BIO Web of Conferences **1**, 00004 (2011)
DOI: 10.1051/bioconf/20110100004

Continuing Robot Skill Learning after Demonstration with Human Feedback

Brenna D. Argall[*][†]

(*) Northwestern University, Evanston, IL, USA
(†) Rehabilitation Institute of Chicago, Chicago, IL, USA
E-mail: brenna.argall@northwestern.edu

Abstract

Though demonstration-based approaches have been successfully applied to learning a variety of robot behaviors, there do exist some limitations. The ability to continue learning after demonstration, based on execution experience with the learned policy, therefore has proven to be an asset to many demonstration-based learning systems. This paper discusses important considerations for interfaces that provide feedback to adapt and improve demonstrated behaviors. Feedback interfaces developed for two robots with very different motion capabilities - a wheeled mobile robot and high degree-of-freedom humanoid - are highlighted.

1. Introduction

Demonstration-based approaches to learning robot control have been applied with success to a variety of behaviors [6, 3, 10, 12], at control levels ranging from low to high. Learning typically consists of behavior demonstration, followed by policy derivation from the recorded dataset via machine learning techniques. Many candidate representations for the learned behavior policy exist; one common representation is a mapping from observations of world state to robot actions. Policy development via demonstration has many advantages, including the generation of focused datasets and intuitive use by human developers. There do however exist some limitations; for example, there might be correspondence issues between the physical embodiments of the teacher and robot, or the demonstrations themselves might be of a poor quality due to suboptimality in the control interface or teachers abilities. For these and other systems, the ability to continue learning after demonstration [1, 7, 8, 9], based on execution experience with the policy originally learned from demonstration, can be a desirable feature.

Learning from experience might take many forms, but key to all is an evaluation of a policys performance; that is, to receive some sort of feedback signal. This paper will identify some key characteristics of interfaces for providing feedback (Sec. 2), highlighting in particular feedback that takes the form of policy corrections provided by a human teacher. Two corrective feedback interfaces, developed for two very different robot platforms, furthermore will be highlighted (Sec. 3).

2 Feedback Type and Interface

There are a multitude of potential formulations for feedback type; for example, the learner might receive a single reward upon reaching a failure state, or the value of a gradient along which a more desirable action may be found. The *amount of information* encoded within the feedback is one important descriptor for a given feedback type. Feedback can encode some, none or all of the translation from policy evaluation to policy update. Other important descriptors include the *continuity* of the feedback - whether its value is discrete/binary and derives from a finite set, or continuous and thus derives from an infinite set - and the *frequency* at which it is provided. Frequency is governed by whether feedback is provided for entire executions, subsets of executions or individual decision points, and the corresponding time duration of each. All of these factors influence both the utility of the feedback to the learner, and its cost to the provider.

The translation from policy execution to performance feedback potentially consists of multiple phases that engage the feedback interface. For example, the execution may need to be presented in a meaningful format for the feedback provider, or the feedback might need to be translated into meaningful information for the learner. One key consideration in the design of a feedback interface therefore is how the execution is

Article available at http://www.bio-conferences.org or http://dx.doi.org/10.1051/bioconf/20110100004

evaluated, which depends both on the source of the evaluation (e.g. automated computation or task expert) and the information required by that source in order to provide an evaluation. Another consideration is how to *associate* the evaluation with the underlying execution. Some feedback forms are loosely tied to the execution data, while others are closely tied, for example an action correction that must strictly associate with the execution point that produced the action. Close feedback-data associations are necessarily influenced by the sampling rate of the policy.

3 Learning from Corrective Feedback following Demonstration

Here the interfaces for providing demonstrations and corrective feedback for two very different robot platforms - a wheeled mobile robot and high degree-of-freedom (DoF) humanoid - are overviewed. Behavior corrections have the advantage of providing an explicit indication of which alternate action should have been taken (or state entered), but at the expense of requiring detailed information from the teacher. By contrast, for example, state rewards are typically directly observble from the environment at no cost to a teacher, but require exploration on the part of the robot learner in order to determine which alternate action should have been executed (or state entered). If a correction interface is designed so that the burden placed on the teacher is minimal, then we have the dual wins of providing highly informative feedback at a low cost to the teacher.

Having the means to provide feedback handles only part of the policy adaptation question, however, and further considerations include (i) how the policy updates in response to feedback, and (ii) how the behavior of the policy is modified. In order to update a policy with feedback, the approach taken in all empirical implementations with the following interfaces is to first generate new behavior examples, and then rederive the policy from the updated set.

3.1 A Wheeled Mobile Robot [1]

The Segway RMP (Fig. 1) is a wheeled differential drive robot, whose dynamics operate as an inverted pendulum [11]. Since the balancing mechanism is proprietary information of Segway LLC, a highly accurate model for the motion of the robot is unavailable to us; thus motivating policy development via demonstration. Furthermore, the dimensionality of the action space is sufficiently low (rotational and translation speeds) to make teleoperation via joysticking fairly straightforward.

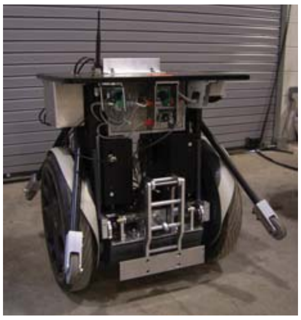

Figure 1: The Segway RMP, a dynamically balancing differential drive mobile robot.

Key challenges to providing corrective feedback within low-level motion control domains are that (i) corrections are continuous-valued and (ii) the policy is sampled at a rapid rate. To address these challenges, advice-operators were introduced as a corrective feedback form suitable for providing continuous-valued corrections, and Focused Feedback for Mobile Robot Policies (F3MRP) as a framework suitable for providing feedback on policies sampled at a high frequency[2]. Concretely defined, an advice-operator is a mathematical computation performed on an observation input or action output. Operators are applied over a learner execution segment, indicated through the F3MRP interface via its visual representation of the ground path taken by the robot. Pairing a modified observation (or action) with the executed action (or observation) represents a corrected mapping. Teacher selection of a single advice-operator and execution segment thus translates into multiple continuous-valued corrections, and therefore is suitable for modifying low-level motion control policies sampled at high frequency.

Advice-operators and the F3MRP interface have been used in multiple algorithms that differ according to how feedback modifies the policy behavior. For all implementations, new behavior examples for the policy update are synthesized by modifying, via advice-operators, data recorded during policy execution by the learner. Initial work used corrections to refine policies learned from demonstration, with empirical validation on the Segway RMP performing a spatial positioning task [2]. Feedback was found to improve policy behavior, and be more effective than providing further teacher demonstrations. Corrective feedback also was used to scaffold simpler policies learned from demonstration into a policy able to execute a more complex task, within a simulated racetrack driving domain [5].

[1] Work with the SegwayRMP was done at the Robotics Institute at Carnegie Mellon University, under the direction of Prof. Manuela Veloso and Dr. Brett Browning.

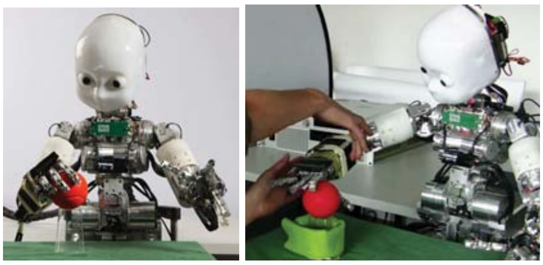

Figure 2: The iCub, a high degree of freedom humanoid (left), whose feedback interface consists of tactile corrections (right).

In this case feedback was furthermore found to enable novel, undemonstrated, policy behavior.

3.2 A High-DoF Humanoid [2]

The iCub (Fig. 2) is a 53-DoF humanoid designed to approximate the size of a human child [14]. In contrast to the Segway, the robot is (for our purposes) non-mobile, has a high number of free degrees of freedom to control during teleoperation (15-DoF per arm), and its movements are governed by an accurate kinematic model. While our teleoperation solution[3] does allow for effective demonstration, it also is strongly subject to sensor drift and requires the human demonstrator to compensate for some correspondence issues, thus motivating the utility of corrections following demonstration.

Here corrective feedback takes a different form, to meet the needs of a different robot platform and domain. In particular, online feedback is provided through a tactile interface located on the body of the robot. Touch is used to indicate incremental position adjustments in Cartesian space during robot arm motion trajectories, as well as to demonstrate adaptation in response to changes in contact signature during object interaction. We posit that tactile feedback naturally extends the idea of teaching robots as humans teach other humans. Moreover, tactile detection can be crucial for safe robot operation around humans, and this detected contact may be exploited to further knowledge transfer from human to robot.

With the iCub interface, policy updates again involve the generation of new behavior examples and rederivation of the policy. Here however new examples are actively generated by immediately adopting the indicated position adjustments during the learner behavior execution, or passively generated by exploiting compliance in the robot joints. Tactile feedback is used to assist in both policy refinement and the reuse of a demonstrated policy when developing a different policy; effectively using the demonstrated policy as prior knowledge for a new behavior. Empirical validation has included grasp positioning on the iCub humanoid [4], as well as grasp adaptation in response to changes in fingertip contact [13].

4 Conclusions

Augmenting demonstration-based policy learning with human feedback has been shown to be an effective approach to policy improvement and adaptation. Feedback that takes the form of behavior corrections can be highly informative, but also potentially costly for the teacher to provide. The potential cost is particularly high in motion control domains with continuous-valued actions and rapidly-sampled policies, and we have overviewed two interfaces that provide solutions to these challenges. Differences between the interfaces are partly due to variations in the motion capabilities of the robot platforms to which they are applied; for instance, the F3MRP approach of plotting the motion path would be less appropriate for high-dimensional manipulator motion paths, while an interface that requires touching the robot would not be practical for a mobile platform. While the primary focus here has been on the challenge of providing highly informative feedback at a low cost to the teacher, an interesting direction for future work would be to explore the option of a generic feedback interface that is hardware-independent, and thus reusable between multiple robot platforms.

Acknowledgements

Thanks to Prof. Aude Billard and Dr. Eric Sauser for their collaboration on the feedback interface for the iCub. Funding for the iCub research was received from the European Community's Seventh Framework Programme FP7/2007-2013 - Challenge 2 - Cognitive Systems, Interaction, Robotics - under grant agreement n^o [231500]-[ROBOSKIN].

Thanks to Prof. Manuela Veloso and Dr. Brett Browning for their collaboration on the feedback interface for the Segway. Funding for the Segway research was partly sponsored by the Boeing Corporation under Grant No. CMU-BA-GTA-1, BBNT Solutions under subcontract No. 950008572, via prime Air Force contract No. SA-8650-06-C-7606, the United States Department of the Interior under Grant No. NBCH-1040007 and the Qatar Foundation for Education, Science and Community Development. The views and conclusions contained in this document are solely those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

[2] Work with the iCub was done in the Learning Algorithms and Systems Laboratory (LASA), at the École Polytechnique Fédérale de Lausanne (EPFL), under the direction of Prof. Aude Billard and in collaboration with Dr. Eric Sauser.

[3] Please see [4] for full details.

References

[1] P. Abbeel and A. Y. Ng. Exploration and apprenticeship learning in reinforcement learning. In *Proceedings of ICML*, 2005.

[2] B. Argall. Mobile robot motion control from demonstration and corrective feedback. In J. Peters and O. Sigaud, editors, *From Motor to Interaction Learning in Robots*. Springer, New York, NY, 2009.

[3] B. Argall, S. Chernova, B. Browning, and M. Veloso. A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, 57(5), 2009.

[4] B. Argall, E. Sauser, and A. Billard. Tactile guidance for policy adaptation. *Foundations and Trends in Robotics*, 1(2), 2010.

[5] B. D. Argall, B. Browning, and M. Veloso. Teacher feedback to scaffold and refine demonstrated motion primitives on a mobile robot. *Robotics and Autonomous Systems*, 59:243–255, 2011.

[6] C. G. Atkeson and S. Schaal. Robot learning from demonstration. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML '97)*, 1997.

[7] D. C. Bentivegna. *Learning from Observation Using Primitives*. PhD thesis, College of Computing, Georgia Institute of Technology, Atlanta, GA, July 2004.

[8] S. Chernova and M. Veloso. Learning equivalent action choices from demonstration. In *Proceedings of IROS*, 2008.

[9] J. Kober and J. Peters. Learning motor primitives for robotics. In *Proceedings of ICRA '09*, 2009.

[10] M. J. Matarić. Sensory-motor primitives as a basis for learning by imitation: Linking perception to action and biology to robotics. chapter 15.

[11] H. G. Nguyen, J. Morrell, K. Mullens, A. Burmeister, S. Miles, K. Thomas, and D. W. Gage. Segway robotic mobility platform. In *SPIE Mobile Robots XVII*, 2004.

[12] N. Ratliff, D. Bradley, J. A. Bagnell, and J. Chestnutt. Boosting structured prediction for imitation learning. *Proceedings of Advances in Neural Information Processing Systems (NIPS '07)*, 2007.

[13] E. L. Sauser, B. D. Argall, G. Metta, and A. G. Billard. Iterative learning of grasp adaptation through human corrections. *Robotics and Autonomous Systems*, In press, 2011.

[14] N. Tsagarakis, G. Metta, G. Sandini, D. Vernon, R. Beira, F. Becchi, L. Righetti, J. Santos-Victor, A. Ijspeert, M. Carrozza, and D. Caldwell. iCub: The design and realization of an open humanoid platform for cognitive and neuroscience research. *Advanced Robotics*, 21, 2007.