2018 IEEE International Conference on Robotics and Automation (ICRA)
May 21-25, 2018, Brisbane, Australia

Human-Driven Feature Selection for a Robotic Agent Learning Classification Tasks from Demonstration

Kalesha Bullard[1], Sonia Chernova[1], and Andrea L. Thomaz[2]

***Abstract*— The state features available to a robot define the variables on which the learning computation depends. However, little prior work considers feature selection in the context of deploying a general-purpose robot able to learn new tasks. In this work, we explore human-driven feature selection in which a robotic agent can identify useful features with the aid of a human user, by extracting information from users about which features are most informative for discriminating between classes of objects needed for a given task (*e.g.* sorting groceries). The research questions examine (a) whether a domain expert is able to identify a subset of informative task features, (b) whether human selected features will enable the agent to classify unseen examples as accurately as using computational feature selection, and (c) if the interaction strategy used to elicit the information from the user impacts the quality of resulting feature selection. Toward that end, we conducted a user study with 30 participants on campus, given a multi-class classification task and one of five different approaches for conveying information about informative features to a robot learner. Our findings show that when features are semantically interpretable, *human feature selection* is effective in LfD scenarios because it is able to outperform computational methods when there is limited training data, yet still remains on-par with computational methods as the training sample size increases.**

I. Introduction

Research on robot learning from demonstration (LfD) focuses on the development of robots capable of learning a wide range of tasks from a small number of human demonstrations. Much of the work in this field is particularly aimed at the development of general-purpose robots capable of performing multiple tasks, such as a household robot able to put away groceries as well as cook a meal, or a service robot able to execute multiple maintenance procedures. Most research in this area assumes that a set of features representing the state of the robot and the surrounding environment are available to the robot, and that these features are then applied to learning new actions (e.g., open cabinet) or new task models (e.g., make coffee) [1].

The state features available to the robotic agent define the variables on which the learning computation depends. However, little prior work considers feature selection in the context of deploying a general-purpose robotic agent able to learn new tasks. Given a new set of demonstrations, which features should the agent use to learn? And is it necessary for the agent to discover informative features *on its own* or is there benefit to soliciting the *help* of a human partner? The feature selection problem is substantial because a general-purpose robotic agent may have the ability to track dozens, or even hundreds, of potential features in its environment, only a handful of which are likely to be relevant for any given task, and incorporating too many unnecessary features leads to poor learning performance. Computational feature selection techniques [2], [3], [4], which rely on identifying statistical patterns in data, may not have sufficient evidence given the small number of training examples encountered in LfD. In fact, in prior work, all LfD papers we surveyed used hand-coded state features, with the exception of [5], in which computational feature selection techniques are applied to identify relevant features based on human demonstrations in the games Frogger and Pong.

Fig. 1: Instances selected by user study participant to teach robot about the specified classes of objects in *sort groceries* task.

Our goal is to enable the robotic agent to decipher informative features for *any* task, using only small number of examples. Thus in this work, we explore interactive feature selection in which an agent can identify relevant features with the aid of a human user. Humans familiar with a target domain typically have the ability to characterize which features are important in decision making, at least at an abstract level. We explore whether non-expert users are able to identify which features are most informative for discriminating between classes of objects needed for a given task, how best to elicit the feature information from the user, and how computational feature selection compares to human-driven feature selection given varying amounts of data. Specifically, we explore three research questions:

1) Is a domain expert able to identify a subset of informative features for discrimination between classes of objects needed for a given task?
2) Do features selected by the domain expert enable the agent to classify unseen examples as accurately as using computational feature selection techniques?
3) Does the way in which information is elicited from the

*This work was supported by the NSF NRI grant.

[1] School of Interactive Computing, Georgia Institute of Technology, Atlanta, Georgia 30332-0250 Email: ksbullard@gatech.edu, chernova@cc.gatech.edu

[2] Department of Electrical & Computer Engineering, The University of Texas at Austin, Austin, Texas 78701 Email: athomaz@ece.utexas.edu

user impact the quality of resulting feature selection?

To address the third research question, we developed three general categories of approaches for allowing humans to communicate feature information to the agent:

1) **Direct Communication** - the user *directly* communicates about features useful for the task by selecting or eliminating from a superset of candidate features
2) **Indirect Inference** - the human teacher selects a small number of instances from each task-relevant object class and the learning agent *indirectly* infers which features are being communicated by the examples shown
3) **Combined Approach** - the human teacher *both* selects a small number of instances from each object class as examples *and* subsequently chooses features being communicated from a superset of candidate features

To study the above research questions we conducted a between-subjects user study with 30 participants. Participants were asked to help a robotic agent discriminate between four classes of objects needed for a household task, by communicating about informative features using one of the interaction strategies above. All users were assumed to have prior knowledge about the task domain. Our findings show that (1) human feature selection outperforms computational methods when there is a small amount of training data and remains on-par with computational methods as the sample size grows, (2) *direct communication* is the most effective strategy for eliciting feature information from users when the task features are intuitive, and (3) when a relatively large amount of training data is available, asking a human teacher to first select a small number of informative instances then *indirectly* inferring the features being communicated leads to the best performance. We also conducted a follow-on study in three additional task domains to examine how reliably users directly communicate informative features; the supplemental findings show people are able to select useful features for a task only when the features are semantically interpretable.

II. Learning Task

A. Problem Statement

The problem of learning task features consists of determining a subset of features for use in building all of the classifiers needed for the task [1]. In our problem formulation, the robotic agent is given a set of task-specific labels to be learned, Y, and the superset of all candidate features, F, associated with the observed state of the world. The agent's goal is to learn how to classify instances of all the labels Y. Feature selection then, either by using a computational approach or by asking questions of a human teacher, can help the agent determine a subset of features $F' \subset F$ that represent a single state space appropriate for all classes Y.

B. Problem Domain

As our running example, we situate a robotic agent within a kitchen setting, learning to sort groceries, as shown in Figure 1, presumably in order to be later put away. We define the *sort groceries* task as teaching the agent to distinguish between four object classes (produce, snacks, food cans & jars, and beverages). For each object instance encountered by the robot, we compute the superset of all candidate features F based on perceptual information extracted from an RGB-D image of the object, the object's relative location to the robot, the robot's joint position information at the time, and audio input from the environment. Table I lists the feature categories and number of features each decomposes into, for a total of 84 low-level features. The agent's goal is to determine which of the listed features are relevant to its task.

TABLE I: High-Level Task Features *(per object instance)*

absolute location (3-dimensional) of object in environment (3)
orientation (yaw) of object on surface (1)
location of object relative to five specified interest points (15)
location of robot's base in environment (3)
pose of robot's two hands (pose quaternion) relative to its body (14)
pose of robot's two hands (pose quaternion) relative to counter (14)
orientation (yaw) of robot's base on ground (1)
robot hand states (open or closed) (2)
position for each joint of robot's 7-dof arms (14)
average color of object (3)
object bounding box size measurements (3)
area of object bounding box (1)
volume of object bounding box (2)
aspect ratio for object bounding box (1)
surface area to volume ratio for object bounding box (1)
compactness of object point cloud (1)
number of SIFT features (measure of visual texture) (1)
max/min/average volume of noise in environment over duration of learning interaction (3)
weight of object (1)

We use the University of Washington RGB-D Object Dataset to obtain a standard set of object images for testing [6]. The object dataset includes over 200,000 images in total, encompassing over 300 objects organized into 51 categories (*e.g.* soda can), with multiple object instances per category (*e.g.* pepsi can, mountain dew can, etc.). For each object instance, the database contains several hundred images captured from different viewpoints and distances from the camera, and some objects in the dataset have been captured under more than one lighting condition. For the sort groceries task, we consider only images related to produce (fruits and vegetables), snacks (food bags, food boxes, and cereal), food cans & jars, and beverages (water bottles and jugs). In addition to using the images, we purchased approximately 60 objects from the dataset to use in the user study.

III. Computational Feature Selection

In this section, we briefly discuss computational feature selection methods and establish a baseline for the impact of feature selection on learning performance in our domain.

A. Algorithm Overview

Feature Selection (FS) aims to eliminate irrelevant and redundant features such that $g : F \rightarrow F'$, where g is the feature selection function. In selecting feature subsets, features are typically evaluated for *relevance* or *usefulness* [2], [3]. There are three classes of computational approaches

for automatic feature selection that have been explored in the literature: *filters*, *wrappers*, and *embedded methods* [4]. Wrappers conduct an exhaustive search through the space of feature subsets to find an optimal solution, thus are computationally intractable given 84 features and 2^{84} candidate feature subsets. Filters are the most computationally efficient class of algorithms but are generally inferior in performance, since they only consider data and not the learning model. Embedded algorithms are considered to be state of the art in computational FS, as they attempt to strike a balance between learning performance and computational efficiency. We test filtering and embedded algorithms in this work. In terms of classifier representation, we selected support vector machines since they typically perform well with sparse data and compared learning performance against two other discriminative classifiers (k-nearest neighbors and random forests)[1]. We observed the best performance using an SVM classifier with a radial basis function kernel. Thus, SVMs are used for all learned models in this work. Below we briefly introduce the feature selection techniques employed.

1) Filtering: Filters take as input the training data and examine the *relevance* of each feature $f \in F$ with respect to each class label $y \in Y$. The **filtering** algorithm **(FI)** employed ranks features based upon information gain IG and selects all features $f \in F$ such that $IG(Y|f) > \tau$ where $\tau = 0$.

2) Embedded Methods: Embedded methods conduct a best first search through the space of feature subsets, evaluating the *usefulness* of each subset $F' \subset F$ with respect to a given predictor. We use two embedded algorithms in this work: **embedded selection (ES)** begins with no features and incrementally adds (*forward selection*) whereas **embedded reduction (ER)** begins with all features and incrementally prunes (*backward elimination*). Both employ a greedy search strategy and evaluate subset F' based upon predictive ability of each feature $f \in F'$ and redundancy between them [8].

B. Evaluation

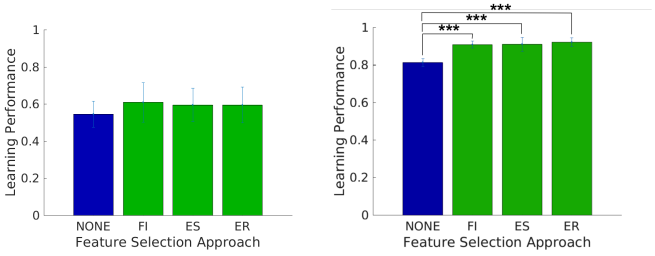
To evaluate the effect of feature selection on our chosen domain, we compare the performance of the above algorithms with an SVM classifier on the sort groceries task. For evaluation, we create a test set, P^{test}, containing 1000 task-relevant images sampled without replacement using stratified random sampling (SRS) from the object dataset. The training set, P^{train}, is similarly sampled to generate $k = 10$ disjoint training samples, $D_{i,\ldots,k}$, each consisting of n sampled images. We use the following evaluation metric to evaluate algorithm performance:

$$E[acc_{\mathbb{D}}(a)] \approx \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n}\sum_{x \in D_i}[1-\delta(h_i^a(x),y)] \qquad (1)$$

where $E[acc_{\mathbb{D}}(a)]$ represents the expected value of the learning accuracy using FS approach a with respect to distribution $\mathbb{D}$, $h_i^a(x)$ is the hypothesis of the learner using a given instance x in training set D_i, y is the ground truth label for instance x, and the quantity $\delta(h_i^a(x),y)$ is 1 if $h_i^a(x) \neq y$ and 0 otherwise.

We aimed to test the FS approaches, given both small and large amounts of training data. We select a *small* training set size that is reasonable for an LfD scenario yet still provides enough examples of each task-relevant object so as to represent some of the diversity of objects that exists within each category. Given the task-relevant categories selected are broad (*e.g.* fruit) and include several types of objects (*e.g.* apples, oranges, lemons, etc.), we decided to test the performance of the learner using at least three object instances per class or n = 12. Additionally, we computed accuracy as a function of the training sample size, in order to understand how much data is needed for learning performance to converge. This happened after about 25 examples per class or n = 100. Thus Figure 2 reports results for $n = 12$ and $n = 100$ training instances, representing relatively *small* and *large* training sample sizes appropriate for our task domain.

[1]From the Weka Software Library [7]



(a) Train Set Size = 12 (b) Train Set Size = 100

Fig. 2: Accuracy of computational FS algorithms for classification of objects in *Sort Groceries* task. Test Set Size = 1000.

C. Results

Given both small and large amounts of training data, FS aids with learning performance. However, the difference in expected performance between a learner using computational FS and a learner using no FS is dependent upon the amount of training data observed. When there is a relatively large amount of training data, as depicted by Figure 2b, using computational FS yields statistically significantly less error than using none, no matter which approach is employed. With only a small amount of training data, overall classification performance is lower and there is no statistically significant difference between expected error of learners with no FS and learners with computational FS.

Hence what we observe is that computational approaches are limited in their ability to improve learning performance when there is a small amount of data, since these approaches are data-driven. Nonetheless in LfD, it is commonly the case that an agent is provided only a small number of examples from the human teacher and can leverage *only* these in order to learn the task. This motivates the need for other techniques for acquiring a subset of discriminative features when there is limited training data available. From this point forward, we use only FI as the computational baseline for small training samples and ER as the baseline for large samples, since each narrowly outperforms the other computational algorithms in the respective scenarios.

IV. Human-Driven Feature Selection

In this work, we are interested in enabling a robotic agent to characterize the essential features of a task when there are very few training examples to observe, since this

is typically the case in LfD scenarios. Given limited data, being more selective about features helps the agent better discriminate between the object classes relevant to the task; however computational methods perform poorly when there is a scarcity of data. In lieu of this, we hypothesize that humans with task domain knowledge can help the agent by providing information about what features they believe to be most informative for the task.

We propose and compare five approaches to determine how human teachers can best aid with the feature selection problem. We group the techniques into three categories based upon interaction style and the type of information they provide: (1) *direct* communication of features, (2) *indirect* inference of features, and (3) *combination* of indirect and direct communication of features.

A. Direct Communication of Features

The first experimental category allows a human teacher to communicate feature information directly, based upon what features seem intuitive to the teacher. We explore two *direct communication* approaches for eliciting information: (a) **human feature selection (HFS)** and (b) **human feature reduction (HFR)**. For both approaches, the human teacher is provided a list of hierarchically arranged candidate features and has the option to choose entire (sub)categories of features to indicate that every feature in the set should be marked or alternatively only choose the individual features within the hierarchical category that are appropriate (*e.g. size features of object*: *volume*, *surface area*, *length*, *width*, and *height*). For human feature selection, the teacher's goal is to provide *only* features they believe to be most useful for the agent in determining which class an unseen object belongs to. In contrast, for human feature reduction, the teacher's goal is to specify features the learning agent should *not* pay attention to (*i.e.* features it should *ignore*) because they will *not* help it determine the class of an unseen object.

B. Indirect Inference of Features

We also realize however that low-level features associated with robot sensors (*e.g.* rgb channel intensities) may not all be intuitive for humans. And people are used to selecting representative examples to characterize a target concept. Thus the second category explored involves *indirect inference* of features: **human instance selection (HIS)**. With this approach, the human teacher's goal is to enable the robotic agent to distinguish between the classes of objects needed for the task by providing a small number of examples of each. The examples selected for each class $c \in C$ are specifically intended to help the agent determine which features are most useful when identifying objects belonging to class c. Then computational FS is used to infer which features were being communicated by the training examples selected. As a note, we only *seed* the training set with the small number of examples selected by the teacher; the rest are automatically generated using SRS.

C. Combined Approach for Conveying Features

Lastly, there are two *combined* approaches, in which the teacher's goal is to *first* select instances, then *follow up* with direct communication of features. The motivation for this category is to allow the teacher to subsequently reflect and explicitly communicate to the learning agent what they were attempting to implicitly highlight through the instances selected prior. We explore: (a) **human instance selection + human feature selection (HIS-FS)** and (b) **human instance selection + human feature reduction (HIS-FR)**.

D. Evaluation and User Study

We sought to explore three research questions in this work: (1) whether a domain expert can identify informative task features, (2) whether human informed feature subsets can perform as well as computational FS, and (3) if the way in which feature information is elicited impacts the quality of the FS. We are especially interested in exploring this for the LfD scenario, where an agent has limited training data available, but access to a human teacher. Toward that end, we have two hypotheses we are testing: (1) humans intuitively understand and are able to characterize informative features of a task for which they have prior knowledge, and (2) humans will do better at characterizing the task *indirectly* (selecting representative instances) than *directly* (enumerating useful features). We hypothesized that people would be more adept at *indirectly* communicating features because some candidate features generated by a robot's sensors may not be as intuitive for people, and with that, we would not necessarily expect the features used by people to map directly to features generated by robot sensors.

For evaluation, we conducted a between-subjects user study with 30 participants on Georgia Tech's campus, to collect data from humans about what features they would teach to help a robotic agent differentiate between the task-relevant object classes. There were three conditions tested (10 participants per condition): (1) feature selection, (2) feature reduction, and (3) instance selection. For the study, all task-relevant objects were grouped by class on a table, but spaced out sufficiently for participants to see and interact with individual objects. All objects purchased corresponded to instances in the object dataset and therefore could be mapped to a corresponding set of images for processing.

In the first two study conditions (HFS and HFR), participants are given the option to interact with the objects in any way they desire in order to help them decide which features to select or prune. They were also asked to do a brief exit survey upon completing the teaching task. For the third study condition (HIS), once three examples per class were selected by the teacher, all twelve examples are brought to the robot's workspace. Figure 1 shows an example of a complete demonstration for all object classes associated with the sort groceries task. Then, instead of completing an exit survey, participants from the third condition were equally subdivided into two sets. Directly following the selection of instances, one set was asked to *additionally* perform feature selection (HIS-FS); the other, feature reduction (HIS-FR). The order was not counterbalanced in this condition. We intentionally requested that each of these participants *first* communicate about features in an *indirect* way by selecting instances, *then* communicate features in a *direct* way by either (a) choosing relevant features or (b) eliminating irrelevant features.

This study provided the data needed for all five human-driven FS approaches discussed on the given task. For the

TABLE II: Source of Training Data and Feature Sets (*per User*)
n = num instances in training sample
k = num training samples

FS Approach	n	k	*Instances*	*Features*
None	12	10	SRS	All
	100	10	SRS	
Computational FS	12	10	SRS	FI
{*ER*}	100	10	SRS	ER
Direct Communication	12	10	SRS	Human
{*HFS, HFR*}	100	10	SRS	
Indirect Inference	12	1	Human	ER
{*HIS*}	100	10	Human + SRS	
Combined Approach	12	1	Human	Human
{*HIS-FS, HIS-FR*}	100	10	Human + SRS	

HIS approach, we only process the first teaching strategy used by participants in the third user study condition. The data for one participant from the instance selection + feature selection subgroup had to be excluded, thereby leaving data from 29 users mapped to the interaction strategies as follows:

- HFS: 10 users
- HFR: 10 users
- HIS: 9 users (4 HIS-FS, 5 HIS-FR)

E. Learning Episode

Now that we have experimental data from users around feature subsets, we need to evaluate the extent to which the features they indicated are useful in learning the various classifiers needed for the task. We evaluate learned models as a function of the number of training instances n, in order to understand how human-driven feature selection compares to computational feature selection, given both the small and large training sample sizes. Table II shows the source for training instances and selection of feature subsets, using each category of FS approach.

Specifically, let A be the set of FS approaches and Y the set of object classes to be learned. Each learning episode consists of training $k|A|$ models, k different learned models for each $a \in A$, in each iteration j of the learning episode, as n increments from $n = 12$ to $n = 100$. The reason we generate k models $\forall a \in A$ is because we randomly generate k disjoint training samples during each iteration j. Thus for each j, $\forall a \in A$, we take the aggregate performance for all k learned models generated by approach a, in order to evaluate the *expected* performance and variance of a. We start the episode with $n = 3|Y| = 12$ examples since that is the number shown by the teacher, and we selected a termination point for the learning episode empirically based upon when learning performance stabilizes. More details are included in the subsections below.

1) Generation of Training Samples: For *indirect* and *combined* approaches, we collected three human demonstrated examples for each classifier needed for the task. Thus at the beginning of a learning episode, the training set is a uniformly distributed sample $D_{i...k,0}$ containing $3|Y|$ object instances, where D_i is the i^{th} training sample, $D_{i,0}$ represents the initial set for the i^{th} training sample, and $|Y| = 4$ for the sort groceries task. For the *computational* and *direct* feature selection approaches, whereby *all* training instances are generated using SRS, this corresponds to $k = 10$ disjoint initial training sets $D_{i...k,0}$, each also containing $3|Y|$ object instances. For the *indirect* and *combined* approaches, where human teachers have selected the instances, there is only one initial training set, and it is composed *only* of the $3|Y|$ examples selected by the human teacher; thus $k = 1$.

After the initial set of $3|Y|$ training examples, all remaining $n-3|Y|$ instances are generated using SRS $\forall a \in A$. Thus for each a, in each subsequent iteration j, a new set of object instances is sampled from D^{train} such that

$$\forall y \in Y, \forall i \mid D_{i,j} \leftarrow D_{i,j} \cup \{o_y\}$$

where o_y is an object instance belonging to class y. Table II summarizes the generation of training data $\forall a \in A$.

2) Selection of Feature Subsets: In each iteration of a learning episode, $|Y|$ new instances are added to the training sample, $\{o_y | \forall y \in Y\}$. Then feature subsets must be selected by each $a \in A$, and a classifier is trained and tested for each feature subset. For the *computational* and *indirect* approaches, feature subsets are also *dynamically* updated in each iteration. The ER algorithm is used to compute a new subset of useful features for both, based upon the *updated* training set. For the *indirect* and *combined* approaches, human teachers have already provided the subset of informative features; therefore the feature subset associated with each of these approaches remains *fixed* throughout the entire learning episode and is used for every training sample $D_i...k$.

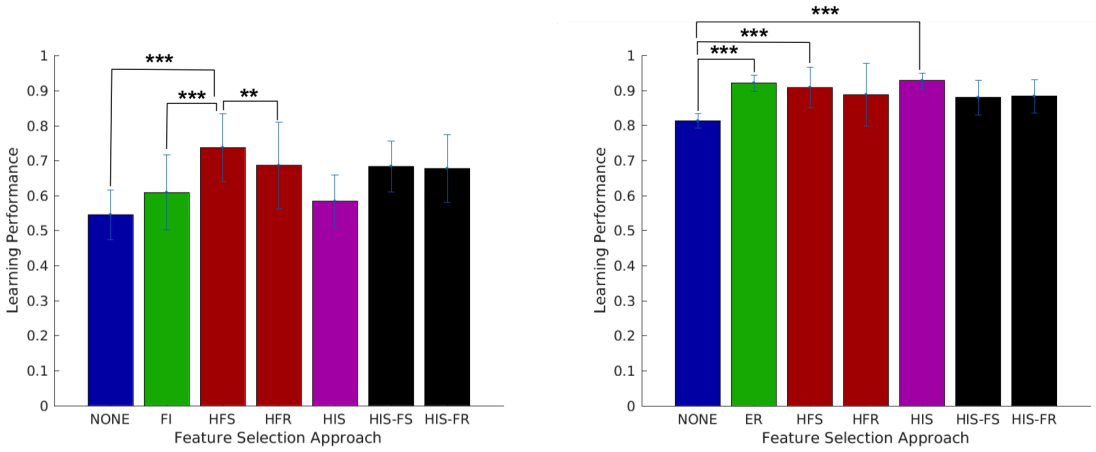
F. Results

Results in figures 3a and 3b reflect learning performance for each $a \in A$ for $n = 12$ and $n = 100$ instances respectively, as computed by Equation 2.

$$E[acc_{\mathbb{D}}(a)] \approx \frac{1}{|U_a|} \sum_{u \in U_a} \left[\frac{1}{k} \sum_{i=1}^{k} \frac{1}{n} \sum_{x \in D_i^u} [1 - \delta(h_i^a(x), y)] \right] \quad (2)$$

where D_i^u represents a training set that may have either been completely randomly generated or at least partially selected by the user $u \in U_a$, the set of users for approach a. Importantly, learning performance is now averaged across all $u \in U_a$. Where $a =$ no FS or $a =$ computational FS, we let $|U_a| = 1$ to denote one oracle that randomly generates instances for each training sample D_i. We used the Mann-Whitney U-test to compute pairwise statistical significance comparisons for each pair of FS approaches. However because there were so many comparisons, the bar graphs in Figure 3 only highlight the statistical significance relationships that juxtapose the best performing human-driven approaches with baseline approaches.

To frame the analysis of results, we are most interested in exploring human-driven FS approaches for LfD scenarios, where a robotic agent has limited training data available, but access to a human teacher. The goal is to determine if we can leverage the human where there may be insufficient evidence for computational FS methods to identify discriminative features. Towards that end, our first hypothesis was that humans intuitively understand and are able to characterize



(a) Train Set Size = 12

(b) Train Set Size = 100

Fig. 3: Learning performance of human-driven FS approaches for classification of objects relevant to *Sort Groceries* task. The task involves four object classes where each training set has an even distribution of the classes. Test Set Size = 1000.

informative features of a task for which they have prior domain knowledge.

Looking at figure 3a, with a small amount of training data, we observe that allowing a human teacher to provide feature information about the task yields a statistically significant increase in expected learning performance as compared to using only a computational approach. Specifically, the HFS interaction strategy appears to be the most effective way of eliciting information about useful features from the human teacher. It also dominates the second best human-driven interaction strategy, HFR. With a large amount of training data, computational FS (ER) and the best human-driven feature selection approaches (HIS and HFS) perform comparably; all three are statistically significantly better than a learner with *no* feature selection. Therefore while humans are not necessarily needed when there is sufficient training data available to the learner, the fact that human approaches are still on par with the best computational approach suggests that the domain knowledge extracted from humans is both *useful* for the learning task and *reliable* as sample size grows.

Thus, the results support our first hypothesis; with both small and large sized training samples, human-driven FS yielded learning performance *at least* comparable with that of the best computational FS method. However, our second hypothesis that people would be better at characterizing the task *indirectly* (selecting representative instances) than *directly* (enumerating useful features) was not supported. Direct selection of features (HFS) significantly outperformed indirect inference of features (HIS) with a small amount of data and was comparable with both HIS and the best computational approach with a large amount of training data. This implies that HFS is the optimal strategy (of those explored) for eliciting feature information from a human.

All statistical significance relationships are shown in Tables III and IV. For each approach $a \in A$, $N = k|U_a|$ where k can be found in Table II. $|U_a|$ for all human-driven approaches is listed at the end of Subsection IV-D. For each pair of approaches (cell) and given value of n, the corresponding table shows the probability p that error(approach a) $<$ error(approach b). Each row shows which FS approaches are dominated by a whereas each column shows which approaches dominate b. So *e.g.*, we observe that the baseline of no FS is dominated by every FS approach when there is a large amount of training data available.

G. Additional Task Domains

Our findings from the first experiment contradicted our second hypothesis that people would be *less* successful in *directly* enumerating informative task features. We believed this to be at least partially attributable to the features in the sort groceries task being quite intuitive for people. Thus we conducted a follow-on study to explore this further; it consisted of an online survey whereby 48 participants were given three tasks: (1) playing Pacman for a reinforcement learning agent [9], (2) autonomous navigation through a crowded environment for a mobile robot [10], and (3) classification of fire, smoke, and thermal reflections for a humanoid firefighting robot [11]. For each task domain, the participant was shown an image of the domain, then asked to check off all features they believed to be most useful for a robotic agent learning to perform the given task.

For each domain, an empirically validated set of useful features is provided by the source referenced, thus used as our baseline for comparison. All participants were recruited from the same population of on-campus students. Table V lists baseline features selected for each domain. Figure 4 illustrates amount of overlap between human-selected features and baseline feature subset for each domain, where a feature was included if selected by at least half of participants.

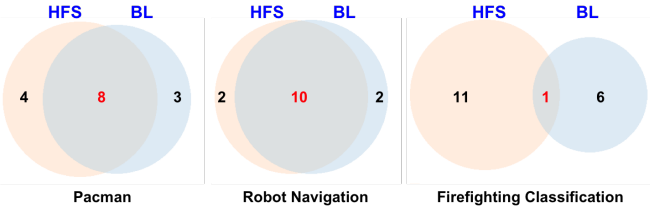


Fig. 4: Venn Diagrams to show amount of overlap between selected feature subsets for each task domain, where BL=baseline set of features empirically validated by source referenced.

TABLE III: Statistical Significance Relationships
where A = *Error(approach a)* and B = *Error(approach b)*
(p-values for $n = 12$)
$* = p < 0.05$; $** = p < 0.01$; $*** = p < 0.001$

$H: A < B$	*b:* None (N=10)	*b:* ER (N=10)	*b:* HFS (N=100)	*b:* HFR (N=100)	*b:* HIS (N=90)	*b:* HIS-FS (N=40)	*b:* HIS-FR (N=50)
a: None	–	0.82	1.0	1.0	0.84	0.99	0.98
a: ER	0.19	–	1.0	0.99	0.41	0.93	0.93
a: HFS	***	***	–	**	***	0.09	0.07
a: HFR	***	***	1.0	–	**	0.33	0.40
a: HIS	0.18	0.60	1.0	1.0	–	0.95	0.96
a: HIS-FS	*	0.09	0.91	0.67	0.07	–	0.06
a: HIS-FR	*	0.08	0.93	0.60	0.06	0.63	–

TABLE IV: Statistical Significance Relationships
where A = *Error(approach a)* and B = *Error(approach b)*
(p-values for $n = 100$)
$* = p < 0.05$; $** = p < 0.01$; $*** = p < 0.001$

$H: A < B$	*b:* None (N=10)	*b:* ER (N=10)	*b:* HFS (N=100)	*b:* HFR (N=100)	*b:* HIS (N=90)	*b:* HIS-FS (N=40)	*b:* HIS-FR (N=50)
a: None	–	1.0	1.0	1.0	1.0	1.0	1.0
a: ER	***	–	0.59	0.28	0.87	*	**
a: HFS	***	0.41	–	*	0.94	***	***
a: HFR	***	0.72	0.96	–	1.0	**	**
a: HIS	***	0.13	0.06	***	–	***	***
a: HIS-FS	***	0.99	1.0	0.99	1.0	–	0.57
a: HIS-FR	***	0.99	1.0	1.0	1.0	0.43	–

TABLE V: Task Domain Selected Features. Firefighting features are with respect to pixel intensities from thermal images of scene.

Pacman	**Navigation**	**Firefighting**
grid width / height	location / orientation of robot	mean
grid cell locations	speed / direction robot is traveling	variance
location of walls	location / orientation of each pedestrian	standard deviation
number of ghosts	num pedestrians per square foot, in robot neighborhood	skewness
location of Pacman / ghosts / food / capsule(s)	speed / direction each pedestrian is traveling, relative to robot	dissimilarity
amount of food remaining	movement of each pedestrian towards / away from / perpendicular to robot	entropy
whether Pacman has been eaten		correlation

The primary insight extracted from this follow-on study is that the only task people were not able to characterize features in a way closely aligned with the computationally validated baseline (*i.e.* firefighting classification) is the one where most of the features selected were difficult to interpret semantically. For example, *skewness of pixel intensities from thermal images of scene* for Firefighting task is more difficult to understand intuitively than *number of ghosts* for Pacman task or *location/orientation of robot* for Navigation task.

V. Discussion

Our overall findings are summarized in Table VI. It highlights the highest performing method(s) for selecting a useful feature subset as we vary two parameters: (1) amount of training data and (2) use of a human teacher.

HUMAN \ DATA	*Small*	*Large*
No	FI	ER
Yes	**HFS**	HIS, ER, HFS

TABLE VI: Experimental Findings

The *bottom row* is what we were interested in exploring in this work. The findings suggests that even without having yet seen any training examples, a robot learner can leverage the knowledge of a domain expert to identify a subset of features useful for constructing a more discriminative task representation. This supports our first hypothesis that humans *are* able to help solve the feature selection problem and are valuable for LfD domains (bottom left cell), where the agent learns from a teacher but training data provided is typically limited. Additionally, we have some insights about successful and unsuccessful ways to extract this feature information, when optimizing for classification accuracy.

In considering all five of the strategies we examined for eliciting feature information from humans teachers, two approaches emerged as most effective: human feature selection (HFS) and human instance selection (HIS). HFS outperformed all other approaches (both computational and human-driven) given a *small* amount of training data; both HIS and HFS were comparable to computational FS given a *large* amount of training data, but neither was able to significantly outperform the best computational FS approach (ER). Thus contrary to what we hypothesized, *direct communication* about features proved to be the most effective overall strategy for users. This was further validated in our follow-on study, providing the insight that users are able to select informative

features *given* that the features can be understood intuitively.

Other insights gleaned were HFS always dominated HFR, and the combined approaches never yielded the best performance, as compared to other human-driven approaches. In future work, we can explore why we observe these trends.

VI. Related Work

Feature Selection for robots and intelligent agents has been previously explored in the literature for several problem domains: mobile robot navigation [12], [13], [14], [10]; simulated autonomous car driving [15], [13], [16]; emotion state classification for a nursing robot [17]; robot soccer and multi-robot domains [18]; gas identification [19] and fire hazard classification [11] for search and rescue; and grasp classification [20], [21]. Nonetheless, most works have looked at enabling a robot to automatically select features using computational algorithms with no human in the loop.

There has also been work within robotics that looks at requesting feature information from a human teacher. Embodied feature queries were introduced by [22] for a robot learning a skill from demonstration by a human teacher. The work focuses on enabling a robot to generate three types of queries using its embodiment where the query types each aim to reduce uncertainty with respect to a different part of the skill learning problem. In contrast, our focus is to examine the efficacy of different natural language question types towards acquiring the same information: a useful set of features to sufficiently characterize the given task. Rosenthal *et.al.* recommend FS as one of the dimensions a robot should request information about when formulating a question, to provide sufficient context to a human [23], but do not explore *how* to elicit this information.

The most closely aligned work has explored the use of automatic feature selection for learning a task policy from human demonstrations [5]. The *abstraction by demonstration* algorithm enables an agent to infer relevant features for the human policy based upon demonstrations given. Though this work is similar in that it seeks to learn informative feature subsets through human provided examples, our work differs in two ways: (1) we seek to determine the most effective approach for eliciting feature information *from* human teachers and (2) we compare the efficacy human-driven FS approaches to computational FS approaches.

VII. Conclusion

Enabling robots to request the most useful features for characterizing a task is an important step toward autonomous task model construction. With only a small amount of data, computational feature selection approaches are limited in their ability to output the most useful features for discriminating between classes of objects needed for a task. Therefore, using computational feature selection as a baseline, this work explored: (1) whether a human teacher is able to characterize the most informative features of a classification task as accurately as computational approaches and (2) the best way to extract this feature information from the teacher. Our results suggest that a human teacher can *directly select* a subset of features that will be informative for discriminating between the task-relevant object classes given that the features are semantically interpretable. And in the case that the learning agent has either no or a small number of training examples, we can expect the subset selected by a human teacher to be more useful for classifying unseen task-relevant objects than that selected by a computational feature selection algorithm.

References

[1] S. Chernova and A. L. Thomaz, *Robot Learning from Human Demonstration*. Morgan and Claypool Publishers, 2014.

[2] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial intelligence*, vol. 97, no. 1, pp. 245–271, 1997.

[3] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1, pp. 273–324, 1997.

[4] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[5] L. C. Cobo, P. Zang, C. L. Isbell Jr, and A. L. Thomaz, "Automatic state abstraction from demonstration," in *Int. Joint Conference on Artificial Intelligence*, vol. 22, no. 1. Citeseer, 2011, p. 1243.

[6] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view rgb-d object dataset," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1817–1824.

[7] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[8] M. A. Hall, "Correlation-based feature subset selection for machine learning," Ph.D. dissertation, University of Waikato, Hamilton, New Zealand, 1998.

[9] U. Berkeley, *Intro to AI Project 3: Reinforcement Learning*, 2014 (accessed 01-March-2017). [Online]. Available: http://ai.berkeley.edu/reinforcement.html

[10] D. Vasquez, B. Okal, and K. O. Arras, "Inverse reinforcement learning algorithms and features for robot navigation in crowds: an experimental comparison," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*. IEEE, 2014, pp. 1341–1346.

[11] J.-H. Kim, S. Jo, and B. Y. Lattimer, "Feature selection for intelligent firefighting robot classification of fire, smoke, and thermal reflections using thermal infrared images," *Journal of Sensors*, vol. 2016, 2016.

[12] C. Diuk, L. Li, and B. R. Leffler, "The adaptive k-meteorologists problem and its application to structure learning and feature selection in reinforcement learning," in *International Conference on Machine Learning*. ACM, 2009, pp. 249–256.

[13] D. Floreano, T. Kato, D. Marocco, and E. Sauser, "Coevolution of active vision and feature selection," *Biological cybernetics*, vol. 90, no. 3, pp. 218–228, 2004.

[14] S. Zhang, L. Xie, and M. D. Adams, "Entropy based feature selection scheme for real time simultaneous localization and map building," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*. IEEE, 2005, pp. 1175–1180.

[15] S. Whiteson, P. Stone, K. O. Stanley, R. Miikkulainen, and N. Kohl, "Automatic feature selection in neuroevolution," in *Proceedings of the 7th annual conference on Genetic and evolutionary computation*. ACM, 2005, pp. 1225–1232.

[16] S. Loscalzo, R. Wright, and L. Yu, "Predictive feature selection for genetic policy search," *Autonomous Agents and Multi-Agent Systems*, vol. 29, no. 5, pp. 754–786, 2015.

[17] M. Swangnetr and D. B. Kaber, "Emotional state classification in patient–robot interaction using wavelet analysis and statistics-based feature selection," *IEEE Transactions on Human-Machine Systems*, vol. 43, no. 1, pp. 63–75, 2013.

[18] D. L. Vail and M. M. Veloso, "Feature selection for activity recognition in multi-robot domains." in *AAAI*, vol. 8, 2008, pp. 1415–1420.

[19] M. Trincavelli and A. Loutfi, "Feature selection for gas identification with a mobile robot," in *IEEE Int. Conf. on Robotics and Automation*. IEEE, 2010, pp. 2852–2857.

[20] L. Y. Chang, N. S. Pollard, T. M. Mitchell, and E. P. Xing, "Feature selection for grasp recognition from optical markers," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*. IEEE, 2007, pp. 2944–2950.

[21] A. Ramisa, G. Alenya, F. Moreno-Noguer, and C. Torras, "Using depth and appearance features for informed robot grasping of highly wrinkled clothes," in *Robotics and Automation (ICRA), 2012 IEEE International Conference on*. IEEE, 2012, pp. 1703–1708.

[22] M. Cakmak and A. L. Thomaz, "Designing robot learners that ask good questions," in *Proc. ACM/IEEE international conference on Human-Robot Interaction*. ACM, 2012, pp. 17–24.

[23] S. Rosenthal, A. K. Dey, and M. Veloso, "How robots' questions affect the accuracy of the human responses," in *IEEE Int. Symp. on Robot and Human Interactive Communication*. IEEE, 2009, pp. 1137–1142.