Representing Spatial Object Relations as Parametric Polar Distribution for Scene Manipulation Based on Verbal Commands

Rainer Kartmann, You Zhou, Danqing Liu, Fabian Paus and Tamim Asfour

***Abstract*— Understanding spatial relations is a key element for natural human-robot interaction. Especially, a robot must be able to manipulate a given scene according to a human verbal command specifying desired spatial relations between objects. To endow robots with this ability, a suitable representation of spatial relations is necessary, which should be derivable from human demonstrations. We claim that polar coordinates can capture the underlying structure of spatial relations better than Cartesian coordinates and propose a parametric probability distribution defined in polar coordinates to represent spatial relations. We consider static spatial relations such as *left of*, *behind*, and *near*, as well as dynamic ones such as *closer to* and *other side of*, and take into account verbal modifiers such as *roughly* and *a lot*. We show that adequate distributions can be derived for various combinations of spatial relations and modifiers in a sample-efficient way using Maximum Likelihood Estimation, evaluate the effects of modifiers on the distribution parameters, and demonstrate our representation's usefulness in a pick-and-place task on a real robot.**

I. Introduction

Humans use spatial relations in natural language to describe tasks to other people, e. g., put the cup to the left of the plate. When such a description is given to a robot, the robot needs to understand and reason about spatial object relations, i. e. what does being left of another object mean. Such understanding of spatial object relations in robotics must also consider the actions that are needed to transfer a given object configuration in the scene to a goal configuration using executable robot skills. These object-action relations are essential for task execution in robotics and research have been conducted to derive representations that bridge the gap between sensorimotor experience and symbolic planning and reasoning. An example for such representations are Object-Action Complexes (OACs), which have been proposed to provide a unifying representation that has the ability to represent and reason about low-level sensorimotor experience as well as high-level symbolic information in a robot architecture and thus to bridge the gap between low-level sensorimotor representations, required for robot perception and control, and high-level representations supporting abstract reasoning, planning and natural language understanding [1]. Other works address specifically the problem of spatial relations learning to endow a robot with the ability to describe a scene's structure by classifying the displacement between two objects [2], to compare spatial relations by learning a distance metric between scenes [3], as well as to encode and recognize actions based on changes of spatial relations using similarity scores [4] and graph networks [5]. Instead of describing or classifying a scene, our goal is the realization of a given spatial relation by manipulating the current scene [6].

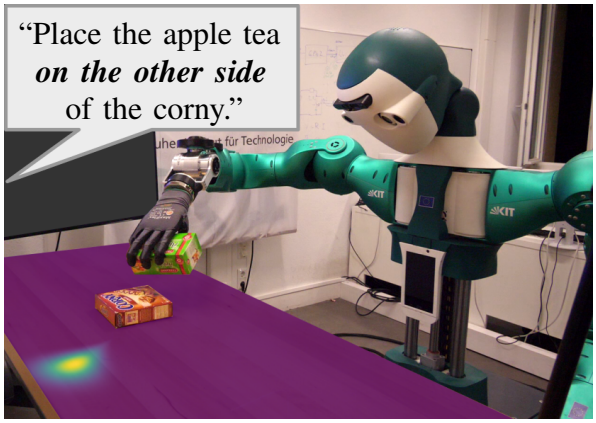


Fig. 1: Scene manipulation based on verbal commands specifying spatial object relations.

When using natural language for scene manipulation, considerable related work addresses the grounding of referring expressions involving spatial relations to physical entities in the current scene. Tellex et al. [7] propose a probabilistic graphical model to map constituents of a natural language command to objects and places as well as robot actions. Shridhar et al. [8] present a framework for grounding reference expressions to objects in an image and resolving ambiguities by using spatial relations. However, instead of limiting possible action parameters to the objects and landmarks present in the scene, we propose a subsymbolic representation of spatial relations able to generate arbitrary placing positions.

Due to the imprecise and subjective nature of spatial relations and their natural language description, a representation of spatial relations needs to be intuitively specified by the human. Since a human can only give a limited number of examples, a sample-efficient way to learn the parameters of said representation is required. This learned representation should generalize to new scene states and objects. Furthermore, the representation should be linked to the motion generation enabling goal-directed scene manipulation.

We consider the representation and derivation of spatial relations of one object to another in a two-dimensional plane for a robot pick-and-place task. In this task, a robot is commanded to move a *target object* relative to a *reference object* according to a spatial relation specified in natural language. For instance, in the command "Move the tea

This work has been supported by the German Federal Ministry of Education and Research (BMBF) under the project OML (01IS18040A).

The authors are with the Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Karlsruhe, Germany. {rainer.kartmann, asfour}@kit.edu

closer to the cup," "tea" is the target object and "cup" is the reference object. The goal is to execute a motion that results in repositioning the target object according to the given command. We consider both *static* and *dynamic* relations, where static relations are relative only to the reference object's location while dynamic relations involve both objects' relative position [4]. In addition, we take into account verbal *modifiers*, which are words that affect the quantitative meaning of a spatial relation in an imprecise manner [9], such as in *roughly left*, *a bit closer* or *not in front of*. We consider spatial relations only from the robot's point of view.

We require a representation of spatial relations to be generative, i. e. allow the generation of multiple target position candidates. To achieve this, we follow the approach of representing spatial relations as probability distributions over valid positions of the target object. However, instead of predicting these distributions over pixels of an image as done in [6], we propose to represent spatial relations using a parametric probability distribution. A parametric distribution offers the benefit of being characterized by a limited number of parameters which can be intuitively interpreted.

The main contribution of this work is a novel, understandable representation of spatial relations as parametric probability distributions based on polar coordinates. This representation can be effectively estimated from demonstration in a sample-efficient manner. We show that the derived representation is suitable to generate target positions which can be used to adapt a movement primitive in a robot pick-and-place task.

II. Related Work

A. Spatial Relations for Scene Understanding

Rosman et al. [2] propose to use the spatial relations *on* and *adjacent* as symbolic scene descriptions, which are determined by classifying the displacement between two objects. Sjöö et al. [10] learn to classify functional relations such as *support force* from geometric object features in simulation. In [11], spatial object relations are extracted from 3D vision using histograms. These works focus on extracting object relations from a given scene, but our goal is to manipulate the scene according to a verbal command. Considerable work has been conducted on using spatial relations for action recognition. In [12], spatial relations are extracted by partitioning the space around an object relative to auxiliary coordinate frames aligned to the workspace. This information is used to classify actions in an RGB-D image sequence. Similar approaches are described in [4], [13], [14] and [15], where Semantic Event Chains (SECs) are derived from human demonstrations to describe and recognize manipulation actions and tasks. Recently, graph networks have been used to segment actions in RGB-D videos based on frame-wise extracted spatial relations [5].

Another part of the literature uses spatial relations to resolve referential expressions in natural language utterances. In an early work, Stopp et al. [16] introduce the applicability of different spatial relations as potential fields around an object based on user-defined cubic splines. In [17], spatial relations are represented as fuzzy memberships to nine image regions around an object to identify an object referred to in a command. Spatial relations are used to enrich a robot's knowledge from daily human instructions in [18]. In these works, no actions were executed on a robot. In the context of robot programming by demonstration, referential expressions are used e. g. in [19] to parametrize primitive actions. However, distances are fixed at all times and action parameters have a limited number of values. In [20], spatial relations are used as a basis to ground abstract spatial concepts such as columns or groups of objects for natural language interaction. Shridhar et al. [21] use spatial relations to resolve ambiguities when referring to similar objects.

The spatial relations used in these works are mainly discriminative and used to classify scenes and actions or to identify referred objects, while we require a generative representation of spatial relations to execute manipulation actions to realize a given spatial relation.

B. Realizing Spatial Relations

The following work deals with the question of how to change a given scene to fulfill given spatial relations. In [3], Mees et al. learn a distance metric to find a relative object pose representing the same relation as known scene examples with potentially different objects. The authors build upon this work in [22], where they describe how gradient descent can be performed on object poses to change a given scene to represent the same relation as a reference scene. Prepositions from natural language commands are incorporated as action parameters in a task representation in [23]. However, fixed positions or offsets are used in action execution. Placement positions are found based on spatial prepositions in [24] by training a multi-class logistic regression to classify random positions on a table. In contrast, our goal is to find generative representations of spatial relations abstracting from examples, which can be directly used to generate target positions. Also, these works do not consider dynamic relations and verbal modifiers. In a work very similar to ours, Mees et al. [6] train a neural network to predict pixel-wise probability maps of placement positions for different spatial relations given an input image of the scene to place an object according to a verbal command. We, however, derive parametric probability distributions representing spatial relations, which can be intuitively interpreted and estimated from few examples. In addition, our representation is not restricted to an input image. While the methods used in [6] could be extended to include dynamic relations and verbal modifiers, these aspects are currently not taken into account.

C. Uncertainty of Spatial Relations and Natural Language

Spatial relations between objects involve some degree of uncertainty. For instance, there is no exact boundary between regions described by *left of* and *in front of*. To cope with uncertainties in spatial relations, a continuous measure of applicability based on user-defined cubic splines is used in [16], while Tan et al. [17] use fuzzy membership of locations to defined regions around the object. Skubic et al. [25] divide

directions around the robot into 16 sub-directions to represent utterances such as "mostly in front but somewhat to the left". Instead, uncertainty and fuzziness of spatial relations are naturally embedded in our probabilistic representations in the form of variance parameters.

As mentioned above, verbal instructions can contain verbal modifiers introducing more uncertainty. For example, a human might prefer saying "Move it a bit closer." rather than "Move it 5 cm closer." In addition, one could instruct the robot to place an object "roughly to the left of" or "not in front of" another. A comprehensive analysis of uncertain information in natural language instructions for robotic applications is provided by [9]. Instead of modeling such modifiers in a special way, we take them into account by estimating probability distributions separately for each combination of spatial relation and modifier and show how this affects the estimated distribution parameters.

D. Learning Spatial Arrangements of Objects

Spatial relations also play a role in the generation of typical or natural arrangements of objects. Jiang et al. [26] predict the correct placements of objects in a 3D scene based on their affordances and potential human poses, i. e. by modeling human-object relations instead of object-object relations. The authors address finding good placement locations with respect to stability and support, i. e. based on relations between objects and the environment, in [27]. However, we are interested in arranging objects based on a spatial relation between them specified by a verbal command. Chang et al. [28] build scene templates from example scenes to generate 3D scenes according to a textual description. They consider typical relative arrangements and support surfaces of objects based on their categories as well as objects which are likely to be present despite not being mentioned explicitly. Although taking into account typical arrangements can be useful, we focus on manipulating a concrete scene to realize a specific spatial relation.

III. Spatial Relations as Polar Distributions

A. Cartesian vs. Polar Space

For the reasons explained above, we aim at modeling spatial relations as parametric probability distributions. The first choice might be to represent a spatial relation as a multivariate Gaussian distribution in Cartesian space. However, this is not suitable for relations such as *close to* and *far away from*, which should cover positions in any direction from the reference object o_{ref}. For these relations, the best option for a Cartesian Gaussian distribution would involve placing the mean inside o_{ref} and adjusting the variances to fit the examples. As a result, positions inside o_{ref} would be considered most likely, which is not desired. In essence, a Cartesian Gaussian distribution fails to represent a "ring" around the reference object. A solution might be to use a Gaussian Mixture Model to construct this ring around the reference object o_{ref}. However, this would require specifying the number of components beforehand.

Instead, we propose to represent a spatial relation not in Cartesian space, but in polar space. As polar coordinates represent positions as direction and distance, we claim that they capture the underlying structure of spatial relations better than Cartesian coordinates. To this end, we define a polar coordinate system (PCS) centered at the reference object, and define a two-dimensional distribution in this polar coordinate space. This way, *far away from* can be represented by covering all angles, but only sufficiently large distances.

B. Polar Distribution

We denote a polar coordinate with distance $d \in \mathbb{R}_{\geq 0}$ and angle $\phi \in [-\pi, \pi]$ as

$$\mathbf{q} = \begin{pmatrix} d & \phi \end{pmatrix}^\top \in \mathbb{R}_{\geq 0} \times [-\pi, \pi]\,. \tag{1}$$

We assume that d follows a Gaussian distribution

$$d \sim \mathcal{N}\left(\mu_d, \sigma_d^2\right), \tag{2}$$

with mean μ_d and variance σ_d^2. As the angle ϕ is periodic over $[-\pi, \pi]$, we utilize the von Mises distribution

$$\phi \sim \mathcal{M}(\mu_\phi, \kappa_\phi)\,, \tag{3}$$

with mean μ_ϕ and concentration κ_ϕ, which is a circular distribution behaving similarly to a Gaussian distribution but wrapping over the interval $[-\pi, \pi]$. For easier comparison, we will refer to the concentration's inverse as variance $\sigma_\phi^2 \coloneqq \kappa_\phi^{-1}$ and use it instead of κ_ϕ.

We define a polar distribution as joint probability distribution $d, \phi \sim \mathcal{P}(\theta)$ with $\theta = (\mu_d, \sigma_d^2, \mu_\phi, \sigma_\phi^2)$ being the distribution's parameters. In our model, we assume that d and ϕ are independent. Therefore, the joint probability density function (p.d.f.) is

$$p(d, \phi) = p(d) \cdot p(\phi)\,. \tag{4}$$

To evaluate a polar distribution in Cartesian space, global Cartesian positions must be transformed to polar space. A local Cartesian position $\mathbf{p}_{\mathrm{loc}} = (x, y)^\top \in \mathbb{R}^2$ is transformed to polar coordinates by

$$\mathrm{polar}(\mathbf{p}_{\mathrm{loc}}) = \begin{pmatrix} d \\ \phi \end{pmatrix} = \begin{pmatrix} \sqrt{x^2 + y^2} \\ \mathrm{atan2}(y, x) \end{pmatrix}. \tag{5}$$

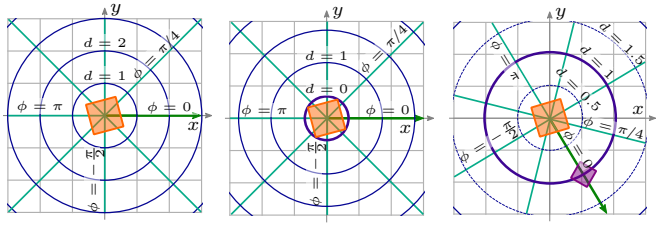
The resulting coordinate system is visualized in Fig. 2a. A global Cartesian position $\mathbf{p}_{\mathrm{glob}} = (x, y)^\top \in \mathbb{R}^2$ is translated to the reference object's local coordinate system before transforming it to polar coordinates:

$$\mathbf{q} = \mathrm{polar}(\mathbf{p}_{\mathrm{glob}} - \mathbf{p}_{\mathrm{ref}}) \tag{6}$$

C. Static vs. Dynamic Relations

As noted in [4], there are two types of spatial relations with respect to time: *static* and *dynamic* spatial relations. Static relations (e. g. *left of*, *close to*) are independent of the target object's current position. In contrast, dynamic relations (e. g. *closer to, on other side of*) depend on both the target and reference objects' positions.

When estimating polar distributions to represent dynamic relations without respecting this dependency, the resulting distributions will cover all angles and distances encountered in the examples, not learning a meaningful representation for



(a) Original (b) Static (c) Dynamic

Fig. 2: Polar coordinate systems (PCS). In all cases, the origin of the PCS is the reference object's position (orange). Its local Cartesian coordinate system (x, y) is shown in gray. The PCS is shown in blue (distance d) and green (angle ϕ). (a) Original definition of PCS. (b) PCS for static relations. Distance $d = 0$ is translated to the object's size r_{ref}. (c) PCS for dynamic relations. Distance $d = 1$ is scaled to the current distance to the target object (shown in purple). Angle $\phi = 0$ is translated to the current direction to the target object.

a specific relation. Therefore, we explicitly incorporate this dependency by aligning the polar coordinate system around the reference object to fulfill two constraints:

1) The distance $d = 1$ corresponds to the current distance between reference and target object.
2) The angle $\phi = 0$ corresponds to the current direction from reference to target object.

This idea is shown in Fig. 2c. Formally, let $\mathbf{p}_{\text{ref}}$ and $\mathbf{p}_{\text{trg}_0}$ be the reference and target objects' initial positions, the dynamic polar coordinate of a global position $\mathbf{p}$ is defined by

$$\text{polar}_{\text{dyn}}\left(\mathbf{p}, \mathbf{p}_{\text{ref}}, \mathbf{p}_{\text{trg}_0}\right) := \begin{pmatrix} d \cdot d_{\text{trg}}^{-1} \\ \phi - \phi_{\text{trg}} \end{pmatrix} \tag{7}$$

with

$$\left(d_{\text{trg}}, \phi_{\text{trg}}\right)^\top = \text{polar}\left(\mathbf{p}_{\text{trg}_0} - \mathbf{p}_{\text{ref}}\right) , \tag{8}$$

$$\left(d, \phi\right)^\top = \text{polar}\left(\mathbf{p} - \mathbf{p}_{\text{ref}}\right) . \tag{9}$$

Note that $\phi - \phi_{\text{trg}}$ in eq. (7) is wrapped over $[-\pi, \pi]$.

As to static relations, while using the original polar coordinates as defined in eq. (6) yielded good results in our experiments, we found that, especially for the *inside* relation, it is useful to let the distance $d = 0$ correspond to the reference object's size $r_{\text{ref}} \in \mathbb{R}_+$:

$$\text{polar}_{\text{sta}}(\mathbf{p}, \mathbf{p}_{\text{ref}}, r_{\text{ref}}) := \begin{pmatrix} d - r_{\text{ref}} \\ \phi \end{pmatrix} \tag{10}$$

with d, ϕ defined as in eq. (9). Note that in this case, the resulting distance $d \in [-r_{\text{ref}}, \infty)$. As a size measure r_{ref}, we use the radius of the object's bounding circle. Fig. 2b illustrates this transformation. With this convention, we found that the estimated distributions generalize better over reference objects of different size. Table I specifies the respective type of each relation.

D. Data Collection and Distribution Estimation

To estimate polar distribution for different relations and modifiers, we collected examples using an interactive data collection tool. The tool generates random verbal commands based on sentence templates as well as random scenes containing two objects. The user is presented with both command and scene and can move the target object to valid positions according to the command.

TABLE I
GENERATED SPATIAL RELATIONS, THEIR TYPES AND COMBINATIONS WITH MODIFIERS

Type	Relations	Modifiers			
		–	*not*	*roughly*, *exactly*	*a bit*, *a lot*
Static	*left*, *right*	✓	✓	✓	
	in front, *behind*	✓	✓	✓	
	near / *close to*	✓	✓		
	away / *far from*	✓			
	inside	✓	✓		
Dynam.	*closer to*	✓			✓
	farther away	✓			✓
	opposite side	✓	✓	✓	

The generated commands include the relations *left*, *right*, *in front*, *behind*, *near*, *far*, *inside*, *closer*, *farther* and *opposite side* and the modifiers *not*, *roughly*, *exactly*, *a bit* and *a lot*. Table I shows the generated combinations. Each relation and modifier can be instantiated by different words. For example, the relation *near* can also be expressed as *close to* and *around*. Similarly, the modifier *not* can as well be instantiated as *anywhere other than*. Examples of generated commands are "Place the milk so that it stands not left of the rusk." and "Move salt roughly to the other side of the popcorn."

The objects are taken from a list of household objects containing their names, shapes and sizes. The shapes include rectangular, circular and elliptic objects. The objects' sizes are randomly varied. The reference object is always placed in the center of the screen; its orientation and the target object's pose are chosen randomly.

Each entered target object position creates one sample. Each sample i consists of the command sentence and the reference and target objects' names, their initial positions $\mathbf{p}^i_{\text{ref}}$, $\mathbf{p}^i_{\text{trg}_0}$ and orientations, as well as the target object's final position $\mathbf{p}^i_{\text{trg}_1}$ and orientation.

Using this tool, we generated data to estimate polar distributions representing spatial relations. In total, we collected 5275 samples from three participants. We partitioned the samples according to relation and modifier in the command using keyword matching. For each combination of relation and modifier, we estimate one single polar distribution. To this end, we rely on established Maximum Likelihood Estimation (MLE). Given the samples of a combination of relation and modifier, we transform the target object's final positions $\mathbf{p}^i_{\text{trg}_1}$ to polar coordinates $\mathbf{q}^i$ according to eq. (7) and eq. (10) and the respective relation's type:

$$\left(d^i, \phi^i\right)^\top \leftarrow \begin{cases} \text{polar}_{\text{sta}}\left(\mathbf{p}^i_{\text{trg}_1}, \mathbf{p}^i_{\text{ref}}, r^i_{\text{ref}}\right) , & \text{static} \\ \text{polar}_{\text{dyn}}\left(\mathbf{p}^i_{\text{trg}_1}, \mathbf{p}^i_{\text{ref}}, \mathbf{p}^i_{\text{trg}_0}\right) , & \text{dynamic} \end{cases}$$

for $i = 1 \ldots N$, with N being the respective number of samples. For static relations, $\mathbf{p}^i_{\text{trg}_0}$ is not used. We perform MLE separately for distances and angles:

$$\mu_d, \sigma_d^2 \leftarrow \text{MLE}_{\mathcal{N}}\left(\left\{d^i\right\}_{i=1\ldots N}\right) \tag{11}$$

$$\mu_\phi, \sigma_\phi^2 \leftarrow \text{MLE}_{\mathcal{M}}\left(\left\{\phi^i\right\}_{i=1\ldots N}\right) \tag{12}$$

The resulting polar distribution is then specified by its parameters $\theta = (\mu_d, \sigma_d^2, \mu_\phi, \sigma_\phi^2)$ as described in Section III-B.

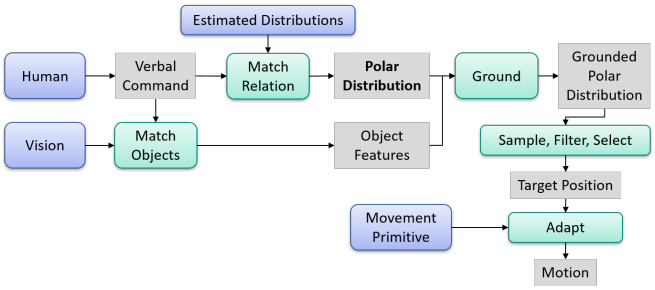


Fig. 3: Using estimated polar distributions in pick-and-place task.

IV. Relational Pick-and-Place Task

Now, we describe how we utilize polar distributions representing spatial relations to place objects according to a verbal command. Fig. 3 shows an overview of the approach.

A. Retrieval and Grounding of Spatial Relation

In our experiments, we only consider known objects which we can localize based on visual features using our previous work in [29]. For each object, we store the class label, relevant geometric information and the visual features needed for localization in the robot's prior knowledge memory. Before receiving a verbal command, we scan the robot's workspace for known objects, and store detected objects and their poses in the robot's working memory.

When given a verbal command, as natural language understanding is not our focus, we perform simple keyword matching to extract the referred objects and desired relation (including a potential modifier). To this end, we search in the verbal command for known object names. We assume that the first object name specifies the target object and the second refers to the reference object. For relations and modifiers, we search for the same keywords as used in Section III. We then retrieve the desired relation's polar distribution parameters θ from the set of previously estimated distributions. Given the referred object entities and relation, we ground the relation's polar distribution to the current scene by aligning the polar coordinate system to the involved objects according to the relation's type as explained in Section III-C.

B. Generation and Selection of Target Position

In order to facilitate collision-free object placing, we exploit the representation by sampling multiple placing positions from the probability distribution. In the following, let $\mathcal{P}(\theta)$ be the used polar distribution, $p(d, \phi)$ its probability density function (p.d.f.), and $\mathbf{p}^i$ and $\mathbf{q}^i$ $(i = 1 \dots n)$ the n sampled Cartesian and polar candidate positions, respectively. An example is shown in Fig. 4.

Sampling enables filtering infeasible placing positions, i. e. those which would collide with other objects already present in the scene or which are not on the table. Let $I_{\text{feas}} \subseteq \{1, \dots, n\}$ be the feasible samples. As sampling can produce positions with low likelihood, which will be less representative for the relation, we evaluate the probability density function at each position and select the candidates

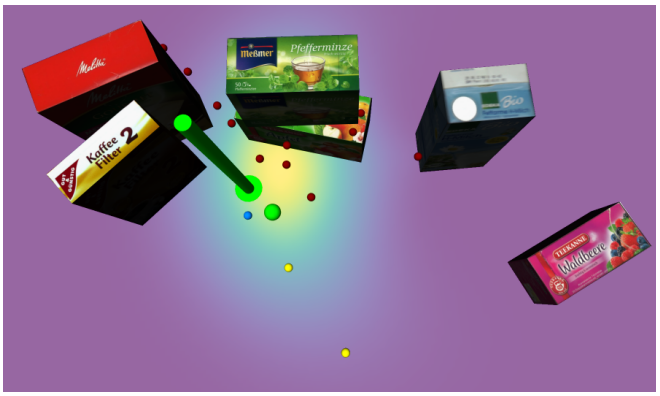


Fig. 4: Sampling, filtering and selection of target position for the command "Put the wild berry tea roughly to the left of the milk." The initial samples drawn from the distribution are shown as small spheres. The red samples have been discarded due to potential collisions. From the remaining samples, the blue and green samples have the highest p.d.f. values. From these two, the green sample, highlighted by an arrow, is selected as it is closer to the target object's current position.

with the highest values:

$$I_{\text{best}} = \left\{ i \in I_{\text{feas}} \;\middle|\; p(\mathbf{q}^i) \geq 0.9 \cdot \max_{j \in I_{\text{feas}}} p(\mathbf{q}^j) \right\} \tag{13}$$

From the remaining positions we select the position $\mathbf{p}^{i^*}$ closest to the target object's current position to avoid unnecessary movement:

$$i^* = \arg\min_{i \in I_{\text{best}}} \left\| \mathbf{p}^i - \mathbf{p}_{\text{trg}} \right\| \tag{14}$$

C. Adaption of Movement Primitive

We use via-points movement primitive (VMP), described in our previous work [30], to represent robot actions as they provide a flexible way to adapt robot actions to different start and goal positions and allow integrating arbitrary via-points. A VMP consists of two parts: an elementary trajectory $h(x)$ and a shape modulation term $f(x)$:

$$y(x) = h(x) + f(x) \tag{15}$$

The elementary trajectory can be any polynomial. In this work, we consider a simple form:

$$h(x) = (y_0 - g)x + g\,, \tag{16}$$

where y_0 is the start and g is the goal. x is the canonical variable that goes from 1 to 0. The shape modulation term is a linear regression model $f(x) = \psi(x)^\top \mathbf{w}$, where $\psi(\cdot)$ is the squared exponential kernels. We learn the weights vector $\mathbf{w}$ from demonstrations. After learning, VMP can adapt to different goals by changing the hyper-parameters g in the elementary trajectory.

We learn VMP using kinesthetic teaching. After learning, we adapt the VMP to different target positions given by the sampling process. In the meantime, we use several via-points to meet some task requirements. As an example, in the pick-and-place task, we use one via-point to avoid collisions with the table while approaching the target object. In this way,

collision-free execution is guaranteed, no matter where the target object is.

V. Evaluation

In this section, we test the following hypotheses: 1) polar distributions are suitable generative representations of spatial relations, 2) the estimated distributions differ meaningfully for different verbal modifiers, 3) estimating polar distributions from examples is sample efficient and 4) the generated spatial relations can be used to change a scene through pick and place tasks using verbal commands on a real robot. To this end, we will provide qualitative results and quantitative analysis of the polar distributions estimated from collected data and demonstrate their usefulness in a real robot experiment.

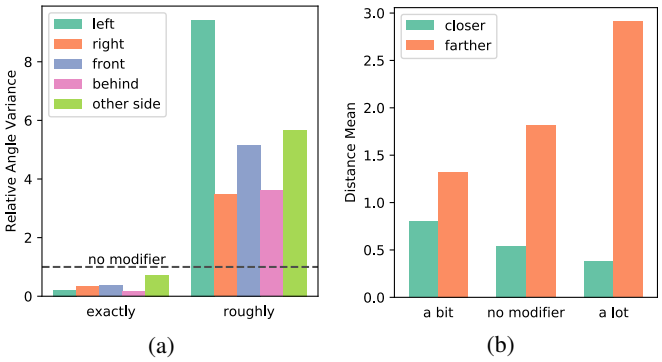
A. Estimated Distributions

First, we provide a qualitative evaluation of the estimated distributions for all relations and modifiers. Table II shows all estimated polar distributions for all considered combinations of relations and modifiers. The images show a top view on an example scene in Cartesian space, where the green milk box is the reference object and the red coffee filters represent the target object. Yellow regions have a high p.d.f. value, while purple regions have a p.d.f. close to zero. For the relations *left of*, *right of*, *in front of*, *behind* and *other side of* it is visible how modifiers *exactly* and *roughly* change the angle variance. Interestingly, for these relations, the distributions for modifier *not* seem to complement the affirmative exemplars mainly in terms of direction, but not in terms of distance. As a consequence, these distributions still represent locations in the area around the reference object. The relations *near*, *far from* and *inside* feature very high angle variance but distinct distance distributions. Especially, it appears that the distributions for *not near* and *far from* are very similar[1]. For the dynamic relations *closer to* and *further from* which take the modifiers *a bit* and *a lot*, it can be seen that the mean direction points to the target object and that the modifiers strongly affect the distance means and variances. Overall, the results indicate that meaningful distributions were estimated for all relations and modifiers.

B. Effects of Modifiers

We conducted a quantitative analysis of the differences between distributions for different modifiers. First, we investigated the estimated angle variances σ_ϕ^2 for *exactly* and *roughly*. To make them comparable for different relations, we normalize each σ_ϕ^2 by the value for the respective relation without a modifier. The results are shown in Fig. 5a. As can be seen, angle variances for *exactly* are lower, and those for *roughly* are higher when compared to no modifier. We found that distance variance σ_d^2 behaves similarly, but there the effect is less strong. We performed a similar analysis for the effects of *a bit* and *a lot* on the distance mean of relations *closer* and *farther*. We did not normalize the values to allow direct comparison between the relations and report the results

[1]In fact, their Bhattacharyya distance D_B is 0.146. For details on D_B see Section V-C.



(a) (b)

Fig. 5: Effects of modifiers on distribution. (a) Relative angle variance for modifiers *exactly* and *roughly*. (b) Distance mean for modifiers *a little* and *much*.

(including no modifier) in Fig. 5b. It can be seen that *a lot closer* results in a distance mean closer to the reference object than *closer* and *a bit closer*. For *farther*, the inverse behavior is observable.

C. Sample Efficiency

To evaluate the sample efficiency of polar distribution estimation, we conducted an experiment as follows: First, for each combination of relation r and modifier m, we randomly draw 100 samples $\mathcal{D}_{r,m}$ from our data set to avoid effects of imbalances. Second, we take k samples from $\mathcal{D}_{r,m}$ and use them to estimate a polar distribution. We do this for $k \in \mathcal{K} = \{2, 3, 5, 10, 25, 50, 75, 100\}$ and repeat each sampling and estimation $R = 10$ times.

Let $\mathcal{P}_{r,m}^{(k,i)}$ be the distribution estimated using $k \in \mathcal{K}$ samples in repetition $i \in \{1, \dots, R\}$, and $\mathcal{P}_{r,m}^*$ be the distribution estimated using $k = 100$ (i. e. all samples). We compared the $\mathcal{P}_{r,m}^{(k,i)}$ with $\mathcal{P}_{r,m}^*$ using two measures:

1) The data's mean likelihood given $\mathcal{P}_{r,m}^{(k,i)}$ relative to its likelihood given $\mathcal{P}_{r,m}^*$.
2) The Bhattacharyya distance of $\mathcal{P}_{r,m}^{(k,i)}$ to $\mathcal{P}_{r,m}^*$.

In the following, each measure is explained and the results are discussed.

1) Relative Mean Likelihood: We define the mean likelihood $\bar{L}(\mathcal{D}, \mathcal{P})$ of data $\mathcal{D}$ given a distribution $\mathcal{P}$ with p.d.f. p as exponential of its mean log-likelihood $\overline{LL}(\mathcal{D}, \mathcal{P})$:

$$\bar{L}(\mathcal{D}, \mathcal{P}) = \exp\big(\overline{LL}(\mathcal{D}, \mathcal{P})\big) = \exp\left(\frac{1}{|\mathcal{D}|} \sum_{\mathbf{q} \in \mathcal{D}} \log p(\mathbf{q})\right)$$

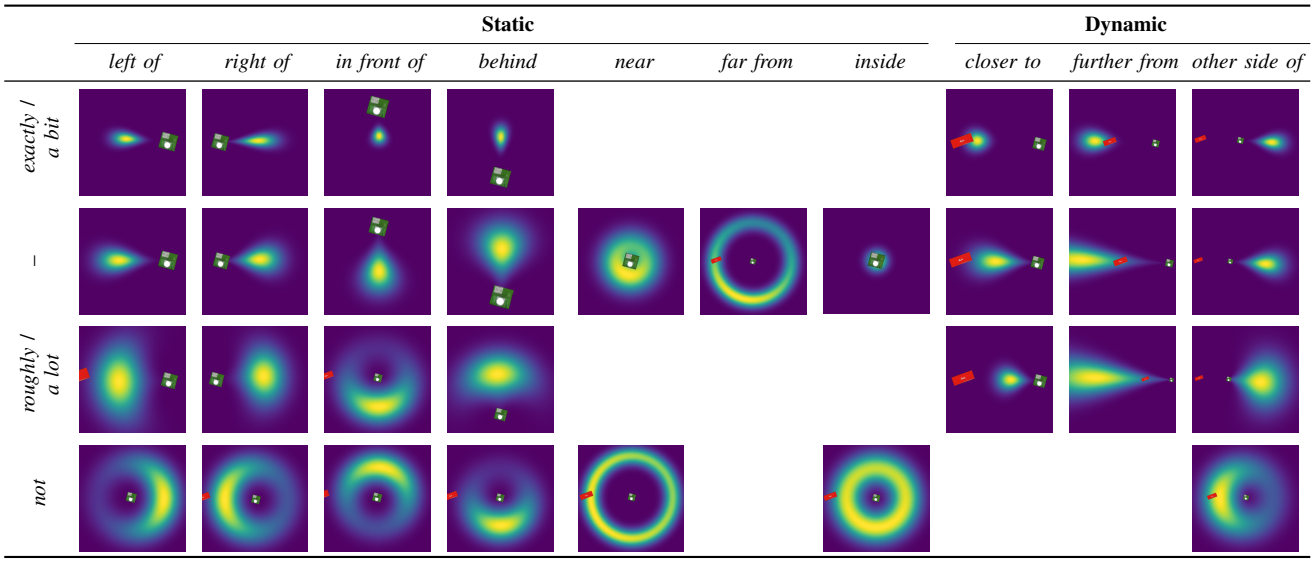
We compute the mean likelihoods of $\mathcal{D}_{r,m}$ given each $\mathcal{P}_{r,m}^{(k,i)}$. In this experiment, $\bar{L}\big(\mathcal{D}_{r,m}, \mathcal{P}_{r,m}^*\big)$ constitutes the optimal performance on $\mathcal{D}_{r,m}$, which is different for each combination of r and m. Therefore, we compute the *relative* mean likelihood of $\mathcal{D}_{r,m}$ given $\mathcal{P}$ as:

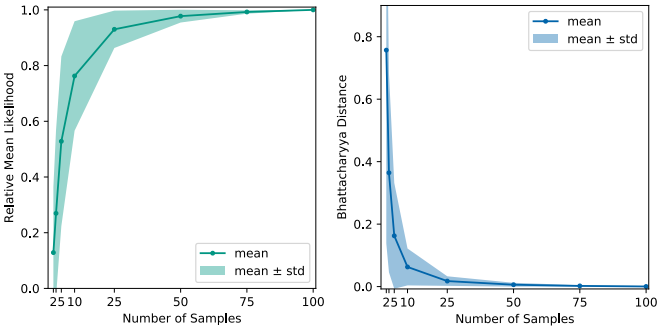
$$\bar{L}_{\text{rel}}(\mathcal{D}_{r,m}, \mathcal{P}) = \frac{\bar{L}(\mathcal{D}_{r,m}, \mathcal{P})}{\bar{L}\big(\mathcal{D}_{r,m}, \mathcal{P}_{r,m}^*\big)} \tag{17}$$

Consequently, $\bar{L}_{\text{rel}}(\mathcal{D}_{r,m}, \mathcal{P}) = 1$ indicates an optimal estimate, while values close to zero indicate worst estimates. We report the relative mean likelihood as mean with standard

TABLE II

ESTIMATED DISTRIBUTIONS FOR EACH COMBINATION OF RELATION AND MODIFIER IN CARTESIAN SPACE. (TOP VIEW OF SCENE, ROBOT'S VIEW IS FROM BOTTOM OF IMAGES.)





(a) Relative mean likelihood (b) Bhattacharyya distance

Fig. 6: Evaluation of sample efficiency as comparison of estimated distributions using k samples versus 100 samples.

deviation aggregated over all relations r, modifiers m and repetitions i in Fig. 6a.

2) Bhattacharyya Distance: The Bhattacharyya distance measures the distance between two probability distributions with density functions p and q, see [31]. It is defined as

$$D_B(p, q) = -\log C_B(p, q) \tag{18}$$

where Bhattacharyya coefficient C_B measures the overlap between two distributions and is defined as

$$C_B(p, q) = \int_{x \in X} \sqrt{p(x) \cdot q(x)}\, dx \ . \tag{19}$$

C_B is close to 1 for distributions with high overlap and 0 if there is no overlap. For two polar distributions with density functions p, q as in eq. (4), the Bhattacharyya coefficient and distance can be written as:

$$C_B(p, q) = C_B(p_d, q_d) \cdot C_B(p_\phi, q_\phi) \ , \tag{20}$$

$$D_B(p, q) = D_B(p_d, q_d) + D_B(p_\phi, q_\phi) \ , \tag{21}$$

where $p_d = p(d)$, $q_d = q(d)$ describe Gaussian distributions and $p_\phi = p(\phi)$, $q_\phi = q(\phi)$ describe von Mises distributions. To compute $D_B(p_d, q_d)$, we use a closed-form solution for Gaussian distributions given in [32]. For $D_B(p_\phi, q_\phi)$, we numerically compute the integral in eq. (19) over the interval $[-\pi, \pi]$ and apply eq. (18). Using eq. (21) we then obtain the Bhattacharyya distance between two polar distributions.

Similar to relative mean likelihoods, we compute the Bhattacharyya distances between each $\mathcal{P}_{r,m}^{(k,i)}$ and $\mathcal{P}_{r,m}^{*}$. The results are shown in Fig. 6b as mean and standard deviation aggregated over all relations, modifiers and repetitions. Note that two identical distributions have a D_B of 0.

3) Discussion of the Results: Both the relative mean likelihood and Bhattacharyya distance quickly approach their respective optimum with increasing number of samples for all combinations of relations and modifiers. Using $k = 10$ samples already yields a relative mean likelihood of $76.27 \pm 0.20\,\%$, and a Bhattacharyya distance of 0.063 ± 0.059 (with a C_B of 0.941 ± 0.052) compared to using 100 samples. When using $k = 25$ samples, relative mean likelihood achieves $93.02 \pm 0.06\,\%$ and the Bhattacharyya distance decreases to 0.018 ± 0.015 ($C_B = 0.983 \pm 0.015$). Using more samples has only marginal effects on the similarity to using all samples. These findings indicate that representing spatial relations as polar distributions introduces a suitable inductive bias which allows to generalize from few samples.

D. Robot Pick-And-Place Task

We demonstrate the usefulness of the generated spatial relation representations in a robot pick-and-place task. We place different, known objects on a table, enter a verbal command as text and follow the procedure described in Section IV to retrieve the desired relation, ground it to the referred objects, sample and select a suitable target position and execute the grasping and placing motions.

The experiments show that the estimated polar distributions can be used to generate suitable target positions. Especially,

in scenes where the distribution mean is blocked by other objects, their generative and fuzzy nature allow us to still find feasible placing locations.

VI. Conclusion

We have presented a parametric probability distribution defined in polar coordinate space representing spatial relations, which can be estimated from examples and used to manipulate a scene in order to fulfill spatial relations specified in verbal commands. We have shown that estimation is sample-efficient and that the estimated parameters differ between verbal modifiers. In future work, we will extend this representation to three dimensions and orientations and investigate how such representations can be derived from few examples shown in real-world, how they can generalize to objects of different shapes and sizes, and how relations can be mapped to and from speech. In addition, we will work on the integration of this work in our work regarding programming by demonstration.

References

[1] N. Krüger, C. Geib, J. Piater, R. Petrick, M. Steedman, F. Wörgötter, A. Ude, T. Asfour, D. Kraft, D. Omrčen, A. Agostini, and R. Dillmann, "Object-Action Complexes: Grounded Abstractions of Sensorimotor Processes," *Robotics and Autonomous Systems*, vol. 59, pp. 740–757, 2011.

[2] B. Rosman and S. Ramamoorthy, "Learning spatial relationships between objects," *International Journal of Robotics Research*, vol. 30, no. 11, pp. 1328–1342, 2011.

[3] O. Mees, N. Abdo, M. Mazuran, and W. Burgard, "Metric Learning for Generalizing Spatial Relations to New Objects," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sept. 2017, pp. 3175–3182.

[4] F. Ziaeetabar, T. Kulvicius, M. Tamosiunaite, and F. Wörgötter, "Recognition and Prediction of Manipulation Actions Using Enriched Semantic Event Chains," *Robotics and Autonomous Systems (RAS)*, vol. 110, pp. 173–188, Dec. 2018.

[5] C. R. G. Dreher, M. Wächter, and T. Asfour, "Learning Object-Action Relations from Bimanual Human Demonstration Using Graph Networks," *IEEE Robotics and Automation Letters*, Oct. 2019.

[6] O. Mees, A. Emek, J. Vertens, and W. Burgard, "Learning Object Placements For Relational Instructions by Hallucinating Scene Representations," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[7] S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. Teller, and N. Roy, "Understanding Natural Language Commands for Robotic Navigation and Mobile Manipulation," in *Twenty-Fifth AAAI Conference on Artificial Intelligence*, Aug. 2011.

[8] M. Shridhar, D. Mittal, and D. Hsu, "INGRESS: Interactive visual grounding of referring expressions," *The International Journal of Robotics Research*, vol. 39, no. 2-3, pp. 217–232, 2020.

[9] M. A. V. J. Muthugala and A. G. B. P. Jayasekara, "A Review of Service Robots Coping With Uncertain Information in Natural Language Instructions," *IEEE Access*, vol. 6, pp. 12 913–12 928, 2018.

[10] K. Sjöö and P. Jensfelt, "Learning spatial relations from functional simulation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sept. 2011, pp. 1513–1519.

[11] S. Fichtl, A. McManus, W. Mustafa, D. Kraft, N. Krüger, and F. Guerin, "Learning spatial relationships from 3D vision using histograms," in *IEEE International Conference on Robotics and Automation (ICRA)*, May 2014, pp. 501–508.

[12] K. Zampogiannis, Y. Yang, C. Fermuller, and Y. Aloimonos, "Learning the Spatial Semantics of Manipulation Actions Through Preposition Grounding," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 1389–1396.

[13] E. E. Aksoy, Y. Zhou, M. Wächter, and T. Asfour, "Enriched Manipulation Action Semantics for Robot Execution of Time Constrained Tasks," in *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, Nov. 2016, pp. 109–116.

[14] F. Ziaeetabar, T. Kulvicius, M. Tamosiunaite, and F. Wörgötter, "Prediction of Manipulation Action Classes Using Semantic Spatial Reasoning," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2018, pp. 3350–3357.

[15] T. R. Savarimuthu, A. G. Buch, C. Schlette, N. Wantia, J. Roßmann, D. Martínez, G. Alenyà, C. Torras, A. Ude, B. Nemec, A. Kramberger, F. Wörgötter, E. E. Aksoy, J. Papon, S. Haller, J. Piater, and N. Krüger, "Teaching a Robot the Semantics of Assembly Tasks," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no. 5, pp. 670–692, May 2018.

[16] 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," in *KI-94: Advances in Artificial Intelligence*, ser. Lecture Notes in Computer Science. Springer, 1994, pp. 39–50.

[17] J. Tan, Z. Ju, and H. Liu, "Grounding Spatial Relations in Natural Language by Fuzzy Representation for Human-Robot Interaction," in *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, July 2014, pp. 1743–1750.

[18] J. Bao, Z. Hong, H. Tang, Y. Cheng, Y. Jia, and N. Xi, "Teach robots understanding new object types and attributes through natural language instructions," in *International Conference on Sensing Technology (ICST)*, vol. 10, Nov. 2016, pp. 1–6.

[19] M. Forbes, R. P. N. Rao, L. Zettlemoyer, and M. Cakmak, "Robot Programming by Demonstration with Situated Spatial Language Understanding," in *IEEE International Conference on Robotics and Automation (ICRA)*, May 2015, pp. 2014–2020.

[20] R. Paul, J. Arkin, N. Roy, and T. M. Howard, "Efficient Grounding of Abstract Spatial Concepts for Natural Language Interaction with Robot Manipulators," in *Robotics: Science and Systems (RSS)*, vol. 12, June 2016.

[21] M. Shridhar and D. Hsu, "Interactive Visual Grounding of Referring Expressions for Human-Robot Interaction," in *Robotics: Science & Systems (RSS)*, 2018.

[22] P. Jund, A. Eitel, N. Abdo, and W. Burgard, "Optimization Beyond the Convolution: Generalizing Spatial Relations with End-to-End Metric Learning," in *IEEE International Conference on Robotics and Automation (ICRA)*, May 2018, pp. 4510–4516.

[23] M. Nicolescu, N. Arnold, J. Blankenburg, D. Feil-Seifer, S. B. Banisetty, M. Nicolescu, A. Palmer, and T. Monteverde, "Learning of Complex-Structured Tasks from Verbal Instruction," in *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, Oct. 2019, p. 8.

[24] S. Guadarrama, L. Riano, D. Golland, D. Göhring, Y. Jia, D. Klein, P. Abbeel, and T. Darrell, "Grounding Spatial Relations for Human-Robot Interaction," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Nov. 2013, pp. 1640–1647.

[25] M. Skubic, D. Perzanowski, S. Blisard, A. Schultz, W. Adams, M. Bugajska, and D. Brock, "Spatial Language for Human–Robot Dialogs," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 34, no. 2, pp. 154–167, May 2004.

[26] Y. Jiang, C. Zheng, M. Lim, and A. Saxena, "Learning to Place New Objects," in *IEEE International Conference on Robotics and Automation (ICRA)*, May 2012, pp. 3088–3095.

[27] Y. Jiang, M. Lim, and A. Saxena, "Learning Object Arrangements in 3D Scenes using Human Context," in *International Conference on Machine Learning (ICML)*, June 2012.

[28] A. Chang, M. Savva, and C. D. Manning, "Learning Spatial Knowledge for Text to 3D Scene Generation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Oct. 2014, pp. 2028–2038.

[29] P. Azad, T. Asfour, and R. Dillmann, "Combining Harris interest points and the SIFT descriptor for fast scale-invariant object recognition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2009, pp. 4275–4280.

[30] Y. Zhou, J. Gao, and T. Asfour, "Learning Via-Point Movement Primitives with Inter- and Extrapolation Capabilities," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019.

[31] A. K. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bulletin of the Calcutta Mathematical Society*, vol. 35, pp. 99–109, 1943.

[32] G. B. Coleman and H. C. Andrews, "Image Segmentation by Clustering," *Proceedings of the IEEE*, vol. 67, no. 5, pp. 773–785, May 1979.