Learning Objects and Grasp Affordances through Autonomous Exploration

Dirk Kraft[1], Renaud Detry[2], Nicolas Pugeault[1], Emre Başeski[1], Justus Piater[2], and Norbert Krüger[1]

[1] University of Southern Denmark, Denmark.
{kraft,nicolas,emre,norbert}@mmmi.sdu.dk
[2] University of Liège, Belgium.
{Renaud.Detry,Justus.Piater}@ULg.ac.be

Abstract. We describe a system for autonomous learning of visual object representations and their grasp affordances on a robot-vision system. It segments objects by grasping and moving 3D scene features, and creates probabilistic visual representations for object detection, recognition and pose estimation, which are then augmented by continuous characterizations of grasp affordances generated through biased, random exploration. Thus, based on a careful balance of generic prior knowledge encoded in (1) the embodiment of the system, (2) a vision system extracting structurally rich information from stereo image sequences as well as (3) a number of built-in behavioral modules on the one hand, and autonomous exploration on the other hand, the system is able to generate object and grasping knowledge through interaction with its environment.

1 Introduction

We describe a robot vision system that is able to autonomously learn visual object representations and their grasp affordances. Learning takes place without external supervision; rather, the combination of a number of behaviors implements a bootstrapping process that results in the generation of object and grasping knowledge.

Learning of objects and affordances has to address a number of sub-aspects related to the object aspect of the problem (**O1**–**O3**) and to the action aspect (**A1**, **A2**):

O1 What is an object, i.e., what is "objectness"?
O2 How to compute relevant attributes (shape and appearance) to be memorized?
O3 How can the object be recognized and how can its pose determined?
A1 What is the (preferably complete) set of actions it affords?
A2 What action is triggered in a concrete situation?

A satisfactory answer to **O1** is given by Gibson [1] as temporal permanence, manipulability and constrained size in comparison to the agent. Note that manipulability can only be tested by acting on the potential object, and hence requires an agent with at least minimal abilities to act upon objects. For **O2** there are requirements discussed in the vision literature. In many systems, in particular in the context of robotics, the object shape is given a priori by a CAD representation and is then used for object identification and pose estimation (see, e.g., Lowe [2]). However, CAD representations are not

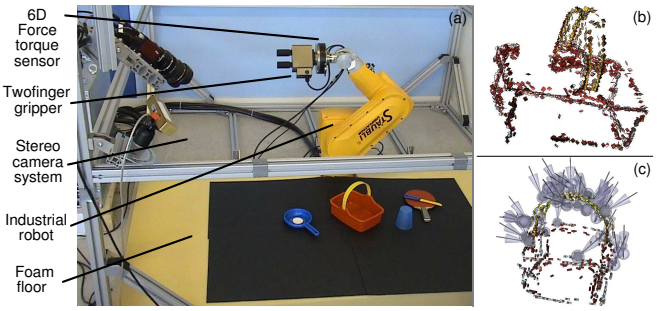

Fig. 1. (a) Hardware setup. **(b, c)** Outcome of the learning process in form of a geometric object model **(b)** and a grasp density **(c)**.

available in a general context and hence for a cognitive system, it is important that it is able to learn object representations from experience. Important issues to be considered when coding objects are that the information memorized (1) is useful for the tasks to be performed with the representations (e.g., for matching or grasping), (2) is efficiently accessible internally, and (3) requires little storage space. **O3** has been addressed widely in the computer vision literature. In particular in the context of grasping, besides the actual recognition, the determination of the pose is of importance since it allows the system to associate learned grasps in object coordinates to an observes object instance.

In cognitive robotics, the automatic association of grasping actions to objects (**A1**, **A2**) is referred as learning *affordances* [3]. For maximum flexibility, it is desirable to represent the set of grasp affordances to the most complete extent possible **A1**. There are attempts to compute such a complete set by analytic means [4] which however in general require a pre-existing 3D surface model. In addition, analytic modeling of the interaction between a gripper and an object surface, besides being very time consuming, is very complex since it involves for example friction parameters that are difficult to estimate. Hence we decided to achieve such knowledge by letting the robot experiment in the real world. The decision on the grasp to be performed in a given situation **A2** involves additional considerations, in particular work-space constraints.

This paper describes a system that approaches the above problems in a way that does not require any explicit prior object or affordance knowledge. Instead, the system generates object and grasping knowledge by pure exploration (see Fig. 1). We present a robot-vision system driven by a number of basic behaviors that generate object and grasp-affordance knowledge within two learning cycles. In the first cycle, a multi-modal visual representation covering geometric as well as appearance information (see Fig. 2) is extracted by actively manipulating a potential object. In the second cycle, the robot "plays" with the object by trying out various grasping options. Successful grasp parameters are associated to the object model, leading to an increasingly complete description of the object's grasp affordance. This is done by largely autonomous exploration with only very little interaction between robot and humans. Only interaction that puts the system into a state from which learning can continue is permitted (e.g., putting the ob-

ject back after playing has pushed it out of the workspace). No high level information such as object identities or demonstrations of ways to grasp it is given to the system.

However, this is not to imply that the system does not make use of any prior knowledge. Quite to the contrary, the system can only perform the complex learning tasks by utilizing a large degree of innate knowledge about the world with which it interacts. However, this knowledge is of rather generic structure. Specifically, the system

- has knowledge about its embodiment and the consequences of its movements in the three-dimensional world (kinematics and the ability to plan motions),
- has a sophisticated early cognitive vision (ECV) system [5–7] that provides semantically rich and structured 2D and 3D information about the world. This system contains prior knowledge about image features and their relations.
- has a set of procedural prior knowledge about how to: *a)* grasp unknown objects based on visual features, *b)* create visual object models based on object motion, *c)* evaluate a grasping attempt, *d)* estimate object pose based on a learned visual model and *e)* generalize from individual grasps to grasping densities.

This paper describes the various sub-modules and their interaction that lead to the autonomous learning of objects and associated grasp affordances. We show that, based on a careful balance of generic prior knowledge and exploratory learning, the system is able to generate object and grasping knowledge while exploring the world it acts upon. While the sub-modules have already been described [7–12], the novel part of this work is the integration of these components into an autonomously learning system.

2 State of the Art

Concerning the aspects **O1**–**O3**, the work of Fitzpatrick and Metta [13] is closely related to our object learning approach since the overall goal as well as the hardware setup are similar: finding out about the relations of actions and objects by exploration using a stereo system combined with a grasping device. We see the main distinguishing feature of this work to our approach in the amount of prior structure we use. For example, we assume a much more sophisticated vision system. Also, the use of an industrial robot allows for a precise generation of scene changes exploited for the extraction of the 3D shape of the object. Similar to this work, we initially assume "reflex-like" actions that trigger exploration. However, since in our system the robot knows about its body and about the 3D geometry of the world and since the arm can be controlled more precisely, we can infer more information from having physical control over the object in terms of an exact association of visual entities based on proprioceptive information. Therefore, we can learn a complete 3D representation of the object (instead of 2D appearance models) that can then be linked to pose estimation. Modayil and Kuipers [14] addressed the problem of detection of objectness and the extraction of object shape in the context of a mobile robot using laser. Here also motion information (in terms of the odometry of the mobile robot) is used to formulate predictions. In this way, they can to extract a 2D cross section of the 3D environment, albeit only in terms of geometric information.

Object grasp affordances (**A1**, **A2**) can emerge in different ways. A popular approach is to compute grasping solutions from the geometric properties of an object,

typically obtained from a 3D object model. The most popular 3D model for grasping is probably the 3D mesh [4], obtained e.g. from CAD or superquadric fitting [15]. However, grasping has also successfully been achieved using models consisting of 3D surface patches [16], 3D edge segments [8,12], or 3D points [17]. When combined with pose estimation, such methods allow a robot to execute a grasp on a specific object. In our system, we start with edge-based triggering of grasping actions [8,12] which is then verified by empirical exploration. This requires a system that is able to perform a large number of actions (of which many will likely fail) in a relatively unconstrained environment, this requires a representation of grasp affordances that translate the grasping attempts into a probabilistic statement about grasp success likelihoods.

Learning grasp affordances from experience was demonstrated by Stoytchev [18, 19]. In this work, a robot discovers successful grasps through random exploratory actions on a given object. When subsequently confronted with the same object, the robot is able to generate a grasp that should present a high likelihood of success.

Means of representing continuous grasp affordances have been discussed by de Granville et al. [20]. In their work, affordances correspond to object-relative hand approach orientations, although an extension is underway where object-relative positions are also modeled [21].

3 The Robot-Vision System

Hardware setup: The hardware setup (see Fig. 1) used for this work consists of a six-degree-of-freedom industrial robot arm (Stäubli RX60) with a force/torque (FT) sensor (Schunk FTACL 50-80) and a two-finger-parallel gripper (Schunk PG 70) attached. The FT sensor is mounted between robot arm and gripper and is used to compute to detect collision. Together with the foam ground, this permits graceful reactions to collision situations which might occur because of limited knowledge about the objects in the scene. In addition, a calibrated stereo camera system is mounted in a fixed position in the scene. The system also makes use of a path-planning module which allows it to verify the feasibility of grasps with respect to workspace constraints and 3D structure discovered by the vision system.

Early cognitive vision system: In this work, we make use of the visual representation delivered by an early cognitive vision system [5–7]. Sparse 2D and 3D features, so-called *multi-modal primitives*, are created along image contours. 2D features represent a small image patch in terms of position, orientation, phase. These are matched across two stereo views, and pairs of corresponding 2D features permit the reconstruction of a 3D equivalent. 2D and 3D primitives are organized into perceptual groups in 2D and 3D (called 2D and 3D contours in the following). The procedure to create visual representations is illustrated in Fig. 2 on an example stereo image pair. Note that the resultant representation not only contains appearance (e.g., color and phase) but also geometrical information (i.e., 2D and 3D position and orientation).

The sparse and symbolic nature of the multi-modal primitives allows for the coding of relevant perceptual structures that express relevant spatial relations in 2D and 3D [22]. Similar relations are also defined for 2D and 3D contours to enable more

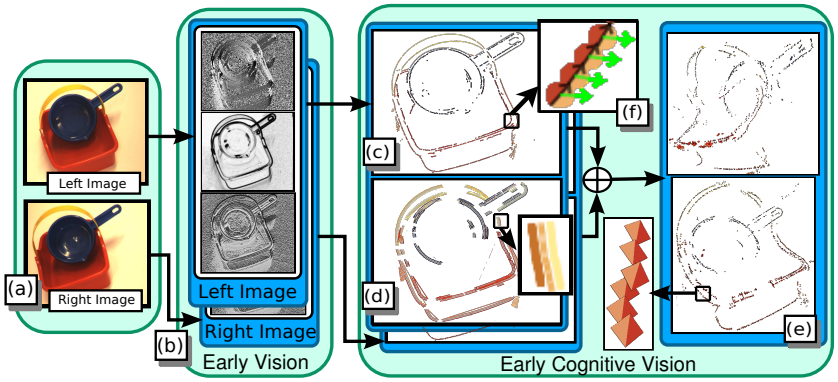

Fig. 2. An overview of the visual representation. **(a)** Stereo image pair, **(b)** Filter responses, **(c)** 2D primitives, **(d)** 2D contours, **(e)** 3D primitives, **(f)** close-up of (c).

global reasoning processes. In our context, the coplanarity and co-colority relations (i.e., sharing similar color structure) permit the association of grasps to pairs of contours. Figure 3(c) illustrates the association of grasp affordances to an unknown object by using appearance and geometrical properties of the visual entities. The formalization of the visual change of a primitive under a rigid-body motion allows for the accumulation of the primitives belonging to the object (see Sect. 4).

4 The First Learning Cycle: Birth of the Object

Within the first learning cycle, the "objectness" of visually-detected structure in the scene **O1** is first tested by trying to obtain physical control over such detected structure and then manipulating it. In case the structure changes according to the movement of the robot arm, a 3D object representation is extracted.

Initial grasping behavior: To gain physical control over unknown objects a heuristic grasp computation mechanism based on [8, 12] is used. Pairs of 3D contours that share a common plane and have similar colors suggest a possible grasp; see Fig. 3(a–c). The grasp location is defined by the position of one of the contours. Grasp orientation is calculated from the common plane defined by the two features and the orientation of the contour at the grasp location. Every contour pair fulfilling this criteria generates multiple possible grasps (see Fig. 3(b) for one such possible grasp definition).

Accumulation: Once the object has been successfully grasped, the system moves it to present it to the camera from a variety of perspectives to accumulate a full 3D symbolic model of the object [7]. This process is based on the combination of three components. First, all primitives are tracked over time and filtered using an Unscented Kalman Filter based on the combination of prediction, observation and update stages. The prediction stage uses the system's knowledge of the arm motion to calculate the poses of all accumulated primitives at the next time step. The observation stage matches the predicted

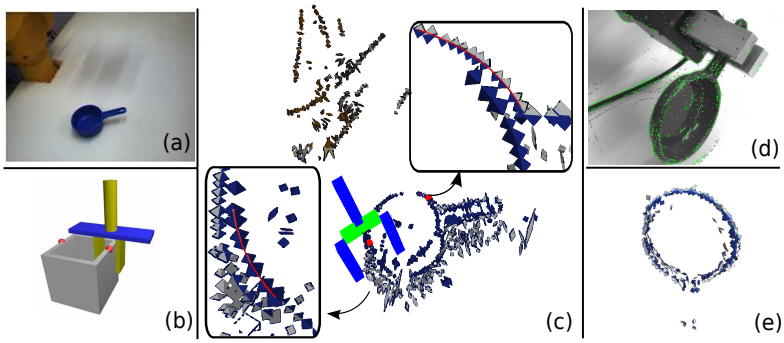

Fig. 3. (a–c) Initial grasping behavior: **(a)** A Scene, **(b)** Definition of a possible grasp based on two contours, **(c)** Representation of the scene with contours generating a grasp. **(d)** A step in the accumulation process where features from the previous scene get matched to the new scene. **(e)** Model extracted by the accumulation process.

primitives with their newly observed counterparts. The update stage corrects the accumulated primitives according to the associated observations. This allows the encoding and update of the feature vector. Second, the confidence in each tracked primitive is updated at each time step according to how precisely the accumulated primitive was matched with a new observation. The third process takes care of preserving primitives once their confidences exceed a threshold, even if they later become occluded for a long period of time. It also ensures that primitives are discarded if their confidence falls below a threshold. New primitives that were not associated to any accumulated primitive are added to the accumulated representation, allowing the progressive construction of a full 3D model. Note that the sparse nature of primitives yields a condensed description.

The learning cycle: Figure 4 (top) shows how the two sub-modules described above interact to generate object models for previously unknown objects. The initial grasping behavior is used to gain physical control over an unknown object. In case no object has been grasped in the process (this is determined using haptic feedback i.e. the distance of the fingers after grasping) another grasping option is executed. After the object has been grasped, the accumulation process is used to generate an object model which is then stored in memory. This process can be repeated until all objects in the scene have been discovered (a naive approach here can be to assume that we have learned all objects if grasping fails for a certain amount of trials). Results of the first learning cycle can be seen in Figs. 1(b), 3(e) and [11].

5 The Second Learning Cycle: Learning Grasp Affordances

In the second learning cycle, the object representation extracted in the first learning cycle is used to determine the pose of the object in case it is present in the scene **O3**. A mechanism such as that triggering the grasps in the first learning cycle generates a large number of grasping options (see Fig. 4 bottom). A random sample of these are then

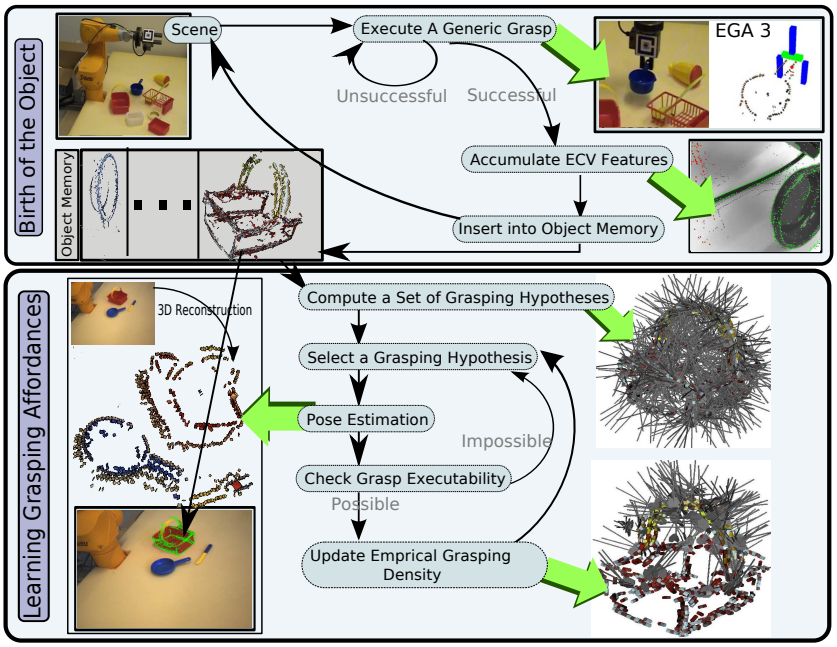

Fig. 4. The two learning cycles and the interaction between them. See text.

tested individually. Successful grasps are then turned into a probability density function that represents a the grasp affordances associated to the object **A1** in the form of the success likelihood of grasp parameters. This grasp density can then be used to compute the optimal grasp in a specific situation **A2** [10]. The second learning cycle is invoked after the first learning cycle has successfully establish the presence and the shape of an object.

Pose estimation: In preparation for pose estimation, a structural object model is built on top of the set of ECV primitives that has been accumulated in the first learning cycle. An object is modeled with a hierarchy of increasingly expressive object parts [9]. Parts at the bottom of the hierarchy represent ECV primitives. Higher-level parts represent geometric configurations of more elementary parts. The single top part of a hierarchy represents the object. A hierarchy is implemented as a Markov tree, where parts correspond to hidden nodes, and relative spatial relationships between parts define compatibility potentials.

An object model can be autonomously built from a segmented ECV reconstruction [9] as produced by the first learning cycle (Sect. 4). Visual inference of the hierarchical model is performed using a belief propagation algorithm (BP; see, e.g., [23]). BP derives a posterior pose density for the top part of the hierarchy, thereby producing a probabilistic estimate of the object pose.

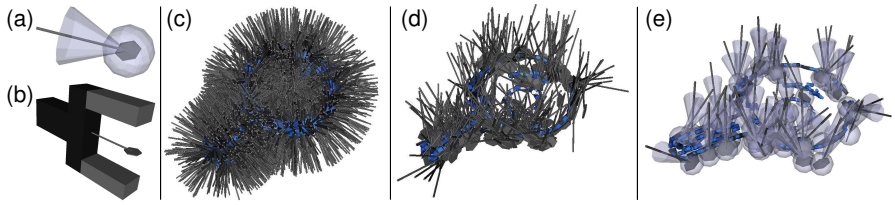

Fig. 5. Grasp density representation. **(a)** illustrates a particle from a nonparametric grasp density, and its associated kernel widths: the translucent sphere shows one position standard deviation, the cone shows the variance in orientation. **(b)** illustrates how the schematic rendering used in the top image relates to a physical gripper. **(c)** Samples from a grasp hypothesis density. **(d)** Samples from an empirical density learned from the hypothesis density in (c). **(e)** A 3D rendering of the kernels supporting the empirical grasp density in (d).

Grasp densities: When formalizing object grasp affordances, we mean to organize and memorize, independently of grasp information sources, the whole knowledge that an agent has about the grasping of an object. By *grasp affordance*, we refer to the different ways of placing a hand or a gripper near an object so that closing the gripper will produce a stable grip. The grasps we consider are parametrized by a 6D gripper pose and a grasp (preshape) type. The gripper pose is composed of a 3D position and a 3D orientation, defined within an object-relative reference frame.

We represent the grasp affordance of an object through a continuous probability density function defined on the 6D object-relative gripper pose space $SE(3)$ [10]. The computational encoding is *nonparametric*: A density is simply represented by the samples we see from it. The samples supporting a density are called *particles* and the probabilistic density in a region of space is given by the local density of the particles in that region. The underlying continuous density is accessed by assigning a kernel function to each particle – a technique generally known as *kernel density estimation* [24]. The kernel functions capture Gaussian-like shapes on the 6D pose space $SE(3)$ (see Fig. 5).

A grasp affordance is attached to the hierarchical model as a new *grasp node* linked to the top node of the network. The potential between grasp node and top node is defined by the grasp density. When an object model is visually aligned to an object instance, the grasp affordance of the object *instance* is computed through the same BP process as used for visual inference. Intuitively, this corresponds to transforming the grasp density to align it to the current object pose, yet explicitly taking the uncertainty on object pose into account to produce a posterior grasp density that acknowledges visual noise.

The learning cycle: Affordances can initially be constructed from a grasp generation method that produces a minimum proportion of successful grasps (e.g., the initial grasping behavior in Sect. 4). In this work we used an approach where we initially use grasp hypotheses at random orientations at the position of the ECV primitives of the object model. We call affordance representations built with any of these weak priors *grasp hypothesis densities* [10]. These are attached to the object hierarchical model, which will allow a robotic agent to execute *random samples* from a grasp hypothesis density under arbitrary object poses, by using the visual model to estimate the 3D pose of the object.

Although grasp hypothesis densities already allow grasping, it is clear that physical experience with an object will add valuable information. We thus use samples from grasp hypothesis densities that lead to a successful grasp to learn *grasp empirical densities*, i.e. grasps that have been confirmed through experience [10]. In this way, we increase grasping performance for the blue pan from 46% to 81%. The process of computing hypothesis densities, pose estimation and execution of random samples from the grasp hypothesis density through which a empirical density is generated is shown in Fig. 4 (bottom).

6 Discussion and Conclusions

The descriptions of the presented sub-modules [7–12] include an evaluation. We therefore only want to reiterate here a few important points that influence the performance and restrict the system. Because of the limitations of the robot system, the objects are limited by size (ca. 5–40 cm) and weight (up to 3 kg). Further restrictions are introduced by the vision system. The objects need to be describable by line-segment features and therefore can not be heavily textured. In addition the used stereopsis process can not reconstruct features on epipolar lines. This can lead to problems for the initial grasping behavior and the pose estimation process, but not the accumulation process.

Besides the blue pan object shown throughout this work we have successfully tested the full system on a toy knife and on a toy basket. The individual sub-components have been tested on more objects.

Autonomous systems benefit from an ability to acquire object and affordance knowledge without external supervision. We have brought together 3D stereo vision, heuristic grasping, structure from motion, and probabilistic representations combining visual features and gripper pose to autonomously segment objects from cluttered scenes and learn visual and affordance models through exploration. This enables an autonomous robot — initially equipped only with some generic knowledge about the world and about itself — to learn about objects and subsequently to detect, recognize and grasp them.

Acknowledgments

This work was supported by the Belgian National Fund for Scientific Research (FNRS) and the EU Cognitive Systems project PACO-PLUS (IST-FP6-IP-027657).

References

1. Gibson, J.: The Ecological Approach to Visual Perception. Houghton Mifflin (1979)
2. Lowe, D.G.: Fitting parametrized 3D-models to images. IEEE Transactions on Pattern Analysis and Machine Intelligence **13**(5) (1991) 441–450
3. Sahin, E., Cakmak, M., Dogar, M., Ugur, E., Ucoluk, G.: To afford or not to afford: A new formalization of affordances toward affordance-based robot control. Adaptive Behavior **15**(4) (December 2007) 447–472
4. Bicchi, A., Kumar, V.: Robotic grasping and contact: A review. In: IEEE Int. Conf on Robotics and Automation. (2000) 348–353

5. Krüger, N., Lappe, M., Wörgötter, F.: Biologically Motivated Multi-modal Processing of Visual Primitives. The Interdisciplinary Journal of Artificial Intelligence and the Simulation of Behaviour **1**(5) (2004) 417–428
6. Pugeault, N.: Early Cognitive Vision: Feedback Mechanisms for the Disambiguation of Early Visual Representation. PhD thesis, Informatics Institute, University of Göttingen (2008)
7. Pugeault, N., Wörgötter, F., Krüger, N.: Accumulated Visual Representation for Cognitive Vision. In Proceedings of the British Machine Vision Conference (BMVC) (2008)
8. Aarno, D., Sommerfeld, J., Kragic, D., Pugeault, N., Kalkan, S., Wörgötter, F., Kraft, D., Krüger, N.: Early reactive grasping with second order 3d feature relations. In: Recent Progress in Robotics: Viable Robotic Service to Human. Springer Berlin / Heidelberg (2008)
9. Detry, R., Pugeault, N., Piater, J.: A probabilistic framework for 3D visual object representation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2009) (accepted).
10. Detry, R., Başeski, E., Krüger, N., Popović, M., Touati, Y., Kroemer, O., Peters, J., Piater, J.: Learning object-specific grasp affordance densities. In: International Conference on Development and Learning. (2009) (to appear).
11. Kraft, D., Pugeault, N., Başeski, E., Popović, M., Kragic, D., Kalkan, S., Wörgötter, F., Krüger, N.: Birth of the Object: Detection of Objectness and Extraction of Object Shape through Object Action Complexes. Special Issue on "Cognitive Humanoid Robots" of the International Journal of Humanoid Robotics **5** (2009) 247–265
12. Popović, M.: An early grasping reflex in a cognitive robot vision system. Master's thesis, The Maersk Mc-Kinney Moller Institute, University of Southern Denmark (2008)
13. Fitzpatrick, P., Metta, G.: Grounding Vision Through Experimental Manipulation. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **361** (2003) 2165 – 2185
14. Modayil, J., Kuipers, B.: Bootstrap learning for object discovery. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) **1** (2004) 742–747
15. Biegelbauer, G., Vincze, M.: Efficient 3D object detection by fitting superquadrics to range image data for robot's object manipulation. In: IEEE International Conference on Robotics and Automation. (2007)
16. Richtsfeld, M., Vincze, M.: Robotic grasping based on laser range and stereo data. In: International Conference on Robotics and Automation. (2009)
17. Huebner, K., Ruthotto, S., Kragic, D.: Minimum volume bounding box decomposition for shape approximation in robot grasping. Technical report, KTH (2007)
18. Stoytchev, A.: Toward learning the binding affordances of objects: A behavior-grounded approach. In: Proceedings of AAAI Symposium on Developmental Robotics. (2005) 17–22
19. Stoytchev, A.: Learning the affordances of tools using a behavior-grounded approach. In: Affordance-Based Robot Control. Volume 4760 of Lecture Notes in Artificial Intelligence (LNAI). Springer (2008) 140–158
20. de Granville, C., Southerland, J., Fagg, A.H.: Learning grasp affordances through human demonstration. In: Proceedings of the International Conference on Development and Learning (ICDL'06). (2006)
21. de Granville, C., Fagg, A.H.: Learning grasp affordances through human demonstration. submitted to the Journal of Autonomous Robots (2009)
22. Baseski, E., Pugeault, N., Kalkan, S., Kraft, D., Wörgötter, F., Krüger., N.: A scene representation based on multi-modal 2D and 3D features. ICCV Workshop on 3D Representation for Recognition 3dRR-07 (2007)
23. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann (1988)
24. Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Chapman & Hall/CRC (1986)