Long-term Knowledge Acquisition in a Memory-based Epigenetic Robot Architecture for Verbal Interaction

Ferdian Pratama[1], Fulvio Mastrogiovanni[2], Sungmoon Jeong[1] and Nak Young Chong[1]

Abstract— **We present a robot cognitive framework based on (a) a memory-like architecture; and (b) the notion of "context". We posit that relying solely on machine learning techniques may not be the right approach for a long-term, continuous knowledge acquisition. Since we are interested in long-term human-robot interaction, we focus on a scenario where a robot "remembers" relevant events happening in the environment. By visually sensing its surroundings, the robot is expected to infer and remember snapshots of events, and recall specific past events based on inputs and contextual information from humans. Using a COTS vision frameworks for the experiment, we show that the robot is able to form "memories" and recall related events based on cues and the context given during the human-robot interaction process.**

I. Introduction

Natural context is involved both when humans interact with each other and when they interact with their environment. In humans, context processing is believed to occur in the hippocampus [2]. Specifically, it refers to those mechanisms to differentiate a particular situation from other situations so that the correct behavior or mnemonic output can be retrieved. Contextual information is considered to be significant and influences our daily behavior [17], [18].

In order to design a robot that proactively understands its environment and engages humans in long-term interaction tasks, the interconnectivity between the various modules within a memory-based, robot cognitive architecture must be enforced, specifically integrating the notion of *context*.

Exhibiting the ability to collect, assess and exploit knowledge progressively in daily interaction with humans is an ideal trait for robots whose behaviors is based on the developmental paradigm. To achieve such a goal, however, few points need to be addressed:

1) currently, no context-based integrated architectures optimized for long-term human-robot interaction tasks (such as those needed in service scenarios) is available to deal with the aforementioned capability,
2) in spite of recent research in memory-based architectures [3], [12], [13], which are based on single memory components, no holistic approach has been devised, which is a fundamental step to provide robots with the necessary flexibility to deal with contextual information.

An analysis of the literature shows that the focus is mainly on single memory components, such as the Working Memory (WM) [7], the Episodic Memory (EM) [3]–[6], [8]–[11], [14], [16] and the Procedural Memory (PM) [15].

In particular, Stachowicz and Kruijff provide an in-depth explanation of both design requirements and formal concepts needed to characterize EM and its storage structure [14]. However, the focus of their work is on the notion of *event*, its properties, and its use in such processes as event recognition, event nesting, and event types of complexity. 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 in [17], [18].

When an attempt is made to design a more comprehensive memory-based robot architecture [3], [12], [13], the goal is restricted to finding a solution to a very specific problem, rather than providing the robot with the capability to develop its own knowledge. Furthermore, neither the relationship between the different 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 describe how a memory-based architecture can benefit from contextual information in long-term human-robot interaction tasks. Specifically, we set-up an interaction scenario where (1) a human shows to a robot a number of scenes involving objects of different color, size and shape on a table, (2) the robot extracts the necessary information to store memory items, and (3) the robot can recollect specific information from its "memories" upon requests from a human.

In particular, we propose an architecture based on the following assumptions.

1) Avoiding the currently widespread mindset that developmental approaches are to be identified with machine learning techniques as the core framework, we posit that continuous knowledge acquisition allows for a progressive evolution of the stored knowledge and its representation, which is based on a continuous interaction with humans.
2) Inspired by state-of-the-art studies in Developmental Psychology [2], [17]–[19], [22]–[24], we argue that an explicit addressing of the role of memory in human-robot interaction processes is crucial in robot knowledge development.

The contribution of the paper is two-fold: on the one hand, we demonstrate the robot's ability to recollect memory elements stored as a result of gaining personal experience, on the basis of specific cues provided by a human; on the

[1] Ferdian Pratama, Sungmoon Jeong, and Nak Young Chong are with the School of Information Science, Japan Advanced Institute of Science and Technology, Ishikawa, Japan {ferdian, jeongsm, nakyoung}@jaist.ac.jp

[2] Fulvio Mastrogiovanni is with the Department of Informatics, Bioengineering, Robotics and Systems Engineering (DIBRIS), University of Genoa, Italy fulvio.mastrogiovanni@unige.it

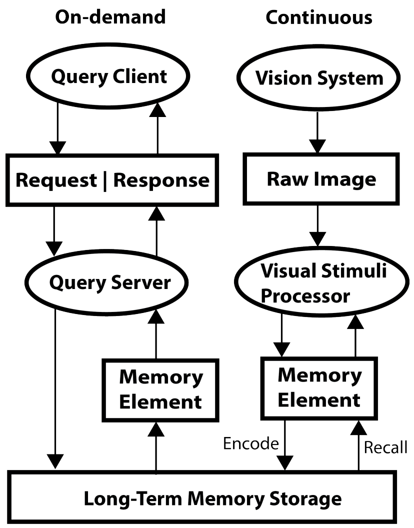


Fig. 1: A graphical representation of the system architecture

other hand, we show that contextual information (in the form of specific memory cues), is fundamental for the retrieval process.

The paper is organized as follows. Section II describes the main concept of the approach. Section III elaborates the conducted experiments with a specific, real-world scenario for the application domain. A discussion and the conclusion sections follow.

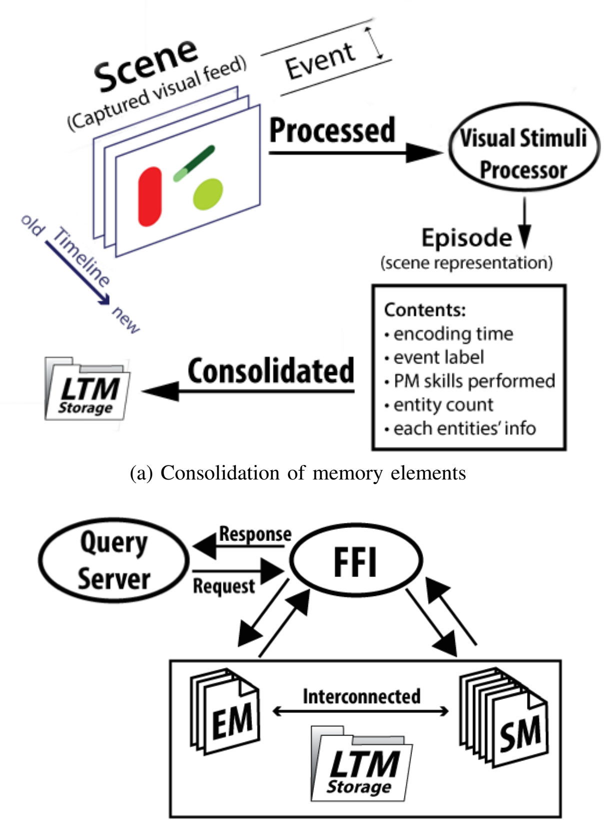
II. System Architecture

The architecture can be represented as a collection of information-exchanging modules. Figure 1 shows the main modules of the architecture. It is based on a client-server architecture and a message passing mechanism. Ovals represent *active* elements (i.e., processing nodes), boxes represent *passive* elements (i.e., messages being exchanged by the modules and the Long-Term Memory storage), whereas arrows represent the information flow within the system.

As it is shown in the right hand side of the diagram, a vision system is responsible for channeling the visual feed of raw images into the Visual Stimuli Processor (VSP). This is where the image processing algorithms infer useful information about the captured scene. The bidirectional arrows connecting the Long-Term Memory (LTM) storage and VSP represent the ability to recall and consolidate any particular memories. The visual feed has a continuous nature. The visual stream is processed and the result (in the form of specific image features) is compared to memory elements that are stored in LTM already. This is important for the whole architecture, due to the fact that storing dynamically growing knowledge in form of image features is much more efficient than storing a continuous visual stream.

The left hand side of the diagram in Figure 1 represents the interaction pipeline using a Human-Computer Interface (HCI) when the robot is required to provide assertions about its stored knowledge, which may occur at any time.

In the current system implementation, humans are able to pose questions regarding any particular events that has been previously experienced by the robot. The request is retrieved from LTM and sent back to the Query Server, which is responsible for retrieving any relevant information inferred from the request. An example of this process is discussed in Section II-B.



(a) Consolidation of memory elements

(b) Retrieval of memory elements

Fig. 2: A more detailed memory processing schematic

A. *Vision-based memory elements consolidation*

Figure 2a depicts the memory consolidation process. Here, visual inputs are processed using color feature extraction, gist [25] and visual attention based on the work described in [26], which are included in VSP. On the one hand, the gist algorithm allows us to extract shape features for scene-wide changes detection (i.e., globally) and for each detected entity (i.e., locally). On the other hand, the visual attention includes a saliency detection algorithm, which allows us to localize each entity detected in the image. Scenes are consolidated based on the absence of movements and, subsequently, the presence of saliency changes. Saliency detection has been considered in the architecture given the widespread belief that it plays a central role in human memory consolidation process and experience segmentation [31], [32]. The study by Jeong *et. al.* [30] provides an evaluation of the relationship between visual attention and motor action from a cognitive neurodynamics perspective.

As it has been pointed out in [22]–[24], humans have the ability to "mentally travel through time" to re-experience their past during event-recollections. Even though a precise understanding of this phenomenon is still subject to significant research efforts, our framework aims at mimicking this feature of the human mind, which undoubtedly plays a central role in everyday purposive behavior. Specifically, only snapshots of the visual feed (the *scenes*) are processed and representation results (the *episodes*) are generated and consolidated, instead of remembering the whole stream of events in a continuous fashion.

The memory model is divided into two main modules, namely the Working Memory (WM) and the LTM. WM is designed on the basis of the Baddeley updated model [19], which includes the Episodic Buffer (EB) component, in addition to the Central Executive (CE), Phonological Loop (PL), and Visuospatial Sketchpad (VSSP) components. LTM is organized as the interplay among three distinct sub-modules, namely the Semantic Memory (SM), Episodic Memory (EM) and the Procedural Memory (PM) components. Their role is described in the following paragraphs. CE consolidates memory elements in LTM, by considering the relation between contextual information in EM, general knowledge in SM, and movements performed in PM, thereby yielding interconnections between EM, SM, and PM.

In human memory, CE is responsible for processing information originating from different sources, coordinating a number of passive subsystems, as well as performing selective attention and inhibition strategies [27]–[29]. Here, we model CE to be able to perform a number of tasks, as follows.

- Explicitly managing memory encoding and decoding processes in such LTM components as EM, PM, and SM, specifically using contextual information.
- Exhibiting familiarity-like information retrieval, i.e., how to identify cues to be used, based on logical processes involving cue analysis and problem awareness [20], [21].
- Manifesting recollection behaviors, i.e., recalling LTM memory elements from the results of familiarity retrieval processes if they match the desired retrieval cues.
- Supervising the PL component (i.e., by analyzing verbal information related to recalled LTM memory items), the VSSP component (related to visual information, i.e., object shapes, colors or locations as perceived in a scene), as well as the interconnection between the two components through the EB.

We can now define the main concepts of the proposed architecture.

Definition 1 (Scene): A scene is a representation of the changes occurring in an input visual stream.

A scene represents the occurrence of an event at a particular time. In short, a *scene* is an event marker, such that an *event* may be bounded by any two arbitrary distinct scenes. This is in contrast with the definition proposed in [14], where a scene corresponds to atomic or complex events. Scenes captured as part of the visual stream are represented as an *episode*, and stored in the EM.

Definition 2 (Factor): A factor $f \in F$ is a single element that forms a SM and PM memory item, where $F = SM \cup PM$. A factor consists of n cue-value pairs, such as $f = \{(r_1, v_1), \ldots, (r_n, v_n)\}$.

SM is represented by five factors of knowledge for a robot involved in long-term human-robot interaction scenarios, namely, entity, person, location, time, and lexical information. Here, we focus on the entity type only. EM corresponds to events directly *experienced* by the robot, represented as *episodes*, which contain information about visually detected entities. PM corresponds to motor skills that are available for the robot to perform. From the WM perspective, LTM is only considered a (possibly complex) memory storage, such that the consolidation and retrieval processes are arranged and managed by CE.

In a complete sensori-motor process, it is believed that the consolidation process involves SM, EM and PM [33], [34]. Whilst SM refers to knowledge about perceived entities, EM is related to scenes, in the sense of Definition 1. In the current set-up, we focus solely on the contribution provided by SM and EM (i.e., only sensory information is used, which does not depend on robot motion processes).

B. Memory elements retrieval

Figure 2b shows a representation of the retrieval process. From Definition 2, we can clearly define the notion of *context*.

Definition 3 (Context): A context c is made up of cue-value pairs corresponding to a particular factor f, and defined as $c = \{(r_{f_1}, v_{f_1}), \ldots, (r_{f_n}, v_{f_n})\}$, where n is the number of desired contextual elements provided by a human during the interaction with the robot.

According to our model, a context can be represented using the previously introduced knowledge elements, namely entity, location, person, time and lexical information. When an event is recollected by retrieving the proper memory elements, a context has the effect of filtering away irrelevant scenes, thereby limiting the number of scenes matching with respect to the context itself. The readers are advised to refer to [1] for more details about the formal definitions of the proposed architecture.

As an example, let us assume to have presented the robot with a scene with three entities, one of which is a blue box. The memory retrieval process may include a question like: “What do you know about a blue box entity when there were three entities presented?”. The question would be translated in a query formally defined as $(cue = shape, value = box)$, whereas the context may be expressed as $(color = blue, count = 3)$.

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 elements, as well as the relationships with other relevant elements [35], [36]. In order to mimic such a capability, we developed a Familiarity Filtering Index (FFI) module, which is used to assess which memory elements can be considered relevant for the memory retrieval process. Specifically, each memory element is indexed with predefined references which represent the whole contents, and the filtering process during memory retrieval is based on the index. The module is part of our current CE implementation, and from a computational perspective, FFI significantly improves the overall system performance.

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 elements that can be recalled afterwards. On the other hand, memory recall exploits contextual information to retrieve events stored in the robot memory, on the basis of the robot *personal experience*. 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 not considered. In order to validate our architecture and the associated hypotheses, we performed tests in a human-robot interaction scenario, where a robot has to observe changes in a scene due to a human operating on a number of objects, which are located on a table. This set-up is described in the next Section.

III. The Target Human-Robot Interaction Scenario

A. Demo scenario

For a better understanding about how the proposed architecture works, we set-up a scenario where a human operates on objects, which are located on a table, and shows these operations to a robot. Objects with different colors and shapes are inserted and removed from the scene. In this scenario, the robot is just a passive observer. The robot perceives the scene using vision. A visual stream is continuously acquired by the robot as long as a human operates on the scene. Within the robot Field of View (FOV), actions performed by humans are visually spotted by the robot through saliency information.

As the visual stream is active, the system processes the input, infers useful information about the detected entities (e.g., color, shape, position, size) using the image processing module, and consolidates them into LTM. In the experiment, when a human replaces one object with another one, a scene change occurs from the perspective of the robot. As a consequence, a new memory element (related to the new spotted entity) is consolidated inside LTM. Actions performed by the human in the experiment are aimed at addressing different memory modules. Specifically, SM is emphasized whenever a novel entity is detected, EM is related to scenes as a whole, whereas PM is addressed during the event occurrence.

The following assumptions are posed: (i) no occlusion is present involving entities in the scene; and (ii) no forgetting mechanism is employed, which means that the knowledge gained by the robot develops *monotonically*.

B. Experimental procedure

The experimental procedure consists of two phases: knowledge acquisition and memory retrieval process.

Knowledge acquisition. Initially, two entities (namely a red can and a green marker pen) are presented in the visible part of the robot workspace (Figure 3a). The robot acquires the scene and consolidates it within LTM. Then, the human presents a novel entity (i.e., a red marker pen, Figure 3b) to the robot. Afterwards, the human replaces one entity (the red pen) with a novel one (a tennis ball, Figure 3c). The exact sequence is as follows:

TABLE I: Input sets for the experiment

			Context			
No.	Cue	Value	Pos	Shape	Color	Count
1	Color	Red	-	-	-	-
2	Color	Red	-	-	-	3
3	Pos	LeftMost	-	-	Green	3
4	Shape	Ball	RightMost	Ball	-	3
5	Shape	Ball	RightMost	Box	-	-

1) ⟨robot⟩ assesses the initial scene with a ⟨red can⟩ and a ⟨green pen⟩
2) ⟨human⟩ puts a ⟨red pen⟩ on the table
3) ⟨robot⟩ assesses the scene
4) ⟨human⟩ takes away the ⟨red pen⟩
5) ⟨robot⟩ assesses the scene
6) ⟨human⟩ puts a ⟨tennis ball⟩ on the table
7) ⟨robot⟩ assesses the scene

During each scene assessment step, the robot remembers the position, color and shape features for each entity.

Memory Retrieval Process. Using a user interface, the human inserts cues, their value and several contexts, and the system retrieves any available data based on both cues and contextual information. In the performed experiment, the following questions are posed to the robot.

1) What entities do you know, which are red?
2) Which entities do you know, which are red, when three entities were present?
3) Which green entity was the leftmost one, when three entities were present?
4) Was the rightmost entity a ball, when three entities were present?
5) Was the rightmost entity a ball, when a box was present?

The questions can be formally translated into a set of contexts, as shown in Table I. It is noteworthy that we artificiously distinguish between Knowledge Acquisition and Memory Retrieval Process. In principle, questions can be posed at any time during the experiment.

Since no a priori knowledge is considered, LTM is initially empty. As the robot experiences and consolidates new scenes, the persistent characteristic of LTM allows it to progressively gather knowledge from its personal experience.

IV. Experimental Evaluation

A. Experimental set-up

The system has been implemented using the ROS framework. Specifically, each module described in Section II has been implemented as a separate ROS node, whereas communication between modules is managed using ROS topics. The experiment is carried out on a Workstation equipped with Intel© Core© i7 CPU 960, 3.20Hz clock frequency

TABLE II: Results for each of the input set

No.	Result
1	Red can, red marker pen
2	Scene 2: red can, red marker pen Scene 3: red can
3	Scene 2: green marker pen Scene 3: green marker pen
4	Yes, in scene 3
5	No

and 12GB of RAM. The visual stream is obtained using a standard USB camera.

As briefly mentioned in the previous Section, visual processing covers both global and local analysis, in terms of shape and color information. The visual attention module includes object detection supported by saliency analysis, which is performed before local analysis. Shape analysis and the localization of each detected entity are covered by the gist algorithm [25] and the visual attention algorithm [26], respectively. Global analysis determines the occurrence of scene changes in terms of color and shape. Local analysis determines both the color and shape of each detected entity in a statistical measure within a particular scene. The query client node is implemented as a HCI. It processes given cue and context, and shows the proper result.

B. Experimental results

Figure 3 depicts the entity detection process based on saliency maps, as well as the corresponding captured scene.

As a result of the experiment, four SM and three EM episodes are maintained after Knowledge Acquisition. The EM and SM data correspond to each scene captured and detected entity, respectively. Instead of four, only three EM memory elements are consolidated. This is due to the fact that during step 5 in Knowledge Acquisition, the system assesses the scene and concludes that there are no differences if compared with the scene consolidated during step 1. Since the two scenes are exactly the same, in order to avoid memory duplication, we do not consider two scenes that are exactly the same but are characterized by a different time stamp. It is noteworthy that this case is very unlikely to happen in the real world.

The first question generally asks about red entities. Red entities that are known to the robot are two, namely a red can and a red marker pen. The can is always present in the robot FOV, whereas the marker pen has been detected in scene 2 (Figure 3). The second is more specific, in that it asks the robot to recall red objects detected only when three different objects were present in the scene. Therefore, scene 1 is not considered. Consistently, two objects are recalled from scene 2 and one object from scene 3. The third question is related to the qualitative position of the green entity, but only when other two entities were present. Again, scene 1 is not considered, whereas scene 2 and scene 3 are used to recall a green marker pen. It is noteworthy that in scene 3, two green entities are present, namely the marker and the ball, but the marker is the leftmost one. As a result from the fourth question, scene 3 is used to recall that a ball was the rightmost object. Finally, a ball is never the rightmost entity, when scenes include a box. Table II shows a summary of these results. It is noteworthy that some information is omitted, such as all the features of the retrieved objects, e.g., color, shape, etc.

C. Interconnectivity analysis

Despite the simplified experiment, we can assess the relationships between contextual information and the memory structure entailed by SM and EM, specifically insofar it impacts on the continuous knowledge acquisition process. Let us consider the results of each question. The result of question 1 shows that no given context is related to all the red entities that are known to SM. This is due to the characteristics of the context that bridges cues and experienced events. Without a specific context, this information becomes general knowledge that is stored as memory elements in SM.

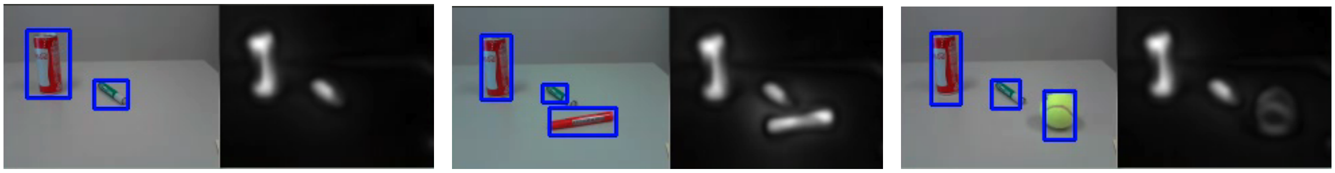
Considering the result of question 2, it is recalled that the red can is present both in scene 2 and scene 3, whereas the red pen is present in scene 2 only. This means that the red can is detected inbetween scene 2 and scene 3, subject to the contextual information related to the fact that three entities are present. The results showcase the robot ability to remember any specific entities during past experiences.

Similarly to the previous case, results of question 3 show that the green pen is detected both in scenes 2 and 3. However, since color and position information are stored as sets of $(mean, var)$ couples in hue space and (x, y) coordinates in the image plane, respectively, it is necessary to provide such information with labels to use a simplified grammar when posing questions. For instance, the *LeftMost* tag is used to indicate the entity characterized by the lower x value, whereas *green* is associated with a specific range of $(mean, var)$ data in hue space.

Results from questions 4 and 5 emphasize the ability of the proposed architecture to assert facts provided that a given context and cue match specific values, if previously experienced by the robot. In particular, a *shape* attribute is given as both context and cue (respectively, associated with *box* and *ball*). In this case, the robot yields a Boolean response. Furthermore, since shapes are labelled as statistical measures of shape feature vectors, they are compared with existing memory elements, specifically using Euclidean distance in the image plane. It is noteworthy that such a labelling process may require either an *a priori* human supervision, or a specifically designed learning approach.

V. Conclusion

In this paper, we present a conceptual design for a long-term knowledge acquisition framework using contextual information within a memory-based architecture. The framework is inspired by current studies in developmental psychology, and adopts a biologically-inspired image processing algorithm. The paper emphasizes the advantages of contextual information through a descriptive, representative experimental scenario. Experimental results also corroborate



(a) Scene 1, captured during *knowledge acquisition* step 1 and 5

(b) Scene 2, captured during *knowledge acquisition* step 3

(c) Scene 3, captured during *knowledge acquisition* step 7

Fig. 3: (Left) Raw Feed, (Right) Saliency Map of each scene

the interconnectivity of each component of the architecture. An analysis of the psychological aspects of memories, the response retrieval of personal past events based on the context and how context may influence the interaction with humans, has been provided. On the basis of these premises, current work is aimed at implementing the architecture (specifically, the sensori-motor part, which is not considered in this paper) on a Baxter dual-arm manipulator as well as integrating a speech-based interface.

References

[1] F. Pratama, F. Mastrogiovanni and N. Y. Chong, "An Integrated Epigenetic Robot Architecture via Context-influenced Long-Term Memory," in *Proc. IEEE Int. Conf. Developmental and Learning and on Epigenetic Robotics. (ICDL-EPIROB)*, pp. 110–116, 2014.

[2] D. M Smith and S. JY Mizumori, "Hippocampal place cells, context, and episodic memory," *Hippocampus*, vol. 16, no. 9, pp. 716–729, 2006.

[3] A. Nuxoll and J. E. Laird, "A cognitive model of episodic memory integrated with a general cognitive architecture," in *Proc. Int. Conf. Cog. Model. (ICCM)*, pp. 220–225, 2004.

[4] A. M. Nuxoll, "Enhancing intelligent agents with episodic memory," Ph.D. dissertation, University of Michigan, 2007.

[5] Nuxoll, Andrew M. and Laird, John E., "Enhancing Intelligent Agents with Episodic Memory", *Cogn. Syst. Res.*, vol 17–18, pp 34–48, 2012.

[6] W. Dodd and R. Gutierrez, "The role of episodic memory and emotion in a cognitive robot," in *Proc. IEEE Int. Workshop Robot Hum. Commun. (ROMAN)*, pp. 692–697, 2005.

[7] J. L. Phillips and D. C. Noelle, "A biologically inspired working memory framework for robots," in *Proc. IEEE Int. Workshop Robot Hum. Commun. (ROMAN)*, pp. 599–604, 2005.

[8] N. S. Kuppuswamy, S. H. Cho, and J. H. Kim, "A cognitive control architecture for an artificial creature using episodic memory,", in *Proc. Int. Joint Conf. SICE-ICASE*, pp. 3104–3110, 2006.

[9] S. Jockel, D. Westhoff, and J. Zhang, "Epirome-a novel framework to investigate high-level episodic robot memory," in *Proc. IEEE Int. Conf. Robot. Biomim. (ROBIO)*, pp. 1075–1080, 2007.

[10] S. Jockel, M. Weser, D. Westhoff and J. Zhang, "Towards an Episodic Memory for Cognitive Robots", in *Proc. of 6th Cognitive Robotics workshop at 18th European Conf. on Artificial Intelligence (ECAI)*, pp. 68–74, 2008.

[11] Z. Kasap and N. Magnenat-Thalmann, "Towards episodic memory-based long-term affective interaction with a human-like robot," in *Proc. IEEE Int. Symp. Robot Hum. Interact. Commun. (Ro-Man)*, pp. 452–457, 2010.

[12] A. F. Morse, J. de Greeff, T. Belpeame, and A. Cangelosi, "Epigenetic Robotics Architecture (ERA)," *IEEE Trans. Auton. Mental Develop.*, vol. 2, no. 4, pp. 325–339, 2010.

[13] F. Bellas, A. Faina, G. Varela, and R. J. Duro, "A cognitive developmental robotics architecture for lifelong learning by evolution in real robots," in *Proc. Int. Joint Conf. Neural Networks*, pp. 1–8, 2010.

[14] D. Stachowicz and G. Kruijff, "Episodic-like memory for cognitive robots," *IEEE Trans. Auton. Mental Develop.*, vol. 4, no. 1, pp. 1–16, 2012.

[15] R. Salgado, F. Bellas, P. Caamano, B. Santos-Diez, and R. Duro, "A procedural long term memory for cognitive robotics," in *Proc. IEEE Workshop Evol. Adapt. Intell. Sys. (EAIS)*, pp. 5762, 2012.

[16] D. G. Tecuci and B. W. Porter, "A generic memory module for event", Ph.D. dissertation, University of Texas at Austin, 2007.

[17] E. E. Smith and S. M. Kosslyn, *Cognitive psychology: Mind and brain*. Pearson Prentice Hall, 2006.

[18] D. R. Godden and A. D. Baddeley, "Context-dependent memory in two natural environments: On land and underwater," Brit. J. Psychol., vol. 66, no. 3, pp. 325–331, 1975.

[19] A. Baddeley, "The episodic buffer: a new component of working memory?," *Trends Cogn. Sci.*, vol. 4, no. 11, pp. 417–423, 2000.

[20] F. Mastrogiovanni, A. Scalmato, A. Sgorbissa, and R. Zaccaria, "Problem awareness for skilled humanoid robots," *International Journal of Machine Consciousness*, vol. 3, no. 1, pp. 91–114, 2011.

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

[22] H. Eichenbaum and N. J. Cohen, *From conditioning to conscious recollection: Memory systems of the brain*, Oxford University Press, 2001.

[23] E. Tulving, "Episodic memory and common sense: how far apart?" *Philos. Trans. R. Soc. Lond.*, vol. 356, no. 1413, pp. 1505–1515, 2001.

[24] E. Tulving, "Episodic memory: from mind to brain," Annu. Rev. Psychol., vol. 53, no. 1, pp. 1–25, 2002.

[25] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," *Prog. Brain Res.*, vol. 155, pp. 23–36, 2006.

[26] S. Jeong, S. Ban and M. Lee, "Stereo saliency map considering affective factors and selective motion analysis in a dynamic environment", *Neural Networks*, vol. 21, pp. 1420–1430, 2008.

[27] A. Baddeley, "Exploring the Central Executive", *The Quarterly Journal of Experimental Psychology Section A*, vol. 49, no. 1, pp 5–28, 1996.

[28] A. Baddeley, "The central executive: A concept and some misconceptions," *Journal of the International Neuropsychological Society* vol. 4, no. 5, pp 523–526, 1998.

[29] F. Collette and M. V. Linden, "Brain imaging of the central executive component of working memory," *Neuroscience & Biobehavioral Reviews*, vol. 26, no. 2, pp 105–125, 2002.

[30] S. Jeong, H. Arie, M. Lee and J. Tani, "Neuro-robotics study on integrative learning of proactive visual attention and motor behaviors", *Cognitive Neurodynamics*, vol. 6, no. 1, pp 43–59, 2011

[31] M. I. Posner, S. E. Petersen, "The attention system of the human brain", *Washington Univ. St. Louis Mo. Dept. of Neurology*, 1989.

[32] S. Kaster and L. G. Ungerleider, "Mechanisms of visual attention in the human cortex", *Ann. rev. neuro.*, vol. 23, no. 1, pp 315–341, 2000.

[33] E. Tulving, "Memory and consciousness", *Canadian Psychology/Psychologie Canadienne*, vol. 26, no. 1, pp 1–12, 1985

[34] L. R. Squire, "Memory systems of the brain: A brief history and current perspective", *Neurobiology of Learning and Memory*, vol. 82, no. 3, pp 171–177, 2004.

[35] R. N. A. Henson, M. D. Rugg, T. Shallice, O. Josephs, and R. J. Dolan, "Recollection and familiarity in recognition memory: an event-related functional magnetic resonance imaging study.", *The Journal of Neuroscience*, vol. 19, no. 10, pp 3962–3972, 1999.

[36] R. N. A. Henson, S. Cansino, J. E. Herron, W. G. K. Robb, and M. D. Rugg "A familiarity signal in human anterior medial temporal cortex?", *Hippocampus*, vol. 13, no. 2, pp 301–304, 2003.