Towards Grasp Transfer using Shape Deformation

Andrey Kurenkov[*], **Viraj Mehta**[*], **Jingwei Ji, Animesh Garg, Silvio Savarese**
Stanford Vision and Learning Lab

Abstract:
Grasping has recently seen a quanta of progress through the combination of model-based methods with deep learning to achieve 95%+ success. However, generalization is achieved through very large datasets and these methods cannot preserve grasp semantics across object instances. We introduce a method for generating grasps for novel objects through the transfer of grasps from similar objects, which can generalize from a small amount of data, does not impose constraints on the generated grasps, and preserves semantically significant grasps. This approach builds on the data-driven grasp generation of Dex-Net 1.0 and the deformation-based 3D reconstruction of DeformNet, by using a 2D image to find an appropriate 3D template, deforming the 3D template along with its known grasps, and using the deformed grasps on the novel object. We train the model on a set of shapes from ShapeNet annotated with grasps, and evaluate the deformed grasps using analytic grasp quality metrics. Our preliminary results suggest that deformation preserves usable grasps as well as semantically relevant grasp locations.

Keywords: grasping, computer vision, deep learning

1 Introduction

Despite it being among the most fundamental and important tasks in robotics, robust object grasping is still difficult due to imprecision in sensing, actuation, and control. Although it is possible to plan grasps using physics-based analytic methods [1], such approaches usually depend on having an external perception system which makes the pipeline both more complex and brittle. Data-driven methods such as Dexnet 1.0 demonstrated that it is possible to improve grasp robustness by leveraging a large dataset of objects with known grasps [2]. But the method still required a pre-processing pipeline to handle perception. This is adequately addressed in an improved Dexnet 2.0 [3] that performs grasp generation end-to-end from depth image input to grasp evaluation. While impressive, this requires a large dataset and was designed specifically to generate parallel-jaw grasps of an object from a top view.

To generalize this method to novel views and preserve grasp semantics – such as grasping a cup by the handle for manipulation – a full retraining of the models would be required. An alternative approach for robust grasping is to first build up a dataset of objects with known grasps, and then generate new robust grasps for a novel object by querying this dataset for relevant objects and doing grasp transfer to adapt the grasps in the dataset to this novel object. This has the benefit of providing robust grasps without expensive physics-based grasp planning, potentially requiring less labeled data or robot experiments, and of preserving semantically meaningful grasps. Other methods have used labelling of parts in meshes and shape segmentation and alignment for transfer [4]. We show that we can perform this transfer through end-to-end image based shape deformation that also outputs point wise correspondence, hence obviating the need for shape segmentation and alignment.

The primary contribution of this paper is a novel method for grasp transfer between similar objects using 3D shape deformation, which is based on combining a large 3D object dataset with known grasps generated using analytic methods and a deep-learning model that can deform a 3D shape from this dataset guided by a 2D image of a novel object not in the dataset.

To generate a dataset of objects with known grasps, we perform antipodal grasp sampling with robust force-closure evaluation on a set of graspable shapes as in Dex-Net 1.0[2]. After training Deformnet

[*] Equal Contribution

1st Conference on Robot Learning (CoRL 2017), Mountain View, United States.

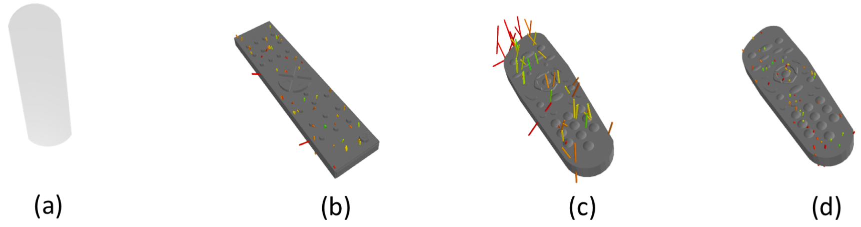

Figure 1: A demonstrative example of our method, showing (a) Input image to DeformNet for the test object, (b) Similar shape template retrieved from the database along with it pre-computed grasps ordered by analytic metric, (c) the deformed grasps overlaid on the known true shape of the test object to be grasped, and (d) the same shape overlaid with the grasps that were generated for it using an analytic method assuming the test object model was known. Although the deformed shape is not shown, it is clear from (c) that the majority of the grasps correctly deform onto the target object and only a subset at the top are unusable.

on this set of shapes exclusively for the task of shape deformation, we can generate grasps for novel objects by using a single RGB image input to first perform shape retrieval from the dataset using a learned image-to-shape embedding, and then deforming the retrieved object as well its known grasps using the FFD layer in an encoder-decoder style network architecture. Generating grasps through deformation based transfer allows generalization across a large variety of objects using only a small set of objects with pre-computed grasps. We provide preliminary results from on small subset of the Shapenet dataset, and show that deformation preserves grasp semantics without significant degradation in quality.

2 Related Work

Grasp Planning Grasp Planning is posed in terms of finding gripper configurations that maximize a success metric for a certain object. Methods for this problem are categorized as either *analytic*, if they model the physics of the grasp to estimate performance [1], or as *empirical*, if they are based on learning statistical models from datasets collected through human annotation, robot experiments, or simulation [5].

Analytic methods typically assume the 3D shape and pose of the object can be known exactly or with known uncertainty, and solve the problem of finding appropriate contact points for a gripper to restrict an object's motion and resist external wrenches[1]. For executing grasps with a robot, such analytic methods are typically used to create a database of 3D objects with known grasps. Point cloud data can then be used to query for relevant objects from the database and for finding the pose of object to grasp, and the highest quality retrieved grasp is executed[2]. Empirical methods typically involve machine learning models that output a grasp quality or directly generate appropriate grasps given an input of 2D or 3D data about a target object. Recently many deep learning approaches have been proposed for both grasp evaluation [3, 6] with separate grasp sampling and for models that directly generate grasps [7, 8]. With the exception of [7], all these methods are based on evaluating parallel-jaw grasps represented as rectangles on a 2D image and are limited in the range of grasps they allow as a result.

Grasp Transfer Aside from training learned models, datasets can be leveraged for grasp transfer by mapping grasps between between similar objects [9, 10], between different grippers [11], and from human-demonstrated grasps to robot grippers [12]. Our method most closely relates to that of Grasp Moduli Spaces [9], which also proposes to transfer grasps through joint shape and grasp deformation, but their deformation method is not based on deep learning and requires a dense point cloud of the full shape to be available for grasping of novel objects. The approach in [10] is also similar to ours in that it uses a similarity metric to retrieve similar grasps from a dataset given 2D images of the target grasps, but is less robust due to only rotating and translating the template grasps and not allowing for full deformation as in our model.

3D Shape Inference Our approach to grasp transfer is unique in that it builds on the recent development of deep learning methods for inferring the 3D shapes of objects. Prior work has explored 3D reconstruction with generative models that output the 3D reconstruction in one of two formats -

occupancy grid (i.e. voxel) [13, 14] and point clouds[15]. Shape retrieval through joint embedding of 2D images and 3D shapes has also been explored as a means for 3D shape inference [16]. Deformnet, the model we use for grasp deformation, finds a balance between the flexibility of a generative approach and the output quality of a database-focused approach. Specifically, the network learns to generate parameters for free-form deformation (FFD) [17]. Because the FFD is not constrained to just deform the shape but also any points on or near the shape, we can easily extend its application beyond shape reconstruction to grasp transfer.

3 Method

3.1 Dataset Generation

Our method relies on having a dataset of objects with known 3D models, so that for each object we can derive its voxelized and point cloud representations, render it from multiple perspectives, and generate usable grasps for it using analytic grasp. We perform voxelization, point sampling, and rendering many views of each object as in the prior DeformNet work. We further enriched by performing grasp generation on the chosen objects as in DexNet 1.0 [2]. This is accomplished by antipodal sampling: we sample a random point on the object surface, form a grasp axis through the point along a randomly sampled direction within the friction cone of the object, calculate the other contact point of the grasp, and keep the grasp if it is in the friction cone. We also store a measure of grasp quality using the Robust Force Closure metric, which has been shown to provide a good measure of the resistance of a given grasp to a wrench while being computationally tractable [18].

3.2 Shape Template Retrieval

To make use of the high-quality 3D CAD models in the existing database, the first step is to retrieve shape templates that have a similar topology to the object in a query image. For this, we use metric learning to learn an embedding that preserves topological similarity between shapes. For the metric, we optimize a loss function over a 2D CNN which forms a smoothed triplet loss [19] over our examples. We define the similarity to be the inverse of the distance in the metric space and retrieve K-nearest neighbors from a query image and randomly pick one in K shapes as the input shape template in the following network for deformation, DEFORMNET. Note that the shape template from database is annotated with grasps. See supplementary material for more details.

3.3 DeformNet

Given a reference image and a shape template that closely matches the input image from the shape template retrieval stage (Sec. 3.2), we want to generate the parameters of a deformation, specifically free-form deformation (FFD) [17], which transforms the shape template into the shape in the reference image. Conventionally, voxel representation is the preferred way to learn 3D reconstruction in learning based reconstruction due to its regularity in space. However, such representation lose details in a 3D shape during quantization which lead to suboptimal reconstruction. Instead of using occupancy grid reconstruction as final output, we benefit from both the voxel and point cloud representation: the network is learned to generate a vector field in grid which is used as the offset for control points in free-form deformation, and this vector field will further determine the deformation of shape template in the form of point cloud. Interpreted as pairs of points, the annotated grasps on the shape template are naturally transformed with shape deformation, generating the output grasps for evaluation. See supplementary material for details including figure for demonstration.

3.4 Loss Function

To make the DEFORMNET end-to-end trainable, we need to define a loss function that optimizes the target task: deforming a shape template to match a target shape. The loss function should measure dissimilarity between the deformed and template shape. To do this, we sample points on the surface of a shape and use the set of points (point cloud) as a surrogate for a shape. Point clouds are easy to manipulate due to their simplicity and efficiency, and we follow [15] by using the Chamfer distance (CD) function on point clouds as the loss function to be minimized. See supplementary material for details on point cloud distance measurement. Besides, we use small regularization in addition to the distance function $\lambda \sum_{i=1}^{N^3} \|v_i\|_1$ to force the network to avoid unnecessary drastic deformation.

4 Evaluation

4.1 Grasp Quality

We present preliminary evaluation of our method on one object category from Shapenet (“Remotes"), by generating deformed grasps for a set of ’remote’ objects and computing the Robust Force Closure metric with the known true shape of each object. Then on we derive three metrics to quantify the performance of our method from different perspectives as seen in Table 1.

We include the top 50 average as a global overview for the effect of deformation on grasps in general, the top 5 average since it is most realistically significant for robot performance to have only several grasps to try, and the top grasp average since it captures the variance involved in individual grasps under a deformation and also the best-case grasping behavior. As expected, the deformed grasps are not as good as grasps generated with the analytic method on the ground truth shapes in all of these measures. However, the grasp generation method requires the mesh of the object that may not always be available when grasping in the wild. Contrarily, *our method only requires a library of templates and an image of the object*. With that context, our method achieves a reasonable retention of grasp quality even after deforming the original mesh.

Source	Top 50	Top 5	Best Grasp
Generated	0.247	0.511	0.59
Deformed (Ours)	**0.131**	**0.328**	**0.40**

Table 1: Quantitative results comparing grasps generated using the known true shape of an object and through deformation using our approach. Deformed grasps get worse evaluation across the board, but notably the best deformed grasps are proportionally closer to non-deformed grasps than the two sums over all grasps.

4.2 Semantic Grasps

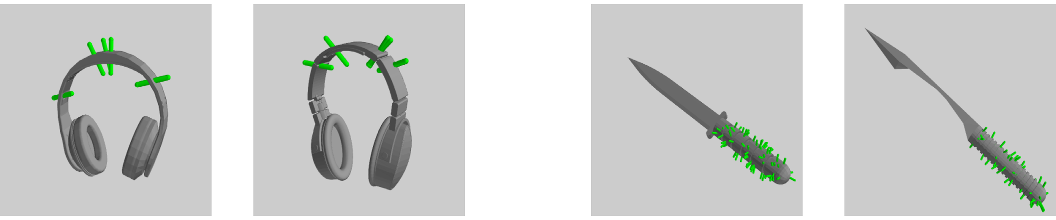

Figure 2: Examples showing deformation can transfer grasps while preserving their semantic significance. In the first example, the grasps on the bridge of the template headphone (left) are transferred to a new pair of headphones (right). Similarly, grasps on the knife handle are transferred to a handle of a new test model.

One strength of deformation for grasping is that it allows for data-efficient semantic grasps. For example, it is reasonable to want to grasp a hammer by the handle and not the head. Generating large amounts of training data for grasps that fit semantic criteria is expensive and time-consuming, but it is feasible to label a few of the templates with these grasps and then learn the deformation in general. As shown in Figure 2, our approach inherently retains the semantic meanings of grasps. This property of the deformation approach cannot be replicated by approaches that generate grasps, and can likely be extended to explicitly deal with object affordances in the future.

5 Discussion and Future Work

We have presented a novel approach for robust robotic grasp transfer based on generating a database of objects with known 3D shapes and grasps, querying this database for appropriate ’template’ shapes using a 2D image of an object to be grasped, and deforming the 3D template along with its known grasps using a deep learning model. Based on evaluation on a small set of shapes within one category using the Robust Force Closure metric, the deformed grasps are generally inferior to grasps generated using analytic methods but are nevertheless still likely to be usable in a robotics context. Having validated that this is a promising approach for grasp transfer, we plan to work towards making this a fleshed out method for grasping through multiple extensive modifications: converting the similarity query and deformation methods to make use of RGBD data, exploring modifications to DeformNet to better generate grasps, generating a much larger dataset with many object categories for training and evaluation, and ultimately testing the grasp execution with a real robot in an end-to-end manner.

References

[1] D. Prattichizzo and J. C. Trinkle. Grasping. In *Springer handbook of robotics*. Springer, 2016.

[2] J. Mahler, F. T. Pokorny, B. Hou, M. Roderick, M. Laskey, M. Aubry, K. Kohlhoff, T. Kröger, J. Kuffner, and K. Goldberg. Dex-net 1.0: A cloud-based network of 3d objects for robust grasp planning using a multi-armed bandit model with correlated rewards. In *Robotics and Automation (ICRA), 2016 IEEE Int'l Conf. on*, pages 1957–1964. IEEE, 2016.

[3] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. *arXiv preprint arXiv:1703.09312*, 2017.

[4] M. Matl, J. Mahler, and K. Goldberg. A global algorithm for transferring parallel-jaw grasps between object mesh sub-segments. In *CASE*, 2017.

[5] J. Bohg, A. Morales, T. Asfour, and D. Kragic. Data-driven grasp synthesis—a survey. *IEEE Transactions on Robotics*, 30(2):289–309, 2014.

[6] D. Seita, F. T. Pokorny, J. Mahler, D. Kragic, M. Franklin, J. Canny, and K. Goldberg. Large-scale supervised learning of the grasp robustness of surface patch pairs. In *Simulation, Modeling, and Programming for Autonomous Robots (SIMPAR), IEEE Int'l Conf. on*, 2016.

[7] 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 Int'l Journal of Robotics Research*, page 0278364917710318, 2016.

[8] L. Pinto and A. Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In *Robotics and Automation (ICRA), 2016 IEEE Int'l Conf. on*, 2016.

[9] F. T. Pokorny, Y. Bekiroglu, and D. Kragic. Grasp moduli spaces and spherical harmonics. In *Robotics and Automation (ICRA), 2014 IEEE Int'l Conf. on*. IEEE, 2014.

[10] R. Detry, C. H. Ek, M. Madry, J. Piater, and D. Kragic. Generalizing grasps across partly similar objects. In *Robotics and Automation (ICRA), 2012 IEEE Int'l Conf. on*, 2012.

[11] A. Paikan, D. Schiebener, M. Wächter, T. Asfour, G. Metta, and L. Natale. Transferring object grasping knowledge and skill across different robotic platforms. In *Advanced Robotics (ICAR), 2015 Int'l Conf. on*, pages 498–503. IEEE, 2015.

[12] F. Heinemann, S. Puhlmann, C. Eppner, J. Élvarez-Ruiz, M. Maertens, and O. Brock. A taxonomy of human grasping behavior suitable for transfer to robotic hands. In *Robotics and Automation (ICRA), 2015 IEEE Int'l Conf. on*, pages 4286–4291. IEEE, 2015.

[13] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. In *European Conf. on Computer Vision*, pages 628–644. Springer, 2016.

[14] R. Girdhar, D. Fouhey, M. Rodriguez, and A. Gupta. Learning a predictable and generative vector representation for objects. In *ECCV*, 2016.

[15] H. Fan, H. Su, and L. Guibas. A point set generation network for 3d object reconstruction from a single image. *arXiv preprint arXiv:1612.00603*, 2016.

[16] Y. Xiang, W. Kim, W. Chen, J. Ji, C. Choy, H. Su, R. Mottaghi, L. Guibas, and S. Savarese. Objectnet3d: A large scale database for 3d object recognition. In *European Conf. Computer Vision (ECCV)*, 2016.

[17] T. W. Sederberg and S. R. Parry. Free-form deformation of solid geometric models. *ACM SIGGRAPH computer graphics*, 20(4):151–160, 1986.

[18] J. Weisz and P. K. Allen. Pose error robust grasping from contact wrench space metrics. In *Robotics and Automation (ICRA), 2012 IEEE Int'l Conf. on*, pages 557–562. IEEE, 2012.

[19] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embedding. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2016.