An Empirical Characterization of Parsimonious Intention Inference for Cognitive-level Imitation Learning

Garrett Katz [1,*], **Di-Wei Huang** [1], **Rodolphe Gentili** [2], **and James Reggia** [1,3]
{gkatz@cs., dwh@cs., rodolphe@, reggia@cs.}umd.edu
*Corresponding author
[1]Department of Computer Science
[2]Cognitive-Motor Neuroscience Laboratory, Department of Kinesiology
[3]Institute for Advanced Computer Studies
University of Maryland, College Park, MD, USA, 20742

Abstract— *Imitation learning is a promising route to better collaboration between humans and artificial agents. It will be most effective if the agent has some cognitive-level "understanding" of a human demonstrator's intentions. Inferring intent is an example of abductive reasoning, wherein an agent explains the available evidence based on causal knowledge. Good explanations should satisfy some notion of parsimony ("Occam's razor"), but the optimal notion of parsimony is often application-specific. We compare several such notions in the context of intention inference, using a robotic imitation learning scenario and the Monroe County Corpus, a standard benchmark in intention inference. Our results suggest that the most popular notions of parsimony in general are not necessarily appropriate for intention inference in particular.*

Keywords: Imitation Learning, Intention Inference, Parsimonious Covering Theory, Artificial Intelligence

Submission type: Regular Research Paper

1. Introduction

Recent decades have seen significant advances in imitation learning (IL), which refers to the acquisition of new skills by an intelligent agent through observation of human demonstrations [1]. We mean the term "intelligent agent" to include both physical cognitive robots and AI software systems. IL holds the promise of intelligent agents that can be taught by human domain experts and other end-users, who have little or no expertise in robotics or computer programming. IL also has the potential to make artificial collaborators more safe, trust-worthy, and usable, aiding and protecting human operators in difficult or dangerous situations, but at the same time keeping a responsible human in the loop to guide behavior. To make this technology effective and robust in uncertain or variable environments, the agent should have some cognitive-level "understanding" of a human demonstrator's intentions. By "intentions," we mean the specific high-level goals for actions that are taken. Understanding these intentions will help the agent achieve the same goals in different situations that require modified actions from what was observed.

Much past work on IL has focused on low-level sensorimotor learning in robots, where the objective is to closely mimick the precise motor trajectories and dynamics relevant to a given skill [2], [3], [4], [5], [6], [7], [8]. These methods are very important and have led to impressive results, but are typically limited to a single class of relatively simple low-level tasks not requiring interactions of sensorimotor processes with high-level planning, reasoning and control. As such, they are not human-competitive on complex tasks that require a sequence of structured actions. Further, the learned procedures are often quite brittle and do not generalize well to situations that are not in the training set, in part because most past IL systems have focused on copying demonstrated actions verbatim rather than trying to "understand" the goals and intentions of the human demonstrator [9].

More recently, some work has been done on IL at a higher cognitive level, but is typically restricted to very simple, simulated, and/or highly constrained task scenarios. There are two prominent branches of research in this area. One branch works within the framework of reinforcement learning, using environments with simple state representations such as cells in discrete two-dimensional grids, or low-dimensional real-valued vectors (e.g., a ⟨position, velocity⟩ pair describing motion along a single coordinate axis) [10], [11], [12], [13]. The other branch uses symbolic representations of goals with internal structure, and in some cases more complex world models, although the tasks are still confined to movement of simple objects on a flat two-dimensional surface [14], [15], [16]. There has also been some work on neural models of goal-directed IL, but focused on low-level goals such as the target position of a reaching movement [17].

This past year we demonstrated that cause-effect reasoning, also known as *abductive inference*, can facilitate cognitive-level IL, enabling generalization of observed procedural skills to new situations on the basis of just one human demonstration [18], [19]. In the first stage of our framework, the agent uses causal reasoning to infer a demonstrator's intentions/goals from their actions, thereby acquir-

ing new high-level procedural knowledge. Next, the agent uses automated planning to carry out the same intentions, but in new situations requiring new low-level actions, thereby enabling some degree of generalization. In this framework, causal intention inference is achieved through a novel extension of Parsimonious Covering Theory (PCT), a model of causal reasoning that has previously been applied in diverse fields such as medical diagnosis, circuit fault localization, natural language processing, and semantic web technology [20], [21], [22], [23]. Automated planning here is performed using a Hierachical Task Network (HTN), a formal model of how high-level intentions/goals can cause lower-level sub-intentions and ultimately observable actions [24]. As is typically the case with HTNs, our framework assumes that these causal relationships have been prespecified by a human domain expert. The key focus of our approach is not the *inductive* problem of learning HTN structures from experience, on which there is substantial past work, e.g. [25], [26] - it is the *abductive* problem of drawing on existing causal knowledge to form plausible explanations for what is observed in a particular instance.

Demonstrations in our framework are recorded using a virtual environment called SMILE, which is a second innovation of our work [27], [28]. In SMILE, objects can be "dragged and dropped," and the demonstrator is effectively an invisible presence. This methodology is pragmatic and user-friendly, and obviates the need for sophisticated human motion capture or mapping between human and robot embodiments. It also provides a realistic setting in which the causal inference algorithm has access to essentially perfect observability of a demonstrator's overt actions and the resulting changes in the external environment. SMILE objects are highly modular and customizable, which extends the reach of our approach beyond robotic control to domains such as emergency response planning and engineering design.

In any causal reasoning problem, there may be more than one valid explanation for the observed evidence. Some of these explanations are better than others, but the formal criterion that is most appropriate is often unknown a priori and potentially domain-specific. The HTN planning model is no exception, since depending on the current world state, the same high-level intentions/goals may cause different action sequences. Similarly, different high-level intentions/goals can potentially cause the same action sequence. This ambiguity leads to a plethora of valid explanations and makes the intention inference problem non-trivial, even when actions and world states are perfectly observable and causal relations are provided as background knowledge. It has long been recognized that when applying abductive inference systems to new problem domains, it is important to conduct a systematic comparison of various technical criteria of plausibility to determine which is optimal [22], [29]. However, in our existing work described above, the causal reasoning system was only tested using one of the simplest such criteria, without considering other more nuanced alternatives from the literature, or new previously unconsidered alternatives.

Here, we perform a systematic comparison of several alternative technical notions of plausibility, old and new, in the context of intention inference. We use PCT as our framework for these experiments, which casts plausibility in terms of *parsimony*, since by design it allows one to easily shift between different criteria for exactly what makes an explanation parsimonious. Our comparison uses two problem domains: the robotic IL domain that we devised in our previous work [18], [19], and the Monroe County Corpus, a benchmark in the automated planning literature that provides an extensive test suite for intention inference in the HTN formalism [30]. We find that the choice of criterion can have a significant impact on performance, but when the optimal criterion is employed, PCT is quite effective.[1] In particular, while there are good theoretical reasons for basing explanation construction primarily on a form of parsimony known as irredundancy [22], we find that in the case of intention inference, criteria other than irredundancy are almost as accurate and qualitatively more precise. Which criterion is optimal also depends on whether plans are always caused by a single top-level intention or might be caused by a top-level *sequence* of intentions. Most significantly, we find that a previously unconsidered criteria that we define below is competitive with and often superior to other previously considered criteria from the literature. All of the code needed to reproduce our results is freely available online.[2]

2. Background: What is Parsimony?

In this section we review the pertinent elements of our formal extension to PCT for intention inference. For full details of the inference algorithms, including theoretical completeness, soundness, and complexity results, we refer the reader to [19]. Our framework posits a set V of intentions and actions, and a *causal relation* $C \subseteq V \times V^*$, where V^* denotes the set of all finite sequences whose entries are elements of V. We write $\langle v \rangle$ to denote any arbitrary sequence in V^*. Each element $(u, \langle v \rangle) \in C$ indicates that u might directly cause the sequence $\langle v \rangle$. In practice, C is represented by a pre-specified knowledge base provided by a human domain expert. Each intention and action is represented by a *schema* that can be parameterized. For example, in the robot domain, the schema `(grasp ?object ?gripper)` (where '`?`' designates an unbound variable) might be grounded by `(grasp block1 left-gripper)`, and in the Monroe domain, `(get-to ?crew ?location)` might be grounded

[1] In PCT, the optimality of a particular explanation is based on a parsimony criterion. But the optimality of a criterion itself is not measured by how parsimonious it is (which would be circular) - it is measured with independent metrics such as accuracy (i.e., how often the parsimonious explanations, according to this criterion, match a known ground truth) or specificity (i.e. how many valid parsimonious explanations there are).

[2] At `www.github.com/garrettkatz/copct`

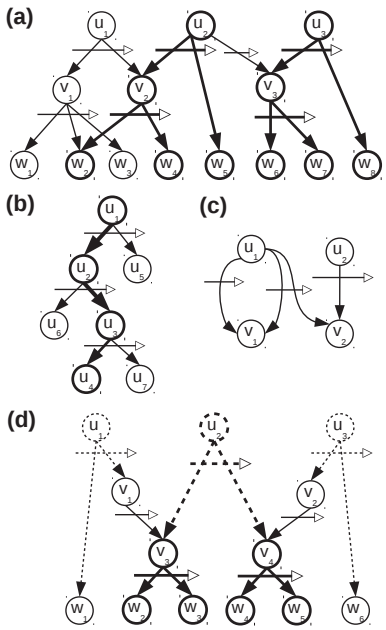

Fig. 1: Examples of causal relations for PCT. A causal link $(u, \langle v \rangle)$ is depicted with vertical edges from u to each entry of $\langle v \rangle$, and horizontal arrows across the edges to each entry in $\langle v \rangle$ to signify the ordering constraint. The meaning of other notation is described in the text.

by `(get-to wcrew1 pittsford-plaza)`. Every intention $u \in V$ is a grounded schema such as these. Causal events in C can be chained together: for example, u might cause $\langle v_1, v_2 \rangle$, and v_1 might cause $\langle w_1, w_2, w_3 \rangle$, and so on, until observable actions are reached. Figure 1 (a)-(d) shows several graphical depictions of causal relations.

Given any ordered forest whose nodes are all in V, and whose ordered parent-child connections are all in C, the ordered root sequence $\langle u \rangle \in V^*$ is called a *cover* of the ordered leaf sequence $\langle w \rangle \in V^*$. For example, in Figure 1(a), $\langle u_2, u_3 \rangle$ is a cover of $\langle w_2, w_4, w_5, w_6, w_7, w_8 \rangle$, with the covering forest shown in bold. $\langle w \rangle$ represents a sequence of observed actions, and $\langle u \rangle$ represents a sequence of hidden intentions that can explain those actions. The internal tree nodes represent intermediate sub-intentions along the causal chains from roots to leaves. A cover $\langle u \rangle$ is *top-level* if it cannot be covered itself by any other higher-level intention sequence. Any top-level cover $\langle u \rangle$ of $\langle w \rangle$ is considered to be a valid *explanation* of $\langle w \rangle$.

In relation to HTN planning, each intention u corresponds to a task, and each causal relationship $(u, \langle v \rangle)$ corresponds to one step from a task to its immediate sub-tasks. A cover corresponds to a top-level task sequence which, if provided as input to the HTN planner, could have resulted in the planned sequence of low-level actions that is actually observed. Multiple $\langle v \rangle$ can be caused by the same u, corresponding to different search paths available for the HTN planner to explore. Multiple u can also cause the same $\langle v \rangle$, corresponding to different tasks that can cause the same sub-tasks, leading to ambiguity in the covering problem.

In both test domains we consider here, this ambiguity is prevalent and results in many valid explanations for a given $\langle w \rangle$. In PCT, this is mitigated by the notion of *parsimony*: The most parsimonious explanations should be considered most plausible; the rest should be discarded. Various technical notions of parsimony have been studied. Theoretically these criteria could be applied to any cover, although in practice, we always limit our attention to the set of *top-level* covers (i.e., valid explanations) in particular, and prune that set using one of these criteria as an additional filter. We compare the following criteria:

- *minimum cardinality* (MC): The *cardinality* of a cover $\langle u \rangle$ is simply the number of elements in the cover.
- *irredundancy* (IR): A cover $\langle u \rangle$ of $\langle w \rangle$ is *irredundant* if no proper subsequence of $\langle u \rangle$ is also a cover of $\langle w \rangle$.
- *maximum depth* (MD): A *causal chain* is a path from a root to a leaf (e.g., the bold-faced path in Figure 1(b)). The covers with the deepest causal chains are considered most parsimonious. The idea is that the roots of deeper chains encode more information (i.e., a larger covering tree) per root node.
- *minimax depth* (XD): For each cover we can measure its shallowest causal chain. Then we can compare this to the shallowest causal chains of other covers. The covers whose shallowest causal chains are deepest overall are considered most parsimonious.
- *minimum parameters* (MP): For each cover we can count the number of distinct parameter values that occur. The covers with the fewest distinct parameter values are considered most parsimonious. For example, in the Monroe County Corpus, the cover

  ⟨`(get-to wcrew1 mendon)`, `(clear-road-wreck hamlin rochester)`⟩

  has four distinct parameter values (`wcrew1`, `mendon`, `hamlin`, and `rochester`), whereas

  ⟨`(get-to wcrew1 mendon)`, `(clear-road-wreck mendon rochester)`⟩

  only has three (`wcrew1`, `mendon`, and `rochester`), so the latter is more parsimonious. This criterion favors more cohesive explanations, as can be seen in these examples: getting the work crew to `mendon` is related to clearing a road wreck near `mendon`, whereas it would be unrelated to clearing a road wreck near `hamlin`.
- *maximum forest size* (FSX): For each cover, we count all nodes in the covering forest. The idea is that covers with the maximal node count encode more information (i.e., more non-root nodes) in the same number of roots.
- *minimum forest size* (FSN): Instead of maximal forest size, minimal forest size is considered the most parsi-

monious, in the sense that the total number of causal links contributing to the explanation is smallest.

Minimum cardinality and irredundancy have been widely used in past abductive AI systems. Some automated planning research has considered minimum forest size, but only as related to planning from high-level tasks *to* low-level actions - not in the opposite direction as we do here [31]. To our knowledge, the other criteria - in particular, minimum parameters - are new and explored here for the first time. Regardless of which criterion is adopted, PCT may still return more than one valid, parsimonious explanation for a given plan, rather than a single "optimal" interpretation, so an empirical comparison is needed.

3. Experimental Methods

3.1 Testing Data and Performance Metrics

For our parsimony comparison we used two domains: our own robotic IL domain [18], [19], and the 3^{rd}-party Monroe County Corpus, a well-known benchmark for intention/goal inference (also known as "plan recognition" in the automated planning literature) [30]. The robotic domain models a tabletop workspace requiring bimanual manipulation of compound object assemblies. Demonstrated skills involve stacking toy blocks to form letters of the English alphabet and maintaining a hard-drive docking station subject to hardware faults. This domain includes observable actions such as grasping and releasing objects, and higher level intentions/goals such as opening drawers, toggling switches, and handing objects between grippers. Causal interpretation of observed demonstrations produces novel top-level intention sequences that were not pre-specified in the knowledge base, constituting newly learned skills. A modest set of test data contains 11 examples of observed demonstrations: 8 for various dock maintenance skills, and 3 for stacking various block configurations. The Monroe domain models an emergency response team based in Monroe County in upstate New York. For example, the intention to clear a car wreck might cause a sequence of sub-intentions such as getting patients into an ambulance and getting the ambulance to a hospital. The sub-intention of getting patients into an ambulance might cause its own sequence, such as getting an EMT into the ambulance and driving the ambulance to the scene. The corpus contains a knowledge base defining all of these causal relationships, and 5000 automatically generated plans (i.e., sequences of observed actions).

In both datasets, every observed plan is annotated with the ground truth top-level intentions from which it was generated, which can be used for quantitative comparison. To evaluate a given criterion, we withheld the ground truth intentions from PCT, invoked PCT on the observed actions, and then compared the parsimonious covers found by PCT with the ground-truth. We used two performance metrics to evaluate each criterion. First, we consider its *accuracy*: how often do the "parsimonious" covers of an observed plan, according to that criterion, include the ground truth explanation? Second, we consider its *specificity*: how many parsimonious covers are found in all? Many more may be found than just the single ground truth. A perfectly accurate and specific criterion would compute a single top-level cover, namely the ground truth one, and no others.

For each criterion, PCT was invoked on every observed plan in each dataset, and the performance metrics for that criterion were calculated. In the large majority of cases, an individual plan was processed in a matter of seconds, although on 162 (3.2%) of the 5000 plans in the Monroe corpus, the PCT interpretation algorithm was still running after 10 minutes and was terminated early. These plans were excluded from computation of the performance metrics.

3.2 Learning Novel Top-Level Intention Sequences in the Monroe Domain

In the original Monroe corpus, each plan was generated from a ground truth consisting of just one top-level intention/goal. In other words, the ground truth "sequence" $\langle u \rangle$ covering a given plan only has a single entry u_1. This has limited relevance to real world scenarios where an agent may have multiple top-level intentions/goals. Particularly, in IL, it is preferable if the agent can learn novel *sequences* of intentions from demonstrations, so that the useful sequences do not all have to be anticipated and hand-coded beforehand in an exhaustive knowledge base. In other words, intention inference should be a *constructive* process that leads to structured explanations, not a pattern classification task.

To provide more challenging problems along these lines, we prepared a modified variant of the Monroe domain by stripping the top-most singleton intentions from the knowledge base. This transformation is illustrated in Figure 1 (d). The dashed nodes and edges depict top-level causes (u_1, u_2, and u_3) defined in the original knowledge base but removed in the modified one. The bold-faced nodes and edges represent a covering tree for a particular observed plan ($\langle w_2, w_3, w_4, w_5 \rangle$). Whereas the singleton u_2 was a top-level cover for this plan using the original causal relation, when using the modified causal relation, this plan can only be covered by a top-level *sequence* (namely, $\langle v_1, v_2 \rangle$). All performance assessments described above were repeated on this more challenging modified corpus.

However, since the original ground-truth covers have been removed from the causal relation, this methodology raises the question of what should be considered the new "ground-truth" covers against which accuracy can be measured. Since full plan trees are provided for each plan in the original corpus, a default "ground-truth" could be the child sequence immediately below the original root. In Figure 1 (d), this would be the sequence $\langle v_2, v_3 \rangle$. Unfortunately, this new "ground-truth" is not guaranteed to be top-level in the modified causal relation. Figure 1 (d) is a counter-example:

$\langle v_2, v_3 \rangle$ has at least one higher-level cover ($\langle v_1, v_2 \rangle$), and in larger examples, it may have many. In cases like this, which were found to be quite common, there is no single incontrovertible ground truth against which to measure accuracy. Consequently, the experiments on the modified dataset were restricted to the plans where this issue did not arise.

4. Results

In the robotic domain, the parsimonious covers always included the ground truth (100% accuracy), except for the forest size criteria: FSN was only 45% accurate, and FSX was in fact 0% accurate.[3] On the other hand, several criteria were quite nonspecific, returning a large number of parsimonious covers in addition to the ground truth. Table 1 details these specificity results. Each column shows the total number of top-level covers retained after pruning by the respective criterion (abbreviations are as defined above). The 11 plans are ordered according to increasing length (i.e., the number of actions in the observed sequence). It was found that only MC and MP remain highly specific when the plan length and total number of top-level covers becomes large. FSN and FSX are omitted because they were so inaccurate.

Table 1: Number of covers found by each criterion on each demonstration in the robotic domain.

Demo	MC	IR	MD	XD	MP
Dock demo 1	2	4	2	4	1
Dock demo 2	2	8	4	8	1
Dock demo 3	2	8	4	8	1
Dock demo 4	1	2	2	2	1
Dock demo 5	2	4	2	4	1
Dock demo 6	2	8	4	8	1
Dock demo 7	2	8	4	8	1
Dock demo 8	1	2	2	2	1
Block demo 1	1	256	256	256	1
Block demo 2	1	1024	1024	1024	1
Block demo 3	1	8192	8192	8192	1

For the original Monroe domain (without top-level intentions stripped from the knowledge base), accuracy of each criterion is shown in Table 2 (center column). In addition to the 162 timeouts, there were a small number of plans (42) in which the top-level covers found by PCT failed to include the ground-truth, even before filtering by any criteria. However, we previously provided evidence that in these examples, the error might in fact be in the corpus itself rather than PCT.[4] Accuracy measurements were restricted to the remaining 4796 plans in the corpus.

[3]This is because the knowledge base includes a top-level `(get-to obj1 obj2)` intention, which accepts grippers as final destinations. So any top-level cover containing `(get-to obj1 obj2)` remains valid when the occurance is replaced with `(get-to obj1 gripper) (get-to gripper obj2)`. The ground truths generally conformed to the former possibility, while the FSX covers conformed to the latter possibility, since there were more nodes in the covering forest.

[4]See `www.cs.umd.edu/~reggia/supplement/index.html`

For most criteria, the set of all top-level covers found for a given plan can be filtered in linear time. The extremal value (minimum cardinality, maximum depth, etc.) can be found in one pass through the covers, and then the most parsimonious ones can be extracted in a second pass. However, filtering by irredundancy is more subtle. We performed irredundancy pruning based on the following proposition: Given two top-level covers t_1 and t_2, if t_1 is a sub-sequence of t_2, then t_2 is redundant. This idea can be used to filter out redundant top-level covers in quadratic time, which proved impractical in some cases. Using another time-out of 5 minutes, the irredundancy filter ran to completion on 3397 of these, so irredundancy metrics are computed relative to this subset.

Table 2: Accuracies on the original and modified corpus.

Criterion	Accuracy (original)	Accuracy (modified)
MC	4796 of 4796 (100.0%)	2859 of 2861 (99.9%)
IR	3397 of 3397 (100.0%)	2470 of 2470 (100.0%)
MD	4796 of 4796 (100.0%)	2856 of 2861 (99.8%)
XD	4796 of 4796 (100.0%)	2861 of 2861 (100.0%)
MP	4464 of 4796 (93.1%)	2724 of 2861 (95.2%)
FSN	1105 of 4796 (23.0%)	634 of 2861 (22.2%)
FSX	2087 of 4796 (43.5%)	2363 of 2861 (82.6%)

We found that minimum cardinality, irredundancy, and depth-based criteria were all perfectly accurate: if the ground truth was included in the full set of top-level covers for a plan, it was retained after filtering by any of these criteria. The minimum parameters criterion was good but not perfect, while forest size-based criteria again performed poorly. These results are fairly unsurprising, given that every ground-truth is a singleton sequence, as described earlier.

More interesting is the comparison of specificity for each criterion. For each plan in the corpus, we recorded the number of top-level covers that were found after filtering by each criterion. Then, for each criterion, we generated a histogram of these counts across all plans in the corpus. These histograms are shown together in Figure 2. Most criteria could be grossly nonspecific in the worst case, retaining thousands of covers or more. This could be attributed to a combinatorial explosion in the number of possible parameterizations of the various intention schema. The Monroe knowledge base is designed in such a way that many different parameterizations of a parent intention can cause the same child intentions. For example, the intention `(clear-road-hazard ?from ?to)` causes the action `(call fema)`. There is a quadratic number of possible assignments of `?from` and `?to` to locations in Monroe County, which would all qualify as valid covers for `(call fema)`. When this occurs multiple times in a sequence, the number of valid covers at the next layer up grows exponentially. In contrast, the minimum cardinality criterion was quite effective at mitigating this issue. In 3898 of the 4838 plans that did not time out (80.6%), it was maximally specific (the ground-truth interpretation was the sole minimum cardinality cover); in 4354 plans (90%), there

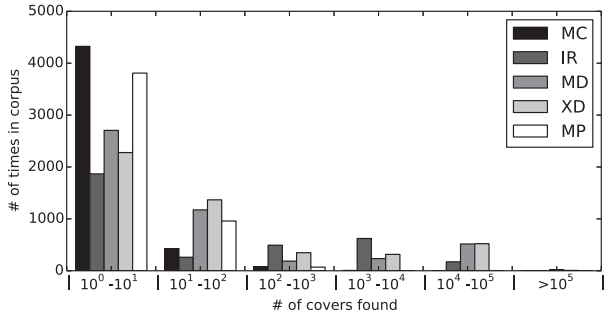

Fig. 2: Histograms showing the number of parsimonious covers found for plans in the original corpus, after filtering by each criterion (see text for details).

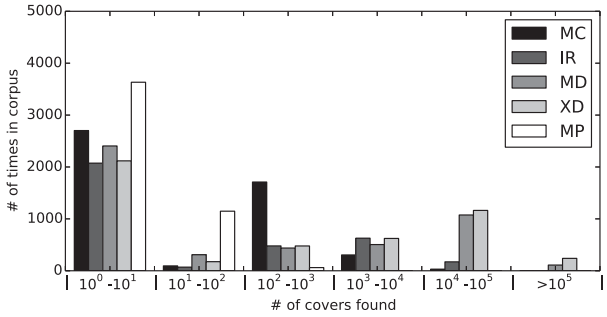

Fig. 3: Histograms showing the number of parsimonious covers found for plans in the modified corpus, after filtering by each criterion (see text for details).

were at most twelve minimum cardinality covers.

Finally, we report performance of PCT on the modified Monroe dataset, in which the top-most intentions were stripped from the knowledge base. As described above, the issue arises in many plans that no unambiguous ground truth is available. Specifically, this was found to occur in 2139 of the 5000 modified plans, which were thus excluded from the accuracy comparison. Specificity was still measured on all 4842 modified plans that did not time out.

Accuracy on the remaining 2861 plans is shown in Table 2 (right column). As in the original corpus, all criteria except those based on forest size are highly accurate. The IR filter again timed out on some examples, so metrics for IR were only calculated on the 2470 plans where it ran to completion.

Analogous to Figure 2, specificity histograms on the modified corpus are shown in Figure 3. In contrast with the original corpus, it was found that like most other criteria, minimum cardinality becomes much less specific when inferring intention sequences, with counts over 100 covers for 42.3% of the plans, and some counts in the thousands in the worst cases, although this still compared favorably with IR. On the other hand, MP proved quite specific on the modified corpus: it was maximally specific in 2756 of 4842 plans (56.9%), and in 90% of the plans, there were at most 16 MP covers found. As can be seen in Table 2, the improved specificity did come with some cost to accuracy.

5. Discussion

The PCT approach to intention inference comes with strong formal guarantees of soundness, concreteness, and complexity [19], but at the cost of assuming detailed background knowledge of the demonstrator, and perfect observability of their low-level actions. However, meeting the perfect observability requirement becomes entirely realistic in practice once a virtual demonstration environment such as SMILE is adopted [27]. Moreover, no more background knowledge is required than already needed by HTNs and other similar hierarchical planning formalisms.

Here we have presented a systematic comparison of several parsimony criteria relevant to this approach, using multiple data sets for testing. In so doing, we have augmented the original Monroe domain to test plan recognition when observed actions can only be explained by a *sequence* of top-level intentions/goals, rather than a single top-level intention. This situation may be more relevant for certain applications, such as IL. Our results suggest that in the context of intention inference, irredundancy is very accurate, but perhaps only because it is so nonspecific. When plans can always be explained by a single top-level intention/goal, minimum cardinality is equally accurate and qualitatively more specific. In the modified Monroe domian, minimum cardinality remains highly accurate, but becomes too nonspecific for practical use in about one fourth of the test cases. Since our newly proposed Minimum Parameters criterion remains highly specific even in these cases, it may sometimes be a preferable criterion, in spite of the 5% accuracy reduction. We surmise that minimizing the number of distinct parameter values in a sequence can mitigate the combinatorial explosion described earlier, since covers with unrelated parameters in the constituent intention schemas will be filtered out. These results are in contrast with other work on abductive inference outside the context of intention inference, where irredundancy is theoretically the gold standard, although there have been other applications where it also produces too many covers to be useful. For example, minimum cardinality was also found to be a preferable criterion to irredundancy in a medical diagnosis application [29].

Even the best criteria identified in this work still suffer from some nonspecificity in the worst case. Finding better criteria for intention inference is an important future research direction. One possibility is to combine the promising criteria - for example, prune first by MP, and then by MC. The incorporation of probabilistic methodology is another promising strategy. In fact, PCT has been extended to incorporate probability theory in the past, and a theoretical analysis led to conclusions about whether a given criterion

would tend to include the most likely explanation [22]. For example, it was shown that the minimum cardinality covers would tend to include the most likely explanation as long as **(a)** the prior probabilities of the possible causes were small and roughly equal, and **(b)** the probabilities that any given u causes any given $\langle v \rangle$ are relatively large. One might argue that in intention inference, a large number of possible parameterizations makes the probability of any particular grounded intention rather low, satisfying **(a)**, and that if any particular grounded intention u causes one of at most a small handful of possible sub-intention sequences $\langle v \rangle$, the causal probability of each is relatively high, satisfying **(b)**. On the other hand, the probabilistic causal model developed for PCT only treated the special case where the causal relation is bipartite (i.e., all causal chains have length 1), and there are no ordering constraints, so it is unclear to what extent those results carry over to intention inference.

At any rate, evaluating probability-based criteria requires that the prior and causal probabilities are already known before the likelihood of any cover can be computed. Although the Monroe corpus does provide prior probabilities for *ungrounded* top-level intention schemas, the parameter distributions and causal probabilities appear difficult to ascertain, in this and possibly other planning domains. Fortunately, much work on HTN planning and plan recognition has produced methods for the *inductive* problems of learning causal relations, and estimating their probability distributions, on the basis of large training sets, in contrast to the *abductive* problem we consider here, namely, constructing an explanation for a particular problem instance when the causal relation is already fully known. Future work should leverage these inductive methods in combination with the probabilistic causal model of PCT, suitably extended to handle causal chaining and ordering constraints.

Lastly, our results may be sensitive to the particular HTN encoding, as authored by a human, of the given domains. Different encodings of the same domains might produce different results. Future work should include experiments that characterize and quantify this sensitivity.

Acknowledgments

Supported by ONR award N000141310597 and by a grant from Lockheed Martin Corp.

References

[1] A. G. Billard, S. Calinon, R. Dillmann, "Learning from humans," *Springer Handbook of Robotics*. Springer, 2016, 1995–2014.

[2] B. Argall, B. Browning, M. Veloso, "Learning mobile robot motion control from demonstrated primitives and human feedback," *Robotics Research*. Springer, 2011, 417–432.

[3] T. DeVautl, S. Forrest, I. Tanimoto, T. Soule, R. Heckendorn, "Learning from demonstration for distributed, encapsulated evolution of auton. outdoor robots," *Proc. of the Comp. Pub. of the Annl. Conf. on Genetic and Evolutionary Computation*. ACM, 2015, 1381–1382.

[4] A. Billard D. Grollman, "Imitation learning in robots," *Encyclopedia of the Sciences of Learning*. Springer, 2012, 1494–1496.

[5] T. Fitzgerald, A. K. Goel, A. L. Thomaz, "Representing skill demonstrations for adaptation and transfer," *AAAI Symposium on Knowledge, Skill, and Behavior Transfer in Auton. Robots*, 2014.

[6] Y. Wu, Y. Su, Y. Demiris, "A morphable template framework for robot learning by demonstration," *RAS*, vol. 62, no. 10, 1517–1530, 2014.

[7] P. Abbeel, A. Coates, A. Y. Ng, "Autonomous helicopter aerobatics through apprenticeship learning," *The Intl. Jrnl. of Robotics Research*, vol. 29, no. 13, 1608–1639, 2010.

[8] J. J. O. Barros, V. M. F. dos Santos, F. M. T. P. da Silva, "Bimanual haptics for humanoid robot teleoperation using ROS and V-REP," *Intl. Conf. on Auton. Robot Sys. and Competitions*. IEEE, 2015, 174–179.

[9] A. Chella, H. Dindo, I. Infantino, "A cognitive framework for imitation learning," *Robotics and Auton. Systems*, vol. 54, no. 5, 403–408, 2006.

[10] D. Verma R. P. Rao, "Imitation learning using graphical models," *European Conf. on Machine Learning*. Springer, 2007, 757–764.

[11] A. L. Friesen R. P. Rao, "Imitation learning with hierarchical actions," *Intl. Conf. on Devel. and Learning*. IEEE, 2010, 263–268.

[12] M. J.-Y. Chung, A. L. Friesen, D. Fox, A. N. Meltzoff, R. P. Rao, "A Bayesian developmental approach to robotic goal-based imitation learning," *PloS One*, vol. 10, no. 11, p. e0141965, 2015.

[13] J. MacGlashan M. L. Littman, "Between imitation and intention learning." *IJCAI*, 2015, 3692–3698.

[14] A. Chella, H. Dindo, I. Infantino, "Learning high-level tasks through imitation," *Intl. Conf. on Intelligent Robots and Sys.* IEEE, 2006, 3648–3654.

[15] B. Jansen T. Belpaeme, "A computational model of intention reading in imitation," *Robotics and Auton. Sys.*, vol. 54, no. 5, 394–402, 2006.

[16] H. Dindo, A. Chella, G. La Tona, M. Vitali, E. Nivel, K. R. Thórisson, "Learning problem solving skills from demonstration," *Intl. Conf. on AGI*. Springer, 2011, 194–203.

[17] E. Oztop, D. Wolpert, M. Kawato, "Mental state inference using visual control parameters," *Cog. Brain Research*, vol. 22, no. 2, 129–151, 2005.

[18] G. Katz, D.-W. Huang, R. Gentili, J. Reggia, "Imitation learning as cause-effect reasoning," *Intl. Conf. on AGI*. Springer, 2016, 64–73.

[19] G. Katz, D.-W. Huang, T. Huage, R. Gentili, J. Reggia, "A novel parsimonious cause-effect reasoning algorithm for robot imitation and plan recognition," *IEEE Trans. on Cog. and Devel. Sys.*, 2017.

[20] V. R. Dasigi J. A. Reggia, "Parsimonious covering as a method for natural language interfaces to expert systems," *AI in Medicine*, vol. 1, no. 1, 49–60, 1989.

[21] C. Henson, A. Sheth, K. Thirunarayan, "Semantic perception," *Internet Computing*, vol. 16, no. 2, 26–34, 2012.

[22] Y. Peng J. A. Reggia, *Abductive inference models for diagnostic problem-solving*. Springer Science & Business Media, 1990.

[23] F. Wen C. Chang, "Probabilistic approach for fault-section estimation in power systems based on a refined genetic algorithm," *IEEE Proceedings-Generation, Transmission and Distribution*, vol. 144, no. 2, 160–168, 1997.

[24] M. Ghallab, D. Nau, P. Traverso, *Automated Planning: Theory & Practice*. Elsevier, 2004.

[25] Q. Yang, R. Pan, S. J. Pan, "Learning recursive HTN-method structures for planning," *Proc. of the ICAPS Workshop on AI Planning and Learning*, 2007.

[26] C. Hogg, U. Kuter, H. Munoz-Avila, "Learning methods to generate good plans." *AAAI*, 2010.

[27] D.-W. Huang, G. E. Katz, J. D. Langsfeld, H. Oh, R. J. Gentili, J. A. Reggia, "An object-centric paradigm for robot programming by demonstration," *Intl. Conf. on Augmented Cog.* Springer, 2015, 745–756.

[28] D.-W. Huang, G. Katz, J. Langsfeld, R. Gentili, J. Reggia, "A virtual demonstrator environment for robot imitation learning," *Intl. Conf. on Technologies for Practical Robot Applications*. IEEE, 2015, 1–6.

[29] S. Tuhrim, J. Reggia, S. Goodall, "An experimental study of criteria for hypothesis plausibility," *Jrnl. of Experimental & Theoretical AI*, vol. 3, no. 2, 129–144, 1991.

[30] N. Blaylock J. Allen, "Generating artificial corpora for plan recognition," *Intl. Conf. on User Modeling*. Springer, 2005, 179–188.

[31] R. Tsuneto, K. Erol, J. Hendler, D. Nau, "Commitment strategies in hierarchical task network planning," *NCAI*, 1996, 536–542.