposed EM design also lacks any correlations between EM and EB, specifically in view of a continuous knowledge acquisition process while interacting with the environment.

The design for EM proposed by Jockel et al. (2007) assumes that an event is hierarchically classified as belonging to one of the following classes: perceptional event, command event, and executive event. In this case, an event is associated with procedural callback procedures. Among the claimed advantages of the architecture, the possibility of storing past experiences in a life-long memory storage component, and the ability to perform *one shot* learning processes. Again, no formal definition of such a notion of event is provided. This is surprising, given the argument that EM essentially consists of sequences of events.

Finally, the EM component designed by Dodd (2005) assumes it to be a medium for robot learning processes. Temporally sequenced records of specific events are stored as memory items called *episodes*. An association is maintained between EM items and the content of SM and WM, as well as task-related information (in a sense, mimicking the availability of PM). Episodes are retrieved from EM using an approach similar to what has been discussed by Anderson (1990) in the context of the ACT-R architecture. The

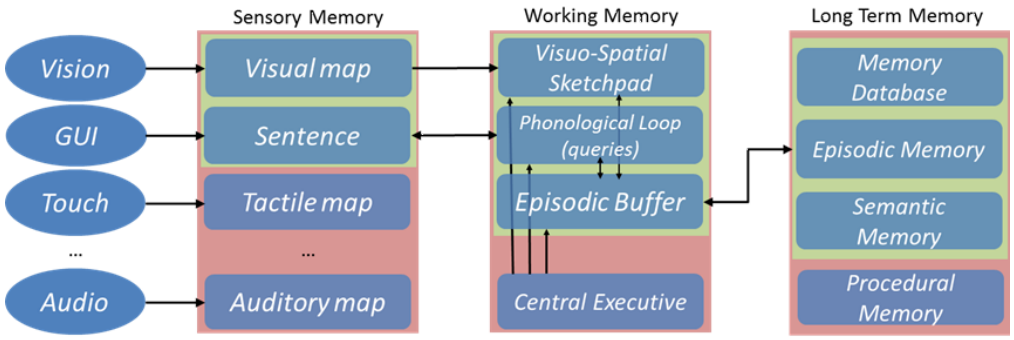


Figure 1. A graphical representation of the proposed memory architecture: parts in light green corresponds to currently implemented components.

main disadvantage of the approach is the difficulty of determining the *correctness* of a retrieved episode. The authors argue that this is due to the lack of a formal *context* definition. Nonetheless, our definition of EM is inspired by these design choices.

From the analysis of the literature, it emerges that two topics are fundamental to design a memory-inspired robot framework, namely a clear design of the architecture (including all its relevant components and interconnections) and an assessment about how contextual information impacts on memory items storage and retrieval.

3. System Architecture

3.1. *Connections with Memory Architectures*

The structure of the proposed memory-inspired architecture is outlined in Figure 1. The general design of the architecture is inspired by the multi store model by Atkinson and Shiffrin (1968) **updated with the WM model by Baddeley (2000)**. Each store can be further divided in subcomponents, according to the current understanding of memory organization in humans and other beings (Baddeley & Hitch, 1974; Wood et al., 2011).

We assume the presence of a number of sensory components feeding different parts of **the Sensory Memory**. Currently, our architecture supports visual maps (in the form of *bitmaps*, but other approaches may be used as well, for instance the framework by Antonelli et al. (2014)), and a simple mechanism to represent *questions* that can be posed to the system (as context-based cues), in a spirit similar to the work by Dayoub et al. (2010), as well as robot *answers* (as familiarity-based cues). Visual maps correspond to the basic representation used by the adopted vision algorithms. In principle, **Sensory Memory** can accommodate other sensory maps, such as tactile and auditory maps (Denei, Mastrogiovanni, & Cannata, 2015; Kallaluri, Even, Morales, Ishi, & Hagita, 2013).

In our current implementation, the visual map is manually segmented as a perceived scene from a continuous stream of visual feed, therefore the robot's and human's hand are not captured within the scene. The visual map is always transferred to be processed in **WM** within the **Visuo-Spatial Sketchpad**. Relevant changes in the perceived **visual feed** constitute ***scenes***. The identification of **scenes** is related to the formation of **EM** memory items **(called *episodes*)** inside **the Episodic Buffer**. This process is managed by a proper computational component **representing the Visuo-Spatial Sketchpad**, which we call Visual Stimuli Processor (**ViSor**). Inside **ViSor**, visual maps are processed using color feature extraction, GIST descriptors (Oliva & Torralba, 2006) and visual attention algorithms based on the work by Jeong, Ban, and Lee (2008), which result is used to feed **the Episodic Buffer**. **In other words, a**

specific image representing the changes within the environment is defined as a *scene*, and scenes are used to form an *episode*. Multiple episodes are encoded as collections of EM items. An *event* consists of several episodes sequentially ordered based on timestamps.

Each memory item is modeled as a collection of cue-value pairs, where cues correspond to features extracted from incoming images. On the one hand, the GIST descriptors algorithm allows us to extract shape features for scene-wide changes detection (i.e., at the global level) and for each detected entity (i.e., at the local level). On the other hand, visual attention includes a saliency detection algorithm, which allows us to localize each entity detected in the image. **Schillaci, Bodiroža, and Hafner (2013) provides an excellent analysis of the influence of saliency in a human-robot interaction domain.**

Once a scene has been captured and processed through the ViSor component, an episode is formed and consolidated into the LTM storage. Here, the scene is captured after the robot performs each pick and place movement. Saliency detection has been considered in the proposed framework given the widespread belief that it plays a central role in the human memory consolidation process and episodic segmentation (Jeong, Arie, Lee, & Tani, 2011; Kaster & Ungerleider, 2000; Posner & Petersen, 1990, 2012).

Currently, only EM and SM have been implemented within LTM, and cue-value pairs in LTM are represented using a relational database. Relevant results (i.e., episodes or SM items) temporarily stored in the Episodic Buffer are compared with the Memory Database, which keeps track of familiar SM items and episodes, and consolidated when either they are not familiar or not listed in the database. The bidirectional arrow connecting **the Episodic Buffer** (specifically, the **ViSor module**) and LTM (specifically, the **Memory Database**) in Figure 1 represents the ability to consolidate and recall memory items.

As postulated by Eichenbaum and Howard (2001) and Tulving (2001, 2002), human memory is characterized by the property of undergoing a continuous, subjective **rehearsal** and active modification, which is how we actually re-experience past events during **memory** recollection. Even though a precise understanding of this phenomenon is still subject to research efforts, our framework aims at mimicking this feature of the human memory, which **without** any doubt plays a central role in everyday behavior.

From a computational point of view, the choice of which information to store inside LTM as a collection of cue-value pairs is an important design parameter for the whole architecture. It is necessary to find a trade-off between the proper selection of image features (i.e., to be stored as cues) best discriminating among different episodes (i.e., having well-separable value spaces), and the need for storing the minimum amount of information (i.e., the size of the **LTM storage**) given the continuous nature of the knowledge acquisition process. **Two main ideas are considered for this matter: (i) considering that every computer science problem is related to the famous "time-space trade-off" regardless of the capacities and availability of computer memories in the present and the future; and (ii) anticipating the increasing needs of storing more information in a single memory item in the future (compared with our currently implemented color and shape information). In particular, although the capacity of computer memory is considered abundant and inexpensive nowadays, having succinct representation of memory items allows for more efficient memory retrieval processes, which eventually allows more memory items to be stored, as well as boosts runtime performance.**

A similar information flow can be determined when the user asks the robot to **recall** previously acquired memory items. Currently, this is done using the cue-value pair based

formalism that is mapped to specific queries in **the Phonological Loop** to be submitted to LTM. The same cue-value based formalism is used to present to a human user the robot accounts related to what has been actually recalled[1].

It is noteworthy that these two information flows are not to be considered in strict alternative. In fact, it is possible to pose questions while the robot is still acquiring new knowledge.

The ability to manage different parts of STM is due to our implementation of **Central Executive**. In human memory, **the Central Executive** is believed to be responsible for processing information originating from different sources, coordinating a number of otherwise passive subsystems, as well as performing selective attention and inhibition strategies (Baddeley, 1996, 1998; Collette & der Linden, 2002). In the current implementation of the architecture, **Central Executive** is designed as a computational process able to perform a number of tasks, as follows:

(1) Managing the encoding processes of **Episodic Buffer** to store relevant visual information computed by **Visuo-Spatial Sketchpad** (e.g., object shapes, colors or locations as perceived in a scene) in the form of cue-value pairs in such LTM components as EM and SM.
(2) Performing familiarity-based information retrieval, i.e., identify relevant cues, based on logical processes involving cue analysis and problem awareness (Mastrogiovanni, Scalmato, Sgorbissa, & Zaccaria, 2011; Mastrogiovanni & Sgorbissa, 2012).
(3) Executing recollection processes, i.e., recalling memory items from LTM using the results of the familiarity-based retrieval process.
(4) Supervising **the Phonological Loop** to analyze cue-value pairs based information related to recalled LTM memory items.

3.2. *Formal Definitions and their Meaning*

In this Section, we define the most important concepts of the proposed architecture, thereby defining the memory model upon which the framework is designed and implemented. We introduce first the notion of *memory item*. We will later use the definition of memory item to formally define elements in SM and EM.

Definition 1 (Memory Item)**:** *A Memory Item $i \in I$ is a set of n cue-value pairs, such as $i = \{(c_1, v_1), \ldots, (c_n, v_n)\}$.*

A memory item is a single element that can be used to represent any of the subcomponents of LTM, such as SM, EM or **Procedural Memory**. In this paper, we do not model **Procedural Memory**. However, it is noteworthy that we explicitly take into account the link between the knowledge represented in SM and EM (Squire & Kandel, 2000). As we discussed in Section 2.1, SM stores general-purpose knowledge about the environment in terms of concepts and their relationships (which are, in a sense, independent from the particular robot and therefore transferable to other robots), whereas EM represents robot experiences (in the form of episodes) anchored to a specific point in space and time (which is typically robot-dependent).

Definition 2 (Entity)**:** *An Entity ϵ is a grounded memory item $i_\epsilon \in E$, with $E \subset I$.*

Entities are a representation of objects in the environment, humans and other agents acting therein. Each entity is mapped to a set of grounded cue-value pairs, where the

[1]Current work is devoted to design and implement a speech-based dialog system grounded with respect to the cue-value pair based formalism.

semantics associated to cues globally define the entity as a type.

Definition 3 (Object)**:** *An Object n is a grounded memory item $i_n \in N$, with $N \subset I$, where i_n is defined in terms of three multi-valued cues, i.e.,* `name`*,* `shape` *and* `color`*, and by a number of Boolean cues, i.e.,* `graspable` *and* `manipulable`*.*

Each memory item corresponding to an object is characterized by specific values associated with its constituent cues. For instance, the `shape` cue can assume one of the values *cube*, *plane*, *disc*, *cylinder* and *sphere*, as well as *custom* for general shapes. As a consequence, a *bluebox* object is modeled as a specific collection of grounded cue-value pairs, as follows: $\{(\texttt{name}, bluebox), (\texttt{shape}, cube), (\texttt{color}, blue), (\texttt{graspable}, true), (\texttt{manipulable}, true)\}$.

Definition 4 (Location)**:** *A Location l is a grounded memory item $i_l \in L$, with $L \subset I$, where i_l is defined in terms of one numerical cue corresponding to a 3-element vector* `pos3d` *and one Boolean cue* `type`*.*

A memory item representing a location can refer to either an *absolute* or *relative* 3D position (expressed using the `type` and `pos3d` cues, respectively), whose semantics depends on the specific Cartesian frame with respect to which the location is expressed. For example, the description of the previously introduced *bluebox* object can be augmented with a description $\{(\texttt{pos3d}, (0.72, 0.13, -0.29)), (\texttt{type}, relative)\}$.

We also introduce a notion of time inspired by a simple linear time logics approach (Emerson & Halpern, 1986), as follows.

Definition 5 (Time Instant)**:** *A Time Instant t is a cue-value pair, with $t = (\texttt{time}, integer)$.*

Time instants **are represented in Unix epoch time, which are** positive *integer* numbers.

Definition 6 (Semantic Memory)**:** *A Semantic Memory SM is a collection of k grounded memory items $\{\mathrm{i}_1, \ldots, \mathrm{i}_k\}$, which can be divided into 5 disjoint sets, such that $SM = \{N, H, L, T, W\}$, where: N represents known (or previously identified) objects, H stores information about humans or other agents the robot interacts with, L is related to entities spatial information (locations), T represents entities temporal information (time instants), whereas W is an association between lexical knowledge and entities.*

We separately model N and H in order to account for inanimate objects and intentional agents, respectively. As previously noted, in this paper we focus on the set N and not on H, which is characterized by the appropriate knowledge to model the objects the robot interacts with.

The representation of object is first *consolidated* as an SM item whenever a *novel* object is detected by **the ViSor module**, through a process which resembles *habituation* (Squire & Kandel, 2000).

Definition 7 (**Episode and Scene**)**:** ***An Episode $\hat{\sigma}$ is a memory item succinctly representing the captured visual changes of the environment, which is a collection of b grounded memory items $\{i_{\hat{\sigma},1}, \ldots, i_{\hat{\sigma},b}\}$, which occur at a time instant t_σ. The visual change occurring at a time instant t_σ is defined as a scene, which is a sequence of visual feed $\{\sigma_1, \ldots, \sigma_b\}$.***

In particular, a scene is an instance of **a captured** image in the visual stream, **which is then later processed using the ViSor module and yields saliency informa-**

tion. The saliency information is further processed through a global and local processing module of color and shape for each detected object, which results are used to form an episode (an EM item). Episodes can employ both memory items related to objects represented therein as well as global descriptors of a scene, such as the number of objects, through the cue `count`. Two subsequent scenes are separated by a significant change in the image saliency level.

Definition 8 (Event Type)**:** *An Event Type ξ is a cue-value pair, with* $\xi = (\texttt{type}, active|passive)$.

Events are classified as being active or passive. An active event originates from one or more actions performed by the robot itself, whereas a passive event either corresponds to actions carried out by humans interacting with the robot or to something that simply happens in the robot workspace and is perceived in a scene. In this paper, we **consider both active and passive events. Although the robot witnesses events that are influenced by its own motions, it should be noted that since PM is not considered at the moment, active events will not influence conducted experiments.**

Definition 9 (Event)**:** *An Event η is a collection of s* ***episodes*** $\{\hat{\sigma}_{\eta,1}, \ldots, \hat{\sigma}_{\eta,s}\}$*, with associated type information.*

An event is defined by two corresponding initial and final scenes **(represented as episodes)**, namely $\hat{\sigma}_{\eta,1}$ and $\hat{\sigma}_{\eta,s}$, as well as by all intermediate scenes. In principle, any two events can be distinct, overlap, or one can include the other, thereby implementing the whole set of relationships between intervals defined by Allen (1983).

Definition 10 (Episodic Memory)**:** *An Episodic Memory EM is a collection of z events* $\{\eta_1, \ldots, \eta_z\}$.

The knowledge retrieval process is based on the notion of context.

Definition 11 (Context)**:** *A Context γ is a collection of any m cue-value pairs, with* $\gamma = \{(c_1, v_1), \ldots, (c_m, v_m)\}$.

Differently from memory items, contexts are not part of the set I, meaning that they do not necessarily correspond to definitions of entities, objects or locations. In our framework, contexts are used in the knowledge retrieval process to recall memory items stored in LTM. As it will be discussed in Section 3.3, humans interacting with the robot can pose a number of questions, which are formally encoded as contexts.

To this aim, cues can be classified as general-purpose and context-dependent, depending on their memory scope. For instance, cues may be appropriate to all the available memory components (e.g., SM and EM), or be related to one component exclusively (e.g., **SM** only).

Definition 12 (General-purpose Cue)**:** *A cue c is general-purpose if it refers to a memory item i that is not specific to any memory component.*

A context using a general-purpose cue may include, for instance, information related to both SM and EM.

Definition 13 (Context-dependent Cue)**:** *A cue c is context-dependent if it refers to a memory item i that is specific to a particular scene observed by the robot.*

As we discussed in Section 2.1, the visual stream in processed by **Visuo-Spatial**

Sketchpad (represented by the implemented ViSor module) to form episodes, which are consolidated in LTM as part of EM through **the Episodic Buffer**. Context-dependent cues are used to retrieve memory items stored in EM.

In a complete sensorimotor process, it is believed that the consolidation process involves SM, EM and **Procedural Memory**, as discussed by Tulving (1985) and Squire (2004).

3.3. *Knowledge Acquisition and Retrieval*

In the proposed framework, LTM is considered as a virtually infinite storage, where information in the form of memory items can be represented indefinitely through synaptic consolidation. On the one hand, SM is expected to store intrinsic properties of objects, which do not change over time, such as their `shape` and `color`. On the other hand, EM stores events described as a collection of **episodes**, each one representing intrinsic as well as extrinsic properties, which may change over scenes, such as the number of objects (i.e., the `count`), as well as their locations and mutual displacements (e.g., `leftMost`, `rightMost`, `front` and `back`).

As discussed in Section 3.2, a context is represented as a collection of cue-value pairs. When an event is **recalled** by retrieving the proper memory items, a context has the effect of filtering away irrelevant **episodes**, thereby limiting the overall number of matching **episodes**. As an example, let us assume to have presented a robot with a scene consisting of three objects, one of which is a *blue box*. The memory retrieval process may include a question like: *What do you know about a blue box when three objects have been presented?* The question is translated in a query defined as a simple cue-value pair (`shape`, *box*), whereas the context may be expressed as {(`color`, *blue*), (`count`, 3)}.

In order to implement the context-based knowledge acquisition and retrieval process, we use a Familiarity Filtering Index (FFI) as a part of **the implemented WM**. The index is meant at mimicking cue familiarity phenomena that can be observed in human cognition. The concept of *familiarity* in humans is exhibited through the ability to recognize an event or an object, even without knowing the details associated with the process leading to the storage of the corresponding memory items, as well as the relationships with other relevant elements, as discussed by Henson, Cansino, Herron, Robb, and Rugg (2003); Henson, Rugg, Shallice, Josephs, and Dolan (1999).

In Section 3.2, we defined an object as a set of cue-value pairs, two of which are `shape` and `color`, respectively. Although the corresponding values are expressed in symbolic form, they are associated with deterministic and statistical information **within WM**: `shape` with GIST descriptors, `color` with mean and variance of object information in the hue space. In the current implementation, shape and color are the features used to compute FFI.

Definition 14 (Familiarity Item)**:** *A Familiarity Item f is a cue-value pair, such as $f_s = (\texttt{shape}, GIST)$ or $f_c = (\texttt{color}, (mean(hue), var(hue)))$.*

Familiarity items define a set of cue-value pairs mapping symbolic representations of perceived objects to their counterparts in feature space.

Definition 15 (Familiarity Filtering Index)**:** *A Familiarity Filtering Index ϕ is a collection of $q + w$ familiarity items, such that $\phi = \{f_{s,1}, \ldots, f_{s,q}, f_{c,1}, \ldots, f_{c,w}\}$.*

FFI is used in both knowledge acquisition and retrieval. During knowledge acquisition, FFI is used to determine whether a detected object is *familiar* to the robot by checking its `shape` and `color` values in feature space. If the values were not present in FFI, then the object would be considered new, otherwise it would be considered familiar. In case

Data: A set $\{f_{c,1}, \ldots, f_{c,w}\}$ of w `color` features
Result: A symbolic *value* for the `color` cue
foreach *`color` feature* $f_{c,i}$ **do**
  $m_i \leftarrow mean(hue(f_{c,i}))$
  $v_i \leftarrow var(hue(f_{c,i}))$
  foreach *familiarity item* $f_{c,j} \in \phi$ **do**
    $m_j \leftarrow mean(hue(f_{c,j}))$
    $v_j \leftarrow var(hue(f_{c,j}))$
    $d_{i,j} \leftarrow dist((m_i, v_i), (m_j, v_j))$
  end
  $d^*_j = min(d_{i,j})$
  $f^*_{c,j} = arg_j(d^*_j)$
  if $d^*_j < \Omega_c$ **then**
    value $\leftarrow$ `color` associated with $f^*_{c,j}$
    $f^*_{c,j} \leftarrow$ update with $f_{c,i}$
  else
    value $\leftarrow$ ask the human for new `colour` name
    $\phi \leftarrow \phi \cup f_{c,i}$
  end
  return *value*
end

Algorithm 1: Knowledge acquisition: familiarity filtering for the `color` cue.

a familiar object was detected, nothing would be consolidated in SM. This mechanism resembles *habituation* (Squire & Kandel, 2000), a memory consolidation strategy widely employed by humans within implicit memory.

Algorithm 1 describes the familiarity filtering process associated with `color` cues during knowledge acquisition. Given **an episode** $\hat{\sigma}$, we assume that **Visuo-Spatial Sketchpad** produces a set of `color` features, each one corresponding to a detected object. The procedure loops on those features. For each feature, it computes mean and variance of the corresponding hue information. Then, for each familiarity item in ϕ, the probabilistic distance between the item and the feature is computed. To this aim, many probabilistic distance measures can be employed, e.g., the Mahalanobis distance is a commonly adopted one. If the smallest distance is below a given threshold Ω_c, which is experimentally tuned **at** 70%, then the feature is considered *familiar*, and the corresponding familiarity item `color` cue is retrieved. Given the association with the feature, the retrieved familiarity item is updated with the new information. It is noteworthy that this implicitly implements a clustering process as an interactive human-robot interaction process. Otherwise, the `color` feature is considered to be a new color, which is added to the set of familiarity items.

Algorithm 2 describes a similar procedure for the `shape` cue. As previously discussed, GIST descriptors are used. Differently from `color` information, here we use a simple Euclidean metric to compute the distance. Again, a threshold Ω_s is experimentally **setup at** 70% to discriminate between familiar and unfamiliar shapes.

The knowledge acquisition process is outlined in Algorithm 3. When **the WM module** detects significant changes in the visual stream, new memory items are consolidated within LTM, involving both EM and SM. Specifically, EM is augmented by novel **episodes**, whereas SM is updated with new **SM** items. The Algorithm assumes the availability of a set of `shape` and `color` cues, and full access to EM and SM. First, an empty **episode** $\hat{\sigma}_i$ is initialized. Then, Algorithm 1 and Algorithm 2 are sequentially called to retrieve symbolic information for `shape` and `color` from image features, respectively. For

Data: A set $\{f_{s,1}, \ldots, f_{s,q}\}$ of q `shape` features
Result: A symbolic *value* for the `shape` cue
foreach *`shape` feature* $f_{s,i}$ **do**
  $GIST_i \leftarrow$ GIST description of $f_{s,i}$
  foreach *familiarity item* $f_{s,j} \in \phi$ **do**
    $GIST_j \leftarrow$ GIST description of $f_{s,j}$
    $d_{i,j} \leftarrow dist(GIST_i, GIST_j)$
  end
  $d^*_j = min(d_{i,j})$
  $f^*_{c,j} = arg_j(d^*_j)$
  if $d^*_j < \Omega_s$ **then**
    value $\leftarrow$ `shape` associated with $f^*_{s,j}$
  else
    value $\leftarrow$ ask the human for new `shape` name
    $\phi \leftarrow \phi \cup f_{s,i}$
  end
  return *value*
end

Algorithm 2: Knowledge acquisition: familiarity filtering for the `shape` cue.

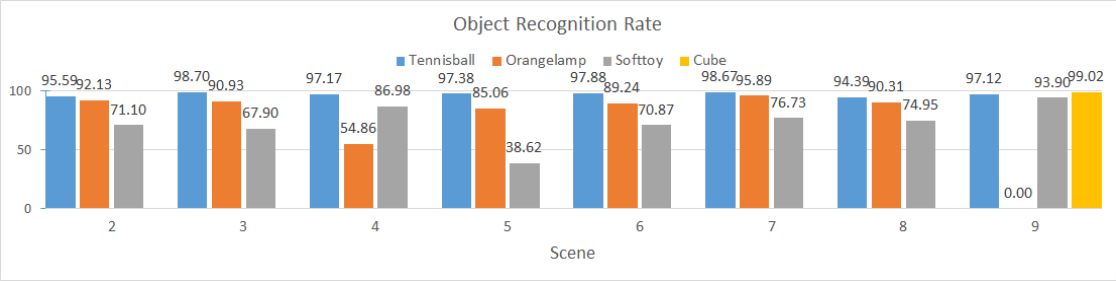


Figure 2. Object recognition rate based on the available knowledge at scene capturing time.

each detected object, `shape` and `color` information is consolidated **as** SM. Finally, the current **episode** $\hat{\sigma}_i$ is built encoding `time`, `name`, `shape`, `color` (and other cues, e.g., `graspable`, `manipulable` or `pos3d`) for each object therein. After the current **episode** $\hat{\sigma}_i$ is **formed, then it is added to LTM storage**.

As an example, let us consider Figure 3(a), where a first scene of a knowledge acquisition process is shown. **The robot performs pick and place movements on several objects available in a particular workspace configuration until the final configuration shown in Figure 3(i) is obtained. Successfully detected objects are enclosed with a bounding box, and subject to local feature extraction process. In scene 1,** proper bounding boxes are associated with detected objects, which are called `soft vinyl toy`, `tennis ball` and `orange lamp`. Since this is the first captured scene, none of the objects are familiar. Therefore, three color familiarity items are created, corresponding to colors `black/yellow`, `light green` and `orange`, and three shape familiarity items are created, labeled as `custom`, `ball`, and `dome`. As a consequence, nine cue-value pairs are added to **LTM storage** (i.e., three **SM items** per object), for `name`, `shape` and `color`, respectively. Furthermore, one scene is stored **in the LTM** as a collection of cue-value pairs for each object **(represented by an episode)**, including also cues that are specific to the scene.

After objects configuration is changed from Scene 1 to Scene 2 (see Figure 3(b)), another scene is captured and processed by **the ViSor module**. **For the statistics, Figure 2 shows the average recognition value from both color and shape,**

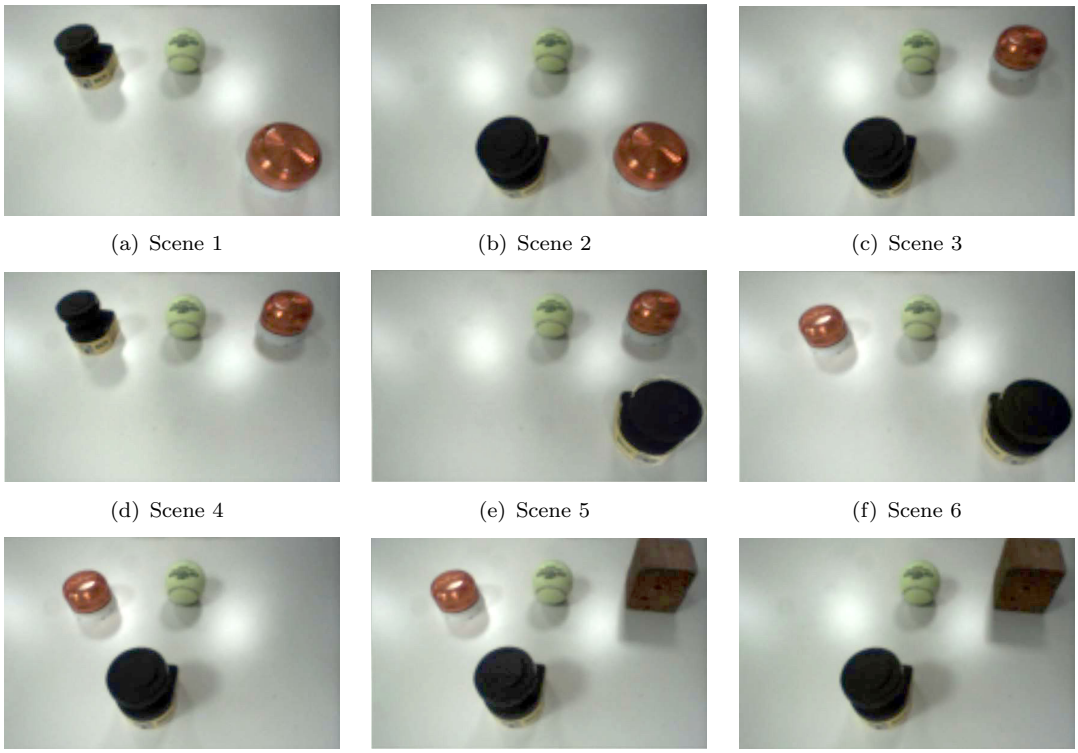
Data: A set $\{(f_s, f_c)_1, \ldots, (f_s, f_c)_o\}$ of o detected objects, represented using `shape` and `color` features, current EM, current SM
Result: An event η to be encoded in EM, a collection of grounded memory items i to be stored in SM

$\hat{\sigma}_i = \emptyset$
$\{f_{s,1}, \ldots, f_{s,o}\} \leftarrow$ Extract from $\{(f_s, f_c)_1, \ldots, (f_s, f_c)_o\}$
$\{(\texttt{shape}, s_1)_1, \ldots, (\texttt{shape}, s_o)_o\} \leftarrow$ Algorithm 1 on $\{f_{s,1}, \ldots, f_{s,o}\}$
$\{f_{c,1}, \ldots, f_{c,o}\} \leftarrow$ Extract from $\{(f_s, f_c)_1, \ldots, (f_s, f_c)_o\}$
$\{(\texttt{color}, c_1)_1, \ldots, (\texttt{color}, c_o)_o\} \leftarrow$ Algorithm 2 on $\{f_{c,1}, \ldots, f_{c,o}\}$
foreach *new detected object* j **do**
  $n \leftarrow$ ask the human for a new object `name`
  $n_j \leftarrow (\texttt{name}, n)$
  SM $\leftarrow$ SM $\cup\ n_j$
  SM $\leftarrow$ SM $\cup\ (\texttt{shape}, s_j)_j$
  SM $\leftarrow$ SM $\cup\ (\texttt{color}, c_j)_j$
end
$t \leftarrow$ current time
$t_{\sigma_i} \leftarrow (\texttt{time}, t)$
$\hat{\sigma}_i \leftarrow \hat{\sigma}_i \cup t_{\sigma_i}$
$s \leftarrow$ ask the human for the scene `name`
$s_{\sigma_i} \leftarrow (\texttt{name}, s)$
$\hat{\sigma}_i \leftarrow \hat{\sigma}_i \cup s_{\sigma_i}$
foreach *detected object* j **do**
  $\hat{\sigma}_i \leftarrow \hat{\sigma}_i \cup n_j$
  $\hat{\sigma}_i \leftarrow \hat{\sigma}_i \cup (\texttt{shape}, s_j)_j$
  $\hat{\sigma}_i \leftarrow \hat{\sigma}_i \cup (\texttt{colour}, c_j)_j$
  ...
end
$\eta \leftarrow$ Create from $\hat{\sigma}_i$
EM $\leftarrow$ EM $\cup\ \eta$

Algorithm 3: Knowledge acquisition process in scenes.

which determines whether an object is familiar. The value for each scene, starting from scene 2, is determined from the robot knowledge available during the time when that particular scene is captured. Now, although all the objects in the scene seem familiar, the statistics shows that `softtoy` in scene 2 has 67.9% **recognition rate. When the recognition value of an object is less than the threshold value, which is set at** 70%**, a new SM item is formed and consolidated. Here, to refer to them as the same object, we set the same name with different ID number as the human feedback, such as `softtoy-2` or `orangelamp-2`, considering that accuracy is not a major issue in our experiments.** Now, all the objects in the scene are familiar to the robot. However, since the objects have been displaced, a new scene representation (in the form of cue-value pairs) is generated, a new **episode** is created and consolidated in EM.

The same pattern applies to all the sequences of scenes in Figure 3 up to Scene 8, when a new object (a wooden cube) is introduced. Since the wooden cube is unfamiliar to the robot, new familiarity items for `shape` and `color` (corresponding to `cube` and `brown`, respectively) are created and stored **as SM items**. Proper information representing the new scene is consolidated **as an episode (EM item)**. Finally, in the last scene of Figure 3(i), the `orange lamp` is removed. **If we look at Figure 2, the recognition value of the cube is** 99.02% **with respect to the available knowledge at that time (i.e., compared with scene 8, since it is the only scene where a cube is**



(a) Scene 1 (b) Scene 2 (c) Scene 3

(d) Scene 4 (e) Scene 5 (f) Scene 6

(g) Scene 7 (h) Scene 8 (i) Scene 9

Figure 3. Workspace configurations exposed to Baxter.

detected). Also, since the `orangelamp` **is removed in scene 9, no** `orangelamp` **is detected (which corresponds to the recognition value of** 0% **in the Figure 2). Therefore,** three objects are detected in scene 9, which are all familiar to the robot. As a consequence, no new information is acquired and consolidated **as** SM, whereas relevant information about the scene is consolidated **as an episode**.

During knowledge retrieval, which in our scenario happens as part of human-robot interaction tasks, FFI is used to recognize whether the input provided by a human (in terms of cue-value pairs arranged in a context) elicits some familiarity with any items that have been consolidated at a previous stage.

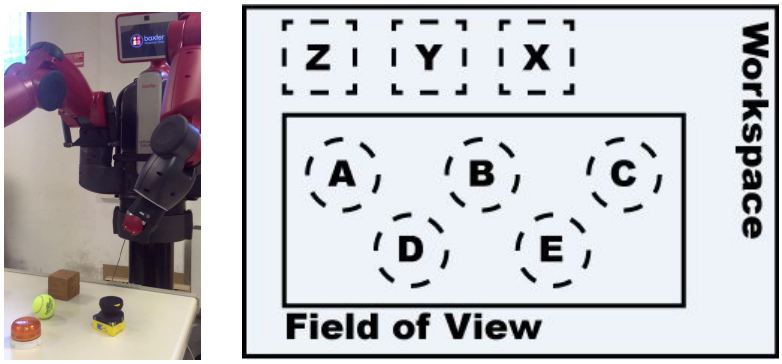
Algorithm 4 shows how familiarity filtering is used during knowledge retrieval. The Algorithm assumes a context is given in the form of a collection of cue-value pairs, and returns a set of **event-related information** Γ^*, initialized as empty. Ideally, the Algorithm iterates on all the **episodes** stored as part of events in **LTM (specifically EM)**. Each cue-value pair is compared with each memory item of each scene, in order to check for familiarity: if the distance between any of these two entities is below a given threshold Ω_f, we say that the corresponding cue is familiar to the robot since it resembles the scene itself. The corresponding familiar event is included in Γ^*.

Two remarks can be made: (i) the semantics associated with distance between cue-value pairs depends on the particular cue, thereby encompassing probabilistic (i.e., involving mean and variance) or deterministic (i.e., Euclidean) distance measures; (ii) different thresholds for familiarity may be used: for instance, one may argue that a single familiar cue does not make a corresponding scene familiar, as well as one single familiar scene does not make the corresponding event familiar.

It is now possible to discuss how the proposed architecture addresses the requirements posed beforehand. On the one hand, the robot is able to encode scenes, and consolidate the associated events into memory items that can be recalled afterwards. On the other

Data: A context γ of m cue-value pairs $\gamma = \{(c_1, v_1), \ldots, (c_m, v_m)\}$
Result: A list of e relevant **event-related information** Γ^*
$\Sigma = \emptyset$
$\Gamma^* = \emptyset$
foreach *event* $\eta_j \in EM$ **do**
  $\hat{\sigma}_j \leftarrow$ retrieve the **episode** corresponding to η_j
  $\Sigma \leftarrow \Sigma \cup \hat{\sigma}_j$
end
foreach *pair* $(c_k, v_k) \in \gamma$ **do**
  foreach ***episode*** $\hat{\sigma}_z \in \Sigma$ **do**
    foreach *memory item* $i_{\hat{\sigma},b} \in \hat{\sigma}_z$ **do**
      $d_{k,b} \leftarrow dist((c_k, v_k), i_{\hat{\sigma},b})$
      if $d_{k,b} < \Omega_f$ **then**
        $\eta_j \leftarrow$ retrieve the episode corresponding to $\hat{\sigma}_z$
        $\Gamma^* \leftarrow \Gamma^* \cup \eta_j$
      end
    end
  end
end
return Γ^*

Algorithm 4: Knowledge retrieval: familiarity filtering during recollection.



(a) Baxter (b) The workspace layout

Figure 4. (a) Our Baxter robot in front of the table. (b) Possible locations of objects on the table.

hand, memory item retrieval exploits contextual information to retrieve events stored in the **robot's** memory, on the basis of the robot *personal experience* in EM. As long as the robot keeps perceiving new scenes, its memory is expected to *grow*, but encoding only relevant events. It is noteworthy that a mechanism usually associated with memory storage, namely *forgetting*, is currently under investigation.

4. Recollection of Personally Experienced Events

4.1. *Scenario*

In order to validate our framework and the associated hypotheses, we developed **two integrated scenarios: the first for robot-environment interaction, the second for human-robot interaction. Our experiments consist of two phases: progressive knowledge acquisition and memory retrieval. During progressive knowledge acquisition, the robot-environment interaction involves a Baxter robot per-**

(a) Initial configuration (b) Final configuration

Figure 5. Configurations of objects on the table in front of the robot.

forming pick and place operations on objects in the workspace, and at the same time, passively observing the changes occurring in a scene (Figure 4(a)). Objects are moved within, inserted in and removed from the robot field of view (Figure 4(b)). The visual stream (a scene) is captured and manually segmented after the pick and place operation is performed, therefore both robot hand and human hand are not captured within the scene. After being captured, the scene is analyzed both globally (i.e., by means of the saliency of detected objects and areas of detected movements) and locally (i.e., determining such information as color, shape, position, and size). This yields an **episode** per scene, which is consolidated in LTM. **The second phase (memory retrieval) represents the human-robot interaction step, in which a human may pose questions to Baxter regarding its past experience.**

In the current procedure, we assume that: (i) no occlusion occurs between objects in the scene; and (ii) no forgetting mechanism is employed, i.e., the knowledge acquired by the robot develops *monotonically*. The system has been implemented using the ROS framework[2]. Each component is implemented as a collection of ROS nodes, whereas the communication between components is managed using ROS topics. The workstation used in our experiments is equipped with an Intel ®Core ™i7-4712MQ CPU, 2.30GHz clock frequency, and 16GB of RAM. The Baxter left hand camera is used as the main perception device. The query component is implemented as a simple user interface. Once provided by a human, the user interface sends the given cue and context data to a processing server, and then shows the result once it is available.

4.2. *Interaction Procedure*

As previously pointed out, we present progressive knowledge acquisition and memory retrieval as two separate stages. Nonetheless, it is possible to execute them concurrently: in this case memory retrieval operates on the currently available knowledge.

4.2.1. Knowledge Acquisition

Initially, three objects (i.e., a black and yellow soft vinyl toy, a tennis ball and an orange lamp) are located in the robot field of view, as shown in Figure 5(a).

To avoid occlusions within the robot field of view five distinct locations on the table, which are labeled as A to E, have been chosen. The locations are arranged in two rows (namely, `front` and `back`, **as depicted in Figure 4(b))**. Based on such a configuration,

[2]**The code is available at** `https://github.com/ferdianap/eris`

objects are initially placed in a subset of available locations, and it is possible to move them into the remaining available locations during the interaction. Outside the robot field of view, a separate area is used to temporarily store other objects to be later introduced in the scene. The corresponding locations are called X, Y and Z.

In our experiment, each object is presented to the robot beforehand, in order to bootstrap the representation using color and shape information. As a consequence, during scene 1, Baxter is able to directly recognize the detected objects using the previously introduced familiarity index. Furthermore, each object presented to the robot is provided with a label. This allows the robot to associate the statistical measurements of each object features with a symbol. In the case of the initial configuration, the label `soft toy` corresponds to the object characterized by the `black/yellow` color and a complex shape labeled `custom1`, the label `orange lamp` is associated with the object characterized by an `orange` color and a `dome` shape, whereas `tennis ball` has a `light green` color and a `sphere` shape. For a robot architecture explicitly adapting to human knowledge, it is noteworthy that such labeling process may require either a human supervision or a specifically designed learning approach.

In the initial configuration (Figure 5(a)), three objects are present in the scene: the `soft toy` in A, the `tennis ball` in B, and the `orange lamp` in E. Two more objects, namely a `wooden cube` and a `purple cube` are placed in Y and Z, respectively.

The robot acquires the scene and consolidates it within LTM. Then, a human performs a sequence of object displacements from the initial position to the final configuration shown in Figure 5(b). During each step, Baxter captures and assesses the scene by determining `position`, `color` and `shape` of each object. The sequence is as as follows:

(1) move `soft toy` from A to D
(2) move `orange lamp` from E to C
(3) move `soft toy` from D to A
(4) move `soft toy` from A to E
(5) move `orange lamp` from C to A
(6) move `soft toy` from E to D
(7) move `wooden cube` from Z to E
(8) move `orange lamp` from A to Z

4.2.2. Memory Retrieval Process

While memory items are stored, it is possible to query the robot memory about them. Currently, this is done using a simple user interface, through which a human can provide memory cues, their values and different contexts. On the basis of both cues and contextual information, the familiarity filter is used to retrieve relevant memories. Examples of questions that can be posed to the robot are **listed in Table 1**.

Table 2 translates the previous questions into formal notation. It is noteworthy that such labels as `leftmost`, `rightmost`, `orange`, etc., are associated with specific numerical ranges that refer to object parameters and the geometry of the scene.

4.3. *Results*

Based on the workspace layout depicted in Figure 4, captured scenes are shown in Figure 3 after each **pick and place** action. Scenes 1 to 7 represent configuration changes related to objects that are present in the robot field of view from the beginning, whereas scenes 8 and 9 represent the insertion and the removal of an object, respectively. As a consequence of this sequence, nine **episodes** and four **SM** memory items are stored in **LTM**.

Table 1. Results corresponding to given questions.

No.	Question	Answer
1	What orange objects do you know?	`orange lamp`
2	What color was the leftmost object, when three objects were presented?	`black/yellow` (scenes 1, 2, 3, 4), `light green` (scene 5), `orange` (scenes 6, 7)
3	How many objects were presented, when the orange lamp was the the rightmost?	3 (scenes 1, 2, 3, 4, 5)
4	When two objects were presented, what shape the leftmost object had?	none
5	When the soft toy was in the front, which object was the rightmost?	`orange lamp` (scenes 2, 3, 5), `soft toy` (scene 6), `tennis ball` (scene 7), `wooden cube` (scenes 8, 9)
6	Is the wooden cube ever at the back, when three objects were presented?	yes (scene 9)

Table 2. Input sets corresponding to possible questions.

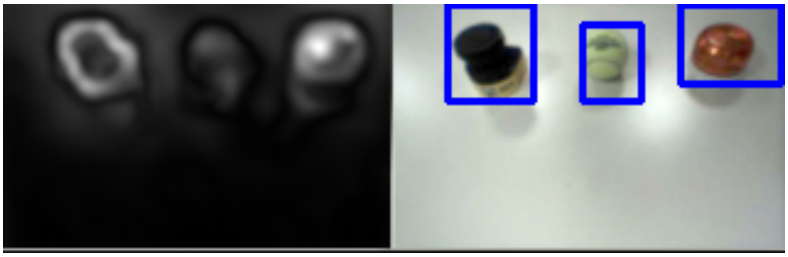
Question	Cue	Value	Position	Shape	Color	Count
1	color	`orange`	-	-	-	-
2	color	-	`leftmost`	-	-	3
3	count	-	-	`dome`	`orange`	-
4	shape	-	`leftmost`	-	-	2
5	position	`rightmost`	`front`	`custom1`	`black/yellow`	-
6	position	`back`	`back`	`cube`	`brown`	3

The first posed question has no context associated with the main cue. It refers to orange objects so far known to the robot. When the sequence of actions performed by the human is completed, only one orange object is known to the robot, namely the `orange lamp`, hence the result in Table 1. The second question is about the color of the `leftmost` object when only three objects are detected. The result shows that an object with a `black/yellow` color was detected from scene 1 to 4, a `light green` object in scene 5, and an `orange` object in scenes 6 and 7. The answer to the third question is that three objects were detected when the `orange lamp` was detected in the `rightmost` location. For the fourth question, none is returned as a result because it never happens that only two objects are detected. The fifth question is about the `rightmost` object that is detected when the `soft toy` is in the `front` row. Finally, the last posed question demonstrates the capability of the system to check the occurrence of a particular event. In this case, the expected values for the cue and the context are provided in input.

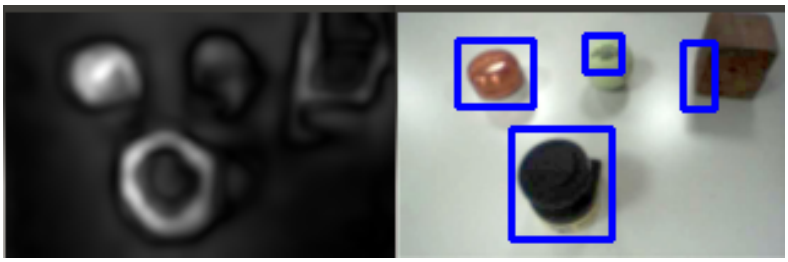
4.4. *Discussion*

Interconnectivity analysis. The purpose of the experiment is to analyze the interconnectivity between memory components, specifically as far as memory items are related to the physical properties of the objects contributing to a particular event (i.e., the SM-EM interconnectivity). In our experiments, since the robot acts only as a passive observer, we do not focus on the SM-PM and PM-EM interconnectivity.

Since the location of objects changes as a consequence of human actions, their geometrical features (e.g., `leftmost` or `rightmost`) and the rows they are located in (i.e., `front` and `back`) change as well. Such changes are enough for use to test the SM-EM interconnectivity.



(a) Correct objects detection in Scene 4



(b) False detection of wooden cube in Scene 8

(c) False detection of wooden cube in Scene 9

Figure 6. Successful and failure cases in captured scenes (left: saliency map, right: object detection).

The result of the first question shows that the system can **recall** memory items even in absence of contextual information. One of the main characteristics of contextual information is its inherent capability to bridge cues and experienced past events. When no specific context is given, **memory recollection only involves general knowledge stored in SM.** Since contextual information is always provided in the other questions, recollection in those cases affects both EM and SM.

The results for the second question are related to many scenes, i.e., the `black/yellow` color is detected in scenes 1 to 4, `light green` in scene 5, and `orange` in scenes 6 and 7. In this case, information stored in EM is fundamental to identify initial and final scenes (i.e., the episodes) related to changes in the provided cue. Therefore, an event in which the `black/yellow soft toy` is in the `leftmost` position is detected to occur from scene 1 to scene 4. A second event related to the `leftmost` position refers to the `tennis ball` in scene 5, whereas a similar event occurs in scenes 6 and 7 involving the `orange lamp`.

Results related to the third and the fourth question can be interpreted in a similar way.

No results are given for the fourth question, because the robot did not experience any events in which only two objects were presented.

In the result for the sixth question, both the cue and its value are provided, as well as the context. This shows the ability of the proposed architecture not only to retrieve, but also to check for the occurrence of an event, which is based on the value given with respect to the context.

Wrong interpretations in scenes. **Figure 6(b) and Figure 6(c) depict a case of**

wrong interpretation of a scene in the image processing phase, where each detected object should be enclosed in a blue rectangle. On the one hand, since saliency-based object detection is employed, several factors affect saliency results, including the mutual distance between objects, lighting conditions, and background colors. On the other hand, the object detection module is based on a number of assumptions and parameters, based on the fact that objects are assumed to correspond to the salient areas in the scene. One obvious drawback of the employed approach is that an object that has the same or a similar color of the background may be very difficult to identify.

As an example, **Figure 6(b) and Figure 6(c) show a false detection of the cube at the right hand side of the scene. In Figure 6(b), the detected region of the `tennisball` is smaller compared to the one in Figure 6(a) or Figure 6(c). We hypothesise these issues to be caused by a number of *a priori* defined and manually-tuned parameters, such as the amount and the size of saliency blob. If we have a closer look at the salience region of the cube in Figure 6(b) and Figure 6(c), we discover it is rather faint compared to the other detected objects within the scene. The fixed values for the parameters cause these issues, which explains that fact that the detected cube region is not as big as it should be. Nevertheless, since our algorithm is modular, modules employed in our architecture can be replaced with more accurate ones, to improve performance.**

As far as a memory-based architecture is concerned, such misinterpretations directly **map** to wrong memories, which may lead to inconsistencies among **episodes**. This is directly related to the well-known *symbol grounding problem* argued by Harnad (1990). Such problems lead to the need to address research challenges related to knowledge revision, rebuttal and forgetting.

5. Conclusions

In this paper, we present and discuss a novel cognitive architecture that allows robots to form *memories* related to their perceptions. We argue that two main characteristics are required to design such an architecture: on the one hand, we adopt a bio-inspired multi-store model that divides memory in components with different capabilities, and we try to mimic those capabilities; on the other hand, we integrate such a design with the use of contextual information, which is fundamental for an efficient memory formation and retrieval. Memories are represented using sets of cue-value pairs capturing relevant features of objects and the robot workspace.

As a case study scenario, we implemented a human-robot interaction where the robot (for the moment, only passively) observes what happens in the environment: humans can displace different objects thereby generating events for the robot to *remember*. Stored memories can be retrieved using a question-answering process.

Current work includes a more structured representation of objects, an improved definition of events, the integration of memories generated by robot's own actions, and mechanisms to allow a robot to revise and change its memories (including forgetting specific events).

References

Allen, C., & Bekoff, M. (1999). *Species of mind: the philosophy and biology of cognitive ethology.* Cambridge, MA, USA: MIT Press.

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, *26*(11), 832–843.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ, USA: Eribaum.

Antonelli, M., Gibaldi, A., Beuth, F., Duran, A., Canessa, A., Chessa, M., ... Sabatini, S. (2014, Dec). A hierarchical system for a distributed representation of the peripersonal space of a humanoid robot. *IEEE Transactions on Autonomous Mental Development*, *6*(4), 259–273.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: a proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: advances in research and theory* (Vol. 2, pp. 89–195). Academic Press.

Baddeley, A. (1996). Exploring the central executive. *The Quarterly Journal of Experimental Psychology: Section A*, *49*(1), 5–28.

Baddeley, A. (1998). The central executive: a concept and some misconceptions. *Journal of the International Neuropsychological Society*, *4*(5), 523–526.

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: advances in research and theory* (Vol. 8, pp. 47–89). New York, NY, USA: Academic Press.

Bellas, F., & Duro, R. J. (2004, August). Multilevel darwinist brain in robots: initial implementation. In *Proceedings of the first international conference on informatics in control, automation and robotics (icinco 2004)*. Setubal, Portugal.

Bellas, F., Faina, A., Varela, G., & Duro, R. J. (2010, July). A cognitive developmental robotics architecture for lifelong learning by evolution in real robots. In *Proceedings of the 2010 international joint conference on neural networks (ijcnn 2010)*. Barcelona, Spain.

Bullemer, P., Nissen, M. J., & Willingham, D. B. (1989). On the development of procedural knowledge. *Journal of Experiemental Psychology: Learning, Memory and Cognition*, *15*(6), 1047–1060.

Collette, F., & der Linden, M. V. (2002). Brain imaging of the central executive component of working memory. *Neuroscience & Biobehavioral Reviews*, *26*(2), 105–125.

Dayoub, F., Duckett, T., & Cielniak, G. (2010, October). Toward an object-based semantic memory for long-term operation of mobile service robots. In *Proceedings of the iros 2010 workshop on semantic mapping and autonomous knowledge acquisition, colocated with the 2010 ieee-rsj international conference on intelligent robots and systems (iros 2010)*. Taipei, Taiwan.

Denei, S., Mastrogiovanni, F., & Cannata, G. (2015). Towards the creation of tactile maps for robots and their use in robot contact motion control. *Robotics and Autonomous Systems*, *63*(3), 293–308.

Dodd, W. (2005). *The design of procedural, semantic, and episodic memory systems for a cognitive robot* (Unpublished doctoral dissertation). Vanderbilt University.

Dodd, W., & Gutierrez, R. (2005, August). The role of episodic memory and emotion in a cognitive robot. In *Proceedings of the 2005 ieee international workshop on robot and human interactive communication (ro-man 2005)*. Nashville, TN, USA.

Eichenbaum, H., & Howard, N. J. (2001). *From conditioning to conscious recollection: memory systems of the brain*. Oxford, USA: Oxford University Press.

Emerson, E. A., & Halpern, J. Y. (1986). "sometimes" and "not never" revisited: on branching versus linear time temporal logic. *Journal of the Association of Computing Machinery*, *33*(1), 151–178.

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: on land and underwater. *British Journal of Psychology*, *66*(3), 325–331.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, *42*, 335–346.

Henson, R., Cansino, S., Herron, J. E., Robb, W., & Rugg, M. D. (2003). A familiarity signal in human anterior medial temporal cortex? *Hippocampus*, *13*(2), 301–304.

Henson, R., Rugg, M. D., Shallice, T., Josephs, O., & Dolan, R. J. (1999). Recollection and familiarity in recognition memory: an event-related functional magnetic resonance imaging study. *The Journal of Neuroscience*, *19*(10), 3962–3972.

Jeong, S., Arie, H., Lee, M., & Tani, J. (2011). Neuro-Robotics study on integrative learning of

proactive visual attention and motor behaviors. *Cognitive Neurodynamics*, *6*(1), 43–59.

Jeong, S., Ban, S., & Lee, M. (2008). Stereo saliency map considering affective factors and selective motion analysis in a dynamic environment. *Neural Networks*, *21*, 1420–1430.

Jockel, S., Weser, M., Westhoff, D., & Zhang, J. (2008, July). Towards an episodic memory for cognitive robots. In *Proceedings of the 6th cognitive robotics workshop (cogrob 2008), co-located with the 18th european conference on artificial intelligence (ecai 2008).* Patras, Greece.

Jockel, S., Westhoff, D., & Zhang, J. (2007, December). Epirome: a novel framework to investigate high-level episodic robot memory. In *Proceedings of the 2007 ieee international conference on robotics and biomimetics (robio 2007).* Sanya, China.

Jones, D. M., Macken, W. J., & Nicholls, A. P. (2004). The phonological store of working memory: is it phonological and is it a store? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(3), 656–674.

Kallaluri, N., Even, J., Morales, Y., Ishi, C., & Hagita, N. (2013, May). Probabilistic approach for building auditory maps with a mobile microphone array. In *Proceedings of the 2013 ieee international conference on robotics and automation (icra 2013).* Karlsruhe, Germany.

Kasap, Z., & Magnenat-Thalmann, N. (2010, September). Towards episodic memory-based long-term affective interaction with a human-like robot. In *Proceedings of the 2010 ieee international symposium in robot and human interactive communication.* Viareggio, Italy.

Kaster, S., & Ungerleider, L. G. (2000). Mechanisms of visual attention in the human cortex. *Annual Review of Neuroscience*, *23*(1), 315–341.

Kuppuswamy, N. S., Cho, S. H., & Kim, J. H. (2006, October). A cognitive control architecture for an artificial creature using episodic memory. In *Proceedings of the sice-icase international joint conference 2006 (sice-iccas 2006).* Busan, Korea.

Mastrogiovanni, F., Scalmato, A., Sgorbissa, A., & Zaccaria, R. (2011). Problem awareness for skilled humanoid robots. *International Journal of Machine Consciousness*, *3*(1), 91–114.

Mastrogiovanni, F., & Sgorbissa, A. (2012). A biologically plausible, neural-inspired planning approach which does not solve "the gourd, the monkey, and the rice" puzzle. *Biologically Inspired Cognitive Architectures*, *2*, 77–87.

Mastrogiovanni, F., & Sgorbissa, A. (2013). A behavior sequencing and composition architecture based on ontologies for entertainment humanoid robots. *Robotics and Autonomous Systems*, *61*(2), 170–183.

Mehl, M. R., & Conner, T. S. (2012). *Handbook for research methods for studying daily life.* New York, MY, USA: The Guildford Press.

Morse, A. F., de Greeff, J., Belpaeme, T., & Cangelosi, A. (2010). Epigenetic robotics architecture (era). *IEEE Transactions on Autonomous Mental Development*, *2*(4), 325–339.

Nuxoll, A. M. (2007). *Enhancing intelligent agents with episodic memory* (Unpublished doctoral dissertation). University of Michigan.

Nuxoll, A. M., & Laird, J. E. (2004, July). A cognitive model of episodic memory integrated with a general cognitive architecture. In *Proceedings of the 2004 ieee international conference on cognitive modeling (iccm 2014).* Pittsburgh, Pennsylvania, USA.

Nuxoll, A. M., & Laird, J. E. (2012). Enhancing intelligent agents with episodic memory. *Cognitive Systems Research*, *17*, 34–48.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: the role of global image features in recognition. *Progress in Brain Research*, *155*, 23–36.

Peters II, R. A., Hambuchen, K. A., & Bodenheimer, R. E. (2009). The sensory ego-sphere: a mediating interface between sensors and cognition. *Autonomous Robots*, *26*, 1–19.

Phillips, J. L., & Noelle, D. C. (2005, August). A biologically inspired working memory framework for robots. In *Proceedings of the 2005 ieee international workshop on robot and human interactive communication (ro-man 2005).* Nashville, TN, USA.

Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience*, *13*, 25–42.

Posner, M. I., & Petersen, S. E. (2012). The attention system of the human brain: twenty years after. *Annual Review of Neuroscience*, *35*, 73–89.

Raaijmakers, J. G. W., & Shiffrin, R. M. (2003). Models versus descriptions: real differences and language differences. *Behavioral and Brain Sciences*, *26*(6), 753–754.

Ratanaswasd, P., Gordon, S., & Dodd, W. (2005, August). Cognitive control for robot task execution. In *Proceedings of the 2005 ieee international workshop on robot and human interactive communication (ro-man 2005).* Nashville, TN, USA.

Salgado, R., Bellas, F., Caamano, P., Santos-Diez, B., & Duro, R. J. (2012, May). A procedural long-term memory for cognitive robotics. In *Proceedings of the 2012 ieee workshop on evolving and adaptive intelligent systems (eais 2012).* Madrid, Spain.

Schillaci, G., Bodiroža, S., & Hafner, V. V. (2013). Evaluating the effect of saliency detection and attention manipulation in human-robot interaction. *International Journal of Social Robotics*, *5*(1), 139–152.

Shaw, J., & Tiggemann, M. (2004). Dieting and working memory: preoccupying cognitions and the role of the articulatory control process. *British Journal of Health Psychology*, *9*(2), 175–185.

Smith, D. M., & Mizumori, S. (2006). Hippocampal place cells, context, and episodic memory. *Hippocampus*, *16*(9), 716–729.

Smith, E. E., & Kosslyn, S. M. (2009). *Cognitive psychology: mind and brain.* Upper Saddle River, NJ, USA: Pearson Prentice Hall.

Spaniol, J., Madden, D. J., & Voss, A. (2006). A diffusion model analysis of adult age differences in episodic and semantic longterm memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(1), 101–117.

Squire, L. R. (2004). Memory systems of the brain: a brief history and current perspective. *Neurobiology of Learning and Memory*, *82*(3), 171–177.

Squire, L. R., & Kandel, E. R. (2000). *Memory: from mind to molecule.* New York, NY, USA: Henry Hold and Company.

Stachowicz, D., & Kruijff, G. M. (2012). Episodic-like memory for cognitive robots. *IEEE Transactions on Autonomous Mental Development*, *4*(1), 1–16.

Tecuci, D. G., & Porter, B. W. (2007, May). A generic memory module for events. In *Proceedings of the 20th international flairs conference (flairs-20).* Key West, FL, USA.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology*, *26*(1), 1–12.

Tulving, E. (2001). Episodic memory and common sense: how far apart? *Philosophical Transactions of the Royal Society of London*, *356*(1413), 1505–1515.

Tulving, E. (2002). Episodic memory: from mind to brain. *Annual Review of Psichology*, *53*(1), 1–25.

Wood, R., Baxter, P., & Belpaeme, T. (2011). A review of long term memory in natural and synthetic systems. *Adaptive Behavior*, *20*(2), 81–103.