How Do We Use Our Hands?
Discovering a Diverse Set of Common Grasps
(Supplementary Material)

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

In this supplementary material, we compare different hand regions harvesting techniques and feature representations. Additional clustering results and some pseudocodes of baselines in the paper are also included at the end.

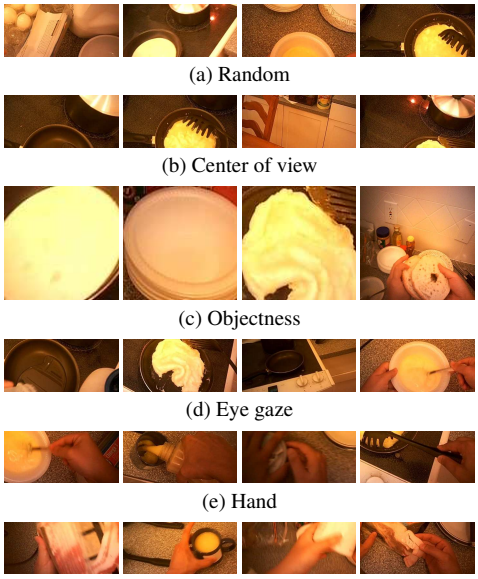
1. Methods for Comparison

Our goal is to discover the dominant modes of hand-object interactions from first-person videos. The approach in our paper has four main components: (1) harvesting candidate hand-object regions, (2) extracting features from the candidate regions, (3) grouping the candidate regions to discover modes of hand-object interactions, and (4) modeling the temporal dynamics to capture higher-order relationships between sequences of candidate hand-object regions. In the paper, we use hand contour to harvest candidate hand-object regions and masked HOG descriptor as feature representation. In this section, we provide the details of the comparing methods.

1.1. Methods for Harvesting Hand-Object Regions

In order to mine 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. To this end, we evaluate six different harvesting methods:

1. Random sampling: a bounding box region is randomly sampled from every frame.
2. Center focus: a bounding box centered in the image is extracted from every frame.
3. Objectness: high scoring bounding box(es) (confidence over 95%) are extracted using a category-independent object detector [1]. These candidate regions are likely to contain an object of any category.
4. Gaze fixation: a bounding box is extracted when an eye gaze fixation point is detected by an eye tracker.



(a) Random
(b) Center of view
(c) Objectness
(d) Eye gaze
(e) Hand
(f) Hand Eye

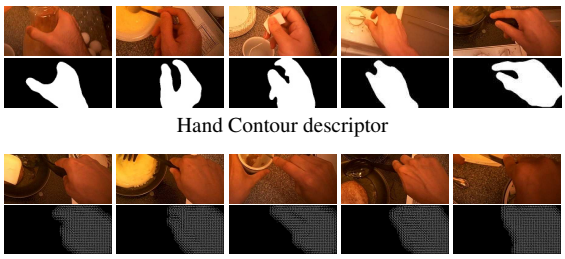
Figure 1. Candidate regions extracted by different methods.

5. Hand contour: a bounding box around the hands is extracted. We detect hands at the pixel level with [2], 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 connect 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 (shifted inward) so that they never exceed the image boundaries.
6. Hand-fixation: a bounding box region is extracted from a hand contour only if there is an eye fixation detected in the potential bounding box.

Methods (2), (4) and (6) are based on an attention-based perspective where we assume that a person's focus of attention will help identify key hand-object regions. Method (3)

*indicates equal contribution

978-1-4673-6964-0/15/$31.00 ©2015 IEEE



Hand Contour descriptor

Masked HOG descriptor

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

is a purely object-centric approach, while methods (5) and (6) take a hand-centric approach.

With the exception of the objectness detector, which generates a bounding box at its own scale, 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. 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 object regions are given in Fig. 1.

Our experimental result (see Table 1) confirms quantitatively our intuition, that the location of the hands are indeed important for learning about hand-object interactions. Based on this result, we use hand locations to harvest hand-object regions in the paper.

1.2. Representations of Hand-Object Interactions

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

We explore six different feature representations to understand which representation best preserves the invariants of a hand-object interaction:

1. A bag-of-words (BOW) over sparsely detected SIFT features is accumulated using a codebook size of 1000 (learned using k-means).
2. A BOW of cuboids [3], a spatio-temporal descriptor often used for action recognition, is generated using spatial interest points [7] and quantized with a codebook size of 2000 (learned using k-means).
3. A global histogram-of-optical-flow (HOF) is used to spatially quantize the motion of the candidate region into a 3×3 grid, with a temporal window size of 3 frames (*i.e.* two temporal bin dimensions) and the optical flow for each bin is quantized into 5 directions.
4. A binary mask of the hand region (Hand Contour) is used to capture the outer shape of the hand (Fig. 2).
5. A large histogram-of-oriented-gradients template (Global HOG) is computed for the entire hand-object region including the hands, objects and background.
6. A large HOG template is generated only for a masked region (Masked HOG), inspired by work in object discovery [5]. 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. 2).

Table 1. Evaluation of candidate hand-object region generation using different bounding box extraction methods.

Harvest Tech.	OP	OM	OU
Hand	**0.967**	**0.911**	**0.956**
Hand eye	0.922	0.867	0.900
Eye gaze	0.838	0.465	0.717
Objectness	0.900	0.456	0.733
Center	0.717	0.131	0.626
Random	0.700	0.156	0.633

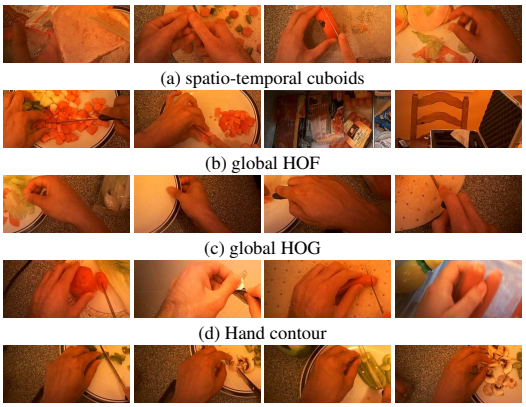
We compared all of the proposed feature representations (Table 2) and found that a joint feature obtained by merging a large gradient histogram template with an extracted hand contour (Masked HOG) is the strongest representation to model the functional proximity between candidate hand-object interactions. This result is interesting because it implies that the appearance of the hand is more important than the appearance of the object. Based on this result, we use Masked HOG as feature representation in the paper.

2. Experimental Evaluation

We compare harvesting techniques and feature representations using the publicly available ego-centric activities dataset of Fathi *et al.* [4]. 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 ego-centric videos contain many naturally occurring interactions with typical kitchen countertop objects.

2.1. Comparing Harvesting Techniques

We first evaluate the ability that each harvesting approach from Section 1.1 has on generating candidate hand-object interaction regions. Since we cannot exhaustively label all hand-object interaction regions from over 7 hours of video, we instead evaluate on a randomly selected set of 90 candidate regions for each harvesting technique, which serve to represent the larger dataset. We use three metrics: object presence (OP), object uniqueness (OU), and object manipulation (OM) defined as:



(a) spatio-temporal cuboids

(b) global HOF

(c) global HOG

(d) Hand contour

(e) Masked HOG

Figure 3. A hand-object interaction cluster discovered using different feature representations (left most image is the cluster center).

$$\text{OP} = \frac{\text{\# visible objects}}{\text{\# candidate regions}}, \quad (1)$$

$$\text{OU} = \frac{\text{\# unique visible objects}}{\text{\# candidate regions}}, \quad (2)$$

$$\text{OM} = \frac{\text{\# objects being manipulated}}{\text{\# visible object regions}}. \quad (3)$$

The OP rate measures the visibility of an object in the bounding box and is the number of regions that contain an object divided by the total number of candidate regions. The OU rate measures the number of unique objects in a cluster and is defined as the number of unique instances divided by the number of candidate regions. A high OU rate means that a single grasp is associated to a diverse set of objects. The objectness detector, for example, is highly sensitive to plates so the OU rate tends to be low. The OM rate is a measure of visible manipulation and is defined as the percentage of times an object region is captured while it is being manipulated by the camera-wearer's hand(s). This is the most important metric as we would like clusters to contain hand interacting with objects.

Table 1 shows that the hand-based region extraction approach generates the best candidate regions with respect to all three metrics. In particular, the hand-based region extraction has a very high OM rate, as the existence of the hands creates a strong prior over the existence of an object that is undergoing manipulation. Somewhat surprisingly, selecting hand detections that also have eye fixations (Hand-eye) does not help in harvesting the hand-object regions. Upon close inspection, we found that eye fixations do not consistently coincide with hand-object interaction. The other baseline approaches perform much worse since they do not search for regions in which a hand is present. For example, while the objectness detector [1] can find object-like regions (thus having high OP rate), it obtains a low OM rate since our video frames often contain many static objects that are not being manipulated.

Table 2. Comparative performance of different features.

Features	Purity	Diversity	Coverage
Masked HOG	**0.727**	**0.727**	141
Hand contour	0.478	0.444	**264**
Global HOG	0.206	0.206	254
BOW (ST cuboids)	0.133	0.133	65
Global HOF	0.122	0.122	24
BOW (SIFT)	0.091	0.091	82

2.2. Comparing Feature Representations

We next evaluate the features from Section 1.2 used to represent each detected hand-object region, by evaluating cluster quality using hand-object regions harvested by hand locations. As the feature representation will have a large effect on the similarity measure, we would like to understand the strengths and limitations of different image representations when attempting to group the regions. In this experiment, we compare several standard approaches used for object and action recognition to represent and cluster the hand-object interaction object regions. The metrics used for evaluation are:

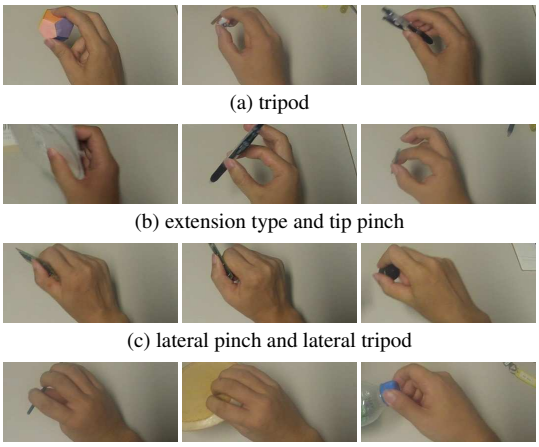
$$\text{purity} = \frac{\text{\# same grasped objects}}{\text{\# candidate regions}}, \quad (4)$$

$$\text{diversity} = \frac{\text{\# unique objects}}{\text{\# visible object regions}}, \quad (5)$$

$$\text{coverage} = \text{\# discovered groups}. \quad (6)$$

Purity and diversity are computed in the following manner. First, each of the clusters are ranked according to the entropy over the video occurrence distribution p(v), $E = -\sum_{m=1}^{M} p(v_m) \log p(v_m)$, where clusters that have object instances for many different videos v are ranked higher and clusters that only contain objects from a single video are ranked lower. The idea is to reward clusters that are likely to contain hand-object interactions of multiple users and objects. Additionally, the instances within a cluster are sorted by their distance to the cluster center. The top 9 instances for the top 10 clusters are scored by human annotation. When each of the objects being grasped are all unique, the purity and diversity score will be equivalent, but when there are near duplicates (same object, same video), only the unique instances are counted to compute the diversity score.

Table 2 shows that the masked HOG features obtains the highest purity and diversity scores. The hand contour feature ranks second showing that the shape of the hand is a strong feature for representing objects in the context of manipulation, which is an expected result. Although the global HOG feature captures both the shape of the hand and commonalities in the shape of interactive objects, we



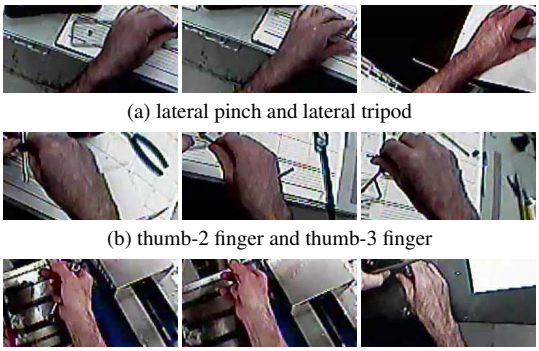
(a) tripod

(b) extension type and tip pinch

(c) lateral pinch and lateral tripod

(d) thumb-4 finger, extension type, lateral tripod

Figure 4. Clustering results of grasps observed multiple times in the UTG dataset.



(a) lateral pinch and lateral tripod

(b) thumb-2 finger and thumb-3 finger

(c) lateral pinch and thumb-3-finger

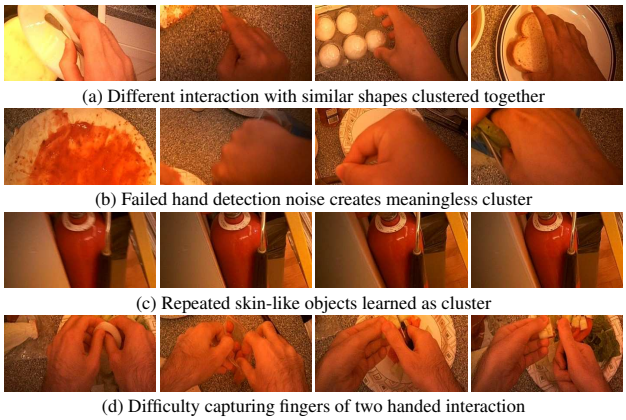
Figure 5. Clustering results of grasps observed multiple times in the YHG dataset.

also found that the global HOG was also greatly affected by background clutter and prevented groups from diversifying across object types. Although the coverage values of the hand contour feature and global HOG feature are high, the low purity values mean that the clusters contain a mix of interactions. The masked HOG feature strikes the best balance between purity and coverage.

Fig. 3 shows examples of regions clustered together using different features. Fig. 3 (c) shows how the shape of the plate or pizza causes dissimilar hand-object interactions to be clustered together. Both the local motion-based spatio-temporal BOWs and the global HOF are the least discriminative as appearance is not explicitly modeled and constant ego-motion prevents the features from modeling subtle motions signatures. Our results confirm that the inclusion of hands as a feature is critical in retaining regions of hand-object interaction.

3. Additional Clustering Results

Figure 4 and Figure 5 shows additional clustering results of UTG and YHG datasets. Some failure cases of



(a) Different interaction with similar shapes clustered together

(b) Failed hand detection noise creates meaningless cluster

(c) Repeated skin-like objects learned as cluster

(d) Difficulty capturing fingers of two handed interaction

Figure 6. Examples of failure cases caused by errors in hand detection and similar hand shapes across multiple interaction.

clustering result on GTEA+ dataset is shown in Figure 6. Figure 7 shows the learned taxonomy tree on the GTEA+ dataset by our DPP-based hierarchical clustering. While the results are affected by meaningless cluster centers caused by failed hand detection, our algorithm is still able to capture Napier's [6] widely adopted 1956 categorizations of grasps into precision and power grasps. For example, the red box, consisting of palmar, index finger extension, lateral pinch and stick, corresponds to the *power grasp*; the blue box, composed of precision thumb-index finger, precision thumb-2 finger, precision thumb-3 finger, and precision thumb-4 finger, represents the *precision grasp*.

4. Baseline Pseudocodes

The pseudocode for the k-means based hierarchical clustering (Section 4.4 in the paper) is shown in Algorithm 1, and pseudocode for nearest representive point clustering (NRP clustering in Section 4.2 of the paper) is shown in Algorithm 2.

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *CVPR*, 2010. 1, 3

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

[3] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VSPETS*, 2005. 2

[4] A. Fathi, Y. Li, and J. M. Rehg. Learning to recognize daily actions using gaze. In *ECCV*, 2012. 2

[5] H. Kang, M. Hebert, and T. Kanade. Discovering object instances from scenes of daily living. In *ICCV*, 2011. 2

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

[7] J. Shi and C. Tomasi. Good features to track. In *CVPR*, 1994. 2

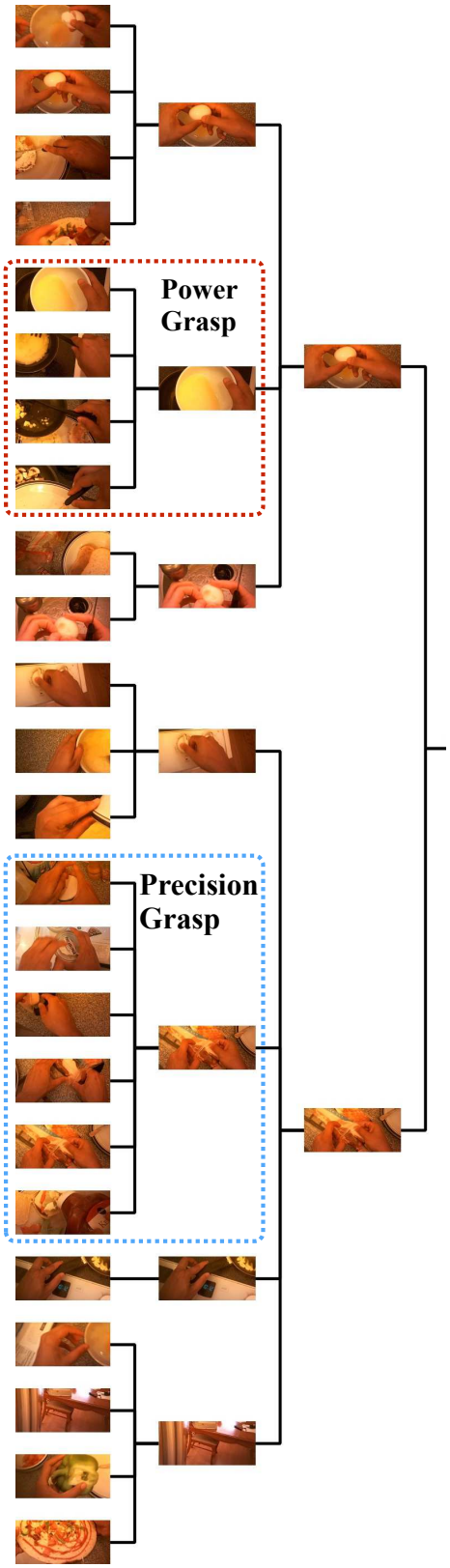


Figure 7. Automatically learned taxonomy tree on the GTEA+ dataset by our DPP-based hierarchical clustering.

Algorithm 1
GetHierarchyCentersKmeans(X)

$L \leftarrow 1, Y_1 \leftarrow \{x_1\}$
for $i = 2 : \tau : |X|$ **do**
  $\ell \leftarrow 1$
  while $\ell \leq L$ **do**
    $y_\ell^* \leftarrow \arg\min_{y \in Y_\ell} d(y, x_i)$
    if $d(y_\ell^*, x_i) > \theta_\ell$ **then**
      $Y_\ell \leftarrow Y_\ell \cup x_i$
      $\ell \leftarrow \ell + 1$
    else
      $n_\ell^* \mathrel{+}= 1$
      $y_\ell^* \mathrel{+}= (x_i - y_\ell^*)/n_\ell^*$
      break
    end if
  end while
  if $\ell == L$ **then**
    $L \leftarrow L + 1, Y_L \leftarrow \{x_i\}$
  end if
end for
return $\{Y_\ell\}$

Algorithm 2
NRPClustering(X)

$Y \leftarrow \{x_1\}$
for $i = 2 : \tau : |X|$ **do**
  $y^* \leftarrow \arg\min_{y \in Y} d(y, x_i)$
  if $d(y^*, x_i) > \theta$ **then**
    $Y \leftarrow Y \cup x_i$
  end if
end for
return Y