Learning Visual Features that Predict Grasp Type and Location

Di Wang
Symbiotic Computing Laboratory
University of Oklahoma
Norman, OK 73019
di@cs.ou.edu

Andrew H. Fagg
Symbiotic Computing Laboratory
University of Oklahoma
Norman, OK 73019
fagg@cs.ou.edu

ABSTRACT

The general goal of a robotic grasping algorithm is to find a set of points of contact between the object and the hand that allows the object to be grasped reliably. Napier (1993) and others (c.f., Cutkosky, 1989; MacKenzie and Iberall, 1994) define a grasp type, or category, in terms of the set of contacts made by the hand, the possible forces at each contact, and the ability of the contacts to resist perturbations. Our goal is for a robot to learn the mapping between images of objects and possible contact points by observing examples of successful grasps. Our hypothesis is that observed grasp type can provide meaningful labels to drive the learning of visual models that capture partial object shapes. In an opposite way, a visual model learned by our algorithm will suggest an approximation to how the hand should be positioned relative to the object in novel scenarios containing novel objects.

Visual feature descriptors such as SIFT (Lowe, 2004) are very discriminative and have been shown to perform well at recognizing specific objects. However, SIFT features tend not to be generalizable across small differences in objects because they capture very specific arrangements of edges and texture. Our goal is to learn visual representations that can be linked to grasping actions, and that do not rely on high-fidelity volumetric or mesh-based models of specific objects. Our approach is to employ visual features that 1) can recognize common components of objects, 2) are locally robust to appearance deformations due to small rotations in 3D, and 3) are robust to common variations in geometries across objects (e.g., the shape of a mug handle). Furthermore, our approach is to define the visual feature categories by clustering examples based on how the object component is actually grasped by a hand.

Given tuples of an example grasp (hand pose and configuration, relative to the object) and the object image during this grasp, our proposed algorithm will learn a set of visual features that will recognize object components in a novel image and suggest appropriate grasps. We assume that tuples corresponding to a single object are labeled as such (i.e., a set of tuples is sampled as an object is manipulated). Although there will be several examples of each object class within the training data set, we make no assumptions about the relationships between these different objects.
REFERENCES


