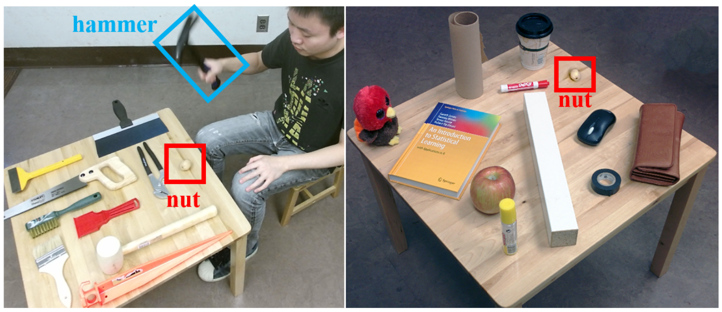
# Understanding Tools: Task-Oriented Object Modeling, Learning and Recognition

Yixin Zhu * Yibiao Zhao * Song-Chun Zhu
Center for Vision, Cognition, Learning, and Art University of California, Los Angeles, CA 90095, USA



**(a) learning** **(b) inference**

Figure 1: Task-oriented object recognition. (a) In a learning phase, a rational human is observed picking a hammer among other tools to crack a nut. (b) In an inference phase, the algorithm is asked to pick the best object (i.e. the wooden leg) on the table for the same task. This generalization entails physical reasoning.

In this paper, we rethink object recognition from the perspective of an agent: how objects are used as “tools" in actions to accomplish a “task". Here a task is defined as changing the physical states of a target object by actions, such as, cracking a nut or painting a wall. A tool is a physical object used in the human action to achieve the task, such as a hammer or brush, and it can be any daily objects and is not restricted to conventional hardware tools. This leads us to a new framework – task-oriented modeling, learning and recognition, which aims at understanding the underlying functions, physics and causality in using objects as tools in various task categories.

Fig. 1 illustrates the two phases of this new framework. In a learning phase, our algorithm observes only one RGB-D video as an example, in which a rational human picks up one object, the hammer, among a number of candidates to accomplish the task. From this example, our algorithm reasons about the essential physical concepts in the task (e.g. forces produced at the far end of the hammer), and thus learns the task-oriented model. In an inference phase, our algorithm is given a new set of daily objects (on the desk in (b)), and makes the best choice available (the wooden leg) to accomplish the task.

From this new perspective, any objects can be viewed as a hammer or a shovel, and this generative representation allows computer vision algorithms to generalize object recognition to novel functions and situations by reasoning the physical mechanisms in various tasks, and go beyond memorizing typical examples for each object category as the prevailing appearance-based recognition methods do in the literature.

Fig. 2 shows some typical results in our experiments to illustrate this new task-oriented object recognition framework. Given three tasks: chop wood, shovel dirt, and paint wall, and three groups of objects: conventional tools, household objects, and stones, our algorithm ranks the objects in each group for a task. Fig. 2 shows the top two choices together with imagined actions using such objects for the tasks.

Our task-oriented object representation is a generative model consisting of four components in a hierarchical spatial-temporal parse graph:

i) An *affordance basis* to be grasped by hand;

ii) A *functional basis* to act on the target object;

iii) An *imagined action* with pose sequence and velocity;

iv) The *physical concepts* produced, e.g. force, pressure.

In the learning phase, our algorithm parses the input RGB-D video by simultaneously reconstructing the 3D meshes of tools and tracking human actions. We assume that the human makes rational decisions in demonstration: picks the best object, grasps the right place, takes the right action (poses, trajectory and velocity), and lands on the target object with the right spots. These decisions are nearly optimal against a large number of compositional alternative choices. Using a ranking-SVM approach, our algorithm will discover the best underlying physical concepts in the human demonstration, and thus the essence of the task.

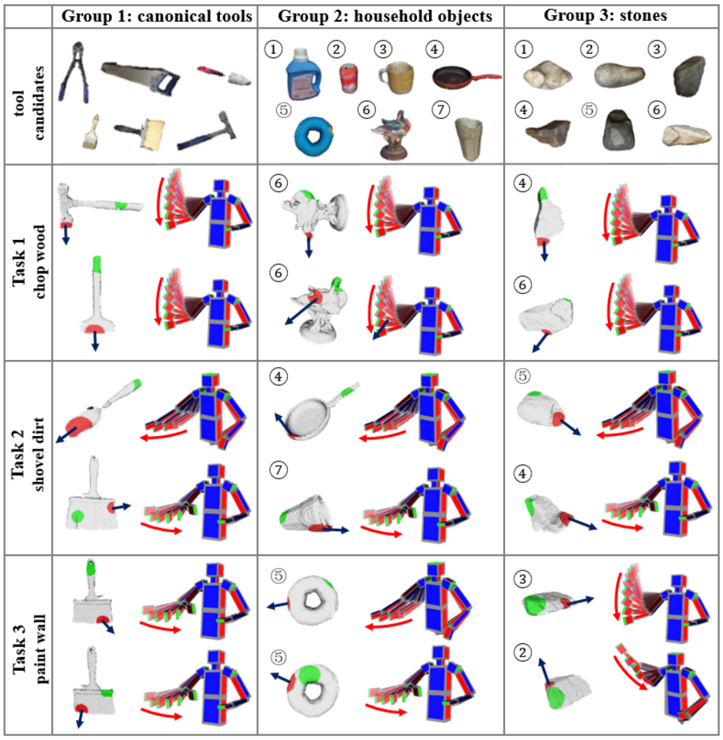


Figure 2: Given three tasks: chop wood, shovel dirt, and paint wall. Our algorithm picks and ranks objects for each task among objects in three groups: 1) conventional tools, 2) household objects, and 3) stones, and output the imagined tool-use: affordance basis (the green spot to grasp with hand), functional basis (the red area applied to the target object), and the imagined action pose sequence.

In the inference stage, our algorithm segments the input RGB-D image into objects as a set of candidates, and computes the task-oriented representation – the optimal parse graph for each candidate and each task by evaluating different combinations. This parse graph includes the best object and its tool-use: affordance basis (green spot), functional basis (red spot), actions (pose sequence), and the quantity of the physical concepts produced by the action.

This paper has four major contributions:

1. We propose a novel problem of task-oriented object recognition, which is more general than defining object categories by typical examples, and is of great importance for object manipulation in robotics applications.

2. We propose a task-oriented representation which includes both the visible object and the imagined use (action and physics). The latter are the ’dark matter’ in computer vision.

3. Given an input object, our method can imagine the plausible tool-use and thus allows vision algorithms to reason innovative use of daily object – a crucial aspect of human and machine intelligence.

4. Our algorithm can learn the physical concepts from a single RGB-D video and reason about the essence of physics for a task.

Yixin Zhu and Yibiao Zhao contribute equally to this work. This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.