Learning Human Types from Demonstration

Stefanos Nikolaidis, Keren Gu, Ramya Ramakrishnan, and Julie Shah
Massachusetts Institute of Technology,
77 Massachusetts Avenue, Cambridge, MA 02139
Email: snikol@alum.mit.edu, kgu@mit.edu, ramyaram@mit.edu, julie_a_shah@csail.mit.edu

Introduction

The development of new industrial robotic systems that operate in the same physical space as people highlights the emerging need for robots that can integrate seamlessly into human group dynamics by adapting to the personalized style of human teammates. This adaptation requires learning a statistical model of human behavior and integrating this model into the decision-making algorithm of the robot in a principled way. We present a framework for automatically learning human user models from joint-action demonstrations that enables the robot to compute a robust policy for a collaborative task with a human, assuming access to demonstrations of human teams working on the task. The robustness of the action selection mechanism of the robot is compared to previous model-learning algorithms in the ability to function despite increasing deviations of human actions from previously demonstrated behavior.

Related Work

For a robot to learn a human model, a human expert is typically required to explicitly teach the robot a skill or specific task (Argall et al. 2009; Atkeson and Schaal 1997; Abbeel and Ng 2004; Nicolescu and Mataric 2003; Chernova and Veloso 2008; Akgun et al. 2012). In this work, demonstrations of human teams executing a task are used to automatically learn human types in an unsupervised fashion. This allows rapid estimation of a human user model, which can be done either offline or online, through the a priori learning of a set of ”dominant” models. This differs from previous approaches (Doshi and Roy 2007) that start with uncertain model parameters and learn them through interaction. Such approaches do not have the limitation of a fixed set of available models, however learning a good model requires a very large amount of data, which can be an issue when using them for practical applications. We present a pipeline to automatically learn the reward function of a Mixed-Observability Markov Decision Process through unsupervised learning and inverse reinforcement learning (Abbeel and Ng 2004). Using MOMDPs to compute personalized policies has been used in prior work (Ong et al. 2010), (Bandyopadhyay et al. 2013), but with the reward structure assumed to be known. Research on POMDP formulations for collaborative tasks in game AI applications (Nguyen et al. 2011; Macindoe, Kaelbling, and Lozano-Pérez 2012; Silver and Veness 2010) also assumed a known human model. Additionally, previous partially observable formalisms (Ong et al. 2010; Bandyopadhyay et al. 2013; Broz, Nourbakhsh, and Simmons 2011; Fern and Tadepalli 2010; Nguyen et al. 2011; Macindoe, Kaelbling, and Lozano-Pérez 2012) in assistive or collaborative tasks represented the preference or intention of the human for their own actions, rather than those of the robot, as the partially observable variable.

Method

Our proposed framework has two main stages, as shown in Figure 1. The training data is preprocessed in the first stage. In the second stage, the robot infers the personalized style of a new human teammate and executes its role in the task according to the preference of this teammate.

The first stage of our framework assumes access to a set of demonstrated sequences of actions from human teams working together on a collaborative task, and uses an unsupervised learning algorithm to cluster the data into dominating human types. The cluster indices serve as the values of a partially observable variable denoting human type, in a MOMDP (Ong et al. 2010). Our framework then employes an inverse reinforcement learning algorithm (Abbeel and Ng 2004) to learn a reward function for each human type, which represents the preference of a human of the given type on a subset of task-related robot actions. Finally, the framework computes an approximately optimal policy for the robot that reasons over the uncertainty on the human type and maximizes the expected accumulated reward.

In the second stage, a new human subject is asked to execute the collaborative task with the robot. The human is first instructed to demonstrate a few sequences of human and robot actions. A belief about his type is then computed according to the likelihood of the human sequences belonging to each cluster. Alternatively, if the human actions are informative of his type —his preference for the actions of the robot —the human type can be estimated online. The robot then executes the action based on the computed policy of the MOMDP, based on the current belief of the human type, at each time step.

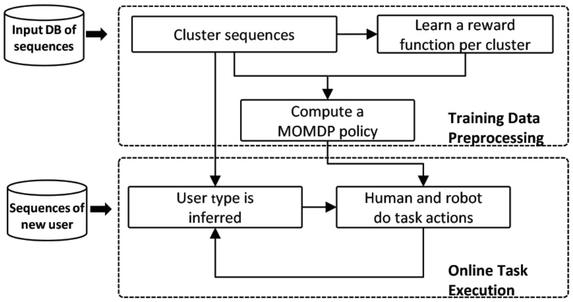

Figure 1: Framework flowchart

Evaluation

In this section, we show the applicability of the proposed framework on a place-and-drill task, using joint-action demonstrations from 18 human subjects. The role of the human was to place screws in one of three available positions, while the robot was to drill each placed screw. The demonstrations were provided during a training phase in which the human and robot switched roles, giving the human the opportunity to demonstrate robot drilling actions to show the robot how he would like the task to be executed. To evaluate our framework, we used leave-one-out cross-validation, by removing one subject and using the demonstrated sequences from the remaining 17 subjects as the training set. In all cross-validation iterations, the human subjects were clustered into two types: a “safe” type, in which each screw was placed before drilling began, and an “efficient” type, in which each screw was drilled immediately after placement. For each type, our framework learns a reward function associated with that type. The number of types and their associated reward functions is then passed to the MOMDP formulation as input.

Each subject left out of the training set for cross-validation - referred to as the “testing subject” - provided three demonstrated sequences of human and robot actions and a probability distribution over its type was calculated. Using this as the initial belief on the human type, and the associated reward function from the inverse reinforcement learning algorithm, a MOMDP/SARSOP (Kurniawati, Hsu, and Lee 2008) solver computed a policy for the robot. We then had the testing subject execute the place-and-drill task with the actual robot, with each performing their predefined roles, during the “task execution phase” (Figure 2).

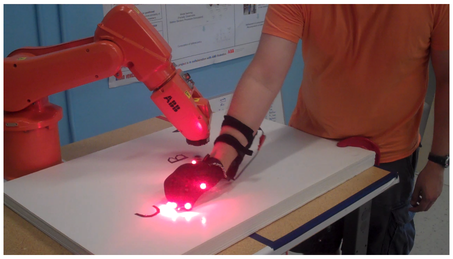

Figure 2: Task execution by a human-robot team on a place-and-drill task.

Robustness and Quality of Learned Policy

We compared the computed policy with a state-of-the-art iterative algorithm for human-robot collaborative tasks, called “human-robot cross-training” (Nikolaidis and Shah 2013), in which the robot learns a human model by switching roles with the human. We used the demonstrated sequences of the testing subject as input for the cross-training algorithm, which computes a policy which matches the human preference during task execution when the human and roboct resume their predefined roles (Nikolaidis and Shah 2013). In the actual human subject data, the human placement actions during task execution were, in most cases, identical to those provided during the demonstrations. Therefore, we simulated the task execution for increasing degrees of deviations from the demonstrated actions of the human, leading the execution to previously unexplored parts of the state-space. We did this by having a simulated human perform a random placement action with a probability ϵ, or the actual action taken by the testing human subject with probability $1 - \epsilon$. For increasing levels of deviations, we computed the accumulated reward for the policy of the proposed framework and the policy computed by the human-robot cross-training algorithm (Figure 3).

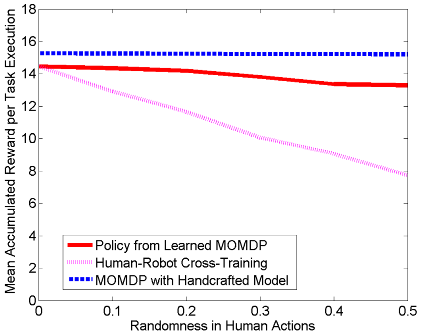

Figure 3: Accumulated reward averaged over 18 iterations of cross-validation (one for each human subject), and over 100 simulated iterations of task execution. The plotted lines illustrate the performance of the different policies. The x-axis represents the probability ϵ of the human performing a random action instead of replaying the actual action taken.

The policy of the human-robot cross-training algorithm performed similarly to the one of the proposed framework if the user did not deviate from his demonstrated placing actions. However, as the deviations increased, the policy from the cross-training algorithm performed worse. On the other hand, the MOMDP agent reasons over the partially observable human type using a reward function that learns from all demonstrated sequences that belong to the cluster associated with that type; therefore, its performance was not affected by these deviations. Figure 3 also shows that the policy computed by the MOMDP using the automatically generated user model has comparable performance to the one using a hand-coded model from a domain expert. The plotted lines denote the accumulated reward, averaged over all iterations of cross-validation.

References

Abbeel, P., and Ng, A. Y. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proc. ICML.* ACM Press.

Akgun, B.; Cakmak, M.; Yoo, J. W.; and Thomaz, A. L. 2012. Trajectories and keyframes for kinesthetic teaching: a human-robot interaction perspective. In *HRI*, 391–398.

Argall, B. D.; Chernova, S.; Veloso, M.; and Browning, B. 2009. A survey of robot learning from demonstration. *Robot. Auton. Syst.* 57(5):469–483.

Atkeson, C. G., and Schaal, S. 1997. Robot learning from demonstration. In *ICML*, 12–20.

Bandyopadhyay, T.; Won, K. S.; Frazzoli, E.; Hsu, D.; Lee, W. S.; and Rus, D. 2013. Intention-aware motion planning. In *Algorithmic Foundations of Robotics X*. Springer. 475–491.

Broz, F.; Nourbakhsh, I.; and Simmons, R. 2011. Designing pomdp models of socially situated tasks. In *RO-MAN, 2011 IEEE*, 39–46. IEEE.

Chernova, S., and Veloso, M. 2008. Teaching multi-robot coordination using demonstration of communication and state sharing. In *Proc. AAMAS*.

Doshi, F., and Roy, N. 2007. Efficient model learning for dialog management. In *Proc. HRI*.

Fern, A., and Tadepalli, P. 2010. A computational decision theory for interactive assistants. In *Interactive Decision Theory and Game Theory*.

Kurniawati, H.; Hsu, D.; and Lee, W. S. 2008. Sarsop: Efficient point-based pomdp planning by approximating optimally reachable belief spaces. In *Robotics: Science and Systems*, 65–72.

Macindoe, O.; Kaelbling, L. P.; and Lozano-Pérez, T. 2012. Pomcop: Belief space planning for sidekicks in cooperative games. In *AIIDE*.

Nguyen, T.-H. D.; Hsu, D.; Lee, W. S.; Leong, T.-Y.; Kaelbling, L. P.; Lozano-Perez, T.; and Grant, A. H. 2011. Capir: Collaborative action planning with intention recognition. In *AIIDE*.

Nicolescu, M. N., and Mataric, M. J. 2003. Natural methods for robot task learning: Instructive demonstrations, generalization and practice. In *Proc. AAMAS*, 241–248.

Nikolaidis, S., and Shah, J. 2013. Human-robot cross-training: computational formulation, modeling and evaluation of a human team training strategy. In *Proc. of HRI*.

Ong, S. C.; Png, S. W.; Hsu, D.; and Lee, W. S. 2010. Planning under uncertainty for robotic tasks with mixed observability. *IJRR* 29(8):1053–1068.

Silver, D., and Veness, J. 2010. Monte-carlo planning in large pomdps. In *Advances in Neural Information Processing Systems*, 2164–2172.