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
{deanh, minghuam, weichium}@andrew.cmu.edu, kkitani@cs.cmu.edu

Abstract

Our aim is to show how state-of-the-art computer vision techniques can be used to advance prehensile analysis (i.e., understanding the functionality of human hands). Prehensile analysis is a broad field of multi-disciplinary interest, where researchers painstakingly manually analyze hours of hand-object interaction videos to understand the mechanics of hand manipulation. In this work, we present promising empirical results indicating that wearable cameras and unsupervised clustering techniques can be used to automatically discover common modes of human hand use. In particular, we use a first-person point-of-view camera to record common manipulation tasks and leverage its strengths for reliably observing human hand use. To learn a diverse set of hand-object interactions, we propose a fast online clustering algorithm based on the Determinantal Point Process (DPP). Furthermore, we develop a hierarchical extension to the DPP clustering algorithm and show that it can be used to discover appearance-based grasp taxonomies. Using a purely data-driven approach, our proposed algorithm is able to obtain hand grasp taxonomies that roughly correspond to the classic Cutkosky grasp taxonomy. We validate our approach on over 10 hours of first-person point-of-view videos in both choreographed and real-life scenarios.

1. Motivation

Why hands? Human hands provide a rich source of information about physical manipulation. However, extracting information about the mechanisms of the hand requires persistent and diligent observation; thus, automating the visual inspection of hand interactions will facilitate large scale analysis and has the potential for significant impact in multiple disciplines.

The earliest work on understanding the mechanism of the hand was highly influential and helped lay the ground work for many disciplines. In this paper, we are particularly interested in work that explored the space of hand manipulation

*indicates equal contribution

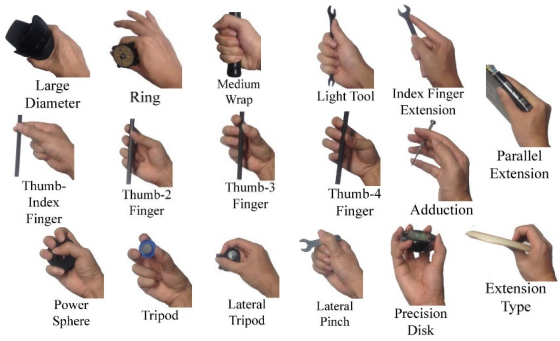


Figure 1. Canonical grasp types [11]

through discrete grasp categories or taxonomies [16, 17, 18, 28, 40, 41, 49, 52, 63]. Early work by Schlesinger [63] classified grasps into six major types based on hand and object properties. Keller [41] observed object-contact patterns in various activities of daily living to understand the statistics of hand usage. The basic findings of Keller in the 1940s played an influential role in the design of the modern artificial hand. Napier's [54] 1956 categorizations of grasps into precision and power grasps was widely adopted by researchers in the medical, bio-mechanical, and robotic fields.

In the context of robotic manipulation, domain constrained categorical representations of grasps such as Cutkosky and Wright's hand grasp taxonomy [11] played an important role in guiding robotic hand design. To this day, robotics researchers painstakingly analyze hours of hand-object interactions, manually categorizing and labeling hand use [6] because of the benefits it brings to understanding robotic arm design. In the early 1990's, Kang and Ikeuchi [37, 38, 39] presented a computational framework for grasp identification, allowing a robotic system to 'understand' a demonstrated human grasp in order to replicate the action with a robotic arm. Their work presented an important paradigm of using the visual classification of a human grasp to automate robotic manipulation; a theme which we will develop in this paper.

The use of grasp categories and taxonomies also play an important role in fields such as gesture recognition [3], com-

puter graphics [61], neuromuscular rehabilitation [70], occupational therapy [34], neuroscience [62], and child development [8]. Clearly, modeling and understanding of hand interactions is an important area of work, and automating the visual analysis of hand interaction can have far reaching consequences beyond the walls of traditional computer vision tasks such as gesture and action recognition.

Why egocentric video? We make the case that *egocentric* video – video footage from a wearable head-mounted camera – is critical for discovering dominant hand-object interactions. The first-person point-of-view (POV) gives us the optimal vantage point to observe how people manipulate and interact with objects in the real world. The first-person POV sees what the camera-wearer sees, is inherently high-resolution, and is less prone to occlusion. Consider a third-person POV camera mounted to a wall; it is very difficult to maintain a detailed view of hands when the person is moving and manipulating very small objects due to low resolution and occlusion. On the other hand, a first-person POV camera can provide detailed information about the contact points of individual fingers even when the subject is mobile.

Automating Prehensile Analysis. 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.

Here is an outline of our procedure: We first harvest candidate hand-object regions by detecting the hand regions (if any) in each frame, and then group the hand regions based on their appearance using a Determinantal Point Process (DPP). Our proposed online DPP clustering algorithm receives harvested hand regions as a temporal stream and creates new clusters based on diversity requirements. Our hierarchical algorithm concurrently learns the secondary structures between hand grasps clusters in a streaming fashion.

Contributions. (1) We develop a robust method for extracting hand regions from first-person video, (2) we propose a novel streaming data clustering algorithm using the Determinantal Point Process, (3) we propose a novel hierarchical extension to the clustering algorithm to learn a grasp taxonomy. To the best of our knowledge, this is the first work to address the problem of unsupervised visual hand grasp discovery at this level of scale.

2. Prior Work

Human-object Interactions: Recently, researchers have started investigating how humans interact with objects and scenes (*i.e.* a functional representation) in order to improve object detection, 3D understanding [24, 31], action recognition [30] or human pose estimation [71]. Functional representations have been learned in completely supervised [13], weakly supervised [12, 57], and unsupervised [26] learning paradigms. In terms of the type of data used, most of these approaches observe human actions or object behavior from *third-person videos* to extract models of the functional categories [26, 43, 55, 67], to perform scene segmentation [68] or to learn actor-object states [23]. Similar to these approaches, we also observe human actions in videos. However, unlike any of these works, we focus on *first-person videos*. Recent work has shown that egocentric cameras have an optimal point-of-view for observing hand-object interactions in unconstrained environments [5]. In egocentric videos, hands and objects are clearly visible and typically centered in the frame. The most related to our paper is work that analyzes visual structures of hand grasps using egocentric videos [7]. However, in stark contrast to our work, it performs supervised grasp classification using known grasp types and a dataset with only those known grasp types. 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.

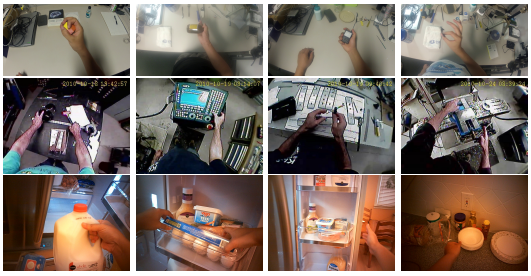


Figure 2. Images from UTG dataset [7], Yale Human Grasping dataset [5], and GTEA+ kitchen activities dataset [22].

Egocentric Sensing: Egocentric visual analysis is an emerging field in computer vision as wearable cameras are becoming readily available. Existing supervised approaches explore event segmentation using visual and motion sensors [66], joint object and action recognition [20, 59], understanding social interactions [21], activity recognition [53, 56, 60], video summarization [46, 50] or gaze prediction [48]. Unsupervised methods discover actions [42] or scenes [32] using low-level visual features extracted from ego-centric videos. In contrast, we aim to discover dominant modes of hand-object interactions using automatically detected hand regions from the first person perspective.

Discovery: Visual object discovery approaches mine for recurring visual patterns in the image collection. Prior work discovers object categories [19, 47, 65] or mid-level discriminative patches [64] by grouping recurring patterns that share similar appearance or context. Unlike previous work, which focuses on the visual properties of objects, we focus on how humans interact with objects (*i.e.* their affordances)

Figure 3. Hand detection of [9] used to harvest candidate regions.

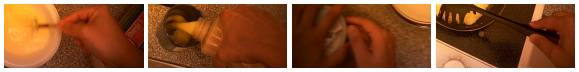


Figure 4. Candidate regions extracted around hands

to discover human-object interaction categories.

Online Clustering: In our setting, we presuppose large (potentially endless) amounts of streaming first-person POV video. It is not practical to store and process the data in batch. In this scenario, online or streaming clustering processes data points as they arrive sequentially. Algorithms such as STREAM [29], BIRCH [72] and variants of Lloyd's algorithm (*e.g.* online k-means [2], leader-follower [15]) have been proposed to solve this problem using a small amount of memory and time. Recent works on online clustering [1, 10] are able to approximate k-means clustering objective with performance guarantees. While these online clustering approaches are efficient, they are not able to enforce explicit diversity between clusters. In this work, we advocate the diversity of the discovered hand-object interactions in order to learn a wider range of grasps more effectively. Determinantal Point Process [45] is a well-known framework for sampling diverse sub-set of points, and has been successfully applied to clustering initialization [35], lexical acquisition [58], document summarization [27], and pose estimation [44]. Inspired by the fast DPP sampling scheme [35], we present a single pass additive clustering algorithm that obtains a greedy set of diverse clusters.

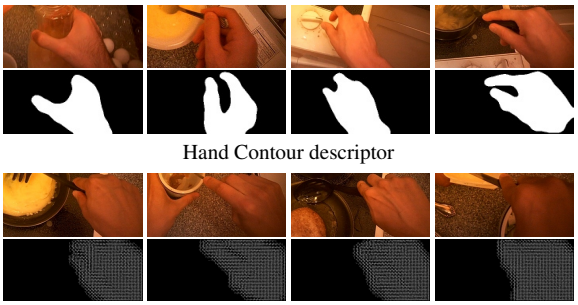
3. Our Approach

Our goal is to discover dominant modes of hand-object interactions from first-person videos. This is accomplished in four steps: (1) harvesting candidate hand-object regions, (2) extracting features from the candidate regions, (3) clustering the regions to discover modes of hand-object interactions, and (4) hierarchical clustering to learn the structure of the discovered modes of hand-object interactions.

3.1. Harvesting Candidate Hand-Object Regions

In order to discover clusters of hand-object interactions, we first need a means to robustly extract the key frames and bounding boxes that capture important hand-object regions. The location of the hands are important for this purpose, and we extract a bounding box around the hands.

We detect hands at the pixel level with [9], 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. The bounding boxes are centered around the hand contour such that the top most pixel of the contour is at the top center position of the box. Regions are also adjusted so that they never exceed the image boundaries.



Hand Contour descriptor

Masked HOG descriptor

Figure 5. Visualization of the hand contour and masked HOG feature descriptor used to represent candidate regions.

We use a fixed size bounding box region of 350×160 pixels. We selected the size heuristically by observing a few qualitative examples. In practice, we found it to encompass various objects in the environment (cups, plates, utensils) well [1]. The box is also wide enough to capture two hands interacting with an object (*e.g.*, peeling an onion with two hands). Examples of harvested regions are given in Fig. 4.

3.2. Representing Hand-Object Interactions

Given an incoming stream of candidate hand-object regions we would like to group similar interactions into the same cluster. Before we can proceed to group the regions, we are faced with the challenge of representation.

We use a large HOG template generated only for a masked region (Masked HOG), inspired by work in object discovery [36]. This representation removes the effect of the background (*i.e.* non-hand regions, including interactive objects) and uses only the contour of the hand to group the regions. We use the following HOG template parameters: 8×8 cell, 8×8 stride, 16×16 blocks with 9 gradient orientation bins (see Fig. 5) [2].

3.3. Grouping Hand-Object Interactions

Now that we have a means of detecting and representing hand-object interactions, we can proceed to cluster them to discover common modes. 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. The Determinantal Point Process can be used 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

[1] We compare this harvesting strategy with five other different methods in the supplementary material.

[2] We compare the Masked HOG descriptor to five other different feature representations including SIFT-BOW and HOF in supplementary material.

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). In the first stage, our algorithm generates diverse candidate cluster centers by fast Determinantal Point Process sampling [35], and in the second stage, it assigns all data instances to each cluster center (or none).

3.3.1 Determinantal Point Process

Given an dataset $D = \{d_1, \ldots, d_M\}$, and a similarity matrix $S \in \mathbb{R}^{M \times M}$ describing the pairwise similarity $S_{ij} = s(d_i, d_j)$ between data items, a DPP defines a probability distribution P_S over the set of all subsets of D. Let $\mathbf{Y}$ be the set of all subsets, P_S is of the form:

$$P_S(\mathbf{Y} = Y) = \frac{\det(S_Y)}{\sum_{Y' \subseteq D} \det(S_{Y'})} = \frac{\det(S_Y)}{\det(S + I)}, \quad (1)$$

where I is the identity matrix, and $S_Y \equiv [S_{ij}]_{i,j \in Y}$ denotes the restriction of S to the entries indexed by elements of Y. It is known that this distribution assign more probability to subsets with larger diversity. A more comprehensive survey of the DPP can be found in [45].

3.3.2 Fast DPP Sampling

The overall idea of fast DPP sampling is to design a rapidly-mixing Markov chain whose stationary distribution is P_S by the Metropolis-Hastings algorithm. In this case, the state space of this Markov chain consists of all possible configurations. In particular, if only the transitions between adjacent states are considered, then the transition is either the insertion or deletion of a single element. The transition probability of insertion is defined as:

$$Pr(Y \to Y \cup \{u\}) = \min\left\{1, \frac{\det(S_{Y \cup \{u\}})}{\det(S_Y)}\right\}, \quad (2)$$

and the deletion probability is defined similarly. It has been proven that if we sample u uniformly from D, then this Markov chain has a stationary distribution P_S.

A critical insight made by Kang [35] is that the determinant ratio can be computed by:

$$\frac{\det(S_{Y \cup \{u\}})}{\det(S_Y)} = \frac{\det(S_Y)(c_u - b_u^\top S_Y^{-1} b_u)}{\det(S_Y)} = c_u - b_u^\top S_Y^{-1} b_u, \quad (3)$$

where $c_u = s(u, u)$ is the self similarity, and $b_u = [s(y_i, u)]_{y_i \in Y}$ is a vector of u's similarity to elements of Y. In this case, the determinant ratio can be computed efficiently (an order faster) by incrementally updating $S_{Y \cup \{u\}}^{-1}$ from S_Y^{-1}. More details can be found in [35].

3.3.3 DPP-based Online Clustering

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 in Section 3.3.2. 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.

Algorithm 1
GetClusterCenters(X)

$Y = \{x_1\}$
for $i = 2 : \tau : |X|$ **do**
  $u \leftarrow x_i$
  $p \leftarrow \min\left\{1, c_u - b_u^\top S_Y^{-1} b_u\right\}$
  $Y \leftarrow Y \cup u$ with prob. p
  Update S_Y^{-1} if necessary
end for
return Y

Algorithm 2
TwoPassClustering(X, θ)

Y = GetClusterCenters(X)
$Z = \{-1, \ldots, -1\}$
for $x_i \in X$ **do**
  $k = \arg\min_j d(x_i, y_j)$
  if $d(x_i, y_k) \leq \theta$ **then**
    $z_i = k$
  end if
end for
return Y, Z

We adapt the procedure in Section 3.3.2 to our streaming setting, in which the new data u comes in sequentially from a stream of $X = \{x_1, \ldots, x_M\}$. We assume that the length of the stream M is much larger than the number of candidate cluster center N, and thus the probability of selecting u from the current candidate cluster centers Y is negligible. In other words, we only consider the insertion of x_i from a data stream X to the set of candidate cluster centers Y.

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. Our clustering approach is summarized in Algorithms 1 and 2.

3.4. Learning a Grasp Taxonomy

Up to now, we have only considered each type of grasp or hand-object interaction separately. However, it seems reasonable that there exists a higher-order relationship between each grasp type. For example, the way we grasp a baseball is different from how we grasp a water bottle, but they are similar when compared to the way we grasp the pen when writing. This higher-order relationship between grasps can be specified by a grasp taxonomy. As state before, grasp taxonomies have been instrumental in guiding robotic hand design as it concisely describes the space of possible grasps [11].

While Cutkosky and Wrights grasp taxonomy [11] has had an important role in the design of robotic manipulation, the original taxonomy was not designed to be comprehensive. The taxonomy was task-based and learned from a machinst. Therefore, there are numerous everyday hand interactions that are not included in the taxonomy. With our framework it is also possible to automatically discover novel grasp concepts outside of Cutokosky's taxonomy. Furthermore, our approach also facilitates the potential to learn the higher-order relationships or taxonomy of our discovered grasps. One possible approach to learn a heirarchical structure is to use agglomerative clustering on top of the

Algorithm 3 GetHierarchyCenters(X)

$L \leftarrow 1, Y_1 \leftarrow \{x_1\}$
for $i = 2 : \tau : |X|$ **do**
  $\ell \leftarrow 1$
  while $\ell \leq L$ **do**
    $Y_\ell \rightarrow Y_\ell \cup x_i$ with prob. p_ℓ^+
    if $Y_\ell \rightarrow Y_\ell \cup x_i$ **then**
      $\ell \leftarrow \ell + 1$
    else
      break
    end if
  end while
  if $\ell == L$ **then**
    $L \leftarrow L + 1, Y_L \leftarrow \{x_i\}$
  end if
end for
return $\{Y_\ell\}$

discovered cluster centers [58]. Instead, 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 $p_\ell^+(u) = Pr(Y_\ell \rightarrow Y_\ell \cup u)$ is defined by equation (2). Here Y_1 corresponds to the cluster centers discovered in Section 3.3.3. Our taxonomy learning approach is summarized in Algorithm 3. 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.

4. Experiments

We evaluate the effectiveness of our approach using egocentric videos from three datasets: (1) the Yale Human Grasping (YHG) dataset [5], (2) the University of Tokyo Grasp (UTG) dataset [7], and (3) the Georgia Tech Egocentric Activities GTEA with Gaze (GTEA+) dataset [22]. The first two datasets contain grasp type labels for the videos, and the GTEA+ dataset is constructed for action recognition without grasp type labels. For experiments, we will first compare different online clustering algorithms quantitatively using labeled datasets. Then we will show the discovered novel grasp concepts and taxonomy from the GTEA+ dataset, which contains numerous everyday grasps that are not included in standard grasp taxonomies. Finally, we evaluate the taxonomy tree learned by the proposed hierarchical DPP clustering.

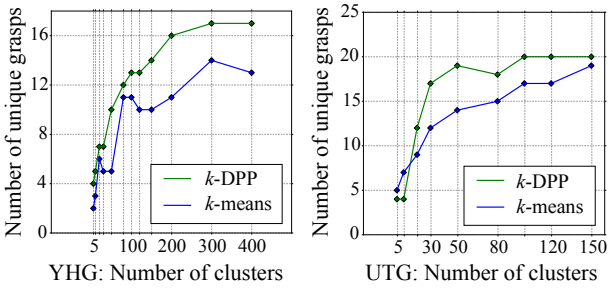


Figure 6. Number of unique grasp types against the number of clusters. DPP discovers unique grasp types more effectively.

4.1. Evaluating Grasp Cluster Diversity

A known property of the DPP is that a sampled subset will be diverse. This feature of the DPP has been proved beneficial in several areas of computer vision and natural language processing [27, 44, 58]. First we confirm that our online DPP clustering can indeed capture the high diversity of human grasps. We use the two labeled datasets (YHG and UTG) to measure performance. The full YHG [5] dataset contains 27.7 hours of tagged video and represents a wide range of manipulative behaviors spanning much of the typical human hand usage [6]. A spreadsheet of the tagged grasp type, object and task parameters, and time information for each successive grasp is also provided. We only use a subset of the first 20 egocentric videos in our experiments, in which the grasp types are well-labeled. The UTG [7] dataset contains 20 grasp types (Fig. 1) based on the statistical result of grasp prevalence provided in [6]. The egocentric videos were recorded by a HD head mounted camera (GoPro Hero2) under controlled environment while subjects performed each grasp type with varying hand poses. In our experiments, we merge the following grasps categories that are distinguished only by finger tip contact or slight changes in apperture: *Thumb-#-Finger* grasps and *Parallel Extension* are merged to *Precision-Thumb-Finger*; *Lateral Tripod* is merged with *Tripod*; *Small Diameter* and *Large Diameter* are merged to *Diameter*.

Although each algorithm is able to determine its own number of clusters, for the diversity experiment, we gradually change (and set) the number of clusters K and count the number of unique grasps that are recovered. We assign each cluster to a canonical grasp type by the majority rule. Higher number of unique grasp types means that the discovered clusters are more diverse with respect to the true grasp categories. In order to control the resulting number of clusters, the insertion/deletion probability of DPP (2) can be

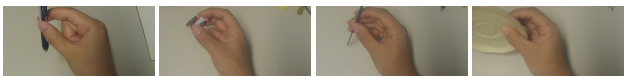


Figure 7. Cluster of Precision-Thumb-Finger grasps observed multiple times in the UTG dataset.

extended to the replacement probability:

$$Pr(Y \to Z \cup \{v\}) = \frac{\det(S_{Z\cup\{v\}})}{\det(S_Y)}, \tag{4}$$

where $Z = Y \backslash \{u\}$. We select the online k-means [2] for comparison. We initialize both the algorithms by uniformly sampling the first K point in the data stream, and update centers iteratively. The results of the YHG and UTG datasets are shown in Figure 6. It is observed that the proposed k-DPP algorithm has better coverage (more diversity) with the same number of clusters when compared to the online k-means algorithm. For example, with 50 discovered clusters, the k-DPP algorithm learns 10 types of canonical grasps while the k-means algorithm only learns 5 canonical grasp types (i.e., many clusters are redundant). This lack-of-diversity property of k-means is also discussed in [33]. Our experiment reinforces the claim that online DPP clustering obtains a more diverse set of clusters when compared to online k-means.

4.2. Evaluating Clustering Quality

Now that we have empirically verified that DPP-based clustering yields more diverse clusters, we want to know if diversity is really beneficial to our online clustering problem. Again, we compare the performances on the labeled YHG [5] and UTG [7] datasets.

We use three standard metrics in [51] to evaluate the clustering quality. The first one is *purity*, which is the classification accuracy. To compute purity, each cluster is assigned to the most frequent grasp label in the cluster, and then the purity is defined by the number of correctly assigned data divided by the total number of data. High purity is easy to achieve with large number of clusters. A trivial case is when each data gets its own cluster, then purity is 1. Other metrics described below will help to analyze results from other perspectives.

This second metric is *normalized mutual information* (NMI), which measures the mutual dependence of the label and the cluster assignment. The minimum of the NMI is 0 if the clustering assignment is independent to the grasp label and 1 if they are perfectly aligned. NMI is defined by the mutual information normalized by the average entropy of label distribution and cluster assignment distribution:

$$NMI(\mathbb{C}, \mathbb{G}) = \frac{I(\mathbb{C}; \mathbb{G})}{[H(\mathbb{C}) + H(\mathbb{G})]/2}, \tag{5}$$

where $\mathbb{C} = \{c_1, \ldots, c_K\}$ is the set of clusters, $\mathbb{G} = \{g_1, \ldots, g_J\}$ is the set of grasp labels, I is the mutual information, and H is the entropy. The mutual information $I(\mathbb{C}; \mathbb{G})$ is defined by:

$$I(\mathbb{C}; \mathbb{G}) = \sum_k \sum_j P(c_k \cap g_j) \log \frac{P(c_k \cap g_j)}{P(c_k)P(g_j)}. \tag{6}$$

High NMI is also easy to achieve when the number of clusters is small - in particular, NMI is 1 if all data are assigned to a single cluster, which is the opposite of purity.

The third metric is based on an alternative information-theoretic view of clustering by interpreting it as a series of decisions, one for each pair of data points in the dataset. When two data points of the same label belong to the same discovered cluster, it is counted as a single true positive. When two data points with different labels belong to different discovered clusters, it is counted as a single true negative. False positive and false negative are defined similarly.

We select three baselines for comparative evaluation:

Leader-Follower clustering. The first baseline is the leader-follower (L-F) algorithm [15], which is a variant of the well-known Lloyd's algorithm to the online setting. It replaces the need to specify the number of clusters by using a sensitivity threshold. The centers are updated by a winner-take-all strategy, where only the nearest centroid to the new data point is updated.

NRP clustering. The second baseline is based on the codebook generation algorithm in [33], which assumes streaming data that arrives in batches of size n. At each iteration, data points belonging to preexisting clusters are first removed. Then a mean-shift procedure is used to discover new cluster centers from the remaining points. In our setting, data arrives one at a time ($n = 1$) so we require no mean-shift procedure – the point either belongs to a preexisting cluster or it becomes the center of a new cluster. This algorithm looks for *local diversity* by removing data points that are already members of existing clusters. In contrast, our proposed DPP method maximizes *global diversity* over the entire set of clusters. We will refer to this simplified algorithm as nearest representive point clustering (NRPC).

Dirichlet Process clustering. The third baseline is the fast Dirichlet Process Mixture (DPM) model inference [69], which has been successfully applied to unsupervised learning of ego-action categories [42]. The nonparametric nature of DPM makes it an ideal candidate for clustering problem with unknown number of clusters, and the computational issue for our large dataset is solved by the online inference framework proposed by Wang and Dunson [69]. For details of DPM and online inference, please refer to [42, 69].

Comparative results for YHG and UTG are shown in Table 1. The proposed DPP clustering performs best for most metrics and is able to find almost all grasp types. The results for **L-F** and **DPM** do not have high purity because the centers can be significantly distorted by outliers [33]. Without the DPP diversity prior or Dirichlet Process prior to control the number of clusters, both **NRPC** and **L-F** yield large numbers of clusters. Interestingly, the purity of DPP is still

Table 1. Comparative Analysis for YHG and UTG

(a) YHG	purity	NMI	P	R	F_1	unique	clusters
L-F [15]	0.630	0.095	0.353	0.079	0.129	14	52
NRP [33]	**0.874**	0.138	**0.677**	0.093	0.164	14	20
DPM [69]	0.512	0.092	0.233	**0.260**	0.246	10	17
Ours	0.836	**0.146**	0.655	0.206	**0.313**	14	13
(a) UTG	Purity	NMI	P	R	F_1	unique	clusters
L-F [15]	0.438	0.088	0.097	0.100	0.099	13	22
NRP [33]	0.535	0.100	0.188	0.046	0.074	13	41
DPM [69]	0.271	0.077	0.043	**0.196**	0.070	7	7
Ours	**0.535**	**0.122**	**0.224**	0.187	**0.204**	12	11

high for both datasets without large number of clusters. Despite a much smaller number of clusters, almost all grasps are discovered (12 out of 13 grasps for UTG; all grasps for YHG). This result shows the advantage of the DPP for effectively discovering hand-object interactions.

4.3. Discovering Novel Hand-object Interaction

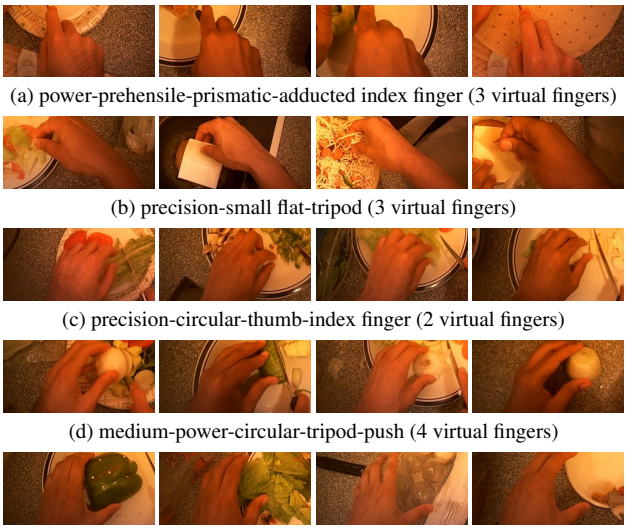
In the previous section, we have shown that the proposed DPP online clustering is the best for finding clusters that are consistent with the grasp labels. We now apply it to daily life egocentric videos. The goal is to automatically discover everyday grasps that may not be included in standard taxonomies [11, 6].

We use the publicly available egocentric activities dataset of Fathi *et al*. [22]. It consists of first-person videos captured by 9 users preparing 8 different dishes (*i.e*. American breakfast, hamburger, Greek salad, pasta, pizza, snack, Tilapia, turkey sandwich) in a kitchen environment for a total of 40 videos. The videos range from 7 to 18 minutes depending on recipe, which amounts to a total of about 0.8 million frames and 7 hours of video. This data is particularly well-suited for our task since egocentric videos contain many naturally occurring interactions with typical kitchen countertop objects (sample images in Fig. 2).

We show examples of discovered interactions in Fig. 8. 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.

4.4. Extension to Taxonomy Tree Learning

Understanding the relationship between grasps types is another dimension across which we can better understand human hand use. Here we evaluate our proposed hierarchical DPP-based clustering by comparing our learned grasp taxonomy to Cutkosky's grasp taxonomy [11].



(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)

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

Figure 8. 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.

The visual grasps taxonomy learned by our approach is given in Fig. 9. Each colored box in Fig. 9 corresponds to an equivalent subtree in Cutkosky's taxonomy. For example, the red box, consisting of power disk and power sphere, corresponds to the *power circular grasp*; the blue box, composed of precision thumb-index finger, precision thumb-3 finger, and precision thumb-4 finger, represents the *precision prismatic grasp*; the green box, including the power light tool and the power medium wrap, refers to the *power prismatic light grasp*. These qualitative examples show that the grasp relationships discovered by our approach correspond well to parts of Cutkosky's taxonomy

To enable a more quantitative analysis for comparing tree structures, we propose a new metric called weighted minimum coverage (WMC) score. The WMC consists of two terms: a tree distance score and a weight factor.

$$\text{WMC}(T, T_{ref}) = \sum_{n_A, n_B \in T_{ref}} w_{AB} \min_{n_i \in A, n_j \in B} |d(n_i, n_j) - d_{ref}(n_A, n_B)|, \tag{7}$$

where n_i and n_j are nodes in the tree T, n_A and n_B are nodes in the reference tree T_{ref}, A and B are labels of n_A and n_B, w_{AB} is the weight for labels A and B, and $d(\cdot, \cdot)$ is the Lowest Common Ancestor [14] distance between two nodes. In our case, the reference tree T_{ref} is the Cutkosky's tree taxonomy, and the labels A and B are grasp labels in the taxonomy. By $n_i \in A$ we mean grasp n_i has grasp label A. The weight w_{AB} is introduced to penalize the size of the

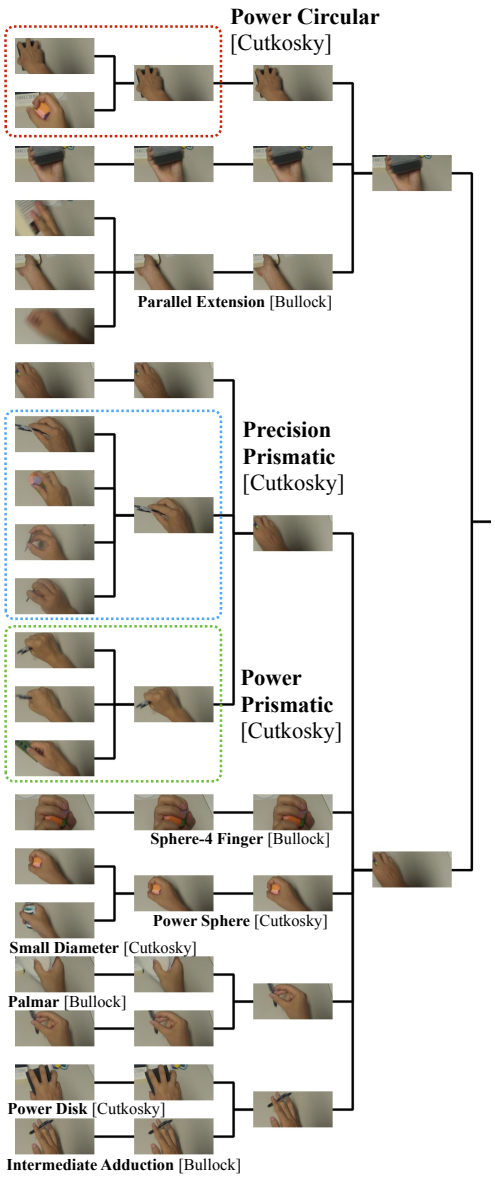


Figure 9. Automatically learned taxonomy tree on the UTG dataset by our DPP-based hierarchical clustering.

sets $\{n_i \in A\}$ and $\{n_j \in B\}$. If the size of these two sets are large, we can find the nodes n_i and n_j that minimize the tree distance score. Therefore, the weight is defined as:

$$w_{AB} = \frac{|\{n_i \in A\}||\{n_j \in B\}|}{|\mathbf{C}|}$$

$$\mathbf{C} = \left\{ \underset{n_i \in A, n_j \in B}{\arg\min} \; |d(n_i, n_j) - d_{ref}(n_A, n_B)| \right\}, \quad (8)$$

where $\mathbf{C}$ is the set of pairs n_i and n_j that minimize the tree distance score. Note that w_{AB} is zero if either $|\{n_i \in A\}|$ or $|\{n_j \in B\}|$ is zero.

Our WMC metric is necessary since traditional tree distance metrics [4, 25] cannot compare tree structures with redundant terminals or non-common terminal nodes .

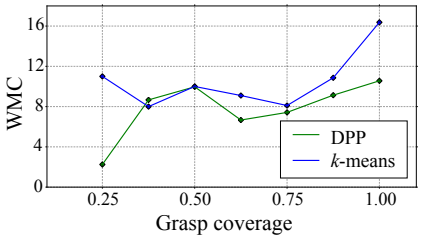


Figure 10. Distance to the Cutkosky's taxonomy tree of the proposed online DPP hierarchical clustering and k-means based hierarchical clustering. The proposed DPP tree learning is more consistent with the standard taxonomy.

Using the WMC score we evaluate the stability of our learned trees over various DPP bandwidth parameters. We also evaluate the performance of a hierarchical k-means algorithm to provide a comparative baseline[3]. We vary the kernel bandwidth of DPP and radius of online k-means to get various trees with different grasp coverage. We define the *grasp coverage* ratio as the number of unique grasps in the tree divided by the total number of grasps defined by Cutkosky's taxonomy. The *grasp coverage* ratio acts as a common axis along which to compare the two algorithms.

The WMC distance with respect to Cutkosky's taxonomy are plotted against the grasp coverage ratio in Fig. 10. As expected the WMC score increases as we increase the grasp coverage ratio for both models. The important observation is that our approach consistently has similar or better (lower) WMC distance.

5. Conclusion

We have presented a new DPP-based online clustering algorithm for discovering diverse and dominant modes of hand-object interaction through the effective use of the egocentric videos. Through extensive experiments, we have showed the effectiveness of the proposed approach in terms of both diversity and standard clustering evaluation metrics. Our evaluation of the discovered interactions shows that we are able to learn a wider range of hand-object interactions in comparison to Cutkosky's classical hand grasps taxonomy. Furthermore, we propose a DPP-based hierarchical clustering framework to automatically learn the higher-order structure or taxonomy of the discovered modes of hand-object interactions. The learned taxonomy is validated both qualitatively and quantitatively consistent with standard hand grasps taxonomy. We believe that this work has taken a solid step forward in data-driven methods for analyzing hand-object interaction and hope that it will lead to significant cross-disciplinary impact on prehensile analysis.

Acknowledgement. This research was supported in part by the JST CREST grant.

[3]Details are included in the supplementary material.

References

[1] N. Ailon, R. Jaiswal, and C. Monteleoni. Streaming k-means approximation. In *NIPS*, 2009.

[2] W. Barbakh and C. Fyfe. Online clustering algorithms. *International Journal of Neural Systems*, 18(03):185–194, 2008.

[3] K. Bernardin, K. Ogawara, K. Ikeuchi, and R. Dillmann. A sensor fusion approach for recognizing continuous human grasping sequences using hidden markov models. *IEEE Transactions on Robotics*, 21(1):47–57, 2005.

[4] P. Bille. A survey on tree edit distance and related problems. *Theoretical computer science*, 337(1):217–239, 2005.

[5] I. Bullock, T. Feix, and A. Dollar. The yale human grasping data set: Grasp, object and task data in household and machine shop environments.

[6] I. M. Bullock, T. Feix, and A. M. Dollar. Finding small, versatile sets of human grasps to span common objects. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 1068–1075. IEEE, 2013.

[7] M. Cai, K. M. Kitani, and Y. Sato. A scalable approach for understanding the visual structures of hand grasps. In *ICRA*, 2015.

[8] J. Case-Smith and C. Pehoski. *Development of hand skills in children*. American Occupational Therapy Association, 1992.

[9] L. Cheng and K. M. Kitani. Pixel-level hand detection in ego-centric videos. In *CVPR*, 2013.

[10] A. Choromanska and C. Monteleoni. Online clustering with experts. In *AISTATS*, 2012.

[11] M. Cutkosky. On grasp choice, grasp models, and design of hands for manufacturing tasks. *Trans. on Robotics and Automation*, 5(3):269–279, 1989.

[12] V. Delaitre, J. Sivic, and I. Laptev. Learning person-object interactions for action recognition in still images. In *NIPS*, 2011.

[13] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for static human-object interactions. In *CVPR Workshops*, 2010.

[14] H. N. Djidjev, G. E. Pantziou, and C. D. Zaroliagis. Computing shortest paths and distances in planar graphs. In *Automata, Languages and Programming*, pages 327–338. Springer, 1991.

[15] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. John Wiley & Sons,, 1999.

[16] S. J. Edwards, D. J. Buckland, and J. McCoy-Powlen. *Developmental & functional hand grasps*. Slack Thorofare, 2002.

[17] J. M. Elliott and K. Connolly. A classification of manipulative hand movements. *Developmental Medicine & Child Neurology*, 26(3):283–296, 1984.

[18] C. E. Exner. Development of hand skills. *Occupational therapy for children*, 5:304–355, 2001.

[19] A. Faktor and M. Irani. Clustering by composition–unsupervised discovery of image categories. In *ECCV 2012*, 2012.

[20] A. Fathi, A. Farhadi, and J. Rehg. Understanding egocentric activities. In *ICCV*, 2011.

[21] A. Fathi, J. Hodgins, and J. Rehg. Social interactions: A first-person perspective. In *CVPR*, 2012.

[22] A. Fathi, Y. Li, and J. M. Rehg. Learning to recognize daily actions using gaze. In *ECCV*, 2012.

[23] R. Filipovych and E. Ribeiro. Recognizing primitive interactions by exploring actor-object states. In *CVPR*, 2008.

[24] D. Fouhey, V. Delaitre, A. Gupta, A. Efros, L. I., and S. J. People watching: Human actions as a cue for single view geometry. In *ECCV*, 2012.

[25] E. B. Fowlkes and C. L. Mallows. A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, 78(383):553–569, 1983.

[26] J. Gall, A. Fossati, and L. van Gool. Functional categorization of objects using real-time markerless motion capture. In *CVPR*, 2011.

[27] J. Gillenwater, A. Kulesza, and B. Taskar. Discovering diverse and salient threads in document collections. In *EMNLP*, 2012.

[28] H. Griffiths. Treatment of the injured workman. *The Lancet*, 241(6250):729–733, 1943.

[29] S. Guha, N. Mishra, R. Motwani, and L. O'Callaghan. Clustering data streams. In *Proceedings of the Annual Symposium on Foundations of Computer Science*, 2000.

[30] A. Gupta, A. Kembhavi, and L. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *PAMI*, 31(10):1775–1789, 2009.

[31] A. Gupta, S. Satkin, A. A. Efros, and M. Hebert. From 3d scene geometry to human workspace. In *CVPR*, 2011.

[32] N. Jojic, A. Perina, and V. Murino. Structural epitome: A way to summarize one?s visual experience. In *NIPS*, 2010.

[33] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *ICCV*, 2005.

[34] N. Kamakura, M. Matsuo, H. Ishii, F. Mitsuboshi, and Y. Miura. Patterns of static prehension in normal hands. *The American journal of occupational therapy*, 34(7):437–445, 1980.

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

[36] H. Kang, M. Hebert, and T. Kanade. Discovering object instances from scenes of daily living. In *ICCV*, 2011.

[37] S. B. Kang and K. Ikeuchi. A framework for recognizing grasps. Technical Report CMU-RI-TR-91-24, Robotics Institute, 1991.

[38] S. B. Kang and K. Ikeuchi. Grasp recognition using the contact web. In *IROS*, 1992.

[39] S. B. Kang and K. Ikeuchi. A grasp abstraction hierarchy for recognition of grasping tasks from observation. In *IROS*, 1993.

[40] I. A. Kapandji. *Funktionelle Anatomie der Gelenke*. Hippokrates-Verlag, 2001.

[41] A. D. Keller. *Studies to determine the functional requirements for hand and arm prosthesis*. Department of Engineering University of California, 1947.

[42] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *CVPR*, 2011.

[43] H. Kjellström, J. Romero, and D. Kragić. Visual object-action recognition: Inferring object affordances from human demonstration. *CVIU*, 115(1):81–90, 2011.

[44] A. Kulesza and B. Taskar. Structured determinantal point processes. In *NIPS*, 2010.

[45] A. Kulesza and B. Taskar. Determinantal point processes for machine learning. *arXiv preprint arXiv:1207.6083*, 2012.

[46] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012.

[47] Y. J. Lee and K. Grauman. Learning the easy things first: Self-paced visual category discovery. In *CVPR*, 2011.

[48] Y. Li, A. Fathi, and J. M. Rehg. Learning to predict gaze in egocentric video. In *ICCV*, 2013.

[49] G. Lister. *The hand: Diagnosis and surgical indications*. Churchill Livingstone, 1977.

[50] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *CVPR*, 2013.

[51] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[52] E. D. McBride. *Disability evaluation and principles of treatment of compensable injuries*, volume 1936. Lippincott, 1963.

[53] T. McCandless and K. Grauman. Object-centric spatio-temporal pyramids for egocentric activity recognition. In *BMVC*, 2013.

[54] J. R. Napier. The prehensile movements of the human hand. *Journal of bone and Joint surgery*, 38(4):902–913, 1956.

[55] A. Pieropan, C. Henrik, and H. Kjellström. Functional object descriptors for human activity modeling. In *ICRA*, 2013.

[56] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012.

[57] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning object class detectors from weakly annotated video. In *CVPR*, 2012.

[58] R. Reichart and A. Korhonen. Improved lexical acquisition through dpp-based verb clustering. In *ACL*, 2013.

[59] M. S. Ryoo, S. Choi, J. H. Joung, J.-Y. Lee, and W. Yu. Personal driving diary: Automated recognition of driving events from first-person videos. In *CVIU*, 2013.

[60] M. S. Ryoo and L. Matthies. First-person activity recognition: What are they doing to me? In *CVPR*, 2013.

[61] R. M. Sanso and D. Thalmann. A hand control and automatic grasping system for synthetic actors. In *Computer Graphics Forum*, volume 13, pages 167–177. Wiley Online Library, 1994.

[62] M. Santello, M. Flanders, and J. F. Soechting. Postural hand synergies for tool use. *The Journal of Neuroscience*, 18(23):10105–10115, 1998.

[63] I. G. Schlesinger. Der mechanische aufbau der künstlichen glieder. In *Ersatzglieder und Arbeitshilfen*, pages 321–661. Springer, 1919.

[64] S. Singh, A. Gupta, and A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012.

[65] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering objects and their location in images. In *ICCV*, 2005.

[66] E. Spriggs, F. De La Torre, and M. Hebert. Temporal segmentation and activity classification from first-person sensing. In *CVPR Workshop*, 2009.

[67] M. Stark, P. Lies, M. Zillich, J. Wyatt, and B. Schiele. Functional object class detection based on learned affordance cues. In *Computer Vision Systems*, pages 435–444. 2008.

[68] M. W. Turek, A. Hoogs, and R. Collins. Unsupervised learning of functional categories in video scenes. In *ECCV*. 2010.

[69] L. Wang and D. B. Dunson. Fast bayesian inference in dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 20(1):196–216, 2011.

[70] S. L. Wolf, P. A. Catlin, M. Ellis, A. L. Archer, B. Morgan, and A. Piacentino. Assessing wolf motor function test as outcome measure for research in patients after stroke. *Stroke*, 32(7):1635–1639, 2001.

[71] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010.

[72] T. Zhang, R. Ramakrishnan, and M. Livny. Birch: an efficient data clustering method for very large databases. In *ACM SIGMOD*, 1996.