robotics Swiss National Centre of Competence in Research

Inverse Reinforcement Learning for Compliant Manipulation in Letter Handwriting

Ajay Kumar Tanwani, Aude Billard

Ecole Polytechnique Federale de Lausanne, Switzerland

Introduction

Inverse Reinforcement Learning (IRL)

- Learn reward function from human demonstrations
- Applications include inferring human goals/preferences, learning policies that generalize to unobserved situations, transferring biological motion to robots

Letter Handwriting Task

- Teach robot the task of letter writing in a compliant way
- Complex task involving grasping force between the hand and the pen, pen-tip interaction force, orientation of hand

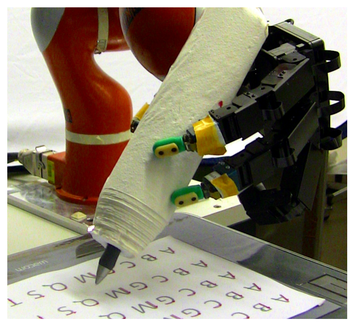
Goals and Objectives

- Humans adapt stiffness during contact with environment
- Humans and robots have different dynamics
- Use reward function in human demonstrations to teach optimal compliance to robot

Different Dynamics

Demonstrations NOT Transferrable

Use reward function to learn variable stiffness profiles for compliant manipulation



Reward Function Transferrable

Same Intention

Methodology: 3-Step Model-Based IRL Framework

1 - Learn Interaction Dynamics

- Collect 3 human demonstrations $\{x_h(t), u_h(t)\}$ for writing each letter $\{a, e, m, o, u, y\}$ to learn the dynamics model:

$$\dot{x} = f_h(x, u)$$

$u \in \mathcal{R}^3 \Rightarrow$ Interaction force $\{F_x, F_y, F_z\}$

$x \in \mathcal{R}^3 \Rightarrow$ Pen-tip position

$\dot{x} \in \mathcal{R}^3 \Rightarrow$ Pen-tip velocity

- f_h is learned from demonstrations using Support Vector Regression

2 - Learn Reward Function

- Reward function scores $\{x_h(t), u_h(t)\}$ on its similarity to known reference letter trajectory $\{x_d(t), u_d(t)\}$

$$r(x_t, u_t) = -(x_t - x_d(t))^T Q (x_t - x_d(t)) - \ldots (u_t - u_d(t))^T R (u_t - u_d(t))$$

$Q = diag[w1\ w2\ w3]$ $R = diag[w4\ w5\ w6]$

- Find Q, R to maximize the probability of $\{x_h(t), u_h(t)\}$ given $\dot{x} = f_h(x, u)$ under the maximum entropy distribution [1]:

$$P(\{x_h(t), u_h(t)\}|Q, R) = \frac{\exp\left(\sum_{t=0}^{T} r(x_h(t), u_h(t))\right)}{\int \exp\left(\sum_{t=0}^{T} r(x_t, u_t)\right) d\pi}$$

$\pi = [u_1, \ldots, u_T]$

3 - Learn Optimal Policy

- Optimal policy $u = \pi(t)$ for the robot dynamics $\dot{x} = f_r(x, u)$ and the learned reward function using iterative Linear Quadratic Regulator [2] gives the time-varying stiffness profiles

$$\pi(t) = u_t + k_t + K_t(x - x_t)$$

$k_t \Rightarrow$ open-loop input correction

$K_t \Rightarrow$ variable stiffness profiles

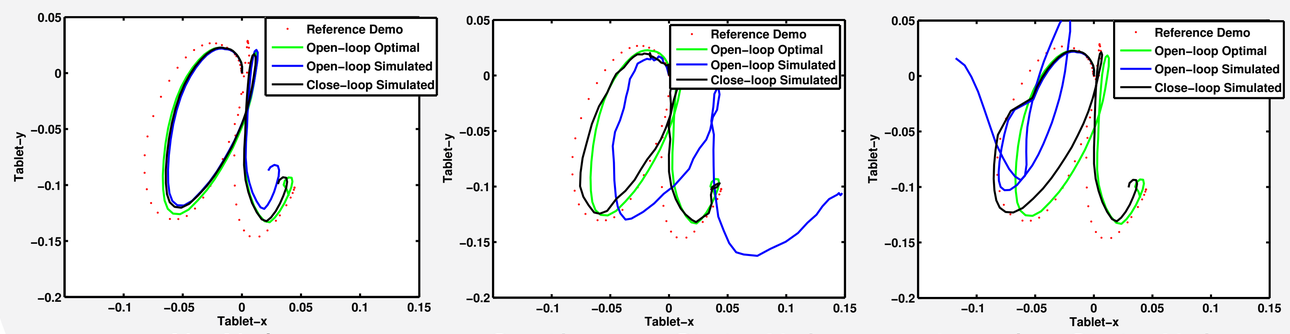
Results

Experimental set-up for letter writing with KUKA robot arm on Tablet

- Learned Reward Function:

$Q = diag[0.53\ 0.26\ 0.0004]$ $R = diag[0.27\ 0.73\ 0.22]$

- Optimal Policy Response: Prelimiary simulation results suggest learning variable stiffness profiles helps to counter perturbations in a compliant manner



No Noise | Continuous State Noise | Impulse State Noise

[1] S. Levine, V. Koltun. Continuous inverse optimal control with locally optimal examples. In *International Conference on Machine Learning 29. 2012.*

[2] W. Li, E. Todorov. Iterative linear quadratic regulator design for nonlinear biological movement systems. In *International Conference on Informatics in Control 1. 2004.*

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

ETH Eidgenössische Technische Hochschule Zürich Swiss Federal Institute of Technology Zurich

USI/SUPSI IDSIA

Universität Zürich UZH

FNSNF Swiss National Science Foundation

The National Centres of Competence in Research are a research instrument of the Swiss National Science Foundation.