Grasping Affordance

Maxwell Gray Ty Trusty Allen Wang Mike Griffin

Abstract

In this work, we propose to find natural human grasps of objects using a twofold approach, similar to prior work in the domain of full human body affordance, in the spirit of Shape2Pose (ref). The first step is to learn the parameters of several specialized neural networks whose outputs will correspond to an energy over the space of hand configurations, whose minima are designed to correspond to "natural" grasps. The second is to find a hand pose that corresponds to a local minimum of the learned energy function. (include results here when we have them)

1. Introduction

Affordance has been studied well in the context of human bodies (reference shape2pose here), and has applications to understanding the semantics and function of novel objects. As far as grasping is concerned, only very limited models have been built for very specific objects (ref the paper about how humans actually grasp things). Both problems are typically formulated as optimization problems that involve geometric information about the object as well as prior knowledge about likely poses. For example, the authors of Shape2Pose (ref) define their energy function as a weighted sum of various metrics, including feature compatibility, a pose prior, and several geometric constraints such as self-intersection and distance from assigned contact points. These approaches do use machine learning to an extent, but notably deep learning has not had nearly the impact in this domain that it has elsewhere (e.g., computer vision) (ref Alexnet or something).

On the other hand, there have been several impressive applications of deep neural networks to computational biology, specifically in learning protein-ligand conformation scoring functions (ref those papers we found). These biological fields share certain aspects with the problem of affordance, in the sense that both problems consist of finding optimal pose parameters for one object in terms of the geometry of another. Therefore, although we do not directly use the approaches found in the biological literature, we take inspiration from this success in formulating our own approach to natural human grasps.

2. Formulation

The energy function we propose takes the form of a weighted sum of three terms:

$$E(\mathbf{q}) = w_{\text{pose}} E_{\text{pose}}(\mathbf{q}) + w_{\text{feat}} E_{\text{feat}}(\mathbf{q}) + w_{\text{dist}} E_{\text{dist}}(\mathbf{q}) \quad (1)$$

where $\mathbf{q}$ is the hand configuration, and E is an object-dependent energy function, $E = E[X]$ for an object X. In practice the hand configuration is a vector of joint angles as well as a root position in space, and we represent the objects as both point clouds and triangular meshes for separate purposes.

2.1. Pose Prior (E_{pose})

The energy E_{pose} plays a similar role to the direct pose prior in the original Shape2Pose work, in that it represents a distribution over natural human poses. However, rather than directly fitting a normal distribution to training poses, we train an Energy-Based Generative Adversarial Network (EBGAN) (ref) with an autoencoder as its discriminator and use the discriminator's reconstruction error for this energy term. The authors of the EBGAN paper note that the discriminating autoencoder can be seen as being regularized by contrastive samples from the generator, leading to a better representation of the input distribution. This intuition, as well as the autoencoder's ability to represent complex distributions, motivates this energy term.

2.2. Feature Compatibility (E_{feat})

Again we draw inspiration from the original Shape2Pose, where feature compatibility is defined in terms of a negative log likelihood that a body part contacts the object at a point p:

$$E_{\text{feat, S2P}} = -\sum_{p \in \text{Body}} \log V_p(m(p)) \quad (2)$$

where V_p is a random forest regression model for the probability that a part p makes contact with the assigned point $m(p)$.

The corresponding energy in our formulation is parameterized by a neural network rather than a random forest. In particular, we use a PointNet++ network (ref) to predict the log probability that each point on the mesh is the nearest point to each joint or end effector on the hand. An important difference between our formulation and that of Shape2Pose is that we do not directly predict contacts in this way; we simply score based on nearest points.

2.3. Distance (E_{dist})

A distance penalty is imposed on the hand for the sum of the distances from each joint and end effector to its nearest point.

$$E_{\text{dist}}(\mathbf{q}) = \sum_{i=1}^{k} ||\mathbf{p}_i - x_i(\mathbf{q})||_2^2 \tag{3}$$

$$\text{where } \mathbf{p}_i = \arg \min_{\mathbf{p}' \in \text{Object}} ||\mathbf{p}' - x_i(\mathbf{q})||_2^2 \tag{4}$$

and $x_i(\mathbf{q})$ is the position of the ith joint, and k is the number of such joints. In this way, despite random initialization, our energy will assign higher scores to configurations near to the object, as we expect a grasp should be.

3. Learning the Energy

3.1. Pose Prior

To learn a prior over poses, we trained an EBGAN model on the Columbia GraspIt dataset (Allen pls explain, maybe add a pic of the histograms you showed us).

3.2. Feature Compatibility

We also trained the PointNet++ on the Columbia dataset, using example grasps to learn nearest-point probabilities (Mike pls explain how you trained and the class imbalance thing if possible).

4. Results

5. Discussion

References

[1] A. Alpher. Frobnication. *Journal of Foo*, 12(1):234–778, 2002.

[2] A. Alpher and J. P. N. Fotheringham-Smythe. Frobnication revisited. *Journal of Foo*, 13(1):234–778, 2003.

[3] A. Alpher, J. P. N. Fotheringham-Smythe, and G. Gamow. Can a machine frobnicate? *Journal of Foo*, 14(1):234–778, 2004.

[4] Authors. The frobnicatable foo filter, 2014. Face and Gesture submission ID 324. Supplied as additional material `fg324.pdf`.

[5] Authors. Frobnication tutorial, 2014. Supplied as additional material `tr.pdf`.