International Journal of Humanoid Robotics

Birth of the Object: Detection of Objectness and Extraction of Object Shape through Object Action Complexes

Dirk Kraft
University of Southern Denmark, Odense, Denmark, kraft@mmmi.sdu.dk

Nicolas Pugeault
University of Edinburgh, Edinburgh, UK and
University of Southern Denmark, Odense, Denmark, npugeaul@inf.ed.ac.uk

Emre Başeski, Mila Popović
University of Southern Denmark, Odense, Denmark
emre@mmmi.sdu.dk, mipop05@student.sdu.dk

Danica Kragić
Royal Institute of Technology, Stockholm, Sweden, dani@kth.se

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

Norbert Krüger
University of Southern Denmark, Odense, Denmark, norbert@mmmi.sdu.dk

We describe a process in which the segmentation of objects as well as the extraction of the object shape becomes realized through active exploration of a robot vision system. In the exploration process, two behavioral modules that link robot actions to the visual and haptic perception of objects interact. First, by making use of an object independent grasping mechanism, physical control over potential objects can be gained. Having evaluated the initial grasping mechanism as being successful, a second behavior extracts the object shape by making use of prediction based on the motion induced by the robot. This also leads to the concept of an 'object' as a set of features that change predictably over different frames.

The system is equipped with a certain degree of generic prior knowledge about the world in terms of a sophisticated visual feature extraction process in an early cognitive vision system, knowledge about its own embodiment as well as knowledge about geometric relationships such as rigid body motion. This prior knowledge allows for the extraction of representations that are semantically richer compared to many other approaches.

Keywords: Early Cognitive Vision, Grasping, Exploration

1. Introduction

According to Gibson[1] an object is characterized by three properties: It

O1 has a certain minimal and maximal size related to the body of an agent,
O2 shows temporal stability, and
O3 is manipulatable by the agent.

Note that all these three properties are defined in relation to the agent (even temporal stability (O2) is relative to the agents lifetime span). Hence, no general agent independent criterion can be given. For an adult, a sofa certainly fulfills all three properties but for a fly, a sofa is more a surface than an object.

The detection of 'objectness' according to the three properties described above is not a trivial task. When observing a scene, usually in a visual system, a number of local features become extracted for which it is unclear whether and to which object they correspond to. Actually, property O3 can only be tested by acting on the scene in case that no prior object knowledge is available.

In many artificial systems, in particular in the context of robotics, the object shape is given by a CAD representation a priori and is then used for object identification and pose estimation (see, e.g., Lowe[2]). However, CAD representations are not available in a general context and hence for any cognitive system, it is an important prerequisite that it is able to learn object representations from experience.

In this paper, we address both problems: We introduce a procedure in which the objectness becomes detected based on the three Gibsonian criteria mentioned above. In addition, the object shape becomes extracted by making use of the coherence of motion induced by the agent after having achieved physical control over something that might turn out to become an object.

Our approach is making use of the concept of Object Action Complexes (OACs) where we assume that objects and actions (here the 'grasping action' and controlled object movement) are inseparably intertwined. Hence, the intention of performing a grasp, the actual attempt to grasp and the evaluation of its success as well as a controlled movement of the object in case of a successful grasp will let the 'objectness' as well as a representation of the object's shape emerge as the consequence of the actions of the cognitive agent[a].

It is worth noting that both aspects, achieving physical control over a *thing*[b] as well as the extraction of object shape is based on a significant amount of prior knowledge, which however is much more generic than a CAD model of an object. More specifically, this prior consists of the system's knowledge about

[a]We note that this extends the notion of 'affordances' by Gibson. According to Gibson: Objects afford actions. While this remains true, it is also — in our hands — the case that an action defines an object. For example the action of drinking defines a cup, where the action of 'placing on top' makes the same (!) thing a pedestal (an upside down cup).
[b]We denote with 'thing' something that causes the extraction of a visual feature but which is not yet characterized as an object since it could be for example also something fixed in the workspace of the robot and hence does not fulfill condition O3.

1) its own body in terms of the shape, the degrees of freedom and the current joint configuration of the robot arm as well as the relative position of the stereo camera system and the robot co-ordinate system,
2) a developed early cognitive system[3,4] that extracts local multi-modal symbolic descriptors (see Fig. 1(a–e)), in the following called primitives, and relations defined upon these primitives expressing statistical and deterministic properties of visual information (see Fig. 2).
3) two behavior modules in terms of two OACs:

   B1 An object independent 'grasping reflex' leads in some cases to successful grasping of potential objects (Fig. 1e shows the end-effector's pose for one successful grasp). Note that here it is less important to have a high success-rate of grasping attempts but that is is more important that a success is actually *measurable* and that it then triggers a second exploration mechanism (see B2).
   The 'grasping reflex' is based on three semantic relations defined within the early cognitive vision system: First, co-planarity of descriptors indicate surfaces and by that possible grasping options. The co-planarity relation is enhanced by a co-linearity and co-colority relation to further enhance the success rate of the 'grasping reflex'.

   B2 After a successful grasp an accumulation module explores the object by looking at different views of the object (see Fig. 1(f,g)) and accumulating this information to determine the objectness of the thing as well as to extract the shape of the object (Fig. 1(h)). This accumulation module is based on prediction based on a rigid body motion relation between primitives. Having gained physical control over an object by the grasping reflex allows for inducing a rigid body motion on the object and by that the object (its objectness as well as its shape) can be characterized by the set of visual descriptors changing according to the induced motion.

The idea of taking advantage of active components for vision is in the spirit of active vision research[5,6]. The grounding of vision in cognitive agents has been addressed for example by a number of groups in the context of grasping[7,8] as well as robot navigation[9].

The work of Fitzpatrick and Metta[7] is the most related one to our approach since the overall goal as well as the hardware set up is similar: Finding out about the relations of actions and objects by exploration using a stereo system combined with a grasping device. We see the main distinguishing feature of this work to our approach in the amount of pre-structure we use. For example, we assume a much more sophisticated vision system that covers multiple visual modalities in a condensed form as well as visual relations defined upon them. This allows us to operate in a highly structured feature space where, instead of pixel-wise representations, we can operate on local symbols for which we can predict changes not only of position but also other feature attributes such as orientation and color. Furthermore, the use

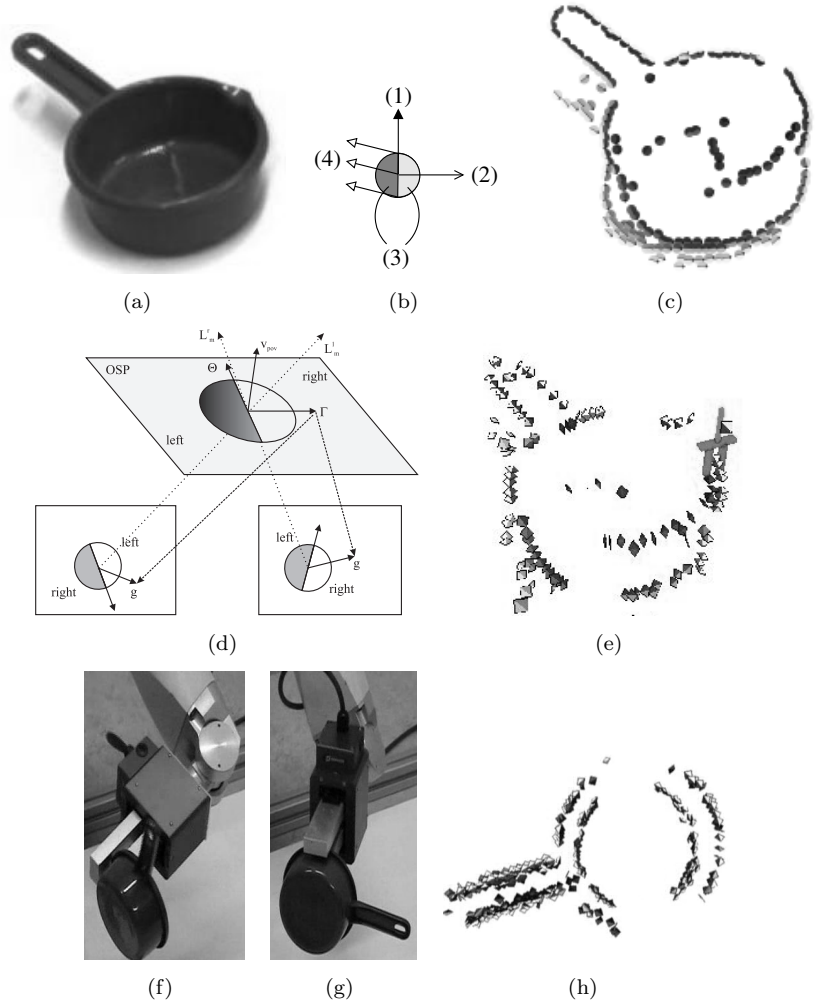


Fig. 1. Overview of the system. (a) Image of the scene as viewed by the left camera at the first frame. (b) Symbolic representation of a primitive wherein (1) shows the orientation, (2) the phase, (3) the color, (4) the optic flow of the primitive. (c) 2D primitives extracted at one object in the scene from (a). (d) Illustration of the reconstruction of a 3D primitive from a stereo pair of 2D primitives. (e) 3D primitives reconstructed from the scene and one grasping hypothesis. (f–g) Two views of robot rotating the grasped object to build its 3D representation. (h) The learned 3D representation of the object.

of a very precise industrial robot allows for a precise generation of changes exploited for the extraction of the 3D shape of the object.

It is not clear what exact prior knowledge can be assumed in the human system. However, there exist strong indications for an innate concept of 3D space as well as for sophisticated feature extraction mechanisms being in place very early

in visual experience. For a discussion of this issue see for example Kellmann and Arterberry[10]. The question of prior knowledge in the context of depth perception and possible consequences for the design of artificial systems is described in Krüger and Wörgötter[11].

Similar to Fitzpatrick and Metta[7], we assume first 'reflex-like' actions that trigger exploration. However, since in our system the robot knows about its body and the 3D geometry of the world and since the arm can be controlled more precisely, these reflexes can make use of more complex visual events. As a consequence we can make use of having physical control over the object and therefore extract rather precise 3D information (in addition to the appearance based information coded in the primitives).

Modayil and Kuipers[9] addressed the problem of detection of objectness and the extraction of object shape in the context of a mobile robot using laser information. Here also motion information (in terms of the odometry of the mobile robot) is used to formulate predictions. In this way, they were able to extract a top-view of the 3D shape of the object however only in terms of geometric information and only in terms of a 2D projection to the ground floor.

The paper is organized as following: In Section 2 the early cognitive vision system is briefly described. In Section 3 and 4 we give a description of the two sub-modules, i.e., the grasping reflex and the accumulation scheme. Sub-aspects of the work have been presented at two workshops[12,13].

2. An Early Cognitive Vision System

In this section, we introduce the visual system in which the detection of 'objectness' as well as the acquisition of the object representation is taking place. The system is characterized by rather structured prior knowledge: First, a scene representation is computed in terms of local symbolic descriptors (in the following called primitives) covering different visual modalities as well as 2D and 3D aspects of visual data (Section 2.1). Second, there are relations defined upon the symbolic descriptors that cover spatial and temporal dependencies as briefly described in Section 2.2. It is only the use of this prior knowledge that allows for the formulation of the two OACs described in Sections 3 and 4.

2.1. *Multi-modal primitives as local scene descriptors*

In this work we use local, multi-modal contour descriptors hereafter called *primitives*[3,4] (see Fig. 1). These primitives give a semantically meaningful description of a local image patch in terms of position as well as the visual modalities orientation, color and phase. The importance of such a semantic grounding of features for a general purpose vision front-end, and the relevance of edge-like structures for this purposes was discussed, e.g., by Elder[14].

The primitives are extracted sparsely at locations in the image which are most likely to contain edges. The sparseness is assured using a classical winner-take-all

operation, ensuring that the generative patches of the primitives do not overlap. Each primitive encodes the image information contained by a local image patch. Multi-modal information is gathered from this image patch, including the position $\boldsymbol{x}$ of the center of the patch, the orientation θ of the edge, the phase ω of the signal at this point, the color $\boldsymbol{c}$ sampled over the image patch on both sides of the edge, the local optical flow $\boldsymbol{f}$ and the size of the patch ρ. Consequently a local image patch is described by the following multi-modal vector:

$$\pi = (\boldsymbol{x}, \theta, \omega, \boldsymbol{c}, \boldsymbol{f}, \rho)^T, \tag{1}$$

that we will name *2D primitive* in the following. The primitive extraction process is illustrated in Fig. 1.

In a stereo scenario, *3D primitives* can be computed from correspondences of 2D primitives (Fig. 1)

$$\Pi = (\boldsymbol{X}, \boldsymbol{\Theta}, \Omega, \boldsymbol{C})^T, \tag{2}$$

where $\boldsymbol{X}$ is the position in space, $\boldsymbol{\Theta}$ is the 3D orientation, Ω is the phase of the contour, and $\boldsymbol{C}$ is the color on both sides of the contour. For details see Pugeault[15].

2.2. *Perceptual relations between primitives*

The sparseness of the primitives allows for the formulation of four *structural relations* between primitives that are crucial in our context since they allow us to relate feature constellations to grasping actions (in the first OAC in Section 3) or visual percepts in consecutive frames (in the second OAC described in Section 4). See Kalkan et al.[16] for more details.

Co-planarity: Two spatial primitives Π_i and Π_j are co-planar iff their orientation vectors lie on the same plane. The co-planarity relation is illustrated in Fig. 2(b). In the context of the grasping reflex described in Section 3, grasping actions become associated to the plane spanned by co-planar primitives.

Collinear grouping (i.e., collinearity): Two 3D 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 the 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. Collinearity of two primitives is illustrated in Fig. 2(a).

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. In Fig. 2(c) a pair of co-color and non co-color primitives are shown. Testing for collinearity and co-colority help to reduce the number of generated grasping hypotheses (see Section 3.2).

Rigid body motion: The change of position and orientation induced by a rigid body motion between two frames at time t and $t+1$ ($\Pi^{t+1} = \mathrm{RBM}(\Pi^t)$) can be computed analytically[17], phase and color can be approximated to be constant.

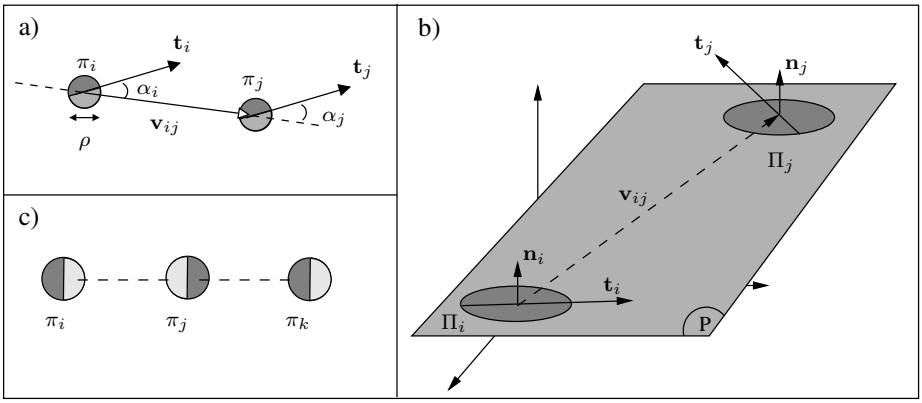


Fig. 2. Illustration of the relations between a pair of primitives. (a) Collinearity of two 2D primitives π_i and π_j. (b) Co-planarity of two 3D primitives Π_i and Π_j. (c) Co-colority of three 2D primitives π_i, π_j and π_k. In this case, π_i and π_j are co-color, so are π_i and π_k; however, π_j and π_k are not co-color.

3. Grasping Reflex

In this section, we describe the first OAC that leads to a physical control over objects. Note that a high success rate is not important in this context, but more that the success can be evaluated by haptic feedback which then gives indications to proceed with another OAC described in Section 4.

3.1. *Elementary grasping actions associated to co-planar primitives*

Coplanar relationships between visual primitives suggest different graspable planes. Fig. 3(a) shows a set of spatial primitives on two different contours l_i and l_j with co-planarity, co-colority and collinearity relations.

Four elementary grasping action (EGA) types will be considered as shown in Fig. 3(b-e). EGA type 1 (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. EGA type 2 (EGA2) is an 'inverted' grasp using the inside of two edges with approach along the surface normal. EGA type 3 (EGA3) is a 'pinch' grasp on a single edge with approach direction perpendicular to the surface normal. EGA type 4 (EGA4) is a wide grasp making contact on two separate edges with approach direction along the surface normal.

EGAs are 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$(x, y, z, k, l, m, \delta)$ where $\mathbf{p} = [x, y, z]$ is the position of the gripper 'center' according to Fig. 3(f); k, l, m are the roll, pitch and yaw angles of the vector $\mathbf{n}$; and δ is the gripper opening, see Fig. 3(f). Note that the gripper 'center' is placed in the 'middle' of the gripper.

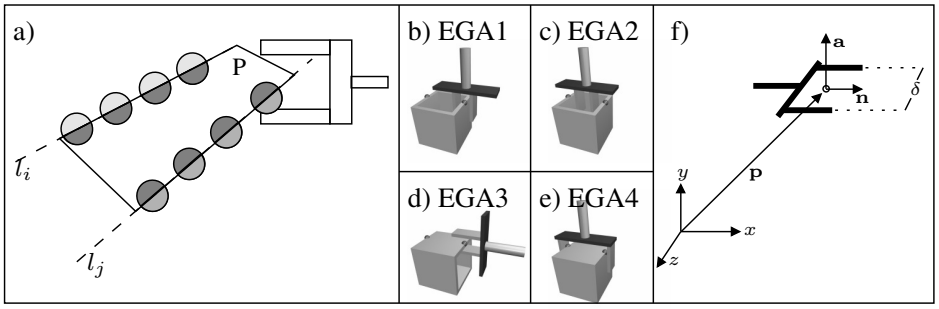


Fig. 3. (a) A set of spatial primitives on two different contours l_i and l_j that have co-planarity, co-colority and collinearity relations; a plane P defined by the co-planarity of the spatial primitives and an example grasp suggested by the plane. (b–e) Elementary grasping action types, EGA1, EGA2, EGA3 and EGA4 respectively. Please note, besides the two primitives (marked as spheres) defining the concrete EGA only the surfaces touched by the gripper are needed for the grasp to be successful. The cube form given here is used to illustrate the differences in the different EGA types and their applicability. (f) Parameterization of EGAs.

These grasp parameters are computed from coplanar pairs of 3D-primitives. Let $\Gamma = \{\Pi_1, \Pi_2\}$ be a primitive pair for which the coplanar relationship is fulfilled. Let Γ_i be the i-th pair and $\mathbf{p}$ the plane defined by the coplanar relationship of the primitives of Γ_i. Let $\Lambda(\Pi)$ be the position of Π and $\Theta(\Pi)$ be the orientation of Π. The parameterization of the EGAs is given with the gripper normal $\mathbf{n}$ and the normal $\mathbf{a}$ of the surface between the two fingers as illustrated in Fig. 3(f). From this, the yaw, pitch and roll angles can be easily computed. For example for EGA1, there will be two possible parameter sets given the primitive pair $\Gamma = \{\Pi_1, \Pi_2\}$. The parameterization is as follows:

$$\begin{aligned} \mathbf{p}_{\text{gripper}} &= \Lambda(\Pi_i), \\ \mathbf{n} &= \nabla(\mathbf{p}), \\ \mathbf{a} &= \mathbf{perp}_{\mathbf{n}}(\Theta(\Pi_i))/ \parallel \mathbf{perp}_{\mathbf{n}}(\Theta(\Pi_i)) \parallel \text{ for } i = 1, 2, \end{aligned} \tag{3}$$

where $\nabla(\mathbf{p})$ is the normal of the plane $\mathbf{p}$ and $\mathbf{perp}_{\mathbf{u}}(\mathbf{a})$ is the projection of $\mathbf{a}$ perpendicular to $\mathbf{u}$. The details of how the other EGAs are computed can be found in Aarno et al.[12].

The main motivation for choosing these grasps is that they represent the simplest possible two fingered grasps humans commonly use which can also be simulated on our robot system. The result of applying the EGAs can be evaluated by the information given by the gripper (Schunk, PT-AP 70) which gives the distance between the two jaws at each instance of time.

For EGA1, EGA3 and EGA4, a failed grasp can be detected by the fact that the gripper is completely closed. 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. EGA2 fails if the gripper is fully opened, meaning that no contact was

made with the object. If none of the above situations is encountered the EGA is considered to be successful.

3.2. *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 six different grasps from the four different categories, it is obvious that the number of suggested actions must be further constrained. 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. Aarno et al.[12] quantified the reduction in EGA hypotheses using collinearity and co-colority in a simulation environment, showing that the number of EGAs can be reduced systematically.

3.3. *Experimental evaluation*

To evaluate the grasping reflex we made experiments within the simulation environment GraspIt[18] and with a real scene. In the GraspIt environment, we evaluated success rates in scenes of different complexity (see Fig. 4(a–d) for a number of successful grasps on two scenes). The success rate was dependent on the scene complexity, ranging from appr. 90%[c] in the case of a simple plane (see Fig. 4(a,b)), to around 25% for scenes of larger complexity (Fig. 4(c,d)).

We then evaluated the exploration strategy on a real scene (see Fig. 4(e)). After reconstructing 3D-primitives from stereo images (Fig. 4(h)), 912 EGAs were generated. However, in a real set-up there are additional constraints such as the definition of a region of interest[d] and the fact that not all EGAs are actually performable due to limited workspace[e]. In addition, grasps leading to collisions with the floor or the wall need to be eliminated. Table 1 shows the effects of the reductions.

In a full exploration sequence, the system attempts to perform one of the 50 remaining EGAs. A failure to grasp an object generally causes changes in the scene, and the whole sequence of capturing images, generating and reducing EGAs would

[c]A success is counted when one of the six EGAs (two instantiations of each EGA1 and EGA3, one of EGA2 and EGA4) generated by a primitive pair has been performed successfully

[d]The region of interest serves two purposes: 1. It represents a computationally cheap way to remove EGAs that would be reduced by the later, more expensive reachability check. 2. It prevents grasping attempts in regions in which no objects should be placed. In our concrete setup, the region of interest is defined as a cube in front of the robot.

[e]Note that the workspace needs to be defined in terms of a 6D pose and that even when a 3D point is reachable, it is not certain that the desired end-effector orientation can be achieved at this point.

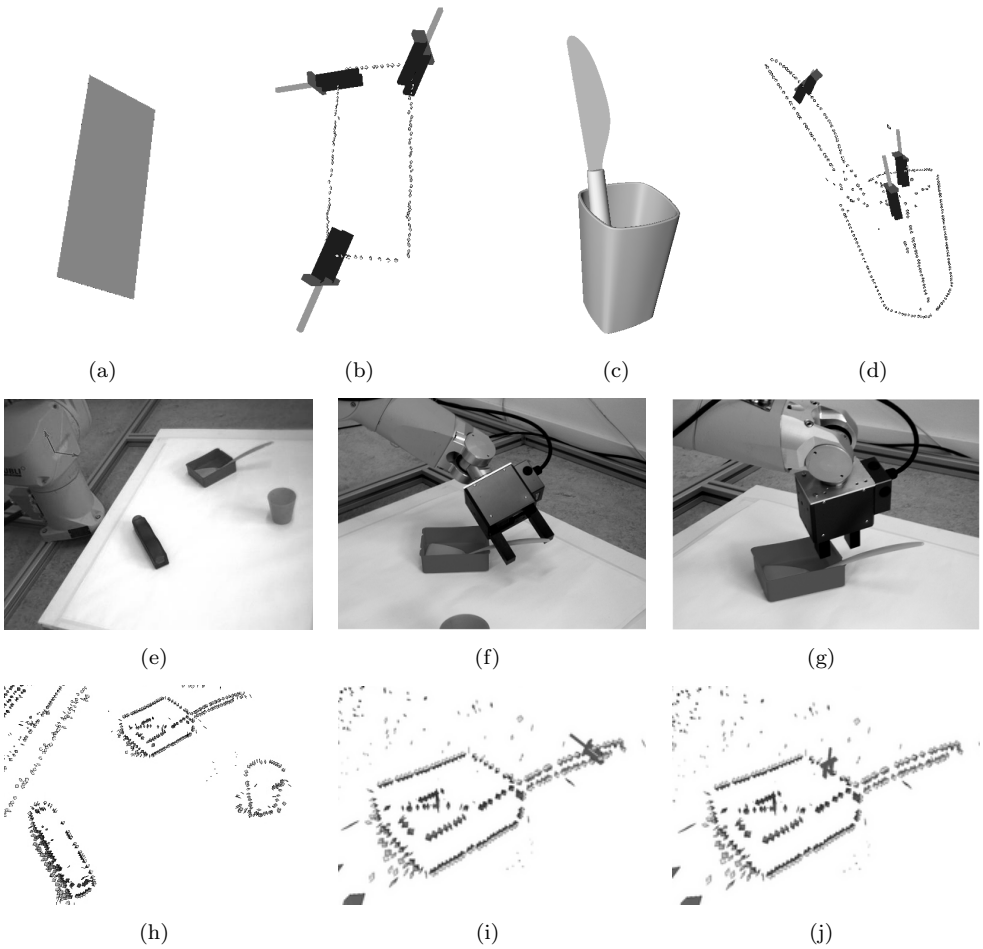


Fig. 4. Evaluation of the grasping reflex in simulation (a–d) and in a real robot environment (e–j). (a,c) Artificial grasping scenes. (b,d) Selected grasping hypotheses generated for the scenes shown in (a,c). (e) The view from the left camera for the real environment. Origin and orientation of the world coordinate frame are illustrated in top left corner. (h) Extracted 3D primitives for the whole scene (see (e)) displayed in the visualization environment. (f,g) The robot arm executing successful grasps. (i,j) The grasps from (f,g) shown in the 3D visualization environment (enlarged).

Table 1. The results of applying reductions to the initial set of EGAs.

Reductions:	Initial number of EGAs	remaining	relative reduction
to region of interest	912	228	75.0%
to reachable configurations	123	105	53.9%
collision free (floor)	105	50	52.4%

be repeated. However, for the purpose of evaluating the whole set of proposed EGAs for a single scene, the objects in this experiment were placed at their original position after each attempted grasp.

In the specific scenario shown in Fig. 4(e), three out of the four objects could be grasped by the reflex. Out of 50 grasps, 7 lead to physical control over objects. In one case, the contact area was too small, leading to an unstable grip, and the accumulation module (see Section 4.1) could not be applied.

4. Detection of Objectness and Object Shape

Having achieved physical control over an object, measured by the distance between the gripper's jaws after closing or opening (in case of EGA2), a second OAC is triggered that makes use of the additional capability of the agent: actively manipulating the object.

If the object's motion within the scene is known, then the relation between this object's features in two subsequent frames becomes deterministic (excluding the usual problems of occlusion, sampling, etc.). This means that a 3D-primitive that is present in one frame is subject to a transformation that is fully determined by the object's motion: generally a change of 3D position and 3D orientation.[f] If we assume that the motion between consecutive frames is reasonably small then a contour will not appear or disappear unpredictably, but will have a life-span in the representation, between the moment it enters the field of view and the moment it leaves it. Assuming having a fully calibrated system and having physical control over the object (as gained by the first OAC described in Section 3), we can compute the 3D-primitives' change in camera coordinates.

These predictions are relevant in different contexts:

Establishment of objectness: The objectness of a set of features is characterized by the fact that they all move according to the robot's motion. This property is discussed in the context of a grounded AI planning system in Geib et al.[19].

Segmentation: The system segments the object from the rest of the scene using its predicted motion.

Disambiguation: Erroneous 3D-primitives can be characterized (and eliminated) by inconsistent motion according to the predictions (see also Krüger et al.[20]).

Learning the object model: A full 3D model of the object can be extracted by merging different $2\frac{1}{2}$D views created by the motion of the end effector.

[f] We neglect the effects of lighting and reflection, and assume that phase and color stay constant.

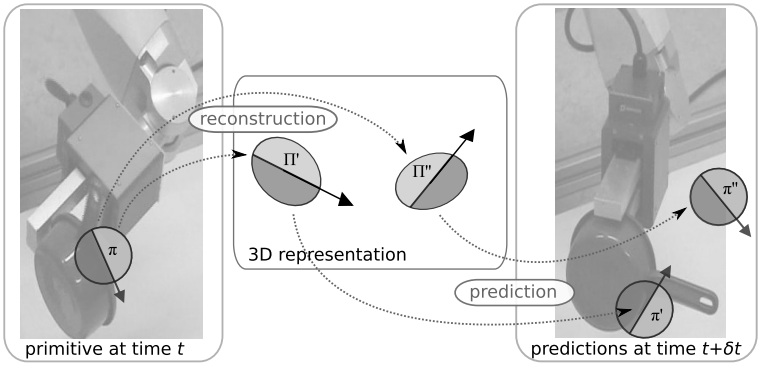


Fig. 5. Example of the accumulation of a primitive (see text).

4.1. *Making predictions from the robot motion*

If we consider a 3D-primitive $\Pi_i^t \in \boldsymbol{\mathcal{S}}_t$ describing an object's contour at time instant t, and assume that the object's motion is known between the two instants t and $t + \Delta t$, then we can predict this primitive's position at time $t + \Delta t$.

The projection of 3D-primitives to the image domain predicts where 2D-primitives should be extracted from each camera's image at time $t + \Delta t$. It is then possible to assess the correctness of a reconstructed 3D-primitive by how reliably it is confirmed by subsequently extracted 2D-primitives.

This prediction/verification process is illustrated in Fig. 5. The left image is taken from a scene at time t; the right image is taken from the same scene, at a later time $t + \delta t$. Assuming that a primitive π is extracted at time t, and lead to two distinct, mutually exclusive, putative 3D reconstructions Π' and Π''. If the object that π describes is subjected to a known motion $M_{t \to t+\delta t}$, then this motion knowledge allows for making predictions on where in the image this primitive should manifest itself at time $t+\delta t$. Mutually exclusive putative 3D-primitives (Π' and Π'') will lead to distinct predictions (π' and π''). When confronting these predictions with the new image at time $t + \delta t$, and the primitives extracted from it, it becomes apparent that π' is confirmed by the newly available information whereas π'' is contradicted, thereby revealing the erroneousness of the hypothesis Π''. Therefore, Π'' is discarded from the representation and thus the ambiguity is reduced.

We then propose to use these predictions to re-evaluate 3D-primitives' confidence depending on their resilience over time. This is justified by the continuity assumption, which states that 1) scene's objects or contours should not appear and disappear abruptly from the field of view (FoV) but move in and out gracefully according to the estimated ego-motion; and 2) a contour's position and orientation at any point in time is fully determined by the knowledge of its position at a previous instant in time and of its motion since.

Consider a primitive Π_i, predicting a primitive $\hat{\Pi}_i^t$ at time t. We write the fact

that this prediction is confirmed by the images at time time t as $\mu_t(\hat{\Pi}_i) = 1$; and the fact that it is not confirmed (i.e., there is no 2D-primitive extracted at time t that is similar to the projection of $\hat{\Pi}_i^t$ on the image plane) as $\mu_t(\hat{\Pi}_i) = 0$. By extension, we code the resilience a primitive Π_i, from its apparition at time 0 until time t as the binary vector:

$$\boldsymbol{\mu}(\Pi_i) = \left(\mu_t(\hat{\Pi}_i), \mu_{t-1}(\hat{\Pi}_i), \cdots, \mu_0(\hat{\Pi}_i)\right)^T. \tag{4}$$

We then apply Bayes formula to evaluate the posterior likelihood that a 3D-primitive is correct knowing its resilience vector:

$$p\left(\Pi_i|\boldsymbol{\mu}(\Pi_i)\right) = \frac{p\left(\boldsymbol{\mu}(\Pi_i)|\Pi\right)p\left(\Pi\right)}{p\left(\boldsymbol{\mu}(\Pi_i)|\Pi\right)p\left(\Pi\right) + p\left(\boldsymbol{\mu}(\Pi_i)|\bar{\Pi}\right)p\left(\bar{\Pi}\right)}. \tag{5}$$

In this formula, Π and $\bar{\Pi}$ are correct and erroneous primitives, respectively. The quantities $p\left(\Pi\right)$ and $p\left(\bar{\Pi}\right)$ are the prior likelihoods for a 3D-primitive to be correct and erroneous. The quantity $p\left(\boldsymbol{\mu}(\hat{\Pi}_i)|\Pi\right)$ (resp. $p\left(\boldsymbol{\mu}(\Pi_i)|\bar{\Pi}\right)$) expresses the probability of occurrence of a resilience vector $\boldsymbol{\mu}(\Pi_i)$ for a correct (resp. erroneous) primitive Π_i.

Furthermore, if we assume independence between the matches $\mu_t(\hat{\Pi}_i)$, then for a primitive Π_i that exists since n iterations and has been matched successfully m times, we have the following relation:

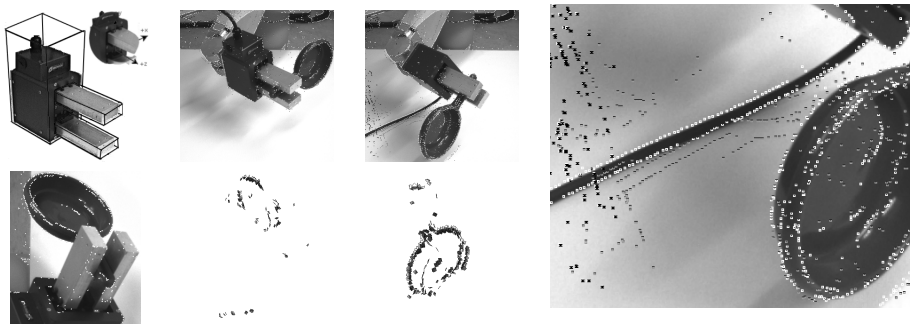
$$\begin{aligned} p\left(\boldsymbol{\mu}(\hat{\Pi}_i)|\Pi\right) &= \textstyle\prod_t p\left(\mu_t(\hat{\Pi}_i)|\Pi\right) \\ &= p\left(\mu_t(\hat{\Pi}_i) = 1|\Pi\right)^m p\left(\mu_t(\hat{\Pi}_i) = 0|\Pi\right)^{n-m}. \end{aligned} \tag{6}$$

In this case the probabilities for μ_t are equiprobable for all t, and therefore if we define the quantities $\alpha = p\left(\Pi\right)$, $\beta = p\left(\mu_t(\hat{\Pi}) = 1|\Pi\right)$ and $\gamma = p\left(\mu_t(\hat{\Pi}) = 1|\bar{\Pi}\right)$ then we can rewrite Eq. (5) as follows:

$$p\left(\Pi_i|\bar{\boldsymbol{\mu}}(\hat{\Pi}_i)\right) = \frac{\beta^m(1-\beta)^{n-m}\alpha}{\beta^m(1-\beta)^{n-m}\alpha + \gamma^m(1-\gamma)^{n-m}(1-\alpha)}. \tag{7}$$

We measured these prior and conditional probabilities using a video sequence with known motion and depth ground truth obtained via range scanner. We found values of $\alpha = 0.46$, $\beta = 0.83$ and $\gamma = 0.41$. This means that, in these examples, the prior likelihood for a stereo hypothesis to be correct is 46%, the likelihood for a correct hypothesis to be confirmed is 83% whereas for an erroneous hypothesis it is of 41%. These probabilities show that Bayesian inference can be used to identify correct correspondences from erroneous ones. To stabilize the process, we will only consider the n first frames after the appearance of a new 3D-primitive. After n frames, the confidence is fixed for good. If the confidence is deemed too low at this stage, the primitive is forgotten. During our experiments $n = 5$ proved to be a suitable value.

The end-effector of the robot follows the same motion as the object. Therefore, this end-effector becomes extracted as well. Since we know the geometry of this



(a) (b) (c) (d)

Fig. 6. Birth of an object. (a) top: bounding boxes of grasper body and its fingers used to eliminate grasper features and grasper coordinate system, bottom: image with eliminated grasper features. (b)–(c) two steps in the accumulation process. Top: 2D projection of the accumulated 3D representation and newly introduced primitives, bottom: accumulated 3D representation. (d) newly introduced and accumulated primitives in detail. Note that, the primitives that are not updated are black (dominant on the left side of the image), the ones that have low confidence are grey and the high confidence ones are white (dominant in the areas close to the cable, the gripper and the object).

Fig. 7. Objects and their related accumulated representation.

end-effector (Fig. 6(a)top), we can however easily subtract it by eliminating the 3D primitives that are inside the bounding boxes that bounds the body of the gripper and its fingers. For this operation, three bounding boxes are defined in the grasper coordinate system. Fig. 6(a)bottom shows the 2D projection of the remaining primitives after the ones produced by the gripper have been eliminated.

4.2. *Experiments*

We applied the accumulation scheme to a variety of scenes where the robot arm manipulated several objects. The motion was a rotation of 5 degrees per frame. The accumulation process on one such object is illustrated in Fig. 6. The top images of Fig. 6(b,c) show the predictions at two frames. The bottom images show the 3D-primitives that were accumulated. The object representation becomes fuller over time, whereas the primitives reconstructed from other parts of the scene are discarded. Fig. 7 shows the accumulated representation for various objects. The hole in the model corresponds to the part of the object occluded by the gripper. Accumulating the representation over several distinct grasps of the objects would yield a complete representation.

5. Conclusion

We introduced a scheme in which two modules in terms of Object Action Complexes (OACs) become combined to extract world knowledge in terms of the objectness of a set of local features as well as the object shape. Although this exploration scheme is completely autonomous, we argued that there is a significant amount of prior knowledge in terms of generic properties of the world built into the system. Starting with a rather sophisticated feature extraction process covering common visual modalities, functional relations defined on those features such as co-planarity, co-linearity, basic laws about Euclidean geometry and the motion of rigid object has been exploited. Furthermore and at least of equal importance was the capability to act on the world that made this process possible. Here the embodiment of the agent is of high importance. The option to grasp and move the objects in a controlled way is rather unique to few species and with high likelihood linked to develop higher cognitive capabilities.

The work described in this paper is part of the EU project PACO-PLUS[21] which aims at a system covering different levels of cognitive processing from low-level processes as described here up to a planning AI level (see Geib et al.[19]). This work introduced describes an important module of such a cognitive system which gives information that higher levels require to start operating. First, it segments the world in objects which are the basic entities that higher level reasoning is based on. Moreover, it generates 3D object representations in a procedural way which then can be used for object identification and pose estimation (see, e.g., Lowe[2] for the use of 3D models for object recognition and Detry and Piater[22] for first steps in directly making use of the extracted representations described in this paper). By the described exploratory procedure, a natural mechanism is given that enlarges the internal world model that then can be used by higher levels for reasoning and planning.

Acknowledgements

We would like to thank Tamim Asfour, Mark Steedman, Christopher Geib and Ron Petrick for fruitful discussions. This work was conducted within the EU Cognitive Systems project PACO-PLUS[21] funded by the European Commission.

References

1. J. Gibson, *The Ecological Approach to Visual Perception* (Boston, MA: Houghton Mifflin, 1979).
2. D. Lowe, Fitting parametrized 3D–models to images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13**(5), 441–450 (1991).
3. N. Krüger, N. Pugeault and F. Wörgötter, Multi-modal primitives: Local, condensed, and semantically rich visual descriptors and the formalization of contextual information, *IEEE Transactions on Pattern Analysis and Machine Intelligence, (also available as Technical report (2007-4) of the Robotics Group, Maersk Institute, University of Southern Denmark)* (under review).
4. N. Krüger, M. Lappe and F. Wörgötter, Biologically motivated multi-modal processing of visual primitives, *Interdisciplinary Journal of Artificial Intelligence & the Simulation of Behaviour, AISB Journal* **1**(5), 417–427 (2004).
5. Y. Aloimonos, I. Weiss and A. Bandopadhay, Active vision, *International Journal of Computer Vision* **2**, 333–356 (1987).
6. R. Rao and D. Ballard, An active vision architecture based on iconic representations, *Artificial Intelligence Journal* **78**, 461–505 (1995).
7. P. Fitzpatrick and G. Metta, Grounding vision through experimental manipulation, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **361**, 2165 – 2185 (2003).
8. L. Natale, F. Orabona, G. Metta and G. Sandini, Exploring the world through grasping: A developmental approach, in *IEEE International Symposium on Computational Intelligence in Robotics and Automation* 2005, pp. 559–565.
9. J. Modayil and B. Kuipers, Bootstrap learning for object discovery, *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* **1**, 742–747 (2004).
10. P. J. Kellman and M. E. Arterberry, *The Cradle of Knowledge* (MIT-Press, 1998).
11. N. Krüger and F. Wörgötter, Statistical and deterministic regularities: Utilisation of motion and grouping in biological and artificial visual systems, *Advances in Imaging and Electron Physics* **131**, 82–147 (2004).
12. D. Aarno, J. Sommerfeld, D. Kragic, N. Pugeault, S. Kalkan, F. Wörgötter, D. Kraft and N. Krüger, Early reactive grasping with second order 3D feature relations, in *Recent Progress in Robotics; ViableRobotic Service to Human, selected papers from ICAR'07*, eds. S. Lee, I. H. Suh and M. S. Kim (Springer-Verlag Lecture Notes in Control and Information Sciences (LNCIS), 2007).
13. N. Pugeault, E. Baseski, D. Kraft, F. Wörgötter and N. Krüger, Extraction of multi–modal object representations in a robot vision system, in *International Conference on Computer Vision Theory and Applications (VISAPP)* 2007.
14. J. H. Elder, Are edges incomplete?, *International Journal of Computer Vision* **34**, 97–122 (1999).
15. N. Pugeault, Early cognitive vision: Feedback mechanisms for the disambiguation of early visual representation, PhD thesis, University of Göttingen, 2008.
16. S. Kalkan, N. Pugeault and N. Krüger, Perceptual operations and relations between 2D or 3D visual entities, Tech. Rep. 2007–3, Robotics Group, Maersk Institute, University

of Southern Denmark (2007).
17. O. Faugeras, *Three–Dimensional Computer Vision* (MIT Press, 1993).
18. A. Miller and P. Allen, GraspIt!: A versatile simulator for robotic grasping, *IEEE Robotics and Automation Magazine* **11**(4), 110–122 (2004).
19. C. Geib, K. Mourao, R. Petrick, N. Pugeault, M. Steedman, N. Krüger and F. Wörgötter, Object action complexes as an interface for planning and robot control, in *Workshop 'Toward Cognitive Humanoid Robots' at IEEE-RAS International Conference on Humanoid Robots (Humanoids 2006)* 2006.
20. N. Krüger, M. Ackermann and G. Sommer, Accumulation of object representations utilizing interaction of robot action and perception, *Knowlede Based Systems* **15**, 111–118 (2002).
21. PACO-PLUS: Perception, Action and Cognition through learning of Object-Action Complexes. EU Cognitive Systems project (IST-FP6-IP-027657), http://www.paco-plus.org/, (2006–2010).
22. R. Detry and J. Piater, Hierarchical integration of local 3D features for probabilistic pose recovery, in *Robot Manipulation: Sensing and Adapting to the Real World, (Workshop at Robotics, Science and Systems)* 2007.