

A Probabilistic Concept Web on a Humanoid Robot

Hande Çelikkanat, Güner Orhan, Sinan Kalkan

Abstract—It is now widely accepted that concepts and conceptualization are key elements towards achieving cognition on a humanoid robot. An important problem on this path is the grounded representation of individual concepts and the relationships between them. In this article, we propose a probabilistic method based on Markov Random Fields to model a *concept web* on a humanoid robot where individual concepts and the relations between them are captured. In this web, each individual concept is represented using a prototype-based conceptualization method that we proposed in our earlier work. Relations between concepts are linked to the co-occurrences of concepts in interactions. By conveying input from perception, action, and language, the concept web forms rich, structured, grounded information about objects, their affordances, words, etc. We demonstrate that, given an interaction, a word, or the perceptual information from an object, the corresponding concepts in the web are activated, much the same way as they are in humans. Moreover, we show that the robot can use these activations in its concept web for several tasks to disambiguate its understanding of the scene.

Index Terms—Concepts, Conceptualization, Concept Web, Markov Random Field

I. INTRODUCTION

In the near future, our daily lives will be populated by robots, and one of the bottlenecks for that will be communication. Communication is simply an exchange of sequences of symbols between two agents, and for this exchange of symbols to work, the symbols should elicit the same meaning in both agents. However, for the robot to be able to *understand* the symbols and develop its skills and world knowledge over time, these symbols should be linked to the sensorimotor experiences of the robot, or the abstractions formed from such experiences, as suggested by Harnad [1]. This problem is so fundamental that it has quickly become a holy grail of cognitive research, and the most probable solution seems to be the embodiment of the mind [2]–[6]. Ample evidence has piled up since then, demonstrating the significant coupling between the sensory and motor cortices on the one side, and the high level functions of conceptualization and language on the other [7]–[10] - for a review, see, *e.g.*, [6], [11], [12].

Abstracting from experience and representing abstracted information in the form of concepts are crucial for cognition. Concepts are representations that allow us to make sense of the world by enabling us to categorize the continuous high-dimensional sensorimotor space. Humans are remarkably good at this task, *i.e.*, the task of abstraction, as well as that of relating concepts to each other. When we look at an object, we not only recognize it almost instantly, but also retrieve the information pertaining to its usage and other categorical

information; *e.g.*, the affordances of the object, the contexts in which it is used, the touch, the taste, the smell, the super-ordinate and sub-ordinate categories to which it is linked, etc.

The fact that an observation activates many different concepts in our brains calls for a representation that incorporates interactions between individual concepts such that they are able to activate one another. This is now widely accepted in the literature - see, *e.g.*, [9], [13], [14]. For example, according to Deacon [15], what makes human cognition unique is its ability to form a web between lexical concepts, since this ability enables us to perform symbolic manipulations over the web itself:

“The symbolic basis of word meaning is mediated [...] by the elicitation of other words (at various levels of awareness). Even if we do not consciously experience the elicitation of other words, evidence that they are activated comes from priming and interference effects [...]” (Quotation belongs to [16])

Similarly, Barsalou [17]–[19] notes:

“...[C]oncepts are not typically processed in isolation but are typically situated in background settings, events and introspections. When representing bicycle, for example, people do not represent a bicycle in isolation but represent it in relevant situations [...] [P]eople situate concepts for the following reason: if the brain attempts to simulate a perceptual experience when representing a concept, it should typically simulate a situation, because situations are intrinsic in perception. At any given moment in perception, people perceive the immediate space around them, including agents, objects and events present. Even when people focus attention on a particular entity or event in perception, they continue to perceive the background situation - the situation does not disappear.” (Quotation belongs to [17])

In this article, we take a similar stance, and study how a web of concepts can be represented and formed by a humanoid robot from its sensorimotor interactions with the environment.

II. CONCEPTS, WEB OF CONCEPTS AND RELATED STUDIES IN ROBOTICS

Below we briefly survey conceptualization theories and existing studies on hierarchies or concept webs. Then, we summarize our contributions in relation to existing findings and the literature.

A. Theories of Concepts

Surprisingly, one can trace the discussion about the theories of concepts back to Aristotle and Plato, in the times of Ancient

Hande Çelikkanat, Güner Orhan, and Sinan Kalkan are with KOVAN Research Lab, Department of Computer Engineering, Middle East Technical University, Ankara, TURKEY

E-mail: {hande.guner.orhan,skalkan}@ceng.metu.edu.tr

Greeks. In Aristotle’s Hylomorphism theory, he stated that *substances* have substantial forms and accidental forms. In this view, substantial forms of a substance carry the essential properties of the substance; *i.e.*, they define what makes the substance. On the other hand, the accidental forms of a substance correspond to properties of the substance that can change without becoming another substance.

Since Aristotle, many scientists have questioned what concepts are and how they can be represented. Over the eras, the following main theories have emerged:

- **The Classical (Rule-based) View:** The classical view holds that categories are separated from one another with strict boundaries, and membership of an exemplar to a category is crisp (either yes or no - see, *e.g.*, [20]). An exemplar belongs to a category if it satisfies the membership rules. As an example, an object is considered a BIRD, if it satisfies the following rule for being a BIRD:

$$\begin{aligned} &has_wings(object) \wedge flies(object) \wedge \\ &lays_eggs(object) \wedge has_beak(object) \wedge \dots \quad (1) \end{aligned}$$

As an example, a sparrow can be considered a BIRD since the rule in Equation 1 is true for a sparrow. However, the rule-based view is believed to be inadequate in answering certain questions. For instance, it fails to explain why a sparrow is a more “typical” bird in our mind, although both sparrow and, *e.g.*, penguin are clearly birds [21].

- **The Prototype View:** The membership of an exemplar in a category is decided by the similarity between the exemplar and the category *prototype* (*e.g.*, [22]) The prototype is the perfect exemplar for the category, and demonstrates what an ideal member should be like. Memberships are not crisp, so an exemplar might be a very typical member of the category (indicated with a high similarity to the prototype), or bear only marginal likeliness and yet be regarded as a member. For instance, a sparrow may be a very prototypical bird, while a penguin (which does not fly) is still allowed as a marginal member. (See also [23] for a geometrical interpretation of the prototype view.)
- **The Exemplar View:** The concepts are stored as a collection of previously encountered exemplars that belong to the category. Therefore, rather than a single prototype to check, a new exemplar is compared to all the stored exemplars of the concept in order to determine membership. It will be the case that more typical features of the concept are implicitly represented through more exemplars included, whereas outliers are still allowed via a smaller number of exemplars. For instance, the BIRD concept will be represented by many flying bird instances, and a small number of non-flying ones.

There are advantages and disadvantages to each of these views, for instance see [24] for a review. Moreover, there exists abundant and contradicting evidence supporting each of these claims. Of course given that categorization and conceptualization are central to our thinking, perhaps we might expect different strategies to be employed for different tasks [25], which may also point towards an underlying hybrid representation [26].

B. Grounding Concepts

Having stated the three main schools of thought on concepts, we now present a number of approaches related to their specific implementation. A major question is what achieves the connection between the concepts in our minds, and our actual sensory-motor experiences in the world. As Harnad [1] points out, symbols that are formed by mere manipulation of other symbols, without being rooted in atomic entities, will not be useful for making sense of the world. A widely-accepted explanation is the embodiment of the agent [3], [5]–[7], [27]–[32]. Being embodied and situated, we experience the world in terms of sensory and motor interactions. These experiences constitute the internal meanings we attribute to concepts. An example is Barsalou’s acclaimed Perceptual Symbol Systems hypothesis [2], suggesting that perceptual input triggers bottom-up patterns of activation in sensorimotor regions in our brains, which are then partially reactivated through top-down mechanisms over the association cortices in order to simulate (perceptual) symbols. Therefore, these distributed but connected brain areas work together to implement a basic conceptual system that leads to the emergence of categorization, propositions, and abstract concepts.

Steels [33] has approached the problem from a functional point of view with his Recruitment Theory of Language. His idea is that language abilities are not formed over dedicated and isolated language areas, but somehow “piggyback” over relatively more primitive cognitive functions. In this scenario, different cognitive modules are recruited arbitrarily for use in communication, with the ones that prove efficient (*i.e.*, by increasing the linguistic power) being kept through developmental stages. In this sense, any region that contributes to the formation of concepts, thereby facilitating interpersonal communication, could have been utilized. Perhaps the most obvious ones, in keeping with Steels theory, would be the primary sensory and motor cortices.

So how does a community reach a common language, in which specific things are considered as instances of the same concept? In [6], Cangelosi devises an environment in which a number of agents evolve together, developing sensorimotor capabilities and related language tools. In time, they gain the ability to communicate to each other the necessary information to find food, such as which items are edible, where they are located, etc. In the course of developing this mode of communication, they necessarily ground the labels in their own sensorimotor experiences. Moreover, these conceptualizations can be transferred from a teacher to a student. Belpaeme and Morse [34], in an attempt to explain how young children learn concepts, compare cross-situational learning of concepts with socially guided learning, to show both are feasible, yet social learning is more efficient. Meanwhile, from a different point of view, Hashimoto and Masumi [35] show that concepts can form as attractors in a dynamic system. The transitions between these attractors correspond to the manipulation of symbols in everyday language use. (An interesting point is that they did not find any regularities in the transitions, therefore this system has no explicit “syntax”.)

The grounding and conceptualization of nouns and adjectives

tives, which are “object classes” and “object properties”, are rather intuitive and well-studied (see for instance [36], [37]); whereas an organized attempt towards the conceptualization and especially generalization of behaviors into verbs is comparatively recent in the literature. We have in our previous work [38], as well as Rudolph *et al.* [39], proposed that behaviors can be generalized, and thus defined, in terms of their effects. The effects provide a generalization over a rather continuous space of actions, which can be virtually indistinguishable in the real world. Take, for instance, a request to “pass the salt”, in which one can use either the left or the right hand, and can manipulate the salt shaker with either a power or a precision grasp. Regardless of the variations, the same command is realized in all these scenarios.

C. Structural Representation in Humans

In a parallel vein, numerous neuroimaging and modeling studies have tried to unveil the exact representation scheme and foci of conceptualization in the human brain. To date, perhaps the most widely-accepted of these theories is the notion of highly-connected webs of modality-specific cortices in the brain, which are activated together in a controlled manner to represent a concept [9], [13], [14], [40], [41]. The initial proposal is in fact quite old, credited to Wernicke and Meynert (see for instance [42] and [43] for a detailed discussion). They have proposed that concepts are made of modality-specific engrams, which reside in their corresponding primary sensory or motor cortices. Since these engrams are fully connected, any hint of the concept, be its name, sound, or taste, would alight the whole web, bringing into mind the holistic knowledge about the concept.

The popularity still claimed by this view is no doubt due to the voluminous neuroimaging evidence collected in the meanwhile, and firmly supporting the fully-connected-web theory of concepts. Goldberg *et al.* [14] and Kellenbach *et al.* [40], for instance, have conducted systematic experiments demonstrating modality-specific cortex activations during semantic retrieval and decision-making tasks. Goldberg *et al.* found that retrieval of tactile knowledge activates the somatosensory, motor, and premotor cortices, while flavor-related knowledge activates the orbitofrontal region, visual information the ventral temporal cortex, and auditory information the superior temporal sulcus [14]. Kellenbach *et al.* [40] similarly showed enhanced posterior inferior temporal cortex activation for color judgments, posterior superior temporal gyrus activation for sound judgments, and right medial parietal cortex activation for size judgments. Moreover, similar results are found outside of the sensory domain as well. In his famous work, Pulvermuller demonstrated that motor and premotor cortices activate somatotopically during language use, specifically tongue-related area (peri-sylvian) showing activation when the subjects are reading the word “lick”, finger-area (lateral) for “pick”, and foot-area (dorsal) for “kick” [9], [13]. The somatotopy discovered is important in terms of implying a systematic activation of the motor and premotor areas in a category-dependent manner. Chao and Martin [44] have also conducted a tool viewing-and-naming task to show selective

activation in left ventral premotor, as well as left posterior parietal cortex. These findings especially make sense when considering that grasping a tool for using it is an integral part of the tool concept, therefore spatial and motor areas are highly relevant to its semantics.

Recently, very strong support for concept web theories has also come from a predictive study [45]. Mitchell *et al.* showed that it is possible to predict the fMRI activation for complex words, such as *celery*, by superposing known fMRI activations from a previously determined set of 25 basic words. These basic words include *eat, taste, see, hear, smell, manipulate, touch, say, and move*, and interestingly, can simply be added together, each multiplied with a weight of its own co-occurrence with the target word in a large text corpus. These findings also support the representation of concepts through a combined web of other concepts.

Given these findings, it is not surprising that many theories on the neural implementation of concepts focus on a connected-web-of-cortical-areas basis (For instance see Pulvermuller [13], Damasio [41], Bryson [32], and Deacon [15]). It is indeed widely held that conceptualization is highly distributed in the brain with the primary areas as its modality-specific pillars. The elegance of this theory is its simplicity in seamlessly integrating both initial experience and subsequent retrieval. (Note also the relevance with [2] and [33].)

Still though, a number of studies have started to question if this is the whole story [43], [46]–[48]. Their focus is whether a connected web of primary cortices is enough to represent concepts, or if there is a dedicated region that orchestrates and connects these low-level cortex activations into a coherent concept meaning. Damasio in [41] and [49] hints a related idea when he mentions high-level, amodal convergence zones, from which the time-locked activation in primary cortices is orchestrated. Lambon Ralph [43] and Patterson *et al.* [46] build their theories on the lesion (and later neuroimaging) studies of Semantic Dementia (SD), a selective and progressive form of dementia, in which the semantic (categorical) knowledge is lost, with other cognitive abilities remaining intact. Quoting an instance recounted by Patterson *et al.* in [46]: “*When we asked one of our patients to name a picture of a zebra, she replied: ‘It’s a horse, ain’t it?’ Then, pointing to the stripes, she added, ‘But what are these funny things for?’*” In the case of semantic dementia, primary cortices and their association areas are intact, therefore the patient can decide the shown picture is a horse-like animal. Moreover she can also detect the stripes visually. However, the concept of a zebra is lost in her mind, therefore she converges on the next close concept that is still available. Another of their examples is a patient (who is competent in all other cognitive facilities) asking, “*What are those things?*” to a herd of sheep. This unusual form of dementia is connected to the degeneration of Anterior Temporal Lobe (ATL) by various studies (see for instance, [50]–[52]. Kellenbach *et al.* also recount an unexpected activation in ATL in a semantic task, which was not specifically searched for in the study, but is meaningfully accounted for in this hypothesis [40]).

The hypothesis proposed by Lambon Ralph [43] and Patterson *et al.* [46] is then whether ATL is a kind of “semantic hub”,

connecting the widespread web of concept into a meaningful entity. Their foundation is the often complex and nonlinear boundaries of concepts. They discuss that, although concepts are collections of features, these features usually bind together in nonlinear and complex ways. One example suggested by Lambon Ralph *et al.* [48] is the case of a single-layer neural network vs. a NN with a hidden layer. The single layer neural network can only bind together linear features, and is therefore unable to classify certain functions. However, one additional layer of representation allows the generation of any function. The progressive nature of semantic dementia also allows for such a test. Lambon Ralph [43] and Patterson *et al.* [46] tested their hypotheses by checking the categorization capabilities of mild and severe semantic dementia patients. They show that, SD patients show under- and over-generalization mistakes when categorizing relatively non-canonical objects (camels with humps, pumpkins as vegetables, etc.) These mistakes are even more prominent in severe SD, where patients tend to ignore all features of objects that are not prototypical of their category. (For instance, drawing ducks as if having four legs, as is typical of the animal category [46].) These interesting findings also suggest that conceptualization is a complex and core cognitive function that needs more investigation.

D. Structural Representation in Robotics

Concepts and their representation have inspired numerous computational and robotic studies as well, some of which try to unveil the mystery by presenting testable models, while others mainly aim to solve the perennial learning and adaptation challenge in robots. One of the most organized attempts of formalizing concepts for use in robotics came in the form of a knowledge processing framework, KNOWROB, proposed by Tenorth and Beetz [53], [54]. Their main point was to develop a system which can process information as efficiently as humans do, by filling in the gaps in conversation with background knowledge. The system can connect to external information sources, such as the Internet or dedicated databases, and possesses manipulators with which it can utilize accessed unformatted information in various tasks freely. Information is kept unformatted (“virtual”) until it is needed, and then can be searched freely for associations. Concepts in KNOWROB can be objects, actions, events, or places, and are organized in a hierarchical manner with more specific concepts inheriting from more general ones. Multiple inheritance is allowed, which enriches membership definitions. Actions are defined as recipes, events as change of states, and all of these are inherited from a general “thing” entity, which is the common ancestor of all nodes (object, action, event, or place) in the ontology. Later on, this system was extended by Palmia [55] with the aim of mutual understanding and cooperation between multiple robots. Another notable example is the utilization of syntactic bootstrapping [56], [57] for the robot to learn, from conversations with humans and online images depicting events, which actions can be used on which kinds of objects, effectively generalizing objects with respect to actions in the process. This approach allows the robots to conduct flexible reasoning on huge amounts of data acquired from the Internet,

as well as to perform error handling and/or guide the supervisor by asking questions if necessary.

On the other side of the fence, there are studies which aim to close the gaps in what we know of human cognition. Baxter *et al.* [58] propose a connected developmental architecture of conceptual memory. The membership of instances to concepts are defined in terms of the Euclidean distance of all features to the concept prototypes. They also learn associative links between different feature spaces in a developmental manner, reminiscent of Hebbian learning. (However, these associative links connect only different modalities of the same concept, and not different but semantically related concepts.) In yet another attempt to bring together different modalities, Morse *et al.* [59] use the “body” of an agent as a “hub” to connect the visual, auditory, and spatial information, enabling the grounding of concepts such as red and cup. [37], [60]–[62] use the formalization of affordances to ground actions.

Another prevalent approach for conceptual representation in robotics is assuming that concept formation occurs in an incremental manner in the form of a hierarchical structure; *i.e.*, a hierarchical representation is assumed of concepts [63]–[66]. In this hierarchy, upper concepts represent the general concepts, whereas lower or terminal concepts refer to the specific properties or instances. The connections imply *is-a* type relations between concepts. Instances can be placed into lower or terminal nodes. The tree structure of the hierarchy also provides an option for branching. Top-down classification of instances depends on selecting the best branch or set of branches to go deeper in a tree, similar to Quinlan’s decision tree approach [67]. One of the earliest attempts is the Elementary Perceiver And Memorizer (EPAM) model [63], [64], which holds nodes with attribute-value pairs in a tree structure. Each edge coming out of a node represents a certain value for a comparison criteria. Leaf nodes correspond to specific *images* of instances. EPAM makes a distinction between classification and prediction tasks as two different processes. This model was later extended by UNiversal MEMory (UNIMEM) [65] to include confidence and feature frequency statistics, nominal values and images in non-terminal nodes. COBWEB [66] has been inspired by these models, as well as CYRUS [68], and introduced an evaluation function which rewards intra-class similarity and inter-class dissimilarity. Finally, CLASSIT [64] enhanced COBWEB by including mean and standard deviation values for attributes. The common point in all these hierarchical methods is that they use the hill-climbing search method and each concept node has its own attribute-value pairs.

The common missing point in all these works is the lack of a global structure of associativity. The connections are restricted to joining either different modalities of the same concept, or a group of concepts stemming from the same ancestor, such as “cup” as a container, and “glass” as another container. Yet, none of these models present a feature of long-distance associativity between seemingly different, but semantically related concepts, such as “water” and “glass”, which should be related by means of the “drinking” action. Moreover, the concepts these studies are generally not grounded: They rely on either ontologies or Internet-based information, or a hand-

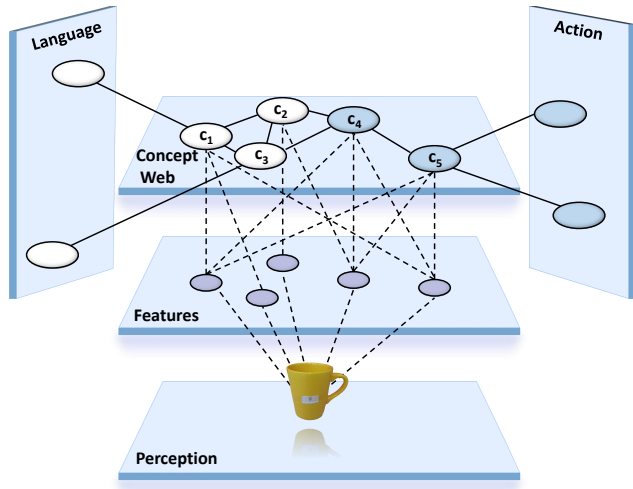


Fig. 1: A schematic presentation of the concept web, which connects related concepts to each other and to their counterparts in the language, action, and perception spaces. Information can flow in from the perception space, through a feature extraction mid-level, or from the language and action spaces as well. A number of nodes are randomly illustrated in white to exemplify active concepts.

designed set of features. Therefore, they lack extendibility and verifiability from a developmental point of view [31].

E. Contributions of the Current Study

The contribution of our study is the modeling of a grounded concept web on a humanoid robot. The web consists of adjective, noun and verb concepts that are abstracted from the sensorimotor interactions of the robot. These concepts are connected to each other via *is-a* type relations, as well as to language, action, and perception. These inter-concept relations are extracted from the interactions of the robot, allowing it to observe co-occurrences between the concepts through these interactions (Figure 1). From our previous analysis in [69], we expect this kind of dependency information to be useful when trying to make sense of the complex real world.

Modeling the concept web requires a multi-label classification problem and our approach to this is a probabilistic one: We use Markov Random Field (MRF) to model the web and the inferences are made using Loopy Belief Propagation (LBP). Our choice for MRF and LBP is due to: (i) The natural resemblance between an MRF and a concept web. In MRF, there are nodes that are fed observations and the nodes regularize other nodes using links between them. This is exactly the requirement for a concept web, which needs to be linked to the perceptual input, and which must spread activation between the concepts. Such a representation will be able to employ the dependency information between the concepts. (ii) The hypothesis that our brains are biological machines that might be encoding information as probability distributions and function using probabilistic inferences (see, *e.g.*, [70], [71]). In fact, it is known that we use probabilistic and statistical mechanisms for language learning and understanding [72]–

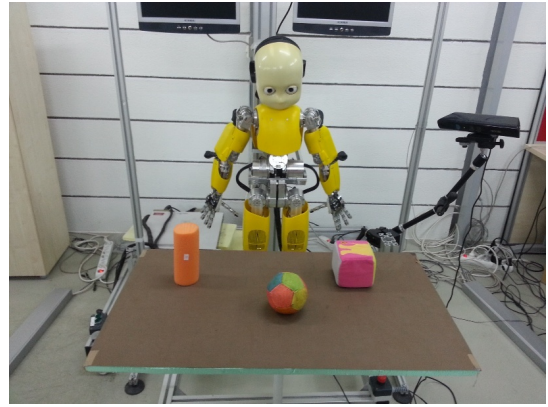


Fig. 2: The experimental setup consisting of the iCub and the Kinect Depth Camera.

[75]. Our approach is in line with findings in Psychology and Neuroscience, which strongly support the philosophical concept web hypothesis by contemporary imaging evidence [9], [13], [14], [40], [45], demonstrating the possible neurological mechanisms for a web structure in the brain. (iii) The ability of MRF to support inference on highly cyclic structures, which is the case in our system due to the densely connected nature of concepts.

Our study is unique in robotics since, even though there are concepts and structured concepts used in robotics studies (as reviewed above), they are hierarchical, hand-designed and usually not grounded. We, on the other hand, focus on modeling the concept web, starting from low-level sensorimotor data, constructing concepts from interactions, and linking those concepts using co-occurrences between concepts in order to build a web of concepts.

An earlier version of this work was published in a conference [69], where our focus was only on whether co-occurrence between nouns and adjectives can improve their prediction accuracies. For the sake of completeness, we also present the extraction process and the representation of individual concepts in our system, using a prototype scheme developed in previous work [38], [69], [76]. However, this specific representation of individual concepts is not central to the contributions of this study, and it can be replaced with any other representation or categorization scheme with no loss of generality. The learning of a concept web from the inter-relations of individual concepts, and showcasing the advantages of using such a representation in different scenarios are the main contributions of this study.

III. EXPERIMENTAL SETUP

We perform our experiments on the iCub humanoid robot platform (Figure 2) [37]. A Kinect device is used to allow iCub to perceive the world and the objects. The 3D point cloud acquired from the Kinect device is processed using the Point Cloud Library (PCL) [77]. The tactile information is collected from the pressure sensors of iCub, placed on each fingertip. Finally, for audio information, we use an external microphone attached to iCub's belly, which allows us to acquire only the



Fig. 4: Objects for each adjective category

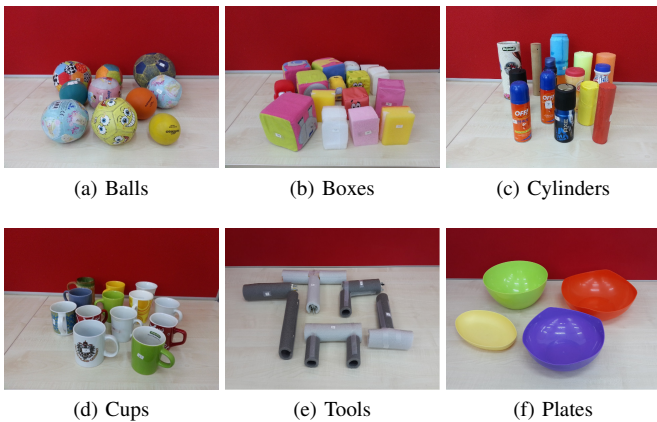


Fig. 3: Objects for each noun category.

relevant audio signal and skip dealing with the noise coming from iCub itself.

A. Objects

We have in total 60 objects, arbitrarily divided into a training (45 objects) and a testing set (15 objects). Each object is labeled with one of the 6 noun groups $\{box, ball, cylinder, cup, plate, tool\}$ (Figure 3), and 5 adjectives out of 5 dichotomic pairs: $\{hard \times soft, noisy \times silent, tall \times short, thin \times thick, round \times edgy\}$ (Figure 4). The mappings between nouns and adjectives are not necessarily 1-to-1, but are semantically sound, e.g., a box can be either tall or short, but it is always edgy. An object cannot be defined with two conflicting adjectives (e.g., both thin and thick) at the same time.

In our framework, since the web is constructed based on co-occurrences, the co-occurrences of concepts explicitly affect our noun and adjective predictions. The co-occurrence values for our dataset are shown at Table I.

B. Behaviors

The robot is equipped with 13 behaviors: *push left, push right, push forward, push backward, move left, move right,*

TABLE I: The frequencies of instances in the dataset in which specified noun and adjective pairs co-occur together (out of 60 objects in the dataset).

	Hard	Soft	Noisy	Silent	Tall	Short	Thin	Thick	Round	Edgy
Box	2	14	2	14	0	16	0	16	0	16
Ball	3	7	7	3	0	10	1	9	10	0
Cylinder	14	0	5	9	10	4	9	5	14	0
Cup	11	0	1	10	0	11	0	11	11	0
Tool	5	0	5	0	5	0	0	5	5	0
Plate	4	0	0	4	4	0	0	4	4	0

TABLE II: Possible applicable set of behaviors with respect to object categories. *arg*: *Left, Right, Forward, Backward*; A: *Applicable*; NA: *Not-Applicable*

	<i>Push</i> (<i>Left, Right, Forward, Backward</i>)	<i>Move</i> (<i>Left, Right, Forward, Backward</i>)	<i>Drop</i>	<i>Grasp</i>	<i>Shake</i>	<i>Knock down</i>	<i>Throw</i>
Box	A	A	A	A	A	A	A
Ball	A	A	A	A	A	A	A
Cylinder	A	A	A	A	A	A	A
Cup	A	A	NA	A	NA	NA	NA
Tool	A	A	A	A	A	A	A
Plate	A	A	NA	A	NA	NA	NA

move forward, move backward, grasp, knock down, throw, drop, and shake.

To further investigate the connections between the objects and the behaviors, we assumed that not all behaviors are applicable to all objects. For instance, cups and plates are assumed to be “fragile”, or possibly containing liquids, therefore they cannot be shaken, dropped, thrown, or knocked down. The list of allowable behaviors for each noun category is shown in Table II.

C. Features and Data Collection

While collecting the data, we perform the following procedures for each object $o \in \mathcal{O}$:

- 1) Place an object o at a random initial position on the table.
- 2) Store the initial visual features \mathbf{e}_v .
- 3) For each applicable behavior $b \in \mathcal{B}$ to object o (Table II):
 - Collect the initial visual features \mathbf{e}_v^b for behavior b .
 - Apply behavior b once.
 - If behavior b is *grasp*, collect audio \mathbf{e}_a , haptic \mathbf{e}_h , and proprioceptive \mathbf{e}_p features while the behavior is in progress, and concatenate them with initial

TABLE III: The visual, audio, haptic and proprioceptive features extracted from the interactions of the robot.

Feature Type	Feature	Position
Visual (\mathbf{e}_v)	Position:(x, y, z)	1-3
	Object dimensions:($width, height, depth$)	4-6
	Object presence $\in \{1, -1\}$	7
	Normal zenith histogram bins	8-27
	Normal azimuth histogram bins	28-47
	Shape index histogram bins	48-67
Audio (\mathbf{e}_a)	13 bins of MFCC (max - min)	68-80
Haptic (\mathbf{e}_h)	Change for index finger	81
	Min values for index finger	82
	Max values for index finger	83
	Mean for index finger	84
	Variance for index finger	85
	Standard deviation for index finger	86
Proprioceptive (\mathbf{e}_p)	Change for index finger	87
	Min values for index finger	88
	Max values for index finger	89
	Mean for index finger	90
	Variance for index finger	91
	Standard deviation for index finger	92

visual features \mathbf{e}_v to obtain entity feature vector $\mathbf{e} = \mathbf{e}_v \cdot \mathbf{e}_a \cdot \mathbf{e}_h \cdot \mathbf{e}_p$.

- Collect the final visual features \mathbf{e}_v^b for behavior b .
- Obtain the effect feature vector for the behavior b by $\mathbf{f} = \mathbf{e}_v^b - \mathbf{e}_v^b$.
- If the final position of the object is out of reachable bounds, randomly reinitialize object position for the next behavior.

From the interactions, we extract the features that are listed in Table III. The first seven visual features come from basic position information and three dimensional properties of the object. The following 40 features come from the zenith and azimuth normal vectors of each point on the object. Shape index histogram bins, forming the majority of the visual features, are created using the shape index property of the object. Initially, the maximum and minimum principal curvatures (Q_{max} and Q_{min}) of the object are calculated. These two curvature measurements define the basic type of a surface (e.g., saddle, plane, etc.). The shape index property is directly dependent on these curvatures and is obtained by $\frac{Q_{max} + Q_{min}}{Q_{max} - Q_{min}}$.

From the raw audio data, we first extract the MFCC (Mel-Frequency Cepstrum Coefficients) features. We chose MFCC because it is widely used for sound classification (see, e.g., [78]). MFCC returns a 13-dimensional feature vector for every 21ms, and we use the difference between the maximum and the minimum of all MFCC vectors for the audio signal extracted from each object. The audio signal is collected during the *grasp* behavior.

Haptic and proprioceptive features are obtained from the index finger¹ of iCub during the *grasp* behavior. For haptic features, the first feature is the difference between the final and the initial value of the tactile sensor. The second and third features are the minimum and maximum of haptic values. The following three features are mean, variance, and standard deviation, for capturing the change of the haptic signal during the interaction. The same features are extracted from the

proprioceptive signals.

We refer to the combination of these features as the *entity feature vector*, and denote it by \mathbf{e} . In other words, \mathbf{e} is the concatenation of visual \mathbf{e}_v , audio \mathbf{e}_a , haptic \mathbf{e}_h and proprioceptive \mathbf{e}_p features (Table III). For verb concepts, following our previous work [38], we try to capture the change in the visual features, which we refer to as the *effect feature vector* (\mathbf{f}). The effect feature vectors are obtained from the differences between the final and initial visual features, i.e., $\mathbf{f} = \mathbf{e}_v^b - \mathbf{e}_v$. See Figure 5 for an illustration of the entity and effect feature vectors.

IV. INDIVIDUAL CONCEPTS

Our experimental framework corresponds to a world inhabited with three kinds of concepts, namely noun, adjective, and verb concepts. Let the set of concepts be denoted $\mathbb{C} = \mathbb{N} \cup \mathbb{A} \cup \mathbb{V}$, with the set of noun concepts $\mathbb{N} = \{box, ball, cylinder, cup, plate, tool\}$, the set of adjective concepts $\mathbb{A} = \{hard \times soft, noisy \times silent, tall \times short, thin \times thick, round \times edgy\}$, and the set of verb concepts $\mathbb{V} = \{push left, push right, push forward, push backward, move left, move right, move forward, move backward, grasp, knock down, throw, drop, shake\}$. As a first step, the robot needs to identify these concepts from its interaction with the world, so that it can impose some structure to its environment. For identifying, and then representing individual concepts, we use a prototype scheme developed in [38], [69], [76], and demonstrated in [79] to be comparable in terms of performance to a number of widely used approaches, including Support Vector Machines, Self Organizing Maps, AdaBoost, etc. For completeness, we present the details of this representation, although it is not an original contribution of this paper. An alternative approach could as well be used for representing individual concepts without fundamentally changing the framework proposed in this article.

A. Conceptualization of a Category

We define concepts by their prototypes (see Section II-A for theories of concepts) following our previous work on prototype-based conceptualization of verbs, nouns and adjectives [38], [69], [76]. The prototypes are extracted from the training instances, which are labeled with human supervision a priori. Each training instance is labeled with exactly 1 noun label and 5 adjective labels (i.e., one of the two antonyms in each pair). In addition, each behavior is applied once to every allowed object, and this interaction is labeled with the related verb concept. Inter-behavior variances are allowed in order to capture the generic effect of the behavior, e.g., the objects are initialized to an arbitrary position before the interaction.

The prototypes are designed to indicate the contribution of each feature to the concept. The noun and adjective prototypes are extracted from the entity features vectors \mathbf{e} of the training data; the verb prototypes are extracted from the effect feature vectors \mathbf{f} of the interactions. Features that contribute consistently positively to the concept are indicated with a '+' sign in the prototype. Similarly, features that contribute negatively

¹From our experience, this finger turned out to provide sufficient information for detecting a grasp.

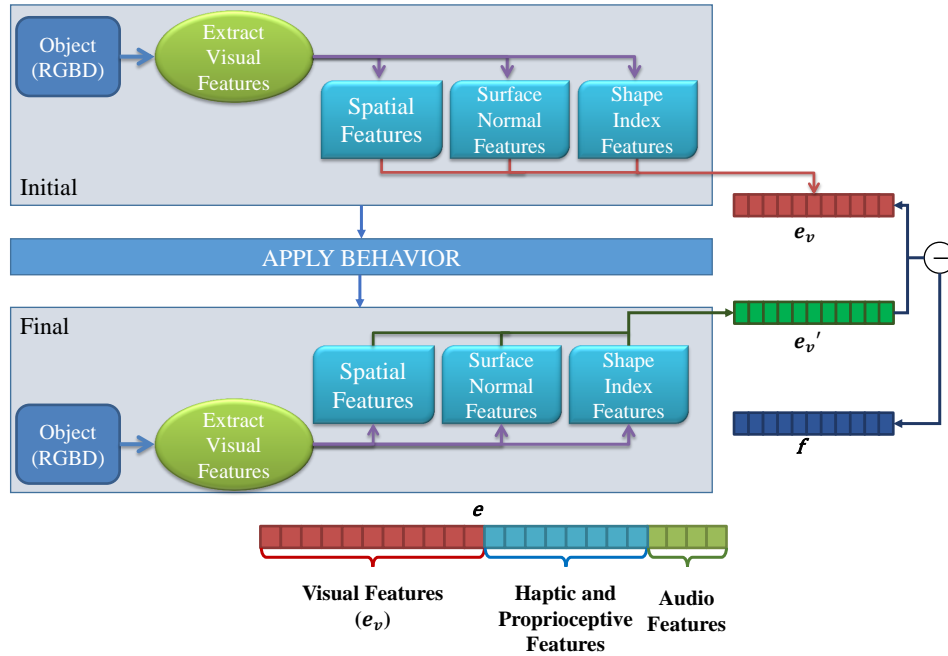


Fig. 5: Extraction of the entity and effect feature vectors. e_v and e'_v are the visual features of an object before and after a behavior is applied. $f = e'_v - e_v$ is the effect feature vector. e is the multi-modal entity feature vector, composed of visual, haptic, proprioceptive, and audio information.

TABLE IV: Extracted prototypes for noun, adjective and verb concepts.

Concepts		Visual Features	Audio Features	Haptic Features	Proprioceptive Features
Nouns	Box	+++++-----+-----*-----+-----+-----	-----*	*-*-*	*****
	Ball	++-----+++-----+-----+-----+-----	*	-+---	*****
	Cylinder	+++-----*+-----+-----+-----+-----	*****-***	+*****	-+++--
	Cup	++-+-----+-----+-----+-----+-----	*-*-*-*-*-*	+-----	-+++--
	Plate	+++++-----+-----*-----*-----*-----+-----	-----+-----	+-----	-**--
	Tool	*****-----+-----+-----+-----+-----	+++++*****	+-----	-+++--
Adjectives	Hard	+++++-----+-----+-----+-----+-----	*****+*****	+*+*+*	-+++--
	Soft	++-+-----+-----+-----+-----+-----	++-+-----+-----	+*+*+*	+++++
	Noisy	*+*+*+-----+-----+-----+-----+-----	+++++*****	+*+*+*	*****
	Silent	+++++*+-----+-----+-----+-----+-----	-----+-----+-----	+*+*+*	*****
	Short	++-+-----*+-----+-----+-----+-----	*-*-*-*-*-*	+*+*+*	*+*+*
	Tall	+++++*+-----+-----+-----+-----+-----	*****+*****	+*+*+*	-+++--
	Thick	+++++*+-----+-----+-----+-----+-----	*****+*****	+*+*+*	*****
	Thin	++-+-----+-----+-----+-----+-----	*-*-*-*-*-*	+*+*+*	-+++--
	Edgy	++-+-----+-----+-----+-----+-----	-----+-----+-----	+*+*+*	*****
	Round	+++++*+-----+-----+-----+-----+-----	*****+*****	+*+*+*	*****
Verbs	Grasp	00+-0*-000+00-----0-----000-00-0-----0-----00++		None	
	Knock Down	0+0000+0-000000-----0-----000-00-0-----0-----++++		None	
	Move Left	0-0000+0-000+00-----0-----000-00-000-0-----++++		None	
	Move Right	0+0000+-00++00-----0-----000-00-000-0-----++++		None	
	Move Forward	-00000+-000+00-----0-----000-00-000000-----++++		None	
	Move Backward	+00000+0-00++00-----0-----000-00-000-0-----++++		None	
	Push Left	000-0*-0-000+00-----0-----000-00-0-----0-----++++		None	
	Push Right	0+0000+-00++00-----0-----000-00-000-0-----++++		None	
	Push Forward	-00000*-000+00-----0-----000-00-0-----0-----++++		None	
	Push Backward	+00-0*-0-000+00-----0-----000-00-000-0-----0+++		None	
	Drop	**+000*0-000000-----0-----000-00-0-----0-----00++		None	
	Throw	*0+000*0-000+00-----0-----000-00-0-----0-----0+++		None	
	Shake	000000*0-000000-----0-----000000-0-00-0-----0++++		None	

are indicated with a '-' sign, and those whose contribution show too much variation are denoted with a '*'. For noun and adjective concepts, '+/-' indicates characteristically high/low values for the associated dimension, whereas '*' indicates irrelevant features that can be discarded from the comparisons regarding the concept. Meanwhile, for verb concepts, '+/-' indicates characteristically increased/decreased features through the application of the behavior, while '*' indicates the changes induced on that dimension are inconsistent. Verb prototypes also include an additional marker '0', which indicates that the feature is not changed significantly by the behavior.

The contributions of each feature are decided by clustering for each concept the mean and variance values of the feature. The mean and variance values are decided regarding the features of all the training instances that are labeled with the specific concept and normalized to allow meaningful comparison. The features are then clustered in the mean-variance space using Robust Growing Neural Gas (RGNG) algorithm [80], which gives the features with (1) high mean and low variance, thereby labeled with '+', and (2) low mean and low variance, thereby labeled with '-', and (3) high variance, thereby labeled with '*'. For verb concepts,

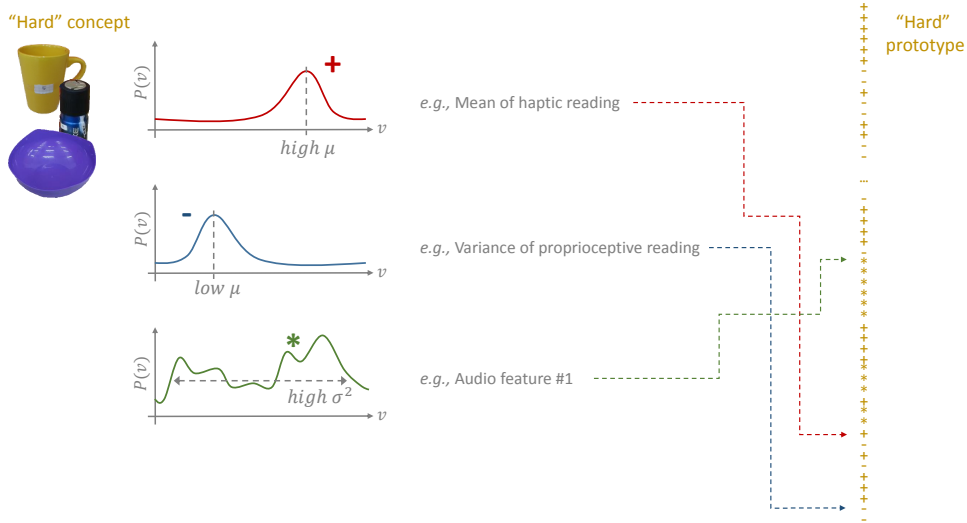


Fig. 6: Schematic visualization of the extraction of a concept prototype. If a feature has a consistently high contribution, marked with a high mean and low variance distribution, it is indicated with a ‘+’ sign. Those with a consistently low contribution, marked with a low mean and low variance distribution, are assigned a ‘-’ sign, whereas those with high variance are marked with a ‘*’ to indicate inconsistent contribution. Sample features are illustrated for the *hard* concept.

Algorithm 1 Derivation of a concept prototype from associated training instances (Adapted from [38]).

for all concepts $c \in \mathbb{C}$ **do**
for all feature dimensions d **do**
 • Compute the mean μ_{cd} :

$$\mu_{cd} = \frac{1}{N} \sum_{i \in \mathbb{I}(c)} i_d, \quad (2)$$

where $\mathbb{I}(c)$ is the set of training instances of concept c , with cardinality $N = |\mathbb{I}(c)|$, and i_d is the d^{th} feature of instance i .

• Compute the variance σ_{cd}^2 :

$$\sigma_{cd}^2 = \frac{1}{N} \sum_{i \in \mathbb{I}(c)} (i_d - \mu_{cd})^2. \quad (3)$$

end for

• Concatenate μ_{cd} ’s and σ_{cd}^2 ’s to obtain the vectors μ_c and σ_c^2 .

end for

for all concepts $c \in \mathbb{C}$ **do**

• Apply Robust Neural Growing Gas algorithm in $\mu_c \times \sigma_c^2$ space:

if $c \in \mathbb{N} \cup \mathbb{A}$ **then**

- Manually assign one of the labels ‘+’, ‘-’, or ‘*’ to the dimension d , considering the cluster that d falls into:

if cluster is high on μ axis and low on σ^2 axis **then**

assign ‘+’ to d

else if cluster is low on both μ and σ^2 axes **then**

assign ‘-’ to d

else if cluster is high on σ^2 axis **then**

assign ‘*’ to d

end if

else

- Manually assign one of the labels ‘+’, ‘-’, ‘*’, or ‘0’ to the dimension d , considering the cluster that d falls into:

if cluster is high on μ axis and low on σ^2 axis **then**

assign ‘+’ to d

else if cluster is low on both μ and σ^2 axes **then**

assign ‘-’ to d

else if cluster is close to 0 on μ axis and low on σ^2 axis **then**

assign ‘0’ to d

else if cluster is high on σ^2 axis **then**

assign ‘*’ to d

end if

end if

end for

negligible mean value and low variance combination is labeled with ‘0’. Figure 6 demonstrates a sample clustering and labeling case for the *hard* concept. The exact procedure is depicted in Algorithm 1, and prototypes extracted and used in this study are shown in Table IV.

B. Category Prediction from Prototypes Only

The prediction procedure takes as input the above prototypes of concepts and the feature vector (\mathbf{e} or \mathbf{f}) of a new object or an interaction, denoted with an \mathbf{x} . When evaluating membership for a concept, only meaningful features (which are labeled with ‘+’, ‘-’ or ‘0’ in the corresponding prototype) are considered. On these meaningful dimensions, the Euclidean distance to the mean values of the concept’s prototype is calculated as follows:

$$D(c, \mathbf{x}) = \frac{1}{|\mathcal{R}_c \setminus \mathcal{R}_c^*|} \sqrt{\sum_{i \in \mathcal{R}_c \setminus \mathcal{R}_c^*} (\mathbf{e}_x^i - \mu_c^i)^2}, \quad (4)$$

where \mathbf{x} is the new instance, \mathcal{R}_c is the set of all feature indices, \mathcal{R}_c^* is the set of indices that are ‘*’-signed (*i.e.*, inconsistent) for concept c ; $|\cdot|$ is the cardinality measure, \mathbf{e}_x^i is the i^{th} feature of instance \mathbf{x} , and μ_c is the mean feature vector of training objects labeled with concept c .

$D(c, \mathbf{x})$ is the closeness of the new instance to the selected concept. We can convert it to the probability estimate of instance \mathbf{x} belonging to the concept c as follows:

$$s_{perc}(c, \mathbf{x}) = \frac{\prod_{r \in \mathbb{C}_i \setminus \{c\}} D(r, \mathbf{x})}{\sum_{r \in \mathbb{C}_i} \left(\prod_{r_t \in \mathbb{C}_i \setminus \{r\}} D(r_t, \mathbf{x}) \right)}, \quad (5)$$

where $\mathbb{C}_i \subset \mathbb{C}$ is either the set of nouns $\mathbb{N} = \{\text{box, ball, cylinder, cup, tool, plate}\}$, the set of one dichotomic pair of

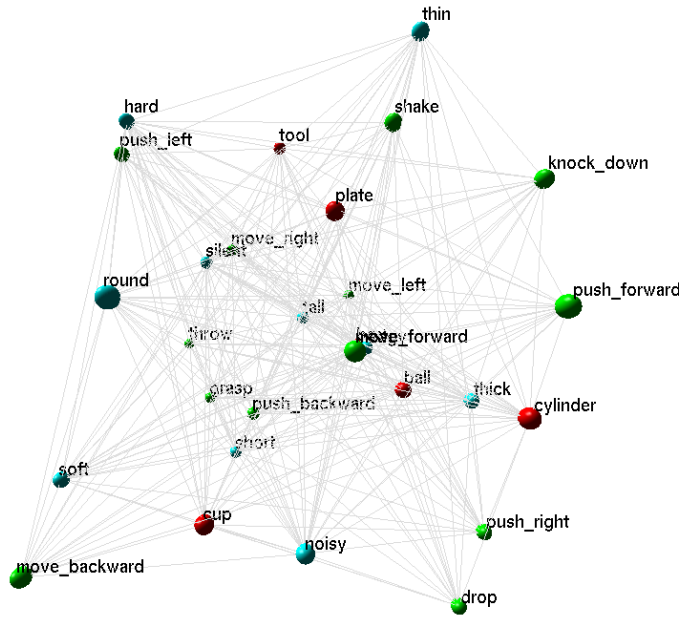


Fig. 7: A snapshot of the concept web iCub has constructed. Connections between related concepts are denoted with gray links. Noun concepts are indicated with red, adjective concepts with blue and verb concepts with green. The graph is created using Ubigraph graph visualization library [81]. [Best viewed in color]

adjectives *e.g.*, $\mathbb{A}_p = \{hard, soft\}$, or the set of verbs $\mathbb{V} = \{push\ left, push\ right, push\ forward, push\ backward, move\ left, move\ right, move\ forward, move\ backward, grasp, knock\ down, throw, drop, shake\}$. Note that this calculation normalizes the belonging probabilities among each concept group. Another important point is that $s_{perc}(\cdot, \cdot)$ depends only on the features extracted from the instance, and does not utilize information about the co-occurrences of the concepts.

V. A PROBABILISTIC WEB OF CONCEPTS

In the previous section, we described the conceptualization of individual categories. In this section, we discuss how we represent the concept web in a probabilistic model, namely, Markov Random Field, which is especially suitable for our purposes due to its ability to conduct inference on densely connected graph structures. Each node of the constructed Markov Random Field corresponds to a concept (noun, adjective, or verb) in our web.

A. Building a Web from Individual Concepts

With $\mathbb{C} = \mathbb{N} \cup \mathbb{A} \cup \mathbb{V}$ the set of all concepts, let us denote W to be the concept web constructed from the interactions of the robot. The web W can be represented as a graph $G(\mathbb{C}, \mathbb{E})$, where each concept $c \in \mathbb{C}$ is treated as a node in W .

The edges \mathbb{E} are established based on the co-occurrences of the concepts. Namely, an edge $E(c_i, c_j) \in \mathbb{E}$, between concepts c_i and c_j , is placed in the web if c_i and c_j have co-occurred in an interaction. The web constructed from the interactions in this study is visualized in Figure 7.

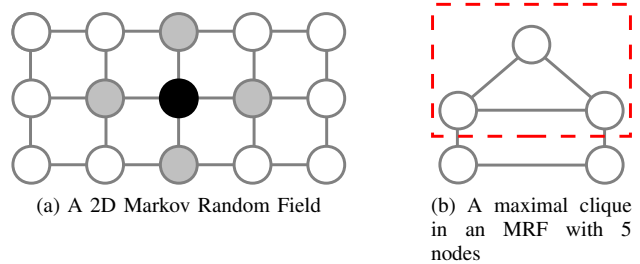


Fig. 8: (a) A sample 2D Markov Random Field. The Markovian property holds in Markov Random Fields, by which a random variable (*i.e.*, the black node), given its immediate neighbors (the gray nodes), is independent of all other random variables. (b) A maximal clique (with 3 nodes) is indicated in an MRF with 5 nodes.

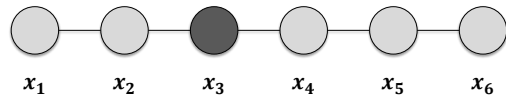


Fig. 9: Sample Markov Random Field chain of variable nodes.

B. Markov Random Field

Markov Random Field (MRF) [82] is a probabilistic graphical model widely used for defining constraints on and between entities in a problem. The entities are represented as nodes and the constraints between the entities are incorporated by the edges connecting them. MRF follows the Markovian property that the state of a node depends only on the neighboring nodes (Figure 8a). Due to these representational constraints,

all probabilistic functions are defined over *maximal cliques*. A *clique* is a subset of nodes that are connected to each other directly, and a *maximal clique* is a clique with the highest number of nodes possible (Figure 8b).

An MRF effectively models the following joint probability distribution:

$$P(\omega) = \frac{1}{Z} \exp(-U(\omega)), \quad (6)$$

where $\omega \in \Omega$ is a possible configuration of the web W , $U(\omega)$ is the energy function of the MRF given a configuration ω , and Z is the normalizing partition function defined over the set of all possible configurations Ω :

$$Z = \sum_{\omega \in \Omega} \exp(-U(\omega)) \quad (7)$$

The energy function $U(\omega)$ is in turn defined as the combination of a data term and a smoothness term:

$$U(\omega) = U_{data}(\omega) + U_{smooth}(\omega), \quad (8)$$

where the data term, $U_{data}(\omega)$, asserts the consistency of the configuration ω with the immediate measurements, while the smoothness term $U_{smooth}(\omega)$ enforces consistency with the a priori knowledge previously encoded into the MRF in the form of clique connections.

C. Belief Propagation in MRF

The potential values in the MRF model demonstrate the correlation between two connected nodes. In belief propagation methods, this correlation is thought of as a *message* from one node to an adjacent one. To demonstrate belief propagation, let us calculate the marginal probability distribution over a node (x_3) for the MRF in Figure 9:

$$p(x_3) = \frac{1}{Z} \sum_{x_2} \psi(x_2, x_3) \sum_{x_1} \psi(x_1, x_2) \sum_{x_4} \psi(x_3, x_4) \sum_{x_5} \psi(x_4, x_5) \sum_{x_6} \psi(x_5, x_6). \quad (9)$$

If we treat the terms in Equation 9 as messages from the adjacent nodes of query node x_3 , then we can re-formulate it, merging these messages as follows:

$$p(x_3) = \frac{1}{Z} \left[\sum_{x_1, x_2} \prod_{i=1}^2 \psi(x_i, x_{i+1}) \right] \cdot \left[\sum_{x_4, x_5, x_6} \prod_{i=3}^5 \psi(x_i, x_{i+1}) \right], \\ = \frac{1}{Z} \mu_{x_2}(x_3) \mu_{x_4}(x_3), \quad (10)$$

where $\mu_{x_2}(x_3)$ is the message to x_3 from x_2 , and $\mu_{x_4}(x_3)$ is the message for x_3 from x_4 .

D. Inferences in Concept Web Using Loopy Belief Propagation

Our concept web W is a cyclic graph by definition, and therefore, making exact inferences given observations is not possible. For such problems, approximate solutions are used, and a widely-used method for this task is Loopy Belief Propagation (LBP) [83]–[85], which iteratively updates the influence of one variable (*i.e.*, concept) on another until convergence.

The influence of one variable on another is called a *message*, and this process is called *message passing*.

LBP re-factorizes the graph into separator nodes and clique nodes - see Figure 10a for an example. Clique nodes are shown as elliptic nodes, whereas separator nodes are symbolized with square nodes. Separator nodes are in fact the concepts in the web, whereas the clique nodes represent the potential of a clique as a single node.

The message passing procedure in LBP differs in many ways when compared to standard belief propagation. For instance, the graph is divided into sub-trees, each of which includes one clique node and the separator nodes connected to it (Figure 10b). After extracting the sub-trees, LBP performs the following until convergence:

- 1) **Update Clique Potentials:** Updating the clique potentials can be thought as message passing from connected separator node to the clique node. Therefore, we can compute the new potentials by multiplying the potentials of separator nodes with the previous value of potentials in the clique node:

$$\mathcal{V}_{\mathcal{K}}^*(\mathbf{x}_{\mathcal{K}}) = \mathcal{V}_{\mathcal{K}}(\mathbf{x}_{\mathcal{K}}) \prod_{x_m \in ne(\mathbf{x}_{\mathcal{K}})} \mathcal{V}_m(x_m), \quad (11)$$

where $\mathbf{x}_{\mathcal{K}}$ is the set of random variables in clique node \mathcal{K} , $ne(\mathbf{x}_{\mathcal{K}})$ is the set of neighboring separator nodes of clique \mathcal{K} , $\mathcal{V}_{\mathcal{K}}(\mathbf{x}_{\mathcal{K}})$ is the previous potential of the clique, and $\mathcal{V}_{\mathcal{K}}^*(\mathbf{x}_{\mathcal{K}})$ is its updated potential.

- 2) **Update Separator Potentials:** After updating the clique potentials, we apply the message passing in the reverse direction. This time, updating the separator potentials is different from updating the clique potentials in that the message from the updated clique node to any one of the connected separator nodes is calculated by summation of the potentials of the clique nodes except the separator node:

$$\mu_{\mathcal{K}^* \rightarrow x_m}(x_m) = \sum_{\mathbf{x}_n \in \mathbf{x}_{\mathcal{K}} \setminus x_m} \mathcal{V}_{\mathcal{K}}^*(\mathbf{x}_n). \quad (12)$$

If the potential of separator node x_m has been updated previously, the new potential value is the multiplication of the previous node potential with the new message from the clique node, divided by the previous one:

$$\phi_s^*(x_m) = \phi_s(x_m) \frac{\mu_{\mathcal{K}^* \rightarrow x_m}(x_m)}{\mu_{\mathcal{K} \rightarrow x_m}(x_m)}. \quad (13)$$

Otherwise, it is directly set to the new value:

$$\phi_s^*(x_m) = \phi_s(x_m) \mu_{\mathcal{K}^* \rightarrow x_m}(x_m). \quad (14)$$

where $\mu_{\mathcal{K}^* \rightarrow x_m}(x_m)$ is the new message, $\mu_{\mathcal{K} \rightarrow x_m}(x_m)$ is the previous message, $\phi_s(x_m)$ is the previous potential of the separator node x_m , and $\phi_s^*(x_m)$ is the updated potential.

- 3) **Iteration:** The previous two steps are iterated for all clique nodes and their separator nodes until the change in the potentials is less than a threshold.

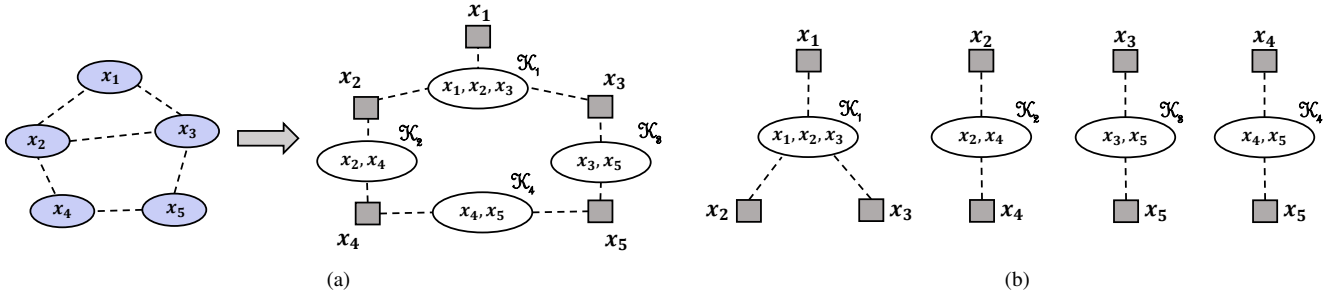


Fig. 10: (a) The representation of an MRF graph as a LBP graph. (b) Divided sub-trees of the graph in Figure 10a.

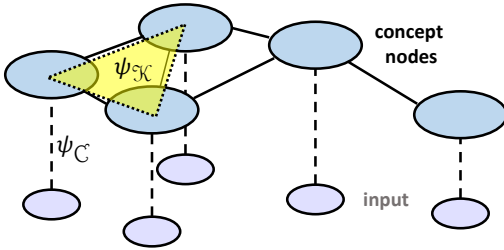


Fig. 11: A schematic visualization of the MRF energy formalization of the concept web. The sum of individual concept potentials ψ_c gives the data term, and the sum of clique potentials ψ_K gives the smoothness term.

E. The Concept Web as a Markov Random Field

For formalizing the concept web as an MRF, we define the energy function of Equation 8 as follows:

$$\begin{aligned} U(\omega) &= U_{data}(\omega) + U_{smooth}(\omega) \\ &= \sum_{c \in \omega} \psi_c(c) + \sum_{\mathcal{K} \in \mathbb{K}} \psi_{\mathcal{K}}(\mathcal{K}, \omega), \end{aligned} \quad (15)$$

with \mathbb{K} denoting the set of all cliques, c is the set of all active concepts in the given configuration ω , ψ_c is the potential of each active concept c , and ψ_K is the potential of each clique K . (See Figure 11 for a schematic visualization.) The data term, expanded as $U_{data}(\omega) = \sum_{c \in \omega} \psi_c(c)$ is responsible with ensuring consistency with the raw perceptions of concepts. Therefore, we define the concept potential ψ_c as:

$$\psi_c(c) = D(c, \mathbf{x}), \quad (16)$$

with \mathbf{x} being the incoming instance, and $D(c, \mathbf{x})$ its distance to the active concept c (Equation 4). The smoothness term, $U_{smooth}(\omega) = \sum_{\mathcal{K} \in \mathbb{K}} \psi_{\mathcal{K}}(\mathcal{K}, \omega)$, asserts consistency with a priori knowledge, which is encoded in terms of edge information as mentioned in Section V-A. The frequently co-occurring concepts are connected by edges, which determine the clique structure of the web. The potential $\psi_{\mathcal{K}}$ of a clique

\mathcal{K} can then be defined as:

$$\psi_{\mathcal{K}}(\mathcal{K}, \omega) = \mathcal{V}(\mathbf{x}_{\mathcal{K}}), \quad (17)$$

where $\mathcal{V}(\mathbf{x}_{\mathcal{K}})$ is the potential of a clique node consisting of the random variables $\mathbf{x}_{\mathcal{K}}$. Finally, through the above definitions, the partition function Z becomes:

$$Z = \sum_{\omega \in \Omega} \exp \left(- \sum_{c \in \omega} \psi_c(c) - \sum_{\mathcal{K} \in \mathbb{K}} \psi_{\mathcal{K}}(\mathcal{K}, \omega) \right), \quad (18)$$

where Ω is the set of all possible configurations. The optimal configuration ω^* is then given by $\arg \min_{\omega} U(\omega)$, obtained through the Loopy Belief Propagation procedure detailed above (Section V-D).

VI. EXPERIMENTAL RESULTS

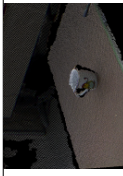





The concept web built from the interactions of iCub is provided in Figure 7. In this section, we demonstrate how this concept web can be helpful for a humanoid robot in three different scenarios: (i) a scenario where the relevant concepts in the web are activated based on perception of an object, (ii) a scenario where the relevant concepts in the web are activated based on a partial perception of an object, with an intended action in mind, and (iii) a scenario where the activation is due to only an intended action in mind, without any specific object singled out.

A. Scenario 1: Perception-driven activation of concepts in the web

In this scenario, iCub is presented with an object, allowed to interact freely with it, and expected to guess what kind of an object it is. It needs to guess both the type of the object (the noun), and its properties (the adjectives). Furthermore, iCub is expected to predict the verbs (the behaviors) that are possibly applicable to the object.

On perceiving the object, iCub first records its visual data through Kinect, then grasps the object to collect haptic, proprioceptive and auditory data (Section III-C). The related features are combined in the entity feature vector \mathbf{e} , and compared against the previously extracted prototypes of nouns and adjectives in order to determine the categories of the object (Section IV-B). These predictions give us an a priori guess about the membership probabilities. These a priori probabilities are in turn used to initialize the activations of

TABLE V: The predicted noun and adjective categories using the concept web, compared with the perception-only guesses. Six objects, one for each noun category, are used for demonstration. Images depict RGB-colored depth images (from the Kinect sensor). The 2nd and 3rd columns depict the predictions of SVM, the 4th and 5th columns show results of SVM applied after ReliefF feature selection, the 6th and 7th columns show prototype-only predictions, and finally the 8th and 9th columns give the concept web estimations. Prediction confidences are indicated in parentheses. Bold text indicates correct decisions whereas stroked text indicates wrong decisions. [Best viewed in color]

Object	SVM		SVM+ReliefF		Prototypes		Concept Web		
	Nouns	Adjectives	Nouns	Adjectives	Nouns	Adjectives	Nouns	Adjectives	
	ball (7%) box (12%) cup (58%) cylinder (14%) plate (4%) tool (5%)	edgy (9%) hard (97%) noisy (12%) short (83%) thick (91%)	ball (3%) box (5%) cup (75%) cylinder (9%) plate (4%) tool (4%)	edgy (4%) hard (98%) noisy (23%) short (94%) thick (91%)	ball (8%) box (13%) cup (43%) cylinder (20%) plate (9%) tool (7%)	edgy (34%) hard (71%) noisy (42%) short (54%) thick (47%)	ball (0%) box (0%) cup (100%) cylinder (0%) plate (0%) tool (0%)	edgy (0%) hard (100%) noisy (0%) short (100%) thick (100%)	round (100%) soft (0%) silent(100%) tall (0%) thin (0%)
	ball (52%) box (10%) cup (12%) cylinder (10%) plate (9%) tool (7%)	edgy (4%) hard (47%) noisy (39%) short (89%) thick (98%)	ball (58%) box (6%) cup (6%) cylinder (10%) plate (12%) tool (8%)	edgy (3%) hard (86%) noisy (19%) short (97%) thick (97%)	ball (26%) box (18%) cup (14%) cylinder (17%) plate (13%) tool (10%)	edgy (42%) hard (53%) noisy (42%) short (64%) thick (57%)	ball (100%) box (0%) cup (0%) cylinder (0%) plate (0%) tool (0%)	edgy (0%) hard (100%) noisy (0%) short (100%) thick (100%)	round (100%) soft (0%) silent(100%) tall (0%) thin (0%)
	ball (6%) box (17%) cup (18%) cylinder (48%) plate (6%) tool (5%)	edgy (7%) hard (97%) noisy (23%) short (36%) thick (33%)	ball (5%) box (9%) cup (13%) cylinder (66%) plate (4%) tool (3%)	edgy (3%) hard (97%) noisy (20%) short (47%) thick (17%)	ball (10%) box (14%) cup (18%) cylinder (41%) plate (9%) tool (8%)	edgy (34%) hard (72%) noisy (30%) short (39%) thick (25%)	ball (0%) box (0%) cup (0%) cylinder (100%) plate (0%) tool (0%)	edgy (0%) hard (100%) noisy (0%) short (100%) thick (0%)	round (100%) soft (0%) silent(100%) tall (0%) thin (100%)
	ball (7%) box (70%) cup (7%) cylinder (10%) plate (3%) tool (3%)	edgy (96%) hard (2%) noisy (16%) short (96%) thick (98%)	ball (4%) box (85%) cup (2%) cylinder (2%) plate (3%) tool (4%)	edgy (97%) hard (0%) noisy (16%) short (97%) thick (100%)	ball (14%) box (42%) cup (11%) cylinder (12%) plate (10%) tool (8%)	edgy (64%) hard (34%) noisy (30%) short (59%) thick (63%)	ball (0%) box (100%) cup (0%) cylinder (0%) plate (0%) tool (0%)	edgy (100%) hard (0%) noisy (0%) short (100%) thick (100%)	round (0%) soft (100%) silent(100%) tall (0%) thin (0%)
	ball (7%) box (18%) cup (13%) cylinder (38%) plate (23%) tool (11%)	edgy (14%) hard (94%) noisy (100%) short (2%) thick (98%)	ball (4%) box (13%) cup (10%) cylinder (17%) plate (46%) tool (10%)	edgy (20%) hard (93%) noisy (100%) short (4%) thick (98%)	ball (11%) box (13%) cup (15%) cylinder (18%) plate (11%) tool (32%)	edgy (48%) hard (65%) noisy (61%) short (39%) thick (57%)	ball (0%) box (0%) cup (0%) cylinder (0%) plate (0%) tool (100%)	edgy (0%) hard (100%) noisy (100%) short (0%) thick (100%)	round (100%) soft (0%) silent(100%) tall (100%) thin (0%)
	ball (12%) box (18%) cup (14%) cylinder (10%) plate (39%) tool (7%)	edgy (13%) hard (62%) noisy (33%) short (52%) thick (99%)	ball (10%) box (15%) cup (8%) cylinder (6%) plate (45%) tool (16%)	edgy (4%) hard (74%) noisy (20%) short (78%) thick (100%)	ball (15%) box (18%) cup (16%) cylinder (17%) plate (21%) tool (13%)	edgy (44%) hard (51%) noisy (44%) short (52%) thick (53%)	ball (0%) box (0%) cup (0%) cylinder (0%) plate (100%) tool (0%)	edgy (0%) hard (100%) noisy (0%) short (0%) thick (100%)	round (100%) soft (0%) silent(100%) tall (100%) thin (0%)

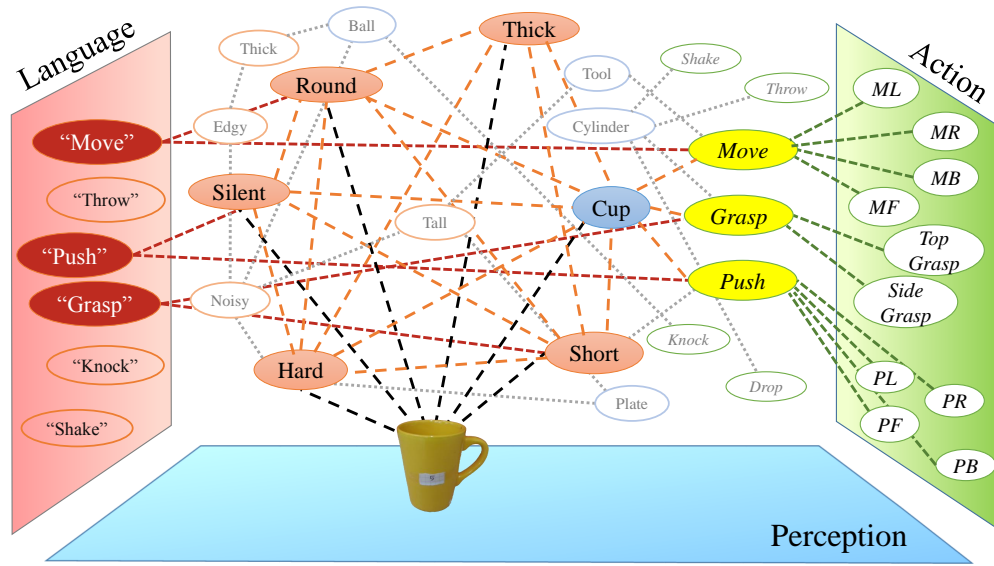


Fig. 12: Schematic presentation of Scenario 1. iCub is presented with a cup, allowed to interact with it freely, and expected to predict the type and properties of the object, as well as what kind of behaviors can be applied on this object. The converged concept web is depicted. The action space and verb concepts are contoured with green, whereas blue and orange colors represent the noun and adjective categories for the object, respectively. The gray and smaller fonts show inactive concepts in the web, while bigger fonts and colored nodes represent activated concepts. There are other concepts and connections that are not shown for clarity. ML: *Move-Left*, MR: *Move-Right*, MB: *Move-Backward*, MF: *Move-Forward*, PL: *Push-Left*, PR: *Push-Right*, PB: *Push-Backward*, PF: *Push-Forward*. [Best viewed in color]

the nodes in the concept web (Note that the concept web architecture, *i.e.*, the connections between the nodes and their joint probabilities, have been determined previously using the training data). Nodes of the unobserved concepts are initialized to 0.5 probability in an unbiased manner. Afterwards the concept web is allowed to propagate its activations. Once convergence is achieved, we expect iCub to (1) refine its initial guesses about the noun and adjective categories of the object, possibly correcting wrong ones, and (2) determine which behaviors are applicable to the object, by propagating activation through the noun and adjective concepts to connected verb concepts.

A sample scenario is presented in Figure 12. A cup is presented to iCub in this case, which iCub correctly detects as a *cup*, and as being *round*, *short*, *hard*, *silent*, and *thick*. It also predicts that it can apply the *grasp* behavior on the object, as well as *move* and *push* behaviors. The rest of the behaviors (*knock down*, *shake*, *throw* and *drop*) are not found applicable to the object.

This scenario depicts the activation of the concept web in a similar fashion to the canonical neurons in the F5 area of monkey brain [86]–[88]. These “visiomotor” neurons are known to fire selectively to certain actions, as well as to the *presentation* of an object to which this action can be potentially applied. This raw recognition of possible action (without necessary recognition of the object *per se*) has been accepted as one of the neurological mechanisms of affordances. The context web approach also results in similar predictive activations in the conceptual representations of the applicable behaviors.

We now apply this scenario systematically to present quantitative results in Table V. Six arbitrarily selected objects, one from each noun category, are presented to the iCub, which is then expected to guess its noun category and adjective categories, and the applicable behaviors. To demonstrate the effectiveness of the approach, the predictions made using the concept web are compared to the prototype-only initial predictions described in Section IV-B, as well as to that of Support Vector Machines, and Support Vector Machines enhanced with ReliefF [89] feature selection. 6 SVMs are trained separately for both the ReliefF and the no-ReliefF cases, 5 of which choose between one of two dichotomic adjectives (*hard* vs. *soft*, *edgy* vs. *round*), and one is responsible with selecting a noun concept (*ball* vs. *box* vs. *cup* vs. *cylinder* vs. *plate* vs. *tool*). Both SVM and SVM+ReliefF cases achieved more than 90% training accuracy, with the exception of the no-ReliefF *noisy* vs. *silent* case with a training accuracy of 82%. In the ReliefF feature selection case, features with weights > 0.1 are accepted, out of a range of $[-1, 1]$.

Table V shows that the concept web predictions are significantly enhanced for both the nouns and the adjectives, as compared to the baseline methods. It is able to correct the wrong predictions of the baselines; whereas for already correctly predicted cases, the prediction confidences are increased. This result is in line with our previous analysis in [76], [79], in which we conclude that an approach which cannot utilize the dependency information between concepts, such as the SVM, SVM+ReliefF, and individual prototypes approaches, would significantly be outperformed by those which can. Therefore,

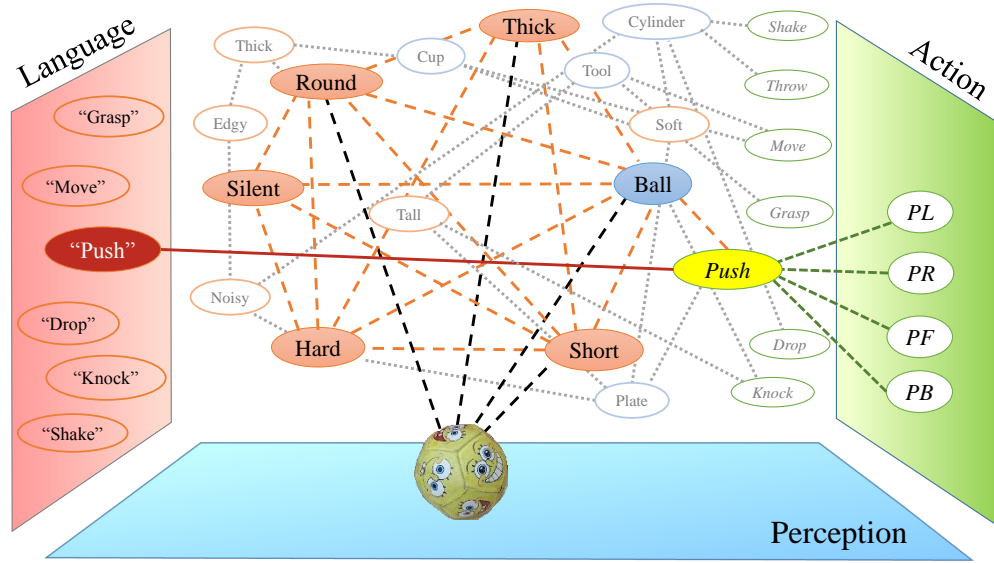



Fig. 13: Schematic presentation of Scenario 2. iCub is presented with a ball, and commanded to apply specifically the *push* behavior on it. The activation enters the system from two different points: From the perception space, through the perceived visual features only (haptic and audio information is not available since iCub is not allowed to grasp the object), and from the *push* verb concept incoming through the language space. When the initially predicted activation has spread and converged, iCub is now able to correctly guess the haptic (*hard*) and auditory (*silent*) properties of the object, which it could not access before. Colored nodes denote activated concepts. Some concepts/connections are not shown for clarity. PL: *Push-Left*, PR: *Push-Right*, PB: *Push-Backward*, PF: *Push-Forward*. [Best viewed in color]

TABLE VI: The predictions as corrected by the activation on the concept web, when there is no direct perceptual access to certain features of the object. The iCub is not allowed to grasp the ball object, and therefore makes initial predictions using only the available visual features (columns 2 and 3). The visual parts of the concept prototypes (*i.e.*, features [1-67]) are used for this comparison. These initial activations are then allowed to spread on the concept web, which converges to the significantly more accurate a posteriori predictions displayed in columns 4 and 5. The haptic and audio predictions are corrected through the spreading of activation. Prediction confidences are indicated in parentheses. Bold text indicates correct decisions whereas stroked text indicates wrong decisions.

Object	Without Concept Web		With Concept Web	
	Nouns	Adjectives	Nouns	Adjectives
	ball (37%) box (14%) cup (12%) cylinder (14%) plate (11%) tool (12%)	edgy (37%) round (63%) hard (45%) soft (55%) noisy (54%) silent(46%) short (59%) tall (41%) thick (54%) thin (46 %)	ball (100%) box (0%) cup (0%) cylinder (0%) plate (0%) tool (0%)	edgy (0%) round (100%) hard (100%) soft (0%) noisy (0%) silent(100%) short (100%) tall (0%) thick (100%) thin (0 %)

the effectiveness of the web-based approach is directly due to its ability to capture second-order conceptual relations, which is ignored by the other methods.

B. Scenario 2: Interaction-driven activation of concepts in the web

In the second scenario (Figure 13), the human partner not only presents iCub with an unknown object, but also commands a single, certain action to be performed. This time, the activation spreads to the concept web from two different entry points.

In the first path, iCub looks at the object, and collects its visual features in a partial entity feature vector (composed of

features [1-67]). Since it is not allowed to grasp the object to investigate it (grasping may not be the required action), haptic, proprioceptive, or auditory features are not available perceptually. This partial entity feature vector (e_v in Section III-C) is compared against the noun and adjective prototypes to predict the corresponding categories for the object (Section IV-B). These predictions are used in turn to activate related concepts in the web. Meanwhile, over a second path, the issued command word (*e.g.*, *grasp*, *push left*, etc.) activates the necessary verb concept through the language space. When the concept web is allowed to propagate activation, knowledge (belief) oscillates between the verb concept and the initially predicted noun and adjective concepts until convergence.

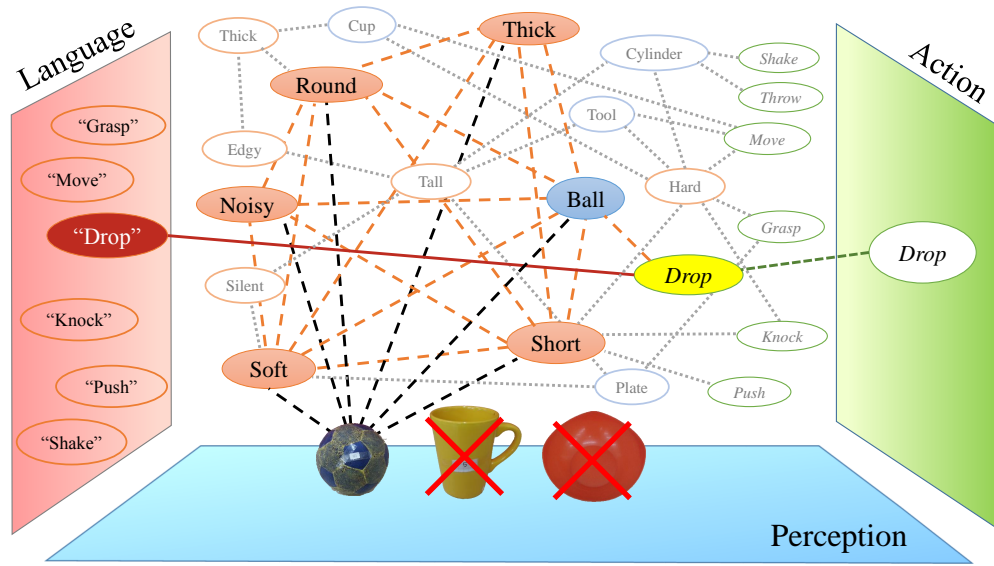


Fig. 14: Schematic representation of Scenario 3. The *ball*, *cup*, and *plate* objects are placed in the environment and the *drop* behavior is commanded to iCub. iCub is expected to select any one of the objects to which the commanded behavior can safely be applied. Activation enters the system from the language space through the commanded verb concept. Since the *drop* behavior is not safe on cups and plates, it spreads towards the *ball* noun and its related adjectives as shown. Colored nodes denote activated concepts. Some concepts/connections are not shown for clarity. [Best viewed in color]

TABLE VII: The selection of objects to which sample commands are applicable. The selection is performed by the spreading activation on the web, which disperses to the related verb concepts as well. *throw* verb concept activates selectively non-cup and non-plate objects. Selection confidences are indicated in parentheses. Images depict RGB-colored depth images (collected via PCL library from the Kinect sensor). [Best viewed in color]

Command	Scene	Existing Objects	Selected Objects
"throw"		cup (25%) box (25%) yellow plate (25%) red plate (25%)	box (100%)
"grasp"		box (16.67%) green cup (16.67%) white cup (16.67%) yellow plate (16.67%) red plate (16.67%) ball (16.67%)	box (16.67%) green cup (16.67%) white cup (16.67%) yellow plate (16.67%) red plate (16.67%) ball (16.67%)

In Figure 13, an example scenario is shown in which iCub is given a ball, and told to apply a "push" behavior on it. Although initially the haptic and auditory information are not available to iCub, these concepts are also active in the converged concept web. The quantitative predictions with and without the concept web are depicted in Table VI.

C. Scenario 3: Action-driven activation of concepts in the web

The final scenario demonstrates how iCub is commanded to perform a certain action in an environment populated with multiple objects (Figure 14). The command does not specify on which object to apply the behavior, therefore iCub must itself choose the appropriate object on which to act. Here we must remember that certain behaviors cannot be applied to all

objects. Therefore, activation must not spread from these verb concepts to inappropriate noun types. After convergence, iCub will pick up a properly activated noun to act upon. If there is more than one appropriate object, a random decision will be made among them.

The entry point of activation in this scenario is from the commanded verb concept. In the sample case in Figure 14, iCub is presented with three objects: a *cup*, a *plate*, and a *ball*. It is then commanded to apply the *drop* behavior. Since the *drop* verb is not connected to the *cup* and *plate* nouns, activation cannot spread to *cup* and *plate*. On the other hand, the *ball* noun is activated through its connection to *drop*. As a result, iCub decides to apply the action to this object. Table VII presents quantitative selection percentages for two sample cases.

This scenario serves as a proof of concept that behaviors can activate related noun concepts, but not unrelated ones. This kind of “reverse” activation spreading can guide the robot’s actions in the world.

VII. CONCLUSION

In this work, we have addressed an important problem in cognitive systems, that of modeling a concept web in a similar fashion to us, humans. The web is constructed based on the co-occurrences of concepts from the interactions of the robot, and modeled using Markov Random Fields. Since the resulting web is a cyclic graph, inferences are made using Loopy Belief Propagation, as is widely done in the literature.

We have demonstrated that, given an observation of an object, our robot can activate in its “brain” the relevant noun concepts, adjective concepts, verb concepts (describing what behaviors can be applied on the object) as well as the words that can be used for the object. Moreover, given an interaction on an object or in fact, an interaction without an object (that would normally take an object), the robot can activate the necessary concepts in the web as well. Being linked to language, perception and motor (action) spaces, the concept web allows activation of relevant information from and to any modality. As we reviewed in detail in Section II-C, such a concept web is very much in line with findings from neuroscience.

Moreover, we showed that such a web allows the robot to make a better interpretation of the environment. By using the co-occurrences from other concepts, wrongly predicted concepts can be corrected, and confidences of correct predictions can be increased.

There are important future directions that can be explored further. For instance, a cognitive model would need to include spatial, temporal, adverb, and social concepts in addition to the noun, adjective, and verb concepts, in order to model the real world more accurately. Moreover, a more realistic model would need to account for super-ordinate, or “higher-order” concepts as well, such as “animal”, or “utensil”. Incorporating a conceptualization mechanism which takes other concepts as its input would allow such reasoning. Finally, the concept web presented in this study is for all intents and purposes a long-term memory model, with no attentional or short-term mechanism. A short-term memory module would enable processing of instances that are inconsistent with what is already known, which can then be either (i) allowed as variations, or (ii) incorporated into the knowledge base as information updates, as necessary.

ACKNOWLEDGMENTS

We thank Erol Şahin for helpful discussions and feedback. For the experiments, we acknowledge the use of the facilities provided by the the Modeling and Simulation Center of METU (MODSIMMER). This work is funded by the Scientific and Technological Research Council of Turkey (TÜBİTAK) through project no 111E287.

REFERENCES

- [1] S. Harnad, “The symbol grounding problem,” *Physica, D*, vol. 42, pp. 335–346, 1990.
- [2] L. Barsalou, “Perceptual symbol systems,” *Behavioral and Brain Sciences*, vol. 22, pp. 577–609, 1999.
- [3] L. Steels, “Evolving grounded communication for robots,” *Trends in Cognitive Science*, vol. 7, no. 7, pp. 308–312, 2003.
- [4] G. Lakoff and M. Johnson, *Philosophy in the flesh: The embodied mind and its challenge to western thought*. New York: Basic Books, 1999.
- [5] 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.
- [6] A. Cangelosi, “Grounding language in action and perception: From cognitive agents to humanoid robots,” *Physics of Life Reviews*, vol. 7, no. 2, pp. 139–151, 2010.
- [7] A. M. Borghi, “Object concepts and embodiment: Why sensorimotor and cognitive processes cannot be separated,” *La Nuova Critica*, vol. 15, no. 4, pp. 447–472, 2007.
- [8] G. Rizzolatti and L. Craighero, “The mirror neuron system,” *Annual Review of Neuroscience*, vol. 27, pp. 169–192, 2004.
- [9] F. Pulvermüller, “Brain mechanisms linking language and action,” *Nature Reviews Neuroscience*, vol. 6, no. 7, pp. 576–582, 2005.
- [10] A. Glenberg, M. Sato, L. Cattaneo, L. Riggio, D. Palumbo, and G. Buccino, “Processing abstract language modulates motor system activity,” *The Quarterly Journal of Experimental Psychology*, vol. 61, no. 6, pp. 905–919, 2008.
- [11] M. H. Fischer and R. A. Zwaan, “Embodied language: a review of the role of the motor system in language comprehension,” *The Quarterly Journal of Experimental Psychology*, vol. 61, no. 6, pp. 825–850, 2008.
- [12] R. A. Zwaan, “The immersed experiencer: Toward an embodied theory of language comprehension,” *Psychology of learning and motivation*, vol. 44, pp. 35–62, 2004.
- [13] F. Pulvermüller, *The neuroscience of language: on brain circuits of words and serial order*. Cambridge University Press, 2002.
- [14] R. F. Goldberg, C. A. Perfetti, and W. Schneider, “Perceptual knowledge retrieval activates sensory brain regions,” *The Journal of Neuroscience*, vol. 26, no. 18, pp. 4917–4921, 2006.
- [15] T. Deacon, “The symbolic species: The co-evolution of language and the human brain,” 1997.
- [16] R. Hudson, “Review of Terrence Deacon, “The symbolic species: The co-evolution of language and the human brain.” London: Penguin, 1997,” *Journal of Pragmatics*, vol. 33, pp. 129–135, 2001.
- [17] L. Barsalou, “Simulation, situated conceptualization, and prediction,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1521, pp. 1281–1289, 2009.
- [18] —, “Situated simulation in the human conceptual system,” *Language and cognitive processes*, vol. 18, no. 5-6, pp. 513–562, 2003.
- [19] W. Yeh and L. Barsalou, “The situated nature of concepts,” *The American Journal of Psychology*, pp. 349–384, 2006.
- [20] J. S. Bruner, J. J. Goodnow, and G. A. Austin, *A study of thinking*. RE Krieger Publishing Company, 1977.
- [21] U. Hahn and N. Chater, “Concepts and similarity,” *Knowledge, concepts and categories*, pp. 43–92, 1997.
- [22] E. H. Rosch, “Natural categories,” *Cognitive psychology*, vol. 4, no. 3, pp. 328–350, 1973.
- [23] P. Gärdenfors, *Conceptual spaces: The geometry of thought*. MIT press, 2004.
- [24] L. Gabora, E. Rosch, and D. Aerts, “Toward an ecological theory of concepts,” *Ecological Psychology*, vol. 20, no. 1, pp. 84–116, 2008.
- [25] J. K. Kruschke and M. K. Johansen, “A model of probabilistic category learning,” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 25, no. 5, p. 1083, 1999.
- [26] Y. Rossee, “Mixture models of categorization,” *Journal of Mathematical Psychology*, vol. 46, no. 2, pp. 178–210, 2002.
- [27] A. Glenberg and K. Kaschak, “Grounding language in action,” *Psychonomic Bulletin & Review*, vol. 9, no. 3, pp. 558–565, 2002.
- [28] D. Jirak, M. M. Menz, G. Buccino, A. M. Borghi, and F. Binkofski, “Grasping language—a short story on embodiment,” *Consciousness and cognition*, vol. 19, no. 3, pp. 711–720, 2010.
- [29] A. Cangelosi, G. Metta, G. Sagerer, S. Nolfi, C. Nehaniv, K. Fischer, J. Tani, T. Belpaeme, G. Sandini, F. Nori *et al.*, “Integration of action and language knowledge: A roadmap for developmental robotics,” *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 3, pp. 167–195, 2010.
- [30] A. M. Borghi, A. M. Glenberg, and M. P. Kaschak, “Putting words in perspective,” *Memory & Cognition*, vol. 32, no. 6, pp. 863–873, 2004.

- [31] A. Stoytchev, "Some basic principles of developmental robotics," *IEEE Transactions on Autonomous Mental Development*, vol. 1, no. 2, pp. 122–130, 2009.
- [32] J. J. Bryson, "Embodiment versus memetics," *Mind & Society*, vol. 7, no. 1, pp. 77–94, 2008.
- [33] L. Steels, "The recruitment theory of language origins," in *Emergence of communication and language*, C. L. Nehaniv, C. Lyon, and A. Cangelosi, Eds. Springer, 2007, pp. 129–150.
- [34] T. Belpaeme and A. Morse, "Word and category learning in a continuous semantic domain: Comparing cross-situational and interactive learning," *Advances in Complex Systems*, vol. 15, no. 3-4, 2012.
- [35] T. Hashimoto and A. Masumi, "Learning and transition of symbols: Towards a dynamical model of a symbolic individual," in *Emergence of communication and language*, C. L. Nehaniv, C. Lyon, and A. Cangelosi, Eds. Springer, 2007, pp. 223–236.
- [36] J. Sinapov, C. Schenck, K. Staley, V. Sukhoy, and A. Stoytchev, "Grounding semantic categories in behavioral interactions: Experiments with 100 objects," *Robotics and Autonomous Systems*, 2012.
- [37] G. Metta, G. Sandini, D. Vernon, L. Natale, and F. Nori, "The iCub humanoid robot: an open platform for research in embodied cognition," in *Proceedings of the 8th workshop on performance metrics for intelligent systems*. ACM, 2008, pp. 50–56.
- [38] S. Kalkan, N. Dağ, O. Yürüten, A. M. Borghi, and E. Şahin, "Verb concepts from affordances," *Interaction Studies*, vol. 15, no. 1, pp. 1–37, 2014.
- [39] M. Rudolph, M. Muhlig, M. Gienger, and H.-J. Bohme, "Learning the consequences of actions: Representing effects as feature changes," in *IEEE International Conference on Emerging Security Technologies (EST)*, 2010, pp. 124–129.
- [40] M. L. Kellenbach, M. Brett, and K. Patterson, "Large, colorful, or noisy? attribute- and modality-specific activations during retrieval of perceptual attribute knowledge," *Cognitive, Affective, & Behavioral Neuroscience*, vol. 1, no. 3, pp. 207–221, 2001.
- [41] A. R. Damasio, "Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition," *Cognition*, vol. 33, no. 1, pp. 25–62, 1989.
- [42] G. H. Eggert, *Wernicke's works on aphasia: A sourcebook and review*. Mouton The Hague, 1977, vol. 1.
- [43] M. A. Lambon Ralph, "Neurocognitive insights on conceptual knowledge and its breakdown," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, no. 1634, 2014.
- [44] L. L. Chao and A. Martin, "Representation of manipulable man-made objects in the dorsal stream," *Neuroimage*, vol. 12, no. 4, pp. 478–484, 2000.
- [45] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just, "Predicting human brain activity associated with the meanings of nouns," *Science*, vol. 320, no. 5880, pp. 1191–1195, 2008.
- [46] K. Patterson, P. J. Nestor, and T. T. Rogers, "Where do you know what you know? The representation of semantic knowledge in the human brain," *Nature Reviews Neuroscience*, vol. 8, no. 12, pp. 976–987, 2007.
- [47] A. Martin, "The representation of object concepts in the brain," *Annual Review of Psychology*, vol. 58, pp. 25–45, 2007.
- [48] M. A. Lambon Ralph, K. Sage, R. W. Jones, and E. J. Mayberry, "Coherent concepts are computed in the anterior temporal lobes," *Proceedings of the National Academy of Sciences*, vol. 107, no. 6, pp. 2717–2722, 2010.
- [49] H. Damasio, D. Tranel, T. Grabowski, R. Adolphs, and A. Damasio, "Neural systems behind word and concept retrieval," *Cognition*, vol. 92, no. 1, pp. 179–229, 2004.
- [50] C. J. Mummary, K. Patterson, C. Price, J. Ashburner, R. Frackowiak, J. R. Hodges *et al.*, "A voxel-based morphometry study of semantic dementia: relationship between temporal lobe atrophy and semantic memory," *Annals of neurology*, vol. 47, no. 1, pp. 36–45, 2000.
- [51] H. Robson, R. Zahn, J. L. Keidel, R. J. Binney, K. Sage, and M. A. L. Ralph, "The anterior temporal lobes support residual comprehension in wernickes aphasia," *Brain*, vol. 137, no. 3, pp. 931–943, 2014.
- [52] W. Simmons and A. Martin, "The anterior temporal lobes and the functional architecture of semantic memory," *Journal of the International Neuropsychological Society*, vol. 15, no. 05, pp. 645–649, 2009.
- [53] M. Tenorth and M. Beetz, "KnowRob: A knowledge processing infrastructure for cognition-enabled robots," *The International Journal of Robotics Research*, vol. 32, no. 5, pp. 566–590, 2013.
- [54] —, "KnowRob: Knowledge processing for autonomous personal robots," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2009*. IEEE, 2009, pp. 4261–4266.
- [55] M. Palmia, "Design and implementation of a system for mutual knowledge among cognition-enabled robots," Master's thesis, 2013.
- [56] M. Tamosiunaite, I. Markelic, T. Kulvicius, and F. Worgotter, "Generalizing objects by analyzing language," in *11th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 2011, pp. 557–563.
- [57] J. C. Trueswell, R. Lila *et al.*, "Learning to parse and its implications for language acquisition," 2009.
- [58] P. Baxter, J. De Greeff, R. Wood, and T. Belpaeme, "Modelling concept prototype competencies using a developmental memory model," *Paladyn*, vol. 3, no. 4, pp. 200–208, 2012.
- [59] A. F. Morse, J. de Greeff, T. Belpaeme, and A. Cangelosi, "Epigenetic robotics architecture (ERA)," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 4, pp. 325–339, 2010.
- [60] E. Uğur, E. Öztöp, and E. Şahin, "Goal emulation and planning in perceptual space using learned affordances," *Robotics and Autonomous Systems*, vol. 59, no. 7, pp. 580–595, 2011.
- [61] E. Uğur and E. Şahin, "Traversability: A case study for learning and perceiving affordances in robots," *Adaptive Behavior*, vol. 18, no. 3-4, pp. 258–284, 2010.
- [62] L. Montesano, M. Lopes, A. Bernardino, and J. Santos-Victor, "Learning object affordances: From sensory-motor coordination to imitation," *IEEE Transactions on Robotics*, vol. 24, no. 1, pp. 15–26, 2008.
- [63] E. A. Feigenbaum and H. A. Simon, "EPAM-like models of recognition and learning," *Cognitive Science*, vol. 8, no. 4, pp. 305–336, 1984.
- [64] J. H. Gennari, P. Langley, and D. Fisher, "Models of incremental concept formation," *Artificial Intelligence*, vol. 40, no. 1-3, pp. 11 – 61, 1989.
- [65] M. Lebowitz, "Experiments with incremental concept formation: UNIMEM," *Machine Learning*, vol. 2, no. 2, pp. 103–138, 1987.
- [66] D. Fisher, "Knowledge acquisition via incremental conceptual clustering," *Machine Learning*, vol. 2, no. 2, pp. 139–172, 1987.
- [67] J. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.
- [68] J. L. Kolodner, "Maintaining organization in a dynamic long-term memory," *Cognitive Science*, vol. 7, no. 4, pp. 243–280, 1983.
- [69] G. Orhan, S. Olgunsoylu, E. Şahin, and S. Kalkan, "Co-learning nouns and adjectives," in *IEEE Third Joint International Conference on Development and Learning and Epigenetic Robotics (ICDL) 2013*. IEEE, 2013, pp. 1–6.
- [70] D. C. Knill and A. Pouget, "The Bayesian brain: the role of uncertainty in neural coding and computation," *Trends in Neurosciences*, vol. 27, no. 12, pp. 712–719, 2004.
- [71] T. Yang and M. N. Shadlen, "Probabilistic reasoning by neurons," *Nature*, vol. 447, no. 7148, pp. 1075–1080, 2007.
- [72] K. A. DeLong, T. P. Urbach, and M. Kutas, "Probabilistic word pre-activation during language comprehension inferred from electrical brain activity," *Nature neuroscience*, vol. 8, no. 8, pp. 1117–1121, 2005.
- [73] A. F. Morse, P. Baxter, T. Belpaeme, L. B. Smith, and A. Cangelosi, "The power of words," in *Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics*, 2011.
- [74] L. Smith and C. Yu, "Infants rapidly learn word-referent mappings via cross-situational statistics," *Cognition*, vol. 106, no. 3, pp. 1558 – 1568, 2008.
- [75] C. Yu and L. B. Smith, "Rapid word learning under uncertainty via cross-situational statistics," *Psychological Science*, vol. 18, no. 5, pp. 414–420, 2007.
- [76] O. Yürüten, E. Şahin, and S. Kalkan, "The learning of adjectives and nouns from affordance and appearance features," *Adaptive Behavior*, vol. 21, no. 6, pp. 437–451, 2013.
- [77] R. B. Rusu and S. Cousins, "3D is here: Point Cloud Library (PCL)," in *IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China, May 9-13 2011.
- [78] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. S. Huang, "Comparing MFCC and MPEG-7 audio features for feature extraction, maximum likelihood HMM and entropic prior HMM for sports audio classification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, 2003, pp. V–628.
- [79] A. C. Bulut, "A multinomial prototype-based learning algorithm," Master's thesis, Middle East Technical University, 2014.
- [80] A. K. Qin and P. N. Suganthan, "Robust growing neural gas algorithm with application in cluster analysis," *Neural Networks*, vol. 17, no. 8-9, pp. 1135–1148, 2004.
- [81] T. Veldhuizen, "Ubigraph: Free dynamic graph visualization software," 2007.
- [82] R. Kindermann and J. L. Snell, *Markov Random Fields and their applications*. American Mathematical Society, 1980, vol. 1.

- [83] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother, "A comparative study of energy minimization methods for Markov Random Fields," in *Computer Vision—ECCV 2006*, pp. 16–29.
- [84] T. Heskes *et al.*, "Stable fixed points of Loopy Belief Propagation are minima of the Bethe free energy," *Advances in neural information processing systems*, vol. 15, pp. 359–366, 2003.
- [85] A. Gouws, "A Python implementation of graphical models," Ph.D. dissertation, Stellenbosch: University of Stellenbosch, 2010.
- [86] A. Murata, L. Fadiga, L. Fogassi, V. Gallese, V. Raos, and G. Rizzolatti, "Object representation in the ventral premotor cortex (area F5) of the monkey," *Journal of neurophysiology*, vol. 78, no. 4, pp. 2226–2230, 1997.
- [87] G. Rizzolatti and L. Fadiga, "Grasping objects and grasping action meanings: the dual role of monkey rostroventral premotor cortex (area F5)," *Sensory guidance of movement*, vol. 218, pp. 81–103, 1998.
- [88] L. Fadiga, L. Fogassi, V. Gallese, and G. Rizzolatti, "Visuomotor neurons: Ambiguity of the discharge or motorperception?" *International journal of psychophysiology*, vol. 35, no. 2, pp. 165–177, 2000.
- [89] I. Kononenko, "Estimating attributes: analysis and extensions of relief," in *Machine Learning: ECML-94*. Springer, 1994, pp. 171–182.



Hande Çelikkanat is currently pursuing her Ph.D. degree in cognitive and developmental robotics in KOVAN Lab., Department of Computer Engineering, Middle East Technical University. She also holds B.Sc. and M.Sc. degrees from the same department, with her M.Sc. thesis titled *Control of a Mobile Robot Swarm via Informed Robots*. Her research interests include the neurological and psychological bases of cognition, especially as related to the development of conceptualization, context, and language, and modeling of these in robots.



Güner Orhan received a B.Sc. degree in the Department of Computer Engineering, Middle East Technical University, 2012. He is currently completing his M.Sc. degree in KOVAN Research Lab., Department of Computer Engineering, METU. The title of the thesis is "Building a Web of Concepts on a Humanoid Robot". His research interests are Developmental Robotics, Cognitive Robotics, Parallel Programming, Computer Vision, and Image Processing.



Sinan Kalkan received his M.Sc. degree in Computer Engineering from Middle East Technical University, Turkey in 2003, and his Ph.D. degree in Informatics from the University of Göttingen, Germany in 2008. After working as a postdoctoral researcher at the University of Göttingen and at Middle East Technical University, he is an assistant professor at Middle East Technical University since 2010. Sinan Kalkan's research interests include biologically motivated Computer Vision and Image Processing and Developmental Robotics.