Integrating Spatial Concepts into a Probabilistic Concept Web

Hande Çelikkanat[1], Erol Şahin[1,2], and Sinan Kalkan[1]
[1]KOVAN Research Lab., Department of Computer Engineering, Middle East Technical University, Turkey
[2]Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA.
Email: {hande,erol,skalkan}@ceng.metu.edu.tr

***Abstract*—In this paper, we study the learning and representation of grounded spatial concepts in a probabilistic concept web that connects them with other noun, adjective, and verb concepts. Specifically, we focus on the prepositional spatial concepts, such as "on", "below", "left", "right", "in front of" and "behind". In our prior work (Celikkanat et al., 2015), inspired from the distributed highly-connected conceptual representation in human brain, we proposed using Markov Random Field for modeling a concept web on a humanoid robot. For adequately expressing the unidirectional (*i.e.*, non-symmetric) nature of the spatial propositions, in this work, we propose a extension of the Markov Random Field into a simple hybrid Markov Random Field model, allowing both undirected and directed connections between concepts. We demonstrate that our humanoid robot, iCub, is able to (i) extract meaningful spatial concepts in addition to noun, adjective and verb concepts from a scene using the proposed model, (ii) correct wrong initial predictions using the connectedness of the concept web, and (iii) respond correctly to queries involving spatial concepts, such as ball-left-of-the-cup.**

***Keywords*—Concepts, Concept Web, Spatial Concepts, Prepositions, Markov Random Field**

I. Introduction

Conceptualization is one of the cornerstones of cognition [1]–[3]. Conceptualizing the complex world allows us to categorize it into meaningful, manageable components, to reason on it, act rationally in it; in short, to impose a structure on complex sensory data. When we conceptualize, we understand: We understand what constitutes a concept, which instances are elements of this concept. Importantly, we also internalize how this specific concept relates to other concepts [4].

A perhaps more advanced question is, how do we conceptualize spatial relations between objects? An understanding of spatial relations of objects in the world is crucial to everyday actions, not only we communicate with each other using them ("Give me the cup on the table."), but we also plan subconsciously using these relations all the time (*e.g.*, Pouring the milk *into* the pot in order to be able to warm it *on* the oven.) There is evidence that the parietal cortex constantly tracks these abundant relations (see, for instance, [5], [6]). Arguably, the only way we could have survived as animals is by carrying an accurate spatial model of the world in our minds: We can close our eyes at any moment and recount the relative positions of the objects around us to an astounding accuracy. But in addition to this instantaneous and automatic spatial modeling, we also have a virtually perfect intuitive *understanding* of the spatial concepts with respect to the laws of physics: We know we cannot (easily) stand on a basketball because it is round, we understand we can place a book beneath the monitor in order to raise it, but not an orange, and so on. Therefore there is more to spatial relations than basic world-modeling: A spatial concept is just like any other concept in that it is most meaningful only when considered in relation to the other concepts in our mind.

978-1-4673-7509-2/15/$31.00 2015 IEEE

In [7], we proposed a densely connected web of concepts as a biologically plausible and robust representation. Such a web allows the considering of concepts in relation to each other, and can guide and correct reasoning and planning in an otherwise too-complex real world [4], [8], [9]. We showed in several scenarios the advantages of such a connected system. However, this initial version was composed only of the noun, adjective, and verb concepts, and it lacked:

Spatial Relations: There was no notion of spatial relations, which resulted in a deficiency of representing scenes holistically, such as two cups are standing on the table next to each other. Such lack of spatial knowledge is fatal, for instance, in planning, *e.g.*, we may not be able to move a ball that is inside a plate by pushing the ball, we may need to first grasp it and take it out of the plate, and then replace it accordingly.

Short-Term Memory: Additionally, there was no equivalent of a short-term memory mechanism in this initial version. The whole model corresponded to the long-term memory: The concept web was adept in representing long-term knowledge and understanding of the world, but it could not model *instantiations* of this knowledge. This was problematic for instance when there were more than one object, since the concept web connected all the active concepts associatively, and a ball (which is round) could not be active together with a box (which is edgy) at the same time. Therefore, we had to resort to an ad-hoc modeling of instantaneous perceptions: The system would *focus* on each object one by one, and extract a separate concept web for it. An superposition of these concept webs would then be accepted as a model of the current scene.

In this work, we aim to overcome these limitations with two improvements: (1) The capability to represent spatial relations, and (2) a short-term memory for individual objects. Combining these two features, the concept web can effectively represent whole scenes. We argue that spatial relations should also be regarded as concepts in a web of concepts, in relation to other concepts, just like a noun or adjective concepts. However, since they are binary *and* directed in nature (since, e.g., a ball can stand on a box, but not vice versa, since the ball is round, therefore the spatial relations have *order*), we present a Hybrid Markov Random Field model, which is a variant of the Markov

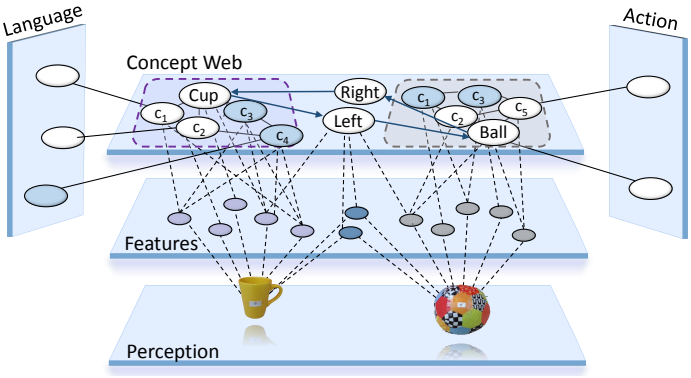


Fig. 1: The concept web combining short-term memories of perceived objects, and the spatial relations. Shaded areas correspond to short-term memory representations of the individual objects. These are fed by the features of the objects, as well as by language and action. The spatial relations in between combine them, and are fed by the relative and the individual features of the objects, as well as by language and action.

Random Field-based concept web model in [7], enhanced to include *directed* connections between spatially-related nodes.

A seminal work on the psychology of spatial conceptualization was conducted by Landau and Jackendorf [10], who argued that people rely hugely on approximations when expressing spatial concepts. For instance, languages provide only the crude descriptions of "in" and "not in" for the important concept of containment, but there is no detailed propositions describing, for instance, "being in a round object", or "being inside and also in contact with the inner surface of an object", etc. Instead, languages tend to elaborate on nouns with details, and abstract over spatial concepts.

In robotics, Fischer [11] investigated the variables affecting people's choice of spatial instructions when interacting with a robot. Stopp *et al.* [12] studied how a robot can anchor verbal spatial descriptions to its physical environment, thus grounding them, proposing a compositional variant of spatial potential fields. Gold *et al.* [13] showed how spatial prepositions, together with pronouns, can be extracted and represented as word trees, depending on entropy and information gain metrics applied on the physical environment. Moratz and Tenbrink [14] developed a system for iterative interpreting of projective relations in human-robot interaction scenarios, in order to enable mutual identification of objects. Hanheide [15] pointed out the qualitative nature of spatial representations in humans (describing something crudely as "on the right" rather than providing exact angle), and proposed using Qualitative Trajectory Calculus in order to formalize the comparative movements of two agents. Tellex *et al.* [16] worked on a robotic forklift scenario, to be controlled by natural language commands, in which they try to learn the parameters for a probabilistic graphical model from a corpus of commands. Meanwhile Golland *et al.* [17] showed that, when trying to minimize the risk of miscommunication between two collaborative agents, *discovering* the spatial relations to describe the environment might be more beneficial than sticking to predetermined descriptions. Inspired by this, Guadarrama *et al.* [18] proposed learning spatial prepositions and object representations simultaneously, combining strategies of template matching, syntactic parsing, and probabilistic analysis.

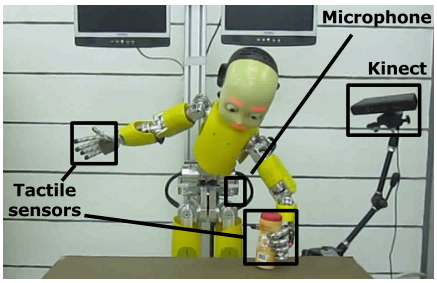


Fig. 2: The experimental setup. iCub senses the environment via its tactile sensors, a microphone and a Kinect.

The closest study to our work is of Anand *et al.* [19], who used spatial relations between noun concepts to guide a visual search via contextual information, using a Markov Random Field. In their work, each object part corresponds to a node in MRF, and spatial neighborhood is used to connect the nodes to each other. Hard-coded, rule-based spatial relations such as "on top of", "in front of" are then integrated into the model as edge potentials to improve the accuracy. Our approach is different in the following aspects: (i) In our model, for each object, a concept web is created and a set of concepts become active. (ii) The spatial prepositional relations are themselves concepts that link concepts activated by different objects.

II. The Experimental Setup

We use the iCub humanoid robot platform [20] in this study. The visual data is collected using a Kinect RGB-D camera. The haptic cues associated with the objects are captured through the tactile sensors on the fingertips.

Object Set: We use 60 objects in the experiments, divided arbitrarily into a training set of 45 objects and a test set of 15 objects. Each object belongs to one of the 6 noun classes: {*box, ball, cylinder, cup, plate, tool*} (Fig. 3). Their properties are divided into 10 adjectives: {*hard* $\times$ *soft, noisy* $\times$ *silent, tall* $\times$ *short, thin* $\times$ *thick, round* $\times$ *edgy*} (Fig. 4). Each training object is labeled with 1 noun and 5 adjectives with human supervision. iCub is given a behavior repertoire of {*grasp, push left, push right, push forward, push backward, move left, move right, move forward, move backward, drop, throw, knock down, shake*}, which are hard-wired primitives.

Data Collection: iCub collects data from each object by interacting with it [21]. 4 kinds of features are collected from each object (Table I-top): (1) Visual features ($\mathbf{e}_v$) are collected from the Kinect sensor. (2) Haptic ($\mathbf{e}_h$), and proprioceptive ($\mathbf{e}_p$) features of the robot hand while interacting with the object are collected by applying a *grasp* behavior it. (3) Audio features of the object ($\mathbf{e}_a$) are collected by applying a *shake* behavior on it. For each object of both the training and test sets, these features are collected and concatenated into the entity feature vector $\mathbf{e}$ of the object. For describing the behaviors, each behavior is applied once on every object, and the difference between the visual features of the object before and after the behavior is recorded. This difference vector is called the effect feature vector $\mathbf{f}$, and describes the changes induced by the behavior.

For developing the spatial concepts {*on, below, left, right, in front of, behind*}, binary features are collected from couples

Fig. 3: The objects used in the experiments, divided to each noun category.

Fig. 4: The objects for each adjective category.

TABLE I: The extracted unary and binary features.

Feature Type	Unary Feature	Position
Visual ($\mathbf{e}_v$)	Position:(x, y, z)	1-3
	Object dimensions:$(width, height, depth)$	4-6
	Normal zenith histogram bins	7-26
	Normal azimuth histogram bins	27-46
	Shape index histogram bins	47-66
Audio ($\mathbf{e}_a$)	13 bins of MFCC (max - min)	67-79
Haptic ($\mathbf{e}_h$)	Change for index finger	80
	Min values for index finger	81
	Max values for index finger	82
	Mean for index finger	83
	Variance for index finger	84
	Standard deviation for index finger	85
Proprioceptive ($\mathbf{e}_p$)	Change for index finger	86
	Min values for index finger	87
	Max values for index finger	88
	Mean for index finger	89
	Variance for index finger	90
	Standard deviation for index finger	91

Feature Type	Binary Feature	Position
Projective ($\mathbf{e}_{proj}$)	Relative x position	1
	Relative y position	2
	Relative z position	3

of objects in the scene during training and testing. Following [10] and [17], we employ binary *projective* features between two objects, which define the relative x, y, z positions of the two objects with respect to each other (Table I-bottom). The projective features are adequate for the projective spatial concepts we deal with in this study. Note that for developing other spatial concepts, such as $\{$*in, out, near, far*$\}$, specialized topological and proximity features, such as the containment and relative distance features would be beneficial [17].

III. Methods

There are two steps of the proposed conceptualization scheme. The first is the detection of active concepts individually from an encountered scene. Noun and adjective concepts are detected separately for each present object, verb concepts are detected from each action applied on an object, and spatial concepts are detected from the binary relations of existing objects with each other. The set of all objects is denoted by $\mathbb{C} = \mathbb{N} \cup \mathbb{A} \cup \mathbb{V} \cup \mathbb{S}$, where $\mathbb{N} = \{$*ball, box, cup, cylinder, plate, tool*$\}$ is the set of noun concepts, $\mathbb{A} = \{$*hard* $\times$ *soft, tall* $\times$ *short, thin* $\times$ *thick, round* $\times$ *edgy, noisy* $\times$ *silent,*$\}$ the set of adjective 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*$\}$ the set of verb concepts, and finally $\mathbb{S} = \{$*on, below, left, right, in front of, behind*$\}$ the set of spatial concepts.

A. Representing Individual Concepts

We represent individual concepts with their prototypes, following our work in [21]–[23]. In [7], [24], this prototyping scheme is shown to be comparable in terms of performance to widely used approaches, such as Support Vector Machines, Self Organizing Maps, etc. One of its advantages is generalization over concepts similar to humans' generalization abilities, especially eminent in spatial concepts [10], [15], *e.g.*, approximating "to the right" as a probability distribution that that covers a feasible region of "right" in the space, without necessarily forcing a 90°-relative direction. Yet this particular representation can as well be replaced with an alternative.

Prototypes of individual concepts are extracted from training instances previously labeled with corresponding concepts through human supervision. For every concept, the set of feature vectors belonging to the instances labeled with the concept are gathered. Each feature dimension is normalized and rendered comparable with other feature dimensions, followed by the extraction of their mean and variance values. Concept prototypes, as strings of {+, -, *, 0} characters, is obtained from these mean and variance values (Table II). Features with a high mean and low variance are denoted with a (+), indicating strongly positive contribution. Features with a low mean value and low variance are denoted with (-), indicating strongly negative contribution. Features with too high variance are indicated with (*), which denotes erratic contribution from the feature dimension. Such features are labeled as *irrelevant*, and excluded from all calculations regarding this concept. Finally, for verb and spatial concepts, features with means around 0 with a low variance are indicated with (0), and denote (i) for verb concepts, unchanging feature value during behavior, and (ii) for spatial concepts, qualitatively similar value between the two related objects. In this study, the experimentally defined thresholds of $\mu_{high} = 0.1, \mu_{low} = -0.1, \sigma^2_{high} = 0.055$ are used for deciding high vs. low mean and variance values.

Noun and Adjective Concepts: The noun and adjective concepts are extracted from training objects that have previously been labeled with them. The entity feature vectors $\mathbf{e}$ are used for extracting these prototypes. Therefore, they indicate all of

TABLE II: Sample extracted prototypes for noun, adjective, verb and spatial concepts.

	Concepts	Visual Features	Audio Features	Haptic Features	Proprioceptive Features
Nouns	Box	+−+++−++−−−++−−−−−−−−−−−−−−++−−+−−−−−−−−∗∗−−−−−−−−−−−−−−−−−−+−−−−	−∗∗∗∗∗−++−∗+∗	−−−−−−	−+∗∗−−
	Ball	−−+−+−+−−−−++−−−−−−−−−−−−−−−+++++−−−−−−−−−−−−−−−−−−−−−−−−+−−−−	∗∗∗∗∗∗∗−∗∗+−+	∗−∗∗∗∗	−+++−−
Adjectives	Hard	+−+∗+++++++++−−−−−−−−−−−−+∗+++∗+−−−−−−−∗−−−−−−−−−−−−−−−−−−+++++	∗∗∗∗∗∗+∗∗+++∗	∗+∗∗∗+∗	−+∗∗−−
	Noisy	−−++++∗−−−∗+++∗−−−−−−−−−−−−+++++∗−∗∗∗∗−−−−−−−−−−−−−−−−−−+−−−−	++++++++++++	∗−∗∗−∗	−+∗∗−−
Verbs	Knock Down	0+0++++00++00+−−−−−−−−−−−−0−0−+0−0−−+0000000−∗−−−−−−−−−−−−−−−++++	None		
	Throw	∗0∗000∗0−000+00−−−−−−−−−−−−0−−−−−000−00−−0−−−−0−−−−−−−−−−−−−−−0++++	None		

Spatial Concepts	Projective Features	Visual Features (First Object)	Visual Features (Second Object)
Left	−0−0	∗∗∗0∗∗∗∗∗∗∗0000000000000∗∗∗0000∗∗∗∗∗∗∗∗∗∗∗∗∗∗0000000000000000∗∗∗∗∗	∗∗∗0∗∗∗∗∗∗∗0000000000000∗0∗00000∗∗∗0∗∗∗∗∗∗∗0000000000000000∗∗∗∗∗
Behind	−00	∗∗∗0∗∗∗+∗∗∗∗0000000000000∗∗∗0000∗∗∗∗∗∗∗∗∗∗∗∗∗∗0000000000000000∗∗∗∗∗	+∗∗0∗∗∗∗∗∗∗0000000000000∗∗00000∗∗∗∗0∗∗∗∗∗∗∗0000000000000000∗∗∗∗∗

(a) (b)

Fig. 5: (a) A standard Markov Random Field, and the expansion into a factor graph. (Adapted from [7].) (b) Expansion of directed cliques in a hybrid Markov Random Field, and which are expanded into two separate clique nodes.

the visual, audio, haptic, and proprioceptive of properties of the related noun or adjective. These prototypes have 91 features.

Verb Concepts: The verb concepts are used for describing the changes induced on a objects through the application of the related action, and are extracted from the effect feature vectors (**f**). Similar to the effect feature vectors, they are composed of only the visual features, and therefore contain 66 characters.

Spatial Concepts: Spatial concepts are binary, and are calculated from couples of objects. A spatial prototype includes: (1) Binary projective features $\mathbf{e}_{proj}$, extracted from the relative positioning of the two objects, which are concatenated with (2) Visual features of the two objects $\mathbf{e}_v^1$ and $\mathbf{e}_v^2$. Therefore, spatial concept prototypes are of length 135. The reason for the inclusion of visual features are due to the implicit relation between the shape of the objects and the physically possible spatial positions: It is difficult to balance something on top a round object, therefore the *on* and *below* relations have semantic ties to the object shapes.

Comparing Individual Concepts: An incoming instance is decided as belonging to certain concepts by comparing its feature vector to the concept prototypes. The irrelevant features of each concept, denoted by (*), are excluded from these comparisons. The closeness of the instance to each concept is decided by a Euclidean distance between the instance and the concept prototype means over the remaining features [7].

B. Concept Web

We represent the concept web as a Markov Random Field, whose nodes correspond to the set of concepts $\mathbb{C} = \mathbb{N} \cup \mathbb{A} \cup \mathbb{V} \cup \mathbb{S}$, and connections denote the co-occurrence information between the concepts. Formally, the concept web W is a graph $W = G(\mathbb{C}, \mathbb{E})$, where each concept $c \in \mathbb{C}$ is treated as a node, and there is an edge $\mathbb{E}(c_i, c_j)$ between concepts c_i and c_j, if c_i and c_j have co-occurred together in a training instance.

Markov Random Field: A Markov Random Field (MRF) [25] is a probabilistic graphical model describing constraints between random variables, which are encoded as connections between nodes. It effectively models the joint probability distribution:

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

with $\omega \in \Omega$ denoting a configuration of the MRF, $U(\omega)$ is the energy of the given configuration, and Z is a normalizing partition function. The energy function in an MRF is always the combination of the data and the smoothness terms:

$$U(\omega) = U_{data}(\omega) + U_{smooth}(\omega), \tag{2}$$

The consistency of the state of the MRF with the incoming measurements is assured through the data term, $U_{data}(\omega)$. The a priori knowledge about the constraints between-nodes, which is coded into the MRF through the edge structure, is asserted through the smoothness term $U_{smooth}(\omega)$.

This energy function must be minimized to find the most likely configuration of the MRF. In highly cyclic MRFs such as ours, an effective method for this is the Loopy Belief Propagation (LBP) algorithm [26], which works over the factor graph notation of the MRF (Fig. 5a), obtained by creating separator nodes from each MRF node, and clique nodes combining connected MRF nodes. LBP algorithm spreads information from the separator nodes to the clique nodes, and then back from the clique nodes to the separator nodes, over multiple iterations.

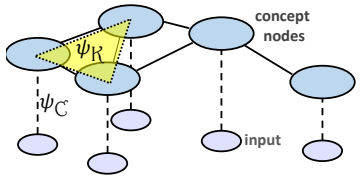


Fig. 6: Visualization of the energy formalization of the MRF. The data term is obtained from the individual concept potentials ψ_c, and the smoothness term is given by the sum of clique potentials ψ_K. Figure taken from [7].

TABLE III: A sample scenario of scene interpretation. Some of the extracted relations for the presented 3D view are indicated.

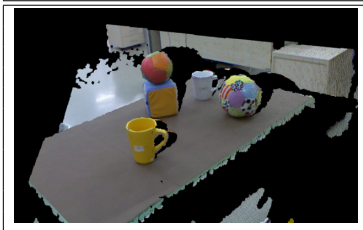
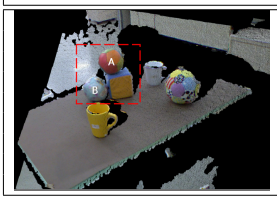
Scene	Extracted Relations
	A ball **on** a box A box **below** a ball A ball **on the right of** a cup A cup **on the left of** a ball A ball **in front of** a cup A cup **behind** a ball A box **on the left** of a cup A cup **on the right** of a box A cup **on the right** of a cup A cup **on the left** of a cup ...

TABLE IV: A sample scenario of the concept web correcting the wrong interpretation of the spatial configuration of ball A.

Scene	Spatial Relations	Without C. Web	With C. Web
	Ball A **on** B	~~18%~~	0%
	Ball A **below** B	15%	0%
	Ball A **left of** B	15%	0%
	Ball A **right of** B	18%	**100%**
	Ball A **in front of** B	~~18%~~	0%
	Ball A **behind** B	16%	0%

Hybrid Markov Random Field: The standard Markov Random field is an undirected graph, which is suitable for our representation of noun, adjective, and verb concepts. However, the spatial concepts require a different scheme. That is because these spatial concepts are *directed* in nature: When object 1 is on the left of object 2, object 2 is *not* on the left of object 1, but on the right of it. Therefore, we propose a variant of Markov Random Field representation, which we call the *Hybrid Markov Random Field*, to model such relations.

Fig. 5b depicts a hybrid Markov Random Field. The difference is in encoding a directed connection via two separate clique nodes in the factor graph. The first clique node denotes information flow in one direction (from concept x_1 to x_2 in the figure), and the second clique node denotes information flowing in the opposite direction (from concept x_1 to x_2 here). The potentials of the two clique nodes are calculated separately, resulting in two "Left" concepts here, each one representing Left of one of the related two objects.

The Concept Web as a Markov Random Field: In [7], we use the following energy formulation for the concept web:

$$U(\omega) = \sum_{c \in \omega} \psi_c(c) + \sum_{\mathscr{K} \in \mathbb{K}} \psi_K(\mathscr{K}, \omega), \tag{3}$$

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 concept c, and ψ_K is the potential of clique K. Fig. 6 visualizes the concept and clique potentials. The data term $\sum_{c \in \omega} \psi_c(c)$ tries to manipulate the solution towards the raw perceptions of concepts. In [7], we define a concept potential ψ_c by $\psi_c(c) = D(c, \mathbf{x})$, with $\mathbf{x}$ being the incoming instance, and $D(c, \mathbf{x})$ its Euclidean distance to concept c [7]. The smoothness term, $\sum_{\mathscr{K} \in \mathbb{K}} \psi_K(\mathscr{K}, \omega)$, tries to manipulate the solution towards the a priori knowledge, encoded in terms of edge information in cliques formed of co-occurrences. A clique potential $\psi_{\mathscr{K}}$ is given by $\psi_K(\mathscr{K}, \omega) = \mathcal{V}(\mathbf{x}_{\mathscr{K}})$, with $\mathcal{V}(\mathbf{x}_{\mathscr{K}})$ denoting the potential of the clique including the nodes $\mathbf{x}_{\mathscr{K}}$. The optimal configuration $arg\ min_{\omega} U(\omega)$ is given by LBP algorithm.

Scene Representation: The representation of a scene in the system is handled in two levels (Fig. 1). On the one side, the attention of the system focuses on each object, and extracts a concept web of the related concepts with the object, acting like short-term memory module. Object-specific short-term memory instantiations are modeled using standard (undirected) MRF representation, since as shown in [7], undirected connections are not only intuitive but also effective in capturing co-dependence between noun, adjective, and verb concepts semantically related together. This is similar to reasoning in humans: Humans are able to understand and reason on single objects situated in an environment, that is, the *identity* of the object is easily extracted and generates its own related concept activations. Yet, simultaneously, the whole scene is considered together, and spatial relations between couples of objects are added by additional MRF links between the short-term memory representations of each object. The hybrid MRF representation is used for modeling the spatial relations due to their directed nature. Moreover, the spatial relations are modeled between the *noun* components of the object representations, since it is natural for humans to address unnamed objects by their nouns, instead of their adjectives, since nouns are more discriminative in communication (*e.g.*, "Pass me the cup next to the kettle", instead of "Pass me the small noisy object next to the tall object.").

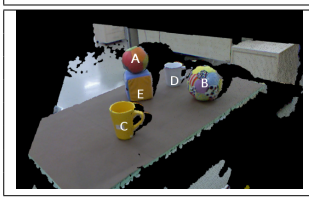
IV. Results

We demonstrate the extended concept web model and the effectiveness of representing spatial relations in a concept web in three different scenarios: (1) Semantic interpretation of an encountered scene as activations in the concept web, (2) correction of wrong predictions through the co-occurrence information coded in the concept web, and (3) using spatial relations to guide object search for human-robot interaction. The training set is composed of 600 arbitrary binary formations of the training objects, which are designed with a priori information: For instance, since it is very difficult to balance an object over a ball, this combination does not exist in the training set, and therefore not represented in the concept web.

1. Scene Interpretation: The interpretation of a sample scene through the proposed system is depicted in Table III. In the 3D Kinect view, there are two cups, two balls, and one box. iCub attends to all objects one by one and extracts their concept webs. Then it examines their spatial relations in the hybrid MRF, resulting in the presented concepts. Systematically, we have also run the system on 5 different world views with 3 to 6 objects in the scene at one moment, resulting in 37 binary relations. In this setup, the system achieves a noun detection rate of 95.2% and a spatial relation detection rate of 91.8%.

2. Correcting Wrong Interpretations: The main strength of the concept web is keeping a priori information about the world, either due to physical laws, or canonical usages of everyday life. Such expectations guide our reasoning, even when our sensors may fail. In the second scenario, we show how the concept web may fulfill a similar function for iCub. In the scene in Table IV, the situation of ball A with respect to ball B is slightly ambiguous: The initial predictions using only the prototypes deduce ball A can be on ball B, on the

TABLE V: A sample scenario of human-robot interaction based on spatial-directions. Objects found by iCub are indicated.

Scene	Queries	Found Objects
	Object(s) **on the right** of the box?	Cup D
	Object(s) **behind** the box?	No such object
	Box is **on the right** of what?	Cup C
	Box is **in front of** what?	None
	Box is **below** of what?	Ball A
	Cup that is **on** another object?	No such object
	Cup that is **on the right** of another object?	Cup D (to the right of ball A, box E, and cup C)
	Ball that is **behind** the box?	No such object

right of it, and also in front of it. In fact, ball A is on another box that is on the right of ball B. Indeed it is not possible to stack balls on top of each other, since they tend to roll easily. Therefore, there is no ball-on-ball example in the training set, and such a clique has not formed in the hybrid MRF, biasing the concept web towards dismissing the wrong prediction.

3. Human-Robot Interaction: In the final scenario, we communicate with iCub using spatial descriptions. iCub instantaneously evaluates the scene, extracting short-term concept webs of the objects and the spatial relations. Then, the human partner points certain object(s) using (i) Two nouns and a relation, *e.g.*, "The cup that is behind the box", (ii) one noun and one relation, *e.g.*, "Object that is on the left of the box", or "Ball is on the left of which object?" Through language, commanded concepts stay active in the hybrid MRF, while the separator node activations of the not-mentioned concepts are reset. Since the hybrid MRF is directional, the separator node activations are reset according to whether the fixed noun(s) in the command are in the first or second noun position. The whole system reiterates until convergence, at the end of which only the concepts that are relevant to *both* the visual scene and the command remain active. A sample case is presented in Table V. Tests of 100 queries performed over 5 real-world scenes demonstrated a performance of 96% for this scenario.

V. Conclusion

In this article, we provided the first steps towards integrating spatial concepts into a probabilistic concept web model proposed in [7] based on Markov Random Fields (MRF). Since classical MRF is based on undirected connections between nodes (concepts), we proposed a hybrid version which can have both undirected and directed relations between concepts. In several scenarios, we demonstrated that such representation is useful for various reasoning tasks.

Acknowledgments

This work is funded by the Scientific and Technological Research Council of Turkey (TÜBİTAK) through project no 111E287. Erol Şahin acknowledges the support of the Marie Curie International Outgoing Fellowship titled "Towards Better Robot Manipulation: Improvement through Interaction" (FP7-PEOPLE-2013-IOF- 628854).

References

[1] G. Lakoff, *Women, fire, and dangerous things: What categories reveal about the mind*. Cambridge Univ Press, 1990.

[2] T. Deacon, "The symbolic species: the co-evolution of language and the human brain," 1997.

[3] L. Barsalou, "Perceptual symbol systems," *Behavioral and Brain Sciences*, vol. 22, pp. 577–609, 1999.

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

[5] L. G. Ungerleider, "Two cortical visual systems," *Analysis of visual behavior*, pp. 549–586, 1982.

[6] H. Damasio, T. J. Grabowski, D. Tranel, L. L. Ponto, R. D. Hichwa, and A. R. Damasio, "Neural correlates of naming actions and of naming spatial relations," *Neuroimage*, vol. 13, no. 6, pp. 1053–1064, 2001.

[7] H. Celikkanat, G. Orhan, and S. Kalkan, "A probabilistic concept web on a humanoid robot," 2015, IEEE Transactions on Autonomous Mental Development (accepted), DOI: 10.1109/TAMD.2015.2418678.

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

[9] M. Bar, "Visual objects in context," *Nature Reviews Neuroscience*, vol. 5, no. 8, pp. 617–629, 2004.

[10] B. Landau and R. Jackendoff, "'What' and 'where' in spatial language and spatial cognition," *Behavioral and brain sciences*, vol. 16, 1993.

[11] K. Fischer, "The role of users concepts of the robot in human-robot spatial instruction," in *Spatial Cognition V Reasoning, Action, Interaction*. Springer, 2007, pp. 76–89.

[12] E. Stopp, K.-P. Gapp, G. Herzog, T. Laengle, and T. C. Lueth, "Utilizing spatial relations for natural language access to an autonomous mobile robot," vol. 861. Springer Science & Business Media, 1994, p. 39.

[13] K. Gold, M. Doniec, and B. Scassellati, "Learning grounded semantics with word trees: Prepositions and pronouns," in *ICDL*, 2007, pp. 25–30.

[14] R. Moratz and T. Tenbrink, "Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations," *Spatial cognition and computation*, vol. 6, 2006.

[15] M. Hanheide, A. Peters, and N. Bellotto, "Analysis of human-robot spatial behaviour applying a qualitative trajectory calculus," in *RO-MAN*, 2012, pp. 689–694.

[16] S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. J. Teller, and N. Roy, "Understanding natural language commands for robotic navigation and mobile manipulation." in *AAAI*, 2011.

[17] D. Golland, P. Liang, and D. Klein, "A game-theoretic approach to generating spatial descriptions," in *EMNLP*, 2010, pp. 410–419.

[18] S. Guadarrama, L. Riano, D. Golland, D. Gouhring, Y. Jia, D. Klein, P. Abbeel, and T. Darrell, "Grounding spatial relations for human-robot interaction," in *IROS*, 2013, pp. 1640–1647.

[19] A. Anand, H. S. Koppula, T. Joachims, and A. Saxena, "Contextually guided semantic labeling and search for three-dimensional point clouds," *The Int. J. of Robotics Research*, pp. 19–34, 2012.

[20] G. Metta, G. Sandini, D. Vernon, L. Natale, and F. Nori, "The iCub humanoid robot: an open platform for research in embodied cognition," in *Workshop on performance metrics for intelligent systems*, 2008.

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

[22] S. Kalkan, N. Dağ, O. Yürüten, A. M. Borghi, and E. Şahin, "Verb concepts from affordances," *Interaction Studies*, vol. 15, pp. 1–37, 2014.

[23] 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.

[24] A. C. Bulut, "A multinomial prototype-based learning algorithm," Master's thesis, Middle East Technical University, 2014.

[25] R. Kindermann, J. L. Snell, *et al.*, *Markov random fields and their applications*. American Mathematical Society Providence, RI, 1980.

[26] T. Heskes, "Stable fixed points of loopy belief propagation are minima of the bethe free energy," in *NIPS*, 2003, pp. 359–366.