Development of Object and Grasping Knowledge by Robot Exploration

Dirk Kraft, Renaud Detry, Nicolas Pugeault, Emre Başeski, Frank Guerin, Justus Piater, and Norbert Krüger

Author post-print
(c) 2010 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/ republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works.

***Abstract*—We describe a bootstrapping cognitive robot system that—mainly based on pure exploration—acquires rich object representations and associated object-specific grasp affordances. Such bootstrapping becomes possible by combining innate competences and behaviours by which the system gradually enriches its internal representations, and thereby develops an increasingly mature interpretation of the world and its ability to act within it. We compare the system's prior competences and developmental progress with human innate competences and developmental stages of infants.**

***Index Terms*—robots with development and learning skills, active exploration of environment, hardware platform for development, using robots to study development and learning**

I. Introduction

THE ability to bootstrap the learning of increasingly rich internal representations of the world is one of the crucial properties of the human cognitive system. This ability allows the system to postulate and verify predictions and, ultimately, to act purposefully—which on the highest level of representation is connected to planning [1]. Such bootstrapping processes are required because, on the one hand, it is impossible to hard-code all relevant objects and actions as well as their properties and effects (see, e.g., [2]); on the other hand, it is known that demanding learning tasks such as object learning, for recognition as well as for grasping, cannot be solved in general terms without a certain amount of prior knowledge coded into the system. The alternative 'tabula rasa' approach (as phrased by [3]) would inevitably fail because of the huge dimensionalities of the learning spaces of the actual problems [4].

The first competence a human (or human-like cognitive agent) needs to learn is to control its own body and how it is linked to its visual system. Infants develop this competence within approximately six months (as measured by reaching success [5, p. 174]). Such learning has also been successfully modelled on humanoid robots (see e.g., [6]). An important learning step that follows such body learning and vision-body calibration is object learning, making use of a rich set of visual descriptors (e.g., textures, shapes, edges) [7], as well as the refinement of reaching and grasping based on initial reaching [8, p. 38] and grasping reflexes [9] (see the discussion in Sect. IV-A). In this paper, we describe an artificial cognitive system in which a similar development takes place. More concretely, mainly driven by pure exploration and without any innate prior object or object-specific grasping knowledge, the system develops increasingly rich and mature internal representations of objects and knowledge of how to grasp them. In parallel, starting from some inchoate innate behavioural patterns, the system develops a set of increasingly complex competences up to a very premature planning ability. Note that our system does not attempt to mimic infant development in the details, but rather in its broad outline, and in particular to develop object and grasp knowledge by a bootstrapping process, where each helps the other. Since the embodiment of our system is different from infants (e.g., high precision of grasping, and 3D reconstruction through fixed and calibrated cameras) some problems infants have to solve (e.g., body learning and learning of association of the visual system to the motor system) can be ignored or simpler solutions can be adopted (e.g., using directly the very precise 3D information for grasping).

It is important to realise what kind of learning is appropriate (and possible) at an early stage of development. Since language is not yet developed, teaching by any explanation is not an option. Moreover, up until about nine months, it is unlikely that imitation plays a crucial role in learning to interact with objects (see Sect. IV-B4). Therefore, exploration is likely to play a decisive role at that stage of development, which is also supported by observations in developmental psychology [10], [11]. Along a similar line of thought, exploration is the main driving force in our system, supported by very weak supervision by a human 'robot-sitter' whose role it is to create learning situations and to avoid self-damage, in a way very similar to infant supervision.

To realise this learning task our system requires a minimal amount of innate knowledge about the world with which it interacts. This knowledge is expressed in (1) the system's embodiment, (2) the machinery for (visual) feature extraction, (3) structural knowledge (statistical machineries and memory system), (4) a number of innate behavioural patterns, and (5) knowledge of the physical world. However, as we show, this knowledge is rather generic, and choices are motivated by neurophysiological knowledge as well as results of developmental psychology (as discussed in Sect. IV-A).

Manuscript received March 1, 2010; revised July 12, 2010. This work was supported by the EU project PACO-PLUS (IST-FP6-IP-027657), the EU Interreg Project IRFO (33-1.2-09), and the Belgian National Fund for Scientific Research (FNRS).

Dirk Kraft, Emre Başeski and Norbert Krüger are with the University of Southern Denmark, Mærsk Mc-Kinney Møller Institute, Campusvej 55, 5230 Odense M, Denmark. (e-mail: {kraft,emre,norbert}@mmmi.sdu.dk)

Renaud Detry and Justus Piater are with the University of Liège, INTELSIG Group, Grande Traverse 10, 4000 Liège – Sart Tilman, Belgium. (e-mail: {Renaud.Detry,Justus.Piater}@ULg.ac.be)

Nicolas Pugeault is with the CVSSP, University of Surrey, GU2 7XH Guildford, UK. (e-mail: n.pugeault@surrey.ac.uk)

Frank Guerin is with the University of Aberdeen, Department of Computing Science, Aberdeen AB24 3UE, Scotland. (e-mail: f.guerin@abdn.ac.uk)

By carefully combining these innate competences and behaviours, we end up with a system that gradually enriches its internal representations. This process leads to an increasingly mature interpretation of the world and the system's ability to act within it, up to (still very simple) planning and plan execution with grounded object and grasp representations. We then compare the system's prior knowledge with equivalent knowledge available to humans, and discuss similarities and differences. In particular, we compare the system's developmental progress with developmental stages of infants in Sect. IV-B.

The modelling of such bootstrapping processes serves (at least) two purposes. First, it is a necessary competence for cognitive robots to learn by interacting with the world; hence, we believe that it is an important capability of future robotic systems. Secondly (and equally importantly) it provides a better understanding of the internal requirements of such complex processes and thus provides insight for the understanding of human development.

A. Related work

Since our work focuses on bootstrapping mechanisms in a cognitive system which makes use of a number of rather complex sub-modules, a large variety of research questions are involved (grasping without object knowledge and object specific grasping, object learning, structure from motion, pose estimation, etc.) that cannot be fully covered in this section; we focus here on work that is relevant in the context of bootstrap learning of grasp representations (Sect. I-A1) or object and action knowledge generation by exploration (Sect. I-A2). For a more detailed discussion of the state of the art for each submodule used in the bootstrapping process, we refer to more technical publications (see, [12]–[16]).

1) Learning grasping affordances: Classical approaches to grasp generation [17], [18] rely on predefined 2D or 3D models. These models have mostly been used to construct successful and stable discrete grasping solutions using analytical methods. Of the more recent, large body of literature on learning how to grasp, the majority focuses on methods that produce a number of discrete grasping solutions [19]. A few recent methods (including that used by our system [14]) instead aim explicitly at producing a continuous, probabilistic characterisation of an object's grasping properties [20], [21]. The latter can naturally be used to produce grasping solutions; additionally, they allow for ranking grasps by providing a likelihood of success for arbitrary grasps. Montesano et al. [21] learnt 2D continuous and probabilistic grasp affordance models for a set of objects of varying shape and appearance, and developed a means of qualifying the reliability of their grasp predictions.

Chinellato et al. [22] describe a relevant manipulation experiment in which categories of grasps (based on features derived from 2D-visual and motor information) are learnt. These categories can be interpreted as symbols representing different grasp types. The features used by Chinellato et al. contain a significant amount of prior information. They are constructed based on the authors' thorough understanding of the grasping process. We do not make use of such specifically-tailored features; the different grasp categories in [22] are object independent while our work learns object-specific grasps.

2) Body, object and action learning by cognitive agents: The work of Fitzpatrick and Metta [6] is closely related to our object learning approach since the overall goal as well as the hardware setup are similar: discovering relations between actions and objects by exploration using a stereo system combined with a grasping device. We see our work's main distinguishing features in the larger amount of prior structure we use and in the more complex and reliable visual and action models we are able to produce. For example, we assume a much more sophisticated vision system. Also, the use of an industrial robot allows for a precise generation of scene changes exploited for the extraction of the 3D shape of the object. Similarly to [6], we initially assume 'reflex-like' actions that trigger exploration (connected to the concept of affordances [23]). However, since in our system the robot knows about its body and about the 3D geometry of the world and since the arm can be controlled more precisely, we can infer more information from having physical control over the object in terms of an exact association of visual entities across successive frames based on proprioceptive information. Therefore, we can learn a complete 3D representation of the object (instead of 2D appearance models) that can then be linked to pose estimation.

Our work is very much related to a general research agenda on bootstrapping cognitive systems outlined by Kuipers. In particular, Modayil and Kuipers [24], [25] addressed the problem of the detection of objectness and the extraction of object shape (of e.g., chairs, trashcan) in the context of a mobile robot using a laser sensor. Motion information (in terms of the odometry of the mobile robot) is used to formulate predictions; in this way, they can extract a 2D cross section of the 3D environment. In our work we did make use of a much richer visual scene representation (leading to full 3D object representation covering geometric and appearance information) as well as a more controlled way to interact with objects (grasping and highly controlled manipulation of objects in contrast to mobile manipulation and merely pushing objects).

Stoytchev [26] shows how to learn to pick up seven different objects by experimentally chaining primitive actions, taken from a discrete set. In contrast, we learn to grasp in a continuous parameter space. Stoytchev's work is done in a dynamic simulation environment, which simplifies some vision problems. Corners of objects are input directly; a different approach would be needed for objects that do not have the relatively simple structure of the seven objects used. Furthermore, objects are not recognised, but are colour coded so that they can be uniquely identified. There is no link between exploration and the learning of an object's visual representation. In contrast to this work, our experiments are done in a realistic environment which tackles real-world vision and we do not ground our grasping actions in primitive actions, but rather use an abstract 6D pose for grasp representation. Stoytchev's later work [27] shows how different tools and their actions can be learnt and how this knowledge can be used

to solve problems. This work is done in a real-world robot scenario while manipulation happens in a planar world. The author also investigates how this learnt tool representation can be updated in case the tool changes (e.g., breaking off of parts). Our work does not explicitly address shape changes; however, our learnt grasp models will adapt to gradually-changing objects during continued interaction. Moreover, added or removed parts will be robustly handled in the same way as clutter and occlusion and will generally be inconsequential, unless the parts in question directly interfere with applied grasps.

This work is based on earlier work [13], [14]. While [14] focuses on technical aspects, the conference publication [13] discusses how individual objects and grasp affordances can be grounded. This article bundles prior work into a bootstrapping system that is described herein for the first time. In this context, we also compare this system with findings from neurophysiology and developmental psychology. In particular, we review some findings that justify the prior knowledge built into the system, and discuss how our system relates to the development of infant grasping abilities in their first year.

II. The Developmental Process

The developmental process has three main stages which themselves can be split again into different sub-stages. This is visualised in Figs. 1 and 2. The first stage involves a first learning cycle (see Fig. 3 (top)) called 'Birth of the object'; the second stage involves a second learning cycle (see Fig. 3 bottom) similar to playing with an object (grasping and dropping the object); the third and final stage can then use the knowledge acquired previously to manipulate objects in a planful way.

During the first 'innate' stage the system ('I' in Fig. 1) merely performs actions triggered by feature-induced affordances (see [28] for a video).[1] The system does not have any concepts of concrete objects and how to grasp them. It only executes a mechanism that tries to perform grasps on something (i.e., not necessarily an object) triggered by specific 3D feature constellations which the early cognitive vision system provides. In case the system experiences a resistance to a total closing of the two-finger gripper, another mechanism triggers a rotational movement which provides the visual system with a set of controlled views of the object, from which it accumulates features that move in agreement with the proprioceptive information (see [29] for a video). The set of these features (provided the predictions of the visual features based on the proprioceptive information can be verified) constitute 'objectness' and object shape. Hence at the end of the first learning cycle the system possesses concrete object knowledge. Based on this it also has now a more mature means to analyse the scene by recognising learnt objects and estimating their pose. This is the second stage reached in 'V' in Fig. 1.

[1]In this paper we refer to two types of affordances: a feature-induced affordance is identified by the visual system based on coplanar contours and requires no prior object knowledge; an object-specific affordance requires an object model, and refers to a specific gripper pose at a specific point on the object model.

During the second stage a more complex behaviour compared to the feature-induced affordance driven behaviour of the first stage is performed. Since the system is now able to estimate the object's pose it can perform focused actions on an object (i.e., grasp it with a certain end-effector pose in relation to the object). In this way the system can systematically test grasp hypotheses to build up grasping experience associated with the object. At the end of this process the system has acquired object shape and appearance and object-specific grasping knowledge; this facilitates entry to the third main stage, which can perform behaviours in which multiple objects can be manipulated in a systematic way, a precursor to planning. This third stage is not described in this paper (see [30] for more detail). Instead, the focus here is on how to bootstrap a system to such a level of knowledge.

We want to note three issues: First, in human development (or in a more faithful developmental robot) these stages are not distinctly sequential. Purely feature-induced affordance-based behaviour, more focused 'playing' and planning will all be available at later stages. Secondly, during play more is learnt than grasping a specific object, for example, fine-tuning of movements, aligning the body with the vision system and learning more sophisticated grasping affordances. This idea of significant overlap between different processes is very much in line with contemporary theories of cognitive development [31, Ch. 4]. Thirdly, the emergence of deliberate planning is not a discontinuous development, but rather a process of gradual increase in the sophistication of skills and their combination; evidence suggests that the basic strategy of applying a means to achieve a goal is available in the first six months, but means-ends performance in manual tasks will not be observed until about 8 months because the required component skills (or "planning operators") are not available earlier [32, p. 40-41]. Later on, as more playing is done, more is learnt about potential planning operators and so more sophisticated planning becomes possible.

Figs. 1 and 2 exemplify the learning process inherent in the interaction of the different internal modules and behaviours interpreted in accordance with the popular model of working memory of Baddeley [33] (see Sect. IV-A3). We realised a procedure that reflects this model and leads to a stepwise enrichment of the internal representations. This enrichment process produces a gradually maturing interpretation of the world that allows for increasingly complex actions. The horizontal axis gives the different types of memory that are involved in the developmental process: Iconic Memory (IM), Visuospatial Sketchpad (VS), Episodic Buffer (EB), Object Memory (OM) and Grasp Memory (GM). Note that IM is part of the sensory store, VS and EB are both part of the short-term memory (STM) while OM and GM are both part of the long-term memory (LTM). The vertical axis represents the time on a coarse scale, corresponding to critical stages of the system's development. The horizontal bar on top of the different memory systems represents competences of the system. These also change (i.e., develop) over time.

The different memory systems have different roles in the developmental process. Iconic memory stores and integrates visual information for only approximately 250 ms [34] and

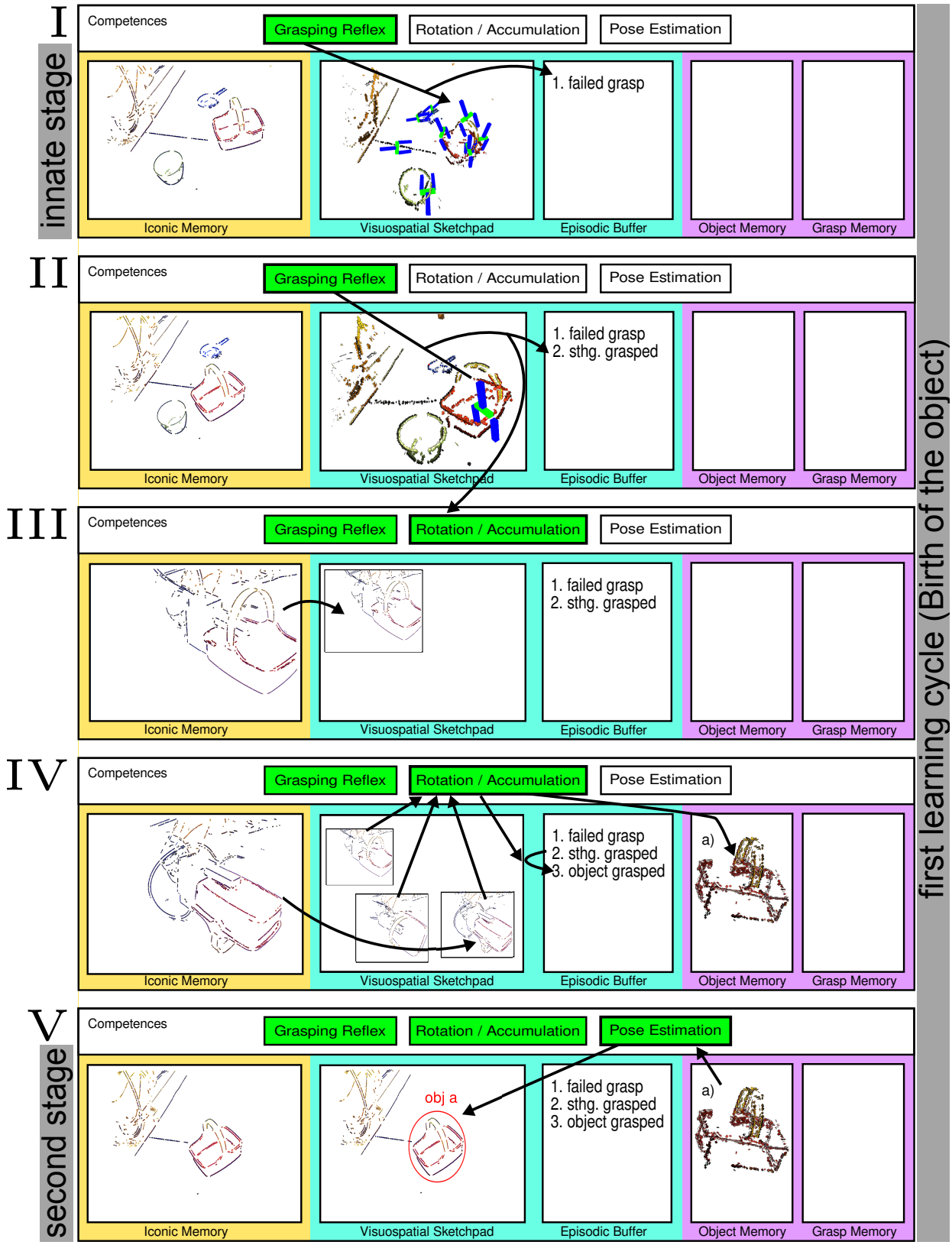


Fig. 1. Illustration of how our processes can be interpreted in accordance with human memory models (using the popular working memory model of Baddeley). See Sect. II for more details. Continued in Fig. 2.

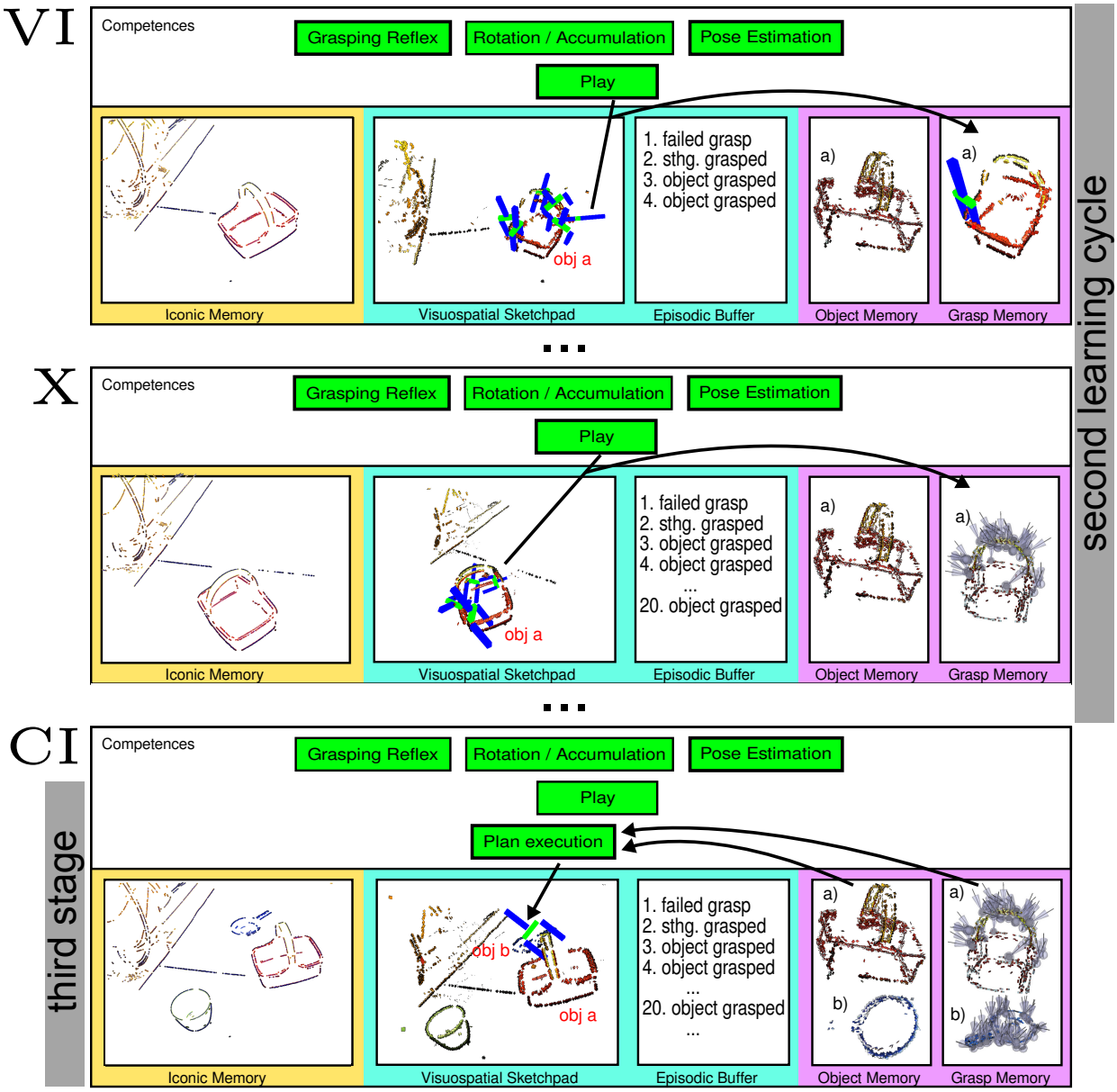


Fig. 2. Continued from Fig. 1.

delivers the information to the visuospatial sketchpad [33, p. 63], where internal representations act on the sensory information (in our case associating grasping hypotheses to visual features, see, e.g., Fig. 1I. Experienced relations between sensory information and action are stored in the episodic buffer [33, p. 139]. The information in the episodic buffer delivers input to learning processes that initialise and refine concepts in the long-term memory—in our case, object-specific grasping affordances. More specifically, we store visual features which triggered a grasp, motor information associated to the grasping attempts, as well as a success evaluation of the grasp attempt. The success evaluation is based on haptic information.

To view the system from Piaget's theoretical perspective [35], we could describe it in terms of the functions of assimilation and accommodation. The accumulation process involves assimilation of the object according to the system's innate idea of objectness, and then rotating it just as it would do with any other object. The accumulation process involves accommodation when the system changes its object memory to accommodate this new, previously-unseen object. The playing process first involves assimilation of the object by a model existing in object memory (pose is estimated in this assimilation process); a grasp attempt is then triggered. The result of the grasp attempt leads to a change in the recorded empirical grasp density; this change corresponds to Piaget's accommodation. This procedure is thus broadly compatible with Piaget's theory where each process involves an element of assimilation and an element of accommodation.

The competences and behavioural patterns are activated in the developmental process. Initially they are directly activated by affordances (e.g., the grasping reflex is used to grasp objects) and later in the developmental process by

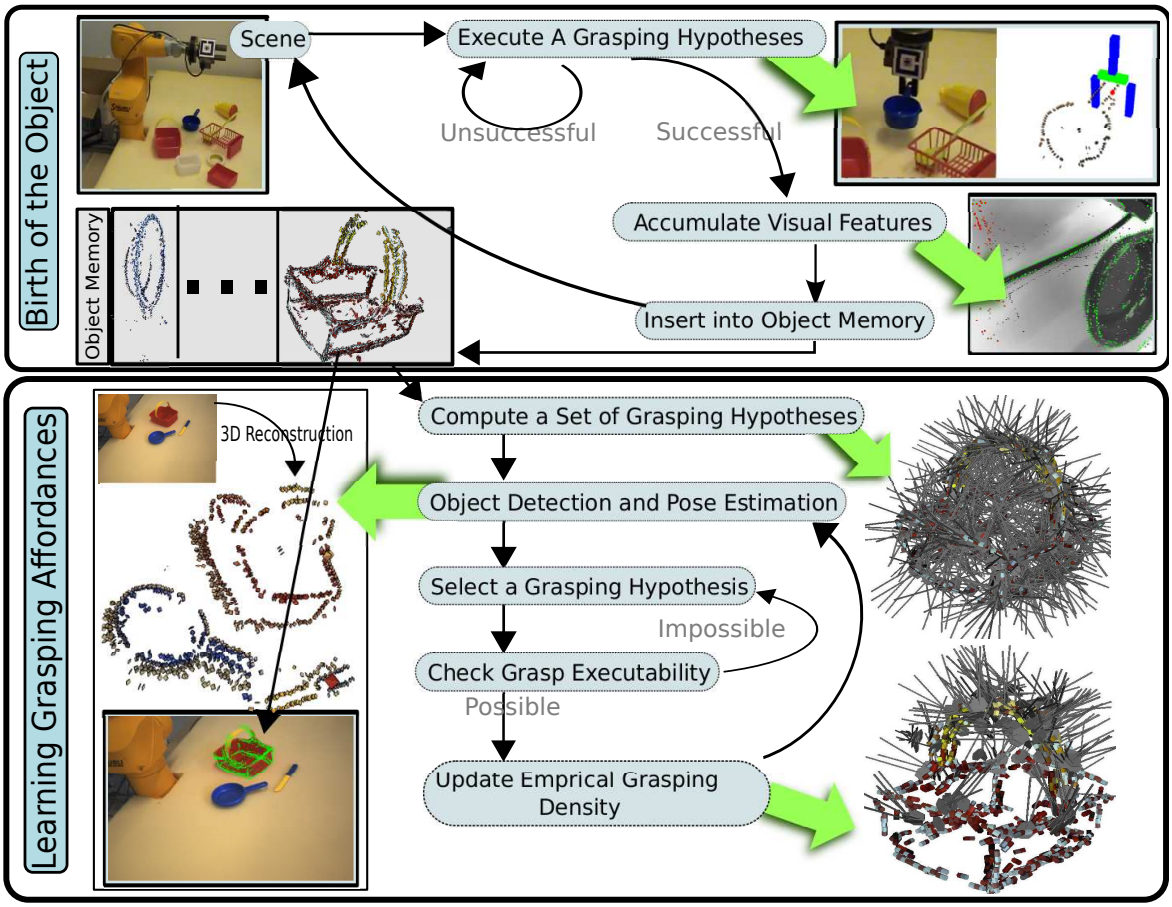


Fig. 3. The two learning cycles. In the first learning cycle (top) a visual object model is learnt after gaining control over the object. The second learning cycle (bottom) uses this model for pose estimation which enables the system to attach attempted grasps to a common coordinate system. These grasps can then be used to construct grasp densities, which form a continuous grasping affordance representation.

more purposeful considerations (e.g., the planner decides what object to grasp in which way). These competences and behavioural patterns are combined over time to form more complex concepts and behaviours, providing an increasingly rich world model: The developing system proceeds from object-independent, stimulus-driven grasping to more purposeful and mature grasping of objects, and finally forms sequences of behaviours to execute purpose-directed plans.

The main developmental steps and the corresponding internal system states (Figs. 1 and 2) are now briefly outlined by means of concrete examples, before we describe the actual sub-modules applied in this process in Sect. III:

I) At the initial stage, a 'grasping reflex' (see Sect. III-A2) is triggered by a visual stimulus in the VS. In this case, the execution of the selected grasping hypothesis leads to a collision between gripper and basket; the grasp is labelled as failed and stored in the EB as a combination of visual features and outcome evaluation. We note that, at this point of development, the only competence available to the system is to trigger one of the feature-induced affordances by the grasping reflex (GR), as indicated by the only green highlighted competence. This constitutes a purely affordance-based, reactive scheme.

II) Another attempt to perform a grasping hypothesis leads to a successful grasp (measured by haptic information); the basket is grasped and the action is labelled accordingly. Note that, by having physical control over 'something', a new (however innately available) competence becomes usable by the system ('Rotation / Accumulation').

III) At this stage, an accumulation process (see Sect. III-A3) is triggered following the successful grasp, during which the object is observed from different perspectives.

IV) After a set of views have been collected, they are incorporated into a common model. An object is born when a sufficient number of the features moving according to the robot's proprioceptive information have been accumulated. The EB is updated accordingly and the object model is stored in OM. This concludes the first learning cycle as described in Sect. III-B1.

V) At this stage, the new entry in the object memory allows for a more complex representation in the visuospatial sketchpad via the use of the pose estimation process (see Sect. III-A4). Now the scene represented in the VS contains not only grasping affordances but also a concrete object that can be localised by a

pose estimation algorithm, as indicated by the green highlighted box.

VI) A new competence combining grasping reflex and pose estimation, generating new entries in EB and GM, is activated. From this point onwards, learning happens by playing, i.e., picking up and dropping the object. This is essentially an iterative combination of the grasping reflex with the pose estimation (described in Sect. III-B2).

X) After grasping the object multiple times, the grasping model (described in detail in Sect. III-A4) becomes sufficiently complete. This concludes the second learning cycle.

XI–C) Additional objects are born and grasping models are learnt by the very same mechanisms.

CI) Based on the learnt object and grasp representations, planning with grounded objects and grasps finally becomes possible. This is described in detail in a separate publication [30].

III. Sub-modules and Learning Cycles

This section presents the technical implementation of the developed system. The implementation is based on an embodiment (see Fig. 4) consisting of a classical six degree of freedom industrial robot arm (Staubli-RX60) with an attached two finger gripper (Schunk PG-70) and a calibrated stereo camera system (Point Grey BumbleBee2). In addition the system uses a Force/Torque sensor (Schunk FTCL-050) to detect collisions between gripper and environment. In this context, a foam floor leads to a slow increase of forces and hence allows for longer reaction times.

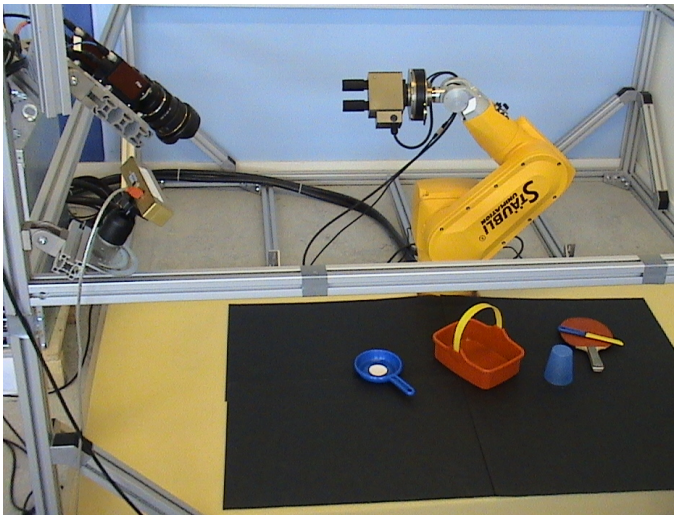


Fig. 4. Hardware setup (description see text).

We now describe in Sect. III-A the modules that are used by the system and how they interact within the two learning cycles (described in Sect. III-B) in a cognitive agent which is able to learn object representations and grasping actions autonomously.

A. Visual and Behavioural Modules

This section details the technical implementation of the a priori competences and behaviours that build the basis for the realised system. Here we want to stress that although many parts of the system are motivated by and have analogies to human information processing, these analogies are not modelled on a neuronal level but represent functional abstractions of processes realised in the human brain by means of neurons (for an in depth discussion of analogies to the human visual system we refer to, e.g., [36]). Moreover, some processes are also realised rather differently in the human brain as discussed in detail in section IV.

1) Early Cognitive Vision System: In this work, we make use of a visual representation delivered by an early cognitive vision system (ECV) [16], [36], [37]. Sparse 2D and 3D features, so-called *multi-modal primitives*, are created along image contours. The 2D features represent a small image patch in terms of position, orientation and phase. These are matched across two stereo views, and pairs of corresponding 2D features permit the reconstruction of a 3D equivalent. The 2D and 3D primitives are organised into perceptual groups in 2D and 3D (called 2D and 3D contours in the following). The procedure to create visual representations is illustrated in Fig. 5 on an example stereo image pair. Note that the resultant representation contains not only geometrical information (i.e., 2D and 3D position and orientation) but also appearance information (e.g., colour).

The sparse and symbolic nature of the multi-modal primitives allows for the coding of relevant perceptual structures that express relevant spatial relations in 2D and 3D [38]. The relations between contours allow us to define grasping hypotheses (see Sect. III-A2). The formalisation of a primitive's visual change under a rigid-body motion enables us to accumulate the 3D primitives belonging to an object (see Sect. III-A3).

2) Feature-induced Grasping Affordances: To gain physical control over unknown objects, a grasp computation mechanism based on previous work [15] is used. Pairs of 3D contours that share a common plane and have similar colours suggest a possible grasp (see Fig. 6a–c). The grasp location is defined by the position of one of the contours. Grasp orientation is calculated from the common plane defined by the two contours and the contour's orientation at the grasp location. Every contour pair fulfilling this criterion generates multiple possible grasps (see Fig. 6a for two such possible grasp definitions and [28] for a video showing the execution of the grasping behaviour in a complex scene). In the context of this behavioural module, the visual features, the feature relations used for defining the grasps as well as a haptically-generated succes evaluation are required and are hence stored in the episodic buffer (for details see [15]).

Here we want to note that the grasping mechanism can be generalised to more complex hands than two-finger grippers. For example in [15] the feature-grasp associations shown in Fig. 6a were mapped to a five finger hand. In general, in [39] it has been shown that the high-dimensional manifold of joint configurations of a five-finger hand can be mapped to a much lower-dimensional subspace that is able to represent most grasping actions. This indicates that a rather limited set of

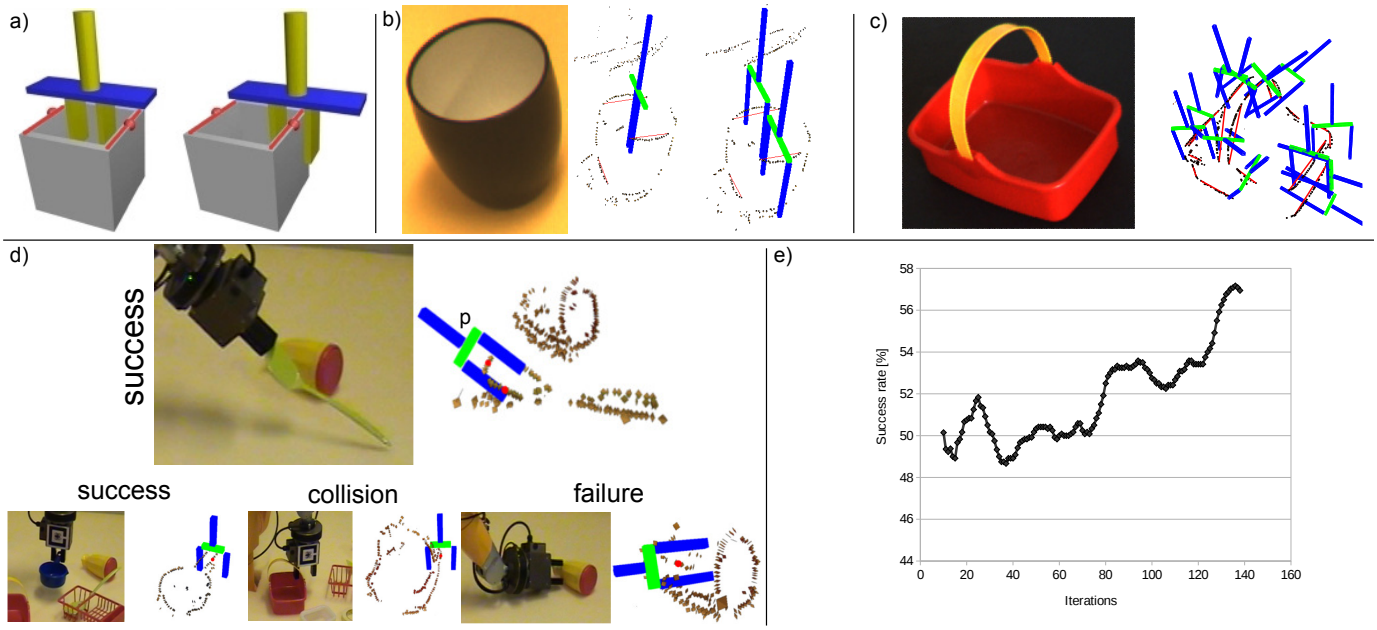


Fig. 6. Grasping reflex. a) Based on two coplanar and co-colour visual contours certain grasps are predicted. b) Concrete situation and the grasps predicted by a specific contour pair. c) More complex scene and a selection of predicted grasps. d) Results based on real grasping attempts stored in the episodic buffer. The gripper pose (p), the position of the centres of the two generating contours (visualised as red dots here) as well as the evaluation result are stored. e) The success rate after learning versus the increase of the learning set.

feature-action associations might already generate a significant variety of grasping actions.

The system's embodiment allows it to detect collisions with its environment (e.g., the object it tries to grasp) and to judge if it successfully grasped 'something'. This allows for an autonomous operation and an autonomous generation of labelled experiences (see Fig. 6d. We have shown in [15] that, based on these labelled experiences, we can learn an improved feature-based grasp generation mechanism. The system uses an artificial neural net to determine which feature relations predict successful grasps. Fig. 6e shows how the success rate increases with the amount of labelled learning data the system can make use of.

3) Accumulation: Once the object has been successfully grasped, the system manipulates it to present it to the camera from a variety of perspectives, in order to accumulate a full 3D symbolic model of the object [16]. This process is based on the combination of three components. First, all primitives are tracked over time and filtered using an Unscented Kalman Filter based on the combination of prediction, observation and update stages. The prediction stage uses the system's knowledge of the arm motion to calculate the poses of all accumulated primitives at the next time step. The observation stage matches the predicted primitives with their newly observed counterparts. The update stage corrects the accumulated primitives according to the associated observations. This allows the encoding and update of the visual primitives. Second, the confidence in each tracked primitive is updated at each time step according to how precisely the accumulated primitive was matched with a new observation. The third process takes care of preserving primitives once their confidences exceed a threshold, even if they later become occluded for a long period of time. It also ensures that primitives are discarded if their confidence falls below a threshold. New primitives that were not associated with any accumulated primitive are added to the accumulated representation, allowing the progressive construction of a full 3D model. Note that the sparse nature of primitives yields a condensed description. Fig. 7a shows how an object model improves qualitatively by applying the accumulation scheme over different frames. Fig. 7b shows the variation of the Kalman gain over frames clearly indicating a convergent behaviour.

4) Statistical Machinery—Pose Estimation and Grasp Densities: The accumulated 3D symbolic reconstruction described above can serve for object pose estimation. Pose estimation is performed using the model of Detry et al. [12]. This model has the form of a hierarchy of increasingly expressive object parts, where bottom-level parts correspond to generic multi-modal primitives. The model is learnt from the accumulated 3D symbolic reconstruction (see Sect. III-A3) of the object, and allows for a probabilistic estimation of the object pose in an arbitrary scene. Visual inference of the hierarchical model is performed using a belief propagation algorithm (BP) [12], [40], [41]. Means of autonomously learning the hierarchical model from an accumulated 3D symbolic model are presented in prior work [12].

The role of pose estimation is to align *object-specific* grasp affordances to arbitrary object poses. Object-specific affordances represent the different ways to place a hand or a gripper near the object so that closing the gripper produces a stable grip. The grasps we consider are parametrised by a 6D gripper pose composed of a 3D position and a 3D orientation. Object-specific affordances are represented probabilistically with *grasp densities*. A grasp density represents the spatial

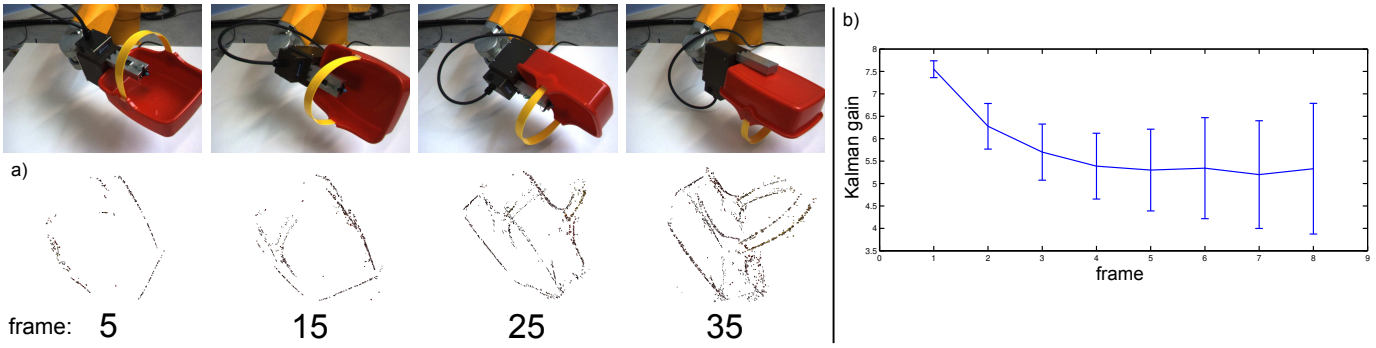


Fig. 7. Illustration of the accumulation process. This accumulation is the product of two interleaved processes. In the generative process, illustrated in panel a), the object is rotated over a number of frames; for every frame the visual features are extracted and accumulated, enriching the object model. The second process, illustrated in panel b), tracks and corrects features over time. The reliance on the predicted features increases while the influence of new observations decreases, as can be seen in the reduction of the gain. Please note that at every step new features (with an initial high Kalman gain) are introduced, which leads to larger error bars. It is important to point out that even though such new, uncertain features continue to be introduced, the overall Kalman gain still decreases. At a Kalman gain of 6.5 new observations and predicted features have approximately the same impact on the update stage. These two sub-processes occur at very different timescales. The accumulation of new information, in panel a), requires large viewpoint variation to be effective, and the object's shape becomes complete only after a half revolution of the object ($\sim$ 35 frames); in contrast, the correction process, in panel b), converges after only a few frames ($\sim$ 5 frames).

distribution of relative object-gripper poses which yield stable grasps; it corresponds to a continuous probability density functions defined on the 6D pose space [14]. The computational representation of grasp densities is non-parametric: A density is represented by a large number of weighted grasp observations. Density values are estimated by assigning a kernel function to each observation and summing the kernels [42]. Letting K_i denote the kernel associated to the i^{th} grasp observation, and letting w_i denote the associated weight, the value of a density at pose x is given by $\sum_i w_i K_i(x)$. The learning of grasp densities—acquiring grasp observations and learning the weights—is detailed in Sect. III-B2. An intuitive illustration of a grasp kernel is given in Fig. 8a and Fig. 8b illustrates a kernel-based grasp density. Grasp densities are registered with the visual reconstruction of the object they characterise, allowing for their alignment to arbitrary object poses through visual pose estimation. The association of grasp densities with the visual model is covered in more detail in prior work [14].

B. The Two Learning Cycles

The modules described in the previous section can be combined into two learning cycles. The first cycle learns visual object models; the second learns object-specific grasping affordances from these models. The issue of learning to recognise affordances which are not specific to known objects is not tackled here but is the subject of ongoing work (see end of Sect. IV-B2).

1) Birth of the object: Fig. 3 (top) shows how the two submodules described above interact to generate visual object models for unknown objects. The initial grasping behaviour (see Sect. III-A2 and steps I–II in Fig. 1) is used to gain physical control over an unknown object. If no object has been grasped in the process (this is determined using haptic feedback, i.e., the distance between the fingers after grasping) another grasping option is executed. When the object has been grasped, the accumulation process (see Sect. III-A3 and steps III–IV in Fig. 1) is activated. If a stable representation emerges after some steps, then the grasped entity possesses temporal permanence. Together with the attributes manipulability and constrained size relative to the agent (which have been established by the agent being able to pick up the entity), the entity fulfils all of Gibson's [43] criteria of objectness. We use the name *Birth of the object* for this transition process, from entity to object. The generated object model is then stored in the Object Memory. This process can be repeated until all objects in the scene have been discovered. Object models resulting from the first learning cycle can be seen in Fig. 3 (top) in the column labelled Object Memory and in [44].

2) Object-specific grasp affordance learning: Affordances can initially be constructed from a grasp generation method that produces a minimum ratio of successful grasps (e.g., the initial feature-induced grasping behaviour in Sect. III-A2). In this work we used an approach where we initially use grasp hypotheses at random orientations at the position of the ECV primitives of the object model (see Fig. 8d). A grasp density model is constructed from these hypotheses by using each hypothesis as a grasp observation; observation weights are uniform. We call the representations built with any of these weak priors *grasp hypothesis densities* [14].

An object's grasp hypothesis density allows for grasping, but yields low success rates. In order to improve success rates, the system uses exploration and the execution of a number of random grasps sampled from the hypothesis density. Successfully-executed grasps are used as observations for building an *empirical* grasp density (see Fig. 8e). The weights associated to these grasps are computed through an importance sampling algorithm [14] in an effort to remove the bias introduced by the grasp hypothesis density. The empirical grasp density yields higher success rates than the grasp hypothesis density, models more accurately the object's properties and reflects the robot's morphology (see Sect. III-A4 and steps VI–X in Fig. 2). In [45] a video of the learning and execution of this grasping behaviour is shown for a variety of objects.

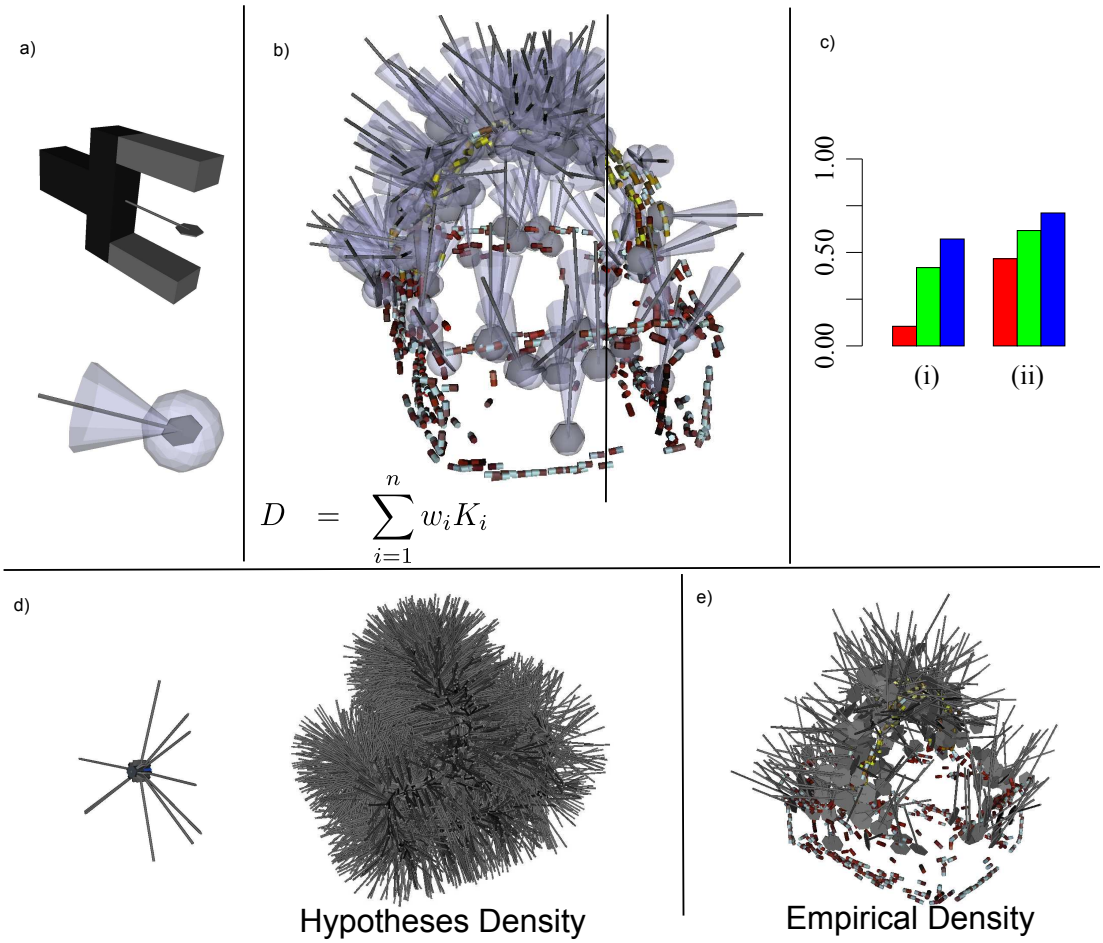


Fig. 8. Grasping affordances are represented using kernel-based grasp densities. a) Iso-probable surface of a 'grasp kernel', and relation between a two-finger grasp and a kernel representing this specific grasp in the model. b) Kernel-based grasp density. The right-hand side shows lighter sampling for illustration purposes. D represents the density, while w_i and K_i represent the individual weights and kernels. c) Grasp success rates for the basket, (i) counting kinematic path planning errors as failures, and (ii) excluding such errors from the score. Red bars correspond to grasps drawn randomly from the the hypothesis density d). Green bars correspond to grasps drawn randomly from the empirical grasp density e). Blue bars correspond to maximum-likelihood grasps from the empirical grasp density e). d) Initial grasping knowledge is modelled with a hypothesis density (right) which is based on kernels placed at the positions of the visual features at random orientations (left). e) Empirical grasp density.

The success rate of grasps sampled randomly from the hypothesis and empirical densities of the plastic basket are shown in Fig. 8c in red and green respectively. Instead of drawing a grasp randomly from a density, a robot may also select the grasp that has the maximum success likelihood. This is done by combining the grasp density with reaching constraints to select the achievable grasp that has the highest success likelihood. The success rates of maximum-likelihood grasps computed with the empirical grasp density of Fig. 8e are shown in blue in Fig. 8c. The process of computing hypotheses densities, pose estimation and execution of random samples from the grasp hypothesis density, through which an empirical grasp density is generated, is shown in Fig. 3 (bottom).

IV. Reflection about similarities and differences to human development

In this section we first discuss the prior knowledge that has been applied in our system and relate it to knowledge about innate structures in humans (Sect. IV-A). We then discuss in Sect. IV-B similarities and differences between our system's development and the development of infants' object representation and grasping abilities.

A. Prior Knowledge

The system's innate knowledge can be distinguished by (1) its embodiment, (2) the machinery for (visual) feature extraction, (3) structural knowledge (statistical machineries and memory system), (4) a number of innate behavioural patterns and (5) knowledge of the physical world. These will be discussed in more detail in the rest of this section.

1) Embodiment: The system has knowledge about its embodiment and the consequences of its movements in the three-dimensional world. In technical terms, the system knows about its body shape and body kinematics and is able to relate its self-motion to the visual information it perceives. In addition it has the ability to plan collision-free motions while respecting the limits of its work space.

Besides the actual control of its body, an important property of the system is that it is able to achieve a high level of control over objects by grasping. Interestingly in this context, Pinker

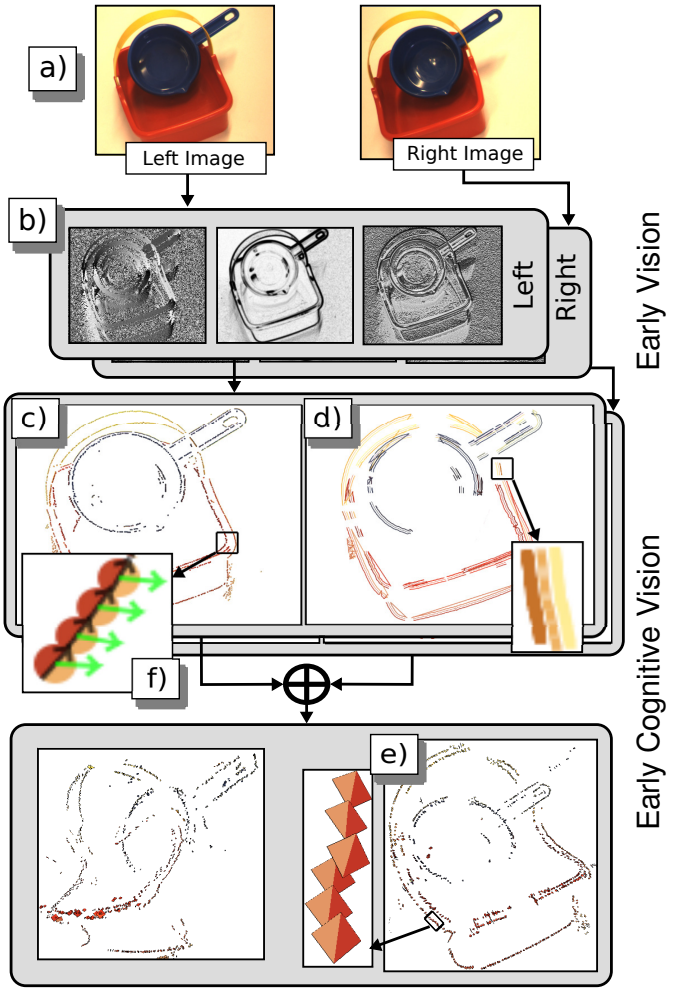


Fig. 5. An overview of the visual representation. **a)** Stereo image pair, **b)** Filter responses, **c)** 2D primitives, **d)** 2D contours, **e)** 3D primitives (note that the surface on which the colour is displayed is for display only and has no geometric meaning, for details see [37]), **f)** close-up of c).

[46, p. 194] speculates that the superior manual dexterity of humans may be one of the factors which facilitated the species's development of higher-level cognitive competences. In the development described here, the robot makes use of very precise predictions on how visual features change according to its ego-motion once the object is grasped. It has also the ability to 'subtract' its hand from visual features generated from the grasped object.

In our system, we decided not to have this knowledge develop. Instead, this body knowledge is hard-coded to allow us to focus exclusively on the problems of object and grasp learning. This is appropriate in the embodiment we chose since an industrial arm allows for a high degree of precision in robot-camera calibration and execution of movements (this is a significant difference from infant precision, and necessarily leads to a number of other differences in our system). It has also been shown that such knowledge can indeed be learnt by exploration (see, e.g., [6]). In humans, the body knowledge and its link to visual information must develop in parallel with object and grasp learning because the body itself is changing due to growth (see Sect. IV-B).

2) Sensory System: The robot is equipped with a sophisticated vision system (which we have called 'Early Cognitive Vision (ECV) system' [36], [37]) which provides semantically rich and structured 2D and 3D information about the world. This system contains prior knowledge about image features and their relations, knowledge about basic geometric laws and how they affect visual features, and most importantly, basic laws of Euclidean geometry used for stereo reconstruction and the prediction of the change of visual features of moving rigid bodies.

There has been a long debate on the innate components of V1, with contradicting evidence [47]–[50]. It has been argued that orientation columns in V1 are innate [47] or develop in some species without visual experience [50]. However, some work points to a large degree of plasticity [49] and it has been shown that filters associated with early feature extraction mechanisms can indeed be learnt [51] (for further discussion see [52]).

Our assumption of having a calibrated stereo system which is also precisely calibrated with the robot arm is an assumption which is not justified by human development, but possible in our specific set-up. It has been argued that infants are able to perceive 3D information from birth, probably based on the convergence cue [53], [54]. Stereo is used after approximately twelve weeks [55]. The stereo machinery starts rather instantaneously, probably caused by maturational change in cortical disparity-sensitive units [54] pointing to a large degree of innate structuring. The actual robot-camera calibration however can be learnt, as demonstrated by [6]; this also reflects the significant change of embodiment taking place in the first year (which is not modelled in our system either).

In summary, the ECV system provides a large degree of semantically-rich information in terms of 2D and 3D feature extraction and spatiotemporal relations between features. There exists evidence that complex feature extraction mechanisms in terms of orientation columns are already present at birth [47] or develop without visual experience of real-world data [48]; however, it is very likely that there are also adaptive and maturational components in this process that are not modelled in our system (see section 3 in [52] for an in-depth discussion).

3) Structural prior knowledge: Our system represents structural knowledge of objects using learnt, statistical models that relate ECV features to each other and to grasp parameters. This allows for robust object detection and pose estimation under uncertainty and noise by probabilistic inference. While it is unclear how this machinery compares to equivalent processes in the brain, there is substantial evidence that many brain processes operate in ways consistent with probabilistic reasoning [56].

For learning, the system makes use of a memory system that has been created in analogy to a model of the human memory system by Baddeley (see [33]) covering different sub-systems: First, an iconic memory which stores the immediate pre-processed sensory data for a short time period (approximately 250 ms, see [34]). Second, a visuospatial sketchpad [33, p. 63] in which internal representations can map on transient visual information. As a consequence of the

developmental process, the interpretation of the visual data in the visuospatial sketchpad becomes more and more complex starting from feature-induced affordances and ending up in complex scene representations in terms of objects and their poses as well as optimal grasping options (compare Fig. 1 step V and Fig. 2 step CI). Third, an episodic buffer [33, p. 139] in which sensory data, actions upon them as well as action consequences are stored. In the technical system, this is required to make instantaneous experimental information temporarily available to later learning stages. Hence, it can be seen as part of the short-term memory. Finally, object representations and grasping affordances (abstracted over long phases of exploration) are stored in a long-term memory.

There is a general notion that an elaborated memory architecture is available to infants. Indeed, it has been argued that limitations of such a memory system lead to severe constraints distinguishing competences of humans and apes. For example, apes are able to create and execute a plan for stacking boxes to reach a banana when the boxes are in their immediate reach. However, they have great difficulties achieving this task when the boxes are not available in the problem area (e.g., the boxes are in another room) [57]. This suggests that they cannot get sufficient access to representation in their long-term memory.

4) Behavioural prior competences: The system has two innate behavioural patterns that drive the bootstrapping process.

First, there is a premature mechanism to grasp unknown objects based on visual feature combinations (coplanar contours). This endows the system with physical control over objects (although with a rather low success likelihood for an individual grasping attempt). When successful, it triggers a second behavioural pattern (discussed below). An important aspect of the grasping mechanism is that individual grasps can be evaluated by haptic information leading to labelled data in the episodic buffer. Based on this information, the success likelihood of the initial (premature) grasping behaviour can be improved.

It is known that infants indeed possess a premature, innate, visually-elicited reach and grasp mechanism [8, p. 38]. This mechanism, together with the tactually-elicited palmar grasp reflex is thought to help bootstrap the learning of a more mature reach and grasp, which begins to appear at about four months (see more detailed discussion in Sect. IV-B).

The second mechanism performs a rotation of the grasped object (inspecting the object from all sides) and allows the system to segment objects and creates visual object models based on ECV features utilising the visual predictions derived by combining visual features and self-controlled object motion. Recent experimental work suggests that six-month-old infants do create 3D representations of objects, while four-month-infants do not, and it is suggested that this may be learnt when the infant begins to sit up, between four to six months, and can thus engage in extensive visual-manual exploration of objects [58], [59].

5) Knowledge of the physical world: The system rotates the grasped object in order to see the other side, and thus to build a complete representation of the object. This ability reflects an implicit expectation that objects do have an unseen side, which rotation can reveal. Evidence from psychology suggests that this may be something that infants need to learn [60]. Infants lack the capability to intentionally rotate an object to find an unseen side until about nine months [61, p. 120], which is quite late compared to grasping (more detail on grasping is given in Sect. IV-B below). Although infants may rotate during random exploration at six months [35], it is doubtful that it is done with the intention of seeing the other side. Instead, it is likely that this exploration helps the infant to develop towards knowledge of other sides. Given the coarse object representations which are likely to be in use up to six months, it is quite probable that such an infant rotating a box may not distinguish the sides as different. The inclusion of this rotation behaviour in the robot therefore represents an implicit knowledge which helps it to bootstrap its object representations.

B. Infant development of object and grasp representations

Human infants, in their first year, progress through significant changes in internal object representations as well as grasping abilities. Compared to our robot system, analogies as well as differences can be discerned.

1) Development of grasping competences: Infant grasping begins with a 'neonatal palmar grasp' reflex present from birth, where the fingers close on stimulation of the palm. This is followed by a *voluntary palmar grasp*, and grasping then progresses through a number of stages [9] leading to a *scissor grasp* at about 8.5 months, which uses the volar sides of the extended thumb and index finger. After some further intermediates this eventually develops into the *pincer grasp* at about twelve months, which uses the volar surfaces of the tips of the thumb and index finger. Development is by no means finished here; the second year will see an improvement in the use of appropriate forces, and the grasp will not approximate adult performance until six to eight years, with further subtle improvements continuing until adolescence [62]. Compared with an infant, our robot system does not develop different grasps and is pre-calibrated for simplicity. The grip used by our robot is two-fingered, and could be mapped to the scissor or palmar grasp, but without mechanical compliance. This simplification is justified in order to make technical aspects simpler; the calibration is simpler, the grip is firm, and also there is less occlusion as there are only two rigid fingers. A much more accurate computational model of infant grasping appears in the work of Oztop et al. [63], however it does not incorporate vision. If a more human-like hand is used in our system, then it may be necessary to extend the accumulation process so that a number of grasps and rotations from different points would be combined (to gather information about parts of the object that were occluded on the first rotation).

In terms of reaching to grasp seen objects, it has been observed that neonates have a very premature reach and grasp mechanism which reduces in frequency over the first two weeks, and is hard to elicit in the period from four to 20 weeks [5, Ch. 6]. The primitive reach motion is visually elicited, and may be ballistic, with no visual feedback to correct the reach motion while it is in progress [8, p. 38]; it has a nine to 40 percent chance of contacting the target [54, p. 250], [5]; furthermore, there is little evidence that the capability to

grasp accurately is present, although some elements such as hand opening are present [8]. The hand opening disappears at about two to three months, and the behaviour becomes more of a swiping with a closed fist, which is then replaced by an open-handed reaching, but with no grasp [64]. Note that palmar grasping without vision is developing in parallel at this time, and there is significant exploration of objects with the hands [65]. The more mature reach and grasp which appears at about 20 weeks (c. five months) has about an 80 percent chance of contacting the target, with the possibility of visual feedback to correct the reach in progress, although the grasp coordination appears to have regressed [5]. It seems that the primitive reach-grasp was undifferentiated (i.e., the reach and grasp are coupled), and by 20 weeks two differentiated (or decoupled) motions have replaced it, which can now be executed independently, or coordinated. Gordon [62] also notes that this type of regression followed by advancement is common: reorganisations can initially result in a decline of the motor skills, before improvement. After this period the coordinated reach and grasp develops rapidly; Bower reports 100 percent accuracy on visually elicited reaching [5, p. 174] at six months and an increasingly developed preadjustment of the hand is seen.

Compared to human grasping development our robot also makes use of an initial premature grasping behaviour which has a rather low success rate; however, after multiple attempts, this leads reliably to situations in which physical control is achieved. However our robot skips some of the infant development where reach and grasp must be decoupled and then coordinated; the starting position of our system is more akin to a six-month-old than a neonate.

One main problem for the young infant is that the motor system and the visual system are not well co-ordinated in the first four months. Infants may regard their hands moving in the second month, but the vision does not guide the hands [35, p. 102]. Subsequent to this, vision augments the activity of the hand. It is not until about the fourth month that proper visually directed grasping will commence, and that the infant will bring the hand into view to grasp seen objects even when the hand is not initially in view [61, p. 110]. After that, the calibration between both systems (comparable to the robot-camera calibration in our system) develops rapidly, leading to a well-calibrated system after six months as described above. For simplicity's sake, our robot is pre-calibrated: the infant must go through a calibration phase because its body is not produced to a precise standard specification and the rapid growth during infancy would necessitate perpetual recalibration in any event; these issues do not apply to our robot.

2) Visual information used to elicit reaching and grasping: There are few studies which address this issue specifically, but these show that five-month-olds seem to organise what they see into manipulable units with internal coherence and external boundaries, and they reach for the boundaries of these perceived objects [66]. Objects adjacent in depth seem to be treated as one object. This is consistent with Bower's finding that an object is at this stage defined by a bounded volume of space [5, p. 126]. There is a developmental progression in the infant's use of information about the size of objects; infants as young as eight weeks make more reaches to a graspable ball than one that is too large [8, p. 43]); five-month-olds tend to reach with two hands regardless of size, seven to eight-month-olds use two hands for large objects more often than for small ones, and at eleven to twelve months reaching closely reflects the object's diameter [67]. A similar pattern appears for the thumb-index finger angle opening during the reach, which increases after seven to eight months, as well as the adjustment of the angle to the object diameter and the proportion of the object within hand opening at touch [67].

With regard to edge detection, some differences between the infant and our robot system can be seen. Edges can arise due to discontinuities in luminance, colour, texture, depth or motion [54, p. 139]. Five-month-old infants use only depth and motion to determine object boundaries, probably because these have higher ecological validity [54, p. 149]; it is not exactly known when infants begin to use luminance and colour, but it is sometime in the second half of the first year [54, p. 149]. In contrast our vision system uses only luminance and colour edges from the beginning, because it detects edges from still 2D images.

Our robotic system recognises potential feature-induced grasping affordances by finding two coplanar contours. Coplanar contours are especially suitable for grasp attempts by our gripper with two rigid fingers, but if a more human-like hand is used this grasp reflex may need to be extended to make grasp attempts of a wider variety of features. The infant does not restrict him/herself to attempting grasps on coplanar contours, but will often attempt to grasp planes which may be at 90 degrees, or more irregular surfaces. The infant will learn that this is not so effective on a wooden block, for example, but can work quite well on a soft material such as a sponge or plastic bag; furthermore, although ineffective for picking up some rigid objects, a poor grasp can be adequate to pull it closer. The infant will thus learn to recognise different affordances, not only those that are good for grasping and lifting, but also those good for grasping and shaking, or pulling. Besides grasping, other exploratory actions (sucking, scratching, shaking, hitting, squeezing, etc.) are performed by infants which are not modelled in our system. This then reflects differences in the richness of the world of the infant vs. robot, and also the richness of its behavioural repertoire, which in turn reflects the richness of its embodiment. Given the simplicity of our robot's embodiment it is reasonable that it is limited to a subset of the infant's repertoire.

Infants' knowledge of grasping seems not to be as object-specific as realised in our robot's grasp densities. Infants are quite successful at grasping during the second-half of the first year, but they do not seem to build whole 3D object models until about 24 months [68]. This means that during the first year infants are probably recognising graspable fragments of objects in a coarse grained way, allowing for a higher degree of abstraction of grasp-object associations generalising across objects (generic affordances). However, our system generates knowledge which creates data for grasp learning in terms of more precise grasp densities on specific object models. Currently, we are working on finding feature-action associations across objects taking the grasp densities as input.

3) Development of object representations: Compared to human development, our robot system starts with a visual system that is much more mature than the infant's system at birth (approximately comparable to a six-month-old). In contrast, in the infant case, as discussed above, it is only after approximately four months that the infant begins to perform visually controlled movements of its arms [61, p. 110]. As a consequence, babies are in a position at the age of four months to produce controlled training data as we did with our robot and camera. Interestingly enough, the infant's concept of objects changes dramatically at this stage of development; babies younger than four months perceive an object as 'something at a certain position' or 'something moving with a certain velocity' [5, p. 227]; the infant does not seem to understand that a stationary object can begin moving without losing its identity, this understanding develops and tends to be present in the four to six months period. After approximately six months the representation of objects starts to be based on shape and colour cues [5], with colour being used considerably later than shape [69]. It is also about this time objects begin to acquire permanency, i.e., objects continue to exist while being occluded [60]. (Note that some later results contradicted this, but more recent results supported it; it is a subject of ongoing debate [70].)

Kaufman et al. [71] describe how the two separate visual processing streams in the infant brain (dorsal or ventral) are responsible for different tasks, and this has interesting implications for the development of object representations. The dorsal route should be primarily used for knowledge relating to grasping, while the ventral would be for representation and recognition of the object; yet these must be integrated to allow grasp knowledge to be associated with an object representation. In contrast our system has one integrated pathway where grasp knowledge and object representations are integrated as soon as they are available. It may be at quite a late age (maybe nine months [71]) that infants can integrate the information from the two streams. Even after this, recent work in psychology suggests that object recognition may undergo a long developmental trajectory; fragmentary representations based on view-dependent images and parts of objects are used up until about 18 months, and then there is a period of rapid change where 3D whole-object geometric representations are built by 24 months [68]. The picture emerging from the psychological literature is rather complex; Kellman and Arterberry explain that 'perception leads to multiple representations that may be recruited for different tasks' [54, p. 262].

During this long development, psychological theories acknowledge the importance of action as the foundation for perceptual and cognitive development [11], [60], and suggest that there should be a strong relationship between self-controlled movements of objects and the development of internal object representations. However, this has only recently begun to be investigated in more detail. The results by Soska et al. (mentioned in Sect. IV-A4 above) showed that self-sitting experience and coordinated visual-manual exploration were the strongest predictors of success on a looking task testing perceptual completion of 3D form [59]. The importance our system gives to active object exploration is very much in line with these theories. Furthermore, our system shows that the ability to create a situation in which an object appears under controlled conditions may help to learn a suitable representation of objects, and the same may be true in biological systems.

In the other direction there is evidence to suggest that 'perceptual competence may precede and guide emerging action systems' [54, p. 247]; therefore there is evidence of a bidirectional interaction between perception and action in development.

Our robot's learning cycle copies this idea in spirit but not in the precise details: in our system (and in infants) visual representation and action knowledge are built up together and bootstrap from each other; i.e., visual recognition of some possible feature-induced affordances facilitates action (to lift and rotate) which facilitates developing a visual representation, which in turn facilitates further development of grasping knowledge. However the robot's cycle is much more rapid than in infancy; a single successful grasp of the robot leads to a full rotation of the object, and the immediate construction of a full 3D model. Thereafter pose can be estimated on any new presentation of the object, and grasp positions can be associated with positions in the model. The competences employed here (3D model construction, pose estimation, rotation behaviour) take much longer to develop in infants [68]. Also, the robot immediately works on whole objects, whereas infants probably represent fragments such as handles or protruding parts before integrating these in full object representations. Infants are however probably going through a similar cycle with their coarser grained fragmentary representations; i.e., they are likely to be recognising a fragment such as a handle in different orientations, and learning about grasping strategies on this fragment.

4) Social Learning: As noted in the introduction, there is no social element in our system, apart from some supervision by a 'robot-sitter'. This is probably a reasonable match with the development of object and grasping knowledge in early infancy where social learning does not seem to play a crucial role. Interactions which require the infant to note an object, another person, and relationships among them, are known as 'triadic' interactions; Carpenter et al. trace their development from nine to 15 months of age, and note that twelve months is an important milestone: 'It is around one year of age that infants first begin to look where adults are looking flexibly and reliably, use adults as social reference points, and act on objects in the way adults are acting on them.' [72, p. 1]. Therefore the omission of social learning from our system probably does not diverge far from the human development of grasping knowledge up to about nine months.

V. Discussion

In this paper, we described a bootstrapping system which acquires rich object representations as well as knowledge about how to grasp these objects by autonomous exploration. This becomes possible by making use of a number of innate competences; these include substantial but generic prior knowledge about the visual perceived world, a sophisticated machinery to deal with uncertainties for pose estimation as well as

grasp representation, an elaborated memory systems and a set of initial behaviours. The system's exploration process can build on these competences to evolve a richer set of competences such as 'playing' and very simple 'planning'. The developmental process of the system has been compared to human development and similarities as well as differences have been described.

The system has been inspired by results from psychology in the broad outline of its developmental approach; however, in the details it is quite different, and part of the reason for this is that the precise details of how infants develop grasping knowledge are not known. The research for this paper has shed light on some gaps in the psychology literature, and thereby opens some interesting research directions which could be pursued by infant studies. For example there is a need to find what visual information is used to elicit and guide grasping; e.g., what exact shapes (e.g. handles or other protuberances) on parts of objects elicit specific grasp attempts with specific pre-adjustment of the hand, to what extent does this generalise across objects, and especially how does this develop throughout the first two years. Further to this, it would be interesting to investigate how this developing knowledge of shape fragments feeds into the process of building 3D object representations, and how it interacts with processes such as object identification and categorisation. Recent results from psychology suggest that recognition via fragments develops early in infancy and that the progression to whole-object shape happens between 18 and 24 months [68], and that action, to manually explore objects, is of prime importance throughout this development [59], [68]. It is the detail of this development which has yet to be worked out.

An important aspect of infant's bootstrapping is the bidirectional interaction between the development of visual representations and action knowledge. In some work on grasping we have seen one direction showing how visual features can be used to learn to identify good grasping points [19], [21]; in the other direction we have seen how exploratory actions can aid with visual segmentation [6], and some recent work has shown how object shape can be learnt via exploratory grasping [73]. The developmental psychology literature however suggests the need for a close bidirectional interaction over a long developmental period, where vision can guide exploratory actions, and these actions in turn help in the development of more advanced visual representations (see Sect. IV-B3). In our system such bidirectional processes take place, e.g., the physical control over objects facilitates the learning of object representations while these learnt representations can be used to create higher level behaviours such as 'playing with the object' to bootstrap object-specific grasping affordances. The potential of such bidirectional processes needs to be further explored.

REFERENCES

[1] D. Vernon, G. Metta, and G. Sandini, "A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computational agents," *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 2, pp. 151–180, Apr. 2007.

[2] R. Brooks, "Intelligence without reason," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 1991, pp. 569–595.

[3] J. Locke, *Essay concerning Human Understanding*, 1690.

[4] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural Comput.*, vol. 4, pp. 1–58, 1992.

[5] T. G. R. Bower, *Development in Infancy*. San Francisco : W.H. Freeman, 1982.

[6] P. Fitzpatrick and G. Metta, "Grounding vision through experimental manipulation," *Philosophical Transactions of the Royal Society A Mathematical Physical and Engineering Sciences*, vol. 361, no. 1811, pp. 2165–2185, Oct. 2003.

[7] E. Spelke, "Principles of object perception," *Cognitive Science*, vol. 14, pp. 29–56, 1990.

[8] J. G. Bremner, *Infancy*. Cambridge, Mass. : Blackwell, 1994.

[9] B. C. L. Touwen, "A study on the development of some motor phenomena in infancy," *Developmental Medicine and Child Neurology*, vol. 13, pp. 435–446, 1971.

[10] D. A. Caruso, "Dimensions of quality in infants' exploratory behavior: Relationships to problem-solving ability," *Infant Behavior and Development*, vol. 16, no. 4, pp. 441–454, Oct 1993.

[11] E. J. Gibson, "Perceptual learning in development: Some basic concepts," *Ecological Psychology*, vol. 12, no. 4, pp. 295–302, Oct 2000.

[12] R. Detry, N. Pugeault, and J. Piater, "A probabilistic framework for 3D visual object representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009.

[13] D. Kraft, R. Detry, N. Pugeault, E. Başeski, J. Piater, and N. Krüger, "Learning objects and grasp affordances through autonomous exploration," in *Computer Vision Systems*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2009, vol. 5815, ch. 24, pp. 235–244.

[14] R. Detry, D. Kraft, A. G. Buch, N. Krüger, and J. Piater, "Refining grasp affordance models by experience," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2010, pp. 2287–2293.

[15] M. Popović, D. Kraft, L. Bodenhagen, E. Başeski, N. Pugeault, D. Kragic, T. Asfour, and N. Krüger, "A strategy for grasping unknown objects based on co-planarity and colour information," *Robotics and Autonomous Systems*, vol. In Press, Corrected Proof, 2010.

[16] N. Pugeault, F. Wörgötter, and N. Krüger, "Accumulated Visual Representation for Cognitive Vision," *In Proceedings of the British Machine Vision Conference (BMVC)*, 2008.

[17] K. Shimoga, "Robot grasp synthesis algorithms: A survey," *Int. J. Robot. Res.*, vol. 15, no. 3, pp. 230–266, 1996.

[18] A. Bicchi and V. Kumar, "Robotic Grasping and Contact: A Review," in *IEEE Int. Conf on Robotics and Automation*, 2000, pp. 348–353.

[19] A. Saxena, J. Driemeyer, and A. Y. Ng, "Robotic grasping of novel objects using vision," *The International Journal of Robotics Research*, vol. 27, no. 2, pp. 157–173, Feb 2008.

[20] C. de Granville, J. Southerland, and A. H. Fagg, "Learning grasp affordances through human demonstration," in *Proceedings of the International Conference on Development and Learning (ICDL'06)*, 2006.

[21] L. Montesano and M. Lopes, "Learning grasping affordances from local visual descriptors," in *Development and Learning, 2009. ICDL 2009. IEEE 8th International Conference on*. IEEE, Jun 2009.

[22] E. Chinellato, A. Morales, E. Cervera, and A. P. del Pobil, "Symbol grounding through robotic manipulation in cognitive systems," *Robotics and Autonomous Systems*, vol. 55, no. 12, pp. 851–859, Dec 2007.

[23] E. Sahin, M. Cakmak, M. Dogar, E. Ugur, and G. Ucoluk, "To afford or not to afford: A new formalization of affordances toward affordance-based robot control," *Adaptive Behav.*, vol. 15, no. 4, pp. 447–472, December 2007.

[24] J. Modayil and B. Kuipers, "Bootstrap learning for object discovery," *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, vol. 1, pp. 742–747, 2004.

[25] ——, "Autonomous development of a grounded object ontology by a learning robot," in *Proceedings of the AAAI Spring Symposium on Control Mechanisms for Spatial Knowledge Processing in Cognitive/Intelligent Systems*. AAAI, 2007.

[26] A. Stoytchev, "Toward learning the binding affordances of objects: A behavior-grounded approach," in *Proceedings of AAAI Symposium on Developmental Robotics*, 2005, pp. 17–22.

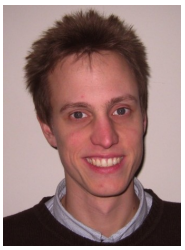
[27] ——, "Learning the affordances of tools using a behavior-grounded approach," in *Towards Affordance-Based Robot Control*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2008, vol. 4760, ch. 10, pp. 140–158.

[28] "Video 'grasping without object knowledge'." [Online]. Available: http://www.mip.sdu.dk/covig/videos/graspingReflexCompressed.divx

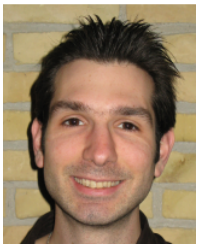
[29] "Video 'grasping and accumulation'." [Online]. Available: http://www.mip.sdu.dk/covig/videos/graspingAndAccumulationNew.wmv

[30] R. P. A. Petrick, D. Kraft, N. Krüger, and M. Steedman, "Combining cognitive vision, knowledge-level planning with sensing, and execution monitoring for effective robot control," in *Proceedings of the Fourth Workshop on Planning and Plan Execution for Real-World Systems at ICAPS 2009*, Thessaloniki, Greece, Sep. 2009, pp. 58–65.
[31] R. S. Siegler, *Emerging Minds : The Process of Change in Children's Thinking*. Oxford University Press, 1998.
[32] P. Willatts, "Development of problem-solving strategies in infancy," in *Childrens Strategies: Contemporary Views of Cognitive Development*, D. Bjorklund, Ed. Lawrence Erlbaum, 1990.
[33] A. Baddeley, *Working Memory, Thought, and Action*. Oxford University Press, 2007.
[34] E. Averbach and G. Sperling, *Information Theory*. Butterworths, 1961, ch. Short term storage of information in vision, pp. 196–211.
[35] J. Piaget, *The Origins of Intelligence in Children*. London: Routledge & Kegan Paul, 1936, (French version published in 1936, translation by Margaret Cook published 1952).
[36] 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.
[37] N. Pugeault, F. Wörgötter, and N. Krüger, "Visual primitives: Local, condensed, and semantically rich visual descriptors and their applications in robotics," *International Journal of Humanoid Robotics (Special Issue on Cognitive Humanoid Vision)*, accepted.
[38] E. Başeski, N. Pugeault, S. Kalkan, L. Bodenhagen, J. H. Piater, and N. Krüger, "Using Multi-Modal 3D Contours and Their Relations for Vision and Robotics," *Journal of Visual Communication and Image Representation*, vol. (accepted), 2010.
[39] M. Ciocarlie and P. Allen, "Hand posture subspaces for dexterous robotic grasping," *Int. J. Rob. Res.*, vol. 28, no. 7, pp. 851–867, 2009.
[40] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.
[41] E. B. Sudderth, "Graphical models for visual object recognition and tracking," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, USA, 2006.
[42] B. W. Silverman. Chapman & Hall, 1986.
[43] J. J. Gibson, *The Ecological Approach to Visual Perception*. Boston, MA: Houghton Mifflin, 1979.
[44] D. Kraft, N. Pugeault, E. Başeski, M. Popović, D. Kragic, S. Kalkan, F. Wörgötter, and 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*, vol. 5, pp. 247–265, 2009.
[45] "Video 'learning grasp affordance densities'." [Online]. Available: http://www.mip.sdu.dk/covig/videos/grasp_densities.avi
[46] S. Pinker, *How the Mind Works*. W. W. Norton & Company, 1999.
[47] T. N. Wiesel and D. H. Hubel, "Ordered arrangement of orientation columns in monkeys lacking visual experience," *The Journal of Comparative Neurology*, vol. 158, no. 3, pp. 307–318, Dec 1974.
[48] I. Gödecke and T. Bonhoeffer, "Development of identical orientation maps for two eyes without common visual experience," *Nature*, vol. 379, pp. 251–255, 1996.
[49] M. Sur, P. Garraghty, and A. Roe, "Experimentally induced visual projections into auditory thalamus and cortex," *Science*, vol. 242, no. 4884, pp. 1437–1441, Dec 1988.
[50] B. Chapman, M. P. Stryker, and T. Bonhoeffer, "Development of orientation preference maps in ferret primary visual cortex," *J. Neurosci.*, vol. 16, no. 20, pp. 6443–6453, 1996.
[51] B. Olshausen and D. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.
[52] 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*, vol. 131, pp. 82–147, 2004.
[53] A. Yonas, L. Pettersen, and C. E. Granrud, "Infants' sensitivity to familiar size as information for distance," *Child Development*, vol. 53, no. 5, pp. 1285–1290, Oct 1982.
[54] P. Kellman and M. Arterberry, *The Cradle of Knowledge*. MIT-Press, 1998.
[55] R. Held, E. Birch, and J. Gwiazda, "Stereo acuity in human infants," *Proceedimgs of the Natianal Acadamy of Sciences, USA*, vol. 77, no. 9, pp. 5572–5574, Sep 1980.
[56] N. Chater, J. Tenenbaum, and A. Yuille, "Probabilistic models of cognition: Conceptual foundations," *Trends in Cognitive Sciences*, vol. 10, no. 7, pp. 287–291, Jul 2006.
[57] W. Köhler, *The Mentality of Apes*. New York: Harcourt Brace and World, 1925.
[58] K. C. Soska and S. P. Johnson, "Development of three-dimensional object completion in infancy," *Child Development*, vol. 79, no. 5, pp. 1230–1236, Sep 2008.
[59] K. C. Soska, K. E. Adolph, and S. P. Johnson, "Systems in development: Motor skills acquisition facilitates 3D object completion," *Developmental Psychology*, vol. 46, no. 1, pp. 129–138, 2010.
[60] J. Piaget, *La contruction du reel chez l'enfant*. Neuchatel: Delachaux and Niestle, 1937.
[61] I. C. Uzgiris and J. M. Hunt, *Assessment in infancy : ordinal scales of psychological development*. Urbana : University of Illinois Press, 1975.
[62] A. Gordon, "Development of the reach to grasp movement," in *Insights into the reach to grasp movement, Advances in Psychology, 105*, K. Bennett and U. Castiello, Eds. Elsevier, 1994.
[63] E. Oztop, N. S. Bradley, and M. A. Arbib, "Infant grasp learning: a computational model," *Exp. Brain Res.*, vol. 158, pp. 480–503, 2004.
[64] D. C. Witherington, "The development of prospective grasping control between 5 and 7 months: A longitudinal study," *Infancy*, vol. 7, no. 2, pp. 143–161, Apr 2005.
[65] A. Streri and J. Féron, "The development of haptic abilities in very young infants: From perception to cognition," *Infant Behavior and Development*, vol. 28, no. 3, pp. 290–304, Sep 2005.
[66] C. von Hofsten and E. S. Spelke, "Object perception and object-directed reaching in infancy." *Journal of Experimental Psychology: General*, vol. 114, no. 2, pp. 198–212, 1985.
[67] J. Fagard, "Linked proximal and distal changes in the reaching behavior of 5- to 12-month-old human infants grasping objects of different sizes," *Infant Behavior and Development*, vol. 23, no. 3-4, pp. 317–329, Mar 2000.
[68] L. B. Smith, "From fragments to geometric shape: Changes in visual object recognition between 18- and 24- months," *Current Directions in Psychological Science*, vol. 18, no. 5, pp. 290–294, Oct 2009.
[69] T. Wilcox, "Object individuation: Infants' use of shape, size, pattern, and color," *Cognition*, vol. 72, no. 2, pp. 125–166, Sep 1999.
[70] L. B. Cohen and C. H. Cashon, "Infant perception and cognition," in *Comprehensive handbook of psychology. Volume 6, Developmental Psychology. II. Infancy*, R. Lerner, A. Easterbrooks, and J. Mistry, Eds. New York: Wiley and Sons, 2003, pp. 65–89.
[71] J. Kaufman, D. Mareschal, and M. H. Johnson, "Graspability and object processing in infants," *Infant Behavior and Development*, vol. 26, no. 4, pp. 516–528, Dec 2003.
[72] M. Carpenter, K. Nagell, and M. Tomasello, "Social cognition, joint attention, and communicative competence from 9 to 15 months of age," *Monographs of the Society for Research in Child Development*, vol. 63, no. 4, pp. i–174, 1998.
[73] D. R. Faria, J. Prado, P. Drews Jr., and J. Diasozto, "Object shape retrieval through grasping exploration," in *Proc.of 4th European Conference on Mobile Robots, ECMR'09*, Mlini/Dubrovnik, Croatia, September 2009.

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 an assistant professor in 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.



Renaud Detry received the engineering degree from the University of Liège in 2006. He is currently working toward the PhD degree at the University of Liège, under the supervision of Professor Justus Piater on multimodal object representations for autonomous, sensorimotor interaction. He has contributed probabilistic representations and methods for multisensory learning that allow a robotic agent to infer useful behaviors from perceptual observations. He is a student member of the IEEE.



Nicolas Pugeault obtained an MSc. from the University of Plymouth in 2002, an Engineer degree from the Ecole Supérieure d'Informatique, Électronique, Automatique (Paris) in 2004, and his PhD from the University of Göttingen in 2008, and is currently a member of the Centre for Vision, Speech and Signal Processing, at the University of Surrey (UK). His research interests include cognitive systems, machine learning and computer vision.

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.

Frank Guerin obtained his Ph.D. degree from Imperial College, London, in 2002. Since August 2003, he has been a Lecturer in Computing Science at the University of Aberdeen. He is interested in infant sensorimotor development, especially means-end behaviour and precursors to tool use. He has previously been interested in multi-agent systems, and game-theoretic aspects thereof, as well as agent communication languages. Dr. Guerin is a member of The Society for the Study of Artificial Intelligence and Simulation of Behaviour, where he has served as a committee member and co-chair of the annual convention.

Justus H. Piater received the PhD degree from the University of Massachusetts, Amherst, in 2001, where he held a Fulbright graduate student fellowship and, subsequently, spent two years on a European Marie Curie individual fellowship at INRIA Rhône-Alpes, France, before joining the University of Liège, Belgium, where he is a professor of computer science and directs the Computer Vision Research Group. His research interests include computer vision and machine learning, with a focus on visual learning, closed-loop interaction of sensorimotor systems, and video analysis. He has published more than 80 technical papers in scientific journals and at international conferences. He is a member of the IEEE Computer Society.

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