Estimation of Surface Geometries in Point Clouds for the Manipulation of Novel Household Objects

Siddarth Jain and Brenna Argall
Northwestern University, Evanston, IL, USA 60208
Shirley Ryan AbilityLab, Chicago, IL, USA 60611
Email: sidd@u.northwestern.edu, brenna.argall@northwestern.edu

Abstract—Real world environments typically include objects with different perceptual appearances. The classification of object point cloud clusters to surface and primitive geometries can be very useful in manipulation planning. In this work, we present and evaluate the utility of 3D surface estimation methods to efficiently classify the local and global surface geometry in the point cloud representations of household objects. We also present a Global Principal Curvature Shape Descriptor (GPCSD) to categorize object point clusters that are geometrically equivalent into similar primitive shape categories and compare its performance to Global Radius-Based Surface Descriptor (GRSD) on a large scale RGB-D object dataset and real-time data from the Kinect. Furthermore, we present a manipulation framework for grasp planning and selection on unknown objects based on geometric surface estimation. The method uses 3D sensor data directly to produce a set of grasps for the objects in a scene based on both the overall shape of the object and its local features. The framework provides kinematically feasible grasp selection and ranking for manipulation planning from the candidate grasps, which is suitable for both autonomous grasp execution and shared-control assistive teleoperation.

I. INTRODUCTION

Environment perception is crucial for the accomplishment of manipulation tasks by robots. In some sense the robot perception system can be trained on object models and thus can specialize to perceive and manipulate a set of objects. However, the great diversity in the types and classes of objects presented by unstructured environments makes it difficult to provide and use accurate and/or comprehensive models of all relevant objects that can be encountered in everyday manipulation tasks.

The set of everyday objects that a robot could encounter in its tasks is vast and diverse as household objects vary in their size, color, texture and shape. However, there exist certain regularities that can be exploited to inform and help robot perception in order to act on novel objects. In particular, the 3D surface geometries of objects contain characteristic information that generalizes over a large variety of objects and thus can provide useful information for recognition as well as everyday manipulation.

Geometrical analysis of a point’s neighborhood can provide discriminative information about the local surface type. The local surface geometries can be characterized by basic surface types in the 3D point clouds of objects, such as corner, cylindrical, edge, planar and spherical. The identification of such local surface types in object point clouds can be utilized to reason about the shape and global surface characteristics of the object—without the need to recognize the specific object instance type or category.

Such local and global surface geometry information can be used in manipulation planning for the robots. Thus, the addition of this semantic information to the 3D representation of object point clouds can enable the robot perception system to operate in a more general way—enabling the robots to perceive and manipulate unseen objects with more flexibility and reliability.

In this paper, we investigate methods for the estimation of local and global surface geometries in the 3D point clouds of household objects. We furthermore present a framework for perception and manipulation based on such geometric characteristics, without the need to recognize a specific object category or instance.

In particular, the utility of Principal Curvatures (PC) and the Radius-based Surface Descriptor (RSD) are examined for their effectiveness in local surface estimation from point clouds of household objects. A global principal curvature based shape descriptor (GPCSD) is generated based on the local PC values. The computation is similar to GRSD, but we use voxel-based labeling based on Principal Curvatures (instead of the RSD). The performance of global descriptors GPCSD and GRSD, in concert with artificial neural networks, in the classification of geometric characteristics of objects are evaluated on a large object dataset and also on real-time sensor data. Furthermore, the utility of the classified shape characteristics are presented in a framework for the manipulation of household objects. The framework provides kinematically feasible grasp selection and ranking for manipulation planning, which is suitable for both autonomous grasp execution and shared-control assistive teleoperation.

The remainder of the paper is organized as follows. Related work is described in Section II. Section III explains the methods for surface estimation. Section IV presents the classification and evaluation of global descriptors. A manipulation framework utilizing shape characteristics is presented in Section V. In Section VI, we present the conclusions.

II. RELATED WORK

In recent years, different descriptors for point clouds have been developed with the goal of addressing problems related
to object recognition and 3D registration. Point Feature Histograms (PFH) [17] were proposed as an informative pose-invariant feature for geometrical surface description. Fast Point Feature Histograms (FPFH) [17] were later deployed to reduce the computational complexity of PFH for real-time applications. The Radius-Based Surface Descriptor (RSD) was proposed to describe the geometry of points in a local neighborhood using surface normals by estimating radial relationships. The utility of principal curvatures estimation that relies on surface normals is shown [17] to represent geometric 3D edges and region growing segmentation in point clouds. These features perform local surface categorization, where each point is associated with a descriptor able to describe the local geometry of that point. By contrast, global feature histograms (for example, GRSD [14]) are high-dimensional representations that capture the notion of objects as whole for the purpose of object recognition. For a more detailed information on features we refer the readers to [1].

The utility of surface classification using local features is mostly demonstrated on data obtained with laser scanners [17], which are less prone to noise. A comparative evaluation of FPFH, RSD and PC local descriptors on 8 indoor scenes captured with a less accurate 3D sensor (ASUS Xtion PRO LIVE) finds FPFH to be better for complex shapes but to have trouble dealing with noisy data. RSD and PC perform similarly, but PC is found to be more robust and smooth for local scene classification.

In this work we consider the local surface labels for geometric classification of points, but more importantly we evaluate the global feature representations from the local surface labels to determine the overall shape characteristics. Furthermore, the majority of the existing evaluations tackle the less general problem of specific object recognition using descriptors [2]. We intend to perform a more general classification of objects into global primitive shapes, which can be used to inform perception for grasp generation, selection and execution on novel objects.

Grasp planning has been widely studied in the literature. For a more detailed review of the field, we refer the readers to survey articles [5] [13]. Similar to this work, there exist approaches that explore shape primitives and depth data in grasp planning. Miller et al. [15] present the use of shape primitives for object grasping and their approach relies on having known object models based on which the grasps can be analyzed in simulation. Jain and Kemp [9] plan overhead grasps using a set of simple heuristics. Ten Pas and Platt [20] propose a geometry based method for generating grasp candidates by sampling from a large set of grasp hypotheses. Hsiao et al. [8] search for orthogonal grasps by fitting a bounding box primitive on object point clouds and produce top and side grasps based on a small set of simple feature weights. Jain and Argall [10] fit primitive shapes to cylindrical and spherical objects in addition to the box shapes for planning top, side and pinch grasps, using surface variation [17] to determine the best fit primitive shape.

In this work, we evaluate and explore more robust surface geometry based approaches for determination of object shape characteristics and present a manipulation planning framework for grasp selection, ranking and execution on novel objects. The determination of shape characteristics is a three step process: (1) computation of local features, (2) aggregating the local features across the entire object (called a global descriptor) and (3) classifying the object, based on the global descriptor, into a primitive shape type. Next, based on the identified primitive shape and estimated pose, grasps are generated on objects, which are further ranked and selected in manipulation planning.

III. SURFACE GEOMETRY ESTIMATION

The following section gives an overview of surface geometry estimation methods for point clouds. First, we discuss local surface geometry estimation approaches (such as RSD and PC local features), and then a global descriptor based approach to determine the overall shape characteristics of object clusters. These shape characteristics then will be used (in Section IV) to classify the object into a primitive shape class.

Methods based on the Radius-based Surface Descriptor (RSD) and Principal Curvatures (PC) have shown good performance on classification of local geometry in point cloud scenes (Sec. II). Thus, we select these local features and compare their performance on point clouds of household objects.

RSD [14] describes the geometry of points in a local neighborhood by estimating their radial relationships. The radius estimation is performed by assuming each point pair to lie on a sphere. By exploiting the relation of the distance $d$ between the points and the angle $\alpha$ between the two point normals, the radius $r$ is estimated,

$$\sqrt{2r \sqrt{1 - \cos(\alpha)}} = d,$$

$$r \approx d / \alpha.$$

The radius approaches infinity for a planar surface and takes on increasingly lower values for surfaces with a higher curvature. The RSD feature for a point, based on its nearest neighbors, consists of a minimum radius and maximum curvature radius $[r_{min}, r_{max}]$ taken from all the point-neighbor spheres.

PC [17] describes a point’s local surface geometry from the eigenvectors and eigenvalues of the principal surface curvatures on that point. For a query point $p_q \in \mathbb{R}^3$, all normals $n_j \in \mathbb{R}^3$, $j = 1...N$, of the set of its $N$ neighborhood points are projected onto the tangent plane of the surface defined by the normal $n_q$ at the query point. The centroid of the projected points $\bar{p} \in \mathbb{R}^3$ and the covariance matrix $C$ from all projections are computed as,

$$\bar{p} = \frac{\sum_{j=1}^{N} p_j}{N},$$

$$C = \frac{1}{N} \sum_{j=1}^{N} (p_j - \bar{p}) \cdot (p_j - \bar{p})^T,$$

where

$$p_j = (I - n_q \cdot n_q^T) \cdot n_j,$$

and $I$ is a $3 \times 3$ identity matrix.
Principal Components Analysis (PCA) [12] is then performed on the point normals of the surface patch in the tangent plane of the given point normal, to find the eigenvalues. The eigenvalues \( \lambda_k \in \mathbb{R} \) and eigenvectors \( \mathbf{v}_k \in \mathbb{R}^3 \) are defined by the following relation and form an orthogonal frame that corresponds to the principal components of the set of neighborhood points:

\[
C \cdot \mathbf{v}_k = \lambda_k \cdot \mathbf{v}_k, \quad k \in \{0, 1, 2\}
\]

where \( \lambda_3 \) corresponds to the maximum curvature and \( \lambda_2 \) to the minimum curvature denoted as \( p_{c_{\text{max}}} \) and \( p_{c_{\text{min}}} \), respectively. The PC feature consists of these minimum and maximum curvatures \( [p_{c_{\text{max}}}, p_{c_{\text{min}}}] \) as well as the principal curvature direction \( \mathbf{v}_3 \), the normalized eigenvector of \( p_{c_{\text{max}}} \) and \( p_{c_{\text{min}}} \).

Using the RSD and PC values, a geometric class label can be assigned to each point in the object cluster. To determine the class label, thresholds on the feature values of the RSD descriptor are used to categorize local surfaces into classes planar, cylindrical, spherical, edge, corner or noisy. In particular, for each voxel with a width of 2.5cm, we compute the minimum and maximum curvature radii \( [r_{\text{min}}, r_{\text{max}}] \) and label the voxel surface by successive checks of the RSD radii in order to categorize it into one of the geometric classes. This is a simple and fast way to categorize local surfaces, and it divides the feature space formed by the radii well enough for the computation of a global feature such as GRSD.

We apply a rule-set of classification thresholds (empirically determined) in order to classify local geometries in the point clouds of household objects. Since RSD and PC are based on a similar geometrical approach (highest and lowest curvature), this thresholds technique is applicable to classify the PC values \( [p_{c_{\text{max}}}, p_{c_{\text{min}}}] \).

A global representation of object type then is made from the local features (geometries). Specifically, for the RSD feature, once all voxels are annotated locally with a geometric class, the Global Radius-based Surface Descriptor (GRSD) [14] is computed, which produces a unique signature for a given object cluster based on the RSD local feature values. GRSD describes the transitions between different surface types (and free space) for an object. Note that in this work we count the transitions between surface types between the occupied voxels instead of along lines between occupied voxels. Each bin counts a transition between a pair of specific surface types, or between a specific surface type and free space. By using five local surface types (planar, cylindrical, spherical, edge and noisy/corner) and considering free space, the number of GRSD histogram bins is 21.

Following a similar approach, for the PC feature, we introduce a global histogram that counts transitions between the voxelized PC labels within an object cluster and also include bins for the number of voxels labeled as planar, cylindrical and spherical (empirically we found this to slightly increase the shape classification performance), making the bin size 24. We call this new assembled descriptor the Global Principal Curvature Shape Descriptor (GPCSD). Figure 1 shows the surface estimation on a cereal box object cluster.

IV. GLOBAL SHAPE CLASSIFICATION AND EVALUATION

As noted above, we are interested in the global feature representations for objects and only compute local surface labels in order to aid in this global classification. We consider three primitive geometry classes to represent many objects that a robot could perceive and manipulate in household environments: box-shaped, cylindrical and spherical.

We select 12 object categories from the large scale RGBD-Washington dataset [13] to evaluate the mapping from global descriptors of object type to the primitive shape classes. For each category, there are four separate object instances in the dataset and each instance has more than 600 views of the object from different angles. Figure 2 shows an instance of each object for all categories along with the ground truth class labels. Note that RGB values from the dataset are only represented for illustration and are not utilized in the descriptor computations. GRSD and GPCSD descriptors are computed for all instances of each object in the 12 categories.
Fig. 3. Normalized confusion matrices for GRSD and GPCSD classification using ANN models on the test set objects. Class labels are Box-shaped (B), Cylindrical (C) and Spherical (S).

An artificial neural network (ANN) is trained on three instances of each object category in the dataset, with one instance being randomly picked to leave out from each category in order to form the test set. The ANN structure has 21 input nodes for the GRSD descriptor (one node for each histogram bin) and 24 for the GPCSD descriptor, one hidden layer with 10 nodes and three output nodes for the class labels. (Note that multiple network structures were tested to prevent overfitting before arriving at this structure.)

Both models perform well on the unseen test set objects and the classification accuracy of the learned GRSD model is 94.93% and that of the GPCSD model is 91.02%. The normalized confusion matrices are shown in Figure 3. The main performance difference between the two approaches appears to be the occasional mislabelling of spherical as cylindrical (1% for GRSD and 8% for GPCSD). These global results are in contrast to findings where PC performs better than RSD for local surface classification on point cloud scenes.

We further test the performance of the learned models on live data from a Kinect sensor. Each object instance from the test set is presented one by one in front of the Kinect at distances of 100 cm, 125 cm and 150 cm from the sensor. We use the same trained models obtained from the RGB-D dataset to query the segmented object clusters. Both models perform well on the live test data, but we find the GRSD model predictions comparatively more robust to distance variations.

V. MANIPULATION PLANNING

We now present a framework that uses shape estimation (Section IV) from real-time sensor data for grasp planning on novel objects. Our framework is overviewed in Figure 4. The first objective is to find a combination of gripper orientations and positions relative to a given object using the sensory data available to the robot. The sensory input to our framework is 3D point cloud data of the scene from a Kinect RGB-D sensor.

We first identify the parts of the input point cloud that are likely to belong to a single object (clusters). Two simplifying assumptions are made: the objects are sitting on a flat surface suitable for manipulation (such as a table), and the minimum distance between two objects in a scene is at least 3cm. We compute a planar fit using Random Sample Consensus (RANSAC)1 to generate a model hypotheses for the flat surfaces and extract the dominant planar surface that provides support for the objects. We then find the individual object clusters $\mathcal{O}$ in the scene by performing Euclidean clustering.

1Our implementation uses the RANSAC implementation provided with Point Cloud Library (PCL).
Next, we compute the global surface descriptor (GRSD) for each segmented object cluster and query the learned ANN model to classify the cluster to a primitive shape category. A model hypothesis of the geometric model parameters for the identified shape category is generated using RANSAC. The obtained model then is refined using a linear least-squares fit for the plane and nonlinear Levenberg-Marquardt optimization for the cylindrical and spherical primitives.

The framework generates, for every observed object point cloud cluster \( O \), a set of possible grasps \( G \). A study involving grasping by human users \([4]\) and pilot experiments of assistive robotic arm teleoperation \([10]\) show that humans tend to grasp with wrist orientations that are orthogonal to the object’s principal axis and its perpendiculars—the hypothesis being that such grasps are more stable. We generate Top, Side and/or Pinch grasps for each object cluster around the modeled shape centroid, by using an established grasp generation heuristics \([10]\). Note that here we represent each grasp \( g \in G \) as a Cartesian position and an approach direction quaternion, in the robot base frame. The grasps can readily be used as a motion planning goal for a path planner, in order to execute the grasp on the desired object.

Figure 5 shows examples of the grasps generated by our method on point clouds of household objects. The primitive shape is estimated for segmented object clusters using the GRSD feature neural network model, and the grasps are generated according to the heuristics described above. Figure 6 shows the performance of our approach on novel objects. The framework is able to identify shape primitives, estimate their pose and generate grasps for a variety of different object types and shapes. Note that these object types do not exist in the set used to train the ANN model—they are not just new instances of a known object type, but rather are entirely new object types.

For manipulation planning, it is desirable to first evaluate grasps that are more likely to be physically reachable by the robotic arm. Since there is likely more than one solution in \( G \), it is preferable to choose grasps such that the planning solution is quickly found. Furthermore, grasp approach directions that minimize the distance between the end-effector and the object centroid are preferred.

We compute metrics with the aim to prefer grasps that result in shorter planning times by incorporating the robot arm kinematics. For this consideration, we compute the condition
number $\kappa$ as the ratio of the smallest and the largest singular value of the Jacobian at configuration $q$,

$$\kappa = \frac{\sigma_{\text{min}, q}}{\sigma_{\text{max}, q}}.$$ 

If the value of the condition number $\kappa$ is close to one, the manipulability ellipsoid is isotropic—indicating that the manipulator has the same motion capability in all task space directions. On the other hand, small values of $\kappa$ indicate closeness to a singular configuration. Therefore, by using $\kappa$ as a quality measure, grasp configurations far away from a singular one can be preferred (Figure 7). For a grasp for which multiple candidate arm configurations $q$ might exist, the average manipulability for a set of collision-free arm configurations $q_i$ that achieve $g$ instead is computed,

$$\kappa = \frac{1}{N} \sum_{i=1}^{N} \frac{\sigma_{\text{min}, q_i}}{\sigma_{\text{max}, q_i}},$$

where $\sigma_{\text{min}, q_i}$ is the smallest singular value and $\sigma_{\text{max}, q_i}$ is the largest singular value of the Jacobian at configuration $q_i$. Note that additionally only collision-free grasps that reach the object are considered by performing environment check.

If the robot is operating with full autonomy, the grasp $g^*$ that maximizes $\kappa$,

$$g^* = \arg \max_{g \in G} \kappa(q_g)$$

is selected, where $q_g$ is the arm configuration at grasp $g$ (obtained via inverse kinematics). Sampling-based planners from the OMPL [19] library then are utilized to plan optimized motions for grasping and manipulating objects.

Another target application domain for this work is to provide assistance in a shared-control framework. In shared control teleoperation [6, 7, 11, 16], both the human user and the robotics autonomy control the robot, in order to achieve the user’s desired goal. The grasps generated by our framework are suitable candidates for the potential user goals and can thus be used readily in the shared control framework.

One key challenge in shared teleoperation is to infer the user’s desired goal in order to inform the robotics autonomy. In our approach the user-preferred grasp $g^*$ out of possible choices $G$ is inferred based on: (1) the Euclidean distance between the robot end-effector position and the grasp position, and (2) the end-effector orientation alignment with the grasp pose. Once the end-effector is in the vicinity of an object, the nearest and best aligned grasp pose then is selected. Figure 8 shows examples of grasp intent inference in shared-autonomy.

VI. Conclusion

In this paper, we have explored methods for local and global surface estimation in point clouds of household objects. A new global descriptor, GPCSD, was introduced and its performance was compared to the GRSD descriptor for the estimation of global shape from object clusters. Evaluation was performed on a large scale RGB-D dataset and also on real-time data from the Kinect using a neural network to classify simple shape primitives. Both descriptors performed well, but GRSD was found to be more robust to distance variations. Furthermore, we presented a manipulation framework for grasp selection and execution on novel objects. The grasp selection method used the 3D sensor data directly to determine a ranked set of grasps for objects in a scene based on the overall shape of the object and its local features. Our results showed that the presented framework is capable of correctly estimating the shape of unseen household objects and producing multiple candidate grasps which can be ranked and selected based on their kinematic feasibility and also on inferred intent for shared-control teleoperation. Our next steps will be to evaluate the manipulation framework in a shared-control system that includes an end-user study.
ACKNOWLEDGMENT

Research reported in this publication was supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) and National Institute of Child and Human Development (NICHD) under award number R01EB019335. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES


