Simultaneous Acquisition of Task and Feedback Models

Manuel Lopes, Thomas Cederborg, Pierre-Yves Oudeyer

To cite this version:

Manuel Lopes, Thomas Cederborg, Pierre-Yves Oudeyer. Simultaneous Acquisition of Task and Feedback Models. Development and Learning (ICDL), 2011 IEEE International Conference on, 2011, Germany. pp.1 - 7, 10.1109/DEVLRN.2011.6037359 . hal-00636166

HAL Id: hal-00636166
https://hal.science/hal-00636166

Submitted on 26 Oct 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Simultaneous Acquisition of Task and Feedback Models

Manuel Lopes, Thomas Cederborg and Pierre-Yves Oudeyer
INRIA, Bordeaux Sud-Ouest, France

Abstract— **We present a system to learn task representations from ambiguous feedback. We consider an inverse reinforcement learner that receives feedback from a teacher with an unknown and noisy protocol. The system needs to estimate simultaneously what the task is (i.e. how to find a compact representation to the task goal), and how the teacher is providing the feedback. We further explore the problem of ambiguous protocols by considering that the words used by the teacher have an unknown relation with the action and meaning expected by the robot. This allows the system to start with a set of known signs and learn the meaning of new ones. We present computational results that show that it is possible to learn the task under a noisy and ambiguous feedback. Using an active learning approach, the system is able to reduce the length of the training period.**

I. INTRODUCTION

Developing robots that can interact and live among people is a very challenging problem due to the complex environments and the difficulties in interacting with humans. Two main paradigms are used to address these problems: *artificial mental development* [1] where complexity is increased guided by a developmental program; and *social learning* where the existence of other people in the environment is exploited by learning new goals/tasks from observation and by using their feedback. In this paper, we will show how a robot can follow a developmental program to learn new tasks and even some rudimentary language skills. The learner will start to acquire new skills from observation and teacher feedback. After some initial learning the learner is able to acquire a model of the teacher feedback behavior that allows to improve the task model. Finally, using such knowledge it is possible to acquire the meaning of new signs that will again increase the quality of the task and feedback model.

Learning from demonstration has provided several examples of efficient learning in robotic systems [2], [3]. Data from a teacher has been used as: initial condition for further self-exploration in robotics [4], information about the task solution [5], information about the task representation [6], among others. Several representations have been used to generalize the demonstration data using reinforcement learning [7], inverse reinforcement learning [8], [6] or regression methods [9], [10], [5]. The different formalisms make use of different information and extract different knowledge, either direct policy information or a reward function that explains the behavior.

Another feature of most of those systems is that the data is provided in a batch perspective where data acquisition is done before the learning phase. Recently it has been suggested that *interactive learning* [4], [11] might be a new perspective on robot learning that combines the ideas of learning by demonstration, learning by exploration and tutor feedback. Under this approach the teacher interacts with the robot and provides extra feedback. Approaches have considered: extra reinforcement signals [7], action requests' [9], [12], disambiguation among actions [10], preferences among states [13], iterations between practice and user feedback sessions [14] and choosing actions that maximize the user feedback [15], [16]. In [17] the authors compare the results when the robot has the option of asking or not the teacher for feedback.

Several studies discuss the different behaviors naive teachers use when instructing robots [7], [18]. An important aspect is that, many times, the feedback is ambiguous and deviates from the mathematical interpretation of a reward or a sample from a policy. For instance, in the work of [7] the teachers frequently gave a reward to exploratory actions even if the signal was used as a standard reward. Also, in some problems we can define an optimal teaching sequence but humans do not behave according to those strategies [18]. The system in [19] automatically learns different interaction protocols for navigation tasks where the robot learns the actions it should make and which gestures correspond to those actions.

In this work we consider a setting where the robot must learn a task description (in the form of a reward function) from interacting with a teacher that provides feedback signals. We extend previous approaches by learning simultaneously how the feedback is being provided and what is the meaning of the teacher's feedback signs. Note that we will call what the teacher says/writes *sign or feedback sign* and the meaning of the sign *feedback*. In a human-robot interaction setting we consider the case where the robot tries an action and then receives a feedback signal from the teacher. Such feedback is not restricted to a pre-defined protocol, with a pre-defined set of signs or words, but should allow for new interaction types and instruction commands. The teachers will also provide signals not expected by the robot. A simple case is when the teacher gives synonyms of feedback words.

Our contributions are: a) a learning by demonstration system that learns a task description based on noisy feedback, b) an interactive learning system with a loosely defined

---

The authors are with the Flowers Team at INRIA, France. Contact email: `manuel.lopes@inria.fr`. Work (partially) supported by INRIA, Conseil Régional d'Aquitaine and the ERC grant EXPLORERS 24007.

protocol in terms of accepted words and their use, and c) an online learning system that estimates simultaneously the task, the feedback protocol and the sign-meaning relations. We assume that the robot is initially equipped with a set of sensory-motor skills and knowledge of some feedback signs. The state space is assumed to be continuous, the set of actions and feedback meanings are finite and the feedback signs can grow infinitely.

The experimental protocol we used is the following. The robot samples a state and tries an action on that state. The teacher has the possibility of providing the robot with a feedback signal. Those signal can refer to the name of the correct action to be used or by explicitly saying if an action is correct or wrong. Our framework is generic and the signal provided by the teacher can refer to the uttered words, gestures, facial expression or even the prosody of speech. By iteratively following this process, the system will learn the task representation. This system is different from typical learning by demonstration systems because the data is acquired in an interactive, and online, setting and not in batch. It is different from previous learning by interaction systems in that the feedback signals received have a much looser protocol and might make use of unknown signs.

In the next Section we provide the details of the algorithm, including a summary of Bayesian inverse reinforcement learning and an active learning extension. Finally we present simulations of our system and conclusions.

II. INVERSE REINFORCEMENT LEARNING WITH AMBIGUOUS FEEDBACK

In this section we present our learning algorithm. Our problem can be divided in three smaller ones: a) learn the task representation; b) learn how the teacher provides the feedback on the executed actions; and c) learn the meaning of novel feedback signals. We remember that the *feedback* is what the teacher means and the *sign or feedback sign* is what it "says/writes/gestures".

A. Bayesian Inverse Reinforcement Learning

We consider a standard *markov decision process* (MDP) and follow the notation of [20]. An MDP is defined by a state and action space X and A respectively, a reward function R and a state transition model P. A policy, $\pi(x, a)$, is a function that attributes a probability of selecting an action in each state and the function $r(x, a)$ gives the reward the agent receives when choosing the action a in state x. The goal of reinforcement learning is to find the optimal policy π^*, that is defined as the ones that maximizes the total discounted reward, i.e. $R = \sum_{t=0}^{\infty} \gamma^t r_t$, with γ a discount factor and r_t the reward received at time t. We define the Q^π-function as the value of taking an action at a given state when following policy π, i.e. $Q^\pi(x, a) = E_\pi \left(r(x, a) + \gamma \sum_y P^a_{xy} \max_b Q^\pi(y, b) \right)$, where $P^a_{xy} = p(x_{t+1} = y|x_t = x, a_t = a)$ is the probability of reaching state y when the current state is x and the chosen action is a.

In our case we are not interested in learning a task by self-exploration but will use data from a teacher to learn the representation of the task the teacher wants the learner to acquire. In this situation we do not have a reward function from which we can get samples but have instead samples from the policy, i.e. we do not have a reward but have actions. This formalism is called the *inverse reinforcement learning* (IRL) problem [21]. The goal is to find the reward function that the teacher is trying to maximize and later on use it to select the best actions.

Using a Bayesian perspective, we follow the *Bayesian IRL* approach (BIRL)[22]. In that setting we consider that, if the teacher is performing the task described by the reward function r, the samples of the demonstration are generated by:

$$p(x, a|r) = \frac{e^{\eta Q(x,a)}}{\sum_b e^{\eta Q(x,b)}}$$

where η is a confidence parameter where high values correspond to the optimal policy and lower values allow samples of non-optimal actions. We assume a uniform state sampling. For numerical purposes it is convenient to rewrite that expression by considering the summed probability of all the optimal actions (A^*) as:

$$p(x, a|r) = \begin{cases} \frac{\sum_{a \in A^*} e^{\eta Q(x,a)}}{\sum_b e^{\eta Q(x,b)}} & if\ a \in A^* \\ \frac{e^{\eta Q(x,a)}}{\sum_b e^{\eta Q(x,b)}} & if\ a \notin A^* \end{cases}$$

To have a normalized probability distribution we have to consider all optimal actions as a single one. To learn the reward we compute the posterior distribution of the reward function after observing a given data vector $D_t = \{A_{0:t}, X_{0:t}\}$:

$$p(R_{t+1}|A_{0:t}, X_{0:t}) \quad \propto \quad p(A_t|R_t, X_t)p(R_t) \qquad (1)$$

for a suitable choice of prior distribution on R, see [22]. The process of computing this posterior distribution is computationally intensive. We implement it with a filtering perspective [23]. We consider that the reward function is a linear combination of basis functions $\phi(x)$ in the following way $R = w^t\phi(x)$. Then, we estimate not the posterior of the parameter w of the mixture, but the posterior of the activation of each feature vector. An intuitive way to see this is to assume that each sample point is generated from a policy corresponding to a single feature vector. Under this perspective the mean of the feature distribution is the best estimation for the reward function.

B. Feedback Model

Now, the learner must infer what the task representation is and how the feedback is being provided. In this section we consider that the signs provided by the teacher have a known relation with the feedback meaning, next subsection will relax this assumption. The difference compared to the standard setting is that the demonstration is not given as a sequence of state-action pairs but as feedback on those pairs. For a given state action pair (x, a) we consider the probability of receiving a given feedback signal f.

If the robot performs the correct action, the teacher might say nothing, might verbalize the correct action to reinforce it or acknowledge that it was the correct action. If the learner performs the wrong action the teacher might say "error", just verbalize the correct action, or say nothing. In all circumstances the learner perceives the feedback with noise and so it can even hear the wrong feedback. Table I shows all the possible feedback protocols that can range from a pure learning from demonstration behavior (protocol 1) to a pure binary reinforcement one (protocol 8). Each protocol is defined with the feedback that the teacher provides the learner when it does the correct action and when it does the wrong action. The teacher might choose to say the correct action (A), say nothing ($\emptyset$), give a confirmation (O) or inform the learner that the selected action is wrong (W). This protocol is ambiguous and the same feedback ($\emptyset$) can either mean correct or incorrect. If more than one correct action is available in a state then the teacher provides, randomly, one of them. To model perceptual errors there is a probability of receiving a random sign instead of the correct one. The only restriction we have in the protocol is that a (W) message after a correct action is made or an acknowledge (O) when a wrong action is executed are only given with low probability. These assumptions model the perceptual noise of the learner and give a small bias that improves the convergence of the algorithm by disambiguating the different protocols. More general protocols could be considered, but for computational efficiency we reduced to a small set that allows the implementation of an efficient filter.

TABLE I

THE 8 FEEDBACK PROTOCOLS CONSIDERED. POSSIBLE FEEDBACK INSTRUCTIONS GIVEN BY THE TEACHER WHEN THE LEARNER DOES THE **CORRECT** OR **WRONG** ACTION ARE: THE ACTION NAME (A), NOTHING ($\emptyset$), CORRECT (O) OR WRONG (W).

Feedback / Action	1	2	3	4	5	6	7	8
Correct	A	A	A	$\emptyset$	$\emptyset$	O	O	O
Wrong	A	$\emptyset$	W	A	W	A	$\emptyset$	W

Each different teacher that will be modeled as a convex combination of these protocols. For the teacher model we will consider a set of parameters M that describe the mixture of protocols in Table I. We do this to be able to explain more teacher behaviors than just the predefined models, this is specially important when we do not know the level of noise on each protocol. As an example, consider $M = [0\ 0.8\ 0\ 0\ 0\ 0.2\ 0\ 0]$, the statistical model for the feedback is as follows:

$$\text{if A is optimal} \quad \begin{cases} p(F = A|A, M) & = 0.8 \\ p(F = O|A, M) & = 0.2 \end{cases}$$

$$\text{if A is non-optimal} \quad \begin{cases} p(F = \emptyset|A, M) & = 0.8 \\ p(F = A|A, M) & = 0.2 \end{cases}$$

This combines 80% of the time a teacher that reinforces the behavior of the learner when it is correct by providing the correct action and says nothing when the action is wrong, and 20% of the time a teacher that confirms that the chosen action was correct or provides it when the learner chooses it wrong.

We have to extend the model in Eq. 1 to include the ambiguous feedback. Our posterior now depends not only on the demonstration but also on the feedback model. By independence we can get the following factored model:

$$\begin{aligned} & p(R_{t+1}, M_{t+1}|A_{0:t}, F_{0:t}) \\ & \quad \propto \quad p(F_t|A_t, R_t, M_t)p(R_t, M_t|A_t) \\ & \quad \propto \quad p(F_t|A_t, R_t, M_t)p(A_t|M_t, R_t)p(R_t, M_t) \\ & \quad = \quad p(F_t|A_t, R_t, M_t)p(A_t|R_t)p(R_t, M_t) \end{aligned} \tag{2}$$

C. Sign-Meaning Model

Another aspect of human-robot interaction systems is that the feedback is often given using a natural interface such as gestures or speech. Most of the times there is an implicit assumption that the vocal signs are assumed to have a known semantics for the learner. Now, we will relax this assumption and allow the teacher to provide instructions to the learner that are unknown. We will define the feedback as the instruction the teacher wants to provide to the learner, as defined in Table I, and the signs as the words actually provided by the teacher. In this way it is possible for the learner to accept new words and learn their meanings. As an example, the teacher might say "good", or "ok", or "correct" and the learner should always understand it as a confirmation, i.e. the different signs all correspond to the same feedback.

Feedback		
Signs		Meanings
Known	up	↑
	down	↓
	left	←
	right	→
	$\emptyset$	CORRECT/WRONG
	ok	CORRECT
	error	WRONG
Unknown	good	?
	bad	?
	⋮	?

Fig. 1. Relation between feedback signs and intended feedback meaning. There are only $Na + 3$ feedback meanings, one corresponding to each available action and the meanings of CORRECT and WRONG. They are fixed and known from the beginning. We assume that there is at least one feedback signal with a known correspondence to a feedback signal, there is the possibility of unknown feedback signs to exist and their relation to the feedback must be learned. For instance the teacher might say good instead of ok. The table shows an example when the agent has 4 available actions (up, down, left and right).

We have to extend the previous feedback model, in Equation 2, to include the uncertainty in the signs received. We will consider a new relation that gives the probability of having a feedback sign g when the teacher wants to provide a given feedback f, $p(g|f, .)$. As the feedback is no longer observed, we have to integrate it out from the observation of the feedback. Finally, we get the following expression:

$$p(G_{t+1}|D_t) = \sum_g p(G_t|F_t)p(g|D_t) \tag{3}$$

This posterior distribution on the sign-meaning vector can also be implemented as a particle filter.

D. Algorithm

The algorithm involves the estimation of three entities from data: the reward, the feedback model and the meanings of the feedback signs. We will use a particle filter to estimate all the variables of interest. To reduce the number of particles we will not represent the full joint distribution but only an approximate of each marginal. We update the weight of each particle taking into account the *maximum a-posteriori* estimate of the other variables. Table II summarizes the algorithm.

TABLE II

ALGORITHM FOR THE JOINT ESTIMATION OF THE TASK REPRESENTATION, FEEDBACK AND SIGN-MEANING MODELS. IT COMBINES THREE PARTICLE FILTERS TO APPROXIMATE THE POSTERIOR DISTRIBUTION OF THE THREE VARIABLES.

```
• Select number of samples $n_r$, $n_g$ and $n_m$
• Sample $n_r$ reward vectors
• Sample $n_g$ sign-meaning parameters
• Sample $n_m$ protocol parameters
    1) Sample state $x$
    2) Choose and execute action $a$
    3) Observe feedback sign $g_t$
    4) Sample feedback from $f_t \sim p(f|g_t)$
    5) Find best feedback parameters $M = argmax_i w_f^{(i)}$
    6) $w_r^{(i)} \leftarrow p(f_t|A_t, R_t^i, M)p(A_t|R_t)w_r^{(i)}$
    7) Resample reward particles
    8) Find best reward parameters $r^* = argmax_i w_r^{(i)}$
    9) $w_f^{(i)} \leftarrow p(f_t|A_t, r^*, M_t)p(A_t|r^*)w_f^{(i)}$
   10) Resample feedback model
   11) $w_g^{(i)} \leftarrow \sum_i p(g_t|f_t)w_g^{(i)}$
   12) Resample sign-meaning model
   13) goto 1
```

E. Active Sampling

The previous algorithm keeps an approximation of the posterior distribution of the reward function. We can use this information to allow the learner to ask the teacher for more informative samples. We do not consider any intrinsic motivation on the system [24] besides that of reducing uncertainty. From the reward distribution it is difficult to decide what state, or action, provides more information. We can follow the active learning extension for IRL as presented in [12], or alternatively [25], to allow the learner to request the most informative samples. In that approach the policy distribution is inferred from the distribution on the rewards. Then, for each state, a measure of the uncertainty is made to select the state where the policy posterior has higher variance. Intuitively, this state is the state where the rewards agree least.

The criteria used is, for each state, the variance of the weighted sum of all policies.

$$I(x) = variance\left(\sum_i w_r^i \pi^i(x, a)\right)$$

The most informative state will be the one where the previous criteria is smaller, meaning that the policy distribution is flat. The action is selected randomly. We note that this exploration strategy just takes into account the uncertainty on the reward. Creating an exploration strategy based on the uncertainty of the sign-meaning estimation does not provide a gain due to the probabilistic model we used. Other sampling criteria that we are going to test is a counter based random sampling of states and actions, and random states with actions sampled with the usual $\epsilon - greedy$ strategy.

III. RESULTS

In this Section we present the results from our algorithm in a set of simulated environments.

A. Navigation Task

We consider a simulated environment with 5 different actions, the number of states in the discretization grid varies in each problem. All results report averages of 20 executions of the algorithm with different parameters. The true reward function to be found by the learner is randomly generated at each experiment, the same occurs for the meaning and feedback models. The reward in this abstract problem can be seen as corresponding to a navigation task and so the reward is the goal location.

Figure 2 compares two situations, the first where the learner estimates the feedback model of the teacher and the second where it does not estimate the feedback model. But, in both cases, considering that *all feedback signs are known*. We can see that learning is faster and with a better quality if a model is estimated and so it shows the importance of our approach. Some protocols are equivalent in terms of speed of learning even without any particular assumption about it. But consider, for instance, what the teacher means when it does not say anything. In some protocols that is equivalent to say it is correct and in some others it means that the action is wrong. Only after knowing this relation can the learner make use of that data to improve its estimation of the task representation.

From Figure 3 we can see that the task can be learned even under a noisy feedback signal, and that we can learn simultaneously the model of the feedback behavior. Around 10% of the feedback signals were noisy. The same figure also compares the different sampling methods. We can see that the active exploration is able to learn faster, with less variance and with a better asymptotic convergence. This situation happens even if the active criteria was developed without taking the noise in the feedback into account.

We now present our full system where there are *some feedback signs with unknown meanings*. The system needs to learn the task, the feedback model and the map of new signs to their meanings. We consider 7 feedback signs whose

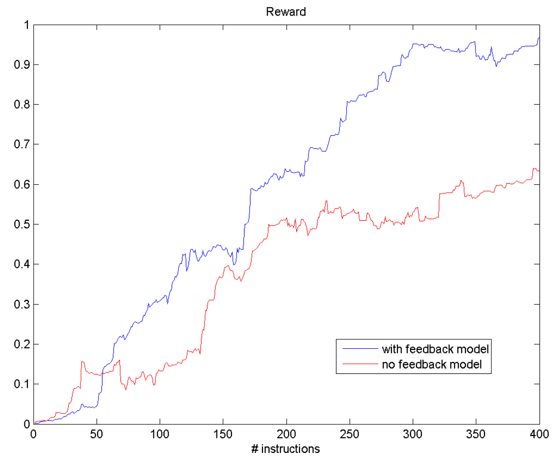

Fig. 2. Comparison of learning of the task model with or without learning the feedback model. Number of states is 400. The figure shows the likelihood of the best estimate of the reward function.

meanings are known, i.e. the five actions plus O and W, and 7 new signs that can map to any of those meanings. Figure 4 shows the results using 100 particles for the estimation of the sign-meaning relations. The first conclusion is that the system can learn all the important variables and, again, the active exploration method learns faster and with less variance than the other methods. Not all the signs meanings are successfully estimated and this situation is caused by a very asymmetrical sampling of the feedback signs. We can observe this in Figure 5. For instance, a teacher that always gives the correct action will never use the signs for correct and wrong and so their meanings will not be learned.

B. *Collecting Objects*

We now consider an environment where the learner can navigate and where there is a probability of finding three different objects. The learner has to learn which objects it should collect, or not, and for each of the object classes learn where they must be delivered. The number of actions is now 7, the 5 navigation ones plus collect and release. The number of feedback signs is now 10, again we assume that we have an initial known set of signs and the teacher will provide 10 new synonyms.

Figures 6 and 7 give the results for a problem with three objects and 64 possible locations. In each execution of the problem the system randomly selects the objects that should be collected and their delivery locations. Results are qualitatively equal to the previous problem.

IV. CONCLUSIONS

Computational approaches in learning by demonstration have evolved a lot in recent years. These methods can now be applied in realistic human-robot interaction settings to effectively provide an intuitive way for untrained teachers to program robots. Under this setting most algorithms have to be adapted to the noise and ambiguity usually present in human dialog. In this work we showed how a learning

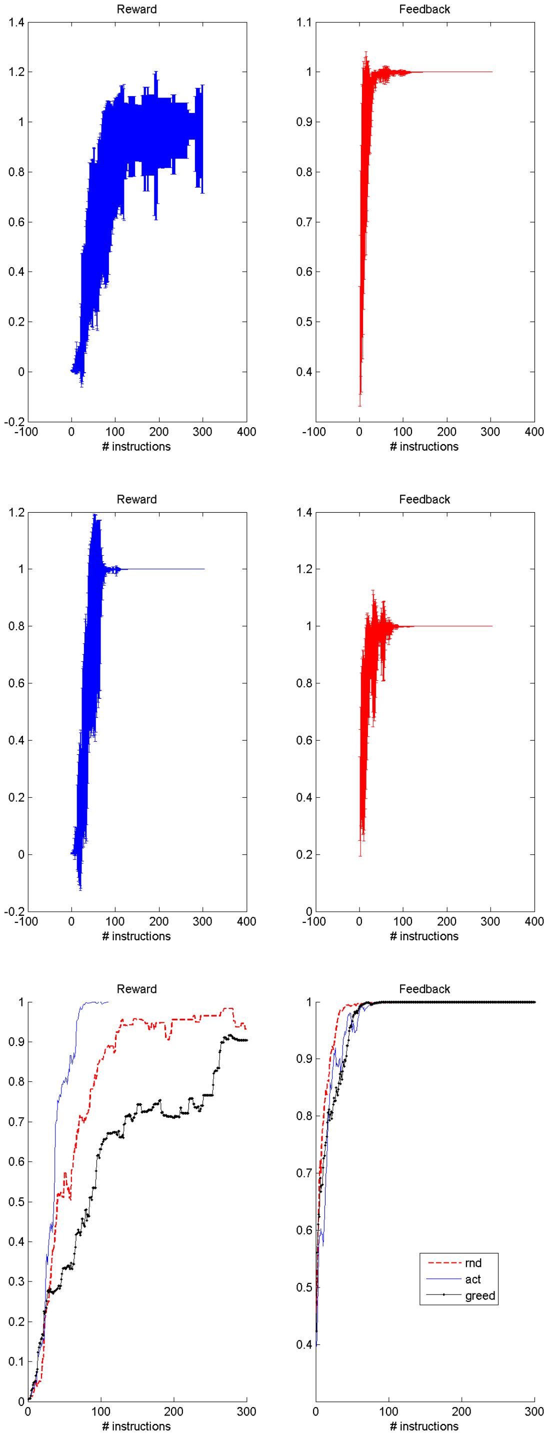

Fig. 3. Simultaneous acquisition of the task and the feedback models with three different exploration methods for a problem with 225 states. The figures show the likelihood of the best model for the reward and the feedback The top figure shows the results for random exploration and the middle one for active exploration. The bottom one compares both, and also one using ϵ-greedy exploration. Results are for 10 runs, the mean are variance bars are shown. The active exploration method learns faster with smaller variance and bias.

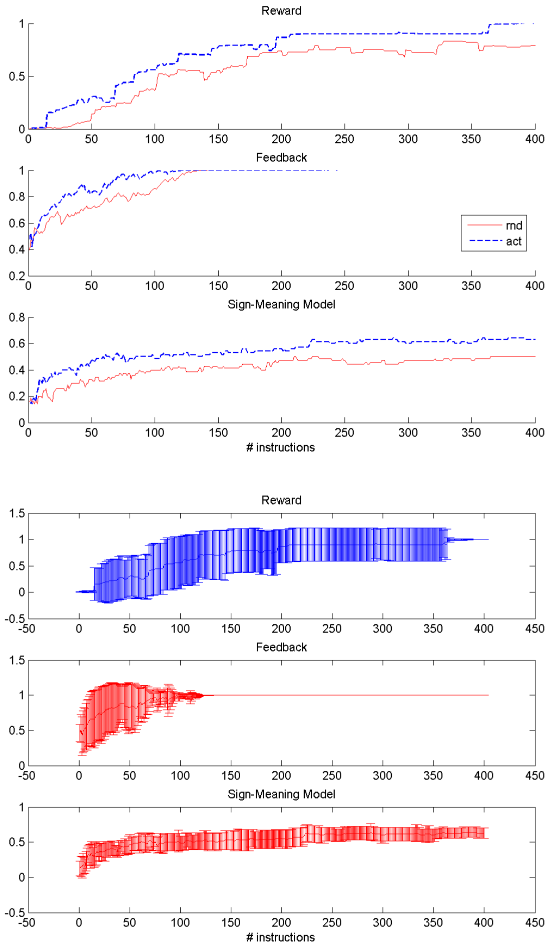

Fig. 4. Learning with 400 states (top) comparison between sampling methods, (bottom) mean and variance for the active method.

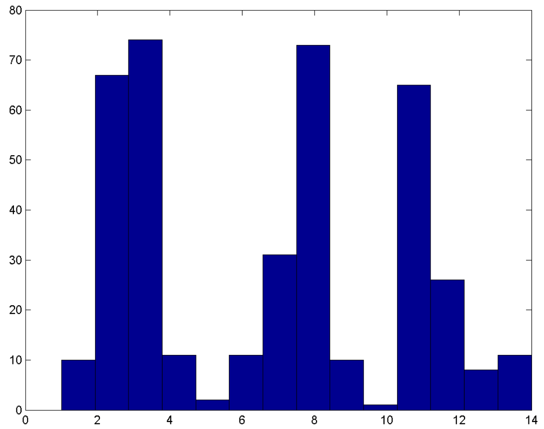

Fig. 5. Histogram of observed feedback signs. We can see that some signs are very rarely used thus making it impossible to estimate their meanings.

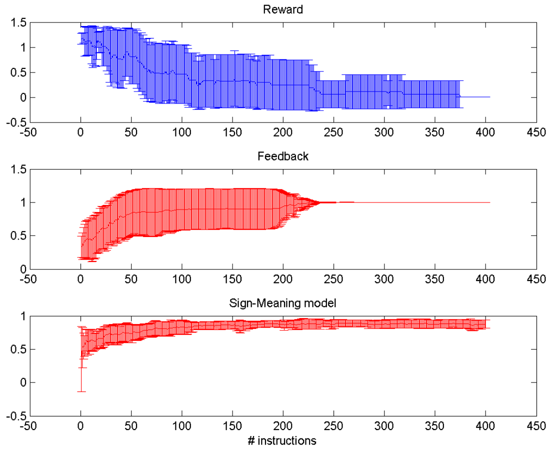

Fig. 6. Mean and variance for the active learning method in the “Object Collecting” Task. The system is able to learn the task, the feedback system and new feedback signs. Top - policy loss; Middle - likelihood of correct feedback model; Bottom - number of correctly assigned signs.

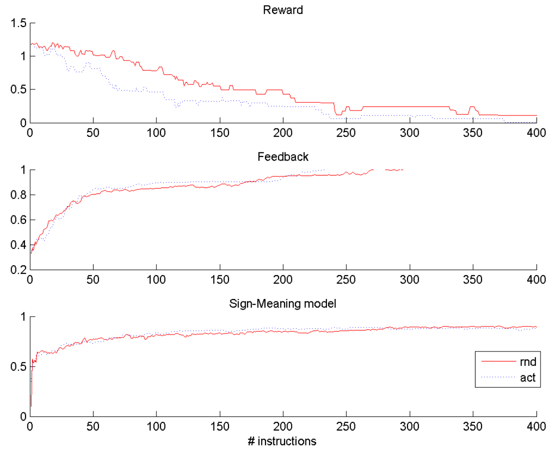

Fig. 7. Comparison between active and randomly sampling in the “Object Collecting” Task. The system is able to learn the task, the feedback system and new feedback signs. Top - policy loss; Middle - likelihood of correct feedback model; Bottom - number of correctly assigned signs.

system can learn a task description when the feedback it gets from the teacher does not follow a *rigid protocol* and is *very noisy* (10% error in correctly recognizing the feedback signs). Having very restricted protocols makes interaction more tiring and does not take advantage of the extra signals that teachers provide to robots. We showed that a learner can estimate simultaneously the feedback protocol and the task representation in a reasonable amount of time and computational complexity.

We took a further challenge and only assumed partial knowledge of the feedback signs. By bootstrapping the systems with some known sign-meaning correspondences, the system could successfully estimate the correspondences of new feedback signs. To further improve the efficiency of the system we presented an active learning approach where the learner asked the teacher for specific information in states where it is more uncertain. This results in faster learning than with a simple random strategy with a smaller variance and bias.

We tested our system in different problems and the qualitative results are consistent among different domains and the system is able to learn all the entities. The results degraded when the noise level increases. The improvement we get from active sampling is dependent on the quality of the posterior estimation and thus it is important to have features that can correctly describe the possible tasks. In terms of the number of signs, in our system we had to assume that some correspondences are known to improve the convergence.

This research can also be compared to language learning research, where we learn synonyms for words labeling actions from a teacher already proficient with the language. The information used to learn these synonyms is not traditionally used in language learning research. It is often the case that a label is used to describe an object or property of an object (and a difficulty often encountered is that it is unknown what property the label refers to), see for example [26]. In our research a label is associated with a desired action that is in fact never seen (the person that knows the language does not perform demonstration of the actions but instead gives the label of the action). The meaning of the label is instead found by building a model of what task the teacher is trying to teach the learner, where a good model of the feedback helps learning a good task model and vice versa (under a set of observations some feedback model task model pairs typically becomes much more probable).

The possible feedback models used spontaneously by people can be more complex than the simple meaning correspondences we assumed [7], [18]. We plan to study how to integrate such complex models in our approach. Also, we want to see if similar approaches can be used when the learner tries to directly estimate the policy. In a real situation the three elements do not need to be learned simultaneously. A robot that interacts repeatedly with the same teacher will be able to reuse part of its knowledge about the used signals and protocols when acquiring new tasks.

REFERENCES

[1] J. Weng, J. McClelland, A. Pentland, O. Sporns, I. Stockman, M. Sur, and E. Thelen, "Autonomous mental development by robots and animals," *Science*, vol. 291, pp. 599 – 600, 2001.

[2] B. Argall, S. Chernova, and M. Veloso, "A survey of robot learning from demonstration," *Robotics and Autonomous Systems*, vol. 57, no. 5, pp. 469–483, 2009.

[3] M. Lopes, F. Melo, L. Montesano, and J. Santos-Victor, "Abstraction levels for robotic imitation: Overview and computational approaches," in *From Motor to Interaction Learning in Robots*, ser. Studies in Computational Intelligence, O. Sigaud and J. Peters, Eds. Springer, 2010, vol. 264, pp. 313–355.

[4] M. Nicolescu and M. Mataric, "Natural methods for robot task learning: Instructive demonstrations, generalization and practice," in *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*. ACM, 2003, pp. 241–248.

[5] S. Calinon, F. Guenter, and A. Billard, "On learning, representing and generalizing a task in a humanoid robot," *IEEE Transactions on Systems, Man and Cybernetics, Part B. Special issue on robot learning by observation, demonstration and imitation*, vol. 37, no. 2, pp. 286–298, 2007.

[6] M. Lopes, F. S. Melo, and L. Montesano, "Affordance-based imitation learning in robots," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'07)*, USA, Nov 2007, pp. 1015–1021.

[7] A. L. Thomaz and C. Breazeal, "Teachable robots: Understanding human teaching behavior to build more effective robot learners," *Artificial Intelligence Journal*, vol. 172, pp. 716–737, 2008.

[8] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Proceedings of the 21st International Conference on Machine Learning (ICML'04)*, 2004, pp. 1–8.

[9] D. Grollman and O. Jenkins, "Dogged learning for robots," in *Robotics and Automation, 2007 IEEE International Conference on*. IEEE, 2007, pp. 2483–2488.

[10] S. Chernova and M. Veloso, "Interactive policy learning through confidence-based autonomy," *J. Artificial Intelligence Research*, vol. 34, pp. 1–25, 2009.

[11] C. Breazeal, A. Brooks, J. Gray, G. Hoffman, J. Lieberman, H. Lee, A. L. Thomaz, and D. Mulanda., "Tutelage and collaboration for humanoid robots," *International Journal of Humanoid Robotics*, vol. 1, no. 2, 2004.

[12] M. Lopes, F. S. Melo, and L. Montesano, "Active learning for reward estimation in inverse reinforcement learning," in *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II*, ser. ECML PKDD '09, 2009, pp. 31–46.

[13] M. Mason and M. Lopes, "Robot self-initiative and personalization by learning through repeated interactions," in *6th ACM/IEEE International Conference on Human-Robot Interaction (HRI'11)*, 2011.

[14] K. Judah, S. Roy, A. Fern, and T. Dietterich, "Reinforcement learning via practice and critique advice," in *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10)*, 2010.

[15] W. Knox and P. Stone, "Interactively shaping agents via human reinforcement: The tamer framework," in *Proceedings of the fifth international conference on Knowledge capture*. ACM, 2009, pp. 9–16.

[16] W. B. Knox and P. Stone, "Combining manual feedback with subsequent mdp reward signals for reinforcement learning," in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'10)*, 2010, pp. 5–12.

[17] M. Cakmak, C. Chao, and A. Thomaz, "Designing interactions for robot active learners," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 2, pp. 108–118, 2010.

[18] M. Cakmak and A. Thomaz, "Optimality of human teachers for robot learners," in *Proceedings of the International Conference on Development and Learning (ICDL)*, 2010.

[19] Y. Mohammad and T. Nishida, "Learning interaction protocols using Augmented Bayesian Networks applied to guided navigation," in *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*. IEEE, 2010, pp. 4119–4126.

[20] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 1998.

[21] A. Y. Ng and S. J. Russell, "Algorithms for inverse reinforcement learning," in *Proc. 17th Int. Conf. Machine Learning*, USA, 2000.

[22] D. Ramachandran and E. Amir, "Bayesian inverse reinforcement learning," in *20th Int. Joint Conf. Artificial Intelligence*, India, 2007.

[23] D. Fox, S. Thrun, W. Burgard, and F. Dellaert, "Particle filters for mobile robot localization," in *Sequential Monte Carlo Methods in Practice*, A. Doucet, N. de Freitas, and N. Gordon, Eds. Springer-Verlag, 2001.

[24] P.-Y. Oudeyer, F. Kaplan, and V. Hafner, "Intrinsic motivation systems for autonomous mental development," *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 2, pp. 265–286, 2007.

[25] R. Cohn, M. Maxim, E. Durfee, and S. Singh, "Selecting Operator Queries using Expected Myopic Gain," in *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. IEEE, 2010, pp. 40–47.

[26] L. Steels, "Experiments on the emergence of human communication," *Trends in Cognitive Sciences*, vol. 10, no. 8, pp. 347–349, 2006.