Contents

1 Introduction *Malthe Høj-Sunesen*

According to ISO 8373 [ISO, 2012], at least two different types robots exist: Industrial robots and service robots. An industrial robot is defined as aa "automatically controlled, reprogrammable, multipurpose manipulator programmable in three or more axes", while a service robot is defined as a "robot that performs useful tasks for humans or equipment excluding industrial automation applications". The classical application of an industrial robot is to have the robot do a predefined behavior repeatedly, while service robots are still very much under development. Due to hardware and software concerns, robots in the industry have previously not seen adaptive behavior, so elements must be aligned in a specific way. Humans, and indeed most animals, are able to look at objects and grasp accordingly. A lot of research is going into making the robot able to understand what it is "looking" at much like humans can, and how to grasp it. This research into grasping[1] objects using only visual cues is the focus point for this report.

[1] For the purposes of this report, *grasping* is to pick up an object.

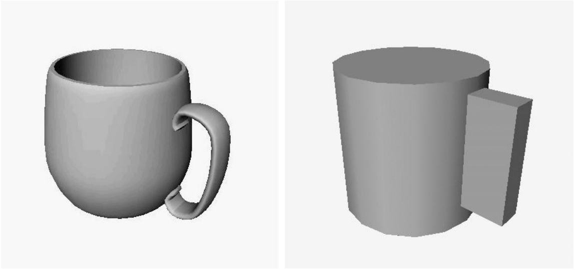


Figure 1: "A mug and its primitive representation"; a cylinder and a box. From [Miller et al., 2003].

2 Motivation *Malthe Høj-Sunesen*

Humans spend years learning how to grasp objects. Babies have a hard time figuring out how to grasp even the most simple objects, and parents solve that problem by giving babies and children plastic cutlery, bouncy, soft toys and always walking around with an eye on each finger. We come to expect of a child to drop toys, knock over glasses, and the like.

A robot is not allowed to fail in the same way. When a robot's hand grasps something we expect it to not let it go — or worse, drop it — before it is supposed to. In a tightly controlled production line that is not a problem. Using embodied AI the parts can be aligned perfectly for the robot and the robot can assemble the parts correctly.

In a not so tightly controlled environment among people it is a bigger problem. If a service robot is supposed to clean up mess left after a human, it is almost guaranteed that the parts are not aligned as a robot could predict. If an industrial robot can figure out the best grasp autonomously for an object it would decrease operator dependency, leading to faster setup and lower costs for the company.

3 Simplifying objects to primitive models *Malthe Høj-Sunesen*

Biederman suggested that elements can be broken down to geons, basic elements describing one feature of an object.

In [Miller et al., 2003] the idea behind geons is used to help a robot simulator find good grasps. The robot knows how to grasp each shape primitive (equivalent to geon). Any object is then reduced to its shape primitives where applicable. This allows a simulator to know which points are good to grasp, resulting in simpler calculations. An example of this reduction can be seen in Figure 1 with shape primitive building bricks shown in Figure 2.

Reducing the visual information in this way will give the simulator a simpler task, as it does not have to simulate thousands of possible grasps but

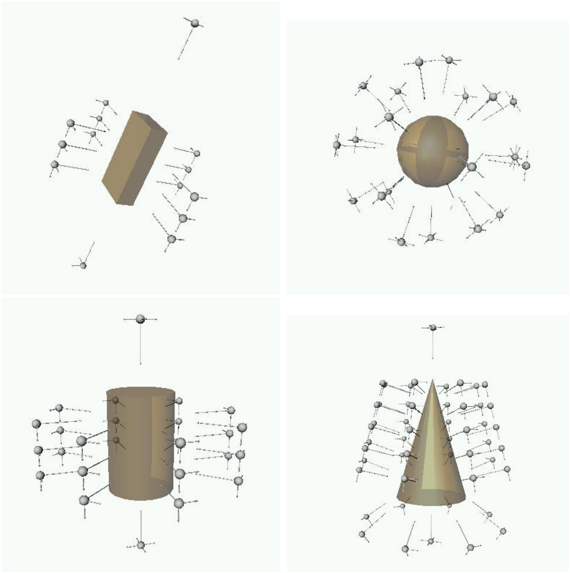


Figure 2: "Examples for grasp generation on single primitives. The balls represent starting positions for the center of the palm. A long arrow shows the grasp approach direction, and a short arrow shows the thumb direction. In most grasp locations, two or more grasp possibilities are shown, each with a different thumb direction." From [Miller et al., 2003].

only the grasps based on the preshape grasps per primitive representation. An example of the found grasps can be seen in Figure 3.

4 Edges and textures *Hsin-Yu Lee*

If we don't want to use shape primitives as our way to generate grasping gestures, there is an alternative way when we want way to recognize the object just base on the stereo images from the cameras. The approach in [Kootstra et al., 2012] tries to imitate the way that human vision system try to recognize unknown things. The analyzing system is called "biological-motivated hierarchical vision system" [Pugeault et al., 2010], which also called "ECV", the "Early Cognitive Vision". Literally, the system was inspired by the primate's vision system. By using this system, the 3D features of edge and surface of the object are naturally aligned together. The basic analyzing process of the system can be seen in Figure 4 on the following page.

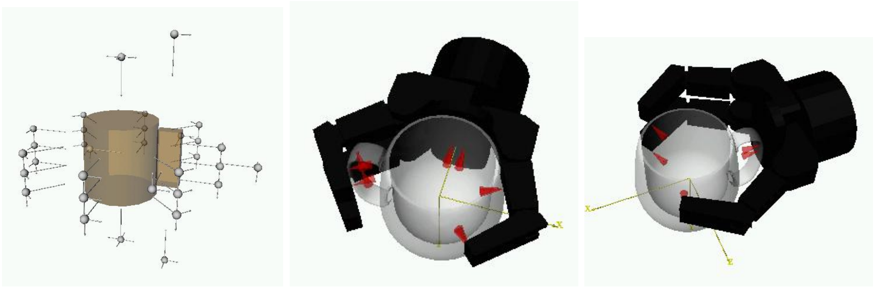


Figure 3: The primitive mug representation and the two best grasps. The red cones indicate point-of-contact. From [Miller et al., 2003].

.

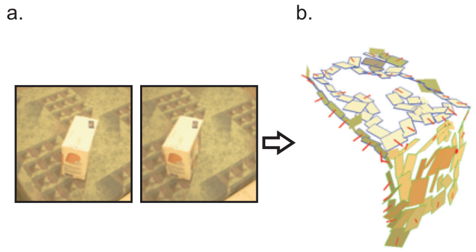


Figure 4: a. The images captured from the stereo cameras. b. The analyzing result of implementing the ECV on the stereo images. From [Kootstra et al., 2012].

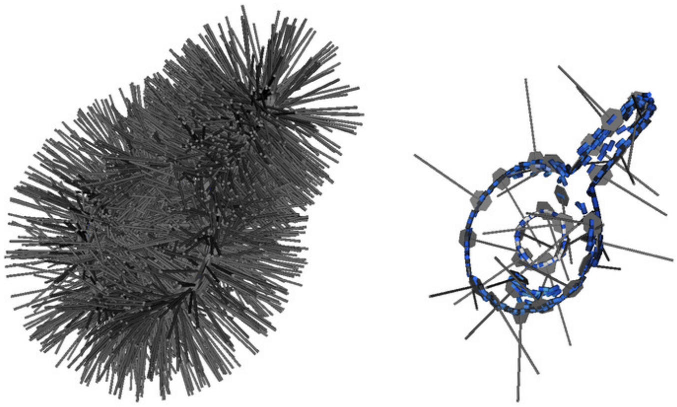


Figure 5: Left: A full graph of the possible grasp positions. Right: Clearer showing of what the sticks mean; the stick is the robotic hand's translation while the paddle at the end is the point where the fingers close. From [Detry et al., 2011]

5 Learning to grasp through attempts *Malthe Høj-Sunesen*

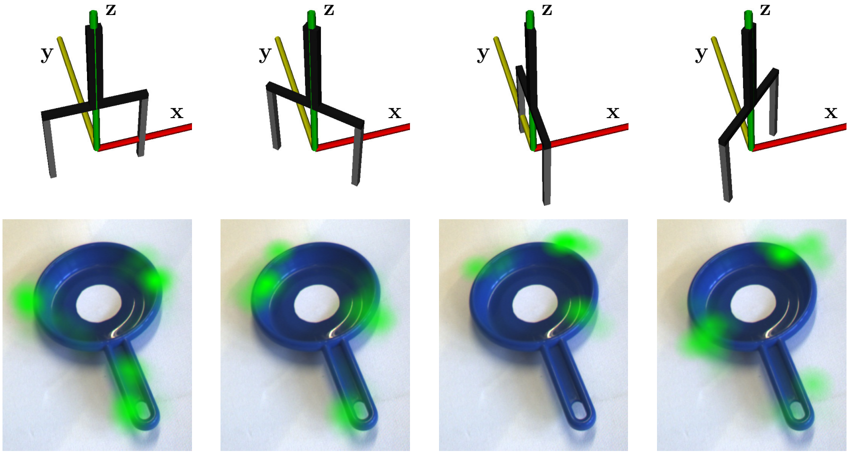
Much like [Miller et al., 2003] in Section 3 tried to emulate how the human vision works according to Biederman, so do [Detry et al., 2011] try to emulate how a child learns to grasp objects. Any parent will tell you that their child did not quite know how to actively grasp[2] toys from the beginning. Where to hold is one of the problems.

The approach in [Detry et al., 2011] is to let a robotic platform learn how to grasp a single object. Using stereo vision, the 3D features of an object can be calculated. The system will then try to calculate where that object can be picked up. An example of where the system calculates a toy pan can be grasped can be seen in Figure 5. s

After calculating where the object can be grasped, the robot will start to grasp the object, time and time again. In [Detry et al., 2011], the robot performed more than 2000 grasps. During the grasp trials the robot and vision system will see if the grasp is stable, ie. if the object is not moving. Using all the trials the robot system can build a model of the possibility that a grasp will be successful. The result is visualized in Figure 6 on the following page.

In the end, the robot has learned how to pick up an object from any angle, and is able to choose the best possible grasp in any situation. This is indeed how children and grown-ups know how to pick up objects as well.

[2] Let alone letting it go again!



(a) Downward grasp (b) 45° rotation around **z** (c) 90° rotation around **z** (d) 135° rotation around **z**

Figure 6: "Various projections of a grasp density generated by a robot". The greener the pixel, the higher the density and thus probability of a good grasp. From [Detry et al., 2011].

6 Elementary grasping actions *Hsin-Yu Lee*

The Elementary Grasping Actions, which also called EGA, is the specific grasping gestures used in the paper [Kootstra et al., 2012] that was implemented once after we find the object by the ECV system. The EGA was also mentioned and defined in the paper [Pugeault et al., 2010]. The Figure 7 on the next page shows the three different ways of grasping object base on edge and surface information separately. Basically, the directions, approaching ways and positions are pre-defined, only slightly changed according to the size of the surface or the distance between two selected contours.

We can simply choose one method to grasp something; we also can use the simulator to help us select the best grasping action through these methods. There are several experiments presented in the paper [Kootstra et al., 2012] shows the performance of each different actions and in the circumstance of using multiple method at the same time.

7 Experiments of grasping actions

7.1 Kootstra et al *Hsin-Yu Lee*

To gain insights in the performance of the different grasp generation methods, they have been tested in two different experimental setups. The experiments result presented in the paper [Kootstra et al., 2012] shows the performance of each grasping method and also the performance of mixed actions.

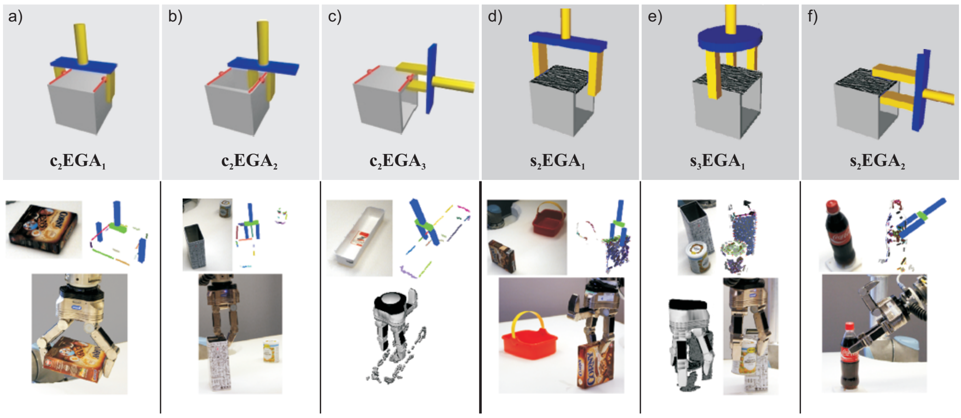


Figure 7: The elementary grasping actions (EGA) are illustrated on the top with some examples of real grasps on the bottom. (a-c) The three contour-based EGAs. The red lines indicate the selected contours. (d-f) The surface-based EGAs. The dark faces show the selected surface. The first letter in the naming scheme marks the type of features used to generate a grasp. 'c' stands for contour and 's' for surface. The first subscript stands for two or three fingers. The last subscript marks the general type of grasp, where '1' is an encompassing grasp, '2' is a pinch grasp from the top and '3' is a pinch grasp from the side of the surface. For each type of grasping action, an example is shown consisting of an original image, a snapshot of the ECV representation along with the selected grasp, and the grasp execution in the simulation/real setup. From [Kootstra et al., 2012]

The Figure 8 on the following page shows part of the experiment result that the grasping has been tested in the real world environment and using the simulator. The bar plots give the distributions of all grasps averaged over the scenes. The results are split up for the different conditions: single objects standing up, single objects laying down, two objects close together, and two objects far apart. Results are labeled with nograsp when a method does not generate any grasps for a given scene. In addition, the consistency of the methods is shown by the black error bars, which give the average standard errors on the proportion of successful grasps (stable+slipped), where the standard error is calculated over the set of different poses within a pose condition of a scene, e.g., the four different poses of the coca-cola bottle standing upright.

It can be seen that, in general, the surface-based grasp methods perform better than the contour-based methods. The encompassing grasps, i.e., EGA_1 grasps, are more successful than the pinch grasps. The three-finger surface-boundary grasp, $\mathrm{sb}_3\mathrm{EGA}_1$ outperforms the two-finger surface-boundary grasp, $\mathrm{sb}_2\mathrm{EGA}_1$. All grasping methods perform better when the single objects are standing upright than when they are laying down. For the double-object scenes, the amount of successful grasps is similar for the objects apart or close together, but in the latter case, there is a higher proportion of collision.

7.1.1 Grasp success as a function *Hsin-Yu Lee*

The average grasp success rates for the different methods lie between 0.2 and 0.6. Although this is the high performance for a system that does not use any direct presetting information from the object, it still means that the robot will often not be able to grasp an object at the first attempt. Assuming that the robot can make a new grasp attempt when the previous grasp fails, we investigate how the success rate increase as a function of the number of grasp attempts.

Figure 9 on the next page shows the grasp success as a function of the number of grasp attempts, N. We define it as success if any of the grasp attempts is successful. The plots show the average success rate for all scenes over 20 randomized runs. And as can be seen in Figure 9 on the following page good grasp results are generally achieved already after a few attempts, especially when the different grasp methods are combined. The combination of sparseness, complementarity, and high performance is the main contribution of our method.

7.2 Detry et al *Malthe Høj-Sunesen*

In order to evaluate the theory, a test setup was made. This setup features *an industrial robot arm, a force torque sensor, a two-finger gripper, and a*

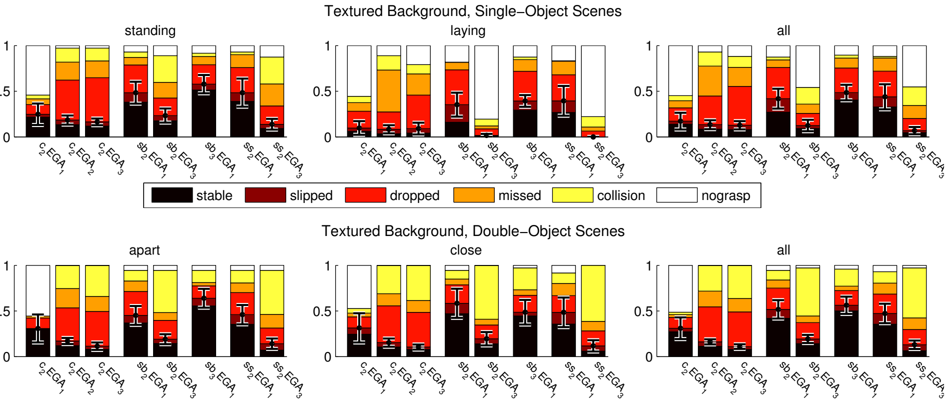


Figure 8: Grasp results for the hybrid real-world and simulated experiments. The stacked-bar plots show the average distribution of all grasps over all scenes. The stable and slipped grasps are considered successful grasps, where the object is held in the hand after lifting. The gray area shows the proportion of scenes where the methods do not suggest any grasps. The black error bars give the average standard errors on the proportion of successful grasps (stable+slipped). The standard error is calculated over the set of different poses within a pose condition of a scene. The error bars show the consistency of the method to different poses of the same object. From [Kootstra et al., 2012].

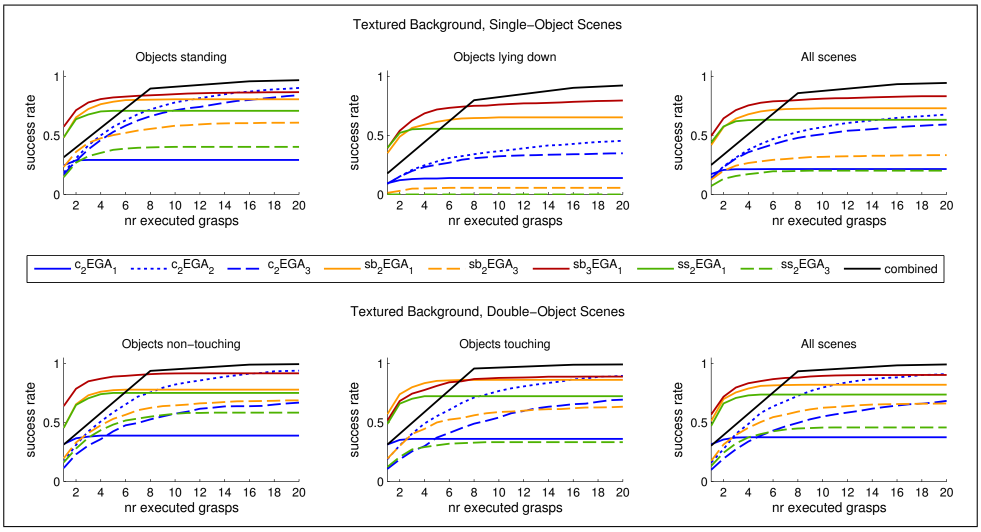


Figure 9: Grasp results for the hybrid real-world and simulated experiments. The plots show the average success rate as a function of the number of grasp attempts for the different grasp methods. The black line shows the performance when the different methods are combined. The fact that the combined results surpass the results of the individual methods shows that they are complementary. From [Kootstra et al., 2012].

stereo camera [Detry et al., 2011]. The robot is controlled by an FSM:

1. Estimate object pose, align grasp density,
2. Produce desired grasp,
3. Submit grasp to grasp planner,
4. Move gripper to pose,
5. Close gripper fingers,
6. Lift object, and
7. Drop object

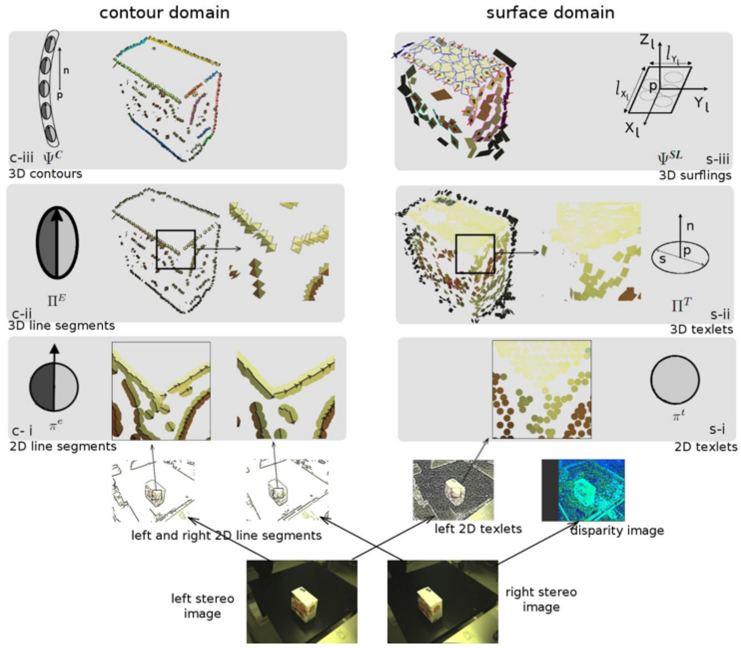


Figure 10: The hierarchical representation of contour and texture information in the ECV system. The stereo images at the bottom are the real world images captured by the camera, the others pictures show the process in the simulator try to find the object based on the edge and surface information. From [Kootstra et al., 2012].

8 ECV system *Hsin-Yu Lee*

Figure 10 shows the hierarchical vision system "ECV" implemented in the paper [Kootstra et al., 2012]. The system recognizes objects by 2D and 3D geometrical and appearance relations between visual entities at the different levels of the hierarchy in two major domains, which are edge and surface. The process of recognizing the edge is: First, use the algorithm of image processing to transfer the stereo images into the 2D line segment images. Second, use the mathematic way to find out the smallest line segments. Then, combine the line segments into a larger segment. Finally, the edge of the objects will come out. It's also similar to the process of forming the surfaces of the object.

References

[Detry et al., 2011] Detry, R., Kraft, D., Kroemer, O., Bodenhagen, L., Peters, J., Krüger, N., and Piater, J. (2011). Learning grasp affordance densities. *Paladyn, Journal of Behavioral Robotics*, 2:1–17.

[ISO, 2012] ISO (2012). Robots and robotic devices — vocabulary. ISO 8373:2008(en), International Organization for Standardization, Geneva, Switzerland.

[Kootstra et al., 2012] Kootstra, G., Popović, M., Jørgensen, J. A., Kuklinski, K., Miatliuk, K., Kragic, D., and Krüger, N. (2012). Enabling grasping of unknown objects through a synergistic use of edge and surface information. *Int. J. Rob. Res.*, 31(10):1190–1213.

[Miller et al., 2003] Miller, A., Knoop, S., Christensen, H., and Allen, P. (2003). Automatic grasp planning using shape primitives. In *Robotics and Automation, 2003. Proceedings. ICRA '03. IEEE International Conference on*, volume 2, pages 1824–1829 vol.2.

[Pugeault et al., 2010] Pugeault, N., Wörgötter, F., and Krüger, N. (2010). Visual primitives: Local, condensed, semantically rich visual descriptors and their applications in robotics. *International Journal of Humanoid Robotics*, 07(03):379–405.