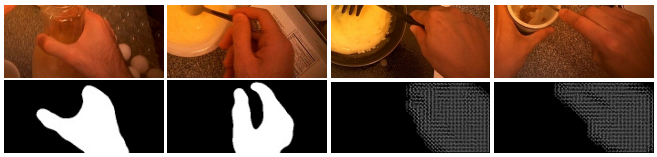
How Do We Use Our Hands? Discovering a Diverse Set of Common Grasps

De-An Huang, Minghuang Ma*, Wei-Chiu Ma*, and Kris M. Kitani
The Robotics Institute, Carnegie Mellon University.

Figure 1: Hand detection of [1] used to harvest candidate regions.



Hand Contour descriptor Masked HOG descriptor

Figure 2: Visualization of the hand contour and masked HOG feature descriptor used to represent candidate regions.

The main contribution of this work is to crystallize how recent developments in egocentric vision and data-driven techniques now make it possible to automate and advance prehensile analysis. Our aim is to show how well-established computer vision techniques can be used to advance basic science regarding the functionality of human hands. Specifically, we will show how to *automatically discover* dominant hand-object interactions from a stream of ego-centric video. While recent works have started investigating how humans interact with objects, to the best of our knowledge, this is the first work to automatically mine large video collections to discover common modes of hand-object interactions for prehensile analysis.

Our goal is to discover dominant modes of hand-object interactions from first-person videos. This is accomplished by clustering the candidate regions to discover modes of hand-object interactions. This is followed by hierarchical clustering to learn the structure of the discovered modes of hand-object interactions.

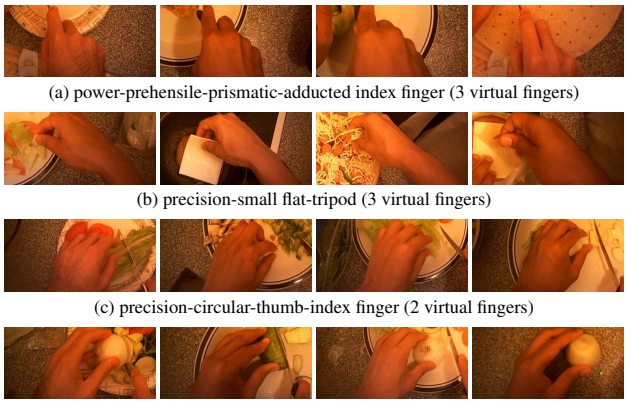
Harvesting and Representing Hand-Object Regions. We detect hands at the pixel level with [1], using code obtained from the authors. It computes a hand probability value for each pixel based on the color and texture of a local surrounding image patch. It then thresholds the probability values and extracts a set of connected components from each frame.

We use a large HOG template generated only for a masked region (Masked HOG) to represent the detected hand-object region. This representation removes the effect of the background and uses only the contour of the hand to group the regions (see Fig. 2).

Grouping Hand-Object Interactions. Classical clustering algorithms will result in multiple clusters learned over high density regions of the data distribution. In the case of hand-object interactions, certain grasp types occur more often than other types and will therefore dominate the type of clusters discovered by classic clustering algorithm. We use the Determinantal Point Process as a sampling prior to enforce diversity between discovered clusters.

To deal with the large number of candidate regions that can be generated by a large video corpus or a continuous stream of ego-centric video (near 10^6 regions in our experiments) and the high dimensionality of the data (near $8K$ dimensions), we present a simple, yet efficient clustering algorithm based on Determinantal Point Process (DPP) [3].

In the first stage, we quickly generate a set of candidate cluster centers $Y = \{y_1, \ldots, y_N\}$ by sampling exemplars through fast DPP sampling [2]. With this scheme, the sampled subset (candidate cluster centers) Y is likely to have larger diversity, and thus can cover the space of hand-object interactions more effectively. In the second stage, each candidate region $x_i \in X$ is assigned to the nearest cluster y_k and the corresponding assignment index k is stored in $Z = \{z_1, \ldots, z_M\}$. If the distance between the data point x_i and the nearest cluster center $y_k \in Y$, is less than a threshold θ, the i-th data point is assigned to $z_i = k$, otherwise it is assigned to -1.



(a) power-prehensile-prismatic-adducted index finger (3 virtual fingers)

(b) precision-small flat-tripod (3 virtual fingers)

(c) precision-circular-thumb-index finger (2 virtual fingers)

(d) medium-power-circular-tripod-push (4 virtual fingers)

Figure 3: Discovered hand-object interactions sorted by distance to centroid. Both inliers and outliers are included to illustrate the purity of the clusters. Labels are manually assigned based on Cutkosky's grasp taxonomy.

Learning a Grasp Taxonomy. We propose a new hierarchical clustering algorithm based on our online DPP-based clustering algorithm, which we describe below. The resulting subset Y obtained using a DPP prior depends highly on the selected similarity function $s(x,y)$. In this work, we use the radial basis function $s(x,y) = \exp(\frac{||x-y||^2}{h^2})$. We observe that with smaller bandwidth h, the size of the resulting subset Y is larger since it is harder for two points x and y to have a high similarity score, and thus the diversity is usually large between a set of points. Empirically, with a sequence of bandwidths $h_1 \leq \ldots \leq h_L$, we find that approximately $Y_1 \supseteq \ldots \supseteq Y_L$, where Y_ℓ is the resulting subset with bandwidth h_ℓ. This gives us a straightforward *online* algorithm to find the centers of each level in the hierarchy. In the streaming setting, given a new data point u, let Y_ℓ be the current cluster centers of level ℓ, then the probability of inserting u to Y_ℓ is defined as $P_\ell^+(u) = \prod_{j=1}^{\ell-1} p_j^+(u)$, where $p_\ell^+(u) = Pr(Y_\ell \rightarrow Y_\ell \cup u)$. Here Y_1 corresponds to the cluster centers Y discovered previously. After the cluster centers for each level are determined, we use nearest neighbors to assign the edges in the taxonomy tree. For each cluster center $y_\ell^i \in Y_\ell$, we find the nearest center $y_{\ell+1}^* \in Y_{\ell+1}$ as its parent.

Discovering Novel Hand-object Interaction. We show examples of discovered interactions in Fig. 3. A subset of the cluster exemplars correspond to Cutkosky's grasp taxonomy, such as cylindrical power grasp or abducted thumb power grasp. However, our method also learns valid grasp concepts outside of Cutkosky's taxonomy: Clamping grasps that are formed by the fingers and another supporting surface, *e.g.*holding down a piece of mushroom against a plate, and tripod precision grasps of flat objects required when pulling off a slice of cheese or bacon from its wrapping. This result is due to the fact that the *task domain* is different. Cutkosky analyzed grasps in small-batch machining operation while we focused on cooking activities. This result suggests that different tasks require a distinct set of hand interactions and also reinforces our claim that an automatic data-driven approach is necessary for large-scale analysis.

[1] Li Cheng and Kris M. Kitani. Pixel-level hand detection in ego-centric videos. In *CVPR*, 2013.

[2] Byungkon Kang. Fast determinantal point process sampling with application to clustering. In *NIPS*, 2013.

[3] Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *arXiv preprint arXiv:1207.6083*, 2012.

* indicates equal contribution.