Represent and Infer Human Theory of Mind for Human-Robot Interaction

Yibiao Zhao, Steven Holtzen, Song-Chun Zhu
Center for Vision, Cognition, Learning, and Autonomy
University of California, Los Angeles

Abstract

This abstract is proposing a challenging problem, i.e. inferring the human mental state (intent and belief) from an observed RGBD video for human-robot interaction. The task is to integrate symbolic reasoning, a field well-studied within A.I. domains, with the uncertainty of computer vision strategies. Traditionally, A.I. relies on first-order logic and closed world assumptions which struggle to take into account the inherent uncertainty of noisy observations within a scene. Computer vision relies on pattern-recognition strategies that have difficulty accounting for higher-level reasoning and abstract representation of world knowledge. By combining these two approaches in a principled way under a probabilistic programming framework, we define new computer vision tasks such as actor intent prediction and belief inference from an observed video sequence.

Our work is largely motivated by the pioneer work of the *Theory of Mind for a Humanoid Robot* (Scassellati 2001) and a series of the cognitive science studies by Baker et al. (Baker, Tenenbaum, and Saxe 2006; Baker, Saxe, and Tenenbaum 2009; Baker and Tenenbaum 2014) about the *Bayesian Theory of Mind* (ToM), which suggests an intentional agent's behavior is based on the *principle of rationality*: the expectation that agents will behave rationally to efficiently achieve their goals given their beliefs about the world. Gergely et al. (Gergely et al. 1995; Gergely, H., and Kirly 2002) showed that infants can infer goals of varying complexity, again by interpreting an agent's behaviors as rational responses to environmental constraints.

Humans perform rational planning according to the present context. There is a strong support for this interpretation of causal inference being intimately related to how humans infer goals and intentions (Baker and Tenenbaum 2014; Baker, Saxe, and Tenenbaum 2009; Baker, Tenenbaum, and Saxe 2006). Inverse planning relies on the "principle of rationality" to make claims about an intentional agent's motions and actions: if one assumes that all actions are made with the goal of efficiently completing a goal, then it is possible to infer that goal by observing the actions. Developmental psychology studies show pre-verbal infants are able to discern rational plans from unrelated sequences of actions (Gergely et al. 1995) and other kinematic properties of human actions (Gergely, H., and Kirly 2002).

Understanding scenes and events is not a simple classification problem. As seen in several data sets (Schuldt, Laptev, and Caputo 2004; Laptev et al. 2008), action understanding algorithms in the field of computer vision have historically been formulated as discriminative learning problems. However, these data-driven algorithms only work for specific action categories with explicit visual patterns. We argue that actions are fundamentally interventions or responses to a dynamic world. As suggested by Pearl (Pearl 2009), agency and action may be intrinsically different from the underlying assumptions of classification at the philosophical level; a tree stump can become a chair if you sit on it. Action and agency in many cases "change" the world, and computer vision and reasoning systems can potentially benefit greatly by incorporating additional knowledge from rational planning and other traditional A.I. procedures.

In this abstract, we consider stochastic inverse action planners as *generative probabilistic programs* (Goodman et al. 2008; Mansinghka et al. 2013) following the generative thinking of cognitive models (Goodman and Tenenbaum ; Tenenbaum et al. 2011). By applying these methods to highly uncertain computer vision tasks we hope to understand scenes and events in a more holistic way than has previously been explored.

Problem Overview

The problem involves an observer cataloging a sequence of actions and attempting to infer the hidden causes and predict subsequent actions for an actor. In Fig.1, a robot with a RGBD camera an observer that is attempting to understand why the human is moving through the room. The hidden cause which explains the human's actions is the theory of mind, which consists of an actor's intentions and beliefs. A representation layer is required in order to bridge the gap between the highly uncertain visual domain and the logical higher-order reasoning domains.

An actor's **belief** about a scene is represented by a mental structure that tracks the location and status of objects. We define the physical layout of a scene as a hierarchical spatial parse tree (SPT), which describes the beliefs of an agent within the scene. The SPT is a spatial organization of the scene; for example, if a cup node is a child of a table

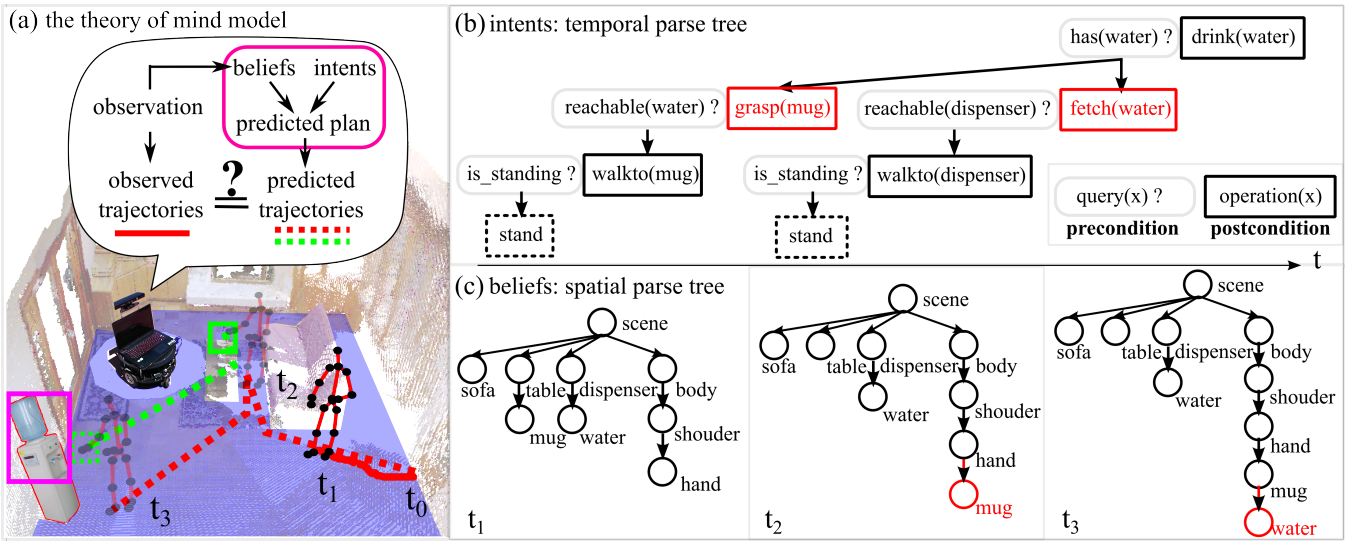

Figure 1: (a) An illustration of the hierarchal theory of mind (Hi-TOM). The human intent and belief are the hidden causes drive the change of the world. With the intent and the belief, the theory of mind, boxed in magenta, predicts plans and synthesizes motion trajectories (dash lines), the algorithm evaluate the likelihood of the Hi-TOM by comparing predicted motion trajectories with partially observed trajectories (solid lines). (b) shows the hierarchical structure of intent is represented by a temporal parse tree. An intent "drink water" is stochastically factorized into smaller plans by checking preconditions of the world as it progress through time. The terminal nodes of parse tree is called operations, which result in the changes to the world state as shown in (c). As time progresses, the spatial parse tree (i.e. actor's beliefs) evolves, informing future decision-making in (b). As a result, the predicted plan encoded by (b) is walkto(mug) $\rightarrow$ grasp(mug) $\rightarrow$ walkto(dispenser) $\rightarrow$ fetch(water) $\rightarrow$ drink(water), resulting in the transferring the mug from the table to the hand at t_2 and the filling of the mug with water at t_3.

node, then the cup is physically supported by the table. This semantic graph structure can be used to answer relational queries about the state of the scene, such as whether or not a mug is reachable (as demonstrated in Fig.1(a)). The observer is uncertain about the precise state of the scene due to the inaccuracies in computer vision models, necessitating the usage of a probabilistic model to describe the state of the scene. For example the status of a cup *has_water*, could be unknown to the observer. A probabilistic program would model this as a `flip` function with stochastic memoization to define the actor's belief.

A similar hierarchical representation, temporal parse tree, is used to model the actor's **intent** as shown in Fig.1(b). The temporal parse tree describes a hierarchical decomposition of events/actions that an actor takes in order to accomplish a goal of*drink_water*. This decomposition process is described by a hierarchical task network (HTN) (Erol, Hendler, and Nau 1994; Nguyen et al. 2007) in this work, which specifies not only, a network of preconditions that must be satisfied in order for an action to be made. Actions are constructed out of composable probabilistic programs, as seen in Fig.1(b). Actions are thus probabilistic programs: given an uncertain representation of the scene from a computer vision algorithm, the program must determine the action that must be used in order to gain the desired effect.

The two main benefits of probabilistic programs are (a) the ability to cleanly separate one's model from one's inference strategy and (b) the ability to represent arbitrarily complex random variables that can then be inferred by an independent model. These two features are crucial for modeling and inferring plans, which are extremely complex and rely on a variety of inference strategies. By separating out the inference strategies, it is significantly easier to adapt to new scenarios in which the previous strategy may no longer work. By being able to represent complex plans succinctly, it makes it easier to express and learn plans from observations.

The problem formulation is thus to infer one aspect of the theory of mind given the other two. For example, given the beliefs and actions of an actor, it is possible to determine that actor's intent under the rational theory of mind; probabilistic programs act as a bridge between the uncertain domain of computer vision and the logical reasoning domain of A.I. planners.

In order to reverse-engineer people' rational planning, we formulate action interpretation tasks in terms of sampling and prediction (approximately) from the posterior distribution:

$$P(Y|X_{obs}) \propto P(Y)\delta_{f(Y,B)}(X_{pred})P(X_{obs}|X_{pred}) \quad (1)$$

In our model, we consider an actor's beliefs and intentions as *generative probabilistic programs* (Goodman et al. 2008; Mansinghka et al. 2013) following the generative thinking of cognitive models (Goodman and Tenenbaum ; Tenenbaum et al. 2011).

Hierarchical mental states Y (namely, a parse tree) are defined on the grammar $P(Y)$. The grammar recursively decomposes the parse tree into goals and sub-goals. Sampling

methods (such as rejection sampling) produce a sequence of actions that satisfy the constraints specified in the grammar.

The relational planner (for example, an HTN) simulates rational behaviors $f(Y, B) \rightarrow X_{pred}$. The variable B represent the background collision map.

An approximate planner (based on the RRT* algorithm (Karaman et al. 2011)) is used to generate approximate rational plans in a complex environment. These two programs span the probabilistic space of all the possible actions.

A predicted mental status Y sampled from the problem state according to $P(Y)$ could be translated into a sequence of actions X_{pred} by an HTN planner. This would be extremely unlikely to exactly match the observed behavior X_{ons}. Instead of requiring exact matches, our formulation relaxes the matching problem by a stochastic likelihood $P(X_{obs}|X_{pred})$. We implement this by applying a Dynamic Time Warping (DTW) algorithm (Vintsyuk 1968), which measures the similarity between two temporal sequences which may vary in time or speed. The DTW algorithm outputs the shortest mean distance as well as matching correspondences between two matched sequences. The shortest mean distance is fed to the stochastic likelihood function in the form of a simple Gaussian function. The Gaussian variance controls the tolerance of the model. The output matching correspondences provide the detailed parsing for each frame of the observed sequence.

Consequences and Future Research

Modeling theory of mind is inherently highly cross-disciplinary, incorporating many aspects of computer vision, planning, first order logic, linguistics, and many other fields. The consequences of fully inferring the theory of mind of an actor in a scene are immense: once one can reason about an actor, it will be possible to understand people's need and motivation, and thus assist human via natural interaction.

Many challenges must be met in order for this model to succeed. While probabilistic programs offer great flexibility in the representation of events, they lack the academic rigor of more well-understood methods like propositional logic and context-free grammars. This makes it difficult to make claims about whether or not an inference algorithm is optimal or correct.

The scope of the problem is potentially vast, making efficient inference important. As the number of plans grows, the number of potential explanations for a sequence of actions grows extraordinarily quickly, especially if the actor is executing multiple plans concurrently. Inferring interleaved plans and higher-order plans remains an untouched problem with far-reaching consequences for linguistics and computer vision.

References

[Baker and Tenenbaum 2014] Baker, C., and Tenenbaum, J. 2014. Modeling human plan recognition using bayesian theory of mind. In Sukthankar, G.; Goldman, R.; Geib, C.; Pynadath, D.; and Bui, H., eds., *Plan, Activity, and Intent Recognition*. Elsevier.

[Baker, Saxe, and Tenenbaum 2009] Baker, C.; Saxe, R.; and Tenenbaum, J. 2009. Action understanding as inverse planning. *Cognition. 2009 Dec* 113:329–349.

[Baker, Tenenbaum, and Saxe 2006] Baker, C.; Tenenbaum, J.; and Saxe, R. 2006. Bayesian models of human action understanding. In *NIPS*, volume 18, 99–106.

[Erol, Hendler, and Nau 1994] Erol, K.; Hendler, J.; and Nau, D. 1994. Htn planning: Complexity and expressivity. In *AAAI*.

[Gergely et al. 1995] Gergely, G.; Nfidasdy, Z.; Csibra, G.; and Br, S. 1995. Taking the intentional stance at 12 months of age. *Cognition* 56:165–193.

[Gergely, H., and Kirly 2002] Gergely, G.; H., B.; and Kirly, I. 2002. Rational imitation in preverbal infants. *Nature* 415:755.

[Goodman and Tenenbaum] Goodman, N. D., and Tenenbaum, J. B. *Probabilistic Models of Cognition (electronic)*.

[Goodman et al. 2008] Goodman, N.; Mansinghka, V.; Roy, D.; Bonawitz, K.; and Tenenbaum, J. 2008. Church: A language for generative models. In *UAI*, 220–229.

[Karaman et al. 2011] Karaman, S.; Walter, M.; Perez, A.; Frazzoli, E.; and Teller, S. 2011. Real-time motion planning using the RRT*. In *ICRA*.

[Laptev et al. 2008] Laptev, I.; Marszalek, M.; Schmid, C.; and Rozenfeld, B. 2008. Learning realistic human actions from movies. In *CVPR*.

[Mansinghka et al. 2013] Mansinghka, V.; Kulkarni, T.; Perov, Y.; and Tenenbaum, J. 2013. Approximate bayesian image interpretation using generative probabilistic graphics programs. In *NIPS*.

[Nguyen et al. 2007] Nguyen, N. T.; Grzech, A.; Howlett, R.; and Jain, L. 2007. Expressivity of strips-like and htn-like planning. In *KES-AMSTA*, volume 4496 of *Lecture Notes in Computer Science*, 121–130. Springer.

[Pearl 2009] Pearl, J. 2009. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition.

[Scassellati 2001] Scassellati, B. M. 2001. *Foundations for a Theory of Mind for a Humanoid Robot*. Ph.D. Dissertation, Massachusetts Institute of Technology.

[Schuldt, Laptev, and Caputo 2004] Schuldt, C.; Laptev, I.; and Caputo, B. 2004. Recognizing human actions: A local svm approach. In *ICPR*.

[Tenenbaum et al. 2011] Tenenbaum, J.; Kemp, C.; Griffiths, T. L.; and Goodman, N. 2011. How to grow a mind: Statistics, structure, and abstraction. *Science* 331 (6022):1279–1285.

[Vintsyuk 1968] Vintsyuk, T. K. 1968. Speech discrimination by dynamic programming. *Cybernetics* 4:52–57.