Enhancing Visual Perception of Shape through Tactile Glances

Mårten Björkman, Yasemin Bekiroglu, Virgile Högman, and Danica Kragic

***Abstract*— Object shape information is an important parameter in robot grasping tasks. However, it may be difficult to obtain accurate models of novel objects due to incomplete and noisy sensory measurements. In addition, object shape may change due to frequent interaction with the object (cereal boxes, etc). In this paper, we present a probabilistic approach for learning object models based on visual and tactile perception through physical interaction with an object. Our robot explores unknown objects by touching them strategically at parts that are uncertain in terms of shape. The robot starts by using only visual features to form an initial hypothesis about the object shape, then gradually adds tactile measurements to refine the object model. Our experiments involve ten objects of varying shapes and sizes in a real setup. The results show that our method is capable of choosing a small number of touches to construct object models similar to real object shapes and to determine similarities among acquired models.**

I. Introduction

One of the reasons that makes the process of autonomous grasping challenging is that object properties required for grasp planning such as shape are commonly not known a priori. In addition, sensory information used to extract this information from the environment, e.g. vision, is prone to error. Processes prior to shape extraction such as scene segmentation are not perfectly accurate due to several issues, e.g., occlusions and noisy measurements. Besides object shape, conceptual high-level object category information is another important input that can be used. In particular, for goal-oriented grasp planning, different instances from the same category can be grasped in a similar way for a particular task. For instance, bottles should be grasped from a side for a pouring task, so as not to block the opening.

Humans interact with the environment using rich sensory information. Studies show that both visual and haptic modalities contribute to the combined percept [1]–[3]. Results from [3] suggest that observers integrate visual and haptic shape information of real 3D objects and that bimodal shape estimates are more reliable than shape estimates that rely on either vision or touch alone.

The goal of our work is to complement visual information with tactile sensing in order to acquire 3D object models. We investigate how to deal with uncertainties in the sensory data to extract object shape and category. Given a scene like the one shown in Fig. 1, with an object in the center of view, our goal is to gain insight on what manipulation actions the object affords. If the shape of the object were known, one could get some idea of what actions to consider, especially if the shape is similar to an object that has already been manipulated. Much information can be provided through stereo vision, using e.g. a Kinect device. Regardless of which stereo vision system is used, however, only one side of an object is seen, i.e., the one side facing the cameras. Without any additional sensory modalities, one can only make qualified guesses of what the occluded side looks like, using assumptions such as symmetry [4], assumptions that may well be incorrect. In this paper we instead propose touch as a means to get additional information. By carefully touching the object, we will show how an object model can be created, a model that provides enough information to categorize the object based on shape.

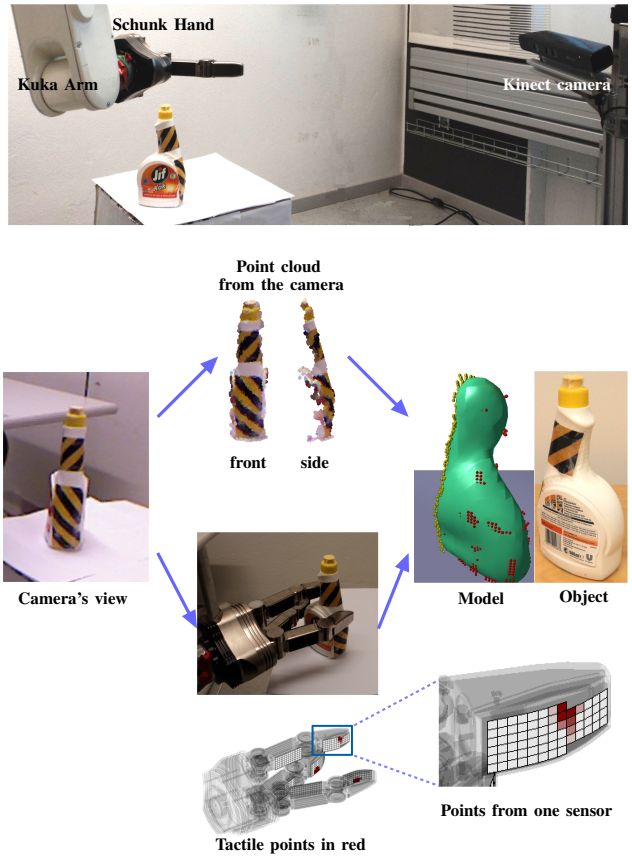


Fig. 1: Extracting object model: We rely on visual measurements from a Kinect and tactile measurements from the fingers. The model is formed based on the point cloud (in yellow) from the camera and the contact points (in red).

M. Björkman, Y. Bekiroğlu, V. Högman, D. Kragic are with the Centre for Autonomous Systems and the Computer Vision and Active Perception Lab, CSC, KTH Royal Institute of Technology, Stockholm, Sweden. Email: {celle|yaseminb|virgile|dani}@kth.se. This work was supported by the Swedish Research Council and the EU projects eSMCs (FP7-IST-270212) and RoboHow.Cog (FP7-ICT-288533).

II. Related work and Contributions

In robotics, object shape estimation has been studied with unimodal data, i.e., only visual [5] or tactile [6] sensing, and bimodal data with visual and tactile sensing combined [7]. Clearly, vision alone delivers useful information about object shape. Krainin et al. [5] proposed a method where a robot picks up and moves an object in front of a sensor. Their approach based on Kalman filters is able to build 3D models of unknown objects using a depth camera observing the robot's hand moving the object. However, they showed that the approach may produce failures with poor alignment in case of a combination of high uncertainty in the object pose, nondistinctive object geometry (completely planar surface) or fairly uniform color and poor lighting conditions. Tactile information can be used to alleviate such problems.

Bierbaum et al. [8] introduced the idea of using Dynamic Potential Fields for tactile exploration to build a contact/tactile point cloud of an unknown object. Their system requires a rough initial estimate about the object position, orientation and dimension, then exhaustively performs grasps in unexplored regions. Faria et al. [9] also builds contact point clouds in an exhaustive way. They however follow a probabilistic approach to store the extracted tactile points in a volumetric map. In their experiments, a human subject wearing a glove with magnetic tracking sensors to obtain fingertip positions performs grasps that follow the contour of objects. Meier et al. [6] followed a similar strategy and used a probabilistic approach, Kalman filters, to build a model of the contact point cloud. Their robot grasps objects at different heights and positions also varying the orientation of the hand. Their results show that the acquired models can successfully be used for classification.

There are approaches that supplement vision with more sensory information especially where visual sensing is weak, e.g., occluded object parts. Maldonado et al. [10] used a proximity sensor to scan the unseen parts of an object by a depth camera without touching the object. They combined the point cloud from the camera and the sensor and built a 3D Gaussian point representation based on the convex hull of the complete point cloud. Their representation simply contains the centroid and the shape of the object through the mean and the covariance matrix of the Gaussian distribution. Dragiev et al. [7] has included laser data in addition to haptic measurements in order to complement vision. They proposed to use Gaussian Process Implicit Surfaces to fuse the uncertain sensory data and showed that this representation can be used to control reaching and grasping such that the hand is moved and oriented towards the object and grasps aligning the fingers according to the object shape. There are also studies that focus on object recognition without explicitly modeling the full 3D shape, but rather representing the objects based on visual [11], tactile [12], [13] or both features [14].

Differently from the aforementioned approaches, we focus on building object models that can be extracted with a small number of actions (touches) in order to understand the category objects belong to, rather than exhaustively trying to explore the whole object. This is an iterative process where the robot executes more touches as it is less confident about the object shape. In summary, our contributions can be listed as follows:

- We incrementally include tactile readings in the shape estimation to further refine the object model that is initialized based on visual measurements only.
- We use a probabilistic approach to shape estimation through Gaussian Process regression to deal with uncertainties in sensory measurements.
- Instead of an exhaustive exploration, we obtain a model of a given object by selecting where to touch next, given the object regions where the shape estimation is most uncertain.
- Our system is able to build models that can be used for shape categorization after a small number of touches.

III. Object modelling

In this section, we describe how objects are represented based on visual and tactile measurements. We introduce Gaussian Process regression modeling of Implicit Surfaces, the strategy to determine how to acquire tactile data and the shape descriptors used for measuring similarities between different objects.

A. Visual Measurements

An observed object is segmented from its background using a segmentation and tracking system that works over sequences of touches. The system uses stereo vision, in our case a Kinect device, in an heterogeneous MRF based framework [15]. The framework uses color and depth information to divide the scene into either planar surfaces, bounded objects or uniform clutter models. The planar and uniform models are automatically initialized, while an ellipsoid used to model the observed object is initiated by a point that is manually placed inside the corresponding image region. From the resulting object segments we get point clouds that serve as starting points for object modelling. Later we will complement these points with tactile readings from touches.

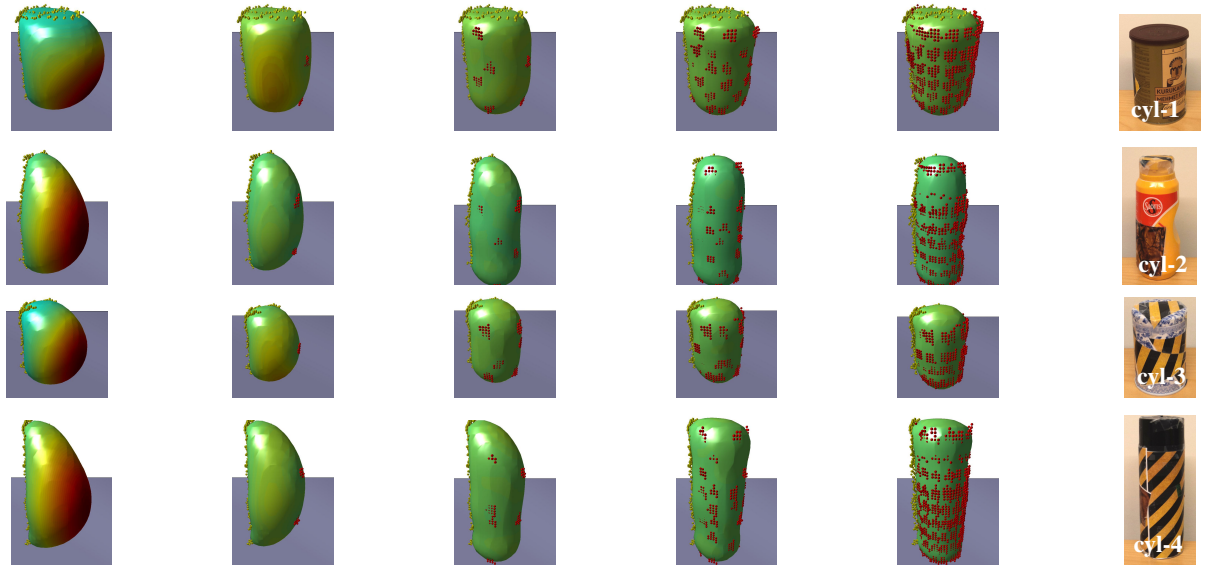
B. Implicit surfaces

From a set of measurements of 3D points $\{\mathbf{x}_i, i = 1...N\}$ that are located on the surface of an object, we now describe how to derive implicit surfaces for representation. The model should later be used for deciding object category based on shape. In our case the measurements originate from stereo vision as well as tactile readings. With a function $f : \mathbb{R}^3 \mapsto \mathbb{R}$, we define an implicit surface by the supporting points $\mathbf{x} \in \mathbb{R}^3$ that satisfy

$$f(\mathbf{x}) = 0.$$

The function $f(\mathbf{x})$ is modelled by Gaussian Process (GP) regression [16], with each observation $y = f(\mathbf{x}) + \epsilon$ assumed to be subjected to zero-mean Gaussian noise, $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$. The shape of the GP is governed by a thin plate covariance function [17]

$$cov(f(\mathbf{x}_i), f(\mathbf{x}_j)) = k(\mathbf{x}_i, \mathbf{x}_j) = 2|r|^3 - 3Rr^2 + R^3,$$



(a) v. only (b) 1 t. (c) 4 t. (d) 12 t. (e) 54 t. Real

Fig. 2: Evolution of object models against the number of touches (t.) for the cylinders: (a) Initial models based solely on visual (v.) measurements depicted by yellow points. The models are oriented to show the back sides that are not visible by the camera. Highest uncertainty is represented by red color and dark green regions correspond to least uncertainty. (b) Models after including tactile measurements from one touch applied to the region with highest uncertainty. Contact points obtained by touching are depicted by red points. (c) Models after 4 touches. (d) Models after 12 touches which were found sufficient to group all the objects confidently. (e) Models after exhaustively exploring the objects which require 54 touches with our setup. (f) The real objects for comparison.

where $r = |\mathbf{x}_i - \mathbf{x}_j|$ and R is a maximum possible value of r. This covariance function has slightly better characteristics than the more frequently used squared exponential function, in particular for rectangular objects where the flatness of surfaces needs to be preserved. Quantitatively, however, we have not observed any significant differences between the two when applied for categorization.

The model is learned from a set of tuples $(\mathbf{x}_i, y_i)$, where $y_i = 0$ for the stereo vision or tactile measurements. Since a physical object, at least those that can be acted on by a robot, occupies a certain volume in 3D space, the implicit surfaces need to be compact (closed and bounded). In order to guarantee this, we place additionally exterior points, for which $y_i = +1$, on the boundaries of the scene and a single interior point that is forced to be inside the closed surface with $y_i = -1$.

In the later experiments, this interior point was chosen as the centroid of the stereo point cloud displaced by 1 cm along the direction of the camera, assuming that this is the smallest object thickness one can expect. With objects assumed to be located within a cube with side lengths $L = 30$ cm and centered at the centroid, the parameter R was set to $\sqrt{3}L$. The only remaining hyperparameter is the expected noise level which is set to $\sigma_n^2 = 0.1$. The value was chosen so as a balance between the smoothness of the surfaces and the noise in the integration of tactile and visual readings.

C. Action selection and cue integration

With GP regression we do not explicitly get a function $f(\mathbf{x})$, but the mean $\bar{f}$ and variance $\mathbb{V}(f)$ of all possible functions that could fit the measurements. The variance can be used as a measure of uncertainty, with higher variance for points far away from already recorded measurements. Examples of implicit surfaces and variances can be seen in Fig. 2a. A surface is given by points for which the mean is zero and the colors illustrate the corresponding variances, with red for points of highest variance. The stereo vision point clouds are shown as yellow points, most of which are occluded by the objects in the figure.

To refine the object models and decrease the uncertainty, the robotic hand is guided towards those points for which the variance is large in order to select a position for touching. We call these touches *ordered* touches in the experiments below, as opposed to *random* touches where new touches are selected in random order. The arm-hand configuration has earlier been calibrated with respect to the camera system, with a precision of a few millimeters. The highest variance point is searched for in a discrete action space defined by the vertical position and the approach angle, both of which are computed with respect to the centroid of the current model. For each possible action, the closest point to each respective tactile sensor pad is found on the implicit surface. The action selected for execution is then based on the maximum variance found among all actions and sensor pads.

Touches are then executed in sequence and the GP model

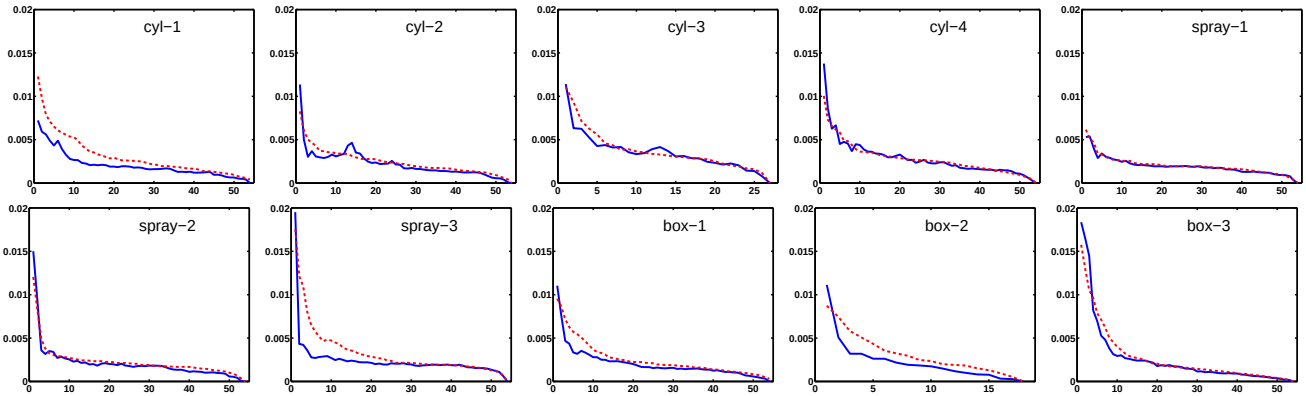


Fig. 3: Convergence of curvature point clusters using ordered (solid) and random (dashed) touches with respect to the final result after up to 54 touches (x-axis). Plots from random touches are based on the average from 10 runs and are thus smoother than those of the ordered touches.

is updated accordingly, as new measurements are integrated with the model. To speed up computations, the measurement set is made sparser to about two thousand points[1]. This is necessary since the computational cost of the GP increases to the cube of number of points. Recorded tactile measurements can be seen as red points in Fig. 2 for an increasing number of touches. In some cases, these points are displaced with respect to the implicit surface, which might happen if the object moves considerably when it is touched. To minimize these displacements, which is critical for long sequences of touches, the object frame is constantly updated for each new touch. This is done by registering the stereo vision point clouds, given by the segmentation system, before and after a touch using the Iterative Closest Point algorithm [18] and transforming new measurements back to the original frame.

D. Shape Descriptors

Representing object shape as a GP or a mesh derived from points on the resulting implicit surface is not straightforward, if the goal is to compare shapes for action selection. Instead we represent the extracted implicit surfaces with shape descriptors that capture information invariant to possible manipulation actions, while discarding redundant information. Two very different objects may afford similar actions, while two seemingly similar objects might not. For example, a rectangular box and a cylinder typically require different grasping strategies, but they may well appear similar when e.g., represented as ellipsoids, if aspect ratios are similar.

In this work, we look at two different rotation and translation invariant shape descriptors; 3D Zernike moments and surface curvatures. Zernike moments have successfully been used for shape retrieval [19] and are attractive due to the flexibility and low number of dimensions required, as well as the fact that Euclidean distances can be used for shape comparison. For the Zernike moments, voxelization is first applied in a 3D grid with voxels of side length $l = 0.75\ cm$, keeping the interior voxels for which the GP means are $\bar{f} \leq 0$ at their center points.

[1]On a 3.2 MHz Core i7 CPU the cost of computing the GP model and associated shape descriptor is about 4 s using PCL, VTK and Eigen.

For comparison using surface curvatures, the Marching Cubes algorithm [20] is first applied to the same grid to find a triangular mesh representing the implicit surface. From this mesh, principal curvatures are then computed [21], with one 2D measurement per vertex point. The shape of an object is thus represented by a sample set of about 500 measurements of curvatures. A kernel based two sample test [22] is used to compare two such representations, using Gaussian kernels with standard deviations of 0.25, which yields a soft decision on the similarity between the sample sets.

IV. Experimental Evaluation

In this section, we first describe our experimental platform and then present results from shape estimation and categorization experiments comparing touch selection strategies and shape descriptors.

The experimental robot platform is composed of an industrial Kuka arm (6 dof), a three-finger Schunk Dextrous hand (7 dof) equipped with tactile sensing arrays, and a Kinect stereo vision camera. The robot can acquire tactile imprints via pressure sensitive tactile pads mounted on the Schunk hand's fingers. Each finger of the hand has 2 tactile sensor arrays composed of 6x13 and 6x14 cells, which yields at most 486 tactile points after one touch. For each touch, the hand is set to a fixed initial joint configuration where the thumb opposes the other two fingers as seen in Fig. 1, then fingers are closed until contact is sensed.

In an earlier study [23] we concluded that the object class was an important factor, if one wants to determine what grasping action to pursue to fulfill particular tasks, tasks such as hand-over, pouring or dish-washing. However, the object class was not derived directly from sensory data, but given manually prior to the experiments. In this work, we aim to automate this process by learning shape-dependent features to replace the manually set object class. Our starting point is thus a set of objects for which we know the respective affordances from earlier experiments. These ten objects can be seen in Fig. 4, with names indicating the similarity in afforded actions. The end goal is to use stereo vision and tactile measurements through a series of touches to

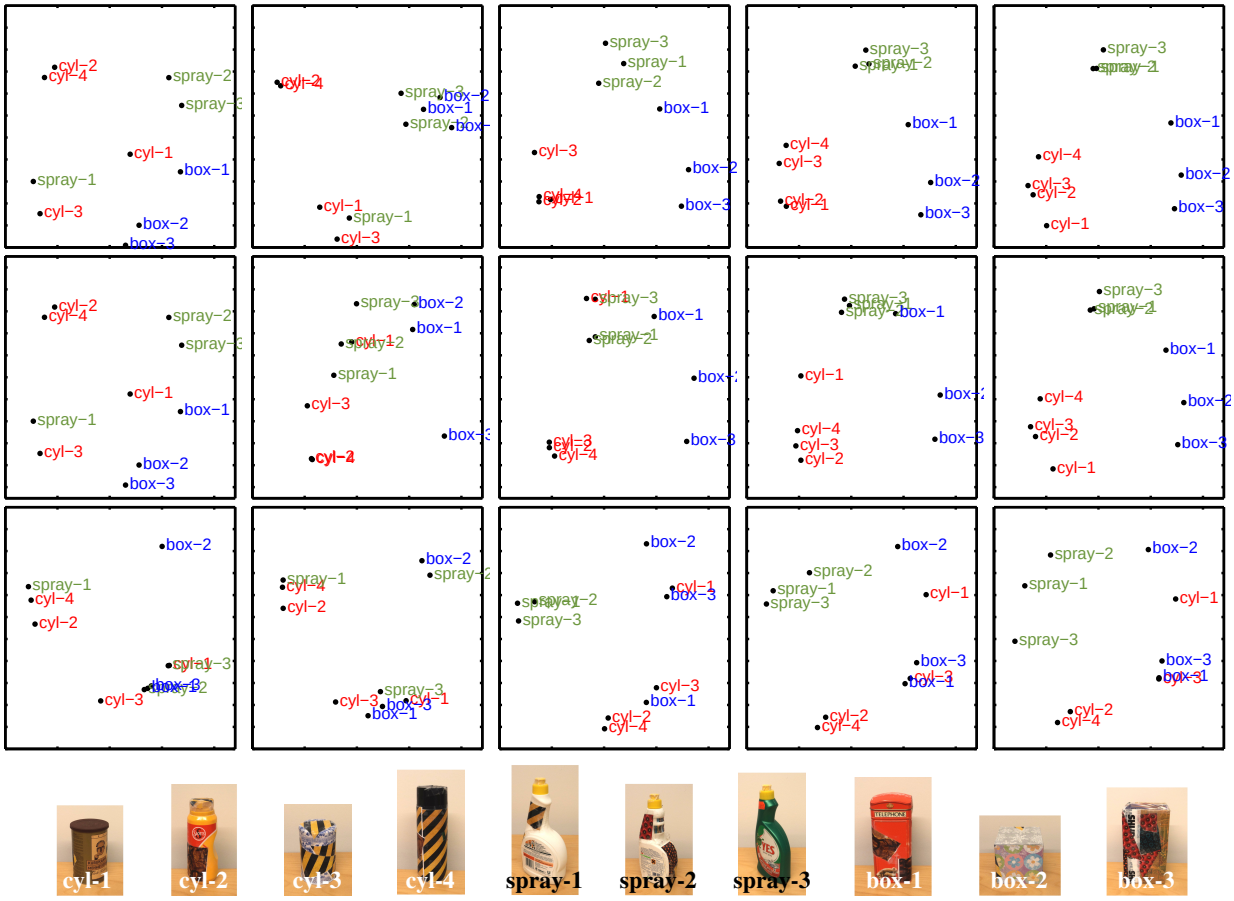


Fig. 4: Spectral embeddings of curvature point clusters, after 0 (left), 1, 4, 12 and 54 (right) touches, using ordered (first row) and random (middle row) touches, as well as with Zernike moments and ordered touches (last row). Ordered touches lead to faster convergence than random touches, and Zernike moments cluster objects more based on similarity in object aspect ratios, than similarities in affording grasp actions.

determine which grasping action the object would afford. The question is: how many touches this would require and what representation should one aim for?

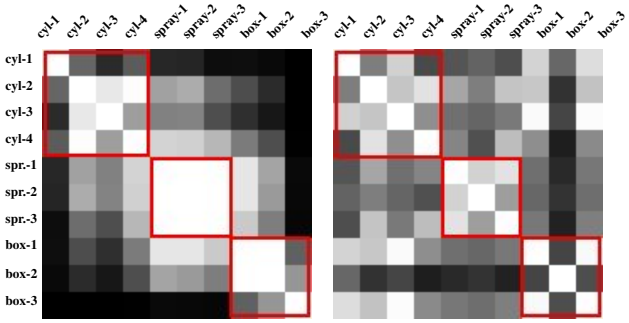
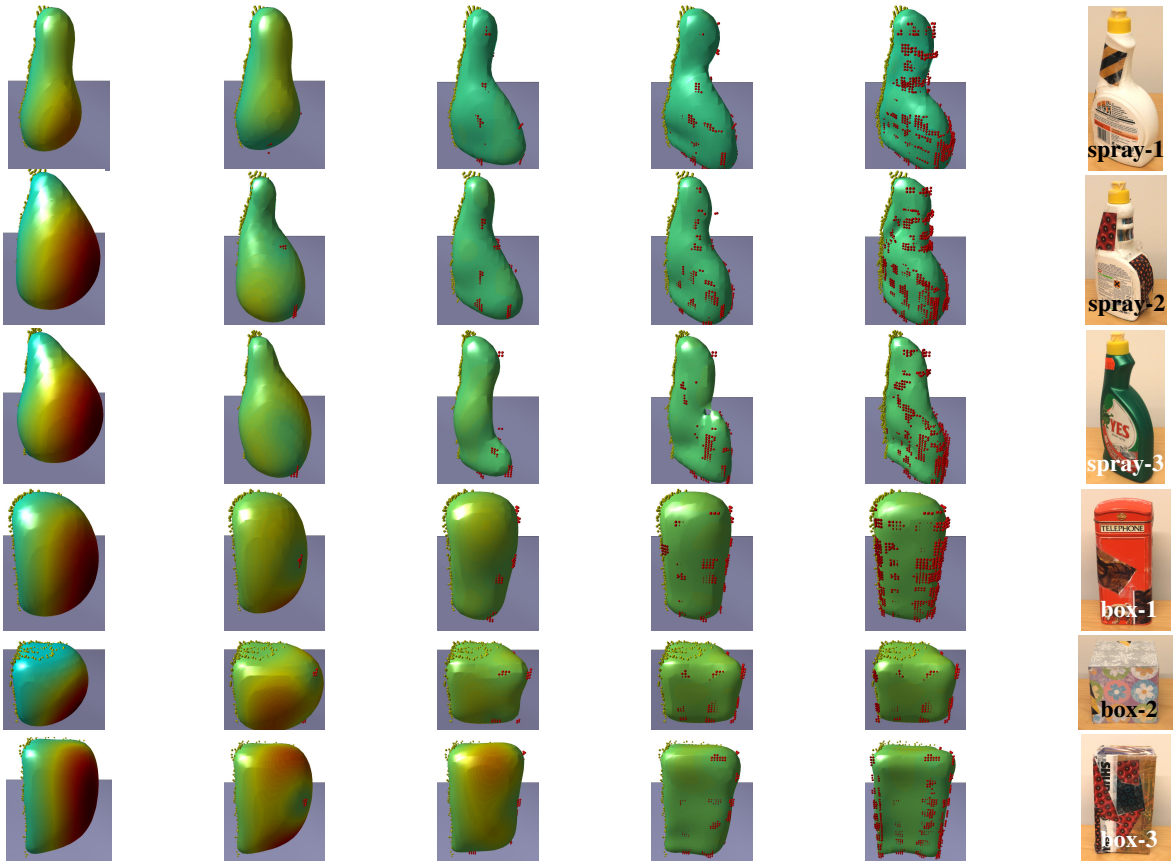


Fig. 5: Similarity matrices using up to 54 touches with columns and rows given by the objects in Fig. 4 using either curvature measures (left) or Zernike moments (right).

A. *Experimental results*

The ten objects were placed on a table-top with the Kinect camera overlooking objects from one side. To fully cover an object with tactile measurements, up to 54 touches (27 for *cyl-2* and 18 for *box-2* due to their lower heights) were performed from the side parallel to the table in a grid of 9 angles (22.5° apart) and 6 heights (spaced at a vertical distance of 2 cm) with respect to the table. The tactile measurements are illustrated as red points in Fig. 2. From the resulting implicit surface model, shape descriptors based on curvatures and Zernike moments (up to order 10) were computed and analyzed.

The convergence of the curvature based descriptors was studied by computing the distances between the descriptors after different numbers of touches and the final one. In Fig. 3 the convergence is shown using either ordered touches computed from points of maximum GP variance or touches selected randomly. The randomly generated sequences of touches were executed 10 times and then averaged. Thus the corresponding curves are slightly smoother than those of the ordered touches. The difference between the two strategies is not consistent. For most objects the difference is small and for some objects random touches are sometimes better, in particular in the beginning. The reason is because ordered pushes are computed from implicit surfaces obtained so far and at an early stage the shapes are still mostly unknown. For



(a) v. only (b) 1 t. (c) 4 t. (d) 12 t. (e) 54 t. (f) Real

Fig. 6: Evolution of object models against the number of touches for the spray bottles and the boxes. See Fig. 2 for details.

the box shaped objects, the first ordered push is usually at a corner edge on the back side of the object, when a preferable push would instead have been on one of the sides. Thus it takes another push or two for the ordered touches to catch up. From the graphs in Fig. 3, as well as from Fig. 2 and 6, it can be concluded that most changes occur during the initial ten touches.

As an illustration of the similarity between different objects, similarity matrices were computed for both curvature and Zernike based shape descriptors, which are given in Fig. 5. From the structures of the two matrices it can be concluded that while the curvatures capture classes relevant for grasping, Zernike moment does not do so to the same degree. In fact, the grouping is quite different for Zernike moments and more related to the aspect ratios of the objects than the curvatures.

This can be more easily illustrated with spectral clustering. Using the method of Ng et al. [24], we computed 2D spectral embeddings from the similarity matrices, embeddings that are shown in Fig. 4 for different numbers of touches. Here the object *cyl-3* is grouped with *box-1* and *box-3* for Zernike moments, due their similar height/width ratios. Whereas the elongated *cyl-2* and *cyl-4* are similar, they are very different from the shorter cylinder *cyl-1*. Even if *box-1* is a bit distant from *box-3* using curvature measures, the three classes can still be trivially found using e.g. k-means clustering. From the embeddings, the benefits of ordered touches can also be seen, compared to the random ones. Already after four touches, the three classes are grouped, even if it is not until 12 touches the *box-1* is closer to the other boxes than the group of spray bottles. The reason for this is that this box is thinner than the other boxes and since the GPs tend to smoothen edges, it is more like a spray bottle after too few touches. The thin plate prior tends to weaken this effect compared to a typical exponential one.

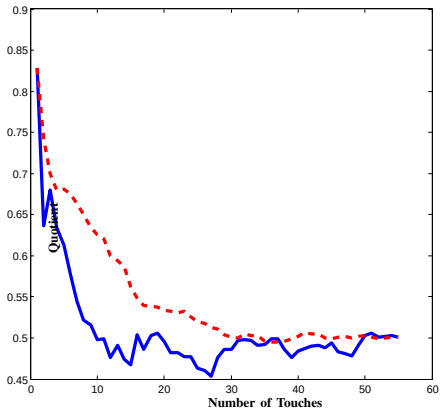


Fig. 7: Evolution of quotient between within- and between-category distances with random (dashed) and ordered (solid) touches using curvatures, as the number of touches increases.

A final illustration of the benefits of ordered touches for shape discrimination can be seen in Fig.7, where the quotient between within- and between-category distances are shown for an increasing number of touches. The quotient stabilizes after only about ten ordered touches, but for random touches at least 25 touches are required. Thus even if the benefits of ordered touches are sometimes limited when studying individual objects, they are considerable for categorization.

V. Conclusions

This paper has presented a method [2] for creation of object models from visual and tactile measurements, with the goal of later applying these for classification and manipulation. From an initial set of visual measurements, an object model is refined by touching the corresponding object on surface points predicted to be most uncertain. Given a curvature based representation of object shape, it was shown that about ten touches are sufficient for objects to be grouped into clusters relevant for manipulation. What remains to be tested in future work, however, is to what extent this representation captures manipulation affordances and can be directly used for action selection, preferably without using an intermediate step of supervised object classification.

A weakness of the current system arises from the fact that GPs have a computational cost proportional to the number of measurement points cubed. To cope with this we currently sample from the total set of points to make the problem computationally tractable. However, there are methods for sparse GPs that choose an optimal subset of points instead [25], [26], which will become a necessity in particular if measurements from additional modalities are later included.

The presented work can be extended in several directions. We intend to investigate more descriptors, other than surface curvatures and Zernike moments, that can be useful for object categorization. We will further integrate the presented approach with a pushing mechanism that can provide additional information on object affordances, e.g. rolling or sliding, potentially leading to more informed decisions about whether more measurements are needed given a particular task. Grasp planners e.g., often need information on object category [23], [27] to plan goal-directed grasps, where objects from the same category can be grasped in a similar way. Hence, we also plan to test the obtained object models for grasping tasks by using them for grasp planning.

References

[1] N. Gaissert and C. Wallraven, "Integrating visual and haptic shape information to form a multimodal perceptual space," in *IEEE World Haptics Conference (WHC)*, June 2011, pp. 451–456.

[2] M. O. Ernst and M. S. Banks, "Humans integrate visual and haptic information in a statistically optimal fashion," *Nature*, vol. 415, no. 6870, pp. 429–433, 2002.

[3] H. B. Helbig and M. O. Ernst, "Optimal integration of shape information from vision and touch," *Experimental Brain Research*, vol. 179, no. 4, pp. 595–606, 2007.

[2] A video that illustrates the method can be found at: `http://www.csc.kth.se/~celle/`

[4] J. Bohg, M. Johnson-Roberson, B. Leon, J. Felip, X. Gratal, N. Bergström, D. Kragic, and A. Morales, "Mind the gap - robotic grasping under incomplete observation," in *IEEE Int. Conf. Robotics and Automation (ICRA)*, May 2011, pp. 686–693.

[5] M. Krainin, P. Henry, X. Ren, and D. Fox, "Manipulator and object tracking for in-hand 3d object modeling," *Int. J. Robotics Research*, vol. 30, no. 11, pp. 1311–1327, September 2011.

[6] M. Meier, M. Schopfer, R. Haschke, and H. Ritter, "A probabilistic approach to tactile shape reconstruction," *IEEE Trans. Robot.*, vol. 27, no. 3, pp. 630–635, June 2011.

[7] S. Dragiev, M. Toussaint, and M. Gienger, "Gaussian process implicit surfaces for shape estimation and grasping," in *IEEE Int. Conf. Robotics and Automation (ICRA)*, May 2011, pp. 2845–2850.

[8] A. Bierbaum, M. Rambow, T. Asfour, and R. Dillmann, "A potential field approach to dexterous tactile exploration of unknown objects." in *IEEE-RAS Int. Conf. Humanoid Robots*, 2008, pp. 360–366.

[9] D. R. Faria, R. Martins, J. Lobo, and J. Dias, "Probabilistic representation of 3d object shape by in-hand exploration," in *IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, 2010, pp. 1560–1565.

[10] A. Maldonado, H. Alvarez-Heredia, and M. Beetz, "Improving robot manipulation through fingertip perception," in *IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, Vilamoura, Algarve, Portugal, 2012, pp. 2947–2954.

[11] A. Ude, D. Omrcen, and G. Cheng, "Making object learning and recognition an active process," *Int. J. Humanoid Robotics*, vol. 5, no. 2, pp. 267–286, 2008.

[12] A. Schneider, J. Sturm, C. Stachniss, M. Reisert, H. Burkhardt, and W. Burgard, "Object identification with tactile sensors using bag-of-features," in *IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, Oct. 2009, pp. 243–248.

[13] Z. A. Pezzementi, E. Plaku, C. Reyda, and G. D. Hager, "Tactile-object recognition from appearance information," *IEEE Trans. Robot.*, vol. 27, no. 3, pp. 473–487, 2011.

[14] P. Allen, "Integrating vision and touch for object recognition tasks," Dept. of Computer Science, Columbia University, Tech. Rep., 1986.

[15] M. Björkman and D. Kragic, "Active 3D segmentation through fixation of previously unseen objects," in *British Machine Vision Conference*, August 2010, pp. 119.1–119.11.

[16] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*. MIT Press, 2006.

[17] O. Williams and A. Fitzgibbon, "Gaussian process implicit surfaces," in *Gaussian Processes in Practice Workshop*, 2007.

[18] P. J. Besl and N. D. McKay, "A method for registration of 3-d shapes," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.

[19] M. Novotni and R. Klein, "Shape retrieval using 3d zernike descriptors," *Computer-Aided Design*, vol. 36, no. 11, pp. 1047–1062, 2004.

[20] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3d surface construction algorithm," in *ACM Siggraph Computer Graphics*, vol. 21, no. 4, 1987, pp. 163–169.

[21] X. Chen and F. Schmitt, "Intrinsic surface properties from surface triangulation," in *European Conference on Computer Vision (ECCV)*. Springer, 1992, pp. 739–743.

[22] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Machine Learning Research*, vol. 13, pp. 723–773, 2012.

[23] Y. Bekiroglu, D. Song, L. Wang, and D. Kragic, "A probabilistic framework for task-oriented grasp stability assessment," in *IEEE Int. Conf. Robotics and Automation (ICRA)*, Karlsruhe, Germany, 2013.

[24] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems*, vol. 2, pp. 849–856, 2002.

[25] L. Csató and M. Opper, "Sparse on-line gaussian processes," *Neural Computation*, vol. 14, no. 3, pp. 641–668, 2002.

[26] M. Titsias, "Variational learning of inducing variables in sparse gaussian processes," in *Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, 2009, pp. 567–574.

[27] M. Madry, D. Song, and D. Kragic., "From object categories to grasp transfer using probabilistic reasoning," in *IEEE Int. Conf. Robotics and Automation (ICRA)*, Minnesota, USA, 2012, pp. 1716–1723.