Towards Affordance Prediction with Vision via Task Oriented Grasp Quality Metrics

Luca Cavalli
Politecnico di Milano
Milan, Italy
luca3.cavalli@mail.polimi.it

Gianpaolo Di Pietro
Politecnico di Milano
Milan, Italy
gianpaolo.dipietro@mail.polimi.it

Matteo Matteucci
Politecnico di Milano
Milan, Italy
matteo.matteucci@polimi.it

***Abstract*—While many quality metrics exist to evaluate the quality of a grasp by itself, no clear quantification of the quality of a grasp relatively to the task the grasp is used for has been defined yet. In this paper we propose a framework to extend the concept of grasp quality metric to task-oriented grasping by defining affordance functions via basic grasp metrics for an open set of task affordances. We evaluate both the effectivity of the proposed task oriented metrics and their practical applicability by learning to infer them from vision. Indeed, we assess the validity of our novel framework both in the context of perfect information, i.e., known object model, and in the partial information context, i.e., inferring task oriented metrics from vision, underlining advantages and limitations of both situations. In the former, physical metrics of grasp hypotheses on an object are defined and computed in known object model simulation, in the latter deep models are trained to infer such properties from partial information in the form of synthesized range images.**

***Index Terms*—task-oriented grasping, robotic grasping, affordance, vision**

I. Introduction

The research community has spent much effort in tackling the problem of grasping novel objects in different settings [1] [2] [3] [4] [5] with the objective of holding objects robustly with robotic manipulators; however, real manipulation tasks go far beyond holding the objects and the quality of a grasp depends on the task it is meant to support. While many quality metrics exist to evaluate the quality of a grasp by itself [6] [7], no clear quantification of the quality of a grasp relatively to a task has been defined. In this paper we propose a framework to extend the concept of quality metric to task-oriented grasping by defining general physical measures for an open set of task affordances. We evaluate both the results provided by such metrics and their applicability in practice by learning to infer them from vision.

More formally, given a grasp G on an object O and a point U on the surface of O (which is the point where we plan to use the object, when the task requires one), we define the affordance function $F_T : (O, G, U) \mapsto \mathbb{R}$ to define the affordance of any possible grasp G and use hypothesis U with respect to task T. The final objective is to optimize for the best grasp, object and use location (O, G, U) that maximizes a given affordance function, as shown in Figure 1. We are interested in finding a set of metrics, as functions of the triplet (O, G, U), to encode significant static and geometric properties of the (O, G, U) system itself. Examples of such metrics are the local geometry of O around U, or the minimum sum of contact forces needed to hold the object O with grasp G under a given gravity vector. A complete description of the metrics used in this work, not to be considered as an exhaustive list, is reported in Section IV-A.

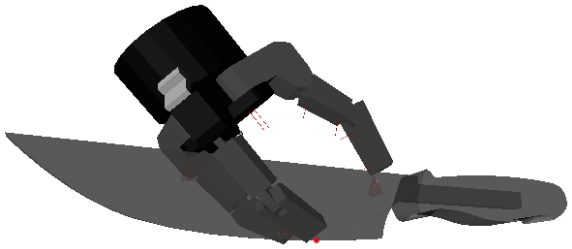


Fig. 1. Best grasp, according to the proposed metrics, for cutting using a common kitchen knife

The proposed approach allows the inference of the affordance of objects without having it bound to their semantic category: semantic information on objects defines their standard use meant for humans, which is not necessarily the only nor even optimal use for robots. Semantics greatly simplifies the task of affordance perception, but it gives no guarantee of optimality, particularly with robot actuators which differ substantially from human hands and arms. Take, for instance, the classical human grasp of a hammer with the wrist direction parallel to the beating direction: actuating such grasp with a Barrett hand [8] which has only a rotational degree of freedom on the wrist would have the same efficacy in beating as a human with a locked wrist, even without taking into consideration the decreased number of fingers and much reduced tangential and torsional friction in contacts.

To validate our framework, we have collected a dataset of grasps and computed elementary graps metrics by using the GraspIt! simulator [9] on the Princeton Shape Benchmark object models [10]. Then we have trained deep models to infer those elementary metrics from range images taken in a simulated camera-in-hand setting to assess the applicability of our framework to more realistic partial-information settings.

In our contribution, we define a framework for quality assessment of task-oriented grasps, we qualitatively validate such framework, we provide a GraspIt! plugin to produce labelled data with minimal to no human intervention, thus

in an extremely scalable way, and we generated a dataset of 400M evaluated grasps on 22 objects of the Princeton Shape Benchmark which we plan to make public in the near future. Moreover, we propose and benchmark models to tackle the problem of learning to infer such metrics from vision. Preliminary results shows direct optimization of affordance functions in simulation produces new and creative grasps which fit the specific actuator in use for the selected task, while direct inference of such metrics from vision is yet an open challenge and there are great margins for further improvement.

II. Related works

Many researchers have worked towards the understanding and formalization of the concept of affordances [11] [12] [13]; they have been inspiring for roboticists to work within the affordances framework to define the autonomous interaction of a robot with an unknown environment. In our work we investigate the broad category of robot affordances focusing on the specific application of task-oriented grasping. Within this context, one of the first approaches towards task-oriented grasping, reported in [14], proposed to encode the task in physical terms (e.g., applying a momentum on a handle to open a door) and then to solve the problem of grasp planning by hardcoding hand postures and their association with tasks; the method has shown good performance in the expected domain, but poor generalization capabilities.

Later works formalized the problem via graphical models, distinguishing task, object features, action features and constraint features. In particular, authors of [15] proposed the use of such formalization and they have been able to effectively learn to infer the likelihood of grasp approach directions with respect to a human-labelled ground truth. The main limitation of this work, in our opinion, is the human intervention, which makes the real definition of the tasks implicit and prevents the scalability of the dataset that can be generated for learning without tedious human teaching.

The direct intervention of human judgment on semantics to evaluate the quality of grasp hypotheses with respect to a given task is nevertheless a common approach to many research works, like [16] and [17] in which authors prove the effectiveness of a human-labelled semantic approach with real robot manipulations. More recently, [18] has proposed to label mesh vertices in simulated objects as being graspable or not according to some task, so that many scene examples can be produced and automatically labeled via simulation. This allowed the system to automatically segment graspable and not graspable regions of objects in cluttered scenes, but still the expressivity of this method is restricted to specifying graspable or not graspable surfaces.

A common limitation of these approaches is the vague definition of task affordances which passes through the human labeling of the ground truth. This entails a great limitation in the size of the data that can be produced for learning and poses questions about the optimality and validity of the labels with respect to the actual task performance with a different actuator than the human hand.

III. Proposed approach

Let $\mathcal{O}$ be the set of possible object surfaces with friction and softness properties defined at each point, let $\mathcal{G}(O)$ defined on an object $O \in \mathcal{O}$ be the set of possible grasps determined by the hand embodiment, degrees of freedom, contact locations on the object and contact nature (e.g., frictionless, hard contact or soft contact), and $\mathcal{U}(O)$ be the set of points on the surface of the same object that can be considered as points of use.

Let $O \in \mathcal{O}$, $G \in \mathcal{G}(O)$, $U \in \mathcal{U}(O)$, then we define the *affordance function* of task T as $F_T(O, G, U) \mapsto \mathbb{R}$ such that $F_T(O_1, G_1, U_1) > F_T(O_2, G_2, U_2)$ if and only if the grasp and use hypothesis (O_1, G_1, U_1) is more suitable than the hypothesis (O_2, G_2, U_2) for task T, thus defining an affordance ordering of an object grasp for task T.

As we want a compact representation of the affordance function, we approximate F_T as a $\tilde{F}_T : \mathbb{R}^n \mapsto \mathbb{R}$ by mapping the triplet (O, G, U) into a metric vector $\phi \in \mathbb{R}^n$ through a function $\Phi(O, G, U) \mapsto \mathbb{R}^n$. This metric vector is a collection of metrics encoding the geometrical and static physical properties of the triplet (O, G, U) which are relevant to approximate F_T. In Section IV we provide some examples of basic metrics ϕ and examples on how they could be used to hardcode $\tilde{F}_T$ for some reference tasks.

A. Achieving Object Semantics Independence

The complete object geometry is generally not available in real world applications, in particular when our long term goal is to infer object affordance from vision with no hardwired semantics. To achieve such goal, we need to frame the problem in the context of uncertain and incomplete information about the object by decoupling the grasp and use location description from the exact object geometry and possibly its semantics.

Recall here that the complete description of a grasp requires the geometry and nature of contact points on the grasped object, and the grasp itself needs to be actuated by a grasping policy. If we assume the grasping policy to be deterministic, then we can define it as a function $GP(p_0, O) \mapsto \mathcal{G}(O)$ that maps an initial state $p_0 \in \mathcal{P}_0$ and an object O into the final grasp $G \in \mathcal{G}(O)$. To decouple from the specific grasp, and its parameters, we fix a grasping policy that allows a sufficient exploration of the grasps space $\mathcal{G}$ via the space $\mathcal{P}_0$ of possible initial states, which we call *pregrasps*.

In particular, we select a simple, but effective, grasping policy defined as follows: from an initial position of the hand with open fingers, we advance towards a fixed direction until the first contact is made, then the fingers are closed until all of them either make contact or are completely closed. Jointly with the use of eigengrasps [19] to describe the degrees of freedom of the hand, this policy allows for a further reduction in the dimensionality of the problem for the translational degree of freedom saved in describing the position of the hand in the space, as approaching the object in a straight line makes the approach direction invariant with respect to the policy. If the eigengrasp parameters are divided into h parameters used to close the grasp and k parameters used to set the initial hand posture, then we can consider $p_0 \in \mathbb{R}^{5+k}$. The

(a) (b) (c)

Fig. 2. The three phases of our grasp policy: (a) the pregrasp parameters determine the initial position and posture of the hand (b) the hand approaches in a straight line until a contact is made (c) fingers close until they make contact or they are completely closed

h eigengrasp parameters used to close the hand are fixed to an open position on the pregrasp, thus they do not contribute to the dimensionality of the pregrasp itself.

On the other hand, use locations are points on the bidimensional surface of each specific object; we decouple them from the specific object by defining use directions and a direction mapping function $DM(d, O) \mapsto \mathcal{U}(O)$ in complete analogy to the grasp decoupling solution. Our direction mapping assumes $d \in \mathbb{R}^2$ to be the spherical coordinates of a directed ray centered in the center of mass of object O and output in the farthest point $U \in \mathcal{U}(O)$ which is the intersection of such ray with the outer object surface.

B. Metrics Inference from Vision

As our goal is to make robots able to use unknown objects in a task consistent way, we need the robot to be able to perceive their affordances via sensors. In particular, we focus on vision being it an extremely common and effective tool to take information from the environment in real applications. We define our inference setting by focusing on the specific case of single range images taken from camera-in-hand perspective: such case provides only local geometry information about the object around the expected location of the planned grasp. We want to learn a model that predicts the elementary metrics $\phi \in \mathbb{R}^n$ of a triplet (O, G, U) with only partial information about the observed object O. Indeed, predicting the vector ϕ would allow to estimate the affordance function F_T for any task T for which we can define an $\tilde{F}_T$.

Let $D : \mathcal{P}_0 \times \mathcal{O} \mapsto \mathbb{R}^{h \times w}$ be the function that maps a pregrasp p_0 on an object O to the depthmap of size $h \times w$ from the camera-in-hand perspective of p_0. Then we want to learn a model $\mathcal{M}^\Phi$ that approximates the mapping from $(p_0, D(p_0, O), d)$ to $\Phi(O, GP(p_0, O), DM(d, O))$ where O is an object, p_0 is a pregrasp, d is a use direction, and Φ is the metric extraction function, under the grasping policy GP and the direction mapping function DM.

We assume to be able to learn model $\mathcal{M}^\Phi$ from a dataset of tuples $(O, p_0, d, \Phi(O, GP(p_0, O), DM(d, O))$ obtained via uniform sampling of p_0 and d values on a set of available objects models and computing the true values of $\Phi(O, GP(p_0, O), DM(d, O))$ via simulation. Details on our data collection setup are explained in Section V-A.

To structure the learning task, we define the model $\mathcal{M}^\Phi$ as the composition of two models: an input value $(p_0, D(p_0, O), d)$ is first classified by a binary classifier $\mathcal{M}^\Phi_C$ to infer whether it represents a "good" grasp worth further evaluation or not. We define "good" grasps those respecting a minimum quality independently from the task, thus employing state of the art grasp quality metrics to generate the ground truth. The samples classified as positive then pass through a regression model $\mathcal{M}^\Phi_R$ that infers the metrics ϕ with the implicit assumption that the grasp is indeed a quality grasp.

For both models $\mathcal{M}^\Phi_C$ and $\mathcal{M}^\Phi_R$ we propose and evaluate the two architectures of the convolutional neural network (CNN) and the PointNet architecture [20]. Both architectures encode the available geometry information (in the form of range image for the CNN or as the equivalent projected point cloud for the PointNet) in a feature vector, we then apply late-fusion of the other input parameters p_0 and d on this feature vector and output the classification label or the inferred regression value with a classical fully connected network.

IV. Task Oriented Grasp Metrics

In this work we concentrate on the sample tasks of beating, cutting and picking, which are defined by their $\tilde{F}_{beat}$, $\tilde{F}_{cut}$ and $\tilde{F}_{pick}$ in Section IV-B. We first define a set of grasp metrics, then define from these the corresponding affordances.

A. Basic Grasp Metrics

We consider the following set of elementary metrics of (O, G, U) which should not be considered as exhaustive:

a) Grasp robustness ($\epsilon \in \mathbb{R}$): is a real number describing the robustness of the grasp. We use the Epsilon metric described in [7] as a builtin in the GraspIt! simulator [9]. Force closure grasps have $\epsilon > 0$ and higher robustness implies a greater minimum perturbance is needed to break the grasp.

b) Rotational inertia ($I \in \mathbb{R}$): quantifies the rotational inertia around the axis of rotation of the wrist of the hand assuming a unitary density of the object and assuming the hand to be integral with the whole object. It does not take into account the mass of the hand itself.

c) Hand effort on impact ($E_i \in \mathbb{R}$): describes the effort of the hand to balance the impact forces after a rotation around the wrist. It assumes a fixed average inertial torque in a small Δt during the impact which is directly proportional to I and a free contact force on the use location towards the normal direction. This metric takes the value of the minimum sum of the contact forces of the hand constrained to the contact friction cones to balance the inertial torque, ∞ if the minimization problem is unfeasible.

d) Hand effort on hold ($E_h \in \mathbb{R}^6$): is a vector of six independent values which quantify the hand effort to balance a different gravity vector. The hand effort is the minimum sum of all contact forces constrained to the contact friction cones that balance a given unitary force of gravity, ∞ if such problem is unfeasible. The six gravity vectors chosen are aligned with the three coordinate axes (once in the same direction, once opposite) of the object mesh as all meshes that we used in the Princeton Shape Benchmark have been designed by humans that gave a semantic meaning to the coordinate axes directions.

e) Momentum discharge efficiency ($\delta \in \mathbb{R}$): quantifies the efficiency of discharging the rotational inertia of the wrist on the object use location. It quantifies the alignment between the inertial torque and the torque generated by a force aligned with the use location normal vector towards the inside of the object surface. It is computed as the dot product of the two normalized vectors, clipped to zero in case of negative values.

f) Force transmitted to use ($U_\tau \in \mathbb{R}$): quantifies the force that can be transmitted to the use location using constrained contact forces. It assumes all contact forces are constrained by their friction cones and have unitary maximum normal forces. It takes the value of the maximum force on the use location towards the use location normal guaranteeing static conditions.

g) Use local geometry ($U_g \in \mathbb{R}$): describes how much the use location has the shape of an edge. It is obtained by fitting a quadratic function on the vertices of the triangles near the use location (including all the triangles that share at least one vertex with the triangle where the use location lies) and extracting the eigenvalues of the hessian matrix of such quadratic function. The two eigenvalues λ_1 and λ_2 are the two principal component curvatures, so we quantify an edge with the expression $(\lambda_1 - \lambda_2)^2$ to identify locations with a great difference in local curvatures.

B. Affordance Functions from Basic Metrics

On top of these metrics we define the affordance functions for $T \in \{beat, cut, pick\}$. In this preliminary study affordance functions have been designed by hand, to validate the feasibility of the framework, in future works we aim to learn them by optimizing task execution efficacy.

a) Beating: The classical beating action of a hammer with a human hand requires dexterous movements of the wrist which would need moving the whole robotic arm to be reproduced on a Barrett hand. For this reason we assume that the beating action will be executed by the robotic actuator by simply rotating the hand clockwise around the wrist. We require that the hold is stable over a minimum threshold and that the rotational energy gets discharged almost entirely on the point of use. We want to maximize the ratio of the energy that we can incorporate into the rotation (assuming a maximum rotational speed) over the actual hand effort of keeping the object stable on the impact.

Input: $\epsilon, \delta, I, E_i, E_h$
Output: $\tilde{F}_{beat}$
1: **if** $\epsilon < \tau_\epsilon$ OR $\delta < \tau_\delta$ OR $\sum_{i=1}^{6} E_h[i] == \infty$ **then**
2: **return** $-\infty$
3: **else**
4: **return** $\frac{I}{E_i}$
5: **end if**

b) Cutting: The action of cutting is extremely complex by itself and varies greatly with different materials and their surface and micro-structural properties. A complete physical study of this particular task is not our objective; we simplify it considering as approximation that greater force provides cuts if executed on a thin enough edge.

Input: ϵ, U_τ, U_g
Output: $\tilde{F}_{cut}$
1: **if** $\epsilon < \tau_\epsilon$ OR $U_g < \tau_{U_g}$ **then**
2: **return** $-\infty$
3: **else**
4: **return** U_τ
5: **end if**

c) Picking: Picking an object (as the first part of the pick-and-place task) only strictly requires a stable grasp for a successful pick. However, different stable grasps may imply very different effort from the hand actuator to balance the force of gravity on the object. For this reason we require a stable grasp and minimize the sum of the contact forces required to balance the force of gravity in the six directions evaluated by the E_h metric. Notice that an unstable grasp will need to have at least one evaluated direction of gravity that the grasp cannot hold, thus we do not check the ϵ metric.

Input: E_h
Output: $\tilde{F}_{pick}$
1: **return** $-\sum_{i=1}^{6} E_h[i]$

V. Framework Validation

To assess the feasibility of the proposed approach we performed a set of experiments focused on the tasks of beating, cutting and picking; this choice has driven the selection of basics metrics to encode in function Φ and the definition of functions $\tilde{F}_{beat}$, $\tilde{F}_{cut}$ and $\tilde{F}_{pick}$ in the previous section. In the validation we aim at:

1) **Validating the framework** by showing that $\text{argmax}_{G,U}\, \tilde{F}_T(\Phi(O, G, U))$ for some selected tasks T provides grasps and use locations that are semantically meaningful with respect to the semantics of task $T \in \{beat, pick, cut\}$
2) **Assessing the feasibility of learning** a model $\mathcal{M}^\Phi$ that can infer basic grasp metrics from partial information about a target object.

(a) (b) (c)

Fig. 3. Pregrasp degree of freedom (a) set to 0, (b) set to 0.25, (c) set to 1

TABLE I
AFFORDANCE FUNCTION THRESHOLDS

Mode name	τ_ϵ	τ_{U_g}	τ_δ
Default	0.3	10	0.95
No robustness required	$-\infty$	10	0.95
Extra robustness required	0.5	10	0.95

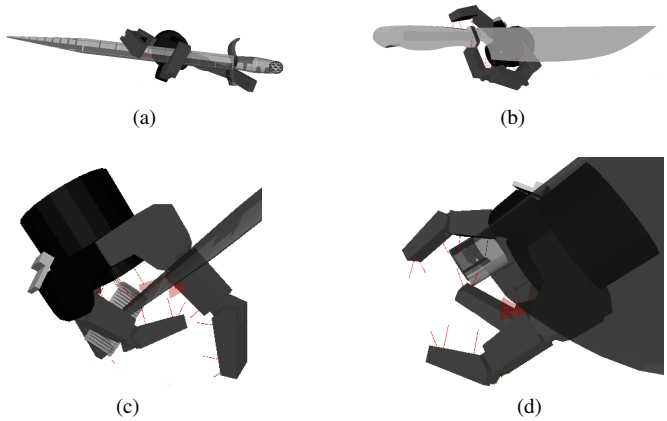
We provide an implementation of the metric extraction function Φ for the selected metrics (which we describe in Section IV-A) as a plugin for the GraspIt! [9] simulator, some of which are formulated as linear programming problems which we solve through the CGAL library [21]. Input object models are selected from the Princeton Shape Benchmark [10] dataset and assumed to be constituted of homogeneous plastic, and grasps are produced using the model of a Barrett hand [8]. All simulated grasps are evaluated and results are logged into a dataset, preserving both stable and unstable grasps. We structure the model $\mathcal{M}^\Phi$ as a classifier that filters stable grasps only and a regressor that estimates the metric vector ϕ of stable grasps from the available partial information.

A. Data collection

We collected a dataset of grasp and use hypotheses to compute metrics for learning purposes. We sample pregrasps and use directions with uniform distribution in their domain and then simulate the grasp and determine the exact use location on the mesh of an object from the Princeton Shape Benchmark [10]. The grasp is simulated with GraspIt! [9] on a Barrett hand [8] with the policy in Figure 2: from an initial hand position and orientation, the manipulator advances in a straight line until a first contact is made with the object, then the three fingers of the Barrett hand are closed independently until a contact is made or the finger is completely closed.

B. Optimization of Task Grasp Metrics via Simulation

We consider only one pregrasp degree of freedom for the Barrett hand to encode the angle between the two joint fingers as shown in Figure 3, thus the domain in which we uniformly randomize the pregrasps is $[-1,1]^5 \times [0,1]$: two values in $[-1,1]$ encode the hand approach direction in normalized spherical coordinates, one value in $[-1,1]$ encodes the hand rotation around its approach axis, two values in $[-1,1]$ are the approach offset on the xy-plane relatively to the bounding box of the considered object and one value in $[0,1]$ is the pregrasp degree of freedom of the Barrett hand; use directions are encoded in normalized spherical coordinates in $[-1,1]^2$. Data collection can be run in parallel on multiple cores and machines, producing millions of data samples each day. Running on 20 cores of an Intel Xeon E5-2630 v4 for a week we could produce a dataset of over 400 million samples of random grasps with metrics, out of which 20 million samples are viable grasps which respect the condition in Section V-C.



(a) (b) (c) (d)

Fig. 4. Optimized grasps from the picking showing the joint pinch strategy. The edge of the blade is pinched between two joints to improve the stability of the grasp; (c) and (d) show the detail of the joint pinch on the blade.

To assess the validity of our framework we extract $\mathrm{argmax}_{G,U}\, \tilde{F}_T(\Phi(O,G,U))$ for our three selected tasks by brute force search on our dataset samples. The parametric thresholds used are the default values reported in Table I. Sample results of this procedure are shown in Figures 4, 5, 6, and 7. Some of the produced grasps do not appear intuitive, such as the grasps produced for moving blades in Figure 4, because they exploit features of their physical actuator which are very different from a human hand. In this particular case we observe that the hand achieves a stable grasp by pinching the edge of the blade between the joint of some finger. This pinch provides multiple contacts with very different normals that provide much greater stability of the grasp on a low friction material. The emergence of such solution is very unlikely to happen from human evaluation of grasps, as the human intuition is heavily biased by the human hand with much more fingers, much higher tangential and torsional friction and more susceptible to damage than the metal Barrett hand assumed in our experiments.

Adaptive behaviours are evident in Figure 5: with thin blades the hand exercises pressure on the edge by rotation, using the handle as a fulcrum, while with larger blades where such technique is not feasible a direct pressure is preferred. Picking grasps (Figure 6 and 4) generally wrap around the center of mass as a direct result of the minimization of the total contact forces for holding against gravity. Beating grasps (Figure 7) display greater variance and generally achieve a stable grasp on the object far from the center of mass to increase the rotational inertia of the object and choose a use location very well aligned with the rotation direction to effectively discharge the rotational energy on the target (the values of δ for the optimal grasps are far nearer to 1 than

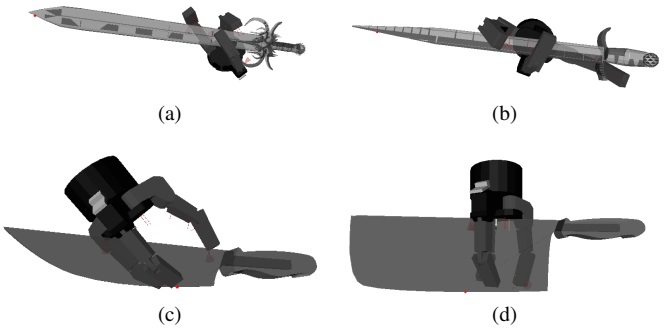


Fig. 5. Optimized grasps from the cutting task showing two different cutting strategies: thin blade tools like (a) and (b) exercise pressure by rotation, wide blade tools like (c) and (d) instead prefer a direct pressure strategy.

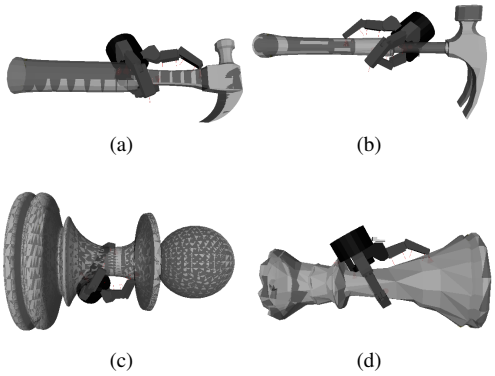


Fig. 6. Optimized grasps from the picking task on sample objects.

the required values for the selected threshold τ_δ). The use location is selected slightly off the center of mass from the opposite side of the hand to balance the beating impulse and produce a torque that contrasts the rotational inertia of the beating movement.

As a further validation test we changed the ϵ threshold requirement for the tasks of beating and cutting according to Table I. The results of such experiment conducted on the model of a typical kitchen knife are shown in Figure 8. The produced grasps for beating with a knife display a similar strategy with respect to the ones for more classical objects as seen in Figure 7: the grasp is as far as possible from the center of mass, switching to the handle only when greater robustness is required. The cutting task on the knife shows less variance, producing robust grasps even when not explicitly required by the fitness function.

C. Learning Grasp Metrics from Vision

For the learning phase we defined exactly what a viable grasp should be to generate the ground truth for the classifier network and define the actual dataset for the regressor. We define a viable grasp sample if:

$$\epsilon > \tau_\epsilon \quad \wedge \quad \sum_{i=1}^{6} E_h[i] < \tau_{E_h} \quad \wedge \quad E_i < \tau_{E_i}$$

where empirically we set $\tau_\epsilon = 0.15$, $\tau_{E_h} = 250$, $\tau_{E_i} = 100$.

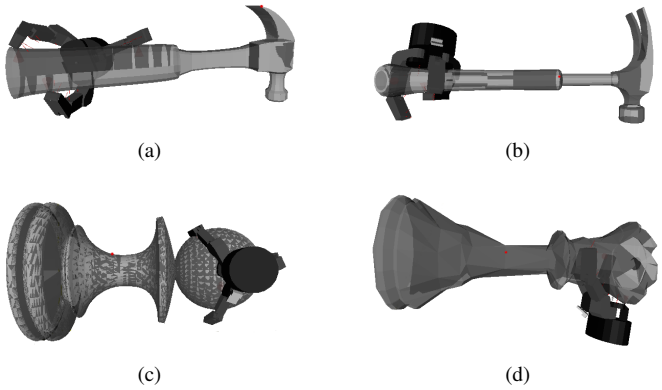
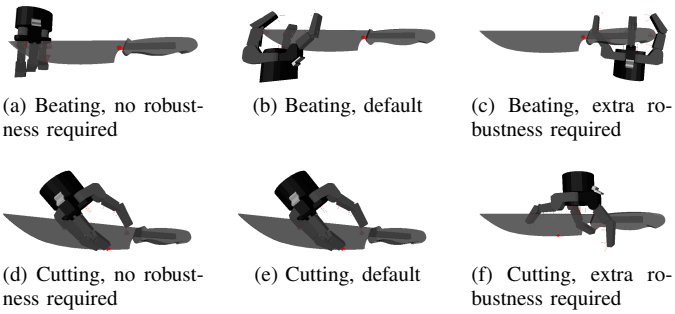


Fig. 7. Optimized grasps from the beating task on sample objects. Notice that the center of mass of hammers in (a) and (b) is on the handle, as the material is assumed homogeneous.



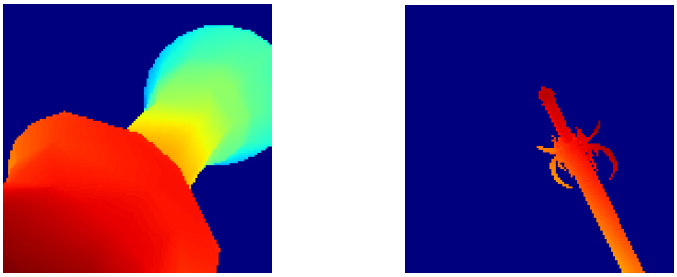
(a) Beating, no robustness required (b) Beating, default (c) Beating, extra robustness required

(d) Cutting, no robustness required (e) Cutting, default (f) Cutting, extra robustness required

Fig. 8. Affordance function parameter variability. The grasp strategy varies slightly while changing the requirement of the robustness of grasp.

Range images are generated using OpenGL with a perspective projection using common parameters. We take the Kinekt 2 depth camera as a reference with a 70x60 field of view angle: we generate subsampled images of resolution 128x128 using a 60 field of view taken from an object-length distance from the center of the object, as shown in Figure 9. We formally consider point clouds equivalent to range images as they are generated to hold the same exact information as the input image. We use Open3D [22] to project the synthesized range image to a point cloud which is cleaned from the background. The resulting point cloud is either randomly subsampled of filled with extra points in the origin to match a standard number of points to build batches for efficient learning. Referring to Princeton Shape Benchmark models, we train on models 1110, 1114, 489, 493, 495, 710, 720, 725, 750, 758 including many different beating tools, cutting tools, screwdrivers, bottles and other non-tool objects, we use as validation set the data synthesized from models 1111, 490, 718, 722, 724 and as test models 1116, 482, 730; validation and test comprise different objects from train but in the same semantic categories.

a) The Classifier: filters viable grasps from input data based on the pregrasp and on the input range image with camera-in-hand perspective. As the overall data distribution



(a) (b)

Fig. 9. Synthesized range images with camera-in-hand perspective.

TABLE II
RESULTS ON REGRESSORS

Model	MSE	Comparison accuracy	Expected GMS
PointNet Full	0.050	0.63	0.63
PointNet Slim	0.049	0.60	0.73
CNN	0.049	0.66	0.82

from our random policy is highly biased towards the negative class (20 times more likely than the positive class), we sample the training data to balance positive and negative samples.

b) The Regressor: infers the metric vector ϕ for later computation of the affordance function relative to the input sample. In this work we trained regressor networks to infer the value of $\sum_{i=1}^{6} E_h[i]$ for testing with the picking affordance function on the picking task. Output values are linearly normalized in the interval $[0, 1]$ from the original domain $[0, \tau_{E_h}]$ granted by the assumption that input values come from the positive class of viable grasps.

The precision-recall curve of the test set for the two trained classifiers are reported in Figure 10. As this classifier model is intended as a filter of good grasp hypotheses, our main metric of interest is the precision of the positive class as this represents the probability that a grasp that passes the filter does really satisfy the expected stability conditions. The recall of the positive class is relevant as well, as it describes the efficiency of the system in missing less good grasp.

The results on the test set of the trained regressors are in Table II. The mean squared error, the classical metric used to assess basic regression, gives an overall score of how near the regression goes to real values, but as our goal is optimization, not estimation, we elaborated on two different metrics. As optimization is mainly built by comparisons, we elaborate the comparison accuracy by sampling random couples of samples and measuring the standard accuracy in telling which input sample corresponds to a greater value of the metric. We define the Global Min Score (GMS): let I_{min} be the input sample that minimizes the predicted output $\mathcal{M}_R^{\Phi}(I_{min})$ over a set S of samples, then the Global Min Score of the model for the set S is the rate of samples that actually have a greater ground truth value than the ground truth value of I_{min}. As this score is very sensible to different choices of S, we sample random subsets of 10% of the available samples in S to plot the probability distribution of the value of GMS.

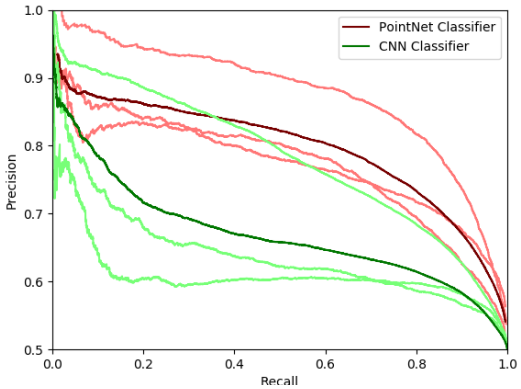


Fig. 10. Recall-precision curve for CNN and PointNet classifiers. Dark lines are the curve on all the objects of the test set for the respective model, light lines are computed on single objects of the same set

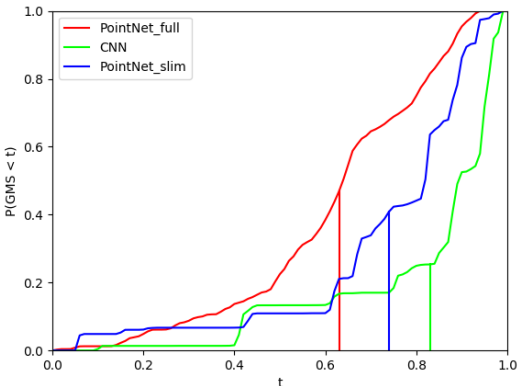


Fig. 11. Global minimum score cumulative distributions for the three regressors. The PointNet full regressor is the standard PointNet, the PointNet slim is the standard PointNet with transformation layers fixed to the identity. The vertical lines of each distribution is the expected value for the GMS.

Figure 11 shows the cumulative GMS distributions and their respective expected value. This is the most specialized measure of performance of our regressor models as it directly quantifies the relative optimality of the selection of the model relatively to the available choices.

VI. Discussion on Limitations and Future Work

In this work we proposed task dependent metrics to evaluate grasps and use hypotheses in a task-oriented setting. Our framework allows for the automatic evaluation of grasps in a simulation environment and to collect labelled grasp data with minimal human interaction; eliminating the need for human intervention in the grasp labeling process allows both for a widely more scalable data collection and clears the biases that humans have in labeling grasps due to the significant differences between human and robotic hands.

We showed that we can easily generate millions of labelled grasps on different objects and that roughly hand-designed affordance functions suffice for the emergence of smart and unintuitive techniques for grasping for the exemplified tasks as in Figures 4 and 5. From this we experimented with

convolutional and PointNet [20] architectures to learn to infer basic grasp stability and holding hand effort from vision and benchmarked their results, which we consider promising.

There are many directions in which this work can be improved and extended to account for its limitations. The most obvious direction is by accounting for more metrics in simulation such as different materials in compound objects, with different friction coefficients and softness values. This would also provide more realistic locations for the center of mass, which is crucial to determine the optimal grasp in many situations such as for the beating with hammers. As a natural follow up, we foresee a validation step on a real robotic arm to prove the generalization capability from simulation to reality. Indeed, we plan to integrate the learned models with a real Barrett hand and test on similar manipulators (with three or even two fingers) to assess the robustness of performance with inaccurate manipulator models.

The current hypothesis of camera-in-hand substantially limits the exploration capability of the algorithm that needs to physically move the arm to evaluate different grasp approach directions. To overcome this limitation we plan to extend the current system to consider a voxel grid describing the current geometrical knowledge of the system about the object. Such voxel grid would represent the probability of emptiness of every voxel and can be built from subsequent range images with known techniques [23]. This would not only allow to integrate knowledge from multiple range images (not necessarily from the hand) but also to rotate the object representation to face any hypothesized approach direction to search for more appropriate directions with the current integrated knowledge.

An important limitation of the current work is the unstructured definition of the affordance functions $\tilde{F}_T$ which at this stage *define* the tasks themselves for the system. A direction of improvement is towards a different description of tasks from which functions $\tilde{F}_T$ can be extracted or learned. Reference [24] extracts a representation of a task in terms of relevant physical quantities from the demonstration of a human choosing a tool and executing the task, such representation can be used to drive the automatic synthesis of the affordance function from a human demonstration. An alternative approach can be the definition of the execution of the task and a performance index of its end effectiveness: this would allow the optimization of the affordance function by reinforcement learning by simulation of the execution of the task itself with no further human intervention.

REFERENCES

[1] H. Dang and P. K. Allen, "Tactile experience-based robotic grasping," in *Workshop on Advances in Tactile Sensing and Touch based Human-Robot Interaction, HRI*, 2012.

[2] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *The International Journal of Robotics Research*, vol. 37, no. 4-5, pp. 421–436, 2018.

[3] D.-J. Kim, R. Lovelett, and A. Behal, "Eye-in-hand stereo visual servoing of an assistive robot arm in unstructured environments," in *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*. IEEE, 2009, pp. 2326–2331.

[4] A. N. Erkan, O. Kroemer, R. Detry, Y. Altun, J. Piater, and J. Peters, "Learning probabilistic discriminative models of grasp affordances under limited supervision," in *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*. IEEE, 2010, pp. 1586–1591.

[5] A. Zeng, S. Song, K.-T. Yu, E. Donlon, F. R. Hogan, M. Bauza, D. Ma, O. Taylor, M. Liu, E. Romo *et al.*, "Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1–8.

[6] M. A. Roa and R. Suárez, "Grasp quality measures: review and performance," *Autonomous robots*, vol. 38, no. 1, pp. 65–88, 2015.

[7] A. T. Miller and P. K. Allen, "Examples of 3d grasp quality computations," in *Proceedings 1999 IEEE International Conference on Robotics and Automation (Cat. No. 99CH36288C)*, vol. 2. IEEE, 1999, pp. 1240–1246.

[8] W. Townsend, "Mcb-industrial robot feature article-barrett hand grasper," *Industrial Robot: An International Journal*, vol. 27, no. 3, pp. 181–188, 2000.

[9] A. T. Miller and P. K. Allen, "Graspit! a versatile simulator for robotic grasping," *IEEE Robotics Automation Magazine*, vol. 11, no. 4, pp. 110–122, Dec 2004.

[10] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser, "The princeton shape benchmark," in *Proceedings Shape Modeling Applications, 2004.* IEEE, 2004, pp. 167–178.

[11] J. Gibson, *The senses considered as perceptual systems*. Boston: Houghton Mifflin, 1966.

[12] C. Michaels, "Affordances: Four points of debate," *ECOLOGICAL PSYCHOLOGY*, vol. 15, pp. 135–148, 04 2003.

[13] E. Şahin, M. Çakmak, M. R. Doğar, E. Uğur, and G. Üçoluk, "To afford or not to afford: A new formalization of affordances toward affordance-based robot control," *Adaptive Behavior*, vol. 15, no. 4, pp. 447–472, 2007.

[14] M. Prats, P. J. Sanz, and A. P. Del Pobil, "Task-oriented grasping using hand preshapes and task frames," in *Robotics and Automation, 2007 IEEE International Conference on*. IEEE, 2007, pp. 1794–1799.

[15] D. Song, K. Huebner, V. Kyrki, and D. Kragic, "Learning task constraints for robot grasping using graphical models," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2010, pp. 1579–1585.

[16] H. Dang and P. K. Allen, "Semantic grasping: Planning robotic grasps functionally suitable for an object manipulation task," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct 2012, pp. 1311–1317.

[17] H. Dang and P. K. Allen, "Semantic grasping: planning task-specific stable robotic grasps," *Autonomous Robots*, vol. 37, no. 3, pp. 301–316, 2014.

[18] R. Detry, J. Papon, and L. Matthies, "Task-oriented grasping with semantic and geometric scene understanding," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 3266–3273.

[19] M. Ciocarlie, C. Goldfeder, and P. Allen, "Dexterous grasping via eigengrasps: A low-dimensional approach to a high-complexity problem," in *Robotics: Science and Systems Manipulation Workshop-Sensing and Adapting to the Real World*. Citeseer, 2007.

[20] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 652–660.

[21] C. Cgal, "Computational geometry algorithms library," 2008.

[22] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3D: A modern library for 3D data processing," *arXiv:1801.09847*, 2018.

[23] B. Curless and M. Levoy, "A volumetric method for building complex models from range images," in *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '96. New York, NY, USA: ACM, 1996, pp. 303–312. [Online]. Available: http://doi.acm.org/10.1145/237170.237269

[24] Y. Zhu, Y. Zhao, and S. Chun Zhu, "Understanding tools: Task-oriented object modeling, learning and recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2855–2864.