Robotics and Autonomous Systems 71 (2015) 118–133

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

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

Model-free incremental learning of the semantics of manipulation actions

Eren Erdal Aksoy*, Minija Tamosiunaite, Florentin Wörgötter

Georg-August-Universität Göttingen, BCCN, Department for Computational Neuroscience, Inst. Physics-3, Friedrich-Hund Platz 1, D-37077 Göttingen, Germany

HIGHLIGHTS

- We addressed the problem of on-line learning of the semantics of manipulations.
- This is the first attempt to apply reasoning at the semantic level for learning.
- Our framework is fully grounded at the signal level.
- We introduced a new benchmark with 8 manipulations including in total 120 samples.
- We evaluated the learned semantic models with 20 long manipulation sequences.

ARTICLE INFO

Article history:
Available online 18 November 2014

Keywords:
Semantics
Manipulation action
Incremental learning
Semantic event chains

ABSTRACT

Understanding and learning the semantics of complex manipulation actions are intriguing and non-trivial issues for the development of autonomous robots. In this paper, we present a novel method for an on-line, incremental learning of the semantics of manipulation actions by observation. Recently, we had introduced the Semantic Event Chains (SECs) as a new generic representation for manipulations, which can be directly computed from a stream of images and is based on the changes in the relationships between objects involved in a manipulation. We here show that the SEC concept can be used to bootstrap the learning of the semantics of manipulation actions without using any prior knowledge about actions or objects. We create a new manipulation action benchmark with 8 different manipulation tasks including in total 120 samples to learn an archetypal SEC model for each manipulation action. We then evaluate the learned SEC models with 20 long and complex chained manipulation sequences including in total 103 manipulation samples. Thereby we put the event chains to a decisive test asking how powerful is action classification when using this framework. We find that we reach up to 100% and 87% average precision and recall values in the validation phase and 99% and 92% in the testing phase. This supports the notion that SECs are a useful tool for classifying manipulation actions in a fully automatic way.

1. Introduction

One of the main problems in cognitive robotics is how to recognize and learn human demonstrations of new concepts, for example learning a relatively complex manipulation sequence like cutting a cucumber. Association-based or reinforcement learning methods are usually too slow to achieve this in an efficient way. They are therefore most often used in combination with supervised learning. Especially the Learning from Demonstration (LfD) paradigm seems promising for cognitive learning [1–5] because humans employ it very successfully. The problem that remains in all these approaches is how to represent complex actions or chains of actions in a generic and generalizable way allowing inferring the essential "meaning" (semantics) of an action irrespective of its individual instantiation.

In our earlier studies we introduced the "Semantic Event Chain" (SEC) as a possible descriptor for manipulation actions [6,7]. The SEC framework analyzes the sequence of changes of the *spatial relations* between the objects that are being manipulated by a human or a robot. Consequently, SECs are invariant to the particular objects used, the precise object poses observed, the actual trajectories followed, or the resulting interaction forces between objects. All these aspects are allowed to change and still the same SEC is observed and captures the "essence of the action" as demonstrated in several action classification tests performed by us [6–9].

[*] Corresponding author.
E-mail address: eaksoye@physik3.gwdg.de (E.E. Aksoy).

http://dx.doi.org/10.1016/j.robot.2014.11.003
0921-8890/© 2014 Elsevier B.V. All rights reserved.

In this paper, we address the problem of on-line, incremental learning of the semantics of manipulation actions observed from human demonstrations. We use the concept of SECs as the main processing tool to encode manipulations in a generic and compact way. Manipulations are continuous in the temporal domain but with event chains we discretize them by sampling only decisive key time points. Those time points represent topological changes between objects and the hand in the scene which are highly descriptive for a given manipulation. Our main intent here is to design a cognitive agent that can infer and learn frequently observed spatiotemporal changes embedded in SECs in an unsupervised manner whenever a new manipulation instance occurs. The learning phase is bootstrapped only with the semantic similarities between SECs, i.e. the encoded spatiotemporal patterns, without using any prior knowledge about actions or objects. Since we use computer vision methods to derive event chains, our approach for incremental learning of semantics is highly grounded in the signal domain. To the best of our knowledge, this is the first attempt to apply reasoning at the semantic level, while being fully grounded at the signal level, to learn manipulations with an unsupervised method. Note, here – on purpose – we do not include any object- or other information to show the power of our methods to fully automatically and in an unsupervised way extract action and object information. Clearly, in praxis, it will often make sense to include whatever additional knowledge is available to further ease action understanding.

The paper is organized as follows. We start with introducing the state of the art. We next provide a detailed description of each processing step; extraction of SEC representations and learning model-SECs for each observed manipulation. In the next section, we discuss experimental results from the proposed framework, which also includes validation and testing of the learned models. We finally conclude with a discussion.

2. State of the art

Learning from Demonstration (LfD) has been successfully applied both at the control [1,2,10] as well as the symbolic level [3–5]. Although various types of actions can be encoded at the control level, e.g. trajectory-level, this is not general enough to imitate complicated actions under different circumstances. On the other hand, at the symbolic level, sequences of predefined abstract action units are used to learn complex actions, but this might lead to problems for execution as many parameters are left out in a symbolic representation. Although our approach with SECs is a symbolic-level representation, SECs can be enriched with additional decisive descriptors (e.g. trajectory, pose, etc.) and do not use any assumption or prior knowledge in the object or action domain. Ideas to utilize relations to reach semantics of actions can be found as early as in 1975. For instance, [11] introduced the first approach about directed scene graphs in which each node identifies one object. Edges hold spatial information (e.g. LEFT-OF, IN-FRONT-OF, etc.) between objects. Based on object movement (trajectory) information, events are defined to represent actions. The main drawback of this approach is that the continuous perception of actions is ignored and is substituted instead by idealized hand-made image sequences. This approach, however, had not been pursued in the field any longer as only now powerful enough image processing methods became available from which object and action information can be extracted.

Still there are only a few approaches attempting to reach the semantics of manipulation actions in conjunction with the manipulated objects [12–18]. The work in [12] is one of the first approaches in robotics that uses the configuration transition between objects to generate a high-level description of an assembly task from observation. Configuration transitions occur when a face-contact relation between manipulated and stationary environmental objects changes. The work presented in [13] represents an entire manipulation sequence by an activity graph which holds spatiotemporal object interactions. The difficulty is, however, that very complex and large activity graphs need to be decomposed for further processing. In the work of [14], segmented hand poses and velocities are used to classify manipulations. A histogram of gradients approach with a support vector machine classifier is separately used to categorize manipulated objects. Factorial conditional random fields are then used to compute the correlation between objects and manipulations. Visual semantic graphs (inspired from our scene graphs) were introduced in [15] to recognize action consequences based on changes in the topological structure of the manipulated object. These visual semantic graphs were further employed together with a context-free manipulation action grammar in [19] to design a cognitive architecture for human manipulation action understanding. In [16] activity trees were presented to recognize actions using a minimal action grammar. The work in [17] suggested a method for hierarchical estimation of contact relationships (e.g. *on* and *into*) between multiple objects. Such contact relations were then employed to execute different daily tasks with robots. Abstract hand movements, such as *moving*, *not moving* or *tool used*, were extracted together with the object information in [18] to further reason about more specific action primitives (e.g. *Reaching*, *Holding*). Recent works such as [20] modeled human activities by employing the human skeleton information as well as roles of manipulated objects. In the modeling process they used the human skeleton information, object segments and their tracks. Likewise, the work in [21] introduced a Bayesian model by using hand trajectories and hand-object interactions while monitoring observed manipulations. In [22] hierarchical models of manipulations were learned with weak supervision from an egocentric perspective without using depth information. Although all those works to a certain extent improve the classification of manipulations and/or objects, none of them extracts key events of individual manipulations and learns a descriptive semantic model in a fully unsupervised manner to represent different manipulation tasks independent from the manipulated objects and their tracks.

In this sense, to our best knowledge, our work is the first study to evaluate and learn the semantics of manipulations in an incremental and model free manner. The concept of semantic event chains has been successfully utilized and extended by others [23–28] not only to represent manipulation actions but also to replicate them by robots. The work in [23] presented active learning of goal directed manipulation sequences, each was recognized using semantic similarities between event chains. Our scene graphs were represented with kernels in [24] to further apply different machine learning approaches. Additional trajectory information was used in [25] to reduce noisy events occur in SECs. Others [26–28] showed execution of various manipulations with different robots by using the key spatiotemporal points provided by SECs.

3. Method

In this method section we will present the core algorithmic components where are complex details will only be given in the Appendix. This should make reading easier, while still everything is present to implement this algorithm if desired.

3.1. Data acquisition

In this work, we investigate eight different manipulation actions: *Pushing*, *Hiding*, *Putting*, *Stirring*, *Cutting*, *Chopping*, *Taking*, and *Uncovering*. Fig. 1(a) shows a sample frame for each manipulation action. All movies used in this study can also be found at www.dpi.physik.uni-goettingen.de/~eaksoye/MANIAC_DATASET.