In IEEE International Conference on Robotics and Automation, ICRA'07, *Workshop: "From features to actions - Unifying perspectives in computational and robot vision"*, April, 2007, Rome, Italy (to be presented)

Early Reactive Grasping with Second Order 3D Feature Relations

Daniel Aarno, Johan Sommerfeld, Danica Kragic
Royal Institute of Technology, Sweden
{bishop, johansom, dani}@kth.se

Nicolas Pugeault
University of Edinburgh, UK
npugeaul@inf.ed.ac.uk

Sinan Kalkan, Florentin Wörgötter
University of Göttingen, Germany
{sinan, worgott}@bccn-goettingen.de

Dirk Kraft, Norbert Krüger
Sydansk University and Aalborg University, Denmark
{norbert, kraft}@mip.sdu.dk

***Abstract*— One of the main challenges in the field of robotics is to make robots ubiquitous. To intelligently interact with the world, such robots need to understand the environment and situations around them and react appropriately, they need context-awareness. But how to equip robots with capabilities of gathering and interpreting the necessary information for novel tasks through interaction with the environment and by providing some minimal knowledge in advance? This has been a longterm question and one of the main drives in the field of cognitive system development.**

The main idea behind the work presented in this paper is that the robot should, like a human infant, learn about objects by interacting with them, forming representations of the objects and their categories that are grounded in its embodiment. For this purpose, we study an early learning of object grasping process where the agent, based on a set of innate reflexes and knowledge about its embodiment. We stress out that this is not the work on grasping, it is a system that interacts with the environment based on relations of 3D visual features generated trough a stereo vision system. We show how geometry, appearance and spatial relations between the features can guide early reactive grasping which can later on be used in a more purposive manner when interacting with the environment.

I. Introduction

For a robot that has to perform tasks in a human environment, it is necessary to be able to learn about objects and object categories. It has been recognized recently that grounding in the embodiment of a robot, as-well as continuous learning is required to facilitate learning of objects and object categories [1], [2]. The idea is that robots will not be able to form useful categories or object representations by only being a passive observer of its environment. Rather a robot should, like a human infant, learn about objects by interacting with them, forming representations of the objects and their categories that are grounded in its embodiment.

Central to the approach are three almost axiomatic assumptions, which are strongly correlated. These also represent the building blocks of our approach toward creating a cognitive artificial agent:

- Objects and Actions are inseparably intertwined; Entities ("things") in the world of a robot (or human) will only become semantically useful "objects" through the action that the agent can/will perform on them. This forms so-called Object-Action Complexes (named OACs) which are the building blocks of cognition.
- Cognition is based on recurrent processes involving nested feedback loops operating on, contextualizing and reinterpreting object-action complexes. This is done through actively closing the perception-action cycle.
- A unified measure of success and progress can be obtained through minimization of contingencies which an artificial cognitive system experiences while interacting with the environment or other agents, given the drives of the system.

To demonstrate the feasibility of our approach, we aim at building a robot system that step by step develop increasingly advanced cognitive capabilities. In this paper, we demonstrate our initial efforts towards this goal by designing a scenario for manipulation and grasping of objects.

One of the most basic interactions that can occur between a robot and an object is for the robot to push the object, i.e. to simply make a physical contact. Already at this stage, the robot should be able to form two categories: physical and non-physical objects, where a physical object is categorized by the fact that interaction forces occur. A higher level interaction between the robot and an object would exist if the robot was able to *grasp* the object. In this case, the robot would gain actual physical control over the object and having the possibility to perform controlled actions on it, such as examining it from other angles, weighing it, placing it etc. Information obtained during this interaction can then be used to update the robots representations about objects and the world. Furthermore, the successfully performed grasps can be used as ground truth for future grasp refinement, [2].

In this paper, we are interested in investigating an initial "reflex-like" grasping strategy that will form a basis for

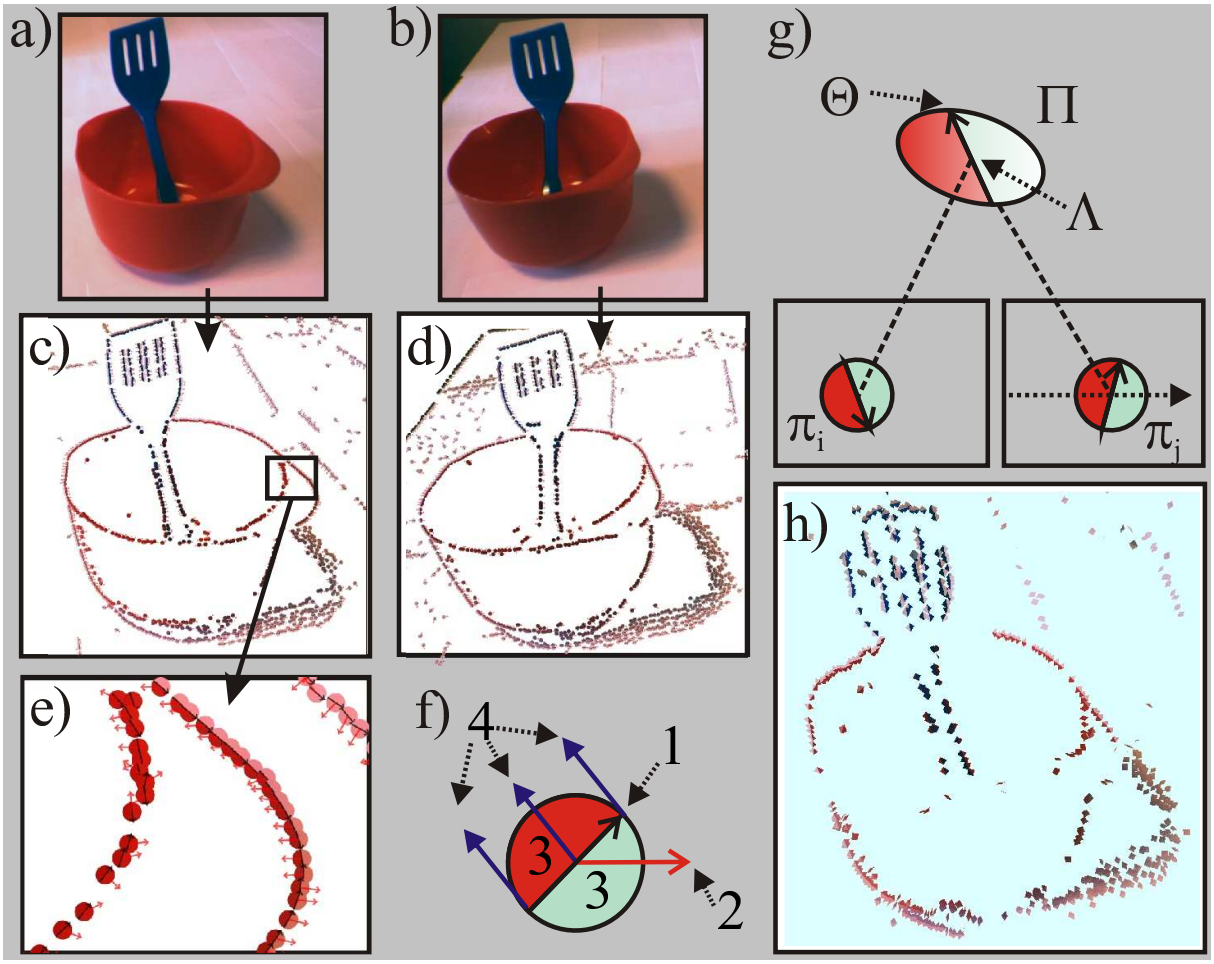


Fig. 1. Illustration of the vision module. a) and b) shows the images captured by the left and right cameras (respectively); c) and d) show the primitives extracted from these two images; in e) a detail of the primitive extraction is shown; f) illustrates the schematic representation of a primitive, where 1. represents the orientation, 2. the phase, 3. the color and 4. the optical flow. g) from a stereo–pair of primitives (π_i, π_j) we reconstruct a 3D primitive Π, with a position in space Λ and an orientation Θ; h) shows the resulting 3D primitives reconstructed for this scenario.

a cognitive robot system that, at the first stage, acquires knowledge of objects and object categories and is able to further refine its grasping behavior by incorporating the gained object knowledge, [3]. The grasping strategy does not require *a-priori* object knowledge, and it can be adopted for a large class of objects. The proposed reflex-like grasping strategy is based on second order relations of multi-modal visual features descriptors, called *spatial primitives*, that represent object's geometric information, e.g. 3D pose (position and orientation) as well as its appearance information, e.g. color and contrast transition etc. [4], see Fig. 1. Co–planar tuples of the spatial primitives allow for the definition of a plane that can be associated to a grasp hypothesis. In addition, these local descriptors are part of semi-global collinear groups [5]. Furthermore, the color information (by defining co–colority in addition to co–planarity of primitive pairs) can be used to further improve the definition of grasp hypotheses. In this paper, we employ the structural richness of the descriptors in terms of their geometry and appearance as well as the structural relations co–linearity, co–planarity and co–colority to derive a set of grasping options from a stereo image.

We note that the purpose of this work is not to develop yet another grasping strategy for a specific setting, but rather to provide low-level grasping reflexes that can be used to generate successful grasps on arbitrary objects. These grasping reflexes are part of a larger framework on cognitive robotics where a robot is equipped only with a set of innate grasps which are used to develop more complex object manipulation abilities through interaction and reinforcement so that 1) more complex feature relations become associated to more precise and successful grasps, and 2) object knowledge becomes acquired and used to further refine the grasping process. We also have to stress out that no scene segmentation is performed, since the system does not even have a concept of an object to start with. In short, the contributions of our work are the generation of a set of grasp suggestions on unknown objects based on visual feedback, grouping of visual primitives for decreasing the size of the grasps and evaluation of grasps using the GraspIt! environment, [6].

In this work, "kitchen-type" objects such as cups, glasses, bowls and various kitchen utensils are considered. However,

our algorithm is not designed for specific object classes but can be applied for any rigid object that can be described by edge–like structures.

This paper is organized as follows. In Section II, we shortly review the related work and in Section III give a general overview of the system. Details about extraction of spatial primitives are presented in Section IV and elementary grasping actions defined in Section V. Results of the experimental evaluation are summarized in Section VI and plans for future research outlined in Section VII.

II. Related Work

The idea to learn or refine grasping strategies is not new. Kamon *et al.* combined heuristic methods with learning algorithms to learn how to select good grasps [7]. Rössler *et al.* used two levels of learners to learn local and global grasp criteria [8], where the local learner learns about the local structure of an object and the global learner learns which of the possible local grasps are best given the object.

There has been a large amount of work presented in the area of robotic grasping during the last two decades [9]. However, much of this work has been dealing with analytical methods where the shape of the objects being grasped is known *a-priori*. This work, referred to as *analytical methods*, has focused primarily on computing grasp stability based on force and form-closure properties or contact-level grasps synthesis based on finding a fixed number of contact locations with no regard to hand geometry, [9],[10]. This problem is important and difficult mainly because of the high number of DOFs involved in grasping arbitrary objects with complex hands. Another important research area is grasp planning without detailed object models where sensor information such as computational vision is used to extract relevant features in order to compute suitable grasps, [11], [12], [13]. In this paper, we denote this approach as *sensor-driven*.

Related to our work, we have to mention systems that deal with automatic grasp synthesis and planning, [14],[15],[16],[17]. This work concentrates on automatic generation of stable grasps given assumptions about the shape of the object and robot hand kinematics. Example of assumptions may be that the full and exact pose of the object is known in combination with its (approximate) shape, [14]. Another common assumption is that the outer contour of the object can be extracted and a planar grasp applied, [16]. Taking into account both the hand kinematics as well as some *a-priori* knowledge about the feasible grasps has been acknowledged as a more flexible and natural approach towards automatic grasp planning [18],[14]. [18] studies methods for adapting a given prototype grasp of one object to another object. The method proposed in [14] presents a system for automatic grasp planning for a Barrett hand [19] by modeling an object as a set of shape primitives, such as spheres, cylinders, cones and boxes in a combination with a set of rules to generate a set of grasp starting positions and pregrasp shapes.

One difference between the analytical and sensor-driven approaches is that the former tend to use complex hands with many DOFs, while the latter use simple ones such as parallel yaw-grippers. One reason for this is that if the reconstruction of the object's shape is not very accurate, using a complex gripping device does not necessarily facilitate grasping performance. For sensor-driven approaches it is also very common to perform only planar grasps where all the contacts between the fingers and the object are confined to a plane. As an example, objects are placed on a table and grasped from above. This simplifies both the vision problem, since only the outer boundary of the object in the image plane has to be estimated, as well as the grasp planning by constraining the search space.

The main differences of our work compared to the above-mentioned work are the following:

- We rely on 3D information based on three dimensional primitives extracted online. This allows us to compute arbitrary grasping directions compared to only planar grasps considered in, e.g. [16].
- The structural richness of the primitives (geometric and appearance based information, collinear grouping) allows for an efficient reduction of grasping hypotheses while keeping relevant ones.
- Our system focuses on generating a certain percentage of successful grasps on arbitrary objects rather than high quality grasps on a constrained set of objects. We will show that with our representations we are able to extract a sufficient number of successful grasping options to be used as initiator of learning schemes aiming at more sophisticated grasping strategies.

III. System Overview

The work presented in this paper serves as a building block for the development of a cognitive robot system. The robot platform considered is comprised of a set of sensors and actuators. The minimum requirements necessary to realize the work presented in this paper is that the sensors are able to deliver a set of visual primitives (section IV) and the configuration of the actuators. The required actuator is a manipulator, comprised of a robotic arm and a gripper device. In this context the term sensor is not necessarily related to a real physical sensing device, but rather an abstract measurement delivered to the system, possibly after performing computations on data sampled from a physical sensor.

The complete system is outlined in Fig. 2. In this paper we are interested in developing grasping reflexes. A grasping reflex is triggered by the vision system. The vision system continuously computes the spatial primitives described in section IV which are feed as sensor input to the set of reflexes and to the cognitives system. If the grasping reflex has not been inhibited by the cognitive system and the sensor stimuli is strong enough, i.e. there are sufficiently many spatial primitives visible, the grasping reflex is performed. This reflex behavior computes a set of possible grasps and tries to perform them. Each grasp evaluated results in a reinforcement signal which can be used by the cognitive system to update its representation of the world. The following

two sections describe the spatial primitives and the rules for generating the grasping actions.

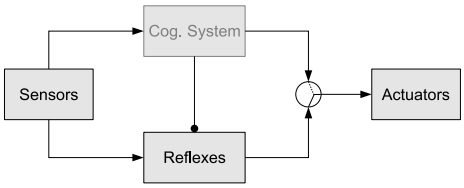


Fig. 2. System overview

IV. Spatial Primitives

The image processing used in this paper is based on multi-modal visual primitives [20], [4], [5]. First, 2D primitives are extracted sparsely at points of interest in the image (in this case contours) and encode the value of different visual operators (hereby referred to as *visual modalities*) such as local orientation, phase, color (on each side of the contour) and optical flow (see Fig. 1.d, 1.e and 1.f). In a second step, the 2D primitives become extended to the spatial primitives used in this work. After finding correspondences between primitives in the left and right image, we reconstruct a spatial primitive, (see Fig. 1.g) that has the following components, (for details see [21], [5]):

$$\Pi = \{\Lambda, \Theta, \Omega, (\mathbf{c}_l, \mathbf{c}_m, \mathbf{c}_r)\},$$

where Λ is the 3D position; Θ is the 3D orientation; Ω is the phase (i.e., contrast transition); and, $(\mathbf{c}_l, \mathbf{c}_m, \mathbf{c}_r)$ is the representation of the color of the spatial primitive, corresponding to the left ($\mathbf{c}_l$), the middle ($\mathbf{c}_m$) and the right side ($\mathbf{c}_r$).

The sparseness of the primitives allows to formulate three *relations* between primitives that are crucial in our context:

- *Co–planarity:*
  Two spatial primitives Π_i and Π_j are co–planar iff their orientation vectors lie on the same plane, i.e.:

  $$cop(\Pi_i, \Pi_j) = 1 - |\mathbf{proj}_{\Theta_j \times \mathbf{v}_{ij}}(\Theta_i \times \mathbf{v}_{ij})|,$$

  where $\mathbf{v}_{ij}$ is defined as the vector $(\Lambda_i - \Lambda_j)$, and $\mathbf{proj}_{\mathbf{u}}(\mathbf{a})$ is defined as:

  $$\mathbf{proj}_{\mathbf{u}}(\mathbf{a}) = \frac{\mathbf{a} \cdot \mathbf{u}}{\| \mathbf{u} \|^2}\mathbf{u}. \quad (1)$$

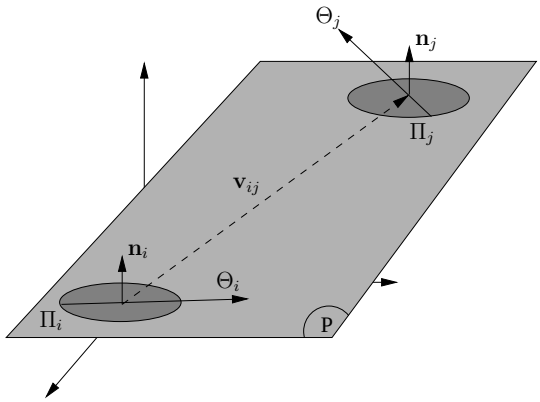
  The co–planarity relation is illustrated in Fig. 3(a).
- *Collinear grouping (i.e., collinearity):*
  Two spatial primitives Π_i and Π_j are collinear (i.e., part of the same group) iff they are part of the same contour. Due to uncertainty in 3D reconstruction process, in this work, the collinearity of two spatial primitives Π_i and Π_j is computed using their 2D projections π_i and π_j. We define the collinearity of two 2D primitives π_i and π_j as:

  $$col(\pi_i, \pi_j) = 1 - \left| sin\left(\frac{|\alpha_i| + |\alpha_j|}{2}\right)\right|,$$

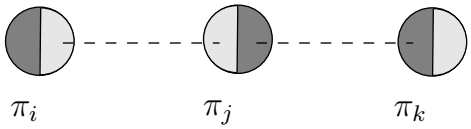
  where α_i and α_j are as shown in Fig. 3(c), see [5] for more details on collinearity.
- *Co–colority:* Two spatial primitives Π_i and Π_j are co–color iff their parts that face each other have the same color. In the same way as collinearity, co–colority of two spatial primitives Π_i and Π_j is computed using their 2D projections π_i and π_j. We define the co–colority of two 2D primitives π_i and π_j as:

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

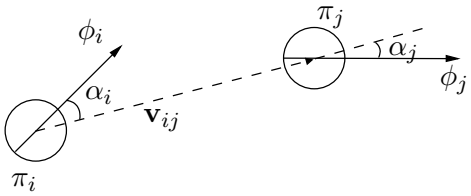
  where $\mathbf{c}_i$ and $\mathbf{c}_j$ are the RGB representation of the colors of the parts of the primitives π_i and π_j that face each other; and, $\mathbf{d}_c(\mathbf{c}_i, \mathbf{c}_j)$ is Euclidean distance between RGB values of the colors $\mathbf{c}_i$ and $\mathbf{c}_j$. In Fig. 3(b), a pair of co–color and not co–color primitives are shown.



(a) Co–planarity of two 3D primitives Π_i and Π_j.



(b) Co–colority of three 2D primitives π_i, π_j and π_k. In this case, π_i and π_j are cocolor, so are π_i and π_k; however, π_j and π_k are not cocolor.



(c) Collinearity of two 2D primitives π_i and π_j.

Fig. 3. Illustration of the relations between a pair of primitives.

Co–planarity in combination with the 3D position allows for the definition of a grasping pose; Collinearity and co–colority allows for the reduction of grasping hypotheses. The use of the relations in the grasping context is shown in Fig. 4.

V. Elementary Grasping Actions

Coplanar relationships between visual primitives suggests different graspable planes. Fig. 4 shows a set of spatial primitives on two different contours l_i and l_j with co–planarity, co–colority and collinearity relations.

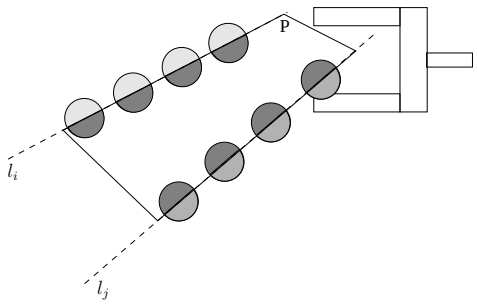


Fig. 4. A set of spatial primitives on two different contours l_i and l_j that have co–planarity, co–colority and collinearity relations; a plane P defined by the co–planarity of the spatial primitives and and an example grasp suggested by the plane.

Five elementary grasping actions (EGA) will be considered as shown in Fig. 5. EGA1 is a "pinch" grasp on a thin edge like structure with approach direction along the surface normal of the plane spanned by the primitives. EGA2 is an "inverted" grasp using the inside of two edges with approach along the surface normal. EGA3 is a "pinch" grasp on a single edge with approach direction perpendicular to the surface normal. EGA4 is similar to EGA2 but its approach direction is perpendicular to the surface normal. Also it tries to go in "below" one of the primitives. EGA5 is wide grasp making contact on two separate edges with approach direction along the surface normal.

The EGAs will be parameterized by their final pose (position and orientation) and the initial gripper configuration. For the simple parallel jaw gripper, an EGA will thus be defined by seven parameters: EGA$(x, y, z, \gamma, \beta, \alpha, \delta)$ where $\mathbf{p} = [x, y, z]$ is the position of the gripper "center" according to Fig. 6; $\gamma,\ \beta,\ \alpha$ are the roll, pitch and yaw angles of the vector $\mathbf{n}$; and δ is the gripper configuration, see Fig. 6. Note that the gripper "center" is placed in the "middle" of the gripper.

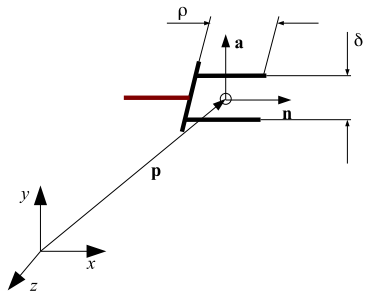


Fig. 6. Parameterization of EGAs.

The main motivation for choosing these grasps is that they represent the simplest possible two fingered grasps humans commonly use. The result of applying the EGAs can be evaluated to provide a reinforcement signal to the system. The number of possible outcomes of each of the EGAs are different and will be explained below.

For all of the EGAs the possibility of an *early failure* exists. That is, the EGA fails before reaching the target configuration. This will result in a reinforcement R_{fe}. Furthermore, it is possible for all EGAs to fail a grasping procedure.

For EGA1, EGA3 and EGA5, a failed grasp can be detected by the fact that the gripper is completely closed. This situation will result in a reinforcement R_{fl}.

For EGA1 and EGA3, the expected grasp is a pinch type grasp, i.e. narrow. Therefore, they can also "fail" if the gripper comes to a halt too early, that is $\delta > \Delta_{min}$. This will result in a reinforcement R_{ft}.

EGA2 fails if the gripper is fully opened, meaning that no contact was made with the object. This gives a reinforcement R_{fh}.

To detect failure of EGA4, a tactile sensor is required on the side of the "fingers". If, after positioning and opening the gripper, there is no contact between the object and the tactile sensor, the EGA has failed. This results in a reinforcement R_{fc}.

If none of the above situations is encountered, a positive reinforcement R_g is given, and the EGA is considered successful.

A. Computing Action Parameters

Let $\Gamma = \{\Pi_1, \Pi_2\}$ be a primitive pair, $\Lambda(\Pi)$ be the position of Π and $\Theta(\Pi)$ be the orientation of Π, also let Γ_i be the i:th pair. From that we can calculate

$$\begin{aligned}
\mathbf{d} &= \Lambda(\Pi_2) - \Lambda(\Pi_1) \\
\mathbf{n}_1 &= \Theta(\Pi_1) \times \mathbf{d} \\
\mathbf{n}_2 &= \Theta(\Pi_2) \times \mathbf{d} \\
sw &= \begin{cases} -1 & \text{if} \quad \mathbf{n}_1 \cdot \mathbf{n}_2 < 0 \\ 1 & \text{else} \end{cases}
\end{aligned}$$

and with those we calculate the plane $\mathbf{p}$

$$\begin{aligned}
\mathbf{P_p} &= \Lambda(\Pi_1) + frac\mathbf{d}2 \\
\mathbf{n_p} &= \frac{\mathbf{n}_1 + sw\mathbf{n}_2}{\parallel \mathbf{n}_1 + sw\mathbf{n}_2 \parallel}
\end{aligned}$$

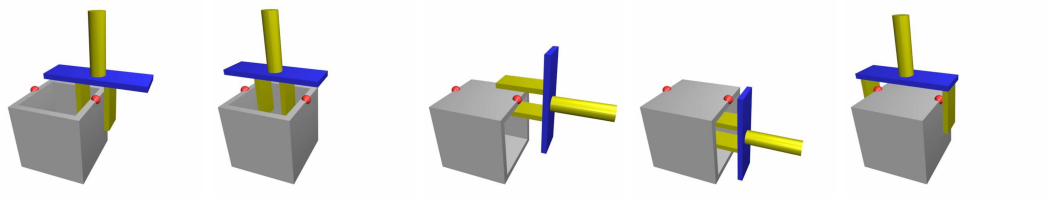
which is used when calculating actions parameters

The parameterization of the EGAs is given with the gripper normal $\mathbf{n}$ and the normal of the surface between the two fingers $\mathbf{a}$ as illustrated in Fig. 6. From this, the yaw, pitch and roll angles can be easily computed.

For EGA1, there will be two possible parameter sets given the primitive pair $\Gamma = \{\Pi_1, \Pi_2\}$. The parameterization is as follows:

$$\begin{aligned}
&\mathbf{p}_{\text{gripper}} = \Lambda(\Pi_i) \\
&\mathbf{n} = \nabla(\mathbf{p}) \\
&\mathbf{a} = \mathbf{perp_n}(\Theta(\Pi_i))/ \parallel \mathbf{perp_n}(\Theta(\Pi_i)) \parallel \quad \text{for } i = 1, 2
\end{aligned}$$

where $\nabla(\mathbf{p})$ is the normal of the plane $\mathbf{p}$ and $\mathbf{perp_u}(\mathbf{a})$ is the projection of $\mathbf{a}$ perpendicular to $\mathbf{u}$. That is $\mathbf{perp_u}(\mathbf{a}) = \mathbf{a} - \mathbf{proj_u}(\mathbf{a})$, where $\mathbf{proj_u}(\mathbf{a})$ is defined according to (1).



(a) EGA1 (b) EGA2 (c) EGA3 (d) EGA4 (e) EGA5

Fig. 5. Elementary grasping actions, EGAs.

For EGA2, there is only one parameter set.

$$\begin{aligned}\mathbf{d} &= \Lambda(\Pi_2) - \Lambda(\Pi_1)\\ \mathbf{p}_{\text{gripper}} &= \Lambda(\Pi_1) + \mathbf{d}/2\\ \mathbf{n} &= \nabla(\mathbf{p})\\ \mathbf{a} &= \mathbf{n} \times \mathbf{d}/ \parallel \mathbf{n} \times \mathbf{d} \parallel\end{aligned}$$

For EGA3, there will be two possible parameter sets for $i = 1, j = 2$ and $i = 2, j = 1$.

$$\begin{aligned}\mathbf{d} &= \Lambda(\Pi_j) - \Lambda(\Pi_i)\\ \mathbf{n} &= \mathbf{d}/ \parallel \mathbf{d} \parallel\\ \mathbf{p}_{\text{gripper}} &= \Lambda(\Pi_i)\\ \mathbf{a} &= \mathbf{n} \times \nabla(\mathbf{p})\end{aligned}$$

For EGA4, there will be two possible parameter sets for $i = 1, j = 2$ and $i = 2, j = 1$. Where ϵ is a step size parameter that will depend on the gripper used.

$$\begin{aligned}\mathbf{d} &= \Lambda(\Pi_j) - \Lambda(\Pi_i)\\ \mathbf{n} &= \mathbf{d}/ \parallel \mathbf{d} \parallel\\ \mathbf{p}_{\text{gripper}} &= \Lambda(\Pi_i) - \nabla(\mathbf{p}) \cdot \epsilon\\ \mathbf{a} &= \mathbf{n} \times \nabla(\mathbf{p})\end{aligned}$$

EGA5 will have the same parameters as EGA2 except that the gripper will be fully opened.

B. Limiting the Number of Actions

For a typical scene, the number of coplanar pairs of primitives is in the order of $10^3 - 10^4$. Given that each coplanar relationship gives rise to 8 different grasps from the five different categories, it is obvious that the number of suggested actions must be further constrained. Another problem is that coplanar structures occur frequently in natural scenes and only a small set of them suggest feasible actions, e.g. objects placed on a table create a lot of 3D line structures coplanar to the table but can not grasped directly by a grasping direction normal to the table. In addition, there exist many coplanar pairs of primitives affording similar grasps.

To overcome some of the above problems, we make use of the structural richness of the primitives. First, their embedding into collinear groups naturally clusters the grasping hypotheses into sets of redundant grasps from which only one needs to be tested. Furthermore, co–colority, gives an additional hypothesis for a potential grasp.

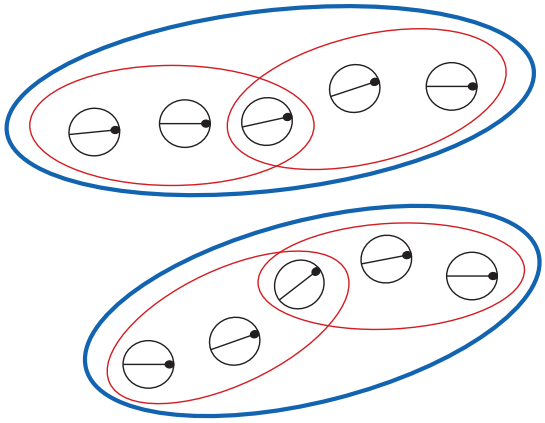


Fig. 7. Small overlapping groups form large groups

1) Using Grouping Information: From the 2D primitives (before stereo reconstruction) collinear neighbors can be found. The collinear neighbors can be mapped to corresponding 3D primitives. These small neighborhoods form the set of *small groups*, $\{g_1, g_2, ..., g_N\}$. The *large groups*, $\{G_1, G_2, ..., G_M\}$, are formed by the grouping of the small groups overlapping each other, Fig. 7 such that if Π_i and Π_j are part of group g_x and Π_j and Π_k is part of group g_y then g_y and g_x is part of the same large group G_z. The result is that the large groups are separated meaning that a primitive that exist in group G_X can not exist in any other group G_Y. Using this grouping information it is possible to add additional constraints on the generation of EGA s.

First, all primitives that are not part of a sufficiently large group G_i are discarded. Secondly, the relations co–planarity and co–colority between small groups of primitives are computed such that primitive $\Pi_i \in g_x$ and $\Pi_j \in g_y$ are only considered to have a co–planarity or co–colority relation if all primitives in g_x are coplanar or cocolor w.r.t all primitives in g_y. Finally, it is possible to constrain the generation of EGAs to only one EGA of each type for each large group.

VI. Experimental Evaluation

Fig. 9, Fig. 10 and Fig. 11 show some of the grasps generated for the scenes evaluated here. Fig. 8 shows visual features generated by the stereo system and a selection of generated actions. Fig. 9 shows a simple plate structure for which the outer contour is generated since the object is

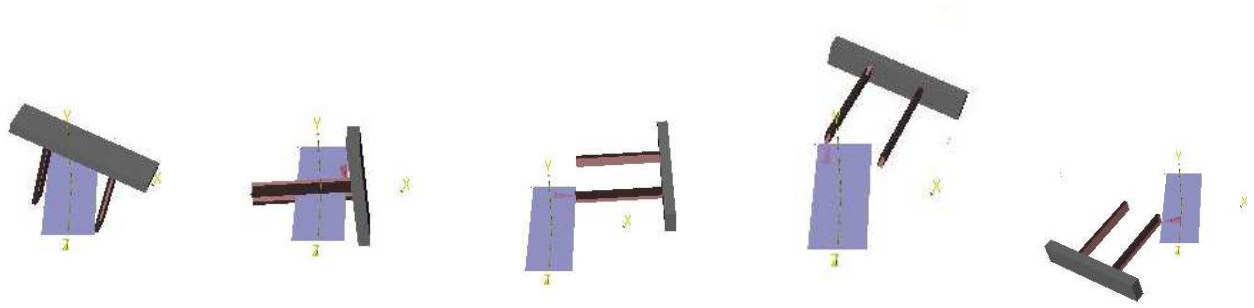


Fig. 9. Examples of tested grasps on a plate (from left): successful grasp using EGA5, and a few early failures using EGA1, EGA3 and EGA5, res5respectively.

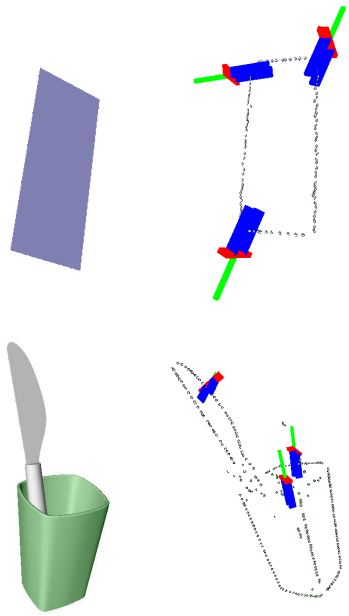


Fig. 8. Two example scenes designed for testing and a selection of the generated actions.

Scene	gr	pl+gr	col+gr	gr+pl+col
Plane	70% (7/10)	83% (5/6)	57% (4/7)	100% (5/5)
Cup	26% (17/66)	38% (14/37)	27% (13/49)	33% (8/24)
Cup/Kn	31% (14/45)	28% (9/32)	31% (11/35)	25% (5/20)
3 objects	8% (33/434)	9% (9/98)	13% (18/139)	15% (8/53)

TABLE I

EXPERIMENTAL EVALUATION OF THE GRASP SUCCESS RATE WHERE THE FOLLOWING NOTATION IS USED: PL (CO-PLANARITY), GR (GROUPING), CL (CO-COLORITY) AND (SUCCESFULL/TESTED) GRASPS.

homogeneous in texture. Fig. 10 shows a scene with a single, but a more complex object than the previous one. Fig. 11 shows two scenes with two (cup and knife) and three objects (box, cup and bottle9.

On each of the scene, after the spatial primitives have been extracted, elementary actions shown in Fig. 5 are tested. There are few reasons for which a certain grasp may fail:

- The system does not have the knowledge of whether the object is hollow or not, so testing EGA2 will results with a collision and thus failure.
- Since no surface is reconstructed, EGA1 will fail for hollow objects which are grasped from “below”.
- If the hand, during the approach, detects a collision on one of the fingers, the grasping process is stopped. In reality, this grasp may happen to be successful anyway if the object is moved so that it is centered between the fingers.

Table I summarizes the results for the generated success rate regarding a number of successful grasps given no knowledge of the object shape. We note that the results are a summary of an extensive experimental evaluation since, given different types and combinations of spatial primitives all generated actions had to be evaluated. It can be seen that for a scene of low complexity (plate) the average number of successful grasps is close to 80%. For more complex scenes this number is dependant on the number and type of objects. It is also important to note not only the percentage but the number of evaluated grasps. Although, in some cases, the success rate is lower when primitives are integrated, there are much fewer hypotheses tested. These results should also be considered together with the results presented in Table II where we show how the integration of grouping, co-colority and co-planarity affects the number of generated hypotheses (affordances). Another thing to point out related to Table I is that most of the unsuccessful grasps happened due to an “early failure” such as that a contact was detected before the grasp was executed. Again, this failure may in some cases result with a successful grasp anyway. Another big source of failure was that there was nothing to lift, i.e. EGA3 could not have been applied.

VII. CONCLUSIONS

Robots should be able to extract more knowledge through their interaction with the environment. The basis for this interaction should not be a detailed model of the environment and lots of *a-priori* knowledge but the robot should be engaged in an exploration process through which it can generate more knowledge and more complex representations. In this paper, we have presented one of the building blocks necessary in such a system.

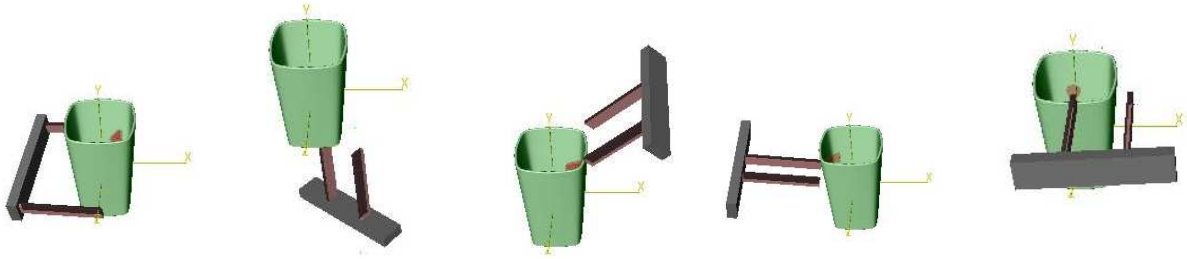


Fig. 10. Examples of tested grasps on a cup (from left): a successful grasp using EGA1, and a few early failures using EGA1, EGA1, EGA2 and EGA3, respectively.

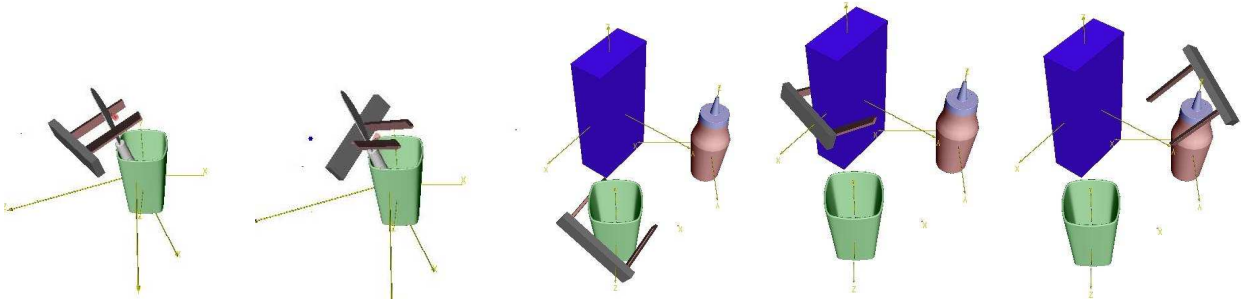


Fig. 11. Examples of successful grasps with two and three objects.

Scene	(no gr)	(no gr)+pl	(no gr)+col	(no gr)+pl+col
Plane	46 224	35 608	38 512	30 224
Cup	172 224	96 112	89 392	56 120
Cup/knife	269 360	140 920	139 136	79 104
3 objects	927 368	303 960	315 336	166 008

Scene	gr	gr+pl	gr+col	gr+pl+col
Plane	80	48	56	40
Cup	528	296	392	192
Cup/knife	360	256	280	160
3 objects	3472	784	1112	424

TABLE II

THE NUMBER OF GENERATED ACTION HYPOTHESES WHERE THE FOLLOWING NOTATION IS USED: NO GR (NO GROUPING), PL (CO-PLANARITY), GR (GROUPING), CL (CO-COLORITY).

In particular, we have designed an early grasping system, based on a set of innate reflexes and knowledge about its embodiment. We relied on 3D information based on primitives extracted online and showed how the structural richness of primitives can be used for an efficient reduction of grasping hypotheses while keeping relevant ones. Rather than dealing with high quality grasps on a constrained set of known objects, we have demonstrated that the system is able of generating a certain percentage of successful grasps on arbitrary objects. This is important for our future research that will develop complex learning schemes aiming at more sophisticated grasping strategies and knowledge representation.

ACKNOWLEDGMENT

This work has been supported by EU through the project PACO-PLUS, FP6-2004-IST-4-27657.

REFERENCES

[1] A. Stoytchev, "Behavior-Grounded Representation of Tool Affordances," in *IEEE International Conference on Robotics and Automation*, pp. 3060–3065, 2005.

[2] P. Fitzpatrick, G. Metta, L. Natale, S. Rao, and G. Sandini, "Learning About Objects Through Action - Initial Steps Towards Artificial Cognition," in *IEEE International Conference on Robotics and Automation*, pp. 3140–3145, 2003.

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

[4] N. Krüger, M. Lappe, and F. Wörgötter, "Biologically motivated multi-modal processing of visual primitives," *The Interdisciplinary Journal of Artificial Intelligence and the Simulation of Behaviour*, vol. 1, no. 5, pp. 417–428, 2004.

[5] N. Pugeault, F. Wörgötter, and N. Krüger, "Multi-modal scene reconstruction using perceptual grouping constraints," in *Proceedings of the 5th IEEE Computer Society Workshop on Perceptual Organization in Computer Vision, (in conjunction with IEEE CVPR 2006)*, 2006.

[6] A. T. Miller and P. Allen, "Graspit!: A versatile simulator for grasping analysis," in *ASME International Mechanical Engineering Congress and Exposition*, 2000.

[7] I. Kamon, T. Flash, and S. Edelman, "Learning Visually Guided Grasping: A Test Case in Sensorimotor Learning," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 28, no. 3, pp. 266–276, 1998.

[8] B. Rössler, J. Zhang, and A. Knoll, "Visual Guided Grasping of Aggregates using Self-Valuing Learning," in *IEEE International Conference on Robotics and Automation*, pp. 3912– 3917, 2002.

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

[10] D. Ding, Y.-H. Liu, and S. Wang, "Computing 3-d optimal formclosure grasps," in *IEEE International Conference on Robotics and Automation*, pp. 3573 – 3578, 2000.

[11] A. Hauck, J. Rüttinger, M. Sorg, and G. Färber, "Visual Determination of 3D Grasping Points on Unknown Objects with a Binocular Camera System," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 272–278, 1999.

[12] M. Rutishauser and M. Stricker, "Searching for Grasping Opportunities on Unmodeled 3D Objects," in *British Machine Vision Conference*, pp. 277 – 286, 1995.

[13] A. Morales, G. Recatalá, P. J. Sanz, and Á. P. del Pobil, "Heuristic Vision-Based Computation of Planar Antipodal Grasps on Unknown Objects," in *IEEE International Conference on Robotics and Automation*, pp. 583– 588, 2001.

[14] A. T. Miller, S. Knoop, and H. I. C. P.K. Allen, "Automatic grasp planning using shape primitives," in *IEEE International Conference on Robotics and Automation*, pp. 1824–1829, 2003.

[15] N. S. Pollard, "Closure and quality equivalence for efficient synthesis of grasps from examples," *International Journal of Robotic Research*, vol. 23, no. 6, pp. 595–613, 2004.

[16] A. Morales, E. Chinellato, A. H. Fagg, and A. del Pobil, "Using experience for assessing grasp reliability," *International Journal of Humanoid Robotics*, vol. 1, no. 4, pp. 671–691, 2004.

[17] R. Platt Jr, A. H. Fagg, and R. A. Grupen, "Extending fingertip grasping to whole body grasping," in *International Conference on Robotics and Automation*, pp. 2677 – 2682, 2003.

[18] N. S. Pollard, "Parallel methods for synthesizing whole-hand grasps from generalized prototypes," *PhD thesis, Dept. of Electrical Engineering and Computer Science, Massachusetts Institute of Technology*, 1994.

[19] http://www.barrett.com/robot/products/hand/handfram.htm.

[20] N. Krüger and F. Wörgötter, "Multi-modal primitives as functional models of hyper-columns and their use for contextual integration," *International Symposium on Brain, Vision and Artificial Intelligence, Lecture Notes in Computer Science, Springer, LNCS 3704*, pp. 157–166, 2005.

[21] N. Krüger and M. Felsberg, "An explicit and compact coding of geometric and structural information applied to stereo matching," *Pattern Recognition Letters*, vol. 25, no. 8, pp. 849–863, 2004.