Learning Object-specific Grasp Affordance Densities

R. Detry*, E. Başeski†, M. Popović†, Y. Touati†, N. Krüger†, O. Kroemer‡, J. Peters‡ and J. Piater*
*University of Liège, Belgium. Email: Renaud.Detry@ULg.ac.be
†University of Southern Denmark.
‡MPI for Biological Cybernetics, Tübingen, Germany.

Abstract—This paper addresses the issue of learning and representing *object grasp affordances*, i.e. object-gripper relative configurations that lead to successful grasps. The purpose of grasp affordances is to organize and store the whole knowledge that an agent has about the grasping of an object, in order to facilitate reasoning on grasping solutions and their achievability. The affordance representation consists in a continuous probability density function defined on the 6D gripper pose space – 3D position and orientation –, within an object-relative reference frame. Grasp affordances are initially learned from various sources, e.g. from imitation or from visual cues, leading to *grasp hypothesis densities*. Grasp densities are attached to a learned 3D visual object model, and pose estimation of the visual model allows a robotic agent to execute *samples* from a grasp hypothesis density under various object poses. Grasp outcomes are used to learn *grasp empirical densities*, i.e. grasps that have been confirmed through experience. We show the result of learning grasp hypothesis densities from both imitation and visual cues, and present grasp empirical densities learned from physical experience by a robot.

I. INTRODUCTION

Grasping previously unknown objects is a fundamental skill of autonomous agents. Human grasping skills improve with growing experience with certain objects. In this paper, we describe a mechanism that allows a robot to learn grasp affordances 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. 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* [21].

The second contribution of this paper 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 VI). These densities are attached to a 3D visual object model learned before-hand [7], 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 imitation 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.

II. RELATED WORK

Object grasps can emerge in many different ways. A popular approach is to compute grasping solutions 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 [13], [17], obtained e.g. from CAD or superquadrics fitting [2]. However, grasping has also successfully been achieved using models consisting of 3D surface patches [20], 3D edge segments [1], or 3D points [11].

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 [13].

Means of representing grasp affordances probabilistically have been discussed in the work of de Granville et al. [5], which is quite closely related in spirit to ours. In this 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 kernels allows us to use significantly simpler learning methods (down to mere resampling of input data, see Section VI-A). 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 VII, and in particular Fig. 7). The affordance densities presented by de Granville et al. correspond to the hypothesis densities developed in this paper.

III. SYSTEM OVERVIEW

The visual object model to which affordances are attached is the part-based model of Detry et al. [7] (Section IV-C). 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. [14], [19] (Section IV-A). 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 V. In Section VI, we demonstrate the learning and refining of grasp densities from two grasp sources. The first source is imitation of human grasps. The second source uses a model-based algorithm which extracts grasping cues from an ECV reconstruction (Section IV-B).

IV. METHODS

A. Early Cognitive Vision

ECV descriptors [14], [19] represent short edge segments in 3D space, each ECV descriptor corresponding to a patch of about 25 square millimeters of *object* surface. To create an ECV reconstruction, pixel patches are extracted along

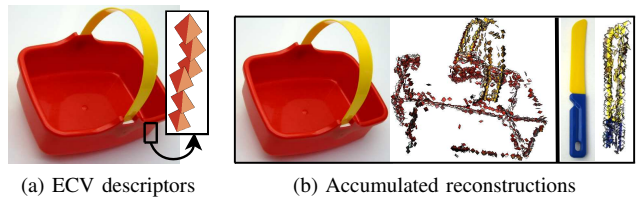(a) ECV descriptors   (b) Accumulated reconstructions
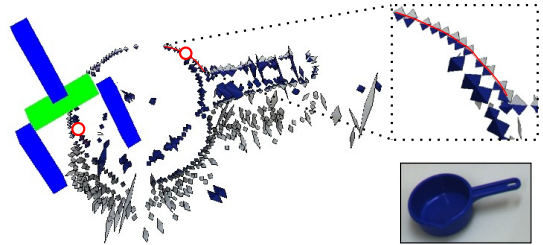
Fig. 1. ECV reconstructions

Fig. 2. Grasp reflex based on visual data.

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 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 [10]. Assuming that the motion adequately spans the object pose space, a complete 3D reconstruction of the object can be generated, eliminating self-occlusion issues [12] (see Fig. 1b).

B. 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 [12] (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.

C. Feature Hierarchies For 3D Visual Object Representation

As explained in Section IV-A, an ECV reconstruction models a scene or an object with low-level descriptors. This section outlines a higher-level 3D object model [7] 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 [9], [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. 1 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 [7].

D. 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 [7]:

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
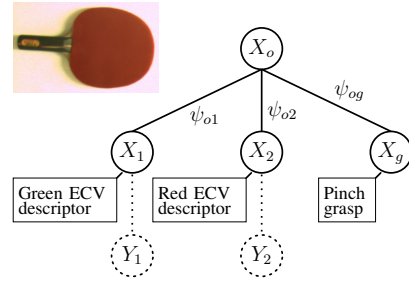

Fig. 3. Multi-sensory modeling of a table-tennis paddle with a 2-level hierarchy. The paddle is represented by feature o (top). Feature 1 represents a generic green ECV descriptor. The rectangular configuration of green edges around the handle of the paddle is encoded in ψ_{o1}. Y_1 and Y_2 are observed variables, which link features 1 and 2 to the visual evidence produced by ECV. X_g is a grasp feature, linked to the object feature through the pinch grasp affordance ψ_{og}.

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 [18], [22]. 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 [6].

V. REPRESENTING GRASP DENSITIES

This section is focused on the probabilistic representation of grasp affordances, and on the integration of grasp affordances within the hierarchical object model. 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.

A. Grasp Features

Within our framework, a grasp affordance is represented with a probability density function defined on $SE(3)$ in an object-relative reference frame. Probabilistically speaking, we store an expression of the joint distribution $\mathbf{p}(X_o, X_g)$, where X_o is the pose distribution of the object, and X_g is the grasp

affordance. This is done by adding a new "grasp" feature to the hierarchical Markov network, and linking it to the top feature (see Fig. 3). The statistical dependency of X_o and X_g is held in a compatibility potential $\psi_{og}(X_o, X_g)$, which exactly corresponds to the grasp density: $\psi_{og}(X_o, X_g)$ holds the relative configuration of grasp affordance and object pose, i.e. the grasp distribution into the reference frame of the top feature.

When an object model has been visually aligned to an object instance (i.e. when the marginal posterior of the top feature has been computed from visually-grounded bottom-up inference), the grasp affordance of the object *instance* is computed through top-down BP inference, by sending a message from X_o to X_g through $\psi_{og}(X_o, X_g)$. Intuitively, this corresponds to transforming the grasp density to align it to the current object pose, yet explicitly taking the uncertainty on object pose into account to produce a posterior grasp density that acknowledges visual noise.

B. Continuous Grasp Densities

From a mathematical point of view, grasp potentials are identical to visual potentials. They can thus be encoded with the same nonparametric density 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 (grasp potentials and grasp random variables) 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

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

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 [15], we define the orientation kernel with the Dimroth-Watson distribution [16]

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

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 [21].
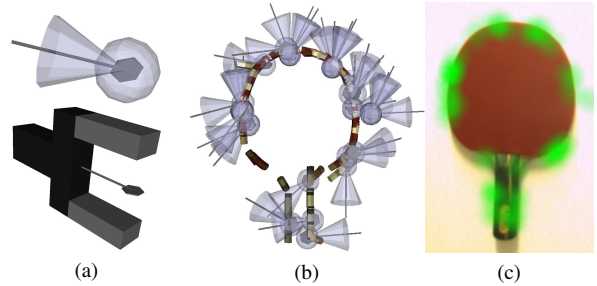

(a)　　　　　(b)　　　　　(c)

Fig. 4. 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 30 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).

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. 4a shows an intuitive rendering of an $SE(3)$ kernel from a grasp density. Fig. 4b and Fig. 4c illustrate continuous densities.

VI. LEARNING GRASP DENSITIES

This section explains how hypothesis densities are learned from source data (Section VI-A), and how empirical densities are learned from experience (Section VI-B).

A. 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.

Since we wish to record object-relative information, densities have to be transformed to the reference frame of the object. Assuming that grasp poses are initially defined in the same reference frame as the visual ECV descriptors, this can be done by aligning the hierarchical model of the object by visual inference, and transforming the particles of each grasp density in the reference frame defined by the pose of the top feature of the aligned model.

A grasp density is integrated into the hierarchical object model through a new primitive feature i. The new feature
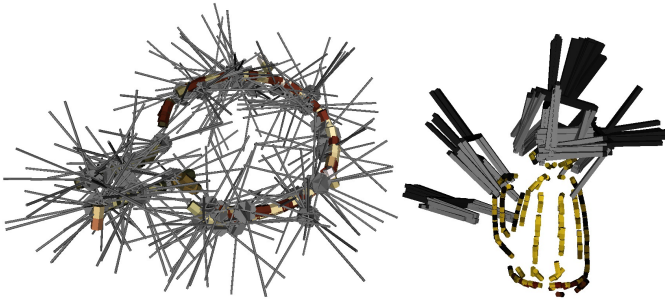
Fig. 5.   Particles supporting grasp hypothesis densities.

Fig. 6.   Barrett hand grasping the toy jug.

Fig. 7.   Samples drawn from grasp empirical densities.

is linked to the top model feature o through a potential $\psi_{io}(X_i, X_o)$ that corresponds to the object-relative density.

B. 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 IV-B. 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 [8] 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. Each empirical density is associated to the object model in the same way as hypothesis densities, through a new feature in the hierarchical network.

VII. RESULTS

This section illustrates hypothesis densities learned from imitation and reflexes, and empirical densities are learned by
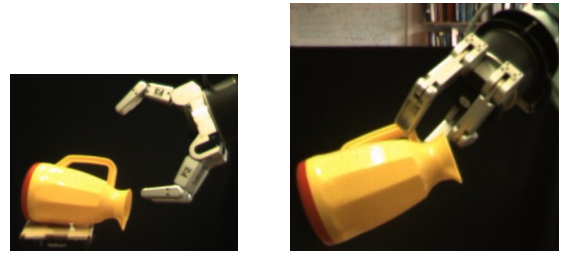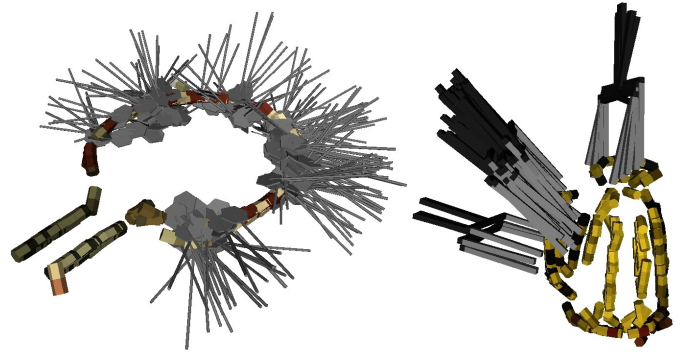
grasping objects with a 3-finger Barrett hand. Densities are built for two objects: the table-tennis paddle of Fig. 3, and a toy plastic jug (Fig. 6). The experimental scenario is described below.

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 IV-B. 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. 5.

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. 6, 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. 4a. 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 VI-B. Samples from the resulting empirical densities are shown in Fig. 7. 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. 7 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.

VIII. Conclusion And Future Work

We presented a framework for representing and learning object grasp affordances, and linking these to a visual object model. The affordance representation is probabilistic and 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 features independently, and connect them through a single top-level model feature. 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 will be directly and exclusively connected to the visual evidence that predicts its applicability, allowing for its generalization across objects.

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 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.

[7] Renaud Detry, Nicolas Pugeault, and Justus H. Piater. Probabilistic pose recovery using learned hierarchical object models. In *International Cognitive Vision Workshop (Workshop at the 6th International Conference on Vision Systems)*, 2008.

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

[9] 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.

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

[11] Kai Huebner, Steffen Ruthotto, and Danica Kragic. Minimum volume bounding box decomposition for shape approximation in robot grasping. Technical report, KTH, 2007.

[12] 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).

[13] 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.

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

[15] 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.

[16] K. V. Mardia and P. E. Jupp. *Directional Statistics*. Wiley Series in Probability and Statistics. Wiley, 1999.

[17] 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.

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

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

[20] Mario Richtsfeld and Markus Vincze. Robotic grasping based on laser range and stereo data. In *International Conference on Robotics and Automation*, 2009.

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

[22] 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.