Object Manipulation Based on Memory and Observation

Li Yang Ku, Mitchell Hebert, Erik Learned-Miller, and Rod Grupen
College of Information and Computer Science, University of Massachusetts Amherst, USA.

Object and tool manipulation are related to the ability to modify one's environments and are common to many intelligent species. This ability is an important part of a complete account of intelligence and will likely be a cornerstone of machine intelligence as well. Moreover, a computational account of object and tool manipulation may provide insight into the mechanisms and processes that give rise to it in nature. Traditional approaches often consider a pipeline of segmentation, object recognition, and pose estimation to be an essential perception stage prior to motor activity. In this work, we propose an alternative approach to object recognition and pose estimation.

In computer vision and robotics, object recognition is often defined as the process of labeling segments in an image or fitting a 3-D model to an observed point cloud. The object models used to accomplish these tasks usually include information about visual appearance and shape. However, what these object recognition systems provide is merely a label for each observed object. The sequence of actions that the robot should perform based on the object label are often manually defined. Without linking actions to object labels these models have limited utility to the robot. In this work, we propose an object model that is composed of a set of viewpoint-specific observations to capture how actions change observation of the object.

In most robotics tasks that require manipulating known objects, pose estimation is often required before planning end effector trajectories. In the Willow Garage grasping pipeline [6], the iterative closest point algorithm is used to check how well a segmented point cloud matches to a stored mesh model. Precomputed grasp points associated with the model are then used to generate a valid motion trajectory. However, object pose estimation is often computationally expensive and inaccurate. In addition, there are many examples of tasks that do not require pose information. This paper proposes using the aspect transition graph object model that skips pose estimation and interacts with objects directly based on memory and observation.

1 Object Model

In the past few decades, experiments in psychophysics and neurophysiology have provided converging evidence that objects are represented in the human brain as collections of viewpoint-specific local features instead of sets of object centered features. It has been shown that when a new object is presented to a human subject, only a small subset of canonical views are retained in memory despite the fact that each viewpoint is presented to the subject for the same amount of time [2] [1]. Experiments on monkeys further confirm that a significant percentage of neurons in the inferior temporal cortex respond selectively to a subset of views of a known object [5].

Closely related to these observations, aspect graphs [3] were first introduced as a way to represent 3-D objects using multiple 2-D views in the field of computer vision. Extending the original concept of aspect graph, we introduce the Aspect Transition Graph (ATG) object model that summarizes how actions change viewpoints and/or the state of the object and, thus, the observation [4]. We define the term "observation" to be the combination of all sensor feedback of the robot at a particular time and the "observation space" as the space of all possible observations. This limits the model to a specific robot, but allows the model to represent object properties other than viewpoint changes alone. Extensions to tactile, auditory and other sensors are possible with this representation. An ATG object model can be used to plan manipulation actions for that object to achieve a specific target aspect. For example, in order for the robot to pick up an object, the target aspect is a view where the robot's end effector surrounds the object. We expect that this view will be common to many such tasks and that it can be the expected outcome of a sequence of open-loop controllers (like moving the end effector to the same field of view as the target object) and closed-loop controllers (like visually servoing features from the hand into the pregrasp configuration relative to the object).

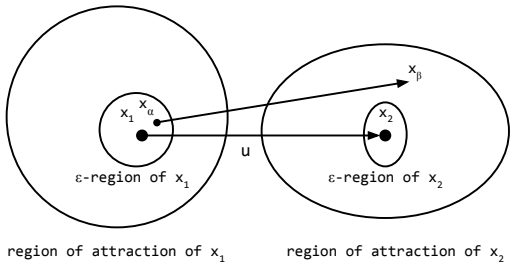

Figure 1: An ATG model containing two aspects x_1 and x_2. The edge labeled u is a model-referenced memorized action that reliably maps the ε-region of x_1 to the interior of the region of attraction of x_2.

An ATG is represented as a directed *multigraph* $G = (\mathcal{X}, \mathcal{U})$, composed of a set of aspect nodes $\mathcal{X}$ connected by a set of action edges $\mathcal{U}$ that capture the probabilistic transition between aspects. An action edge U is a triple (X_1, X_2, A) consisting of a source node X_1, a destination node X_2 and an action A that transitions between them. Note that there can be multiple action edges (associated with different actions) that transition between the same pair of nodes.

Figure 1 shows an example of an ATG model that contains two aspects x_1, x_2 and one action edge u connecting the two aspects in the observation space. An aspect is represented as a single dot in the figure. The ellipses around x_1, x_2 represent the ε-region of the corresponding aspect. Inside the ε-region, the observation is close to the target aspect, and is considered to have "converged". The ε-region is task dependent; a task that requires higher precision such as picking up a needle will require a smaller ε-region. Each aspect x is located in the ε-region but does not have to be in the center. The location and shape of the ε-region also depends on the given task since certain dimensions in the observation space might be less relevant when performing certain tasks.

The larger ellipses surrounding the ε-regions are the region of attraction of the closed-loop controller referenced to aspects x_1 and x_2. Observations within the region of attraction converge to the ε-region of the target aspect by running a closed-loop controller that does not rely on additional information from the object model. In our experiment, a visual servoing controller is implemented to perform gradient descent to minimize the observation error. The region of attraction for using such a controller is the set of observations from which a gradient descent error minimization procedure leads to the ε-region of the target aspect.

The arrow in Figure 1 that connects the two aspects is an action edge (x_1, x_2, a) that represents a memorized action. Action a is an open-loop controller that causes aspect transitions. Instead of converging to an aspect, open-loop controllers tend to increase uncertainty in the observation space. Under situations when there is no randomness in observation, action execution and the environment, executing action a from aspect x_1 will transition reliably to aspect x_2.

In a system which actions have stochastic outcomes, the arrow in Figure 1 that connects the observation x_α within the ε-region of x_1 to observation x_β represents a scenario where action a is executed when x_α is observed. We define ε_u as the maximum error between the aspect x_2 and the observation x_β when action a is executed while the current observation is within the ε-region of aspect x_1. ε_u can be caused by a combination of kinematic and sensory errors generated by the robot or randomness in the environment. If the region of attraction of the controller that converges to aspect x_2 covers the observation space within ε_u from x_2, by running the convergent controller we are guaranteed to converge within the ε-region of aspect x_2

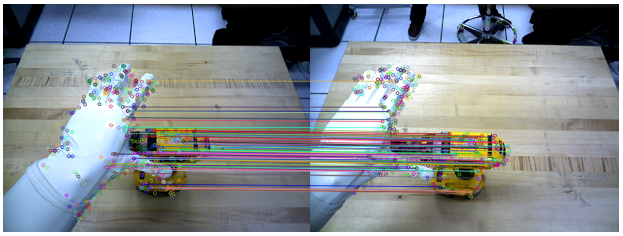

Figure 2: Visual servoing. The target aspect is on the left and the current observation is on the right. A line in between represents a pair of matching keypoints. The goal is to converge current observation to the target aspect.

under such environment. As long as the transitioned observation is within the region of attraction of the next aspect we can guarantee convergence to the desired state even when open-loop controllers are within the sequence.

We call an Aspect Transition Graph model *complete* if the union of the regions of attraction over all aspects cover the whole observation space and a path exists between any pair of aspects. A complete ATG object model allows the robot to manipulate the object from any observation to one of the aspects. Complete ATG object models are informative but often hard to acquire and do not exist for irreversible actions. On the other hand, it is not always necessary to have a complete ATG to accomplish a task. For example, a robot can accomplish most drill related tasks without modeling the bottom of the drill. Therefore, we define an Aspect Transition Graph object model to be *sufficient* if it can be used to accomplish all required tasks with the object. In this work, we will focus on sufficient ATG object models.

2 Manipulation

Object manipulation of known objects often require estimating the object pose before calculating grasping points and trajectories. Instead of storing a 3D model that contains invariant features in the object frame our aspect transition graph (ATG) model stores a set of viewpoint-specific observations. Since available actions from one observation to another are modeled in action edges in an ATG model, object pose estimation is not necessarily required in order to interact with an object. Given a sufficient ATG model, object pose estimation may be skipped and the robot can directly interact with objects based on memorized observations.

In our experiment a visual servoing controller is used to converge from an observation within the region of attraction to the ε-region of the corresponding aspect. The visual servoing controller is used to control the end effector of the robot to reach a pose relative to a target object using visual sensor feedback. Unlike many visual servoing approaches, our visual servoing algorithm does not require a set of predefined visual features on the end effector or target object nor does it require an inverse kinematic solution for the robot. A visuomotor Jacobian, defined as the derivative of each feature's location and orientation in the image plane with respect to the robot's joint configuration, is learned online using Broyden's method. To acheive convergence, the only information required is the current observation and the target aspect. Figure 2 shows an example where the visual servoing algorithm tries to converge from the current observation on the right to the target aspect on the left.

In this work, an action edge in an aspect transition graph model represents a memorized action that performs a movement relative to a point in observation or an end effector pose of the robot. For example, in one of the drill grasping tasks shown in Figure 3 the robot is trained to move its hand first to a pre-grasping pose before moving to a pose that contacts the object to increase accuracy; the first action edge in this ATG model represents an end effector movement relative to the center of the paritally observed point cloud such that the end effector becomes visible at the pre-grasping location. The second action edge in this ATG model represents an end effector movement relative to the last end effector pose such that the end effector reaches a grasping pose based on the memorized movement. Visual servoing is executed after each open-loop action to minimize the error between the current observation and the target aspect.

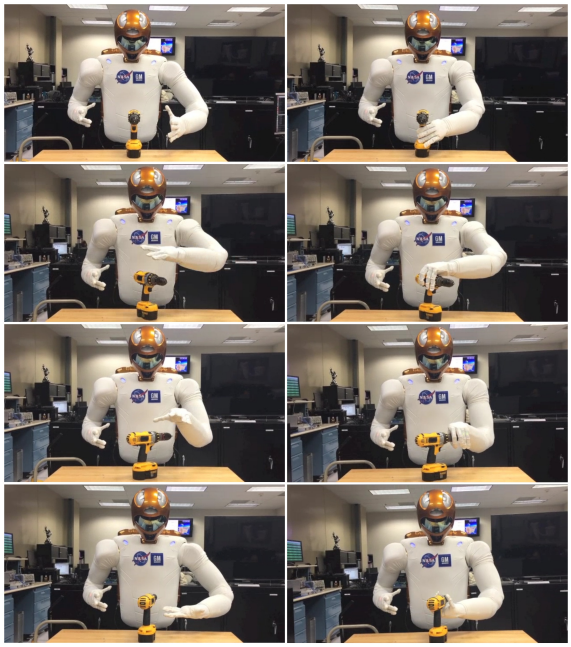

Figure 3: Robonaut 2 grasping the drill posed at different orientations. Image pairs in the same row represents the intermediate and final states of one drill grasping trial.

3 Experiment

In this work, we tested our approach on a tool grasping task on Robonaut 2. The goal is to have Robonaut 2 use the drill directly with its left hand or grasp the drill from the top or the side of the drill with its left hand so that it can adjust the drill to a better grasping pose for the right hand. An aspect transition graph model of a drill is first created from a teleoperator demonstration. Five different grasping trajectories for five different drill orientations ranging from 0 to 180 degrees are shown to the robot. One goal aspect that represents successfully grasping the drill is created and used to connect all five grasping demonstrations. A grasping test is then performed on 21 random drill poses ranging from 0 to 180 degrees and within 10 cm from the original training position. Our approach successfully grasped the drill 19 out of 21 times in this experiment. One of the two failures was caused by the planner failing to generate a valid trajectory to an intermediate aspect and the other was caused by failing to reach an intermediate aspect. Figure 3 shows examples of Robonaut 2 grasping the drill oriented at different poses during testing.

4 Acknowledgment

This material is based upon work supported under Grant NASA-GCT-NNX12AR16A and a NASA Space Technology Research Fellowship. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Aeronautics and Space Administration.

5 References

[1] Heinrich H Bülthoff and Shimon Edelman. Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences*, 89(1):60–64, 1992.

[2] Shimon Edelman and Heinrich H Bülthoff. Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. *Vision research*, 32(12):2385–2400, 1992.

[3] Jan J Koenderink and Andrea J van Doorn. The internal representation of solid shape with respect to vision. *Biological cybernetics*, 32(4):211–216, 1979.

[4] Li Yang Ku, Erik G Learned-Miller, and Roderic A Grupen. Modeling objects as aspect transition graphs to support manipulation. *International Symposium on Robotics Research*, 2015.

[5] Nikos K Logothetis, Jon Pauls, and Tomaso Poggio. Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, 5(5):552–563, 1995.

[6] Melonee Wise and Matei Ciocarlie. ICRA Manipulation Demo, 2010. URL http://wiki.ros.org/icra_manipulation_demo. [Online; accessed 19-September-2015].