Desk Organization: Effect of Multimodal Inputs on Spatial Relational Learning

Ryan Rowe*, Shivam Singhal*, Daqing Yi, Tapomayukh Bhattacharjee, and Siddhartha S. Srinivasa

***Abstract*— For robots to operate in a three dimensional world and interact with humans, learning spatial relationships among objects in the surrounding is necessary. Reasoning about the state of the world requires inputs from many different sensory modalities including vision (V) and haptics (H). We examine the problem of desk organization: learning how humans spatially position different objects on a planar surface according to organizational "preference". We model this problem by examining how humans position objects given multiple features received from vision and haptic modalities. However, organizational habits vary greatly between people both in structure and adherence. To deal with user organizational preferences, we add an additional modality, "utility" (U), which informs on a particular human's perceived usefulness of a given object. Models were trained as generalized (over many different people) or tailored (per person). We use two types of models: random forests, which focus on precise multi-task classification, and Markov logic networks, which provide an easily interpretable insight into organizational habits. The models were applied to both synthetic data, which proved to be learnable when using fixed organizational constraints, and human-study data, on which the random forest achieved over 90% accuracy. Over all combinations of {H, U, V} modalities, UV and HUV were the most informative for organization. In a follow-up study, we gauged participants preference of desk organizations by a generalized random forest organization vs. by a random model. On average, participants rated the random forest models as 4.15 on a 5-point Likert scale compared to 1.84 for the random model.**

I. Introduction

Researchers have developed robotic systems that can perform a variety of household tasks ranging from automated cleaning robots [1], to aiding in kitchen tasks [2], [3], to robots assisting the disabled and elderly in their everyday lives [4]. The knowledge of encoding spatial relations of objects is relied on by many in-home tasks such as organizing a desk or a shelf, cleaning an area, or retrieving requested objects.

Spatially organizing objects in turn requires an algorithm to model a person's preferences as well as an object's attributes in order to infer spatial relationships suited to a particular task. In this paper, we investigate this problem further by focusing on the task of autonomous desk organization. Through this task, we are particularly interested in exploring the role of an object's multimodal physical attributes and a person's organizational preferences in learning and inferring spatial relationships.

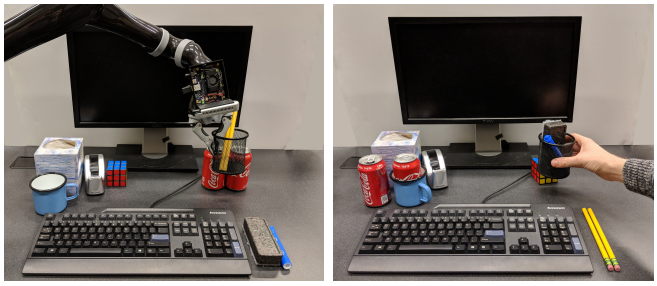


Fig. 1: A sample desk organization task.

Desk organization presents a challenge as the location of objects on an organized desk is a function of not only the object's physical attributes but also a person's preferences of what is "organized" and also their perceived usefulness of each object.

Multimodal learning, which leverages information from inputs from multiple modalities such as vision and haptics, can provide interesting insights into the roles these inputs play in the task of spatially organizing objects. In this paper, we consider physical attributes such as color, shape, size, weight, and rigidity of an object as well as its functional attributes such as its utility in the context of a user's preference in order to infer spatial relations.

We approached learning these spatial relationships from multimodal inputs using both Markov Logic Networks (MLN) [5] and Random Forests [6]. For visual modality, we trained a MaskRCNN [7] over the objects used during experimentation as a proof of concept to detect objects as well as to identify their color and shape. For haptic modality, we used a haptic sensor to obtain values for rigidity. Some physical attributes such as weight (heavy or light), size (large or small), and functional attributes (utility) are subjective and depend on the user. Therefore, we performed human-subject studies and obtained values for these attributes from human responses. Our models use these attributes to learn spatial relationships that a particular user prefers when organizing their desk, then predicts object placements and relationships for new desk situations.

Our results indicate that people do follow latent patterns when organizing objects on a desk. We selected MLN as it provides easily interpretable results in the form of weighted formula, which showed that utility was the most informative modality when determining spatial relationships between objects, followed by vision and lastly haptics. We compared Markov logic networks to a random forest which trades interpretability for increased accuracy, lower training time,

*These authors contributed equally to the work. All the authors are with Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, Washington 98195 {rfrowe, shivam42, dqyi, tapo, siddh}@cs.washington.edu

This work was funded by the National Institute of Health R01 (#R01EB019335), National Science Foundation CPS (#1544797), National Science Foundation NRI (#1637748), the Office of Naval Research, the RCTA, Amazon, Honda, and the UW Allen School Postdoc Research Award. We thank Aditya Mandalika for help with robot experiments.

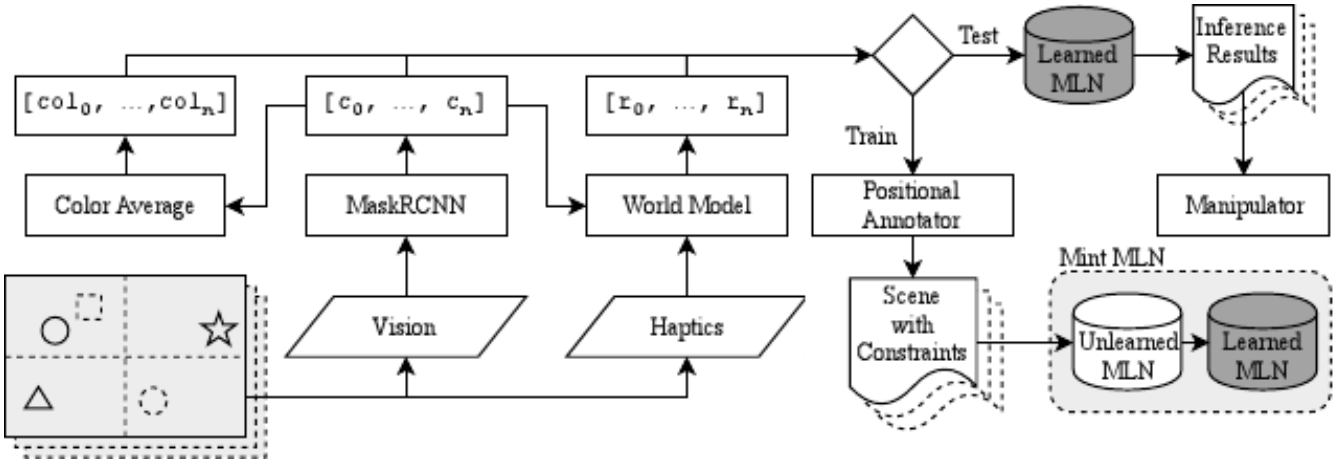


Fig. 2: Example pipeline for the desk organization problem. Utilizes MaskRCNN for object detection on visual input which is used to extract clusters (c_i). From this visual input, properties such as color and shape can be extracted by annotation models. Using the detected clusters, haptic properties can be gathered through manipulation with a robotic end-effector equipped with haptic sensors.

and better abstraction of organizational concepts.

Our contributions include:

1) Analyzed the role of multimodal inputs in spatial relational learning
2) Modeled user's preferences in the context of spatial relational learning using a desk organization task
3) Added utility modality to model individual differences in object preferences

II. Related Work

A. Multimodal Spatial Reasoning

Spatial reasoning is a fundamental skill that supports robotic understanding of human intent [8] and execution of daily tasks [9]. There has been extensive work done in understanding spatial reasoning for robotics [10]–[12]. A typical organizational directive, such as "place the eraser to the left of the keyboard" not only relies on the robot's ability to properly categorize objects ("eraser" and "keyboard") but also on its learned spatial knowledge associated with prepositions ("left of") [13]. Observing how humans define spatial relations in tasks allows spatial relation learning from humans, which has now attracted attention in the fields of computer vision and audition [14]–[17] as well as robotics [18]–[20]. Maintaining spatial concepts between objects and places [20] as a form of knowledge enhances the robustness in navigating a human-living environment [19] by self-localization of the objects in the domain (e.g. water bottle, box, mouse, etc.).

Our focus is to analyze how multimodal inputs specifically facilitate spatial reasoning for robotics. When a robot physically interacts with an environment, there is an opportunity to collect data from multiple modalities such as vision, haptics, textures, gestures, language, and audio [18], [21]. The mutual enhancement between multiple modalities inspires our multimodal learning [22]. Multimodal learning in robotics does not only provide more information to spatial learning [23], but also explores the association between different perceptions [18]. Spatial relational learning falls under the multimodal task of translation, namely, using vision, haptic, and/or audio data to ground the natural language which describes the spatial relations.

B. Relational Models

Markov Logic Networks (MLN), which are probabilistic graphical models, are a common method to represent relational information about the objects in the world [5], in this case, spatial relations (e.g., the coffee cup is behind the monitor) and object attribute relations (e.g., the water bottle is a hard blue cylinder). In addition to MLNs, others have tried neural networks, add-or graphs, and support vector machines to varying degrees of success [14], [18], [24]. There are some similar methods which use attention augmented networks to learn spatial relations [25], [26]. MLNs combine the ability of a Bayesian network to encode arbitrary probability distributions with the power of first-order logic [5]. We use the MLN, as such models require less data to train, allow for the addition of hand-crafted features (e.g., for spatial relations), and offer more interpretability than others [24]. Similar to [27], we use the MLN to encode cross-modal relationships essentially serving as the fusion step in multimodal learning. First order formulae in our MLNs are used to encode spatial-modality relationships between object attributes and spatial relationships while the graph structure allows for cross-modal interaction. In our case, MLN groundings are attribute predicates and domains (such as color, shape, etc.) along with an object and attribute value (blue, rectangle, etc.). Each predicate can, as in [28], have an associated annotator which generates groundings in the associated domain. These annotators are separate, independent models which are given detected clusters and produce annotations in their expert domain, such as color, shape, etc. Figure 2 shows how a MaskRCNN can be used in conjunction with visual annotators, which operate on image input, as well haptic annotators which require visual input to detect and classify rigidity and weight by robotic manipulation.

We also use an ensemble learning algorithm, namely

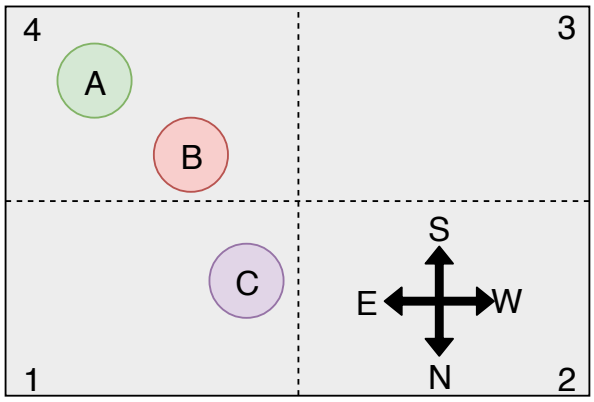


Fig. 3: An example scene with 3 objects, A, B, and C, showing their positions relative to a desk (which quadrant the object is in) and relative to other objects (by cardinal direction).

random forest, to compare with the MLN [6]. We vectorized the attribute predicates and domains. We were motivated to try these as an alternative to the more interpretable MLNs. By turning spatial reasoning into a classification problem with discrete spatial relations, random forests can be used for spatial reasoning. They have been used in various fields for multi-modal learning. Some examples include classification of Alzheimer's disease [29], automatic job-candidate screening based on video CV's [30], and news article classification [31]. The features used depend on the problem at hand: [29] uses, amongst others, MRI volumes and voxel-based FDG-PET signal intensities. On the other hand, the job-candidate used videos [30], while the news article one used n-gram textual features and a representative image [31]. Even so in the case of missing/incomplete data. So, random forests seem to be a pragmatic model to use for our problem.

III. A Tale of Two Models for Desk Organization

We define the desk organization problem as spatial inference according to the properties of objects [32], [33]. For a given desk organization "scene", we assume that there are K objects $\mathbb{O} = \{o_1, \cdots, o_K\}$. We also assume that we can extract features $F(o_i)$ of an object o_i using multiple modalities. These features are used to infer the spatial information $\mathbb{R}$ about how the object should be organized with other objects. We model the human preference as $P(\mathbb{R} \mid F(o_1), \cdots, F(o_K))$, which then defines the desk organization problem as

$$\arg\max_{R} P(\mathbb{R} \mid F(o_1), \cdots, F(o_K)). \tag{1}$$

so that a robot can organize the objects based on inferred spatial information.

In this paper, we focus on two forms of spatial relations. Objects can be located relative to the desk and relative to each other. Spatial relations between objects are encoded as cardinal directions between unique object pairs. For example, in Figure 3, object A is southeast of object B (and correspondingly object B is northwest of object A). Spatial relations between an object and the desk are described by the quadrant $\in \{1, 2, 3, 4\}$ in which the object resides. In Figure 3, object A is in quadrant 4. This quadrant-based spatial relation serves two purposes. First, it models how people position objects on desks as very few (if any) are placed in the center of a desk (where the quadrants would intersect) and most objects are placed in the corners of the desk. Second, it simplifies the relational space that the model must learn as the relative relations are known between any two objects in different quadrants (e.g., any object in quadrant 1 is necessarily north of another in quadrant 4). In addition to cardinal directions between objects, we allow for smaller objects to be placed "in" larger objects. Thus, we define two types of spatial information for the desk organization problem, which are

- $R^{Quad}(o_i) \in \{1, \cdots, 4\}$ tells which quadrant object o_i is in.
- $R^{Rel}(o_i, o_j) \in \{E, NE, N, NW, W, SW, S, SE, IN, NONE\}$ tells the spatial relation of object o_i relative to o_j.

Without any loss of generality, we impose conditional independence assumption to decompose the problem. We can solve equation (1) by factorizing as

$$\begin{aligned}\arg\max_{\{R^{Quad}, R^{Rel}\}} &\prod_{i=1}^{K} P\left(R^{Quad}(o_i) \mid F(o_1), \cdots, F(o_K)\right) \\ &\prod_{i=1}^{K} \prod_{j=i+1}^{K} P\left(R^{Rel}(o_i, o_j) \mid F(o_1), \cdots, F(o_K)\right)\end{aligned} \tag{2}$$

so that we can solve each factor independently. Using equation (2), we can have two inference problems for quadrants and relations. The conditional independence allows us to solve equation (2) by:

$$\begin{aligned}&\hat{R^{Quad}}(o_i) = \arg\max_{R^{Quad}} P\left(R^{Quad}(o_i) \mid F(o_1), \cdots, F(o_K)\right) \\ &\hat{R^{Rel}}(o_i, o_j) \\ &= \arg\max_{R^{Rel}} P\left(R^{Rel}(o_i, o_j) \mid F(o_1), \cdots, F(o_K)\right)\end{aligned} \tag{3}$$

Figure 4 illustrates the two models that represent equation (2). For an object o_i, we evaluate which quadrant it shall be in, $R^{Quad}(o_i)$, by the features of the object $F(o_i)$ and in the context of the features defined over all other objects $\{F(o_k) \mid k \in \{1, \cdots, K\} \setminus \{i\}\}$. Similarly, we evaluate the spatial relation of o_i to reference object o_j where $i \neq j$ by features of both objects, $F(o_i)$ and $F(o_j)$, and in the context of the features defined over all other objects $\{F(o_k) \mid k \in \{1, \cdots, K\} \setminus \{i, j\}\}$.

We analyze our models' ability to represent a human's preference in this desk organization problem in the following aspects:

- *Accuracy:* How many spatial relations can the model correctly predict?
- *Generalization:* Can this model generalize to different people's organizational preferences?
- *Interpretability:* Can we determine from the model what rules a human uses when organizing objects?
- *Satisfaction:* Is the person satisfied with the desk organization produced by the model?

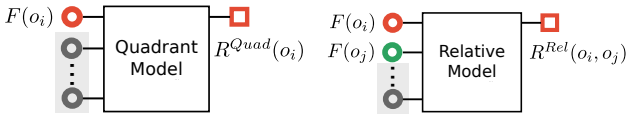


Fig. 4: Models for spatial position and relation inference.

In order to model the prediction defined in equation (3), we choose two canonical models, Random Forest [34] for classification precision and Markov Logic Network [5] for interpretability.

A. Modeling using Markov Logic Networks

The graph nature of MLN allows it to directly combine both quadrant and relative spatial relations in one model and allows for the querying of each type of relation independently. In order for the MLN to learn, it must first enumerate all predicate groundings using the specified domains. In order to limit the size of this space, we construct attribute domains COLOR and SHAPE to use a handful of unique values while SIZE, WEIGHT, and RIGIDITY use a binary classification. In the MLN, we define the following domains:

QUAD	=	{1,2,3,4}
DIR	=	{E, NE, N, NW, W, SW, S, SE, IN, NONE}
COLOR	=	{RED, BLUE, BLACK, GREEN, YELLOW, OTHER}
SHAPE	=	{RECTANGLE, CYLINDER, CUBE, OTHER}
SIZE	=	{SMALL, LARGE}
WEIGHT	=	{LIGHT, HEAVY}
RIGIDITY	=	{SOFT, HARD}
UTILITY	=	{1,2,3,4,5,6,7}

Initially, we provide the MLN with an unlearned set of formula expressed in first-order logic in terms of the predicates and domains above. We include formulae relating COLOR, SHAPE, SIZE, WEIGHT, RIGIDITY, and UTILITY to DIR and QUAD. During training, these formulae are expanded to include all possible groundings, after which, during the training process each formula receives a weight.

B. Modeling using Random Forests

Although less interpretable, random forests have more powerful representational abilities. As described earlier, we use two independent random forest models to capture all spatial relations. One model captured the quadrants for each object (RF_{quad}), and the other model captured the relative spatial relations between objects (RF_{rel}). The architecture of the 2 models is the same. The models use Gini impurity [35] as their criterion for splitting nodes. We keep the trees of the forests fully grown and unpruned, as our dataset was not so big as to put a cap on the memory consumption of the model. We define 20 estimators (20 decision trees comprising the forest) in each random forest.

We decided to split the spatial reasoning task into 2 models, which follow equation (3); they deal with vectorized representations of features and spatial relations. Quadrant relations, and relative spatial relations operate in 2 separate domains, or classes, namely QUAD and DIR as defined in § III-A. So, following the domain design in MLN, we would need to enable multi-class classification. We describe in more detail how the two forests worked together to perform scene generation as well as how the accuracies of the two forests were computed.

For simplicity, the MLN predicate syntax as described in Section III-A was also used to describe scenes and objects with the Random Forest. For each object, the domain groundings (which act as features for the classifier) are converted into one-hot vectors. The vectors are concatenated to product input vectors for the random forest. In our usage, we restrict the set of desk scenes where $K = 7$. To generate scenes, RF_{quad} is used to assign a quadrant to each of the 7 objects and then RF_{rel} assigns cardinal relational directions between every pair of objects in the same quadrant.

IV. Experiments and Results

A. Experimental Setup

For experimentation, we selected 17 objects that fit in the domains described in Section III-A. These include typical desk objects such as a mouse, a box of paperclips, and a cellphone as well as more unusual objects such as an eraser cube, empty soda cans, and a Rubik's cube. These objects were chosen such that they shared properties in some modalities but differed in others. For example, the Rubik's cube and eraser cube both have the SHAPE of CUBE and the COLOR of OTHER, but differ in their RIGIDITY, where the Rubik's cube is HARD and the eraser is SOFT. Some objects, such as the phone, only have one set of attributes while others, such as the dry erase marker, can have different attributes depending on which marker is used (red, blue, etc.). With our object set defined, we experimented with *synthetic data*, *human data* from multiple studies, and *real-world robot demonstrations* in order to determine the representational power of our models.

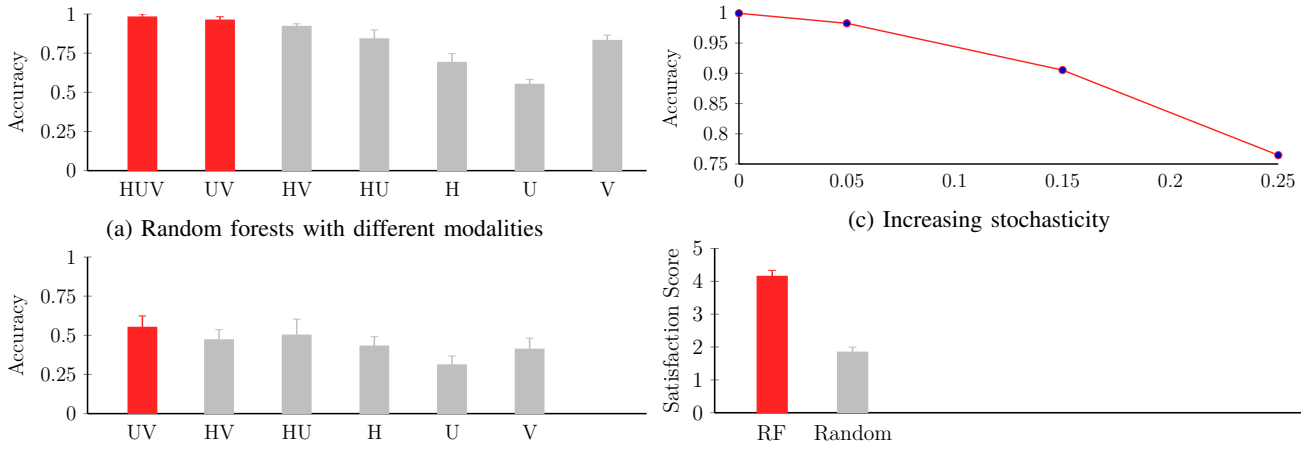
B. Synthetic data

1) Experimental Procedure: We generated the synthetic data using a scene generator by picking objects at random with replacement from the set of all objects and positioning them on a desk programmatically according to a predefined list of constraints.

In collecting synthetic data, we generated 30 scenes. These scenes consisted of 6 to 9 objects chosen uniformly with replacement from the set of all objects. A quadrant annotator then produced a list of quadrant predicate groundings (both QUAD and DIR), given the set of input objects, which described the position of every object. The groundings were chosen based on the UTILITY, COLOR, and SHAPE of the provided objects.

2) Results: Using 5-fold cross-validation, MLNs were trained over the simulation set and used to predict object relations. We measured accuracy in terms of number of relational groundings (e.g., DIR(o_0, o_5, N)) which were correctly predicted. Note that this relation means o_0 is to the North of o_5. Over the simulation set, MLN achieved 99% accuracy, thus showing organizational spatial relations can be learned by one of our models.

To simulate uncertainty in annotation, "stochasticity" was added to the simulation. Several sets of true positional



(a) Random forests with different modalities

(c) Increasing stochasticity

(b) MLNs with different modalities

(d) Participant satisfaction

Fig. 5: (a,b,c): Average fold accuracy during 5-fold cross-validation over 30 synthetic sample scenes. (d): Average participant satisfaction, rated on a 5-point Likert scale on follow up survey images organized by both trained random forest and a uniformly random quadrant model.

predicate groundings were generated with stochasticities $p = 0.05$, 0.15, and 0.25. It means that, with probability p each object attribute was modified to a value other than its original (e.g., from COLOR(o_i, RED) to COLOR(o_i, BLUE)). The introduction of this parameter simulates a noisy or untrusted attribute annotator in order to determine (1). Through 5-fold cross-validation with MLN, we found that the model was relatively robust in learning the simulated organizational strategy given small amounts of noise. As seen in Figure 5c, accuracy, hence learnability, decreases drastically with increasing stochasticity.

C. Human data: Initial study

1) Experimental Procedure: For the initial study, we collected the human data during a human-participant study with 11 participants. The participants designed 30 scenes (similar to the experiments with synthetic data) by picking 7 objects with replacement. For each of the 30 scenes, we instructed each participant to organize the 7 scene objects on a desk divided into 4 quadrants only with the instruction being that objects may not span across quadrants and may not be on top of one another.

We also asked the study participants to subjectively characterize each unique object's WEIGHT and SIZE in {LIGHT, HEAVY} and {SMALL, LARGE} categories respectively. These results were averaged over all participants to determine the canonical WEIGHT and SIZE of each object. Each participant was also asked to rate the utility of each object on a 7-point Likert scale. These results were not averaged together; instead, when organizing a desk for a particular participant, each object was assigned the utility that the participant in question responded with. RIGIDITY was determined by measuring the stiffness of each object with a spring scale and choosing a threshold such that half of the objects were SOFT and half were HARD. Each object was measured on the surface where a human would normally grasp it.

We trained two models over the same human dataset: 330 scenes, each with 7 objects, from 11 surveys with 5-fold cross-validation. In order to account for differences in an individual's organizational preferences, the dataset was partitioned by participant so models predicting for participant n had only been trained on scenes from participant n. As described in the beginning of Section IV, each participant's UTILITY ratings for each object were used during training. In order to generate the true spatial relations for these scenes, during the study photos were taken of each scene from an overhead camera after organization. A positional annotator used hand-annotated masks for each object to determine which quadrant each was in, the pairwise cardinal relations for each object, and the IN relation if two masks sufficiently overlapped.

2) Results: Models were trained for each unique combination of the available modalities, resulting in 7 models: HUV, HV, UV, HU, H, U, V. Accuracy during cross-validation was measured for the MLN as described in Section IV-B. For the random forest, the RF_{quad} and RF_{rel} cross-validation accuracies were averaged together with weights K and $\binom{K}{2}$ respectively where K is the number of objects in the scene. This was done to make the results comparable with those of the MLN, as K and $\binom{K}{2}$ represent the ratio of QUAD and DIR formula respectively.

As seen in Figure 5a, the Random Forest is able to achieve very high accuracy in cross-validation with nearly 95% in average when using haptics, vision, and utility. With this decomposition, we can also observe that vision alone achieves 83% accuracy in average, followed then by haptics and utility. Here we see the benefit of multimodal learning, as even with the introduction of one additional modality, haptics, to vision, we see the accuracy increase from 83% to 92% in average. The addition of our new modality, utility, increased accuracy further to 98%. Both of these are statistically significant increases, with p-values less than 0.05, calculated via t-test. The t-statistic for the first is 16.945 (79.30 degrees of freedom), and 27.399 (85.89 degrees of freedom) for the second. When using both haptics and vision, the introduction of utility lead to a 6% increase in accuracy. This, along with the 55% accuracy when using

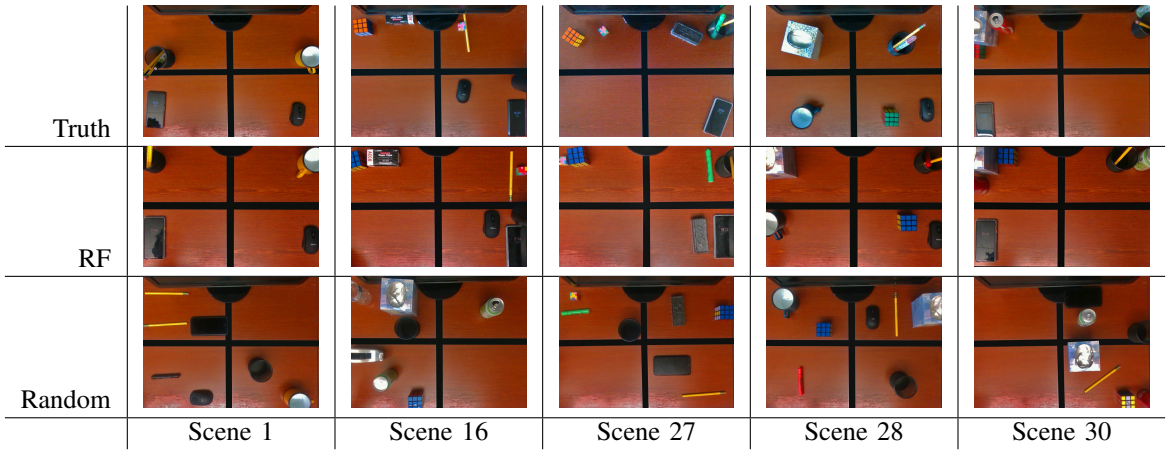


Fig. 6: Five scenes from five different study participants used in the follow-up survey. Each scene was organized by the participant (Truth), a random forest model (RF), and a uniformly random positioner (Random).

utility alone, indicates that humans often take an object's usefulness into consideration when organizing items on a desk.

Although the MLN does surpass the RF in interpretability, it's representational abilities are far inferior. As explained in Section III-A, training the MLN requires expanding the provided formula with all combinations of domain groundings. As a result, the number of QUAD formula is linear with domain size and number of modalities while the number of DIR formula is multiplicative. Due to this fact, the parameter space increases exponentially and convergence takes a long time (much more than the random forest). So, we omit this combination. Similarly, 10% of non-HUV MLNs also failed to converge due to overflow and are therefore excluded from the data in Figure 5b. From this figure, we can again see that haptics and vision are among the most informative modalities, followed by utility. However, the overall accuracy of the MLN model is much lower, with the highest accuracy being 71% when using HV on participant 2's desk organizations. Despite this, we again see that the interaction of multiple modalities yields higher overall predictive performance. Comparing utility with haptics plus utility, we see that the difference is statistically significant at $p < 0.05$ (t-statistic is 11.924 with 84.842 degrees of freedom). Note that these values are still much better than randomly guessing the positions of each of the objects. For example, in any one scene, there can be 7 unique objects. Each of these must be assigned 1 of 4 quadrants uniformly at random: the probability of being exactly correct is $1/4^7 = 0.00006$.

3) Insights: Interpretability of MLNs: Because MLN is programmed from first-order logics, the trained weight of the first-order formulae supports its interpretability. We can examine the weighted formula in the trained models to gain insights into one's organizational preferences. For example, with the highest weight of 16161, the formula:

$$\text{UTILITY}(o_1, 5) \wedge \text{UTILITY}(o_2, 3) \wedge \text{DIR}(o_1, o_2, \text{NW})$$

informs us that utility is taken into account when positioning objects and that this participant prefers more useful objects to be positioned in front and to the right of less useful objects. For another participant, we see with weight 10130, that:

$$\text{COLOR}(o_1, \text{BLUE}) \wedge \text{COLOR}(o_2, \text{OTHER}) \wedge \text{DIR}(o_1, o_2, \text{SE})$$

This again gives us insight into the organizational preferences of this survey participant. Mainly that they tend to position BLUE objects behind and to the left of OTHER objects.

D. Human data: Follow-up Study

1) Experimental Procedure: After the initial user study, which provided us with *human data* to train the MLN and RF models on, we produced a numerical measure, namely accuracy, of the model's performance. However, human organizational habits are very subjective and multiple different arrangements of objects on a desk may be considered "organized" by different people or even the same person. To deal with this issue, we designed a follow-up survey to see how well our accuracy metrics matched "human satisfaction". This was largely motivated by our overarching goal: to build robots which can effectively operate in a three-dimensional world and interact with humans.

We sent follow-up surveys to each of the 11 participants of the original study. For each survey, we randomly chose 5 of the 30 scenes from the original study. For each of the 5 scenes, we included in the survey 3 images of that scene:

1) the scene as organized by the participant during the study.
2) the scene as organized randomly.
3) the scene as organized according to a Random Forest model's predictions.

The original scenes are used as a reminder of what the study involved and how they originally organized each scene. They serves as a "calibration", so that their organizational schemes don't drift too far from the original study. For the random guesser, the 7 objects from the original scene were positioned uniformly randomly along the x and y axis of the desk plane. The randomly chosen location then inherently yielded the quadrant and relative spatial positions of each object. The purpose of these images is to provide a

comparison to the human organizational scheme from before and the random forest organizational scheme. The scenes as organized by the random forest were included to measure our model's ability to learn a participant's organizational scheme. However, the random forest models (namely RF_{quad} and RF_{rel}) used to generate this scene version were trained on scene data from *all* participants. That is to say, the models for the follow-up survey were trained on $11 \times (30-5) = 275$ scenes, and then the scene generation was performed on 5 scenes per survey.

In this capacity, the models are "general", as opposed to the "personal" models used in Section IV. Generalized models were chosen as opposed to tailored models to determine if, in addition to human satisfaction, there were cross-person patterns in how different people organized their desks and if the models could pick up on these patterns. This was directly motivated by real-life constraints of deploying agents with good enough *priors* to perform personalized tasks, such as desk organization, out of the box before they have a chance to tailor their internal models to a particular person.

2) Results: Generalization and Satisfaction: In Figure 6, we show 5 of the 30 scenes organized in three different ways for 5 of the participants. The random forest was able to successfully learn patterns in organizational choices that humans make. For example, in scene 28 it learned to place the mouse in the 4th quadrant and east of the Rubik's cube. It also successfully positioned the pencil and dry erase marker inside of the pen cup in scene 1.

Although the RF models used for generation of these organized scenes were generalized, and therefore different from the models that achieved high cross-validation accuracy in Section IV-C, study participants rated the random forest organizations quite highly in terms of satisfaction. As seen in Figure 5d, participants were more satisfied with the scenes as organized by the random forest compared to those organized by the random guesser. The difference is statistically significant at $p < 0.05$ (t-statistic 68.784 with 105.726 degrees of freedom).

E. Robot Demonstration

We used HERB 3.0 [36] to organize 5 soda cans in simulation (visualized using RViz), and in real. See Figure 7. HERB is the Home Exploring Robot Butler; it is the robotic platform used for testing our models. HERB 3.0 has a mobile base and 2 Barrett 7−DOF WAM arms with Barrett hands; only arms and hands were used in manipulating objects in our experiment. It is also equipped with multiple laser rangefinders and cameras placed in various configurations (e.g. base, neck, etc.) to allow versatile perception capabilities; these were not used. Seeing as how our major goals were to analyze the spatial reasoning capabilities of the two models, and the advantage multimodal learning yields to it, We kept most of our work in "simulation" using cylindrical objects as a proof of concept. We also demonstrated the performance using an actual robot, HERB, performing a desk organization task, as a way of more tangibly demonstrating the above results.

This trial demonstrates that with the proper modeling of spatial relations and using motion planning algorithms,

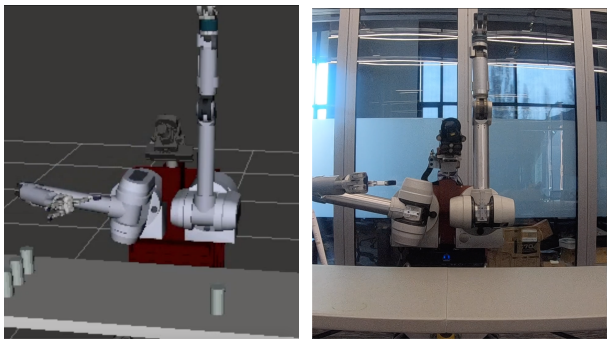


Fig. 7: Robot trials in simulation and real-life. Video in [37]

robots performing such tasks as desk organization are feasible in home environments. As learned models mentioned above produce positions of objects according to spatial relations learned from human demonstrations, the focus of robot integration is moving objects from any initial start position to a goal position provided by our learned model. Thus, we designed our trial consisting of a start and goal position for each of the 5 soda cans used in the experiments. The start positions were randomly selected on the right side of the table. The goal positions are extracted from running the random forest model. HERB was positioned near the table and used its right arm to successfully move all the soda cans to their goal positions [37].

V. Discussion

Random forests performed well but the weights of the model are difficult to interpret. Markov Logic Networks are mainly for performance and interpretability, as they perform well in simulation, and in the survey to some extent. It is easy to understand what an MLN learned as it simply consists of weighted first order logic formulae, which we designed to be very indicative of what features are being used to determine spatial relations.

The models do have their limitations as well. One limitation of MLNs is intractability. If an MLN has too many formulae, inference becomes intractable, or takes prohibitively long. One limitation of the random forest model is that it requires fixed input dimensionality, meaning one trained on 8-object scenes cannot be used for inference on 9-object scenes. A realistic, deployable model would need to be able to handle scenes with any number of objects. Another limitation we faced with Random Forests was the presence of conflicts, that is, sometimes the random forests would suggest in one scene two conflicting relations. For example, say RF_{quad} assigns QUAD(o_1, 3) and QUAD(o_2, 3) for two objects. RF_{rel} could, in some instances, predict DIR(o_1, o_2, N) and DIR(o_2, o_1, N). This occurred because the RF_{rel} does not take into account previous predictions in the scene; it predicts each relation independent of all other ones in the scene. However, we corrected for this by ignoring the conflicts in a manner consistent with how the program would have done it; once one relation is predicted, all other conflicting relations are null and void. Note that, in order to measure the capabilities of these models for capturing spatial relations, we "turned off" the noise, by using pristine, hand-labeled annotations from the surveys as opposed to running

the MaskRCNN on them. In this, we make an assumption that the MaskRCNN can accurately detect each object and attribute annotators produce perfect annotations.

In the future, we plan to perform realistic robotic manipulation tasks in real homes. This would involve the creation of visual attribute annotators for color, shape, and size while haptic annotators would require use of a robotic arm and haptic sensor to pick up each object, as detected by MaskRCNN, to determine its rigidity and weight. This would involve some pipeline, similar to 2, which could also potentially involve natural language input to supplement or replace missing or inaccurate property annotations.

REFERENCES

[1] P. Fiorini and E. Prassler, "Cleaning and household robots: A technology survey," *Autonomous robots*, vol. 9, no. 3, pp. 227–235, 2000.

[2] R. Dillmann, "Teaching and learning of robot tasks via observation of human performance," *Robotics and Autonomous Systems*, vol. 47, no. 2-3, pp. 109–116, 2004.

[3] M. Beetz, F. Stulp, B. Radig, J. Bandouch, N. Blodow, M. Dolha, A. Fedrizzi, D. Jain, U. Klank, I. Kresse, *et al.*, "The assistive kitchena demonstration scenario for cognitive technical systems," in *Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE International Symposium on*. IEEE, 2008, pp. 1–8.

[4] Y. S. Choi, T. Deyle, T. Chen, J. D. Glass, and C. C. Kemp, "A list of household objects for robotic retrieval prioritized by people with als," in *Rehabilitation Robotics, 2009. ICORR 2009. IEEE International Conference on*. IEEE, 2009, pp. 510–517.

[5] M. Richardson and P. Domingos, "Markov logic networks," *Machine learning*, vol. 62, no. 1-2, pp. 107–136, 2006.

[6] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[7] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2980–2988.

[8] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Transactions on Circuits and Systems for Video technology*, vol. 18, no. 11, p. 1473, 2008.

[9] P.-C. Chung and C.-D. Liu, "A daily behavior enabled hidden markov model for human behavior understanding," *Pattern Recognition*, vol. 41, no. 5, pp. 1572–1580, 2008.

[10] M. Mansouri and F. Pecora, "A representation for spatial reasoning in robotic planning," 2013.

[11] C. Landsiedel, V. Rieser, M. Walter, and D. Wollherr, "A review of spatial reasoning and interaction for real-world robotics," *Advanced Robotics*, vol. 31, no. 5, pp. 222–242, 2017. [Online]. Available: https://doi.org/10.1080/01691864.2016.1277554

[12] W. G. Kennedy, M. D. Bugajska, M. Marge, W. Adams, B. R. Fransen, D. Perzanowski, A. C. Schultz, and J. G. Trafton, "Spatial representation and reasoning for human-robot collaboration," in *AAAI*, 2007.

[13] R. Paul, J. Arkin, N. Roy, and T. M Howard, "Efficient grounding of abstract spatial concepts for natural language interaction with robot manipulators," 2016.

[14] S. Li, G. Kulkarni, T. Berg, A. Berg, and Y. Choi, "Composing simple image descriptions using web-scale n-grams," in *CoNLL 2011 - Fifteenth Conference on Computational Natural Language Learning, Proceedings of the Conference*, 2011, pp. 220–228.

[15] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S.-C. Zhu, "I2t: Image parsing to text description," *Proceedings of the IEEE*, vol. 98, no. 8, pp. 1485–1508, 2010.

[16] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and VQA," *CoRR*, vol. abs/1707.07998, 2017. [Online]. Available: http://arxiv.org/abs/1707.07998

[17] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek, "Areas of attention for image captioning," *CoRR*, vol. abs/1612.01033, 2016. [Online]. Available: http://arxiv.org/abs/1612.01033

[18] J. Thomason, J. Sinapov, M. Svetlik, P. Stone, and R. J. Mooney, "Learning multi-modal grounded linguistic semantics by playing "i spy"," in *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI-16)*, New York City, 2016, pp. 3477–3483. [Online]. Available: http://www.cs.utexas.edu/users/ai-lab/pub-view.php?PubID=127564

[19] A. Taniguchi, T. Taniguchi, and T. Inamura, "Simultaneous estimation of self-position and word from noisy utterances and sensory information," *IFAC-PapersOnLine*, vol. 49, no. 19, pp. 221 – 226, 2016, 13th IFAC Symposium on Analysis, Design, and Evaluation ofHuman-Machine Systems HMS 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2405896316321188

[20] S. Isobe, A. Taniguchi, Y. Hagiwara, and T. Taniguchi, "Learning relationships between objects and places by multimodal spatial concept with bag of objects," in *International Conference on Social Robotics*. Springer, 2017, pp. 115–125.

[21] M. Turk, "Multimodal interaction: A review," *Pattern Recognition Letters*, vol. 36, pp. 189–195, 2014.

[22] T. Nakamura, T. Araki, T. Nagai, and N. Iwahashi, "Grounding of word meanings in latent dirichlet allocation-based multimodal concepts," *Advanced Robotics*, vol. 25, no. 17, pp. 2189–2206, 2011. [Online]. Available: https://doi.org/10.1163/016918611X595035

[23] E. A. Sisbot, L. F. Marin, and R. Alami, "Spatial reasoning for human robot interaction," in *Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on*. IEEE, 2007, pp. 2281–2287.

[24] T. Baltrusaitis, C. Ahuja, and L. Morency, "Multimodal machine learning: A survey and taxonomy," vol. abs/1705.09406, 2017. [Online]. Available: http://arxiv.org/abs/1705.09406

[25] T. Zhang, Y.-T. Li, and J. P. Wachs, "The effect of embodied interaction in visual-spatial navigation," *ACM Trans. Interact. Intell. Syst.*, vol. 7, no. 1, pp. 3:1–3:36, Dec. 2016. [Online]. Available: http://doi.acm.org/10.1145/2953887

[26] D. Kontogiorgos, "Multimodal language grounding for improved human-robot collaboration: Exploring spatial semantic representations in the shared space of attention," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ser. ICMI 2017. New York, NY, USA: ACM, 2017, pp. 660–664. [Online]. Available: http://doi.acm.org/10.1145/3136755.3137038

[27] D. Skočaj, A. Vrečko, M. Mahnič, M. Janíček, G.-J. M. Kruijff, M. Hanheide, N. Hawes, J. L. Wyatt, T. Keller, K. Zhou, *et al.*, "An integrated system for interactive continuous learning of categorical knowledge," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 28, no. 5, pp. 823–848, 2016.

[28] D. Nyga, F. Balint-Benczedi, and M. Beetz, "Pr2 looking at things - ensemble learning for unstructured information processing with markov logic networks," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, May 2014, pp. 3916–3923.

[29] K. Gray, P. Aljabar, R. Heckemann, A. Hammers, and D. Rueckert, "Random forest-based similarity measures for multi-modal classification of alzheimer's disease," *NeuroImage*, vol. 65, 10 2012.

[30] H. Kaya, F. Grpinar, and A. A. Salah, "Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video cvs," pp. 1651–1659, July 2017.

[31] D. Liparas, Y. HaCohen-Kerner, A. Moumtzidou, S. Vrochidis, and I. Kompatsiaris, "News articles classification using random forests and weighted multimodal features," pp. 63–75, 2014.

[32] T. W. Malone, "How do people organize their desks?: Implications for the design of office information systems," *ACM Transactions on Information Systems (TOIS)*, vol. 1, no. 1, pp. 99–112, 1983.

[33] C. Pantofaru, L. Takayama, T. Foote, and B. Soto, "Exploring the role of robots in home organization," in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. ACM, 2012, pp. 327–334.

[34] A. Liaw, M. Wiener, *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[36] S. Srinivasa, D. Berenson, M. Cakmak, A. C. Romea, M. Dogar, A. Dragan, R. A. Knepper, T. D. Niemueller, K. Strabala, J. M. Vandeweghe, and J. Ziegler, "Herb 2.0: Lessons learned from developing a mobile manipulator for the home," *Proceedings of the IEEE*, vol. 100, no. 8, July 2012.

[37] "Robot demonstration video," https://sites.google.com/cs.washington.edu/mintdeskorg/home,[Online; Retrieved on 10th May, 2019].