View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Open Research Exeter

Learning and Using Context on a Humanoid Robot Using Latent Dirichlet Allocation

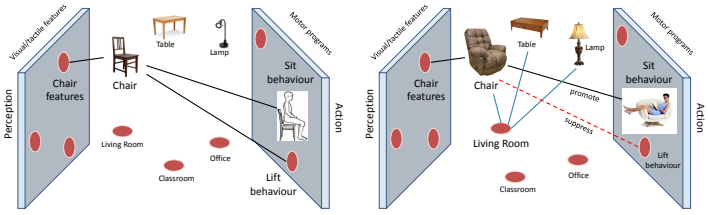
Hande Çelikkanat*, Güner Orhan*, Nicolas Pugeault†, Frank Guerin‡, Erol Şahin* and Sinan Kalkan*
*KOVAN Research Lab., Computer Engineering, Middle East Technical University, Turkey
Email: {hande,guner.orhan,erol,skalkan}@ceng.metu.edu.tr
†CVSSP, University of Surrey, UK, Email: n.pugeault@surrey.ac.uk
‡Computing Science, University of Aberdeen, Scotland, Email: f.guerin@abdn.ac.uk

***Abstract*—In this work, we model context in terms of a set of concepts grounded in a robot's sensorimotor interactions with the environment. For this end, we treat context as a latent variable in Latent Dirichlet Allocation, which is widely used in computational linguistics for modeling topics in texts. The flexibility of our approach allows many-to-many relationships between objects and contexts, as well as between scenes and contexts. We use a concept web representation of the perceptions of the robot as a basis for context analysis. The detected contexts of the scene can be used for several cognitive problems. Our results demonstrate that the robot can use learned contexts to improve object recognition and planning.**

I. Introduction

Context is the totality of the information characterising the situation of a cognitive system; e.g. it can include objects, persons, places, and temporally extended information related to ongoing tasks, but also information not directly related to these tasks. In natural cognitive systems, behaviour is a response to not only a given stimulus, but also the stimulus in the context of other stimuli. It is known that the way in which an organism responds to one stimulus depends also on other, apparently irrelevant, stimuli which constitute what we call context. In fact natural systems not only consider these "apparently irrelevant" stimuli, but they also use them to their advantage. Yeh & Barsalou [1] show how humans tested on a variety of cognitive tasks get a significant performance boost by taking account of contextual information. This applies to tasks that are highly relevant for cognitive robots, such as recognising objects and events, categorisation, retrieval of relevant information and skills, language understanding, problem solving and reasoning. In artificial cognitive systems, unfortunately, extraneous stimuli tend to be viewed as something to be avoided or factored out. This is partly a legacy from the closed world assumptions of classical Artificial Intelligence.

Rather than viewing context as inconvenient, we believe that we can use it to our advantage. This advantage accrues because contextual information promotes relevant information and behaviours, while it suppresses irrelevant ones. This can be explained by an example from Yeh & Barsalou [1] using the notion of situated concepts. Situated concepts are concepts that are associated (in memory) with contextual information derived from the situations in which the concept was previously experienced. For example, the concept of chair is linked to the situations where it has occurred (including locations such as offices, living rooms, classrooms, and associated actions such as reclining), as illustrated in Figure 1(b). Activating the concept of chair then allows us to draw immediate inferences about where to find one, and likely adjacent objects, e.g. tables and lamps, which may be useful in planning. Irrelevant inferences are suppressed: in a living room context a chair can be expected to be large and soft, not small and hard. Possible actions are also activated and instantiated in a context dependent fashion (e.g. reclining on a living room chair); unlikely actions (lifting) are suppressed. All of this is based on likelihood, but it works, because the statistics of the associated contextual knowledge have been built up through extensive experience in real-life scenarios.



(a) Existing Cognitive Systems (b) Our proposal

Fig. 1. (a) Existing cognitive systems have concepts which have links to perceptual features and motor actions which were programmed by a designer or trained in context-free environments. (b) We propose a system that learns links between concepts and sensorimotor primitives that are based on the statistics of its interactions in real-life environments (i.e. *in context*). For clarity, only a few links and concepts are shown.

For developmental roboticists it is not a new idea to think that abstract symbolic representations must be linked to the perceptual states of the robot (some people call this "grounding" [2]). In this article, we propose to extend that by linking the interpretation and the processing of an object or an event by also using the sensorimotor data from the other entities in the environment. The robot first builds a web of concepts that get activated by the multi-modal data coming from the scene. We define and model context as related to a set of active concepts in this web using Latent Dirichlet Allocation [3].

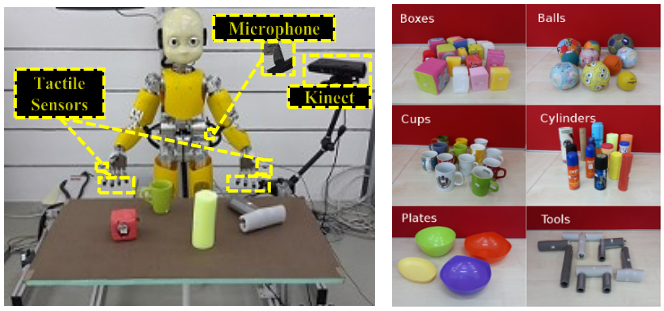
A. Context in Cognitive Science and Robotics

It is a matter of consensus across fields that context processing is an essential part of embodied cognition (e.g., psychology

[1], language [4]–[6], AI [7], robotics [8], [9] and computer vision [10]. Shank and Abelson [7], for example, argued that reasoning about situations in daily life relies on "scripts" that inform reasoners about the prototypical features of these situations. A restaurant, for example, tends to come with a menu, dishes, a waiter, a chef, and so on. This work has gone on to influence today's formal ontologies. Probably the earliest research on context focused on linguistic phenomena, studying how the understanding of an expression (e.g. a personal pronoun like "it") is affected by the rest of the sentence or text [5]. Later research applied these ideas to other aspects of communication, including speech (e.g. pitch accent) and body language (e.g. [11]). Even more drastically, the notion of a context has been extended to all symbolic systems (e.g., [12]). Perhaps most notably, McCarthy [9] and his colleagues proposed the rectification of context in classical (logical) AI, arguing that Artificial Intelligence needs to put the notion of a context centre stage. In McCarthy's view, intelligent machines "must construct or choose a limited context containing a suitable theory whose predicates and functions connect to the machine's inputs and outputs in an appropriate way" [13]. This work gave rise to a wave of theoretical work focusing on issues like the problem of "lifting" information about one knowledge base to another (e.g., if a universally quantified proposition is known to be true in context X then what is the strongest proposition known to be true in a context Y that is related to X in some well- understood way).

Work in all these traditions continues to inspire Cognitive Science and AI. But times have changed: the rise of embodied cognition theories in the 90's, for instance, has offered a different perspective on context, based on a perceptual and action-based rather than symbolic approach [8]. This perceptual perspective is particularly relevant for robotics, where contexts typically need to be acquired from perception (i.e., they cannot be programmed in advance). Barsalou, for example, has advocated the necessity for concepts to be situated [1], [14]; in other words, for an abstract concept to be related to concrete contexts. Coventry et al. [6] studied the difference between geometric and functional contexts in the use of spatial prepositions ("over" vs. "above") and of linguistic quantifiers ("few" vs. "many" vs. "several").

Robotic science has achieved significant success in terms of both theory and applications in the past five decades; however, research involving context has focused on the environmental aspect only, i.e., situation awareness, which involves perceiving and interpreting what is happening in the environment. Robotic studies have investigated situation awareness in urban search for rescue tasks [15], home security [16] and elderly people's living environments [17], object recognition in daily activities [18]. LDA has already been used in robotics; e.g., for learning concepts and their labels [19], or autonomous drive annotation [20]. In computer vision, the notion of context has grown in prominence over the last decade, both explicitly and implicitly. Explicitly, the study of visual gist [21] showed that holistic encodings of the visual input could carry a large amount of information for intelligent systems allowing



(a) The iCub humanoid robot (b) Used objects

Fig. 2. (a) The iCub humanoid robot platform. (b) The objects we use in our experiments. [Best viewed in color]

scene identification [21], [22], urban scene detection [23], and autonomous navigation [24]. Also, the importance of context in visual detection and recognition tasks has become prevalent in recent years: action recognition [25], object categorisation [26], and detection [10]. Implicitly, the now popular data-driven, machine learning-based approach to vision led to algorithms that efficiently extract all predictive information from the visual data, effectively making heavy use of context to reach high performance (see [27] for a criticism).

In summary we see in existing works various piecemeal efforts to tackle particular facets of context in particular application domains. In contrast, and following the intuitions of Yeh & Barsalou [1], we argue that a principled approach is needed to learn, represent and process context in a developing cognitive system. If such can be achieved then the benefit will not be for just one task, but across all areas of cognition. However, such a computational or robotic implementation has not been attempted yet. In this study, we make an initial attempt at such a formalization.

II. The Experimental Setup

We use the iCub humanoid robot platform in this work. iCub operates on a tabletop environment (Figure 2(a)), observing and exploring objects. It collects visual information through a Kinect sensor. Audio, haptic and proprioceptive information are collected by applying previously learned actions on the objects. Its repertoire of objects consists of 60 real-life or hand-made objects (Figure 2(b)-(g)). The performed behaviors are grasping and shaking for collecting audio information, and checking coating material, knocking them over if they are of convenient height, and pushing them forwards, backwards, leftwards, and rightwards. Eventually, ten sets of features are extracted for each object, before and after each action.

For processing the visual data from the Kinect, we use the Point Cloud Library [28], to extract 60 raw features of shape (surface normal and shape index features), as well as the position and the size of the object. Haptic information comes from iCub's tactile sensors, proprioceptive information from the finger joints. Audio information is collected from a microphone. In total we use a feature vector of length 92 [29].

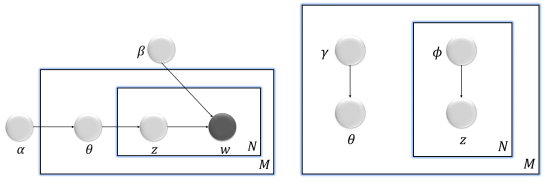
III. METHODS

We relate context to a set of concepts that are active in a web of concepts that the robot learns. Context is extracted as a latent topic variable using Latent Dirichlet Allocation (LDA), a state-of-the-art method used widely in natural language processing for modeling topic in text documents. First, we allow the robot to acquire information about its surrounding, by performing exploratory movements in the environment. At the end of this open-ended learning, the robot gathers information about the current scene, represented in the form of a concept web. The connections in the web carry information about commonly co-occurring concepts. For instance, the activation of the *cup* concept can activate the related *grasping* concept. Then, given the scenes that the robot has encountered, it analyzes its perceptions with LDA, leading to the emergence of hidden contexts in the scene. The detected contexts in turn provide feedback to the concept web in order to affect currently active concepts on the spot.

A. Latent Dirichlet Allocation (LDA)

LDA [3] is a generative probabilistic model for inference of hidden (latent) variables, called topics, from documents. A natural application is inferring the topics of text documents. Each document $d \in \mathcal{D}$ is composed of a set of words $w_1, ..., w_N$ ($w_i \in \mathcal{W}$), and is assumed to be a finite mixture over a (known) set of topics $t_1, ..., t_M$ ($t_i \in \mathcal{T}$). Therefore, a document can be represented by its topic probability distribution $p(t|d)$. Each topic has a probability distribution of generating various words in the document, $p(w|t)$. However, the memberships between the words and the topics are not strongly defined: A word can be generated by multiple topics.

LDA assumes that a collection of documents (called a corpus) is defined by a Dirichlet prior α, and each topic has an a priori fixed word distribution β. A document d is "generated" by choosing $\theta_d \sim \text{Dir}(\alpha)$ to be the topic distribution for document, then for each word position n choosing a topic $z_n \sim \text{Multinomial}(\theta_i)$, and then a word w_n from the probability $p(w_n|z_n, \beta)$. A graphical visualization is provided in Figure 3(a).



(a) (b)

Fig. 3. (a) Plate notation of LDA. The outer plate represents documents, the inner plate represents each word position. The parameters α and β are corpus-level, meaning they need to be sampled only once when generating a corpus. θ_i is document level, therefore needs to be sampled once for each document. The variables z_{in} and w_{in} are word-level and should be sampled once for every word position in every document. (b) The variational distribution used to approximate LDA. Free variational parameters γ and ϕ are introduced to remove the coupling between θ and β. Both figures based on [3].

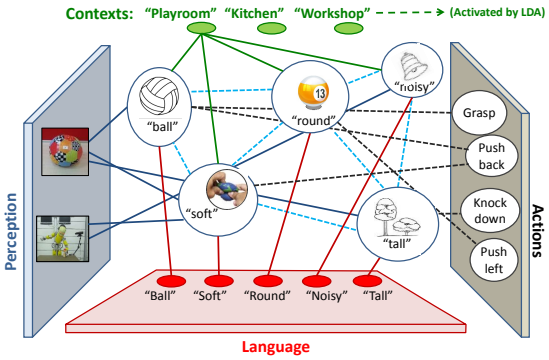


Fig. 4. A sample concept web. Concepts connect together perceptions, actions, and language. Frequently co-occurring concepts have stronger links. 'Contexts' as detected by LDA are connected to the web as a set of higher-level nodes. Only a subset of concepts and connections are shown for clarity.

Given a corpus, LDA tries to find the parameters α and β that maximize the log likelihood of the data. Once we estimate them, we can make various inferences, such as the probability of a given document, the probability distribution of topics of a document, or the likely topic of each word given its document.

However, due to the coupling between the parameters θ and β, this distribution is intractable to compute. Blei et al. [3] offer alternative approximate solutions, among which we follow the variational inference method. The idea is to introduce free variational parameters γ and ϕ, which allow us to get rid of the coupled parameters (Figure 3(b)). Minimizing the Kullback-Liebler divergence (D) between the actual and variational distributions gives the optimal values for γ and ϕ:

$$(\gamma^*, \phi^*) = \arg\min_{(\gamma,\phi)} \quad \text{D}\left(p(\theta, \vec{z}|\gamma, \phi) \;||\; p(\theta, \vec{z}|\vec{w}, \alpha, \beta)\right). \quad (1)$$

The minimization problem can be solved iteratively by using Expectation-Maximization. We have used the LDA source code package provided as companion to [3].

B. Building a Web of Concepts

We model the robot's perceptions of the world in a *concept web*. Basically a concept web is a convenient representation of what we are momentarily perceiving over what we already know about the semantic relations in the world. The robot connects what it sees, be it entities, properties, or applied actions, to previously deduced concepts $c \in \mathcal{C}$ using the prototype approach developed in previous work [29], [30]. A concept can be a noun ($c \in \mathcal{N}$), adjective ($c \in \mathcal{A}$), or verb ($c \in \mathcal{V}$), with $\mathcal{C} = \mathcal{N} \cup \mathcal{A} \cup \mathcal{V}$. The concepts that are highlighted by the immediate perceptions are regarded activated ($c \in \mathcal{C}_{act}$), which in turn proceed to activate other related concepts (Algorithm 1). *Relatedness* of concepts is determined by neighborhood relations, and is denoted by a set of links $l \in \mathcal{L}$. The activation spreads over the web until convergence. Eventually, the converged activation map of the concept web forms an informed representation of the scene, and is used as input to the LDA algorithm.

1) Mapping Perception to Concepts: We represent concepts with a condensed prototype-based representation [29], [30],

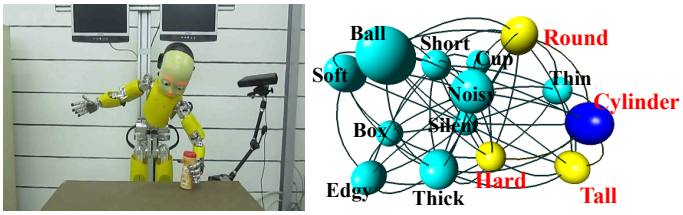


Fig. 5. A sample scene/interaction, with a single cylinder object, and the concept web formed by this scene. The active concepts are indicated by dark blue (for nouns) and yellow (for adjectives). [Best viewed in color]

and check membership by a simple comparison against the prototype of a concept. Instead of the prototype-approach, one could use any classifier to predict the labels $(\mathcal{N}, \mathcal{A}, \mathcal{V})$ from observations; however, we use the prototype-based approach since it performs better [29]. For the sake of concision, the details and example prototypes are provided in Appendix I.

The raw perception vector from each object is matched against prototype vectors of each concept, learned a priori. We use 6 noun concepts ($\mathcal{N}$ ={*ball, box, cup, cylinder, plate, tool*}), 10 adjective concepts ($\mathcal{A}$ ={*hard, soft, tall, short, thin, thick, round, edgy, noisy, silent*}) and 6 verb concepts ($\mathcal{V}$ ={*moving left, moving right, moving forward, moving backward, knocking down, making noise*}). These verb concepts are abstractions over the behaviors ($\mathcal{B}$={*p*ush left, grasp and move left, push forward, grasp and move forward, push right, grasp and move right, knock down, push backward, grasp and move backward}) based on their effects [30]. (The behaviors are hard-coded for our setup.)

2) Learning the Concept Web: For utilizing a priori discovered semantic information about the world, we organize known concepts in a concept web (Figure 4). Nodes in the web denote concepts, The links between the web depends on the relatedness of the concepts. Concepts that co-occur frequently have stronger connections to each other. Co-occurrence is calculated during training by the conditional probabilities, thus the strength of the link l_{BA} from concept A to B is:

$$l_{BA} \propto p(B|A), \tag{2}$$

where $p(B|A)$ is the conditional probability of B given A (determined statistically from the training set, with $p(B|A)$ not necessarily equal to $p(A|B)$).

3) Spreading of Activation over the Concept Web: Operating on a previously learned concept web, current perceptions of the robot are mapped to the web to invoke related concepts. The perceptually activated concepts also spread their activation to their neighbors, depending on the strength of the connection in between. A previously inactive node c_n that is located hop steps away from an already active node c_{act} is activated if

$$\text{activation}(c_{act}) \times p(c_n|c_{act}) \times 2^{(1-hop)} > \tau, \tag{3}$$

where activation$(\cdot)$ is the activation of a node, and τ is the threshold value for activation. The algorithm is provided in Algorithm 1. Spreading continues on the web until convergence. Fully converged web is used as the input for deciding on the context. A sample concept web for a specific scene, after the convergence of activations, is depicted in Figure 5.

Algorithm 1 Spreading of Activation in Concept Web.

C_{act}: list of active nodes (initially perceptually activated with an activation of 1.0)
procedure ACTIVATENEIGHBORS(c_{act})
  // c_{act}: active concept
  for all $c_n \in neighborhood\{c_{act}\}$ **do**
    activation$(c_n) \leftarrow$ **activation**$(c_{act}) \times p(c_n|c_{act}) \times 2^{(1-hop)}$
    if **activation**$(c_n) > \tau$ **then**
      activate(c_n)
      $C_{act} \leftarrow C_{act} \cup \{c_n\}$
    end if
  end for
end procedure
procedure SPREADACTIVATION( )
  while $C_{act} \neq \emptyset$ **do**
    $c_{act} \leftarrow C_{act}$.pop()
    ActivateNeighbors(c_{act})
  end while
end procedure

C. Formalizing and Learning Context

In the robot's environment, the sensory input it collects comes from a "scene" at a time. This scene gives rise to the perceived context, effectively facilitating and optimizing the robot's actions on the spot. A given scene, populated by a variety of objects, might be comprised of one or more contexts, which are in turn semantically connected to the objects in the scene.

Therefore, in this paper, we model a scene as equivalent of an LDA *document* $d \in \mathcal{D}$, whereas the concepts evoked by the objects in the scene correspond to *words*, $w_i \in \mathcal{W}$. Any scene will have an associated topic probability distribution, with each *topic* ($t_i \in \mathcal{T}$) corresponding to a specific context. A context will then have a probability distribution of generating each concept ($c \in \mathcal{C}$), which can be a noun, adjective, or verb $\mathcal{C} = \mathcal{N} \cup \mathcal{A} \cup \mathcal{V}$. The collection of all the scenes the robot encounters will constitute the *corpus*. The relations between LDA terms and ours are summarized in Table I for clarity.

We model the scene, i.e, the *document*, as a concept web, described above. Our scenes conform the bag-of-words assumption of LDA, the order of the objects in the scene is not important, only their existence counts. This assumption carries over to the concept web, the existence of activation, but not its particular order, is important.

When the robot encounters a scene, it infers a new concept web, and uses it as an LDA document. LDA estimates the context distribution for the overall scene, as well as the likely contexts of each individual active concept node. The estimated contexts are then imposed on the concept web, as separate active nodes higher in the hierarchy. (Note than since we do not have an attentional mechanism, we focus on the objects one by one, and collect all the perceived concepts in one concept web.)

One of the important points is that the contexts of each scene is determined by the system in an unsupervised manner, through the implicit detection of co-occurrences of concepts by LDA via approximating the variational parameters. The training phase does not include any supervised labeling of the scenes in term of expected contexts. LDA only requires the

TABLE I
THE CORRESPONDENCE BETWEEN THE LDA TERMS AND OUR NOTATION.

LDA	Our Notation
document $d \in \mathcal{D}$	a single scene (represented as a concept web)
corpus $\mathcal{D}$	all scenes encountered during training phase
word $w_i \in \mathcal{W}$	an active concept c_{act} in the concept web (can be a noun, adjective, or verb: $c_{act} \in \mathcal{C} = \mathcal{N} \cup \mathcal{A} \cup \mathcal{V}$)
topic $t_i \in \mathcal{T}$	a 'context', either Kitchen, Playroom, or Workshop

TABLE II
USED CONTEXTS AND THEIR PREVALENT CONCEPTS. NOTE THAT THERE IS NO STRICT ONE-TO-ONE MEMBERSHIPS.

Kitchen			Playroom			Workshop		
cup	short	thin	ball	edgy	silent	tool	edgy	tall
plate	hard	thick	box	soft	thick	cylinder	hard	thin
round	silent		round	noisy		round	silent	thick

number of contexts to be detected as input. Then assigns each scene with its likely contexts in an unsupervised way. The detected contexts are not named but are given arbitrary ids, therefore the assignment of their names is done afterwards via human inspection, for ease of discussion. The contexts the robot detected in this work, as later named by us, are presented in Table II. Figure 4 demonstrates how detected contexts are afterwards joined to the concept web.

D. Using Context

Once the context is determined, the robot can make use of this high-level knowledge for several tasks, such as object recognition, scene understanding, planning etc. In our particular implementation, we model 'context' as a higher-level node in the concept web, which can be connected to any other node of the concept web, as if a regular node (Figure 4). Therefore, an activated context node can spread its activation to other nodes, effectively guiding reasoning of the robot further.

IV. RESULTS

We present two different scenarios to demonstrate how context can be used by a robot. In the first, we show the effect of context on categorizing objects and their properties. We also present the quantitative results of our context detection mechanism in a number of pure and mixed contexts. In the second scenario, we show how including context might be beneficial in a typical planning scenario.

For learning to detect the context, the robot is trained with 50 scenes, each of which is selected randomly from one of 3 contexts. Each scene is filled with objects belonging to the noun category of its designated contexts. The adjective categories related to the contexts follow as a natural result of choosing objects from these noun categories. The LDA module is given the 50 scenes without explicitly specifying any a priori information about their expected contexts. As mentioned in Section III-C, the detected contexts therefore are decided in an unsupervised manner.

A. Scenario 1: Using Context for Scene and Object Categorization

The first scenario shows how context can be used, for understanding the scenes and using this for improving object categorization. Figure 6 depicts the RGB-colored depth images for 6 objects, one for each category, as seen by the Kinect sensor. The second column shows the predicted related noun and adjective concepts and their confidence values. The predictions are generally correct, however they have low confidence values, and unrelated concepts are also always predicted with non-negligible confidences. There are also two wrong predictions (adjectives).

After putting objects in a context and including the contextual information with LDA, contextual activation inhibits the prediction of unrelated categories. For instance, the noun category probabilities of ball, box, plate and tool are zeroed out, as well as the probability of edginess, for the first object (a cup). Moreover, wrong adjective predictions are corrected using context.

For evaluation, Figure 7 also presents the quantitative performance of our framework in correctly deducing the context in pure and mixed scenes. 3 pure scenes (one for each of the Kitchen, Playroom, and Workshop contexts), as well as 3 mixed scenes are presented to the robot, after its training with 50 pure scenes. As we see, the robot can correctly predict the context in the pure scenes and mixed scenes with high confidence.

B. Scenario 2: Using Context in Planning

The second scenario demonstrates how context can facilitate action planning in real-time. The robot is given an initial state of an object, and is told to bring it to an expected final state. It is expected to find a viable action plan via forward planning. However, planning trees have large branching factors, being often too large for real-time processing. We argue that context may help in simultaneous pruning of the tree in order to enable efficient processing. The robot first predicts the context of the given object, and then uses only the relevant behaviors in that context.

Two sample scenarios are presented in Figure 8. The robot is expected to move objects to target positions on the table. It follows a breadth-first forward planning scheme. At each node, the action's predicted effect feature vector is compared against the goal state (similar to [30]). If the goal state is not yet reached, 9 child nodes are expanded for the current node.

The pruning of the nodes is done simultaneously during the expansion phase. For instance, for an object that belongs to the Kitchen context, it is decided (by a human user) that the actions *knock down* and *push x* are not feasible, since they can result in spilling of liquids in containers. On the other hand, for an item of the Playroom context, which is a ball with 50% probability, pushing will cause it to roll down from the table, and again should not be attempted. Therefore, these nodes are eliminated in these contexts as soon as they are added to the planning tree, and not expanded any further. The resulting planning trees are shown to have 4^3 and 5^3 nodes instead of 9^3, with a reduction of 665 and 604 nodes respectively.

V. CONCLUSION

We have presented a framework for robots to learn and use the context(s) of a given scene. Our formulation builds

TABLE III
SAMPLE PROTOTYPES FOR NOUNS ($\mathcal{N}$), ADJECTIVES ($\mathcal{A}$) AND VERBS ($\mathcal{V}$)

	Concepts	Visual Features	Audio Features	Haptic Features	Proprioceptive Features
$\mathcal{N}$	Box	+−+++−++−−−++−−−−−−−−−−−−−++−−+−−−−−−−−∗∗−−−−−−−−−−−−−−−−−−−+−−−−	−∗∗∗∗∗−++−∗++	−−−−−−	−+∗∗−
	Ball	−−+−+−+−−−−++−−−−−−−−−−−−−++++−−−−−−−−−−−−−−−−−−−−−−−−−−−+−−−−	∗∗∗∗∗∗∗−∗∗+−+	∗−∗∗∗∗	−+++−
$\mathcal{A}$	Hard	+−+∗++++++++−−−−−−−−−−−−−−+∗++++−−−−−−−∗−−−−−−−−−−−−−−−−−−−+++++	∗∗∗∗∗∗∗∗∗∗∗∗	∗+++++	−+∗∗−
	Noisy	−−+++∗+−−−∗++∗−−−−−−−−−−−−−+++++∗−∗∗∗∗−−−−−−−−−−−−−−−−−−−−+−−−	+++++++++++	∗−∗∗−∗	−+∗∗−
$\mathcal{V}$	Move Left	0−0∗∗∗∗0−0+0++−−−−−−−−−−−−−−0−−−00−0−−+00−000000−−−−−−−−−−−−−−−+++++	None		
	Knock Down	0+0++++00++00+−−−−−−−−−−−−−0−0−+0−0−−+0000000−∗−−−−−−−−−−−−−−−+++++	None		

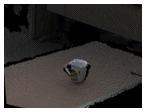
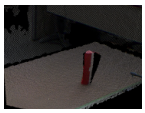
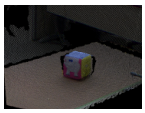
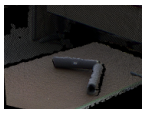
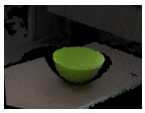
Object	Predicted Nouns w/o Context (% confidence)	Predicted Adjectives w/o Context (% confidence)	In a Context	Predicted Context	Predicted Nouns w/ Context (% confidence)	Predicted Adjectives w/ Context (% confidence)
	ball (9%) box (14%) **cup (34%)** cylinder (21%) plate (6%) tool (13%)	edgy (42%) **round (58%)** **hard (60%)** soft (40%) noisy (38%) **silent(62%)** **short (54%)** tall (46%) thick (48%) **thin (52 %)**		**Kitchen (100%)** Playroom (0%) Workshop (0%)	ball (0%) box (0%) **cup (85%)** cylinder (0%) plate (15%) tool (0%)	edgy (0%) **round (100%)** **hard (60%)** soft (25%) noisy (38%) **silent(62%)** **short (54%)** tall (46%) thick (48%) **thin (52 %)**
	**ball (25%)** box (14%) cup (14%) cylinder (16%) plate(11%) tool (15%)	edgy (43%) **round (57%)** hard (45%) **soft (55%)** **noisy (51%)** silent(49%) **short (55%)** tall (45%) **thick (57%)** thin (43%)		Kitchen (0%) **Playroom (100%)** Workshop (0%)	**ball (64%)** box (36%) cup (0%) cylinder (0%) plate (0%) tool (0%)	edgy (43%) **round (57%)** hard (0%) **soft (100%)** **noisy (51%)** silent(49%) **short (100%)** tall (0%) **thick (100%)** thin (0%)
	ball (11%) box (15%) cup (22%) **cylinder (29%)** plate (7%) tool (14%)	edgy (39%) **round (61%)** **hard (61%)** soft (39%) noisy (39%) **silent (61%)** short (48%) **tall (52%)** thick (41%) **thin (59%)**		Kitchen (0%) Playroom (0%) **Workshop (100%)**	ball (0%) box (0%) cup (0%) **cylinder (67%)** plate (0%) tool (32%)	edgy (0%) **round (100%)** **hard (100%)** soft (0%) noisy (34%) **silent (66%)** short (0%) **tall (100%)** thick (41%) **thin (59%)**
	ball (13%) **box (28%)** cup (15%) cylinder (16%) plate (9%) tool (15%)	**edgy (58%)** round (42%) hard (42%) **soft (58%)** noisy (43%) **silent (57%)** **short (56%)** tall (44%) **thick (57%)** thin (43%)		Kitchen (0%) **Playroom (100%)** Workshop (0%)	ball (%32) **box (68%)** cup (0%) cylinder (0%) plate (0%) tool (0%)	**edgy (58%)** round (42%) hard (0%) **soft (100%)** noisy (43%) **silent (57%)** **short (100%)** tall (0%) **thick (100%)** thin (0%)
	ball (10%) box (17%) cup (16%) cylinder (19%) plate (8%) **tool (27%)**	edgy (45%) **round (55%)** **hard (53%)** soft (47%) ~~noisy (54%)~~ silent (46%) short (41%) **tall (58%)** **thick (52%)** thin (48%)		Kitchen (0%) Playroom (0%) **Workshop (100%)**	ball (0%) box (0%) cup (0%) cylinder (41%) plate (0%) **tool (59%)**	edgy (0%) **round (100%)** **hard (100%)** soft (0%) noisy (49%) **silent (51%)** short (0%) **tall (100%)** **thick (52%)** thin (48%)
	ball (13%) box (12%) cup (12%) cylinder (14%) **plate (29%)** tool (16%)	edgy (47%) **round (53%)** hard (49%) ~~soft (51%)~~ noisy (46%) **silent (53%)** short (49%) **tall (51%)** **thick (53%)** thin (47%)		**Kitchen (100%)** Playroom (0%) Workshop (0%)	ball (0%) box (0%) cup (29%) cylinder (0%) **plate (71%)** tool (0%)	edgy (0%) **round (100%)** **hard (66%)** soft (34%) noisy (46%) **silent (53%)** short (49%) **tall (51%)** **thick (53%)** thin (47%)

Fig. 6. Object categorization as facilitated by the help of context understanding. Images depict RGB-colored depth images (collected via PCL library from the Kinect sensor). The use of bold text indicates correct decisions. The striked-through text indicates wrong decisions. [Best viewed in color]

Scene	Existing Objects	Predicted Context (% contribution)	Scene	Existing Objects	Predicted Context (% contribution)
Pure Contexts			**Mixed Contexts**		
	2 cups,2 plates	**Kitchen (100%)**		2 plates, 1 cup, 1 box	**Kitchen (67%), Playroom (33%)**
	3 boxes, 3 balls	**Playroom (100%)**		2 plates, 2 cup, 1 ball, 1 box	**Kitchen (58%), Playroom (42%)**
	2 tools,3 cylinders	**Workshop (100%)**		1 tool, 2 cylinders, 1 plate, 1 cup	**Workshop (55%), Kitchen (45%)**

Fig. 7. Predicted contexts in pure and mixed topic scenes. Images depict RGB-colored depth images (collected via PCL library from the Kinect sensor). The use of bold text indicates correct decisions. [Best viewed in color]

on the Latent Dirichlet Allocation approach, in which each context is modeled as a latent topic variable. The encountered scenes are represented as concept webs, with frequently co-occurring concepts sharing stronger links. A number of such representations for different scenes are examined by the robot for learning to distinguish the prominent context(s). The

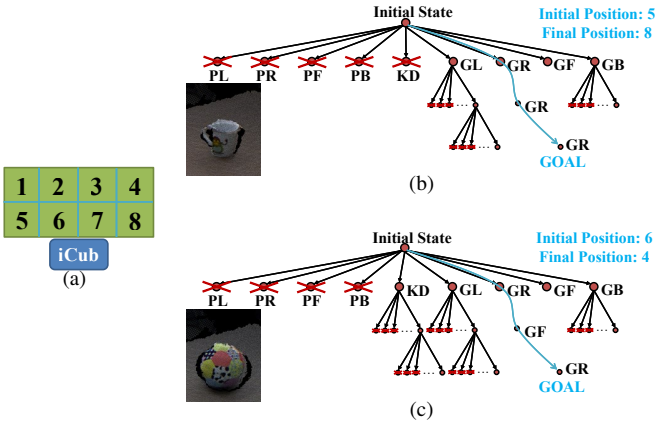


Fig. 8. Pruning of forward planning trees by integrating contextual information. (a) iCub's workspace schematized. (b-c) Two planning scenarios. The branches that are pruned due to being irrelevant for the current context are shown with crosses. The behavior abbreviations stand for: PL: Push left, PR: Push right, PF: Push forward, PB: Push backward, KD: Knock down, GL: Grasp and move left, GR: Grasp and move right, GF: Grasp and move forward, GB: Grasp and move backward (b) First planning scenario. iCub is expected to move a cup from position 5 to position 8. Since pushing and knocking actions are dangerous in the kitchen context, these nodes are pruned without further expansion. (c) Second planning scenario. iCub must bring a ball from position 6 to position 4. This time pushing actions are pruned, since pushing a ball causes it to roll down from the table.

obtained contextual information is fed back to the concept web for enhancing the robot's performance in two tasks. We have provided preliminary results that the robot can use the context model to increase its object recognition and planning performance.

ACKNOWLEDGMENTS

We would like to thank Angelo Cangelosi, Anna Borghi and Honghai Liu for fruitful discussions on the integrating context in cognitive systems. This work is partially funded by the Scientific and Technological Research Council of Turkey (TÜBİTAK) through project no 111E287.

APPENDIX I

A feature with a low variance and highly positive contribution is marked with a '+', low variance and highly negative contribution with '-', negligible contribution with '0', and too much variance to be consistent with '*'. Eventually we obtain 25 prototype strings, one for each concept, formed of +,-,0,* characters. A number of sample prototypes are depicted in Table III. The raw feature vector of each observation (extracted from different modalities and concatenated) is matched against each prototype to determine concepts for this observation. Each object belongs to 1 noun category, to several adjective categories, and if applicable (any action has been performed), to several verb categories. For details, the reader is referred to [29], [30].

REFERENCES

[1] W. Yeh and L. W. Barsalou, "The situated nature of concepts," *The American journal of psychology*, pp. 349–384, 2006.

[2] A. Cangelosi and T. Riga, "An embodied model for sensorimotor grounding and grounding transfer: Experiments with epigenetic robots," *Cognitive science*, vol. 30, no. 4, pp. 673–689, 2006.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[4] J. Barwise and J. Perry, *Situation and Attitudes*. MIT Press, 1983.

[5] H. Kamp and U. Reyle, *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory*. Kluwer Academic (Dordrecht; Boston), 1993.

[6] K. R. Coventry, A. Cangelosi, S. N. Newstead, and D. Bugmann, "Talking about quantities in space: Vague quantifiers, context and similarity," *Language and Cognition*, vol. 2, no. 2, pp. 221–241, 2010.

[7] R. C. Schank and R. P. Abelson, "Scripts, plans, goals and understanding," *Hillsdale, NJ: Lawrence Erlbaum*, 1977.

[8] R. A. Brooks, "Intelligence without representation," *Artificial intelligence*, vol. 47, no. 1, pp. 139–159, 1991.

[9] J. McCarthy, "Artificial intelligence, logic and formalizing common sense," *Philosophical logic and artificial intelligence*, 1989.

[10] A. Torralba, "Contextual priming for object detection," *Int. Journal of Computer Vision*, vol. 53, no. 2, pp. 169–191, 2003.

[11] K. Van Deemter, B. Krenn, P. Piwek, M. Klesen, M. Schröder, and S. Baumann, "Fully generated scripted dialogue for embodied agents," *Artificial Intelligence*, vol. 172, no. 10, pp. 1219–1244, 2008.

[12] V. Akman and M. Surav, "Steps toward formalizing context," *AI magazine*, vol. 17, no. 3, p. 55, 1996.

[13] J. McCarthy, "From here to human-level ai," *Artificial Intelligence*, vol. 171, no. 18, pp. 1174 – 1182, 2007.

[14] L. W. Barsalou, "Simulation, situated conceptualization, and prediction," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1521, p. 1281, 2009.

[15] B. Larochelle, G.-J. Kruijff, N. Smets, T. Mioch, and P. Groenewegen, "Establishing human situation awareness using a multi-modal operator control unit in an urban search & rescue human-robot team," in *Int. Symp. on Robot and Human Interactive Comm.*, 2011, pp. 229–234.

[16] A. Gregoriades, S. Obadan, H. Michail, V. Papadopoulou, and D. Michael, "A robotic system for home security enhancement," *Aging Friendly Technology for Health and Independence*, pp. 43–52, 2010.

[17] G. Tsuruma, H. Kanai, T. Nakada, and S. Kunifuji, "Dangerous situation awareness support system for elderly people with dementia," in *Int. Conf. on Human Computer Interaction*. ACTA Press, 2007, pp. 62–67.

[18] D. Nyga, F. Balint-Benczedi, and M. Beetz, "Pr2 looking at things: Ensemble learning for unstructured information processing with markov logic networks," in *ICRA*, 2014.

[19] T. Araki, T. Nakamura, T. Nagai, S. Nagasaka, T. Taniguchi, and N. Iwahashi, "Online learning of concepts and words using multimodal lda and hierarchical pitman-yor language model," in *IROS*, Oct 2012, pp. 1623–1630.

[20] T. Bando, K. Takenaka, S. Nagasaka, and T. Taniguchi, "Automatic drive annotation via multimodal latent topic model," in *IROS*, Nov 2013, pp. 2744–2749.

[21] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *IJCV*, vol. 42, no. 3, pp. 145–175, 2001.

[22] L. W. Renninger and J. Malik, "When is scene identification just texture recognition?" *Vision research*, vol. 44, no. 19, pp. 2301–2311, 2004.

[23] N. Pugeault and R. Bowden, "Learning pre-attentive driving behaviour from holistic visual features," *ECCV*, pp. 154–167, 2010.

[24] C. Ackerman and L. Itti, "Robot steering with spectral image information," *IEEE Trans. on Robotics*, vol. 21, no. 2, pp. 247–251, 2005.

[25] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *CVPR*. IEEE, 2009, pp. 2929–2936.

[26] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky, "Exploiting hierarchical context on a large database of object categories," in *CVPR*. IEEE, 2010, pp. 129–136.

[27] N. Pinto, D. D. Cox, and J. J. DiCarlo, "Why is real-world visual object recognition hard?" *PLoS computational biology*, vol. 4, 2008.

[28] R. B. Rusu and S. Cousins, "3d is here: Point cloud library (pcl)," in *ICRA*, 2011, pp. 1–4.

[29] G. Orhan, S. Olgunsoylu, E. Sahin, and S. Kalkan, "Co-learning nouns and adjectives," in *ICDL*. IEEE, 2013, pp. 1–6.

[30] S. Kalkan, N. Dag, O. Yürüten, A. M. Borghi, and E. Sahin, "Verb concepts from affordances," *Interaction Studies*, vol. 15, no. 1, 2014.