Vision, Modeling, and Visualization (2017)
M. Hullin, R. Klein, T. Schultz, A. Yao (Eds.)

Data Driven Synthesis of Hand Grasps from 3-D Object Models

S. Majumder[1] H. Chen[2] and A.Yao[1]

[1]Institut für Informatik II, Computer Graphik, Universität Bonn, Germany
[2]RWTH Aachen University, Germany

Abstract

Modeling and predicting human hand grasping interactions is an active area of research in robotics, computer vision and computer graphics. We tackle the problem of predicting plausible hand grasps and the contact points given an input 3-D object model. Such a prediction task can be difficult due to the variations in the 3-D structure of daily use objects as well as the different ways that similar objects can be manipulated. In this work, we formulate grasp synthesis as a constrained optimization problem which takes into account the anthropomorphic and kinematic limitations of a human hand as well as the local and global geometric properties of the interacting object. We evaluate our proposed algorithm on twelve 3-D object models of daily use and demonstrate that our algorithm can successfully predict plausible hand grasps and contact points on the object.

1. Introduction

Given an object, *grasp synthesis* refers to the problem of finding a plausible grasp configuration that satisfies a set of criteria relevant for interacting with the object. Modeling and predicting human hand grasps is an active and popular area of research as it has applications in robotics [SDN08], computer vision and computer graphics [Liu09]. Existing grasp synthesis algorithms can be broadly divided into two categories : *analytic* [SEKB12] and *data-driven* [BMAK14]. Given an input object model, analytic approaches determine the contact locations on the object and grasping pose through kinematic and dynamic formulations. Analytic approaches are known to be computationally expensive as a certain number of conditions have to be satisfied for a successful grasp [SEKB12]. Contrary to analytic approaches, the data-driven paradigm places more emphasis on learning models that capture the relationship between the object's shape and features and the grasping pose by training on annotated examples. As 3-D data acquisition devices and modeling tools became more widely available, research in data-driven direction gained more traction within the community [Shi96, BMAK14]. In this work, we also adopt a data-driven approach which models the hand-object interaction and automatically synthesizes 3-D hand grasps when presented with an object model (refer to Figure 1).

We are motivated by the energy minimization approach of [KCGF14], which automatically predicts human pose and contact points when given the 3D structure of an object such as a bicycle or a fitness machine. The energy minimization incorporates local affordance features as well as global constraints such as symmetry of the human body and human pose priors. We adopt a similar approach for synthesizing realistic hand grasps given a 3-D object model. However, the model in [KCGF14] cannot be directly applied to the grasp synthesis problem because unlike the human body, the human hand is not symmetric. Furthermore, grasp stability is an important factor to consider when synthesizing hand grasps for object interaction, *i.e.* physically possible hand grasps are not always natural nor plausible in real life due to a lack of object stability.

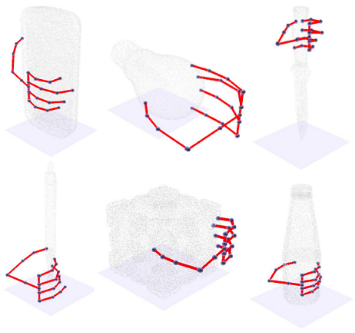


Figure 1: *Given a 3D object model as input, we predict a plausible hand pose and contact points on the object surface.*

We propose an energy-minimization approach for the task of 3-D grasp synthesis and summarize our contributions as follows. *First*, we relax the symmetry constraints of [KCGF14] by proposing a modified energy term that reflects the part-wise reflectional symmetry of the human hand. *Secondly*, we propose a novel energy term which leads to the synthesis of stable grasps. Stability of synthesized hand grasps is a feature that is often found only in *analytic* approaches but with our proposed energy term, we are able to incorporate this desirable property into a data-driven paradigm. *Third*, to speed up the computation of the energy-minimization, we propose

DOI: 10.2312/vmv.20171258

a simplified hand kinematic model with 22 degrees of freedom and 7 contact parts. The proposed hand model predicts similar grasps as more complex models with 21 contact parts but in only a fraction of the time. *Finally*, to validate our approach we present a dataset covering 6 grasp types and 12 types of objects with complete annotations for the hand contact points and the 3-D hand model.

2. Related Works

Existing grasp synthesis algorithms can be broadly divided into analytic and data-driven approaches. We give only a short overview and refer the reader to existing surveys [SEKB12, BMAK14] for more details. Instead, we primarily focus on approaches which cast the hand grasp synthesis as an energy minimization.

2.1. Analytic Approaches

Analytic approaches focus on the analysis of kinematics, stability and/or dynamic formulations. Several of these approaches aim to synthesize stable grasps [Liu00, DLW00, LLD04]. These approaches are often dependent on an ideal background such as simplified contact models [Ngu88], Coulomb friction [HPK13] and rigid body modeling [SK16, MLSS94]. When applied to real world scenarios, synthesized grasps may be improper (*anthropomorphically not possible*) [PT08] due to ambiguities and imperfections unaccounted for in the formulations.

2.2. Data-Driven Approaches

Data-driven or empirical approaches rely on learning from examples and predict graspable regions based on object geometric features [Sax09, LLS15]. These examples can either be provided in the form of generated labeled training data, human demonstration or through trial-and-error. A standard data-driven approach samples grasp candidates given an object and then ranks them according to some metric [BMAK14]. The approach in [MCFdP04] learns a vision-based grasp system by repeating a large number of grasping actions on different objects. In [SDN08], a simple logistic regressor is learned based on large amounts of synthetic training data to predict grasps without the need for satisfying any kinematic or stability constraints. More recently, there has been focus on the relationship between grasp prediction and object features [BK10, HCCJ10]. In comparison to analytic approaches, data-driven approaches pay more attention to the aspects of the object representation and perceptual processing. As a result, the data-driven approaches may generate grasps which are improper, as pointed out in [KEK09].

2.3. Grasp Synthesis as Energy Minimization

Several approaches, both analytic and data-driven, have cast grasp synthesis as an energy minimization problem [Lia, JGT11, CGA07, HWA*12]. Jia *et al.* in [JGT11] proposed a two-finger grasping approach for deformable objects by minimizing the object's potential energy under external squeezing forces. Ciocarlie *et al.* in [CGA07] use to simulated annealing to minimize an energy term based on local geometric features such as distances between the contact points and object surface, and angular differences between surface normals at the contact locations and the closest point on the object.

The works to date have formulated the energy minimization based on either only the synergy of the pose and geometric features [CGA07, HWA*12] or on force and stability analysis [JGT11]. We propose an optimization framework which incorporates both the compatibility of hand poses with local geometric features of the object as well the stability of the object on application of a particular grasping pose *jointly* during the energy minimization. This allows us to synthesize physically possible hand grasps which ensures object stability during the interaction.

3. Approach

Our proposed approach proceeds in two stages : learning a hand-object interaction model and using this learned model to infer the grasping pose when presented with an input shape.

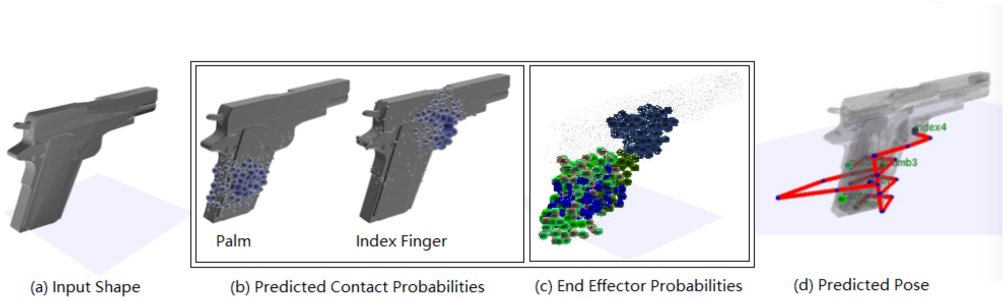
For learning, the input is a collection of 3-D shapes with manually annotated contact points and poses represented by the joint angles. Our goal is to learn an interaction model that is able to measure the quality of a pose given an input object shape. The interaction model incorporates terms learned from examples to model the local geometry of contact points and the joint angles for hand interaction poses, and it includes penalty terms for deviations from the part-wise reflectional symmetry of the human hand, intersections with the shape and penalizes unstable grasping poses.

For inference, the input is a novel shape, and the output is a set of joint angles and contact points parameterizing the most likely hand interaction pose. The key algorithm in this stage searches the combinatorial space of hand poses to find the ones with lower energies (meaning higher compatibility) according to the interaction model. First, possible contact points on the object are sampled; this constrains the search space for possible hand poses. We then sample large number of poses from the learned joint angle distributions. The distribution of the hand parts and the sample points are then aligned using a rigid transformation. For each *aligned* pose-contact points pair, the exact value of the objective function is evaluated. The pose with the lowest energy is selected as the final solution. An overview of our approach is given in Figure 2.

3.1. Kinematic Model of Hand Skeleton

Estimating an accurate kinematic model of the human hand is rendered difficult by its anatomical complexity. Consequently, simplifying assumptions are often made in analytic solutions to ease the implementation or speed up computations [BBD12]. The human hand has 27 degrees of freedom (DOFs) : 4 in each of the four fingers, 3 for extension and flexion, 1 for abduction and adduction; the thumb has 5 DOFs and remaining 6 DOFs for the rotation and the translation of the wrist [AD09, ES03].

We make the following simplifying assumptions on top of the 27 DOF model. First, in contrast to the standard model, we simplify the role of the thumb to behave like the any other finger. Also, it has one fewer joint and thus has 3 DOFs instead of 4 (the DOF for all other fingers). Secondly, for our experiments, we assume that the input object model is presented in an upright position. Thus, we remove the DOFs corresponding to rotation and are left with 3 DOFs (corresponding to translations in the *xyz* plane). In total, our kinematic hand model has 22 DOFs as shown in Figure 3.



(a) Input Shape (b) Predicted Contact Probabilities (c) End Effector Probabilities (d) Predicted Pose

Figure 2: *Grasp Synthesis Pipeline : (a) Given an input 3D shape, (b) we first classify the surface for possible contact points corresponding to each key part of our kinematic hand model, (c) find the probability distribution for each contact part by sampling hand poses from training examples, and (d) predict the grasping pose by minimizing energy terms corresponding to (b) and (c).*

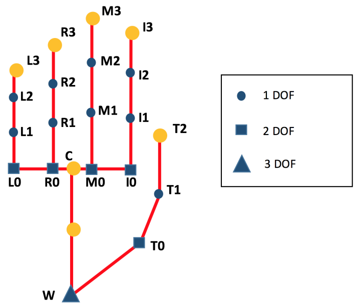


Figure 3: *Hand skeleton model with* 22 *degrees of freedom. The circles (in yellow) indicate the key parts of our proposed hand model which make physical contact with the object.*

The joint angles of the hand specify only the pose which it has to assume when interacting with objects. To fully determine the grasp, we also need to predict where the hand makes physical contact with the object of interest. We refer to the hand parts that establish contact as *key parts*. For precise grasp predictions, all the finger tips, finger joints and points connecting the base of each finger and thumb to the root of our kinematic hand model (denoted as C in Figure 3), as well as C itself can be assigned as a *key part*. This totals 21 key parts - $\{L_{i=0\to 3}, R_{i=0\to 3}, M_{i=0\to 3}, I_{i=0\to 3}, T_{i=0\to 2}, W, C\}$. However, this imposes a heavy computational burden. For our work, we identify 7 locations on the human hand as *contact parts* as shown in Figure 3. Later in our experiments, we observed that having 7 contact parts instead of 21 leads to minimal loss of precision as observed during our experiments.

3.2. Modeling Hand-Object Interaction

In this work, we cast the grasp synthesis problem as an energy minimization problem and adopt a framework similar to [KCGF14]. Based on the observation that local geometric features are often insufficient for predicting contact points, Kim *et al.* in [KCGF14] proposed a framework that allows the incorporation of both local constraint (in the form of key part-local geometric feature compatibility) and global constraints stemming from anthropomorphic limitations and properties of the human pose into a single energy minimization framework. Similar in concept, we want the inferred grasp to interact with the object in a believable manner, *i.e.* make contact with the object in "graspable" areas. This is in turn dependent on the local geometric properties of the object. Besides, we want to ensure that the synthesized grasp does not intersect the object surface. Furthermore, we add the functional constraint of 'do not drop the object' in order to synthesize stable grasps. In our proposed interaction model, each individual energy terms addresses one of the aforementioned issues.

In the learning stage, we model the hand-object interaction for a class of shapes. Our goal is to build a model that can be used to evaluate the interaction between a shape S and a hand grasp represented by a rigid transformation T, joint angles $\hat{\theta} = \{\theta_1, \ldots, \theta_n\}$ where n is the number of joints, key hand parts P (tip of each finger, center of the palm, and the base of the four fingers), and contact point assignments $m : P \to S \cup \{ground, unassigned\}$. Some hand parts may be unassigned and rest in free space : $p \to unassigned$, or may be placed on the ground plane : $p \to ground$.

Our proposed model searches over a space of all plausible hand grasps, and picks a grasp minimizing the following objective:

$$\begin{aligned}\mathcal{E}(T,\hat{\theta},m,S) &= w_{dist}\mathcal{E}_{dist}(T,\hat{\theta},m,S) + w_{feat}\mathcal{E}_{feat}(m,S)\\ &\quad + w_{pose}\mathcal{E}_{pose}(\hat{\theta}) + w_{stab}\mathcal{E}_{stab}(T,m,\hat{\theta},S) \qquad (1)\\ &\quad + w_{isect}\mathcal{E}_{isect}(T,\hat{\theta},S)\end{aligned}$$

$\mathcal{E}_{dist}$ and $\mathcal{E}_{feat}$ are local energy terms assigned to the key parts; $\mathcal{E}_{dist}$ penalizes key parts that do not make physical contact with the object while $\mathcal{E}_{feat}$ penalizes contact assignments when the corresponding key part is incompatible with the local surface geometry. The remaining energy terms define global pose constraints: $\mathcal{E}_{pose}$ penalizes implausible poses, $\mathcal{E}_{stab}$ penalizes unstable grasping poses, and $\mathcal{E}_{isect}$ penalizes surface intersections.

3.2.1. Contact Distance [KCGF14]

If a hand part is assigned to a surface point on the 3-D object, we want the hand part to establish physical contact with the object. To ensure this, we penalize large separations between the object and the assigned contact part. The energy term is given by

$$\mathcal{E}_{dist} = \sum_{p\in P,\; m_p \neq\, unassigned} \|T\mathbf{p}_\theta - m_p\|^2, \qquad (2)$$

where $\mathbf{p}_\theta$ is the position of key hand parts $p \in P$ given joint angles θ and m_p denotes the contact point assignments for each key part p. Parts assigned to the ground are measured by separation in height.

3.2.2. Feature Compatibility [KCGF14]

The feature compatibility measures how likely it is for a surface point on the object to be in physical contact with a particular hand part. Given training shapes $S_1, S_2, ..., S_M$ with annotated ground truth contacts $m_i : P \rightarrow S_i$, we learn a regression model $V_p : S \rightarrow [0,1]$ for each part $p \in P$ which estimates the probability that it will be placed on a point on a query surface S. The model relies on local geometric features to predict which regions are compatible with which hand part: for instance, large flat/cylindrical surfaces are meant for the palms and small homogeneous surfaces (such as trigger or button) are meant for more assertive parts such as the thumb and index finger.

Using the iterative farthest-point algorithm, $1000 \cdot A$ points $C_{S_i} = \{c_1, c_2, \cdot, c_K\}$ are sampled on each shape S_i, where A is the shape's surface area in square centimeters. Geometric features such as local neighborhoods, local symmetry axes, curvature, shape diameter function, and a histogram of distances are computed at these points.

Next, for each body part p and training shape S_i, we can compute a normalized measure V_p^i which is 1 at the ground truth contact point m_p^i and decays to zero. We define $V_p^i(c_j)$ at sample point c_j as

$$V_p^i(c_j) = \exp\left(\frac{-g(c_j, m_p^i)^2}{\tau^2}\right),$$

where $g(,)$ is the geodesic distance and τ is a tuning parameter. τ is chosen in a way such that $V_p^i(c_j)$ is 0.4 at a geodesic distance of 2 cm.

For each hand key part p, we train a random regression forest with 30 trees to estimate V_p. When predicting the pose, the regression forest is used to predict feature compatibility at each candidate contact point assigned to a hand part. The overall compatibility is measured by the energy term $\mathcal{E}_{feat}$ given by,

$$\mathcal{E}_{feat} = \sum_{p \in P} -\log V_p(m_p) \tag{3}$$

For parts mapped to the ground plane or left unassigned, the feature compatibility is estimated from training data statistics with $V_p(ground) = M_{ground}/M$ where M_{ground} is the number of times part p was placed on the ground or left unassigned. A lower bound of 0.1 is set to avoid infinite energies.

3.2.3. Pose Prior and Symmetry

The pose prior helps to distinguish between plausible (*anthropomorphically possible*) poses from implausible ones [KCGF14]. Similar to [HEKL*13] we use a Gaussian Mixture Model (GMM) to learn a probabilistic encoding of finger joint angle distributions. We use the same hand skeletal model in all examples. Each hand pose is represented by a 26 dimensional $\hat{\theta}$ - 22 degrees of freedom and 4 parameters for the location and rotation.

First, we use standard k-means clustering to group all input training poses into $\mathcal{L}$ clusters. In most cases, we set $\mathcal{L} = 3$. Then, for each cluster l_k (where $k = \{1, 2, \cdots, \mathcal{L}\}$), we use a Gaussian with learned mean $\mu_i^{l_k}$ and standard deviation $\sigma_i^{l_k}$ to represent the variation of the θ_i - the i-th joint angle. Note that the distribution of each joint angle is modeled independently

The human hand lacks (reflective) symmetry like the full human body. However, by analyzing the detailed grasp taxonomy of [LFNP14], we observe that 56 of the 73 grasp types have symmetric behavior across the middle, ring, and pinkie fingers. To interact with objects with triggers or buttons, people often use the thumb (the remote control) or the index finger (the gun and the spray bottle), which makes their pose different than that of the middle, ring and pinkie fingers. As such, we relax the constraints of [KCGF14] and incorporate a 3-finger symmetry in the pose prior energy term. We set the joint angles of the ring and little finger to be symmetric with the corresponding joint angles on the middle finger. For each symmetric pair $(\theta_i, \theta_i^{sym})$, the deviation of the joint angles is represented with a Gaussian : $|\theta_i - \theta_i^{sym}| \sim \mathcal{N}(\mu_i^{sym}, \sigma_i^{sym})$, where a smaller σ_i^{sym} indicates that the middle, ring and pinkie fingers are aligned in an symmetrical manner in a grasp.

The pose-prior energy term is now given by

$$\mathcal{E}_{pose} = min_{l \in \mathcal{L}} \sum_{i}^{26} \frac{|\theta_i - \mu_i^l|^2}{(\sigma_i^l)^2} + \frac{(|\theta_i - \theta_i^{sym}| - \mu_i^{sym})^2}{(\sigma_i^{sym})^2} \tag{4}$$

The first term in the summation penalizes the deviations of the inferred joint angle and the joint angle distribution learned from the examples. It prefers poses which are similar to the ones observed during training. The second term in the summation penalizes inferred poses which violate the symmetrical behavior observed during training.

3.2.4. Stability

A grasped object is defined to be in equilibrium if the sum of all forces and the sum of all moments acting on it are equal to zero [Shi96]. However, an equilibrium grasp can both be stable or unstable. A grasp is said to be stable when the grasped object is in equilibrium (no net forces and torque) and it should be possible to increase the grasping force's magnitude to prevent any displacement due to an arbitrary applied force [VI12, Cut89]. *Force closed grasps* are a subset of equilibrium grasps which have the important property of being stable [SEKB12]. Force closure is an important property in grasping and has an extensive literature [MLSS94, Ngu88]. In grasp synthesis, we want to generate grasps not only with plausible poses and contact points, but also ensure that objects of interaction are stable. We introduce a novel energy term which ensures the predicted grasp results in *force closure* by restricting the motion of the object through the contact forces exerted by the hand. For simplicity, we assume that all contacts between the fingertips and the objects are point contacts which can only exert a normal force through the point of contact and a frictional force along the surface in a direction perpendicular to that of the normal force.

In the simplest scenario, we assume that 2 contact points are required to make an object stable (Figure 4 (a)). Furthermore, we assume that the forces exerted through the contact points are equal in magnitude and they are applied at points diametrically opposite to each other. The frictional force is given by $F_f = \mu(F_1 + F_2)$ where μ is the coefficient of friction of the surface. If the magnitude of F_f is bigger than the force exerted by gravity, then the object remains stable in the z-axis. Furthermore, if F_1 and F_2 are equal and opposite in direction (180° between them), then they cancel each other out in the xy-axis.

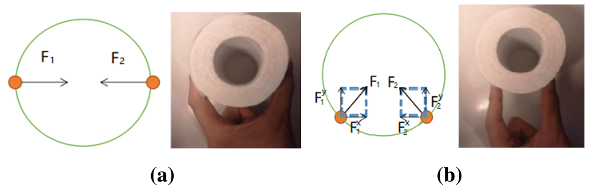


Figure 4: *(a) Normal forces F_1 and F_2 of equal magnitude applied at diametrically opposite points into the surface cancel each other out. (b) As the angle between the normal forces F_1 and F_2 decrease, the y-components increase making the object unstable.*

Now consider an angle between the force vectors smaller than 180° (Figure 4 (b)). There are still two forces in operation: F_1 and F_2. F_1^x and F_1^y are the components in x and y directions of force F_1. Likewise, F_2^x and F_2^y for F_2. Even if F_1^x and F_2^x are in equilibrium in x direction, the object is not in mechanical equilibrium in the direction of y because of the additive nature of F_1^y and F_2^y. Furthermore, as the angle between F_1 and F_2 gets smaller than 180 degrees, the y-components increase, making the object more unstable.

Based on these observations, we propose an energy term $\mathcal{E}_{stab}$ which favors hand parts to be assigned to sampled contact points which are 180° apart. Let $p,q \in P$ denote the p-th and q-th key part respectively. Also, let $a_p = T\mathbf{p}_\theta$ indicate the position of the key part p given joint angle θ and $b_p = m_p$ denote the contact point for hand part p. The force vector for key part p can now be denoted by the vector $\overrightarrow{a_p b_p}$ and likewise $\overrightarrow{a_q b_q}$ for key part q.

The energy term $\mathcal{E}_{stab}$ is then given by

$$\mathcal{E}_{stab} = \sum_{p,q \in P, p \neq q} \frac{\overrightarrow{a_p b_p}\ \overrightarrow{a_q b_q}}{|\overrightarrow{a_p b_p}||\overrightarrow{a_q b_q}|}. \tag{5}$$

During interaction with objects of daily use, we often assume a grasping pose where a pair or more of fingers are placed at large angles with respect to each other in order to impart stability. We can classify 7 key hand parts into 2 groups. One includes the tips of index, middle, ring and little fingers. The tip of the thumb, the center of palm and the center of the forward half palm form the other group. In common conditions, contact points which generate the forces with big angle are respectively from those 2 groups. For example, the tip of thumb and the tip of the index finger when interacting with disk shaped objects [VI12].

3.2.5. Intersection [KCGF14]

The intersection energy term helps us to avoid impossible grasps where the hand skeleton intersects the object. We assume the hand is represented as a skeleton, with linear bones $B = b_1, b_2, ..., b_K$ connecting the joints joints. For each link b_i, we check for $I_S(b_i)$ - the intersection with the shape S. Intersections within a small distance of the shape and the assigned contact part are ignored. Higher penalties are applied when the bone intersects the surface orthogonally. The intersection energy is given as the sum of maximal per-link penalties:

$$\mathcal{E}_{isect} = \sum_{b_i \in B} max_{q \in I_s(b_i)} |\text{normal}(q) \cdot \text{direction}(b_i)|. \tag{6}$$

3.3. Inferring the Grasping Pose

During inference, the key challenge is to efficiently sample the combinatorial search space spanned by the hand pose and the contact points. Instead of jointly minimizing over this search space, we observe that it is possible to sample high-probability contact assignments m and high-probability poses $\hat{\theta}$ independently, since they contribute to two separate terms $\mathcal{E}_{feat}(m)$ and $\mathcal{E}_{pose}(\hat{\theta})$ respectively.

3.3.1. Sampling contact points

The contact points m_p for each body part $p \in P$ are sampled independently, by picking candidate points on the shape whose compatibility energy $\mathcal{E}_{feat}(m,S)$ with p is lower than the cost of leaving them unassigned to any contact point.

3.3.2. Sampling plausible poses

We sample plausible hand poses with low energy $\mathcal{E}_{pose}$ by directly sampling the joint angle Gaussian distributions from the pose prior. In our experiments we sample $5,000$ poses in a fraction of a second. The space is discretized into grids of 1cm^3 voxels. Each voxel stores a portion of the pose prior energy of the sampled pose corresponding to the key part lying in this voxel. $\mathcal{E}_{pose}$ can be computed by adding over the partial energy in the voxels containing the individual hand key parts. Joints contributing to multiple partial energies have their contributions averaged over the overlapping paths [KCGF14]. Note that discretization introduces some approximation error at the cost of reduced complexity.

3.3.3. Pose-Contact Point Alignment

Next, for every sampled contact point m_p for the corresponding part $p \in P$, 32 rotations are considered around the up axis in an attempt to align the part distribution grids with respect to the surface. The anchor and the rotation define the rigid transformation T, which aligns part distribution grids to the surface.

Given the aligned grid, we estimate a lower bound on the feature and pose energy terms, as well as the corresponding pose, by greedily assigning body parts to contact points. Each successive assignment m_p^i) is chosen to be the one that least increases $\mathcal{E}_{feat} + \mathcal{E}_{pose}$. The 3 finger symmetry term is measured with respect to the previously assigned $i-1$ points, and the pose prior is bounded from below by the entry in the aligned voxel containing the assigned contact point.

Finally, in order to infer the best pose, we need to compute the full energy function, which requires knowledge of the exact joint angles $\hat{\theta}$. All candidate poses which were sampled previously are sorted in order of increasing estimated lower bound on energy, and for each pose we $\mathcal{E}_{dist} + \mathcal{E}_{pose}$ is minimized. Following [KCGF14], $\hat{\theta}$ is solved for iteratively until the energy $\mathcal{E}_{dist} + \mathcal{E}_{pose}$ stops decreasing. Given $\hat{\theta}$, we solve for $\mathcal{E}_{stab} + \mathcal{E}_{isect}$ and rank the predicted poses according to the lowest values of the energy function.

4. Dataset

For evaluating our proposed approach, we need a hand-object interaction dataset which contains detailed hand annotation consisting of 3-D joint locations and deformation and the 3-D object model along with the contact points. We looked through numerous datasets but they were not suitable for our approach because of missing 3-D object models [XC13, AWK15], missing 3-D joint locations and deformations [FRE*13]. In [TSLP14], the dataset consists of raw depth data of grasps with objects and it is difficult to construct the rigid hand skeleton model and also the 3-D model of the object. Even though, segmentation masks for the objects are provided, it is still not suitable for feature analysis because of the lack of precision. Most importantly, none of the mentioned datasets contain contact points annotations. Consequently, we recorded and release a new dataset to validate our proposed approach.

4.1. Grasp Taxonomy

We studied the detailed grasp taxonomy of [LFNP14], where typical human hand grasps are classified into 73 different grasps based on different object geometries and hand shapes. We select a subset of 6 highly distinct grasp types for our experiments. For each grasp type, we annotate the contact points and joint angles for $2-4$ different object categories. In total, we annotate the contact points for 111 object models spanning over 12 different object categories - ‘bottle’, ‘mug’, ‘knife’, ‘sword’, ‘phone’, ‘cube’, ‘bulb’, ‘fruit’, ‘gun’, ‘pen/pencil’, ‘spoon/fork’, ‘coin/chess pieces’. Sample annotations for each grasp type from our dataset are shown below in Figure 5. Note that the grasp types are named based on the geometry of the object with which they interact. The annotated 3-D grasps sizes are chosen to approximate real hand sizes and likewise for the 3-D object shapes.

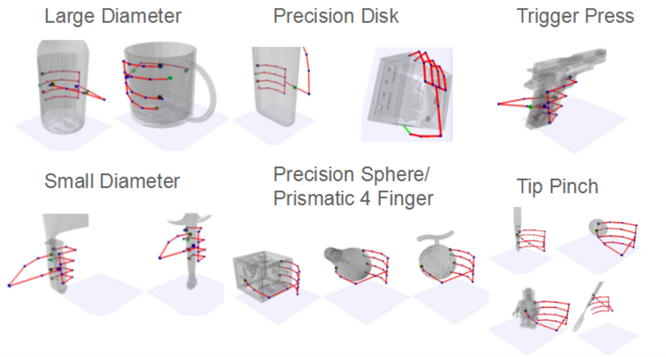


Figure 5: *Example of Dataset*

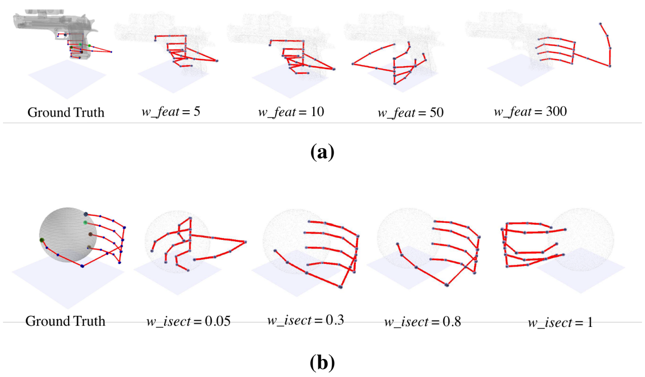
5. Experiments

In this section, we validate our proposed approach on the aforementioned dataset. We perform experiments to select appropriate weights for each energy term and to assess the correctness of the predicted grasp. We also demonstrate the improvements in grasp prediction with the introduction of our newly proposed energy terms. Finally, we show that having a simplified hand model of 7 key parts finstead of 21 leads to minimal loss in accuracy.

We ran a leave-one-out experiment for each grasp type, i.e for each grasp, we train on all models except one and predict the pose for the omitted model. In order to quantitatively evaluate the correctness of the poses predicted by our algorithm, we measure the distances between all predicted and ground-truth joint positions for each object model. In the following plots, on the vertical axis we list the fraction of joints whose error is less than the distance threshold listed on the horizontal axis (ranging from 0 to 25 millimeters).

5.1. Weights of Energy Terms

First, we want to select appropriate weights for each of the energy terms. We run several experiments by varying the weights for each of the individual energy terms while keeping the weights for the other terms fixed. We observed that the feature compatibility and intersection energy terms, $\mathcal{E}_{feat}$ and $\mathcal{E}_{isect}$, have the most impact on the quality of the synthesized grasp as shown in Figure 6. Based on the experimental evaluations on our dataset, we set $w_{dist} = 1000, w_{feat} = 10, w_{isect} = 0.3, w_{pose} = 10$ and $w_{stab} = 500$.



(a)

(b)

Figure 6: *Variation of synthesized grasp quality for different values of (a) w_{feat} and (b) w_{isect}.*

5.2. Correctness of Prediction

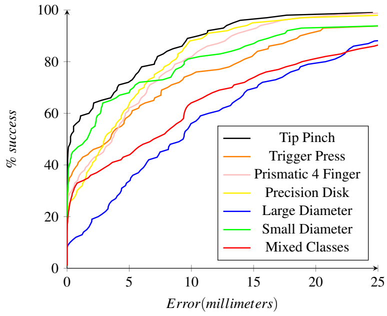


Figure 7: *Prediction accuracy on single versus mixed classes.*

From Figure 7, we observe that when the distance threshold reaches 10 millimeters, we achieve a correctness higher than 50% for all 6 grasp types. Except for the large-diameter grasp type, we reach more than 60% and even 80% correctness for certain grasps. Prediction for the *tip pinch* grasp has the best performance while the most difficult grasp to predict is for *large diameter* objects. Performance on the other four grasp types are similar. We speculate two possible reasons for the variation in performance. First,

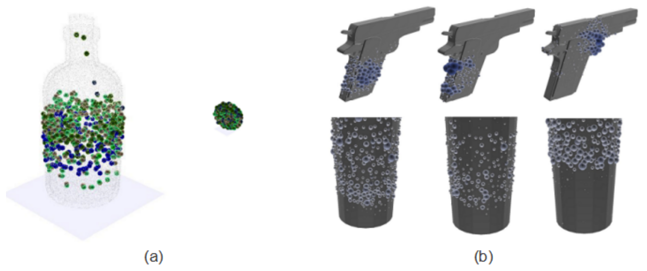


Figure 8: *(a) There are many more candidate contact points on objects with larger surface areas than objects with smaller surface areas. (b) Candidate contact points on homogeneous surfaces tend to be similar with respect to one another in terms of local geometrical features.*

the graspable area of large diameter objects (bottles and mugs) are comparatively larger than other object types which results in many candidate contact points, as opposed to smaller objects like the coin (*tip pinch*) which will have fewer candidate contact points. Secondly, majority of the candidate contact points for grasping a cylinder are on the curved side surface, where geometric features are similar. In comparison, an object like the gun has several distinct geometric features which are unique to specific hand parts. Consequently, estimating grasping pose for the category 'tip pinch' results in the best performance whereas the accuracy drops significantly (∼20% for the 0-10 mm threshold) while estimating the pose for 'large diameter'.

Figure 8(b) shows candidate contact points for the palm center, the tip of the thumb and the tip of index finger. For a gun, contact points for different hand parts are easily recognized and clearly located in 3 parts, whereas for a bottle, it is difficult to distinguish the hand parts, causing interference on the prediction.

We find that the correctness of mixed class (leave-one-out over all grasp types combined) is still higher than 60% when the threshold is 10 millimeters. The mixed class is better than the single class of large diameter but worse than the other 5 single classes, leaving us to speculate that learning the interaction model on the mixed classes have an interfering and adversarial effect on each other.

5.3. Modified Energy Terms

We compare the accuracy of the synthesized hand grasps with and without our proposed stability term and modified symmetry in the pose prior using the mixed classes. We plot the comparison in Figure 9, demonstrating that there is an improvement of ∼5% and ∼10% on an average in the prediction with the addition of the stability and the modified symmetry term respectively. The qualitative improvements in the synthesized grasp from having the new energy term for stability is shown with an example (Figure 10).

5.4. Simplified Kinematic Hand Model

In our prediction pipeline, the total number of key part plays an important role on the precision of the synthesized grasps. Having more key parts lead to improved prediction but at the cost of increased computational complexity. We compare two different models - one with 7 key parts and the other model which considers all the joints and the finger tips as contact points - 21 in total (Figure 3). We use leave-one-out method to test the system on a our proposed dataset.

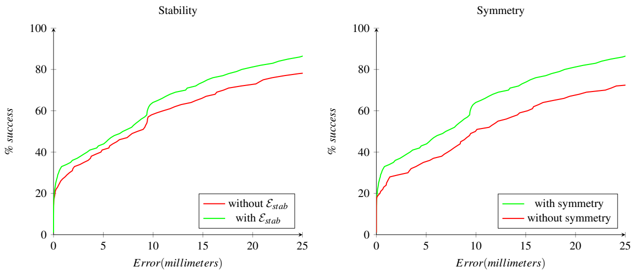


Figure 9: *Improved grasp synthesis accuracy due to the addition of the new energy term for stability and the modified energy term for symmetry.*

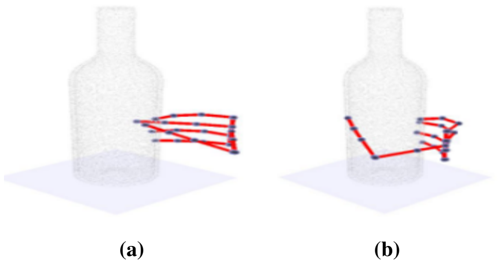


Figure 10: *(a) Without $\mathcal{E}_{stab}$ (b) With $\mathcal{E}_{stab}$. The addition of the new energy term leads to more realistic grasp synthesis.*

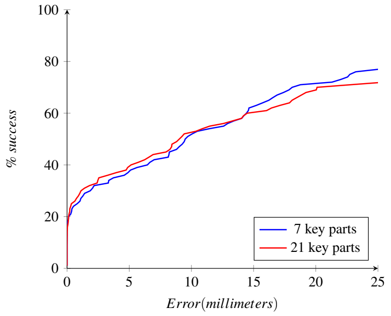


Figure 11: *Prediction accuracy for* 7 *key parts vs.* 21 *key parts.*

As can be seen from Figure 11, for lower distance thresholds (0-10 mm) the increment in precision is minimal ∼2-5%. On the other hand, average grasp synthesis estimation for 7 key-parts varies from ∼3s for small objects such as coin/chess pieces to ∼550–600s for objects with large surface areas and homogeneous features such as mug or bottle. The average estimation time rises by a factor of 25 when using the 21 key-part hand model. Thus throughout our experiments, we reported results for the 7 key-part model as it allows to keep the estimation time tractable.

6. Conclusion and Future Work

In this work, we proposed a data-driven energy minimization-based approach for grasp synthesis. Our method predicts grasping poses consistent with the local geometric features of the object, with a part-wise reflective symmetry of the hand and ensures object stability under interaction. We evaluate our proposed approach on a newly proposed dataset with 6 grasp types containing 111 annotated object models spread over 12 object categories. Our experiments show that our approach is able to synthesize grasps where $60\% - 80\%$ of the hand parts are correctly placed within a distance of 10 millimeters. Upon correctness and runtime analysis, we noticed that prediction accuracy is directly dependent on the scale of the object.

As future work, we would like to create our own large scale dataset, complete with grasp (parameterised as joint angles) and contact point annotations, for objects of daily use. We would also like to explore the *form closure* and *force closure* properties of hand grasps in detail.

References

[AD09] Agur A. M., Dalley A. F.: *Grant's atlas of anatomy*. Lippincott Williams & Wilkins, 2009. 2

[AWK15] A. Wetzler R. S., Kimmel R.: Rule of thumb: Deep derotation for improved fingertip detection. In *Proceedings of the British Machine Vision Conference (BMVC)* (2015). 6

[BBD12] Bullock I. M., Borràs J., Dollar A. M.: Assessing assumptions in kinematic hand models: a review. In *Biomedical Robotics and Biomechatronics (BioRob), 2012 4th IEEE RAS & EMBS International Conference on* (2012). 2

[BK10] Bohg J., Kragic D.: Learning grasping points with shape context. *Robotics and Autonomous Systems* (2010). 2

[BMAK14] Bohg J., Morales A., Asfour T., Kragic D.: Data-driven grasp synthesis – a survey. *IEEE Transactions on Robotics* (2014). 1, 2

[CGA07] Ciocarlie M., Goldfeder C., Allen P.: Dimensionality reduction for hand-independent dexterous robotic grasping. In *Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on* (2007). 2

[Cut89] Cutkosky M. R.: On grasp choice, grasp models, and the design of hands for manufacturing tasks. *IEEE Transactions on robotics and automation* (1989). 4

[DLW00] Ding D., Liu Y.-H., Wang S.: Computing 3-d optimal form-closure grasps. In *Robotics and Automation, 2000. Proceedings. ICRA'00. IEEE International Conference on* (2000). 2

[ES03] ElKoura G., Singh K.: Handrix: animating the human hand. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation* (2003). 2

[FRE*13] Feix T., Romero J., Ek C. H., Schmiedmayer H., Kragic D.: A Metric for Comparing the Anthropomorphic Motion Capability of Artificial Hands. *Robotics, IEEE Transactions on* (2013). 6

[HCCJ10] Hsiao K., Chitta S., Ciocarlie M., Jones E. G.: Contact-reactive grasping of objects with partial shape information. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on* (2010). 2

[HEKL*13] Huang B., El-Khoury S., Li M., Bryson J. J., Billard A.: Learning a real time grasping strategy. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on* (2013). 4

[HPK13] Hang K., Pokorny F. T., Kragic D.: Friction coefficients and grasp synthesis. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on* (2013). 2

[HWA*12] Hammond F. L., Weisz J., Andrés A., Allen P. K., Howe R. D.: Towards a design optimization method for reducing the mechanical complexity of underactuated robotic hands. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on* (2012). 2

[JGT11] Jia Y.-B., Guo F., Tian J.: On two-finger grasping of deformable planar objects. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on* (2011). 2

[KCGF14] Kim V. G., Chaudhuri S., Guibas L., Funkhouser T.: Shape2pose: Human-centric shape analysis. *ACM Transactions on Graphics (TOG)* (2014). 1, 3, 4, 5

[KEK09] Kawaguchi K., Endo Y., Kanai S.: Database-driven grasp synthesis and ergonomic assessment for handheld product design. *Digital Human Modeling* (2009). 2

[LFNP14] Liu J., Feng F., Nakamura Y. C., Pollard N. S.: A taxonomy ofeveryday grasps in action. *Humanoid Robots (Humanoids), 201414th IEEE-RAS International Conference on* (2014). 4, 6

[Lia] Liarokapis M. V.: Directions, methods and metrics for mapping human to robot motion with functional anthropomorphism: A review. 2

[Liu00] Liu Y.-H.: Computing n-finger form-closure grasps on polygonal objects. *The International journal of robotics research* (2000). 2

[Liu09] Liu C. K.: Dextrous manipulation from a grasping pose. In *ACM Transactions on Graphics (TOG)* (2009). 1

[LLD04] Liu Y.-H., Lam M.-L., Ding D.: A complete and efficient algorithm for searching 3-d form-closure grasps in the discrete domain. *IEEE Transactions on Robotics* (2004). 2

[LLS15] Lenz I., Lee H., Saxena A.: Deep learning for detecting robotic grasps. *The International Journal of Robotics Research* (2015). 2

[MCFdP04] Morales A., Chinellato E., Fagg A., del Pobil A. P.: Using experience for assessing grasp reliability. *International Journal of Humanoid Robotics* (2004). 2

[MLSS94] Murray R. M., Li Z., Sastry S. S., Sastry S. S.: *A mathematical introduction to robotic manipulation*. 1994. 2, 4

[Ngu88] Nguyen V.-D.: Constructing force-closure grasps. *The International Journal of Robotics Research* (1988). 2, 4

[PT08] Prattichizzo D., Trinkle J. C.: Grasping. In *Handbook of Robotics*. 2008. 2

[Sax09] Saxena A.: *Monocular depth perception and robotic grasping of novel objects*. Tech. rep., 2009. 2

[SDN08] Saxena A., Driemeyer J., Ng A. Y.: Robotic grasping of novel objects using vision. *The International Journal of Robotics Research* (2008). 1, 2

[SEKB12] Sahbani A., El-Khoury S., Bidaud P.: An overview of 3d object grasp synthesis algorithms. *Robotics and Autonomous Systems* (2012). 1, 2, 4

[Shi96] Shimoga K. B.: Robot grasp synthesis algorithms: A survey. *The International Journal of Robotics Research* (1996). 1, 4

[SK16] Siciliano B., Khatib O.: *Springer handbook of robotics*. 2016. 2

[TSLP14] Tompson J., Stein M., Lecun Y., Perlin K.: Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics* (2014). 6

[VI12] Venkataraman S. T., Iberall T.: *Dextrous robot hands*. 2012. 4, 5

[XC13] Xu C., Cheng L.: Efficient hand pose estimation from a single depth image. In *ICCV* (2013). 6