Learning to grasp from point clouds

Plinio Moreno Jonas Hornstein
José Santos-Victor
Instituto Superior Técnico & Instituto de Sistemas e Robótica
1049-001 Lisboa - Portugal
{plinio, jhornstein, jasv}@isr.ist.utl.pt

Abstract

We study how to encode the local information of the point clouds in such a way that a robot can learn by experimentation the *graspability* of objects. After learning, the robot should be able to predict the *graspability* of unknown objects. We consider two well known descriptors in the computer vision community: Spin images and shape context. In addition, we consider two recent and efficient descriptors: the Point Feature Histogram and the Viewpoint Feature Histogram. We evaluate the discriminative properties of the descriptors on a synthetic scenario and a simulator scenario, classifying the points with a standard learning algorithm, the Support Vector Machines. The results suggest the addition of more global information to the local descriptors.

1 Introduction

Previous works have shown promissing results on learning grasping points from local visual descriptors [13, 12, 8, 17, 4, 11]. The learning approaches aim to build a model from exploration of the objects that is able to generalize across objects. The majority of these approaches rely on local visual features computed on a stero pair in order to obtain sparse 3D features, which are utilized to estimate the graspable points of an object [13, 8, 17]. The recent development of range cameras have brought a denser 3D data, which have been utilized to build global object models for robotic grasping [11]. In difference to previous works, our aim is to build local object models using dense 3D data in order to learn the graspable and non-graspable shapes of objects.

Previous work at IST [12] addressed learning on an image-based approach by estimating the (full) probability distribution of grasping at each pixel given the output of a large set of visual descriptors. The image-based descriptors encode the similarity of the region around the pixel to several type of textures. We want to follow a similar approach in this work, taking into account the richer perception capabilities of the range camera mounted on the mobile platform. In this paper, we explore local descriptors of point clouds in order to discriminate between graspable and non-graspable regions.

In the context of 3d image registration and matching, several mesh-based and point cloud-based descriptors have been proposed. On the point cloud type, the most popular descriptors include the spin images [9] and the shape context [3]. In the context of robotics, the Point Feature Histogram (PFH) [15] and the Viewpoint Feature Histogram (VFH) [16] have shown both efficiency and robustness. We consider these four features in order to build local descriptors of the graspable and non-graspable regions.

The proper scenario for learning how to discriminate the local features must consider the following steps: (i) Acquisition of the point cloud, (ii) selection of a reference (reaching) point, (iii) execution of the grasping action several times and (iv) saving the grasping result and its corresponding features. This exploratory setup provides data for the application of a learning algorithm, which should be able to predict the grasping outcome on unseen objects. Since such an experimental setup is time consuming and the actual scenario is available just in a simulated world, we present results on two setups: On the first one, the data samples come from: (a) previously acquired point clouds from the ROS household database and (b) their grasping labels are defined by educated guesses. On the second one, the data samples come from: (a) partial views of the objects acquired in the ORCA simulator [1] and (b) their grasping labels come from the repetition of the grasping action four times. On one hand, If the grasping is successful across all the repetitions, a new positive sample is stored. On the other hand, if the grasping is unsuccessful on all the repetitions, a negative sample is stored.

The two experimental scenarios allow us to compare the feature capabilities on a perfect sensor acquisition vs. the partial view of the simulation. We apply the Support Vector Machines (SVM) learning algorithm [7] with three

different kernels: (i) linear and (ii) polynomial of degree two. We consider five objects on the synthetic scenario and four objects on the ORCA scenario, performing the leave-one-out validation in order to compute the True Positive (TP) rate and the False Positive (FP) rate.

2 Local descriptors

We consider the following 3D local descriptors for point clouds: Spin images [9], shape context [3, 10], PFH [15] and VFH [16]. The spin images is one of the most referred descriptors in the computer vision community, while the shape context is a very general descriptor that can be applied to points of any dimension and has shown good matching capabilities in both 2D and 3D [5]. The Point Feature Histogram (PFH) and Viewpoint Feature Histogram (VFH) have been related to grasping applications due to their efficiency and better matching results when compared to spin images [15, 16].

Spin image descriptor [9] represents the neighborhood N of a reference point in a cloud by fitting an oriented coordinate system at the point. The local system of cylindrical coordinates at the reference point p is defined using the normal and tangent plane: the radial coordinate α defined as the perpendicular distance to the line through the surface normal $n(p)$, and the elevation coordinate β, defined as the signed perpendicular distance to the tangent plane. The spin descriptor is a histogram of points in the support region represented in α, β coordinates. The support region is defined by limiting the range of the values of α and β and requiring that the angle between the normals is below a threshold, which considers self occlusion artifacts. The histogram is usually represented as a 2D image that at pixel (i, j) contains the number of points at the region parametrized by a range $\alpha \in [\alpha_1 \quad \alpha_2], \beta \in [\beta_1 \quad \beta_2]$. Thus, the spin image parameters are: (i) The number of neighbors considered, (ii) the α range and (iii) the β range.

The concept of shape context descriptor was first introduced in [3] for image analysis, though it is directly applicable to 3D shapes [10]. The shape context describes the structure of the shape as relations between a point to the rest of the points in the region. Given the coordinates of a point p on the shape, the shape context descriptor is constructed as a histogram of the direction vectors from p to the rest of the point, $|q - p|_2, \quad q \in N$. Typically, a log-polar histogram is used where the angle between the points is sampled uniformly and the distance between the points is sampled logarithmically. In the case of 3D points, the log-polar histogram has tree coordinates: the inclination angle, the azimuth angle and the distance (sampled logarithmically) so the parameters of the shape context are: (i) the number of bins for both angles and (ii) the number of distance bins.

The PFH encodes the statistics of the shape of a point cloud by accumulating the geometric relations between all the point pairs. Given a pair of points in the neighborhood and their normals, the PFH accumulates four dimensional histogram of (see Figure 1): (i) the cosine of the angle α, (ii) the cosine of the angle ϕ, (iii) the angle θ and (iv) the distance between the points (normalized to the bounding box of the neighborhood). The PFH parameters are : (i) the dimensions considered to compute the histogram and (ii) the number of bins for each dimension.

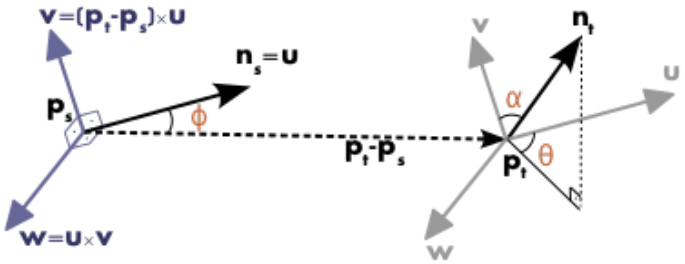

Figure 1: PFH computation illustration (from [14])

The VFH aims to remove the viewpoint invariance of PFH, by augmenting the PFH feature with the relation between the camera's point of view to the point cloud of an object. The view vector is provided by the direction of the principal axis of the camera and the viewpoint component is computed by collecting a histogram of the angles that the viewpoint direction makes with each normal. In difference to the multi-dimensional histogram of the original PFH, the PFH component of VFH computes one dimensional histograms of each of the four values mentioned above, between: (i) the viewpoint direction located at the object's central point and (ii) each of the

normals on the surface. Since our approach relies on a reference point and its normal, we compute the PFH component of VFH between: (i) the normal of the reference point and each of the normals of the neighborhood. Thus, we compute the PFH component of the VFH feature in order to describe a neighborhood, which we refer to as VFH in the remainder of the paper.

3 Learning procedure

We follow the reaching point concept presented by [8, 13] that relates some of the parameters of the gripper (end effector position and approaching orientation) to a particular point on the object (point location and its normal vector). Each reaching point can be classified as graspable or non-graspable according to the values of its local descriptor. We apply a traditional learning algorithm, the SVM classifier [7], in order to discriminate between graspable and non-graspable regions. A robot that learns by exploration the graspability of (local) neighborhoods of objects, should consider the following steps: (i) Data acquisition, (ii) reaching point selection (iii) feature computation and (iv) grasping trial execution. We consider two scenarios for the visual descriptor evaluation: (i) A first scenario where the data acqusition is simulated and the grasp labels are provided by educated guesses and (ii) a second scenario where all the steps are executed on the ORCA simulator.

3.1 Synthetic scenario

This scenario assumes that the 3D point data for each object has been acquired previously with a set of scans around the object, so we have complete point clouds. In addition, the grasping labels come from educated guesses so the learning by exploration is not really performed. Thus, this " perfect scenario" serves as a comparison baseline to the more realistic setup provided by the ORCA simulator.

In order to reduce the large amount of points of all the objects, we apply an interest point detector on the point clouds: the Laplace-Beltrami operator [2]). The local maxima of the Laplace-Beltrami operator selects points where there are local changes in curvature. These changes in curvature occurs in saddle points, hills, valleys and protuberances of the surface. In the case of a gripper, some of the interest points are graspable such as the hills and protuberances. In addition, the interest point selection reduces the complexity of the classification problem. Figure 2 illustrates the interest points selected by the local maximum of the Laplace-Beltrami operator for the objects considered in this paper.

We compute the descriptors at the reference points selected by the Laplace-Beltrami operator and employ the labels provided by educated guess. On this setup, we learn to discriminate between the graspable and non-graspable points.

3.2 ORCA scenario

This scenario simulates the complete learning cycle described on section 3, in the ORCA simulation environment [1]. In each experiment, the robot performs a grasping trials on a large number of reaching points. The objects that were chosen for these experiments were all different types of IKEA cups, ranging from small espresso cups to large size coffee cups. The setup is shown in Figure 3.

A reaching point, is defined as a fixed distance from a given point on the object, in the direction of the surface normal at that given point. The 3D data is acquired from a simulated range camera (with the parameters of the Kinect sensor), placed on the robot platform. Figure 4 shows an example of the depth image acquired from the Kinect. This image is then segmented in order to remove background objects such as the table, and the depth is converted to a point cloud describing the object, see Figure 5.

To reduce the number of reaching points, the point cloud is subsampled and at each of the chosen points, the surface normal is estimated from the neighboring points.

By using inverse kinematics, a trajectory is generated that allows the robot to approach the reaching point along the direction of the object's surface normal. The robot stops when the base of the gripper is placed at the reaching point. From this point the robot closes the gripper and moves the arm upwards while maintaining a closed grip. At the end of the movement, the distance between the fingers of the grippers is used to evaluate if the grasp was successful. If the distance is under a small threshold the object is assumed to have fallen out of the grip and the grasp is unsuccessful. An example of a successful grasp is shown in Figure 6 and an unsuccessful grasp in Figure 7.

Figure 2: The white diamonds display the location of the interest points from the local maxima of the Laplacian-Beltrami operator

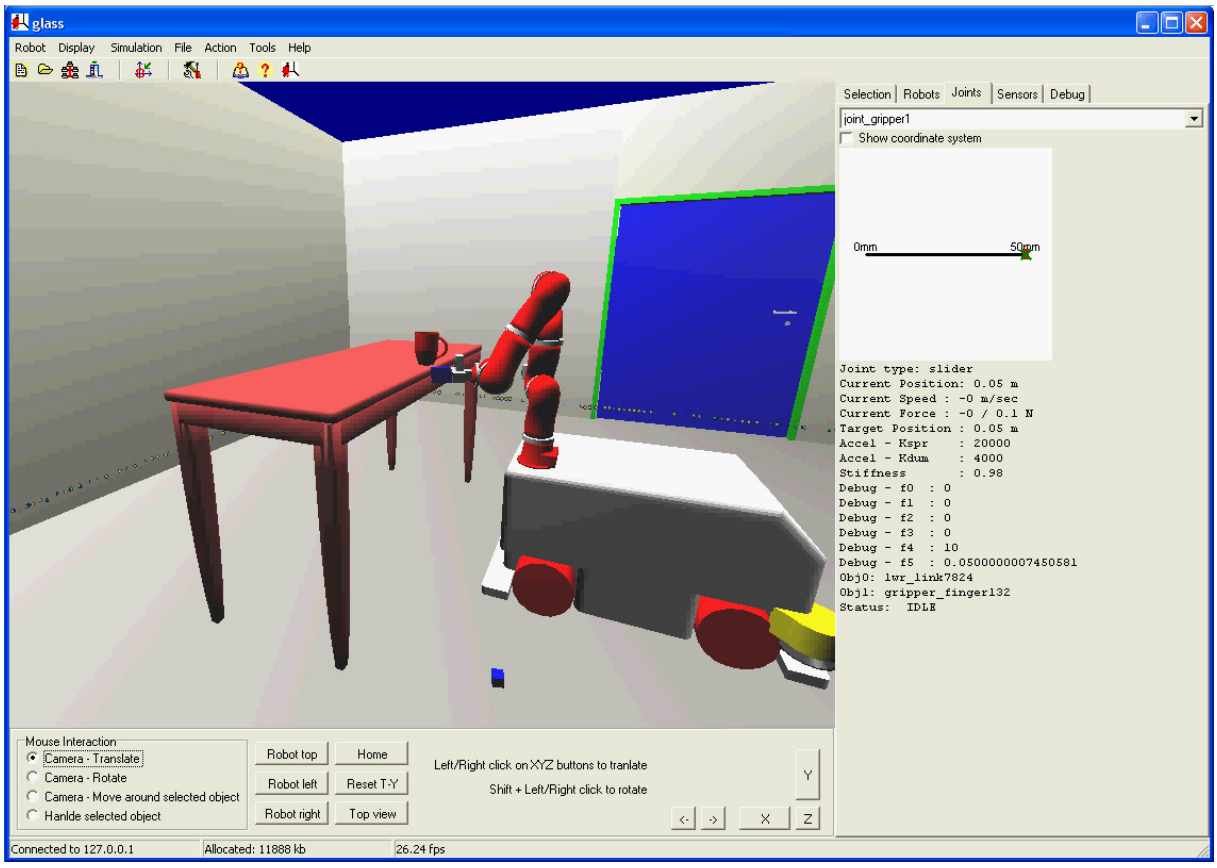

Figure 3: The ORCA simulator

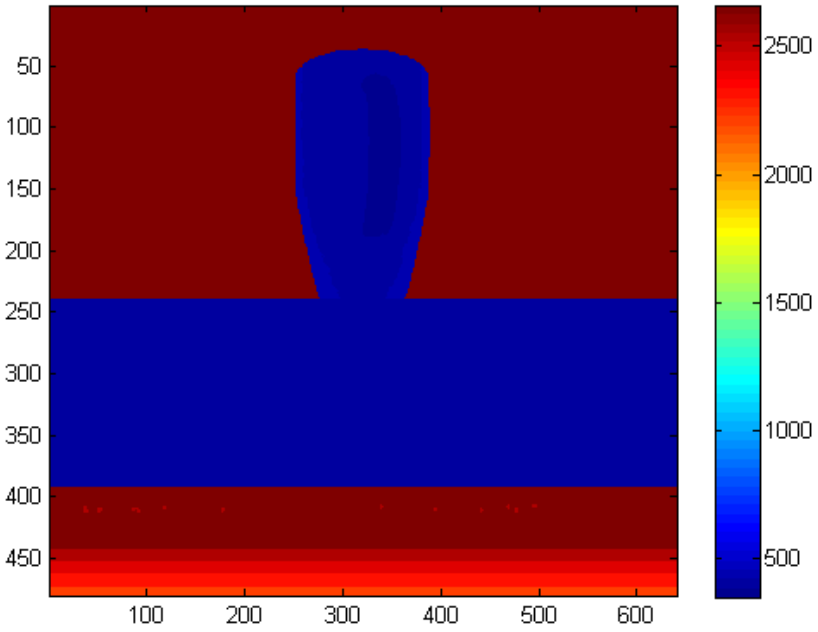

Figure 4: Depth image example

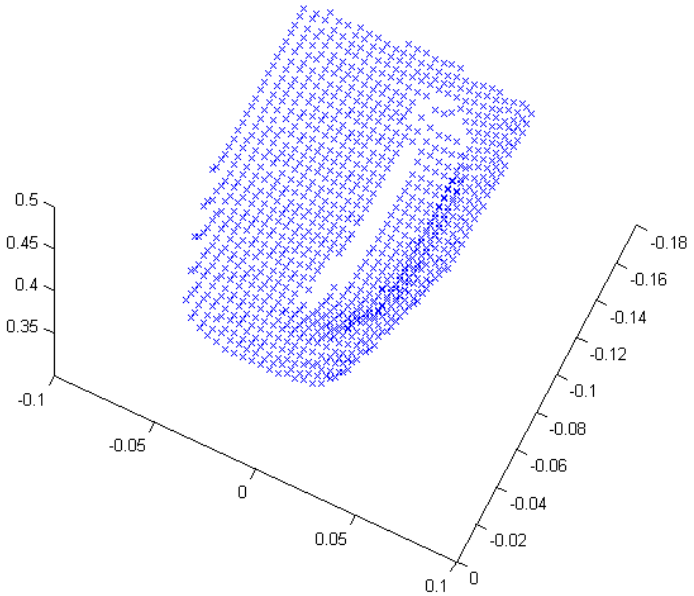

Figure 5: Point cloud example

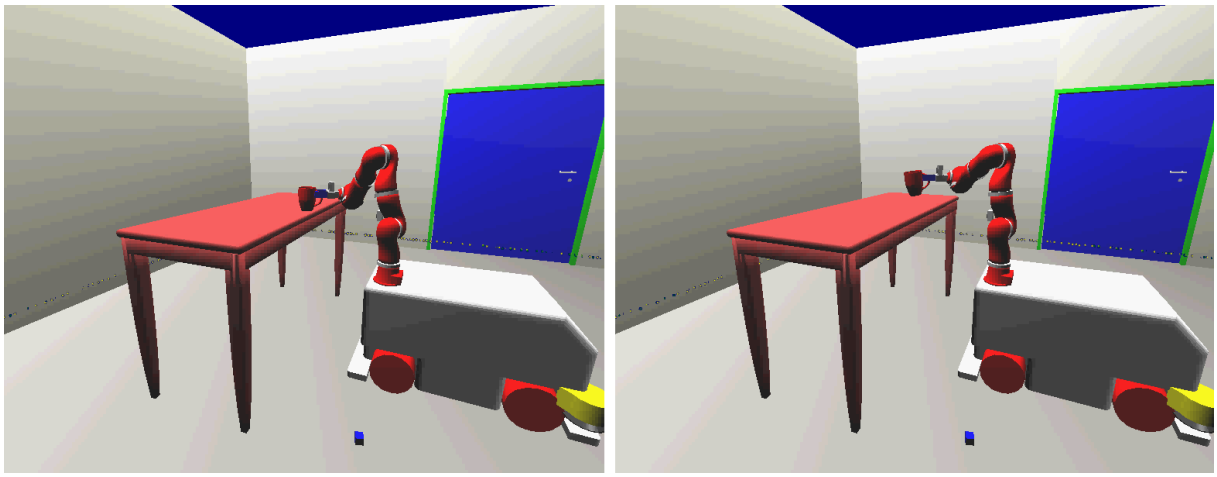

Figure 6: Successful grasp

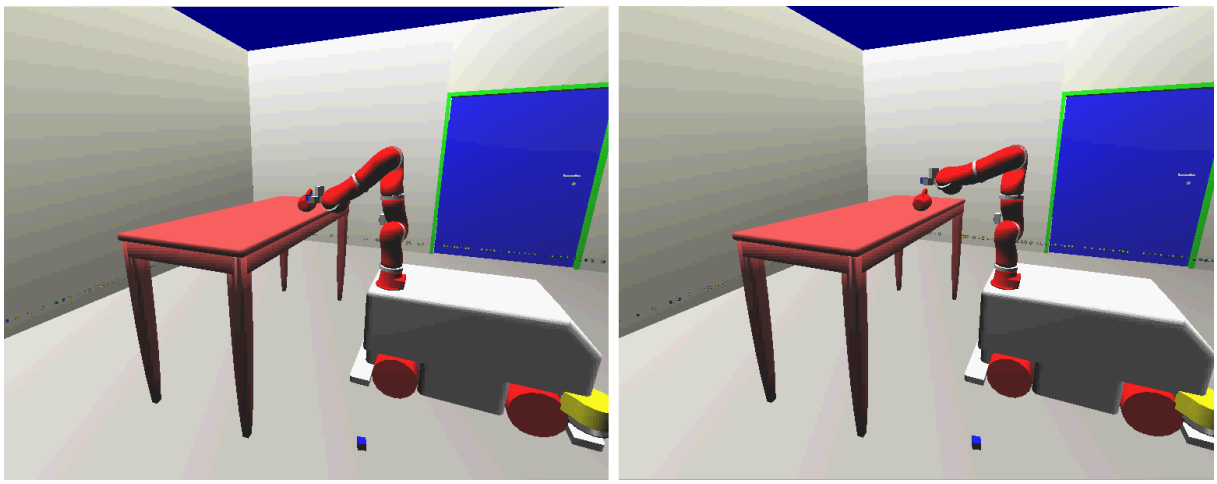

Figure 7: Unsuccessful grasp

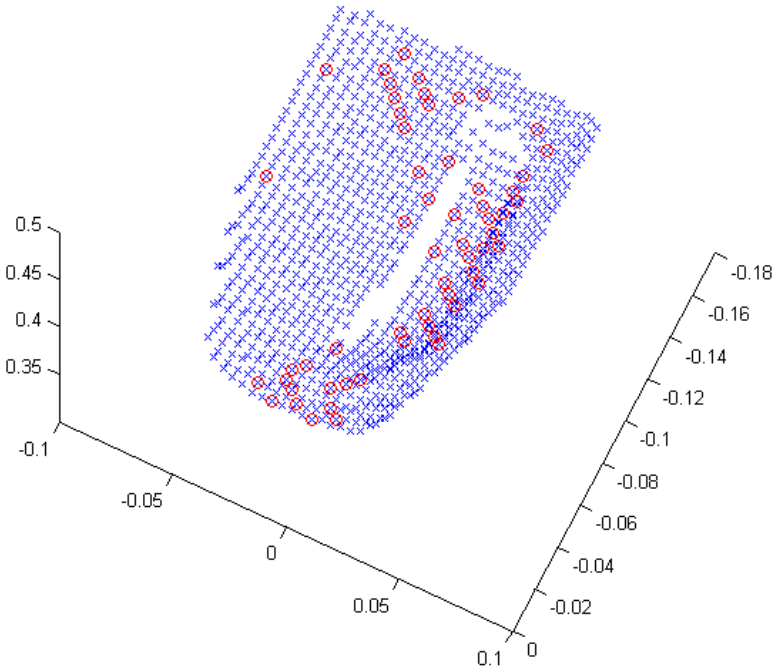

Figure 8: Grasping result. Points that resulted in a successful grasp are shown with a red circle.

This is repeated for each of the reaching points. The result of this experiment is shown in Figure 8. As can be seen in the figure, most of the successful grasps are found when grasping at the handle. In total six different objects were used in the experiments, and each object was tested with eight different orientations, resulting in a dataset of 477 successful grasping attempts and 5317 unsuccessful attempts.

4 Results

We evaluate the classification performance by applying the leave-one-out validation method and computing the True Positive (TP) and False Positive (FP) rates for each experiment. The leave-one-out test allow us to evaluate the generalization capabilities of the classifier. Regarding the grasping prediction on unseen objects, the critical issue is to have almost zero FP because we do not want the robot to try to grasp a non-graspable object. The features' parameters for all the experiments are:

- PFH: Three angular features (the distance is not considered) and three bins for each dimension. A sample $x \in \mathbb{R}^{3^3=27}$.
- VFH: All four features and three bins for each dimension. A sample $x \in \mathbb{R}^{4\times3=12}$.
- Spin images: Ten bins for each angular dimension. A sample $x \in \mathbb{R}^{10^2=100}$.
- Shape context: Four angular bins and four distance bins. A sample $x \in \mathbb{R}^{4^3=64}$.

4.1 Synthetic scenario

On the initial experimental round on this scenario, we have compared the reduction of the data samples by applying an interest point detector vs. the complete data. On the complete data scenario, the learning algorithm was not able to generalize on the testing set so all the points were classified as graspable. On the interest point selection scenario, we have found generalization capabilities. Thus, we just present the results of the reduced set of reaching points computed by the Laplace-Beltrami operator [2].

We choose five common objects available from the household object database [6], shown in Figure 2: a champaign glass, a wine glass, a cup, a espresso cup and a dish. We evaluate the generalization capabilities across similar objects in two tests: (i) We select the cups and glasses and divide them into two sets for the leave-one-out setup and (ii) use the classifiers learnt in the previous step to classify the reaching points of a very different object, the dish. These two tests are shown on the left and right columns of Table 1 and Table 2.

		4 object	5 object
PFH	TP (%)	15.13	8.41
	FP (%)	4.76	4.24
VFH	TP (%)	1.43	3.72
	FP (%)	0.56	0.39
PFH + VFH	TP (%)	19.41	11.53
	FP (%)	6.65	5.28
Spin Images	TP (%)	18.09	10.34
	FP (%)	8.72	8.26
Shape Context	TP (%)	5.13	5.28
	FP (%)	2.34	2.41

Table 1: Results of synthetic scenario using the SVM with linear kernel. The top three features are: PFH + VFH, Spin Images and PFH

		4 object	5 object
PFH	TP (%)	25.23	15.03
	FP (%)	9.43	9.47
VFH	TP (%)	9.85	6.16
	FP (%)	4.22	3.1
PFH + VFH	TP (%)	25.53	16.68
	FP (%)	7.57	6.58
Spin Images	TP (%)	24.1	21
	FP (%)	16.76	15.39
Shape Context	TP (%)	14.98	14.47
	FP (%)	13.39	12.43

Table 2: Results of synthetic scenario using the SVM with polynomial kernel of degree 2. The top three features are: PFH + VFH, PFH and Spin images

We remark that independently of the type of kernel and testing set, the top three features are: PFH + VFH, PFH and Spin images. However, the TP rate is not very high and the FP is very large for grasping purposes. Figure 9 illustrates the classification result of one of the tests.

4.2 ORCA scenario

We choose four cups in order to evaluate the learning procedure, shown in Figure 10. We aim to evaluate the generalization capabilities of the features under the partial view constraint, which has a large impact on the local descriptor. By applying the leave-one-out setup on the data acquired on ORCA, we observe that the top three features are: PFH + VFH, PFH and shape context. However, the generalization capabilities of the features computed on partial point clouds is reduced.

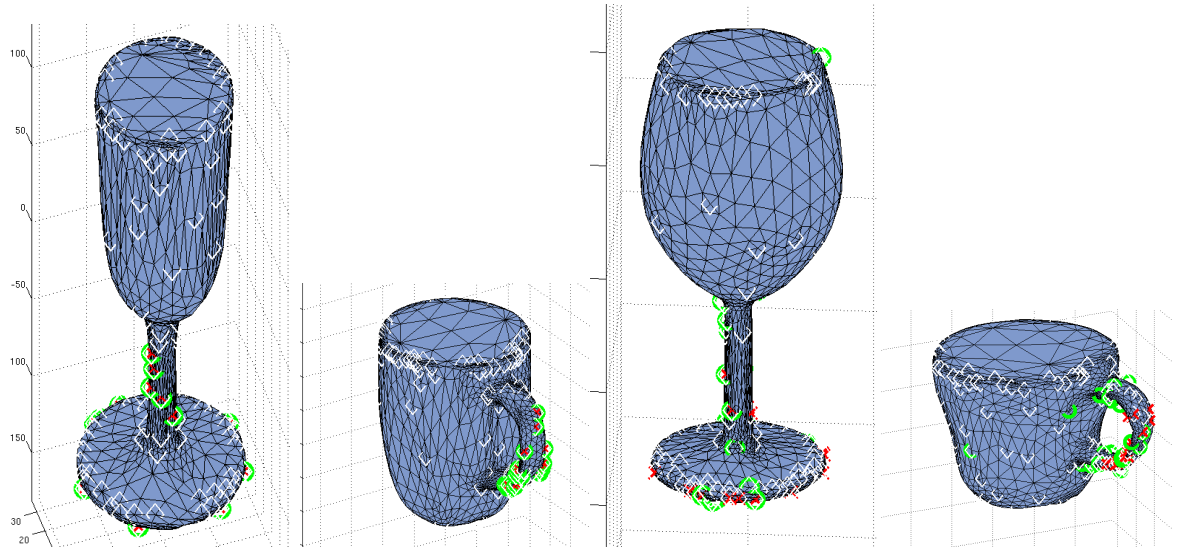

Figure 9: The white diamonds display the locations of the interest points using the Laplacian operator on the point clouds, the red crosses show the hand-labeled grasping points and the green circles show the points classified as graspable by the SVM learning algorithm. The left hand side images show the classification on the training set, where all the grasping points are correctly detected. The right hand side images show the classification on the testing set, where most of the graspable points are correctly detected with very few false positives.

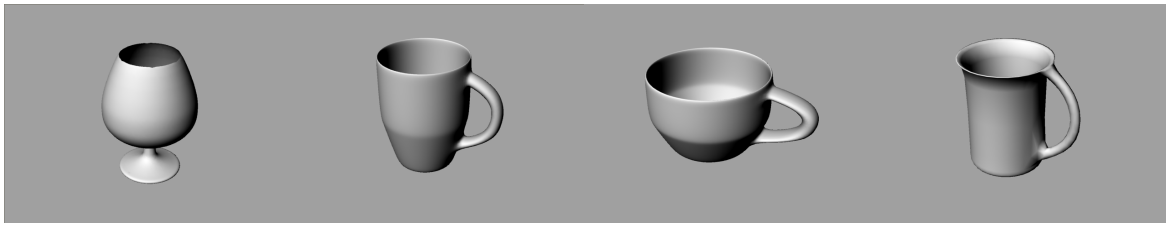

Figure 10: Objects used in the ORCA simulation environment

	True Positives (%)	False Positives (%)
PFH	27.5	21.13
VFH	9.43	9.47
PFH + VFH	29.4	19.3
Spin Images	0	11.84
Shape Context	17.5	9.93

Table 3: Results of ORCA scenario using the SVM with polynomial kernel of degree 2

5 Conclusions

We explore the application of point cloud descriptors on the classification of graspable vs. non-graspable points. Our approach fits into the learning by experimentation paradigm, where the robot gathers the grasping labels from its own experience. We consider four point cloud descriptors: PFH, VFH, spin images and shape context. As an initial approach, we select a small set of kitchen objects to perform the experiments and apply the SVM learning algorithm. The results show that the PFH gathered with the VFH feature provide promissory results which may be improved by considering more global information of the objects.

Acknowledgements

This work was supported by EU Project First-MM (FP7-ICT-248258)

References

[1] H. Baltzakis. Orca simulator. Available from: http://www.ics.forth.gr/~xmpalt/research/orca/.

[2] Mikhail Belkin, Jian Sun, and Yusu Wang. Constructing laplace operator from point clouds in rd. In *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '09, pages 1031–1040, Philadelphia, PA, USA, 2009. Society for Industrial and Applied Mathematics. Available from: http://dl.acm.org/citation.cfm?id=1496770.1496882.

[3] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(4):509–522, apr 2002. doi:10.1109/34.993558.

[4] Jeannette Bohg and Danica Kragic. Learning grasping points with shape context. *Robotics and Autonomous Systems*, 58(4):362–377, 2010. Available from: http://www.sciencedirect.com/science/article/pii/S0921889009001699, doi:10.1016/j.robot.2009.10.003.

[5] A. M. Bronstein, M. M. Bronstein, and M. Ovsjanikov. *3D Imaging, Analysis, and Applications*, chapter "3D features, surface descriptors, and object descriptors". Springer.

[6] Matei Ciocarlie. Household objects database. Available from: http://www.ros.org/wiki/household_objects_database.

[7] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995. 10.1007/BF00994018. Available from: http://dx.doi.org/10.1007/BF00994018.

[8] A.N. Erkan, O. Kroemer, R. Detry, Y. Altun, J. Piater, and J. Peters. Learning probabilistic discriminative models of grasp affordances under limited supervision. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 1586 –1591, oct. 2010. doi:10.1109/IROS.2010.5650088.

[9] Andrew E. Johnson and Martial Hebert. Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 21(5):433–449, 1999.

[10] Marcel Körtgen, G. J. Park, Marcin Novotni, and Reinhard Klein. 3d shape matching with 3d shape contexts. In *The 7th Central European Seminar on Computer Graphics*, April 2003.

[11] Alexis Maldonado, Ulrich Klank, and Michael Beetz. Robotic grasping of unmodeled objects using time-of-flight range data and finger torque information. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Taipei, Taiwan, October 18-22 2010.

[12] L. Montesano and M. Lopes. Learning grasping affordances from local visual descriptors. In *IEEE 8TH International Conference on Development and Learning*, China, 2009.

[13] Mila Popović, Dirk Kraft, Leon Bodenhagen, Emre Başeski, Nicolas Pugeault, Danica Kragic, Tamim Asfour, and Norbert Krüger. A strategy for grasping unknown objects based on co-planarity and colour information. *Robot. Auton. Syst.*, 58:551–565, May 2010. Available from: http://dx.doi.org/10.1016/j.robot.2010.01.003, doi:http://dx.doi.org/10.1016/j.robot.2010.01.003.

[14] R. Rusu. Point cloud library. Available from: http://pointclouds.org/documentation/tutorials.

[15] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. In *The IEEE International Conference on Robotics and Automation (ICRA)*, Kobe, Japan, 05/2009 2009.

[16] Radu Bogdan Rusu, Gary Bradski, Romain Thibaux, and John Hsu. Fast 3d recognition and pose using the viewpoint feature histogram. In *Proceedings of the 23rd IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Taipei, Taiwan, 10/2010 2010.

[17] A. Saxena, J. Driemeyer, and A. Y. Ng. Robotic Grasping of Novel Objects using Vision. *The International Journal of Robotics Research*, 27(2):157, 2008.