Visual Grasp Affordances From Appearance-Based Cues

Hyun Oh Song
Mario Fritz
Chunhui Gu
Trevor Darrell

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2013-16
http://www.eecs.berkeley.edu/Pubs/TechRpts/2013/EECS-2013-16.html

March 4, 2013

Visual Grasp Affordances From Appearance-Based Cues

Hyun Oh Song, UC Berkeley
Mario Fritz, MPI Informatics
Chunhui Gu, UC Berkeley
Trevor Darrell, ICSI, UC Berkeley

Abstract

In this paper, we investigate the prediction of visual grasp affordances from 2D measurements. Appearance-based estimation of grasp affordances is desirable when 3-D scans are unreliable due to clutter or material properties. We develop a general framework for estimating grasp affordances from 2-D sources, including local texture-like measures as well as object-category measures that capture previously learned grasp strategies. Local approaches to estimating grasp positions have been shown to be effective in real-world scenarios, but are unable to impart object-level biases and can be prone to false positives. We describe how global cues can be used to compute continuous pose estimates and corresponding grasp point locations, using a max-margin optimization for category-level continuous pose regression. We provide a novel dataset to evaluate visual grasp affordance estimation; on this dataset we show that a fused method outperforms either local or global methods alone, and that continuous pose estimation improves over discrete output models.

1. Introduction

Affordances are believed to be one of the key concepts that enables an autonomous agent to decompose an infinite space of possible actions into a few tractable and reasonable ones. Given sensor input, resemblance to previous stimuli–both at an instance and category level–allows us to generalize previous actions to new situations. Gibson [6] defined affordances as "action possibilities" that structure our environment by functions of objects that we can choose to explore. In the context of robotics, this concept has attained new relevance, as agents should be able to manipulate novel objects. Early models proposing a computational approach for predicting affordance functions started from a geometric paradigm [19]. A number of different implementations [17, 18, 13] of this idea have been attempted, but often suffer from the fact that matching primitives in real-wold settings can be challenging. In this paper we explore the direct inference of grasp affordances using monocular cues.

Research in the robotics field has for some time developed grasp strategies for known objects based on 3-D knowledge on an instance basis [7, 9]. In cases where clutter or material properties preclude extraction of a reliable point cloud for a target objects, appearance-based cues are desirable. Recently, methods for generalizing grasps using 2-D observations have been proposed [15, 16, 10, 12, 2]. This new class of methods reflects the traditional goal of inference of grasp affordance.

But typically, these "graspiness" measures have been computed strictly locally (especially [15]), without leveraging any larger image context. Models which find grasp points based only on classifiers defined on local texture models cannot capture category or instance-level bias, and therefore may break an object (fragile wine glass grasped from the top), trigger an unintended side-effect (grasping spray bottle at the trigger), damage the gripper (not grasping potentially hot pot at handle) or simply acquire an unstable grasp. We propose a method for combining such local information with information from object-level pose estimates; we employ category-level continuous pose regression to infer object pose (and from that, grasp affordances). Also, we develop a grasp inference method using pose estimates from a max-margin regression technique, and show this strategy can significantly improve performance over discrete matching methods.

Previous methods have not, to our knowledge, addressed pose regression for inferring grasp affordances. This is mainly a result of the difficult interaction of intra-object category variation interleaved with changing pose, which makes it hard to learn and generalize across instances and view-points in a robust manner. In fact only recently, pose estimation under category variation has been attempted for discrete view-point classes [14, 8, 11]. In order to leverage larger contexts for improved grasp affordance, stronger models for pose estimation are needed; we employ continuous, category-level pose regression.

Our work provides the following contributions: 1) we combine texture-based and object-level appearance cues for grasp affordance estimation; 2) we evaluate max-margin pose regression on the task of category-level, continuous pose estimation; and 3) we collect and make available a new dataset for image-based grasp affordance prediction re-

search.

2. Method

We develop a method for grasp affordance estimation that uses two paths: a local path inspired by the framework of [15], which models grasp affordance using a local texture cue, and a global path, that utilizes object-level regression to estimate object pose, and then regresses from object pose to grasp points. For the global path, we extend the framework proposed in [8] to the task of category-level continuous outputs, as those are what is needed in our task. Figure 1 illustrates how the two pipelines interact in our framework.

When the global and the local visual affordance detectors are fused properly, it exceeds either method alone, as shown in experiments below. Informally, we consider the global detector to be exploiting object-level information to get the estimate "in the ballpark", where the local detector could bring the final estimate to be aligned to a good edge based on the local "graspiness".

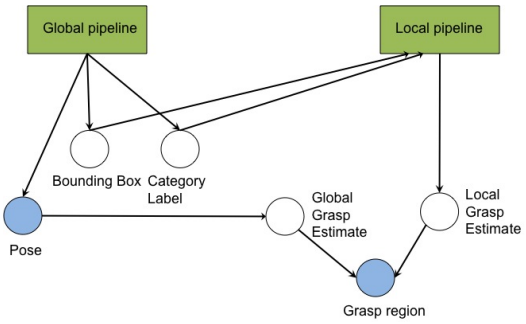


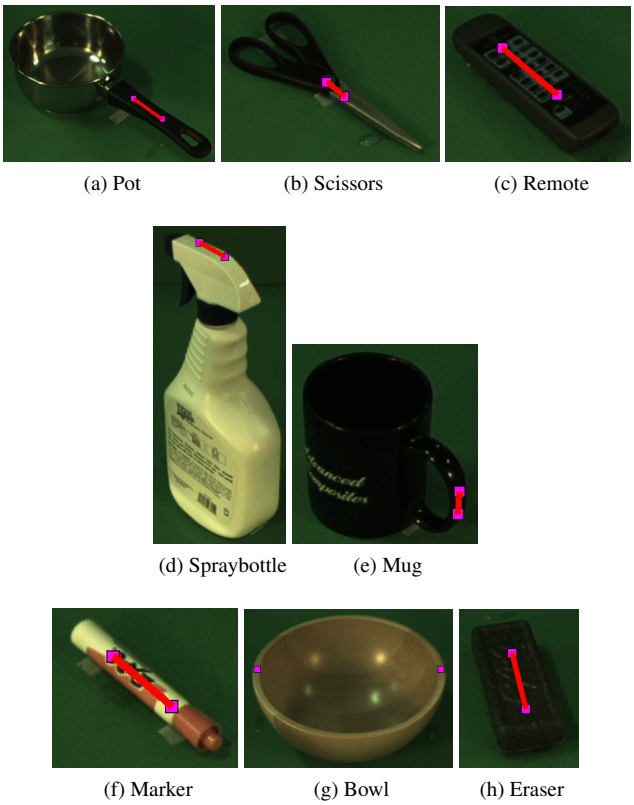
Figure 1: The block diagram of the complete system

2.1. Local grasp region detection

Saxena et al. [15] trains a local grasp point detector that looks at local patches and classifies them either as a valid grasp point or not. They propose a binary classification model trained on local patches extracted from synthetic supervised data; the model identified grasp points from a local descriptor that is similar to a multi-scale texture filter bank, but with some differences (see [15]). Informally, and in our experience, the model learns a set of local edge structures that compose a good grasp point such as the handle of a mug reasonably well.

Since local grasp measures operate based on local appearance, they lack specificity when run on entire images. In their operational system this is mitigated by restricting response to the known (or detected) bounding box of the target object in the image. They also employ a triangulation step for verification with stereo sensors, which we do not apply here as our data set is monocular.[1] The pure local

[1] We are interested both in detecting grasp affordances with robotic sensors, and also doing so from general image and video sources, so as to improve scene understanding in general and/or to improve estimation of other objects or agents in the world. E.g., we can constrain our estimate of the pose or motion of a person if we can infer how he or she is holding an object, or how they will grasp it if they are approach the object. (C.f., [20]).



(a) Pot (b) Scissors (c) Remote

(d) Spraybottle (e) Mug

(f) Marker (g) Bowl (h) Eraser

Figure 2: Examples of triangulated grasp annotations.

method cannot capture category or instance-level bias such as a human demonstration of a grasp strategy for a category of interests.

Figure 2 shows example images annotated with grasp regions. We define grasp regions as where humans or robotic agents would stably grasp objects. The cooking pot shown in Figure 2 (a), for example, the ground truth grasp regions are the mid part along the elongated direction of the handle. The marker in Figure 2 (f), excludes regions near the cap and the tail of the marker from the grasp region for better grasp stability. Along with the "grasp region" attributes, we also annotated "grasp scale" attributes for all the training instances that are used in feature extraction stages. More explanations on the annotation attributes are given in the following subsections.

We address some important technical details of the local method in [15] and propose modifications employed in our local grasp measure classifier which lead to more reliable

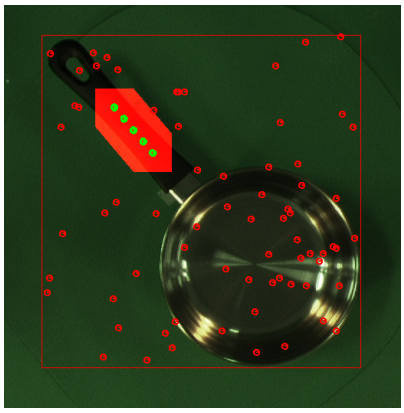


Figure 3: Example of supervised key point sampling

responses.

2.1.1 Supervised key point sampling

One of the most common techniques for sampling key points for feature extraction is sampling evenly in a grid structure as implemented in [15]. However, this method is very susceptible to binning effects and ambiguities in training data. The binning effect is when small object displacement can cause very different samples and is an inherent problem in grid structures. We avoided this problem by uniformly sampling positive patches along the ground truth grasp bands as shown as green circles in Figure 3.

The label ambiguities can occur if key points that are very close together get sampled and assigned to different labels. In the binary classification sense, these ambiguous data can be interpreted as inseparable data points in feature dimensions that adversely effect the separating hyperplane. Our approach is to utilize an additional annotation which we call the "grasp scale" attribute of the grasp annotations to define a convex hull around the ground truth grasp region and randomly sample negative key points outside the convex hull. Figure 3 illustrates the convex hull as red polygon and randomly chosen negative key points as red circles.

2.1.2 Category dependent descriptor scale

While the method above determines key point sampling locations, the scale of the descriptor turns out to be an important factor in order to obtain reliable local grasp measures. This relates to the aperture problem as encountered in the scale of local features. Having a small local scale results in features that encode edge type responses and tend to be reproducible. For larger scales, we add more context which makes the feature more unique and therefore also more discriminative. The best scale will therefore naturally vary from object class to object class. E.g. with a set of fixed size descriptors (aperture), it's impossible to capture both the parallel edges from narrow handle of mugs and wide handle of cooking pots. This holds true for the largest context descriptors also. We again utilize the "grasp scale" attribute of the grasp annotations and set descriptor scales dependent on the attribute.

2.2. Global grasp region regression

Our global path is based on a method for category-level continuous pose regression. We start with the model in [8], which reports results on continuous pose regression over trained instances and on discrete pose estimation from category level data. We extend it here to the case of category-level continuous pose estimation, which to our knowledge has not been previously reported by this or any other method[2].

A multi-scale window scanning is applied to localize objects in the image, searching across category and pose: following [8] and [5], we define a score function, $S_{\mathbf{w}}(x)$ of a image window x evaluated under a set of viewpoints as following

$$S_{\mathbf{w}}(x) = \max_{v \in \mathcal{V}, \Delta\theta} f(v, \Delta\theta) \quad (1)$$

$$= \max_{v \in \mathcal{V}, \Delta\theta} \left(w_v + g_v^T \Delta\theta\right)^T \psi_v(x) - d(\Delta\theta) \quad (2)$$

$$\theta(x) = \theta_{v^*} + \Delta\theta^* \quad (3)$$

where $\mathbf{w} = \{w_v\}$ are the different viewpoint templates, $\psi_v(x)$ is the feature vector at location x, g_v are the Jacobian matrices of the templates w_v over θ at discrete viewpoint v, $\Delta\theta$ are the offset viewpoint angles of x with respect to the canonical viewpoint angles θ_v, $d(\cdot)$ is a quadratic loss function that confines $\theta(x)$ to be close to θ_v. Denote $\Delta\theta$ by their elements $[\Delta\theta_1, \Delta\theta_2, \Delta\theta_3]^T$, then $d(\Delta\theta) = \sum_{i=1}^{3} d_{i1}\Delta\theta_i + d_{i2}\Delta\theta_i^2$. In Eqn. (3), v^* and $\Delta\theta^*$ are obtained when the score function reaches its maximum. The variables w_v, g_v, θ_v and d_{i1}, d_{i2} are learned from training data. Given positive examples $\{x_1, x_2, \ldots, x_P\}$ with continuous pose labels $\{\theta_1, \theta_2, \ldots, \theta_P\}$ we can express the above criteria efficiently as

$$f(v, \Delta\theta) = \left(w_v + g_v^T \Delta\theta\right)^T \psi_v(x) - d(\Delta\theta) \quad (4)$$

$$= \tilde{w_v}^T \tilde{\psi_v}(x) \quad (5)$$

where $\tilde{w_v}$ and $\tilde{\psi_v}(x)$ are stacked versions of the weights and parameters. See [8] for details.

Given pose estimates, we can directly infer grasp points. amount. The global affordance detector works by regressing upon the pose of an object to a 2D affordance in the

[2]Except of course for face pose estimation, for which there is a considerable literature; we consider here the case of multi-category recognition and regression methods.

image plane. (The local detector simply identifies points in the image that have the local appearance of graspable region; this is complementary information.)

We marginalize over the pose estimate in order to obtain a robust grasp point prediction:

$$\mathbf{g}^* = \arg\max_{\mathbf{g}} \sum_{\boldsymbol{\theta}} P(\mathbf{g}|\boldsymbol{\theta}, c) P(\boldsymbol{\theta}|c) \tag{6}$$

where $\mathbf{g}$ is the global grasp affordance, $\boldsymbol{\theta}$ is the pose estimate and c is the category label.

2.3. Fused grasp region estimates

The position of the final estimate is based on fusion of local and global paths. Position and orientation estimates are represented as a probability density function over location and angle, respectively, and multiple hypotheses can be returned as appropriate. The local and global paths each provide a probability map over estimated grasp location in the image. We return the fused estimates, taking the entrywise product of the two probability map to be the fused estimate.

$$[\mathbf{u}; \mathbf{v}]^* = \arg\max \left(\mathcal{N}\left(\mathbf{g}, \sigma\mathbf{I}\right) \circ \mathbf{L}\right) \tag{7}$$

where $[\mathbf{u}; \mathbf{v}]$ is the fused grasp region, $\mathbf{g}$ is the global grasp affordance, σ is a smoothing parameter, $\mathbf{L}$ is the grasp likelihood map from the local pipeline.

Figure 4 shows some examples where our fusion scheme successfully recovers from failures in either the local or the global pipeline. Figure 4 (a) and (d) show the output of the global pipeline and Figure 4 (b) and (e) show the top scoring patches from the local measure. The first row shows erroneous global grasp estimate due to incorrect pose estimate getting corrected by fusion step owing to correct local estimate. The second row shows the global pipeline not begin affected by poor local estimate during the fusion step.

3. Experiments

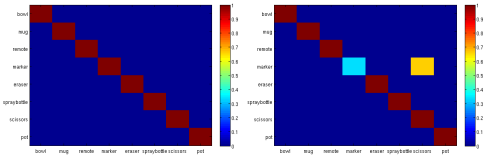
3.1. New Dataset for Evaluating Visual Grasp Affordance Prediction under Categorial Variation

Existing datasets with pose annotated visual categories only address discrete view point classes [14]. We are only aware of a single exception [11], which only has a single category (car) and also doesn't lend itself to the investigation of grasp affordances.

Therefore we propose a new dataset consisting of 8 categories (markers, erasers, spray bottles, bowls, mugs, pots, scissors and remote controllers) common to office and domestic domain for each of which we imaged 5 instances. The training set shows the instances under 259 viewpoint variations yielding a training set of total size of 1295 images per categories. All the images in the dataset also have the grasp affordance annotations with grasp region and scale attributes mentioned before.

Scenes	Methods	Category averaged detections
Clean scene	Ours	96.9 %
	3D [1]	65.6 %
Cluttered scene	Ours	81.3 %
	3D [1]	3.10 %

Table 1: Detection accuracy comparison on both scenes



(a) Confusion table for the uncluttered scene (b) Confusion table for the cluttered scene

Figure 5: Categorization confusion tables for detected object

As for test sets, we collected two sets of data. On the first set, we collected 8 instances per categories of previously unseen objects both in an uncluttered desk and a cluttered desk. On this dataset, we evaluate our detection performance against an established baseline system using 3D modalities [1]. The other testset contains 18 viewpoint variations per categories as well as significant scale changes of previously unseen instances in cluttered background. We show experimental results on detection, categorization, pose estimation and grasp affordance estimation.

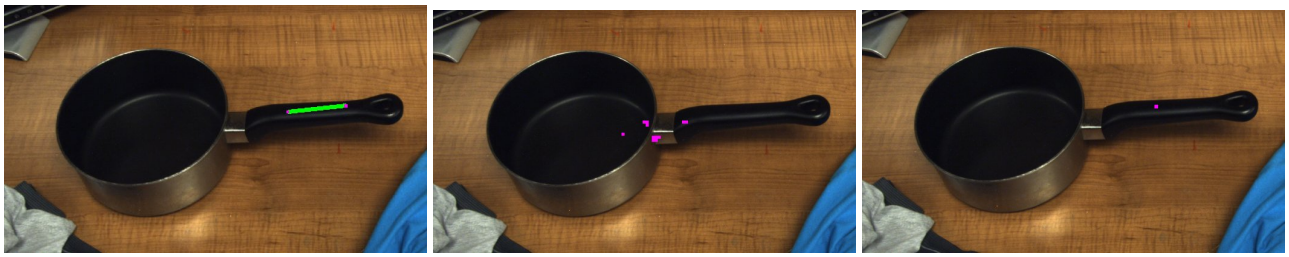
At time of publication this database will be made available to the public. We intend to keep this database growing in order to maintain it as a challenge for fine grained category pose estimation.

3.2. Detection performance comparison against 3D baseline

We chose the top first detections across all the categories and followed the standard detection criteria in vision communities where the ratio between the intersection and the union of the predicted and ground truth bounding boxes are thresholded at 50% [4]. Table 1 shows the detection accuracies on both the clean and cluttered desk scenes compared against the 3D baseline [1]. Also, figure 5 shows the categorization confusion tables of our system for the both the clean and the cluttered scenes evaluated on correct detections.

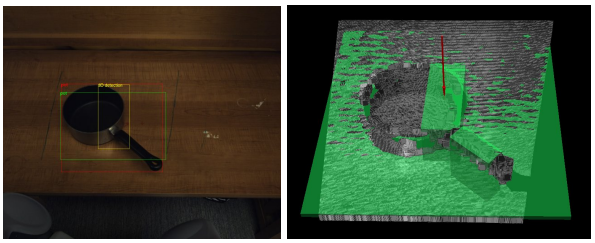
Figure 6 shows some failure cases of the baseline 3D detection system [1]. In Figure 6 (b),(d) show failed 3D detection bounding cubes and Figure 6 (a),(c) show overlayed detections. The red bounding boxes are the ground truth,

(a) Incorrect global estimate (b) Correct local estimate (c) Fused estimate

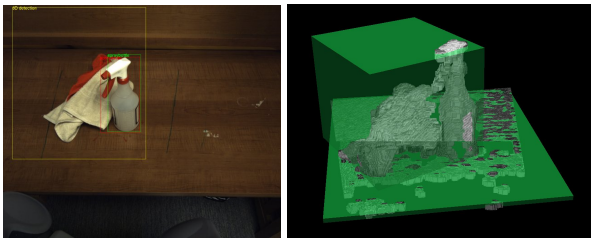


(d) Correct global estimate (e) Incorrect local estimate (f) Fused estimate

Figure 4: Individual failures corrected by the probabilistic fusion. Best viewed when zoomed in.



(a) Overlayed detections (b) 3D detection



(c) Overlayed detections (d) 3D detection

Figure 6: Examples of failure cases of 3D [1] versus 2D detection

the green bounding boxes are output of our system and the yellow boxes are the 3D detection overlayed onto the image plane.

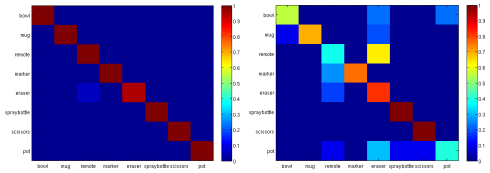
Generally, when textured light is shed on dark colored or weakly reflective objects, the color contrast from the textured light is very small causing very sparse point cloud. The sparsity then segregates points cloud into multiple groups causing multiple 3D detections. This scenario could be detrimental when a precise object size has to be known to place the picked-up object to another location. Also, when there is a background clutter, point cloud of the clutter objects gets easily aggregated with the foreground object causing a aggregate 3D detection. However, a 2D scanning window based framework can handle this to a better extent as shown in Table I. Finally, 3D point cloud based detection fails when objects have not enough protrusion from the table e.g., scissors.

3.3. Detection, Categorization, Pose estimation and Grasp affordance estimation results on heavily cluttered scene

We now report the experimental result on the the second test data set with more viewpoint and scale variations and clutter mentioned above. To examine how the detection and categorization effects the pose estimation and grasp affordance estimation task, we separately carried out the experiment in two scenarios where the ground truth bounding boxes and category labels were supplied versus not supplied.

3.3.1 Detection and categorization

We applied the same detection evaluation scheme in the previous experiment where the top first detection among all locations of a given image among all the categories were thresholded at 50% [4].



(a) Our method (b) Nearest neighbor baseline

Figure 7: Categorization confusion tables for detected objects

The detection rates of the top first detections for bowls, mugs, remotes, markers, erasers, spray bottles, scissors and pots were 100.00(%), 100.00(%), 50.00(%), 88.89(%), 44.44(%), 94.44(%), 50.00(%), 66.67(%) and 55.56(%) respectively. The average rate across all the categories rate was 81.25(%). Figure 7 shows the confusion table for the categorization performance on correct detections.

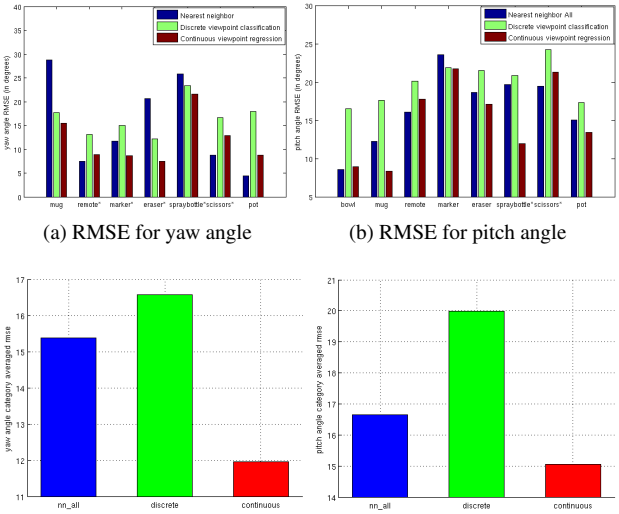
3.3.2 Multi-Category Pose Prediction

We evaluate current approaches to 3d pose estimation and investigate how they translate to our desired setting of angle accurate predictions while performing generalization to previously unseen test objects. As a baseline method we looked at a nearest neighbor approach where we compute HOG [3] of given test images and compare among all the 1295 images per categories(stored as HOG [3] templates) with L2 distance metric. Additionally we evaluate [8] as it is to our knowledge the state-of-the-art on the popular 3d (discrete) pose database proposed in [14] both in discrete viewpoint classification mode and in continuous viewpoint regression mode.[3].

Figure 8 and Figure 9 shows the performance in root mean squared error of the roll and pitch angle we obtain using the proposed dataset when the object location and category labels were given and not given respectively. As expected we observe a significant drop when comparing the angle accurate results from [8] to our setting where we evaluate both on cross-instance and cross-category generalization.

Part of the increased error originated from the geometric symmetry of those classes. Ambiguity due to symmetry is an inherent problem and we see it amplified in the category scenario, where additional texture and labelings on objects is less likely to be informative for the viewpoint angle. The most common symmetry is 180° symmetry in yaw angles for elongated objects such as remotes, markers, erasers and scissors. In manipulation tasks with pinch grasps however, this symmetry becomes trivial because pinch grasp grippers

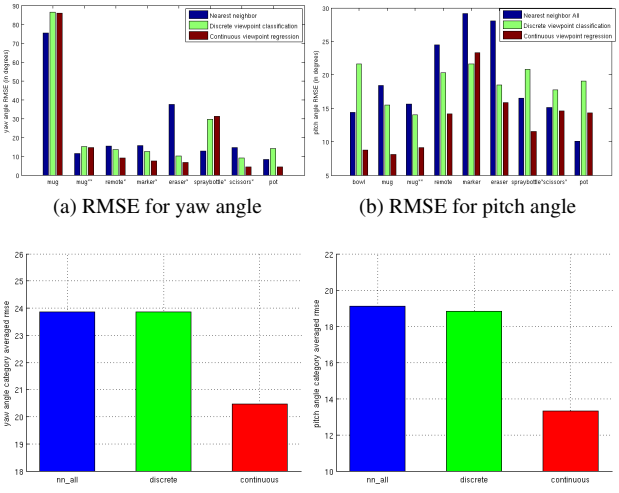
[3]Code was provided by the authors



(a) RMSE for yaw angle (b) RMSE for pitch angle

(c) Category averaged rmse of yaw (d) Category averaged rmse of pitch

Figure 8: Accuracy of pose prediction given object locations and category labels. The top two plots show RMSE for yaw and pitch angle, respectively, bottom two show category averaged RMSE for yaw and pitch, respectively.



(a) RMSE for yaw angle (b) RMSE for pitch angle

(c) Category averaged rmse of yaw (d) Category averaged rmse of pitch

Figure 9: Accuracy of pose prediction without object locations and category labels. The top two plots show RMSE for yaw and pitch angle, respectively, bottom two show category averaged RMSE for yaw and pitch, respectively.

have 180° symmetry as well. (i.e. executing yaw angle 10 and 190° would result in the same gripper configuration). Exploiting this symmetry, we wrapped the yaw angle estimates for the above categories at 180°.

For mug category however, the handle is the only descriptive part that defines the yaw angle. Even if the detec-

	Local (px)	Global (px)	Fused (px)	Fused (cm)
Bowls	62.33	17.61	9.99	0.41
Mugs	61.97	12.83	8.38	0.35
Remotes	33.35	6.82	8.46	0.35
Markers	18.22	5.68	3.67	0.15
Erasers	56.03	12.84	19.02	0.79
Spray Bottles	153.70	44.19	19.34	0.80
Scissors	10.41	17.11	12.05	0.50
Pots	177.41	47.43	34.02	1.41
Average	71.68	20.56	14.37	0.59

Table 2: Affordance prediction given groundtruth bounding box

	Local (px)	Global (px)	Fused (px)	Fused (cm)
Bowls	65.10	39.47	28.50	1.18
Mugs	38.06	134.91	109.23	4.51
Mugs**	20.58	38.13	19.84	0.82
Remotes	38.20	8.54	9.91	0.41
Markers	13.22	7.98	4.72	0.19
Erasers	46.15	13.31	17.81	0.74
Spray Bottles	114.23	35.04	33.26	1.37
Scissors	7.95	10.52	5.07	0.21
Pots	181.34	46.43	29.13	1.20
Average	58.31	37.15	28.61	1.18

Table 3: Affordance prediction without bounding box and category label

tion threshold of 50% was satisfied, the handle was sometimes not included in the detected bounding box rendering the pose regression task ill-defined. Therefore, we increased the detection threshold for mugs at 70% so that it's guaranteed that the handle is included in the detection. Figure 9 (a) shows the yaw angle pose estimation results for mugs both when the threshold were at 50% and at 70% (mug** designates angle estimates when mugs were thresholded at 70%).

3.3.3 Visual Grasp Affordance Prediction

We now evaluate the accuracy of our joint method for grasp affordance prediction. Again, we use the proposed dataset where we have annotated graph affordances.

We investigate the same two scenarios as in the pose estimation experiment. The first assumes that a bounding box was provided by a bottom up segmentation scheme - as it could be available in a robotic setting by 3d sensing or a form of background subtraction. The second scenario will run our full detection pipeline and all further processing is based on this output.

As a first baseline we compare to the results from purely local measures (tagged "Local(px)"). The approach "Global(px)" only uses the global path by predicting grasp affordances regressing from the predicted the poses conditioned on the corresponding predicted category labels. Then, we present the fused approach (tagged "Fused(px)"). Finally, we converted the mean pixel deviation from the fused estimate into real world metric distances by working out the perspective projection using the previously recorded depth measurements (tagged "Fused(cm)").

Table 2 shows the average distance in pixels between the predicted grasp affordance and the ground truth annotation when the bounding box is assumed to be known while Table 3 shows results employing the full processing pipeline. We observe consistent improvements on the average results going from the purely local cue, switching to the global pipeline and finally fusing local and global in our combined approach. Overall, we reduced the average distance obtained by local model by more than half in pixel deviations. [15] reports 1.80 cm metric distance error when the object locations were known. We report 0.59 cm and 1.18 cm metric distance error when the object locations were known and not known.

Figure 10 presents example predictions of our framework on randomly chosen test objects. The magenta patches represent the points among the fused probability maps where the likelihoods are the highest (patches were blown up to help the visualization) The red boxes and thick axes represent ground truth bounding boxes and axes. Respectively, the green boxes and the thin axes represent the predicted object locations and pose.

4. Conclusion

Appearance-based estimation of grasp affordances is desirable when other (e.g., 3-D) sensing means cannot accurately scan an object. We developed a general framework for estimating grasp affordances from 2-D sources, including local texture-like measures as well as object-category measures that capture previously learned grasp strategies. Our work is the first to combine texture-based and object-level monocular appearance cues for grasp affordance estimation. Further, we provided a novel evaluation of max-margin pose regression on the task of category-level continuous pose estimation. On a novel dataset for visual grasp affordance estimation we show that a fused method outperforms either local or global methods alone, and that continuous pose estimation improves over discrete output models.

References

[1] Tabletop Object Detector. Willow Garage, Robot Operating System, 2011.
Available online at http://www.ros.org/wiki/tabletop_object_detector. 4, 5

[2] J. Bohg and D. Kragic. Learning grasping points with shape context. *Robotics and Autonomous Systems*, 2009. 1

(a) Bowl (b) Mug (c) Remote (d) Marker

(e) Eraser (f) Spray bottle (g) Scissors (h) Pot

Figure 10: Examples predictions of our framework

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 6

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, June 2010. 4, 5

[5] P. F. Felzenszwalb, D. McAllester, and D. Ramana. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008. 3

[6] J. J. Gibson. The theory of affordance. In *Percieving, Acting, and Knowing*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1977. 1

[7] C. Goldfeder, M. Ciocarlie, H. Dang, and P. K. Allen. The columbia grasp database. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2009. 1

[8] C. Gu and X. Ren. Discriminative mixture-of-templates for viewpoint classification. In *ECCV*, 2010. 1, 2, 3, 6

[9] K. Huebner, K. Welke, M. Przybylski, N. Vahrenkamp, T. Asfour, D. Kragic, and R. Dillmann. Grasping known objects with humanoid robots: A box-based approach. In *International Conference on Advanced Robotics*, 2009. 1

[10] Q. Le, D. Kamm, A. Kara, and A. Ng. Learning to grasp objects with multiple contact points. In *IEEE Int. Conf. on Robotics and Automation*, 2010. 1

[11] M. Ozuysal, V. Lepetit, and P. Fua. Pose estimation for category specific multiview object localization. In *CVPR*, 2009. 1, 4

[12] N. Ratliff, J. A. Bagnell, and S. Srinivasa. Imitation learning for locomotion and manipulation. In *IEEE-RAS International Conference on Humanoid Robotics*, 2007. 1

[13] E. Rivlin, S. J. Dickinson, and A. Rosenfeld. Recognition by functional parts. *Computer Vision and Image Understanding: CVIU*, 62(2):164–176, 1995. 1

[14] S. Savarese and L. Fei-Fei. 3d generic object categorization, localization and pose estimation. In *ICCV*, Rio de Janeiro, Brazil, October 2007. 1, 4, 6

[15] A. Saxena, J. Driemeyer, and A. Y. Ng. Robotic grasping of novel objects using vision. *Journal of Robotics Research*, 2008. 1, 2, 3, 7

[16] A. Saxena, L. Wong, and A. Ng. Learning grasp strategies with partial shape information. In *AAAI*, 2008. 1

[17] L. Stark and K. Bowyer. Achieving generalized object recognition through reasoning about association of function to structure. *PAMI*, 13(10):1097–1104, 1991. 1

[18] L. Stark, A. Hoover, D. Goldgof, and K. Bowyer. Function-based recognition from incomplete knowledge of shape. In *WQV93*, pages 11–22, 1993. 1

[19] P. H. Winston, B. Katz, T. O. Binford, and M. R. Lowry. Learning physical descriptions from functional definitions, examples, and precedents. In *AAAI*, 1983. 1

[20] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, San Francisco, CA, June 2010. 2