Ontology-based 3D Pose Estimation for Autonomous Object Manipulation

Rigas Kouskouridas, Theodora Retzepi, Eleni Charalampoglou and Antonios Gasteratos
Democritus University of Thrace
Department of Production and Management Engineering
Vas. Sofias 12, Building I, Xanthi Greece 67100
Emails: rkouskou@pme.duth.gr tretzepi@ee.duth.gr echaral@ee.duth.gr gasteratos@ieee.org

***Abstract*—In this paper a novel solution to the problem of guiding a robotic gripper in order to perform manipulation tasks, is presented. The proposed approach consists of two main modules corresponding to the training and testing sessions, respectively. During training, we employ an ontology-based framework with a view to the establishment of a database holding information regarding several geometrical attributes of the training objects. An accurate estimation of the 3D pose of an object-target is obtained during the testing phase and through the efficient exploitation of the established database. The most common solution to the 3D pose estimation problem implies extensive training sessions that are based on oversampled datasets containing several instances objects captured under varying viewpoints. However, such an approach engenders high complexity accompanied by large computational burden. We address this issue by proposing an ontology-based framework and a fuzzy-based approach that is able to efficiently interpolate between two known instances of the trained objects. Experimental results justify both our theoretical claims and our choice to adopt an ontology-based solution.**

I. Introduction

An expansive range of domestic tasks for service robots, such as collecting objects, loading or unloading a dishwasher and opening doors, is based on object handling. The problem of particle manipulation is specifically challenging in unstructured environments that may include a range of several objects with varying shapes and sizes. Although humans are capable of excelling under such variations, a robotic application with the sufficient trade offs between performance and computational burden has yet to be built. In the last few decades, increasing importance has been gained to the problem of grasping unknown objects in a fully automatic way, mainly due to the wide-spread use of service and rehabilitation robotics [1], [2], [3], [4]. A psychological, biological and engineering focus has given to the manipulation task but is still considered as not being fully solved. Despite the existence of abundant available approaches for certain cases, there is still no general valid solution. According to the literature, the approaches dedicated to object manipulation are categorized into two major streams, engineering-based methodologies and vision-based strategies.

Regarding the first category and given the importance of grasping for robots, a range of approaches have been proposed. Up to the last decade, most of these techniques relied on complete and accurate 3D models of the objects, in order mechanical operations accompanied by conventional methods to be employed. Building accurate models for an efficient representation of objects constitutes a very challenging task that often is sufficiently accomplished via laser scanning. A system for grasping 3D objects with unknown geometry using a Salisbury robotic hand was presented in [5],where each object was placed on a motorized and rotated table under a laser scanner in order a set of 3D points to be generated. These points were combined to be form a 3D model. A framework of automatic grasping of unknown objects via a laser-range scanner and a simulation environment was developed in [6]. Furthermore, a method for the adequate accomplishment of industrial bin picking tasks was presented in [7]. The authors proposed a system that provides accurate 3D models of objects that are further exploited in order to perform precise grasping operations. However, the proposed super quadrics based object modeling approach can only be used for rotationally symmetric objects. Moreover, a technique to calculate possible grasping points for unknown objects with the help of the flat top surfaces of the objects based on a laser-range scanner system was published in [8]. Additionally, surface properties, such as friction and compliance are of basic importance in the grasping process. Nevertheless, a global metric cannot easily describe such attributes, whilst they are often modeled as being uniform for a whole object. An alternative approach to the object manipulation problem is the use of statistical learning methods. For instance, de Granville et al. [9] examined the problem of representing the orientation of a hand as it approaches an object, and determined the feasibility of extracting canonical grasps from a human demonstration. Canonical grasps were represented using clustering procedures based on a combination of distributions. Another approach [10] involves combining analytical and empirical methods by segmenting an object into a set of super quadratics and then learning which ones are more suitable for grasping. According to the literature, an acceptable solution to the object manipulation problem could be given by integrating geometrical attributes of objects based on CAD models. Relatively simple CAD wire structures of objects are used by model-based methods [11], [12].

In this paper we present a new solution to the problem of training a robotic gripper in order to execute manipulation tasks. The suggested technique is composed of two units, namely the training and the testing modules. During training

978-1-4577-1775-8/12/$26.00 ©2012 IEEE

we adopt an ontology-based approach that is responsible for building a database holding information about the trained objects. While other related projects depend on oversampled datasets producing high computational burden, the proposed work establishes compact and abstract representations of trained models to be exploited in the grasping procedure. Furthermore, since the liaison between input and output spaces is a non-linear one, we solve the interpolation problem by employing a fuzzy-based strategy. We have comparatively studied the performance of our method against other related works, whilst experimental results justify our choices. An accurate estimation of the 3D pose of an object target is obtained during the testing phase and through the efficient exploitation of the established database.

II. Related Work

During the last few years, ontology-based architectures have found solid ground in computer vision and especially in image understanding applications. In one of the early works in this field [13] a method for building a scalable system capable of examining images and accurately classifying the latter based on their visual content was presented. Mailot et. al [14] proposed a visual concept ontology accompanied by a dedicated knowledge acquisition toll to enable knowledge-based low-level cognitive vision. Ontologies have also been studied in medical imaging frameworks in order to facilitate machine-based reasoning that depends upon additional interpretive semantics. In particular, in [15] an approach that emphasizes in designing and implementing a system to formally annotate medical images captured to aid the diagnosis and management of breast cancer, was presented. Additionally, in [16] an ontology-based architecture proved to be of fundamental importance in the task of formulating the information needed for adequate image retrieval.

Earlier work on grasping using vision feedback was based on modeling an object as a set of primitive shapes, such as spheres, cylinders, cones and boxes that establish a set of rules for generating grasp configurations. A Support Vector Machine (SVM) solution capable of learning a grasp quality measure that corresponds to the grasping parameters representing the degrees of freedom of a hand, was presented in [17]. There are two important issues that arise during the designing process of an image-based grasping routine. First, adequate information is required by the robot to identify various objects in the environment. This data must be unique so that a target-object to be recognized and robustly invariant to image transforms. Secondly, specific and stable features have to be defined in image plane to facilitate the robot guidance to grasp the target object based on an image-based visual servo controller design. The majority of image understanding frameworks incorporate appearance-based approaches that use global or local features to describe individual targets. Several methods have been proposed for local-feature detection, e.g Harris corner detector [18], Shi-Tomasi features [19], Scale Invariant Feature Transform (SIFT) [20] etc. All object recognition and localization systems depend on successful extraction of sufficient number of features for distinguish different objects. With a view to the efficient estimation of the location of a target in the working space, in [21], a path planning-based method that emphasizes in the calculation of a free trajectory for the end-effector from its current location to that of the targets was presented. The advantage of such a strategy is that the trajectory of the end-effector can be optimized according to certain criterion. The use of Rapidly-Exploring Random Trees (RRTs) algorithm was presented in [22] to cope with multiple possible grasping strategies according to the objects shape and orientation. In [23], a 3D object recognition and pose estimation method based on combining photometric SIFT features and geometric ones in a sequence of imagery scenes was proposed. Particle filtering techniques are then applied and particles representing possible poses of object are generated for each feature.

To the best of our knowledge, the closest works to our paper are presented in [24] and [25]. In [24] the authors present a novel computer vision technique for objects depth estimation suitable for adoption by any two-part algorithm, i.e. a technique employing both a detector and a descriptor. It is based on the observation that the features extracted from any two-part algorithm correspond to spots on the object's surface and their center of mass is related to the one of the objects. Thus, by extracting these features at known positions of the sought object, its distance from the camera can be estimated. The proposed technique was tested on the two most common two-part algorithms, namely the SIFT and the SURF [26], and was found to outperform with the first one. Furthermore, its efficiency depends on the distribution of object's features over its surface and as a result, the algorithm fails to estimate object's distance from the camera in case all the object's features detected are located on an occluded part of it. This method is capable of estimating only 3 degrees of freedom (DoF) whereas our method provides accurate 6 DoF estimations. Additionally, in [25] each model pose in the working database represents an entry in a hash table designed to predict the 3D parameters of the model. Additionally, the affine projection parameters relating the trained model to the testing image are estimated by a least-squares solution. However, Lowe's work accounts only for affine transformations where, in most of the cases, affine is accompanied by projective ones. Furthermore, the least-squares solution assumes a linear relation between input and output models. To address the interpolation issue we employ a fuzzy-based interpolation structure that accounts also for non-linear relations between input and output spaces.

III. Ontologies

It is essential to design and apply imaging modules as well as link percepts to motor commands via an ontology-based framework, to distinguish grasping points over objects based on their shape, size, and 3D orientation. An ontology is defined as a representation of a set of distinguishable features that enable the efficient modeling of knowledge domain [27]. The ultimate purpose of an ontology is to a) build a semantic network that comprising classes accompanied by

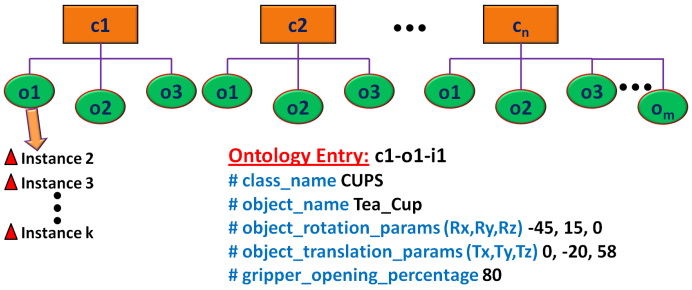

Fig. 1. The proposed ontology-based framework establishes a sophisticated database with intra-classes relations that are further exploited during the testing session of our system. Each class, e.g. cups (c1), is divided into several sub-classes, e.g. tea-cup (c1-o1) or coffee mug (c1-o2), that refereed as objects. The latter are shot under varying viewpoints that correspond to numerous instances of the object (c1-o1-i1).

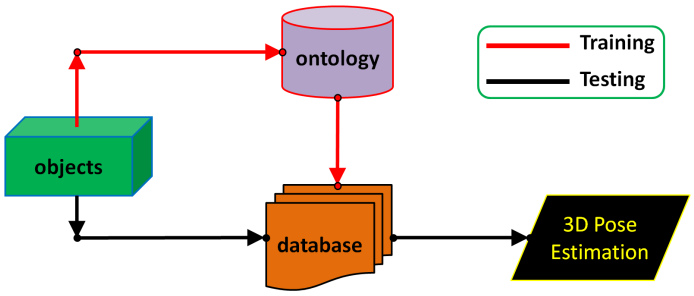

Fig. 2. This figure illustrates the outline of the proposed method. Initially, during the training session, we capture several objects under varying viewpoints in order to establish the ontology-based database. The testing session of our system is responsible for providing an accurate 3D pose estimation of a target in a scene. The latter is accomplished through the ontology architecture that, among others, incorporates vital attributes, i.e. rotation parameters, gripper's opening percentage, to be given to the gripper in order to perform a grasp.

respective properties and b) account for inter-relationships among class members. The extracted clear features express a high-level structure that encapsulates information about the actual meaning and physical attributes of the available data. It is apparent that through an ontology-based framework, one can establish a sophisticated structure that efficiently characterize both hierarchical and relational models, without however, the loss of information regarding the actual knowledge of classes, members and their bindings. According to [28], ontologies are characterized by an efficient data handling scheme that enables sufficient blending of several datasets accompanied by their internal properties, i.e. liaisons to other databases. The high adoption rate of ontology-based systems into several applications is mainly due to their capability to create a semantic language that accredits efficient communication modules between input knowledge (user's environment) and output operations (machine level). Several tasks in robotics and imaging depend upon the sufficient translation of this semantic language in order to be performed. Finally, according to [28], a system that incorporates an ontology-based architecture is empowered by three major advantages: a)Re-usability: An ontology could be accessed several times to provide information regarding the classes and their members; b)Search: As an advanced metadata database, an ontology could be adjusted to serve as an index for efficient data handling; c) Knowledge acquisition: Based on the ability of ontology to remarkably transform environment's data into machine operations, a system may increase both its execution time and reliability.

IV. Algorithm Description

The main idea underlying our algorithm is the establishment of an ontology-based structure that allows, through its efficient exploitation, the efficient accomplishment of 3D object pose estimation tasks. Our ontology-based database consists of images of several objects captured under varying rotation and translation parameters with respect to the camera. The proposed architecture of the ontology-based framework is depicted in Fig. 1.The aforementioned process could be apprehended as the training session of the proposed system where vital geometrical attributes of the trained objects are stored, through ontologies, in an advanced database. An accurate estimation of the 3D pose of a testing target is obtained during the testing phase of the algorithm that incorporates the efficient exploitation of knowledge obtained through ontologies. The outline of the proposed method is illustrated in Fig. 2. In the following passage we analytically present the main components of the training and testing sessions, respectively.

Training Session: This phase of the algorithm is devoted to the construction of the database of our system. As stated above, we decided to adopt an ontology-based architecture that is able to link together, in a sophisticated manner, several geometrical attributes of the trained objects. These characteristics are of fundamental importance during the grasping process, whilst corresponding to the rotation and translation parameters accompanied by the respective grippers opening percentage. It is apparent that, the training of our system is performed off-line and it is a time demanding procedure. Several objects were shot, whilst their groundtruth measurements were stored in the database. In Fig.1 the ontology-based dataset, which holds information regarding the object class, e.g. c1, the object itself, e.g. c1-o1 and its additional instances c1-o1-i1, is presented.

Testing Session: During this process we put a testing object in front of a robotic manipulator and within its working volume. The ultimate goal of our method is to provide to the gripper with an accurate estimation of the 3D pose of the target with a view to the adequate accomplishment of a manipulation task. Towards this end, we propose a fuzzy-based architecture that enables sufficient interpolation capacities between two known instances of the target. While other related projects either utilize extensive matching operations (e.g. the test image is compared with all training ones) or lack interpolation capabilities, the proposed architecture address this issues by encompassing minimum matching procedures accompanied by efficient interpolation capacities. This is accomplished by matching the testing image with limited training instances of the object, resulting in minimum execution time.

V. Experimental Results

Initial experimental results provide evidence of efficient accomplishment of visual servoing tasks. The following figures Fig. 3 and Fig. 4, represent the operations executed through the training and testing sessions, respectively. Fig. 5 illustrates the robotic arm that is employed throughout the experiments. We have comparatively studied the performance of our method against the related works of Lowe [25] and Kouskouridas et. al [24]. The first stands for a method that adopts a least squares solution to solve for the 3D object pose estimation problem. On the other hand, the approach presented in [24] is capable of estimating only the translation parameters (only 3 DoFs) of an object. In imaging applications, a very challenging aspect comes to light in cases where the testing object is partially occluded. Unlike human beings that are capable of simultaneously recognizing an object and estimating its pose in such conditions, computer vision algorithms fall short to achieving similar performance. We address this issue by expanding our ontology with images of partially occluded objects that are introduced in the existing database with the percentage of obstruction lying in the range [0-95].

Fig. 3. During the training session several objects are captured under varying viewpoints. Additionally, we utilize the SCORBOTE-ER Vplus robotic arm, which is a vertical articulated robot with 6 DoFs, in order perform object manipulation tasks. Throughout this process we store in the ontology-based framework the corresponding rotation and translation parameters accompanied by the respective opening percentage of the gripper.

Additionally, we adopt the evaluation criterion presented in [29] so as to comparatively appraise the performance of the proposed 3D pose module against partial occlusions. The adopted evaluation metric considers as successful a measurement of the 3D pose of an object in cases where the error of the computed rotation parameters is less than 5^0. Fig. 6 depicts the superiority of our work against other related projects, whilst providing evidence of being more tolerant to partial occlusions. The proposed ontology-based 3D pose estimation technique proved to hold occlusion invariance capacities compared to the works of Kouskouridas et.al [24] and Lowe et al.[25].

Regarding the fuzzy-based interpolation module, the two instances that produce the maximum feature correspondences between training and testing images, are used as input values

Fig. 4. Initial experimental results regarding only the 3D pose estimation module of the proposed method. The accurate estimation of the 3D configuration of an object is fed to the controller of the SCORBOTE-ER Vplus robotic arm in order to perform a grasp.

Fig. 5. The SCORBOTE-ER Vplus robotic arm that is employed through the experimental setups of the proposed method.

of a Fuzzy Inference System (FIS). The output of the latter corresponds to an accurate interpolated estimation of the 3D pose of the testing object. The inputs of the FIS share identical membership functions (MF), whilst they are guided by a set of fuzzy rules to ensure that the manipulator will adequately grasp the testing object.

In Fig. 7 images of 6 different objects belonging to 2 different classes are illustrated. These images are fed into the ontology-based architecture during training in order to acquire the adequate mapping between input and output spaces. The proposed framework provides the pillars of efficient object manipulation through its ontology-based architecture while indicative visual results are shown in Fig 8. It is apparent that,

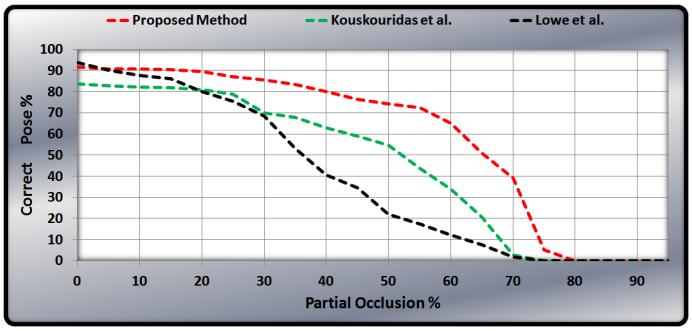

Fig. 6. The performance of our method against partial occlusions is comparatively evaluated with the works of Kouskouridas et al. [24] and Lowe et al. [25].

objects belonging to the same class share common grasping points, i.e. a car is grasped at its center of mass and a cup at its handle.

VI. Conclusions

We have presented a sophisticated framework that provides a solid solution to the problem of vision-based object manipulation. While other contemporary frameworks crave extensive supervision during training along with large databases holding images of several objects under varying viewpoints, our ontology-based structure decreases computational burden and minimizes execution times though its advanced architecture. During training we acquire several images of different objects that are stored along with their 3D pose groundtruth measurements and gripper's opening percentage into the database. Experimental results justified our choice to adopt an ontology-based approach, while providing evidence of low error rate. As far as future work is concerned, we aim at introducing more info into the ontology-based system such as gripper's pressure feedback and utilize machine learning techniques in order to efficiently represent the non-linear liaison between input and output spaces.

References

[1] R. Alicia CASALS, E. PORTELL, C. Xavier, and J. CONTIJOCI, "Capdi: A robotized kitchen for the disabled and elderly," *Assistive technology on the threshold of the new millennium*, vol. 6, p. 346, 1999.

[2] C. Martens, N. Ruchel, O. Lang, O. Ivlev, and A. Graser, "A friend for assisting handicapped people," *Robotics & Automation Magazine, IEEE*, vol. 8, no. 1, pp. 57–65, 2001.

[3] O. Ivlev, C. Martens, and A. Graeser, "Rehabilitation robots friend-i and friend-ii with the dexterous lightweight manipulator," *Restoration of Wheeled Mobility in SCI Rehabilitation*, vol. 17, pp. 111–123, 2005.

[4] A. Remazeilles, C. Leroux, and G. Chalubert, "SAM: a robotic butler for handicapped people," in *Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE International Symposium on*. IEEE, 2008, pp. 315–321.

[5] S. Stansfield, "Robotic grasping of unknown objects: A knowledge-based approach," *The International journal of robotics research*, vol. 10, no. 4, pp. 314–326, 1991.

[6] B. Wang, L. Jiang, J. Li, and H. Cai, "Grasping unknown objects based on 3d model reconstruction," in *Advanced Intelligent Mechatronics. Proceedings, 2005 IEEE/ASME International Conference on*. IEEE, 2005, pp. 461–466.

[7] F. Boughorbel, Y. Zhang, S. Kang, U. Chidambaram, B. Abidi, A. Koschan, and M. Abidi, "Laser ranging and video imaging for bin picking," *Assembly Automation*, vol. 23, no. 1, pp. 53–59, 2003.

[8] M. Richtsfeld and M. Zillich, "Grasping unknown objects based on 21/2d range data," in *Automation Science and Engineering, 2008. CASE 2008. IEEE International Conference on*. IEEE, 2008, pp. 691–696.

[9] C. de Granville, J. Southerland, and A. Fagg, "Learning grasp affordances through human demonstration," in *Proceedings of the International Conference on Development and Learning (ICDL06)*, 2006.

[10] S. El-Khoury and A. Sahbani, "A new strategy combining empirical and analytical approaches for grasping unknown 3d objects," *Robotics and Autonomous Systems*, vol. 58, no. 5, pp. 497–507, 2010.

[11] P. Azad, T. Asfour, and R. Dillmann, "Combining appearance-based and model-based methods for real-time object recognition and 6d localization," in *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*. IEEE, 2006, pp. 5339–5344.

[12] K. Yamazaki, M. Tomono, T. Tsubouchi, and S. Yuta, "A grasp planning for picking up an unknown object for a mobile manipulator," in *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*. IEEE, 2006, pp. 2143–2149.

[13] C. Breen, L. Khan, and A. Ponnusamy, "Image classification using neural networks and ontologies," in *Database and Expert Systems Applications, 2002. Proceedings. 13th International Workshop on*. IEEE, 2002, pp. 98–102.

[14] N. Maillot, M. Thonnat, and A. Boucher, "Towards ontology-based cognitive vision," *Machine Vision and Applications*, vol. 16, no. 1, pp. 33–40, 2004.

[15] B. Hu, S. Dasmahapatra, P. Lewis, and N. Shadbolt, "Ontology-based medical image annotation with description logics," in *Tools with Artificial Intelligence, 2003. Proceedings. 15th IEEE International Conference on*. IEEE, 2003, pp. 77–82.

[16] E. Hyvönen, A. Styrman, and S. Saarela, "Ontology-based image retrieval," in *Towards the semantic web and web services, Proceedings of XML Finland 2002 Conference*, 2002, pp. 15–27.

[17] R. Pelossof, A. Miller, P. Allen, and T. Jebara, "An svm learning approach to robotic grasping," in *Robotics and Automation, 2004. Proceedings. ICRA'04. 2004 IEEE International Conference on*, vol. 4. IEEE, 2004, pp. 3512–3518.

[18] C. Harris and M. Stephens, "A combined corner and edge detector," in *Alvey vision conference*, vol. 15. Manchester, UK, 1988, p. 50.

[19] J. Shi and C. Tomasi, "Good features to track," in *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*. IEEE, 1994, pp. 593–600.

[20] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[21] T. Takahama, K. Nagatani, and Y. Tanaka, "Motion planning for dual-arm mobile manipulator-realization of tidying a room motion," in *Robotics and Automation, 2004. Proceedings. ICRA'04. 2004 IEEE International Conference on*, vol. 5. IEEE, 2004, pp. 4338–4343.

[22] Y. Hirano, K. Kitahama, and S. Yoshizawa, "Image-based object recognition and dexterous hand/arm motion planning using rrts for grasping in cluttered scene," in *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on*. IEEE, 2005, pp. 2041–2046.

[23] S. Lee, E. Kim, and Y. Park, "3d object recognition using multiple features for robotic manipulation," in *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*. IEEE, 2006, pp. 3768–3774.

[24] R. Kouskouridas, A. Gasteratos, and E. Badekas, "Evaluation of two-part algorithms for objects' depth estimation," *Computer Vision, IET*, vol. 6, no. 1, pp. 70–78, 2012.

[25] D. Lowe, "Object recognition from local scale-invariant features," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, vol. 2. Ieee, 1999, pp. 1150–1157.

[26] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Computer Vision–ECCV 2006*, pp. 404–417, 2006.

[27] T. Gruber *et al.*, "A translation approach to portable ontology specifications," *Knowledge acquisition*, vol. 5, no. 2, pp. 199–220, 1993.

[28] N. Shadbolt, W. Hall, and T. Berners-Lee, "The semantic web revisited," *Intelligent Systems, IEEE*, vol. 21, no. 3, pp. 96–101, 2006.

[29] S. Hinterstoisser, S. Benhimane, and N. Navab, "N3m: Natural 3d markers for real-time object detection and pose estimation," *ICCV*, pp. 1–7, 2007.

Fig. 7. A small portion of the ontology-based database. Here 6 different objects belonging to 2 different classes (e.g. cars and cups) are presented.

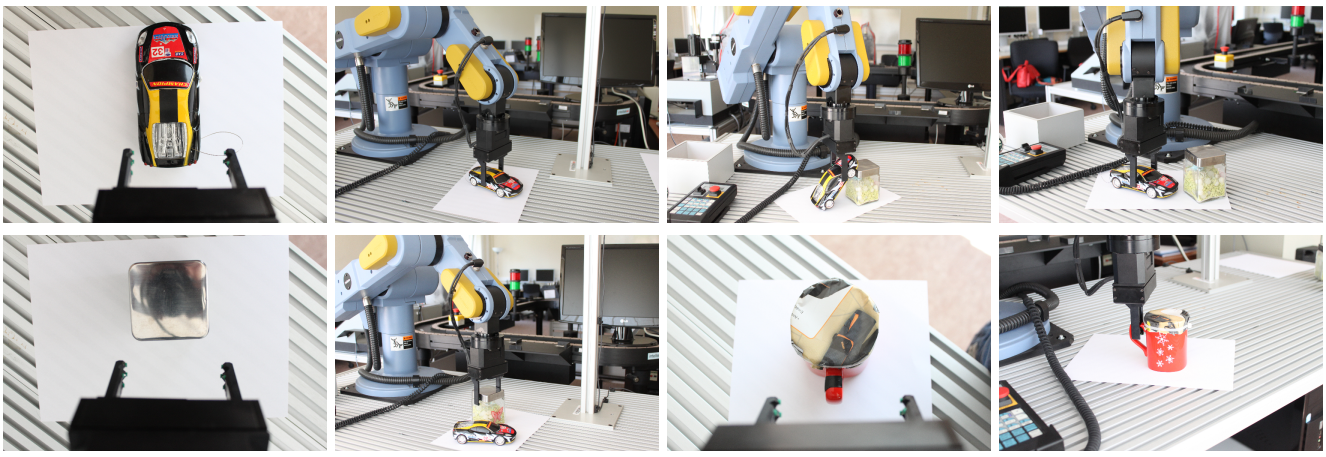

Fig. 8. The proposed ontology-based architecture provides the pillars of efficient accomplishment of object manipulation tasks. Through ontologies we calculate the grasping point of the testing object with a view to cars to be grasped at their center of mass and cups at their handle.