Scene-Dependent Intention Recognition for Task Communication with Reduced Human-Robot Interaction

Kester Duncan[1], Sudeep Sarkar[1], Redwan Alqasemi[2], and Rajiv Dubey[2]

[1] Computer Sci. & Eng. Dept., Univ. of South Florida, Tampa FL, U.S.A.*
[2] Mechanical Eng. Dept., Univ. of South Florida, Tampa FL, U.S.A.†

Abstract. In order for assistive robots to collaborate effectively with humans, they must be endowed with the ability to perceive scenes and more importantly, recognize human intentions. These intentions are often inferred from observed physical actions and direct communication from fully-functional individuals. For individuals with reduced capabilities, it may be difficult or impossible to perform physical actions or easily communicate. Therefore, their intentions must be inferred differently. To this end, we propose an intention recognition framework that is appropriate for persons with limited physical capabilities. This framework determines and learns human intentions based on scene objects, the actions that can be performed on them, and past interaction history. It is based on a Markov model formulation entitled Object-Action Intention Networks, which constitute the crux of a computer vision-based human-robot collaborative system that reduces the necessary interactions for communicating tasks to a robot. Evaluations were conducted on multiple scenes comprised of multiple possible object categories and actions. We achieve approximately 81% reduction in interactions overall after learning, when compared to other intention recognition approaches.

Keywords: intention recognition, human robot interaction, intelligent robots, robot vision systems

1 Introduction

Assistive robotic technologies can provide the elderly, persons with disabilities, and injured veterans with opportunities to achieve higher levels of independence and quality of life as they reduce dependence on caregivers and increase self-sufficiency. However, there are many challenges to developing such robotic technologies. Seeing a bottle, reaching for it, and picking it up to pour its contents are relatively easy tasks for a fully-functional human; however, this task is difficult for both individuals with reduced capabilities and robots. For a robot to complete such tasks, it must possess the perceptive ability to effectively process

* [kkduncan, sarkar]@cse.usf.edu
† [alqasemi, dubey]@usf.edu

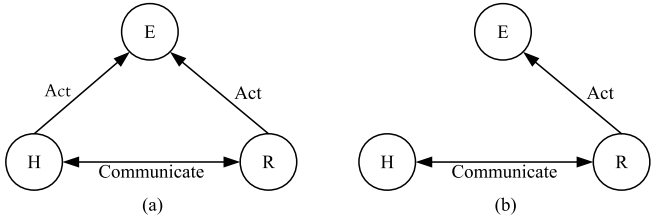


Fig. 1. Configuration of human-robot interactive systems using intention recognition. Circles with 'H' represent humans, 'R' represents robots, and 'E' represents the environment. For approaches grouped under (a), the human and robot both perform tasks on the environment whereas for (b), the environment is acted on only by the robot.

the individual's environment, namely the scene. This involves addressing key computer vision problems such as object segmentation, object categorization, and object pose estimation, along with manifold mechanical tasks. These issues have been considerably addressed in the computer vision and robotics literatures [13], [15], [4], [3], [16], [7] and in this work we leverage on recent advances to present a computer vision-based human robot collaborative system.

Another challenge exists when there is a lack of full communication. It is relatively easy for a fully-functional individual to directly express their intent, yet for persons with reduced capabilities, this may prove to be quite difficult or impossible depending on their communication ability. As a result, the problem of recognizing one's intention is brought to the forefront [19] as it is requisite for successful communication and collaboration [20]. These intentions are inferred from actions carried out or from changes in the environment [8], [19], [2]. In order to build robots that are competent assistants, we must endow them with this intelligent ability in order to understand the action that is to be performed on the environment, e.g. pick up the red cup vs. throw away the brown box. In so doing, we can reduce the need for direct human-robot interaction and thereby maximize robotic task performance.

As an example, consider a robot that helps a person incapable of communicating physically, choose a task to perform at a breakfast table. The robot is capable of scanning the table, and for every item that is found, a group of possible tasks that can be performed with it is recorded in a list. Subsequently, the robot asks the individual to indicate the task they want to perform. Suppose there are n possible tasks on the list. For the extreme case where the user desires the n^{th} task, the robot would have had to prompt the user n times in order to know what to do. With intention recognition, it is possible to reduce the length of this list by only considering items the individual would most likely want performed and thereby reduce the amount of time necessary for communicating their intent.

In the literature, intention recognition approaches can be classified according to two main configurations hinged on human-robot interactions as depicted in Figure 1:

1. *Intention recognition via observation of the user's physical actions and the environment* (Fig. 1a): the human and the robot may both perform tasks on the environment and there is some form of interaction between them.
2. *Intention recognition via observation of the user's environment* (Fig. 1b): the human interacts with the robot that in turn interacts with the environment.

For the first configuration (Fig. 1 (a)), the human can act directly on the environment. Therefore, both the human and the environment can be observed to infer intentions. Sensors are used to observe the user's physical actions and determine their intentions [10], [9], [22]. For example, Kelley et al. [10] presented an approach that observed an individual using an RGB-D camera and with a neural network-based method, they were able to predict their actions by analyzing their hand positions in relation to objects in the scene. Consequently, the main underlying goal of this category of approaches is to develop effective socially-interactive robotic systems and the target population is usually able-bodied individuals.

Conversely, for the second configuration, as depicted in Fig. 1 (b), the human cannot act directly on the environment; they act on it via the robot. Sensors are used to observe the environment and the robot interacts with the human for determining their intentions [6], [21], [1]. For example, Demeester et al. [6] presented a system that estimated the intent of a user using the sensor readings of their environment and the user's commands so as to take corrective action during wheelchair maneuvering. The system in turn provided assistance that was tailored to the user's driving ability.

Our survey of the state of the art finds that the second category is not fully explored. It is according to this configuration that our proposed work belongs because it is more appropriate for dealing with persons with disabilities wherein it may be quite difficult for them to perform activities on the environment. The main goal of this category of approaches is to develop robotic systems that function effectively in human environments in order to work collaboratively for achieving common goals. In light of this, we present a novel intention recognition framework used within a computer vision-based human-robot collaborative system that allows persons with disabilities to perform tasks with reduced robot interaction.

2 Overview

An activity is primarily a sequence of steps involving objects and actions carried out to accomplish a specific task. As depicted in the example of Figure 2, the activity 'Drink a soda' consists of steps whereby a soda can (*object*) is picked up (*action*) by a robotic arm and poured (*action*) into a cup (*object*) followed by moving the cup to facilitate drinking. In this work, we focus on recognizing these

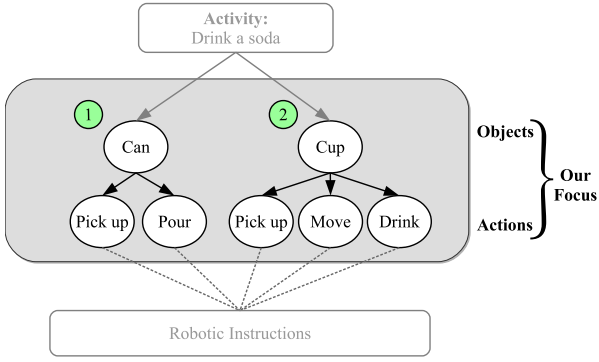


Fig. 2. Scope of our work: we determine the object-action pair that represents a user's intention at a particular step of an activity and attempt to reduce the amount of human-robot interaction necessary to accomplish this step.

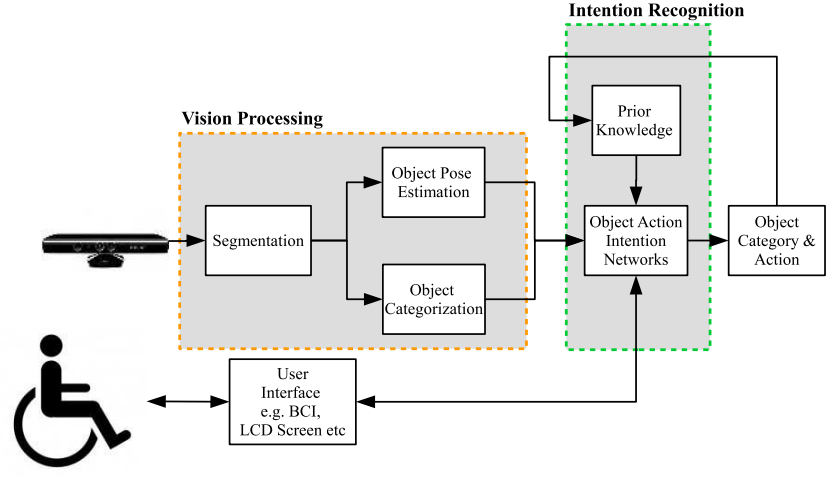


Fig. 3. Schematic of our complete human-robot collaborative system. Object category and pose information along with a user's past decision history are incorporated in our approach for intention recognition.

individual steps from captured visual data with minimum human interaction.

A schematic for our complete human-robot collaborative system is depicted in Figure 3. The system is partitioned into vision processing and intention recognition components. For vision processing, the robotic component first performs object segmentation on the captured RGB-D data to extract objects, which are represented as point cloud clusters. The position and orientation of these objects as well as their category identities are then ascertained via pose estimation and

object categorization respectively. This information is required as input to our intention recognition component.

Our intention recognition problem is therefore cast as determining the object and action pair that the user chooses but has not yet communicated for execution by a robot for manipulation tasks. This is done using our object-action intention networks which are probabilistic graphical models that capture prior object-action knowledge with the capability of adapting to a user's preferences. They are based on: (1) a learned decision history of the human when queried by our system, and (2) an analysis of the scene from the captured RGB-D data such as which object categories are present, object properties (e.g. distance from camera, color), and the relationships between objects and the actions that can be performed with them (see Fig. 3).

3 Vision Processing

Our intention recognition framework is fully dependent on the scene content for determining the user's goal, therefore the scene must be processed accordingly. This is accomplished via object segmentation, object categorization, and object pose estimation.

3.1 Segmentation

Given an RGB-D point cloud of a scene, we first perform planar segmentation using a RANSAC-based approach, then we extract candidate object point clusters from the plane found with the largest candidate object footprint as outlined by Rusu et al. in [18]. These resultant clusters are further processed as described in subsequent sections.

3.2 RGBD Object Categorization

We employ a categorization method based on multiple cues: intensity, 2D contour shape, and 3D shape. This allows us to keep a balance between discrimination and generalization. From a 2D projection of an object point cloud, we extract SIFT [12] and HOG [5] features for appearance and contour shape respectively. The 3D shape properties are obtained by using Fast Point Feature Histograms (FPFH) [17]. For the final object representation, the Bag-of-Words (BoW) model is employed. For classification, we train Support Vector Machines (SVMs) on the BoW features obtained. For cue integration, we adopt the ensemble of classifiers paradigm whereby we train another SVM on the class confidence outputs provided by each feature classifier. The final classification decision is then made by choosing the category with the strongest support. We use the aforementioned methods in this work because they have demonstrated good performance for categorization tasks [14].

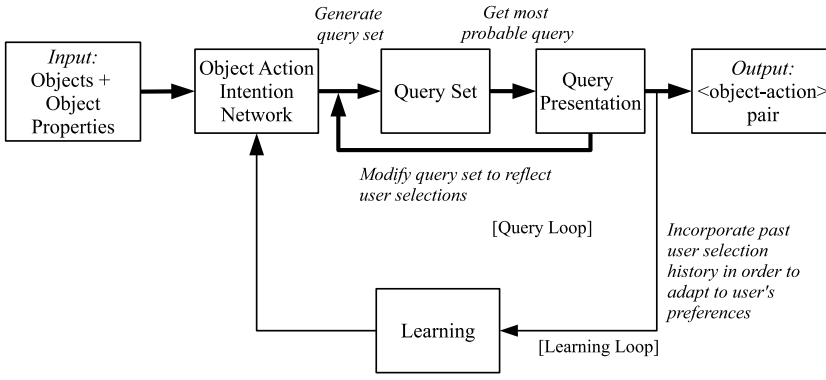


Fig. 4. An overview of our intention recognition approach which uses an object-action intention network, a set of queries for interacting with the user, and a learning process to improve predictions.

3.3 3D Object Pose Estimation using Superquadrics

We estimate the pose (location and orientation) of object point clusters using the low latency method described by Duncan et al. in [7]. By employing superquadrics, which are compact parametric shapes with tri-axis symmetry, this method is able to determine object positions without a model base. It also recovers these superquadrics in a rapid manner by using an effective multi-scale voxelization scheme.

4 Intention Recognition Framework

An overview of our intention recognition framework is shown in Figure 4. From vision processing, object information is used as input for construction of our object-action intention network. Based on the network, a set of queries is generated and proposed to the user. The form of this proposal can vary depending on the type of human-robot interface used e.g. object bounding box for touch screen, verbal questions for speech recognition. For generalization, a query is simply a yes-or-no question. In an ideal situation, the first query proposed to the user coincides with the user's intention. If this does not occur, the query set is modified and another query is presented until the user's intention is communicated as depicted by the query loop in Figure 4. The user's selections are learned via the learning loop in order to adapt to the user's preferences. Notably, the ability of our system to simultaneously infer a user's intention and learn their preferences over time differentiates it from the state of the art. With learning, we are able to improve the intention predictions which in effect reduces the need for many rounds of user interaction. The individual components of this framework are unveiled in the following sections.

4.1 Object-Action Intention Networks

We formulate our intention recognition problem by using probability theory as follows. For each 3D scene that is captured and processed, there are n objects[3]. These objects are represented by the binary random variables $\mathbf{O} = \{O_1, \ldots, O_n\}$ where their values o_i indicate whether the user wants to manipulate the object or not. Associated with each object are binary action variables $\mathbf{A} = \{A_1, \ldots, A_m\}$ whose values a_j indicate whether the user wants to perform the action on the object or not. In addition to these object and action variables are object feature variables $\mathbf{F} = \{F_1, \ldots, F_{cn}\}$, which are also binary random variables representing some intrinsic property of the object e.g distance from camera, color etc. There can be c feature variables per object and they can potentially bias an object for selection by the user, e.g. user's preference for red objects.

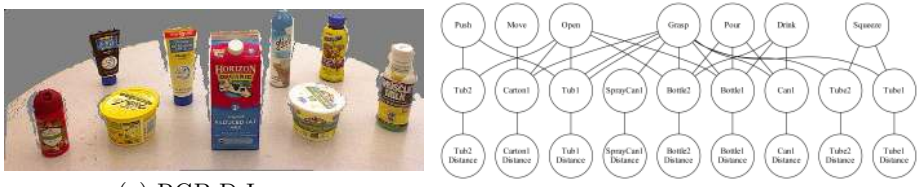
Under this formulation, our task is to infer the most probable object that the user wants to manipulate as well as the most probable action that the user intends to perform on the object. Thus, our goal is to find the highest-probability joint assignment of object and action variables of the form $P(o_i = yes, a_j = yes)$, which represents the intent of the user. With a joint distribution of these variables, we can answer questions about the observed scene ranging from the standard conditional probability query $P(\mathbf{O} = \mathbf{o}, \mathbf{A} = \mathbf{a} \,|\, \mathbf{F} = \mathbf{f})$, to finding the most probable assignment to some subset of variables[4]. We are particularly interested in determining the maximum a posteriori (MAP) probability, whereby the task is to infer the most likely assignment to the variables in $\mathbf{O}$ and $\mathbf{A}$ given the evidence $\mathbf{F} = \mathbf{f}$: $\arg\max_{\mathbf{o},\mathbf{a}} P(\mathbf{o}, \mathbf{a} \,|\, \mathbf{f})$.

We employ Markov Networks [11] as the foundation of our Object-Action Intention Networks so that we can model and encode the relationships between the object and action variables as directional influence between them cannot be naturally ascribed. These networks are undirected graphical models that efficiently capture the joint distribution P over the set of random variables by exploiting existing independence properties that exist between them. The set of object and action variables comprise the corresponding set of object and action nodes in the network. The edge links between an object and action node signifies that the action can be performed on the object and that a direct probabilistic relationship exists between them. The joint distribution is quantitatively parameterized in the network using factors Φ, which are compatibility functions mapping the values of a set of random variables $\mathbf{d}$ to positive real numbers $\mathbb{R}^+$. Using the joint distribution, we can acquire the Maximum A Posteriori probability (MAP) as shown in Equation 1. We believe that the MAP is likely to reflect a user's preferences over time.

$$MAP(\mathbf{O}, \mathbf{A} \,|\, \mathbf{f}) = \arg\max_{\mathbf{o},\mathbf{a}} P(\mathbf{o}, \mathbf{a}, \mathbf{f}) \tag{1}$$

[3] In this work, we use the term objects to refer to instances of generic object categories.

[4] We use capital letters, e.g. $\mathbf{X}$, to denote random variables and small letters e.g. $\mathbf{x}$ to denote values taken by $\mathbf{X}$.



(a) RGB-D Image

(b) Object-Action Intention Network

Fig. 5. Automatically configured Object-Action Intention Network in (b) for a real 3D scene as captured by an RGB-D camera shown in (a). The object category instances present in the scene and their distances from the camera are used to automatically generate the network along with the associated actions that can be performed on them.

In this work, we only utilize an object's distance from the camera as a contributing object feature (i.e. $c = 1$) for reasoning and these variables indicate whether an object is near or far from the camera based on a dynamically-determined distance threshold. The dependencies of these object, action, and feature variables are captured by the network with the overall joint probability distribution $P(o_1, \ldots, o_n, a_1, \ldots, a_m, \mathbf{f}_1, \ldots, \mathbf{f}_{cn})$. Objects in the scene are resolved via object categorization and the object distances from the camera are calculated via object pose estimation as described earlier. An example of an object-action intention network is shown in Figure 5 (b) for the scene shown in Figure 5 (a). The representational complexity of these networks is $\mathcal{O}(nm)$, where n is the number of objects and m is the number of possible actions. The computational complexity is proportional to maximum size of the cliques in these networks.

Query Selection The user is prompted with a series of *queries* based on the marginal probabilities of variables and factors in the network. A query in its most generic form is a yes-or-no question involving an object variable, an action variable, or a combination of both[5]. For every network that is constructed, there is a set of s generated queries $\mathbf{Q} = \{Q_1, ..., Q_s\}$ sorted according to their probabilities. The goal is to get the first query Q_1 of $\mathbf{Q}$ to match the intent of the human. For this to be achieved, the user's selections must be learned and this is described in a subsequent section. Moreover, an individual query Q_t may represent an attempt to determine if the user wants to manipulate an object $Q_t = \{O_i\}$, perform an action $Q_t = \{A_j\}$, or perform an action on a specific object $Q_t = \{O_i, A_j\}$. The user's response leads to a modification of the set $\mathbf{Q}$, eventually resulting in one that only contains the query corresponding to the user's intention. Feedback from the user is treated as observations or evidence in the network and the network is updated accordingly (see Figure 4).

[5] A query can be mapped to different types of human-robot interfaces that vary in terms of communication bandwidth e.g. touch screen (more queries) vs. brain computer interface (BCI) (less queries).

4.2 Incremental Intention Learning

To implement the modification of the probability distribution over all object-action pairs in response to user's preference, we cast our learning framework as a form of Recursive Bayesian Incremental Learning described as follows - Let $\boldsymbol{\theta}$ represent the collection of multinomial parameter vectors for the network factors defined over all object and action variables, hence $\theta_i = \phi_i$, where ϕ_i is the i^{th} factor in a network. Furthermore, let $\mathcal{D}^k = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ explicitly represent k observed user choices where $\mathbf{x}_i = \{No, No, \ldots, o_i = Yes, \ldots, No, a_j = Yes, \ldots, No\}$ indicating a user's selection of the i^{th} object and the j^{th} action as captured via our object action intention network (the No's coincide with the objects and actions that were not selected). Our aim is to determine the most likely object-action pair based on information we have already acquired. Therefore, by using Bayes formula, the posterior probability for the distribution over all object and action variables satisfies the recursive relation given in Equation 2.

$$P(\boldsymbol{\theta} \,|\, \mathcal{D}^k) \;\propto\; P(\mathbf{x}^k \,|\, \boldsymbol{\theta}) P(\boldsymbol{\theta} \,|\, \mathcal{D}^{k-1}) \tag{2}$$

Equation 2 allows us to incrementally learn a user's preferences as they repeatedly interact with our system and data are collected. Given that $P(\boldsymbol{\theta} \,|\, \mathcal{D}^0) = P(\boldsymbol{\theta})$, we can use this equation repeatedly to produce the sequence of probabilities $P(\boldsymbol{\theta})$, $P(\boldsymbol{\theta} \,|\, \mathbf{x}_1)$, $P(\boldsymbol{\theta} \,|\, \mathbf{x}_1, \mathbf{x}_2)$, and so on. The term $P(\mathbf{x}^k \,|\, \boldsymbol{\theta})$ in Eq. 2 is known as the likelihood function and it represents the probability of the observed data given the parameters values $\boldsymbol{\theta}$. Its value is given in Equation 3 whereby the probability of each datum $\mathbf{x}_i$ given the parameters $\boldsymbol{\theta}$ is proportional to the product of the factors ϕ_i defined over the subset of random variable values $\mathbf{d}_i$.

$$P(\mathbf{x}^k \,|\, \boldsymbol{\theta}) \propto \prod_{i=1}^{k} \phi_i(\mathbf{d}_i \mid \boldsymbol{\theta}) \tag{3}$$

$$\propto \prod_{i=1}^{k} \prod_{j=1}^{s} \theta_{ij}^{N_{ij}} \tag{4}$$

Each object-action factor ϕ_i has s values (in our case $s = 4$), therefore θ_{ij} represents the j^{th} value of the i^{th} factor as shown in Eq 4. N_{ij} represents the selection of the respective factor value which can be 1 or 0 in our case.

The term $P(\boldsymbol{\theta} \,|\, \mathcal{D}^{k-1})$ in Equation 2 represents the prior probability distribution based on the set of previously observed data samples. The prior is updated as data are collected, thereby producing the posterior distribution which then serves as the prior for the subsequent observation. We assume that the distributions under consideration are of the Dirichlet form [11]. This translates into each factor ϕ_i defined over a subset of variable values $\mathbf{d}_i$ being Dirichlet. As a result, the posterior distribution is also Dirichlet, which allows us to update them using sufficient statistics from the data. What follows is that the maximum *a posteriori* estimate for $\boldsymbol{\theta}$ is given according to sufficient statistics as shown in Eq. 5.

$$\hat{\theta}_{ij} = \frac{N_{ij} + \alpha_{ij} - 1}{\sum_{j=1}^{s} N_{ij} + \sum_{j=1}^{s} (\alpha_{ij} - 1)} \tag{5}$$

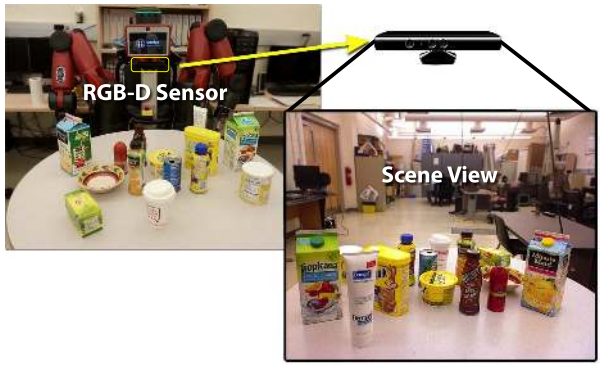


Fig. 6. Experimental set up: RGB-D scenes of common household objects are captured and processed. Objects are extracted, categorized, and their poses estimated and this is used to determine the most likely object and action that an individual desires.

With equivalent α_{ij}'s, we are left with the form shown in Eq. 6 which is our learning equation.

$$\hat{\theta}_{ij} = \frac{\lambda N_{ij} + 1}{\sum_{j=1}^{s} \lambda N_{ij} + s} \tag{6}$$

The Dirichlet hyperparameter α_{ij} stores the prior count observed for the value j of factor ϕ_i whereas N_{ij} represents its current count and $\lambda = \frac{1}{\alpha - 1}$. Thus, the parameters $\boldsymbol{\theta}$ are updated as more information becomes available. As a result, when the user selects an object to manipulate and an action to perform on that object, the value of the corresponding parameter value θ_{ij} is updated.

5 Experimental Validation

The experiments detailed in this section were performed on RGB-D scenes of common household objects captured by an RGB-D sensor where all of the objects are located on table-tops (see Figure 6). We outline the evaluation set up, outline the baseline approaches used for comparison, then we present our results.

5.1 Evaluation Set Up

To determine the user's desired intention, human-robot interaction must take place and we refer to these rounds of interactions as *sessions*. Thus, at the end of each session, the object the human wants to manipulate and the action they want to perform on that object is determined. The main performance metric we use for our evaluations is the number of human-robot interactions per session for arriving at the correct intention. The ideal scenario occurs when this value is 1, which indicates that the first query proposed to the user is their actual intent. Experiments were conducted to ascertain this value over multiple sessions

Table 1. Examples of scenes used for evaluation. The scenes differ by a combination of different objects and varying object positions

using multiple instances of objects from 11 categories with at least 4 possible actions per object. We evaluate the performance of our framework under significant scene changes between sessions and its response after being trained with a predetermined group of intentions.

Baseline Intention Recognition Approaches To the best of our knowledge, no algorithm exists in the state of the art with which a direct comparison can be performed. Therefore we constructed 3 approaches for this purpose:

- *Random*: randomly selects an object-action pair from the set of all possible configurations of objects and actions.
- *Scene-Random*: randomly selects an object-action pair from the set of possible configurations of objects and actions afforded by the scene.
- *Scene-Probability*: selects the highest probable object-action pair based on the total number of object-action pairs possible in the scene.

5.2 Performance under significant scene changes

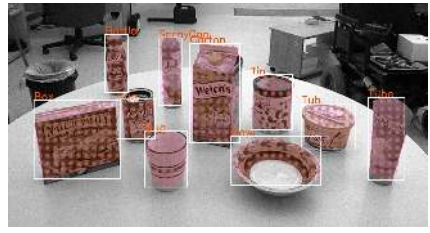
In this section, we demonstrate how our framework performs under considerable scene changes. As an example, consider the differences between scenes capturing a bathroom counter-top versus a kitchen counter-top. Our test scenarios involve at most 10 randomly-selected and randomly-positioned objects. For each session, each object is either placed in a different position or replaced with another. One possible intention (object-action pair) is chosen for 20 consecutive sessions and the number of interactions required to communicate this intention per session is calculated. This results in a total of 340 test scenes (see Table 1 for some examples) Consider Figure 7 which lists the top 5 queries of query sets generated for the 1st, 10th, and 20th sessions for one of the runs of this experiment. The goal intention for this run was `Drink-from-Bottle`. As time progressed, queries involving the object and action comprising this intention are ranked higher as the system learns and accomodates for the user's selections. By the 20th session, the first query proposed to the user actually coincides with the desired intention. The aforementioned procedure is followed for all the intentions considered in this work and the average number of interactions required per session is determined.

Session 1
Q1. Do you want to grasp something?
Q2. Do you want to move something?
Q3. Do you want to use **bottle**1?
Q4. Do you want to grasp **bottle**1?
Q5. Do you want to move **bottle**1?

Session 10
Q1. Do you want to use **bottle**1?
Q2. Do you want to **drink** from something?
Q3. Do you want to **drink** from **bottle**1?
Q4. Do you want to grasp **bottle**1?
Q5. Do you want to grasp **bottle**1?

Session 20
Q1. Do you want to **drink** from **bottle**1?
Q2. Do you want to **drink** from something?
Q3. Do you want to use the **bottle**1?
Q4. Do you want to grasp **bottle**1?
Q5. Do you want to grasp something?

(a) Top 5 queries for the 1st, 10th, and 20th sessions.



(b) Session 1 scene after vision processing

Fig. 7. Example run for the desired intention `Drink-from-Bottle`. Notice how queries involving both `drink` and `bottle` are ranked higher than others over time

Figure 8 displays the result of the complete experiment and compares our results with those acquired for the baseline approaches[6]. The figure shows that despite considerable modifications to the scene, our framework manages to reduce the number of human-robot interactions over time and consistently outperforms the other methods. By the 20th session, it takes approximately 1 interaction to determine what the user desires and the number of interactions were reduced by an average of 81%. We have discovered that as a result of our intention recognition formulation, placing objects closer to the camera increased their likelihood for selection and vice versa. Therefore, by moving desired objects away from the camera, the average number of interactions increased. This phenomenom is usually observed in the first couple of sessions (see how the Scene-Probability method outperforms our method for the 1st session in Figure 8). However, this effect is significantly reduced after multiple sessions and is of negligible impact as a result of learning.

[6] It should be noted that for all of the experiments in this work, the object-action pairs of the constructed networks are initialized with equivalent probabilities which are then altered over time due to incremental learning. Also, multiple instances of an object category can be present in the scene at the same time.

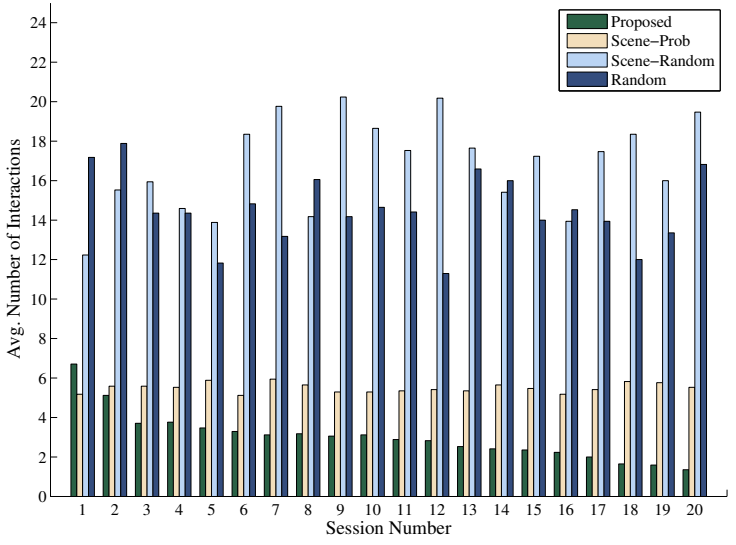


Fig. 8. Intention recognition results for significant scene changes. Each session corresponds to a different scene where the object composition of the scene varies. Notice how over time our method consistently reduces the amount of human-robot interactions required for communicating tasks to the robot.

5.3 Performance after intention group training

Consider the following scenario: a person wakes up, brushes their teeth, drinks a cup of coffee, and eats a bowl of cereal before they head off to work. They repeat this sequence of events every morning. For humans, it takes relatively no effort to determine this person's morning routine after some time. For instance, if the individual's spouse wants to help them get to work faster, all they have to do is put toothpaste on the toothbrush, make coffee, and prepare the cereal ahead of time because they are cognizant of their spouse's routine. For this reason, this section presents the results of training on a group of intentions over time then determining the amount of interaction required to choose one of them from the group. This is somewhat analogous to learning the person's morning routine as previously described. Ideally, the selection likelihood of the intentions in the group should be higher than all other possible intentions, therefore the amount of interaction required to select one of them should be small.

The experiment is performed on scenes where the objects and their positions vary over the span of at least 50 sessions. Each intention in the group is selected at most 10 times in no particular order given a conducive scene. At the conclusion of this "training" period, one of these intentions is randomly chosen and the average number of interactions required for choosing it is calculated. Figure 9 illustrates the results of this experiment. It shows that our framework reduces the

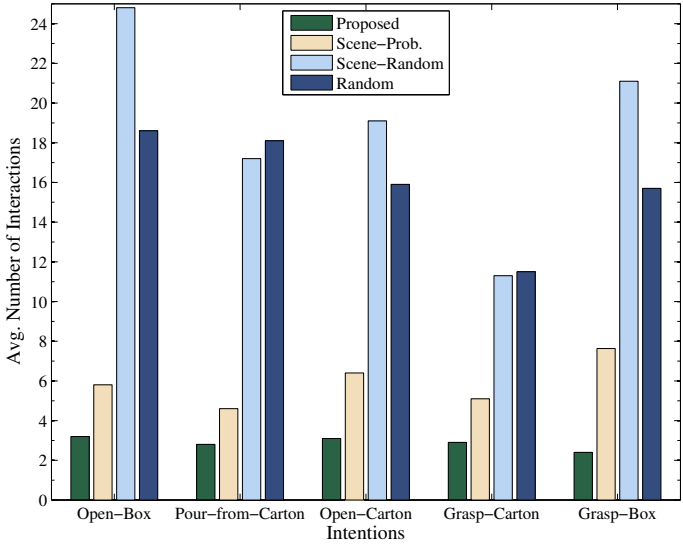


Fig. 9. Group Training: a group of intentions are learned over time then the amount of interaction required to choose one of them from the group is determined.

necessary amount of interaction for all intentions tested and that it consistently outperforms the baseline methods. This behavior is desired because we want our framework to be able to capture a user's preferences over time in order to simultaneously reduce human interaction and maximize robot task performance but at that the same time be flexible enough to adapt to new information.

6 Conclusion

In this paper, we have presented a vision-based human-robot collaborative system that enables the recognition and learning of human intentions. At the core of this system is our object-action intention recognition framework that is only dependent on scene information for inferring intentions rather than on the observation of human physical actions, which is the commonly-accepted approach. This is our principal contribution to the state-of-the-art as it is appropriate for assistive robotic systems for persons with limited physical capabilities. We have demonstrated through our evaluations that our framework is capable of adapting to a user's preferences and reduces the amount of interaction necessary for communicating tasks to a robot.

References

1. Carlson, T., Demiris, Y.: Collaborative Control for a Robotic Wheelchair: Evaluation of Performance, Attention, and Workload. IEEE Transactions on Systems,

Man, and Cybernetics, Part B: Cybernetics 42(3), 876–888 (2012)
2. Charniak, E., Goldman, R.P.: A Bayesian model of plan recognition. Artificial Intelligence 64, 53–79 (1993)
3. Collet, A., Martinez, M., Srinivasa, S.S.: The MOPED framework: Object recognition and pose estimation for manipulation. The International Journal of Robotics Research 30(10), 1284–1306 (Apr 2011)
4. Collet, A., Berenson, D., Srinivasa, S.S., Ferguson, D.: Object recognition and full pose registration from a single image for robotic manipulation. In: 2009 IEEE International Conference on Robotics and Automation. pp. 48–55 (May 2009)
5. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: IEEE Conference on Computer Vision and Pattern Recognition. vol. 1, pp. 886–893. IEEE (2005)
6. Demeester, E., Huntermann, A., Vanhooydonck, D., Vanacker, G., Brussel, H.V., Nuttin, M.: User-adapted plan recognition and user-adapted shared control: A Bayesian approach to semi-autonomous wheelchair driving. Autonomous Robots 24, 193–211 (2007)
7. Duncan, K., Sarkar, S., Alqasemi, R., Dubey, R.: Multi-scale Superquadric Fitting for Efficient Shape and Pose Recovery of Unknown Objects. In: International Conference on Robotics and Automation (2013)
8. Heinze, C.: Modelling intention recognition for intelligent agent systems. Ph.D. thesis, The University of Melbourne, Melbourne, Australia (2003)
9. Kelley, R., Tavakkoli, A., King, C., Ambardekar, A., Nicolescu, M., Nicolescu, M.: Context-Based Bayesian Intent Recognition. IEEE Transactions on Autonomous Mental Development 4(3), 215–225 (Sep 2012)
10. Kelley, R., Wigand, L., Hamilton, B., Browne, K., Nicolescu, M., Nicolescu, M.: Deep networks for predicting human intent with respect to objects. In: Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction - HRI '12. p. 171 (2012)
11. Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. MIT Press, 1 edn. (2010)
12. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
13. Madry, M., Song, D., Ek, C.H., Kragic, D.: "Robot bring me something to drink from": Object Representation For Transferring Task Specific Grasps. In: IEEE International Conference on Robotics and Automation (2012)
14. Madry, M., Song, D., Kragic, D.: From Object Categories to Grasp Transfer Using Probabilistic Reasoning. In: IEEE International Conference on Robotics and Automation (2012)
15. Meger, D., Forssén, P.E., Lai, K., Helmer, S., McCann, S., Southey, T., Baumann, M., Little, J.K., Lowe, D.G.: Curious George: An attentive semantic robot. Robotics and Autonomous Systems 56, 503–511 (2008)
16. Ramík, D.M., Madani, K., Sabourin, C.: From visual patterns to semantic description: A cognitive approach using artificial curiosity as the foundation. Pattern Recognition Letters 34, 1577–1588 (2013)
17. Rusu, R., Blodow, N., Beetz, M.: Fast Point Feature Histograms (FPFH) for 3D Registration. In: IEEE International Conference on Robotics and Automation. pp. 3212–3217. IEEE (2009)
18. Rusu, R., Blodow, N., Marton, Z.: Close-range scene segmentation and reconstruction of 3D point cloud maps for mobile manipulation in domestic environments. In: International Conference on Intelligent Robots and Systems. pp. 3–8 (2009)

19. Tahboub, K.A.: Intelligent Human-Machine Interaction Based on Dynamic Bayesian Networks Probabilistic Intention Recognition. Journal of Intelligent and Robotic Systems 45(1), 31–52 (March 2006)
20. Tavakkoli, A., Kelley, R., King, C., Nicolescu, M., Nicolescu, M., Bebis, G.: A Vision-Based Architecture for Intent Recognition. In: Advances in Visual Computing, pp. 173–182. Springer (2007)
21. Vanhooydonch, D., Demeester, E., Nuttin, M., Brussel, H.V.: Shared Control for Intelligent Wheelchairs: An Implicit Estimation of the User Intention. In: 2003 International Workshop on Advances in Service Robotics. pp. 176–182 (2003)
22. Zhu, C., Sun, W., Sheng, W.: Wearable Sensors based Human Intention Recognition in Smart Assisted Living Systems. In: International Conference on Information and Automation. pp. 954–959 (2008)