Grasping Familiar Objects using Shape Context

Jeannette Bohg and Danica Kragic
Centre for Autonomous Systems
Computer Vision and Active Perception Laboratory, KTH, Stockholm, Sweden.
{bohg,danik}@csc.kth.se

***Abstract*— We present work on vision based robotic grasping. The proposed method relies on extracting and representing the global contour of an object in a monocular image. A suitable grasp is then generated using a learning framework where prototypical grasping points are learned from several examples and then used on novel objects. For representation purposes, we apply the concept of shape context and for learning we use a supervised learning approach in which the classifier is trained with labeled synthetic images. Our results show that a combination of a descriptor based on shape context with a non-linear classification algorithm leads to a stable detection of grasping points for a variety of objects. Furthermore, we will show how our representation supports the inference of a full grasp configuration.**

I. Introduction

Robotic grasping of unknown objects is an open problem in the robotic community. The modeling process depends on the object representation, the embodiment of the robot and the task itself making the amount of potential grasps huge. Although humans master this skill easily, no suitable representations of the whole process have yet been proposed in the neuroscientific literature, making it thus difficult to develop robotic systems that can mimic human grasping behavior. However, there is some valuable insight. [1] propose that the human visual system is characterized by a division into the dorsal and ventral pathway. While the dorsal stream is mainly responsible for the spatial vision targeted towards extracting action relevant visual features, the ventral stream is engaged in the task of object identification. This dissociation also suggests two different grasp choice mechanisms dependent on whether a known or unknown object is to be picked up. Support for this thesis can be found in behavioral studies by [2]. The authors claim that in the case of novel objects, our actions are purely guided by affordances as introduced by [3]. In case of known objects, semantic information (e.g. through grasp experience) is needed to grasp them appropriately according to their function. However as argued in [1], [4] this division of labor is not absolute. In case of objects that are similar to previously encountered ones, the ventral system helps the dorsal stream in the action selection process by providing information about prehensile parts along with their afforded actions. Considering the related work in the area of robotic grasping, we can make the general observation that there is a trade-off between the quality of an inferred grasp and the applicability of the method in a real world scenario. The more precise, accurate and detailed an object model is, the more suitable it is for grasp planning based on criteria such as e.g. stability.

In our approach, we formulate the basic requirements for an object representation. First, it has to be suitable to be extracted from sensory data such as stereo cameras. Second, it has to be rich enough to allow for the inference of the most important grasp parameters. In our case that is the *approach vector* [5] and the wrist orientation of the robotic hand. We see precise shape, texture and weight to be handled by a subsequent fine controller based on tactile-feedback and corrective movements as presented in our previous work, [6]. Thus, we introduce a method that applies an object representation fulfilling these requirements. We detect a grasping point based on the global shape of an arbitrary object in a monocular image [7]. This results in relating the 2D form of an object to a single point in left and right images. The 3D coordinates can then be inferred from stereo geometry. The advantage of using global object shape over local appearance lies in the fact that it can be used to define the appropriate approach vector for a 3D grasping point. We further apply a supervised learning algorithm, thus providing a methodology for grasping objects of *similar* shape. The contributions of our approach are:

i) We apply the concept of shape context to the task of robotic grasping which to the best of our knowledge has not yet been applied for that purpose. The approach is different from the one taken in [8] where only local appearance is used instead of global shape.
ii) We are inferring full grasp configurations for arbitrarily shaped objects from a stereo image pair. These are the main difference to the work presented in [9], [10] where either only planar objects are considered or three views from an object have to be obtained by moving the camera.
iii) We analyze how stable our algorithm is for a general tabletop scenario in the presence of background clutter without having trained with examples of that specific scenario as e.g. done in [8].

The remainder of this paper is organized as follows: In the next section, we present related work. In Sec. III, the method of applying shape context to grasping is introduced. In Sec. IV we evaluate our method. Sec. V concludes the paper and gives an outlook on future work.

II. Related Work

Vision based object grasping can be divided in grasping of:

- *Known Objects*: These approaches consider grasping of *a-priori* known objects. The goal is then to estimate object's pose and retrieve a suitable grasp, e.g. from an experience database.
- *Unknown Objects*: Approaches that fall into this category commonly represent the shape of an *unknown* object and apply rules or heuristics to reduce the number of potential grasps.
- *Familiar Objects*: These approaches try to re-use grasp experience that was gathered beforehand on specific objects to pick up objects that look similar to them. Objects can be *familiar* in different ways, e.g., in terms of shape, color or texture and are likely to be graspable in a similar way.

A. Grasping Known Objects

Some of the approaches approximate the object's shape with a number of primitives such as cylinders and boxes [11] or superquadrics (SQ) [12]. [13] reduce the number of candidate grasps by randomly generating a number of them dependent on the object surface. The method of [14] treats the problem of finding a suitable grasp as a shape matching problem between the hand and the object. All these approaches are developed and evaluated in simulation. However, [5] and [15] combine real and simulated data for the purpose of grasping *known* objects, i.e. their 3D model is available. In a monocular image a known object is recognized and its pose within the scene is estimated. Given that information, an appropriate grasp configuration can be selected from a grasp experience database. While [5] still apply the selected grasp in simulation, [15] ported this approach to the robotic platform described in [16]. [17] consider known deformable objects. The objects are detected in monocular images and visible object parts serve as a basis for planning a stable grasp under consideration of the global object shape. However, all these approaches are dependent on an a-priori known dense or detailed object model either in 2D or in 3D.

B. Grasping Unknown Objects

In reality, it is very difficult to infer full and accurately object geometry. There are various ways to deal with this sparse, incomplete and noisy data. [18], [19] approximate an object with shape primitives that provide cues for potential grasps. [18] decompose a point cloud derived from a stereo camera into a constellation of boxes. The simple geometry of a box reduces the number of potential grasps significantly. [19] approximate the rough object shape with a quadric whose minor axis is used to infer the wrist orientation, the object centroid serves as the approach target and the rough object size helps to determine the hand pre-shape. The quadric is estimated from multi-view measurements of the rough object shape in monocular images.

C. Grasping Familiar Objects

A promising direction in the area of grasp planning is to re-use experience to grasp *familiar* objects. Many of the objects surrounding us can be grouped together into categories of common characteristics. There are different possibilities what these commonalities can be. In the computer vision community, objects within one category usually share characteristic visual properties. In robotics however, and specifically in the area of manipulation, the goal is to enable an embodied, cognitive agent to interact with these objects. In this case, objects in one category should share common affordances i.e. they should be graspable in a similar way. The difficulty then is to find a representation that can encode this common affordance and is grounded in the embodiment and cognitive capabilities of the agent.

Our approach, and also the methods that are going to be mentioned in the following, try to learn from experience how different objects can be grasped given various representations. This is different from the above mentioned systems in which unknown objects are grasped and the difficulty lies in finding appropriate rules and heuristics. In the following, we will present related work that tackle the grasping of familiar objects.

First of all, there are approaches that rely on 3D data only. [20] segment a given point cloud into parts and approximate each part by a superquadric. Their parameters are then fed into an artificial neural net (ANN) in order to classify it as prehensile or not. [21] directly use a single SQ to find a suitable grasp configuration for a Barrett hand. The experience for doing that is provided by an SVM. [22] build upon a database of 3D object annotated with the best grasps that can be applied to them. To infer a good grasp for a new object, very basic shape features are extracted to classify it as similar to an object in the database. However, all these methods are presented in simulation only. Similar to the method presented in this paper, there are experience based approaches that avoid the difficulty of 3D reconstruction by relying mainly on 2D data. [8] proposed a system that infers a point at which to grasp an object directly as a function of its image. However, if more complex goals are considered that require subsequent actions, e.g. pouring something from one container into another, methods that have a notion about 3D and about what an object constitutes are necessary. In this paper, we propose a method like that.

III. Methodology

Our approach is based on the hypothesis that visual attributes of objects afford specific actions. The action considered here is a stable grasp of an object. The visual attribute is the object's shape context calculated based on its contour in a monocular image. We assume that we can apply grasping experience gathered from a set of known objects to grasp yet unknown objects that have similar shaped prehensile parts. To that end, we use a supervised learning technique that provides the robot with that sort of experience from the database of synthetic images.

In our approach, we use a stereo image pair to perform scene segmentation resulting in hypotheses of several objects. Shape context is then computed on each of the hypotheses. Further, 2D points are determined at which each of the hypotheses can be grasped. The model for this is

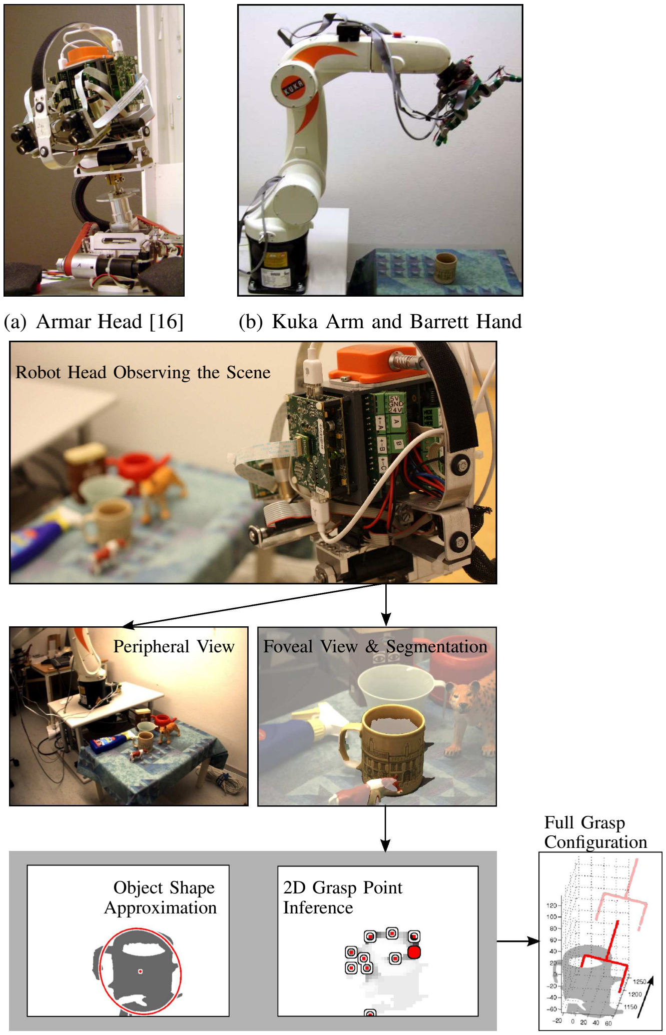

(a) Armar Head [16]
(b) Kuka Arm and Barrett Hand
Robot Head Observing the Scene
Peripheral View
Foveal View & Segmentation
Object Shape Approximation
2D Grasp Point Inference
Full Grasp Configuration

Fig. 1. Components of the Stereo Vision based Grasp Inference System. Peripheral and foveal views, object shape approximation, grasp point inference and full grasp configuration.

computed beforehand through offline training on an image database reported in [8]. The points in the left and in the right image are associated to each other to infer a 3D grasping point via triangulation. In parallel to the grasping point detection, the segments are analyzed in terms of rough object pose. By integrating the 3D grasping point with this pose, a full grasp configuration can be determined and executed. In the following sections, the individual steps of the system are explained in more detail. A detailed flow chart of the whole system is given in Fig. 1 along with the used hardware.

A. Scene Segmentation

The system starts by performing the *figure-ground segmentation*. In our case, we have no knowledge about *what* the object actually is so we approach this problem through reasoning on what constitutes an object in a scene.

The advantage of using an active stereo head (see Fig. 1(a)) lies in its capability to fixate on certain objects of interest. Once the system is in fixation, zero-disparities can be employed as a cue for figure-ground segmentation through different segmentation techniques such as e.g. watersheding [23]. The assumption made is that continuity in depth indicates a coherent object. However, in Fig. 2 it can be observed that the ground on which the object in fixation stands is usually also classified as foreground.

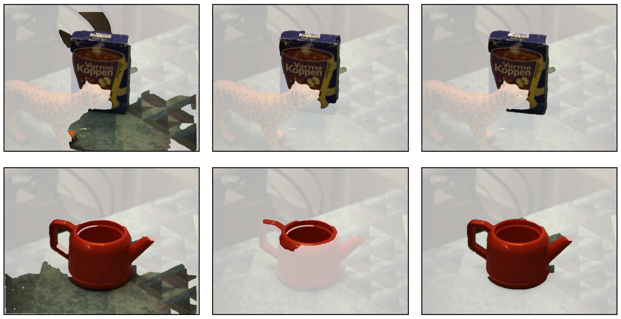

(a) (b) (c)

Fig. 2. Segmentation results for a) segmentation based on zero disparities, b) using additionally a table plane assumption and c) using additionally the hue.

The environment in which we expect service robots to perform are dominated by surfaces that are parallel to the ground. In order to overcome the previously mentioned segmentation problem, we can include the assumption that a dominant plane is present in the scene. In our examples, this plane represents the table plane objects are placed on. For that purpose, we fit a planar surface to the disparity image. The probability for each pixel in the disparity image to belong to that plane or not depends on its distance to it. plane.

An additional assumption that can be introduced into the system is that objects are usually either uniformly colored or textured. By introducing this cue in conjunction with the table plane assumption, we can stabilize the figure-ground segmentation. In this way, we can overcome disparity holes and instabilities due to an uncertain table plane detection, as shown in Fig. 2.

B. Representing Relative Shape

In this section, we propose a representation that fulfills the mentioned requirements for an object representation: it is rich enough to infer necessary grasp parameters and can be extracted from real world sensors. We assume that the object is segmented in the image. Intuitively seen, how to grasp an object depends to a large extent on its global shape. However, we need a local descriptor that relates this global property to each single point on the object. Our idea on how to exploit this object attribute is to apply the concept of shape context that was up till now mainly used for shape matching and object recognition, [7].

The basis for the computation of shape context is an edge image of the object. N samples are taken with a uniform distribution from the contour. For each point we consider the vectors that lead to all the other sample points. These vectors relate the global shape of the object to the considered reference point. For each point, we create a log-polar histogram with angle and radius bins to comprise

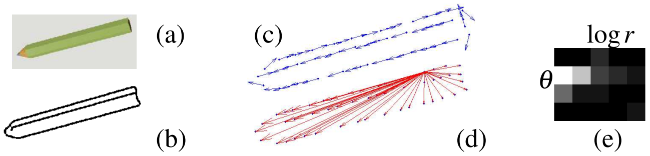


Fig. 3. Example for deriving the shape context descriptor for the image of a pencil. (a) Input image of the pencil. (b) Contour of the pencil derived with the Canny operator. (c) Sampled points of the contour with gradients. (d) All vectors from one point to all other sample points. (e) Histogram with four angle and five log-radius bins comprising the vectors depicted in (d).

this information into a compact descriptor. The logarithmic division ensures that the influence of nearby samples is emphasized. An example for this whole process is shown in Fig. 3.

The shape context is computed on a result of running a Canny edge detector on the segmented image. We do not consider single contour points but instead subdivide the image into rectangular patches (in our case 10×10 pixels). A descriptor for each patch serves as the basis to decide whether it is a grasping point or not. This descriptor is simply composed of the accumulated histograms of all sample points on the object's contour that lie in that patch. Typically only few sample points will be in a 10×10 pixel wide window. This turned out not to be sufficient for the classification task. We therefore calculated the accumulated histograms in three different spacial scales centered at the current patch and concatenated them to form the final feature descriptor of dimension 120.

C. Classification of 2D Grasping Points

The goal of the presented approach is to identify a point of an object at which it can be grasped. An image of that object serves as input data. In our case, we consider input patches and classify them based on the object's shape as either being grasp candidates or not. This decision is made based on experience obtained during the training of the classifier. We use a supervised approach, where a non-linear SVM classifier with the Radial Basis Function kernel is employed. Thus, for training, we follow the methodology of [8] but utilize a non-linear classifier.

D. Approximating Pose and Generating Grasps

Objects are to be manipulated with the three-fingered Barrett hand in a pinch grasp configuration. Our goal is to approach a 3D grasping point with the palm of the hand in a specific wrist orientation. Given a 2D grasping point in the left image of the stereo camera, we can determine its 3D position if we also know its position in the right image. In order to infer the orientation of the Barrett hand we have to at least roughly estimate the pose of the unknown object. The question is how to derive it from stereo images without relying on 3D reconstruction.

Here, we approximate arbitrarily shaped objects by fitting an ellipse to the segmented object in the image plane. The major and minor axis of this ellipse in 2D serve as the basis to obtain a rough estimate of the 3D object pose. For this purpose, we detect three points in the left image: the centroid of the segment, an object point on the major axis and an object point on the minor axis. Via stereo matching we can find the corresponding points in the right image and thus obtain three 3D points that define a tilted plane. The objects pose is then associated with the three dimensional position of its segment centroid and the orientation of the plane. The assumption we make is that a single plane can in general roughly approximate the orientation of an object.

After we run the classifier on each image of the stereo image pair, we associate the resulting 2D grasping hypotheses to each other in order to obtain a 3D point via triangulation. For this purpose we create a set $B_l = \{b_{(i,l)} | i = 1 \cdots m\}$ of m image patches i in the left image that are local maxima regarding the classifier response $P(g_{(i,l)} = 1 | D_{(i,l)})$ and whose adjacent patches in the 8-neighborhood carry values close to that of the center patch. We apply stereo matching to obtain the corresponding patches $B_r = \{b_{(i,r)} | i = 1 \cdots m\}$ in the right image. Let $P(b_{(i,l)} | D_{(i,l)})$ and $P(b_{(i,r)} | D_{(i,r)})$ be the probability for each image patch in set B_l and B_r to be a grasping point given the respective feature descriptors $D_{(i,l)}$ or $D_{(i,r)}$. Assuming Bayesian independence between corresponding patches in the left and right image, the probability $P(b_i | D_{(i,l)}, D_{(i,r)})$ for a 3D point b_i to be a grasping point is determined by

$$P(b_{(i,l)} | D_{(i,l)}, D_{(i,r)}) = P(b_{(i,l)} | D_{(i,l)}) * P(b_{(i,r)} | D_{(i,r)}). \quad (1)$$

According to this measure, we can rank the 3D grasping points. The best patch is then

$$b = \arg\max_i P(b_{(i,l)} | D_{(i,l)}, D_{(i,r)}). \quad (2)$$

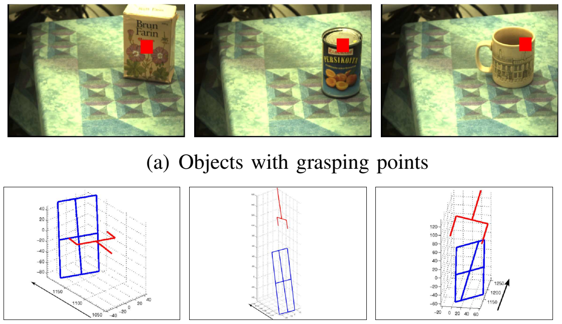
Orientation of the Hand: Given a 3D grasping point and an object pose, we define three possibilities to choose the approach vector:

i) vector a_{ma} defined by the major axis of the ellipse in 3D,
ii) vector a_{mi} defined by the minor axis in 3D or
iii) normal vector n_p of the plane p spanned by these two vectors.

Which of them is chosen depends on the position of the 2D grasping point within the 2D ellipse. Let $x_{b_{(i,l)}}$ be the vector defined from the grasping point $b_{(i,l)}$ to the center of mass c_l of the segment in the left image. Let x_{mi} be the vector from c_l to the point on the minor axis of the 2D ellipse lying on the segment boundary. Let x_{ma} be the vector defined equivalently for the major axis. Let ϕ be a given threshold for the distance between $b_{(i,l)}$ and c_l. If $|x_{b_{(i,l)}}| < \phi$ then the hand will approach the grasping point b_i with a vector n_p. The wrist orientation will be determined by aligning the vector between the thumb and two fingers with a_{mi}. If

$$\frac{|x_{ma} \cdot x_{b_{(i,l)}}|}{|x_{ma}|} > \frac{|x_{mi} \cdot x_{b_{(i,l)}}|}{|x_{mi}|}, \quad (3)$$

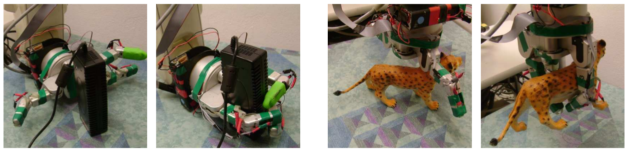
i.e. $x_{b_{(i,l)}}$ is better aligned with x_{ma} than with x_{mi}, a_{ma} will be chosen as approach direction towards b_i. The wrist orientation will be fixed by aligning the vector between the



(a) Objects with grasping points

(b) Grasp Configurations

Fig. 4. Examples for generated grasp configurations. (a) Right image of the stereo camera with grasp point labeled. (b) Related grasp configuration with a schematic gripper and the plane with the axes approximating the object pose. the viewing direction is indicated by the arrow.



(a) Side Grasp

(b) Top Grasp

Fig. 5. An example for the execution of a side grasp and a top grasp on our robotic platform.

thumb and two fingers with n_p. In case $x_{b_{(i,l)}}$ is better aligned with x_{mi} than with x_{ma}, a_{mi} will be chosen as approach vector. The wrist orientation will be fixed in the same way as for the previous case. Examples of grasp configurations are given in Fig. 4. Fig. 5 shows an example for the top and for the side grasp.

IV. Experimental Results

Our method relies on scene segmentation. The quality of the segmentation is affected by the cues that are integrated and on how the considered environment complies to the assumptions, e.g, dominant plane, uniformity in color or texture. In this section, we evaluate how the relative shape based representation is affected by different segmentation qualities. For that purpose, we collected images of differently textured and texture-less objects, e.g. boxes, cans, cups, elongated objects, or toys, composed in scenes of different levels of complexity. This ranges from single objects on a table to several objects occluding each other. We present some representative examples of the grasping point inference methods when applied to different sets of objects.

Ideally, we would like to achieve two things. First, the grasping points that are inferred in two images of the same object given different qualities of segmentation have to correspond to each other. This is important because later on we would like to match grasping points in a stereo image pair to obtain a 3D point. Second, the quality of the inferred grasping points should only be minimally affected by the background complexity.

In Fig. 6(a), we show the results of the grasping point classification for a red texture-less teapot. The applied model is trained on mugs and pencils. The left column shows the segmented input of which the first one is always the ground truth segment. The right shows the grasping points generated by our descriptor. For comparison, the middle column shows the results performed with the local appearance based method suggested in [8]. The red dots label the detected grasping points. They are the local maxima in the resulting probability distribution, where up to ten highest valued local maxima are selected. Grasping point classification is shown for a case when the teapot is the only object in the scene and when it is partially occluded. The detection of grasping points is quite stable when facing decreasing quality of segmentation and occlusion. In Fig. 6(b) (last row), even though there is a second handle now in the segmented region, the rim of the teapot is still detected as graspable and the general resulting grasping point distribution looks similar to the cases in which the handle was not yet in the segment. This means, that the object that is currently in fixation by the vision system, the one that dominates the scene, produces the strongest responses of the grasping point model even in the presence of other graspable objects.

In Fig. 6(c) and 6(d), we applied the models trained on mugs and cups to images of a can and a cup. We can observe that our method is less affected by texture due to the incorporation of the global shape. Finally in Fig. 6(e) and 6(f), we applied models to objects that are not similar to any object that the grasping point models were trained on. We observe several peaks in the distribution, suggesting the ability of the model to generalize over different shapes.

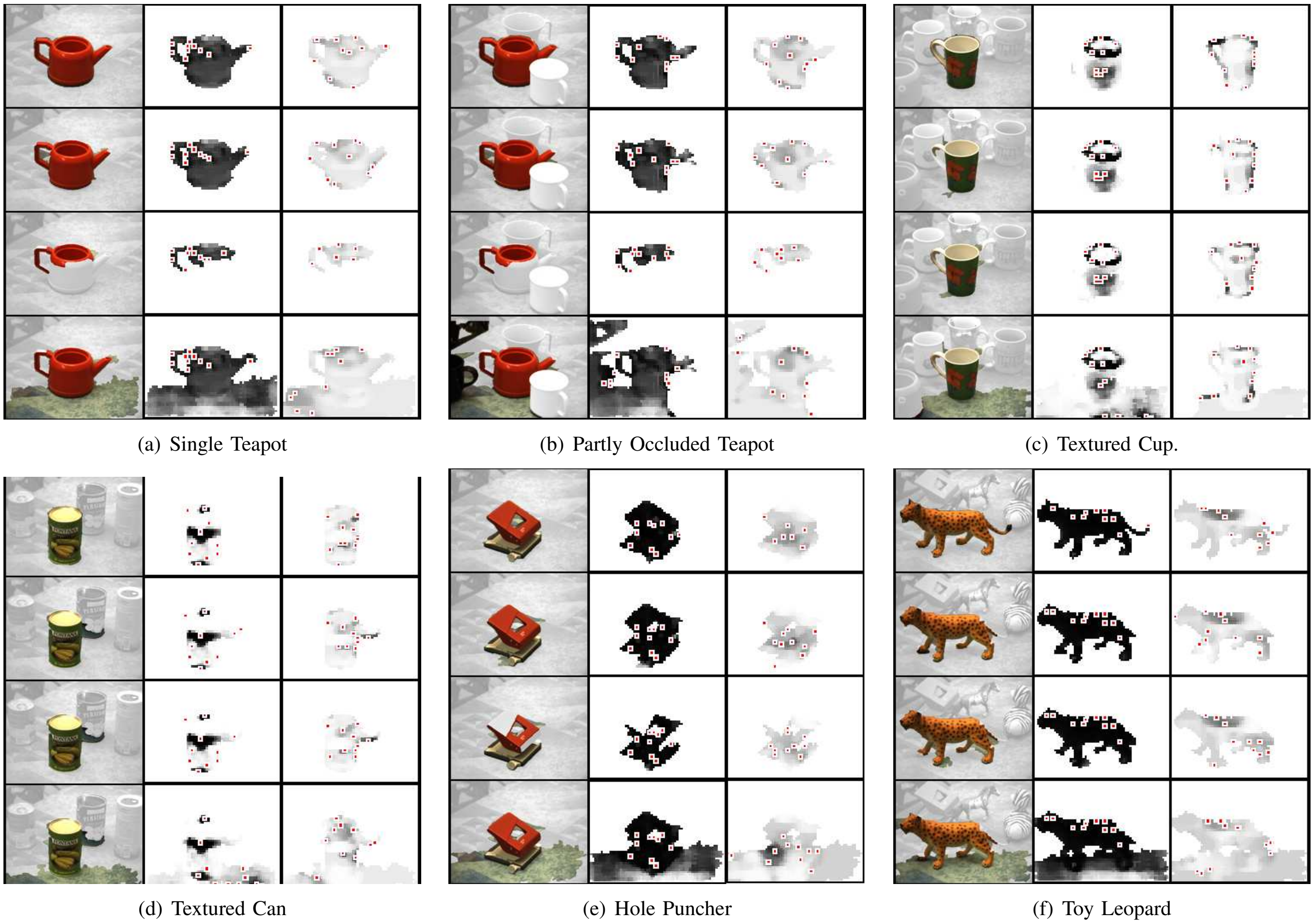
V. Conclusions

Grasping of unknown objects in natural environments is an important and unsolved problem in the robotic community. In this paper, we have developed a method for detecting a grasping point on an object by analyzing it in a monocular image and reconstructing the suitable 3D grasping representation based on a stereo view. We argued that for the purpose of grasping a yet unseen object, its global shape has to be taken into account. Therefore, we applied shape context as a visual feature descriptor that relates the object's global shape to a single point. According to the proposed terminology, our approach falls into the type of systems that work on grasping familiar objects where the familiarity in our case is defined through similarity in shape.

Evaluation in the real scene has proven the applicability of our approach in the presence of clutter and provides further insight into the difficulty of the object grasping process. We see several aspects to be evaluated in the future work. We will continue to further develop the method but integrate it more on the stereo level for generating the grasping point hypotheses. In addition, we will consider other type of representations that take into account several aspects of 2D-3D information.

Acknowledgment

The authors would like to thank Mårten Björkman for providing the segmentation system and Xavi Gratal Martínez



(a) Single Teapot

(b) Partly Occluded Teapot

(c) Textured Cup.

(d) Textured Can

(e) Hole Puncher

(f) Toy Leopard

Fig. 6. Grasping point model trained on labeled synthetic objects applied to segmented objects in real scenes. The darker a pixel, the higher the probability of being a grasping point.

for his great help with the hardware system. This project has been supported by the EU IST-FP7-IP GRASP (2008-2012).

References

[1] M. Goodale, "Separate Visual Pathways for Perception and Action," *Trends in Neurosciences*, 1992.

[2] A. Borghi, *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thinking*. Cambridge University Press, 2005, ch. Object Concepts and Action.

[3] J. Gibson, *The Ecological Approach to Visual Perception*. Lawrence Erlbaum Associates, 1979.

[4] M. J. Webster, J. Bachevalier, and L. G. Ungerleider, "Connections of Inferior Temporal Areas TEO and TE with Parietal and Frontal Cortex in Macaque Monkeys," *Cerebral cortex*, 1994.

[5] S. Ekvall and D. Kragic, "Learning and Evaluation of the Approach Vector for Automatic Grasp Generation and Planning," in *ICRA*, 2007.

[6] J. Tegin, S. Ekvall, D. Kragic, B. Iliev, and J. Wikander, "Demonstration based Learning and Control for Automatic Grasping," *Jrnl. of Intelligent Service Robotics*, 2008, to appear.

[7] S. Belongie, J. Malik, and J. Puzicha, "Shape Matching and Object Recognition Using Shape Contexts," *PAMI*, 2002.

[8] A. Saxena, J. Driemeyer, J. Kearns, and A. Y. Ng, "Robotic Grasping of Novel Objects," *NIPS*, vol. 19, 2007.

[9] A. Morales, E. Chinellato, A. Fagg, and A. del Pobil, "Using Experience for Assessing Grasp Reliability," *IJHR*, 2004.

[10] J. Speth, A. Morales, and P. J. Sanz, "Vision-Based Grasp Planning of 3D Objects by Extending 2D Contour Based Algorithms," in *IROS*, 2008.

[11] A. T. Miller, S. Knoop, H. I. Christensen, and P. K. Allen, "Automatic Grasp Planning Using Shape Primitives," in *ICRA*, 2003.

[12] C. Goldfeder, P. K. Allen, C. Lackner, and R. Pelossof, "Grasp Planning Via Decomposition Trees," in *ICRA*, 2007.

[13] C. Borst, M. Fischer, and G. Hirzinger, "Grasping the Dice by Dicing the Grasp," in *IROS*, 2003.

[14] Y. Li and N. Pollard, "A Shape Matching Algorithm for Synthesizing Humanlike Enveloping Grasps," *ICHR*, 2005.

[15] A. Morales, P. Azad, T. Asfour, D. Kraft, S. Knoop, R. Dillmann, A. Kargov, C. Pylatiuk, and S. Schulz, "An Anthropomorphic Grasping Approach for an Assistant Humanoid Robot," in *International Symposium on Robotics*, 2006.

[16] T. Asfour, K. Regenstein, P. Azad, J. Schröder, A. Bierbaum, N. Vahrenkamp, and R. Dillmann, "ARMAR-III: An Integrated Humanoid Platform for Sensory-Motor Control," in *ICHR*, 2006.

[17] J. Glover, D. Rus, and N. Roy, "Probabilistic Models of Object Geometry for Grasp Planning," in *ICRA*, 2008.

[18] K. Hübner and D. Kragic, "Selection of Robot Pre-Grasps using Box-Based Shape Approximation," in *IROS*, 2008.

[19] C. Dunes, E. Marchand, C. Collowet, and C. Leroux, "Active Rough Shape Estimation of Unknown Objects," in *ICRA*, 2008.

[20] S. El-Khoury and A. Sahbani, "Handling Objects By Their Handles," in *IROS Workshop on Grasp and Task Learning by Imitation*, 2008.

[21] R. Pelossof, A. Miller, P. Allen, and T. Jebera, "An svm learning approach to robotic grasping," in *ICRA*, 2004.

[22] N. Curtis and J. Xiao, "Efficient and Effective Grasping of Novel Objects through Learning and Adapting a Knowledge Base," in *ICRA*, 2008.

[23] M. Björkman and J. Eklundh, "Foveated figure-ground segmentation and its role in recognition," in *BMVC*, 2005.