A Strategy for Grasping Unknown Objects based on Co-Planarity and Colour Information

Mila Popović[1], Dirk Kraft[1], Leon Bodenhagen[1], Emre Başeski[1],
Nicolas Pugeault[2], Danica Kragic[3], Tamim Asfour[4], and Norbert Krüger[1]

[1] *Cognitive Vision Lab,*
The Mærsk Mc-Kinney Møller Institute,
University of Southern Denmark,
Campusvej 55, DK-5230 Odense, Denmark
Email: {mila, kraft, lebo, emre, norbert}@mmmi.sdu.dk

[2] *Centre for Vision, Speech and Signal Processing*
University of Surrey
GU2 7XH, Guildford, UK
Email: n.pugeault@surrey.ac.uk

[3] *Centre for Autonomous Systems*
Computational Vision and Active Perception Lab
Royal Institute of Technology
SE-100 44, Stockholm, Sweden
Email: dani@kth.se

[4] *Karlsruhe Institute of Technology*
Institute for Anthropomatics
Humanoids and Intelligence Systems Lab
Adenauerring 2, 76131 Karlsruhe, Germany
Email: asfour@kit.edu

Abstract

In this work, we describe and evaluate a grasping mechanism that does not make use of any specific object prior knowledge. The mechanism makes use of second-order relations between visually extracted multi–modal 3D features provided by an early cognitive vision system. More specifically, the algorithm is based on two relations covering geometric information in terms of a co-planarity constraint as well as appearance based information in terms of co-occurrence of colour properties. We show that our algorithm, although making use of such rather simple constraints, is able to grasp objects with a reasonable success rate in rather complex environments (i.e., cluttered scenes with multiple objects).

Moreover, we have embedded the algorithm within a cognitive system that allows for autonomous exploration and learning in different contexts. First, the system is able to perform long action sequences which, although the grasping attempts not being always successful, can recover from mistakes and more importantly, is able to evaluate the success of the grasps autonomously by haptic feedback (i.e., by a force torque sensor at the wrist and proprioceptive information about the distance of the gripper after a gasping attempt). Such labelled data is then used for improving the initially hard-wired algorithm by learning. Moreover, the grasping behaviour has been used in a cognitive system to trigger higher level processes such as object learning and learning of object specific grasping.

Key words: Vision based grasping, cognitive systems, early cognitive vision

1. Introduction

The capability of robots to effectively grasp and manipulate objects is necessary for interacting with the environ-

ment and thereby fulfil complex tasks. These capabilities need to be implemented and evaluated in natural environments, considering both known and unknown objects. Considering important requirements for the next generation of service robots such as robustness and flexibility, robots should be able to work in unknown and unstructured environments, be able to deal with uncertainties in feature acquisition processes, and should perform fast and reliable. These requirements also assume that the robots are able to deal with initially unknown objects as well as to be able to learn from experience. The work introduced here describes an algorithm for grasping unknown objects as well as the improvement of this algorithm through learning. The basic idea is the modelling and generation of *elementary grasping actions* (see figure 1) – simple perception-action pairs suitable for generation of grasps where very little or no information about the objects to be grasped is known *a-priori*.

The body of work in the area of robotic grasping is significant, see, e.g., [1–10]. We distinguish approaches based on the level of *a-priori* object information used to model the grasping process. In particular, objects to be grasped may be assumed known, that is, both the shape and the appearance of the object are known and used to associate specific grasping strategies to them through exploration, (see, e.g., [2,3]) or different types of supervised learning (see, e.g., [9,10]). When objects are assumed to be unknown (as in our case), the assumptions of the system naturally need to be much more general in order to generate suitable grasping hypotheses (see, e.g., [4]).

Grasping objects in unconstrained environments without any specific prior object knowledge is as discussed above a very difficult problem. Hence, a performance close to 100% is not to be expected. Although humans can solve this problem, it needs to be acknowledged that this skill only develops after years of learning (see, e.g., [11]) and hence is likely to make use of a vast amount of experience with a variety of objects. However, once the object is known to the system, a much higher performance is achievable. The grasping behaviour described in this paper has been used to generate such object knowledge and to learn grasping based on this knowledge, i.e., to build up general world knowledge by learning.

In this work, we describe a grasping behaviour which is connected to two types of learning. First, the behaviour itself can be improved through learning. While performing initial grasping exploration, labeled training data is automatically generated. Learning is achieved by combining known input parameters, and the labeled outcomes of grasping attempts. This learning scheme is described in more detail in section 7.

For the second kind of learning, the initial behaviour can be used to *establish* more sophisticated grasping making use of object knowledge. We give a brief description of this second kind of learning below and refer for details to [12,13]. More specifically, we use the initial object independent grasping behaviour as described in this paper to constitute object shape representations (see also [12]) and

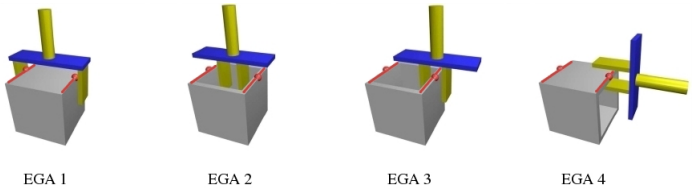

Fig. 1. Elementary grasping actions (EGAs), figure adapted from [15]. The red lines indicate 3D contours that have been reconstructed from stereo images. When contours are connected by relations of co-planarity and co-colourity they trigger generation of EGAs. The red dots symbolize the 3D primitives in the middle of each contour. Those are called 'parent primitives' and are used to compute the EGAs. In case of EGA 2, the gripper fingers are initially closed and the grasp is accomplished by opening fingers and thus applying force to the concave objects from inside out. EGA types 3 and 4 each generate two actions, one for each parent primitive. See also figures 11 and 13.

associate grasping affordances to those and hence realise a grasping based on object prior knowledge. More concretely, the early cognitive vision system (see figure 2) briefly described in section 3 is able to extract 3D representations (see figure 3) of objects by accumulating information over different frames (see [14]). However, a pre-requisite for using this accumulation mechanism is that the robot has physical control over the object (see figure 3a), allowing the robot to perform movements leading to predictable visual transformations. Based on these predictions, filtering processes can be used to eliminate wrong 3D features, leading to reliable 3D object representations (see figure 3b). The physical control is achieved by means of the grasping behaviour described in this paper. Then the robot can perform a controlled movement (mostly a rotation) of the object, that can be used in the accumulation algorithm to extract complete and reliable object representations (as shown in figure 3b) that are then stored in an object memory.

Moreover, the computation of grasping hypotheses based on co–planar contours can also be performed for these stored accumulated representations (see figure 3c) and be tested after a successful pose estimation has been performed (as done, in e.g., [16,17]). This mechanism has been used in the PACO-PLUS system to learn 'grasp densities' which associate grasping affordances to learned object models (see figure 3d and [13]). [1] By that, the grasping mechanism introduced in this paper, which does not require any object prior knowledge has been used to bootstrap a system which generates an object specific grasping mechanism with a higher success rate due to the larger prior being incorporated.

The paper is organised as follows. We first give an overview of the state of the art in robot grasping in section 2, where we also outline distinguishing features of our

[1] Here we only briefly describe the role of the initially object independent grasping behaviour for object knowledge based grasping. Its use for such grasping requires additional complexities such as object memory, pose estimation and a probabilistic representation of object–grasp associations that are beyond the scope of this paper. These are are fully treated in a separate publication (see [13]).

approach in comparison to existing work. In section 3 we describe the visual representations from which the grasps become computed. In section 4, we describe how the grasps are computed from the visual features. The experimental setup and the evaluation of the grasping strategy is outlined in section 5. Section 6 describes implementation on the humanoid robot ARMAR-III. In section 7, we discuss the aspects of fine-tuning the grasping strategy by learning based on the data acquired by autonomous exploration.

2. Related work

In this section, we present an overview of the current research in the area of robotic grasping and relate it to our work.

One area of research in the field of object grasping are analytical approaches (see, e.g., [18,5–7]) that model the interaction between a gripper and an object to compute promising grasps. When contact points between the robot hand and the object are determined and the coefficients of friction between the two materials are known, it is possible to calculate a wrench space - i.e., 6D space of forces and torques that can be applied by the grasp. A force-closure grasp can resist all object motions provided that the gripper can apply sufficiently large forces. These forces can be measured by tactile sensing (see e.g. [8]) and grasp quality can be computed as objective functions which can be further enhanced by optimising the parameters of a dexterous hand (see, e.g., [19,20]). In most of those approaches it is assumed that either the shape properties of the object are known or that these can be easily extracted using visual information which can be difficult in realistic settings.

Related to the analytical approaches are considerations on the robot embodiment. Since robot hands often have many degrees of freedom, the search space of possible grasp configurations is very large. Analytical approaches are therefore usually used together with some heuristics which guide and constrain the optimisation process. For example, heuristically-based grasp generators often include some grasp preshape types (see, e.g., [21,22,4]) based on human grasping behaviour. Domain specific knowledge, e.g. workspace constraints, hand geometry, task requirements or perceptual attributes are also used (see, e.g., [23,24,20]). In addition, simulations can further speed up the learning process (see, e.g., [25,26]).

In industrial applications, the association of grasps to known objects is often done manually or by guiding the gripper directly to an appropriate pose during a training phase where the object is in a known pose. Learning by demonstration (see, e.g., [27–29]) can be a very efficient tool to associate grasps to known objects, in particular when dealing with humanoid robots. Once prior knowledge is present in terms of a 3D object model and defined grasping hypotheses (see, e.g., [30]), the grasping problem is basically reduced to object recognition and pose estimation.

Another approach is learning by exploration. In the recently published work [13], grasp densities become associated to 3D object models which allow for memorising object–grasp associations with their success likelihoods. In this context, a number of learning issues become relevant such as active learning (see, e.g., [31]) and the efficient approximation of grasp quality surfaces from examples (see, e.g., [9]). An interesting approach, which can be positioned in between grasping with and without object prior knowledge, is the decomposition of a scene into shape primitives to which grasps become associated (see, e.g., [20,21]).

Grasping unknown objects is acknowledged to be a difficult problem which varies in respect to the complexity of objects and scenes. Many projects (see, e.g., [32,4,10])

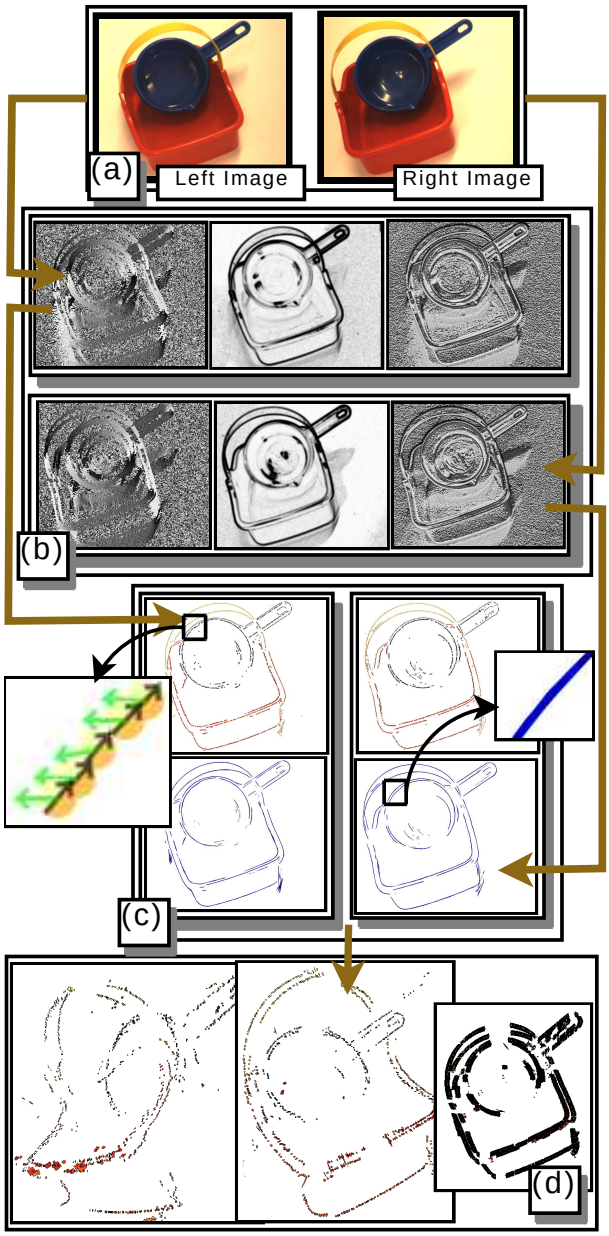

Fig. 2. Different type of information that is available in the representation. **(a)** Original stereo images. **(b)** Filter responses. **(c)** 2D primitives and contours. **(d)** 3D primitives from two different view points and 3D contours.

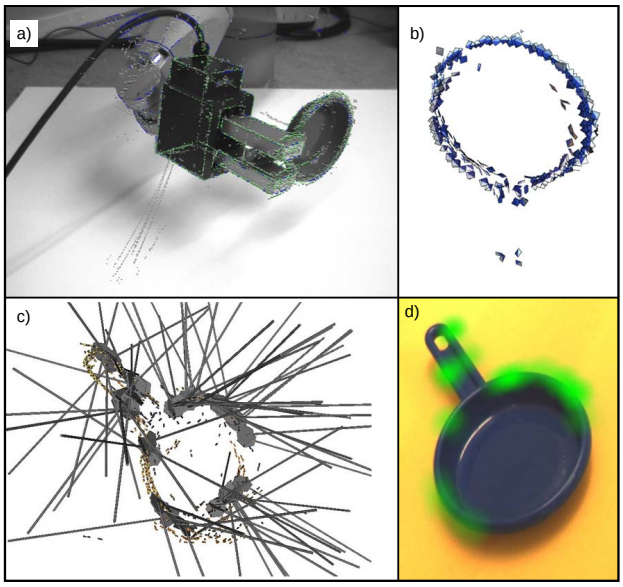

Fig. 3. a) An image of the object held by the robot where tracked 3D primitives are indicated as local line segments. b) Illustration of the object model consisting of successfully tracked 3D-primitives. Note that the hole at the handle originates from the fact that the gripper holds the object at the handle and hence the handle is occluded for the vision system. c) Grasping hypotheses generated by our algorithm extracted based on the learned representation shown in b). d) Projection of grasp densities (note that although the grasping density is a 6D manifold, only 3D positions are shown) extracted from empirically tested grasping hypotheses found being successful.

share the following sequence of steps S1–S4:

S1 Extracting relevant features
S2 Grasp hypotheses generation
S3 Ranking of grasp hypotheses
S4 Execution of the best candidate grasp

The complexity of a system depends on the choice of sensors, the diversity of considered objects, the scene configuration and the kind of a-priori knowledge assumed. A number of examples relies on visual sensors and a simple gripper with 2 or 3 fingers [33–35,4]. In [35–37] the 2D contours of an object are used as a relevant feature, and grasp planning as well as quality evaluation are based on approximating the centre of mass of the object with the geometrical centre of the contour. Often the camera is positioned above the scene, pointing vertically down and in some cases several object contours were captured from different angles [33]. Most contemporary vision based approaches assume a simple situation where the scene consists of one object placed against a white background, such that the segmentation problem is minimal. Other approaches use range scanning sensors, [38–40]. This is an attractive choice, since they provide detailed geometrical model of an object. When a detailed 3D model is available, the grasp planning does not differ a lot from the case of grasping known objects.

Some recent work considers also the generation of grasping hypotheses based on local features rather than the object shape model [10]. The algorithm is trained via supervised learning, using synthetic images as training set. From two or more images in which grasping hypotheses are generated, the system performs approximate triangulation to derive 3D position of the grasping point. The work of [41] makes use of explicit information in terms of 2D position and orientation to learn feature combinations indicative for grasping. The tasks of computing such feature combinations can be linked to the concept of 'affordances' proposed by Gibson [42]: The occurrence of a certain feature combination potentially triggers a certain grasping action indicating the "graspability" of the object. A challenging task is to learn such object affordances in a cognitive system (see, e.g., [43]). Our work does not rely on object specific prior knowledge but it can generate the grasp hypotheses based on the current relationship between scene features. In particular, our system uses 3D features which can provide more optimal grasps in terms of approaching the object and orienting the hand accordingly.

Once a contact with the object is made, tactile information can be used to further optimise the grasp (see, e.g., [8,44,45]). In [44], a data–base that matches tactile information patterns to successful grasps is used to guide the grasping process. Self-Organizing Maps are used for the interpolation of grasp manifolds associated to shape primitives. In [8], so called 'Contact Relative Motions' (CRMs) triggered by tactile information are are used to translate the grasp synthesis problem into a control-problem with the aim of finding the shortest sequences of CRMs to achieve stable grasps. Our prior work presented in [45] shows how tactile feedback can be used for implementation of corrective movements and closed loop grasp adaptations.

Note that some initial work on our approach described here has previously been presented at a conference [15] where the system was tested only in simulation and thus did not deal with any real-world problems. In the work presented here, we have implemented the grasping system on an actual hardware consisting of a stereo vision system and a robot arm. As a consequence of the extensive evaluation done here, it was required to make a number of significant modifications compared to [15]. Moreover, we have introduced an adaptive component in our approach and discuss the work in the context of a concrete cognitive system.

2.1. *Contributions and relation to prior work*

As outlined in the previous section, grasping of unknown objects in unconstrained environments is a hard problem due to the small amount of prior knowledge that can be assumed. To create a system that solves this problem in a general way with high success rate, a number of strategies need to become combined and learning needs to be an integral part of such a system. Our algorithm provides a strategy based on 3D edges and other visual modalities and can be seen as being complementary to strategies based on 2D features or 3D descriptions extracted by range scanners. Here, we point out specific contributions of our work related to the existing grasping approaches.

D1 **Weak prior knowledge:** Our grasping strategy is based on a weak prior information of objects to be grasped: In particular, it is based on the existence of co-planar pairs of 3D edges. We will show that such basic cues can already lead to a large amount of successful grasps in complex scenes (see D3) and hence can be used in a bootstrapping process of a cognitive system in which stronger bias is developed by experience as described in the introduction and [12].

D2 **3D representation:** Our approach makes use of the fact that 3D information is independent from transformations in space. The prior knowledge we use generates a full 3D pose and hence we can also grasp objects that are tilted in any 3D orientation (see figure 13).

D3 **Error recovery:** Because of the weak prior, we can not expect our approach to work with a success rate close to 100%. We prefer to generate a certain percentage of successes on arbitrary objects rather than high quality grasps on a constrained set of objects. However, for this the system needs to be able to continue in case of unexpected events and non successful grasps (see figures 9, 10 and 13). By that we are able to work on rather complex scenes with multiple objects and no pre-segmentation with a reasonable success rate (see [46] for a movie).

D4 **Autonomous success evaluation:** We can confirm the success by means of haptic information. By that, we are able to build up an episodic memory (see figure 12) of evaluated grasping attempts, containing: 1) the grasping hypotheses, 2) the visual features that generated them, and 3) a success evaluation. These triplets are used as a ground truth for further learning based on neural networks to refine the pre-wired grasping strategy.

D5 **Realisation on different embodiments:** We show that the grasping behaviour can be adapted to different embodiments. More specifically, we applied it with a two-finger gripper as well as a humanoid with a five finger hand, (see section 6).

3. Visual representation

The grasping behaviour described in this work is based on the early cognitive vision system [47,48]. We use a calibrated stereo camera system to create sparse 2D and 3D features, namely multi-modal primitives (described in section 3.1), along image contours. In this system, we compute local information covering different visual modalities such as 2D/3D orientation, phase, colour, and local motion information. This local information is then used to create semi-global spatial entities that are called contours (described in section 3.2). In section 3.3 two perceptual relations, co-planarity and co-colourity are defined between primitives and between contours, and later used in calculation of grasping hypotheses. Note that primitives, contours and their perceptual relations are particularly important in the context of this work, since the grasping hypotheses defined in section 4 are based on them.

3.1. *Multi-modal primitives*

2D primitives represent a small image patch in terms of position $\mathbf{x}$, orientation θ, phase ϕ and three colour values $(\mathbf{c_l}, \mathbf{c_m}, \mathbf{c_r})$ describing the colour on the left and right side of the edge as well as on a middle strip in case a line structure is present. They are denoted as $\pi = (\mathbf{x}, \theta, \phi, (\mathbf{c_l}, \mathbf{c_m}, \mathbf{c_r}))$. Pairs of corresponding 2D features across two stereo views afford the reconstruction of a 3D primitive encoded by the vector

$$\Pi = (\mathbf{X}, \Theta, \Phi, (\mathbf{C_l}, \mathbf{C_m}, \mathbf{C_r}))$$

in terms of a 3D position $\mathbf{X}$ and a 3D orientation Θ as well as phase and colour information generalised across the corresponding 2D primitives in the left and right image (for details, see [48]).

Figure 2 illustrates what kind of information exists on different levels of the feature extraction. The process starts with a pair of stereo images (figure 2 (a)). Then the filter responses (figure 2 (b)) are calculated which give rise to the multi-modal 2D primitives and contours (figure 2 (c)). After finding corresponding 2D feature pairs across two stereo views, the 2D information is used to create 3D primitives and 3D contours (figure 2 (d)).

3.2. *Contours*

Collinear and similar primitives are linked together by using the perceptual organisation scheme described in [49] to form structures denoted as *contours*. Since the linking is done according to geometrical and visual good continuation, contours represent parts of a scene as geometrically and visually smooth curves. As their building blocks, contours are also multi-modal entities containing visual modalities such as mean colour and phase. Therefore, they do not only contain geometrical but also appearance based information. In figure 4, 3D contours of an example scene are presented.

3.3. *Relations between primitives and contours*

The sparse and symbolic nature of the multi-modal features gives rise to perceptual relations defined on them that express spatial relations in 2D and 3D (e.g., co-planarity, co-colourity). The co-planarity relation (see figure 5b) between two spatial 3D primitives Π_i and Π_j is defined as:

$$cop\left(\Pi_i, \Pi_j\right) = \frac{\Theta_j \times \mathbf{V}_{ij}}{|\Theta_j \times \mathbf{V}_{ij}|} \bullet \frac{\Theta_i \times \mathbf{V}_{ij}}{|\Theta_i \times \mathbf{V}_{ij}|}$$

where $\mathbf{V}_{ij}$ is the vector connecting the two primitives positions.

Two 3D primitives are defined to be co-colour if their parts that face each other have the similar colour. Note that the co-colourity of two 3D primitives is computed using their 2D projections. We define the co-colourity (see figure 5 (a)) of two 2D primitives π_i and π_j as:

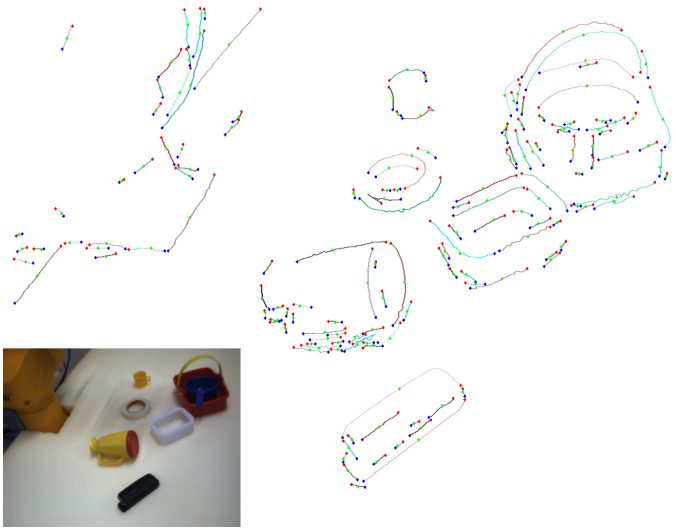

Fig. 4. 3D contours extracted from the scene that is shown in the bottom left (left image). Red dots indicate the first primitive in a contour, green the middle, and blue the last primitive in the contour.

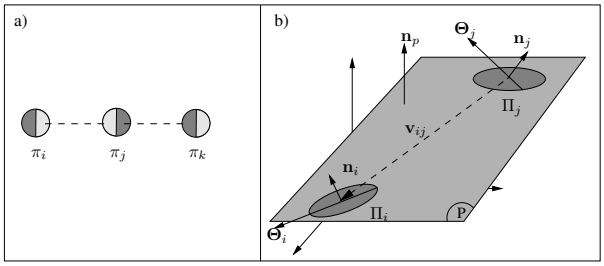

Fig. 5. Illustration of the perceptual relations between primitives. **a)** Co-colourity of three 2D primitives π_i, π_j and π_k. In this example, π_i and π_j are co-colour, so are π_j and π_k; however, π_i and π_k are not co-colour. **b)** Co-planarity of two 3D primitives Π_i and Π_j. $\mathbf{n}_i$ and $\mathbf{n}_j$ are normals of the planes that are defined as cross products of individual primitives orientations Θ_i, Θ_j and the orientation of the connecting line $\mathbf{V}_{ij}$, (see sec. 4.1). $\mathbf{n}_p$ is the normal of a common plane defined by combining the two normals $\mathbf{n}_i$ and $\mathbf{n}_j$.

$$coc(\pi_i, \pi_j) = 1 - \mathbf{d_c}(\mathbf{c}_i, \mathbf{c}_j),$$

where $\mathbf{c}_i$ and $\mathbf{c}_j$ are the RGB representation of the colours of the parts of the primitives π_i and π_j that face each other; and $\mathbf{d_c}(\mathbf{c}_i, \mathbf{c}_j)$ is the Euclidean distance between RGB values of the colours $\mathbf{c}_i$ and $\mathbf{c}_j$. Note that $\mathbf{d_c}(\mathbf{c}_i, \mathbf{c}_j)$ returns a value between 0 and 1 since the components of the RGB colour space is represented within the interval 0 and 1.

Since contours represent larger portions of scenes than local features, contours and their relations can give a more global overview of the scene. The contour relations used in this work are straightforward extensions of primitive relations. The definitions of co-colourity and co-planarity between contours are supported by an algorithm for associating primitives between contours (see [50] for details).

4. Grasping Strategy

The grasping behaviour proposed in this work is a low-level procedure that allows for the robot manipulator to grasp unknown objects. As explained in section 3, the early cognitive vision system extracts multi-modal visual feature descriptors from stereo images. Multi-modal relations between primitives support perceptual grouping into contours. Second order relations, co-planarity and co-colourity, between contours indicate possible co-planar edges originating from the same object, or even the same surface in a scene. The grasping behaviour is based on four basic grasping actions that can be performed on a pair of such contours using a parallel gripper.

In the early cognitive vision system, edges are represented as 3D contours. As described in section 3, 3D contours are sets of linked 3D primitives. Pairs of contours that are both co-planar and co-colour are called "similar contours". In the middle of each of the two contours in a such pair, one representative 3D primitive is chosen. These primitives are called 'parent primitives' and contain the information about respective contour's position and orientation. Figure 1 shows the four types of elementary grasping actions (EGAs) defined by two parent 3D primitives. Parent primitives are chosen in the middle of contours in order to aim at some level of stability when grasping, although one can not reason about it in a rigorous manner as the system does not perform segmentation and does not have any object model at this stage, see figure 4.

It is important to notice that in a real scene only some of the four suggested grasps are meaningful. For example, if an object in the scene is not concave, only grasps of type EGA 1 can be successfully performed. Since the information provided by the initial image representation is not sufficient to determine which of the grasping actions are suitable, the system suggests grasps of all four EGA types. Suggested grasping actions are therefore called *grasping hypotheses*. The term is also appropriate since grasping actions can fail because of other factors (such as uncertainties in the position and the orientation of the gripper that come from the uncertainty of the visual reconstruction, from limitations of the manipulator, or from an unforeseen collision with the environment) even if the intended action was reasonable.

4.1. *Elementary Grasping Actions (EGAs)*

Two parent primitives Π_i, Π_j produce a set of parameters used for defining the four EGAs. The parameters (see figure 5) are given as follows:

- position and orientation of the common plane p defined by co-planar parent primitives. It is denoted by position $\mathbf{P}_p$ of the point in the common plane half way between $\mathbf{X}_i$ and $\mathbf{X}_j$ and orientation $\mathbf{n}_p$ of the plane normal
- distance between parent primitives: $d_p = \| \mathbf{V}_{ij} \|$ (figure 5)
- direction connecting the parent primitives: $\mathbf{D} = \frac{\mathbf{V}_{ij}}{d_p}$

– individual primitives orientations Θ_i and Θ_j

This section starts with the definition of the common plane p and then proceeds to show how specific EGA types are constructed.

The common plane p is represented by $\mathbf{P}_p$ and $\mathbf{n}_p$ which are calculated as:

$$\mathbf{n}_p = \pm \frac{\Theta_i \times \mathbf{D} + \Theta_j \times \mathbf{D}}{\| \Theta_i \times \mathbf{D} + \Theta_j \times \mathbf{D} \|} \qquad (1)$$
$$\mathbf{P}_p = \frac{\mathbf{X}_i + \mathbf{X}_j}{2}$$

where $\mathbf{X}_i$ is the position of the i_{th} 3D primitive in the scene. Note that we assure that $(\Theta_i \times \mathbf{D}) \cdot (\Theta_j \times \mathbf{D}) > 0$ by choosing the direction of the Θ_j, so that vectors $\Theta_i \times \mathbf{D}$ and $\Theta_j \times \mathbf{D}$ point into similar directions.

The plus-minus sign on the right hand side of the equation above indicates that the direction of the normal of the averaged plane is also arbitrary. It is important to know which direction of the plane normal to use in order to predict meaningful grasps. The initial scene representation does not provide this information. Nevertheless, it is intuitively clear to the human viewer why the top side of the box in figure 1 (EGA 1) should be grasped from above. This observation can be expressed mathematically. The normal of the visible side of a surface always forms an obtuse angle to the vector originating from the point of view and pointing to the surface (figure 6). When this observation is

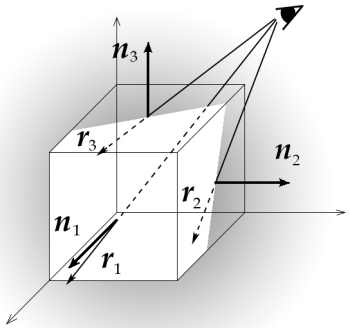

Fig. 6. Choosing the correct surface normal. $\mathbf{n}_1$, $\mathbf{n}_2$, and $\mathbf{n}_3$ are outward surface normals marking the sides of the cube visible on the illustration. The two sides visible from the marked point of view have surface normals $\mathbf{n}_2$, and $\mathbf{n}_3$. $\mathbf{r}_1$, $\mathbf{r}_2$ and $\mathbf{r}_3$ are camera rays, vectors originating from the marked point of view and pointing to the surface normals.

turned around, it follows that visible surfaces should adopt the direction of the normal that forms an obtuse angle to the camera ray in order to give expectable grasps. Another aspect of this observation concerns camera placement. Visible features of an objects should be the ones reachable by the manipulator. This kind of reasoning is applicable for EGA 1, EGA 2 and EGA 3 cases, while EGA 4 type of grasp does not depend on the direction of the plane normal but still requires that only one direction is adopted as the opposite direction would only duplicate already existing hypotheses.

Using the argumentation above, we adopt a heuristics where only one direction of normal is used for generating EGAs. The advantages are that the number of produced hypotheses is dramatically reduced (number of EGA 1, 2 and 3 grasps is halved), and in the majority of cases, the wrong hypotheses are excluded.

4.1.0.1. *Mathematical formulation of EGAs* A grasp is defined by the position and the orientation of its tool reference frame (Tool Centre Point (TCP) reference frame) in relation to, for example, the Robot's Base reference frame (figure 7), and the initial distance d between gripper fingers.

If the origin and the orientation of the TCP reference frame are defined as in figure 7 such that $\mathbf{Z}_{TCP}$ (Z axis of TCP frame) is parallel to the gripper's fingers, $\mathbf{X}_{TCP}$ axis connects the fingers, and $\mathbf{Y}_{TCP} = \mathbf{Z}_{TCP} \times \mathbf{X}_{TCP}$, and the origin is placed between two fingers, on some negative $\mathbf{Z}_{TCP}$ distance (depth of the grasp) from fingertips, then elementary grasping actions are given with expressions as follows.

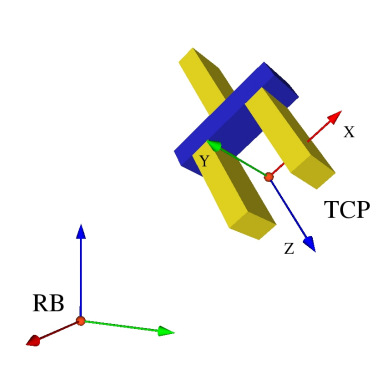

Fig. 7. The figure shows the Tool Centre Point (TCP) reference frame. It is given in respect to the robot's base (RB) frame. The position and orientation of the TCP reference frame is used when defining elementary grasping actions.

EGA 1:

$$\mathbf{P}_{TCP} = \mathbf{P}_p$$
$$\mathbf{Z}_{TCP} = -\mathbf{n}_p \qquad (2)$$
$$\mathbf{X}_{TCP} = \mathbf{D}$$
$$d_p < d \leq d_{max}$$

Initial finger distance d should be bigger than the distance between parent primitives d_p, so that grasping position can be approached without colliding with the object. It is limited by the maximum fingers opening distance d_{max}. The $\mathbf{X}_{TCP}$ can have the opposite direction as well ($-\mathbf{D}$) when using a parallel gripper, as the gripper has reflection symmetry across ZY plane.

EGA 2: is a grasp that is designed for concave objects, it has same the position and the orientation as EGA 1 but the initial finger distance is zero and fingers are opened in order to grasp an object (figure 1, EGA 2).

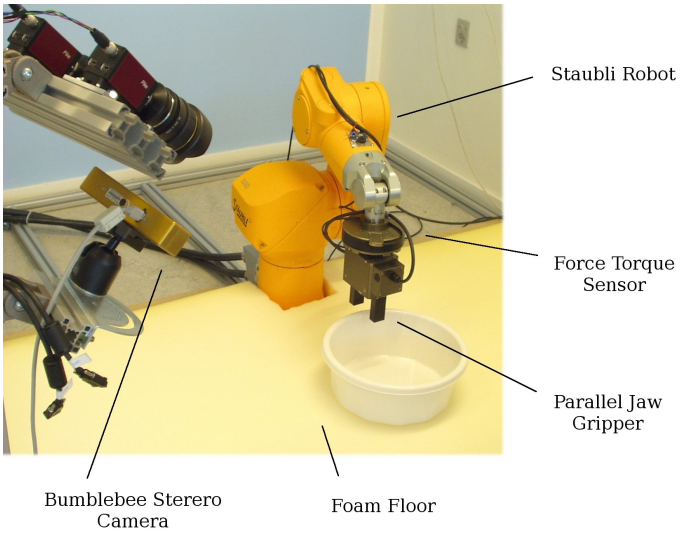

Fig. 8. Experimental setup.

Since the grasping tool is a simple parallel gripper, EGA 1 and 2 will be successful only when the parent primitives individual orientations are orthogonal to the line connecting them, meaning that the two parent primitives should be positioned opposite to each other:

$$|\Theta_i \cdot \mathbf{D}| < C \quad \wedge \quad |\Theta_j \cdot \mathbf{D}| < C \tag{3}$$

where C is a positive real number smaller than one. If this is not the case, the grasp is unstable or not possible.

Both *EGA 3* and *4* give two grasping actions, one for every parent contour. EGA 3 and 4 use the individual orientations of the parent primitives (projected to the common plane) as $\mathbf{Y}_{TCP}$ direction and do not rely on the orientation of the connecting line. This is why orthogonality to the connecting line is not a requirement. The calculations are analogue to the case of EGA 1 (equation 3).

5. Experiments

This section gives a description of the experimental setup (section 5.1) and explains the testing procedure (section 5.2). Qualitative and quantitative results are given in section 5.3 and then become discussed in section 5.4.

5.1. *Experimental setup*

This section gives a description of the hardware (section 5.1.1) and software elements (section 5.1.2) used in the experimental setup.

5.1.1. *Hardware*

The hardware setup consists of a Staubli RX60 six degrees of freedom industrial robot arm, a fixed Bumblebee2 colour stereo camera, a FTACL 50-80 Schunk Force Torque sensor and a PowerCube 2-finger-parallel gripper tool mounted on the Force Torque sensor, (figure 8). The floor is covered with flexible foam layer. The stereo camera has a fixed position with respect to the robot. A common frame of reference is derived through a robot-camera calibration procedure.

The force torque sensor is used for active collision detection. The sensor is mounted between the wrist and the tool of the robot, and it measures forces or torques acting on the tool. By comparing forces and torques that can be expected from the influence of the gravitational force alone with those that are actually measured by the sensor, it is possible to detect any external collision or force that acts on the tool, (see figure 9).

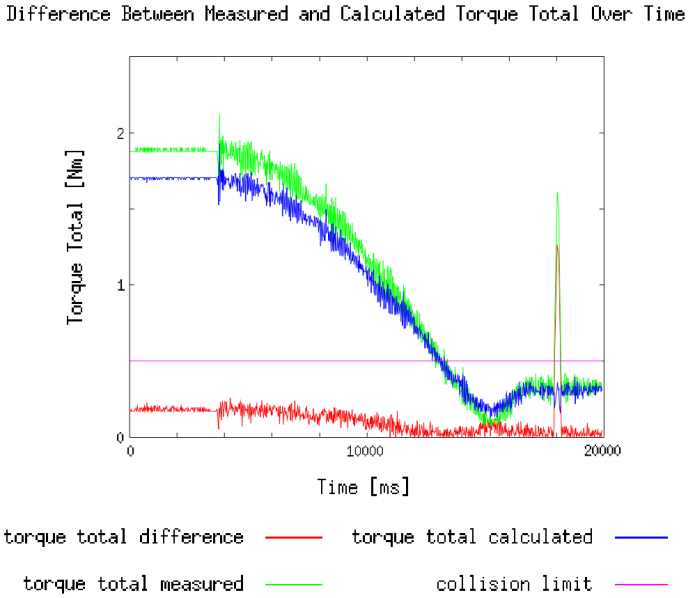

Fig. 9. The graph shows the total measured and the total calculated torques, and the difference between measured and calculated values as a function of time for a sample grasping attempt where a collision happened. Figure 13 shows an example collision situation and the corresponding grasping hypotheses.

The control application for executing the grasping attempts is run on a PC machine under Linux operating system. The system uses a Modbus interface to communicate to the Staubli robot and RS232 serial communication to communicate to the gripper and the force torque sensor. A firewire interface connects the camera to a Windows PC machine that exchange information with the control application through a TCP/IP connection.

5.1.2. *Software*

The implementation is based on three distinct software environments CoViS, RobWork [51] and Orocos [52]. CoViS is a cognitive vision system that is modelling early cognitive functions of biological visual systems, (section 3). It is being developed by the Cognitive Vision Group at University of Southern Denmark. RobWork is a framework for simulation and control of robotics with emphasis on industrial robotics and their applications. Orocos Real-Time Toolkit (RTT) is a C++ framework for implementation of

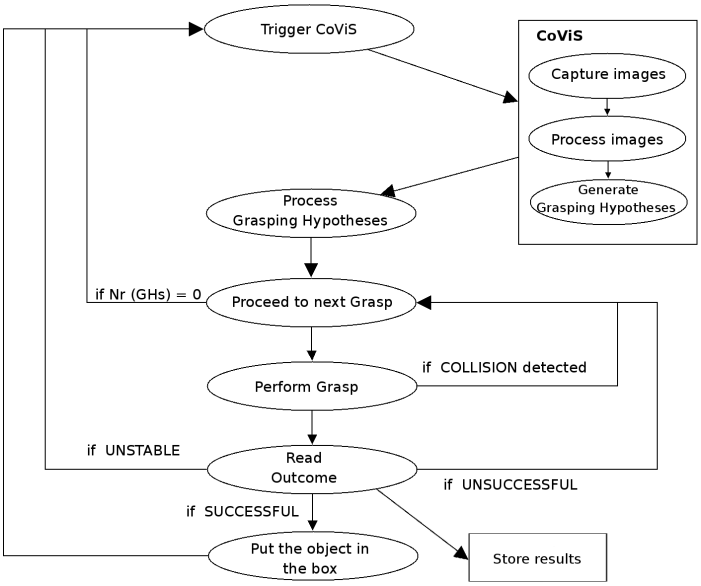

Fig. 10. State diagram showing work flow of exploration using the grasping behaviour.

(realtime and non-realtime) control systems. RobWork and Orocos are integrated into a single control application that communicates to the CoViS application using a TCP/IP connection.

5.2. *Testing*

Figure 10 shows the state diagram for the grasping procedure. The procedure starts by capturing and processing images,and producing grasping hypotheses (GHs). GHs are then processed, certain number of grasping actions are tested and results are stored.

The system tries to find a maximum of five grasps that are reachable by the robot and can be accessed with a collision free movement of the robot. The search is stopped when five feasible grasps are found or when there are no more grasping hypotheses available. The selected grasps are scheduled for execution. Collision free trajectories are calculated using the RRT-connect (Rapidly-exploring Random Trees) motion planner [53] with PQP (Proximity Query Package) collision detection strategy [54]. Figure 11d. shows the simulation environment used for motion planning. It includes the 3D models of the robot (kinematic and geometric) and the floor, and for each new scene it imports the reconstructed contours of the objects present in a region of interest in front of the robot. Since models of objects in the scene are not complete, the calculated path is collision free "to the best of knowledge". This does not present a problem since the system is able to automatically detect collisions and recover from them.

The number of grasping hypotheses generated by CoViS can vary from several to several thousand for each scene, depending on the scene's complexity and the quality of the reconstruction. As only few grasping hypotheses can be tested, (the scene will eventually become affected by robot actions), it is necessary to adopt a criteria for selecting those.

In this work, grasping hypotheses are ranked by the amount of the verticality of the grasp, or more precisely:

$$R_S = Z_{TCP} \cdot (-Z_W)$$

where the ranking score R_S is in the interval [-1,1]. Z_{TCP} is the orientation of the Z axis of the TCP frame (see figure 7) expressed in the World reference frame, and $(-Z_W)$ is the vector pointing vertically down. Hence, grasps where the gripper fingers are pointing down vertically have the highest rank. This heuristics is motivated by the observation that most objects are, due to the gravity, accessible from the top when placed freely in a scene. Thus the computationally expensive motion planning algorithm is less likely going to report a collision when planning for vertical grasps. Having in mind the great number of generated grasping hypotheses, the ranking function has been introduced to guide the search for feasible ones. Also, when approaching an object from the top, it is less likely that the object will slide away. However, the non-vertical grasps are performed as well (see figure 13). For eliminating such simple heuristics by learning see section 7.

Execution of a selected grasp can result in *successful*, *unstable* or *unsuccessful* grasping attempt or can report a *collision* in which case the robot stops and returns to the initial position. This evaluation is done autonomously by measuring the distance between the fingers. More precisely, we say that a grasp is

- *unsuccessful* if the distance between the fingers after closing/opening is 0/maximal,
- *unstable* if the distance is larger than 0 or smaller than maximum during the closing/opening of the finger but 0/maximal after having picked up the object,
- *successful* if the distance is larger than 0 or smaller than maximum during the closing/opening of the finger as well as after picking up the object.

Moreover, *collisions* are detected by the force torque sensor. Since there is no prior information about objects in a scene, the gripper is using a constant, predefined force for all grasping attempts.

5.3. *Results*

The experimental evaluation presented in this section is designed as an exploratory case analysis. The aim is to illustrate different aspects of the system's behaviour, its capabilities and weaknesses. Two types of experiments were performed.

In the first group of experiments (described in section 5.3.1), a test scene contains a single object. The robot attempts to remove it from the scene by using the grasping behaviour. Fourteen objects have been used in the evaluation (figure 12). The size and the shape of the objects are chosen so that grasping is connected to different degrees of difficulty.

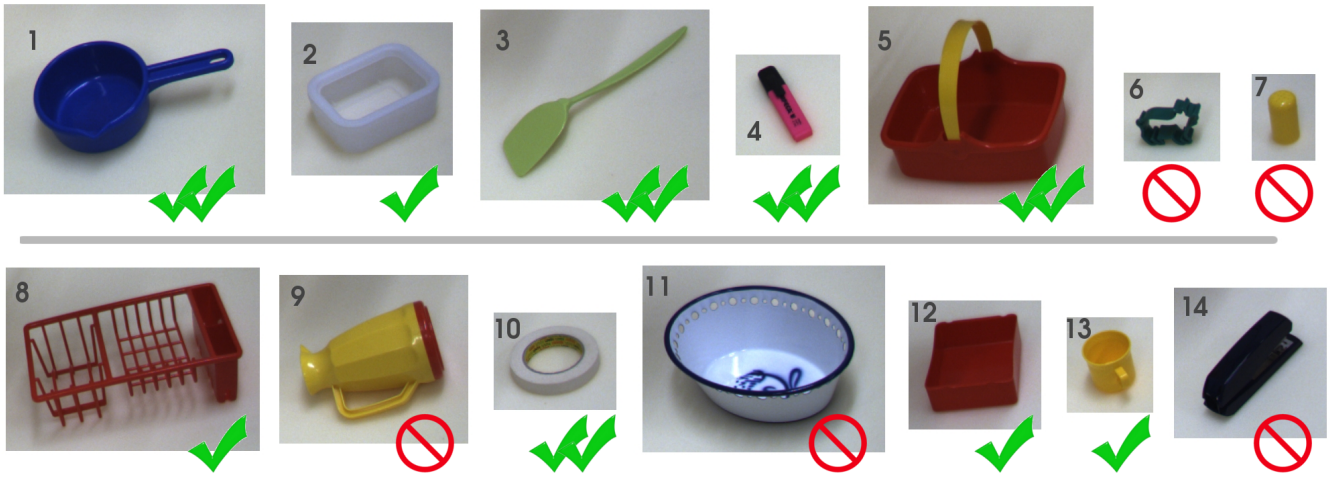

Fig. 12. Office and toy kitchen objects used in the experiments. Objects are of mostly uniform colours, and their size and the shape is suitable for grasping with the parallel jaw gripper. The marks illustrate average success rate in grasping individual objects measured in the experiments (see also table 3).

The second group of experiments (described in section 5.3.2) were performed on five complex scenes [2] containing a selection of the very same objects investigated in section 5.3.1 which are however distributed randomly with high degree of clutter and occlusion (see figure 16). The goal was to remove as many objects as possible from the scene. A short video showing an experimental setting similar to complex scenes described here is available at [46] (snapshots of the video are shown in figure 13).

5.3.1. *Single objects*

Each of the fourteen objects has been presented to the system in several different positions and orientations that vary in terms of grasping difficulty. Experiments performed with the first object are described in detail. Results on other experiments are given briefly.

Object 1

Three experiments were performed with object 1 (see figure 14). In the first experiment, the object was successfully grasped in the first attempt with the grasp of EGA 3 type. The same happened in the second experiment and the successful grasping hypotheses are shown in figure 11c. In the third experiment, the object was not grasped because it turned out to be unreachable by the robot. The object was also placed further away from the camera system than in the first two experiments, which gave a lower quality reconstruction and thus a fewer number of grasping hypotheses.

As mentioned in section 5.2, the ranked list of grasping hypotheses (GHs) is processed top-down. The processing stops when a certain number (five here) of accessible GHs have been found, or when there are no more GHs available (see section 4.1). In order to give an illustration of a typical processing outcome, full sets of GHs have been processed for the first two experimental situations, and the results are shown in table 1. The order of the conditions that a grasping hypothesis has to fulfil in order to be accepted is identical to the order in the table, (e.g., GHs are first checked for reachability, then it is checked whether the position of the tool during grasping is collision free and if both of those conditions are satisfied the system will try to calculate a collision free trajectory).

Experimental situation	1	2
number of grasping hypotheses (GH)	66	373
number of accepted GHs	11	37
number of unreachable GHs	46	243
number of GHs where tool is in collision	9	93
number of GHs where collision free path was not found	0	0

Table 1
The results of processing full sets of GHs for the first two experimental situations (see figure 14). Finding a collision free path is an easy task due to the fact that the scene is not complex.

As can be seen from table 1, only a small percentage of the computed grasping hypotheses become actually performed. Most computed grasping hypotheses can be disregarded by constraints that can be computed beforehand.

Table 2 shows some intermediate values from the grasping hypotheses generation program for the first two scenes of figure 14. The number of contours, contour pairs and the number of similar contour pairs are derived from the whole image representation. Parent primitive pairs are then assigned to the pairs of similar contours. A parent pair is discarded if any of the two primitives does not belong to a certain region of interest in front of the robot. Background features that originate from the robot and the edge of the ground surface (figure 14) generate a lot of undesirable similar contours and that is why the number of discarded par-

[2] We show results on three of these scenes. Results on the other two scenes are described in [55]

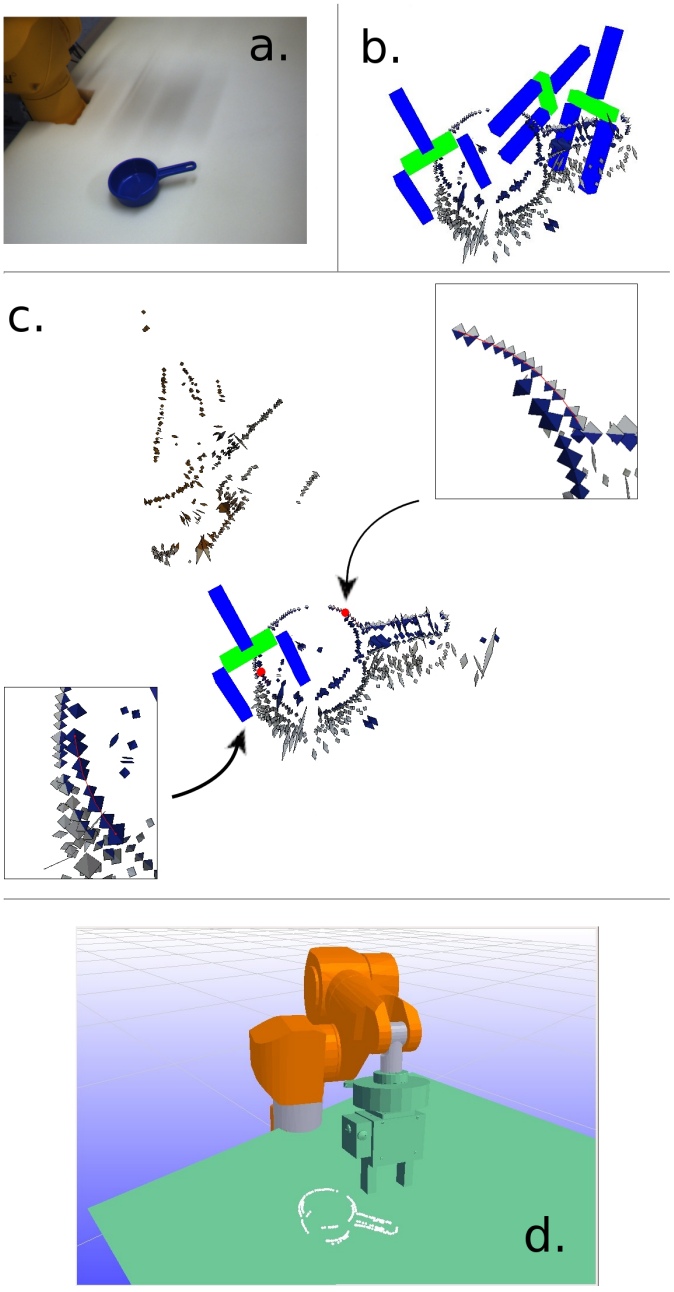

Fig. 11. a) The image taken during one of the experiments (section 5.3.1) captured by the left camera. b) Some grasping hypotheses generated for that scene, displayed in a visualisation environment. c) A successful grasping hypotheses (EGA 3) where parent contours are magnified. The primitives in the top left corner come from the robot and the background. d) RobWork simulation environment shows 3D models of Staubli robot and floor. Additionally, the information about 3D edges in the scene is provided by the vision system. The 3D contours are composed of 3D primitives, which are modeled as small cubes. The models are used for planning collision free motions of the robot and for the visualisation purposes. It is important to notice that extracted contours do not contain as many outliers, compared to full reconstruction in b.

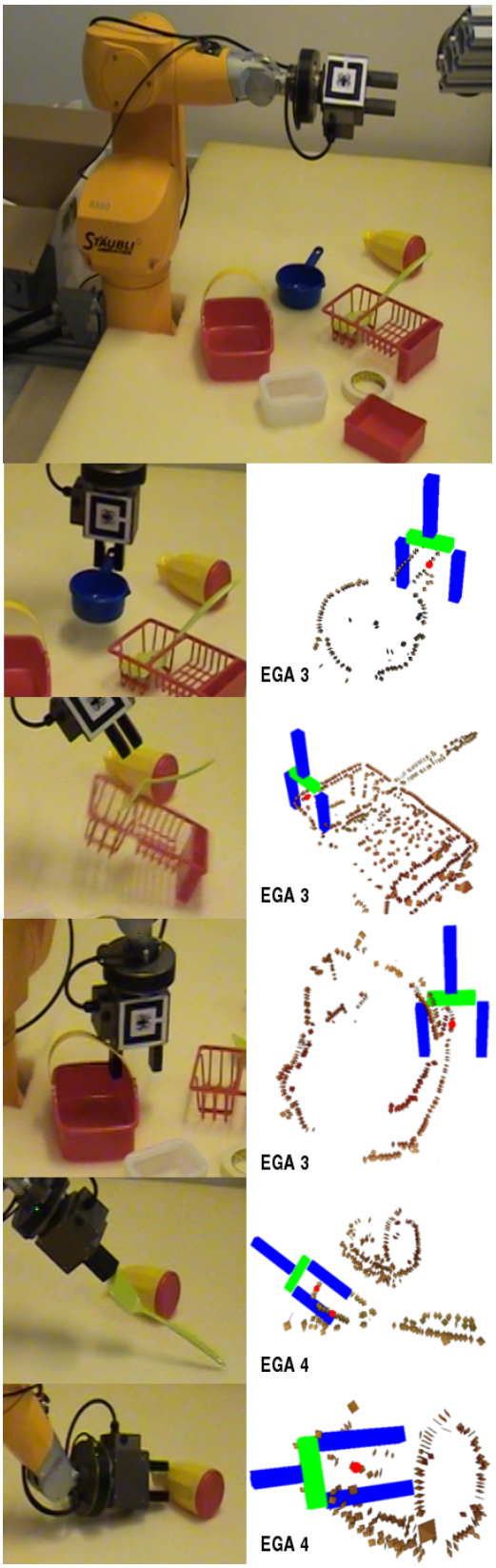

Fig. 13. Five grasping outcomes from the video available at [46]. From top to bottom: Successful, Unstable, Collision, Successful and Fail cases. A snapshot from the video (left) and the corresponding grasping hypotheses viewed in a visualisation environment (right) is shown for each case.

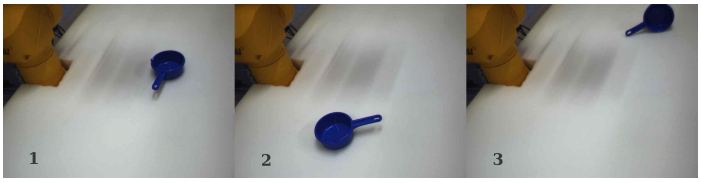

Fig. 14. Three experimental situations for the object 1. Figure shows the original images used for acquiring image representations, captured by the left camera. The darker areas at the middle of all three images are shadow cast by the robot when in the initial position.

Experimental situation	1	2
number of contours	27	30
number of all contours pairs	351	435
number of similar contours pairs	201	241
number of accepted parents pairs	17	94
number of discarded parents	184	147
number of GHs	66	373

Table 2
Intermediate values from grasping hypotheses generation program.

ent pairs is high. This however does not explain why there is a significant difference between number of good parent pairs and consequently generated grasping hypotheses in the two cases. The difference arises because the representation of object 1 contains less detail in the first case, as it is further away from the camera.

As mentioned in the introduction of this section, it is important to notice that individual experimental situations were designed to demonstrate different aspects of the system's performance and are not suitable for direct statistical analysis. However, we still present a weak numerical comparison of the experimental results on different objects. Figure 15 shows experimental situations for the 14 objects. Table 3 gives the corresponding distribution of different grasping outcomes.

One of the factors that influences the outcome of a grasping attempt is the placement of the object with respect to the camera since reconstructed primitives have uncertainties that vary with the distance from the image centre and with the distance from the camera. Small objects that are placed too far away also do not have a good enough reconstruction for triggering grasps. Object 11 turned out to be too heavy to be lifted from the ground. Objects 3, 4, and 10 have edges that are positioned very close to the floor so that small errors in the vertical direction can cause collisions with the floor. In some cases (object 12 - situation 3, object 5 - situation 3, object 2 - situation 2) the object's opening was not available for vertical (top down grasps) which are ranked highest, so that potentially successful grasps with non-vertical orientations where not chosen. In few cases shadows triggered grasping attempts.

5.3.2. *Multiple objects*

In the second evaluation stage, grasping hypotheses were tested on three complex scenes. For each scene, the robot

object nr.	1	2	3	4	5	6	7	8	9	10	11	12	13	14
successful grasps(%)	67	30	50	50	50	0	0	25	0	50	0	33	17	6
unstable grasps (%)	0	0	3	0	0	0	0	18	0	0	100	33	0	11
collisions (%)	0	42.5	16	0	39	33	17	41	77	11	0	0	29	83
unsuccessful grasps (%)	33	27.5	31	0	11	0	33	16	23	39	0	33	29	0
no grasps (%)	0	0	0	50	0	66	50	0	0	0	0	0	25	0

Table 3
The results of experiments with single objects.

random scene	1	2	3
number of grasping attempts	30	30	30
successful grasps	6	4	5
unstable grasps	5	2	3
collisions	18	12	16
unsuccessful grasps	1	12	6

Table 4
The results of experiments with complex scenes 1, 2 and 3.

performed 30 grasping actions in order to remove as many objects as possible from a scene.

Figure 16 show three complex scenes. Photos on the left show initial situations and photos on the right show the same scenes after performing 30 grasping attempts. As can be seen by comparing the changes, even in these complex scenes with many objects, strong occlusions and clutter, a good number of grasping attempts have been successfully performed.

Table 4 gives results of the experiments for complex scenes. The relative success of the grasping behaviour depends on the number of the attempts taken into account. The 30 grasping attempts were usually enough for the system to perform all possible successful grasps. We experienced that in case the system continues working after this point, the number of the unsuccessful, collision and unstable outcomes grows. It happens because the remaining objects are not in the suitable position, (unreachable or graspable edges are not visible), or due to the ranking criteria, some nonsuccessful grasps repeatedly become favoured so that other possibilities are not explored. The initial ranking criteria is eliminated in the learning stage, see section 7.

In a complex scene, grasping hypotheses can be defined with edges from two different objects. The use of the co-colourity relation (i.e., two primitives sides facing each other have the same colour, see section 3.3) make it likely that the parent primitives are from the same object. However, the outer colour of edges of the two objects is usually the colour of the floor surface and if the two edges are co-planar at the same time, a grasping hypotheses will be created. In most cases, this is not a disadvantage as GHs originating from different objects often give good results.

Fig. 15. Experimental situations for objects 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13 and 14. Photos captured by the left camera.

5.4. *Discussion of Experiments*

The grasping experiments performed on single objects as well as those performed on cluttered scenes with many objects showed that there is a consistency in graspability of specific objects. In other words, some objects are grasped easily and consistently whenever they are in suitable position and image processing produces a good representation. Other objects are grasped just occasionally. This depends on how well individual object's features (weight, size, shape, colour, material) pair with the type of gripper used in the experiments. On the other hand, it depends on how suitable the object's features are for the kind of image processing used, i.e. how difficult it is to extract good co-colour and co-planar contours. For small or distant objects, the reconstruction was often poor. In these cases, images with higher resolution or making use of a visual attention mechanism could improve the performance.

The gripper used in this setup limits grasps of EGA 1 and EGA 2 types only to small objects. Large objects are mostly grasped by the edges with grasps of type EGA 3, if they are concave. Although object 9 could be grasped by the handle, this did not happen because the algorithm does not identify the handle as a specifically good grasping position. Here object dependent grasping knowledge (see section 1) acquired by supervised learning (e.g, by imitation learning (see, e.g., [28])) might become an important option for improvement.

Table 5 gives the distribution of EGA grasp types for the successfully performed grasps in the experiments with sin-

Fig. 16. Complex scenes 1, 2 and 3. Images on the left show initial scenes. Images on the right show the corresponding scenes after 30 grasping attempts with our grasping behaviour.

	percentage
EGA 1	10%
EGA 2	5%
EGA 3	50%
EGA 4	35%

Table 5
Distribution of EGA types for successful grasps in single objects experiments.

gle objects. EGA types 3 and 4 are represented more often because EGA types 1 and 2 have two additional constrains, (the distance between grasped edges has to be within the opening range of the gripper, and the edges have to be mirror symmetric).

The current system has an open loop - "look-and-move" type of control. The drawback of this is the high sensitivity to calibration errors. The accuracy of grasping operation depends directly on the accuracy of the visual sensor, the robot end-effector, and the robot-camera calibration. This could be avoided by visual servoing. However, in our calibrated set up this was not necessarily required.

The exploration procedure could be additionally enhanced by making use of tactile sensors and based on that, reactive grasping strategies (see, e.g., [44]). Our simple grasping strategy could then serve as an initial "approach" planner. This would potentially reduce the number of unstable grasps and would also give rich feedback for learning (see also section 7). On the other side, underactuated (coupled joints) robotic hands with adaptability and/or robotic hands with compliance [56,57] do not require complicated sensing and control, and are robust in situations where the object shape is not fully known. The following section 6 discusses mapping EGAs to more complex grippers.

6. Realisation on the humanoid robot ARMAR-III

The grasping method described in this paper has been developed and implemented for a simple parallel griper. Despite the use of this unsophisticated grasping tool, and the small amount of prior knowledge, the method has performed well. The initial choice of gripper has proven to be good for the proof of the concept, providing results without the need for overhead complexity. However, the method can be extended to more complex hands. This section discusses mapping EGAs to multi fingered hands and describes the implementation on the humanoid robot ARMAR-III [58].

The concept of "virtual fingers" [59] presents a good basis for extending the grasping strategy to more complex hands. Instead of separately analysing degrees of freedom of individual fingers, grasps can be described in terms of "virtual fingers" - abstract functional units that represent one or more fingers, or hand surfaces, acting together in applying an opposition force [24]. For example, both long precision grasp, where the thumb is opposed to one or more other fingers, and lateral pinch grasp, where opposition lies between the thumb and the side of the index finger can be characterized as grasps where only two virtual fingers participate [60]. Having in mind the idea behind the EGAs — to grasp objects using grasps that are proposed by the occurrence of certain combinations of visual features — it is clear that execution of EGAs can be carried out by any hand capable of performing oppositional forces. The advantage of the multi fingered hand is in that it can accommodate different sizes of objects, having a "repertoire" of possible oppositions in different scales — from opposing two fingertips, to opposing several fingers to the palm of the hand. Another advantage is that more then one finger can act in a unison, stabilizing the grasp. The current definition of EGAs does not extend to grasps where more then two virtual fingers are involved.

A mapping to a concrete multi fingered hand should take into account the lack of grasping symmetry that is present in a parallel gripper. A decision has to be made on how to assign virtual fingers. This can be dealt with in more then one way. For example, if an EGA 3 or EGA 4 pinch grasp is achieved by opposing the thumb to remaining fingers, it is better to let the thumb aim for the assumed opening in the object, as the opening might be too narrow to accommodate many fingers. When trying to perform an EGA 1 encompassing grasp, the system could choose to place the thumb on the shorter of the two parent contours, thus leaving the longer one to receive as many other fingers as possible. Further similar considerations can be made in rela-

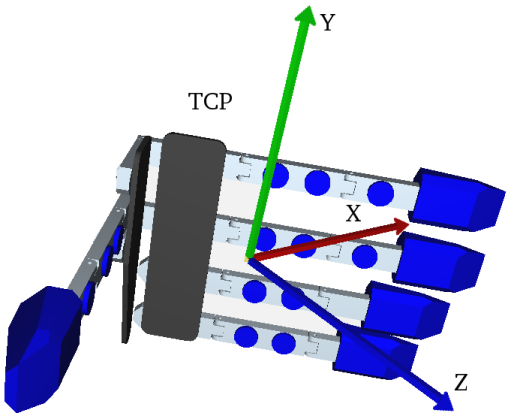

Fig. 17. A model of the five fingered anthropomorphic hand of the humanoid robot ARMAR-III. The shown TCP reference frame allows for a direct mapping to the EGAs produced for the two finger gripper (see figure 7). Please note that only one of the two possible orientations for a single EGA is shown (the not shown one is achieved by rotating the hand 180° around the z axis).

tion to properties of the parent contours, or of the specific robotic hand used.

For the experiments with the humanoid robot ARMAR-III a five fingered anthropomorphic hand [61] was used. The concrete version of the hand used has six independent degrees of freedom and eleven joints, the fingers are pneumatically actuated and no position control is used. For our purposes we consider the thumb as one and the combination of index and middle finger as the other virtual finger. No considerations were made about how to assign virtual fingers in regard to visual features, all grasping hypotheses were given in both possible orientations. The thumb is hereby located between index and middle finger. Figure 17 shows the hand with the Tool Centre Point (TCP) reference coordinate system which is selected similar to the one for the two finger gripper (see figure 7). This definition allows for an easy transfer of EGAs to this new hand. Grasps of type EGA 2 were not used on the ARMAR-III. Figure 18 shows the robot performing two grasping actions.

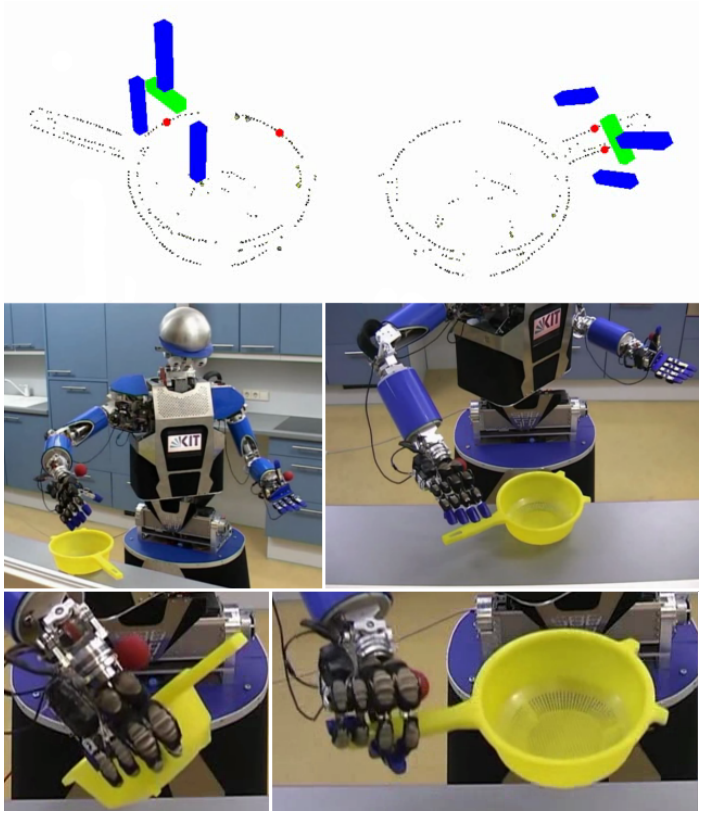

Fig. 18. Two executed grasps on ARMAR-III taken from the video [62]. From top to bottom: Predicted grasps, snapshots of the robot performing approaching movements, final configuration. Predicted grasps are shown from the point of view of the robot.

7. Refinement of Initial Grasping Behaviour

The grasping behaviour introduced in this paper is an important part of the cognitive system developed with the project PACO-PLUS [63]. Of particular importance is that the success of the action can be evaluated autonomously by the system. In our case, haptic information from the gripper can be used to distinguish between successful, unstable and unsuccessful grasps as described in section 5.2. Hence, some kind of an episodic memory (see, e.g., [64]) can be build up autonomously that can then be used for further refinement of the grasping behaviour. In that context, we have defined a learning procedure that allows for improving the grasping behaviour by making use of the grasping attempts and their evaluation stored in the episodic memory.

The exploration behaviour described in this paper performs multiple autonomously evaluated grasping attempts, stored in the episodic memory. This memory can be used as input for learning since it preserves information that gives indications about likelihoods of success or failure. For example, if the parent primitives are more distant than the maximal distance between fingers allowed by the gripper, EGA 1 and 2 are not executable any more. Another example is the uncertainty of the reconstruction of the primitives that depends on the distance of the object to the camera as well as the 'eccentricity' (i.e., the distance to the principal ray of the camera) of the parent primitives (for an exact analysis of the uncertainty see [65]). Hence, it is possible to learn the relation between these parameters and the success likelihood of a grasping attempt.

More specifically, we have used such parameters extracted from the evaluated grasping attempts as a basis for the learning algorithm. A grasp is associated with two parent primitives, Π_i and Π_j. The 3D positions of those are denoted X_i and X_j. The position of the grasp is denoted P_{TCP}. C_L and C_R denote the positions of the optical centre of the left and right camera. The following features, illustrated in figure 19, have been computed:

F1 The height, h. It is given by the z-component of P_{TCP}.
F2 The angle, φ_v, between the normal, n_p, (equation 2) and a vector in the world reference frame pointing vertically up
F3 The distance between parent primitives, d_p.

F4 The distance to the camera, d_{cam}.

F5 The angle, φ_{cam}, defined by a ray from the optical centre to TCP and vector pointing normally out of the image plane.

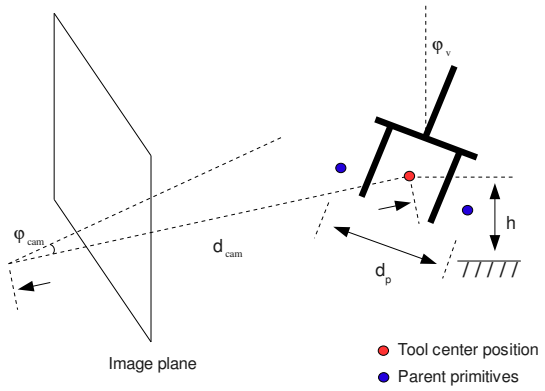

Fig. 19. Features used for learning.

F1–F3 refer to the robots ability to grasp the object independently of where the object is positioned. F4 and F5 are related to the relative position of camera and object which are the reason for large variation of the quality of the 3D reconstruction which will have an effect on the quality of the computed EGA. The grasping attempts used for learning are shown in figure 20 with respect to φ_{cam}, d_{cam} and the outcome of the grasps. It can be observed that the success of the grasp (indicated as a green cross) depends on F4 and F5. All features are independent of each other and can be computed using the 3D positions of the parent primitives and the camera calibration parameters. In addition, some EGAs might be more robust to reconstruction errors or wrong interpretations of data. Therefore the type of the grasp needs to be taken into consideration as well. The learning is implemented using a radial basis function network (for details of the structure of the network as well as for a detailed analysis of the features F1–F5 we refer to [66]).

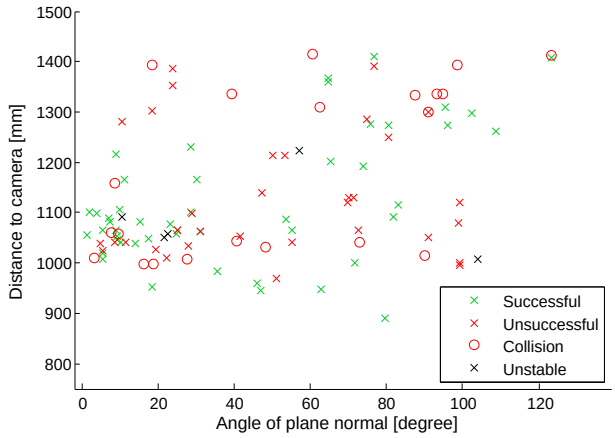

Fig. 20. Distribution of recorded grasps. Unstable grasps have not been used for learning.

The effect of the learning is tested by randomly dividing the data set stored in the episodic memory into a training set and a test set, containing about 117 resp. 20 evaluated grasps. Then the 10 grasps which get the highest score from the network are selected from the test set and the amount of successful grasps, the success ratio, is determined. For comparison, 10 grasps have been selected randomly from the test set and also their success ratio has been determined. This procedure has been repeated 20 times and the average success ratios have been computed in order to compensate for the size of the overall data set. The average success ratio increased from below 35 when grasps have been selected randomly to more than 45 when the trained neural network has been used.

For further development on the learning part we refer to [67] where relations between 3D contours has been integrated and used as features as they are pose invariant and semantically richer.

8. Summary and conclusion

We have described a grasping mechanism that does not make use of any specific object prior knowledge. Instead, the mechanism makes use of second order relations between visually extracted 3D features representing edge structures. We showed that our algorithm, although making use of such rather simple constraints, is able to grasp objects with a reasonable success rate in rather complex environments. Meanwhile, the grasping mechanism has also been used on a humanoid robot.

Moreover, we have described the role of our grasping behaviour within a cognitive system. The system is able to evaluate the success of the grasps autonomously by haptic feedback. By this it can create ground truth in terms of labelled data that has been used for improving the initially hard-wired algorithm by learning. Moreover, the grasping behaviour has been used to trigger higher level processes such as object learning and learning of object specific grasping [12,13].

Grasping without prior object knowledge is a task in which multiple cues need to be merged. In this way, we see our 3D approach as complementary to other mechanism based on 2D information (such as, e.g., [68,10]) or 3D surface information (such as, e.g., [21]).

9. Acknowledgements

This work has been funded within the PACO-PLUS project (IST-FP6-IP-027657). We would like to thank Renaud Detry for providing figure 3c) and d). We also thank Morten Kjærgaard for initial work on the OROCOS control application.

References

[1] A. Bicchi, V. Kumar, Robotic grasping and contact: A review, in: Proceedings of IEEE International Conference on Robotics and Automation, 2000, pp. 348–353.

[2] L. Natale, F. Orabona, G. Metta, G. Sandini, Exploring the world through grasping: a developmental approach, in: Computational Intelligence in Robotics and Automation, 2005. CIRA 2005. Proceedings. 2005 IEEE International Symposium on, 2005, pp. 559–565.

[3] G. Recatalá, E. Chinellato, A. P. D. Pobil, Y. Mezouar, P. Martinet, Biologically-inspired 3D grasp synthesis based on visual exploration, Autonomous Robots 25 (1-2) (2008) 59–70.

[4] K. Huebner, S. Ruthotto, D. Kragic, Minimum Volume Bounding Box Decomposition for Shape Approximation in Robot Grasping, in: Proceedings of the 2008 IEEE International Conference on Robotics and Automation, 2008, pp. 1628–1633.

[5] C. Borst, M. Fischer, G. Hirzinger, Grasp Planning: How to Choose a Suitable Task Wrench Space, New Orleans, LA, USA, 2004, pp. 319 – 325.

[6] D. Ding, Y.-H. Liu, S. Wang, The synthesis of 3-d form-closure grasps, Robotica 18 (1) (2000) 51–58.

[7] M. Buss, H. Hashimoto, J. B. Moore, Dextrous hand grasping force optimization, Robotics and Automation, IEEE Transactions on 12 (3) (1996) 406–418.

[8] R. Platt, Learning Grasp Strategies Composed of Contact Relative Motions, in: IEEE-RAS International Conference on Humanoid Robots, 2007.

[9] R. Pelossof, A. Miller, P. Allen, T. Jebara, An SVM Learning Approach to Robotic Grasping, Robotics and Automation, 2004. Proceedings. ICRA '04. 2004 IEEE International Conference on (2004) 3512–3518 Vol.4.

[10] A. Saxena, J. Driemeyer, J. Kearns, C. Osondu, A. Y. Ng., Learning to grasp novel objects using vision, In 10th International Symposium of Experimental Robotics (ISER).

[11] P. J. Kellman, M. E. Arterberry, The Cradle of Knowledge: Development of Perception in Infancy (Learning, Development, and Conceptual Change), The MIT Press, 1998.

[12] D. Kraft, N. Pugeault, E. Başeski, M. Popović, D. Kragic, S. Kalkan, F. Wörgötter, N. Krüger, 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(accepted).

[13] R. Detry, M. Popović, Y. P. Touati, E. Başeski, N. Krüger, J. Piater, Autonomuous Learning of Object-specific Grasp Affordance Densities, 8th International Conference on Development and Learning.

[14] N. Pugeault, F. Wörgötter, N. Krüger, Accumulated visual representation for cognitive vision, in: British Machine Vision Conference, 2008.

[15] D. Aarno, J. Sommerfeld, D. Kragić, N. Pugeault, S. Kalkan, F. Wörgötter, D. Kraft, N. Krüger, Early Reactive Grasping with Second Order 3D Feature Relations, in: IEEE International Conference on Robotics and Automation (ICRA), Workshop: From features to actions - Unifying perspectives in computational and robot vision, 2007.

[16] R. Detry, N. Pugeault, J. H. Piater, Probabilistic Pose Recovery Using Learned Hierarchical Object Models, in: International Cognitive Vision Workshop (Workshop at the 6th International Conference on Vision Systems), 2008.

[17] R. Detry, J. Piater, A probabilistic framework for 3d visual object representation, IEEE PAMI.

[18] A. Bicchi, On the closure properties of robotic grasping, International Journal of Robotics Research 14 (1995) 319–334.

[19] H. Maekawa, K. Tanie, K. Komoriya, Tactile feedback for multifingered dynamic grasping, Control Systems Magazine, IEEE 17 (1) (1997) 63–71.

[20] J. Tegin, S. Ekvall, D. Kragic, B. Iliev, J. Wikander, Demonstration based Learning and Control for Automatic Grasping, in: Proc. of the International Conference on Advanced Robotics, 2007.

[21] A. T. Miller, S. Knoop, H. Christensen, P. K. Allen, Automatic grasp planning using shape primitives, in: Proceedings of the IEEE International Conference on Robotics and Automation, 2003, Vol. 2, 2003, pp. 1824–1829.

[22] K. K. Aydin, Fuzzy logic, grasp preshaping for robot hands, in: ISUMA '95: Proceedings of the 3rd International Symposium on Uncertainty Modelling and Analysis, IEEE Computer Society, Washington, DC, USA, 1995, pp. 520 – 523.

[23] M. R. Cutkosky, On grasp choice, grasp models, and the design of hands for manufacturing tasks, IEEE Transactions on Robotics and Automation 5 (3) (1989) 269–279.

[24] T. Iberall, Human prehension and dexterous robot hands, The International Journal of Robotics Research 16 (3) (1997) 285–299.

[25] A. T. Miller, P. K. Allen, GraspIt!: A Versatile Simulator for Robotic Grasping, Robotics & Automation Magazine, IEEE 11 (4) (2004) 110–122.

[26] J. Jørgensen, H. Petersen, Usage of simulations to plan stable grasping of unknown objects with a 3-fingered Schunk hand, in: Workshop on robot simulators: available software, scientific applications and future trends, ICRA, 2008.

[27] S. Ekvall, D. Kragic, Integrating object and grasp recognition for dynamic scene interpretation, in: IEEE International Conference on Advanced Robotics, 2005. ICAR'05, 2005, pp. 331–336.

[28] J. Steil, F. Röthling, R. Haschke, H. Ritter, Situated robot learning for multi-modal instruction and imitation of grasping, Robotics and Autonomous Systems Special Is (47) (2004) 129–141.

[29] R. Dillmann, M. Kaiser, A. Ude, Acquisition of elementary robot skills from human demonstration, in: In International Symposium on Intelligent Robotics Systems, 1995, pp. 185–192.

[30] Scape Technologies, http://www.scapetechnologies.com/.

[31] M. Salganicoff, L. H. Ungar, R. Bajcsy, Active learning for vision-based robot grasping, Machine Learning 23 (2-3) (1996) 251–278.

[32] R. Mario, V. Markus, Grasping of unknown objects from a table top, in: Workshop on Vision in Action: Efficient strategies for cognitive agents in complex environments,ViA08, 2008.

[33] M. J. Taylor, A. Blake, Grasping the Apparent Contour, in: ECCV '94: Proceedings of the Third European Conference-Volume II on Computer Vision, Springer-Verlag, London, UK, 1994, pp. 25–34.

[34] G. Bekey, H. Liu, R. Tomović, W. Karplus, Knowledge-Based Control of Grasping in Robot Hands Using Heuristics from Human Motor Skills, IEEE Trans. Robotics and Automation vol. 9, no. 6 (1993) 709–722.

[35] E. Chinellato, R. B. Fisher, A. Morales, A. P. D. Pobil, Ranking planar grasp configurations for a three-finger hand., in: ICRA, 2003, pp. 1133–1138.

[36] A. Morales, P. J. Sanz, A. P. D. Pobil, A. H. Fagg, Vision-based three-finger grasp synthesis constrained by hand geometry, Robotics and Autonomous Systems 54 (6) (2006) 496–512.

[37] D. P. Perrin, C. E. Smith, O. Masoud, N. Papanikolopoulos, Unknown Object Grasping Using Statistical Pressure Models, in: Proceedings of the 2000 IEEE International Conference on Robotics and Automation, ICRA 2000, April 24-28, 2000, San Francisco, CA, USA, 2000, pp. 1054–1059.

[38] G. Taylor, L. Kleeman, Grasping unknown objects with a humanoid Robot, Proceedings 2002 Australasian Conference on Robotics and Automation (2002) 191–196.

[39] F. Ade, M. Rutishauser, M. Trobina, Grasping unknown objects, in: Proceedings of Dagstuhl Seminar: Environment Modeling and Motion Planning for Autonomous Robots, World, 1995, pp. 445–459.

[40] B. Wang, L. Jiang, J. LI, H. Cai, Grasping unknown objects based on 3D model reconstruction, Advanced Intelligent Mechatronics. Proceedings, 2005 IEEE/ASME International Conference on (2005) 461 – 466.
[41] A. Jefferson, J. Coelho, J. H. Piater, R. A. Grupen, Developing haptic and visual perceptual categories for reaching and grasping with a humanoid robot, Robotics and Autonomous Systems 37 (2-3) (2001) 195–218.
[42] J. J. Gibson, The Ecological Approach to Visual Perception, Lawrence Erlbaum Associates.
[43] G. Fritz, L. Paletta, M. Kumar, G. Dorffner, R. Breithaupt, E. Rome, Visual Learning of Affordance Based Cues, in: Simulation of Adaptive Behavior, Vol. 4095, 2006, pp. 52–64.
[44] J. Steffen, R. Haschke, H. Ritter, Experience-based and Tactile-driven Dynamic Grasp Control, Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on (2007) 2938–2943.
[45] J. Tegin, S. Ekvall, D. Kragic, B. Iliev, J. Wikander, Demonstration based learning and control for automatic grasping, in: Int. Conf. on Advanced Robotics, Jeju, Korea, 2007.
[46] Grasping Video, http://www.mip.sdu.dk/covig /videos/graspingReflexCompressed.divx.
[47] N. Krüger, M. Lappe, F. Wörgötter, Biologically Motivated Multi-modal Processing of Visual Primitives, The Interdisciplinary Journal of Artificial Intelligence and the Simulation of Behaviour 1 (5) (2004) 417–428.
[48] N. Pugeault, Early Cognitive Vision: Feedback Mechanisms for the Disambiguation of Early Visual Representation, VDM Verlag Dr. Müller, 2008.
[49] N. Pugeault, F. Wörgötter, N. Krüger, Multi-modal Scene Reconstruction Using Perceptual Grouping Constraints, in: Proc. IEEE Workshop on Perceptual Organization in Computer Vision (in conjunction with CVPR'06), 2006.
[50] S. Kalkan, Multi-modal Statistics of Local Image Structures and its Applications for Depth Prediction, Ph.D. thesis, University of Goettingen, Germany (2008).
[51] RobWork, http://www.mip.sdu.dk/robwork/.
[52] The Orocos Real-Time Toolkit, http://www.orocos.org/rtt.
[53] J. J. Kuffner, J. Steven, M. Lavalle, Rrt-connect: An efficient approach to single-query path planning, in: In Proc. IEEE Intl Conf. on Robotics and Automation, 2000, pp. 995–1001.
[54] E. Larsen, S. Gottshalck, M. Lin, D. Manocha, Fast proximity queries with swept sphere volumes, Tech. Rep. TR99-018, Department of Computer Science, University of North Carolina (1999.).
[55] M. Popović, An Early Grasping Reflex in a Cognitive Robot Vision System, Master's thesis, Cognitive Vision Lab, The Mærsk Mc-Kinney Møller Institute, University of Southern Denmark (2008).
[56] A. Dollar, R. Howe, Simple, reliable robotic grasping for human environments, in: Technologies for Practical Robot Applications, 2008. TePRA 2008. IEEE International Conference on, 2008, pp. 156–161.
[57] L. Jiang, D. Sun, H. Liu, An inverse-kinematics table-based solution of a humanoid robot finger with nonlinearly coupled joints, Mechatronics, IEEE/ASME Transactions on 14 (3) (2009) 273–281.
[58] T. Asfour, K. Regenstein, P. Azad, J. Schroder, A. Bierbaum, N. Vahrenkamp, R. Dillmann, ARMAR-III: An integrated humanoid platform for sensory-motor control, in: Humanoid Robots, 2006 6th IEEE-RAS International Conference on, 2006, pp. 169–175.
[59] M. Arbib, T. Iberall, D. Lyons, Coordinated control programs for movements of the hand, Experimental brain research (1985) 111–129.
[60] M. R. Cutkosky, R. D. Howe, Human grasp choice and robotic grasp analysis (1990) 5–31.
[61] I. Gaiser, S. Schulz, A. Kargov, H. Klosek, A. Bierbaum, C. Pylatiuk, R. Oberle, T. Werner, T. Asfour, G. Bretthauer, R. Dillmann, A new anthropomorphic robotic hand, in: Humanoid Robots, 2008. Humanoids 2008. 8th IEEE-RAS International Conference on, 2008, pp. 418–422.
[62] Grasping Reflex on the Humanoid Robot ARMAR-III, http://www.mip.sdu.dk/covig/videos/ GraspReflexHumanoid.avi.
[63] PACO-PLUS: Perception, Action and Cognition through learning of Object-Action Complexes, iST-FP6-IP-027657, Integrated Project (2006-2010).
[64] A. D. Baddeley, Essentials of Human Memory, Psychology Press, Taylor and Francis, 1999.
[65] N. Pugeault, S. Kalkan, E. Başeski, F. Wörgötter, N. Krüger, Reconstruction uncertainty and 3D relations, in: Proceedings of Int. Conf. on Computer Vision Theory and Applications (VISAPP'08), 2008.
[66] L. Bodenhagen, Project in Artificial Intelligence, http://www.mip.sdu.dk/covig/publications/reportAIP.pdf (2008).
[67] L. Bodenhagen, D. Kraft, M. Popović, E. Başeski, P. E. Hotz, N. Krüger, Learning to grasp unknown objects based on 3d edge information, IEEE International Symposium on Computational Intellegience in Robotics and Automation.
[68] J. Pauli, H. Hexmoor, M. Mataric, Learning to recognize and grasp objects, Autonomous Robots 5 (1998) 239–258.

Mila Popović received her B.Sc. degree in Astrophysics from the University of Belgrade, Serbia and her M.Sc. degree in Computer Systems Engineering from the University of Southern Denmark, Denmark, in 2004 and 2008 respectively. She is currently a Ph.D. student in the Mærsk McKinney Møller Institute, University of Southern Denmark. Her research interests include robotic grasping and cognitive robotics.

Dirk Kraft obtained a diploma degree in computer science from the University of Karlsruhe (TH), Germany in 2006 and a Ph.D. degree from the University of Southern Denmark in 2009. He is currently a research assistant at the Mærsk McKinney Møller Institute, University of Southern Denmark where he is working within the EU-project PACO-PLUS. His research interests include cognitive systems, robotics and computer vision.

Leon Bodenhagen received his M.Sc. in Computer Systems Engineering in 2009 from the University of Southern Denmark, where he now is a Ph.D. student at the Mærsk McKinney Møller Institute. His research interests include cognitive systems and cognitive robotics.

Emre Başeski received his B.S. and M.S. degree in Computer Engineering from the Middle East Technical University, Turkey, in 2003 and 2006 respectively. He is currently a Ph.D. student in the Mærsk McKinney Møller Institute, University of Southern Denmark. His research interests include machine vision and learning.

Nicolas Pugeault obtained an M.Sc. from the University of Plymouth in 2002 and an Engineer degree from the Ecole Supérieure d'Informatique, Électronique, Automatique (Paris) in 2004. He obtained his Ph.D. from the University of Göttingen in 2008, and is currently working as a Reseach Fellow at the University of Surrey, United Kingdom.

Danica Kragic received her B.S. degree in mechanical engineering from the Technical University of Rijeka, Croatia, and her Ph.D. degree in computer science from the Royal Institute of Technology (KTH), Stockholm, Sweden in 1995 and 2001, respectively. She is currently a professor in computer science at KTH and chairs the IEEE RAS Committee on Computer and Robot Vision. She received the 2007 IEEE Robotics and Automation Society Early Academic Career Award. Her research interests include vision systems, object grasping and manipulation and action learning.

Tamim Asfour received his diploma degree in electrical engineering and his Ph.D. degree in computer science from the University of Karlsruhe, Germany in 1994 and 2003, respectively. Currently, he is leader of the humanoid robotics research group at the institute for Anthromopmatics at the Karlsruhe Institute of Technology (KIT). His research interests include humanoid robotics, grasping and manipulation, imitation learning, system integration and mechatronics.

Norbert Krüger is a Professor at the Mærsk McKinney Møller Institute, University of Southern Denmark. He holds a M.Sc. from the Ruhr-Universität Bochum, Germany and his Ph.D. from the University of Bielefeld. He is a partner in several EU and national projects: PACO-PLUS, Drivsco, NISA, Handyman.

Norbert Krüger is leading the Cognitive Vision Lab which is focussing on computer vision and cognitive systems, in particular the learning of object representations in the context of grasping. He has also been working in the areas of computational neuroscience and machine learning.