Robotics and Autonomous Systems 61 (2013) 1323–1334

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

A syntactic approach to robot imitation learning using probabilistic activity grammars

Kyuhwa Lee *, Yanyu Su, Tae-Kyun Kim, Yiannis Demiris

Personal Robotics Lab, Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2BT, UK

HIGHLIGHTS

- We present a syntactic approach to robot imitation learning.
- It captures reusable task structures in the form of probabilistic activity grammars.
- We aim to learn with a reasonably small number of samples under noisy conditions.
- We evaluate on both synthetic and two real-world humanoid robot experiments.
- Our method shows improvement on imitation learning when compared with other methods.

ARTICLE INFO

Article history:
Received 12 March 2013
Received in revised form
27 July 2013
Accepted 5 August 2013
Available online 19 August 2013

Keywords:
Robot imitation learning
Probabilistic grammars
Activity representation

ABSTRACT

This paper describes a syntactic approach to imitation learning that captures important task structures in the form of probabilistic activity grammars from a reasonably small number of samples under noisy conditions. We show that these learned grammars can be recursively applied to help recognize unforeseen, more complicated tasks that share underlying structures. The grammars enforce an observation to be consistent with the previously observed behaviors which can correct unexpected, out-of-context actions due to errors of the observer and/or demonstrator. To achieve this goal, our method (1) actively searches for frequently occurring action symbols that are subsets of input samples to uncover the hierarchical structure of the demonstration, and (2) considers the uncertainties of input symbols due to imperfect low-level detectors.

We evaluate the proposed method using both synthetic data and two sets of real-world humanoid robot experiments. In our Towers of Hanoi experiment, the robot learns the important constraints of the puzzle after observing demonstrators solving it. In our Dance Imitation experiment, the robot learns 3 types of dances from human demonstrations. The results suggest that under reasonable amount of noise, our method is capable of capturing the reusable task structures and generalizing them to cope with recursions.

1. Introduction

Humans are capable of learning novel activity representations despite noisy sensory input by making use of the previously acquired contextual knowledge, since many human activities often share similar underlying structures. For example, when we observe a hand transferring an object to another place where a grasping action cannot be seen due to some occlusions, we can still infer that a grasping action occurred before the object was lifted.

Similarly, in the process of language acquisition, a child learns more complex concepts and represents them by using previously learned vocabularies. Analogously, the structure of an activity can be represented using a formal grammar, where symbols (or vocabularies) represent the smallest meaningful units of action components, i.e. primitive actions. We are interested in learning reusable action components to better understand more complicated tasks that share the same structures under noisy environments.

The learning of reusable action components is one of the crucial tools for robot imitation learning (also called robot programming by demonstration), which has become an important paradigm, as it enables a robot to incrementally learn higher-level knowledge from human teachers. Our approach shares the concept of imitation learning presented in the Handbook of Robotics (Chapter 59) [1], as well as in [2–5] where a robot learns a new task directly from human demonstration without the need of extensive reprogramming.

There are several important issues in imitation learning: *what* to imitate, *how* to imitate, *who* to imitate, *when* to imitate and how to judge if imitation was successful [6]. In this paper, we mainly

* Corresponding author. Tel.: +44 7540328405.
E-mail address: lee.kyuh@gmail.com (K. Lee).

0921-8890/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.robot.2013.08.003

focus on the issue of *what* to imitate, which is an actively investigated area, where a robot needs to understand the goal or intention of actions, as done similarly in [7–11]. It is also known that humans tend to interpret actions based on goals rather than motion trajectories [12,13]. Another active research area, which studies on solving problems of *how* to imitate, focuses on learning the trajectories of joints (e.g. [14–19]). Although this is not our main focus, we address this issue in our Dance Imitation experiment (Section 5.3).

We are inspired by the work done in [20] which has the same motivation about hierarchical learning. In their work, the authors designed a set of primitive actions which are then used as building blocks, i.e. basic vocabularies, to represent higher-level activities. However, it does not deal with more complex concepts such as recursions which we will deal with here. In this respect, we choose Stochastic Context-Free Grammars (SCFGs) as our representation framework due to (1) robustness to noise as a result of the probabilistic nature, (2) compactness on representing hierarchical and recursive structures, and (3) generation of human-readable output which can be intuitively interpreted for users even without deep technical knowledge. It is worth noting that "context-free" in SCFG is used as a contrast to "context-sensitive", which is another type of grammars, i.e. it does not mean that it lacks the contextual knowledge. Although some other commonly used techniques such as Hidden Markov Models (HMMs) require lower computational complexity, they are often relatively less expressive, and cannot easily represent structures with repetitions and recursions. For example, the recursive activity $a^n b^n$, where a = Push, b = Pull (equal number of Push and Pull operations.), cannot be represented using HMMs. SCFGs extend Context-Free Grammars by adding rule probabilities, a notion similar to state transition probabilities in HMMs. We are especially interested in the real-world applications where noise cannot be avoided. Hence, in our case we consider the symbol probabilities as well as the rule probabilities.

In this paper, we present a method on learning activity grammars from human demonstrations which can be used as a prior to better recognize more complex tasks that share the same underlying components with ambiguity. We assume that (1) the system can detect meaningful atomic actions which are not necessarily noise-free, and (2) extensive complete datasets are not always available but numerous examples of smaller component elements could be found.

2. Related works

A large amount of effort has been spent to understand tasks using context-free grammars (CFGs). In [21], Ryoo defines a game activity representation using CFGs which enables a system to recognize events and actively provide proper feedback to the human user when the user makes unexpected actions. In [22], Ivanov defines SCFG rules to recognize more complicated actions, e.g. music conducting gestures, using HMM-based low-level action detectors. In [23], a robot imitates human demonstrations of organizing objects using SCFG-based task-independent action sequences. For other interesting areas that utilize CFGs as the underlying framework, e.g. computational biology and speech recognition, please refer to [24]. Aloimonos et al. [25] give the detailed explanations about various useful applications that use linguistic approaches including human motoric action representations.

The aforementioned studies consider cases where the proper grammar rules are given in advance. As opposed to manually defining the grammar rules to represent a task, there are also several approaches aiming at constructing (i.e. inducing) grammars from data. In an early work, Nevill-Manning et al. [26] presented the SEQUITUR algorithm which can discover the hierarchical structures among symbols. Solan et al. [27] presented the ADIOS algorithm which induces CFGs and context-sensitive

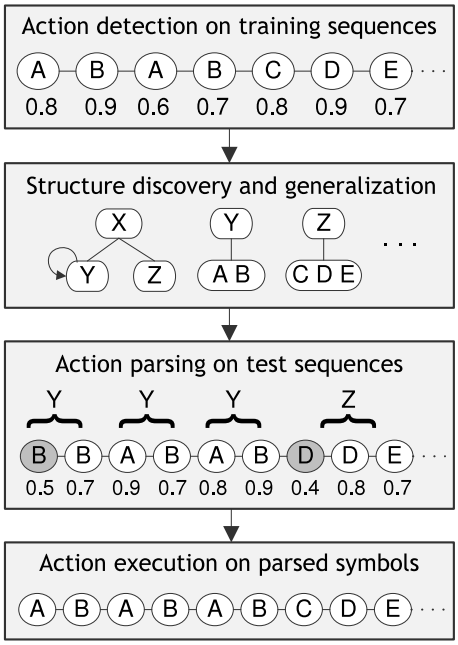

Fig. 1. Overview of our approach to imitation learning with an example. The input training sequences are converted into streams of symbols with probability, respectively indicated by circles and numbers below, from which the original structure is uncovered using grammatical representations. The acquired knowledge is used to better recognize unforeseen, more complex activities (test sequences) that share the same structure components.

grammars as well, with some restrictions (e.g. no recursions) using graphical representations. Stolcke and Omohundro [28] presented a SCFG induction technique, which more recently has been extended by Kitani et al. [29] to remove task-irrelevant noisy symbols to cope with more realistic environments. Lee et al. [30] apply SCFG learning algorithm to discover the optimal number of symbols required to represent a task. In [31], Ogale et al. construct a SCFG grammar based on frequency of human pose pairs, i.e. bigrams, considering slightly varying viewpoints. However, it does not have a generalization step which differs from our method.

Compared to the conventional learning techniques, our method has two distinctive features: (1) our method actively searches for frequently occurring substrings from the input stream that are likely to be meaningful to discover the hierarchical structures of activity; (2) we take into account the uncertainty values of the input symbols computed by low-level atomic action detectors. Fig. 1 gives an overview of our approach with an example for illustrative purpose. Similar to Ivanov's work [22] where they augmented the conventional SCFG "parser" by considering the uncertainty values of the input symbols, we extend the conventional SCFG "induction" technique by considering the uncertainty values of the input symbols.

In [28], Stolcke and Omohundro proposed a technique on merging states which generalizes SCFG rules to deal with unforeseen input with arbitrary lengths, e.g. symbols generated using recursive representations. They introduce two operators, chunking and merging, which convert an initial naive grammar to a more general one. The method assumes that input terminal symbols are deterministic, i.e. all symbols are equally reliable and do not contain any uncertainty values. Our method is different in that it takes into account the uncertainty (or probability) values of input symbols and explicitly searches for regularities using an n-gram-like frequency table within each input sample. This allows our method to learn a better grammar that reflects the noise term included in the observation.

More recently, Kitani et al. [29] presented a framework of discovering human activities from video sequences using a SCFG