"Robot bring me something to drink from": object representation for transferring task specific grasps

Marianna Madry Dan Song Carl Henrik Ek Danica Kragic

***Abstract*— In this paper, we present an approach for task-specific object representation which facilitates transfer of grasp knowledge from a known object to a novel one. Our representation encompasses: (a) several visual object properties, (b) object functionality and (c) task constrains in order to provide a suitable goal-directed grasp. We compare various features describing complementary object attributes to evaluate the balance between the discrimination and generalization properties of the representation. The experimental setup is a scene containing multiple objects. Individual object hypotheses are first detected, categorized and then used as the input to a grasp reasoning system that encodes the task information. Our approach not only allows to find objects in a real world scene that afford a desired task, but also to generate and successfully transfer task-based grasp within and across object categories.**

I. Introduction

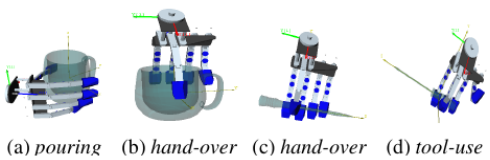
Perception of and interaction with an object is one of the key requirements for a robot acting and interacting in the environment. In this paper, we develop and evaluate an object representation that allows for linking of object, task and action information for the purpose of transferring grasping knowledge between objects that afford the same task and fulfill the same functionality. The aspect of defining the relationship between an object and its functionality is related to the concept of affordances [1], [2], [3]. In this paper, we expand this notion by incorporating requirements imposed by a given task, e.g. when pouring from an object, the fingers should not occlude the opening of the object, as presented in Fig. 1. In this context, an object representation needs to have capacity to accommodate constrains on a type of action (grasp) applied to an object.

The authors are with Computer Vision and Active Perception Lab, Center for Autonomous Systems, KTH-Royal Institute of Technology, Sweden, e-mails: madry,dsong,chek,danik@csc.kth.se. This work was supported by GRASP, IST-FP7-IP-215821 and Swedish Foundation for Strategic Research. We thank *Jeannette Bohg* for a help.

There are clearly several general requirements for an object representation, especially in a context of transferring a plausible action between different objects. It is important that an object representation remains "constant under various transformations" [3], namely has an ability to generalize over inter-class variation, as presented in Fig. 2. At the same time, many objects that facilitate different functions are hard to distinguish due to similar physical properties. An example can be a mug and a roll of toilet paper while similar in appearance clearly only the former affords pouring, see Fig. 3. This directly translates to the question: *Which object attributes should be modeled to maintain balance between discrimination and generalization properties of the representation?*

Although several object representations that relate object and action have been proposed, they encode object attributes in a data domain using relatively simple features [4], [5], [6], [7] and describe object attribute based on a single modality [2], [8], [7], [9]. Intuitively, the most effective approach is to capture various object properties using different modalities, for example object shape based on 2D and 3D data.

Our aim is to leverage on recent advances in the object representation to show how this can facilitate transfer of grasp knowledge to a novel object. We approach this problem in two steps, see Fig. 4. First, we obtain semantic information about an object category defined by its physical attributes (appearance, color, shape) using RGB images (2D) and point cloud data (3D). Evidences for the object category, obtained separately for each of the attributes, are integrated to keep balance between discrimination and generalization. Second, we use a probabilistic model to infer: (a) object function, e.g. if it afford pouring, and (b) make detailed decisions on sensorimotor level, e.g. plan grasps that afford pouring.



(a) *pouring* (b) *hand-over* (c) *hand-over* (d) *tool-use*

Fig. 1. Grasping a cup: (a) *pouring* and (b) *hand-over* task (hand should not block the opening), and a screwdriver: (c) *hand-over* and (d) *tool-use* task (hand should grasp the handle).

Fig. 2. A class of objects that afford the same task (pouring) and is characterized by high inter-class variations. It imposes a requirement for an object representation to have high generalization power.

Fig. 3. Examples of physically similar objects that afford different tasks. It imposes a requirement for an object representation to have high discrimination power.

In summary, we demonstrate that, by using an approach that combines information about an object, functionality and action, the robot can not only choose the objects in a real 3D scene that afford the assigned task, but also plan the grasp such that it satisfies the constraints posed by the task. Thus, grasp knowledge can be transferred between objects that belong to the same category, though the details of the geometry and physical properties vary.

II. Related Work

The notion of object categories is useful for task-related grasping: for humans, it is natural to use and manipulate objects based on their functionality and current task [1]: when using a screwdriver as a tool, the fingers should be placed at the handle, see Fig. 1. The knowledge of how to grasp an object can be transferred between objects that belong to the same category. In [10] and [6], we developed a probabilistic framework using Bayesian network to represent the task-related grasping. The task requirements are encoded through the conditional dependencies between a task variable and a set of object and grasp features. This work was done in a simulation environment and the inference engine assumed the object class unknown. Learning of the network structure in [6] revealed the importance of a proper choice of an object representation for an accurate transfer of task specific grasp to a novel object.

Many household objects belonging to the same functional category differ significantly in physical properties (Fig. 2) and objects affording different tasks are alike in color and shape (Fig. 3). Thus, an object representation needs, at the same time, to ensure discrimination between the categories and handle high within-class variations. We predict that a representation encoding various object properties originating from different modalities (e.g. 2D and 3D data) will be the most effective. In [11] authors presented an hierarchical classification system where 3D descriptor is used to narrow choice of objects to those of similar shape specified by 2D descriptor. However, in this approach the grasp hypothesis is given for an object instance, not category.

Several works have been devoted to relate an object with a performed task at the object category level. Grasp affordances for local regions of an object have been inferred based on object appearance in 2D (SIFT descriptor [2], [9], saturation channel filters [12]), or geometric shape properties in 3D (hexahedron [7], edge segments [8]). The approach presented in [8] was further extended to represent spatial relationship between local edge position and orientation in a hierarchical manner [13]. Finally, global 3D object attributes located in the input data domain (object size, volume, convexity, symmetry) have been used to learn object affordances based on both action and function [4], [5], [6].

III. Object Representation

We combine recent advances in object representation to enable transfer of task-constrained grasp knowledge between objects that belong to the same category defined by their physical properties. Our Object Categorization System (OCS) integrates descriptors of appearance (e.g. texture) , color and shape using both RGB images (2D) and point cloud (3D) data. As shown in Fig. 4 (middle column), we train a separate OCS for each descriptor and then fuse evidences from a few OCSs to obtain the final categorization. We discuss as follows: feature extraction, classification and aspects of integration of the single cue OCSs.

A. Feature Extraction

There are many descriptors that encode object appearance (SIFT [14], textones [15]), color (opponentSIFT [16]) and contour shape (HoG [17]). Studies on 2D cue integration [18] show that contour- and shape-based methods are adequate for handling the generalization requirements needed for object categorization, however they are not robust to occlusions. On the other hand, appearance- and color-based descriptors have been successfully applied in object instance recognition and detection [14], [15]. However, their performance drops significantly with clutter and illumination changes. Also, different 3D shape descriptors have been proposed [19]. Only a few of them are applicable to real 3D data that covers only the visible part of the object: spin images [20], Fast Point Feature Histograms (FPFH) [21], or Radius-based Surface Descriptor (RSD) [22].

Motivated by the fact that the object representation should have high discrimination and generalization power, in order to be robust to real world condition and diverse for cue integration, we extract from a segmented part of an image multiple 2D descriptors encoding different object attributes: appearance (SIFT), color (opponentSIFT), contour shape (HoG). The final object representation for 2D descriptors follows a concept of the spatial pyramid [23]. The 3D shape properties of an object are obtained by applying the FPFH [21] and RSD [22] descriptors to a point cloud. It was shown that the normal-based descriptors obtain high performance for an object categorization [24]. To obtain the final object representation, the Bag-of-Words (BoW) model [25] is employed.

B. Classification

For classification, we use SVMs with a χ^2 kernel successfully applied in previous studies [16][17][11] for the histogram-based object representations. For cue integration, we need information about the confidence with which an object is assigned to a particular class. Several studies have been devoted to find confidence estimates for large margin classifiers [26]. In principle, they interpret the value of the discriminative function as a dissimilarity measure between the sample and the optimal hyperplane. In this work, we use the One-against-All strategy for M-class SVMs which was shown to be superior to other methods [27].

C. Cue integration

Various cue integration approaches have been applied to object classification based on 2D data [26][28]. In contrast to the *low level integration* that operates directly on feature

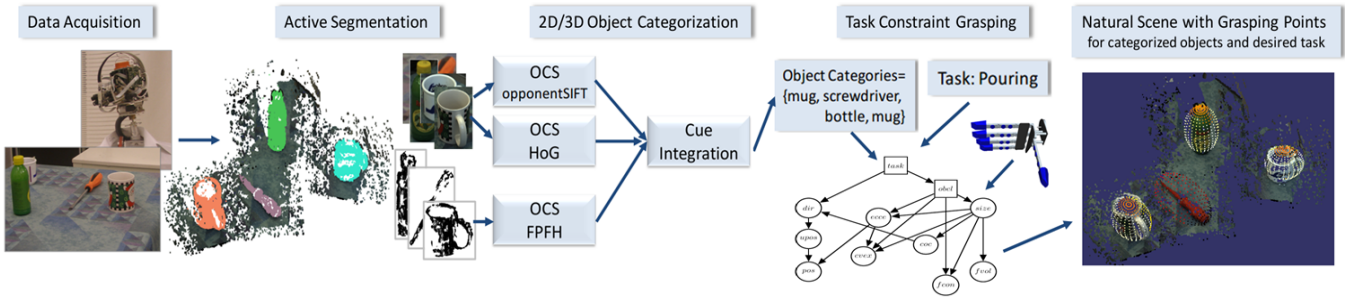


Fig. 4. Visual Object Category-based grasp generation for an arbitrary scene: objects are first segmented and categorized using our 2D-3D Object Categorization Systems (OCSs). Then, grasping hypotheses are generated taking the task into account. The image is best viewed in color.

vectors, the *high level integration*, that is commonly accomplished by an ensemble of classifiers or experts, have been shown to be more robust to noisy cues. Further, the classifier outputs can be combined using *linear* [26] or *nonlinear* [28] techniques. However, nonlinear methods requires a larger training dataset to the estimate relatively complex relationship between parameters. In a case of a limited amount of training data this may lead to a drop in performance. We have observed this behavior for our application. Results can be found in [29].

In this work, taking a high level approach we combine information from the single cue OCSs. Total support for each class is obtained as a weighted sum, product or max function of the evidences provided by individual classifiers and function parameters are estimated during training. The final classification decision is made by choosing the class with the strongest support.

For details about the described Object Categorization System, we direct the reader to our previous work [29].

IV. Encoding Task Constraints

In our previous work [10][6][30], we developed a probabilistic framework for embodiment-specific grasp representation. We model the system as a efficient Bayesian network exploiting conditional dependencies between task, object, action and constraints. The model is trained using a synthetic database of objects, generated grasps, and the task labels provided by a human. We refer the reader for the detailed process of data generation to [10].

Both the structure and the parameters of the BN are learned from the database. The BN structure encodes dependencies among the set of task-related variables, and the parameters encode their conditional probability distributions. Fig. 5 shows the learned structure of the BN with the features listed in Table I. Once trained, the model can be used to infer conditional distribution of any subset of variables based on a partial or complete observation of others. This allows us to select object (e.g. by $P(obcl|task)$) and plan grasp (e.g. by $P(pos, dir|task)$) in a task-oriented manner.

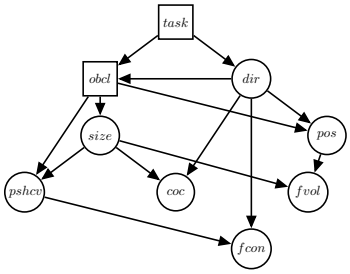


Fig. 5. The structure of the Bayesian network task constraint model.

V. Experimental Evaluation

We first describe the dataset and experimental setup followed by the evaluation of several descriptors, encoding various object properties, and their integration. Finally, we demonstrate the results of transferring a prior grasp information to a novel object.

A. Database

Most available 2D-3D object datasets contains a limited choice of object categories suitable for our purpose [11][31][32]. Therefore, we collected a new database – the Stereo Object Category (SOC) database [29], where a number of objects with similar physical properties, afford different tasks, see Fig. 3. In order to capture variations in appearance, shape and size within each class, various objects were selected for each category.

The SOC database contains RGB-D data collected using the 7-joint Armar III robotic head equipped with two foveal and peripheral cameras. To differentiate an object and background, an active segmentation method was used [33]. The database includes 14 object categories: *ball, bottle, box, can, car-statuette, citrus, mug, 4-legged animal-statuette, mobile, screwdriver, tissue, toilet-paper, tube and root-vegetable*, each with 10 different object instances per category. For each object, both 2D (RGB image) and 3D (point cloud) data were collected from 16 different views around the object (separated by 22.5°). Additionally, there is a choice of data collected in natural scenes. A few subjects were asked to randomly place between 10 to 15 objects from 14 different

TABLE I

Features used for the Task Constraint Bayesian network.

Name	Dimension	States	Description
$task$	-	5	Task Identifier
$obcl$	-	7	Object Category
$size$	3	6	Object Dimensions
dir	4	15	Approach Direction (Quaternion)
pos	3	17	Grasp Position
$fcon$	11	3	Final Hand Configuration
$pshcv$	3	3	Grasp Part Shape Vector
coc	3	8	Center of Contacts
$fvol$	1	4	Free Volume

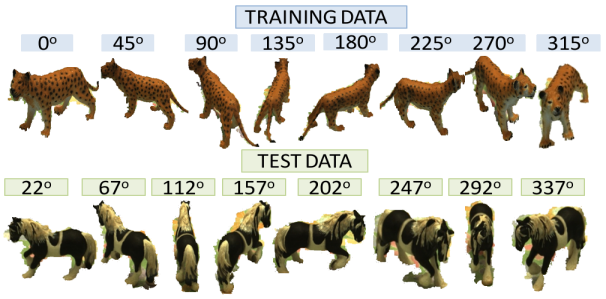


Fig. 6. An experimental setup where eight views per object are selected to train an object representation and the remaining eight views for its evaluation. Data for all objects and natural scenes can be viewed at: `http://www.csc.kth.se/~madry/research/stereo_database`.

TABLE II

RESULTS FOR THE FEATURE SELECTION EXPERIMENTS.[1]

Descriptor	SIFT	oppSIFT	HoG	RSD	FPFH
Av.Categ.Rate	83.3%	83.9%	63.3%	58.9%	69.7%
σ	2.9%	3.4%	2.8%	1.8%	1.9%

categories on a table. As a result, objects poses, scale and degree of occlusion vary significantly. We recorded data for 10 natural scenes, see examples in Fig. 8.

B. Experimental Setup

For each experiment, we performed cross-validation with the data divided into four sets: (1) training, (2) validation of OCS parameters, (3) validation of the cue integration parameters, and (4) testing. Objects were randomly selected for each set with the ratio 4:1:1:4 objects per category. In total, 56 objects were used for training and testing, and 14 objects for validations. Rotation of the objects in training and testing differs. As depicted in Fig. 6, eight views per object are selected to train the representation and the remaining eight views for its evaluation. Our aim is to test the performance of the system for categorization and not object instance recognition, an object used in the training phase was never again used for evaluation.

C. Object Representation

We built five identical single cue OCSs, one for each descriptor, to evaluate the performance of the descriptors for encoding different object properties: appearance (SIFT), color (opponentSIFT), contour shape (HoG) and 3D shape (RSD and FPFH).

1) Feature Selection: As presented in Table II, the best performance was obtained for SIFT and opponentSIFT. This indicates that appearance and color information is less affected by viewpoint changes than shape information. Further, FPFH yielded a higher categorization rate than HoG as a result of degradation of an object shape by projective transformation in 2D data. FPFH showed to be more descriptive feature than RSD.

2) Feature Integration: When combining different features, the best performance was obtained for integration of the three descriptors: opponentSIFT+HoG+FPFH. Confusion matrices presented in Fig. 7 show that by capturing diverse object properties (appearance, contour and 3D shape) origination from different sensors (2D and 3D) not only significantly improve robustness of the categorization system, but is essential to discriminate between similar objects that afford different tasks. Such classification is very challenging for a system based on a single cue.

3) Natural Scenes: We evaluated performance of the 2D-3D object representation on 10 natural scenes. For categorization we chose the best classifier trained following the procedure described in Section V-B. The final label was found by choosing the category with the highest confidence value. The categorization results for a few scenes together with a confidence vector for each object are presented in Fig. 8. The system yielded a high categorization rate of 91.7% in spite of occlusions or inaccurate segmentation. It is capable to operate in a very challenging scenario.

D. Object Category-based Task-constrained Grasping

In this section, we present results for transferring task specific grasp experience to a novel object. As shown in Fig 9, the robot faces a scene containing several unknown objects. For each segmented object hypothesis a category label defined by object physical properties is assigned. In the given scene, 13 objects were found, all correctly classified. Next, the robot needs to decide: (1) which object should be grasped given the assigned task, and (2) how to perform the grasp to fulfill the task requirements. The probabilistic reasoning system is trained on a grasp database that includes stable grasps generated on a set of synthetic object models using the hand model from the humanoid robot Armar [34]. The object models are extracted from the Princeton Shape Benchmark [31] with 3-8 models per category, Five tasks were labeled: *hand-over*, *pouring*, *dishwashing*, *playing* and *tool-use*. The total training set includes 1227 cases with 409 cases per grasping task.

1) Grasp Transfer: We infer the most suitable grasp parameters given the object category $obcl$ and assigned task $task$. A grasp is parameterized by multiple variables: dir, $fcon$ and pos, see Tabel I. We illustrate the results on a grasp position pos, i.e. a direction from which the hand is placed relative to the object. For each object, we sample a set of points on an ellipsoid which size is determined by the pos data, and infer the likelihood of each pos point conditioned on $obcl$ and $task$, $P(pos|obcl, task)$. The resulting likelihood maps for $obcl = mug$ and $task = pouring$ are presented in Fig. 9.

The pos variable is represented in the synthetic object local coordinate system. In order to transfer grasp information to an arbitrary object in the scene, it is necessary to convert the pos data from the local object frame to the world coordinates. This transformation requires the knowledge of the size, position and orientation of the object. In this paper, we assume orientation to be known. The size and position are determined by estimating a minimum bounding sphere of the filtered 3D point cloud. We assume that the diameter of the bounding sphere corresponds to the largest object dimension.

[1] In [29] results have been obtained for a different set of experiments.

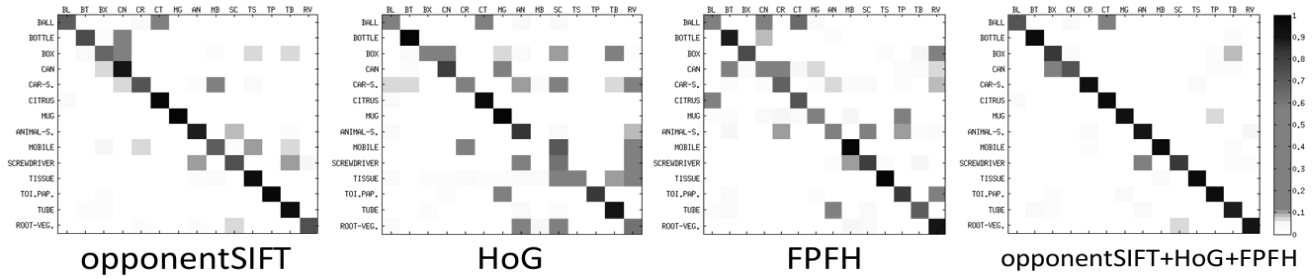


Fig. 7. Confusion matrices obtained for: (a) color (opponentSIFT), (b) contour shape (HoG), (c) 3D shape (FPFH) descriptor, and (d) integrated opponentSIFT+HoG+FPFH (linear combination method, sum rule). Object representation based on integrated descriptors increased simultaneously categorization rates for several objects classes characterized by similar properties, comparing to the single cue OCSs, such as: (a) shape, as *screwdriver* and *root-vegetable* where only the former can be used as a tool, *ball* and *citrus* where only the former affords playing; (b) appearance, *bottle* vs. *can*.

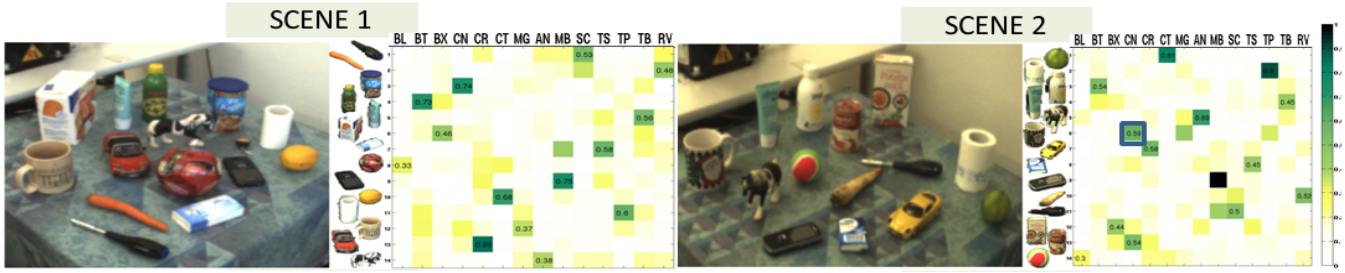


Fig. 8. Categorization results for natural scenes. For each object in a scene, the confidence values over 14 categories are shown. All objects were correctly classified except an object marked using a blue square in confidence vector.

Several examples of grasp transfer to the real objects are presented in Fig. 9.

2) Task-constrained Grasping in a Real Scene: Fig. 10 shows the results of the experiment for a natural scene. We show the likelihood maps for each object using colored sample points of $P(pos|task, obcl)$. For the *pouring* task, the likelihoods of the sample points around the mugs and bottle are clearly higher than for other objects indicating that they are the only objects affording the task. Similarly, *screwdriver* is the only objects that can be used as a tool. For the *hand-over* task, all objects have high likelihood. This indicates that by using the representation that relates object category and functionality, we can successfully select the object according to their task accordance. For the objects that afford *pouring*, for example *mug*s, the likelihood maps show darker color on the top of the object. This is because the robot hand should not block the opening of an object when pouring a liquid. When using the *screwdriver* as a tool, the network favors the position around the tip of the screwdriver whereas leaving the handle part for regrasp.

VI. Conclusions

We presented a framework capable of transferring of grasping knowledge between objects that share similar physical attributes and/or have the same functionality. We demonstrated that choosing an object representation that encodes diverse objects properties (color, contour, 3D shape and appearance) and integrates information from different visual sensors (2D and 3D), not only significantly improve robustness of the categorization system, but assures relevant balance between discrimination and generalization in the representation. This means that we can distinguish objects that both belong to the same functional category, but significantly differ in physical properties, and objects that afford different tasks, but are alike in color and shape. To summarize, the proposed framework enables reasoning and planning of goal-directed grasps in real world scenes with multiple objects allowing to execute the command "Robot bring me something to drink from".

References

[1] J. G. Greeno, "Gibson's Affordances," *Psychological Review*, 1994.
[2] G. Fritz, L. Paletta, R. Breithaupt, E. Rome, and G. Dorffner, "Learning predictive features in affordance-based robotic systems," in *IROS*, 2006.
[3] E. Sahin, M. Cakmak, M. Dogar, E. Ugur, and G. Ucoluk, "To afford or not to afford: A new formalization of affordances towards affordance-based robot control," *ISAB*, 2007.
[4] L. Montesano, M. Lopes, A. Bernardino, and J. Santos-Victor, "Learning object affordances: From sensory–motor coordination to imitation," *T-RO*, 2008.
[5] A. Saxena, L. Wong, and A. Y. Ng, "Learning Grasp Strategies with Partial Shape Information," in *AAAI*, 2008.
[6] D. Song, C.-H. Ek, K. Huebner, and D. Kragic, "Multivariate discretization for bayesian network structure learning in robot grasping," in *ICRA*, May 2011.
[7] 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 *ICAR*, 2009.
[8] M. Popović, D. Kraft, L. Bodenhagen, E. Başeski, N. Pugeault, D. Kragic, T. Asfour, and N. Krüger, "A strategy for grasping unknown objects based on co-planarity and colour information," *RAS*, 2010.
[9] C. Goldfeder, M. Ciocarlie, J. Peretzman, H. Dang, and P. Allen, "Data-driven grasping with partial sensor data," in *IROS*, 2009.
[10] D. Song, K. Huebner, V. Kyrki, and D. Kragic, "Learning Task Constraints for Robot Grasping using Graphical Models," in *IROS*, 2010.
[11] Z.-C. Marton, D. Pangercic, R. B. Rusu, A. Holzbach, and M. Beetz, "Hierarchical object geometric categorization and appearance classification for mobile manipulation," in *Humanoids*, 2010.
[12] L. Montesano and M. Lopes, "Learning grasping affordances from local visual descriptors," in *ICDL*, 2009.
[13] R. Detry, N. Pugeault, and J. Piater, "A probabilistic framework for 3D visual object representation," *TPAMI*, 2009.

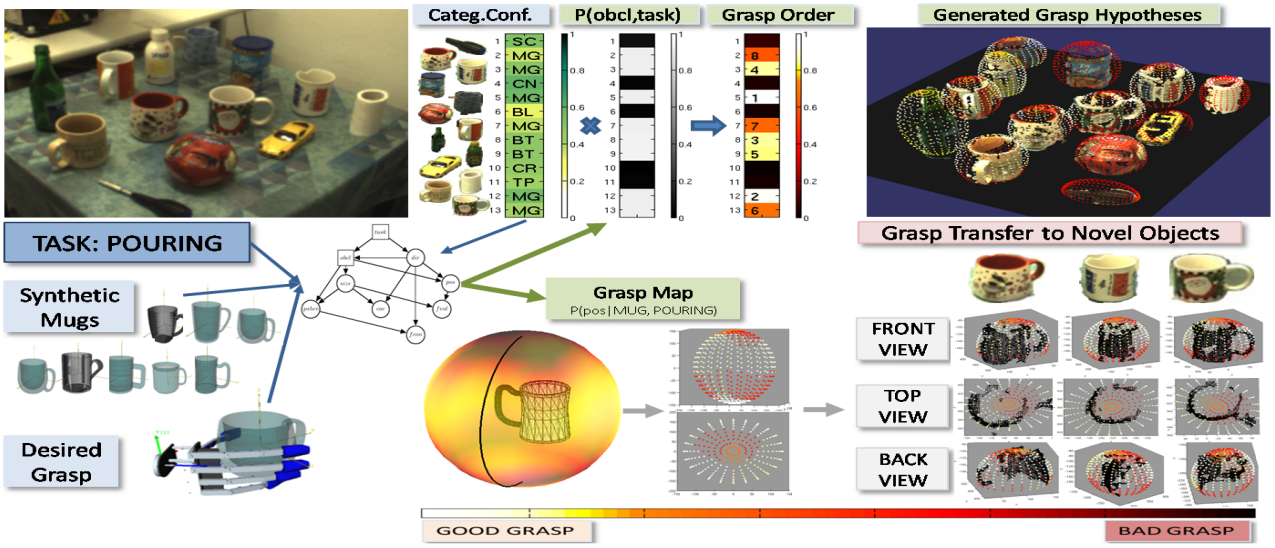
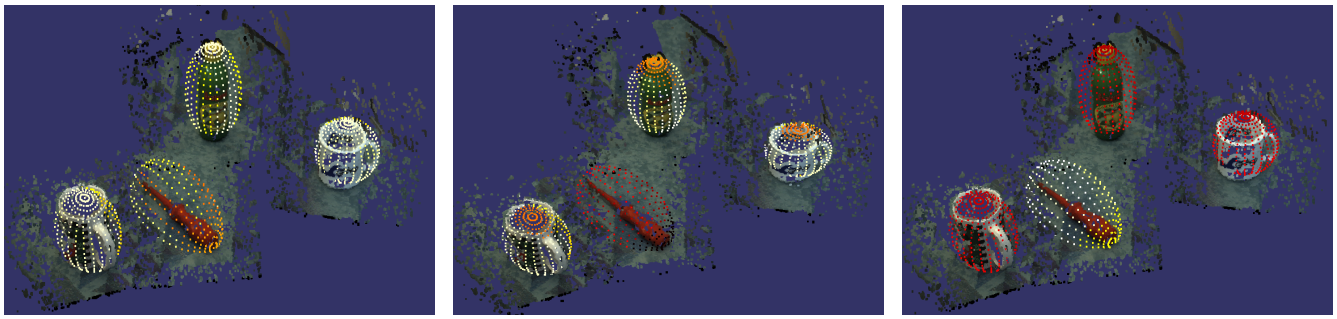


Fig. 9. Grasp transfer from a synthetic object model to real objects in a scene (best viewed in color). The grasping points with the highest value of $P(pos|obcl, task)$, indicated by the brightest color, implies the best grasp position for the task. For each object classified as a *mug*, a set of grasping points is presented in the front (camera), top and back view. By transferring the grasp map, we are able to generate grasp points for the back (not visible) part of an object without reconstructing the full object shape. The order in which objects should be grasped given a task is determined as a product $C \cdot P(obcl, task)$ where C is object categorization confidence.



(a) Hand-over (b) Pouring (c) Tool-use

Fig. 10. Grasp hypotheses and associated probabilities for three different tasks: (a) hand-over, (b) pouring and (c) tool-use. The grasping probability around an object is indicated by color of a point. The brighter is the point, the higher is the probability. The images are best viewed in color. Experimental results for a number of natural scenes and five different tasks are available on our website `http://www.csc.kth.se/~madry/research/madry12icra`.

[14] G. D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, 2004.

[15] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *CVPR*, 2008.

[16] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *PAMI*, 2010.

[17] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.

[18] B. Leibe and B. Schiele, "Analyzing appearance and contour based methods for object categorization," in *CVPR*, 2003.

[19] J. W. H. Tangelder and R. C. Veltkamp, "A survey of content based 3D shape retrieval methods," in *SMI*, 2004.

[20] A. Johnson, "Spin-images: A representation for 3-D surface matching," Ph.D. dissertation, Carnegie Mellon University, 1997.

[21] R. B. Rusu, N. Blodow, and M. Beetz, "Fast Point Feature Histograms (FPFH) for 3D Registration," in *ICRA*, 2009.

[22] Z.-C. Marton, D. Pangercic, N. Blodow, J. Kleinehellefort, and M. Beetz, "General 3D modelling of novel objects from a single view," in *IROS*, 2010.

[23] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, 2006.

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

[25] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on Statistical Learning in Computer Vision (ECCV)*, 2004, pp. 1–22.

[26] M. E. Nilsback and B. Caputo, "Cue integration through discriminative accumulation," in *CVPR*, 2004.

[27] A. Pronobis and B. Caputo, "Confidence-based cue integration for visual place recognition," in *IROS*, 2007.

[28] A. Pronobis, O. M. Mozos, and B. Caputo, "SVM-based discriminative accumulation scheme for place recognition," in *ICRA*, 2008.

[29] M. Madry, D. Song, and D. Kragic, "From object categories to grasp transfer using probabilistic reasoning," in *ICRA*, 2012, to appear.

[30] D. Song, C. H. Ek, K. Huebner, and D. Kragic, "Embodiment-Specific Representation of Robot Grasping using Graphical Models and Latent-Space Discretization," in *IROS*, 2011.

[31] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser, "The Princeton Shape Benchmark," in *SMI*, 2004, pp. 167–178.

[32] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view RGB-D object dataset," in *ICRA*, 2011.

[33] M. Bjorkman and D. Kragic, "Active 3D scene segmentation and detection of unknown objects," in *ICRA*, 2010.

[34] T. Asfour, K. Regenstein, P. Azad, J. Schröder, A. Bierbaum, N. Vahrenkamp, and R. Dillmann, "ARMAR-III: An integrated humanoid platform for sensory-motor control," in *Humanoids*, 2006.