This chapter is part of "O. Sigaud and J. Peters, editors. *From Motor Learning to Interaction Learning in Robots*, volume 264 of *Studies in Computational Intelligence*. Springer, 2010".

http://dx.doi.org/10.1007/978-3-642-05181-4_19

Learning Continuous Grasp Affordances by Sensorimotor Exploration

R. Detry[1], E. Başeski[2], M. Popović[2], Y. Touati[2], N. Krüger[2], O. Kroemer[3], J. Peters[3] and J. Piater[1]

Abstract We develop means of learning and representing object grasp affordances probabilistically. By *grasp affordance*, we refer to an entity that is able to assess whether a given relative object-gripper configuration will yield a stable grasp. These affordances are represented with *grasp densities*, continuous probability density functions defined on the space of 3D positions and orientations. Grasp densities are registered with a visual model of the object they characterize. They are exploited by aligning them to a target object using visual pose estimation. Grasp densities are refined through experience: A robot "plays" with an object by executing grasps drawn randomly for the object's grasp density. The robot then uses the outcomes of these grasps to build a richer density through an importance sampling mechanism. Initial grasp densities, called *hypothesis* densities, are bootstrapped from grasps collected using a motion capture system, or from grasps generated from the visual model of the object. Refined densities, called *empirical* densities, represent affordances that have been confirmed through physical experience. The applicability of our method is demonstrated by producing empirical densities for two object with a real robot and its 3-finger hand. Hypothesis densities are created from visual cues and human demonstration.

R. Detry and J. Piater
University of Liège, Belgium
e-mail: Renaud.Detry@ULg.ac.be

E. Başeski, M. Popović, Y. Touati and N. Krüger
University of Southern Denmark

O. Kroemer and J. Peters
MPI for Biological Cybernetics, Tübingen, Germany

O. Sigaud, J. Peters (Eds.): From Mot. Learn. to Inter. Learn. in Rob., SCI 264, pp. 451–465.
springerlink.com

1 Introduction

Grasping previously unknown objects is a fundamental skill of autonomous agents. Human grasping skills improve with growing experience with certain objects. In this chapter, we describe a mechanism that allows a robot to learn grasp affordances [11] of objects described by learned visual models. Our first aim is to organize and memorize, independently of grasp information sources, the whole knowledge that an agent has about the grasping of an object, in order to facilitate reasoning on grasping solutions and their likelihood of success. A *grasp affordance* corresponds to the the different ways to place a hand or a gripper near an object so that closing the gripper will produce a stable grip. We represent the affordance of an object for a certain grasp type through a continuous probability density function defined on the 6D gripper pose space $SE(3)$, within an object-relative reference frame. The computational encoding 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 [28].

The second contribution of this chapter is a framework that allows an agent to learn initial affordances from various grasp cues, and enrich its grasping knowledge through experience. Affordances are initially constructed from human demonstration, or from a model-based method [1]. The grasp data produced by these *grasp sources* is used to build continuous *grasp hypothesis densities* (Section 5). These densities are registered with 3D visual object model learned beforehand [8], which allows a robotic agent to execute *samples* from a grasp hypothesis density under arbitrary object poses, by using the visual model to estimate the 3D pose of the object.

The success rate of grasp samples depends on the source that is used to produce initial grasp data. However, no existing method can claim to be perfect. For example, data collected from human demonstration will suffer from the physical and mechanical difference between a human hand and a robotic gripper. In the case of grasps computed from a 3D model, results will be impeded by errors in the model, such as missing parts or imprecise geometry. In all cases, only a fraction of the hypothesis density samples will succeed; it thus seems necessary to also learn from experience. To this end, we use samples from grasp hypothesis densities that lead to a successful grasp to learn *grasp empirical densities*, i.e. grasps that have been confirmed through experience.

A unified representation of grasp affordances can potentially lead to many different applications. For instance, a grasp planner could combine a grasp density with hardware physical capabilities (robot reachability) 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 the use of continuous grasp success likelihoods to infer robustness requirements

on the execution particular grasp: if a grasp is centered on a narrow peak, pose estimation and servoing should be performed with more caution than when the grasp is placed in a wide region of high success likelihood.

2 Related Work

Object grasps can emerge in many different ways. One can for instance learn 2D image patches that predict stable grasps. For example, Saxena et al. [27] have trained a classifier on a set of 2D images that were hand-labeled with good grasping points. Good grasping points are then identified in several views of an object and triangulated to infer their 3D position.

Grasping solutions can also emerge from the geometric properties of an object, typically obtained from a 3D object model. The most popular 3D model for grasping is probably the 3D mesh [15, 22], obtained e.g. from CAD or superquadrics fitting [2]. However, grasping has also successfully been achieved using models consisting of 3D surface patches [26], 3D edge segments [1], or 3D points [13]. When combined with an object pose estimation technique, the previous methods allow a robot to execute a grasp on a specific object. This involves object pose estimation, computation of a grasp on the aligned model, then servoing to the object and performing the grasp [15].

In learning a continuous grasp affordance, 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 $\mathbf{p}(O|G)$ or learning $\mathbf{p}(G|O)$. The former allows one to directly compute a probability of success. 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. 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 computationally feasible.

This chapter develops a method for learning success-conditional grasp densities, closely related in spirit to the work of de Granville et al. [5]. In their work, affordances correspond to object-relative hand approach orientations, although an extension where object-relative positions are also modeled is under way [4]. The aim of the authors is to build compact sets of canonical grasp approaches from human demonstration; they mean to compress a large number of examples provided by a human teacher into a small number of clusters. An affordance is expressed through a density represented as a mixture of position-orientation kernels; machine learning techniques are used to compute mixture and kernel parameters that best fit the data. This is quite different from our approach, where a density is represented with a much larger number of simpler kernels. Conceptually, using a larger number of ker-

nels allows us to use significantly simpler learning methods (down to mere resampling of input data, see Section 5.1). Also, the representation of a grasp cluster through a single position-orientation kernel requires the assumption that hand position and orientation are independent within the cluster, which is generally not true. Representing a cluster with many particles can intrinsically capture more of the position-orientation correlation (see Section 6, and in particular Fig. 6). The affordance densities presented by de Granville et al. correspond to the hypothesis densities developed in this chapter.

Instead of considering success-conditional grasp probabilities, Montesano et al. [23] formalize grasp affordances as success probabilities $\mathbf{p}(O|I)$, where I is a local image patch. A robot thus learns a mapping from 2D image patches to grasp success probabilities, where a grasp is parametrized by its 2D gripper position. From a procedural viewpoint, the method of Montesano et al. differs from ours in its explicit exploitation of failed grasps, whereas in our work, empirical densities are learned from successful grasps only. We note that, in a probabilistic sense, our learning method does take failed grasps into account, through the absence of learning data in regions where grasps were sampled and failed. However, we agree that making active use of failed trials may increase robustness, and we intend to evaluate this option in future work. Another promising avenue lies in active learning of grasp options, as demonstrated by Kroemer et al. [16].

Learning grasp affordances from experience was also demonstrated by Stoytchev [29, 30]. In his work, a robot discovers successful grasps through random exploratory actions on a given object. When subsequently confronted with the same object, the robot is able to generate a grasp that should present a high likelihood of success.

In this chapter, learning may be bootstrapped with grasp data provided by a motion capture system, a process that constitutes a simple form of imitation learning. For a broader discussion of imitation learning, we refer the reader to two dedicated chapters within this collection [20, 19].

The system developed in this chapter is build on a set of existing methods which are described in Section 3. The visual object model to which affordances are attached is the part-based model of Detry et al. [8] (Section 3.3). An object is modeled with a hierarchy of increasingly expressive object parts called *features*. The single top feature of a hierarchy represents the whole object. Features at the bottom of the hierarchy represent short 3D edge segments for which evidence is collected from stereo imagery via the Early-Cognitive-Vision (ECV) system of Krüger et al. [17, 25] (Section 3.1). In the following, we refer to these edge segments as *ECV descriptors*. The hierarchical model grounds its visual evidence in ECV reconstructions: a model is learned from segmented ECV descriptors, and the model can be used to recover the pose of the object within an ECV representation of a cluttered scene.

The mathematical representation of grasp densities and their association to hierarchical object models is discussed in Section 4. In Section 5, we demonstrate the learning and refining of grasp densities from two grasp sources. The

(a) ECV descriptors (b) Accumulated reconstructions

Fig. 1: ECV reconstruction. Each ECV descriptor is rendered with a small plane patch. Patch normals are not part of ECV descriptors; they are arbitrarily defined for the purpose of 3D rendering.

first source is imitation of human grasps. The second source uses a model-based algorithm which extracts grasping cues from an ECV reconstruction (Section 3.2).

3 Methods

This section briefly describes the methods that are brought together for modeling the visual percepts of an object, and for bootstrapping hypothesis densities from visual cues. These sophisticated methods have proved essential for a robust execution of grasps on arbitrary objects in arbitrary poses.

3.1 Early Cognitive Vision

ECV descriptors [17, 25] represent short edge segments in 3D space, each ECV descriptor corresponding to a circular image patch with a 7-pixel diameter. To create an ECV reconstruction, pixel patches are extracted along image contours, within images captured with a calibrated stereo camera. The ECV descriptors are then computed with stereopsis across image pairs; each descriptor is thus defined by a 3D position and 3D edge orientation. Descriptors may be tagged with color information, extracted from their corresponding 2D patches (Fig. 1a).

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 [12]. Assuming that the motion adequately spans the object pose space, a complete 3D-edge reconstruction

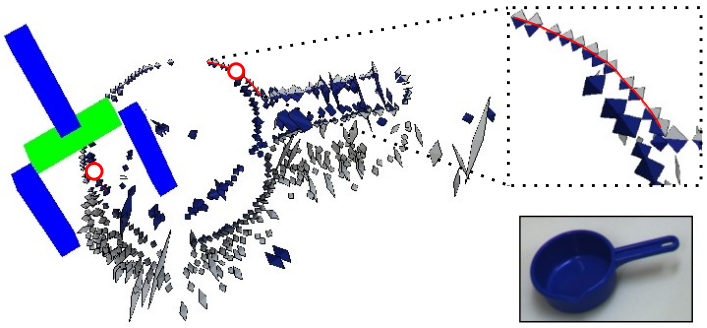


Fig. 2: Grasp reflex based on visual data. Each ECV descriptor is rendered with a small plane patch. Patch normals are not part of ECV descriptors; they are arbitrarily defined for the purpose of 3D rendering.

of the object can be generated, eliminating self-occlusion issues [14] (see Fig. 1b).

3.2 Grasp Reflex From Co-planar ECV Descriptors

Pairs of ECV descriptors that are on the same plane and which have color information such that two similar colors are pointing towards each other can be used to define grasps. Grasp position is defined by the location of one of the descriptors. Grasp orientation is calculated from the normal of the plane linking the two descriptors, and the orientation of the descriptor at which the grasp is located [14] (see Fig. 2). The grasps generated by this method will be referred to as *reflexes*. Since each pair of co-planar descriptors generates multiple reflexes, a large number of these are available.

3.3 Feature Hierarchies For 3D Visual Object Representation

As explained in Section 3.1, an ECV reconstruction models a scene or an object with low-level descriptors. This section outlines a higher-level 3D object model [8] that grounds its visual evidence in ECV representations.

An object is modeled with a hierarchy of increasingly expressive object parts called *features*. Features at the bottom of the hierarchy (*primitive* features) represent ECV descriptors. Higher-level features (*meta*-features) represent geometric configurations of more elementary features. The single top feature of a hierarchy represents the object.

Unlike many part-based models, a hierarchy consists of features that may have several *instances* in a scene. To illustrate this, let us consider a part-based model of a bike, in which we assume a representation of wheels. Traditional part-based models [10, 3] would work by creating two wheel parts – one for each wheel. Our hierarchy however uses a single *generic* wheel feature; it stores the information on the existence of *two* wheels *within* the wheel feature. Likewise, a primitive feature represents a *generic* ECV descriptor, e.g. any descriptor that has a red-like color. While an object like the basket of Fig. 1b produces hundreds of red ECV descriptors, a hierarchy representing the basket will, in its simplest form, contain a single red-like primitive feature; it will encode internally that this feature has many instances within a basket object.

A hierarchy is implemented in a Markov tree. Features correspond to hidden nodes of the network; when a model is associated to a scene (during learning or detection), the pose distribution of feature i in the scene is represented through a random variable X_i. Random variables are thus defined over the pose space, which exactly corresponds to the Special Euclidean group $SE(3) = \mathbb{R}^3 \times SO(3)$. The random variable X_i associated to feature i effectively links that feature to its instances: X_i represents as one probability density function the pose distribution of all the instances of feature i, therefore avoiding specific model-to-scene correspondences.

The geometric relationship between two neighboring features i and j is encoded in a compatibility potential $\psi_{ij}(X_i, X_j)$. A compatibility potential represents the pose distribution of all the instances of the child feature in a reference frame defined by the parent feature; potentials are thus also defined on $SE(3)$.

The only observable features are primitive features, which receive evidence from the ECV system. Each primitive feature i is linked to an observed variable Y_i; the statistical dependency between a hidden variable X_i and its observed variable Y_i is encoded in an observation potential $\phi_i(X_i)$, which represents the pose distribution of ECV descriptors that have a color similar to the color of primitive feature i.

Density functions (random variables, compatibility potentials, observation potentials) are represented nonparametrically: a density is represented by a set of particles [8].

3.4 Pose Estimation

The hierarchical model presented above can be used to estimate the pose of a known object in a cluttered scene. Estimating the pose of an object amounts to deriving a posterior pose density for the top feature of its hierarchy, which involves two operations [8]:

1. Extract ECV descriptors, and transform them into observation potentials.

2. Propagate evidence through the graph using an applicable inference algorithm.

Each observation potential $\phi_i(X_i)$ is built from a subset of the early-vision observations. The subset that serves to build the potential $\phi_i(X_i)$ is the subset of ECV descriptors that have a color that is close enough to the color associated to primitive feature i.

Evidence is propagated through the hierarchy using a belief propagation (BP) algorithm [24, 31]. BP works by exchanging *messages* between neighboring nodes. Each message carries the belief that the sending node has about the pose of the receiving node. In other words, a message allows the sending feature to probabilistically vote for all the poses of the receiving feature that are consistent with its own pose – consistency being defined by the compatibility potential through which the message flows. Through message passing, BP propagates collected evidence from primitive features to the top of the hierarchy; each feature probabilistically votes for all possible object configurations consistent with its pose density. A consensus emerges among the available evidence, leading to one or more consistent scene interpretations. The pose likelihood for the whole object is eventually read out of the top feature; if the object is present twice in a scene, the top feature density should present two major modes. The global belief about the object pose may also be propagated from the top node down the hierarchy, reinforcing globally consistent evidence and permitting the inference of occluded features.

Algorithms that build hierarchies from accumulated ECV reconstructions are discussed in prior work [7].

4 Representing Grasp Densities

This section is focused on the probabilistic representation of grasp affordances. By *grasp affordance*, we refer to the different ways to place a hand or a gripper near an object so that closing the gripper will produce a stable grip. The grasps we consider are parametrized by a 6D gripper pose composed of a 3D position and a 3D orientation.

From a mathematical point of view, grasp densities are identical to the visual potentials of Section 3.3. They can thus be encoded with the same nonparametric representation. Density evaluation is performed by assigning a kernel function to each particle supporting the density, and summing the evaluation of all kernels. Sampling from a distribution is performed by sampling from the kernel of a particle ℓ selected from $\mathbf{p}(\ell = i) \propto w^i$, where w^i is the weight of particle i.

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 $SE(3)$ point x into a translation

λ and a rotation θ by

$$x = (\lambda, \theta), \quad \mu = (\mu_t, \mu_r), \quad \sigma = (\sigma_t, \sigma_r),$$

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, $\mathbf{N}(\cdot)$ is a trivariate isotropic Gaussian kernel, and $\mathbf{\Theta}(\cdot)$ is an orientation kernel defined on $SO(3)$. Denoting by θ' and μ'_r the quaternion representations of θ and μ_r [18], we define the orientation kernel with the Dimroth-Watson distribution [21]

$$\mathbf{\Theta}(\theta; \mu_r, \sigma_r) = \mathbf{W}(\theta'; \mu'_r, \sigma_r) = C_w(\sigma_r) e^{\sigma_r (\mu_r'^{\top} \theta')^2} \tag{2}$$

where $C_w(\sigma_r)$ is a normalizing factor. This kernel corresponds to a Gaussian-like distribution on $SO(3)$. The Dimroth-Watson distribution inherently handles the double cover of $SO(3)$ by quaternions [5].

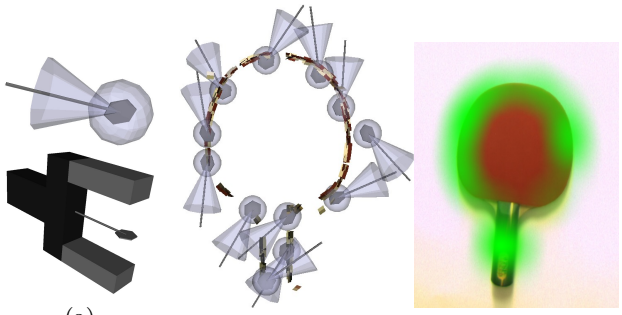
The bandwidth σ associated to a density should ideally be selected jointly in $\mathbb{R}^3$ and $SO(3)$. However, this is difficult to do. Instead, we set the orientation bandwidth σ_r to a constant allowing about 10° of deviation; the location bandwidth σ_t is then selected using a k-nearest neighbor technique [28].

The expressiveness of a single $SE(3)$ kernel (1) is rather limited: location and orientation components are both isotropic, and within a kernel, location and orientation are modeled independently. Nonparametric methods account for the simplicity of individual kernels by employing a large number of them: a grasp density will typically be supported by a thousand particles. Fig. 3a shows an intuitive rendering of an $SE(3)$ kernel from a grasp density. Fig. 3b and Fig. 3c illustrate 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 3.4), the object grasp density can be similarly aligned, and one can readily draw grasps from the density and execute them onto the object. A deeper integration of the visual model with grasp densities has been covered in prior work [6].

5 Learning Grasp Densities

This section explains how hypothesis densities are learned from source data (Section 5.1), and how empirical densities are learned from experience (Section 5.2).



(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 table-tennis paddle (for clarity, only 12 kernels are rendered). Fig. (c) indicates with a green mask of varying opacity the values of the location component of the same grasp density along the plane of the paddle (orientations were ignored to produce this last illustration).

5.1 Hypothesis Densities From Examples

Initial grasp knowledge, acquired for instance from imitation or reflex, is structured as a set of grasps parametrized by a 6D pose. Given the nonparametric representation, building a density from a set of grasps is straightforward – grasps can directly be used as particles representing the density. We typically limit the number of particles in a density to a thousand; if the number of grasps in a set is larger than 1000, the density is *resampled*: kernels are associated the particles, and 1000 samples are drawn and used as a representation replacement.

5.2 Empirical Densities Through Familiarization

As the name suggests, hypothesis densities do not pretend to reflect the true properties of an object. Their main defect is that they may strongly suggest grasps that might not be applicable at all, for instance because of gripper discrepancies in imitation-based hypotheses. A second, more subtle issue is that the grasp data used to learn hypothesis densities will generally be afflicted with a source-dependent spatial bias. A very good example can be made

from the reflex computation of Section 3.2. Reflexes are computed from ECV visual descriptors. Therefore, parts of an object that have a denser visual resolution will yield more reflexes, incidentally biasing the corresponding region of the hypothesis density to a higher value. The next paragraph explains how grasping experience can be used to compute new densities (*empirical* densities) that better reflect gripper-object properties.

Empirical densities are leaned from the execution of *samples* from a hypothesis density, intuitively allowing the agent to familiarize itself with the object by discarding wrong hypotheses and refining good ones. Familiarization thus essentially consists in autonomously learning an *empirical* density from the outcomes of sample executions. A simple way to proceed is to build an empirical density directly from successful grasp samples. However, this approach would inevitably propagate the spatial bias mentioned above to empirical densities. Instead, we use importance sampling [9] to properly weight grasp outcomes, allowing us to draw samples from the physical grasp affordance of an object. The weight associated to a grasp sample x is computed as $\mathbf{a}(x) \,/\, \mathbf{q}(x)$, where $\mathbf{a}(x)$ is 1 if the execution of x has succeeded, 0 else, and $\mathbf{q}(x)$ corresponds to the value of the continuous hypothesis density at x. A set of these weighted samples directly forms a grasp empirical density that faithfully and uniformly reflects intrinsic gripper-object properties.

5.3 Updating Empirical Densities In Long-Term Interaction

In long-term interaction, a robot is constantly producing new evidence which should ideally be used to continuously enhance empirical densities. The methodology presented above can easily be adapted for basic long-term learning. Essentially, the solution stems from observing that the IS learning of empirical densities may in fact use any arbitrary function as hypothesis density. In an initial learning cycle, the hypothesis density is computed from grasp cues. Let us call this initial hypothesis density the *bootstrap* density. In the next learning cycle, the empirical density from the first cycle may be linearly combined with the bootstrap density to form a new hypothesis density that represents a trade-off between exploration of new possibilities and safe reproduction of known experience. Once enough samples from the new hypothesis density have been experienced, the empirical density can be replaced by an updated representation. In long-term interaction, hypothesis densities are thus successively computed as weighted sums of the current empirical density and the bootstrap density. Giving a high weight to the empirical density triggers top-down learning, i.e. refining globally known affordance. Conversely, focusing on the bootstrap density corresponds to bottom-up learning, i.e. integrating new low-level evidence into the model.

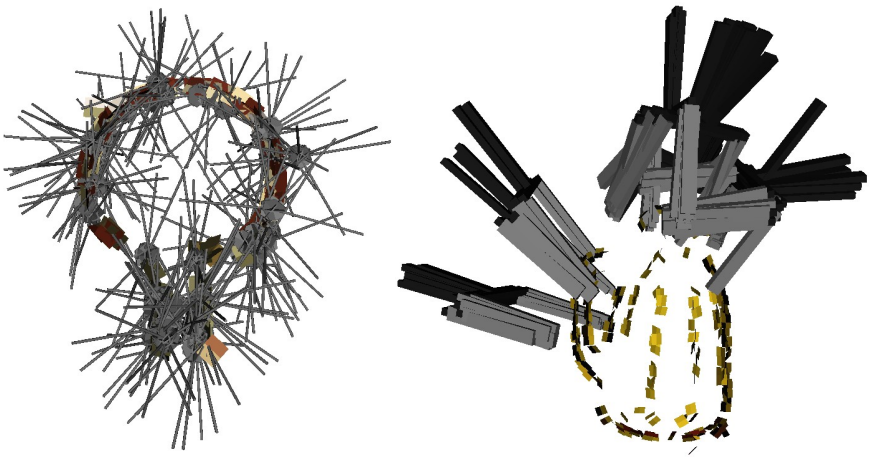


Fig. 4: Particles supporting grasp hypothesis densities.

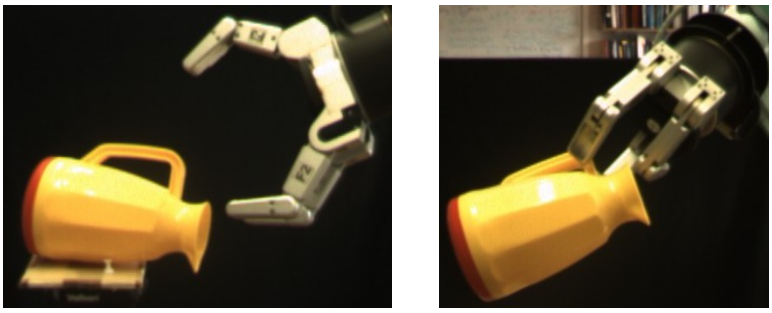


Fig. 5: Barrett hand grasping the toy jug.

6 Results

This section illustrates hypothesis densities learned from imitation and reflexes, and empirical densities are learned by grasping objects with a 3-finger Barrett hand. Densities are built for two objects: the table-tennis paddle of Fig. 3, and a toy plastic jug (Fig. 5). The experimental scenario is described below.

For each object, the experiment starts with a visual hierarchical model, and a set of grasps. For the paddle, grasps are generated with the method described in Section 3.2. Initial data for the jug was collected through human demonstration, using a motion capture system. From these data, a hypothesis density is built for each object. The particles supporting the hypothesis densities are rendered in Fig. 4.

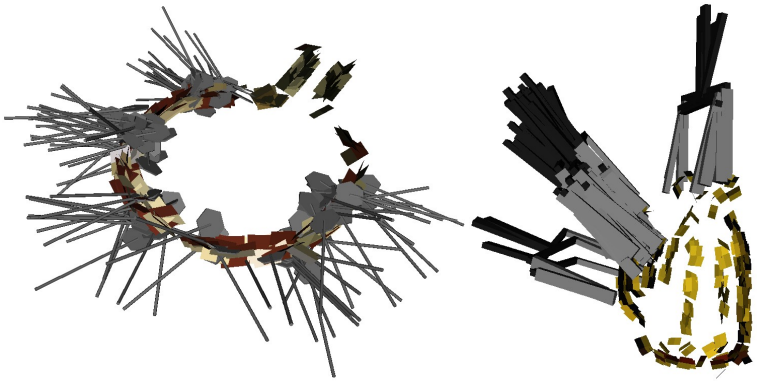


Fig. 6: Samples drawn from grasp empirical densities.

In order to refine affordance knowledge, feedback on the execution of hypothesis density samples is needed. Grasps are executed with a Barrett hand mounted on an industrial arm. As illustrated in Fig. 5, the hand preshape is a parallel-fingers, opposing-thumb configuration. The reference pose of the hand is set for a pinch grasp, with the tool center point located in-between the tips of the fingers – similar to the reference pose illustrated in Fig. 3a. A grasp is considered successful if the robot is able to firmly 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. Sets of 100 and 25 successful grasps were collected for the paddle and the jug respectively. This information was then used to build a grasp empirical density, following the procedure described in Section 5.2. Samples from the resulting empirical densities are shown in Fig. 6. For the paddle, the main evolution from hypothesis to empirical density is the removal of a large number of grasps for which the gripper wrist collides with the paddle body. Grasps presenting a steep approach relative to the plane of the paddle were also discarded, thereby preventing fingers from colliding with the object during hand servoing. None of the pinch-grasps at the paddle handle succeeded, hence their absence from the empirical density.

While grasping the top of the jug is easy for a human hand, it proved to be very difficult for the Barrett hand with parallel fingers and opposing thumb. Consequently, a large portion of the topside grasps suggested by the hypothesis density are not represented in the empirical density. The most reliable grasps approach the handle of the jug from above; these grasps are strongly supported in the empirical density.

The left image of Fig. 6 clearly illustrates the correlation between grasp positions and orientations: moving along the edge of the paddle, grasp approaches are most often roughly perpendicular to the local edge tangent. The nonparametric density representation successfully captures this correlation.

7 Conclusion And Future Work

We presented a framework for representing and learning object grasp affordances probabilistically. The affordance representation is nonparametric: an affordance is recorded in a continuous probability density function supported by a set of particles.

Grasp densities are initially learned from visual cues or imitation, leading to grasp hypothesis densities. Using the visual model for pose estimation, an agent is able to execute *samples* from a hypothesis density under arbitrary object poses. Observing the outcomes of these grasps allows the agent to learn from experience: an importance sampling algorithm is used to infer faithful object grasp properties from successful grasp samples. The resulting *grasp empirical densities* eventually allow for more robust grasping.

Importance Sampling is a batch learning method, that requires the execution of a large number of grasps before an empirical density can be built. Learning empirical densities *on-line* would be very convenient, which is a path we plan to explore next.

We currently learn visual and grasp models independently. Yet, a part-based representation offers an elegant way to *locally* encode visuomotor descriptions. One of our goals is to learn visual parts that share the same grasp properties across different objects. This way, a grasp *feature* would be directly and exclusively connected to the visual evidence that predicts its applicability, allowing for its generalization across objects.

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). We thank Volker Krüger and Dennis Herzog for their support during the recording of the human demonstration data.

References

1. Daniel Aarno, Johan Sommerfeld, Danica Kragic, Nicolas Pugeault, Sinan Kalkan, Florentin Wörgötter, Dirk Kraft, and Norbert Krüger. Early reactive grasping with second order 3D feature relations. In *The IEEE International Conference on Advanced Robotics*, 2007.
2. G. Biegelbauer and M. Vincze. Efficient 3D object detection by fitting superquadrics to range image data for robot's object manipulation. In *IEEE International Conference on Robotics and Automation*, 2007.
3. G. Bouchard and B. Triggs. Hierarchical part-based visual object categorization. In *Computer Vision and Pattern Recognition*, volume 1, pages 710–715, 2005.

4. C. de Granville and A. H. Fagg. Learning grasp affordances through human demonstration. *Submitted to the Journal of Autonomous Robots*, 2009.
5. Charles de Granville, Joshua Southerland, and Andrew H. Fagg. Learning grasp affordances through human demonstration. In *Proceedings of the International Conference on Development and Learning (ICDL'06)*, 2006.
6. 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.
7. Renaud Detry and Justus H. Piater. Hierarchical integration of local 3D features for probabilistic pose recovery. In *Robot Manipulation: Sensing and Adapting to the Real World (Workshop at Robotics, Science and Systems)*, 2007.
8. Renaud Detry, Nicolas Pugeault, and Justus Piater. A probabilistic framework for 3D visual object representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2009.
9. A. Doucet, N. de Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer, 2001.
10. Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient matching of pictorial structures. In *Conference on Computer Vision and Pattern Recognition (CVPR 2000)*, pages 2066–, 2000.
11. James J. Gibson. *The Ecological Approach to Visual Perception*. Lawrence Erlbaum Associates, 1979.
12. R.I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.
13. Kai Huebner, Steffen Ruthotto, and Danica Kragic. Minimum volume bounding box decomposition for shape approximation in robot grasping. Technical report, KTH, 2007.
14. D. Kraft, N. Pugeault, E. Başeski, M. Popović, D. Kragic, S. Kalkan, F. Wörgötter, and N. Krüger. Birth of the Object: Detection of Objectness and Extraction of Object Shape through Object Action Complexes. *Special Issue on "Cognitive Humanoid Robots" of the International Journal of Humanoid Robotics*, 2008. (accepted).
15. Danica Kragic, Andrew T. Miller, and Peter K. Allen. Real-time tracking meets online grasp planning. In *Proceedings of the 2001 IEEE International Conference on Robotics and Automation*, pages 2460–2465, 2001.
16. Oliver Kroemer, Renaud Detry, Justus Piater, and Jan Peters. Active learning using mean shift optimization for robot grasping. In *International Conference on Intelligent RObots and Systems*, 2009.
17. N. Krüger, M. Lappe, and F. Wörgötter. Biologically Motivated Multi-modal Processing of Visual Primitives. *The Interdisciplinary Journal of Artificial Intelligence and the Simulation of Behaviour*, 1(5):417–428, 2004.
18. James Kuffner. Effective sampling and distance metrics for 3D rigid body path planning. In *Proc. 2004 IEEE Int'l Conf. on Robotics and Automation (ICRA 2004)*. IEEE, May 2004.
19. S. Lallee, E. Yoshida, A. Mallet, F. Nori, L. Natale, G. Metta, F. Warneken, and P.F. Dominey. Human-robot cooperation based on interaction learning. In *From Motor to Interaction Learning in Robots*. Springer, 2009.
20. Manuel Lopes, Francisco Melo, Luis Montesano, and Jose Santos-Victor. Cognitive processes in imitation: Overview and computational approaches. In *From Motor to Interaction Learning in Robots*. Springer, 2009.
21. K. V. Mardia and P. E. Jupp. *Directional Statistics*. Wiley Series in Probability and Statistics. Wiley, 1999.
22. A. T. Miller, S. Knoop, H. Christensen, and P. K. Allen. Automatic grasp planning using shape primitives. In *Proceedings of the IEEE International Conference on Robotics and Automation, 2003*, volume 2, pages 1824–1829, 2003.

23. L. Montesano and M. Lopes. Learning grasping affordances from local visual descriptors. In *International Conference on Development and Learning*, 2009.
24. Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.
25. Nicolas Pugeault. *Early Cognitive Vision: Feedback Mechanisms for the Disambiguation of Early Visual Representation*. Vdm Verlag Dr. Müller, 2008.
26. Mario Richtsfeld and Markus Vincze. Robotic grasping based on laser range and stereo data. In *International Conference on Robotics and Automation*, 2009.
27. 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.
28. B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC, 1986.
29. Alexander Stoytchev. Toward learning the binding affordances of objects: A behavior-grounded approach. In *Proceedings of AAAI Symposium on Developmental Robotics*, pages 17–22, Stanford University, Mar 21-23 2005.
30. 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.
31. Erik B. Sudderth, Alexander T. Ihler, William T. Freeman, and Alan S. Willsky. Nonparametric belief propagation. In *Computer Vision and Pattern Recognition*, Los Alamitos, CA, USA, 2003. IEEE Computer Society.