2010 IEEE International Conference on Robotics and Automation
Anchorage Convention District
May 3-8, 2010, Anchorage, Alaska, USA

Refining Grasp Affordance Models by Experience

Renaud Detry, Dirk Kraft, Anders Glent Buch, Norbert Krüger, Justus Piater

Abstract— **We present a method for learning object grasp affordance models in 3D from experience, and demonstrate its applicability through extensive testing and evaluation on a realistic and largely autonomous platform. *Grasp affordance* refers here to relative object-gripper configurations that yield stable grasps. These affordances are represented probabilistically with *grasp densities*, which correspond to continuous density functions defined on the space of 6D gripper poses. A grasp density characterizes an *object*'s grasp affordance; densities are linked to visual stimuli through registration with a visual model of the object they characterize. We explore a batch-oriented, experience-based learning paradigm where grasps sampled randomly from a density are performed, and an importance-sampling algorithm learns a refined density from the outcomes of these experiences. The first such learning cycle is bootstrapped with a grasp density formed from visual cues. We show that the robot effectively applies its experience by downweighting poor grasp solutions, which results in increased success rates at subsequent learning cycles. We also present success rates in a practical scenario where a robot needs to repeatedly grasp an object lying in an arbitrary pose, where each pose imposes a specific reaching constraint, and thus forces the robot to make use of the entire grasp density to select the most promising achievable grasp.**

I. Introduction

In cognitive robotics, the concept of affordances [7], [14] characterizes the relations between an agent and its environment through the effects of the agent's actions on the environment. Affordances have become a popular formalization for cognitive control processes, while bringing valuable insight on how cognitive control can be done. Within the field of robotic grasping, methods formalized as *grasp affordances* have recently emerged [2], [21], [3], [11]. Grasp affordances generally allow for an assessment of the success (effect) of a grasp solution (action) on a particular object (environment).

Grasping skills can be programmed into a robot in many different ways, starting with completely hard-wired kinematics, and ranging over a wide variety of methods of increasing autonomy and adaptivity. Amongst these, providing a robot with the means of learning grasping skills *from experience* appears particularly appealing – even beyond the conveniently autonomous aspect of the process: First, in performing manipulation tasks, a robot produces valuable information about its environment, and making use of that information seems only natural. Secondly, learning from experience directly involves the body of the robot, therefore producing a model intimately adapted to its morphology.

R. Detry and J. Piater are with the University of Liège, Belgium. Email: Renaud.Detry@ULg.ac.be.

D. Kraft, A. G. Buch and N. Krüger are with the University of Southern Denmark.

A generally accepted aspect of the theory of affordances is that it relates the opportunities provided by the environment to the abilities of the agent, instead of expressing a property of the environment alone [17], [14]. Learning from experience thus appears as a natural way of discovering grasp affordances. The main contribution of this paper is the application of a method for learning grasp affordances probabilistically from experience [3] and its thorough evaluation. Evaluation is conducted on a realistic, largely autonomous platform, through the collection of a large grasp dataset – more than 2000 grasps tested on a robot.

In this work, affordances express relative object-gripper configurations that yield stable grasps. They are represented probabilistically with *grasp densities* [3], which correspond to continuous density functions defined on the space of 6D gripper poses $SE(3)$. A grasp density characterizes an *object*'s grasp affordance; densities are linked to visual stimuli through registration with a visual model of the object they characterize.

Grasp densities are learned and refined through experience. Intuitively, the robot "plays" with an object in a sequence of grasp-and-drop actions. Grasps are selected randomly from the object's grasp density. After each (successful) grasp, the object is dropped to the floor. When a satisfying quantity of data is available, an importance-sampling algorithm [5] produces a refined grasp density from the outcomes of the set of executed grasps. Learning is thus organized in *cycles* of batches of grasps.

In theory, the grasp density of an object that has never been grasped could be initialized to a uniform distribution. Unfortunately, the success rate of completely random grasps is extremely low and cannot allow for reasonable learning time; in the experiments presented in this paper, initial grasp densities are bootstrapped from visual cues. Throughout the paper, these densities constructed from visual cues will be called *bootstrap densities*. By contrast, densities which are the result of experience will be referred to as *empirical densities*. Within each learning cycle, the density used by the robot to produce grasps will be called *hypothesis density*. In the first cycle, the hypothesis density is a bootstrap density. In subsequent cycles, the hypothesis density will typically correspond to the empirical density learned during the previous cycle.

Experiments are run on the robotic platform of Fig. 1. A simple control algorithm drives the grasp-and-drop protocol on the robot. The pose of the object is recovered by visual pose estimation on the imagery provided by the camera, using a previously-learned visual model [4]. Path planning automatically excludes most of the grasps that would pro-

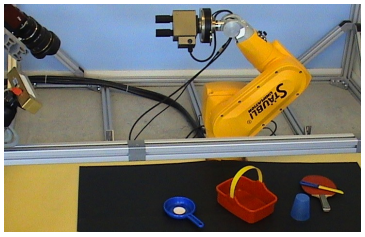


Fig. 1: Experiment platform (industrial arm, force-torque sensor, two-finger gripper, stereo camera, foam floor)

duce a collision with the ground, and some of the grasps that would collide with the object. Success is assessed by measuring the distance between the fingers of the gripper after the arm has been brought up. The resulting system is largely autonomous and forms a realistic setup. We show that the robot successfully exploits its actions to downweight poor grasp solutions, which is reflected in the higher success rates achieved in subsequent learning cycles. We finally quantify the success rate of our method in a practical scenario where a robot needs to repeatedly grasp an object lying in an arbitrary pose, where each pose imposes specific reaching constraints, and thus forces the robot to make use of the entire grasp density to select the most promising grasp within the achievable region.

II. Related Work

As discussed above, learning grasp affordances from experience has become an interesting and popular paradigm. In the work of A. Stoytchev [18], [19], a robot discovers successful ways of grasping tools through random exploratory actions. When subsequently confronted with the same object, the robot is able to generate a grasp that should present a high likelihood of success. Montesano et al. [11] learned 2D continuous and probabilistic grasp affordance models for a set of objects of varying shape and appearance, and developed means of qualifying the reliability of their grasp predictions. Detry et al. [3] presented a method for learning continuous and probabilistic grasp affordance models in 3D along with preliminary experimental results.

We note that one problem in learning from experience is that it is usually slow. The main alternative to learning from experience is learning from a human teacher [2], [15], which is typically much faster. However, with this paradigm, the model is not necessarily adapted to the robot morphology.

A large body of literature on learning how to grasp focuses on methods that produce a number of *discrete* grasping solutions [15], [1]. A few recent methods instead explicitly aim at producing a *continuous*, probabilistic characterization of the grasping properties of an object [2], [3], [11]. The latter can naturally be used to produce grasping solutions; additionally, they allow for *ranking* grasps, i.e. provide a likelihood of success for an arbitrary grasp.

In learning a continuous characterization of object grasping properties probabilistically, one has a choice between learning success probabilities or learning success-conditional grasp densities. Denoting by O a random variable encoding grasp outcomes (success or failure), and by G a random variable encoding grasp poses, this translates to learning $P(O = \text{success}|G)$ or learning $P(G|O = \text{success})$. The former allows one to directly compute a probability of success; it will generally be learned through discriminative methods from positive and negative examples (successful and failed grasps). The latter allows for grasp sampling, while still providing direct means of computing *relative* success probabilities – e.g. grasp a is twice as likely to succeed as grasp b. Success-conditional grasp densities are generative models computed from positive examples only. We note that one can theoretically be computed from the other using Bayes' rule. However, depending on the means of function representation, this process may prove either too costly or too noisy to be feasible in practice.

The learning of success-conditional grasp densities has been discussed in the work of de Granville et al. [2], where grasp densities are defined on hand approach orientations. Instead of considering success-conditional grasp probabilities, Montesano et al. [11] model grasp affordances as success probabilities. Formally, they learn a representation of $P(O|I)$, where I is a local image patch. A grasp action consists in servoing the robot hand to a selected 2D position, approaching the object from the top until contact is made, and closing the hand. A robot thus learns a mapping from 2D image patches to grasp success probabilities, where a grasp is parametrized by its 2D hand position.

The most important application of grasping research is in generating a grasping solution from visual percepts of an object. Grasping research may thus be pertinently classified on the relationship a method entertains with visual perceptions. In the field of robotics, *visual perception* encompasses a wide spectrum of representations: At the lower level, a scene may be described in terms of a large number of point elements, such as image pixels or depth maps. At the other end of the spectrum, a scene may be represented by instances of object models and their 2D or 3D poses. The gap between these two extremes is filled, bottom-up, by visual features of varying size and complexity, and, top-down, by object models that are recursively formed of visual parts of decreasing size and complexity. Intuitively, grasping methods that link to lower-level visual percepts can easily generalize across objects. These methods typically learn a continuous mapping from local visual descriptors to the probability of success of a grasp [11]. Grasp parameters are deduced from the visual descriptor, e.g. 2D grasping coordinates from the 2D position of the descriptor [11], or 3D grasp position from stereo matching of 2D grasping points [15]. On the other hand, methods that link to higher-level visual entities benefit from an increased geometric robustness. These will generally allow the encoding of richer grasp parameters such as 3D relative position and orientation. They typically learn a mapping from objects to grasp parameters and grasp probabilities [2], [3]; grasps are registered with the visual object model. They are aligned to an object pose through visual pose estimation.

In this paper, we develop and evaluate a method for

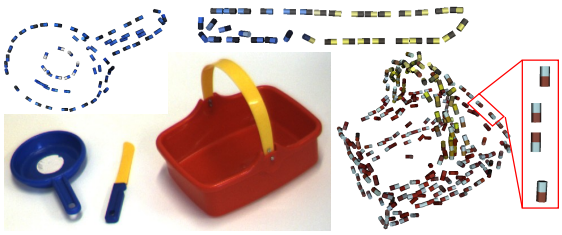


Fig. 2: ECV (accumulated) reconstructions. Each cylinder corresponds to an ECV descriptor. The axis of a cylinder is aligned with the direction of the modeled edge. Each cylinder bears the two colors found on both sides of the edge in 2D images.

learning by experience continuous tree-dimensional success-conditional grasp densities [3]. Densities encode object grasps; they are linked to a high-level object model. The contribution of this paper include original means of bootstrapping densities, a technique for exploiting local density maxima, and a thorough evaluation of the resulting system through a realistic robot setup and scenario. By that, we can demonstrate the full potential of the concept of grasp densities in a setting in which a motion planner interacts with them. In particular we show that besides increasing the success rate for physically executed grasps, the number of grasps to be tested by the motion planner reduces by a factor of ten.

III. Vision

This section presents the vision methods used in this work. We first introduce 3D object-edge reconstructions which are used in Section V for bootstrapping densities. These reconstructions also serve as a basis for a hierarchical object model introduced next.

A. Early Cognitive Vision (ECV)

ECV descriptors [10], [13] represent short edge segments in 3D space, each ECV descriptor corresponding to a circular image patch with a 7-pixel diameter. They are computed by combining 2D edge extraction in image pairs and stereopsis across the pairs. Each descriptor is defined by a position (3 degrees of freedom – DOF) and edge tangent (2 DOF), therefore living in $\mathbb{R}^3 \times S^2_+$ where S^2_+ is a 2-hemisphere. Descriptors may be tagged with color information, extracted from their corresponding 2D patches (Fig. 2).

ECV reconstructions can further be improved by manipulating objects with a robot arm, and *accumulating* visual information across several views through structure-from-motion techniques [8]. Assuming that the motion adequately spans the object pose space, a complete 3D reconstruction of the object can be generated, eliminating self-occlusion issues [9] (see Fig. 2).

B. Pose Estimation

The visual models we use for pose estimation have the form of a hierarchy of increasingly expressive object parts [4], where bottom-level parts correspond to generic ECV descriptors. Visual inference of the hierarchical model is performed using a belief propagation algorithm (BP) [12], [20], [4]. BP derives a probabilistic estimate of the object pose, which in turn allows for the alignment of the grasp model to the object. Means of autonomously learning the hierarchical model and the underlying accumulated ECV reconstruction are presented in prior work [4], [9].

IV. Grasp Densities

This section explains how grasp densities probabilistically model grasp affordances, and how importance sampling is used to learn empirical densities.

A. Mathematical Representation

We are interested in modeling object-relative gripper configurations. The grasps we consider are thus parametrized by a 6D gripper pose composed of a 3D position and a 3D orientation. Grasp densities are continuous probability density functions defined on the 6D pose space $SE(3)$; they model the likelihood of success of any grasp $x \in SE(3)$. Their computational representation is nonparametric: A density is represented by a large number of weighted samples called *particles*. The probabilistic density in a region of space is given by the local density of the particles in that region. The underlying continuous density function is accessed through *kernel density estimation* [16], by assigning a kernel function to each particle supporting the density. Density evaluation at a given pose x is performed by summing the evaluation of all kernels at x. Sampling from a distribution is performed by sampling from the kernel of a particle drawn at random.

Grasp densities are defined on the Special Euclidean group $SE(3) = \mathbb{R}^3 \times SO(3)$, where $SO(3)$ is the Special Orthogonal group (the group of 3D rotations). We use a kernel that factorizes into two functions defined on $\mathbb{R}^3$ and $SO(3)$. Denoting the separation of an $SE(3)$ pose x into a translation λ and a rotation θ by $x = (\lambda, \theta)$, $\mu = (\mu_t, \mu_r)$, $\sigma = (\sigma_t, \sigma_r)$, we define our kernel with

$$\mathbf{K}(x; \mu, \sigma) = \mathbf{N}(\lambda; \mu_t, \sigma_t)\, \mathbf{\Theta}(\theta; \mu_r, \sigma_r) \tag{1}$$

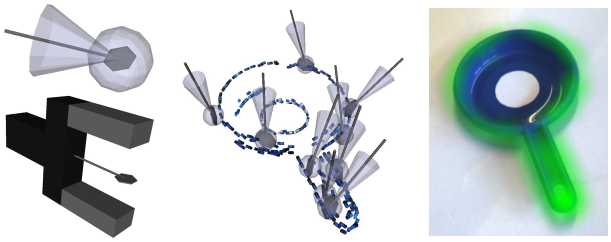
where μ is the kernel mean point, σ is the kernel bandwidth, $\mathbf{N}$ is a trivariate isotropic Gaussian kernel, and $\mathbf{\Theta}$ corresponds to a pair of antipodal von-Mises Fisher distributions which forms a Gaussian-like distribution on $SO(3)$ [6], [20]. Formally, the value of $\mathbf{\Theta}$ is given by

$$\mathbf{\Theta}(\theta; \mu_r, \sigma_r) = C_4(\sigma_r) \frac{e^{\sigma_r\, \mu_r^T \theta} + e^{-\sigma_r\, \mu_r^T \theta}}{2} \tag{2}$$

where $C_4(\sigma_r)$ is a normalizing constant.

The position bandwidth σ_t is fixed to 10 mm; the orientation bandwidth σ_r allows rotations of about 5°. For a more detailed mathematical description and motivation of $SE(3)$ kernels and kernel density estimation, we refer the reader to the work of Sudderth et al. [20] and Detry et al. [4]. Fig. 3 illustrates $SE(3)$ kernels and continuous densities.

Grasp densities are defined in the same reference frame as visual features. Once visual features have been aligned to an object pose (Section III-B), the object grasp density can



(a) (b) (c)

Fig. 3: Grasp density representation. The top image of Fig. (a) illustrates a particle from a nonparametric grasp density, and its associated kernel widths: the translucent sphere shows one position standard deviation, the cone shows the variance in orientation. The bottom image illustrates how the schematic rendering used in the top image relates to a physical gripper. Fig. (b) shows a 3D rendering of the kernels supporting a grasp density for a toy pan (for clarity, only ten kernels are rendered). In Fig. (c), the opacity of the green mask is proportional to the value of a grasp density for the pan (orientations were ignored for this 2D projection).

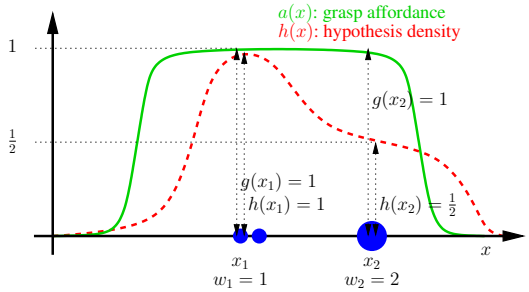


Fig. 4: 1D illustration of the importance-sampling weight computation. Although grasps such as x_2 are less likely to be executed than grasps like x_1, the weight associated to x_2 is twice as large as that associated to x_1.

be similarly aligned, and one can readily draw grasps from the density and execute them on the object. The association of grasp densities with the visual model is covered in more detail in prior work [3].

B. Learning Algorithm

Learning is organized in cycles, within each of which the robot exploits its current grasping knowledge and importance sampling [5] to produce a refined empirical density. Importance sampling is a technique that allows one to draw samples from an unknown *target* distribution by properly weighting samples from a preferably similar *proposal* distribution. The target distribution $t(x)$ cannot be sampled from, but it can be evaluated. Therefore, samples are drawn from the known proposal distribution $p(x)$, and the difference between the target and the proposal is accounted for by associating to each sample x a weight given by $t(x)\,/p(x)$.

Let us model with $g(x)$ the outcome of grasp x as

$$g(x) = \begin{cases} 1 & \text{if grasp at } x \text{ succeeds,} \\ 0 & \text{if grasp at } x \text{ fails.} \end{cases}$$

In the context of this paper, the target distribution corresponds to a perfect model of the object grasp affordance $a(x)$. An empirical density could be build from a set of samples from a; yet sampling a cannot be done directly. However, by approximating $a(x) \simeq g(x)$, we can produce binary evaluations of $a(x)$ by executing grasps. Importance sampling thus allows us to indirectly draw samples from $a(x)$, by drawing samples from a hypothesis density $h(x)$, and weighting each sample $\hat{x}$ as $g(\hat{x})/h(\hat{x})$. Fig. 4 illustrates the concept of importance sampling in a simple one-dimensional case.

V. Experiments

This section demonstrates the applicability of our method to learning empirical densities, quantifies the efficacy of the various learning cycles, and estimates the efficacy of empirical densities in a typical grasping scenario.

Section V-A explains the process of executing a set of grasp *trials*, and details the nature of recorded data. Section V-B presents the application of this process for both learning empirical densities and estimating their efficacy in practical scenarios. Results are discussed in Section V-C.

A. Setup

Our robotic platform is composed of an industrial robotic arm, a force-torque sensor, a two-finger gripper, and a stereo camera. The force-torque sensor is mounted between the arm and the gripper. The arm and the camera are calibrated to a common world reference frame. The execution of a set of grasp trials is driven by a finite state machine (FSM), which instructs the robot to grasp and lift an object, then drop the object to the floor and start again. The floor around the robot is covered with foam, which allows objects to lightly bounce during drop-off. This also allows the gripper to push slightly into the floor and grasp thin objects lying on the foam surface (e.g. the knife of Fig. 2).

The FSM is initially provided with an object model, which includes a visual model as described in Section III-B, and a grasp density registered with the visual model. The FSM then performs a set of grasp trials, which involve the following operations:

i. Estimate the pose of the object and align the grasp density,
ii. Produce a grasp from the aligned grasp density,
iii. Submit the grasp to the path planner,
iv. Servo the gripper to the grasp pose,
v. Close the gripper fingers,
vi. Lift the object,
vii. Drop the object.

Pose estimation (i) is performed by means detailed in Section III-A and Section III-B. Depending on the purpose, grasps (ii) are drawn either randomly, or from a local maximum of the density.

The path planner has a built-in representation of the floor and the robot body. Its representation of the floor is defined a few centimeters below the foam surface, to allow the gripper

to grasp thin objects as explained above. The planner is provided with a gripper pose (ii) and the 3D scene reconstruction extracted during pose estimation (i). Because the 3D scene reconstruction does not cover self-occluded parts of the object, a 3D *accumulated* reconstruction of the object is also provided (Section III-A). The path planner computes a collision-free path to the target gripper configuration. It can avoid self-collisions and most ground collisions from its built-in knowledge of the arm and workspace; it can also avoid some object collisions from the 3D-edge scene reconstruction and the aligned object reconstruction. When no path can be found, the path planner is able to produce a detailed error report.

During servoing (iv) and grasping (v), measures from the force-torque sensor are compared to a model of the arm dynamics, allowing for automatic collision detection. Closure success is verified after grasping (v) by measuring the gap between the fingers, and after lifting (vi) by checking that the fingers cannot be brought closer to each other. The object is finally dropped to the floor from a height of about 50 cm and bounces off to an arbitrary pose.

Robot assessments are monitored by a human supervisor. Pose estimation will sometime fail, e.g. because the object fell out of the field of view of the camera, or because of a prohibitive level of noise in the stereo signal. Pose estimates are visualized in 3D; if pose estimation fails, the trial is aborted and the supervisor moves the object to another arbitrary pose. After path planning, the supervisor has a chance to abort a grasp that would clearly fail. During servo, grasp and lift, he can notify undetected collisions. Despite this supervision, the resulting system is largely autonomous: The role of the supervisor is limited to notifying wrong robot assessments; pose estimates and grasps are never tuned by hand.

If the robot properly executes the operations mentioned above and lifts the object, the trial is a success. When an operation produces an error, the trial is a failure, and the FSM starts over at step ii, or at step i if the error involved an object displacement. Errors can come from a pose estimation failure, no found path, supervisor notification of bound-to-fail grasp, collision (notified either from the force-torque sensor or from the supervisor), or empty gripper (v and vi). We define two mutually-exclusive error classes. The first class, denoted by E_p, includes errors arising from a path-planner–predicted collision with the ground or the object. The second class, E_r, correspond to object collisions, ground collisions, or void grasps, either asserted by the supervisor, or physically occurring on the robot. Errors E_r also include cases where the object drops off the gripper during lift-up. The FSM keeps track of errors by counting the number of occurrences e_r and e_p of errors of class E_r and E_p. Pose estimation failures and cases where the path planner cannot find an inverse-kinematics solution at all (e.g. object out of reach) are ignored because these are not intrinsically part of the concept of grasp densities. Naturally, the number s of successful grasps is also recorded.

The execution of a complete grasp trial typically takes 40 to 60 seconds. Through the process described above, the robot will effectively learn grasp affordances offered by an object lying on a flat surface in a natural pose.

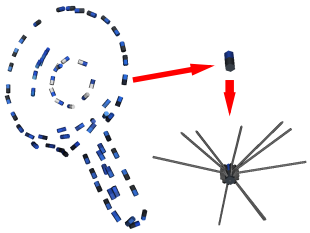


Fig. 5: Bootstrapping grasp densities from ECV descriptors

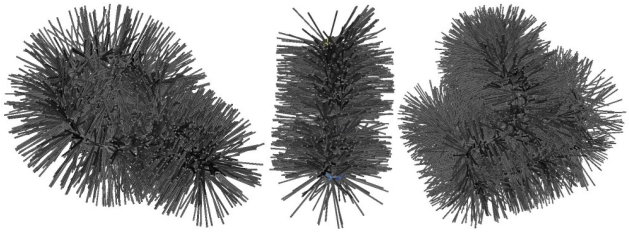


Fig. 6: Particles supporting bootstrap densities

B. Protocol

This section applies the process of the previous section to learn and evaluate empirical densities. Experiments were run on the three objects of Fig. 2, selected for their differences in shape and structure, which offer a large variety of grasping possibilities. Visual models were acquired by performing a 3D reconstruction of the object edges (Section III-A), and organizing the resulting ECV descriptors in a hierarchy (Section III-B).

Grasp densities were bootstrapped from the ECV reconstructions of the objects, through a process that is intentionally kept simple in order to limit the amount of bias introduced into the system. As explained in Section III-A, an ECV reconstruction represents object edges with short 3D segments. Object edges appeared as natural candidates for grasping, and an interesting way to bias grasp learning. We thus define bootstrap densities as functions yielding a high value around object edges. Bootstrap densities are, just like other densities, represented nonparametrically. They are formed by generating sets of $SE(3)$ particles from ECV descriptors. Mathematically, ECV descriptors live in $\mathbb{R}^3 \times S^2_+$; an ECV descriptor thus cannot fully define an $SE(3)$ grasp. Therefore, we create a *set* of $SE(3)$ particles for each ECV descriptor, effectively covering the ECV orientation degree of freedom (See Fig. 5 and Fig. 6).

One known weakness of importance sampling is its slow convergence when the target distribution has heavier tails than the proposal. This is unfortunately the case with bootstrap densities, since they only cover a part of the object affordance – pinch grasps are applicable to parts that are not supported by a visual edge. For this reason, we slightly modify the importance-weight computation (Section IV-B), effectively using $g(x)/\left(h(x)+C\right)$ instead of $g(x)/h(x)$,

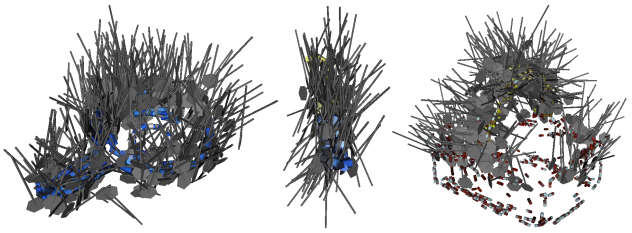


Fig. 7: Samples from empirical densities learned during the first cycle

Batch		s	e_r	e_p	r_{rp}	r_r
Pan	*cycle 1*	200	370	1631	0.091	0.351
	cycle 2	100	86	114	0.333	0.538
	test	75	39	24	0.543	0.658
Knife	*cycle 1*	100	131	751	0.102	0.433
	cycle 2	100	153	157	0.244	0.395
	test	63	71	89	0.283	0.470
Basket	*cycle 1*	151	173	1121	0.104	0.466
	cycle 2	100	62	77	0.418	0.617
	test	64	26	22	0.571	0.711

TABLE I: Success/error counts and success rates. (See also Fig. 8.)

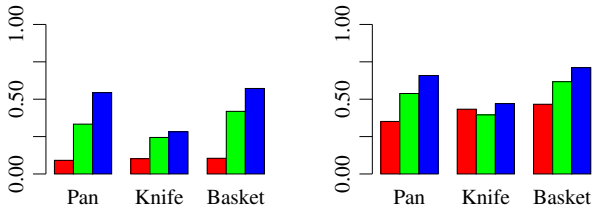
which amounts to using a hypothesis density that contains a uniform component of value C from which grasps always fail.

To each object of Fig. 2, we applied two learning cycles. In the first cycle, the robot uses the object bootstrap density b as hypothesis to learn an empirical density g_1. In the second cycle, the hypothesis density corresponds to g_1, and the robot learns a second empirical density g_2. The purpose of the second cycle is to provide a quantitative comparison of the grasping knowledge expressed by bootstrap and empirical densities through the success statistics of both processes; g_2 is not used thereafter.

We tested the performance of our method in a usage scenario in which it has to successively allow a robot to perform the grasp that has the highest likelihood of success within the robot's region of reachability. However, expressing the region of $SE(3)$ that the robot can reach is not trivial, and goes beyond the scope of this paper. Our usage scenario thus implements each grasp trial by randomly drawing a set of grasps from an empirical density, and sorting these grasps in decreasing order of likelihood according to that empirical density. The grasps are sequentially submitted to the path planner and the first feasible grasp is selected. The empirical density used in the usage scenario is g_1, in order to provide a direct comparison with the statistics collected during the second learning cycle.

C. Results and Discussion

The empirical densities produced during the first learning cycle are shown in Fig. 7. Comprehensive quantitative results are displayed in Table I. Columns s, e_r, and e_p correspond to the statistics collected during the experiment. The last two



(a) r_{rp} (b) r_r

Fig. 8: Success rates. Red, green, and blue bars respectively illustrate rates for the first cycle, second cycle, and test. Numerical rates can be read from Table I.

columns show success rates computed as

$$r_{rp} = \frac{s}{s + e_r + e_p}, \qquad r_r = \frac{s}{s + e_r}.$$

Rows titled *cycle 1* and *cycle 2* correspond to the first and second learning cycles. Rows titled *test* correspond to the usage scenario defined above. Fig. 8 shows success rates graphically.

Fig. 7 shows that the empirical densities learned in the first cycle are a much better model of grasp affordances than the bootstrap densities of Fig. 6. The global success rate r_{rp} (Fig. 8a) for the two learning cycles provides a quantitative comparison of the grasping knowledge expressed by bootstrap and empirical densities. The empirical densities produced during the first cycle allow the robot to collect, during the second cycle, a similar amount of positive examples with a much smaller number of trials. The red bars in Fig. 8a show that grasps generated from modes of an empirical indeed have a higher chance of success than randomly sampled grasps.

Fig. 8b shows success rates in which planner-detected errors E_p are ignored. Since planner-detected errors largely amount to ground-collisions, Fig. 8b shows that a large portion of the knowledge acquired by the robot models which side of the object most often faces up, hence encouraging the robot to produce grasps approaching to that side. This situation is pushed to the limit with the knife: All grasps suggested by its bootstrap density would effectively work for a free-floating knife, i.e. all grasps that do not collide with the ground have the same chance of success. When ignoring E_p errors, the success rate for the first and second cycles of the knife are almost identical.

Our results make a number of issues explicit. For all objects we can reduce the number of grasps that need to be considered by the motion planner by an average factor of 10. This is an important result, since path planning is generally slow, and ground plane information may not always be available to the planner. The average success rate of grasps performed by the robot (ignoring those rejected by the planner) grows from 42% to 52%. In test scenarios, the success rate of robot grasps is 61% in average. These numbers are quite encouraging, given that we tested our system under realistic settings: Visual models, which are learned autonomously [4], [9], do not exhaustively encode relevant object features. During pose estimation, estimates

that are considered successful are nevertheless affected by errors of the order of 5–10 mm in position and a few degrees in orientation. The path planner approximates obstacles with box constellations that may often be imprecise and over-restrictive. Inverse kinematics can perform only up to the precision of the robot-camera calibration. When grasping near the floor, the force-torque sensor may issue a collision detection for a grasp that has worked before, because of a different approach dynamic. For the pan, and in particular for the knife, we have a very difficult grasping situation, given the short distance between the object and the ground. As a consequence, small errors in pose estimates can lead to collisions even with an optimal grasp. Therefore, the error counts in Table I do not exclusively reflect issues related to grasp densities.

We showed that comprehensive grasp affordance models can be acquired by largely autonomous learning. The concept of grasp densities served as a powerful tool to represent these affordances and exploit them in finding an optimal grasp in a concrete context.

VI. Conclusion And Future Work

We have presented a method for learning *three-dimensional* probabilistic grasp affordance models in the form of grasp densities, and demonstrated their applicability through extensive testing on a largely autonomous platform.

Grasp densities are learned from experience with an importance-sampling algorithm: samples from an object affordance are indirectly drawn by properly weighting samples from an approximating hypothesis density. Densities are bootstrapped from a 3D object-edge reconstructions, yielding bias towards edge grasps.

We assembled an experiment setup which efficiently implements a realistic learning environment: The robot handles objects appearing in arbitrary poses, and deals with the noise inherent to autonomous processes. We have collected a large amount of data quantifying the progress made from bootstrap to empirical densities. We also evaluated empirical densities in a realistic usage scenario, where the robot effectively selects the grasp with the highest success likelihood amongst the grasps that are within its reach. Result are particularly convincing given the low level of control on the overall experiment process.

In this paper, grasp densities characterize *object* grasps; they are registered with a visual model of the object. Yet, affordances generally characterize object-robot relations through a minimal set of properties, which means that object properties not essential to a relation should be left out. This in turn allows e.g. for generalization of affordances between objects that share the same grasp-relevant features. Ultimately, instead of registering densities with a whole object, we aim to relate them to visual object *parts* that predict their applicability. The part-based model of Section III-B offers an elegant way of *locally* encoding visuomotor descriptions, allowing for generalization of grasps across objects that share the same parts.

Acknowledgments

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

References

[1] A. Bicchi and V. Kumar. Robotic grasping and contact: a review. In *IEEE International Conference on Robotics and Automation*, 2000.

[2] Charles de Granville, Joshua Southerland, and Andrew H. Fagg. Learning grasp affordances through human demonstration. In *International Conference on Development and Learning*, 2006.

[3] Renaud Detry, Emre Başeski, Norbert Krüger, Mila Popović, Younes Touati, Oliver Kroemer, Jan Peters, and Justus Piater. Learning object-specific grasp affordance densities. In *International Conference on Development and Learning*, 2009.

[4] Renaud Detry, Nicolas Pugeault, and Justus Piater. A probabilistic framework for 3D visual object representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2009.

[5] A. Doucet, N. de Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer, 2001.

[6] R. A. Fisher. Dispersion on a sphere. In *Proc. Roy. Soc. London Ser. A.*, 1953.

[7] James J. Gibson. *The Ecological Approach to Visual Perception*. Lawrence Erlbaum Associates, 1979.

[8] R.I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[9] 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. *International Journal of Humanoid Robotics*, 5:247–265, 2009.

[10] N. Krüger, M. Lappe, and F. Wörgötter. Biologically Motivated Multimodal Processing of Visual Primitives. *The Interdisciplinary Journal of Artificial Intelligence and the Simulation of Behaviour*, 1(5):417–428, 2004.

[11] L. Montesano and M. Lopes. Learning grasping affordances from local visual descriptors. In *International Conference on Development and Learning*, 2009.

[12] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[13] Nicolas Pugeault. *Early Cognitive Vision: Feedback Mechanisms for the Disambiguation of Early Visual Representation*. Vdm Verlag Dr. Müller, 2008.

[14] Erol Sahin, Maya Cakmak, Mehmet R. Dogar, Emre Ugur, and Gokturk Ucoluk. To afford or not to afford: A new formalization of affordances towards affordance-based robot control. *Adaptive Behavior*, 2007.

[15] A. Saxena, J. Driemeyer, and A. Y. Ng. Robotic Grasping of Novel Objects using Vision. *The International Journal of Robotics Research*, 27(2):157, 2008.

[16] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC, 1986.

[17] T. Stoffregen. Affordances as properties of the animal environment system. *Ecological Psychology*, 15(2):115–134, 2003.

[18] Alexander Stoytchev. Toward learning the binding affordances of objects: A behavior-grounded approach. In *AAAI Symposium on Developmental Robotics*, 2005.

[19] Alexander Stoytchev. Learning the affordances of tools using a behavior-grounded approach. In E. Rome et al., editors, *Affordance-Based Robot Control*, volume 4760 of *Lecture Notes in Artificial Intelligence (LNAI)*, pages 140–158. Springer, Berlin / Heidelberg, 2008.

[20] Erik B. Sudderth. *Graphical models for visual object recognition and tracking*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2006.

[21] J. D. Sweeney and R. Grupen. A model of shared grasp affordances from demonstration. In *International Conference on Humanoid Robots*, 2007.