Learning to detect visual grasp affordance

Hyun Oh Song, Mario Fritz, *Member, IEEE,* Daniel Goehring, and Trevor Darrell

***Abstract*—Appearance-based estimation of grasp affordances is desirable when 3-D scans become 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. Finally, we demonstrate our autonomous object detection and grasping system on the Willow Garage PR2 robot.**

***Note to Practitioners*—Learning grasp affordances for autonomous agents such as personal robots is a challenging task. We propose an unified framework which first detects target objects, infers grasp affordance of the target object, and executes robotic grasp. Our method is mainly based on 2D imagery data which can be more robust when 3D scans are unavailable due to background clutter and material properties such as surface reflectance. One of the future extensions would be to automate the training phase so that robots can actively learn object models by interacting with objects as opposed to having a human in the loop collecting and annotating training images.**

***Index Terms*—Object detection, Machine learning, Pose estimation, Affordance, Grasping, Autonomous agent.**

I. 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 [10] defined affordances as "action possibilities" that structure our environment by functions of objects that we can choose to explore. In particular, grasp affordance captures the set of feasible grasp strategies which might be available to the agent when presented with previously unseen objects.

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 [32]. A number of different implementations [24, 28, 29] 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.

Fig. 1: PR2 robot grasping a previously unseen cooking pot placed on a cluttered scene fully autonomously.

Research in the robotics field has for some time developed grasp strategies for known objects based on 3-D knowledge on an instance basis [11, 13]. 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 or 2.5-D observations have been proposed [2, 15, 16, 18, 19, 23, 26, 27]. This new class of methods reflects the traditional goal of inference of grasp affordance.

But typically, these "graspiness" measures have been computed strictly locally [15, 18, 26], without identifying the object to be grasped and thus doesn't leverage any larger image context. Models which find grasp points based only on local texture classifier 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), simply acquire an unstable grasp [9], or be incapable of recognizing and fetching specified objects of interest. 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

H. Song and T. Darrell are with the Department of Computer Science, UC Berkeley, CA 94720, USA. (email: song@eecs.berkeley.edu; trevor@eecs.berkeley.edu)

M. Fritz is with Max-Planck-Institute for Informatics, Campus E1 4, 66123 Saarbrucken, Germany. (email: mfritz@mpi-inf.mpg.de)

D. Goehring is with Arminallee 7, Institut fuer Informatik, Freie Universitaet Berlin, 14195 Berlin, Germany. (email: daniel.goehring@fu-berlin.de)

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. Only recently, pose estimation under category variation has been attempted for discrete view-point classes [12, 20, 22, 25]. 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 propose a fully autonomous robotic object grasping system by combining 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 research.

II. Related work

Learning visual affordances for object grasping has been an active area of robotics research. This area of research has been approached from several fronts, including: 3D model based methods [3, 5, 21], learning local graspable 2D or 2.5D patches [2, 14, 15, 18, 26, 27, 30]. However, there has been less attention towards learning to grasp objects by first recognizing a semantic object category, estimating object pose and applying category specific grasp strategies learned from supervised training.

[30] took a step towards this approach of learning category specific grasp affordances and proposed a method using a code book based detection [17] model to estimate object grasp affordances from 2D images. However, the experiments were limited to only one object class and the object pose was not estimated by the algorithm requiring hard coded grasp poses.

Recently [12, 22] proposed max-margin pose estimation methods based on the state of the art object detection system [7, 8] enabling simultaneous detection and pose estimation. However, the detection and pose estimation performance haven't been evaluated when the object is not centered in the image and object category is unknown.

Overall, in contrast to the previous approaches, our system performs combined end to end inference of recognition, pose estimation, affordance estimation and grasping. We demonstrate fully autonomous object detection and grasping on PR2 robot.

III. Methods

We develop a method for grasp affordance estimation that uses two paths: a local pipeline inspired by the framework of [26], which models grasp affordance using a local texture cue, and a global pipeline, that utilizes object-level regression to estimate object pose, and then regresses from object pose to grasp regions. For the global path, we extend the framework proposed in [12] to the task of category-level continuous outputs, as those are what is needed in our task. Figure 2 illustrates how the two pipelines interact in our framework.

The local and global grasp estimates are fused in a probabilistic framework. In the experimental section, we will show that this integrated model outperforms its individual components. 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".

In the following subsections, we address components of the system in detail. Section III-A discusses the key ingredients in the local pipeline. Then, section III-B explains the global pipeline. Finally, section III-C describes the probabilistic fusion process of the two pipelines.

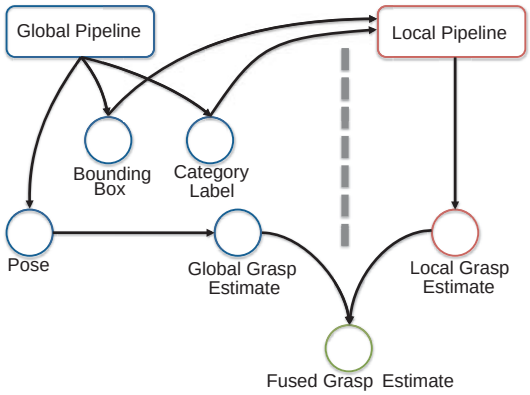


Fig. 2: The block diagram of the complete system. Local and global information is fused to a joint grasp estimate. In addition, the local pipeline is improved by bounding box and category label predictions from the global pipeline.

A. Local grasp region detection

Saxena et al. [26] train a local grasp patch detector that looks at local patches and classifies them either as valid grasp patches 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 [26]). Our analysis shows that the model learns a set of local edge structures that compose a good grasp patch 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 method

[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., [33]).

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

Figure 3 shows example images annotated with "grasp region" attributes (handle of cooking pot, mid part of markers, etc.). We define grasp regions as where humans or robotic agents would stably and safely grasp objects. 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 [26] and propose modifications employed in our local grasp measure classifier which lead to more reliable responses. Subsection III-A1 provides the algorithm and visualizations from feature extraction steps to inference and post processing steps of the local method and subsections III-A2 and III-A3 introduces improvements to the algorithm.

1) Feature extraction and inference: Figure 4 (left) shows a visualization of a SIFT (Scale-invariant feature transform) descriptor at [31] three different scales on an example keypoint. Each bin shows local statistics of gradient orientations. For training, we train a patch level binary classifier w_L with positive and negative training data as shown in Figure 6. For inference, we first compute a SIFT representation of the test image Ψ and convolve it with the learned classifier w_L to get a local grasp affordance map L followed by Gaussian smoothing as shown in Eq. 1.

$$L = \Psi * w_L * k\left(0, \sigma I\right), \tag{1}$$

where $k\left(0, \sigma I\right)$ is a truncated zero mean Gaussian blur kernel. Figure 4 (middle) shows the classifier response for all uniformly sampled test keypoints. Regions with classification confidence greater than 0.5 are overlaid in the red channel of the image. Figure 4 (right) visualizes the result where Gaussian smoothing is applied on the classifier output. We set the standard deviation of the Gaussian kernel equal to the keypoint grid sample size throughout the experiment. We tried both SVM (support vector machine) and logistic regression classifiers and they showed negligible difference in performance.

Extracting good (easy to learn and unambiguous) training patches from real camera images requires more precautions than extracting the data from synthetic graphics data [26]. Some of the issues that arise from working with real sensor imagery involve: alignment difficulties in experiments, incorrect annotations, wide varieties across object instances, presence of texture on object surfaces, and realistic lighting conditions, etc. Subsections III-A2 and III-A3 addresses some of these issues in more depth.

2) 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 [26]. 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 data samples and is an inherent problem when data is sampled in grid structures. Figure 5 shows difficulties of this approach. We avoided this problem by uniformly sampling positive patches along the ground truth grasp bands as shown as green circles in Figure 6.

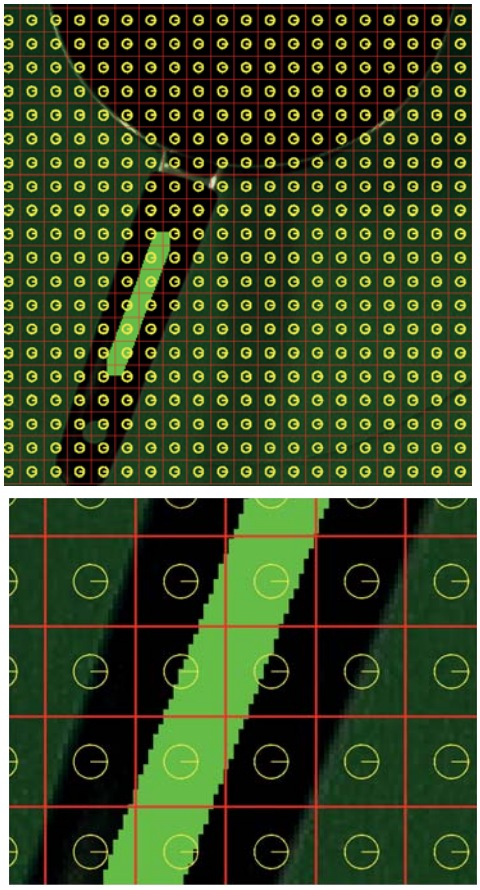


Fig. 5: Issues of keypoint even-sampling strategy. (Top) Visualization of evenly sampled keypoints for a cooking pot. Yellow circles denote each keypoints. Green band represents ground truth grasp region annotation in the center of the handle. (Bottom) Close up view near the annotated region. Due to binning effect, it becomes ambiguous which keypoints should get assigned with positive/negative grasp regions labels. Arbitrary assignment causes outliers in the training process.

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 6 illustrates the convex hull as red polygon and randomly chosen negative key points as red circles.

3) 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.

Fig. 3: Randomly sampled examples of our dataset with grasp annotations. Grasp region attribute is defined by two end points (magenta patches in the figures). We take convex hull of the two points as valid grasp region except for the bowl category.

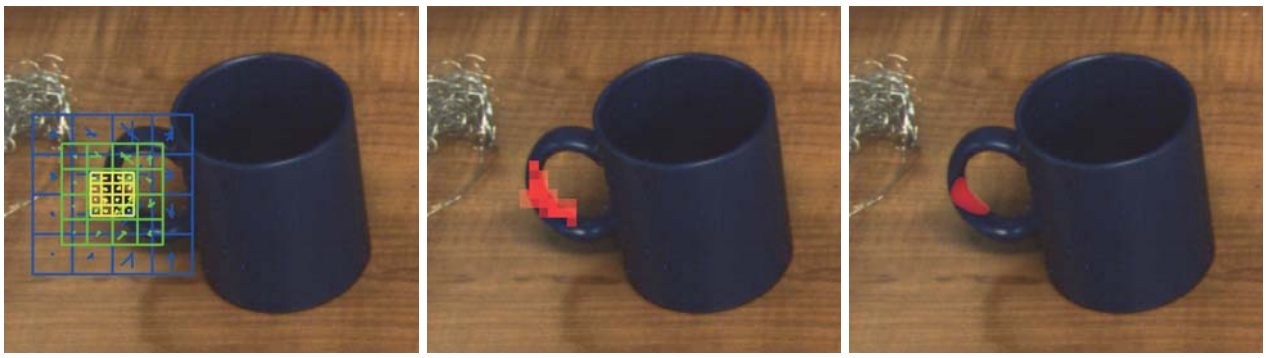


Fig. 4: Local grasp region detection. (left) SIFT descriptor on a key point. Descriptor scales are color coded with yellow, green and blue (middle) Red patches indicate thresholded classifier output. (right) Visualization after Gaussian smoothing.

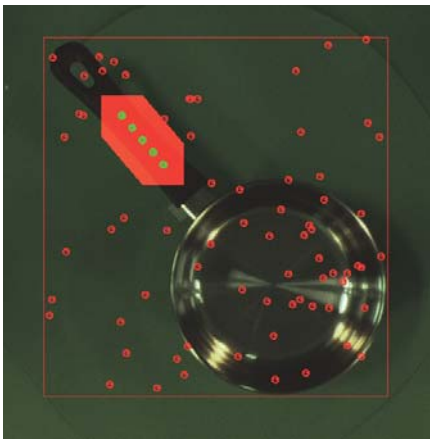


Fig. 6: Example of supervised key point sampling. Positive key points(green markers) are sampled along "grasp region" annotation while negative key points(red markers) are randomly sampled strictly outside the convex hull defined by "grasp scale" annotation.

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. Note that at test time, the grasp scale is derived from the bounding box and object class provided by the global pipeline as shown in Figure 2.

B. Global grasp region regression

Our global path is based on a method for category-level continuous pose regression. We start with the model in [12], 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 for

general object classes[2]. In subsection III-B1 we review the pose estimation model and in subsection III-B2, the conversion process from pose to grasp affordance estimate is discussed.

1) Pose Estimation: A multi-scale window scanning is applied to localize objects in the image, searching across category and pose. First, we define discretized canonical viewpoints in the viewing hemisphere as illustrated in Figure 7. Then, following [12] and [7, 8], we define a score function, $S_{\mathbf{w}}(\mathbf{x}) \in \mathbb{R}$ of a image window $\mathbf{x} \in \mathbb{R}^m$ evaluated under the set of viewpoints as following,

$$\begin{aligned} S_{\mathbf{w}}(\mathbf{x}) &= \max_{v \in \mathcal{V}, \Delta\boldsymbol{\theta}} f(\mathbf{x}, v, \Delta\boldsymbol{\theta}) \\ &= \max_{v \in \mathcal{V}, \Delta\boldsymbol{\theta}} \left(\mathbf{w}_v + J_v^T \Delta\boldsymbol{\theta}\right)^T \psi_v(\mathbf{x}) - d(\Delta\boldsymbol{\theta}) \\ \boldsymbol{\theta}(\mathbf{x}) &= \boldsymbol{\theta}_{v^*} + \Delta\boldsymbol{\theta}^* \end{aligned} \tag{2}$$

where $v = \{1, \ldots, V\}$ correspond to viewpoint indices sampled from the viewing hemisphere at V different locations, $\mathbf{w}_v \in \mathbb{R}^m$ are learned viewpoint templates. $\psi_v(\mathbf{x}) \in \mathbb{R}^m$ is the SIFT feature vector computed on window $\mathbf{x}$ of the input image.

$\boldsymbol{\theta}_v \in \mathbb{R}^3$ is the supervised Euler angle annotation at viewpoint index v. $\Delta\boldsymbol{\theta}$ represents small deviation angle from the supervised annotation angle $\boldsymbol{\theta}_v$. The final pose estimate is the deviation corrected angle $\boldsymbol{\theta}(\mathbf{x})$.

$J_v \in \mathbb{R}^{3 \times m}$ is the Jacobian matrix of the viewpoint template $\mathbf{w}_v$ over the angle $\boldsymbol{\theta}_v$. The motivation of the Jacobian term is that we want to slightly deform the learned canonical view templates by $\Delta\boldsymbol{\theta}$. Explicitly, the Jacobian linearization of vector $\mathbf{w}_v$ about the canonical view angle $\boldsymbol{\theta}_v$ with respect to the three Euler angles $[\theta_1, \theta_2, \theta_3]^T$ can be written as:

$$J_v^T = \begin{bmatrix} \frac{\partial \mathbf{w}_v(1)}{\partial \theta_1} & \frac{\partial \mathbf{w}_v(1)}{\partial \theta_2} & \frac{\partial \mathbf{w}_v(1)}{\partial \theta_3} \\ \vdots & \vdots & \vdots \\ \frac{\partial \mathbf{w}_v(m)}{\partial \theta_1} & \frac{\partial \mathbf{w}_v(m)}{\partial \theta_2} & \frac{\partial \mathbf{w}_v(m)}{\partial \theta_3} \end{bmatrix} \tag{3}$$

The input to the pose estimation algorithm is the test view x and the output is the pose estimate $\boldsymbol{\theta}(\mathbf{x}) = \boldsymbol{\theta}_{v^*} + \Delta\boldsymbol{\theta}^*$ where $\boldsymbol{\theta}_{v^*}$ denotes the best matching discrete viewpoint and the $\Delta\boldsymbol{\theta}^*$ denotes the slight deformation from the viewpoint to the actual test view. $d(\cdot)$ is a quadratic loss function that confines $\boldsymbol{\theta}(\mathbf{x})$ to be close to $\boldsymbol{\theta}_v$. Denote $\Delta\boldsymbol{\theta}$ by their elements $[\Delta\theta_1, \Delta\theta_2, \Delta\theta_3]^T$, then

$$d(\Delta\boldsymbol{\theta}) = \sum_{i=1}^{3} d_{i1} \Delta\boldsymbol{\theta}_i + d_{i2} \Delta\boldsymbol{\theta}_i^2 \tag{4}$$

In Eqn. (2), v^* and $\Delta\boldsymbol{\theta}^*$ are obtained when the score function reaches its maximum. The variables $\mathbf{w}_v$, J_v, $\boldsymbol{\theta}_v$ and d_{i1}, d_{i2} are learned from training data. Given positive examples $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_P\}$ with annotated pose labels $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_P\}$ we can express the above criteria compactly as a dot product between reparameterized weight vector and feature vector as follows,

[2]But see the extensive literature on face pose estimation

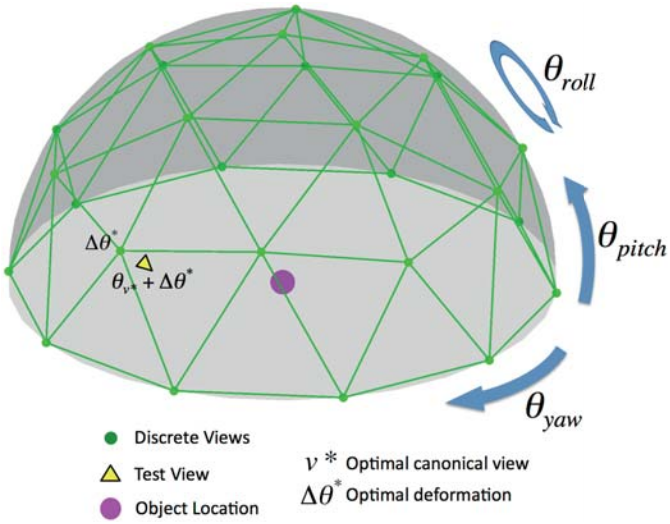


Fig. 7: Illustration of the discretized viewing hemisphere and the pose estimation algorithm. Input to the algorithm is the test view $\mathbf{x}$ and the output is the pose estimate $\boldsymbol{\theta}(\mathbf{x}) = \boldsymbol{\theta}_{v^*} + \Delta\boldsymbol{\theta}^*$.

$$f(\mathbf{x}, v, \Delta\boldsymbol{\theta}) = \tilde{\mathbf{w}}_v{}^T \tilde{\psi}_v(\mathbf{x}) \tag{5}$$

where $\tilde{\mathbf{w}}_v$ and $\tilde{\psi}_v(\mathbf{x})$ are structured as following,

$$\begin{aligned} \tilde{\mathbf{w}}_v &= \left[\mathbf{w}_v^T, vec(J_v)^T, d_{11}, \ldots, d_{32}\right]^T \\ \tilde{\psi}_v(\mathbf{x}) &= [\psi_v(\mathbf{x}), \Delta\theta_1 \psi_v(\mathbf{x}), \Delta\theta_2 \psi_v(\mathbf{x}), \Delta\theta_3 \psi_v(\mathbf{x}), \\ &\quad -\Delta\theta_1, -\Delta\theta_2, -\Delta\theta_3, -\Delta\theta_1^2, -\Delta\theta_2^2, -\Delta\theta_3^2]^T \end{aligned} \tag{6}$$

where $vec(\cdot)$ operator forms a vector from the input matrix by stacking columns of the input matrix on top of each other. We discuss the training and inference procedures for pose estimation below.

a) Training procedure: We solve the following optimization problem in Eq. 7 to jointly train all the view point templates $\mathbf{w}_1, \ldots, \mathbf{w}_V$ in max-margin framework.

$$\begin{aligned} \min_{\tilde{\mathbf{w}}_1, \ldots, \tilde{\mathbf{w}}_V} \frac{1}{2} \sum_{v=1}^{V} ||\tilde{\mathbf{w}}_v||_2^2 + C_N \sum_{n=1}^{N} \sum_{v=1}^{V} \max\left(0, 1 + \tilde{\mathbf{w}}_v^T \tilde{\psi}_v(\mathbf{x}_n)\right) \\ + C_P \sum_{p=1}^{P} \max\left(0, 1 - \tilde{\mathbf{w}}_{v(p)}^T \tilde{\psi}_{v(p)}(\mathbf{x}_p)\right) \end{aligned} \tag{7}$$

where $\mathbf{x}_p$ and $\mathbf{x}_n$ denote positive and negative training data, and $v(p)$ denotes the supervised viewpoint label for a positive data $\mathbf{x}_p$. C_N and C_P are regularization parameters which controls the tradeoff between the classification performance

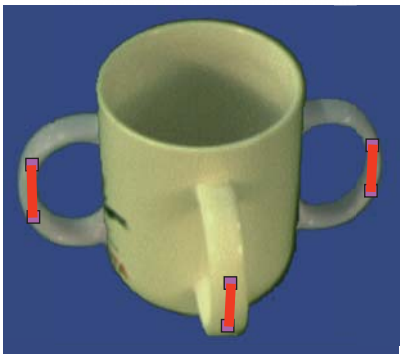


Fig. 8: Overlaid samples of training data for one object instance of mug category as yaw angle is varied at $0°, 180°, 270°$. Pitch and roll angles are fixed at $45°$ and $0°$ respectively. We use nonparametric locally weighted regression to learn the mapping between object pose to grasp labels (illustrated with magenta patches)

on the training data and the norm of the model parameters $\{\tilde{\mathbf{w}}_i\}_{i=1}^{V}$. The intuition behind the optimization problem in Eq. 7 is that we want all the view point templates to score low for all the negative data while the supervised template $\tilde{\mathbf{w}}_{v(p)}$ scores high for the corresponding positive data.

b) Inference procedure: Having learned the viewpoint templates $\mathbf{w}_1, \ldots, \mathbf{w}_V$, we can perform sliding window style object detection which assigns score $S_{\mathbf{w}}(\mathbf{x})$ at every image locations $\mathbf{x}$. After thresholding the score, we can infer the viewpoint estimate $\boldsymbol{\theta}(\mathbf{x})$ of the object hypothesis.

$$\begin{aligned}[v^*,\ \Delta\boldsymbol{\theta}^*] &= \underset{v,\ \Delta\theta}{\operatorname{argmax}} f(\mathbf{x}, v, \Delta\boldsymbol{\theta}) \\ \boldsymbol{\theta}(\mathbf{x}) &= \boldsymbol{\theta}_{v^*} + \Delta\boldsymbol{\theta}^*\end{aligned} \tag{8}$$

The intuition behind Eq. 8 is that we want to infer the most likely pose of an object hypothesis location $\mathbf{x}$ by maximizing over possible discrete view point labels v and angle deformation $\Delta\boldsymbol{\theta}$.

2) Pose to Grasp Affordance Regression: Given pose estimates, $\boldsymbol{\theta}(\mathbf{x})$ we can directly infer grasp regions. The global affordance prediction step works by regressing upon the pose of an object to a 2D affordance in the image plane. (The local detector simply identifies points in the image that have the local appearance of graspable region; this is complementary information.) Regressing from pose and category information, the global pipeline infers the grasp affordance annotations in Figure 3. This can be formulated as learning multidimensional regression functions $h(\boldsymbol{\theta}; c)$ that maps a 3D pose estimate $\boldsymbol{\theta}$ to a grasp label $\mathbf{g}$ in pixel coordinates given a category label c and assigns probability estimate on the likelihood $P(\mathbf{g}|\boldsymbol{\theta}, c)$. Explicitly, $\mathbf{g} = [\mathbf{g}_1, \mathbf{g}_2]$ where $\mathbf{g}_1 \in \mathbb{R}^2$, $\mathbf{g}_2 \in \mathbb{R}^2$ are individual end points in grasp region labels in pixel coordinates illustrated as magenta patches in Figure 3. We use locally weighted regression to learn the regression functions from the training data for each categories. Figure 8 illustrates a sample trajectory of $\mathbf{g}$ as $\boldsymbol{\theta}$ is varied. Then, we marginalize over the candidate pose estimate in order to obtain a robust grasp point prediction:

$$\mathbf{g}^* = \underset{\mathbf{g}}{\operatorname{argmax}} \sum_{\boldsymbol{\theta}} P(\mathbf{g}|\boldsymbol{\theta}, c) P(\boldsymbol{\theta}|c), \tag{9}$$

where $\mathbf{g}^*$ is the predicted most likely grasp region, angle $\boldsymbol{\theta} = [\theta_{pitch}, \theta_{yaw}, \theta_{roll}]^T$ is the Euler angle pose estimate and c is the category label. Then the global grasp affordance map G is determined by the following procedure.

$$G = \text{convHull}(\mathbf{g}) * k(0, \sigma I) \tag{10}$$

where we take the convex hull of the predicted grasp estimates and convolve with the truncated zero mean Gaussian kernel $k(0, \sigma I)$ with standard deviation σ set equal to the one used in the local pipeline.

C. 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 and taking argmax to be the fused estimate.

$$\mathbf{a}^* = \text{argmax}(\mathbf{G} \circ \mathbf{L}), \tag{11}$$

where $\circ$ denotes matrix Hadamard product, $\mathbf{a}^*$ is the fused grasp region with maximal confidence, $\mathbf{G}$ is the global grasp affordance, $\mathbf{L}$ is the grasp likelihood map from the local pipeline.

Figure 9 shows some examples where our fusion scheme successfully recovers from failures in either the local or the global pipeline. Figure 9 (a) and (d) show the output of the global pipeline and Figure 9 (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.

D. Generating 3D grasp points

Grasping an object requires knowledge of the 3D coordinates (x, y, z) of a grasping point, and the 3 orientation angles $(\theta_{yaw}, \theta_{pitch}, \theta_{roll})$ to specify the gripper orientation. The (x, y, z) coordinates of the grasping point are obtained by projecting the (u, v) coordinates of a pixel to a calibrated range sensor. For our experiments we used an Asus Xtion camera mounted on the head of a PR2 robot.

Finding the gripper angles requires constraining the $(\theta_{yaw}, \theta_{pitch}, \theta_{roll})$ orientation angles. The first two angles $(\theta_{yaw}, \theta_{pitch})$ is calculated by the pose estimation algorithm, as illustrated in Fig. 7. In this work we assume that an object has $0°$ roll angle and therefore can be grasped either from the top or from the side. This is true for most of household

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

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

Fig. 9: Individual failures corrected by the probabilistic fusion. Best viewed when zoomed in.

objects that have to stand upright on a tabletop surface. The information about what approach to use is provided during training.

Choosing an overhead or a side grasp simultaneously constrains both θ_{roll} and θ_{pitch}, thus specifying all the needed parameters to determine the desired gripper position. Executing the grasp then requires planning a collision free motion to a pre-grasping position, and closing the gripper around the specified target. We exploited the redundancy of the PR2 arm (7 DOF), to find grasping postures that do not collide with the tabletop surface.

IV. Experiments

We performed two sets of experiments. Experiments in section IV-B compares the detection performance between approach and 3D baseline. The set of experiments in section IV-C are designed to extensively evaluate various aspects of our approach in terms of detection, categorization, pose estimation, grasp affordance prediction, and robotic grasping.

A. Dataset for Evaluating Visual Grasp Affordance Prediction under Categorial Variation

Datasets for learning 2D and 2.5D grasp locations exist [18, 26]. However the number of images and pose varieties in the dataset are quite limited (total of 1035 images for 9 object categories) in order for one to learn object detector models from. Furthermore, pose annotations for the images are not provided in the dataset.

Existing datasets with pose annotated visual categories only address discrete view point classes [25]. We are only aware of a single exception [20], which only has a single category (car) and also doesn't lend itself to the investigation of grasp affordances.

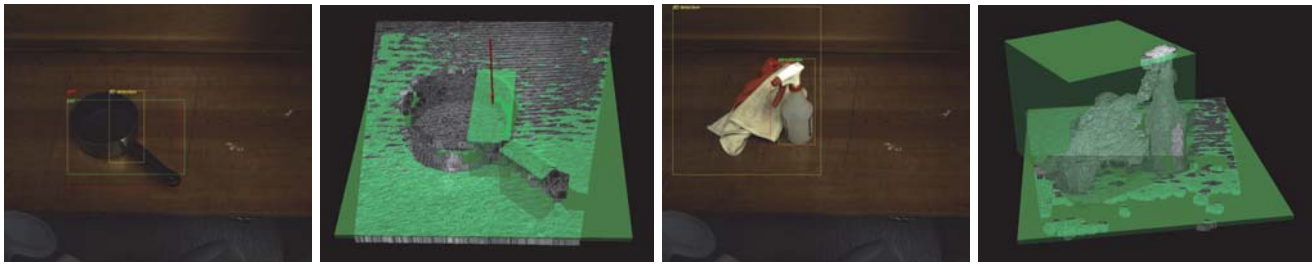
Therefore we propose a new dataset consisting of 8 object categories (markers, erasers, spray bottles, bowls, mugs, pots, scissors and remote controllers) common to office and domestic domain for each of which we imaged 5 instances at 1280×960 resolution. The training set shows the instances under 259 viewpoint variations (*pitch* angle: $0 \sim 90^\circ$ sampled at 15° each, *yaw* angle: $0 \sim 350^\circ$ sampled at 10° each) yielding a training set of total size of 10360 images. All the images in the dataset also have the grasp affordance annotations with grasp region and scale attributes mentioned before. Figure 3 shows subset of our dataset.

As for test sets, we collected two sets of data. On the first set, we collected 8 instances per category 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.

B. Detection performance comparison against 3D baseline

We chose the highest scoring detections in the image across all the categories and followed the standard object detection criteria where a predicted bounding box is considered a true positive detection if the ratio between the intersection and the union of the predicted and ground truth bounding box is more than 50% [6]. Table 1 shows the detection accuracies on both the clean and cluttered desk scenes compared against the 3D baseline [1].

Figure 10 shows some failure cases of the baseline 3D detection system [1]. In Figure 10 (b),(d) show failed 3D



(a) Overlayed detections (b) 3D detection (c) Overlayed detections (d) 3D detection

Fig. 10: Example failure cases of 3D [1] versus 2D detection. (Left column) Red, green, and yellow bounding boxes indicate ground truth, 2D detection, and 3D detection bounding boxes respectively. (Right column) Visualization of the 3D detection.

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

TABLE I: Detection accuracy comparison on both scenes. "3D" indicates [1] and "Ours" is the proposed method

detection bounding cubes and Figure 10 (a),(c) show overlayed detections. The red bounding boxes are the ground truth, 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 a 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, a point cloud of the clutter objects gets easily aggregated with the foreground object causing an erroneous oversized 3D detection. However, a 2D scanning window based framework can handle this more robustly as shown in Table I. Finally, 3D point cloud based detection fails when objects have not enough protrusion from the table e.g., scissors.

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

We now report experimental results on the the second test data set with substantially more viewpoint and scale variations and clutter as mentioned above. Section IV-C1 shows results on object detection and categorization. Section IV-C2 reports root mean squared error on pose estimation while jointly inferring object locations and category labels. Finally, section IV-C3 shows our joint visual grasp affordance estimation results.

1) Detection and categorization: We applied the same detection evaluation scheme in the previous experiment where the highest scoring detection among all locations of a given image among all the categories were considered a true positive if the bounding box overlap criterion is more than 50% [6]. For comparison, we also experimented with a baseline method where a closest matching (via ℓ_2 distance metric in SIFT feature space) training instance among the database of 10360 annotated training images are found and the labels of the nearest neighbor instance are then returned as predictions.

Mean detection accuracy was 72.22%. Figure 11 shows the confusion table for the categorization performance on correct detections (predicted bounding box overlaps more than 50% with the ground truth box). Figure 11 (Top) shows that our method confuses the eraser category as the remote control category in some cases, but generally chooses right object category labels compared to the nearest neighbor baseline (shown in Figure 11, Bottom).

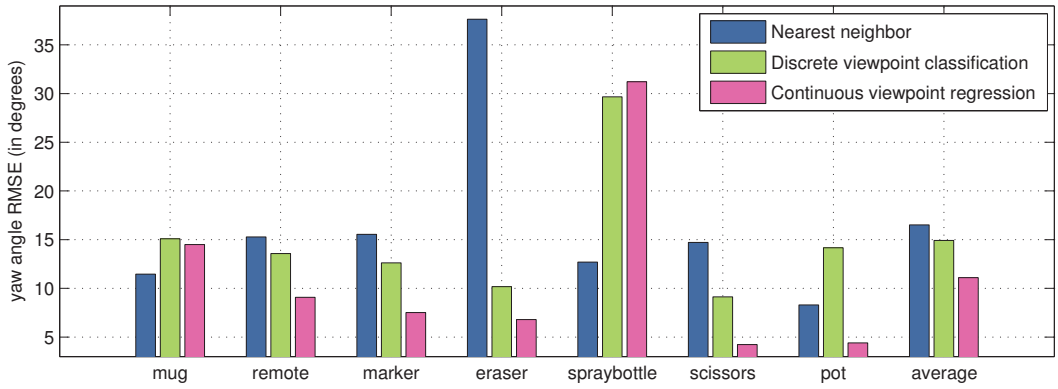
2) Multi-Category Pose Prediction: We evaluate current approaches to 3d pose estimation and investigate how they translate to our desired setting of angle accurate predictions while performing generalization to previously unseen test objects. As a baseline method we looked at a nearest neighbor approach where we compute HOG (histogram of gradients) [4] features of given test images and compare among all the 1295 images per categories(stored as HOG [4] templates) with L2 distance metric. Additionally we evaluate [12] as it is to our knowledge the state-of-the-art on the popular 3d (discrete) pose database proposed in [25] both in discrete viewpoint classification mode and in continuous viewpoint regression mode.[3].

Figure 12 shows the performance in root mean squared error of the roll and pitch angle estimations we obtain using the proposed dataset when the object location and category labels are unknown and jointly inferred as well as the object pose. As expected we observe a moderate drop when comparing the angle accurate results from [12] to our setting where we evaluate both on cross-instance and cross-category generalization. However, we can see that continuous viewpoint regression method improves the pose estimation performance over other methods on most object categories.

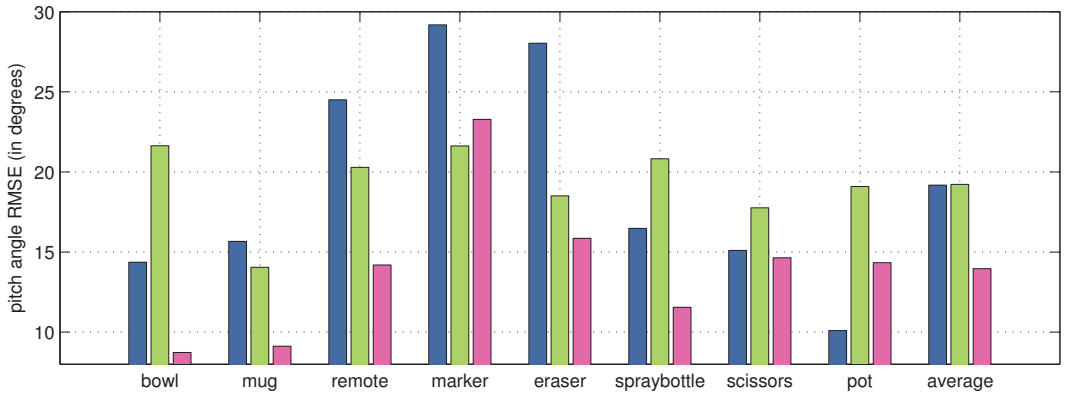
3) Visual Grasp Affordance Prediction: We now evaluate the accuracy of our joint method for grasp affordance prediction. Again, we use the proposed dataset where we have annotated graph affordances.

We investigate two scenarios as in the pose estimation experiment. The first assumes that a bounding box was provided

[3]Code was provided by the authors



(a) RMSE for yaw angle



(b) RMSE for pitch angle

Fig. 12: Accuracy of pose prediction without object locations and category labels. The top and the bottom plots show RMSE for yaw and pitch angles respectively. The last three bars show the class averaged results. Note that the yaw angle for the bowl category is omitted due to yaw angle symmetry.

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.

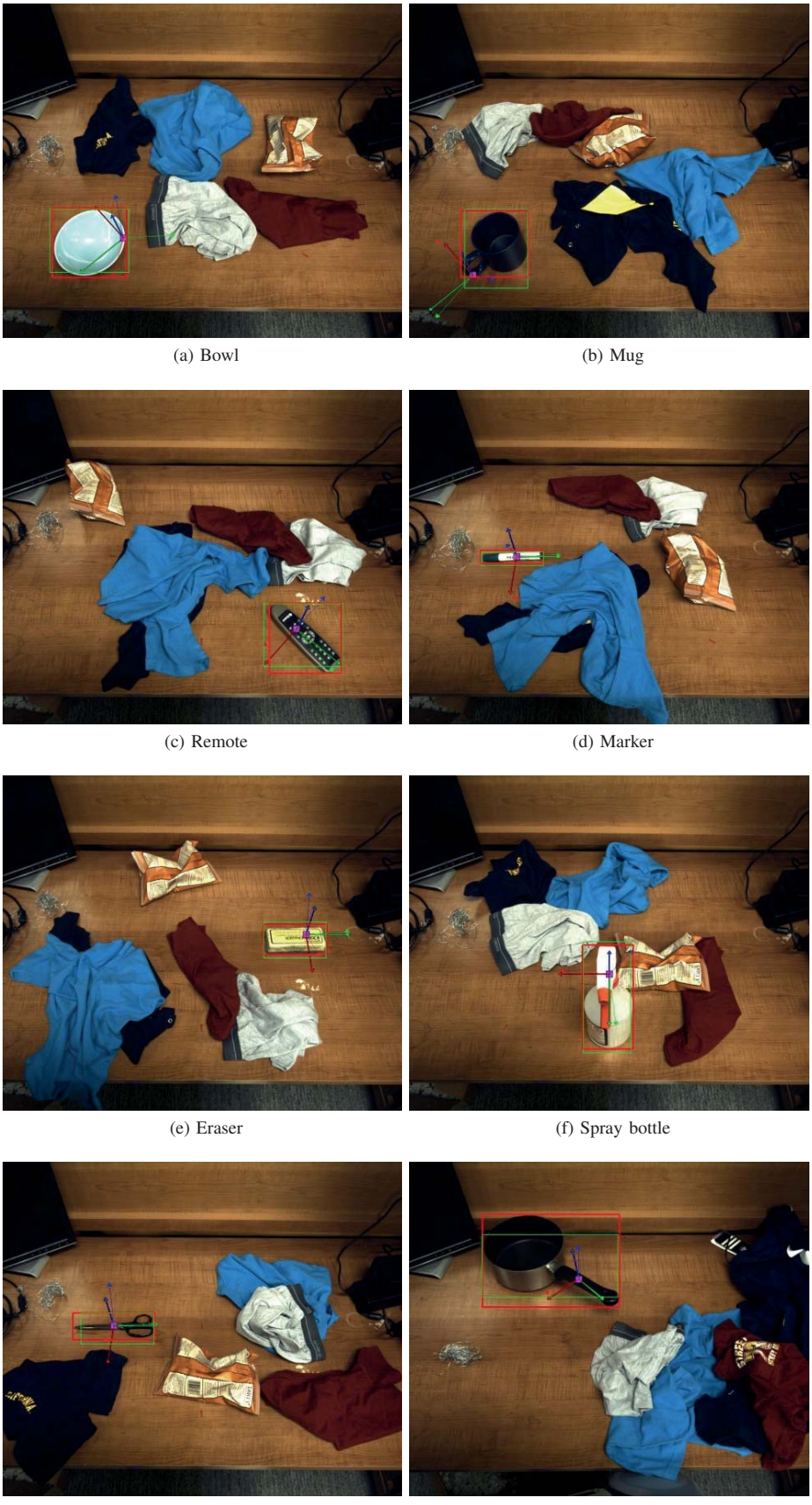
As a first baseline we compare to the results from purely local measures (tagged "Local(px)"). The approach "Global(px)" only uses the global path by predicting grasp affordances regressing from the predicted the poses conditioned on the corresponding predicted category labels. Then, we present the fused 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 about a factor of four. For comparison, [26] reports 1.80 cm metric distance error when the object locations were known. We report 0.59 cm and 0.77 cm metric distance error when the object locations were known and not known.

	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 II: Affordance prediction given groundtruth bounding box

Figure 13 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.



(a) Bowl (b) Mug

(c) Remote (d) Marker

(e) Eraser (f) Spray bottle

(g) Scissors (h) Pot

Fig. 13: Examples predictions of our framework

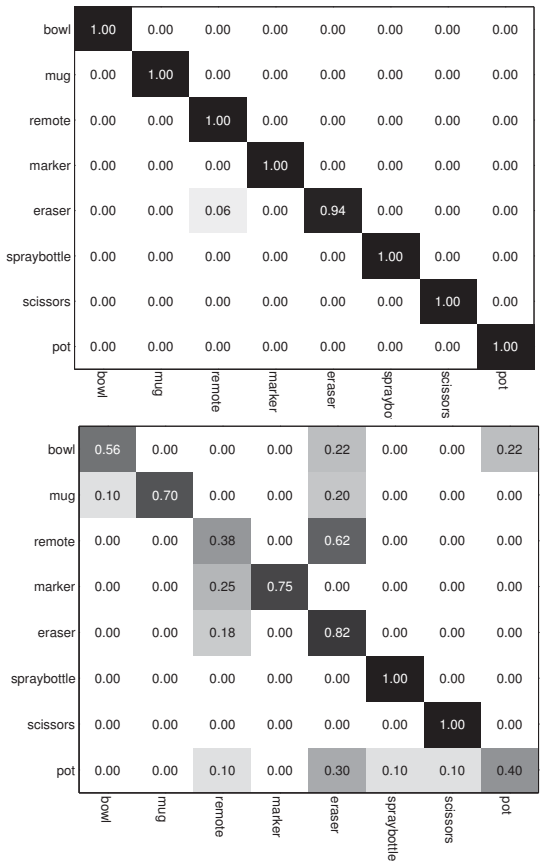


Fig. 11: Categorization confusion matrices for correctly detected objects. (Top) Our detection method. (Bottom) Nearest neighbor baseline

	Local (px)	Global (px)	Fused (px)	Fused (cm)
Bowls	65.10	39.47	28.50	1.18
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	60.85	24.93	18.53	0.77

TABLE III: Affordance prediction without bounding box and category label

4) Robot grasping experiments: For the robot grasping experiment, we placed previously unseen test objects on a cluttered table in front of PR2 robot. We designed the experiment to test how well the robot can grasp test objects in a fully autonomous setting where the robot has to first localize a test object, classify which object category it belongs to, infer the object pose, estimate the grasp affordance, and execute the grasp in collision free path.

Whenever the robot picked up the correct object at the correct position which matches the supervised grasp annotation shown in Figure 3, the experiment was counted as a success. The results of the experiments are shown in Table IV.

The visual inference (detection, categorization, pose estimation) were mostly correct for mugs but small affordance error in localizing the mug handle caused the robot to unstably grasping the handle causing grasp failures. For small and flat objects (markers, scissors) both mislocalization due to the background clutter and affordance estimate error contributed equally to grasp failures.

	Grasp success (%)
Bowls	86
Mugs	50
Remote	86
Markers	50
Erasers	86
Spray Bottles	62
Scissors	45
Pots	80
Average	65

TABLE IV: Grasp success rate

We made a video demonstration of the PR2 robot grasping the mentioned test objects at:

http://www.youtube.com/watch?v=C3HU1Tb5hF4

V. Conclusion

Appearance-based estimation of grasp affordances is desirable when other (e.g., 2.5-D or 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 localize the target object and infer grasp affordance by combining texture-based and object-level monocular appearance cues. Further, we provided a novel evaluation of max-margin pose regression on the task of category-level continuous pose estimation and a method for inferring grasp affordance from the pose estimate.

Our analysis is made possible by a novel dataset for visual grasp affordance and angle accurate pose prediction for indoor object classes. We will make our code and the dataset public to the research community to further stimulate research in this direction.

References

[1] Wilowgarage 3D Tabletop Object Detector. Tabletop object detector. willow garage, robot operating system, 2011. URL http://ros.org/wiki/tabletop_object_detector.

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

[3] N. Curtis and J. Xiao. Efficient and effective grasping of novel objects through learning and adapting a knowledge base. In *ICRA*, 2008.

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

[5] S. El-Khoury and A. Sahbani. Handling objects by their handles. In *IROSW*, 2008.

[6] 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, jun 2010.

[7] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008.

[8] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9):1627–1645, 2010.
[9] C. Ferrari and J. Canny. Planning optimal grasps. In *ICRA*, 1992.
[10] J. J. Gibson. The theory of affordance. In *Perceiving, Acting, and Knowing*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1977.
[11] C. Goldfeder, M. Ciocarlie, H. Dang, and P. K. Allen. The columbia grasp database. In *ICRA*, 2009.
[12] C. Gu and X. Ren. Discriminative mixture-of-templates for viewpoint classification. In *ECCV*, 2010.
[13] 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.
[14] Y. Jiang, M. Lim, C. Zheng, and A. Saxena. Learning to place new objects in a scene. In *IJRR*, 2012.
[15] D. Katz, A. Venkatraman, M. Kazemi, D. Bagnell, and A. Stentz. Perceiving, learning, and exploiting object affordances for autonomous pile manipulation. In *RSS*, Berlin, Germany, June 2013.
[16] Q. Le, D. Kamm, A. Kara, and A. Ng. Learning to grasp objects with multiple contact points. In *ICRA*, 2010.
[17] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *ECCVW*, 2004.
[18] I. Lenz, H. Lee, and A. Saxena. Deep learning for detecting robotic grasps. In *RSS*, Berlin, Germany, June 2013.
[19] W. Meeussen, M. Wise, S. Glaser, S. Chitta, C. McGann, P. Mihelich, E. Marder-Eppstein, M. Constantin Muja, V. Eruhimov, T. Foote, J. Hsu, R. Bogdan Rusu, B. Marthi, G. Bradski, K. Konolige, B. Gerkey, and E. Berger. Autonomous door opening and plugging in with a personal robot. In *ICRA*, pages 729–736, 2010.
[20] M. Ozuysal, V. Lepetit, and P. Fua. Pose estimation for category specific multiview object localization. In *CVPR*, 2009.
[21] R. Pelossof, A. Miller, P. Allen, and T. Jebera. An svm learning approach to robotic grasping. In *ICRA*, 2004.
[22] B. Pepik, S. Michael, G. Peter, and B. Schiele. Teaching 3d geometry to deformable part models. In *CVPR*, Providence, RI, USA, 2012.
[23] N. Ratliff, J. A. Bagnell, and S. Srinivasa. Imitation learning for locomotion and manipulation. In *IEEE-RAS International Conference on Humanoid Robotics*, 2007.
[24] E. Rivlin, S. J. Dickinson, and A. Rosenfeld. Recognition by functional parts. *CVIU*, 62(2):164–176, 1995.
[25] S. Savarese and L. Fei-Fei. 3d generic object categorization, localization and pose estimation. In *ICCV*, Rio de Janeiro, Brazil, October 2007.
[26] A. Saxena, J. Driemeyer, and A. Y. Ng. Robotic grasping of novel objects using vision. *IJRR*, 2008.
[27] A. Saxena, L. Wong, and A. Ng. Learning grasp strategies with partial shape information. In *AAAI*, 2008.
[28] L. Stark and K. Bowyer. Achieving generalized object recognition through reasoning about association of function to structure. *PAMI*, 13(10):1097–1104, 1991. ISSN 0162-8828.
[29] L. Stark, A.W. Hoover, D.B. Goldgof, and K.W. Bowyer. Function-based recognition from incomplete knowledge of shape. In *WQV93*, pages 11–22, 1993.
[30] M. Stark, P. Lies, M. Zillich, J. Wyatt, and B. Schiele. Functional object class detection based on learned affordance cues. In *ICVS*, 2008.
[31] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008.
[32] P. H. Winston, B. Katz, T. O. Binford, and M. R. Lowry. Learning physical descriptions from functional definitions, examples, and precedents. In *AAAI*, 1983.
[33] 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.

Hyun Oh Song is a postdoctoral fellow in the Computer Science Department at Stanford University. He received Ph.D in Computer Science at UC Berkeley in 2014. His graduate study was supported by Samsung Lee Kun Hee Scholarship Foundation (Now Samsung Scholarship Foundation). In 2013, he spent time at LEAR group, INRIA as a visiting student researcher. His research interests are in machine learning, optimization, and computer vision with an application focus in learning with parsimony for large scale object detection and discovery. Broadly, he's interested in solving challenging problems in artificial intelligence.

Mario Fritz is a senior researcher at the Max-Planck-Institute for Informatics and is heading a group on "Scalable Learning and Perception". He is also junior faculty at the Saarland University. His research interest are centered around computer vision and machine learning but extend to natural language processing, robotics and more general challenges in AI. From 2008 to 2011, he did his postdoc with Prof. Trevor Darrell at the International Computer Science Institute as well as UC Berkeley on a Feodor Lynen Research Fellowship of the Alexander von Humoldt Foundation. He did his PhD between 2004 and 2008 at the TU Darmstadt under the supervision of Prof. Bernt Schiele after he got his master in computer science at the University of Erlangen-Nuremberg

Daniel Goehring works in the Robotics and Intelligent Systems Group of Prof. Raul Rojas on self-driving cars. His research interests include object detection and classification, tracking, localization and sensor fusion for single or multiple robots. He was a research scholar at Prof. Darrell's Vision group at ICSI, Berkeley from 2012-2014, from 2010-2011 he joined Prof. Rojas self-driving car team at Freie Universitat Berlin, where he lead the Navigation group. He received his PhD degree in 2009 and his Diploma in 2004 from Humboldt-Universitat zu Berlin, where he worked on state estimation for four- and two-legged soccer playing robots in the A.I. group of Prof.Hans-Dieter Burkhard.

Trevor Darrell is on the faculty of the CS Division of the EECS Department at UCB and is the vision group lead at ICSI. Darrells group develops algorithms for large-scale perceptual learning, including object and activity recognition and detection, for a variety of applications including multimodal interaction with robots and mobile devices. His interests include computer vision, machine learning, computer graphics, and perception-based human computer interfaces. Prof. Darrell was previously on the faculty of the MIT EECS department from 1999-2008, where he directed the Vision Interface Group. He was a member of the research staff at Interval Research Corporation from 1996-1999, and received the S.M., and PhD. degrees from MIT in 1992 and 1996, respectively. He obtained the B.S.E. degree from the University of Pennsylvania in 1988, having started his career in computer vision as an undergraduate researcher in Ruzena Bajcsy's GRASP lab.