Object Decomposition based on Graspability

Anonymous WACV submission

Paper ID 478

Abstract

Objects are composed of a configuration of parts which are mostly designed for a certain functionality. For example a spatula is designed for scooping through a grasp from the handle. Therefore, recognizing and representing functionalities into object representation is of a great importance. This representation must provide us with an efficient inference of functionalities. Therefore, the representation must be discriminative enough. We introduce here a novel compositional model based on grasping functinality. Objects are represented based on graspablity of their regions. Moreover, the grasps and their robustness are encoded into the representation. This allows us for efficient and robust search for graspability on objects. We evaluated our method in a robotic grasping scenario and achieved a promising accuracy for novel objects.

1. Introduction

Learning the visual representation of objects for the discrimination among different object categories as well as its generalization to new object instances is of a great importance. Compositional models for object representation has been successful at this task [3, 12, 9, 2]. Unfortunately, they make use of arbitrary parts, whereas objects are composed of functional parts. This work will be about incorporating functionalities such as grasping into the object representation.

We are interested in a discriminative and robust representation of object based on the intended functionalities. As an example, for grasping functionality, we would like to recognize graspable versus non-graspable regions, the area which our grasp is more robust because the it does not change much. We are also interested in segmenting an object into regions which have the same grasping behavior in terms of robustness and scale.

The contributions of the present work is a novel object segmentation method that is based on grasping affordances. Our segmentation provides us with graspable and non-graspable regions as well as the robustness of the grasp in each region. All the grasping information is integrated into our representation without a need for an additional inference method. We show in our experimental results that our decomposition provides us with graspable regions which are used for robotic grasping experiment.

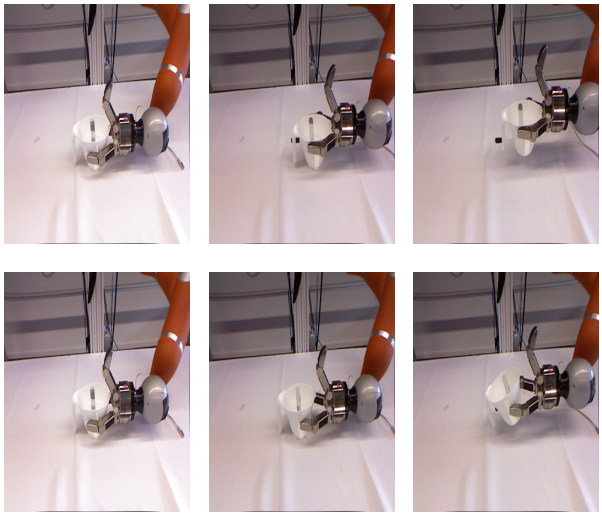

Figure 1. Objects can be represented by the regions which are graspable and the robustness of the grasp. The container is grasped from the middle of the body. If the robot gripper moves up or down, the grasp will not be changed. But if the robot gripper moves outward, the grasp might be lost. The information on robustness of a grasp in certain directions and the way that the robot grasps an object are very important in robotic grasping and must be encoded into the object representation.

2. Related Work

Detecting affordances in robotics has been highly motivated by the visual characteristics of objects. Recent work [7, 6, 1, 4, 5], associated grasping to small patches that the robot touches during grasping. These patches are represented based on their visual features such as geometrical properties of the object surface, surface normals and curvatures, or deep features. The learned features are then used to recognize graspable versus non-graspable regions.

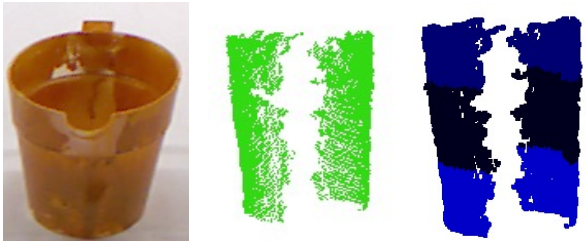

Figure 2. From left to right: pitcher object in RGB-D and its segmented regions. Regions are colored based on grasp robustness. The darker the color, the more robust the region for grasping.

Often, these approaches provide fairly good detection results but their search space is quite huge.

Part-based methods can overcome this problem. Not only do they reduce the search space during recognition, but they also provide a framework for the generalization of affordances such as grasping among different object categories. In the work discussed in [11], object parts are segmented based on geometrical properties such as local convexity, and grasping is associated with the object parts and their categories. Detecting affordances that go beyond grasping based on object parts is explored in [8]. In this work, object parts are labeled based on their affordances during training. Conventional deep features are extracted from pairs of supervoxels in each object part. Conditional Random Fields (*CRF*) are then used to predict affordances for objects. Even though, these approaches provide promising results, still inference happens either in a very large object level or very small supervoxel pairs. Moreover, they do not consider the relation between different affordances which can consequently affect visual segmentation of objects and decrease inference time. We present here a method for decomposition of objects based on part functionalities. Our decomposition starts at a reasonable level of granularity that is neither supervoxels nor object level. We then propose a method for representing as well as inferring functionalities based on visual information

3. Object Representation from Grasping

The structure of an object has much to do with its functionalities. We are interested in a representation based on functional parts. The main functionality which we have considered in this work is grasping. Our object representation relies on differentiating between graspable and non-graspable regions, including for the former the robustness of grasping in those regions. As an example, the pitcher which is shown in Figure 2 can be grasped more robust from the middle than the bottom or the top. Hence, the pitcher is segmented into three regions as is shown with different colors. In addition to this, given a region we will include information on how to grasp it. This knowledge is encoded into

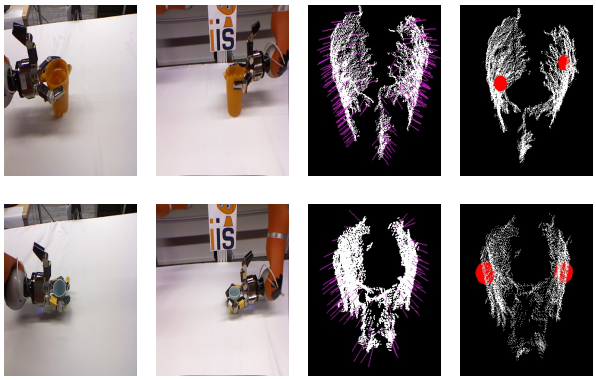

Figure 3. Kinesthetic teaching for learning grasps. Robotic arm is manually guided to the grasp area and the object is captured by two calibrated kinects.

the region representation. Next, we explain these two main criteria in our grasp-based object representation, namely 1) including grasp information and 2) encoding grasp robustness.

For the first criterion, we performed a set of robotic grasping experiments with two finger grasps as shown in Figure 3. Since in one view both fingers of the gripper are not visible, we used two views with calibrated kinects. Next, we consider the grasped region of the object based on the contact points. We can constraint our representation by noting that for our objects, 1) the robot gripper does not penetrate into the object, 2) the contact normals are collinear and 3) the grasped regions are convex. The minimal ellipse that does not penetrate into the region will then determine a grasp. One of the principle axes of the ellipse is determined by the contact points. The other axis is specified by the plane normal to the gripper pose.

Our grasp representation based on the ellipses contains enough information on how to grasp the region, but the robustness of the grasp is still not encoded. In order to do this, we performed experiments in which the robot tries to grasp the object by moving from the contact points. We then fitted an ellipse to the moved grasp and computed the difference of the area between the ellipse and the original fitted one. The maximal value for area difference between ellipses while achieving successful grasps will be our threshold for sensitivity t.

We model the movement by motion along the principal axes of the ellipse and the third axis which is perpendicular to the them. Grasp robustness is represented in six dimensions, for moving outwards and inwards along each axis. We moved at certain steps along each direction. In each step, we computed the difference of areas. We then represented the probability of grasping along a certain direction. If we denote an ellipse with e, each direction with m_i and each step with s_j, the probability of grasping $p(g|m_i, e)$ is

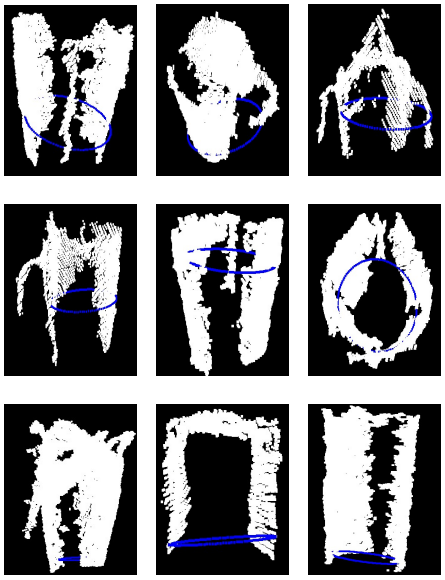

Figure 4. Clusters obtained after spectral clustering. Each row shows one cluster and each column one instance in the cluster.

computed as follows,

$$p(g|m_i,e) = \sum_{s_j} p(g,s_j|m_i,e) \quad (1)$$

$$= \sum_{s_j} \frac{p(g,s_j,m_i,e)}{p(m_i,e)} \quad (2)$$

$$= \sum_{s_j} \frac{p(g|s_j,m_i,e)p(s_j|m_i,e)p(m_i,e)}{p(m_i,e)} \quad (3)$$

$$= \sum_{s_j} p(g|s_j,m_i,e)p(s_j|m_i,e) \quad (4)$$

where $p(s_j|m_i,e) = \frac{1}{S}$ is distributed uniformly as the number of steps. $p(g|s_j,m_i,e)$ indicates the grasping probability for each step along each axes. In order to compute this probability, we make use of the robustness threshold t which was obtained through experiments,

$$p(g|s_j,m_i,e) = \begin{cases} 1, & \text{if } \|e_{s_j,m_i} - e\| < t \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

The next step is to group the ellipses with the same grasping ellipses and robustness into the same region. Equivalent classes of ellipses will then be learned, through clustering from the training data. For this purpose, we used spectral clustering algorithm duo to it performs overall balancing between clusters and prevents singleton clusters. The results of this clustering algorithm on our data is shown in Figure 4.

4. Inferring Graspable Regions in Novel Objects

The first step in decomposing a novel object into graspable regions is obtaining the candidate contact points. The grasped contact normals have two properties (Section 3), 1) the contact normals are collinear and 2) the contact normals point in opposite directions and outward. First, the collinearity property is determined by the inner product of the normalized connecting axis between contact points l and the contact normals n,

$$\text{colli}(l,n) = \begin{cases} l \cdot n, & \text{if } l \cdot n < t_{\text{colli}} \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

where t_{colli} is the threshold for collinearity which is determined during training. The second constraint is specified by the angle between contact normals $\langle n_1, n_2 \rangle$. The angle must be greater than an angle threshold t_θ learned during training.

We included these two constraints into a score function. The score between two contact points c_1, c_2 is computed as,

$$\underset{c_1,c_2}{\text{score}} = \alpha \frac{\langle n_1, n_2 \rangle}{\pi} + (1-\alpha) \frac{-\text{colli}(l,n_1) - \text{colli}(l,n_2)}{2} \quad (7)$$

We set $\alpha = 0.5$ in our experiments. For a novel object, we first compute its supervoxels and consider only their mean normal vectors. We then consider each pair of contact points based on the computed supervoxels and examine the aforementioned score value. We then keep the contact points whose scores are higher than their neighbors.

Given candidate the contact points, the next step is to fit an ellipse to them. Since, the input data are 3D pointclouds, there are many ellipses which can be fitted. For selecting the best fitted ellipse for our goal, we considered two criteria, namely 1) area of the ellipse and 2) elevation of the plane for the fitted ellipse. For the second criterion, we enforce that they are graspable ellipses. Since at certain elevations the robot cannot grasp the object. During the grasping experiments, we collected statistics on the relation between the elevation of the plane which passes through the contact normals and the gripper pose. We then use this information for assigning higher probability to certain elevations.

We select the ellipse e^* which has the maximum fitting value as given in Eqn 8, where ϕ indicates the elevation of the fitted plane. $p(\phi|g)$ is the probability of an elevation for grasping, which is obtained from our grasping experiments. r_e is the absolute value of the estimated principle axis of the ellipse and max_r is the maximum value of the principle axis of the ellipse from all the fitted ellipses, likewise min_r.

$$e^* = argmax_e \, \text{fit}(e) \tag{8}$$

$$\text{fit}(e) = \alpha p(\phi|g) + (1-\alpha)(1 - \frac{r_e - min_r}{max_r - min_r}) \tag{9}$$

After fitting ellipses, we compute the robustness of a grasp by moving along the principle axes of the ellipse (Section 3). Finally, we obtain a feature vector based on the ellipse parameters and the motion along the axes. Since, we are interested in graspable regions in an object, non-graspable ones are filtered out. For this purpose, a classifier based on an SVM with RBF kernel is trained to discriminate between graspable versus non-graspable ellipses represented by the aforementioned feature vector. We then assign each graspable ellipse to the closest cluster which determines the type of the ellipse. The graspable ellipses might have overlapping areas, which we need to remove to decrease the uncertainty in that particular region. In order to do this, we compute a matrix which indicates whether each two ellipse overlap or not. From the set of overlapping ellipses, we select the one which has the maximum confidence. The confidence is defined as the confidence on the class of the ellipse C_e and is computed based on the mean of a cluster μ_{C_e} as follows,

$$e^* = \underset{e}{\text{argmax}} \, p(C_e|e) \tag{10}$$

$$p(C_e|e) = \frac{\exp(-\|e - \mu_{C_e}\|)}{\sum_C \exp(-\|e - \mu_C\|)} \tag{11}$$

Next, we merge adjacent ellipses which belong to the same cluster into a region. This gives us the decomposition of an object based on grasping. Each region has a probability for grasping which is computed based on the robustness of its class for grasping $p(g|c)$. We will consider only the mean of a cluster μ_c for computing this probability. In order to compute this probability, we marginalize over the ellipse movement directions m_i,

$$p(g|c) = \sum_{m_i} p(g, m_i|c) \tag{12}$$

$$= \sum_{m_i} \frac{p(g, m_i, c)}{p(c)} \tag{13}$$

$$= \sum_{m_i} \frac{p(g|m_i, c)p(m_i, c)}{p(c)} \tag{14}$$

$$= \sum_{m_i} p(g|m_i, c)p(m_i|c) \tag{15}$$

$$\tag{16}$$

$p(g|m_i, c)$ determines the probability of grasping in a certain movement direction m_i for cluster c (already encoded in the feature vector). $p(m_i|c)$ is uniformly distributed, $p(m_i|c) = \frac{1}{M}$, where M indicates the number of axes movements, which in our case is six.

When the regions and their probabilities are determined, object regions will be used for grasping. Given the region probabilities, we select the most probable region. A region is composed of a number of ellipses and the gripper position for grasping, which can be computed from the ellipses. The ellipses which are on the boundary of the regions are more sensitive for grasping than those which are on the middle. Therefore, we consider only the middle ellipses in each region for grasping.

After computing the best ellipse in the most probable region, the gripper pose is obtained. The translation of the ellipse is already given by the position of the principal axis of the ellipse which is determined after the ellipse fitting. An ellipse gives us rotation information in 2D and we need to consider another axis which is perpendicular to the principal axes in order to compute the 3D orientation. considering e_x as the connecting axis for contact points c_1, c_2, and e_z as the second axis n which is determined by the ellipse fitting procedure, the third one e_y is computed as the cross product between two principal axes as follows,

$$e_x = c_1 - c_2 \tag{17}$$

$$e_z = -n \tag{18}$$

$$e_y = e_x * e_z \tag{19}$$

This axis e_y can have two directions, we consider the one that goes with he gravity axis g_v. Considering an upright positioned object, this axis should be in the same direction as the gravity axis, otherwise it will be inverted,

$$\text{sign}(e_y) = \begin{cases} -1, & \text{if } e_y \cdot g_v < 0 \\ 1, & \text{otherwise} \end{cases} \tag{20}$$

$$e_y = \text{sign}(e_y)e_y \tag{21}$$

5. Experimental Results

We evaluated our approach on two dataset, IKEA kitchen object and YCB object dataset [10]. The grasping labels for training are available on (grasp database[1]). The experimental setup for grasping experiments consists of a robot with two KUKA 7-DoF Light-Weight Robot 4+ arms with servo-electric 3-Finger Schunk SDH-2 dexterous hands. There are two kinects for capturing RGB-D data which are located in opposite of each other.

For grasp learning purpose, we performed 92 grasps on 24 different objects from IKEA kitchen objects and YCB

[1] https://iis.uibk.ac.at/public/GraspAnnotateDataset/

Figure 5. Training objects for grasping experiment.

Figure 6. Test objects for grasping experiment.

dataset as shown in Figure 5. We also labeled the non-graspable contact points. Our grasp learning procedure for training grasps as well as grasp movement robustness is provided as a video in our supplementary materials.

We used the training data for learning ellipse classes for our compositional method. Some qualitative results of our decomposition can be seen in Table 8. We then used the decomposition to infer graspable regions on novel objects. As mentioned earlier, our method selects the regions with the highest grasp probabilty and the ellipse which is located in the middle of the region. We evaluated our grasping experiment on 19 objects which are not seen during training as shown in Figure 6. The qualitative results of our grasping experiments are given in Table 1 also as a video in our supplementary materials. The quantitative results of our grasping experiments is given in Table 5. As can be seen, our method succeeds even on very different objects, convex and non-convex and has a promising performance. Also, due to the compositional nature of our method, computing grasps after obtaining regions is very efficient.

Discussion The main source of error is the noise of input pointcloud due to the reflection or transparency. Which

Planning Success	Grasp Ellipses Success	Overall Performance
89%	78%	57%

Table 2. Quantitative Results for Grasping Experiments.

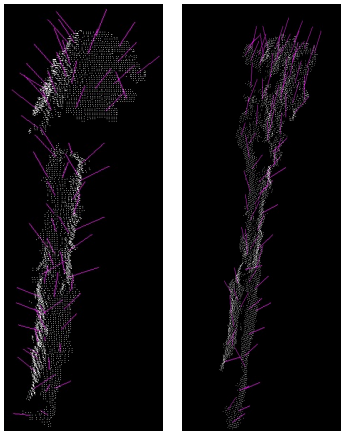

Figure 7. Grasp failures due to the low-resolution of input pointclouds and not a robust normal estimation.

then results in a not robust estimation of surface normals. As shown in Figure 7, due to this problem, the grasps for spatulas could not be computed. We considered to replace normals with principle curvatures which are more robust. Our algorithm suffers from the resolution problem of supervoxels. That is for small parts such as handles, we obtain only one supervoxel, hence no collinear pairs of contact can be found. Therefore, we fail in representing small parts which are also important for grasping. We consider to use 3D edges in addition to improve the results.

6. Conclusion

The contribution of our work is a novel object decomposition based on graspability. Our representation carries enough information for grasping as well as the robustness of a grasp in a certain region of an object. This decomposing allows for an efficient grasp inference on novel objects which is of a great importance on robotic applications.

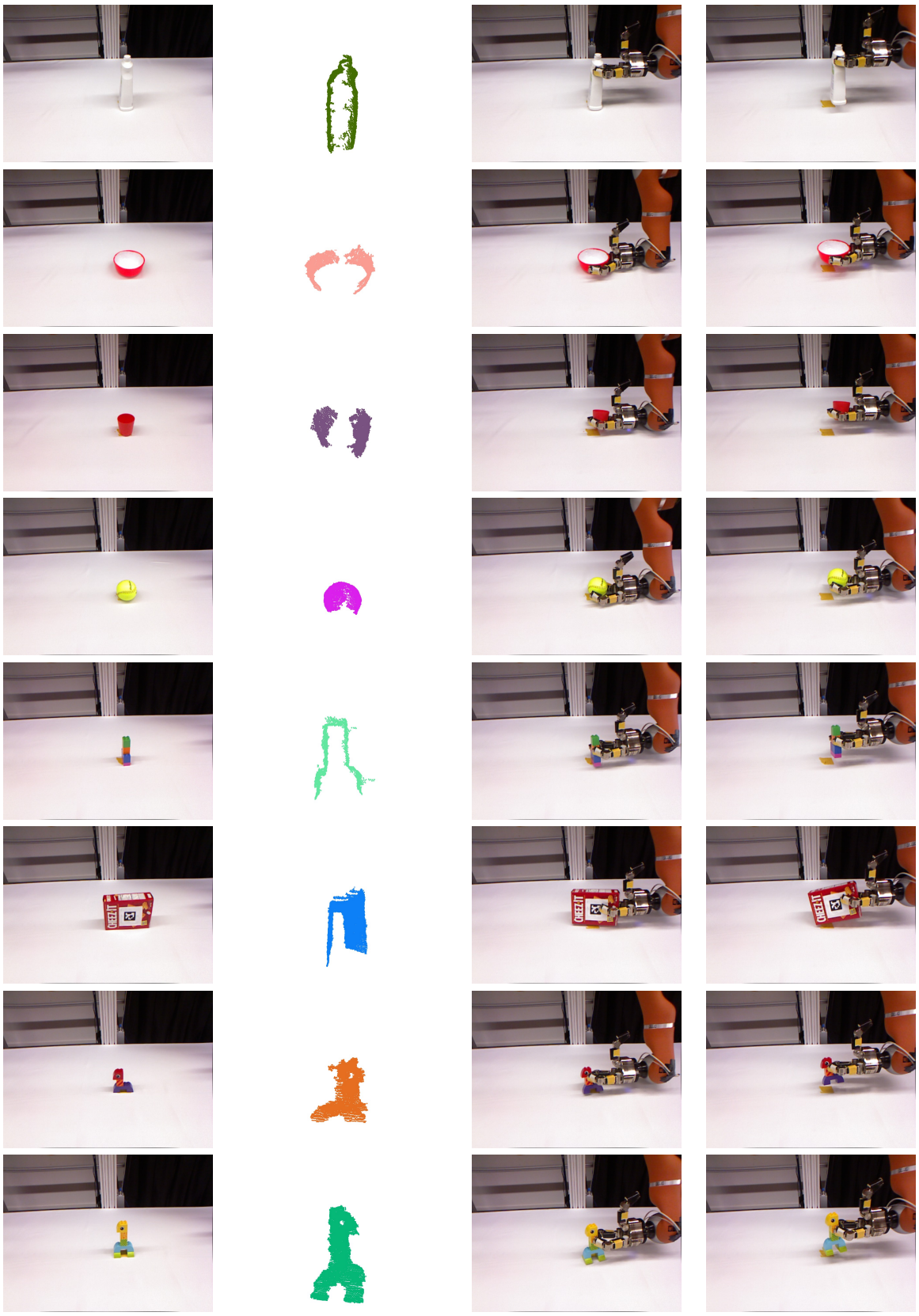

Table 1. Robotic Grasping Experiment Results. Each row shows grasping for a single object. From left to right, RGB image, point cloud, robot arm in grasping and lifting the objects.

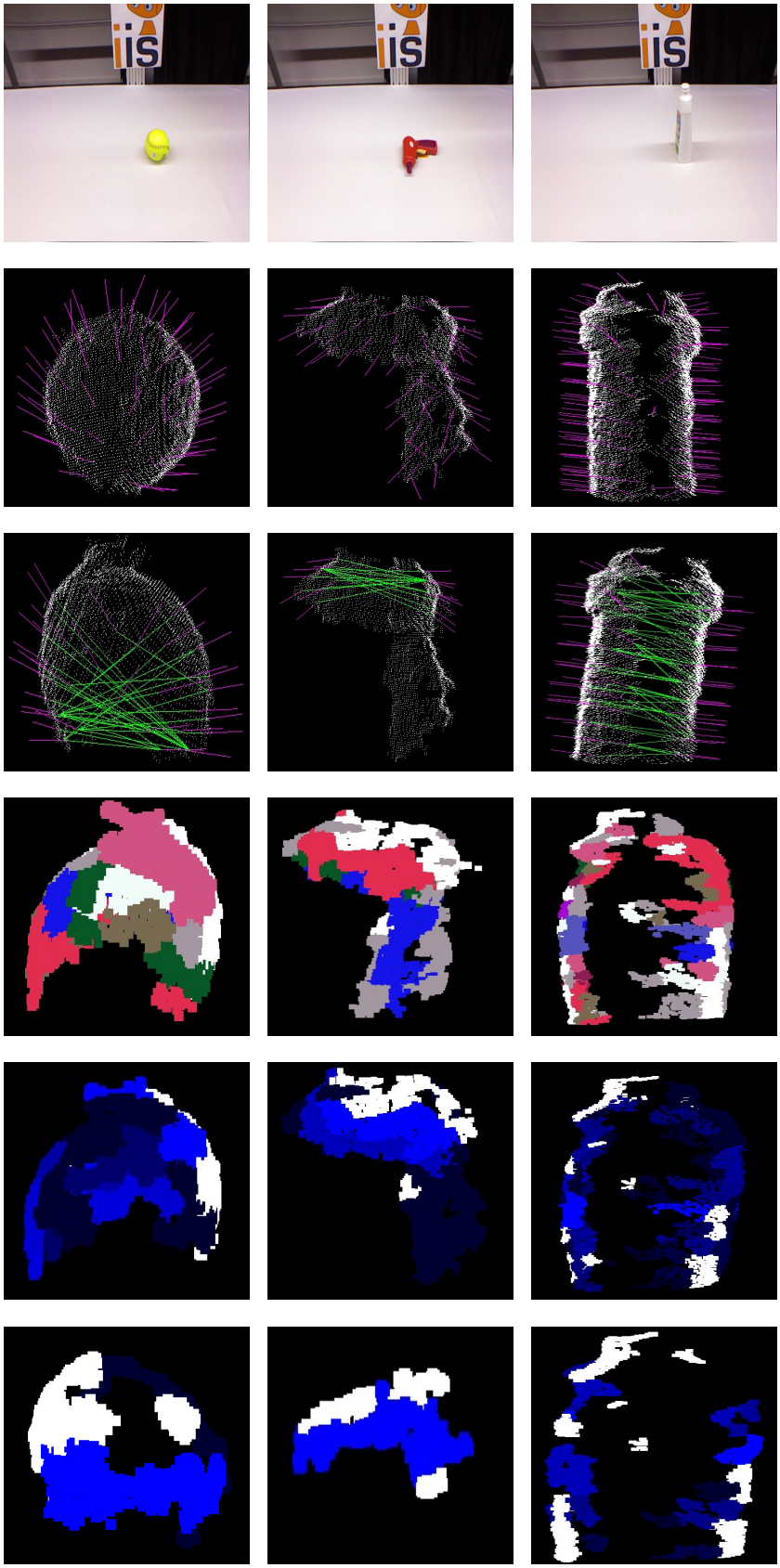

Figure 8. Object Decomposition Results. From top to bottom: input object, normals estimated on supervoxesl, candidate lines as contact points, fitted graspable ellipses and probability of grasp for all the ellipses, probability of grasp after ellipse elimination. Lighter colors show higher probable grasps. Overlapping ellipses are not removed from the first rows for a better understanding.

References

[1] A. Boularias, O. Kroemer, and J. Peters. Learning robot grasping from 3-d images with markov random fields. IEEE, Sept. 2011.

[2] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32, 2010.

[3] S. Fidler and A. Leonardis. Towards scalable representations of object categories: Learning a hierarchy of parts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[4] M. Kopicki, R. Detry, M. Adjigble, R. Stolkin, A. Leonardis, and J. L. Wyatt. One shot learning and generation of dexterous grasps for novel objects. *The International Journal of Robotics Research*, 2015.

[5] S. R. Lakani, M. Popa, A. J. Rodríguez-Sánchez, and J. H. Piater. CPS: 3d compositional part segmentation through grasping. In *12th Conference on Computer and Robot Vision, CRV 2015*, 2015.

[6] Q. V. Le, D. Kamm, A. F. Kara, and A. Y. Ng. Learning to grasp objects with multiple contact points. In *ICRA*. IEEE, 2010.

[7] I. Lenz, H. Lee, and A. Saxena. Deep learning for detecting robotic grasps. *IJRR*, 2015.

[8] A. Myers, C. L. Teo, C. Fermüller, and Y. Aloimonos. Affordance detection of tool parts from geometric features. In *ICRA*, 2015.

[9] B. Ommer and J. Buhmann. Learning the Compositional Nature of Visual Object Categories for Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32, 2010.

[10] A. Singh, J. Sha, K. Narayan, T. Achim, and P. Abbeel. Bigbird: A large-scale 3d database of object instances. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2014.

[11] S. Stein, F. Worgotter, M. Schoeler, J. Papon, and T. Kulvicius. Convexity based object partitioning for robot applications. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 3213–3220, May 2014.

[12] L. Zhu and A. L. Yuille. A hierarchical compositional system for rapid object detection. In *Advances in Neural Information Processing Systems*, 2005.