Robotics and Autonomous Systems 62 (2014) 487–496

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

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

# Object–object interaction affordance learning

Yu Sun *, Shaogang Ren, Yun Lin

*Department of Computer Sci & Eng, University of South Florida, 4202 E. Fowler Ave, Tampa, FL 33613, United States*

ARTICLE INFO

*Article history:*
Received 7 April 2012
Received in revised form
10 October 2013
Accepted 9 December 2013
Available online 14 December 2013

*Keywords:*
Action recognition
Robot learning
Learn from demonstration
Object classification
Graphical model

ABSTRACT

This paper presents a novel object–object affordance learning approach that enables intelligent robots to learn the interactive functionalities of objects from human demonstrations in everyday environments. Instead of considering a single object, we model the interactive motions between paired objects in a human–object–object way. The innate interaction-affordance knowledge of the paired objects are learned from a labeled training dataset that contains a set of relative motions of the paired objects, human actions, and object labels. The learned knowledge is represented with a Bayesian Network, and the network can be used to improve the recognition reliability of both objects and human actions and to generate proper manipulation motion for a robot if a pair of objects is recognized. This paper also presents an image-based visual servoing approach that uses the learned motion features of the affordance in interaction as the control goals to control a robot to perform manipulation tasks.



## 1. Introduction

Object categorization and human action recognition are important capabilities for an intelligent robot. Traditionally, these two problems are treated separately. However, manipulation skills and object affordance are highly related for humans. Therefore, seeking an approach that can connect and model the motion and features of an object in the same frame is considered a new frontier in robotics. With the boom in learning from demonstration techniques in robotics [1–3], more and more researchers are trying to model object features, object affordance, and human action at the same time. Most of the work builds the relationship between single object features and human action or object affordance [4–6].

In daily life, when we are performing tasks, we pay most of our attention to object states or object interactions. For example, when we are writing on paper with a pen, we focus our attention on the pen point, which is the interaction part between the pen and the paper. Moreover, object interaction can directly reveal an object's functions. For instance, when we put a book into a schoolbag, the putting motion tells us that the schoolbag is a container for books. There are endless interactive examples with paired objects in our daily lives. Fig. 1 shows several objects on a table that have an inter-object relationship: a CD and a CD case, a pen and a piece of paper, a spoon and a cup, and a cup and a teapot. In this paper, we attempt to capitalize on the strong relationship between paired objects and interactive motion by building an object relation model and associating it with a human action model in the human–object–object way to characterize inter-object affordance.

The interactive motions of these objects are better defined if we know the interactive pairs. For example, in daily life, we move a teapot in many different ways, such as putting it on a table, storing it on a shelf, and washing it. However, if we have a teapot and a teacup in a scene, water-pouring motion is more likely to occur. Likewise, if we recognize a pouring motion and a teacup, it is very likely that the object associated with the pouring motion is the teapot. We define the interactive motion between paired objects as the object–object-interaction affordance that is connected to both objects. Object–object-interaction affordance is not only useful for object and motion recognition, but also important for robotic learning, as robots can learn object–object-interaction affordance as a manipulation skill that is intrinsic to the paired objects.

Object affordance cognition is one of many core capabilities that a robot needs to gain before it can intelligently perform tasks in the real world. However, this challenging problem has been explored only recently in limited works. Many of the current works model object affordance with interaction between a single object and human action and then use the mutual relationship to improve the recognition of each other. Gupta and Davis [4] recently achieved inspiring success in using single object–action to improve the recognition rate of both the object and human motion. Jiang et al. [7] encoded human preferences about object placements along with the geometric relationship between objects and their placing environments. Kjellstrom et al. [5] used conditional random field (CRF) and factorial conditional random field (FCRF) to model the object type and human action relationship and estimated the 3D hand pose to represent human action, which includes open, hammer, and pour actions. Yao and Li [8] modeled the mutual

* Corresponding author. Tel.: +1 8139747508.
*E-mail address:* yusun@cse.usf.edu (Y. Sun).

0921-8890/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.robot.2013.12.005

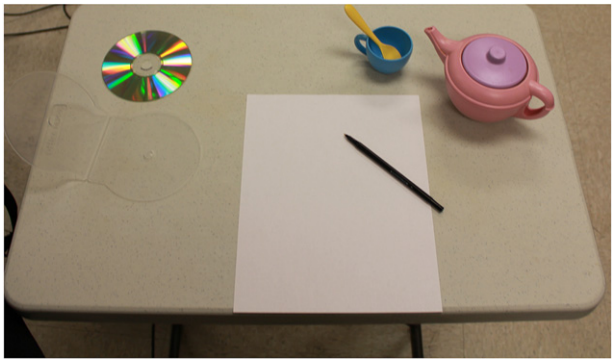

**Fig. 1.** Several objects on a table have inter-object relationships.

context information between human poses and objects in still images using a structure-learning method to model the human and object interaction and achieved a state-of-the-art result in object and human pose detection in static images. Most recently, Gall et al. [6] recovered human action from depth images and used it to represent object function and affordance. In their work, objects were classified according to the involved human action in an unsupervised way based on high-level features.

Some recent works have tried to infer object affordance from object low-level features or 3D shapes. Stark et al. [9] obtained object affordance cues from human hand and object interaction in training images and then detected an object and determined its functions according to its affordance cue features. Grabner et al. [10] proposed a novel way to determine object affordance using computer graphical simulation. With 3D object shapes, their system "imagines" an actor performing actions on objects in a scene to determine the objects' affordances. In [10], first the 3D geometry of a single indoor image was recovered, and then the affordances of the objects were inferred from the joint space of human poses and scene geometry by modeling the physical interaction.

In the robotics community, several works obtained and used object–action relation without considering many low-level object features. In [11], concrete object recognition was not considered, and objects were categorized solely according to object interaction sequences. Objects were segmented out from a number of video sequences, and an undirected semantic graph was used to represent the space interaction relationship between objects. With a sequence of graphs, their work was able to represent object temporal and spatial interactions in an event. With the semantic graphs, they constructed an event table and a similarity matrix. The similarity between two sequences of object interaction events could be obtained according to the similarity matrix. The objects could further be categorized according to their roles in the interactions, and the obtained semantic graphs might be used to represent robotic tasks.

In summary, most current works focus on object–action interaction or low-level object affordance features. Few investigate the affordance relationship between objects. This paper presents a way to model inter-object affordance and then use the inter-object affordance relationship to improve object and action recognition.

Studies in neuroscience and cognitive science on objects' affordance [12] indicate that the mirror neurons in human brains congregate visual and motor responses [13–15]. Mirror neurons in the F5 sector of the macaque ventral premotor cortex fire both during observation of interacting with an object and during action execution, but do not discharge in response to simply observing an object [16,17]. Recently, Yoon et al. [18] studied the affordances associated to pairs of objects positioned for action and found an interesting so-called "paired object affordance effect". The effect was that the response time by right-handed participants is faster if the two objects were used together when the active object (supposed to be manipulated) was to the right of the other object. Borghi et al. [19] further studied the functional relationship between paired objects and compared it with the spatial relationship and found that both the position and functional context are important and related to the motion; however, the motor action response is faster and more accurate with the functional context than the spatial context. The study results in neuroscience and cognitive science indicate that there are strong connections between the observation and the motion, and functional relationships between objects are directly associated with the motor actions. A comprehensive review of models of affordances and the canonical mirror neuron system can be found in [20].

Inspired by the studies above, we propose to capitalize on the connection between the observation of functional-related objects and active functional motion actions to address the skill-learning problem in robotics. In this paper, we simplify the functional-related objects with piece-wise functional-related paired objects and model the inter-object manipulation motions as the inter-object affordance and associate it with paired-object recognition. The goal is to allow robots to learn inter-object affordance motions from humans and then trigger the robot to generate the correct manipulation motion when observing the paired objects.

To model the functional relationship of the paired objects and their relationship with the manipulation motion, this paper presents a graphical model that connects the paired objects and the manipulation motions. The graphical model intuitively represents the functional connectivity of the objects, such as a teapot and a cup or a book and a schoolbag, and extends that connectivity to manipulation motions. A Bayesian Network is employed to model these relationships, in which the paired objects, the interact action, and the consequence of the object interaction are included as a node in the graphical model.

In addition, we developed a method to recognize the paired objects and human motion by analyzing the interactive motion and the statistical knowledge learned from training data. We also constructed a method to leverage object recognition accuracy from videos with the recognition of human interactions, and vice versa. With hand motion trajectory and statistical knowledge learned from training data, the detection accuracy of the interactive objects is significantly improved. With the recognition of the objects, the interactive motions carried out by humans are recognized with much higher accuracy as well.

The interactive motions associated with the paired objects can be learned as the affordance in interaction with statistical models such as Gaussian mixture models. The learned motion can then be directly used to control a robot to perform the proper manipulation motion when the robot sees the paired objects. This paper presents an image-based visual servoing approach that uses the learned motion features in the interaction affordance as the control goals to control the robot to perform the manipulation task instead of manually programming the motion.

We recruited 6 subjects, evaluated our approach with 5 pairs of objects in experiments, and recorded the interactive motion in 50 video sequences.

## 2. Model human–object–object-interaction affordance

Fig. 2 illustrates the workflow of our framework. We first obtained the initial likelihood of the objects' manipulation and reaction. The object initial likelihoods were estimated with a sliding window object detector, which is based on the Histogram of Oriented Gradients (HoG). We estimated the initial likelihood of human action based on the feature of human hand motion trajectory. The human hand was tracked in the whole process, and the hand motion was segmented according to the velocity changing. With motion segmentation and possible object locations, the interactive