

Task-based Grasp Adaptation on a Humanoid Robot[★]

Jeannette Bohg^{**} Kai Welke^{*} Beatriz León^{***} Martin Do^{*}
Dan Song^{****} Walter Wohlking[†] Marianna Madry^{****}
Aitor Aldóma[†] Markus Przybylski^{*} Tamim Asfour^{*}
Higinio Martí^{***} Danica Kragic^{****} Antonio Morales^{***}
Markus Vincze[†]

^{*} *Humanoids and Intelligence Systems Lab, Institute for
Anthropomatics, KIT, DE*

^{**} *Autonomous Motion Lab, MPI for Intelligent Systems, DE*

^{***} *Robotic Intelligence Laboratory, Department of Computer Science
and Engineering, UJI, ES*

^{****} *Computer Vision and Active Perception lab, KTH, SE*

[†] *Automation and Control Institute, TUW, AT*

Abstract: In this paper, we present an approach towards autonomous grasping of objects according to their category and a given task. Recent advances in the field of object segmentation and categorization as well as task-based grasp inference have been leveraged by integrating them into one pipeline. This allows us to transfer task-specific grasp experience between objects of the same category. The effectiveness of the approach is demonstrated on the humanoid robot ARMAR-IIIa.

Keywords: Robotic Grasping and Manipulation, Task-based Grasp Synthesis, Visual Servoing, Attention, Segmentation, Object categorization, System Integration

1. INTRODUCTION

State-of-the-art approaches towards autonomous robot grasping can be roughly divided into grasping known, unknown or familiar objects (Bohg, 2011). For known objects, the problem reduces to recognition and pose estimation as for example in Ciocarlie et al. (2010); Huebner et al. (2009). Given this, object-specific grasps can be retrieved from a database. For unknown objects, usually heuristics (commonly using object shape features in relation to the robot hand) are used to generate and rank grasps as for example in Hsiao et al. (2010). If the goal is to transfer prior grasp experience to novel object, the main challenge is to find a similarity metric that yields high values for two objects that can be grasped in a similar way. Commonly, these similarity metrics are defined on relatively low-level features such as 2D and 3D shape or appearance, e.g. in Saxena et al. (2008). Comparatively little work has been done on transferring grasp experience between objects that are similar in terms of more high-level features like object parts (Detry et al., 2012) or categories (Madry et al., 2012; Marton et al., 2011). Only Marton et al. (2011) have demonstrated their method on a robot.

In this paper, our aim is to leverage on some recent advances on object categorization considering both 2D and 3D data to show how it can facilitate robotic grasping on a humanoid platform. In detail, this paper makes the following contributions:

- Attention mechanism using geometric information,
- Integration of complete categorization pipeline (segmentation, 2D/3D categorization, pose estimation)
- Integration of pipeline with task-based grasp inference system, grasp and motion planning
- Grasp execution using visual servoing without prior object models

The remainder of this paper is outlined as follows. In the next section, we will review related grasping pipelines that have been demonstrated on a robotic platform. This is followed by a description of all the modules in the proposed integrated system. Section 4 demonstrates and discusses the whole grasping pipeline.

2. RELATED WORK

Recently, several fully integrated robot systems that are able to grasp objects from a table top have been proposed. They differ in the amount of prior information the robot is assumed to have about the object and in how the inferred grasps are executed.

Most of these systems assume that the models of the objects are known to the robot and have grasp hypotheses associated to them. Ciocarlie et al. (2010) propose a robust grasping pipeline in which known object models are fitted to point cloud clusters using standard ICP (Besl and McKay, 1992). Knowledge about the table plane, the assumption that objects are rotationally symmetric and are always standing upright helps in reducing the search space of potential object poses. Objects that have been

[★] This work was supported by the EU through the project GRASP, IST-FP7-IP-215821.

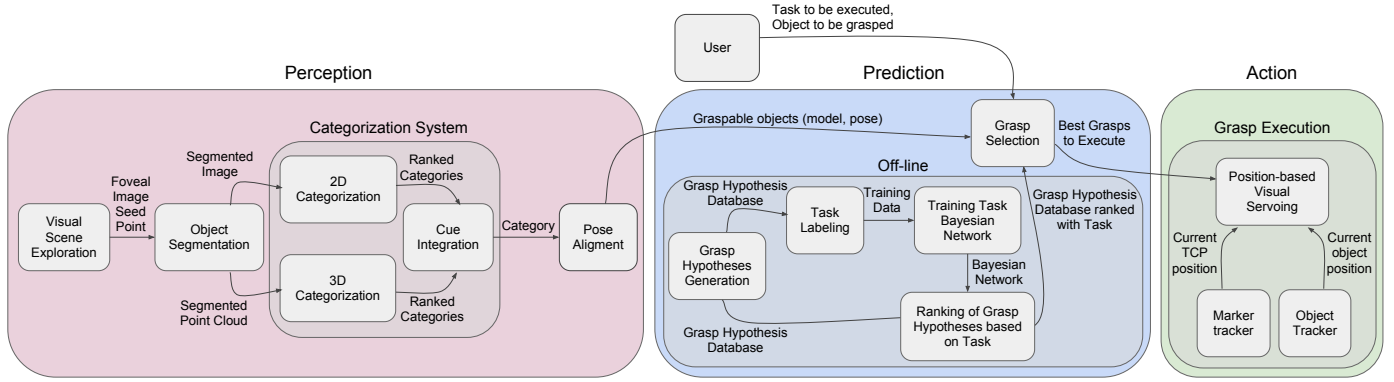


Fig. 1. Grasping Pipeline

detected but not identified are grasped using a reactive approach proposed by Hsiao et al. (2010). Gratal et al. (2010) employ an active vision system to detect and recognize objects with a subsequent pose estimation step. Object models are simplified to be either boxes or cylinders also assuming that they are always standing upright. Visual servoing is employed to robustly grasp an object from the top. Similarly, Huebner et al. (2009) propose a grasping pipeline on ARMAR-IIIa (Asfour et al., 2006) for picking up known objects from a sideboard. For object recognition and pose estimation, the method proposed by Azad et al. (2007) is applied. Grasp hypotheses for each object are synthesised off-line using a box-based shape approximation of object CAD models (Huebner, 2012).

Very few grasping systems have approached the problem of transferring grasp experience between objects of the same category. Marton et al. (2011) present and demonstrate an approach similar to ours in that an object categorization system is used in a robotic grasping framework. Furthermore, it combines 2D and 3D descriptors. However, the category of an object is not used to infer a suitable grasp. Instead, a 3D descriptor helps to narrow down the choice of categories to those of similar shape, and then a 2D descriptor is applied to look up a specific object instance. One of the grasp hypothesis associated to that object instance is then executed.

In this paper, we demonstrate how integration of 2D and 3D cues for object categorization can facilitate robot grasping. Specifically, we do not require exact models of the object to be grasped. A model of the same category annotated with grasp hypotheses is sufficient. Furthermore, we show how the detected category helps to infer a grasp that is task-specific.

Robotic Platform The proposed pipeline is demonstrated on the humanoid robot ARMAR-IIIa (Asfour et al., 2006) consisting of seven subsystems: head, left arm, right arm, left hand, right hand, torso, and a mobile platform. The head has seven DoF and is equipped with two eyes, which have a common tilt and an independent pan. For the visual perception of the environment, the humanoid’s active head features two stereo camera systems, one with a wide-angle lenses for peripheral vision and one with a narrow-angle lenses for foveal vision. For grasping and manipulation, the robot provides a 3 DoF torso and two arms with 7 DoF each. The arms follow an anthropomorphic design: 3 DoF for each shoulder, 2 DoF in each elbow and 2 DoF

in each wrist. Each arm is equipped with a pneumatic-actuated five-fingered hand. The robot moves on a wheeled holonomic platform.

3. THE GRASPING PIPELINE

An overview of the proposed grasping pipeline is shown in Fig. 1. It is subdivided into three major building blocks. Given visual input from stereo cameras, the perception block is responsible for detecting an object and estimating its category and pose. The prediction block takes in the object pose and category and infers a ranked list of grasps according to a specific task. The best grasp is then executed on the robotic platform.

3.1 Perception

Attention The success of visual processing such as classification rate and reliability of pose alignment strongly depends on the quality of the visual sensor data provided as input. Consequently, providing as detailed views of the objects as possible is beneficial for all processing steps. On the humanoid platform ARMAR-IIIa such detailed views can be obtained by making use of the active head and the foveal camera system (Asfour et al., 2008). The fixation of the objects in the foveal camera pairs involves mechanisms of attention allowing to determine points of interest in the scene and mechanisms for the execution of gaze shifts.

Attention points are computed based on geometric information. We apply semi-global matching (Hirschmüller, 2005) to the wide-field stereo views for obtaining a dense 3D reconstruction of the scene. It is assumed that most objects are placed on flat surfaces thereby simplifying the detection of interest points. We process the resulting 3D data using plane fitting similar to Rusu et al. (2009). After removing the detected support surface, the remaining 3D points are clustered in an unsupervised manner using the growing neural gas method (Fritzke, 1995). Each cluster center serves as a point of interest.

To bring these attention points into the view of the foveal cameras, the kinematic model of the eye system is calibrated off-line (Welke et al., 2008). The inverse kinematics problem is solved using a differential kinematics approach for 6 DoF of the Karlsruhe Humanoid Head, where the redundancy is exploited in order to keep natural poses similar to the approach proposed in Ude et al. (2006).

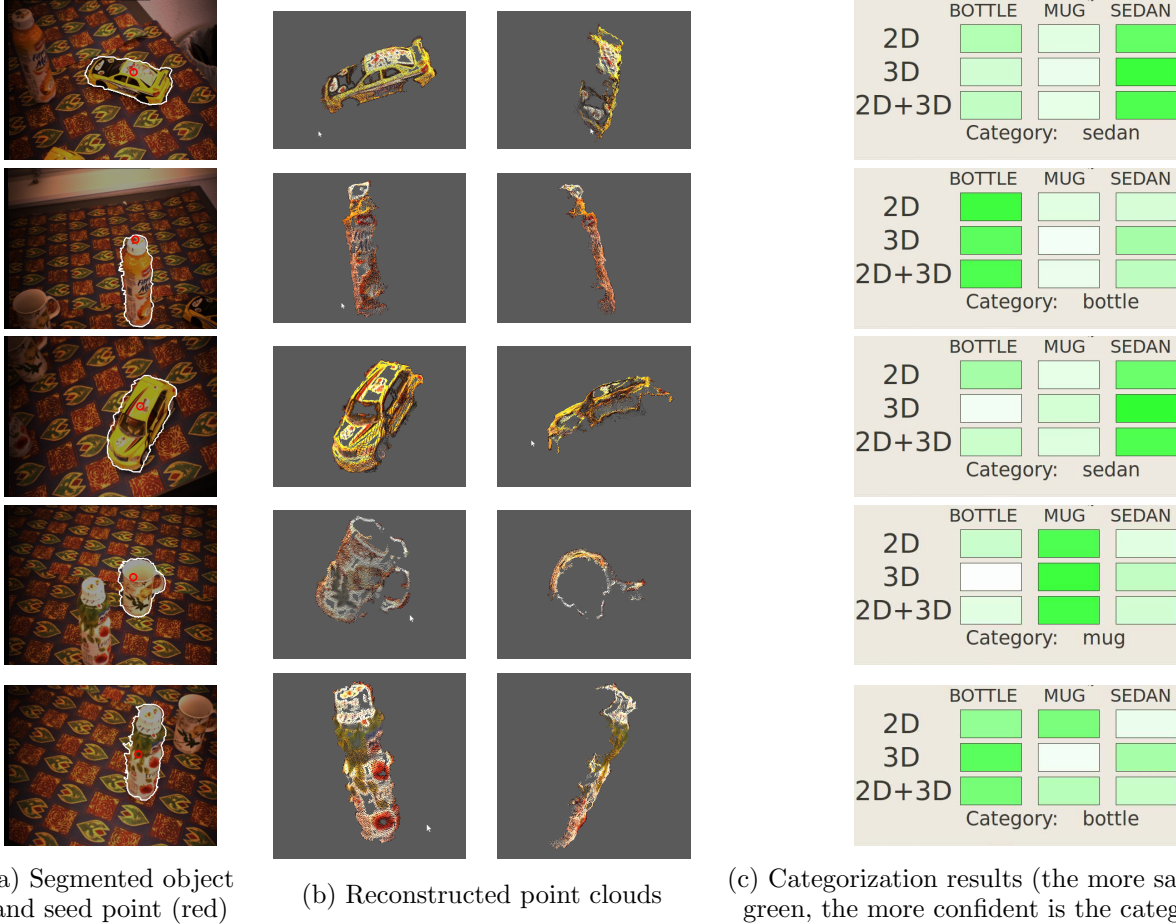


Fig. 2. Intermediate results of the categorization pipeline.

Segmentation Once the previously detected points of interest are visible in the foveal cameras of the active head, we can refine the preliminary 3D clusters. We use a recent 3D object segmentation approach by Björkman and Kragic (2010). It relies on three possible hypotheses: figure, ground and a flat surface. The segmentation approach is an iterative two-stage method that first performs pixel-wise labeling using a set of model parameters and then updates these parameters in the second stage.

To initialize this process, we assume that points close to the previously detected points of interest are likely to belong to the object. We therefore project these attention points from the wide field cameras to the foveal view. An imaginary 3D ball is placed around them and everything within the ball is initially labeled as foreground. For the flat surface hypothesis, RANSAC (Fischler and Bolles, 1981) is applied to find the most likely plane. The remaining points are initially labeled as background points.

Fig. 2 (a) shows examples of segmented objects with the corresponding seed point. The resulting filtered segmented point clouds are shown in Fig. 2 (b).

Categorization System The object segmentation delivers both, a segmented RGB image and a segmented 3D point cloud. Each kind of data encodes different characteristics of the object that are complementary. In this paper, we propose to fuse these different cues to achieve a more robust object categorization. Specifically, we run two dis-

tinct *object categorization systems* (OCS) in parallel of which one is processing 2D cues and the other 3D cues only. To obtain a final category, the evidence provided by each system is merged. As complementary cues for an object category, we have defined appearance (SIFT (Lowe, 2004)), color (opponentSIFT (van de Sande et al., 2010)) and 3D shape (Wohlkinger and Vincze, 2011).

In the case of the 2D OCS (Madry et al., 2012), the object is represented by spatial pyramids (Lazebnik et al., 2006). For classification, we use the One-against-All strategy for M -class SVMs with a χ^2 kernel. The confidence with which an object is assigned to a particular class is calculated based on the distance of a sample to the SVM hyperplane (Pronobis and Caputo, 2007).

For the 3D OCS (Wohlkinger and Vincze, 2011), the classification task is posed as a shape matching problem. Given the segmented object point cloud, the 3D OCS finds the most similar 3D model and view from a database of synthetically generated 2.5D views of CAD models gathered from the web. The *Shape Distribution on Voxel Surfaces* descriptor (Wohlkinger and Vincze, 2011) encodes the global 3D shape by means of histograms of point-to-point distances and is calculated directly on the point cloud. The output of the 3D OCS system is a ranked list of classes and object views associated to each class.

Cue integration is based on an algebraic combination of classifier outputs. The total support for each class

is obtained as a linear weighted sum of the evidences provided by individual OCSs. The final decision is made by choosing the class with the strongest support, as visualized in Fig. 2c. The class name, the 3D model and the best matching views are handed onto the next step, the pose alignment module.

Pose Alignment Whether a grasp is suitable for an object of a specific category and task, is highly dependent on the relative alignment between hand and object. Therefore, we need to estimate the exact pose of the object in the scene.

From the previous categorization step, we are given a set of object prototype models of a specific category that are most similar to the current object of interest. To determine the best matching model from this set, we need to align each of them with the segmented point cloud and compare the values of a similarity metric. This is achieved by applying the approach by Aldoma and Vincze (2011). The method was originally proposed for aligning geometrically similar 3D shapes. It is based on the assumption that objects belonging to the same category usually share at least one stable pose. Therefore by using the stable poses of the two models to be aligned, the search space over the full space of rigid body motions can be reduced to a set of 2D problems. From the transformation between stable planes, 5 DoF are directly estimated while the other two (rotation about plane’s normal and scale) can be efficiently obtained by cross-correlation on the log-polar space after being transformed to the frequency domain. The stable planes for the object models in the database are computed off-line. For the query object, we assume that it is currently resting in one of its stable poses. We therefore use the normal of the table for alignment.

3.2 Prediction

At this point in the processing pipeline, we have the following information available about an object in the scene: i) its category, ii) the most similar object model from a database and its estimated pose and iii) a specific task. Given this, our goal is to infer a ranked list of grasps.

We approach this problem, by *off-line* generating a set of task-ranked grasp hypotheses. This involves i) the generation the grasp hypotheses and ii) their ranking according to task and object’s category. In the *on-line* process, the most similar object model and the given task serve as a look-up in this database to retrieve the highest ranked grasp. This is visualized in Fig. 1 in the *Prediction* block. In the following section, these building blocks are described in more detail.

Grasp planning and selection is performed using the OpenRAVE simulator (Diankov, 2010). The simulated model of the ARMAR-IIIa platform shown in Fig. 3 was created using the Robot Editor available in the OpenGRASP toolkit (León et al., 2010).

Grasp Hypothesis Generation This process is performed off-line using the grasp planning method proposed by Przybylski et al. (2011) and is based on the medial axis transform (MAT) (Blum, 1967). The MAT can represent arbitrary three-dimensional shapes. It is constructed by

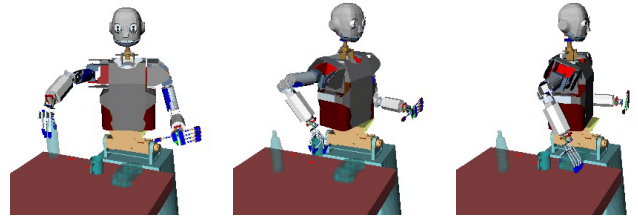


Fig. 3. Simulated model of the ARMAR-IIIa humanoid robot in OpenRAVE grasping different objects

inscribing spheres of maximum diameter into an object’s shape, where each of theses spheres have to touch the object’s surface at two or more different points. The object is then described as a set of spheres, where each sphere has a position, radius and an angle between its centroid and the closest object boundary point as parameters (Miklos et al., 2010). For the actual grasp planning process, we sort the inscribed spheres into a grid structure with respect to their Cartesian coordinates. Candidate grasps for a sphere in this grid are generated by estimating the symmetry properties of the sphere centers in the query sphere’s vicinity. These symmetry properties are then used to choose approach point, approach direction and hand orientation such that the fingers can wrap around the object. Each candidate is tested for stability using the common ϵ -measure for force-closure (Ferrari and Canny, 1992).

Task Constraint Model and Task Ranking We model the conceptual task requirements for a given hand through conditional dependencies between the task T and a set of variables including object features O , grasp action parameters A and constraint features C . This is captured in a Bayesian Network (BN). As described in more detail in our previous work (Song et al., 2010, 2011), we learn both the parameters and the structure of this BN. The necessary training data for this process is generated from a synthetic database of objects on which grasps have been generated as described in the previous section. Each of these object-grasp pairs is visualized to a human expert who labels it with one or more tasks that this grasp would be suitable for. This off-line training process is summarized in the flow chart of Fig. 1.

After training, the BN encodes the joint distribution of all the variables $P(T, O, A, C)$. Fig. 4 shows the learned task constraint model. The variables are explained in Table 5. From Fig. 4, we notice that one of the object features, object category $obcl$, is directly influenced by the task variable. This indicates the importance of the object category information in determining its functional affordance of grasping tasks, hence reinforces the importance of the object categorization in the “perception” block.

Given this BN, we can then infer the conditional distribution of one variable given observations of all or a subset of other variables. For example, we can infer $P(obcl|task)$ to decide on which object category the given task can be best performed. We can also infer $P(A|task, obcl)$ to provide grasp ranking given a task and the category of the current object of interest.

Grasp Selection based on Object Category and Task The result of the off-line process provides the robot with

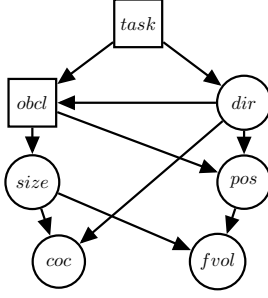


Fig. 4. The structure of the Bayesian network task constraint model for the ARMAR hand.

Groups	Name	Dim	Description
<i>T</i>	<i>task</i>	-	Task Identifier
<i>O</i>	<i>obcl</i>	-	Object Category
	<i>size</i>	3	Object Size
<i>A</i>	<i>dir</i>	4	Hand Orientation (Quaternion)
	<i>pos</i>	3	Grasp Position
<i>C</i>	<i>coc</i>	3	Center of Contacts
	<i>fvol</i>	1	Free Volume

Fig. 5. Features in the task Bayesian networks.

databases of grasp hypotheses for each object annotated with a task-probability. Given this, we can select the best grasp to be executed on-line.

The perception part of the pipeline outputs the perceived objects in the scene with their calculated poses in the environment and they are loaded into OpenRAVE. A task is then specified by the user. The grasp hypotheses stored for that object are ordered by their probability for the specific task. The best ranked hypotheses are then tested to ensure that they are reachable under the current robot configuration and that the robot trajectory is collision free. Examples for this process on different objects are shown in Fig. 3

3.3 Action

Once the best rated, reachable grasp hypothesis has been determined, it is executed by the humanoid platform ARMAR-IIIa. Since the joints in the humanoid’s arm are wire-driven and the head is actively controlled, even small inaccuracies in the kinematic chain from head to the hand involving at least 10 DoF may cause significant displacement errors when positioning the hand. To guarantee a stable grasp execution, position-based Visual Servoing is employed to align the hand with the grasp hypothesis that is defined relative to the object pose. Therefore, both the target object and the hand of the robot are tracked.

Target Object Model Acquisition Since we do not assume knowledge of the exact target object model, tracking the pose of the target is realized by acquiring a model from the current scene. For the acquisition process, the robot takes exactly the same pose as for observing the object during the scene exploration. The target object model is build using 3D points estimated at Harris interest points from the peripheral images with the grasp pose in the left camera frame as reference.

Grasp Approach and Execution During the grasp approach phase which involves moving the whole torso of the robot, the head moves simultaneously to maintain

good visibility of the target and the robot hand. The head and torso movements are compensated by tracking the target using the previously acquired target model. For this purpose, we make use of the Kanade-Lukas-Tomasi optical flow algorithm (?) for tracking the 3D points in the images.

Similar to our previous work (Vahrenkamp et al., 2008), we exploit an artificial marker on the humanoid’s wrist to simplify the tracking of the robot hand. Using the tracked position of the artificial marker and the kinematic model of the arm and the hand, the 6 DoF pose of the end-effector frame is estimated. The position-based Visual Servoing approach iteratively minimizes the distance between the end-effector pose and the grasp pose defined by the selected grasp hypothesis.

Once the target pose is reached, the robot closes its hand and lifts the object in an open-loop manner.

4. EXPERIMENTS

4.1 Setup

Our experimental setup features a scene which contains several object instances of different categories placed on a sideboard. Fig. 6 shows an example scene on which we will illustrate and discuss the proposed pipeline. Since only stereo vision is used for perception, we covered the table by a textured table cloth. This facilitates the detection of the dominant plane in the scene, an assumption used by several modules in the perception block of the whole grasping pipeline (attention, segmentation and pose estimation).

In this paper, we consider objects of three different categories: cars, bottles and mugs. The 2D OCS system was trained on two object instances per category. For each object, RGB images were collected from eight equidistant views around the object. For training the 3D OCS, seven car, eight bottle and ten mug models each in two scales were used. Eighty views per model were generated. In Madry et al. (2012); Wohlkinger and Vincze (2011), each OCS is demonstrated for a much broader class of categories. The restriction to only three classes in this paper is due to the extensive task-labeling process as well as to restrictions on the size and weight of the objects to be grasped.

We are considering four different tasks: hand-over, pouring, dish-washing and playing labeled on 1956 stable grasp hypotheses for the 50 object models. Since some hypotheses are good for multiple tasks, we have 2166 training examples in total.

4.2 Discussion

Perception An example output for the generation of points of interest in a scene can be found in Fig. 7. It is based on a rough clustering of the scene into background and a set of objects. Dependent on the scene geometry, some objects might be split into several clusters so that during the scene exploration phase, they are re-visited. This is, however, not a problem since the initial clusters are refined by running object segmentation in the foveal view. Duplicated attention points are filtered by thresholding the overlap of different clusters.

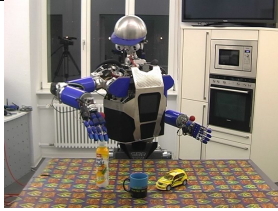
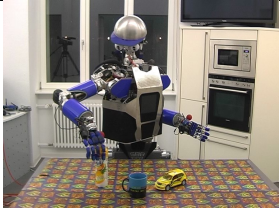
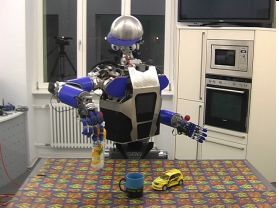
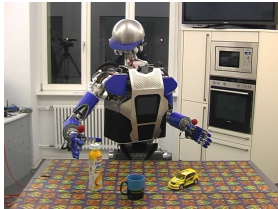
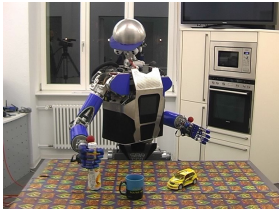
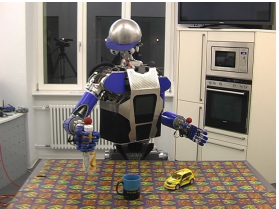
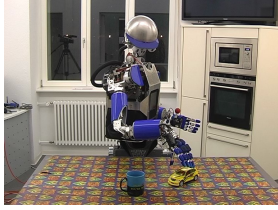
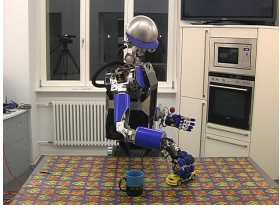
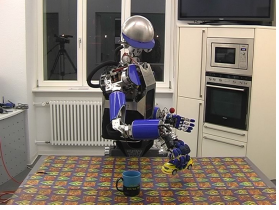
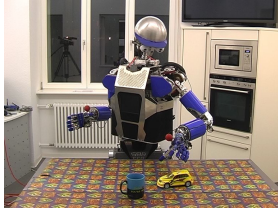
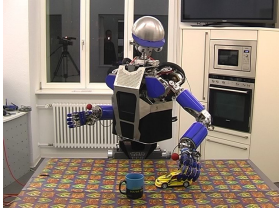
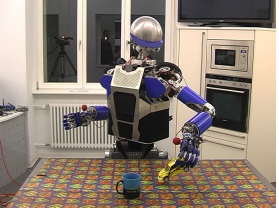

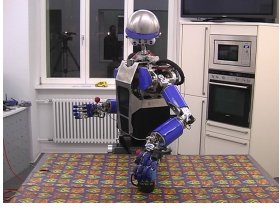
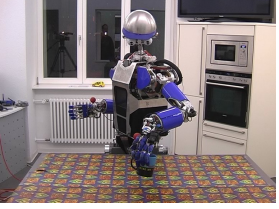
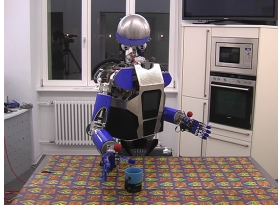
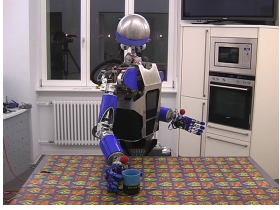
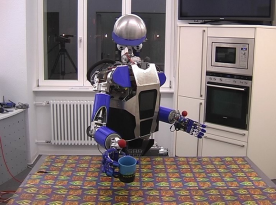
	Task	Approach	Grasp	Lift
Bottle	Hand-over			
	Pouring			
Sedan	Hand-over			
	Playing			
Cup	Hand-over			
	Dish-washing			

Fig. 6. Example grasp executions of grasps given different tasks. Video available at www.youtube.com/watch?v=rXNwBurCnTc.

Different from our previous work (Björkman and Kragic, 2010) where object segmentation is applied on a fixating vision system, here the vergence angle between the left and right camera is kept static. As seed points, we are using the cluster centers in the wide field cameras projected to the foveal view. Therefore, we have shown that the segmentation method does not require a fixating system where seed and fixation point are equal.

The points of interest are usually on the object center. However, they might be positioned closer to the object boundary as for example for the second object in Fig. 2. The good segmentation result suggests that the method is robust to the placement of the seed points because it simultaneously keeps a foreground, background and flat

surface model. Segmentation will lead to erroneous results like over- or under-segmentation if the seed points are not placed on an object or placed on the object's boundary, if objects are touching each other or if fewer seed points than objects are given.

The more precise the resulting segmented point cloud, the better the quality of the object categorization will be. Given this, it was found that the 2D and 3D OCSs are complementing each other well, resulting in a more robust object categorization.

The input of the OCS to the pose estimation consists of a set of objects instead of just one. In theory, the object view matched in the database by the 3D OCS should provide



Fig. 7. The attention system generates a gaze pattern (blue) which is composed of a set of fixation points (red). The fixation points are used as seed points for object segmentation in the foveal images.

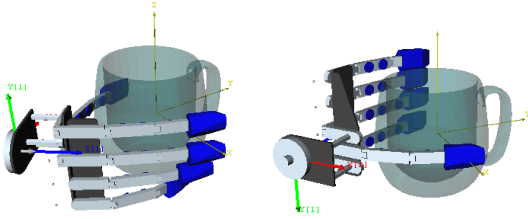


Fig. 8. Comparison of best ranked grasp for a mug according to the task *pouring* (left) or *dish-washing* (right).

the pose of the object relative to the camera. However, the shape descriptor is not discriminative regarding the roll angle around the principal axis of the camera, has no scale and operates on discretized views of the model. For exact alignment of the 3D model to the sensed data, we rely on the pose estimation step on a set of well matching views.

Prediction In the example scene in Fig. 6, we directly compared the execution of grasps on the same object in the scene given different tasks. You can observe that grasps for the task *hand-over* are usually top grasps that leave a major part of the object uncovered. The grasps for *playing*, *pouring* and *dish-washing* are as expected oriented towards the functionality of the object itself in relation to the task. For the dish-washing task, the best ranked grasp on the mug is visualized in Fig. 8 (right). This grasp was however not reachable in this scene and therefore the next best grasp for this combination of object and task has been selected. It is very similar to pouring from this mug as visualized in Fig. 8 (left).

5. CONCLUSIONS

In this paper, we presented a grasping pipeline that allows autonomous robot grasping according to an object's category and a given task. Several state-of-the-art modules performing scene exploration through gaze shifts, segmentation, object categorization and task-based grasp selection were integrated. We showed how this allows the robot to transfer task-specific grasp experience between objects of the same category. The effectiveness of the pipeline was demonstrated on the humanoid robot ARMAR-IIIa.

To increase the robustness of grasp execution, we have designed and implemented an *overcomplete* pipeline where

the task of different modules overlap. This holds for attention, segmentation, categorization and pose estimation.

However, information in this pipeline only flows into one direction (from left to right) without any intermediate failure detection. If this was the case, repeated execution of perceptual processes could be requested to improve the input to a module. Furthermore, this repetition could be based on more information that was already obtained in a later stage of the pipeline.

Another potential improvement of the robustness of the pipeline could be achieved by not only executing the reaching motion in a closed loop manner but also the grasp itself. From the perception block of the pipeline, we know the geometry of the object quite well. This would allow us to adopt an approach similar to Pastor et al. (2011) for on-line comparison of actual and expected tactile sensor readings and adaptation of the grasp if necessary.

REFERENCES

- Aldoma, A. and Vincze, M. (2011). Pose alignment for 3d models and single view stereo point clouds based on stable planes. *Int. Conf. on 3D Imaging, Modeling, Processing, Visualization and Transmission*, 0, 374–380.
- Asfour, T., Regenstien, K., Azad, P., Schröder, J., Vahrenkamp, N., and Dillmann, R. (2006). Armar-iii: An integrated humanoid platform for sensory-motor control. In *IEEE/RAS Int. Conf. on Humanoid Robots*, 169–175.
- Asfour, T., Welke, K., Azad, P., Ude, A., and Dillmann, R. (2008). The Karlsruhe Humanoid Head. In *IEEE/RAS Int. Conf. on Humanoid Robots*. Daejeon, Korea.
- Azad, P., Asfour, T., and Dillmann, R. (2007). Stereo-based 6d object localization for grasping with humanoid robot systems. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 919–924.
- Besl, P.J. and McKay, N.D. (1992). A method for registration of 3-d shapes. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 14, 239–256.
- Björkman, M. and Kragic, D. (2010). Active 3D scene segmentation and detection of unknown objects. In *IEEE Int. Conf. on Robotics and Automation*.
- Blum, H. (1967). *Models for the Perception of Speech and Visual Form*, chapter A transformation for extracting new descriptors of shape, 362–380. MIT Press, Cambridge, Massachusetts.
- Bohg, J. (2011). *Multi-Modal Scene Understanding for Robotic Grasping*, chapter 7. Generation of Grasp Hypotheses, 101–152. KTH.
- Ciocarlie, M., Hsiao, K., Jones, E.G., Chitta, S., Rusu, R.B., and Sukan, I.A. (2010). Towards reliable grasping and manipulation in household environments. In *Int. Symposium on Experimental Robotics*. New Delhi, India.
- Detry, R., Ek, C.H., Madry, M., Piater, J., and Kragic, D. (2012). Generalizing grasps across partly similar objects. In *IEEE International Conference on Robotics and Automation*. To appear.
- Diankov, R. (2010). *Automated Construction of Robotic Manipulation Programs*. Ph.D. thesis, Carnegie Mellon University, Robotics Institute.
- Ferrari, C. and Canny, J. (1992). Planning optimal grasps. In *IEEE Int. Conf. on Robotics and Automation*, 2290–2295.

- Fischler, M. and Bolles, R. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. of the ACM*, 24, 381–395.
- Fritzke, B. (1995). A growing neural gas network learns topologies. In *Advances in Neural Information Processing Systems 7*, 625–632. MIT Press.
- Gratal, X., Bohg, J., Björkman, M., and Kragic, D. (2010). Scene representation and object grasping using active vision. In *IROS'10 Workshop on Defining and Solving Realistic Perception Problems in Personal Robotics*.
- Hirschmüller, H. (2005). Accurate and efficient stereo processing by semi-global matching and mutual information. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 807–814.
- Hsiao, K., Chitta, S., Ciocarlie, M., and Jones, E.G. (2010). Contact-reactive grasping of objects with partial shape information. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*.
- Huebner, K. (2012). BADGr – A toolbox for Box-based Approximation, Decomposition and GRASPing. *Robotics and Autonomous Systems*, 60(3), 367–376.
- Huebner, K., Welke, K., Przybylski, M., Vahrenkamp, N., Asfour, T., Kragic, D., and Dillmann, R. (2009). Grasping known objects with humanoid robots: A box-based approach. In *Int. Conf. on Advanced Robotics*.
- Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *IEEE Conf. on Computer Vision and Pattern Recognition*.
- León, B., Ulbrich, S., Diankov, R., Puche, G., Przybylski, M., Morales, A., Asfour, T., Moio, S., Bohg, J., Kuffner, J., and Dillmann, R. (2010). OpenGRASP: A toolkit for robot grasping simulation. In *Int. Conf. on Simulation, Modeling, and Programming for Autonomous Robots*.
- Lowe, G.D. (2004). Distinctive image features from scale-invariant keypoints. *Int. Journal of Computer Vision*, 60(1), 91–110.
- Madry, M., Song, D., and Kragic, D. (2012). From object categories to grasp transfer using probabilistic reasoning. In *IEEE Int. Conf. on Robotics and Automation*. To appear.
- Marton, Z.C., Pangercic, D., Blodow, N., and Beetz, M. (2011). Combined 2D-3D Categorization and Classification for Multimodal Perception Systems. *Int. Jour. of Robotics Research*. Accepted.
- Miklos, B., Giesen, J., and Pauly, M. (2010). Discrete scale axis representations for 3d geometry. In *ACM SIGGRAPH*. ACM, New York, NY, USA.
- Pastor, P., Righetti, L., Kalakrishnan, M., and Schaal, S. (2011). Online Movement Adaptation based on Previous Sensor Experiences. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 365 – 371. San Francisco, USA.
- Pronobis, A. and Caputo, B. (2007). Confidence-based cue integration for visual place recognition. In *Int. Conf. on Intelligent Robots and Systems (IROS)*, 2394–2401.
- Przybylski, M., Asfour, T., and Dillmann, R. (2011). Planning grasps for robotic hands using a novel object representation based on the medial axis transform. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*.
- Rusu, R., Holzbach, A., Diankov, R., Bradski, G., and Beetz, M. (2009). Perception for mobile manipulation and grasping using active stereo. In *IEEE/RAS Int. Conf. on Humanoid Robots*. Paris.
- Saxena, A., Wong, L., and Ng, A.Y. (2008). Learning Grasp Strategies with Partial Shape Information. In *AAAI Conf. on Artificial Intelligence*, 1491–1494.
- Song, D., Huebner, K., Kyrki, V., and Kragic, D. (2010). Learning Task Constraints for Robot Grasping using Graphical Models. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*.
- Song, D., Ek, C.H., Huebner, K., and Kragic, D. (2011). Embodiment-specific representation of robot grasping using graphical models and latent-space discretization. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 980–986.
- Tomasi, C. and Kanade, T. (1991). Detection and tracking of point features. Technical Report CMU-CS-91-132, Computer Science Department, Pittsburgh, PA.
- Ude, A., Gaskett, C., and Cheng, G. (2006). Foveated vision systems with two cameras per eye. In *IEEE Int. Conf. on Robotics and Automation*, 3457–3462.
- Vahrenkamp, N., Wieland, S., Azad, P., Gonzalez, D., Asfour, T., and Dillmann, R. (2008). Visual Servoing for Humanoid Grasping and Manipulation Tasks. In *IEEE/RAS Int. Conf. on Humanoid Robots*, 406–412.
- van de Sande, K.E.A., Gevers, T., and Snoek, C.G.M. (2010). Evaluating color descriptors for object and scene recognition. *Pattern Analysis and Machine Intelligence*, 32(9), 1582–1596.
- Welke, K., Przybylski, M., Asfour, T., and Dillmann, R. (2008). Kinematic calibration for saccadic eye movements. Technical report, Institute for Anthropomatics, Universität Karlsruhe.
- Wohlkinger, W. and Vincze, M. (2011). Shape Distributions on Voxel Surfaces for 3D Object Classification From Depth Images. *IEEE Int. Conf. on Signal and Image Processing Applications*.