2014 IEEE International Conference on Robotics & Automation (ICRA)
Hong Kong Convention and Exhibition Center
May 31 - June 7, 2014. Hong Kong, China

Single Image 3D Object Detection and Pose Estimation for Grasping

Menglong Zhu[1], Konstantinos G. Derpanis[2], Yinfei Yang[1], Samarth Brahmbhatt[1]
Mabel Zhang[1], Cody Phillips[1], Matthieu Lecce[1] and Kostas Daniilidis[1]

***Abstract*— We present a novel approach for detecting objects and estimating their 3D pose in single images of cluttered scenes. Objects are given in terms of 3D models without accompanying texture cues. A deformable parts-based model is trained on clusters of silhouettes of similar poses and produces hypotheses about possible object locations at test time. Objects are simultaneously segmented and verified inside each hypothesis bounding region by selecting the set of superpixels whose collective shape matches the model silhouette. A final iteration on the 6-DOF object pose minimizes the distance between the selected image contours and the actual projection of the 3D model. We demonstrate successful grasps using our detection and pose estimate with a PR2 robot. Extensive evaluation with a novel ground truth dataset shows the considerable benefit of using shape-driven cues for detecting objects in heavily cluttered scenes.**

I. Introduction

In this paper, we address the problem of a robot grasping 3D objects of known 3D shape from their projections in single images of cluttered scenes. In the context of object grasping and manipulation, object recognition has always been defined as simultaneous detection and segmentation in the 2D image and 3D localization. 3D object recognition has experienced a revived interest in both the robotics and computer vision communities with RGB-D sensors having simplified the foreground-background segmentation problem. Nevertheless, difficulties remain as such sensors cannot generally be used in outdoor environments yet.

The goal of this paper is to detect and localize objects in single view RGB images of environments containing arbitrary ambient illumination and substantial clutter for the purpose of autonomous grasping. Objects can be of arbitrary color and interior texture and, thus, we assume knowledge of only their 3D model without any appearance/texture information. Using 3D models makes an object detector immune to intra-class texture variations.

We further abstract the 3D model by only using its 2D silhouette and thus detection is driven by the shape of the 3D object's projected occluding boundary. Object silhouettes with corresponding viewpoints that are tightly clustered on the viewsphere are used as positive exemplars to train the state-of-the-art Deformable Parts Model (DPM) discriminative classifier [1]. We term this shape-aware version S-DPM.

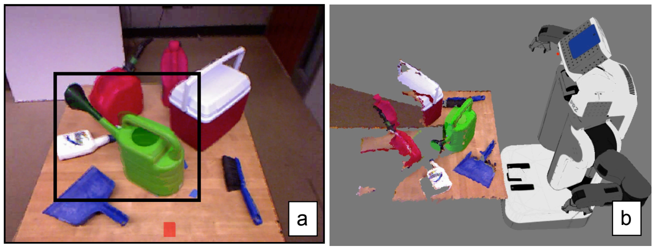

Fig. 1: Demonstration of the proposed approach on a PR2 robot platform. a) Single view input image, with the object of interest highlighted with a black rectangle. b) Object model (in green) is projected with the estimated pose in 3D, ready for grasping. The Kinect point cloud is shown for the purpose of visualization.

This detector simultaneously detects the object and coarsely estimates the object′s pose. The focus of the current paper is on instance-based rather than category-based object detection and localization; however, our approach can be extended to multiple instance category recognition since S-DPM is agnostic to whether the positive exemplars are multiple poses from a single instance (as considered in the current paper) or multiple poses from multiple instances.

We propose to use an S-DPM classifier as a first high recall step yielding several bounding box hypotheses. Given these hypotheses, we solve for segmentation and localization simultaneously. After over-segmenting the hypothesis region into superpixels, we select the superpixels that best match a model boundary using a shape-based descriptor, the chordiogram [2]. A chordiogram-based matching distance is used to compute the foreground segment and rerank the hypotheses. Finally, using the full 3D model we estimate all 6-DOF of the object by efficiently iterating on the pose and computing matches using dynamic programming.

Our approach advances the state-of-the-art as follows:

- In terms of assumptions, our approach is among the few in the literature that can detect 3D objects in single images of cluttered scenes independent of their appearance.
- It combines the high recall of an existing discriminative classifier with the high precision of a holistic shape descriptor achieving a simultaneous segmentation and detection reranking.
- Due to the segmentation, it selects the correct image contours to use for 3D pose refinement, a task that was previously only possible with stereo or depth sensors.

[1] These authors are with the GRASP Laboratory, Department of Computer Information Science, University of Pennsylvania, 3330 Walnut Street, Philadelphia, PA USA. {menglong, yinfeiy, samarthb, zmen, codyp, mlecce, kostas}@cis.upenn.edu

[2] Konstantinos G. Derpanis is with the Department of Computer Science, Ryerson University, 245 Church Street, Toronto, Ontario Canada. kosta@scs.ryerson.ca

978-1-4799-3685-4/14/$31.00 ©2014 IEEE

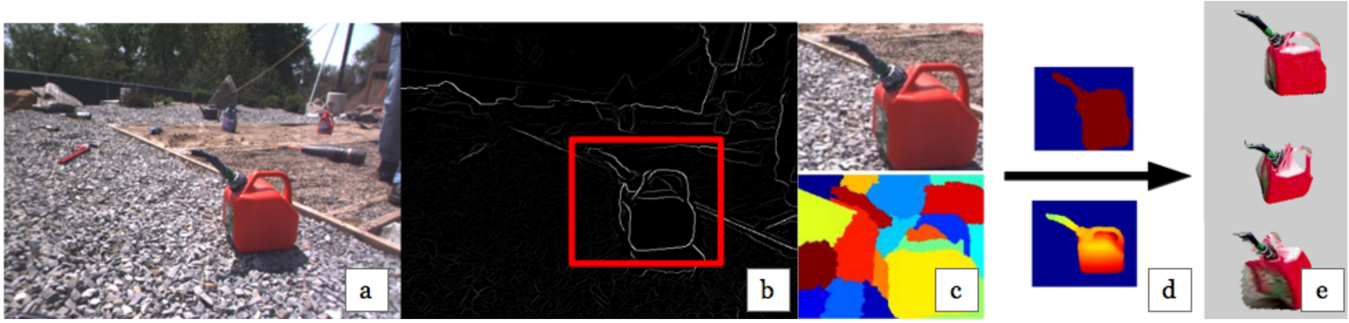

Fig. 2: Overview of the proposed approach. From left-to-right: a) The input image. b) S-DPM inferences on the gPb contour image yielding an object detection hypothesis. c) The hypothesis bounding box (red) is segmented into superpixels. d) The set of superpixels with the closest chordiogram distance to the model silhouette is selected. Pose is iteratively refined such that the model projection aligns well with the foreground mask silhouette. e) To visualize the pose accuracy, the side of the 3D model facing the camera is textured with the corresponding 2D pixel color; three textured synthetic views of the final pose estimate are shown.

In the video supplement, we demonstrate our approach with a (PR2) robot grasping 3D objects on a cluttered table based on a single view RGB image. Figure 8 shows an example of the process. We report 3D pose accuracy by comparing the estimated pose rendered by the proposed approach with a ground truth point cloud recovered with a RGB-D sensor. Such grasping capability with accurate pose is crucial for robot operation, where popular RGB-D sensors cannot be used (e.g., outdoors) and stereo sensors are challenged by the uniformity of the object's appearance within their boundary. We also document an extensive evaluation on outdoor imagery with diverse backgrounds. The dataset contains a set of 3D object models, annotated single-view imagery of heavily cluttered outdoor scenes[1], and indoor imagery of cluttered tabletops in RGB-D images.

II. Related Work

Geometry-based object recognition arguably outdates appearance-based approaches. A major advantage of these approaches is their invariance to material properties, viewpoint and illumination. We first survey approaches that use a 3D model, either synthetic or obtained from 3D reconstruction. Next, we describe approaches using multiple view exemplars annotated with their pose. We close with a brief description of 2D shape-based approaches and approaches applied to RGB-D test data.

Early approaches based on using explicit 3D models are summarized in Grimson's book [3] and focus on efficient techniques for voting in pose space. Horaud [4] investigated object recognition under perspective projection using a constructive algorithm for objects that contain straight contours and planar faces. Hausler [5] derived an analytical method for alignment under perspective projection using the Hough transform and global geometric constraints. Aspect graphs in their strict mathematical definition (each node sees the same set of singularities) were not considered practical enough for recognition tasks but the notion of sampling of the view-space for the purpose of recognition was introduced again in [6] which were applied in single images with no background. A Bayesian method for 3D reconstruction from a single image was proposed based on the contours of objects with sharp surface intersections [7]. Sethi et al. [8] compute global invariant signatures for each object from its silhouette under weak perspective projection. This approach was later extended [9] to perspective projection by sampling a large set of epipoles for each image to account for a range of potential viewpoints. Liebelt et al. work with a view space of rendered models in [10] and a generative geometry representation is developed in [11]. Villamizar et al. [12] use a shared feature database that creates pose hypotheses verified by a Random Fern pose specific classifier. In [13], a 3D point cloud model is extracted from multiple view exemplars for clustering pose specific appearance features. Others extend deformable part models to combine viewpoint estimates and 3D parts consistent across viewpoints, e.g., [14]. In [15], a novel combination of local and global geometric cues was used to filter 2D image to 3D model correspondences.

Others have pursued approaches that not only segment the object and estimate the 3D pose but also adjusts the 3D shape of the object model. For instance, Gaussian Process Latent Variable Models were used for the dimensionality reduction of the manifold of shapes and a two-step iteration optimizes over shape and pose, respectively [16]. The drawback of these approaches is that in the case of scene clutter they do not consider the selection of image contours. Further, in some cases tracking is used for finding the correct shape. This limits applicability to the analysis of image sequences, rather than a single image, as is the focus in the current paper.

Our approach resembles early proposals that avoid appearance cues and uses only the silhouette boundary, e.g., [6]. None of the above or the exemplar-based approaches surveyed below address the amount of clutter considered here

[1] The annotated dataset and 3D models are available at the project page: http://www.seas.upenn.edu/~menglong/outdoor-3d-objects.html

and in most cases the object of interest occupies a significant portion of the field of view.

Early view exemplar-based approaches typically assume an orthographic projection model that simplifies the analysis. Ullman [17] represented a 3D object by a linear combination of a small number of images enabling an alignment of the unknown object with a model by computing the coefficients of the linear combination, and, thus, reducing the problem to 2D. In [18], this approach was generalized to objects bounded by smooth surfaces, under orthographic projection, based on the estimation of curvature from three or five images. Much of the multiview object detector work based on discrete 2D views (e.g., [19]) has been founded on successful approaches to single view object detection, e.g., [1]. Savarese and Fei-Fei [20] presented an approach for object categorization that combines appearance-based descriptors including the canonical view for each part, and transformations between parts. This approach reasons about 3D surfaces based on image appearance features. In [21], detection is achieved simultaneously with contour and pose selection using convex relaxation. Hsiao et al. [22] also use exemplars for feature correspondences and show that ambiguity should be resolved during hypothesis testing and not at the matching phase. A drawback of these approaches is their reliance on discriminative texture-based features that are hardly present for the types of textureless objects considered in the current paper.

As far as RGB-D training and test examples are concerned, the most general and representative approach is [23]. Here, an object-pose tree structure was proposed that simultaneously detects and selects the correct object category and instance, and refines the pose. In [24], a viewpoint feature histogram is proposed for detection and pose estimation. Several similar representations are now available in the Point Cloud Library (PCL) [25]. We will not delve here into approaches that extract the target objects during scene parsing in RGB-D images but refer the reader to [26] and the citations therein.

The 2D-shape descriptor, chordiogram [2], we use belongs to approaches based on the optimal assembly of image regions. Given an over-segmented image (i.e., superpixels), these approaches determine a subset of spatially contiguous regions whose collective shape [2] or appearance [27] features optimize a particular similarity measure with respect to a given object model. An appealing property of region-based methods is that they specify the image domain where the object-related features are computed and thus avoid contaminating objected-related measurements from background clutter.

III. Technical approach

An overview of the components of our approach is shown in Fig. 2. 3D models are acquired using a low-cost depth sensor (Sec. III-A). To detect an object robustly based *only* on shape information, the gPb contour detector [28] is applied to the RGB input imagery (Sec. III-B). Detected contours are fed into a parts-based object detector trained on model silhouettes (Sec. III-C). Detection hypotheses are over-segmented and shape verification simultaneously computes the foreground segments and reranks the hypotheses (Sec. III-E). Section III-D describes the shape descriptor used for shape verification. The obtained object mask enables the application of an iterative 3D pose refinement algorithm to accurately recover the 6-DOF object pose based on the initial coarse pose estimate rendered by the object detector (Sec. III-F).

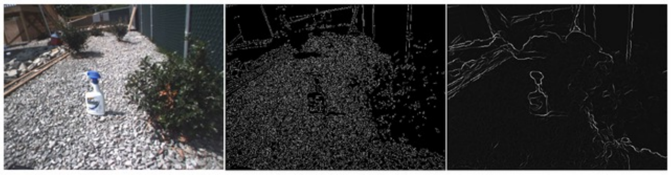

Fig. 3: Comparison of the two edge detection results on same image. (left-to-right) Input image, Canny edge and gPb, respectively.

A. 3D model acquisition and rendering

3D CAD models have been shown to be very useful for object detection and pose estimation both in 2D images and 3D point clouds. We utilize a low-cost RGB-D depth sensor and a dense surface reconstruction algorithm, KinectFusion [29], to efficiently reconstruct 3D object models from the depth measurements of real objects. The 3D object model is acquired on a turntable with the camera pointing in a fixed position. After the model is reconstructed with the scene, we manually remove the background and fill holes in the model.

To render object silhouettes from arbitrary poses, we synthesize a virtual camera at discretized viewpoints around the object center at a fixed distance. Each viewpoint is parameterized by the azimuth, a, elevation, e, and distance, d, of the camera relative to the object. Viewpoints are uniformly sampled on the viewsphere at a fixed distance and at every ten degrees for both the azimuth and elevation.

B. Image feature

Our approach to shape-based recognition benefits from recent advances in image contour detection. In unconstrained natural environments, the Canny edge detector [30] generally responds uniformly to both object occlusion boundaries and texture. One can falsely piece together the silhouette of a target object from a dense set of edge pixels. The state-of-the-art contour detection algorithm gPb [28] computes the likelihood of each pixel being an object contour and thus suppresses many edges due to texture/clutter. Figure 3 shows an example of Canny edge detection and gPb on the same input image. Compared to Canny edges, gPb suppresses ubiquitous edge responses from background clutter.

Given detected contours in the image, we seek to localize the subset of contour pixels that best represent the object silhouette. We will show that for cluttered scenes, discriminative power is essential to achieve high recall with desired precision.

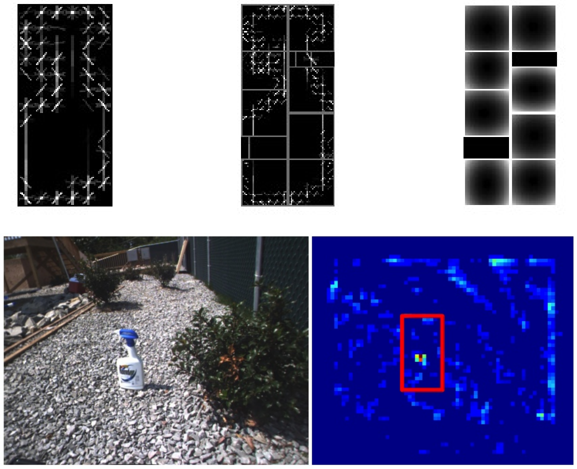

Fig. 4: Spray bottle detection using S-DPM. (first row, left-to-right) Root appearance model, part appearance models centered at their respective anchor points and the quadratic deformation cost; brighter regions indicate larger penalty cost. (second row) Input image and detection response map of the spray-bottle; red, yellow and blue indicate large, intermediate and low detection responses, respectively.

C. Object detection

The Deformable Parts Model (DPM) [1] is arguably the most successful object detector to-date. DPM is a star-structured conditional random field (CRF), with a root part, F_0, capturing the holistic appearance of the object and several parts $(P_0, \ldots, P_n)$ connected to the root where $P_i = (F_i, v_i, s_i, a_i, b_i)$. Each model part has a default relative position (the anchor point), v_i, with respect to the root position. Parts are also allowed to translate around the anchor point with a quadratic offset distance penalty, parameterized by the coefficients a_i and b_i. The anchor points are learned from the training data and the scales of the root and parts, s_i, are fixed. The detection score is defined as:

$$\sum_{i=0}^{n} F_i \cdot \phi(H, p_i) - \sum_{i=1}^{n} a_i \cdot (\tilde{x_i}, \tilde{y_i}) + b_i \cdot (\tilde{x_i}^2, \tilde{y_i}^2), \qquad (1)$$

where $\phi(H, p_i)$ is the histogram of gradients (HOG) [31] feature extracted at image location p_i, and $(\tilde{x_i}, \tilde{y_i})$ is the offset to the part anchor point with respect to the root position p_0. At test time, the root and part model weights are each separately convolved with the HOG feature of the input image. Due to the star structure of the model, maximizing the above score function at each image location can be computed efficiently via dynamic programming. To deal with intra-class variation, DPM is generalized by composing several components, each trained on a subset of training instances of similar aspect ratio. We refer to [1] for more details.

To simultaneously detect an object and coarsely estimate its pose from the edge map using only model silhouette shape information, we train a shape-aware modified version of DPM, which we term S-DPM. Each component of the learned S-DPM corresponds to a coarse pose of the object. More specifically, the silhouettes of the synthetic views of the object are clustered into 16 discrete poses by grouping nearby viewpoints. A S-DPM component is trained based on the silhouettes of a coarse pose cluster used as positive training data and silhouettes of other poses and objects and random background edges used as negatives. Figure 4 shows an example of a trained spray bottle model. During inference, each of the model components are evaluated on the input contour imagery and the hypotheses with a detection score above a threshold are retained. Detections of different components are combined via non-maximum suppression. This step retains high scoring detections and filters out neighboring lower scoring ones whose corresponding 2D bounding box overlaps with that of the local maximum by greater than 50% (PASCAL criteria [32]). The coarse pose of the object is determined by the maximum scoring component at each image location.

D. Shape descriptor

We represent the holistic shape of each S-DPM detected object with the chordiogram descriptor [2]. Given the object silhouette, this representation captures the distribution of geometric relationships (relative location and normals) between pairs of boundary edges, termed chords. Formally, the chordiogram is a K-dimensional histogram of all chord features on the boundary of a segmented object. A chord is a pair of points (p, q) on the boundary points. Chord feature $d_{pq} = (l_{pq}, \psi_{pq}, \theta_p - \psi_{pq}, \theta_q - \psi_{pq})^\top$ is defined by chord vector length l_{pq}, orientation ψ_{pq} and normals θ_p and θ_q of the object boundary at p and q. The edge normal direction points towards the segment interior to distinguish the same edge with different foreground selection of bordering superpixels. Figure 5 shows two examples of chord features and their corresponding chordiogram feature bins when the bordering foreground superpixels differ. The chordiogram is translation invariant since it only relates the relative position of boundary pixels rather than the absolute position in the image.

E. Shape verification for silhouette extraction

We use the chordiogram descriptor for two tasks: (i) to recover the object foreground mask (i.e., the silhouette) for accurate 3D pose estimation and (ii) to improve detection precision and recall by verifying that the shape of the foreground segmentation resembles the model mask.

The fact that S-DPM operates on HOG features provides flexibility in dealing with contour extraction measurement noise and local shape variance due to pose variation. However, S-DPM only outputs the detections of the object hypotheses rather than the exact location of the object contour. Even in the object hypothesis windows, the subset of edge pixels that correspond to the object silhouette is not apparent. In addition, contour-based object detection in cluttered scenes is susceptible to false detections caused by piecing together irrelevant contours.

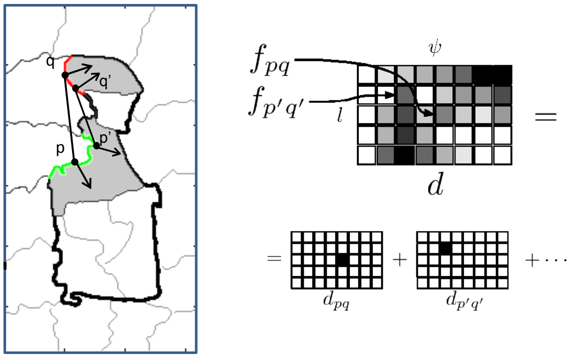

Fig. 5: Chordiogram construction. The bold boundary in the image on the left denotes the correct superpixel boundary of the object. Gray highlighted regions denote the foreground superpixels under consideration. At the two chords, pq and $p'q'$, the features, f_{pq} and $f_{p'q'}$, fall into different bins of the histogram, i.e., the chordiogram shown on the right. At each boundary point, the foreground selection of bordering superpixels defines the normal direction.

To recover exact object contour pixel locations and reduce false positives, an additional shape matching step is required on top of the object hypotheses. Here, we propose using the collective shape of a subset of superpixels within each hypothesis region to verify the presence of an object.

For each detection hypothesis region, superpixels are computed directly from gPb [28]. Searching over the entire space of superpixel subsets for the optimal match between the collective shape of the superpixels and the object model is combinatorial and impractical. Instead, we use a greedy algorithm to efficiently perform the search. In practice, with limited superpixels to select from, our greedy approach recovers the correct region with high probability. Figure 6 shows example results of shape verification. The greedy algorithm begins with a set of connected superpixels as a seed region and greedily searches over adjacent superpixels, picking the superpixel that yields the smallest χ^2 distance to the chordiogram of model silhouette. Intuitively, if we have a set of superpixels forming a large portion of the object with a few missing pieces, adding these pieces yields the best score. The initial seeds are formed by choosing all triplets of adjacent superpixels, and limiting examination to the top five seeds that yield the smallest χ^2 distance. The connectivity graph of superpixels is a planar graph with limited node degrees. The complexity of finding triplets in such a planar graph is $O(N \log N)$ in the number of nodes.

Once the correct foreground superpixels are selected, the detection bounding box is re-cropped to reflect the recovered foreground mask. Empirically, this cropping step yields a better localization of the detection result over the S-DPM, as measured in terms of precision and recall, see Sec. IV Edges of the foreground mask are extracted and used in the subsequent processing stage for accurate 6-DoF *continuous* pose estimation.

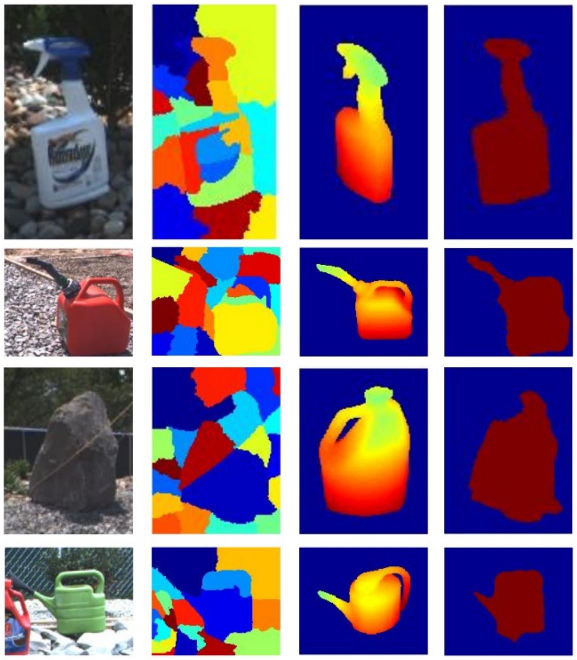

Fig. 6: Shape descriptor-based verification examples. (left-to-right) Detection hypothesis window of the object, superpixel over-segmentation of the hypothesis region, visualization of the coarse object pose from the object detector and selected foreground mask.

F. Pose refinement

Robotic grasping requires an accurate estimate of an object's 3D pose. To improve upon the coarse pose estimate provided by the S-DPM, we perform a final iterative pose refinement step to recover the full *continuous* 6-DoF pose. This step is restricted to the region of the verified superpixel mask.

Our iterative refinement process consists of two steps: (i) determining the correspondence between the projected occluding boundary of the 3D model and the contour points along object segmentation mask, and (ii) estimating an optimal object pose based on the correspondences.

The contour correspondences are estimated using dynamic programming (DP) to ensure local matching smoothness. Given the initial (coarse) pose output from the object detection stage, the 3D object model is rendered to the image and its corresponding projected occluding boundary is extracted. Each point on the contour is represented by a descriptor capturing close-range shape information. The 31-dimensional contour descriptor includes the gradient orientation of a contour point (the central point) and the gradient orientations of the nearest 15 points on each side of the central point along the contour. The gradient orientation of the central point is subtracted from all elements of the descriptor, which gives in-plane rotation invariance. The matching cost between each pair is set to be the l_2 distance of the feature descriptor extracted at each point. DP is then used to establish the correspondences between contour points.

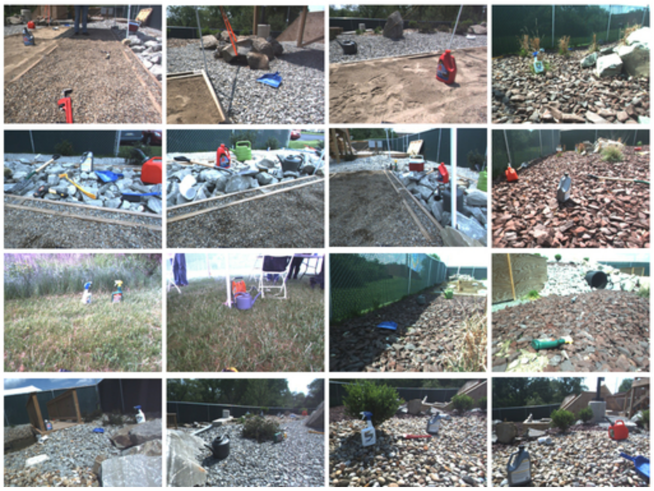

Fig. 7: Representative images from the introduced outdoor dataset. The dataset was captured using a ground robot and includes diverse terrains, e.g., rocks, sand and grass, with illumination changes. Portions of the terrain are non-flat. Objects are scattered around the scene and typically do not occupy a major portion of the scene.

To estimate the refined pose we use the motion field equation [33]:

$$u(x, y) = \frac{1}{Z}(xt_z - t_x) + \omega_x(xy) - \omega_y(x^2 + 1) + \omega_z(y)$$
$$v(x, y) = \frac{1}{Z}(yt_z - t_y) - \omega_x(y^2 + 1) - \omega_y(xy) + \omega_z(x),$$

where $u(x, y), v(x, y)$ denote the horizontal and vertical components of the displacement vectors, respectively, between the model and matched image contour points, computed by DP, $Z(x, y)$ denotes the depth of the 3D model point for the current pose estimate and the Euler angles $(\omega_x, \omega_y, \omega_z)$ and 3D translation vector (t_x, t_y, t_z) denote the (locally) optimal motion of the object yielding the refined pose. The motion update of the current pose is recovered using least squares. This procedure is applied iteratively until convergence. In practice, we usually observe fast convergence with only three to five iterations. The running time of the pose refinement is about one second on an Intel 2.4GHz i7 CPU.

IV. Experiments

Outdoor detection evaluation We introduce a challenging outdoor dataset for 3D object detection containing heavy background clutter. This dataset was collected from a moving robot and consists of eight sequences containing a total of 3403 test images; the dimensions of each image are 512 × 386. Figure 7 shows a set of representative imagery from the introduced dataset. The scenes contain a variety of terrains (e.g., grass, rock, sand, and wood) observed under various illumination conditions. The dataset represents the task of a robot navigating a complex environment and searching for objects of interest. The objects of interest are mostly comprised of textureless daily equipment, such as a watering pot, gas tank, watering can, spray bottle, dust pan, and liquid container. For each frame, 2D bounding boxes that tightly outline each object are provided. Further, the dataset includes the corresponding 3D model files used in our empirical evaluation.

On the outdoor dataset, we performed a shape-based object detection evaluation. We compared four methods, DOT [34], S-DPM with only the root model, full S-DPM with root and parts, and the full S-DPM plus shape verification (proposed approach), on a detection task on the introduced dataset. Both DOT and S-DPM used the same training instances from Sec. III-A with a slight difference. For S-DPM, we trained one model component for each of 16 discrete poses. For DOT, we used the same quantization of the viewsphere but trained with 10 different depths ranging from close to far in the scene. During testing, S-DPM is run on different scales by building an image pyramid. The input to both methods were the same gPb thresholded images. In all our experiments, the threshold is set to 40 (gPb responses range between 0 and 255), where edges with responses below the threshold were suppressed. The default parameters of gPb were used. We did not observe a noticeable difference in the detection and pose estimate accuracy with varying the gPb parameter settings.

Table III shows a comparison of the average precision for detection on the outdoor dataset. The proposed approach consisting of the full S-DPM plus shape verification achieves the best mean average precision. It demonstrates that using shape verification improves detection due to the refinement of the bounding box to reflect the recovered silhouette. Full S-DPM outperforms both the root only S-DPM and DOT. This shows the benefit of the underlying flexibility in S-DPM.

Table top evaluation We evaluated our pose refinement approach under two settings. First, we recorded an indoor RGB-D dataset, with multiple objects on a table, from a head mounted Kinect on a PR2 robot. The RGB-D data is used as ground truth. We evaluated using three objects, watering can, gas tank, watering pot, placed at two different distances from the robot on the table and two different poses for each distance. For each scene, the target object was detected among all objects on the table and segmented using shape verification, and then the 6-DoF pose was estimated, as described in Sec. III-F. The model point cloud was projected into the scene and Iterative Closest Point (ICP) [35] was performed between the model point cloud and the Kinect point cloud. We report ICP errors for both rotation and translation in Tables I and II, resp. Errors in the rotations and translations are small for different angles and different depth. Translation errors in the X and Y directions are smaller than in Z direction. Since Z is the depth direction, it is most affected by the 3D model acquisition and robot calibration. Both measurements show our method is robust and suitable for grasping task.

In addition, using the object pose estimated from our approach, we demonstrate with a PR2 robot successful

	watering pot	gas tank	watering can	spray bottle	dust pan	liquid container	average AP
S-DPM full+shape	0.686	0.645	0.523	0.515	0.429	0.506	**0.5507**
S-DPM full	0.688	0.610	0.547	0.507	0.387	0.509	0.5413
S-DPM root only	0.469	0.535	0.433	0.436	0.295	0.436	0.4340
DOT	0.407	0.412	0.340	0.089	0.111	0.188	0.2578

TABLE III: Average precision on the introduced outdoor dataset.

		Estimated Rotation			Error		
		Roll	Pitch	Yaw	Roll	Pitch	Yaw
watering can	dist1	1.65	48.44	-145.37	0.99	3.57	-1.63
		5.50	50.73	-22.37	-3.20	-3.92	-0.07
	dist2	-4.33	41.93	48.78	-3.20	-3.92	-0.07
		2.44	49.60	-54.82	-0.12	1.95	-1.92
watering pot	dist1	-0.43	59.20	-73.00	-1.25	-0.28	1.36
		0.69	51.90	156.86	-1.82	-0.63	-3.48
	dist2	-10.43	66.93	38.28	-1.078	-6.67	-2.43
		-0.633	52.24	-131.94	-0.21	1.14	-0.88
gas tank	dist1	-0.15	50.58	-136.17	1.43	2.73	-4.58
		2.84	50.15	-51.15	-2.63	3.20	2.79
	dist2	-2.44	48.24	129.43	-3.57	0.02	-2.14
		-7.40	45.22	109.90	-1.55	-1.79	-1.03

TABLE I: Estimated absolute rotation of the object and error in degrees.

		Estimated Translation			Error		
		X	Y	Z	X	Y	Z
watering can	dist1	-46.5	-82.3	-1023.6	-1.14	-0.7	-2.8
		-57.1	-86.4	-1023.2	-1.2	2.8	-7.2
	dist2	-85.1	183.2	-1182.9	3.6	3.6	4.8
		-114.9	186.0	-1200.3	4.5	2.2	-5.1
watering pot	dist1	16.4	-154.0	-1020.9	2.8	1.0	0.06
		-117.6	-112.4	-1028.3	0.4	0.2	2.2
	dist2	-6.8	32.7	-1051.2	2.0	-2.9	-3.5
		-106.5	-6.6	-1053.1	-0.5	-0.2	-1.9
gas tank	dist1	-23.8	21.2	-1061.2	-1.8	-0.9	-3.2
		19.5	-116.0	-958.8	-0.4	1.7	-3.2
	dist2	-77.0	6.7	-1064.6	0.4	-0.9	-2.0
		-111.3	178.9	-1200.8	0.6	-0.4	-1.4

TABLE II: Estimated absolute translation of the object and error in centimeters.

detections and grasps of various objects from a cluttered table. In Fig. 8, we show qualitative results of the PR2 successfully grasping various objects on a cluttered table.

V. Conclusion

We presented an integrated approach for detecting and localizing 3D objects using pure geometric information derived from a database of 3D models. We create an initial set of hypotheses with a state-of-the-art parts-based model trained on clusters of poses. Detection hypotheses are segmented and reranked by matching subsets of superpixels with model boundary silhouettes using the chordiogram descriptor. The resulting segmentation enables the refinement of 3D pose in a small number of steps. Due to the holistic nature of the chordiogram-based superpixel selection, our approach is resistant to clutter. We demonstrate the grasps of textureless objects in difficult cluttered environments in the video supplement.

Acknowledgment Financial support through the following grants NSF-IIP-0742304, NSF-OIA-1028009, ARL MAST CTA W911NF-08-2-0004, and ARL Robotics CTA W911NF-10-2-0016, NSF-DGE-0966142, and NSF-IIS-1317788 is gratefully acknowledged.

References

[1] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *PAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.

[2] A. Toshev, B. Taskar, and K. Daniilidis, "Shape-based object detection via boundary structure segmentation," *IJCV*, vol. 99, no. 2, pp. 123–146, 2012.

[3] W. Grimson, *Object recognition by computer: The role of geometric constraints*. Cambridge, MA: The MIT Press, 1990.

[4] R. Horaud, "New methods for matching 3-D objects with single perspective views," *PAMI*, vol. 9, no. 3, pp. 401–412, May 1987.

[5] G. Häusler and D. Ritter, "Feature-based object recognition and localization in 3D-space, using a single video image," *CVIU*, vol. 73, no. 1, pp. 64–81, January 1999.

[6] C. Cyr and B. Kimia, "3D Object Recognition Using Shape Similarity-Based Aspect Graph," in *ICCV*, 2001, pp. 254–261.

[7] F. Han and S. Zhu, "Bayesian reconstruction of 3D shapes and scenes from a single image," in *Int. Workshop on High Level Knowledge in 3D Modeling and Motion*, 2003.

[8] A. Sethi, D. Renaudie, D. Kriegman, and J. Ponce, "Curve and surface duals and the recognition of curved 3D objects from their silhouettes," *IJCV*, vol. 58, no. 1, pp. 73–86, June 2004.

[9] S. Lazebnik, A. Sethi, C. Schmid, D. Kriegman, J. Ponce, and M. Hebert, "On pencils of tangent planes and the recognition of smooth 3D shapes from silhouettes," in *ECCV*, 2002, pp. III: 651–665.

[10] J. Liebelt, C. Schmid, and K. Schertler, "Viewpoint-independent object class detection using 3D Feature Maps," in *CVPR*, 2008, pp. 1–8.

[11] J. Liebelt and C. Schmid, "Multi-view object class detection with a 3D geometric model," in *CVPR*, 2010, pp. 1688–1695.

[12] M. Villamizar, H. Grabner, J. Andrade-Cetto, A. Sanfeliu, L. Van Gool, F. Moreno-Noguer, and K. Leuven, "Efficient 3D object detection using multiple pose-specific classifiers," in *BMVC*, 2011.

[13] D. Glasner, S. Vitaladevuni, and R. Basri, "Contour-based joint clustering of multiple segmentations," in *CVPR*, 2011, pp. 2385–2392.

[14] B. Pepik, M. Stark, P. Gehler, and B. Schiele, "Teaching 3D geometry to deformable part models," in *CVPR*, 2012, pp. 3362–3369.

[15] Q. Hao, R. Cai, Z. Li, L. Zhang, Y. Pang, F. Wu, and Y. Rui, "Efficient 2D-to-3D correspondence filtering for scalable 3d object recognition," in *CVPR*, 2013.

[16] V. A. Prisacariu, A. V. Segal, and I. Reid, "Simultaneous monocular 2D segmentation, 3D pose recovery and 3D reconstruction," in *ACCV*. Springer, 2013, pp. 593–606.

[17] S. Ullman and R. Basri, "Recognition by linear combinations of models," *PAMI*, vol. 13, pp. 992–1006, 1991.

[18] R. Basri and S. Ullman, "The alignment of objects with smooth surfaces," *CVGIP*, vol. 57, no. 3, pp. 331–345, May 1993.

[19] C. Gu and X. Ren, "Discriminative mixture-of-templates for viewpoint classification," in *ECCV*, 2010, pp. V: 408–421.

[20] S. Savarese and L. Fei-Fei, "3D generic object categorization, localization and pose estimation," in *ICCV*, 2007, pp. 1–8.

[21] N. Payet and S. Todorovic, "From contours to 3D object detection and pose estimation," in *ICCV*, 2011, pp. 983–990.

[22] E. Hsiao, A. Collet, and M. Hebert, "Making specific features less discriminative to improve point-based 3D object recognition," in *CVPR*. IEEE, 2010, pp. 2653–2660.

[23] K. Lai, L. Bo, X. Ren, and D. Fox, "A scalable tree-based approach for joint object and pose recognition," in *AAAI*, 2011.

[24] R. Rusu, G. Bradski, R. Thibaux, and J. Hsu, "Fast 3D recognition and pose using the viewpoint feature histogram," in *IROS*, 2010, pp. 2155–2162.

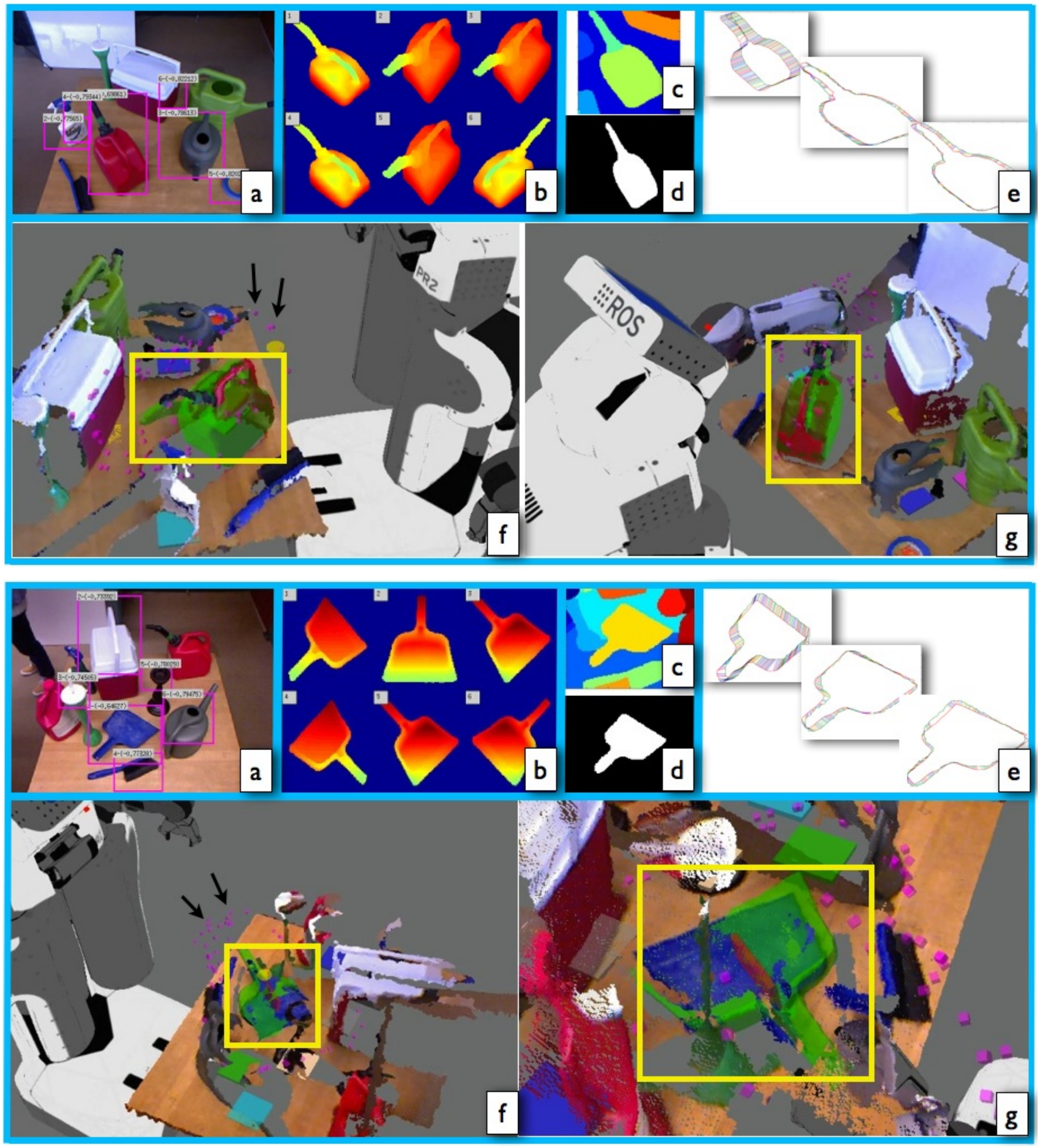

Fig. 8: PR2 grasping process for two example input images. Top panel for gas tank and bottom for dust pan. a) S-DPM detection bounding boxes ordered by the detection score in decreasing order. b) Corresponding pose output from S-DPM for each detection. c) Segmentation of top scored detection window. d) Foreground mask selected by shape verification. e) Three iterations in pose refinement, alignments (shown in color) between curves are computed using DP. f) Visualization of PR2 model with the Kinect point cloud. Notice that the estimated model given in light green is well aligned with the point cloud. Grasping points are indicated by arrow. g) Another view of the same scene.

[25] R. B. Rusu and S. Cousins, “3D is here: Point cloud library (PCL),” in *ICRA*, 2011, pp. 1–4.

[26] H. Koppula, A. Anand, T. Joachims, and A. Saxena, “Semantic labeling of 3D point clouds for indoor scenes,” in *NIPS*, 2011.

[27] S. Vijayanarasimhan and K. Grauman, “Efficient region search for object detection,” in *CVPR*, 2011, pp. 1401–1408.

[28] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, “Contour detection and hierarchical image segmentation,” *PAMI*, vol. 33, no. 5, pp. 898–916, 2011.

[29] S. Izadi, R. Newcombe, D. Kim, O. Hilliges, D. Molyneaux, S. Hodges, P. Kohli, J. Shotton, A. Davison, and A. Fitzgibbon, “Kinectfusion: real-time dynamic 3D surface reconstruction and interaction,” in *ACM SIGGRAPH*, vol. 23, 2011.

[30] J. Canny, “A computational approach to edge detection,” *PAMI*, no. 6, pp. 679–698, 1986.

[31] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *CVPR*, 2005, pp. I: 886–893.

[32] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge,” *IJCV*, vol. 88, no. 2, pp. 303–338, June 2010.

[33] B. K. P. Horn, *Robot Vision*. the MIT Press, 1986.

[34] S. Hinterstoisser, V. Lepetit, S. Ilic, P. Fua, and N. Navab, “Dominant orientation templates for real-time detection of texture-less objects,” in *CVPR*, 2010, pp. 2257–2264.

[35] P. Besl and N. McKay, “A method for registration of 3-D shapes,” *PAMI*, vol. 14, no. 2, pp. 239–256, February 1992.