Demonstration, Tactile Correction and Multiple Training Data Sources for Robot Motion Control

Brenna D. Argall **Eric L. Sauser** **Aude G. Billard** [*]

Abstract

This work considers our approach to robot motion control learning from the standpoint of multiple data sources. Our paradigm derives data from human teachers providing *task demonstrations* and *tactile corrections* for policy *refinement* and *reuse*. We contribute a novel formalization for this data, and identify future directions for the algorithm to reason explicitly about differences in data source.

1 Introduction

The problem of learning for robot control naturally lends itself to learning from multiple sources, as frequently numerous sources for training data are available. For example, a robot might observe the world through multiple sensing modalities, or have multiple teachers demonstrating the target behavior under development. We present in this paper multiple novel training data sources that arise from our approach that learns motion control from a human teacher, who provides both *demonstrations* and *tactile corrections* for the purposes of policy *refinement* and *reuse*. Data thus derives from a variety of teachers and teaching techniques, as well as multiple target behaviors under development.

Our work addresses the challenge of robot motion control, which is fundamental to many robotics applications. Development initially was motivated by the goal of reducing the requirements placed on a policy developer, by increasing policy robustness through refinement and transferring existing knowledge through reuse. One consequence of our policy development techniques is the presence of several novel data sources. We are interested in addressing the different sources explicitly within our algorithm, and this paper provides a first step in that direction.

To begin, we present our algorithm and its initial empirical validation on a high degree of freedom humanoid. We then frame our work to date within the context of multiple sources, formalizing what it means to be a distinct data source under our paradigm. Many of the design decisions related to explicitly addressing the multiple data sources of our characterization are highlighted, and potential formulations are hypothesized.

2 Learning from Demonstration and Tactile Corrections

Within *Learning from Demonstration (LfD)*, teacher executions of a desired behavior are recorded and a policy, or mapping from world state to robot action, is derived from the resultant dataset. LfD has seen success on a variety of robotics applications, and has many attractive characteristics for both teacher and learner [3]. Even so, policy development typically still is non-trivial, and to have a robot learn from its execution performance, or *experience*, can be a valuable policy improvement tool [1]. Our work employs *tactile* corrections to modify a policy learned through demonstration, for the purposes of both policy refinement and reuse.

[*] B. D. Argall, E. L. Sauser and A. G. Billard are with the Learning Algorithms and Systems Laboratory (LASA), École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland. [brennadee.argall,eric.sauser,aude.billard]@epfl.ch

2.1 The Tactile Policy Correction Algorithm

Under our *Tactile Policy Correction (TPC)* algorithm [2], a policy is initially derived from task demonstration by a teacher. We formally define the world to consist of actions $A \in \mathbb{R}^{\ell}$ and observations $Z \in \mathbb{R}^{(m+n)}$ of world state. An observation $\mathbf{z} \in Z$ consists of two components, $\mathbf{z} = (\mathbf{z}_{\varphi}, \mathbf{z}_{\neg\varphi})$, where $\mathbf{z}_{\varphi} \in \mathbb{R}^{m}$ describes the robot pose, and $\mathbf{z}_{\neg\varphi} \in \mathbb{R}^{n}$ describes any other observables that are of interest to the policy. We define a *demonstration* to consist of a sequence of T observations $\{\mathbf{z}^j\}_{j=1}^{T}$, recorded during teacher execution of the task. The collected set $D = \{\mathbf{z}^j\}_{j=1}^{N}$ of demonstrations, totaling N recorded observations, is then provided to the robot learner. From this set a policy $\pi : Z \rightarrow A$ is derived. We decompose policy execution into two steps: pose prediction and action selection. Pose prediction is accomplished via regression techniques, based on state observations. Action selection is accomplished via a robot-specific controller.

Following demonstration and policy derivation, the robot executes with its policy and receives tactile corrections from the human teacher. Tactile corrections are used within two capacities, either to refine the existing policy or to build a new policy bootstrapped on the demonstrated policy. The robot learner translates tactile feedback into an incremental shift $\boldsymbol{\delta}^t \in \mathbb{R}^m$ in the robot pose. The form taken by the feedback is platform-specific, depending both on the tactile sensors employed to detect human touch and how the sensor feedback is processed. The robot controller is then passed the new *adjusted* pose, for which the incremental shift $\boldsymbol{\delta}^t$ is added to the current robot pose $\boldsymbol{\varphi}^t \in \mathbb{R}^m$. The timestep concludes with the recording of observation $\mathbf{z}^t = \left(\mathbf{z}^t_{\varphi}, \mathbf{z}^t_{\neg\varphi}\right)$, that encodes within $\mathbf{z}^t_{\varphi}$ the current pose $\boldsymbol{\varphi}^t$ corrected by tactile feedback, into the set of corrected execution points D_c.

Upon completion of the entire execution, policy π is rederived from demonstration set D *and* the set of corrected executions D_c; the corrected execution thus is treated as a new demonstration for the policy. Policy derivation gives greater importance to the corrected data, through a weight w^j

$$w^j \;=\; \begin{cases} \left(1 - \frac{N}{N+N_c}\right)\left(1 - \bar{w}(\mathbf{z}^j_{\neg\varphi})\right) & \mathbf{z}^j \in D \\ \left(1 - \frac{N_c}{N+N_c}\right)\bar{w}(\mathbf{z}^j_{\neg\varphi}) & \mathbf{z}^j \in D_c \end{cases}$$

for point $\mathbf{z}^j \in D \cup D_c$, where N is the number of datapoints in D, and N_c the number in D_c, and $\bar{w}(\mathbf{z}^j_{\neg\varphi})$ is a global weight function dependent on execution time τ, which we encode within $\mathbf{z}^j_{\neg\varphi}$.

The TPC algorithm is agnostic to the regression techniques used for pose prediction and action selection. In our empircial validation, pose prediction is accomplished through the regression technique of *GMM-GMR*, and action selection through an inverse kinematic controller.

2.2 Empirical Validation

We have implemented [2] the TCP algorithm on a small 57-DoF humanoid, the iCub robot. In the validation task, the robot learns to position the end-effector of its 7-DoF arm for the grasping of different objects at various locations. Task demonstration was performed via teleoperation of the robot arm by a human teacher, during which the robot recorded observations from its own sensors; observations $\mathbf{z}^t \in \mathbb{R}^8$ consisted of the robot pose $\boldsymbol{\varphi}^t$ (i.e. $\mathbf{z}_{\varphi}$) and time τ^t (i.e. $\mathbf{z}_{\neg\varphi}$). Tactile corrections were provided through a touch pad interface that allowed for translational and rotational adjustments at the wrist and hand of our 7-DoF manipulator (Fig. 1A). Corrections were employed to refine the demonstration policy, and to build other policies able to accomplish alternate tasks; namely, to arrive at a different location or with a different end-effector orientation.

Policy refinement resulted in improved grasp quality, measured as the number of fingers in contact with the object, as well as an increase in the number of successful grasps able to lift the object from the table. Policy reuse enabled the development of policies able to execute *undemonstrated* position-object combinations (Fig. 1B,C).

3 Multiple Data Sources within TCP

We now consider our TCP algorithm from the viewpoint of multiple data sources, and contribute a novel formulation by which to characterize source. We further identify the design decisions currently under consideration for an algorithm that explicitly addresses data source under our formulation.

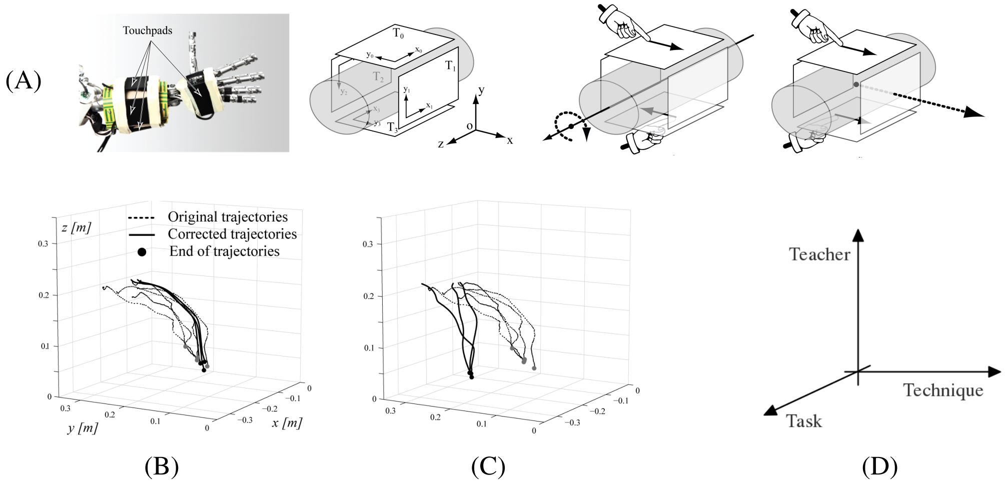

Figure 1: *Top:* (A) Schematic of the touch pads controlling the robot wrist and hand. Fingers sliding on opposite pads can result in both rotational or translational arm motions. *Bottom:* Grasp trajectories in Cartesian space, with the corrected trajectories showing policy (B) refinement and (C) reuse. (D) The axes of our data source characterization.

3.1 Formalism

We formalize the definition of a data source under our paradigm according to three axes (Fig. 1D). The first axis is the *teacher*, where multiple teachers might be involved in policy development. The second axis is the control *technique*, where different mechanisms might control the platform from which data is being recorded. The third axis is *task*, where data intended to accomplish different tasks are assumed to derive from different sources. Given n demonstration *Teachers*, m control *Techniques* and ℓ *Tasks*, the total possible number of sources that might provide data is $n \times m \times \ell$.

The empirical implementation just presented utilized two different human teachers (*Teacher A, B*), one for demonstration and one for correction. Similarly two control techniques were employed (*Technique A, B*): teleoperation of the robot for the purpose of demonstration, and kinesthetic repositioning through a tactile interface for the purpose of correction. Policies for three tasks were developed (*Task A, B, C*): to grasp a ball at the demonstrated location (*refinement*) and at a new location (*reuse*), and to grasp a bottle (*reuse*).

Under our proposed data source formalism, the work of the previous section therefore had the potential to produce data from 12 sources ($2 \times 2 \times 3$).[1] The current TCP algorithm however does not distinguish between *Teachers*, *Techniques* or *Tasks*, [2] and rather distinguishes between only old versus new data, with new data being weighted more heavily during policy rederivation.

3.2 Open Questions

Data derived from each source potentially differs in quality, as well as task appropriateness. For example, one teacher might be more adept at demonstrating the task and thus provide higher quality demonstration data, or one task might be more similar to the task under development and thus its correction data more appropriate for reuse. Our goal is to develop an algorithm able to determine these differences and fuse the multiple data sources automatically. There are many open design decisions for the development of such an algorithm, a handful of which we identify here.

One design decision is the choice of data source representation. We have delineated three axes along which our data can differ in source, but whether or not to include all of these axes within a given problem formulation is open for debate. Instead of all three, one could choose to split along only two

[1] Only 4 of the 12 sources produced data however (*demonstration*: TeacherA-TechniqueA-TaskA, *refinement*: TeacherB-TechniqueB-TaskA, *reuse*: TeacherB-TechniqueB-TaskB and TeacherB-TechniqueB-TaskC).

[2] Any separation along these lines was coincidental or external to the algorithm: the old-new data split happened to also separate the two *Teachers* and *Techniques*, and different *Tasks* were learned separately.

axes; for example, to consider each unique *Teacher-Technique*[3] combination to be a distinct source, regardless of task. One could even choose to split along a single axis only; for example along the *Technique* axis, and thus to have models that are agnostic to differences in teacher or task. Note that to split along no axes would be to treat all data as if from a single source.

To what extent the algorithm reasons about the different sources, along with the amount of information that it is provided to do so, is a second decision. For example, the learner might be provided with a model of the relationships between data sources, and labels indicating from which source each datapoint derives. Perhaps the sources are partly labeled (e.g. by teacher only) or entirely unlabeled, and models are provided that enable the learner to infer data source from sensor readings.

Given a set of identified sources, a third decision is the manner of integration between the multiple data sources. For example, one approach could bootstrap a policy for a new task by incorporating data from prior *Task* sources based on nearness within the observation space (as in [4]). Another approach could attach some sort of reliability measure to different sources, and then use this measure during integration (similar to [6]). A reliability measure would require a mechanism by which the performance of each source is evaluated, for example accumulated state reward or credit from a human teacher. Depending on the data source representation, a measure that discriminates between sources could indicate data quality (e.g. between different *Teachers*) or appropriateness (e.g. transfer to another *Task*), or both, and thus be used for the purposes of policy refinement or reuse. Another option could leverage the multiple data sources against each other, for example to extract the important characteristics of a particular *Task* by comparing data produced from different *Teacher-Technique* combination sources (similar to [5]).

A fourth design decision is the level - individual datapoints, entire sources - at which integration occurs. For example, consider that the data sources are to be combined into a single prediction. At one extreme, a model is developed for each source and the predictions from all models are combined into a single prediction. At the other extreme, data from all sources are combined into a single model, which then makes the single prediction.

3.3 Conclusions

We have considered our approach for policy development, via demonstration and tactile corrections, from the standpoint of multiple data sources, and contributed a novel formulation by which to characterize data source within this approach. Some of our criteria are particular to LfD (e.g. multiple teachers), while other criteria are particular to our correction paradigm (e.g. control technique). Future work will develop an algorithm that reasons explicitly about these different sources.

Acknowledgments

EU Projects IST-2004-004370 (RobotCub), IST-04169 (feelix-growing) and FET-7 grant number 231500 (RobotSkin) funded this research.

References

[1] B. Argall, S. Chernova, M. Veloso, and B. Browning. A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, 57(5):469–483, 2009.

[2] B. D. Argall, E. L. Sauser, and A. G. Billard. Tactile guidance for policy refinement and reuse. In *Under submssion*.

[3] A. Billard, S. Callinon, R. Dillmann, and S. Schaal. Robot programming by demonstration. In B. Siciliano and O. Khatib, editors, *Handbook of Robotics*, chapter 59. Springer, New York, NY, USA, 2008.

[4] F. Orabona, C. Castellini, B. Caputo, A. E. Fiorilla, and G.Sandini. Model adaptation with least-squares svm for adaptive hand prosthetics. In *Proceedings of ICRA '09*, 2009.

[5] P. K. Pook and D. H. Ballard. Recognizing teleoperated manipulations. In *Proceedings of ICRA '93*, 1993.

[6] S. Thrun and T. M. Mitchell. Lifelong robot learning. *Robotics and Autonomous Systems*, 15, 1995.

[3] Note that from the LfD perspective of correspondence between the teacher and learner, each *Teacher-Technique* combination maps uniquely to the learner platform. To consider different *Teacher-Technique* combinations to be distinct sources thus has the benefit of acknowledging different correspondence mappings. Whether to then explicitly model them is an open question; this rarely happens within the LfD literature.