A Generative Framework for Multimodal Learning of Spatial Concepts and Object Categories: An Unsupervised Part-of-Speech Tagging and 3D Visual Perception Based Approach

Amir Aly, Akira Taniguchi, Tadahiro Taniguchi

A Generative Framework for Multimodal Learning of Spatial Concepts and Object Categories: An Unsupervised Part-of-Speech Tagging and 3D Visual Perception Based Approach

Amir Aly[1] and Akira Taniguchi[2] and Tadahiro Taniguchi[3]

***Abstract*— Future human-robot collaboration employs language in instructing a robot about specific tasks to perform in its surroundings. This requires the robot to be able to associate spatial knowledge with language to understand the details of an assigned task so as to behave appropriately in the context of interaction. In this paper, we propose a probabilistic framework for learning the meaning of language spatial concepts (spatial prepositions) and object categories based on visual cues representing spatial layouts and geometric characteristics of objects in a tabletop scene. The model investigates unsupervised Part-of-Speech (POS) tagging through a Hidden Markov Model (HMM) that infers the corresponding hidden tags to words. Spatial configurations and geometric characteristics of objects on the tabletop are described through 3D point cloud information that encodes spatial semantics and categories of referents and landmarks in the environment. The proposed model is evaluated through human user interaction with Toyota HSR robot, where the obtained results show the significant effect of the model in making the robot able to successfully engage in interaction with the user in space.**

I. INTRODUCTION

The growing omnipresent role of robots in the social life of human users requires high level cognitive functions that could allow them to efficiently work with humans in different tasks. Developing the robot spatial intelligence to discover its physical environment involves using multimodal sensory information to semantically interpret and reason about spatial relationships between world referents and landmarks [19, 28]. This embodied spatial cognition bridges between spatial knowledge and language so as to make the robot able to understand and express spatial concepts through language during interaction [20, 22].

Grounding language in perception has been long considered as a major challenge both in cognitive science and artificial intelligence. Harnad [15] defined the "Symbol Grounding" problem, which refers to assigning a meaning to each meaningless symbol (e.g., a new word) in a structure through sensorimotor interaction with the environment. An early initiative to investigate the important effect of visual cues on understanding spoken language was discussed in Tanenhaus et al. [37]. Similarly, Siskind [33] discussed a primary logic-based model for grounding language (verbs of motion) in perception. Roy [29] successfully developed a model to visually ground words describing a task, where the system learns the visual semantics of phrases through a "show-and-tell" training procedure. A further step towards making robots able to easily engage in conversations was discussed in Roy et al. [30], where they proposed an architecture to provide perceptual and affordance representations of grounded words. Another interesting study was discussed in Matuszek et al. [21], where they proposed a probabilistic joint learning model that uses a categorial grammar for creating compositional meaning representations of language and visually perceived objects in space.

Similarly, grounding spatial concepts has been extensively studied in the related literature during the last years. An early study about understanding spatial concepts and events of objects in a movie was discussed in Regier [27], who developed a computational connectionist model to learn spatial prepositions for static and dynamic objects, and to ground semantics of language spatial terms. The connectionist learning model takes as input the trajectories between objects for several movie frames and outputs labels representing the corresponding spatial prepositions. Similarly, Cangelosi et al. [4] integrated a computational connectionist model that encodes the dynamics of a visual scene in a neural representation using an Elman network. Afterwards, the model uses a dual-route vision-language network to estimate spatial terms that best describe the perceived scene. These two previous approaches; however, do not investigate spatial concepts and relationships through a language-based analysis. On the other hand, Tellex et al. [40] presented a probabilistic graphical model for grounding spatial relationships of natural language commands in an open environment. This model is trained on images indicating spatial relationships, route directions, and mobile manipulation, paired with command texts representing the required robot actions in space. Dawson et al. [10] employed a probabilistic framework for understanding utterances representing spatial relationships between static referents and landmarks in a virtual space. The proposed graphical model learns correspondences between sentences and spatial relationships by observing an interacting human describing a scene (and pointing to a location) repeatedly. This model is trained over a limited range of utterances so that the parser can not handle spatial relationships representing more than one landmark (e.g., BETWEEN object A and

[1]Amir Aly is a senior researcher at the Emergent Systems Laboratory, Ritsumeikan University, Japan amir.aly@em.ci.ritsumei.ac.jp

[2]Akira Taniguchi is a PhD candidate at the Emergent Systems Laboratory, Ritsumeikan University, Japan a.taniguchi@em.ci.ritsumei.ac.jp

[3]Tadahiro Taniguchi is a full professor at the Emergent Systems Laboratory, Ritsumeikan University, Japan taniguchi@em.ci.ritsumei.ac.jp

object B). This restricted relational vocabulary of the model constitutes a limitation with complex spatial configurations. Guadarrama et al. [14] presented an interesting system for learning the meaning of spatial relationships through visual perception of objects in space. This robot system integrates several modules of different functionalities, such as: object segmentation, action-template matching that compares an utterance to manually constructed templates representing specific actions of the robot, and syntactic supervised parsing to segment the different parts of speech, like spatial prepositions and nouns representing objects from input texts.

Although these studies, among many others, discuss interesting approaches for language grounding whether through 2D or 3D visual perception, they all lack the ability to tag parts of speech in an unsupervised manner so as to understand syntactic structure of speech - and hence to infer the meaning of a sentence - in a developmentally plausible approach. *This could open the door to study unsupervised grammar induction* so as to make a robot understand syntactic dependencies between words composing a verbal instruction, which is a future scope of this study [16]. Additionally, this approach could allow for addressing another important problem in cognitive science and artificial intelligence, which is the "Symbol Emergence" problem [26, 39]. Investigating this bottom-up development of symbols has attracted much attention of researchers during the last decade [25], especially after Steels [35] argued that the "Symbol Grounding" problem had been, conceptually, solved.

To the best of our knowledge, no similar study in the related literature investigated the integration of unsupervised Part-of-Speech (POS) tagging with grounding spatial concepts and object categories in language through 3D visual perception, which constituted our inspiration for this work. The rest of the paper is structured as follows: Section (II) presents a detailed description of the proposed framework, Section (III) illustrates the experimental design, Section (IV) provides a description of the experimental results, and finally, Section (V) concludes the paper.

II. SYSTEM ARCHITECTURE

The proposed system in this paper is coordinated through four different sub-systems: (1) Speech recognition system (Google HTML5 API speech recognition toolkit), which recognizes the instructions of the human tutor during interaction with the robot, (2) 3D Object segmentation system, which segments objects and planes into point clouds so as to determine their spatial relationships and geometric characteristics, (3) Part-of-Speech (POS) tagging system, which grammatically tags the recognized sentences in an unsupervised manner (i.e., *it assigns numerical tags to words without using any pre-tagged corpus or tagging dictionary, and the role of the system would be to ground the meaning of these numerical tags through visual perception*), and finally (4) Multimodal graphical model for grounding spatial concepts and object categories of sentences through visual perception. The inputs and outputs of these sub-systems are summarized as follows:

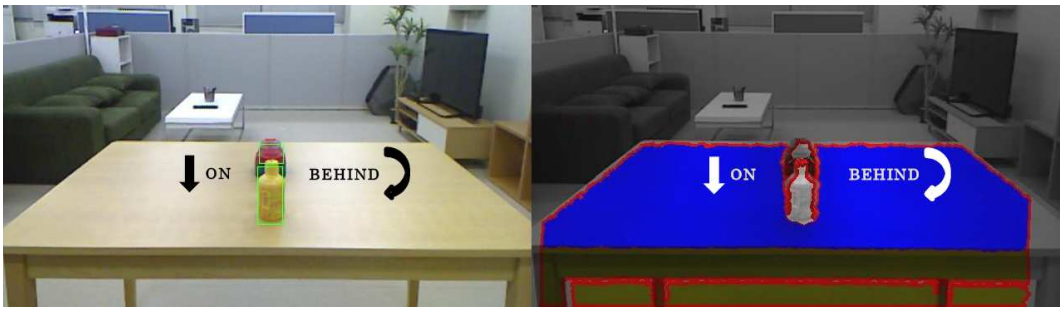
1) **3D object segmentation**:
   - **Output**: Centroid coordinates and geometric features of objects.
2) **Part-of-Speech (POS) tagging**:
   - **Input**: Sentence(s) (i.e., instructions of the human tutor to the robot).
   - **Output**: Numerical tags - representing syntactic categories of words - to be grounded by the graphical learning model.
3) **Multimodal graphical learning model**:
   - **Input**: The outputs of points 1 & 2.
   - **Output**: In case of a test sentence and a corresponding spatial layout of objects, the model determines the spatial relationship between the referent and the landmark (Section IV). The following subsections shed light on each sub-system in detail.

A. Unsupervised Object Segmentation in 3D Point Cloud

Object discovery and segmentation in complex environment have been extensively studied in the computer vision community [24]. Different approaches for 3D point cloud segmentation have been investigated in the related literature. Edge-based methods determine the boundaries of regions in point cloud through detecting the points with a big change in intensity [17]. Region-based methods focus on clustering nearby points of similar properties using neighboring information [18]. Graph-based methods search on point cloud segmentation through graph representations of points [36]. Model-based methods efficiently create point clusters based on geometric criteria so that points with similar mathematical representations would create together one segment (e.g., points representing a plane) [32]. In this study, a model-based method is employed for segmenting objects lying in a plane into separate point clouds in an unsupervised manner [7, 8]. Unlike the other highlighted methods, this model is fast, reliable, and does not require much prior knowledge about the environment, such as object models and the number of regions to process, as with the region-based methods.

The integrated visual perception methodology to the proposed model detects the major plane in the environment (i.e., floor or tabletop) using the RANSAC algorithm [11], and tracks it through consecutive frames. Having calculated the representative equation of the major horizontal plane, orthogonal planes to the detected major plane touching at least one image border are considered as wall planes. After filtering out the corresponding points to floor (or tabletop) and wall planes in the processed point cloud, the remaining points are voxelized and clustered into blobs representing object candidates. Finally, the algorithm excludes very small and large blobs, in addition to blobs with very close centroids to a detected wall or at a border of the depth image. Figure (1) shows the segmentation results of objects lying in a tabletop in different spatial configurations.

In this research study, both objects and the major tabletop plane they are lying in are segmented into point clouds, where the centroids represent their (x, y, z) coordinates with



(a) Spatial concepts: ON and BEHIND

(b) Spatial concepts: BESIDE, ON, and BETWEEN

Fig. 1: Different spatial concepts of objects in a tabletop scene expressed through 3D point cloud information

reference to the robot camera. Each point cloud is characterized through the Viewpoint Feature Histogram (VFH) descriptor [31] that encodes both the geometry and viewpoint of an object while being invariant to pose and scale. Having known the locations of objects on the tabletop, the robot employs a graphical learning model (Section II-C) in order to ground the meaning of spatial concepts and object categories through cross-situational learning [34], as indicated in Section (III).

B. Unsupervised Part-of-Speech Tagging

Part-of-Speech (POS) tagging is the process of marking words in sentences with grammatical attributes (e.g., noun, verb, preposition, adjective, etc.). The related literature reveals different supervised, semi-supervised, and unsupervised approaches towards syntactically tagging words. Supervised tagging methods employ pre-tagged training corpora in order to create tagging dictionaries indicating possible tags of words and word-tag frequencies, which are used to tag test words through appropriate models, such as the rule-based tagging model of Brill [2] and the stochastic tagging models of Church [6] and Cutting et al. [9]. However, these models would not be, probably, able to tag new words as this process requires using complete dictionaries of language, which is a difficult condition to fulfill. Semi-supervised tagging methods try to overcome this obstacle, partially, as they do not require large and high quality pre-tagged corpora, where the statistical models can estimate appropriate tags for new word sequences [41]. On the other hand, unsupervised tagging methods do not need any training corpus, where they induce grammatical tags for word sequences through rule-based models [3] or statistical models [5]. Consequently, integrating this unsupervised approach for tagging word sequences (i.e., *without using any pre-tagged training corpus*[1]) to our model would be an efficient step towards creating a developmentally plausible system for grounding spatial concepts and object categories of speech during human-robot interaction in space.

A Part-of-Speech (POS) tagging model assigns a grammatical tag $\mathrm{T} = (t_1, \ldots, t_n)$ to each word in the sequence $\mathrm{w} = (\omega_1, \ldots, \omega_n)$. The first-order Bayesian Hidden Markov Model (HMM) employs words as observations and tags as hidden states (Figure 2) [12]. The probability distribution of tag states for the word sequence w is defined as follows:

$$\mathbb{P}(t_1, \ldots, t_n) = \prod_{i=1}^{n} \mathbb{P}(t_i \,|\, t_{i-1}) \tag{1}$$

where the transition probability to the tag t_i is conditioned on the tag t_{i-1}. This could encode the intuitive grammar that parts of speech might follow, like having a noun after a determiner. Emission distributions of numerical tags over words are defined through the probability $\mathbb{P}(\omega_i \,|\, t_i)$ of the word ω_i being conditioned on the tag t_i. The generative transition and emission parameters of the proposed HMM model (ϕ, θ) for each tag state are characterized through multinomial distributions with Dirichlet priors $(\alpha_\phi, \alpha_\theta)$ (where K denotes the number of tag states):

$$\begin{aligned} t_i \,\big|\, t_{i-1} = t \sim Mult\,(\phi_t) \quad , \quad \phi_t \,\big|\, \alpha_\phi \sim Dir\,(\alpha_\phi) \\ \omega_i \,\big|\, t_i = t \sim Mult\,(\theta_t) \quad , \quad \theta_t \,\big|\, \alpha_\theta \sim Dir\,(\alpha_\theta) \end{aligned} \tag{2}$$

[1] For example, the POS tagging system could assign these numerical tags to words of the sentence: (Push,**2**) (the,**7**) (Ball,**4**) (Beside,**9**) (the,**7**) (Cup,**4**) so that the framework grounds the meaning of these tags through visual perception.

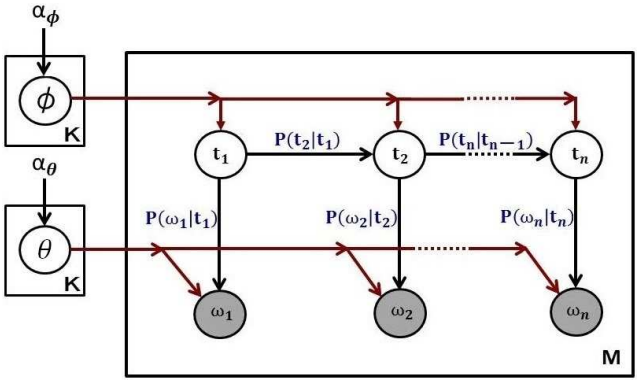


Fig. 2: Graphical representation of the Hidden Markov Part-of-Speech tagging model

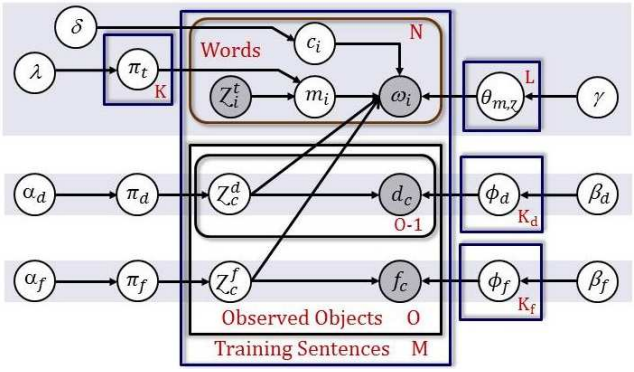


Fig. 3: Graphical representation of the learning model. The index "i" denotes the order of words and their corresponding syntactic tags, the index "d" denotes a spatial dyadic layout between a referent and a landmark, and the index "f" denotes object geometric features.

For an unannotated training corpus containing a set of m sentences $\mathrm{W} = \{w_1, \ldots, w_m\}$, the POS tagging model tries to induce the most likely numerical tag set $\mathrm{T} = \{T_1, \ldots, T_m\}$ for each sentence in the corpus that maximizes the following expression:

$$\mathbb{P}(T, W) = \prod_{(T,w) \in (T,W)} \Big(\mathbb{P}(T, w \,|\, \phi, \theta) \Big) = \prod_{(T,w) \in (T,W)} \Big(\prod_{i=1}^{n} \mathbb{P}(t_i \,|\, t_{i-1}, \phi_t) \mathbb{P}(\omega_i \,|\, t_i, \theta_t) \Big) \tag{3}$$

Inferring the latent tag variables employs the Gibbs sampling algorithm [13, 23], which produces a set of samples from the posterior distribution $\mathbb{P}(T \,|\, W)$ (i.e., it loops over the possible tag assignments to words that could maximize Equation 3) expressed as follows (where "-i" denotes all samples except the i-th sample):

$$\mathbb{P}(T_i, T^{(i)} \,\big|\, T_{-i}, W, T^{(-i)}, w, \alpha_\phi, \alpha_\theta) \tag{4}$$

C. Multimodal Graphical Learning Model

Grounding spatial concepts and object categories through visual perception employs the probabilistic learning model illustrated in Figure (3). This model, basically, investigates dyadic spatial relationships between referents and landmarks (i.e., between two objects or between an object and the tabletop). In this study, we focus on four prepositions that encode dyadic spatial relationships: {BESIDE, IN, ON, and BEHIND} (referent, landmark), in addition to one preposition that encodes a complex spatial relationship: {BETWEEN} (landmark, referent, landmark) (i.e., a triadic layout, which might be considered as a composite of dyadic spatial relationships). We investigate the ability of the model to learn complex spatial relationships that employ more than one landmark after being divided into sub-groups of corresponding dyadic relationships in order to reduce the complexity of the probabilistic learning model.

The employed Gaussian mixture learning model in this study - inspired by the model of Taniguchi et al. [38] - is illustrated in Figure (3). The observed state ω_i represents each token in the word sequence $\mathrm{w} = (\omega_1, \ldots, \omega_n)$, and the observed state z_i^t represents the corresponding tags to words (Section II-B). The observed state d_c represents a dyadic spatial configuration of a referent and a landmark expressed through their centroid coordinates (i.e., the state $d_c = f(P_{A_{x,y,z}}, P_{B_{x,y,z}})$, where A and B denote the referent and the landmark) (Section II-A). The observed state f_c represents the geometric characteristics of observed objects expressed through the point-cloud-based VFH descriptor (each training sentence describes a spatial layout involving O objects) (Section II-A). For a complex spatial relationship with more than one landmark (e.g., a triadic layout), the possible dyadic relationships between objects could be expressed as follows: *Observed Objects* $O \times (O-1)$ (Figure 3). The definitions of the learning model parameters are summarized in Table (I)[2]. The probabilistic distributions that characterize the different channels of the model are defined as follows:

$$\begin{cases} \theta_{m,z_{L_1}} & \sim & Dir\,(\gamma) & , & \mathrm{L}_1 = (1, \ldots, \mathrm{L}) \\ \phi_{d_{K_1}} & \sim & GIW\,(\beta_d), & & \mathrm{K}_1 = (1, \ldots, \mathrm{K_d}) \\ \phi_{f_{K_2}} & \sim & GIW\,(\beta_f), & & \mathrm{K}_2 = (1, \ldots, \mathrm{K_f}) \\ \pi_{t_{K_3}} & \sim & Dir\,(\lambda) & , & \mathrm{K}_3 = (1, \ldots, \mathrm{K}) \\ \pi_d & \sim & Dir\,(\alpha_d) & & \\ \pi_f & \sim & Dir\,(\alpha_f) & & \\ c_i & \sim & Cat\,(\delta) & & \\ m_i & \sim & Cat\,(\pi_{z_i^t}) & & \\ \omega_i & \sim & Cat\,(\theta_{m,z}) & & \\ z_c^d & \sim & Cat\,(\pi_d) & & \\ z_c^f & \sim & Cat\,(\pi_f) & & \\ d_c & \sim & Gauss\,(\phi_{z_c^d}) & & \\ f_c & \sim & Gauss\,(\phi_{z_c^f}) & & \end{cases} \tag{5}$$

The inference of the latent variables in the learning model employs the Gibbs sampling algorithm [13] (Table II), which repeatedly creates samples from the posterior distributions of the model parameters indicated as follows:

[2] The mathematical definitions of the probabilistic distributions of the proposed learning model in this study are illustrated, in detail, in Bishop [1].

$$
\begin{cases}
\phi_d & \sim & \mathbb{P}\left(\phi_d \,\middle|\, d_c\,,\beta_d\right)\\
\phi_f & \sim & \mathbb{P}\left(\phi_f \,\middle|\, f_c\,,\beta_f\right)\\
\pi_t & \sim & \mathbb{P}\left(\pi_t \,\middle|\, m\,,\mathcal{Z}^t\,,\lambda\right)\\
\pi_d & \sim & \mathbb{P}\left(\pi_d \,\middle|\, \mathcal{Z}_c^d\,,\alpha_d\right)\\
\pi_f & \sim & \mathbb{P}\left(\pi_f \,\middle|\, \mathcal{Z}_c^f\,,\alpha_f\right)\\
\mathcal{Z}_c^d & \sim & \mathbb{P}\left(\mathcal{Z}_c^d \,\middle|\, d_c\,,\pi_d\,,w\right)\\
\mathcal{Z}_c^f & \sim & \mathbb{P}\left(\mathcal{Z}_c^f \,\middle|\, f_c\,,\pi_f\,,w\right)\\
\theta_{m,z} & \sim & \mathbb{P}\left(\theta_{m,z} \,\middle|\, W\,,m\,,c\,,\mathcal{Z}_c^d\,,\mathcal{Z}_c^f\,,\gamma\right)\\
c_i & \sim & \mathbb{P}\left(c_i \,\middle|\, \omega_i\,,\theta_{m,z}\,,m_i\,,\mathcal{Z}_c^d\,,\mathcal{Z}_c^f\,,\delta\right)\\
m_i & \sim & \mathbb{P}\left(m_i \,\middle|\, \omega_i\,,\mathcal{Z}_i^t\,,\theta_{m,z}\,,c_i\,,\mathcal{Z}_c^d\,,\mathcal{Z}_c^f\,,\pi_t\right)
\end{cases}
\tag{6}
$$

Having calculated the latent variables, the model learns the correspondences between syntactic categories of words and visual cues so as to successfully collaborate with the human tutor in different spatial tasks.

TABLE I: Definitions of the learning model parameters in the different channels: word sequence, spatial dyadic layout, and object geometric features

Parameter	Definition
λ	Hyperparameter of the Dirichlet distribution π_t
c_i	Index of spatial relationship (Object A $\leftrightarrows$ Object B) of each word
m_i	Modality index $\in$ {Preposition, Object, Others} of each word
$\theta_{m,z}$	Word distribution over modalities
L	Number of word distribution categories over modalities = K_d + K_f + 1
γ	Hyperparameter of the Dirichlet distribution $\theta_{m,z}$
α_d	Hyperparameter of the Dirichlet distribution π_d
β_d	Hyperparameter of the Gaussian Inverse Wishart distribution ϕ_d
K_d	Number of categories in the spatial layout modality
α_f	Hyperparameter of the Dirichlet distribution π_f
β_f	Hyperparameter of the Gaussian Inverse Wishart distribution ϕ_f
K_f	Number of categories in the object features modality
$\mathcal{Z}_c^d$	Index of spatial layout distributions
$\mathcal{Z}_c^f$	Index of object features distributions

TABLE II: Inference of the latent variables of the graphical learning model

Algorithm 1 Inference of the model latent variables

procedure Gibbs_Sampling($\{d_c\}, \{f_c\}, W, \{Z^t\}$)
  Initialization of $\theta, \phi_d, \phi_f, \pi_t, \pi_d, \pi_f, \{Z_c^d\}, \{Z_c^f\}, \{c_i\}, \{m_i\}$
  for $j = 1$ to *iteration_number* **do**
    Equation (6)
  end for
  return $\theta, \phi_d, \phi_f, \pi_t, \pi_d, \pi_f, \{Z_c^d\}, \{Z_c^f\}, \{c_i\}, \{m_i\}$
end procedure

III. EXPERIMENTAL SETUP

In this section, we introduce the experimental design and the scenario of interaction between a human tutor and Toyota HSR robot (Figure 4)[3] in front of a table (the major landmark), where objects in different spatial configurations lie in. We used 5 objects of different categories (i.e., TOY, BALL, BOX[4], BOTTLE, and CUP) as referents and landmarks, and 5 prepositions (i.e., BESIDE, BEHIND, IN, ON, and BETWEEN) to indicate the referent-landmark spatial relationships (Figure 1). Over and above, the robot performs 5 actions on the objects within a human-robot interaction context: {TOUCH, HOLD, PUSH, RAISE, and THROW} (robot, object)[5].

Fig. 4: The robot in front of a table at home environment and 5 objects: TOY, BALL, BOX, BOTTLE, and CUP

During the cross-situational learning phase [34], the human tutor uses a total of 60 sentences in order to teach the robot different spatial configurations of objects, such as: "RAISE the BOTTLE ON the TABLE" and "HOLD the BALL BETWEEN the TOY and the CUP". The experimental scenario is described as follows:

- The human tutor trains the robot on randomly-ordered spatial layouts of different objects lying in a tabletop through visual cues and descriptive sentences.
- For each observed scene in the training phase, the robot uses the visual perception system in order to segment objects into point clouds so as to determine their geometric characteristics, and to define the relative spatial relationships between the centroid of each object's point cloud to those of the other objects and the tabletop (Section II-A).
- For each descriptive sentence, the robot uses the POS tagging system to define numerical tags representing syntactic categories of words (e.g., "(Push,**2**) (the,**7**) (Ball,**4**) (Beside,**9**) (the,**7**) (Cup,**4**)") (Section II-B).
- The robot employs the learning model to ground the numerical tags of spatial concepts and object categories (Section II-C).
- During the test phase, the human tutor uses a total of 30 sentences in order to evaluate the robustness of the learning phase. At this phase, the robot performs the 5 programed actions on objects - mentioned earlier[5] -

[3]The Human Support Robot (HSR) is developed by Toyota for assisting people in their daily life activities. It has a full-motion lightweight body with a total of 11 degrees of freedom. The robot is equipped with stereo, Asus Xtion, and wide-angle cameras, a display screen, in addition to an array of sensors including a force-torque sensor, a laser range sensor, and an IMU sensor. The robot has one arm with a gripper that allows it to grasp objects at different heights efficiently [Toyota HSR Robot Website].

[4]The object "BOX" is used only as a landmark in this study.

[5]In the context of this study, these action verbs were directly modeled and programmed on the robot for generating object-directed behaviors. The robot used the calculated distances to objects through point cloud information (Section II-A) to control its joins and the end-effector so as to perform predefined behaviors representing the indicated action verbs. In the future, we would consider making the robot able to learn - in the training phase - body behavior of the human tutor while doing different actions so as to generate adaptive behaviors to the context of interaction on its own.

TABLE III: Modality estimation results of the different parts of speech during the test phase

Modality Index (/Sentences)			
Correct (Parts of Speech)	Wrong (Parts of Speech)		
	Action Verbs	And	Others (Box, the)
79.7%	14.5%	4.3%	1.5%

TABLE IV: Modality estimation results of spatial preposition and object-referring words in the different sentences during the test phase

Preposition	Modality Index (/Sentences)		Object	Modality Index (/Sentences)	
	Correct	Wrong		Correct	Wrong
BESIDE	6	0	TOY	15	0
BEHIND	6	0	BALL	8	0
IN	6	0	BOX	13	2
ON	6	0	BOTTLE	15	0
BETWEEN	6	0	CUP	10	0

based on its learning experience of spatial concepts and object categories.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The system was trained offline on a total of 60 sentences describing different spatial configurations of referents and landmarks in space, and was tested on other 30 sentences. The framework was evaluated on its ability to determine the modality of each word in a test sentence to be: *Preposition* (dyadic layout), *Object* (point-cloud-based object geometric features), or *Others*, and to determine the referent and the landmark(s) of the sentence so as to define the direction of the spatial relationship (i.e., Object A $\leftrightarrows$ Object B) that the robot needs to understand in order to be able to perform a spatial task.

Estimating the modality of each word in a sentence during the test phase showed a reasonable performance of the probabilistic learning model. Tables (III and IV) reveal that the modalities of the majority of parts of speech of the test corpus were correctly estimated. Action verbs were totally considered by the framework to be objects, which is a logical result considering that the graphical model did not contain a learning modality for the human body behavior. This interesting rational result shows that the learning model assigned action verbs the same modality index of objects upon which they were performed, despite the fact that they could have been considered - even partially - to be dyadic-layout-referring words (i.e., prepositions) according to the word probability distribution illustrated in Figure (5), however the model estimated the most relevant modality.

A similar tendency was shown with the coordination word "AND", which was totally considered to be a dyadic-layout-referring word (i.e., preposition) (Table III and Figure 5) instead of belonging to the modality "Others". This last interesting finding represents an important insight towards understanding complex spatial relationships that employ more than one landmark as in the case of the preposition BETWEEN (landmark, referent, landmark). In this case, the preposition "BETWEEN" was correctly considered to be a dyadic-layout-referring word (similarly to all the other prepositions, as indicated in Table IV and Figure 5), in addition to the coordination word "AND" that encodes a relationship between the two landmarks. Consequently, this model was able to successfully perceive the number of dyadic spatial relationships composing a complex relationship with more than one landmark.

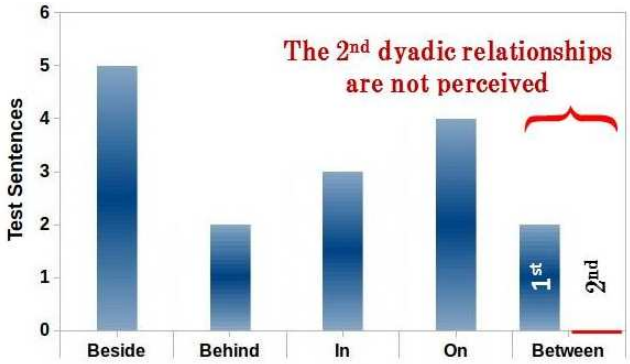


Fig. 6: Correct referent-landmark spatial relationships for the different prepositions

On the other hand, defining the referent and the landmark(s) in sentences was successful in dyadic spatial relationships, while being the contrary in complex relationships as indicated in Figure (6). The figure shows that in case of the preposition "BETWEEN", the framework was not able to correctly perceive the first and second referent-landmark relationships, unlike the cases of the other prepositions (Figure 7). This means that this framework can perceive the number of dyadic relationships within complex layouts, but still needs development in order to be able to perceive the correct referent-landmark arrangements within complex spatial configurations. This point constitutes a follow up of this current research study so as to make the robot able to efficiently collaborate with human users in space.

(a) The robot determines the referent and the landmark in the sentence

(b) The robot successfully raises the bottle

Fig. 7: The robot successfully performs the required action on the BOTTLE located beside the CUP based on point cloud information

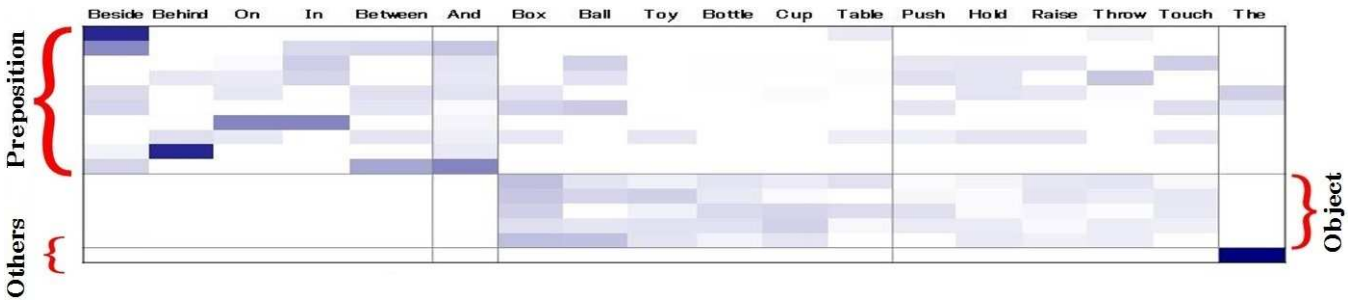


Fig. 5: Probability distribution of words over the different modalities (the dark blue color represents high probability distribution)

V. CONCLUSION AND FUTURE WORK

This paper discusses a multimodal framework for grounding spatial concepts and object categories of language through visual perception during interaction between a human tutor and Toyota HSR robot. The illustrated system learns, in an unsupervised manner, the meaning of the numerical syntactic tags of parts of speech based on point cloud information that encodes dyadic spatial relationships between referents and landmarks, in addition to their geometric characteristics.

The proposed model succeeded in determining referent and landmark-referring words in most of the sentences that describe dyadic spatial layouts so that the robot was able to perform the required actions on objects based on their point cloud information (Figure 7). On the other hand, the system was not successful in understanding triadic spatial relationships between one referent and two landmarks through analyzing their corresponding in-between dyadic relationships. However, the framework raised an interesting insight on grounding complex spatial relationships, where the probability distribution of the coordination word "AND" was considered to be a dyadic-layout-referring word. This means that the system is able to understand that a complex spatial relationship expressed through the preposition BETWEEN (landmark, referent, landmark) encodes two spatial relationships. However, the proposed model needs further development in order to appropriately define referent-landmark arrangements in complex spatial layouts.

To meet this target, and for our future work, we would extend the proposed learning model in this study in order to be able to successfully learn referent-landmark relationships in complex spatial configurations. Besides, we are considering adding other modalities to the learning model in order to ground the meaning of color-referring words through visual perception, and to learn the dynamics of human actions (i.e., associated body movements to action verbs) based on visual cues so as to manipulate objects in space autonomously and adaptively to the context of interaction[5].

REFERENCES

[1] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. 4

[2] E. Brill. A simple rule-based part of speech tagger. In *Proceedings of the 3rd Conference on Applied Natural Language Processing (ANLC)*, Trento, Italy, 1992. 3

[3] E. Brill and M. Pop. Unsupervised learning of disambiguation rules for Part-of-Speech tagging. In S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann, and D. Yarowsky, editors, *Natural Language Processing Using Very Large Corpora*, volume 11 of *Text, Speech, and Language Technology*, pages 27–42. Springer, 1999. 3

[4] A. Cangelosi, K.R. Coventry, R. Rajapakse, D. Joyce, A. Bacon, L. Richards, and S.N. Newstead. Grounding language in perception: A connectionist model of spatial terms and vague quantifiers. In A. Cangelosi, G. Bugmann, and R. Borisyuk, editors, *Modeling Language, Cognition, and Action: Proceedings of the 9th Neural Computation and Psychology Workshop*, pages 47–56. World Scientific, 2005. 1

[5] C. Christodoulopoulos, S. Goldwater, and M. Steedman. Two decades of unsupervised POS induction: How far have we come? In *Proceedings of the 15th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 575–584, Cambridge MA, USA, 2010. 3

[6] K.W. Church. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the 2nd Conference on Applied Natural Language Processing (ANLC)*, Austin TX, USA, 1988. 3

[7] C. Craye, D. Filliat, and J.F. Goudou. Environment exploration for object-based visual saliency learning. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Stockholm, Sweden, 2016. 2

[8] C. Craye, D. Filliat, and J.F. Goudou. RL-IAC: An exploration policy for online saliency learning on an autonomous mobile robot. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Daejeon, South Korea, 2016. 2

[9] D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun. A practical Part-of-Speech tagger. In *Proceedings of the 3rd Conference on Applied Natural Language Processing (ANLC)*, Trento, Italy, 1992. 3

[10] C.R. Dawson, J. Wright, A. Rebguns, M.V. Escarcega, D. Fried, and P.R. Cohen. A generative probabilistic framework for learning spatial language. In *Proceedings of the 3rd Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics (ICDL)*, Osaka, Japan, 2013. 1

[11] M.A. Fischler and R.C. Bolles. Random Sample Consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24 (6):381–395, 1981. 2

[12] J. Gao and M. Johnson. A comparison of Bayesian estimators for unsupervised Hidden Markov Model POS taggers. In *Proceedings of the 13th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 344–352,

Honolulu HI, USA, 2008. 3

[13] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6 (6):721–741, 1984. 4

[14] S. Guadarrama, L. Riano, D. Golland, D. Gohring, Y. Jia, D. Klein, P. Abbeel, and T. Darrell. Grounding spatial relations for human-robot interaction. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Tokyo, Japan, 2013. 2

[15] S. Harnad. The symbol grounding problem. *Physica D*, 42: 335–346, 1990. 1

[16] W.P. Headden, D. McClosky, and E. Charniak. Evaluating unsupervised part-of-speech tagging for grammar induction. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, Manchester, United Kingdom, 2008. 2

[17] X.Y. Jiang, U. Meier, and H. Bunke. Fast range image segmentation using high-level segmentation primitives. In *Proceedings of the 3rd IEEE International Workshop on Applications of Computer Vision (WACV)*, Sarasota Fl, USA, 1996. 2

[18] K. Koster and M. Spann. MIR: An approach to robust clustering application to range image segmentation. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 22 (5), 2000. 2

[19] B. Landau and R. Jackendoff. 'What' and 'where' in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16:217–238, 1993. 1

[20] C. Liu, J. Walker, and J.Y. Chai. Ambiguities in spatial language understanding in situated human robot dialogue. In *Proceedings of the AAAI Fall Symposium: Dialog with Robots*, Arlington VA, USA, 2010. 1

[21] C. Matuszek, N. FitzGerald, L. Zettlemoyer, L. Bo, and D. Fox. A joint model of language and perception for grounded attribute learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, Edinburgh, Scotland, 2012. 1

[22] R. Moratz, T. Tenbrink, J. Bateman, and K. Fischer. Spatial knowledge representation for human-robot interaction. In C. Freksa, W. Brauer, C. Habel, and K.F. Wender, editors, *Spatial Cognition III*, pages 263–286. Springer, 2003. 1

[23] G. Neubig. Simple, correct parallelization for blocked Gibbs sampling. Technical report, Nara Institute of Science and Technology, 2014. 4

[24] A. Nguyen and B. Le. 3D point cloud segmentation: A survey. In *Proceedings of the 6th IEEE International Conference on Robotics, Automation, and Mechatronics (RAM)*, Manila, Philippines, 2013. 2

[25] P.Y. Oudeyer. *Self-Organization in the Evolution of Speech*, volume 6. Oxford University Press, Oxford, UK, 2006. 2

[26] K. Plunkett, C. Sinha, M.F. Moller, and O. Strandsby. Symbol grounding or the emergence of symbols? vocabulary growth in children and a connectionist net. *Connection Science*, 4 (3–4):293–312, 1992. 2

[27] T. Regier. *The Human Semantic Potential: Spatial Language and Constrained Connectionism*. MIT Press, Cambridge MA, USA, 1996. 1

[28] B. Rosman and S. Ramamoorthy. Learning spatial relationships between objects. *International Journal of Robotics Research*, 30(11):1328–1342, 2011. 1

[29] D. Roy. Learning visually-grounded words and syntax for a scene description task. *Computer Speech and Language*, 16 (3):353–385, 2002. 1

[30] D. Roy, K-Y. Hsiao, and N. Mavridis. Conversational robots: Building blocks for grounding word meanings. In *Proceedings of the International Workshop on Learning Word Meaning from Non-Linguistic Data (HLT-NAACL)*, 2003. 1

[31] R.B. Rusu, G. Bradski, and J. Hsu R. Thibaux. Fast 3D recognition and pose using the viewpoint feature histogram. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2155–2162, Taipei, Taiwan, 2010. 3

[32] R. Schnabel, R. Wahl, and R. Klein. Efficient RANSAC for point-cloud shape detection. *Computer Graphics Forum*, 26 (2):214–226, 2007. 2

[33] J.M. Siskind. Grounding language in perception. *Artificial Intelligence Review*, 8:371–391, 1994-95. 1

[34] K. Smith, A.D.M. Smith, and R.A. Blythe. Cross-situational learning: An experimental study of word-learning mechanisms. *Computer Graphics Forum*, 35(3):480–498, 2011. 3, 5

[35] L. Steels. The symbol grounding problem has been solved, so what's next? In M. de Vega, A. Glenberg, and A. Graesser, editors, *Symbols and Embodiment: Debates on Meaning and Cognition*, pages 223–244. Oxford University Press, 2008. 2

[36] J. Strom, A. Richardson, and E. Olson. Graph based segmentation of colored 3D laser point clouds. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Taipei, Taiwan, 2010. 2

[37] M.K. Tanenhaus, M.J. Spivey-Knowlton, K.M. Eberhard, and J.C. Sedivy. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634, 1995. 1

[38] A. Taniguchi, T. Taniguchi, and A. Cangelosi. Multiple categorization by iCub: Learning relationships between multiple modalities and words. In *Proceedings of the International Workshop on Machine Learning Methods for High-Level Cognitive Capabilities in Robotics, in conjunction with the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Daejeon, South Korea, 2016. 4

[39] T. Taniguchi, T. Nagai, T. Nakamura, N. Iwahashi, T. Ogata, and H. Asoh. Symbol emergence in robotics: A survey. *Advanced Robotics*, 30(11–12), 2016. 2

[40] S. Tellex, T. Kollar, S. Dickerson, M.R. Walter, A.G. Banerjee, S. Teller, and N. Roy. Approaching the symbol grounding problem with probabilistic graphical models. *AI Magazine*, 32(4):64–76, 2011. 1

[41] K. Toutanova and M. Johnson. A Bayesian LDA-based model for semi-supervised part-of-speech tagging. In *Proceedings of the 20th International Conference on Neural Information Processing Systems (NIPS)*, pages 1521–1528, Vancouver, Canada, 2007. 3