Landmark-Enhanced Heuristics for Goal Recognition in Incomplete Domain Models

Ramon Fraga Pereira[*] and **André Grahl Pereira**[**] and **Felipe Meneguzzi**[*]

[*] Pontifical Catholic University of Rio Grande do Sul (PUCRS), Brazil
ramon.pereira@acad.pucrs.br
felipe.meneguzzi@pucrs.br

[**] Federal University of Rio Grande do Sul (UFRGS), Brazil
agpereira@inf.ufrgs.br

Abstract

Recent approaches to goal recognition have progressively relaxed the assumptions about the amount and correctness of domain knowledge and available observations, yielding accurate and efficient algorithms. These approaches, however, assume completeness and correctness of the domain theory against which their algorithms match observations: this is too strong for most real-world domains. In this paper, we develop goal recognition techniques that are capable of recognizing goals using *incomplete* domain theories by considering different notions of planning landmarks in such domains. We evaluate the resulting techniques empirically in a large dataset of incomplete domains, and perform an ablation study to understand their effect on recognition performance.

1 Introduction

Goal recognition is the problem of identifying the correct intended goal of an observed agent, given a sequence of observations as evidence of its behavior in an environment and a domain model describing how the observed agent generates such behavior. Approaches to solve this problem vary on the amount of domain knowledge contained in the behavior model assumed to be used by the observed agent (Sukthankar et al. 2014), as well as the level of observability and noise in the observations used as evidence (Sohrabi, Riabov, and Udrea 2016). While recent research has progressively relaxed the assumptions about the accuracy and amount of information available in observations required to recognize goals (E.-Martín, R.-Moreno, and Smith 2015; Sohrabi, Riabov, and Udrea 2016; Pereira, Oren, and Meneguzzi 2017), they all assume a complete and accurate model of the agent under observation. Such a strong assumption restricts goal recognition to applications where the requirements of completeness and correctness can be met by high-quality domain engineers. Ideally, we would like to be able to use imperfect domain models, either because the domain engineer made mistakes or because a learning algorithm generated the domain from noisy data.

Real-world domains often have two potential sources of uncertainty: (1) ambiguity in domain engineering either because of a noisy domain acquisition process or the nature of the actions being modeled; and (2) ambiguity from how imperfect sensor data reports features of the environment. The former stems from a possibly incomplete understanding of the actions being modeled, but more importantly, the inherently noisy and imperfect way in which automated domain acquisition through machine learning algorithms (Asai and Fukunaga 2018; Amado et al. 2018) we envision being the main source of real-world domain models. The latter stems from the potential unreliability in the interpretation of actions using real-world noisy data with learned sensor models being used to classify objects to be used as features (*e.g.*, logical facts) of the observations (Granada et al. 2017), so it is useful to model a domain with such feature as optional.

We develop heuristic approaches to goal recognition that use an enhanced notion of landmarks to cope with incomplete planning domain models (Weber and Bryce 2011; Nguyen, Sreedharan, and Kambhampati 2017) and provide five key contributions. First, we formalize goal recognition in incomplete domains (Section 2) combining the standard formalization of Ramírez and Geffner (2009; 2010) for plan recognition and that of Nguyen, Sreedharan, and Kambhampati (2017) for planning in incomplete domains. Second, we adapt an algorithm from (Hoffmann, Porteous, and Sebastia 2004) to extract *definite* and *possible* landmarks in incomplete domain models (Section 3). Third, we develop a notion of *overlooked* landmarks that we can extract online as we process (*on the fly*) observations that we can use to match candidate goals to the multitude of models induced by incomplete domains. Fourth, we develop two heuristics that account for the various types of landmarks as evidence in the observations to efficiently recognize goals avoiding running expensive planners for incomplete domains (Section 4). Finally, we build a new dataset for goal recognition in incomplete domains based on an existing one (Pereira and Meneguzzi 2017) by removing varying amounts of information from complete domain models and annotating them with possible preconditions and effects that account for uncertain and possibly wrong information (Section 5). We use this dataset to compare our approaches with a baseline (Pereira, Oren, and Meneguzzi 2017) and evaluate them through an ablation study over the various types of landmarks showing that *overlooked* landmarks become increasingly important to the accuracy of the heuristics as domain incompleteness increases.

2 Problem Formulation

2.1 Incomplete STRIPS Domain Models

To formalize incomplete domain models, we adapt the formalism of incomplete domain models from Nguyen, Sreedharan, and Kambhampati (2017), defined as $\widetilde{\mathcal{D}} = \langle \mathcal{R}, \widetilde{\mathcal{O}} \rangle$. Here, $\mathcal{R}$ is a set of predicates with typed variables, and $\widetilde{\mathcal{O}}$ is the set of definitions of incomplete operators, each of which comprised of a six-tuple $\widetilde{op} = \langle pre(\widetilde{op}), \widetilde{pre}(\widetilde{op}), \mathit{eff}^+(\widetilde{op}), \mathit{eff}^-(\widetilde{op}), \widetilde{\mathit{eff}}^+(\widetilde{op}), \widetilde{\mathit{eff}}^-(\widetilde{op}) \rangle$, where: $pre(\widetilde{op})$ and $\mathit{eff}(\widetilde{op})$ have the same semantics as in the STRIPS (Fikes and Nilsson 1971) domain models; and *possible* preconditions $\widetilde{pre}(\widetilde{op}) \subseteq \mathcal{R}$ that *might* be required as preconditions, as well as $\widetilde{\mathit{eff}}^+(\widetilde{op}) \subseteq \mathcal{R}$ and $\widetilde{\mathit{eff}}^-(\widetilde{op}) \subseteq \mathcal{R}$ that *might* be generated as *possible* effects respectively as add or delete effects. An incomplete domain $\widetilde{\mathcal{D}}$ has a *completion set* $\langle\langle \widetilde{\mathcal{D}} \rangle\rangle$ comprising all possible domain models derivable from the incomplete one. There are 2^K possible such models where $K = \sum_{\widetilde{op} \in \widetilde{\mathcal{O}}} (|\widetilde{pre}(\widetilde{op})| + |\widetilde{\mathit{eff}}^+(\widetilde{op})| + |\widetilde{\mathit{eff}}^-(\widetilde{op})|)$, and a single (unknown) ground-truth model $\mathcal{D}^*$ that actually generates the observed state. An incomplete planning problem derived from an incomplete domain $\widetilde{\mathcal{D}}$ and a set of typed objects Z is defined as $\widetilde{\mathcal{P}} = \langle \mathcal{F}, \widetilde{\mathcal{A}}, \mathcal{I}, G \rangle$, where: $\mathcal{F}$ is the set of facts (instantiated predicates from Z), $\widetilde{\mathcal{A}}$ is the set of incomplete instantiated actions from $\widetilde{\mathcal{O}}$ with objects from Z, $\mathcal{I} \subseteq \mathcal{F}$ is the initial state, and $G \subseteq \mathcal{F}$ is the goal state. Example 1 from Weber and Bryce (2011) illustrates an abstract incomplete domain and problem.

Example 1 *Let $\widetilde{\mathcal{P}}$ be an incomplete planning problem, where: $\mathcal{F} = \{p, q, r, g\}$; $\widetilde{\mathcal{A}} = \{\widetilde{a}, \widetilde{b}, \widetilde{c}\}$, where:*

- $pre(\widetilde{a}) = \{p, q\}, \widetilde{pre}(\widetilde{a}) = \{r\}, \widetilde{\mathit{eff}}^+(\widetilde{a}) = \{r\}, \widetilde{\mathit{eff}}^-(\widetilde{a}) = \{p\}$
- $pre(\widetilde{b}) = \{p\}, \mathit{eff}^+(\widetilde{b}) = \{r\}, \mathit{eff}^-(\widetilde{b}) = \{p\}, \widetilde{\mathit{eff}}^-(\widetilde{b}) = \{q\}$
- $pre(\widetilde{c}) = \{r\}, \widetilde{pre}(\widetilde{c}) = \{q\}, \mathit{eff}^+(\widetilde{c}) = \{g\}$
- $\mathcal{I} = \{p, q\}$; *and* $G = \{g\}$.

2.2 Goal Recognition in Incomplete Domains

Goal recognition is the task of recognizing an agents' goals by observing their interactions in an environment. Whereas most planning-based goal recognition approaches assume complete domain model (Ramírez and Geffner 2009; 2010; E.-Martín, R.-Moreno, and Smith 2015; Sohrabi, Riabov, and Udrea 2016; Pereira and Meneguzzi 2016; Pereira, Oren, and Meneguzzi 2017), we assume that the *observer* has an *incomplete* domain model while the *observed agent* is planning and acting with a *complete* domain model. To account for such uncertainty and incompleteness, the domain model available to the observer contains possible preconditions and effects (as defined in Section 2.1). Definition 1 formalizes goal recognition over incomplete domain models.

Definition 1 (Goal Recognition Problem) *A goal recognition problem with an incomplete domain model is a quintuple $\widetilde{\mathcal{T}} = \langle \widetilde{\mathcal{D}}, Z, \mathcal{I}, \mathcal{G}, Obs \rangle$, where: $\widetilde{\mathcal{D}} = \langle \mathcal{R}, \widetilde{\mathcal{O}} \rangle$ is an incomplete domain model (with possible preconditions and effects); Z is the set of typed objects in the environment, in which $\mathcal{F}$ is the set of instantiated predicates from Z, and $\widetilde{\mathcal{A}}$ is the set of incomplete instantiated actions from $\widetilde{\mathcal{O}}$ with objects from Z; $\mathcal{I} \in \mathcal{F}$ an initial state; $\mathcal{G}$ is the set of possible goals, which include a correct hidden goal G^* (i.e., $G^* \in \mathcal{G}$); and $Obs = \langle o_1, o_2, ..., o_n \rangle$ is an observation sequence of executed actions, with each observation $o_i \in \widetilde{\mathcal{A}}$. Obs corresponds to the sequence of actions (i.e., a plan) to solve a problem in a complete domain in $\langle\langle \widetilde{\mathcal{D}} \rangle\rangle$.*

A solution for a goal recognition problem in incomplete domain models $\widetilde{\mathcal{T}}$ is the correct hidden goal $G^* \in \mathcal{G}$ that the observation sequence Obs of a plan execution achieves, specifically, the correct hidden goal is the intended goal that the observed agent wants to achieve. As most keyhole goal recognition approaches, observations consist of the actions of the underlying plan, more specifically, we observe incomplete actions with possible precondition and effects, in which some of the preconditions might be required and some effects might change the environment. While a full (or complete) observation sequence contains all of the action signatures of the plan executed by the observed agent, a partial observation sequence contains only a sub-sequence of actions of a plan and thus misses some of the actions actually executed in the environment. Our approaches are not limited to use just actions as observations and can also deal with logical facts as observations, *i.e.*, state observations, like (Sohrabi, Riabov, and Udrea 2016).

3 Extracting Landmarks in Incomplete Domain Models

In planning, landmarks are facts (or actions) that must be achieved (or executed) at some point along all valid plans to achieve a goal from an initial state (Hoffmann, Porteous, and Sebastia 2004). Landmarks are often used to build heuristics for planning algorithms (Richter, Helmert, and Westphal 2008; Richter and Westphal 2010). Whereas landmark-based heuristics extract landmarks from complete and correct domain models in the planning literature, we extend the landmark extraction algorithm of Hoffmann *et al.* in (2004) to extract *definite* and *possible* landmarks in incomplete STRIPS domain models. This algorithm uses a Relaxed Planning Graph (RPG), which is a leveled graph that ignores the delete-list effects of all actions, thus containing no mutex relations (Hoffmann and Nebel 2001). This algorithm builds the RPG and then extracts a set of *landmark candidates* by back-chaining from the RPG level in which all facts of the goal state G are possible, and, for each fact g in G, it checks which facts must be true until the first level of the RPG. For example, if fact B is a landmark and all actions that achieve B share A as precondition, then A is a landmark candidate. To confirm that a landmark candidate is indeed a necessary condition, and thus a landmark, the algorithm builds a new RPG removing actions that achieve the landmark candidate (called *achiever* actions) and checks the solvability over this modified problem, which can be done in polynomial time (Blum and Furst 1997). If the modified problem is unsolvable, then the landmark candidate is a necessary landmark, and the actions that achieve it are necessary to solve the original problem.

Like most planning approaches in incomplete domains (Weber and Bryce 2011; Nguyen and Kambhampati 2014; Nguyen, Sreedharan, and Kambhampati 2017), we reason about possible plans with incomplete actions (observations) by assuming that they succeed under the *most optimistic* conditions: (1) possible preconditions do not need to be satisfied; (2) possible and known delete-effects are ignored; and (3) possible add effects are assumed to occur. Thus, we adapt the extraction algorithm from Hoffmann, Porteous, and Sebastia (2004) to extract landmarks from incomplete domain models by building an Optimistic Relaxed Planning Graph (ORPG) instead of the original RPG. An ORPG is leveled graph that deals with incomplete domain models by assuming the most *optimistic* conditions. While the optimistic assumption may lead our landmark extraction algorithm to consider an infeasible goal possible, it never rules out a possible goal. The ORPG allows us to extract the *definite* and *possible* landmarks from Definitions 2 and 3.

Definition 2 (Definite Landmark) *A definite landmark L_D is a fact (landmark) that is extracted from a known add effect $\mathit{eff}^+(a)$ of an achiever a (action) in the ORPG.*

Definition 3 (Possible Landmark) *A possible landmark L_P is a fact (landmark) that is extracted from a possible add effect $\widetilde{\mathit{eff}}^+(a)$ of an achiever a (action) in the ORPG and is such that $L_P \cap L_D = \emptyset$.*

Figure 1 shows the ORPG for Example 1. The set of *definite* landmarks is $\{p, r, g\}$, and the set of *possible* landmarks is $\{q\}$. The classical landmark extraction algorithm from Hoffmann, Porteous, and Sebastia, returns $\{p, r, g\}$ as landmarks ignoring q as a fact landmark because it does not assume the most *optimistic* condition that possible add effects always occur, disregarding action a as a possible achiever. During the landmark extraction process, our algorithm selects first the action achievers that have known add effect ($\mathit{eff}^+(a)$) to a candidate landmark, prioritizing achievers of *definite* landmarks over achievers of *possible* landmarks. For example, consider the ORPG in Figure 1, in this graph the action achievers of r are the actions a and b, and our algorithm selects first the action b because it has a known add effect to r, and so on.

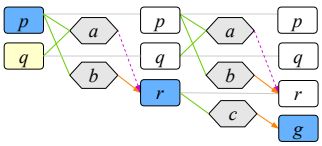


Figure 1: ORPG for Example 1. Green arrows represent preconditions, Orange arrows represent add effects, and Purple dashed arrows represent possible add effects. Light-Blue boxes represent *definite* landmarks and Light-Yellow boxes represent *possible* landmarks. Hexagons represent actions.

4 Heuristic Goal Recognition Approaches using Enhanced Landmarks

Under the optimistic assumption we use in our technique, an incomplete action $\widetilde{a}$ instantiated from an incomplete operator $\widetilde{op}$ is applicable to a state S iff $S \models \mathit{pre}(\widetilde{a})$ and results in a new state S' such that $S' := (S/\mathit{eff}^-(a)) \cup (\widetilde{\mathit{eff}}^+(\widetilde{a}) \cup \mathit{eff}^+(a))$. Thus, a valid plan π that achieves a goal G from $\mathcal{I}$ in an incomplete planning problem $\widetilde{\mathcal{P}}$ is a sequence of actions that corresponds to an *optimistic* sequence of states. The $[\widetilde{a}, \widetilde{b}, \widetilde{c}]$ sequence of actions is a valid plan to achieve goal state $\{g\}$ from the initial state $\{p, q\}$ from Example 1. It corresponds to the *optimistic* state sequence: $s_0 = \{p, q\}, s_1 = \{p, q, r\}, s_2 = \{q, r\}, s_3 = \{q, r, g\}$. The number of completions for this example is $|\langle\langle\widetilde{\mathcal{D}}\rangle\rangle| = 2^5$, i.e. 2 possible preconditions, 1 possible add effect and 2 possible delete effects.

Key to our goal recognition approaches is observing the evidence of achieved landmarks during observations to recognize which goal is more consistent with the observations in a plan execution. To do so, our approaches combine the concepts of *definite* and *possible* with that of *overlooked* landmarks (Definition 4).

Definition 4 (Overlooked Landmark) *An overlooked landmark L_O is an actual landmark, a necessary fact for all valid plans towards a goal from an initial state, that was not detected by approximate landmark extraction algorithms.*

Most landmark extraction algorithms extract only a subset of landmarks for a given planning problem, and to overcome this problem, we aim to extract *overlooked* landmarks by analyzing preconditions and effects in the observed actions of an observation sequence. Since we are dealing with incomplete domain models, and it is possible that some incomplete planning problems have few (or no) *definite* and/or *possible* landmarks, we extract *overlooked* landmarks from the evidence in the observations as we process them in order to enhance the set of landmarks useable by our goal recognition heuristics. This *on the fly* landmark extraction checks if the facts in the known preconditions and known and possible add effects are not in the set of extracted *definite* and *possible* landmarks, and if they are not, we check if these facts are *overlooked* landmarks. To do so, we use a function that builds a new ORPG by removing all actions that achieve a fact (*i.e.*, a potentially *overlooked* landmark) and checks the solvability of this modified problem. If the modified problem is indeed unsolvable, then this fact is an *overlooked* landmark. We check every candidate goal G in $\mathcal{G}$ using this function to extract additional (*overlooked*) landmarks. Example 2 illustrates how we extract *overlooked* landmarks *on the fly* for a given candidate goal and an observed action.

Example 2 *Consider the goal state defined in Example 1 as a candidate goal $G = \{g\}$, for a recognition problem with initial state $\mathcal{I} = \{p, q\}$, and an observed action a. Assume that a landmark extraction algorithm extracts $L = \{p, q, g\}$ for initial state $\mathcal{I}$ and candidate goal G. Given the observed action a, we check if the facts in the preconditions and known/possible effects of a are in L, and these facts are: $\{p, q, r\}$. Since r is not in L, r can be an overlooked landmark. To check if r is an overlooked landmark, we build ORPG removing all actions that achieve r (i.e., actions a and b) and check whether G remains solvable in this ORPG. In this case, the goal atom g is unreachable (and G is unsolvable) because r is a necessary fact to achieve g. Figure 1 il-*

lustrates that g is unachievable without actions a and b, and consequently without r, because there is no action that adds this fact. Thus, r is an overlooked landmark that was not extracted by the extraction algorithm, but it was extracted on the fly from an observed action during plan execution.

4.1 Enhanced Goal Completion Heuristic

By combining our new notions of landmarks we develop a goal recognition heuristic for recognizing goals in incomplete domain models. Our heuristic estimates the correct goal in the set of candidate goals by calculating the ratio between achieved *definite* ($\mathcal{AL}_G$), *possible* ($\widetilde{\mathcal{AL}}_G$), and *overlooked* ($\mathcal{ANL}_G$) landmarks and the amount of *definite* ($\mathcal{L}_G$), *possible* ($\widetilde{\mathcal{L}}_G$), and *overlooked* ($\mathcal{NL}_G$) landmarks. The estimate computed using Equation 1 represents the percentage of achieved landmarks for a goal from observations.

$$h_{\widetilde{GC}}(G) = \left(\frac{\mathcal{AL}_G + \widetilde{\mathcal{AL}}_G + \mathcal{ANL}_G}{\mathcal{L}_G + \widetilde{\mathcal{L}}_G + \mathcal{NL}_G} \right) \quad (1)$$

4.2 Enhanced Uniqueness Heuristic

Most recognition problems contain multiple candidate goals that share common fact landmarks, generating ambiguity that jeopardizes the goal completion heuristic. Clearly, landmarks that are common to multiple candidate goals are less useful for recognizing a goal than landmarks that occur for only a single goal. As a consequence, computing how unique (and thus informative) each landmark is can help disambiguate similar goals for a set of candidate goals (Pereira, Oren, and Meneguzzi 2017). Our second heuristic approach is based on this intuition, which we develop through the concept of *landmark uniqueness*, which is the inverse frequency of a landmark among the landmarks found in a set of candidate goals. Intuitively, a landmark L that occurs only for a single goal within a set of candidate goals has the maximum uniqueness value of 1. Equation 2 formalizes the computation of the *landmark uniqueness value* for a landmark L and a set of landmarks for all candidate goals $K_{\mathcal{G}}$.

$$L_{Uniq}(L, K_{\mathcal{G}}) = \left(\frac{1}{\sum_{\mathcal{L} \in K_{\mathcal{G}}} |\{L | L \in \mathcal{L}\}|} \right) \quad (2)$$

Using the concept of *landmark uniqueness value*, we estimate which candidate goal is the intended one by summing the uniqueness values of the landmarks achieved in the observations. Unlike our previous heuristic, which estimates progress towards goal completion by analyzing just the set of achieved landmarks, the landmark-based uniqueness heuristic estimates the goal completion of a candidate goal G by calculating the ratio between the sum of the uniqueness value of the achieved landmarks of G and the sum of the uniqueness value of all landmarks of a goal G. Our new uniqueness heuristic also uses the concepts of *definite*, *possible*, and *overlooked* landmarks. We store the set of *definite* and *possible* landmarks of a goal G separately into $\mathcal{L}_G$ and $\widetilde{\mathcal{L}}_G$, and the set of *overlooked* landmarks into $\mathcal{NL}_G$. Thus, the uniqueness heuristic effectively weighs the completion value of a goal by the informational value of a landmark so that unique landmarks have the highest weight. To estimate goal completion using the *landmark uniqueness value*, we calculate the uniqueness value for every extracted (*definite*, *possible*, and *overlooked*) landmark in the set of landmarks of the candidate goals using Equation 2. Since we use three types of landmarks and they are stored in three different sets, we compute the landmark uniqueness value separately for them, storing the landmark uniqueness value of *definite* landmarks $\mathcal{L}_G$ into $\Upsilon_{\mathcal{L}}$, the landmark uniqueness value of *possible* landmarks $\widetilde{\mathcal{L}}_G$ into $\Upsilon_{\widetilde{\mathcal{L}}}$, and the landmark uniqueness value of *overlooked* landmarks $\mathcal{NL}_G$ into $\Upsilon_{\mathcal{NL}_G}$. Our uniqueness heuristic is denoted as $h_{\widetilde{UNIQ}}$ and formally defined in Equation 3.

$$h_{\widetilde{UNIQ}}(G) = \left(\frac{\sum_{\mathcal{A}_L \in \mathcal{AL}_G} \Upsilon_{\mathcal{L}}(\mathcal{A}_L) + \sum_{\widetilde{\mathcal{A}}_L \in \widetilde{\mathcal{AL}}_G} \Upsilon_{\widetilde{\mathcal{L}}}(\widetilde{\mathcal{A}}_L) + \sum_{\mathcal{ANL} \in \mathcal{ANL}_G} \Upsilon_{\mathcal{NL}_G}(\mathcal{ANL})}{\sum_{L \in \mathcal{L}_G} \Upsilon_{\mathcal{L}}(L) + \sum_{\tilde{L} \in \widetilde{\mathcal{L}}_G} \Upsilon_{\widetilde{\mathcal{L}}}(\tilde{L}) + \sum_{\mathcal{NL} \in \mathcal{NL}_G} \Upsilon_{\mathcal{NL}_G}(\mathcal{NL})} \right) \quad (3)$$

5 Experiments and Evaluation

We evaluate our approaches empirically in two ways. First, we compare the recognition performance of our approaches against the two landmark-based approaches of Pereira, Oren, and Meneguzzi (2017), which we use as baselines. We then empirically evaluate the effect of the various types of landmarks to recognition performance.

5.1 Datasets and Setup

We used openly available goal and plan recognition datasets (Pereira and Meneguzzi 2017) for our experiments. These datasets contain thousands of recognition problems comprising large and non-trivial planning problems (with optimal and sub-optimal plans as observations) for 15 planning domains, including domains and problems from datasets that were developed by Ramírez and Geffner (2009; 2010). All planning domains in these datasets are encoded using the STRIPS fragment of PDDL. Domains include those modeled after realistic applications (*e.g.*, DWR, ROVERS, LOGISTICS), as well as hard artificial domains (*e.g.*, SOKOBAN). Each recognition problem in these datasets contains a complete domain definition, an initial state, a set of candidate goals, a correct hidden goal in the set of candidate goals, and an observation sequence. An observation sequence contains actions that represent an optimal or sub-optimal plan that achieves a correct hidden goal, and this observation sequence can be full or partial. A full observation sequence represents the whole plan that achieves the hidden goal, *i.e.*, 100% of the actions having been observed. A partial observation sequence represents a plan for the hidden goal, varying in 10%, 30%, 50%, or 70% of its actions having been observed. To evaluate our recognition approaches in incomplete domain models, we modify the domain models of these datasets by adding annotated possible preconditions and effects. Thus, the only modification to the original datasets is the generation of new, incomplete, domain models for each problem, varying the percentage of incompleteness in these domains. We vary the percentage of incompleteness of a domain from 20% to 80%. For example, consider that a com-

plete domain has, for all its actions, a total of 10 preconditions, 10 add effects, and 10 delete effects. A derived model with 20% of incompleteness needs to have 2 possible preconditions (8 known preconditions), 2 possible add effects (8 known add effects), and 2 possible delete effects (8 known delete effects), and so on for other percentages of incompleteness. Like (Nguyen and Kambhampati 2014; Nguyen, Sreedharan, and Kambhampati 2017), we generated the incomplete domain models by following three steps involving randomly selected preconditions/effects: (1) move a percentage of known preconditions and effects into lists of possible preconditions and effects; (2) add possible preconditions from delete effects that are not preconditions of a corresponding operator; and (3) add into possible lists (of preconditions, add effects, or delete effects) predicates whose parameters fit into the operator signatures and are not precondition or effects of the operator. These steps yield three different incomplete domain models from a complete domain model for each percentage of domain incompleteness with different possible lists of preconditions and effects. We ran all experiments using a single core of a 12 core Intel(R) Xeon(R) CPU E5-2620 v3@2.40GHz with 16GB of RAM, with a time limit of 2 minutes and memory limit of 2GB.

5.2 Experimental Results: ROC Space

Our approaches recognize goals at low recognition time for most planning domains, taking at most 2.7 seconds, including the process of extracting landmarks, among all goal recognition problems, apart from IPC-GRID and SOKOBAN, which took substantial recognition time. So, 1092 (20% of domain incompleteness) out of 4368 problems for IPC-GRID and SOKOBAN do not exceed the time limit of 2 minutes (for both our approaches and the baseline). SOKOBAN exceeds the time limit of 2 minutes for most goal recognition problems because this dataset contains large problems with a huge number of objects, leading to an even larger number of instantiated predicates and actions. For example, as domain incompleteness increases (*i.e.*, the ratio of possible to definite preconditions and effects), the number of possible actions (moving between cells and pushing boxes) increases substantially in a grid with 9x9 cells and 5 boxes as there are very few known preconditions for several possible preconditions. As a basis of comparison, state-of-the-art planners (Nguyen and Kambhampati 2014; Nguyen, Sreedharan, and Kambhampati 2017) for incomplete domain models take substantially more time than our 2-minute timeout to generate a single plan for domains that our approach recognizes in less than 2 seconds. For example, CPISA (Nguyen, Sreedharan, and Kambhampati 2017) takes $\approx$ 300 seconds to find a plan with 25 steps in domains (*e.g.*, SATELLITE) with 2 possible preconditions and 3 possible add effects, whereas our dataset contains much more complex incomplete domains and problems. The average number of possible complete domain models $|\langle\langle\widetilde{\mathcal{D}}\rangle\rangle|$ is huge for several domains, showing that the task of goal recognition in incomplete domain models is quite difficult and complex. For instance, the average number of possible complete domains in this dataset varies between 9.18 (SOKOBAN with 20% of domain incompleteness) and 7.84^{15} (ROVERS with 80% of domain incompleteness).

We adapt the *Receiver Operating Characteristic* (ROC) curve metric to show the trade-off between true positive and false positive results. An ROC curve is often used to compare true positive and false positive predictions of the experimented approaches. Here, each prediction result of our goal recognition approaches represents one point in the space, and thus, instead of a curve, our graphs show the spread of our results over ROC space. The diagonal line in ROC space represents a random guess to recognize a goal from observations. This line divides the ROC space in such a way that points above the diagonal represent good classification results (better than random guess), whereas points below the line represent poor results (worse than random guess). The best possible (perfect) prediction for recognizing goals are points in the upper left corner (0,100).

Figure 2 shows ROC space graphs corresponding to recognition performance over the four percentages of domain incompleteness we used in our experiments. We aggregate multiple recognition problems for all domains and plot these results in ROC space varying the percentage of domain incompleteness. We compare our enhanced heuristics ($h_{\widetilde{GC}}$ and $h_{\widetilde{UNIQ}}$) using the various types of landmarks (*definite*, *possible*, and *overlooked*) against the baseline (h_{gc} and h_{uniq}) (Pereira, Oren, and Meneguzzi 2017), that uses just the landmarks extracted by a classical landmark extraction algorithm, *i.e.*, ignoring the incomplete part of the domain model (possible preconditions and effects). Although the true positive rate is high for most recognition problems at most percentages of domain incompleteness, as the percentage of domain incompleteness increases, the false positive rate also increases, leading to several problems being recognized with a performance close to the random guess line. This happens because the number of extracted landmarks decreases significantly as the number of known preconditions and effects diminishes, and consequently, all candidate goals have few (if any) landmarks. For example, in several cases in which domain incompleteness is 60% and 80%, the set of landmarks is quite similar, leading our approaches to return more than one candidate goal as the correct one. Thus, there are more returned goals during the recognition process as incompleteness increases. These results show that our enhanced approaches perform better and are more accurate than the baselines, aggregating most points in the left corner, while the points for the baseline approaches are closer to (and sometimes below) the random guess line.

5.3 Ablation Study: The Impact of Landmarks

Since the key contribution of our approaches are based on the new types of landmarks (*definite*, *possible*, and *overlooked*), as opposed to the traditional landmarks from planning heuristic, we wanted to objectively measure the effect of these new types of landmark on the recognition performance of our heuristics. Thus, we performed an ablation study that consists of measuring the performance of our approaches using some possible combinations of landmark types. This ablation study evaluates only linear combinations of landmark types, and some combinations are not included in this paper because they had poor performance and

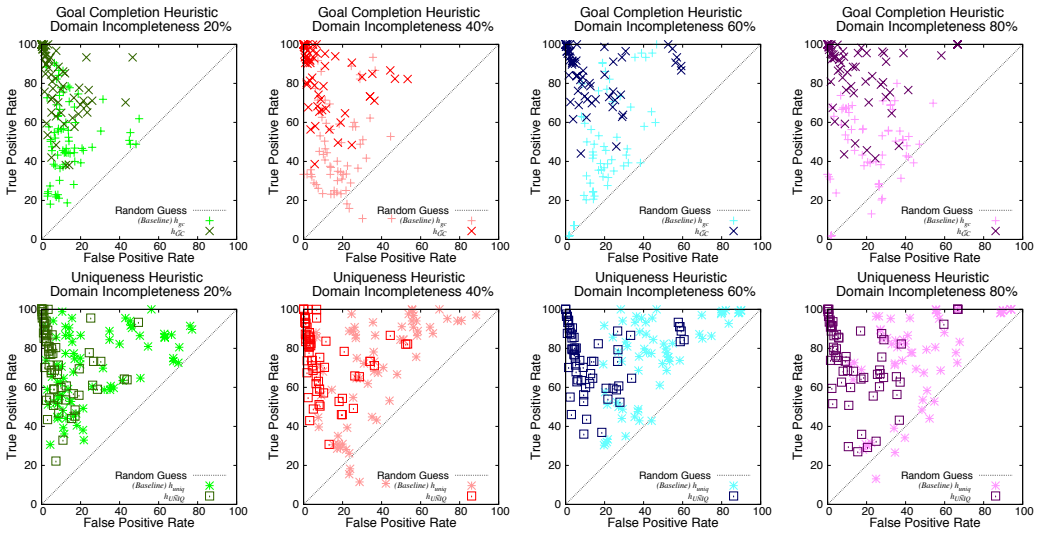


Figure 2: ROC space for all four percentage of domain incompleteness, comparing our enhanced heuristic approaches against the baseline for both Goal Completion and Uniqueness Heuristic.

seem to be not relevant for the ablation study. We evaluate our approaches using the standard metrics of *precision* (ratio of correct positive predictions among all predictions) and *recall* (ratio between true positive results and total true positive and false negative results). In order to present a unified metric, we report the *F1-score* (harmonic mean) of *precision* and *recall*. To perform our ablation study, we use the *correlation* (C) between the averages of *F1-score* (F_1) and the absolute number of the various types of landmarks (*definite* D, *possible* P, and *overlooked* O landmarks) over all domains and problems, and *Spread* S, representing the average number of returned goals.

In what follows, the *correlation* is a real value in $[-1, 1]$ such that -1 represents an *anti-correlation* between the landmarks and the *F1-score*, 0 represents no correlation between the landmarks and the *F1-score*, whereas a value of 1 represents that more landmarks correlate to a higher *F1-score*. We computed the *correlation* between the averages of landmarks and *F1-scores* over all domains and degrees of incompleteness in Table 1 and plot these correlations as a function of the level of incompleteness in Figures 3a–3f as opposed to how traditional landmarks affect performance on the baseline approaches (*correlation* varying between 0.32 and 0.67 for a lower *F1-score*, Table 1, lines 1 and 2). Figures 3a and 3d show how all types of landmark correlate to performance in the complete technique (represented by $\widetilde{GC}$ (*D+P+O*) and $\widetilde{UNIQ}$ (*D+P+O*)). At low levels of incompleteness we have larger *F1-scores* and larger numbers of *definite* landmarks, leaving a smaller number of *overlooked* landmarks to be inferred *on the fly*. Under these conditions, *overlooked* landmarks start off at a slight anti-correlation with performance. As the level of incompleteness of the domain description increases, the number of *definite* landmarks decreases, but their correlation to performance increases, as the more landmarks we can infer, the better performance we can achieve. The number of *overlooked* landmarks remains broadly the same over time, as they are tied to the observations more than they are tied to the domain description, and their correlation to performance monotonically increases as the incompleteness increases. This indicates that *overlooked* landmarks play an increasingly important role in recognition performance as incompleteness increases, giving more information to our enhanced heuristics. The number of *possible* landmarks also varies with incompleteness, initially increasing as the number of possible effects increases, to subsequently decrease as the number of possible effects leads to less bottlenecks in the state-space to yield landmarks. As the number of possible effects increases, so does their unreliability as sources of landmarks, which is reflected in their decreasing *correlation* to performance.

As we ablate landmarks, performance drops most substantially when we remove either *definite* or *possible* landmarks from the heuristics (Figures 3c and 3f), indicating their importance to recognition accuracy. We can see that using *overlooked* landmarks exclusively, in $\widetilde{GC}$ (O) and $\widetilde{UNIQ}$ (O) provides a close approximation of the performance of the technique using all landmark types. This is strong evidence that *overlooked* landmarks are the most important contribution of our technique.

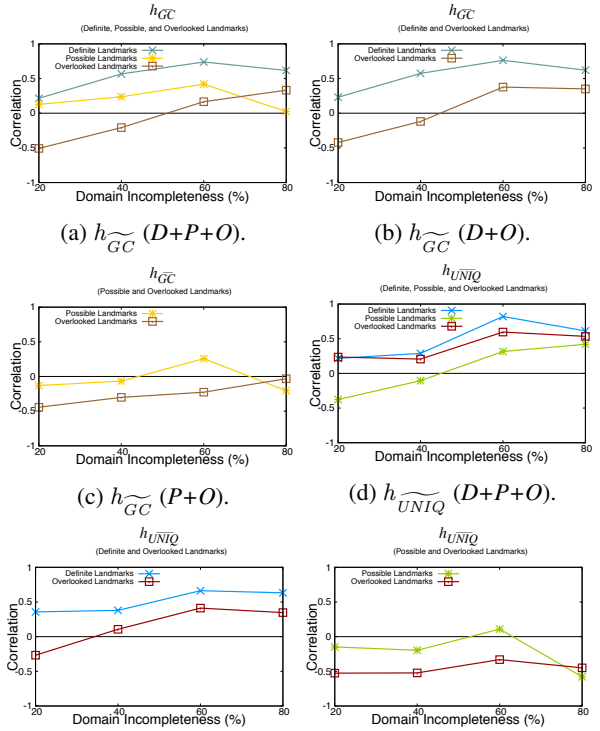
Figure 4 compares all approaches with respect to *F1-score* averages varying the domain incompleteness. Thicker lines represent the approaches that have higher *F1-scores*: our enhanced heuristics ($\widetilde{GC}$ and $\widetilde{UNIQ}$) that combine the use of the new types of landmarks, *i.e.*, *D+O*, *D+P+O*, and O, showing that using *overlooked* landmarks substantially improves goal recognition accuracy.

Table 2 reports the results for all evaluated approaches and domains varying not only domain incompleteness but also the percentage of observability of the observations, showing the averages for all types of landmarks, *F1-score*, and *correlation*. Namely, each inner table in Table 2 represents the results for each percentage of observability (10%, 30%, 50%,

	Domain Incompleteness 20%						*Domain Incompleteness 40%*						*Domain Incompleteness 60%*						*Domain Incompleteness 80%*					
	D	P	O	S	F_1	$CD/CP/CO$	D	P	O	S	F_1	$CD/CP/CO$	D	P	O	S	F_1	$CD/CP/CO$	D	P	O	S	F_1	$CD/CP/CO$
Baseline (gc)	11.5	0	0	1.53	0.44	0.35/0/0	4.9	0	0	2.27	0.36	0.42/0/0	3.5	0	0	3.32	0.31	0.65/0/0	2.9	0	0	4.85	0.34	0.67/0/0
Baseline ($uniq$)	11.5	0	0	1.48	0.44	0.32/0/0	4.9	0	0	1.95	0.34	0.36/0/0	3.5	0	0	2.83	0.32	0.33/0/0	2.9	0	0	4.12	0.33	0.40/0/0
$\widetilde{GC}$ (D+P+O)	11.6	1.4	21.2	1.06	0.74	0.22/0.13/-0.51	9.6	2.1	19.0	1.17	0.73	0.56/0.23/-0.20	7.1	2.3	19.4	1.31	0.67	0.73/0.41/0.16	6.8	1.2	19.1	1.38	0.65	0.61/0.02/0.33
$\widetilde{GC}$ (D+O)	11.6	0	25.5	1.06	0.75	0.23/0/-0.42	9.6	0	24.9	1.23	0.74	0.57/0/-0.11	7.1	0	25.8	1.38	0.68	0.76/0/0.37	6.8	0	22.5	1.66	0.66	0.62/0/0.35
$\widetilde{GC}$ (P+O)	0	1.4	42.8	6.11	0.32	0/-0.13/-0.44	0	2.1	35.3	5.93	0.34	0/-0.06/-0.30	0	2.3	29.7	5.86	0.33	0/0.25/-0.22	0	1.2	27.9	5.84	0.27	0/-0.20/-0.03
$\widetilde{GC}$ (P)	0	1.42	0	7.24	0.22	0/0.29/0	0	2.13	0	7.18	0.29	0/0.07/0	0	2.23	0	7.01	0.26	0/0.56/0	0	1.18	0	6.99	0.19	0/0.06/0
$\widetilde{GC}$ (O)	0	0	47.5	1.12	0.71	0/0/-0.21	0	0	41.4	1.23	0.70	0/0/0.05	0	0	35.9	1.39	0.65	0/0/0.42	0	0	31.2	1.68	0.63	0/0/0.34
$\widetilde{UNIQ}$ (D+P+O)	11.6	1.4	21.2	1.05	0.69	0.21/0.23/-0.37	9.6	2.1	19.0	1.16	0.67	0.28/0.21/-0.10	7.1	2.3	19.4	1.29	0.63	0.82/0.59/0.31	6.8	1.2	19.1	1.37	0.61	0.61/0.034/0.41
$\widetilde{UNIQ}$ (D+O)	11.6	0	25.5	1.05	0.68	0.35/0/-0.26	9.6	0	24.9	1.20	0.65	0.38/0/0.11	7.1	0	25.8	1.38	0.61	0.66/0/0.41	6.8	0	22.5	1.41	0.62	0.63/0/0.34
$\widetilde{UNIQ}$ (P+O)	0	1.4	42.8	5.50	0.32	0/-0.14/-0.52	0	2.1	35.3	5.71	0.35	0/-0.19/-0.53	0	2.3	29.7	5.49	0.33	0/0.10/-0.33	0	1.2	27.9	5.65	0.29	0/-0.57/-0.45
$\widetilde{UNIQ}$ (P)	0	1.42	0	6.87	0.28	0/-0.27/0	0	2.13	0	6.15	0.27	0/-0.42/0	0	2.23	0	5.92	0.27	0/-0.19/0	0	1.18	0	5.81	0.27	0/-0.55/0
$\widetilde{UNIQ}$ (O)	0	0	47.5	1.05	0.71	0/0/-0.11	0	0	41.4	1.17	0.70	0/0/0.03	0	0	35.9	1.28	0.64	0/0/0.39	0	0	31.2	1.36	0.62	0/0/0.33

Table 1: Performance results for ablation study, comparing the baseline approaches against our heuristics using various combinations of landmark types.

70%, and 100%) over the evaluated dataset. It is possible to see that our complete technique (D+P+O) when applied for both heuristics outperforms the other combinations (including the baseline approaches) in all variations of domain incompleteness and observability. Thus, from these results, we can conclude the combination of all types landmarks (*definite*, *possible*, and *overlooked*) yields better results in comparison to other combinations over the dataset.



(a) $h_{\widetilde{GC}}$ (D+P+O). (b) $h_{\widetilde{GC}}$ (D+O).

(c) $h_{\widetilde{GC}}$ (P+O). (d) $h_{\widetilde{UNIQ}}$ (D+P+O).

(e) $h_{\widetilde{UNIQ}}$ (D+O). (f) $h_{\widetilde{UNIQ}}$ (P+O).

Figure 3: Correlation of landmarks to performance. Note that this does not report absolute *F1-score*.

6 Conclusions and Future Work

We have developed novel goal recognition approaches that deal with incomplete domain models that represent *possible* preconditions and effects besides traditional models where such information is assumed to be *known*. The main contributions of this paper include the formalization of goal

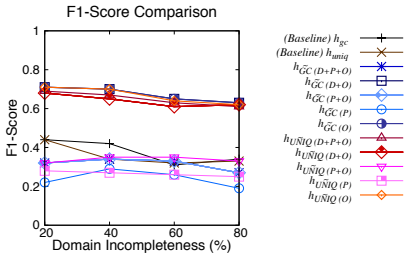


Figure 4: *F1-score* average of all evaluated approaches over the dataset varying the domain incompleteness.

recognition in incomplete domains, two (enhanced) heuristic approaches for such goal recognition, novel notions of landmarks for incomplete domains, and a dataset to evaluate the performance of such approaches. Our novel notions of landmarks include that of *possible* landmarks for incomplete domains as well as *overlooked* landmarks that allow us to compensate fast but non-exhaustive landmark extraction algorithms, the latter of which can also be employed to improve existing goal and plan recognition approaches (Pereira and Meneguzzi 2016; Pereira, Oren, and Meneguzzi 2017). Experiments over thousands of goal recognition problems in 15 planning domain models show two key results of our approaches. First, these approaches are fast and accurate when dealing with incomplete domains at all variations of observability and domain incompleteness. Our use of novel heuristics frees us from using full fledged incomplete-domain planners as part of the recognition process. Approaches that use planners for goal recognition are already very expensive for complete domains and are even more so in incomplete domains, since they often generate plans taking into consideration many of the possible models, and even then they often fail to generate robust plans for these domains. Second, our ablation study shows that our additional notions of landmarks have substantial impact on the accuracy of our approach over simply ignoring the uncertain information from the domain model, as we use in the baseline approach. Importantly, the ablation study shows that overlooked landmarks contribute substantially to the accuracy of our approach. We envision such techniques to be instrumental in using learned planning models (Asai and Fukunaga 2018) for goal recognition (Amado et al. 2018).

Acknowledgements: We thank Miquel Ramírez for the invaluable discussions about previous versions of this paper. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior Brasil (CAPES) - Finance Code 001. Felipe acknowledges support from CNPq under project numbers 407058/2018-4 and 305969/2016-1.

10% of Observability

ggssg	Domain Incompleteness 20%						Domain Incompleteness 40%						Domain Incompleteness 60%						Domain Incompleteness 80%					
	D	P	O	S	F_1	$CD/CP/CO$	D	P	O	S	F_1	$CD/CP/CO$	D	P	O	S	F_1	$CD/CP/CO$	D	P	O	S	F_1	$CD/CP/CO$
Baseline (gc)	11.5	0.0	0.0	1.77	0.33	-0.04/0/0	4.9	0.0	0.0	2.41	0.28	-0.02/0/0	3.5	0.0	0.0	2.45	0.25	0.34/0/0	2.9	0.0	0.0	2.79	0.26	0.36/0/0
Baseline ($uniq$)	11.5	0.0	0.0	1.36	0.33	-0.11/0/0	4.9	0.0	0.0	1.67	0.29	-0.02/0/0	3.5	0.0	0.0	1.42	0.28	0.08/0/0	2.9	0.0	0.0	1.90	0.28	0.09/0/0
$\widetilde{GC}$ (D+P+O)	11.6	1.4	1.3	1.94	0.46	-0.11/-0.08/-0.61	9.6	2.1	1.2	2.24	0.44	0.16/-0.13/-0.47	7.1	2.2	1.4	2.86	0.40	0.31/0.16/-0.26	6.8	1.1	1.3	3.27	0.39	0.21/-0.40/-0.24
$\widetilde{GC}$ (D+O)	11.6	0.0	1.7	2.03	0.45	-0.13/0/-0.55	9.6	0.0	1.7	2.51	0.44	0.15/0/-0.42	7.1	0.0	1.9	3.23	0.40	0.28/0/-0.16	6.8	0.0	1.6	3.58	0.38	0.20/0/-0.23
$\widetilde{GC}$ (P+O)	0.0	1.4	3.0	4.30	0.30	0/-0.16/-0.37	0.0	2.1	2.4	4.11	0.31	0/-0.14/-0.26	0.0	2.2	2.0	4.08	0.29	0/0.01/-0.27	0.0	1.1	1.8	4.26	0.28	0/-0.55/-0.18
$\widetilde{GC}$ (P)	0.0	1.4	0.0	2.31	0.24	0.11/0/0	0.0	2.1	0.0	2.29	0.22	-0.07/0/0	0.0	2.2	0.0	2.08	0.22	0.28/0/0	0.0	1.1	0.0	2.20	0.21	-0.29/0/0
$\widetilde{GC}$ (O)	0.0	0.0	3.4	3.54	0.43	0/0/-0.39	0.0	0.0	2.9	3.92	0.42	0/0/-0.34	0.0	0.0	2.6	4.38	0.38	0/0/-0.16	0.0	0.0	2.1	4.51	0.37	0/0/-0.19
$\widetilde{UNIQ}$ (D+P+O)	11.6	1.4	1.3	1.87	0.40	-0.12/-0.02/-0.46	9.6	2.1	1.2	2.01	0.40	0.08/-0.16/-0.39	7.1	2.2	1.4	2.34	0.35	0.24/0.12/-0.23	6.9	1.1	1.3	2.91	0.35	0.13/-0.57/-0.26
$\widetilde{UNIQ}$ (D+O)	11.6	0.0	1.7	2.00	0.38	-0.13/0/-0.43	9.6	0.0	1.7	2.39	0.38	0.06/0/-0.39	7.1	0.0	1.9	3.01	0.35	0.16/0/-0.26	6.9	0.0	1.6	3.12	0.34	0.12/0/-0.31
$\widetilde{UNIQ}$ (P+O)	0.0	1.4	3.0	4.38	0.30	0/-0.23/-0.47	0.0	2.1	2.4	3.98	0.32	0/-0.33/-0.46	0.0	2.2	2.0	3.79	0.29	0/-0.15/-0.34	0.0	1.1	1.8	3.73	0.29	0/-0.62/-0.35
$\widetilde{UNIQ}$ (P)	0.0	1.4	0.0	1.25	0.26	-0.26/0/0	0.0	2.1	0.0	1.32	0.27	-0.42/0/0	0.0	2.2	0.0	1.32	0.27	-0.20/0/0	0.0	1.1	0.0	1.32	0.27	-0.55/0/0
$\widetilde{UNIQ}$ (O)	0.0	0.0	3.4	3.37	0.43	0/0/-0.42	0.0	0.0	2.9	3.80	0.42	0/0/-0.34	0.0	0.0	2.6	4.24	0.38	0/0/-0.17	0.0	0.0	2.1	4.40	0.37	0/0/-0.19

30% of Observability

	Domain Incompleteness 20%						Domain Incompleteness 40%						Domain Incompleteness 60%						Domain Incompleteness 80%					
	D	P	O	S	F_1	$CD/CP/CO$	D	P	O	S	F_1	$CD/CP/CO$	D	P	O	S	F_1	$CD/CP/CO$	D	P	O	S	F_1	$CD/CP/CO$
Baseline (gc)	11.5	0.0	0.0	1.46	0.33	0.14/0/0	4.8	0.0	0.0	1.64	0.28	0.29/0/0	3.5	0.0	0.0	1.90	0.25	0.55/0/0	2.9	0.0	0.0	2.16	0.26	0.59/0/0
Baseline ($uniq$)	11.5	0.0	0.0	1.33	0.33	0.15/0/0	4.8	0.0	0.0	1.59	0.29	0.20/0/0	3.5	0.0	0.0	1.72	0.28	0.21/0/0	2.9	0.0	0.0	2.07	0.28	0.30/0/0
$\widetilde{GC}$ (D+P+O)	11.5	1.4	4.4	1.39	0.66	0.09/0.05/-0.64	9.6	2.1	4.0	1.55	0.63	0.48/0.16/-0.45	7.1	2.2	4.4	1.88	0.57	0.70/0.54/-0.04	6.8	1.1	4.2	2.15	0.55	0.53/-0.08/0.10
$\widetilde{GC}$ (D+O)	11.6	0.0	5.5	1.41	0.65	0.09/0/-0.56	9.6	0.0	5.4	1.64	0.63	0.50/0/-0.35	7.1	0.0	6.0	2.01	0.57	0.69/0/0.16	6.8	0.0	5.1	2.29	0.55	0.54/0/0.15
$\widetilde{GC}$ (P+O)	0.0	1.4	9.5	4.88	0.32	0/0/-0.41	0.0	2.1	7.6	4.49	0.34	0/0/-0.29	0.0	2.2	6.5	4.42	0.32	0/0/-0.18	0.0	1.1	6.0	4.66	0.30	0/0/-0.23
$\widetilde{GC}$ (P)	0.0	1.4	0.0	1.91	0.29	0/0.17/0	0.0	2.1	0.0	1.92	0.28	0/0.01/0	0.0	2.2	0.0	1.83	0.28	0/0.52/0	0.0	1.1	0.0	1.77	0.27	0/-0.34/0
$\widetilde{GC}$ (O)	0.0	0.0	10.7	1.97	0.64	0/0/-0.35	0.0	0.0	9.2	2.16	0.61	0/0/-0.19	0.0	0.0	8.2	2.50	0.54	0/0/0.18	0.0	0.0	6.8	2.69	0.53	0/0/0.13
$\widetilde{UNIQ}$ (D+P+O)	11.6	1.4	4.4	1.28	0.61	0.06/0.15/-0.52	9.6	2.1	4.0	1.36	0.58	0.28/0.14/-0.40	7.1	2.2	4.4	1.64	0.52	0.62/0.60/-0.02	6.8	1.1	4.2	1.94	0.50	0.53/-0.08/0.12
$\widetilde{UNIQ}$ (D+O)	11.6	0.0	5.5	1.28	0.59	0.08/0/-0.51	9.6	0.0	5.4	1.47	0.55	0.28/0/-0.28	7.1	0.0	6.0	1.84	0.50	0.53/0/0.09	6.8	0.0	5.1	2.12	0.51	0.55/0/0.04
$\widetilde{UNIQ}$ (P+O)	0.0	1.4	9.5	4.06	0.32	0/0/-0.50	0.0	2.1	7.6	3.78	0.33	0/0/-0.53	0.0	2.2	6.5	3.50	0.32	0/0/-0.34	0.0	1.1	6.0	3.26	0.29	0/0/-0.38
$\widetilde{UNIQ}$ (P)	0.0	1.4	0.0	1.25	0.26	0/-0.26/0	0.0	2.1	0.0	1.31	0.26	0/-0.42/0	0.0	2.2	0.0	1.31	0.27	0/-0.19/0	0.0	1.1	0.0	1.31	0.27	0/-0.55/0
$\widetilde{UNIQ}$ (O)	0.0	0.0	10.7	1.80	0.63	0/0/-0.34	0.0	0.0	9.2	2.01	0.61	0/0/-0.16	0.0	0.0	8.2	2.31	0.54	0/0/0.19	0.0	0.0	6.8	2.55	0.52	0/0/0.14

50% of Observability

	Domain Incompleteness 20%						Domain Incompleteness 40%						Domain Incompleteness 60%						Domain Incompleteness 80%					
	D	P	O	S	F_1	$CD/CP/CO$	D	P	O	S	F_1	$CD/CP/CO$	D	P	O	S	F_1	$CD/CP/CO$	D	P	O	S	F_1	$CD/CP/CO$
Baseline (gc)	11.5	0.0	0.0	1.36	0.44	0.38/0/0	4.8	0.0	0.0	1.50	0.36	0.43/0/0	3.5	0.0	0.0	1.74	0.31	0.63/0/0	2.9	0.0	0.0	1.95	0.34	0.71/0/0
Baseline ($uniq$)	11.5	0.0	0.0	1.27	0.45	0.32/0/0	4.8	0.0	0.0	1.48	0.34	0.41/0/0	3.5	0.0	0.0	1.56	0.32	0.31/0/0	2.9	0.0	0.0	1.74	0.32	0.46/0/0
$\widetilde{GC}$ (D+P+O)	11.6	1.4	8.8	1.24	0.76	0.29/0.17/-0.45	9.6	2.1	7.9	1.31	0.76	0.63/0.37/-0.20	7.1	2.2	8.5	1.53	0.70	0.85/0.55/0.24	6.8	1.1	8.2	1.76	0.67	0.65/0.05/0.35
$\widetilde{GC}$ (D+O)	11.6	0.0	10.9	1.26	0.76	0.34/0.0/-0.35	9.6	0.0	10.6	1.33	0.76	0.66/0.0/-0.04	7.1	0.0	11.4	1.59	0.69	0.84/0.0/0.46	6.8	0.0	9.8	1.84	0.67	0.65/0.0/0.39
$\widetilde{GC}$ (P+O)	0.0	1.4	18.6	5.19	0.18	0/-0.11/-0.37	0.0	2.1	14.9	4.73	0.20	0/-0.04/-0.28	0.0	2.2	12.6	4.82	0.19	0/0.34/-0.17	0.0	1.1	11.7	4.87	0.16	0/-0.50/-0.28
$\widetilde{GC}$ (P)	0.0	1.4	0.0	1.94	0.31	0/0.20/0	0.0	2.1	0.0	1.96	0.35	0/0.11/0	0.0	2.2	0.0	2.02	0.33	0/0.65/0	0.0	1.1	0.0	1.70	0.30	0/-0.27/0
$\widetilde{GC}$ (O)	0.0	0.0	20.7	1.54	0.74	0/0/-0.07	0.0	0.0	17.8	1.59	0.74	0/0/0.10	0.0	0.0	15.7	1.88	0.66	0/0/0.45	0.0	0.0	13.3	2.08	0.64	0/0/0.31
$\widetilde{UNIQ}$ (D+P+O)	11.6	1.4	8.8	1.13	0.72	0.32/0.26/-0.32	9.6	2.1	7.9	1.18	0.70	0.33/0.37/-0.07	7.1	2.2	8.5	1.38	0.64	0.83/0.64/0.36	6.8	1.1	8.2	1.60	0.62	0.63/0.06/0.40
$\widetilde{UNIQ}$ (D+O)	11.6	0.0	10.9	1.14	0.70	0.39/0.0/-0.21	9.6	0.0	10.6	1.25	0.67	0.45/0.0/0.17	7.1	0.0	11.4	1.52	0.62	0.68/0.0/0.44	6.8	0.0	9.8	1.78	0.63	0.66/0.0/0.33
$\widetilde{UNIQ}$ (P+O)	0.0	1.4	18.6	4.02	0.32	0/-0.13/-0.48	0.0	2.1	14.9	3.83	0.35	0/-0.17/-0.51	0.0	2.2	12.6	3.72	0.33	0/0.18/-0.29	0.0	1.1	11.7	3.21	0.29	0/-0.57/-0.41
$\widetilde{UNIQ}$ (P)	0.0	1.4	0.0	1.25	0.26	0/-0.26/0	0.0	2.1	0.0	1.31	0.27	0/-0.42/0	0.0	2.2	0.0	1.31	0.27	0/-0.19/0	0.0	1.1	0.0	1.31	0.27	0/-0.55/0
$\widetilde{UNIQ}$ (O)	0.0	0.0	20.7	1.39	0.74	0/0/-0.02	0.0	0.0	17.8	1.46	0.74	0/0/0.11	0.0	0.0	15.7	1.68	0.66	0/0/0.46	0.0	0.0	13.3	1.93	0.64	0/0/0.32

70% of Observability

	Domain Incompleteness 20%						Domain Incompleteness 40%						Domain Incompleteness 60%						Domain Incompleteness 80%					
	D	P	O	S	F_1	$CD/CP/CO$	D	P	O	S	F_1	$CD/CP/CO$	D	P	O	S	F_1	$CD/CP/CO$	D	P	O	S	F_1	$CD/CP/CO$
Baseline (gc)	11.5	0.0	0.0	1.35	0.50	0.49/0/0	4.8	0.0	0.0	1.46	0.41	0.54/0/0	3.5	0.0	0.0	1.70	0.33	0.67/0/0	2.9	0.0	0.0	1.90	0.37	0.71/0/0
Baseline ($uniq$)	11.6	0.0	0.0	1.23	0.49	0.45/0/0	4.8	0.0	0.0	1.37	0.36	0.49/0/0	3.5	0.0	0.0	1.45	0.33	0.42/0/0	2.9	0.0	0.0	1.67	0.34	0.50/0/0
$\widetilde{GC}$ (D+P+O)	11.5	1.4	14.4	1.09	0.88	0.38/0.27/-0.14	9.6	2.1	12.8	1.14	0.87	0.62/0.31/0.12	7.1	2.2	13.6	1.30	0.83	0.72/0.32/0.27	6.8	1.1	13.2	1.43	0.80	0.65/0.11/0.48
$\widetilde{GC}$ (D+O)	11.6	0.0	17.5	1.11	0.87	0.38/0/-0.04	9.6	0.0	17.0	1.15	0.87	0.62/0/0.20	7.1	0.0	18.0	1.33	0.82	0.78/0/0.51	6.8	0.0	15.6	1.48	0.80	0.66/0/0.48
$\widetilde{GC}$ (P+O)	0.0	1.4	29.7	5.62	0.31	0/-0.09/-0.40	0.0	2.1	24.1	5.03	0.35	0/-0.04/-0.29	0.0	2.2	20.4	5.13	0.35	0/0.25/-0.25	0.0	1.1	19.0	5.11	0.33	0/-0.48/-0.34
$\widetilde{GC}$ (P)	0.0	1.4	0.0	2.10	0.30	0/0.26/0	0.0	2.1	0.0	2.07	0.31	0/0.10/0	0.0	2.2	0.0	2.19	0.28	0/0.59/0	0.0	1.1	0.0	1.72	0.27	0/-0.21/0
$\widetilde{GC}$ (O)	0.0	0.0	33.0	1.25	0.84	0/0/0.038	0.0	0.0	28.5	1.28	0.84	0/0/0.30	0.0	0.0	24.9	1.47	0.78	0/0/0.52	0.0	0.0	21.4	1.63	0.77	0/0/0.47
$\widetilde{UNIQ}$ (D+P+O)	11.6	1.4	14.4	1.09	0.83	0.46/0.37/-0.09	9.6	2.1	12.8	1.10	0.80	0.35/0.34/0.21	7.1	2.2	13.6	1.23	0.76	0.84/0.56/0.46	6.8	1.1	13.2	1.37	0.74	0.64/0.18/0.56
$\widetilde{UNIQ}$ (D+O)	11.6	0.0	17.5	1.09	0.81	0.61/0/0.17	9.6	0.0	17.0	1.15	0.77	0.45/0/0.46	7.1	0.0	18.0	1.32	0.75	0.70/0/0.57	6.8	0.0	15.6	1.46	0.76	0.68/0/0.52
$\widetilde{UNIQ}$ (P+O)	0.0	1.4	29.7	4.20	0.32	0/-0.09/-0.49	0.0	2.1	24.1	3.97	0.35	0/-0.14/-0.51	0.0	2.2	20.4	3.91	0.34	0/0.18/-0.31	0.0	1.1	19.0	3.13	0.29	0/-0.56/-0.46
$\widetilde{UNIQ}$ (P)	0.0	1.4	0.0	1.25	0.26	0/-0.27/0	0.0	2.1	0.0	1.31	0.27	0/-0.42/0	0.0	2.2	0.0	1.31	0.27	0/-0.20/0	0.0	1.1	0.0	1.31	0.27	0/-0.55/0
$\widetilde{UNIQ}$ (O)	0.0	0.0	33.0	1.14	0.83	0/0/0.16	0.0	0.0	28.5	1.19	0.83	0/0/0.26	0.0	0.0	24.9	1.35	0.77	0/0/0.49	0.0	0.0	21.4	1.54	0.75	0/0/0.46

100% of Observability

	Domain Incompleteness 20%						Domain Incompleteness 40%						Domain Incompleteness 60%						Domain Incompleteness 80%					
	D	P	O	S	F_1	$CD/CP/CO$	D	P	O	S	F_1	$CD/CP/CO$	D	P	O	S	F_1	$CD/CP/CO$	D	P	O	S	F_1	$CD/CP/CO$
Baseline (gc)	11.5	0.0	0.0	1.42	0.53	0.58/0/0	4.8	0.0	0.0	1.57	0.42	0.62/0/0	3.5	0.0	0.0	1.77	0.36	0.66/0/0	2.9	0.0	0.0	1.89	0.38	0.72/0/0
Baseline ($uniq$)	11.5	0.0	0.0	1.39	0.53	0.59/0/0	4.8	0.0	0.0	1.54	0.38	0.54/0/0	3.5	0.0	0.0	1.63	0.35	0.48/0/0	2.9	0.0	0.0	1.83	0.35	0.52/0/0
$\widetilde{GC}$ (D+P+O)	11.6	1.4	21.2	1.04	0.96	0.35/0.23/0.32	9.6	2.1	19.0	1.07	0.94	0.46/0.31/0.34	7.1	2.2	19.6	1.21	0.91	0.58/0.23/0.41	6.8	1.1	19.1	1.27	0.88	0.57/0.25/0.60
$\widetilde{GC}$ (D+O)	11.6	0.0	25.5	1.05	0.95	0.35/0/0.30	9.6	0.0	24.7	1.08	0.93	0.46/0/0.36	7.1	0.0	25.7	1.23	0.90	0.61/0/0.53	6.8	0.0	22.5	1.29	0.88	0.57/0/0.60
$\widetilde{GC}$ (P+O)	0.0	1.4	42.8	6.10	0.32	0/-0.09/-0.47	0.0	2.1	35.2	5.47	0.36	0/-0.03/-0.37	0.0	2.2	29.7	5.39	0.35	0/0.27/-0.23	0.0	1.1	27.8	5.44	0.32	0/-0.46/-0.40
$\widetilde{GC}$ (P)	0.0	1.4	0.0	2.46	0.24	0/0.29/0	0.0	2.1	0.0	2.38	0.23	0/0.12/0	0.0	2.2	0.0	2.32	0.23	0/0.58/0	0.0	1.1	0.0	1.88	0.22	0/-0.17/0
$\widetilde{GC}$ (O)	0.0	0.0	47.4	1.12	0.91	0/0/0.12	0.0	0.0	41.4	1.13	0.90	0/0/0.39	0.0	0.0	35.9	1.28	0.86	0/0/0.57	0.0	0.0	31.2	1.39	0.84	0/0/0.58
$\widetilde{UNIQ}$ (D+P+O)	11.6	1.4	21.2	1.01	0.92	0.36/0.37/0.31	9.6	2.1	19.0	1.03	0.88	0.15/0.25/0.49	7.1	2.2	19.6	1.09	0.85	0.65/0.39/0.56	6.8	1.1	19.1	1.17	0.83	0.53/0.26/0.65
$\widetilde{UNIQ}$ (D+O)	11.6	0.0	25.5	1.02	0.90	0.46/0/0.53	9.6	0.0	24.7	1.07	0.87	0.22/0/0.62	7.1	0.0	25.7	1.16	0.85	0.58/0/0.59	6.8	0.0	22.5	1.23	0.85	0.55/0/0.60
$\widetilde{UNIQ}$ (P+O)	0.0	1.4	42.8	5.40	0.33	0/-0.07/-0.50	0.0	2.1	35.2	5.16	0.36	0/-0.12/-0.51	0.0	2.2	29.7	5.06	0.34	0/0.18/-0.29	0.0	1.1	27.8	5.01	0.30	0/-0.54/-0.54
$\widetilde{UNIQ}$ (P)	0.0	1.4	0.0	2.22	0.26	0/-0.27/0	0.0	2.1	0.0	2.25	0.27	0/-0.42/0	0.0	2.2	0.0	2.30	0.27	0/-0.19/0	0.0	1.1	0.0	2.31	0.27	0/-0.55/0
$\widetilde{UNIQ}$ (O)	0.0	0.0	47.4	1.05	0.89	0/0/0.33	0.0	0.0	41.4	1.09	0.90	0/0/0.32	0.0	0.0	35.9	1.23	0.85	0/0/0.51	0.0	0.0	31.2	1.35	0.82	0/0/0.53

Table 2: Performance results for ablation study, comparing the baseline approaches against our heuristics using some possible combinations of landmark types, separated by observability (10%, 30%, 50%, 70%, and 100%).

References

Amado, L.; Pereira, R.; Aires, J. P.; Magnaguagno, M.; Granada, R.; and Meneguzzi, F. 2018. Goal recognition in latent space. In *Proceedings of the 2018 International Joint Conference on Neural Networks*, IJCNN'18, 4841–4848. Washington, DC, USA: IEEE.

Asai, M., and Fukunaga, A. 2018. Classical planning in deep latent space: Bridging the subsymbolic-symbolic boundary. In *The Thirty-Second AAAI Conference on Artificial Intelligence*, 6094–6101.

Blum, A. L., and Furst, M. L. 1997. Fast Planning Through Planning Graph Analysis. *Journal of Artificial Intelligence Research (JAIR)* 90(1-2):281–300.

E.-Martín, Y.; R.-Moreno, M. D.; and Smith, D. E. 2015. A Fast Goal Recognition Technique Based on Interaction Estimates. In *Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI 2015)*.

Fikes, R. E., and Nilsson, N. J. 1971. STRIPS: A new approach to the application of theorem proving to problem solving. *Journal of Artificial Intelligence Research (JAIR)* 2(3):189–208.

Granada, R.; Pereira, R. F.; Monteiro, J.; Barros, R.; Ruiz, D.; and Meneguzzi, F. 2017. Hybrid activity and plan recognition for video streams. In *The AAAI 2017 Workshop on Plan, Activity, and Intent Recognition*.

Hoffmann, J., and Nebel, B. 2001. The FF Planning System: Fast Plan Generation Through Heuristic Search. *Journal of Artificial Intelligence Research (JAIR)* 14(1):253–302.

Hoffmann, J.; Porteous, J.; and Sebastia, L. 2004. Ordered Landmarks in Planning. *Journal of Artificial Intelligence Research (JAIR)* 22(1):215–278.

Nguyen, T. A., and Kambhampati, S. 2014. A Heuristic Approach to Planning with Incomplete STRIPS Action Models. In *Proceedings of the Twenty-Fourth International Conference on Automated Planning and Scheduling, ICAPS 2014, Portsmouth, New Hampshire, USA, June 21-26, 2014*.

Nguyen, T.; Sreedharan, S.; and Kambhampati, S. 2017. Robust Planning with Incomplete Domain Models. *Artificial Intelligence* 245:134 – 161.

Pereira, R. F., and Meneguzzi, F. 2016. Landmark-Based Plan Recognition. In *Proceedings of the 22nd European Conference on Artificial Intelligence (ECAI 2016)*.

Pereira, R. F., and Meneguzzi, F. 2017. Goal and Plan Recognition Datasets using Classical Planning Domains. Zenodo.

Pereira, R. F.; Oren, N.; and Meneguzzi, F. 2017. Landmark-Based Heuristics for Goal Recognition. In *Proceedings of the Conference of the Association for the Advancement of Artificial Intelligence (AAAI 2017)*.

Ramírez, M., and Geffner, H. 2009. Plan Recognition as Planning. In *Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence (IJCAI 2009)*.

Ramírez, M., and Geffner, H. 2010. Probabilistic Plan Recognition Using Off-the-Shelf Classical Planners. In *Proceedings of the Conference of the American Association of Artificial Intelligence (AAAI 2010)*.

Richter, S., and Westphal, M. 2010. The LAMA Planner: Guiding Cost-based Anytime Planning with Landmarks. *Journal of Artificial Intelligence Research (JAIR)* 39(1):127–177.

Richter, S.; Helmert, M.; and Westphal, M. 2008. Landmarks Revisited. In *Proceedings of the Twenty-Third National Conference on Artificial Intelligence - Volume 2*, AAAI'08, 975–982. AAAI Press.

Sohrabi, S.; Riabov, A. V.; and Udrea, O. 2016. Plan Recognition as Planning Revisited. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI 2016)*.

Sukthankar, G.; Goldman, R. P.; Geib, C.; Pynadath, D. V.; and Bui, H. H. 2014. *Plan, Activity, and Intent Recognition: Theory and Practice*. Elsevier.

Weber, C., and Bryce, D. 2011. Planning and Acting in Incomplete Domains. In *Proceedings of the 21st International Conference on Automated Planning and Scheduling, ICAPS 2011, Freiburg, Germany June 11-16, 2011*.