Visual Descriptor Learning for Predicting Grasping Affordances

In the recent decade a change has been happening as robots to an increased extent have started to emerge in the world of humans. Examples of such are robot vacuum cleaners, hotel receptionists etc. In these examples, the task of interacting in the highly dynamic world of humans has been addressed. However, the task they are performing is still rather constrained. A robot vacuum cleaner is for instance not able to move a chair if it is in the way, instead it senses it and moves around it. This requires strategies to handle such dynamic situations. Although vacuuming is a rather constrained task, it still outlines some of the many complexities that exist in our world.

A key aspect for robots to interact in the world of humans is the ability to get an understanding of its local environment given sensor information. Such sensors are cameras that see the world or tactile sensors that feel the world. A special kind of sensor which is popular these years is range sensors that in addition to a colour image provide additional depth information. This means that it can see 3D structures in the world. This type of sensor is used throughout this work.

In this thesis, the challenge of making robots to interact in the world of humans is addressed by the task of grasping unknown objects. This task involves representing the object that the robot sees with its sensors and understand it in a way such that a suggestion can be made of how to grasp it.

The contributions from this thesis stem from three works that all relate to the task of grasping unknown objects but with particular focus on how to represent the visual data in a good way.

In the first work, the problem of representing what the robot perceives from sensors was investigated. The aim was to find suitable properties of 3D areas on objects that can indicate a good place to grasp. As a part of this investigation areas of different sizes were used as well as areas with added knowledge such that it is close to a boundary of the object. The results from this investigation provided insights into the importance of selecting an appropriate visual representation when using it for a specific task such as grasping.

Given this knowledge, a visual descriptor was developed with the aim of interpreting the visual data acquired from sensors to meaningful visual structures. The descriptor allowed for describing 3D areas of an object by its curvature, if it was a double sided structure such as “walls” and whether the area was close to the edge of the object, for example at the rim of a cup. Given these properties of local areas on objects, a system was trained that enabled grasping of a new object.

In the final work of the thesis, a mechanism was suggested for learning areas on objects in a hierarchical way. Small areas were combined pairwise to create larger areas. This process was performed N times such that N layers of area descriptors were found. Based on these descriptors, grasps were learned that enabled grasping of novel objects.

The contributions from this work and the gained insights into visual representations might help move robots a bit closer in being able to understand and deal with the complexity of our world.