Appearing in *Paladyn. Journal of Behavioral Robotics*, 2(1):1–17, 2011.
This copy is for personal use only. The original publication is available at
www.springerlink.com.
http://dx.doi.org/10.2478/s13230-011-0012-x

Learning Grasp Affordance Densities

R. Detry, D. Kraft, O. Kroemer, L. Bodenhagen, J. Peters, N. Krüger, and J. Piater

R. Detry is with the Centre for Autonomous Systems, Kungliga Tekniska högskolan (KTH),
Stockholm, Sweden.
Email: detryr@kth.se
Web: http://www.csc.kth.se/~detryr/
D. Kraft, L. Bodenhagen and N. Krüger are with the University of Southern Denmark.
O. Kroemer and J. Peters are with the Darmstadt University of Technology, Germany, and the
MPI for Biological Cybernetics, Tübingen, Germany.
J. Piater is with the University of Innsbruck, Austria.

Abstract

We address the issue of learning and representing object grasp affordance models. We model grasp affordances with continuous probability density functions (*grasp densities*) which link object-relative grasp poses to their success probability. The underlying function representation is nonparametric and relies on kernel density estimation to provide a continuous model. Grasp densities are learned and refined from exploration, by letting a robot "play" with an object in a sequence of grasp-and-drop actions: the robot uses visual cues to generate a set of grasp hypotheses, which it then executes and records their outcomes. When a satisfactory amount of grasp data is available, an importance-sampling algorithm turns it into a grasp density. We evaluate our method in a largely autonomous learning experiment, run on three objects with distinct shapes. The experiment shows how learning increases success rates. It also measures the success rate of grasps chosen to maximize the probability of success, given reaching constraints.

Keywords: Robot learning, grasping, probabilistic models, cognitive robotics.

1 Introduction

Grasping previously unknown objects is a fundamental skill needed by autonomous agents for manipulation. This paper addresses the issue of vision-based grasp learning, i.e., acquiring the capability to compute grasping parameters from visual observations. Visual observations tell the agent about the spatial configuration of the objects that surround it. These observations serve as a basis for computing grasping parameters, i.e., for computing the position and orientation to which the robot must bring its hand in order to robustly grasp an object.

In traditional industrial approaches, vision-based grasping is implemented by designing 3D object models with computer-aided design (CAD), and by manually defining a few model-relative grasp approaches. These models are specific to a single object. By contrast, learning approaches are designed to allow agents to acquire grasping skills on their own, adapting to new objects without re-programming.

The first contribution of this paper is a model that describes the grasping properties of a particular object, for the purpose of reasoning on grasping solutions and their feasibility. Concretely, the model encodes the different ways to place a hand or a gripper near an object so that closing the gripper produces a stable grip.

The second contribution of this paper is a method for learning grasp models from experience. The agent learns by "playing" with objects. It repeatedly tries to grasp an object, and, whenever a grasp succeeds, it drops the object before it tries to grasp it again. By visually observing objects during grasps, the robotic agent is able to extract the relation between object positions and good gripper configurations. As the amount of experience grows, the agent becomes increasingly efficient at computing grasps from visual evidence. Leaning is evaluated on a realistic, largely autonomous platform, through the collection of a large grasp dataset – more than 2000 grasps tested on a robot.

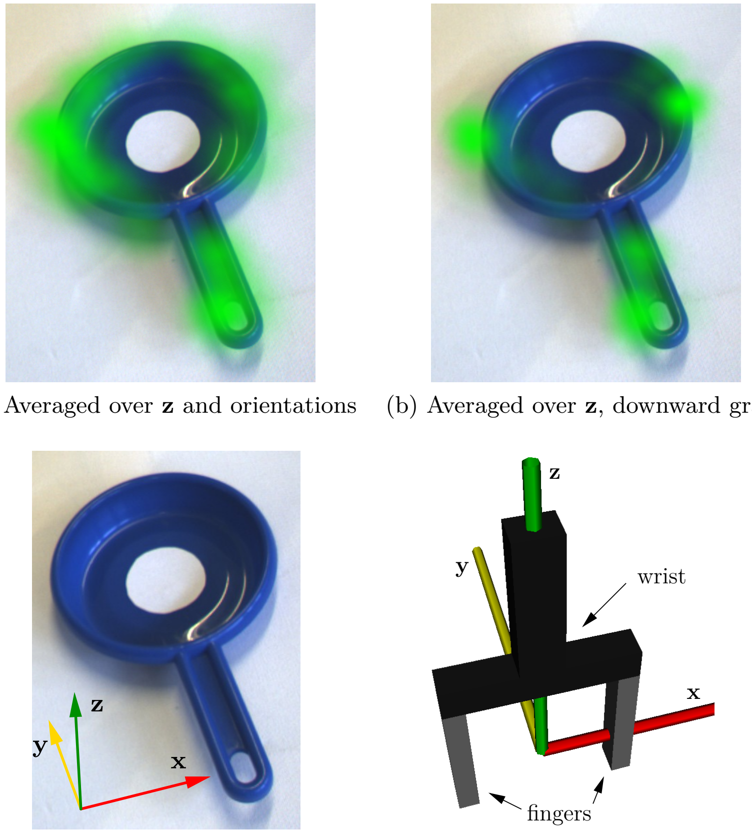
The remainder of this introduction provides a technical overview of our work, with a description of the model in Section 1.1, and a description of the model acquisition in Section 1.2.

1.1 Grasp Affordance Model

In cognitive robotics, the concept of affordances [12, 27] 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 human/autonomous manipulation processes, while bringing valuable insight on how manipulation can be done. Within the field of robot grasping, methods formalized as *grasp affordances* have recently emerged [4, 33, 5, 22]. When modeling grasp affordances, the perception of the environment is most often visual. Agent parameters usually correspond to arm and gripper parameters. Grasp affordance models represent the success (effect) of grasp solutions (action) applied to an object (environment).

This work aims to model object-specific grasp affordances. Also, although grasping can be seen as an effect in itself, objects are generally grasped for a specific purpose. In this paper, as in previous work [5], we focus on affordances for the single task of gripping for lifting up. The affordance model is linked to visual percepts through a visual model of object structure: a visually-recovered estimate of the pose of an object (its 3D position and orientation) aligns the grasp model to the object, and the aligned grasp model then serves as a basis for reasoning on grasp feasibility.

Our model encodes object-relative gripper configurations and the probability of their grasping success. Grasps are parametrized by the 6D pose of the gripper relative to the object. Grasp affordances are represented with probability density functions defined on the 6D space of object-relative gripper poses. These functions, referred to as *grasp densities*, continuously represent relative object-gripper configurations that lead to a successful grasp. Grasp densities allow us to compute, for each gripper-object configuration, the probability of succeeding in lifting-up the object when the gripper is brought to the given configuration, and its fingers are closed. The continuity and flexibility of our representation is illustrated for a toy object in the image shown in Figure 1a, in which the intensity of the green mask is proportional to the grasp success probability. For the purpose of formally commenting on this illustration, we define a 3D reference



(a) Averaged over **z** and orientations

(b) Averaged over **z**, downward grasp

(c) 3D reference frame

(d) Orientation of a "downward" grasp

Figure 1: Figure (a) and Figure (b) present 2D illustrations of grasp densities. In Figure (a), the green opacity of a pixel is proportional to the probability of success of a grasp centered at the corresponding (x, y) position, averaged over all z positions and over all orientations. Axes are defined in Figure (c). The **y** axis is parallel to the handle of the pan. The **z** axis is normal to the plane defined by the main disk of the pan. Naturally, this image illustrates only a part of the information contained in the object's grasp density. Figure (b) shows another projection, where the mask's opacity is proportional to the probability of a "downward grasp" – the two-finger gripper is oriented in a way such that the line between its fingers is orthogonal to the handle of the toy pan, see Figure (d). In this case, only the handle and segments of the circle are graspable, as other segments of the circle will collide with the fingers of the gripper.

frame such that the **xy** plane contains the main circle of the pan, and the **z** axis points upwards (Figure 1c). In Figure 1a, the green opacity of a pixel is proportional to the probability of success of a grasp centered at the corresponding (x, y) position, averaged over all z positions and all orientations. Naturally, this image illustrates only a part of the information contained in the object's grasp density. Figure 1b shows another projection, where the green opacity is proportional to the probability of a "downward grasp" – the two-finger gripper is oriented in a way such that the line between its fingers is orthogonal to the handle of the toy pan (see Figure 1d). In this case, only the handle and segments of the circle are graspable, as other segments of the circle will collide with the fingers of the gripper. We note that although this paper mainly discusses affordances learned with a two-finger gripper, our method is applicable to other kinds of manipulators, as shown in Section 6.1, and discussed in Section 6.3.

The computational encoding of grasp densities is nonparametric. A density is represented by a large number of weighted samples often 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 [31]. We thus make only few assumptions on the shape of grasp affordances. Such grasp densities can accurately represent complex distributions, as illustrated in Figure 1 where grasps are distributed along the edge of the toy pan.

Our model can potentially be used in many different applications. The most interesting one is probably the computation of an optimal grasp in a specific context. For instance, a grasp planner can combine a grasp density with hardware physical capabilities (robot reaching capabilities) and external constraints (obstacles) in order to select the grasp that has the largest chance of success, within the subset of achievable grasps. Another possibility, is to use the continuous grasp success probability to tell which pose parameters matter the most for the execution of a particular grasp. For instance, the grasp model of a cylinder should allow the agent to realize that grasp positions are allowed to vary along the main axis of the cylinder; while, for example, the rotation angle of the gripper around the approach vector should be carefully adjusted.

1.2 Acquisition Through Autonomous Exploration

Research has shown that grasp affordance models can be acquired successfully from human demonstrations [4, 33, 5] and by means of exploration [5, 22]. When learning from a teacher, the learning process is often faster, as the agent is guided directly towards efficient grasping poses. Learning from exploration requires time for finding good grasping poses. However, it can usually be achieved in a largely autonomous way. As it directly involves the body of the robot, it produces a model intimately adapted to the robot morphology. In this paper, we present a mechanism for acquiring grasp densities from exploration. Intuitively, the robot "plays" with an object in a sequence of grasp-and-drop actions. The robot executes grasps suggested by visual cues. After each (successful) grasp, the object is dropped to the floor. When a sufficient quantity of data is available, an importance-sampling algorithm [9] produces an *empirical* grasp density from the outcomes of the set of executed grasps.

We demonstrate the applicability of our method in a large experiment in which we

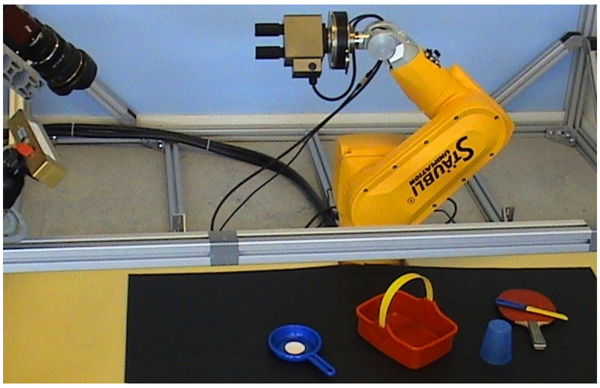


Figure 2: Experiment platform (industrial arm, force-torque sensor, two-finger gripper, stereo camera, foam floor).

learn and test affordance models for three objects on a realistic platform, through the collection of a large grasp dataset – more than 2000 grasps tested on a robot. We show that the success rate of grasps suggested by empirical grasp densities is larger than the success rate of grasps suggested by visual cues, which reveals the value of the learning process. We also 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. 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 region of reach. Experiments are conducted on a largely autonomous platform: the processes involved in the learning scenario (visual pose estimation, path planning, collision detection, success assessment) do not require human intervention.

2 Related Work

In classical robotics, grasping solutions are computed from the geometric properties of an object, which are typically obtained from a 3D shape model. The most popular 3D model for grasping is the 3D mesh, which allows for studying contact forces, and for finding a force-closure optimal grasp therefrom [30, 1, 17, 21]. However, contact-force analysis requires rather accurate shape models. These models often need to be produced by CAD, and they should ideally be augmented by surface-friction and weight-distribution characteristics. As such models are difficult to obtain, grasping has been studied with simpler models, consisting, for example, of 3D edge segments [24], or 3D points [13]. Yet, as these models convey only a fraction of the information generally required for making an accurate grasp quality assessment, the robustness of the resulting grasps is limited.

In order to provide autonomous systems with grasping capabilities, researchers have become increasingly interested in designing robots that *learn* how to grasp. Many learning paradigms have been applied, such as supervised learning [29], learning from demonstrations [10, 4, 33], exploration [22, 5], and active learning [28, 18]. Saxena et al. [29] have explored a supervised learning approach to model the appearance of 2D image patches that predict stable grasps. The authors [29] trained a classifier on a set of 2D images that were hand-labeled with good grasping points. The classifier then identified grasping points in stereo views of an object, and triangulation eventually produced 3D grasp positions.

Grasping strategies have also been learned from human demonstrations. Learning from demonstrations has led, for example, to the identification of clusters in grasp preshape sequences [10] and hand approaches for grasping specific objects [4, 33]. de Granville et al. [4] have modeled grasp affordances with mixture models defined on the space of object-relative gripper orientations. The aim of the authors [4] was to build compact sets of canonical grasp approaches from human demonstrations, by compressing a large number of examples (provided by a human teacher) into a small number of clusters. An affordance was expressed through a density represented as a mixture of position-orientation kernels, where machine learning techniques were used to compute the mixture weights and kernel parameters that best fitted the data. Sweeney and Grupen [33] associated demonstrated grasps to object shapes approximated by ellipsoids. By fitting the same ellipsoids to visual observations, the authors [33] allowed a robot to apply the demonstrated grasps to novel objects similar in shape to those used for training.

Mappings from visual observations to grasping strategies have also been learned from exploration [22, 5]. In the work of Montesano et al. [22], a robot learned a mapping from local 2D image patches to grasp success probabilities by executing exploratory grasp actions. A grasp action consisted of transporting the robot hand to a selected position in a 2D plane, approaching the object from the top until contact is made, and closing the hand. A grasp was thus parametrized by the position of the hand in a 2D plane.

In grasp learning, an important issue is the limited availability of training examples, which is due to the time cost of executing a grasp with a robot. This issue motivates active learning approaches [28, 18], which aim at maximizing the usefulness of tried grasps by defining a balance between the refinement of promising grasps and the exploration of uncertain regions. In the work of Kroemer et al. [18], an active gripper-pose learner is combined with a vision-based grasp-approach controller, which collaborate to model both where and how to grasp an object.

A number of methods for vision-based grasping, learn a mapping from local visual features to grasp parameters [22, 29]. One advantage of these methods is generalization, i.e., a model learned from a set of objects can potentially be applicable to a novel object that shares similarities with the objects from which the model was learned. However, the geometric information provided by a local feature detector is limited. It is thus difficult to link gripper orientation parameters to a single local feature. Other methods link grasp parameters to visual object models [4, 5]. These methods benefit from an increased geometric robustness, which allows for the encoding of richer grasp parameters, such as the 3D positions and orientations of precise pinch grasps. The latter approach is the one we follow in this paper.

This paper provides a complete and consolidated description of the concept of grasp densities [5, 6, 7], and provides the theoretical underpinnings of applications that have already been published elsewhere [15, 18]. It goes beyond our previous work [6] by presenting the method's mathematical foundation in detail, and by evaluating the feasibility of exploiting learned grasp densities to grasp objects. This paper also aims to provide an intuition of what 6D grasp densities look like, and an intuition of the properties they model. To this end, Section 3 presents sets of 2D projections of grasp models in which the relations between a model and the relative object-gripper geometry becomes explicit.

To date, the concept of grasp densities has been used as technology in a number of contexts, such as active learning [18] and the bootstrapping of world knowledge in a cognitive architecture [15]. In the context of active learning, we studied how an active gripper-pose learner, and a vision-based grasp-approach controller, can collaborate to model both *where* and *how* to grasp an object [18]. The active learner encapsulated a reinforcement learner which reflected on specific characteristics of previously-executed grasps (e.g., slippage) to decide where to grasp next. When a grasping point was selected, a low level controller adapted previously-demonstrated trajectories to this particular grasping point and to the local object shape around it. This work [18] built on the material of the present paper by using grasp densities for initializing grasping knowledge by demonstrations. Exploration was subsequently handled by the reinforcement-learning algorithm, which aimed at refining a number of specific grasps, in order to discover locally-optimal policies. By contrast, the work discussed in the present paper aims at representing the whole space of grasping possibilities that an object offers to the robot.

The material of the present paper was also exploited to design an artificial cognitive agent capable of bootstrapping grasping knowledge from exploration [15], in order to acquire a set of visuo-motor competences without relying on prior knowledge on the shape, appearance, or ways of grasping an object. In the presence of novel objects, the agent executed so-called “grasp reflexes” onto object edges detected by a 3D edge reconstruction method. Many of these reflexes lead to nothing, because many edges came, for example, from floor patterns or from ungraspable object parts. However, eventually, the agent succeeded at binding an object to its gripper. The agent then rotated the object in front of a camera, and computed a complete 3D reconstruction of object edges. Once the agent had a 3D edge reconstruction of an object, it learned ways of robustly grasping the object by exploring a large number of grasping configurations, using 3D model-based pose estimation to memorize grasps relatively to the object. This work [15] focused on the integration of several technologies. Grasp reflexes were computed using the work of Popovic et al. [24]. Three-dimensional edge reconstructions, and the accumulation of 3D data across views, was handled by the method of Pugeault et al. [25]. The sub-problem of learning accurate grasping plans once the shape and appearance of the object had been learned was solved using the learning methods proposed in the present paper. In order to learn grasping plans, the agent explored grasps suggested by the 3D edge reconstruction of the object. The knowledge acquired with these experiments was then represented with grasp densities. In the work of Kraft et al. [15], the discussion is limited to the *application* of grasp densities to the development of a cognitive agent, focusing on the developmental aspect of the agent's bootstrapping behavior and on its comparison to humans'. By contrast, the aim of the present paper is to provide a clear mathematical foundation for the grasp learning method.

3 Grasp Affordance Model

This section defines our probabilistic grasp affordance model, and its linking to visual percepts.

In the following, we will refer to the 6D pose of an open gripper as *grasp pose*, or simply *grasp*. We model object affordances for grasping and lifting up objects. A grasp x is successful if placing the gripper at x and closing its fingers allows the robot to stably lift up the object. Our grasp affordance model consists of a probability density function defined on the group of 6D poses $SE(3)$. The *grasp density* of an object o is proportional, for any gripper pose x, to the probability of grasp success when o lies in a reference pose and a gripper is brought and closed at x. Grasp densities are registered to a visual object model, i.e., the visual model and the grasp model are defined in the same reference frame. Grasp densities can thus be aligned to arbitrary object poses by visual pose estimation. Grasp densities do not model continuous preshape configurations. However, separate densities can model affordances for distinct complex preshapes, for example, one density for power grasp affordance, and one for pinch grasps.

3.1 Kernel Density Estimation

Grasp densities are modelled nonparametrically through kernel density estimation (KDE) [31]. KDE is a technique which allows one to model a continuous density function from a set of observations $\{\hat{x}_i\}$ drawn from it, by representing the contribution of the i^{th} observation with a local kernel function $\mathbf{K}_{\hat{x}_i,\sigma}$ centered at $\hat{x}_i$. The kernel function is generally symmetric with respect to its center point. The amplitude of its spread around the center point is controlled by the bandwidth parameter σ. KDE belongs to a class of nonparametric estimation techniques, which make few assumptions on the shape of the estimated function.

A grasp density is encoded by a set of grasp observations, which we will refer to as *particles*. The continuous value of a grasp density is computed from its particle set through KDE. For conciseness, particles are often weighted, which allows one to denote, for example, a pair of identical particles by a single particle of double mass. In the following, the weight associated to a particle $\hat{x}_i$ is denoted by w_i.

KDE models the continuous density d over points x as the weighted sum of the evaluation of all kernels at x

$$d(x) = \sum_{i=1}^{n} w_i \, \mathbf{K}_{\hat{x}_i,\sigma}\left(x\right), \tag{1}$$

where n is the number of particles encoding d. Samples can be drawn from this density as follows:

1. First, a particle $\hat{x}_i$ is selected by drawing i from $P(i = \ell) \propto w_\ell$. (The probability of selecting a given particle is proportional to its weight, which effectively gives a higher chance to particles with a larger weight.)

2. Then, a random variate x is generated by sampling from the kernel $\mathbf{K}_{\hat{x}_i,\sigma}\left(x\right)$ associated to $\hat{x}_i$.

Grasp particles belong to the Special Euclidean group $SE(3) = \mathbb{R}^3 \times SO(3)$, where $SO(3)$ is the group of 3D rotations, and $\mathbb{R}^3 \times SO(3)$ denotes the (semidirect) product of 3D translations and 3D rotations. We denote the separation of grasp and kernel parameters into positions and orientations by $x = (\lambda, \theta)$, $\mu = (\mu_t, \mu_r)$, $\sigma = (\sigma_t, \sigma_r)$. The kernel we use is defined with

$$\mathbf{K}_{\mu,\sigma}(x) = \mathbf{N}_{\mu_t,\sigma_t}(\lambda)\, \mathbf{\Theta}_{\mu_r,\sigma_r}(\theta)\,, \tag{2}$$

where μ is the kernel mean point, σ is the kernel bandwidth, $\mathbf{N}$ is a trivariate isotropic Gaussian kernel, and $\mathbf{\Theta}$ is an orientation kernel defined on $SO(3)$. The orientation kernel $\mathbf{\Theta}$ is defined with the unit-quaternion representation of 3D rotations and the von-Mises Fisher distribution on the 3-sphere in $\mathbb{R}^4$ [11]. Because unit quaternions form a double cover of the rotation group, $\mathbf{\Theta}$ has to verify $\mathbf{\Theta}(q) = \mathbf{\Theta}(-q)$ for all unit quaternions q. We thus define $\mathbf{\Theta}$ as a pair of antipodal von-Mises Fisher distributions [11, 32]:

$$\mathbf{\Theta}_{\mu_r,\sigma_r}(\theta) = \frac{1}{2} C_4(\sigma_r) \left(\mathrm{e}^{\sigma_r\, \mu_r^T \theta} + \mathrm{e}^{-\sigma_r\, \mu_r^T \theta} \right), \tag{3}$$

where $C_4(\sigma_r)$ is a normalizing constant. The $SE(3)$ kernel $\mathbf{K}$ is simulated by drawing samples from $\mathbf{N}$ and $\mathbf{\Theta}$ independently. Efficient simulation methods are available for both normal distributions [2] and von-Mises Fisher distributions [34]. The bandwidths σ_t and σ_r are fixed by hand to allow deviations of approximately 10 mm and 5° respectively, which follows quite naturally from the size and morphology of the hand and the objects to be grasped.

The expressiveness of a single $SE(3)$ kernel (2) is rather limited: location and orientation components are both isotropic, and within a kernel, locations and orientations are modeled independently. We account for the simplicity of individual kernels by employing a large number of them, i.e., a grasp density will typically be supported by several hundreds of particles.

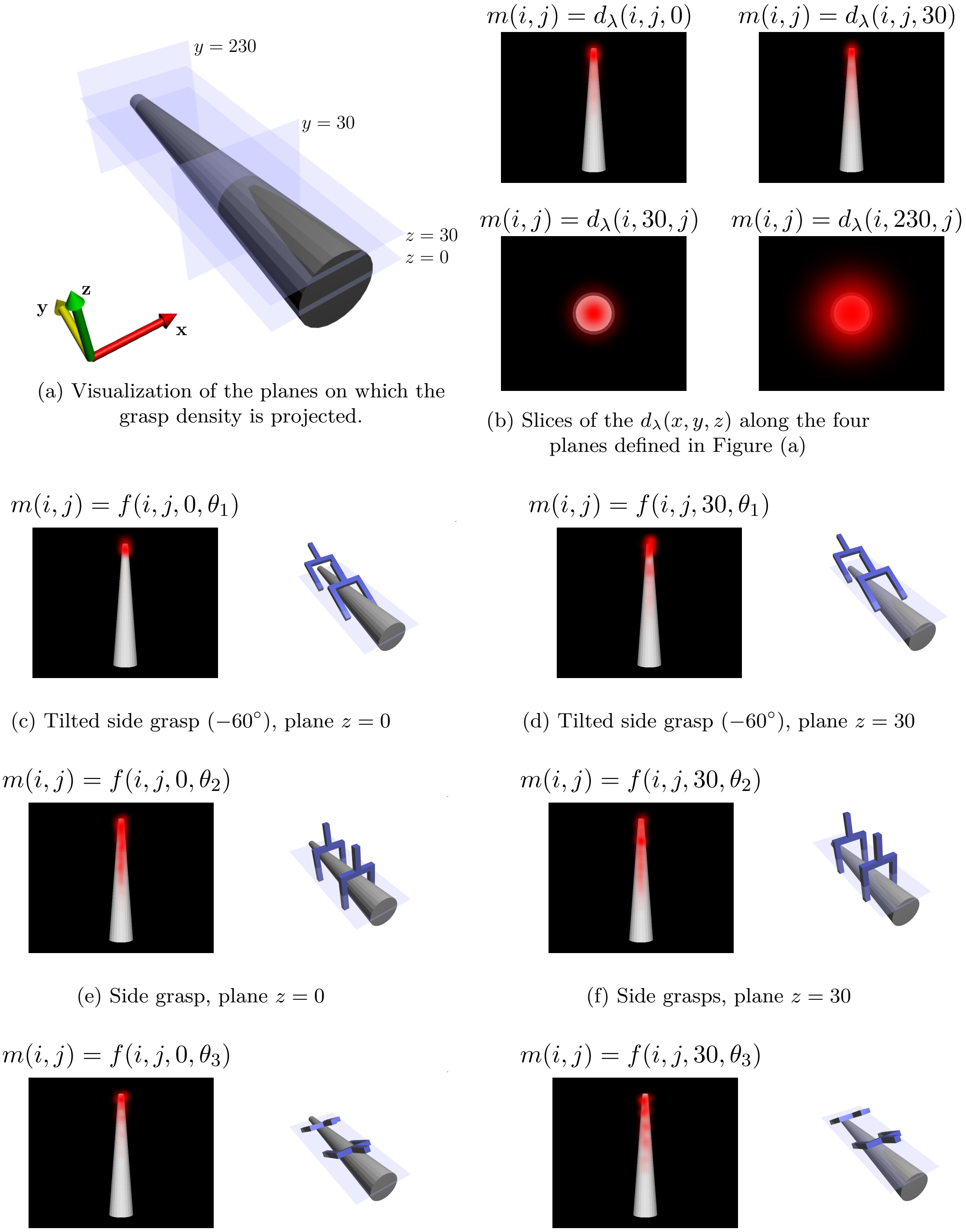
We note that the grasp model discussed in this paper may be supported by means other than the mixture model described above. For instance, nonlinear dimensionality reduction [20], and in particular manifold learning techniques such as kernel PCA or Isomap, could possibly prove helpful. However, in this paper, we focus only on the mixture-model representation.

3.2 Illustrations

Figure 3 and Figure 4 make the 6D domain of grasp densities explicit. In these figures, we illustrate various slices of grasp densities. These illustrations define a correspondence between image pixel coordinates and coordinates within a 3D plane, and they represent the values of a grasp density along that 3D plane by overlaying a mask of varying opacity onto the image. For instance, in Figure 3c, the value of the mask at pixel (i, j) is defined as

$$m(i,j) = d\left(\left[(i,j,0)\,, \theta_1\right]\right),$$

where d is the illustrated grasp density, and θ_1 is a fixed orientation shown on the right side of Figure 3c. This figure thus illustrates a slice of d along the hyperplane defined by $z = 0$ and $\theta = \theta_1$.



(a) Visualization of the planes on which the grasp density is projected.

$m(i,j) = d_\lambda(i,j,0)$ $m(i,j) = d_\lambda(i,j,30)$

$m(i,j) = d_\lambda(i,30,j)$ $m(i,j) = d_\lambda(i,230,j)$

(b) Slices of the $d_\lambda(x,y,z)$ along the four planes defined in Figure (a)

$m(i,j) = f(i,j,0,\theta_1)$

(c) Tilted side grasp (−60°), plane $z = 0$

$m(i,j) = f(i,j,30,\theta_1)$

(d) Tilted side grasp (−60°), plane $z = 30$

$m(i,j) = f(i,j,0,\theta_2)$

(e) Side grasp, plane $z = 0$

$m(i,j) = f(i,j,30,\theta_2)$

(f) Side grasps, plane $z = 30$

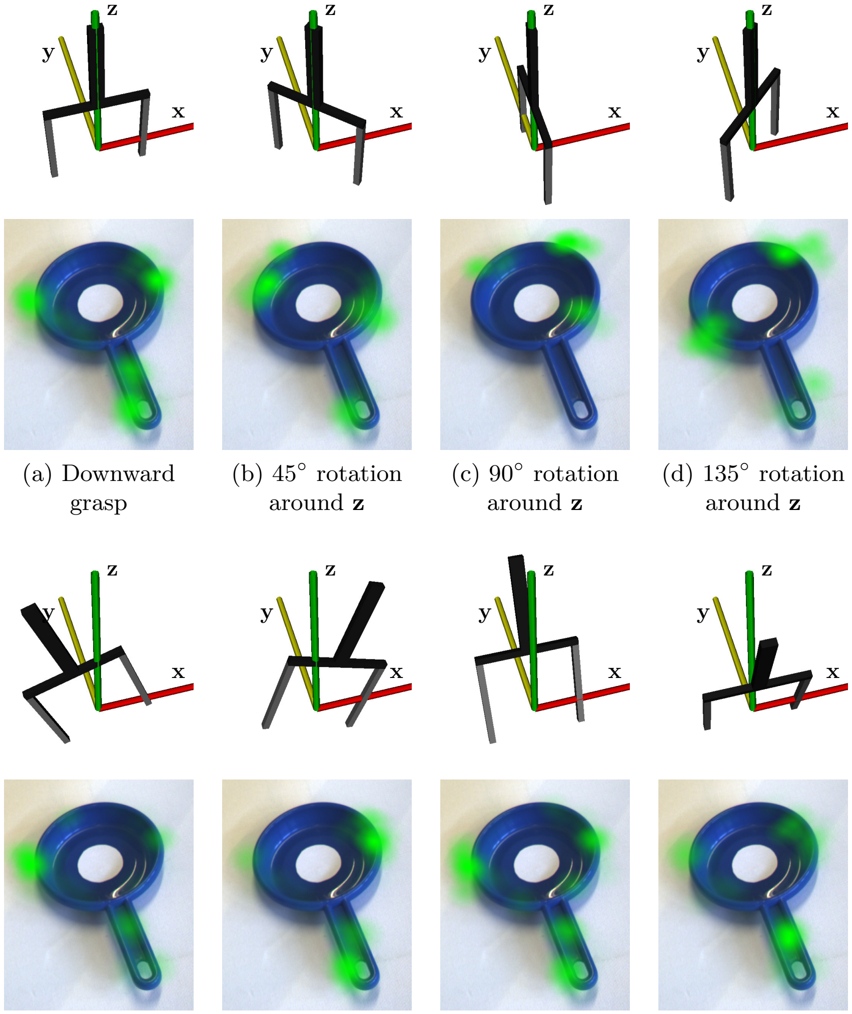
$m(i,j) = f(i,j,0,\theta_3)$

(g) Tilted side grasp (60°), plane $z = 0$

$m(i,j) = f(i,j,30,\theta_3)$

(h) Tilted side grasp (60°), plane $z = 30$

Figure 3: Various projections of a grasp density d learned from simulated grasping actions. The value of each projection is illustrated with a red mask of varying opacity. Figure (b) shows slices of the orientation-marginal density $d_\lambda(x,y,z) = \int d([(x,y,z),\theta])\mathrm{d}\theta$ along the four planes defined in Figure (a). The other figures show slices of d for three fixed orientations θ_1, θ_2, and θ_3 which are illustrated on the right side of each subfigure. For clarity, we define $f(x,y,z,\theta) = d([(x,y,z),\theta])$. Each subfigure defines the quantity to which the opacity of the mask m at pixel (i,j) is proportional.



(a) Downward grasp

(b) 45° rotation around **z**

(c) 90° rotation around **z**

(d) 135° rotation around **z**

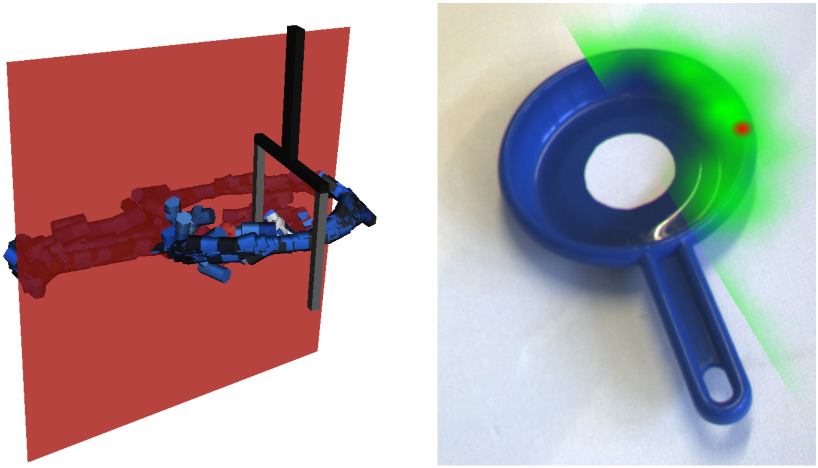
(e) −20° tilt around **y**

(f) 20° tilt around **y**

(g) −20° tilt around **x**

(h) 20° tilt around **x**

Figure 4: Various projections of a grasp density d generated by a robot (see Section 5). In each subfigure, the green opacity of a pixel of the bottom image is given by $m(i,j) = \int d([i,j,z],\theta)\mathrm{d}z$, where θ is a fixed gripper orientation defined in the top image of the same subfigure. The base orientation is the "downward" grasp of Figure 1d; the corresponding probabilities are shown in Figure (a). Figure (b) to Figure (d) show probabilities for grasps whose orientation is such that the angle between the handle of the pan and the normal of the gripper finger plane amounts to 45°, 90° and 135° respectively. Figure (e) to Figure (h) show success probabilities for various tilted grasps.



(a) Best grasp within region of reach

(b) Constrained grasp density (green) and optimal grasp position (red)

Figure 5: Grasping under kinematic constraints. We hypothesize that the object lies at the limit of the robot's workspace. The reaching limit is shown in red in Figure (a). Cutting the object's grasp density along the reaching limit allows the robot to find the most promising grasp within the reachable region. (Note that the red dot in Figure (b) does not correspond to a maximum of the green opacity mask. The reason for this is that the green mask shows the constrained grasp density integrated over $\mathbf{z}$ positions and over orientations, analogously to Figure 1a, whereas the red dot shows the maximum in $SE(3)$.)

Figure 3 shows the grasp density of a simple conic object. Figure 3b shows slices, along four different planes in the position space, of a marginalization of the density over orientations (the planes are shown in Figure 3a). At the narrow end of the cone, the opening of the gripper is substantially larger than the cone section. When aiming at the narrow end, there is a relatively wide range of gripper positions that will lead to a stable grasp. As the grasping point moves towards the wide end of the cone, the gripper needs to be increasingly centered on the main axis of the cone. As a result, in Figure 3b, the $\mathbf{xz}$ spread of the grasp density is much wider within the plane $y = 230$ than in $y = 30$.

By contrast to Figure 3b, which shows the grasp density marginalized over orientations, Figure 3c to 3h show the grasp density for specific, fixed orientations. Figure 3e and 3f illustrate side grasps, while Figure 3c, 3d, 3g, and 3h illustrate tilted side grasps.

The grasp density shown in Figure 3 was acquired in simulation [14], with a method similar to the one described in the next section.

Figure 4 shows a grasp density learned from the experimental data presented in Section 5. Figure 4a to Figure 4d show probabilities for grasps whose orientation is such that the angle between the handle of the pan and the normal of the gripper finger plane amounts to 0°, 45°, 90° and 135° respectively. There is a slight lack of symmetry in the affordance model presented in Figure 4a to Figure 4d. However, as learning continues, the model illustrated in Figure 4 should present increasing symmetry around the center of the toy pan. Figure 4e to Figure 4h show success probabilities for various tilted grasps. Figure 4e clearly shows that when the gripper is tilted to the left, grasps on the right side of the pan are less likely to succeed, as the leftmost finger may hit the bottom of the pan. Conversely, in Figure 4f in which the gripper is tilted to the right, grasps to the pan's left side are less likely to succeed.

Figure 5 illustrates how reach limits can be taken into account to select the best-achievable grasp.

4 Acquisition

This section presents the acquisition of grasp densities. The *initial* grasp density of a new object is built from visual cues. The robot then explores a set of grasps drawn from this initial density. When a satisfactory number of grasps have been executed, collected data are used to construct an *empirical* density.

The following sections offer a formal motivation for the acquisition process. Section 4.1 explains how successful grasp trials can be used to create a grasp density. Section 4.2 discusses the influence of the initial density on learning. Ideally, the robot should explore grasps uniformly around the object, in order to give an equal chance to every grasping configuration. However, most grasping configuration have a very low probability of success. As a result, uniform exploration is infeasible in practice. Instead, our exploration strategy builds on importance sampling to create an empirical density from the grasps generated from a vision-based initial density. Section 4.3 explains how we build initial grasp densities from visual cues.

4.1 Exploratory Learning

We denote by O a random variable which models grasp outcomes. O can be either *success* (“$\texttt{s}$”) or *failure* (“$\texttt{f}$”). We denote by X object-relative gripper poses. The statistical relations between gripper poses and grasp success are captured by the joint probability of O and X, which we denote by $P(O, X)$. The joint probability of O and X can be decomposed as

$$P(O|X)P(X) = P(O, X) = P(X|O)P(O). \tag{4}$$

$P(O = o|X = x)$ is the probability of obtaining outcome o when placing the robot’s hand at pose x with respect to the object. It depends on two variables, one of which is discrete ($o \in \{\texttt{s}, \texttt{f}\}$), and the other one continuous ($x \in SE(3)$). For instance, $P(O = \texttt{s}|X = x)$ is the probability of successfully grasping the object when placing the gripper at x with respect to the object. $P(X = x)$ is the prior probability of grasping an object at x. This probability defines the way grasps are chosen during exploration. If $P(X)$ is uniform, grasps will be executed uniformly in $SE(3)$. $P(X|O)$ is the outcome-conditional grasp pose probability. For instance, given that a grasp on an object has succeeded, $P(X = x|O = \texttt{s})$ gives the probability that this grasp has been applied at pose x. $P(O)$ is the prior probability of grasping success. In the following, we denote the probability density function associated to a continuous random variable V by $p_V(v)$, such that

$$P(V \in D) = \int_D p_V(v)\mathrm{d}v$$

for any arbitrary subset D of the codomain of V. For discrete variables, the equivalent of the probability density function is called *probability mass function*. We denote the probability mass function associated to a discrete variable U by $p_U(u)$, such that

$$P(U = u) = p_U(u)$$

for any u belonging to the codomain of U. Following that notation, Eq. 4 can be written as

$$p_{O|X=x}(o)p_X(x) = p_{O,X}(o, x) = p_{X|O=o}(x)p_O(o). \tag{5}$$

Grasp affordances are usually modeled in one of two ways. They can be modeled with an empirical representation of pose-conditional grasp success probabilities

$$f(x) = p_{O|X=x}(\texttt{s}) = P(O = \texttt{s}|X = x).$$

Grasp affordances can also be modeled with an empirical representation of success-conditional grasp densities $d(x) = p_{X|O=\texttt{s}}(x)$. In this work, affordances are modeled with grasp densities. (The differences between the two approaches are discussed in Section 4.5.) Grasp densities characterize success-conditional gripper pose probabilities, i.e., the distribution of robust grasping solutions around the object. From a theoretical viewpoint, a direct way of learning a grasp density is to obtain pose samples from $p_{X|O=\texttt{s}}(x)$, that is, obtain a set of pose examples whose distribution follows the density $d(x) = p_{X|O=\texttt{s}}(x)$. Unfortunately, this density cannot be sampled by direct empirical means. Nevertheless, two important observations can be made at this point:

- By executing a grasp at pose x, we generate an outcome sample that follows $p_{O|X=x}(o)$. Indeed, by definition of $p_{O|X}(o)$, grasps repeatedly executed at pose x will succeed with a frequency equal to $p_{O|X=x}(o)$.
- As the grasp success prior $p_O(\mathtt{s})$ is constant over grasp poses, a grasp density is proportional to $p_{O,X}(\mathtt{s}, x)$.

These observations provide us with a procedure for obtaining samples from $p_{X|O=\mathtt{s}}(x)$:

1. We generate a set of samples from
$$p_{O,X}(o, x) = p_{O|X=x}(o)p_X(x)$$
by repeatedly selecting a grasp x_i from $p_X(x)$, and observing its execution outcome o_i. We denote this sample set by S, with
$$S = \{(o_i, x_i)\}_{i\in[1,n]} \,. \tag{6}$$
The definition of $p_X(x)$ (and how to sample from it) is discussed in the next section.
2. We then generate a set T of samples from $p_{O,X}(\mathtt{s}, x)$ by selecting the successful samples in S:
$$T = \{x_i : (\mathtt{s}, x_i) \in S\} \,. \tag{7}$$
3. Since $p_O(\mathtt{s})$ is a constant, samples drawn from $p_{O,X}(\mathtt{s}, x)$ are distributed as $p_{X|O=\mathtt{s}}(x)$ (see Eq. 5).

As the elements in T are distributed as $p_{X|O=\mathtt{s}}(x)$, they form a nonparametric grasp density representation. As a result, we can form an empirical grasp density from successful grasp examples.

4.2 Grasp Pose Prior

The only missing element of the learning procedure described above is the definition of the pose prior $p_X(x)$. One possibility is to use a prior which directs exploration to promising object regions. While this procedure allows the robot to quickly acquire grasping skills, it produces a grasp density which not only depends on the physical properties of the object (e.g., shape, friction, mass, which are modeled by $p_{O|X}(o)$), but also depends on the chosen prior since

$$p_{X|O=\mathtt{s}}(x) \propto p_{O|X=x}(\mathtt{s})p_X(x). \tag{8}$$

The other possibility is to make the prior uniform in a region of $SE(3)$ surrounding the object. Theoretically, this is an attractive option, as it avoids introducing bias into the learning process. Unfortunately, as grasp poses sampled from a uniform distribution on $SE(3)$ have a very low chance of success, executing the sampling procedure defined above with a uniform pose prior will yield a prohibitively slow convergence rate. This problem can theoretically be addressed using importance sampling [9], as explained in

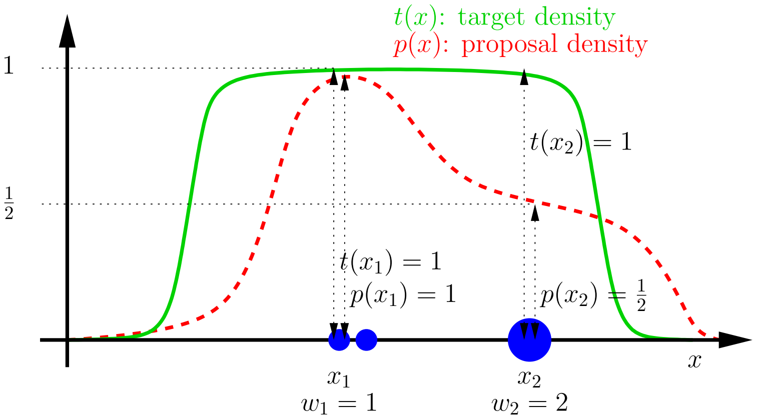


Figure 6: Importance sampling weight computation. Although points such as x_2 are less likely to be drawn than points like x_1, the weight associated to x_2 is twice as large as that associated with x_1.

the next paragraphs. However, a uniform pose prior is difficult to achieve in practice; we will return to this issue in Section 4.4.

Importance sampling is a technique that allows one to draw samples from a *target* density by properly weighting samples drawn from a (preferably similar) *proposal* density. The target density $t(x)$ is difficult to sample from, but it can be evaluated. Therefore, samples are drawn from the proposal density $p(x)$, and the difference between the target and the proposal is accounted for by associating to each sample x a weight given by $t(x)/p(x)$. Figure 6 illustrates the effect of importance sampling in a simple 1-dimensional case.

Let the pose prior $p_X(x)$ of Eq. 5 be a uniform distribution, and let $h(x)$ be a density function which yields a high value in promising grasping regions. The paragraphs below describe an algorithm for learning grasp densities with a uniform pose prior while executing grasps sampled from $h(x)$. The algorithm is based on the importance sampling algorithm.

1. Instead of executing grasps sampled from the uniform prior $p_X(x)$, we execute grasps sampled from $h(x)$, a reasonable number of which end in success. The resulting sample set
$$S' = \{(o_i, x_i)\}_{i\in[1,n]} \tag{9}$$
follows $p_{O|X=x}(o)h(x)$.

2. We generate a set T' of samples from $p_{O|X=x}(\mathtt{s})h(x)$ by selecting the successful samples in S':
$$T' = \{x_i : (\mathtt{s}, x_i) \in S'\}\,. \tag{10}$$

3. Importance sampling allows us to construct a sample set T'' which follows $p_{O,X}(\mathtt{s}, x)$ from the samples in T'. Letting $p_{O,X}(\mathtt{s}, x)$ be the target density and letting

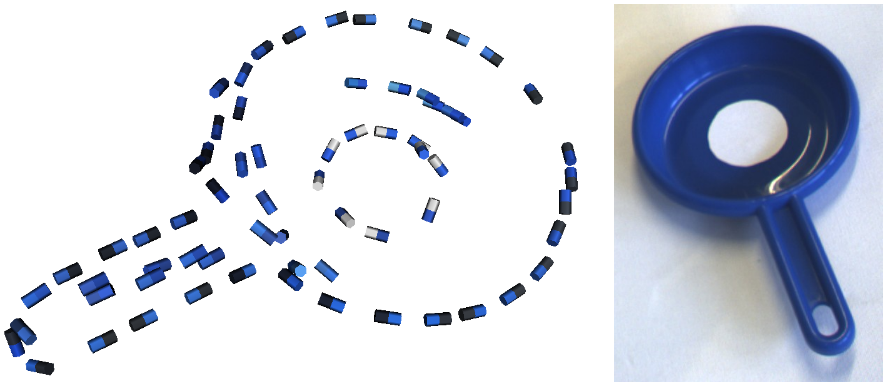


Figure 7: ECV reconstruction (left) of a toy pan (right). Each cylinder corresponds to an ECV descriptor. The axis of a cylinder is aligned with the direction of the modeled edge. Each cylinder bears the two colors found on both sides of the edge in 2D images. For clarity, the reconstruction only shows a fraction of the descriptors available for this object.

$p_{O|X=x}(\mathtt{s})h(x)$ be the proposal density, associating an importance weight

$$\begin{aligned} w_i &= \frac{p_{O,X}(\mathtt{s}, x_i)}{p_{O|X=x_i}(\mathtt{s})h(x_i)} && (11) \\ &= \frac{p_{O|X=x_i}(\mathtt{s})p_X(x_i)}{p_{O|X=x_i}(\mathtt{s})h(x_i)} && (12) \\ &= \frac{p_X(x_i)}{h(x_i)} && (13) \end{aligned}$$

to each sample x_i in T' yields a sample set T'' that follows $p_{O,X}(\mathtt{s}, x)$. Since $p_X(x)$ is uniform, the importance weights can be computed as

$$w_i = \frac{1}{h(x_i)}. \qquad (14)$$

4. Since $p_{O,X}(\mathtt{s}, x)$ is proportional to $p_{X|O=\mathtt{s}}(x)$ (cf. Eq. 5), T'' also follows $p_{X|O=\mathtt{s}}(x)$, and hence forms a grasp density. This grasp density is evaluated with Eq. 1, using the weights defined in Eq. 14.

The grasp density generated by this process will converge to $p_{X|O=\mathtt{s}}(x)$ when the number of samples becomes large, and an adequate design of $h(x)$ can potentially yield a reasonable convergence rate towards the limit case. As mentioned above, we use visual cues to direct exploration. The construction of $h(x)$ from visual cues is explained in the next section.

4.3 Creating Initial Densities From Visual Cues

The two-finger gripper of Figure 2 is best suited for precision pinches. As object regions which afford pinch grasps are often characterized by visual edges, we design initial densities to direct exploration towards edges.

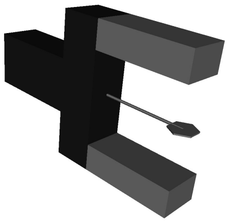
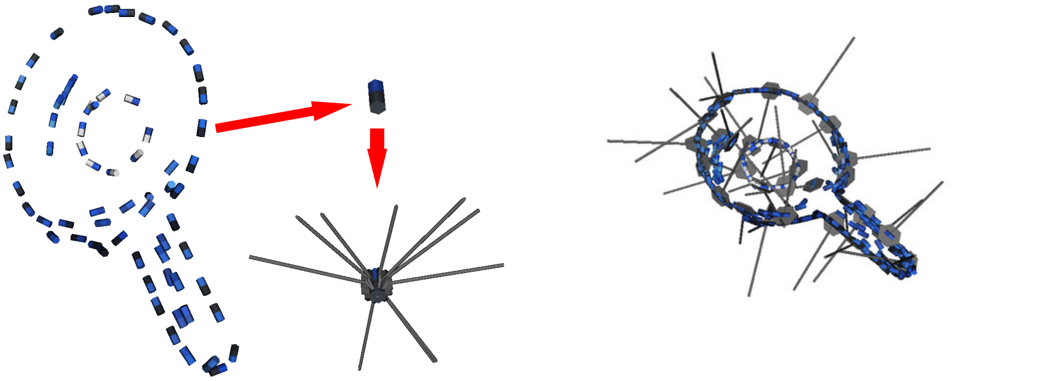


Figure 8: For clarity, we often render grasp poses with a small paddle-like object. This image shows how the paddle object relates to a physical two-finger gripper.



(a) Defining a set of grasp particles from one ECV descriptor

(b) Some of the particles supporting an initial density

Figure 9: Building initial grasp densities from ECV descriptors. Each descriptor gives rise to a set of grasp poses. This figure makes use of the compact grasp-pose rendering introduced above (Figure 8).

Initial densities are computed from 3D reconstructions of object edges. These reconstructions are provided by the model of Krüger and Pugeault [26], which represents short edge segments in 3D space. These segments, called *early-cognitive-vision (ECV) descriptors*, are computed by combining 2D edges extracted in stereo image pairs. Each descriptor is defined by a 3D position and edge-tangent orientation, therefore living in $\mathbb{R}^3 \times S^2$. Descriptors are tagged with color information, which is extracted from their corresponding 2D patches (Figure 7). 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. Assuming that the motion adequately spans the object pose space, a complete 3D reconstruction of the object can be generated, eliminating self-occlusion issues [16].

Constructing an initial density works by defining a large set of grasps onto object edges. ECV descriptors have two degrees of freedom in orientation, whereas a grasp orientation has three. Each descriptor thus leads to a set of grasps for which the third orientation parameter is uniformly distributed, as sketched in Figure 9. The resulting grasps are then directly used as particles supporting the initial density.

4.4 Discussion

One weakness of importance sampling is its slow convergence when the target density has heavier tails than the proposal, i.e., when the ratio proposal–target is globally smaller in the extreme regions of the densities. As we have discovered, it is difficult to design initial densities that reasonably cover promising areas, while excluding non-graspable object parts. For this reason, in the experiments presented below, we compute importance weights with

$$\frac{1}{h(x)+C}. \tag{15}$$

The case $C = 0$ corresponds to $p_X(x)$ being a uniform pose prior. Setting C to a large value amounts to defining $p_X(x) \simeq h(x)$. In the experiments presented below, C is set to the peak value of the kernels supporting $h(x)$ divided by the number of kernels, which compromises between a high convergence rate and an unbiased grasp density.

4.5 Grasp Densities vs. Success Probabilities

Grasp affordances are usually modeled either by grasp success probabilities, or by success-conditional grasp densities. Let us write Eq. 5 again:

$$p_{O|X=x}(o)p_X(x) = p_{O,X}(o,x) = p_{X|O=o}(x)p_O(o). \tag{16}$$

Grasp success probabilities, which are written as $p_{O|X=x}(\mathtt{s})$, correspond to a discriminative grasp model. Success-conditional grasp densities $p_{X|O=\mathtt{s}}(x)$ form a generative model. An important difference between the two approaches lies in their use of training data, as grasp densities are learned from successful grasps only, while success probabilities will generally be learned from both positive and negative grasp examples. In practice, successful grasps are strong cues for the object's grasp affordance. In contrast, failed grasps are not necessarily related to *object* properties, as failures may be caused by obstacles, or by issues related to the robot body (e.g., reaching constraints). In this paper, we focus on the evaluation of grasp densities learned from successful grasps only. However, the data presented in Section 5 can potentially be used to train a discriminative model. While negative grasp examples are not always related to object properties, they may nonetheless improve the computation of robust grasps. One of our future aims is to test a discriminative model learned from the data collected for this paper, and compare its performances to those of the generative model.

5 Exploratory Learning Experiments

In this section, we demonstrate the applicability of our method to learning empirical densities, and we estimate the efficacy of empirical densities in a typical grasping scenario. For this purpose, we have developed a mostly autonomous robotic platform, which allows for the exploration of a large number of grasps (in almost arbitrary robot-object configurations) with minimal human intervention.

Our grasp models are registered with a visual object model that allows the estimation of the object's 6D pose (i.e., 3D position and orientation). Grasp models can thus be visually aligned to arbitrary spatial configurations of the object they characterize. In this work, we use a visual model that represents edges in 3D [8]. This model has the form of a hierarchy of increasingly expressive object parts, where bottom-level parts correspond to groups of the ECV descriptors described in Section 4. Visual inference is performed by extracting an ECV scene reconstruction from a pair of stereo images and propagating this evidence through the hierarchy using a belief propagation algorithm (BP) [23, 32, 8]. 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, have been presented in previous work [8, 16].

Path planning is implemented with a probabilistic planner [19], which has built-in knowledge of the robot body and its workspace. It is provided with an ECV reconstruction of the scene from which it is able to detect potential object collisions. The ECV scene reconstruction is augmented with an aligned, accumulated, ECV reconstruction, which allows the planner to prevent collisions with occluded parts of the object to be grasped. Collisions that are not foreseen by the planner are automatically detected by a force-torque sensor coupled with a model of the gripper movement dynamics.

Beyond the practical advantage of conducting large experiments, the autonomy of our platform further demonstrates the applicability of our method to grasp learning within a minimally supervised environment. It also demonstrates the robustness of the method to the relatively high level of noise introduced in the autonomous resolution of pre-grasping problems, i.e., pose estimation, path planning, and collision detection.

We present an experiment in which the robot learns empirical densities for three objects. In order to measure the value of learning, we compare the success rate of grasps sampled randomly for initial (vision-based) densities to the success rate of grasps sampled from empirical densities. We also present an experiment in which the robot repeatedly executes the most promising grasp under reaching constraints.

These experiments involve testing sets of grasp *trials*. Section 5.1 explains the process of executing a set of grasp trials, and it details the nature of the recorded data. Section 5.2 presents the application of this process for both learning empirical densities and estimating their efficacy in practical scenarios. Results are discussed in Section 5.3.

5.1 Grasp Trials

Our robotic platform is composed of an industrial robotic arm, a force-torque sensor, a two-finger gripper, and a stereo camera (see Figure 2). 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 to 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 drops. The foam floor also allows the gripper to push slightly into the floor and grasp thin objects lying on the foam surface.

The FSM is initially provided with an object model, which consists of a grasp density registered with a visual model, as described above. 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 (either a random sample, or the best achievable grasp, depending on the experiment)

iii. Submit the grasp to the path planner,

iv. Move 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 above. The path planner has a built-in representation of the floor and 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 ECV reconstruction of the scene. It computes a collision-free path to the target gripper configuration through its built-in knowledge and given information about the scene. When no path can be found, it produces a detailed error report.

While the gripper is approaching the object (iv), and subsequently while grasping the object (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 sometimes fail, for example, 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 arm movement, 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, and no explicit guidance is given.

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, a supervisor notification of bound-to-fail grasp, a collision (notified either from the force-torque sensor or from the supervisor), or an empty gripper (v and vi).

The impact of learning on the robot's grasping skills can be quantified by comparing the grasp success rate of grasps sampled from initial densities to the success rate of grasps sampled from empirical densities. The impact of learning on a *grasp model* is quantified by the success rate of the grasps *suggested by the densities.* It is also interesting to measure the impact of learning on the robot's ability to *autonomously grasp the object*, which is quantified by the success rate of the grasps *physically executed by the robot.* As the path planner may prevent grasps from being executed, the grasps physically executed by the robot are a subset of the grasps suggested by the densities.

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, corresponds to object collisions, ground collisions, or void grasps, either physically generated by the robot, or asserted by the supervisor. E_r errors also include cases where the object drops out of 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 success rate of the grasps suggested by a grasp density is given by

$$r_{rp} = \frac{s}{s + e_r + e_p}. \tag{17}$$

This rate quantifies the quality of the density. The success rate of the grasps physically executed by the robot is

$$r_r = \frac{s}{s + e_r}. \tag{18}$$

This rate quantifies the robot's autonomous ability to grasp the object using the density. This rate does not take errors e_p into account, as the corresponding grasps are rejected autonomously before physical execution.

The execution of a complete grasp trial takes about 40–60s. The time-consuming processes are pose estimation, path planning, and arm movements. The runtime of grasp sampling is negligible – sampling a grasp takes a microsecond on average.

Through the process described above, the robot will effectively learn *pick-up* grasp affordances, offered by an object lying on a flat surface in a natural pose. We note that while the gripper is being closed, the object may shift or rotate. Hence, different gripper poses may lead to the same grasp. Whether or not the object moves during a grasp is not taken into account in this experiment, i.e., a gripper pose is considered to yield a successful grasp as soon as it allows the robot to firmly lift up the object.

5.2 Evaluation

We conducted experiments with the three objects of Figure 10, selected for their differences in shape and structure, which offer a large variety of grasping possibilities.

Initial grasp densities were built from the ECV reconstructions as explained in Section 4, yielding the models illustrated in Figure 11. The robot learned an empirical density for each object. It also tried sets of grasps randomly sampled from each empirical density.

We tested the efficacy of our method in a usage scenario in which the robot successively performs grasps that have the highest probability of success within its region

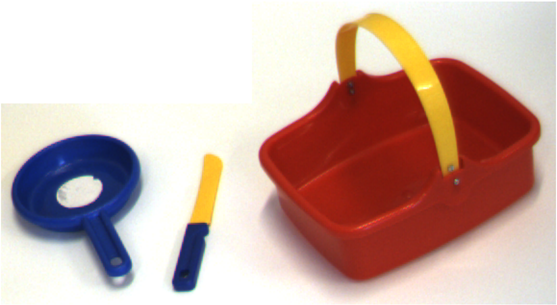


Figure 10: Object library, composed of a pan, a knife and a basket

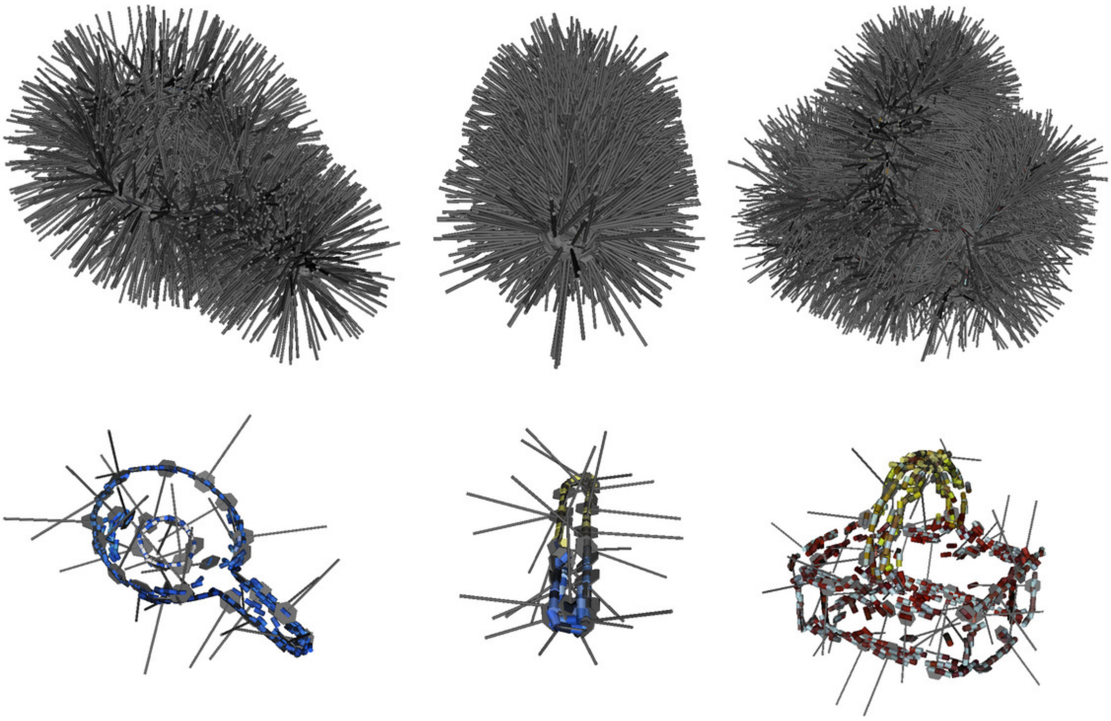


Figure 11: Particles supporting the initial densities of the pan (left-side images), knife (middle images) and basket (right-side images). For clarity, each density is shown twice: the top images show a large number of particles, while bottom images show only a few. As explained in the text, the feasibility of grasps represented by initial densities is limited. For many grasps, the gripper wrist will inevitably collide with the object. Other grasps approach edges which cannot be grasped, such as those at the bottom of the basket.

of reach. 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 implements each grasp trial by randomly drawing a set of grasps from an empirical density, and sorting these grasps in decreasing order of probability according to that empirical density. The grasps are sequentially submitted to the path planner and the first feasible grasp is executed.

5.3 Results

Empirical densities are shown in Figure 12, 13, and 14. Comprehensive quantitative results are displayed in Figure 15. Columns titled s, e_r, and e_p correspond to the statistics collected during the experiment. The last two columns show the success rates defined in Eq. 17 and Eq. 18. Rows titled *initial densities* and *empirical densities* show the success rates of grasps sampled from initial and empirical densities, respectively. Rows titled *best achievable grasp* correspond to the usage scenario in which the robot repeatedly performs the grasp that has the highest chance of success within its region of reach. Figure 16 shows success rates graphically.

Figure 12 shows that the empirical densities are a much better model of grasp affordances than the initial densities of Figure 11. The global success rates r_{rp} (see Figure 16a) provide a quantitative comparison of the grasping knowledge expressed by initial and empirical densities. The empirical densities allow the robot to collect a number of positive examples similar to the number of positive examples collected from initial densities but with a much smaller number of trials. The red bars in Figure 16a confirm that grasps generated from modes of an empirical density have a higher chance of success than randomly sampled grasps.

Figure 16b shows success rates in which planner-detected errors E_p are ignored. From initial to empirical densities, the physical success rate r_r increases less than r_{rp}, which indicates that the robot has partly learned to do the work of the planner, i.e., to avoid grasps which may lead to a collision with the object or with the ground.

Our results make a number of issues explicit. For all objects, we reduced the workload of the motion planner by an average factor of ten (a significant result, as path planning is computationally expensive). The average success rate of grasps performed by the robot (ignoring those rejected by the planner) grows from 42% to 52%. In “best-achievable grasp” scenarios, the success rate of robot grasps is 61% on average. These numbers are quite encouraging, given that we tested our system in real-world settings. For instance, visual models, which are learned autonomously [8, 16], 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 difficult grasping situation, given the short distance between grasping points 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 Figure 15

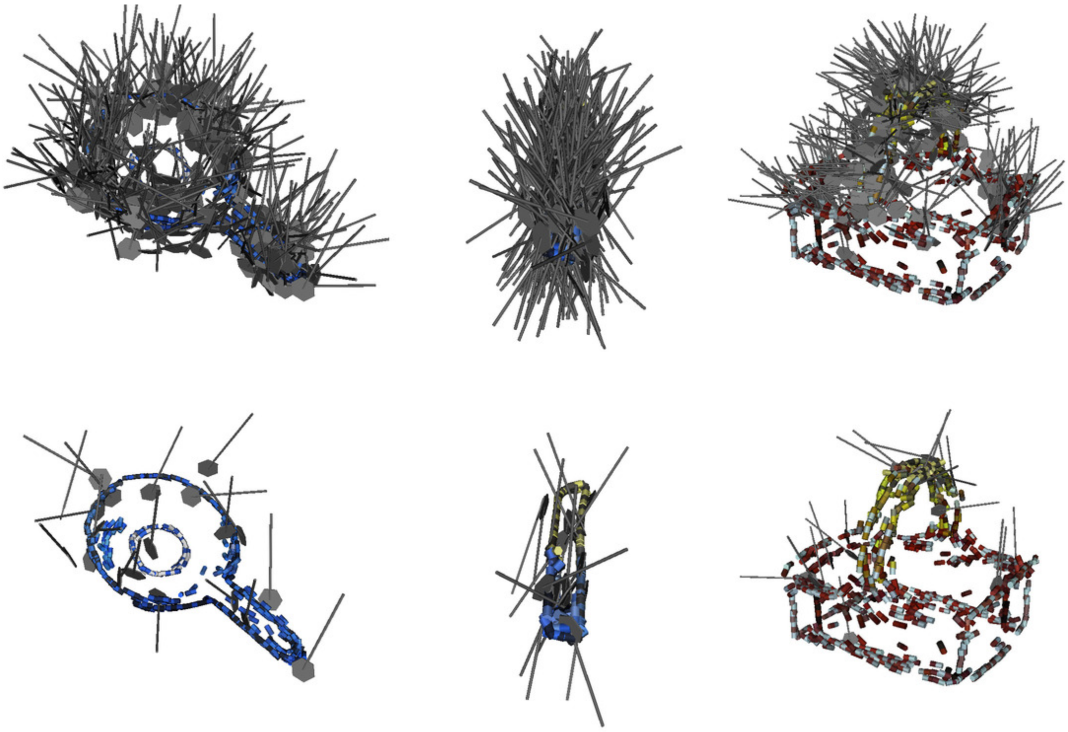
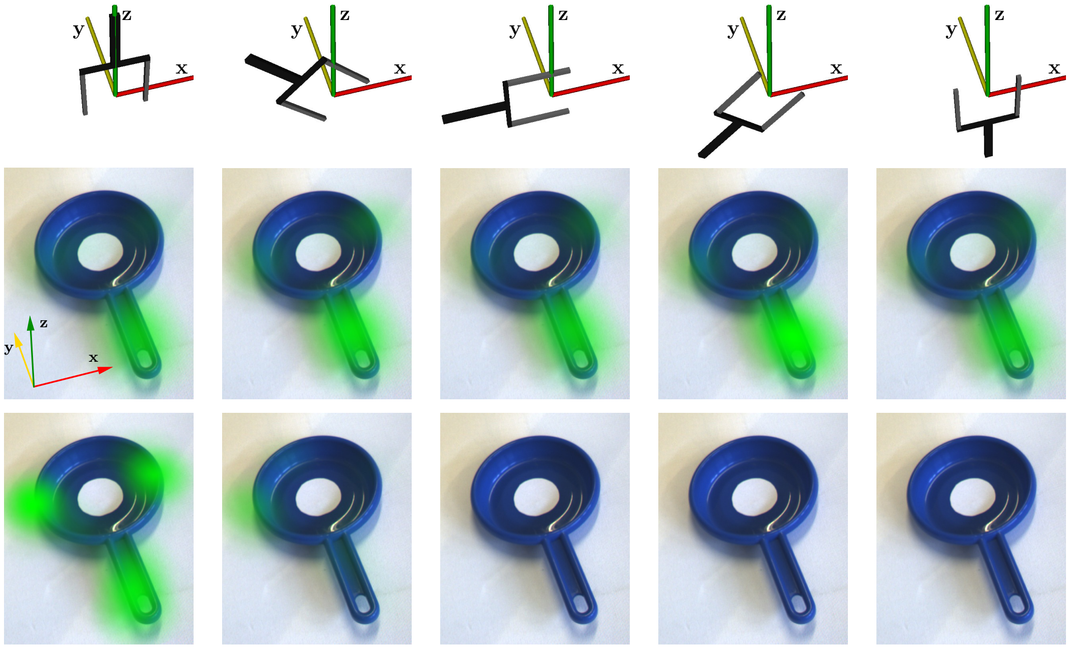
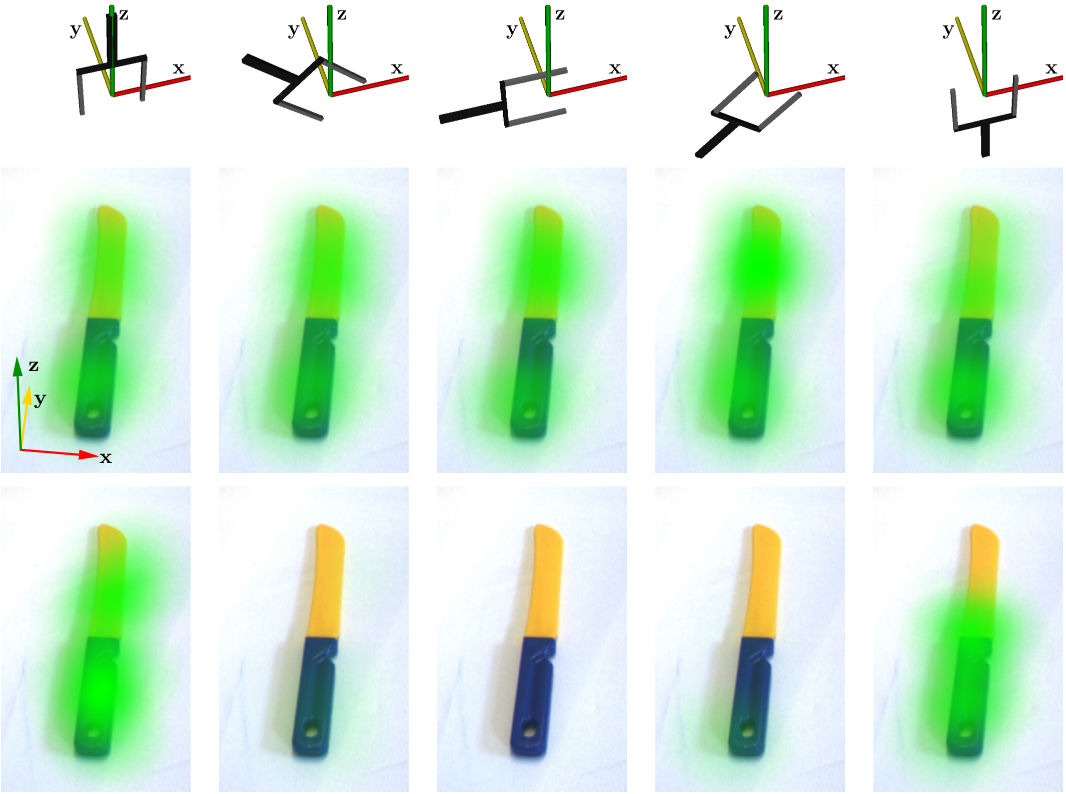


Figure 12: Samples from the empirical densities. For clarity, each density is shown twice: top images show a large number of samples, while bottom images show only a few.



(a) Downward grasp (b) 45° around $\mathbf{y}$ (c) 90° around $\mathbf{y}$ (d) 135° around $\mathbf{y}$ (e) 180° around $\mathbf{y}$

Figure 13: Various projections of the initial density (middle row) and empirical density (bottom row) of the pan. In each subfigure, the green opacity of a pixel is given by $m(i, j) = \int d([i, j, z], \theta)\mathrm{d}z$, where θ is a fixed gripper orientation defined in the top row. Gripper orientations are defined with respect to the reference frame shown in the middle row. The $\mathbf{y}$ axis is parallel to the handle of the pan. The $\mathbf{z}$ axis is normal to the plane defined by the main disk of the pan. In this figure, the initial density suggests that the handle of the pan is graspable with any of the considered gripper orientations. The empirical density, however, suggests that, in order to pick up the object, the only possible grasps are those approaching from the top onto either the handle or two specific points of the circle.

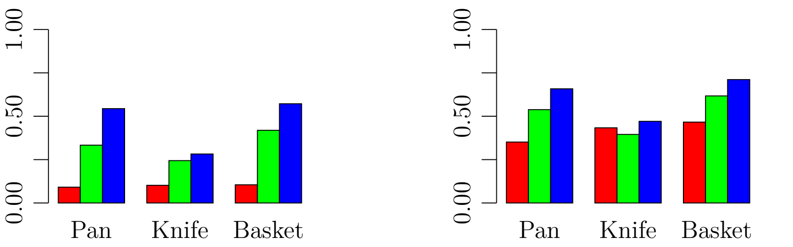


(a) Downward grasp (b) 45° around **y** (c) 90° around **y** (d) 135° around **y** (e) 180° around **y**

Figure 14: Various projections of the initial density (middle row) and empirical density (bottom row) of the knife. In each subfigure, the green opacity of a pixel is given by $m(i, j) = \int d([i, j, z], \theta)\mathrm{d}z$, where θ is a fixed gripper orientation defined in the top row. Gripper orientations are defined with respect to the reference frame shown in the middle row. The **y** axis is aligned with the knife. The **x** axis is parallel to the plane defined by the blade of the knife. This figure shows that while the initial density suggests that the knife is graspable with any of the considered gripper orientations, the empirical density suggests that, in order to pick up the object, the only possible grasps are those approaching normally to the blade or handle planes. Because the knife can be grasped when lying like shown on these pictures, and also when flipped 180° around **y**, the empirical density suggests that the knife could be grasped from either sides.

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

Figure 15: Success/error counts and success rates. See also Figure 16.



(a) Success rate r_{rp} computed from both physical and planner-detected failures

(b) Success rate r_r computed from physical failures only

Figure 16: Success rates. Red, green, and blue bars respectively illustrate rates for grasps sampled from initial densities, sampled from empirical densities, and best-achievable grasps. Numerical rates can be found in Figure 15.

do not exclusively reflect issues related to grasp densities.

We showed that comprehensive grasp affordance models can be acquired by mostly 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.

6 Discussion

This section reviews some of the decisions that lead to the grasp density method (Section 1 and 2), and discusses plans and ideas for future work. In Section 6.1, we discuss how our method can potentially incorporate learning by demonstration. Section 6.2 sets a path towards cross-object generalization. Section 6.3 explains how complex hand preshape information could be integrated into our framework.

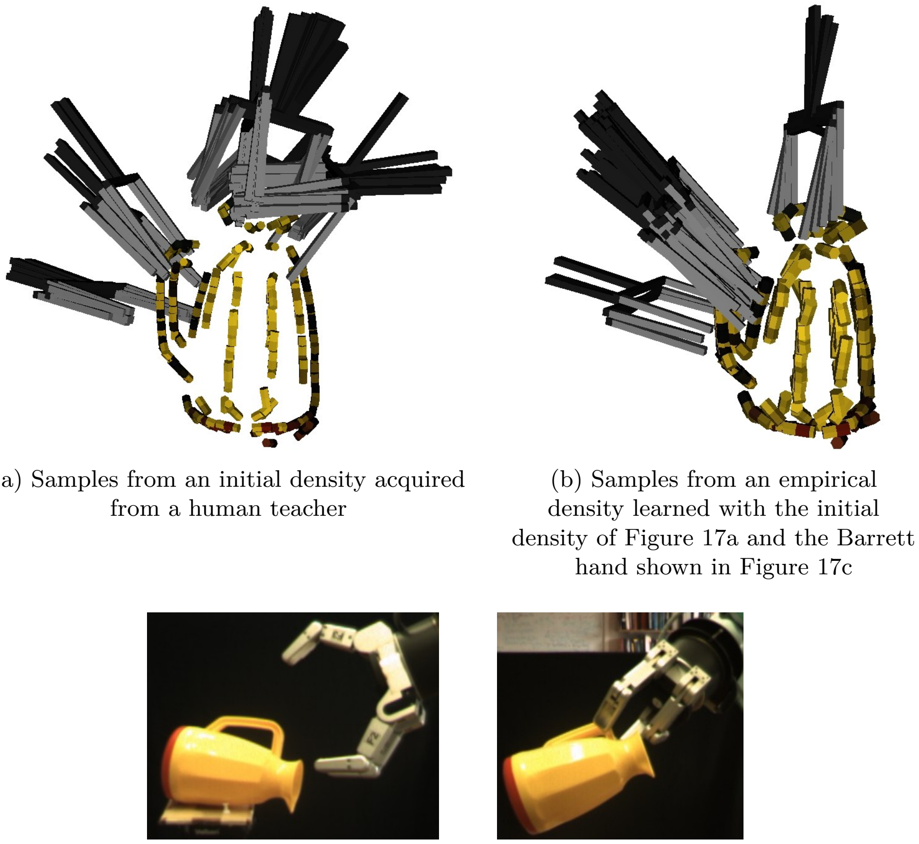
6.1 Combining Demonstration And Exploration

In Section 2, we argued that learning from demonstration, while usually faster than exploratory learning, fails to produce models that intimately fit to the robot morphology. One reasonable compromise that naturally comes to mind, is to learn an initial model from a teacher, then refine this model through exploration: the teacher guides the robot towards promising grasping regions, then lets exploratory learning adapt the model to the robot morphology. This paradigm is easily implemented within the grasp density framework by learning initial densities from a teacher. The paragraphs below present results which illustrate how our framework performs at combining demonstration and exploratory learning.

Building a grasp density from a set of grasp poses provided by a human teacher is fundamentally easy, as one may simply use the grasp poses as particles supporting the density's nonparametric representation. Figure 17a shows samples from an initial density created from a set of 50 grasps performed by a human and recorded with a motion capture system.

Using the density of Figure 17a as initial density, we have let a platform composed of an industrial arm and a Barrett hand (Figure 17c) learn an empirical density. As illustrated in Figure 17c, the hand preshape was a parallel-finger, opposing-thumb configuration. A grasp was considered successful if the robot was able to stably lift up the object, success being asserted by raising the robotic hand while applying a constant, inward force to the fingers, and checking whether at least one finger is not fully closed. This learning process yielded a 20% success rate, eventually providing us with 25 successful grasps. From these grasps, we built the empirical density shown in Figure 17b.

While the object and platform differences that exist between this experiment and that of the previous section unfortunately prevent us from drawing quantitative conclusions, we feel that learning from demonstration did provide an overall faster learning experience. Nevertheless, Figure 17 indicates that there are substantial differences between the demonstrated initial density and the refined empirical density. These differences are caused by the discrepancies between a human hand and the mechanical hand. For example, while grasping the top of the jug is easy for a human hand, it proved to be



(a) Samples from an initial density acquired from a human teacher

(b) Samples from an empirical density learned with the initial density of Figure 17a and the Barrett hand shown in Figure 17c

(c) Barrett hand grasping the plastic jug

Figure 17: Acquiring grasp densities for a plastic jug by combining exploratory and demonstration learning

difficult for the Barrett hand with parallel fingers and opposing thumb. Consequently, a large portion of the topside grasps suggested by the initial density are not represented in the empirical density. The most reliable grasps approach the handle of the jug from above, and these grasps are indeed strongly supported in the empirical density. We conclude that, while the human teacher can focus the robot towards interesting points, the discrepancies between a human hand and mechanical hands support the need for embodied learning.

6.2 Generalization

In this paper, we modeled *object* grasp densities, and link these to a visual model of the whole object. This allowed our system to suggest grasps onto occluded or partly-occluded object parts, and made it robust to visual noise, allowing it to work in situations where small object parts would be undetectable. Estimating the 6D pose of the object also permitted the alignment of precise 6D pinch grasp poses.

However, grasp affordances ideally characterize object-robot relations through a minimal set of properties, meaning that object properties not essential to a relation should be left out. This in turn allows, for example, for generalization of affordances between objects that share the same grasp-relevant features. Ultimately, instead of associating densities with a whole object, we aim to relate them to visual object *parts* that predict their applicability, allowing for generalizing grasps across objects that share the same parts. Grasp densities offer elegant means of discovering object parts for which the visual structure is a robust predictor of grasping properties. From a set of available 3D models, one could arbitrarily segment parts (e.g., the handle of the pan), detect these parts in all available visuo-motor object models, and see whether, throughout all detected part poses, there exists a strong correlation between the grasp density of the part and the grasp density of the object models at detected poses. The resulting generic parts would speed up the learning of the grasp density of an object that has never been grasped, by helping form an initial density from the generic parts that match the object's visual structure.

6.3 Preshape Model

In Section 3, we specified that we learned grasp densities with a fixed hand preshape. Affordances for differently preshaped grasps (power grasps or precision pinches) could then be represented with a different density for each preshape. Although discretizing the preshape space seems reasonable for simple two-finger grippers, it can quickly become inadequate for more complex robotic hands. In applications where preshapes play an important role, one may benefit from including continuous preshape-related variables within the latent grasping model, and simultaneously learn hand poses and preshapes, that lead to successful grasps. However, this approach also has its limitations, as the 6D space of hand poses is already relatively large, and exploring additional dimensions comes at the price of a reduced pose exploration. We note that low-dimensional, dexterous hand control has recently been successfully implemented on a robotic platform [3]. This work constitutes an interesting prospect, bringing means of modeling preshapes continuously, while limiting the number of additional latent dimensions.

7 Conclusion

We have presented a method for learning and representing object grasp affordances probabilistically, and we have demonstrated its applicability through a large experiment on an autonomous platform.

Object grasp affordances are modeled with grasp densities, which capture the success probability of object grasps. These densities are represented continuously in 6D using kernel density estimation. Grasp densities are refined through experience. Grasps drawn from an initial density are evaluated by the robot, and successful grasps are used to create an empirical density. Initial densities are built from 3D object-edge reconstructions, which directs exploration towards edge grasps.

We assembled an experiment setup which efficiently implements a realistic learning environment, in which 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 which quantifies the progress made from initial to empirical densities. We have also evaluated empirical densities in a realistic usage scenario, where the robot effectively selects the grasp with the highest success probability amongst the grasps that are within its reach. Result are particularly convincing given the low level of external control on the overall experimental process.

Acknowledgments

We warmly thank Emre Başeski, Mila Popović, Younes Touati, and Anders Glent Buch for their precious help with the ECV descriptor implementation, robot kinematics, and robot supervision. We also thank Volker Krüger and Dennis Herzog for their support during the recording of the human demonstration data.

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] G. E. P. Box and M. E. Muller. A note on the generation of random normal deviates. *The Annals of Mathematical Statistics*, 29(2):610–611, 1958.

[3] M. T. Ciocarlie and P. K. Allen. Hand posture subspaces for dexterous robotic grasping. *Int. J. Rob. Res.*, 28(7):851–867, 2009.

[4] C. de Granville, J. Southerland, and A. H. Fagg. Learning grasp affordances through human demonstration. In *IEEE International Conference on Development and Learning*, 2006.

[5] R. Detry, E. Başeski, N. Krüger, M. Popović, Y. Touati, O. Kroemer, J. Peters, and J. Piater. Learning object-specific grasp affordance densities. In *IEEE International Conference on Development and Learning*, pages 1–7, 2009.

[6] R. Detry, E. Başeski, M. Popović, Y. Touati, N. Krüger, O. Kroemer, J. Peters, and J. Piater. Learning continuous grasp affordances by sensorimotor exploration. In O. Sigaud and J. Peters, editors, *From Motor Learning to Interaction Learning in Robots*, pages 451–465. Springer-Verlag, 2010.

[7] 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*, pages 2287–2293, 2010.

[8] R. Detry, N. Pugeault, and J. Piater. A probabilistic framework for 3D visual object representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(10):1790–1803, 2009.

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

[10] S. Ekvall and D. Kragic. Interactive grasp learning based on human demonstration. In *IEEE International Conference on Robotics and Automation*, 2004.

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

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

[13] K. Huebner, K. Welke, M. Przybylski, N. Vahrenkamp, T. Asfour, D. Kragic, and R. Dillmann. Grasping known objects with humanoid robots: A box-based approach. In *International Conference on Advanced Robotics*, 2009.

[14] J. A. Jørgensen, L.-P. Ellekilde, and H. G. Petersen. Robworksim – an open simulator for sensor based grasping. In *Proceedings for the joint conference of ISR 2010 (41st International Symposium on Robotics) and ROBOTIK 2010 (6th German Conference on Robotics)*, 2010.

[15] D. Kraft, R. Detry, N. Pugeault, E. Başeski, F. Guerin, J. Piater, and N. Krüger. Development of object and grasping knowledge by robot exploration. *IEEE Transactions on Autonomous Mental Development*, 2(4):368–383, 2010.

[16] 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.

[17] D. Kragic, A. T. Miller, and P. K. Allen. Real-time tracking meets online grasp planning. In *IEEE International Conference on Robotics and Automation*, pages 2460–2465, 2001.

[18] O. Kroemer, R. Detry, J. Piater, and J. Peters. Combining active learning and reactive control for robot grasping. *Robotics and Autonomous Systems*, 58(9):1105–1116, 2010.

[19] J. J. Kuffner and S. M. Lavalle. RRT-Connect: An efficient approach to single-query path planning. In *IEEE International Conference on Robotics and Automation*, 2000.

[20] J. Lee and M. Verleysen. *Nonlinear dimensionality reduction*. Springer Verlag, 2007.

[21] A. T. Miller, S. Knoop, H. Christensen, and P. K. Allen. Automatic grasp planning using shape primitives. In *IEEE International Conference on Robotics and Automation*, volume 2, pages 1824–1829, 2003.

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

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

[24] 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*, 2010.

[25] 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*, 2010. (to appear).

[26] 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*, 2010.

[27] E. Sahin, M. Cakmak, M. R. Dogar, E. Ugur, and G. Ucoluk. To afford or not to afford: A new formalization of affordances towards affordance-based robot control. *Adaptive Behavior*, 2007.

[28] M. Salganicoff, L. H. Ungar, and R. Bajcsy. Active learning for vision-based robot grasping. *Mach. Learn.*, 23:251–278, May 1996.

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

[30] K. Shimoga. Robot grasp synthesis algorithms: A survey. *The International Journal of Robotics Research*, 15(3):230, 1996.

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

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

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

[34] A. T. A. Wood. Simulation of the von Mises-Fisher distribution. *Communications in Statistics—Simulation and Computation*, 23(1):157–164, 1994.