From Object Categories to Grasp Transfer Using Probabilistic Reasoning

Marianna Madry, Dan Song and Danica Kragic

***Abstract*— In this paper we address the problem of grasp generation and grasp transfer between objects using categorical knowledge. The system is built upon an i) active scene segmentation module, able of generating object hypotheses and segmenting them from the background in real time, ii) object categorization system using integration of 2D and 3D cues, and iii) probabilistic grasp reasoning system. Individual object hypotheses are first generated, categorized and then used as the input to a grasp generation and transfer system that encodes task, object and action properties. The experimental evaluation compares individual 2D and 3D categorization approaches with the integrated system, and it demonstrates the usefulness of the categorization in task-based grasping and grasp transfer.**

I. INTRODUCTION

Household environments pose serious challenges to robotic perception and manipulation: objects are difficult to locate and manipulate due to the unstructured settings, variable lighting conditions and complex appearance properties. Although some excellent examples of finding and manipulating a *specific* object in a scene have been reported in the literature [1][2], the aspect of *generalization* have not been addressed properly. No system is capable of flexibly and robustly, in realistic settings, finding objects that *fulfill* a certain functionality thus executing tasks such as "**Robot, give me something to hammer with.**" or "**Robot, bring me something to drink from.**"

The aspect of function is related to that of affordances [3] and has been addressed frequently in works that learn relations between objects and actions [4][5][6][7]. However, none of these consider the aspect of *task* in their model: what the agent wants to do with an object will affect the type of action (grasp) to apply. In this case, the task will be constraining the action space - not just any grasp can be applied on the object, see Fig. 1. Another closely related example is finding something to *hammer-with* or *pour-to* that relates to the notion of functional categories that have been addressed to a limited extent in computer vision [8][9].

In this paper, we present work on encoding object categorical knowledge with task and action related reasoning. Knowledge of object category facilitates action (grasp) transfer: i) detecting an object that affords pouring may be pursued at the categorical level, or ii) knowing how to grasp an object that affords pouring may be transferred to another object that belongs to the same category. We build upon our previous work [10][11][12], where we developed a probabilistic grasp reasoning system. The system models the conditional dependencies between the tasks, actions and objects taking into account the constraints posed by each. In previous work, we concentrated on theoretical problems of structure learning in graphical models without considering the aspect of *real* sensory information extracted in natural scenes.

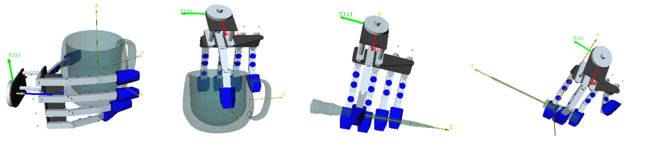

(a) *pouring* (b) *hand-over* (c) *hand-over* (d) *tool-use*

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).

In this paper we present an integrated approach to task-oriented grasp reasoning and categorization, with the novel aspect of grasp transfer. The contributions of the proposed system are that:

- we enable a robot to choose the objects in a 3D scene that afford the assigned task while
- planning the grasp that satisfies the constraints posed by the task;
- grasp knowledge can be transferred between objects that belong to the same category, even under considerable differences in appearance and physical properties.

Our system integrates 2D and 3D visual information and captures different object properties (appearance, color, shape) thus making the categorization process robust in a real-world scenario. We show that the system can successfully discriminate between objects sharing similar properties but affording different tasks, such as a carrot and a screwdriver that are structurally similar but fulfill different functions (see Fig. 4).

The paper is organized as follows: In Sec. II we present the probabilistic reasoning framework and in Sec. III object categorization system. Sec. IV outlines the experimental evaluation and Sec. V concludes the paper.

A. The system

Our system consists of three main parts, see Fig. 2:

- Visual Front End: here, an active robot head equipped with foveal and peripheral cameras gives input to the real-time scene segmentation system [13];
- Categorization: the system provides information about object class using various object properties such as appearance, color and shape;
- Reasoning system: the probabilistic grasp reasoning system, that encodes task-related grasping [10][11].

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

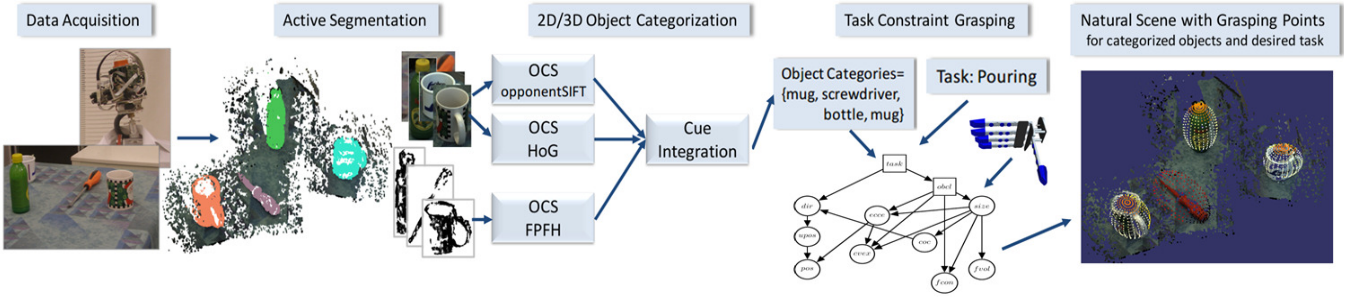


Fig. 2. Visual Object Category-based grasp generation for an arbitrary scene: objects are first segmented and categorized using our 2D-3D Object Categorization Systems (OCSs). Then, grasping hypotheses are generated taking the task into account. The image is best viewed in color.

We start by providing the necessary details for our probabilistic reasoning system.

II. Encoding Task Constraints

In the previous work [10][11][12], we have developed a probabilistic framework for embodiment-specific grasp representation. We model the conceptual task requirements using a Bayesian network through conditional dependencies between task, object, action and constraints posed by each. The model is trained using a synthetic database of objects, grasps generated on them, and the task labels provided by a human. The data generation is based on the toolbox BADGr [14] providing 3D object shape approximation, grasp planning, execution and also grasp-related feature extraction and task labeling. 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. 3 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 one variable 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.

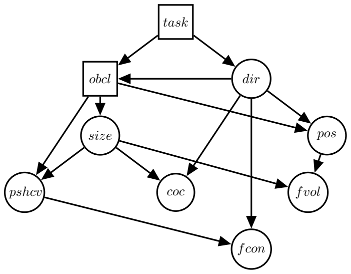


Fig. 3. The structure of the Bayesian network task constraint model.

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

Our previous work was done in simulation and the inference engine assumed the object class unknown. Learning of the network structure in [11] revealed the importance of the categorical information. This motivated us to integrate the object categorization module with the task-constraint grasp reasoning system.

III. 2D-3D Object Categorization System

Many household objects that afford different tasks have similar shape or appearance properties making them hard to discriminate. For example, a mug and a roll of toilet paper are alike in shape, but only the former object affords pouring a liquid to (see Fig. 4). Thus, our Object Categorization System (OCS) integrates visual descriptors capturing different object properties such as appearance, color, shape using both RGB images (2D) and point cloud data derived from disparity maps (3D).

As shown in Fig. 1, we first build a single cue OCS for each feature descriptor which are then integrated for the final categorization. All single cue OCSs implement the following methodology: (a) data acquisition (Sec. III-A), (b) feature extraction (Sec. III-B), and (c) classification (Sec. III-C). The methods used to integrate these single cue OCSs will be described in Sec. III-D.

A. Scene Segmentation

Prior to categorization, object hypotheses are first generated using a multi-cue scene segmentation system [13]. The method relies on attentional mechanisms to direct cameras towards regions of interest, subsequently grouping areas close to the center of fixation as the foreground. Points of the disparity maps are then labeled as either the object (foreground), supporting plane (flat surface) or the background.

The segmented point cloud is further processed to remove outliers and equalize point density. We rely on the statistical outlier removal and voxel grid filters from PCL [15]. The resulting point cloud contains approx. 2000 points representing the visible part of the object. Our system does not require reconstruction of the whole object from its partial view as in [16][17]. Such reconstruction methods often assume objects to be symmetrical which is not always the case.

B. Feature Extraction

The object representation is crucial for achieving robust categorization. Several descriptors have been proposed in the field of computer vision to encode object appearance

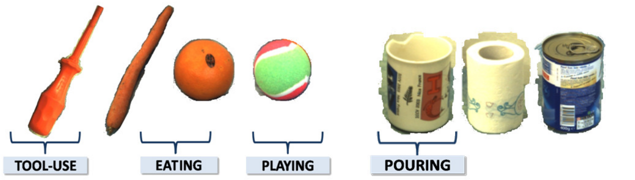


Fig. 4. Examples of physically similar objects that afford different tasks.

(SIFT [18], textones [19]), color (opponentSIFT [20]) and contour shape (HoG [21]). Studies on 2D cue integration [22] 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 [18], [19]. However, their performance drops significantly in case of clutter and illumination changes. In object retrieval and computer graphics, a number of 3D shape descriptors have been proposed [23]. Only a few of them are applicable to real 3D data that covers only the visible part of the object: spin images [24], RSD [16], FPFH [25][26].

Motivated by the fact that the object representation should have high discriminative power, 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 [27]. The 3D shape properties of an object are obtained by applying the FPFH descriptor [25] to each 3D point in the segmented point cloud. It was shown that the normal-based descriptors obtain high performance for the task [26]. To obtain the final object representation, a *bag-of-words* BoW model [28] is employed.

C. Classification

Motivated by the histogram-based object representation (BoW), we use for classification SVMs with a χ^2 kernel successfully applied in previous studies [20][21][17]. For the purpose of cue integration, we need information about the confidence with which an object is assigned to a particular class. Several studies were devoted to find confidence estimates for large margin classifiers [29]. 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 [30]. The confidence measure for a sample $\mathbf{x}$ is calculated as:

$$C(\mathbf{x}) = D_j * (\mathbf{x}) - \max_{j=1\ldots M, j \neq j*} \{D_j(\mathbf{x})\} \qquad (1)$$

where $D_j(\mathbf{x})$ is equal to the difference between the average distance of the training samples to the hyperplane and the distance from $\mathbf{x}$ to the hyperplane.

D. Cue Integration

Various cue integration approaches have been applied to object recognition and categorization based on 2D data [29][31]. In contrast to the *low level integration* that operates directly on feature 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* [29] or *nonlinear* [31] techniques.

Our object categorization system takes a high level approach integrating evidences from the single cue OCSs. We use methods based on an combination of classifier outputs. We evaluate both the linear and nonlinear algebraic techniques.

In case of the linear techniques, the total support for each class is obtained as a linear weighted sum, product or max function $F(\cdot)$ of the evidences provided by individual classifiers. The final decision is made by choosing the class with the strongest support. Let us assume that d_{ij} is an evidence provided by classifier i for a category j, and w_i is a weight for classifier i (both are normalized to sum up to one for all L classifiers and M categories), then the class with the strongest support $j_0 \in \{1, \ldots, M\}$ is chosen as:

$$j_0 = \arg \max_{j=1,\ldots,M} \frac{F(d_{1j}, \ldots, d_{Lj}; w_1, \ldots, w_L)}{\sum_{j=1}^{M} F(d_{1j}, \ldots, d_{Lj}; w_1, \ldots, w_L)}. \qquad (2)$$

The weights $w_i|_{i=1,\ldots,L}$ are estimated during training.

In case of the nonlinear techniques, we have used an approach where an additional SVM classifier is trained to model the relation between evidences provided by the different single cue OCSs [31]. The outputs from the single cue OCSs are concatenated to build a feature vector that is fed to the subsequent SVM classifier. During training, parameters of the nonlinear function $F(\cdot)$, equal to the classifier kernel function, are estimated. We evaluated the performance of the following three nonlinear functions: (a) radial basis function (RBF), (b) χ^2 function, and (c) histogram intersection.

Linear methods are simple and have low computational complexity. However, to infer weights $w_i|_{i=1,\ldots,L}$, an exhaustive search over parameter values is needed which becomes an intractable task for a large number of cues. The nonlinear methods owing to more complex function may better adapt to the varying properties of the cues. However, they also require a larger training dataset which may be infeasible for real world scenarios.

IV. EXPERIMENTAL EVALUATION

First, we present the dataset and experimental setup in Sec. IV-A and IV-B. Then, we study robustness of different 2D and 3D descriptors in Sec. IV-C followed by a systematic evaluation of several 2D-3D integration strategies in Sec. IV-D. Further, we demonstrate grasp generation on novel objects based on categorical information. We also show how the grasp knowledge can be transfered between objects that belong to the same category. Finally, we study performance of the integrated system in realistic scenario for multiple objects, scenes and tasks in Sec. IV-E.

A. Database

Most of the available 2D-3D object databases contain unsuitable object classes to demonstrate the *category*-based

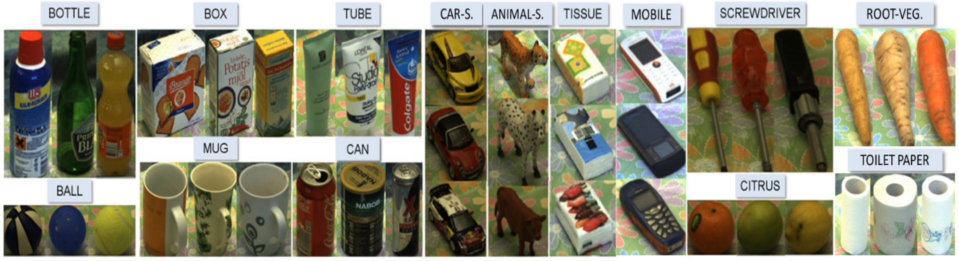


Fig. 5. Examples of objects used to create the database presented in Section IV-A. Different objects were chosen for each category in order to capture variations in appearance, shape and size within each class. The data for all the 140 objects can be viewed at our web site `http://www.csc.kth.se/~madry/research/stereo_database/index.php`.

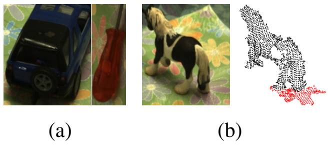


Fig. 6. Examples of imperfect segmentation in both 2D and 3D: (a) only a part of an object is detected, or (b) the segmentation mask contains background points (background points are marked in red).

TABLE II

RESULTS FOR THE FEATURE SELECTION EXPERIMENTS FOR *Setup-50*.

Descriptor	SIFT	opponentSIFT	HoG	FPFH
Av.Categ.Rate σ	86.2% 4.5%	**86.8%** 3.3%	75.1% 1.8%	65.8% 2.7%

task-directed grasping [17][32][33]. We collected a new database with objects chosen from everyday categories. The dataset contains a number of objects that are similar in shape and appearance, but afford different tasks (e.g. ball/orange, orange/carrot, carrot/screwdriver). There are 14 categories: *ball, bottle, box, can, car-statuette, citrus, mug, 4-legged animal-statuette, mobile, screwdriver, tissue, toilet-paper, tube and root-vegetable* (see Fig. 5), each with 10 different object instances per category (in total 140 objects). For each object, the 2D (RGB image) and 3D (point cloud) data were collected from 16 different views around the object (separated by 22.5°) using the 7-joint Armar III robotic head presented in Fig. 2. To differentiate the object and background we used the active segmentation method [13] that generated good results in ca. 90% of cases. Typical examples of imperfect segmentation are shown in Fig. 6. Additionally, in order to evaluate performance of the categorization and grasp generation systems in the real environment, we collected data for 10 natural scenes. Five subjects were asked to randomly place between 10 to 15 objects from 14 different categories on a table. In the scenes, different lightning condition and occlusions of objects are present. Several scenes are shown in Fig. 11 and 13.

B. Experimental Setup

The database was divided into four sets used for: (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, data for 56 objects were used for training and testing, and data for 14 objects for subsequent validations. Due to the fact that we aim to test performance of the system for the object categorization and not object instance recognition, an object that was presented to the system during the training phase was not used to evaluate its the performance.

For training, we selected 8 views per object separated by 45° (Fig. 7 top row). We also used 8 images per object for testing, however we varied a number of *unknown* viewpoints between 0 and 8. Fig. 7 (bottom row) presents a test setup where half of the views is unknown. This setup reflects the best the real condition and we called it *Setup-50*. The results are reported for a single object view and information provided by different views was not fused. To average the results each experiment was repeated five times for randomly chosen object instances. We report the average categorization rate and standard deviation (σ).

C. Feature Selection for Object Categorization

We built four 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 (FPFH). The SIFT and opponentSIFT were extracted using a grid detector, and HoG descriptor using the Canny edge detector. The final object representation for the 2D descriptors follows a concept of the spatial pyramid, and for the 3D descriptor BoW model.

In order to assess the performance of the descriptors under different viewpoints, we varied a number of *unknown* viewpoints in the test set between 0 and 8. The results are illustrated in Fig. 8. All 2D descriptors obtained rather high categorization rate when the viewpoint was known (0 views), but the performance dropped significantly as the viewpoint varies. The highest performance was obtained for opponentSIFT which indicates that color information is less influenced by the viewpoint changes than shape information (HoG). The 2D descriptors yielded higher categorization rates than the 3D descriptor. However, the performance of the 3D descriptor is only slightly affected by the viewpoint changes. Additionally, we attach the numerical results for *Setup-50* in Table II.

D. Cue Integration

In this section, we present results from combining 2D and 3D categorization. The best performance of **92%** was

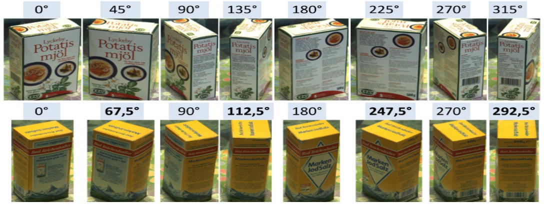
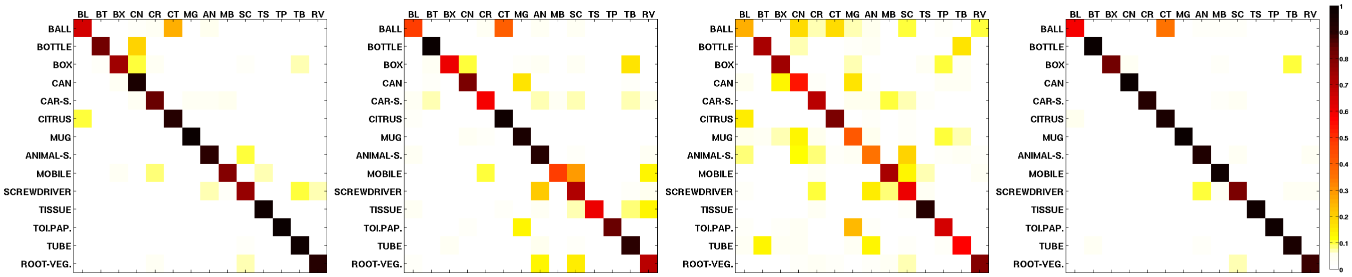


Fig. 7. *Setup-50*. Objects from eight different viewpoints selected to train the system (top row) and evaluate its performance (bottom row).



(a) opponentSIFT (b) HoG (c) FPFH (d) opponentSIFT+HoG+FPFH

Fig. 9. 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). The images are best viewed in color.

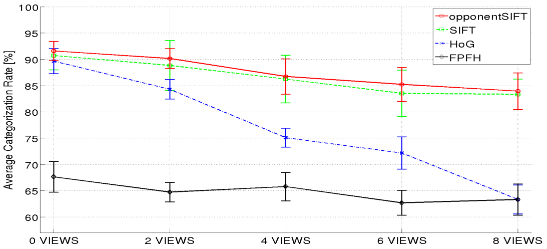
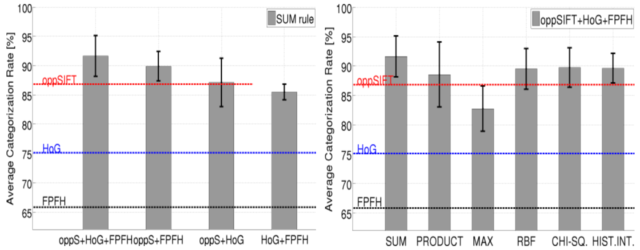


Fig. 8. Performance of descriptors under varying viewpoint.



(a) Features (b) Integration methods

Fig. 10. Average categorization rate for: (a) different pairs/triples of features (for linear combination method, sum rule), (b) different linear and nonlinear combination methods (for opponentSIFT+HoG+FPFH).

obtained for integration of the three descriptors: opponentSIFT+HoG+FPFH using the linear combination method. When comparing to the best single cue OCS (based on opponetSIFT), the combination of 2D and 3D features improved performance of the system in average by **5%**. The confusion matrix obtained for this experiment is presented in Figure IV-D (d). The results show that capturing diverse object properties (appearance, contour and 3D shape) and integration of information from different visual sensors (2D and 3D) not only significantly improve robustness of the categorization system, but are essential to discriminate between similar objects that afford different tasks. The integrated system is able to correctly classify objects that are alike in shape or appearance, but are to be used for different purpose (see Fig. 4). For example, it correctly categorizes objects of similar: (a) shape, such as *screwdriver* and *root-vegetable* where only the former can be used as a tool, *ball* and *citrus* where only the former affords playing, or *mug*, *can* and *toilet-paper* where only the former affords pouring a liquid; (b) appearance: *citrus* vs. *root-vegetable*, *bottle* vs. *can*. Such classification is very challenging for a system based on a single cue.

1) Detailed Results: The categorization results for different choices of features and cue integration methods are presented in Fig. 10. The results confirm that descriptors need to be complementary, i.e. capture different object properties and originate from different sensors which motivates the use of multiple sensors capturing various characteristics of the objects. The best categorization rate is obtained for fusion of all three features (opponentSIFT+HoG+FPFH). The second best for the combination of descriptors that capture different object attributes and originate from different channels: 2D color and 3D shape descriptor (opponentSIFT+FPFH). Further, for the color and shape descriptor from the same channel (opponentSIFT+HoG) and for the two shape descriptors (HoG+FPFH). The same trend in performance is observed for both the linear and nonlinear combination methods. This is evidence of selective properties of our system.

In case of the linear algebraic methods, we tested the weighted sum, product and max rule. For all combinations of features, the approach based on the sum and product rule improved the performance of the system in comparison to the best single cue OCS (opponentSIFT), and the sum rule was superior to the product rule. The max rule that in case of two classifiers is equivalent to the majority voting, yielded the lowest categorization rate further supporting the notion of complementarity. In case of the nonlinear algebraic methods, we evaluated the RBF, χ^2 and histogram integration functions. All the nonlinear functions provided a comparable performance. In our study, the linear algebraic integration methods outperformed the nonlinear methods. A small set of data was used to train the SVM classifier for the nonlinear methods. We can draw the conclusion that in case of a limited amount of data, the simpler fusion methods are more efficient.

2) Natural Scenes: We evaluated performance of the 2D-3D integrated OCS on 10 natural scenes where each contains 10-15 objects randomly placed on a table. To categorize objects, we chose the best classifier trained following the procedure described in Section IV-B. It is important to note that an object presented to the system during training was not used to evaluate its performance. For each object in the scene, we estimated a confidence vector over the 14 object categories. The final label was found by choosing the category with the highest support. The categorization results for a few scenes together with a confidence vector for each object are presented in Fig. 11. We can observe that the system is capable to operate in a very challenging scenario. For 10 natural test scenes, it yielded a high categorization rate of **91.7%** in spite of occlusions (see Scene: 3, Objects: 2, 6, 13, 14) or inaccurate segmentation (S: 4, O: 1, 5). The most difficult remained the discrimination between the *ball* and *citrus* category (S: 3, O: 8) what matches the trend presented in the confusion matrices in Fig. 8.

E. Object Category-based Task-constrained Grasping

In this section, we summarize the results of an integrated system considering categorization for task-constrained object grasping. Our experimental scenario considers multiple objects grasp planning constrained by the assigned tasks. In addition, we take robot embodiment into account. The robot is presented with a scene containing several unknown objects, see Fig. 12. First, object hypothesis are segmented from the background. Secondly, each hypothesis is fed into our object categorization system. In the given scene, 13 objects were found, all correctly classified. The confidence value of each object provides evidence for the order in which objects should be grasped.

Next, given the assigned task, the robot needs to decide: (1) which object should be grasped, and (2) how to grasp it to fulfill the task requirements. For this purpose, we use the embodiment-specific task constraint model. The model 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 [32]. Each category includes 4 different object shapes that are scaled to 2 sizes. 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: Our goal is to infer the most suitable grasp parameters for an object in the 3D scene given the assigned task $task$ and the categories of the objects $obcl$. A grasp is parameterized by multiple variables: dir, $fcon$ and pos. In this paper we only illustrate the results on pos. The reason is that pos represents from which direction the hand is placed relative to the object, therefore is a very intuitive variable to exhibit task constraints. For each object, we sample a set of points on an ellipsoid the size of which 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. 12. The point that has the highest P (indicated by the brightest color) implies the best grasp position for the task.

The pos variable in the BN 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 object size, position and orientation in a scene. In this paper, we assume that the orientation of the object is known. The size and position are determined by estimating a minimum bounding sphere of the filtered 3D point cloud (outliers and background points are removed). We assume that the diameter of the bounding sphere corresponds to the largest object dimension. Several examples of grasp transfer to the real objects are presented in Fig. 12. For each object in the scene that was classified as a *mug*, a set of grasping points, that create a grasping map, is presented in the front (camera), top and back view. It is important to note that 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.

2) Task-constrained Grasping in a Real Scene: Fig. 13 shows the results of the experiment for natural scenes. We show the likelihood maps for each object using colored sample points of $P(pos|task, obcl)$ for five tasks: *hand-over*, *tool-use*, *pouring*, *playing* and *dishwashing*. Order in which objects should be grasped given a task is determined as $P(obcl, task) \cdot C$, where C is object categorization confidence introduced in Eq. 1.

In Fig. 13, we see that for the *pouring* task (Row: 3, Column: 2), the likelihoods of the sample points around mugs and bottles are clearly higher than for other objects indicating that they are the only objects affording the task. Similarly, *screwdriver*s are the only objects that can be used as a tool (R:2, C:1 and R:4, C:1), and *cars* and *balls* to play (R:2, C:2). For the *hand-over* task, all objects have high likelihood. This indicates that by using the object category information and the task constraint BN, we can successfully select the object according to their task affordance.

For the objects that afford *pouring*, for example *mug*s in Scene 2 (R:3, C:2, Objects 6 and 9) 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 (R:2, C:1, Object 2), the network favors the position around the tip of the screwdriver whereas leaving the handle part for regrasp.

V. CONCLUSIONS AND FUTURE WORK

Robots grasping objects in unstructured environments need the ability to select grasps for unknown objects and transfer this knowledge to other objects based on their category and functionality. Although for pure categorization 2D information may be sufficient, 3D information is required for grasping and manipulation of objects, and thus can be also used for categorization. In this work, the categorization system is integrated with a task constrained model for goal-directed grasp planning. We showed that the object category information can be efficiently used to infer the task affordance of the observed objects. The proposed system allows for reasoning and planning of goal-directed grasps in real-world scenes with multiple objects.

We have presented the 2D-3D object categorization system that is built upon an active scene segmentation module. The system allows to generate object hypotheses and segment them from the background in real time. Experimental evaluation showed that the proposed system achieved high categorization rate (up to 92%), significantly better than the

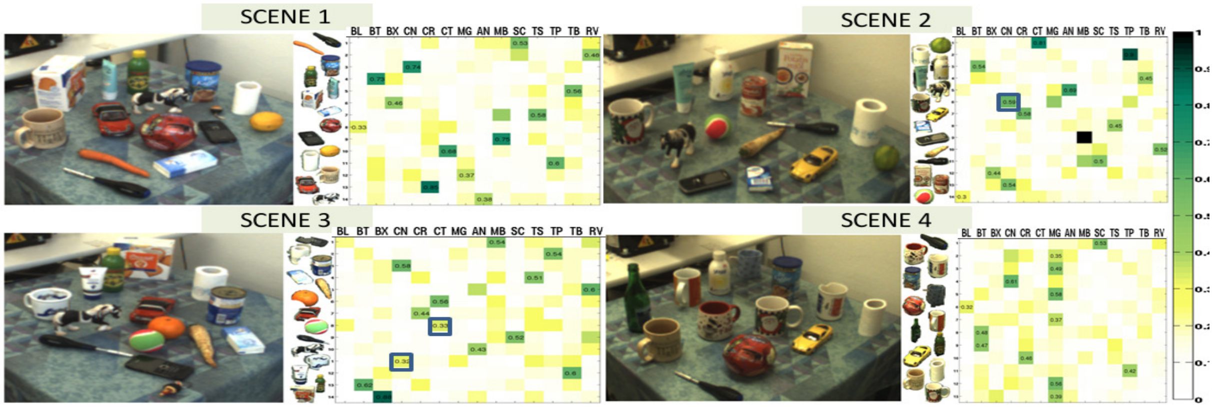


Fig. 11. Categorization results for natural scenes. For each object in a scene, confidence values over 14 categories are shown. All objects were correctly classified except three objects marked using a blue square in confidence vector.

classic single cue SVM for the same task. Moreover, the cue integration method proposed in this paper is very efficient and capable to model situations where limited amount of data is available. The results show that capturing diverse object properties (appearance, color, shape) and integration of information from different visual sensors (2D and 3D), not only significantly improve robustness of the categorization system, but are essential to discriminate between similar objects that afford different tasks.

The current system focuses on vision-based, task-oriented grasp planning. The next step in performing a manipulation action is execution of a stable grasp. One avenue for future research is to integrate this system with the tactile sensing based on-line stability estimation system in [35]. The aim will be to condition the choice of grasps based on the multiple sensory signals available to a robot prior to and while manipulating the object.

References

[1] K. Hsiao, S. Chitta, M. Ciocarlie, and E. G. Jones, "Contact-reactive grasping of objects with partial shape information," in *IROS*, 2010.

[2] K. Welke, J. Issac, D. Schiebener, T. Asfour, and R. Dillmann, "Autonomous acquisition of visual multi-view object representations for object recognition on a humanoid robot," in *ICRA*, 2010.

[3] J. G. Greeno, "Gibson's Affordances," *Psychological Review*, vol. 101, no. 2, pp. 336–342, 1994.

[4] G. Fritz, L. Paletta, R. Breithaupt, E. Rome, and G. Dorffner, "Learning predictive features in affordance-based robotic systems," in *IROS*, 2006.

[5] 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*, vol. 15, no. 4, pp. 447–472, 2007.

[6] D. Kraft, E. Baseski, M. Popovic, N. Kruger, N. Pugeault, D. Kragic, S. Kalkan, and F. Worgotter, "Birth of the object: Detection of objectness and extraction of object shape through object action complexes," *IJHR*, vol. 5, no. 2, pp. 247–265, 2008.

[7] H. Kjellstrom, J. Romero, and D. Kragic, "Visual object-action recognition: Inferring object affordances from human demonstration," *CVIU*, vol. 114, no. 1, pp. 81–90, 2011.

[8] L. Stark and K. Bowyer, "Achieving generalized object recognition through reasoning about association of function to structure," *PAMI*, vol. 13, pp. 1097–1104, 1991.

[9] S. Oh, A. Hoogs, M. Turek, and R. Collins, "Content-based retrieval of functional objects in video using scene context," in *ECCV*, 2010.

[10] D. Song, K. Huebner, V. Kyrki, and D. Kragic, "Learning Task Constraints for Robot Grasping using Graphical Models," in *IROS*, 2010.

[11] D. Song, C.-H. Ek, K. Huebner, and D. Kragic, "Multivariate discretization for bayesian network structure learning in robot grasping," in *ICRA*, May 2011.

[12] 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.

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

[14] K. Huebner, "BADGr—A toolbox for box-based approximation, decomposition and GRasping," *RAS*, 2012.

[15] *Point Cloud Library*, http://pointclouds.org, Last visited: Aug 2011.

[16] 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.

[17] D. Marton, Z-C.and Pangercic, R. B. Rusu, A. Holzbach, and M. Beetz, "Hierarchical object geometric categorization and appearance classification for mobile manipulation," in *Humanoids*, 2010.

[18] G. D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 1, pp. 91–110, 2004.

[19] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *CVPR*, 2008, pp. 1–8.

[20] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582–1596, 2010.

[21] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, vol. 1, 2005, pp. 886–893.

[22] B. Leibe and B. Schiele, "Analyzing appearance and contour based methods for object categorization," in *CVPR*, 2003, pp. 409–415.

[23] J. W. H. Tangelder and R. C. Veltkamp, "A survey of content based 3D shape retrieval methods," in *SMI*, 2004, pp. 145–156.

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

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

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

[27] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, vol. 2, 2006.

[28] 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.

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

[30] A. Pronobis and B. Caputo, "Confidence-based cue integration for visual place recognition," in *IROS*, 2007, pp. 2394–2401.

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

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

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

[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.

[35] Y. Bekiroglu, K. Huebner, and D. Kragic, "Integrating grasp planning with online stability assessment using tactile sensing," in *ICRA*, 2011.

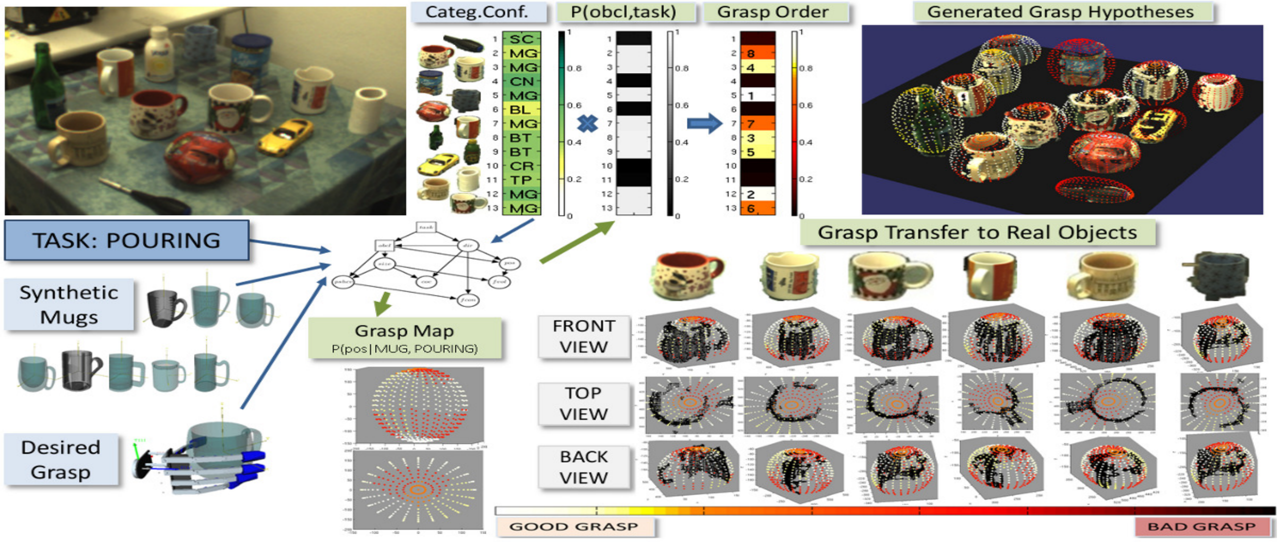


Fig. 12. Grasp transfer from a synthetic object model to real objects in a scene. The grasping points with a high value of $P(pos|obcl, task)$ (good grasping points) are represented by bright color in the heat maps.

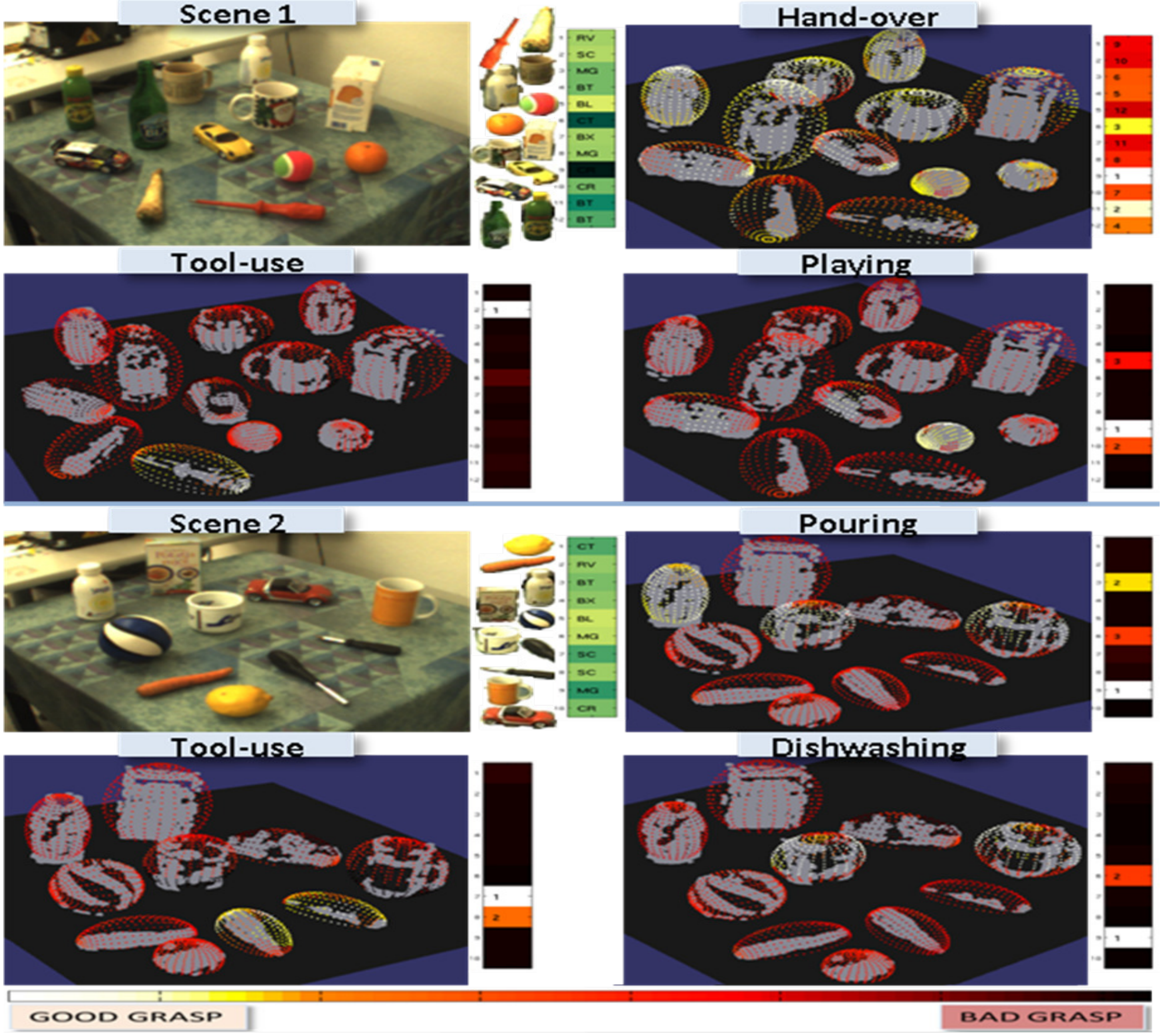


Fig. 13. Generated grasp hypotheses and associated probabilities for the natural scenes. The grasping probability around an object is indicated by a color of the point (the brighter is the point, the higher is the probability). For each scene, we specify which objects should be grasped first (bar on the right side of a scene grasp map). The images are best viewed in color. **Additional experimental results** for natural scenes together with the movies that present accurate 3D information are available on our website `http://www.csc.kth.se/~madry/research/madry12icra`.