An Ego-vision System for Hand Grasp Analysis

Minjie Cai, Kris M. Kitani, and Yoichi Sato

***Abstract*—This paper presents an egocentric vision (ego-vision) system for hand grasp analysis in unstructured environments. Our goal is to automatically recognize hand grasp types and to discover the visual structures of hand grasps using a wearable camera. In the proposed system, free hand-object interactions are recorded from a first-person viewing perspective. State-of-the-art computer vision techniques are used to detect hands and extract hand-based features. A new feature representation which incorporates hand tracking information is also proposed. Then grasp classifiers are trained to discriminate among different grasp types from a pre-defined grasp taxonomy. Based on the trained grasp classifiers, visual structures of hand grasps are learned using an iterative grasp clustering method. In experiments, grasp recognition performance in both laboratory and real-world scenarios are evaluated. The best classification accuracy our system achieves is 92% and 59% respectively. System generality to different tasks and users is also verified by the experiments. Analysis in real-world scenario shows that it is possible to automatically learn intuitive visual grasp structures that are consistent with expert-designed grasp taxonomies.**

***Index Terms*—Hand grasp, wearable system, egocentric vision, recognition.**

I. Introduction

GRASP is commonly defined as every hand postures used for holding an object stably during hand manipulation tasks. Understanding the way how humans grasp object is important in different domains ranging from robotics [1], prosthesis [2], hand rehabilitation [3], to motor control analysis [4] and many others. In robotics, the study of hand function provides critical inspiration for robotic hand development [1]. In rehabilitation, statistical information about daily hand grasp usage is an important factor of the evaluation criterion for injured hand recovery [3].

Traditional approaches to grasp analysis have been developed primarily in controlled laboratory settings which often include hand-contact sensors or calibrated cameras. However, there are many limitations in such structured environments. Intrusive sensors may inhibit free hand-object interactions; calibrated camera system requires hand interactions are recorded in a limited workspace. As a result, hand grasp in real-world environments has seldom been studied.

Our goal is to develop a fully automatic and non-contact system for analyzing hand grasp usage in daily activities. In particular, we propose an ego-vision system for recognizing hand grasp types and learning visual grasp structures using a wearable camera. There are many benefits from the proposed system. First, it overcomes the constraints of other modes of hand sensing by allowing for continuous recording of natural hand activities. Furthermore, it provides an ideal egocentric view for grasp analysis since the hand-object interactions are often visible in the center of the visual field. Most of all, an ego-vision system enables us to study hand grasp in the real-life setting at a large scale that is impossible before.

Our system incorporates advances of computer vision techniques that can be used as a tool to advance studies in prehensile analysis. In particular, we adopt state-of-the-art approaches for egocentric hand detection, in order to deal with the new challenges of egocentric vision such as unconstrained hand movements and rapidly changing imaging conditions (*e.g.*, illumination and background) due to extreme ego-motion. Based on detected hand regions, features are examined and extracted which encode appearance and motion of the hand interactions, and grasp classifiers are trained for discrimination among different grasp types. Finally, the trained grasp classifiers are used to measure the visual similarities between hand grasps and learn an appearance based grasp hierarchy, which we call the *visual structures of hand grasps*. The experiments show that it is possible to learn intuitive visual structures automatically from data which are consistent with an expert-designed grasp taxonomy.

This paper extends our prior work [5] as follows: 1) We extensively evaluate the system performance by examining state-of-the-art feature representation used in object and action recognition. 2) We propose a new feature representation which achieves best classification accuracy and is robust to unreliable hand detection. 3) We greatly expand the UT Grasp Dataset and evaluate the system generality to different tasks and users. 4) We quantitatively evaluate the consistence of the automatically learned grasp structures with expert-designed grasp taxonomies.

The rest of the paper is organized as follows. Section II presents related work. Section III describes the architecture and main components of our ego-vision system. Performance evaluation of the system is shown in Section IV. Section V discusses the system performance and possible extensions. Section VI concludes the paper.

II. Related Work

A. Human Grasp Taxonomy

Grasp taxonomies have been studied for decades to better understand the use of human hands [6][2][7][8][1][9][10]. Early work by Schlesinger [6] classified hand grasps into 6 major categories based on hand shape and object properties. In 1956, Napier [7] proposed a classification scheme for power and precision grasps based on requirements of the manipulation task, which has been widely adopted by researchers

Manuscript received November 23, 2015; revised May 31, 2016 and November 23, 2016; accepted February 26, 2017.

Minjie Cai and Yoichi Sato are with the Institute of Industrial Science, The University of Tokyo, Tokyo, Japan (email: {cai-mj; ysato}@iis.u-tokyo.ac.jp)

Kris M. Kitani is with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA (email: kkitani@cs.cmu.edu)

This research was funded in part by the JST CREST grant.

in the medical, biomechanical and robotic fields. Through observation of manufacturing tasks, Cutkosky provided a comprehensive grasp taxonomy [1] which has played an important role in guiding robotic hand design. Recently, Huang *et al*. [11] proposed an unsupervised method to discover appearance-based grasp taxonomies. In their method, hand images with similar appearance are clustered together as distinct grasp types.

The human grasp taxonomy proposed by Feix *et al*. [10] is the most complete to date as argued and has been widely used in grasp analysis in recent years [12], [13], [14]. Considerable efforts have been devoted in obtaining the statistics of human hand usage based on manual annotation [13][15][16]. However, the annotation process requires many hours of visual inspection by skilled annotators. As it becomes easier to acquire large amounts of video data, it is clear that the manual approach is not scalable to larger datasets. In this work, however, we propose an ego-vision system that is able to support automatic grasp analysis with large amounts of video data.

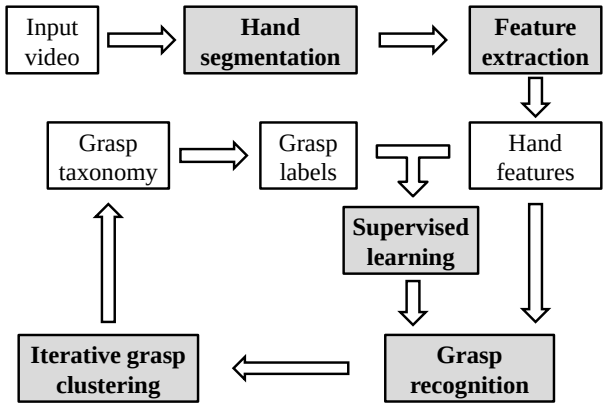


Fig. 1. Outline of the proposed system. The highlighted blocks are the main processing components of the system which will be introduced in Section III. Input video captured from a wearable camera is processed by hand segmentation and feature extraction to extract feature representation for hand images. Ground-truth grasp labels and extracted hand features are used as input of the supervised learning to train grasp classifiers for recognizing different grasp types. Grasp taxonomy is a collection of grasp types which are predefined or generated from iterative grasp clustering.

B. Automated Grasp Analysis

Approaches for automatic hand grasp analysis have been developed primarily in structured environments. Hand tracking devices such as data gloves or inertial sensors have been used to obtain detailed measurements of joint angles and positions of the hand [17][18][19][20]. Santello *et al*. [17] used Principle Component Analysis (PCA) to analyze finger coordination of imagined hand grasp using joint angle data from a data glove. However, the main limitation of hand tracking devices is that they must be worn on the hand and thus inhibit free hand interactions.

Visual sensing of the hands manipulating the objects [21][22][23][24][25] allows a non-contact markerless tracking of hand-object interactions. Romero *et al*. [24] proposed a non-parametric estimation method to track hand poses interacting with objects by performing a nearest neighbor search in a large synthetic dataset. However, most visual tracking systems require that hand interactions are recorded in a structured environment. Yang *et al*. [26] trained a convolutional neural network to classify hand grasp types on unstructured public dataset. However, it only considers a small number of grasp types trained on static hand images. In our work, the proposed system can handle a more complete set of grasp types from real-life hand manipulation tasks.

C. Hand Detection In Egocentric Vision

With the portability and ideal egocentric view provided by wearable cameras, egocentric vision has recently become a popular topic in computer vision community. Li and Kitani [27] first addressed hand detection problem in the context of egocentric video. They proposed a pixel-level hand detection method which can adapt to changing illuminations. Li *et al*. [28] studied the eye-hand coordination in egocentric video and used information from hand detection to predict where the eyes look. Baraldi *et al*. [29] proposed to use dense trajectories with hand segmentation for hand gesture recognition and proved the effectiveness of dense trajectories in egocentric paradigm. Rogez *et al*. [30] recently presented promising results on discrete hand pose recognition from RGB-D data. However, these discrete poses have no direct semantic correspondence to hand grasp types. Our prior work [5] first developed techniques to recognize hand grasp types in everyday hand manipulation tasks recorded with a wearable RGB camera and provided promising results with appearance-based features. Saran *et al*. [31] used detected hand parts as intermediate representation to recognize fine-grained grasp types. The intermediate representation outperforms low-level appearance-based representation when hand parts can be well detected. This work further extends our prior work by incorporating hand tracking information to tackle unreliable hand segmentation in real-world scenario.

III. Grasp Learning System

We aim to automate the hand grasp analysis for daily manipulation tasks. To achieve this goal, we propose an ego-vision system which can recognize different hand grasp types and learn visual grasp structures automatically from large scale of data recorded with a wearable camera. The outline of our system is illustrated in Fig. 1. The input to the system is egocentric video recording daily manipulation tasks. Based on state-of-the-art hand detection techniques we segment hand regions from egocentric videos. Then we extract grasp-related features for training discriminative grasp classifiers. Finally, we use an iterative clustering method to learn visual structures of hand grasps.

A. Hand Segmentation

The detection of hands from egocentric videos is an important pre-processing stage of hand grasp analysis but also a challenging task. In egocentric videos, the background and hand appearance are rapidly changing due to frequent camera

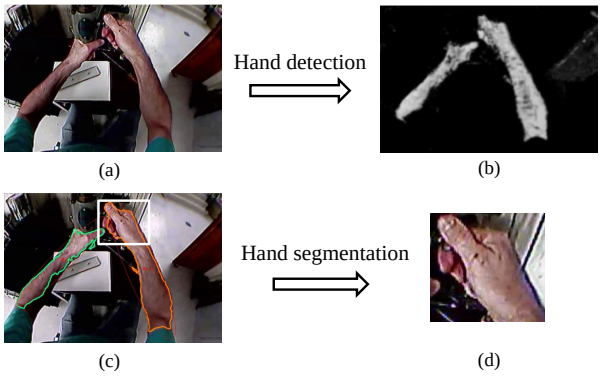


Fig. 2. Example of hand segmentation. (a) Image from egocentric video (b) Pixel-wise hand probability map (c) Candidate hand regions (d) Hand region segmented within a bounding box

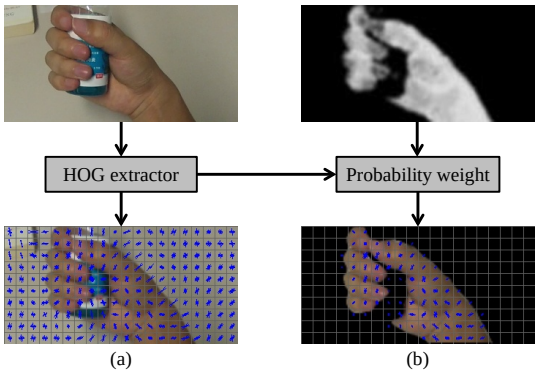


Fig. 3. Visualization of hand-shape related features. (a) Histogram of Oriented Gradient (HoG) (b) Hand probability weighted HoG (HHoG)

motion. Recent work on egocentric hand detection has shown that robust hand detection performance can be achieved if the hand model is adaptable to changes in imaging conditions [27]. Therefore, we train a multi-model hand detector composed by a collection of hand pixel classifiers indexed by global image appearance. Given a test image, the global appearance represented by a color histogram is computed as a visual probe, for every frame, in order to recommend the n-best hand pixel classifiers. Based on the multi-model hand detector, a probability map is generated for each image as illustrated in Fig. 2(b). The value of each pixel represents the likelihood of being a hand pixel in the original image.

Hand regions of a test image are segmented based on the corresponding hand probability map. Candidate hand regions with arms are first obtained by binarizing the probability map with a threshold. Regions under a certain area proportion are discarded and at most two candidate regions are retained. Fig. 2(c) shows two candidate hand regions painted with green and orange contours. In present study we only consider the right handed grasp. The left hand is suppressed by simply selecting the candidate hand region which is right-most. If no hand region is detected, the image is discarded. The hand region is finally segmented with a fixed size bounding box (Fig. 2(d)). To remove the unwanted arm part, ellipse parameters (length of long/short axis, angle) are fitted to the candidate hand region. The arm part is approximately removed by shortening the length of long axis to 1.5 times of the length of short axis. A fixed size bounding box is drawn by fixing the top-center of the bounding box to the top-center of the arm-removed hand region. The size of the bounding box is determined heuristically for each video and takes advantage of the fact that the distance between the hands from the head-mounted camera is consistent throughout the video.

Moreover, a temporal tracking method [32] is utilized to handle the case of two overlapping hands. Briefly speaking, the position and movement of each candidate hand region is stored and used in hand segmentation of the next video frame. Thus two overlapped hands can be separated by using tracking information of each hand before overlapping.

B. Feature Representation

In expert-defined grasp taxonomies, different grasp types are often identified by different hand shapes, object context and types of hand-object interactions. Therefore, we examine and extract features for hand regions addressing different aspects of hand grasp.

1) Hand Shape: Hand shape is represented by Histogram of Oriented Gradient (HoG) [33] computed from a hand region. The HoG feature is an image descriptor based on collected local distribution of intensity gradients and has been widely used in object detection. It is computed by first dividing a hand region into a grid of smaller regions (cells) and then computing histogram of gradient orientations in each cell. Cell histograms within a larger region (blocks) are then accumulated and normalized to make the block descriptor less sensitive to varying illumination. Finally, the resulting block histograms are concatenated to form a HoG feature descriptor. We use a cell size of 8×8 pixels with 9 orientation bins, and a block size of 16×16 pixels. A visualization of example HOG feature is shown in Fig. 3(a).

Two variants of HoG features are examined. The first is the global HoG feature described above. The second is hand probability weighted HoG (HHoG). HHoG effectively suppresses gradients from the background. As shown in Fig. 3(b), HoG features corresponding to non-hand regions are removed by weighting each block histogram with squared hand probability at the center of the block.

2) Visual Context: We extract features from local keypoints in order to capture the visual context of the grasped object. In particular, we extract Scale Invariant Feature Transform (SIFT) [34] for each detected keypoints. Example keypoints are visualized in Fig. 4 where the scale and orientation of each keypoint are illustrated with a green circle and a red radius. Histogram of gradients around each keypoint is computed as a keypoint descriptor. Note that keypoints are detected around the object and the part of the hand in contact with the object. We used a Bag-of-Words (BoW) approach to obtain a feature descriptor which is composed by the frequency of different keypoint patterns. A codebook of 100 keypoint patterns is generated using k-means clustering over all keypoint descriptors.

3) Convolutional Neural Network: Unlike HoG and SIFT which are hand crafted feature representation composed by

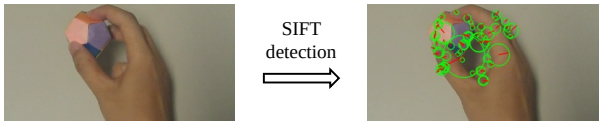


Fig. 4. Visualization of SIFT keypoints. The circle and the line segment starting from the center of the circle indicate the region scale and principle orientation of each keypoint respectively.

orientation histograms, Convolutional Neural Network (CNN) is a biologically inspired hierarchical model which is believed to be able to extract high level feature representation as human brain does. With the advancement of hardware computing capacity and efficient training algorithms, the use of deep and large scale of CNNs becomes feasible and has achieved substantially higher accuracy in different visual recognition domains [35][36][37]. CNN has also been utilized for recognizing grasp types in static images [26] where a five-layer CNN is trained with nearly 5000 image patches. However the amount of labeled data is insufficient for training a large CNN.

In this work we combine a large CNN model pre-trained on a large auxiliary dataset (ImageNet) with domain-specific fine-tuning on a small hand grasp dataset, similar to the work of Girshick *et al*. [38]. Here we are interested in CNN-based feature representation. We extract a middle layer feature output as the feature representation of a hand region by forward propagating the hand region through the trained CNN model.

4) Dense Hand Trajectories: The dense trajectories proposed by Wang *et al*. [39] has been widely used as feature representation for action recognition, and proven to achieve state-of-the-art performance on many video datasets of third person view. To apply it to grasp recognition in egocentric videos, it is important to focus on the regions where hand interactions occur and remove irrelevant information from the background. Motion-based background subtraction doesn't work well in first person video since the background is moving due to camera motion and is hard to reliably estimate and remove the camera motion as illustrated in Fig. 5(c). In this work, we propose a feature representation called "Dense Hand Trajectories (DHT)" which uses hand detection as a spatial prior to extract dense trajectories most related to hand interactions.

We first briefly introduce the extraction of dense trajectories [39] following which the proposed DHT is presented. At each frame, feature points are densely sampled on a grid spaced by 5 pixels at multiple spacial scales. Points in homogeneous area are removed since it is impossible to track them without any structure. Feature points at each spacial scale are tracked separately using a dense optical flow algorithm [40]. Each trajectory is composed by feature points tracked for consecutive frames with trajectory length set to $L = 15$ frames. The main difference of our proposed DHT from [39] is that we use detected hand regions as a spatial prior to weight the trajectories. Specifically, we define a variable H for each tracked trajectory to count the times of passing through the hand regions as illustrated in Fig. 6. At each frame t, a trajectory with a starting feature point sampled within the hand region is initialized with $H = 1$ as indicated by the trajectory

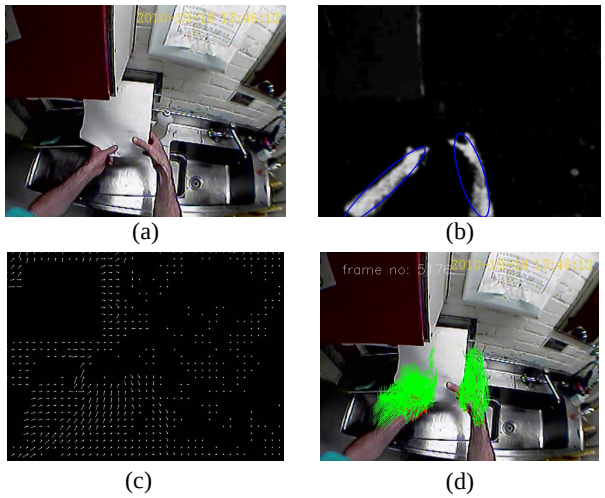


Fig. 5. Example of dense hand trajectories. (a) Image from egocentric video (b) Hand probability map (c) Visualization of optical flow after removing the camera motion (d) Visualization of dense hand trajectories about the hand region

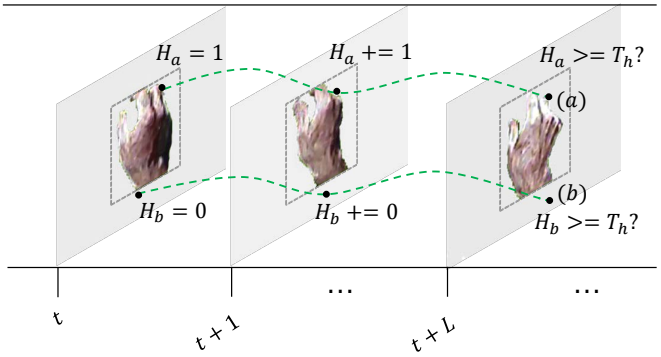


Fig. 6. Illustration of our approach to extracting dense hand trajectories. The detected hand regions are used as spatial prior to weight trajectories which pass through the hand regions. Variable H is used to count the times of being tracked within the hand regions for each trajectory. At the end of tracking (L indicates tracking length), trajectories with H less than a certain threshold T_h are considered as non-hand trajectories and removed.

(a), otherwise is initialized with $H = 0$ as indicated by the trajectory (b). At each subsequent frame during the tracking procedure, H is increased by 1 for all trajectories of which the feature points being tracked are within the hand regions. At the end of tracking, trajectories with H less than a certain threshold T_h are considered as non-hand trajectories and thus removed. In our experiments, we set $T_h = L/2$.

There are two stages of feature extraction based on dense hand trajectories. At the first stage, descriptors are computed for each trajectory. At the second stage, descriptors of trajectories are pooled together and further encoded for each frame. We compute four descriptors same as in [41], which are Displacement, HoG, Histograms of Optical Flow (HOF), and Motion Boundary Histograms (MBH). Length of descriptors are 30 for Displacement, 96 for HOG, 108 for HoF and 192 for MBH. These descriptors contains information of both hand motion and hand appearance in the space-time volume along the trajectory. We use Fisher vector to encode

pooled trajectory descriptors for each frame. Fisher vector has shown performance improvement over bag-of-features for image/video classification in recent researches. For details of Fisher vector encoding, one can refer to [42]. We first use Principal Component Analysis (PCA) to reduce the dimension of each descriptor type to $D = 16$, and randomly sample a subset of $300,000$ descriptors to estimate the Gaussian Mixture Model (GMM) with number of Gaussians set to $K = 256$, as in [42]. The dimension of each descriptor type after Fisher vector encoding is $2DK$. Each frame is represented by concatenation of Fisher vectors of different descriptor types.

C. Grasp Recognition And Abstraction

We train one-versus-all multi-class grasp classifiers for the grasp types defined in Feix's taxonomy [10]. We use this taxonomy since it is one of the most complete one in existence and has been widely applied to hand manipulation analysis [15][16]. Probability calibration [43] is conducted for each classifier in order to produce comparable scores. During testing, each video frame with detected hands is classified independently and assigned with a grasp type of which the classifier outputs the highest score.

We define a correlation index for measuring the visual similarity between different pairs of grasp types based on classification results. The correlation index $C_{i,j}$ between grasp type i and grasp type j is defined as:

$$C_{i,j} = \frac{1}{2}(\frac{m_{i,j}}{n_i} + \frac{m_{j,i}}{n_j}) \quad (1)$$

where $m_{i,j}$ denotes the number of samples from grasp type i misclassified as grasp type j and vice versa. n_i, n_j are the number of samples from grasp type i and grasp type j, respectively.

Based on the correlation index, we implement an iterative grasp clustering algorithm by iteratively clustering two most similar grasp types. The algorithm is described in Algorithm 1. This process automatically learns a dendrogram of grasp types, that is, the visual structures of hand grasps.

Algorithm 1 Iterative Grasp Clustering

Initialize: $N \Leftarrow$ the number of grasp types, consider each grasp type as a single-member grasp cluster
while $N > 1$ **do**
  Step1: Train grasp classifiers for each grasp cluster
  Step2: Perform grasp classification, compute correlation index for each pair of grasp clusters
  Step3: Merge two grasp clusters with biggest correlation index into one grasp cluster, $N \Leftarrow N - 1$
end while

IV. Experiments

To examine the effectiveness of different visual features for recognizing grasp types, we collected a new dataset in a laboratory environment (we call it "UT Grasp Dataset"). Only a subset of grasp types in Feix's taxonomy are considered in the dataset, since not all the grasp types are commonly used in everyday activities. We select 17 distinct grasp types as shown in Fig. 7 based on the statistical result of grasp prevalence provided by Bullock *et al.* [15]. We have also trained a classifier for non-grasp type using hand images when the hand is not holding any object (*e.g.*, when the hand is approaching the object). Five subjects were asked to grasp different objects placed on a desktop after brief demonstration of how to perform each grasp type. There are five unique sets of objects which are commonly used in different tasks (cleaning, cooking, office work, bench work, and entertainment). Each subject performed all 17 grasp types on one object set in one video recording. The same grasping was performed twice at different time. In total, we recorded 50 trials (50 video recordings) of hand grasp data with five subjects and five object sets. Each recording lasts about five minutes and the total video data is over four hours. Videos were recoded by a head mounted camera (GoPro Hero2) at 30 fps and downsized to 960×540 pixels per frame. Fig. 8 (top 2 rows) shows example images from UT Grasp Dataset.

Fig. 7. Grasp taxonomy [10] used in the experiment. 17 grasp types commonly used in daily manipulation tasks [15] are selected.

To evaluate our system in real-world environments, we also conducted experiments on a public human grasping dataset [44]. 20 video sequences recording a machinist's daily work are used (we call it "Machinist Grasp Dataset"). The total length of video data is nearly 2.5 hours. The video quality of the Machinist Grasp Dataset is relatively low with image resolution of 640×480 pixels. Fig. 8 (bottom 2 rows) shows some example images. Grasp types have been annotated by experienced raters. We focus on the same 17 grasp types as in UT Grasp Dataset which are frequently used through out all sequences.

We have examined six different features in our system as described in Section III-B. Four features (HoG, HHoG, SIFT, CNN) rely on hand regions of fixed size. In the experiments, hand regions are segmented with bounding boxes of 160×160 pixels for UT Grasp Dataset and 128×128 pixels for Machinist Grasp Dataset. Both HoG and HHoG are

Fig. 8. Images samples from UT Grasp Dataset (top 2 rows) [5] and Machinist Grasp Dataset (bottom 2 rows) [44].

computed on hand regions after resizing to 160×160 pixels and the feature dimension is 2916. The feature dimension of SIFT is 100 since it is encoded using BoW with 100 dictionary entries. Features based on CNN are extracted from hand regions using the Caffe implementation [45] of the CNN model proposed by Krizhevsky *et al*. [35]. Each hand region is forward propagated through five convolutional layers and a fully connected layer and the output feature dimension is 4096. Another two features are based on dense trajectories. Improved Dense Trajectories (IDT) proposed by Wang and Schmid [41] improves dense trajectories by removing camera motion estimated by computing homography from matched feature points between two consecutive frames. Our proposed DHT also removes camera motion. The difference is that we discard feature matches within detected hand regions since the hand motion is inconsistent with camera motion. Both IDT and DHT are encoded using Fisher vector with same parameters and the feature dimension is 32768.

Linear SVMs are trained for each grasp type using the visual features mentioned above. We use the implementation of LIBSVM [46] for training. At test time, each frame with detected hand region is assigned to a grasp type of which the classifier obtains the highest score. The classification accuracy is used for evaluating the grasp recognition performance.

A. Grasp Recognition On UT Grasp Dataset

We applied our approach to UT Grasp Dataset to see how visual features can discriminate among different grasp types in controlled environments.

1) Cross-Trial Performance: To evaluate grasp recognition performance for specific user (subject) and task over different trials, we train grasp classifiers for each subject and object set on one trial and test them on another trial. Recognition performance of different features are shown in Fig. 9(a). The average and standard deviation of accuracy is computed

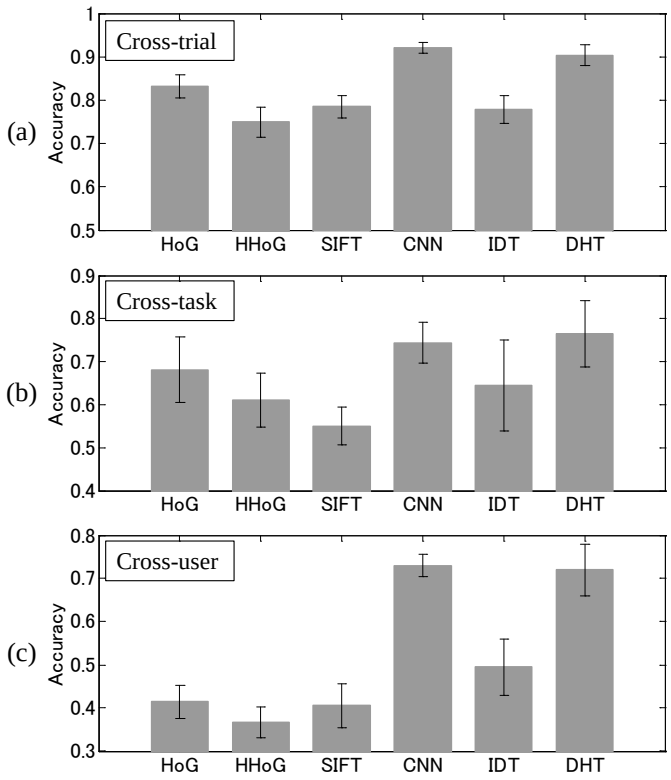


Fig. 9. Grasp recognition performance of different features on UT Grasp Dataset. The figure shows performance statistics (average and standard deviation of classification accuracy) under three different experimental settings: (a) Cross-trial (b) Cross-task and (c) Cross-user.

from the classification accuracy on all subjects and object sets. CNN-based feature achieves best average accuracy of 0.92. As for the four appearance-based features (HoG, HHoG, SIFT, CNN), the superior performance of CNN demonstrates the advantage of high-level biology-inspired features in accurate classification. Performance from SIFT indicates local appearance-based feature alone is less discriminative than global features. Although the separation between hand and object in HHoG seems intuitive and well-motivated, HHoG performs worse than HoG. This is partly due to the hand segmentation noises, and also because HoG encodes additional information about the grasped object. As for the two trajectory-based features, better performance of the proposed DHT over IDT proves the effectiveness of removing unrelated information from the background. Although DHT has slightly worse performance than CNN, we believe this is because hand appearance is consistent in different trials and motion information contained in DHT doesn't help in the controlled environment. Experimental results show that it is possible to construct high performance task-specific grasp classifiers for specific users.

2) Cross-Task Performance: To evaluate system generality across different tasks (simulated by different object sets), we use a leave-one-task-out cross-validation scheme. Specifically, we train grasp classifiers on four object sets and test on the rest object set and iterate the process five times. The average and standard deviation of accuracy is computed from classification

TABLE I
GRASP RECOGNITION PERFORMANCE ON MACHINIST GRASP DATASET. PRECISION (P) AND RECALL (R) ARE SHOWN FOR TOP NINE PREVALENT GRASP TYPES. NUMBER WITHIN PARENTHESES ASIDE EACH GRASP TYPE INDICATES SAMPLE PROPORTION.

	MW (.20)		LP (.19)		LT (.12)		T3F (.11)		TIF (.11)		T4F (.07)		T2F (.05)		PS (.04)		IFE (.03)		Total
	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	Accu.
HoG	.35	.37	.48	.67	.38	.53	.17	.14	.17	.23	.29	.06	.15	.09	.08	.06	.86	.69	.34
HHoG	.32	.37	.39	.49	.38	.58	.18	.14	.20	.20	.09	.02	.06	.02	.06	.06	.33	.42	.29
SIFT	.19	.21	.43	.62	.26	.41	.00	.00	.04	.01	.00	.00	.00	.00	.20	.06	.27	.69	.24
CNN	.59	.56	.59	.74	.64	.77	.26	.23	.31	.35	.26	.25	.21	.21	.41	.28	.70	.73	.49
IDT	.63	.60	.68	.84	.80	.95	.19	.16	.39	.46	.33	.28	.20	.19	.76	.69	.83	.73	.54
DHT	.69	.71	.65	.86	.88	.94	.24	.22	.46	.49	.40	.38	.32	.40	.69	.63	.95	.77	.59

accuracy on all object sets. From Fig. 9(b), we can see the DHT-based feature achieves best average accuracy of 0.764. Compared to the performance obtained in the cross-trial case (Fig. 9(a)), average accuracy of the cross-task case degrades by nearly 15%, and the standard deviation of accuracy becomes larger. This is reasonable since objects used in different tasks have different appearance which undermines the discrimination ability of appearance-based classifiers. Still, experimental results demonstrate the system's ability to generalize across different tasks.

3) Cross-User Performance: To evaluate system generality across different users, we use a leave-one-subject-out cross-validation scheme. The average and standard deviation of accuracy is computed from classification accuracy on all subjects. As illustrated in Fig. 9(c), best performance is achieved from CNN-based feature and DHT-based feature with average accuracy of 0.73 and 0.72 respectively. The performance degrades nearly 20% in the cross-user case compared to that obtained in the cross-trial case. Two important reasons can explain the performance degradation. One reason is that the skin color and size of hands of different users are different. Another reason is that different users prefer different grasping styles even in performing the same grasp type. Taking *Writing Tripod* for example, one user prefers to grip the pen-like tool between the index and middle fingers, which is uncommon to other users who distribute pressure evenly on three fingers–the thumb, index and middle fingers. Although current subject size is not sufficient enough to fully validate the system's ability to generalize to large population, the potential of training general grasp classifiers which can be applied to other users is demonstrated.

B. Grasp Recognition On Machinist Grasp Dataset

We applied our approach to Machinist Grasp Dataset to evaluate the system performance in real-world environments.

Grasp recognition performance of different features on Machinist Grasp Dataset using 5-fold cross validation is shown in Table I. Sample proportion of each grasp type is also shown in the table as the prevalence of different grasp types is non-uniform. Due to space limitation, results of nine most prevalent grasp types and total accuracy are illustrated. Abbreviation is used for each grasp type and is composed by first letters of the full name. Our proposed DHT achieves highest accuracy of 0.59 compared to other features. It is reasonable that DHT works better than IDT since irrelevant trajectory information from background has been removed. CNN-based feature improves the accuracy by over 0.15 compared to HoG, which verifies the superiority of biology-inspired high-level features over hand-crafted features. Also it is obvious that trajectory-based features (DHT, IDT) outperform appearance-based features (CNN, HoG), partly because hand motion information is also captured in trajectory-based features which enhances the discrimination ability.


(a) (b)

Fig. 10. Examples of unreliable hand detection. (a) Incomplete hand detection with fingers missing due to extreme lighting condition (b) False detection from background with similar skin color

We believe the robustness to unreliable hand detection of trajectory-based features is another important reason why they outperform appearance-based features. Hand detection in real-world scenarios is sometimes unreliable due to extreme lighting conditions (*e.g.*, overexposure) and cluttered background. Fig. 10 shows some examples of bad detection. Grasp recognition relying on appearance-based features is heavily influenced by unreliable hand detection. To evaluate such influence, we also compared the classification accuracy under different hand detection conditions as shown in Table II. For ideal detection, we manually select image samples in which automatic hand detection results are acceptable and nearly 25% of instances are removed. For real detection, we use all image samples. There is a performance drop from ideal detection to real detection for HoG and CNN, which indicates appearance-based features are sensitive to hand detection. However, IDT and DHT are robust to hand detection with even slight performance improvement under real detection. We believe the reason resides on the feature tracking procedure through which IDT and DHT are extracted. And more training data under real detection further improves the recognition performance.

Although our system achieves promising performance with accuracy of 0.59 compared to 0.2 (the percentage of the most prevalent grasp type *Medium Wrap*) at the chance level, it

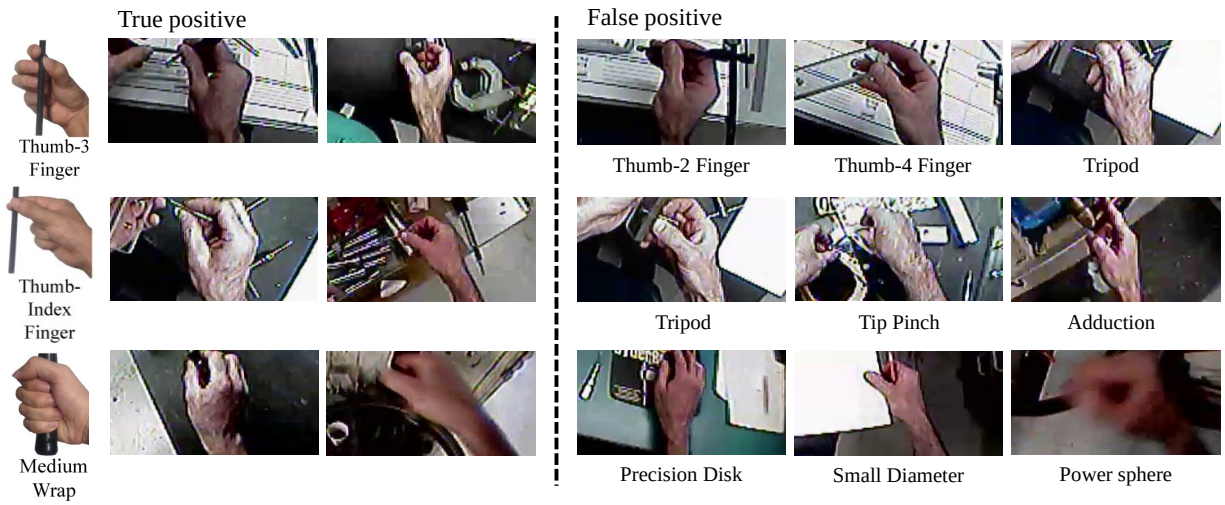


Fig. 11. Examples of true positives and false positives from grasp recognition on Machinist Grasp Dataset. Image crops on the left side are examples of true positives of *Thumb-3 Finger*, *Thumb-Index Finger*, and *Medium Wrap*. Image crops on the right side are examples of false positives with original grasp types indicated under each image.

TABLE II
PERFORMANCE INFLUENCES BY HAND DETECTION. FOR IDEAL DETECTION, IMAGE SAMPLES WITH IDEALLY DETECTED HAND REGIONS ARE USED. FOR REAL DETECTION, ALL IMAGES SAMPLES ARE USED.

	Ideal detection	Real detection
HoG	0.408	0.339
HHoG	0.325	0.294
SIFT	0.271	0.238
CNN	0.524	0.485
IDT	0.523	0.543
DHT	0.579	0.592

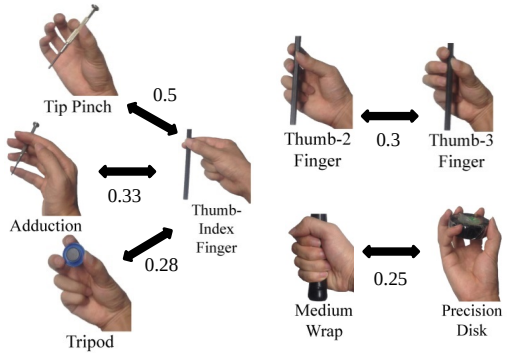


Fig. 12. Top 5 pairs of grasp types with highest correlation index.

fails to work well for some visually similar grasp types. As shown in Table I, precision and recall of some grasp types (*e.g.*, *Thumb-2 Finger* and *Thumb-3 Finger*) are relatively low. Some examples of failure cases are shown in Fig. 11. Two columns of image crops on the left side show true positives of a grasp type of which the prototype is also illustrated. Three columns of image crops on the right side show false positives with their original grasp types indicated under each image. As shown in these examples, some grasp types are extremely difficult to differentiate, even for human annotators. Taking *Thumb-3 Finger* for example, both of the first true positive and the first false positive show the machinist's hand holding a tool. It is hard to tell how many fingers are used in holding the tool only from visual perception.

The visual similarity between some pairs of grasp types (*e.g.*, *Thumb-2 Finger* and *Thumb-3 Finger*) poses big challenges in training discriminative grasp classifiers based on visual features. Distinguishing between such fine-grained grasp types would require more advanced techniques to extract detailed information such as the exact finger positions and contact surfaces.

C. Learning The Visual Structures Of Grasps

Here we show how the correlation between visually trained grasp classifiers can be used to discover the visual structure of hand grasps. Based on Equation 1, correlation index is computed for all grasp pairs using the classification results obtained on Machinist Grasp Dataset. We have removed bad hand detection samples from training data in order to make the correlation between classifiers more likely reflect the visual similarity of hand grasps. Top 5 grasp pairs with highest correlation index are shown in Fig. 12.

Following the iterative grasp clustering algorithm described in Algorithm 1, a dendrogram of grasp types is obtained by iteratively clustering two most correlated grasp types after each iteration of supervised learning. A dendrogram is a binary tree which gives a complete graphical description of the hierarchical clustering. The final constructed grasp dendrogram based on DHT is shown in Fig. 13. The original grasp types from Feix's taxonomy are located at the leaf nodes (level-0). Grasp types with the higher correlation are clustered at lower levels, while those dissimilar with each other are clustered later at higher levels of the dendrogram. We observe that grasp types are clustered in a manner consistent with known divisions of power and precision grasps in expert-designed grasp taxonomies [1][10]. With the exception of *Precision Disk* and *Extension Type*, the division between

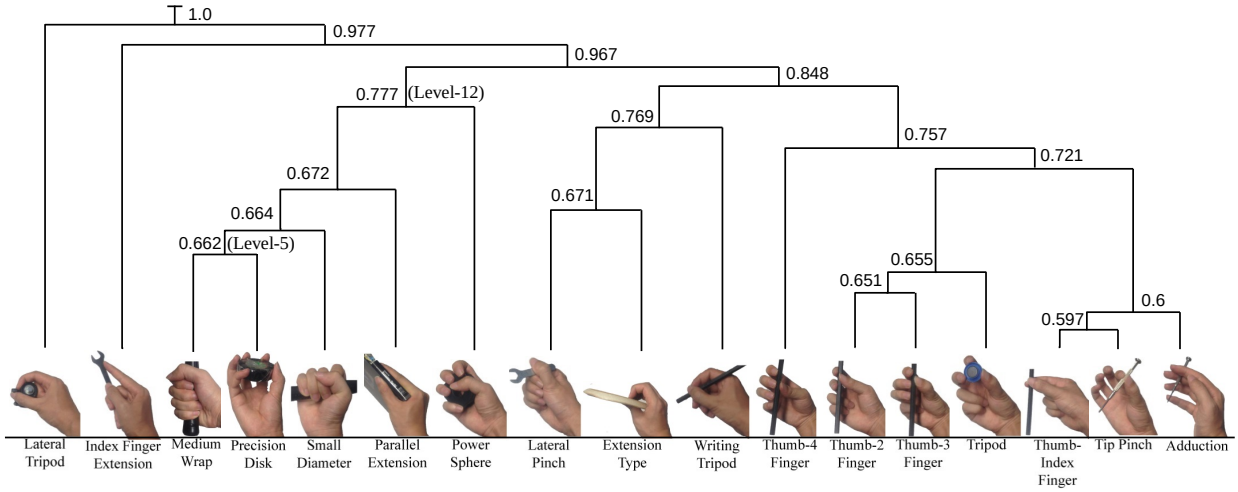


Fig. 13. Automatically learned grasp dendrogram (taxonomy tree). Classification accuracy obtained at different clustering levels are shown.

power and precision grasps is preserved until level-12 (the 12-th iteration) of the grasp dendrogram. There are five groups of grasp types remained at level-12. The group with grasp types ranging from *Medium Wrap* to *Power Sphere* represents the power grasps characterized by stably holding an object with palm and five fingers. In contrast, the group ranging from *Thumb-4 Finger* to *Adduction* represents the precision grasps often used to flexibly manipulate an object with dexterous finger articulation. Another interesting group represented by *Lateral Pinch* and *Writing Tripod* stands intermediately between power and precision grasps where both stability and dexterity are addressed. These qualitative examples show that our approach can discover grasp structures consistent with parts of the expert-designed taxonomy.

The more important observation however is that intuitive grasp structures have been learned automatically from data. While classical grasp taxonomies have been designed through manual introspection, the shared uncertainty among visual classifiers can also be used to learn intuitive structures over human grasps. To have a quantitative comparison between different hierarchical grasp taxonomies, we propose a new metric called Normalized Common Distance (NCD) score. The NCD score is computed as:

$$NCD(T_a, T_b) = \frac{1}{N} \sum_{\substack{l_A \in T_a, T_b \\ l_B \in T_a, T_b \\ A \neq B}} \left| \frac{d_a(l_A, l_B)}{H_a} - \frac{d_b(l_A, l_B)}{H_b} \right|$$

where l_A and l_B are leaf nodes with labels of A and B respectively, H_a and H_b are depth of tree T_a and T_b, $d(*,*)$ is the Lowest Common Ancestor (LCA) [47] distance between two nodes, and N is the number of all possible pairs of (l_A, l_B). In our case, a tree is a hierarchical grasp taxonomy and labels of its leaf nodes are grasp types from the taxonomy. Taking DHT-based tree (Fig. 13) for example, the tree has a depth of 8, and two leaf nodes with label *Medium Wrap* and *Power Sphere* has LCA distance of 5. The proposed NCD score can be used for comparing tree structures with different depth and branches. The NCD score has a minimal value of 0, and a upper bound value of 2, with smaller value indicating higher similarity.

We learned grasp taxonomy trees automatically based on three different features (HoG, CNN, DHT) and compared them with a reference taxonomy tree (Cutkosky's grasp taxonomy). We also compared between the automatically learned taxonomy trees themselves. The NCD scores are shown in Table III. The reference taxonomy tree has the smallest NCD score with the DHT-based one than with other ones, indicating the DHT-based taxonomy tree is most similar to Cutkosky's taxonomy tree. Another important observation is that the automatically learned taxonomy trees are actually very similar to each other as indicated by the NCD scores between themselves.

TABLE III
QUANTITATIVE COMPARISON BETWEEN DIFFERENT GRASP TAXONOMY TREES MEASURED BY NCD SCORE.

Tree pair	NCD score
(T_{ref}, T_{hog})	0.358
(T_{ref}, T_{cnn})	0.418
(T_{ref}, T_{dht})	0.353
(T_{hog}, T_{cnn})	0.200
(T_{cnn}, T_{dht})	0.324
(T_{dht}, T_{hog})	0.304

D. Recognition Using Grasp Abstractions

Based on the learned grasp taxonomy tree (Fig. 13), it is possible to "cut" the tree at different levels to obtain different sets of grasp clusters. Furthermore, each slice (abstraction) level can be interpreted as a new grasp taxonomy. By learning grasp classifiers for grasp taxonomies at different abstraction levels, we can achieve a trade-off between more detailed classification and more robust classification. To better show this trade-off, grasp classification accuracy at each abstraction level is also given in Fig. 13. By cutting a higher level of the tree to define a smaller grasp taxonomy, we can achieve more reliable grasp classification. For example, at level-12 of the tree, we will be able to differentiate 5 grasp types with an accuracy

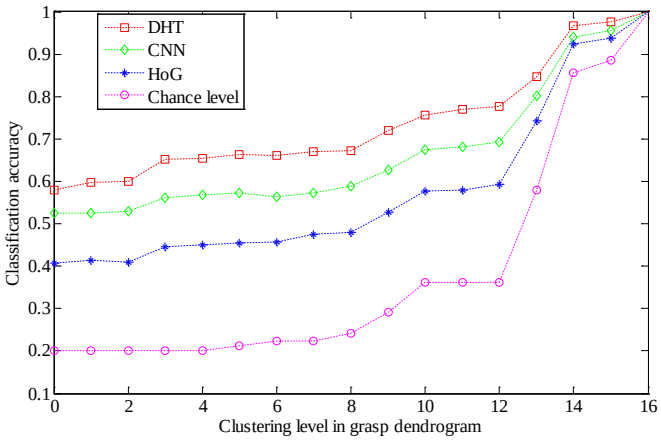


Fig. 14. Grasp recognition performance at different levels of grasp abstractions.

of 0.78. On the other hand, cutting at level-5 allows us to differentiate 12 grasp types with an accuracy of 0.66. Thus, the learned visual structure gives researchers the flexibility of finding a good balance between better performance and more detailed grasp analysis.

The variation of grasp recognition performance at different levels of the grasp taxonomy trees based on HoG, CNN and DHT are shown in Fig. 14. The chance level (percentage of the most prevalent grasp type in the selected abstraction level) is also drawn to demonstrate the bottom-line performance. As expected, the classification accuracy for all three features grows up steadily as we increase the abstraction level. From level-12 the accuracy increases dramatically since big grasp clusters are merged together and the chance of misclassification becomes much lower. Moreover, the big performance gap among the three features at lowest level (fine-grained classification) becomes smaller as abstraction level increases and inter-class ambiguity diminishes.

V. Discussions

As illustrated in the experiments, there is a big performance gap between grasp recognition in the laboratory setting and in the real-world setting. Also there is visual similarity between some pairs of grasp types, making it hard for the visually trained classifiers to reliably distinguish between fine-grained grasp types. Nevertheless, the visual similarity between different grasp types is explored to learn intuitive visual structures of hand grasps.

In the following sections, we first discuss the influences of different environments on system performance and the key issues to be addressed. Then, we discuss the insights and possible applications from the automatically learned visual structures of hand grasps.

A. System Performance Under Different Environments

In general, the proposed system achieves reliable grasp recognition performance in controlled environments, where hands can be reliably detected and each grasp type is correctly performed and clearly recorded. Specifically, the system achieves average accuracy of 0.92 in the cross-trial case, where training data and test data record one subject grasping the same set of objects at different time. The average accuracy drops to 0.764 in the cross-task case. The changing object appearance is the main reason of performance degradation since the objects being grasped in the test data never appear in the training data. The accuracy further drops to 0.73 in the cross-user case, which demonstrates that hand appearance and grasping styles of different users also have an impact on the system performance.

The system performance degrades significantly in real-world environments, where hands in real-life manipulation tasks are recorded and the video quality is relatively low. Specifically, average accuracy drops from 0.904 (we compare with the cross-trial case in UT Grasp Dataset since the data in Machinist Grasp Dataset is recorded from single subject) to 0.59 when the proposed DHT is used. It should be noted that the system performance degrades much worse when prevalent appearance features (HoG and CNN) are used. For HoG, accuracy drops from 0.831 to 0.34. And for CNN, accuracy drops from 0.92 to 0.49.

We believe there are three key issues to be addressed for real-world applications of the system. One major issue is reliable hand detection in real-world environments. Although the DHT is proposed to address the problem of false hand detection, future work is desired to fundamentally improve the hand detection. The second issue is more diverse grasp taxonomies. Most of existing grasp taxonomies have been designed for rigid objects with consistent shapes. Therefore, it is hard for human raters to reliably annotate the grasp types with soft objects (*e.g.* towel) or with objects of irregular shapes. The third issue is the visual similarity between different grasp types. In present work, visual structure of hand grasps has been learned to provide a trade-off strategy between more detailed classification and more robust classification. However, to improve the discrimination ability for fine-grained grasp classification, other modes of sensing data such as depth information might also be desired to infer more detailed grasp information.

B. Visual Structure Of Hand Grasps

As mentioned above, the learned grasp structures provide researchers with a compromise solution between more robust classification and more detailed classification. It depends on actual situation that which level of grasp abstraction to use for training grasp classifiers. For applications in which only the classification of power and precision grasps is cared about, higher abstraction level with less grasp types can be selected to achieve better performance without affecting the application goal. The chance level is another important factor to be considered in the selection of abstraction level, for the actual recognition power is reflected in the ratio of classification accuracy versus chance level. Specifically, taking Fig. 14 for example, the abstraction level above level-12 would better not be used as the chance level rises dramatically after merging big clusters.

The learned visual structure can also be used to refine grasp annotations. There are often two reasons behind high

correlation of two grasp types. One reason is that the two grasp types are intrinsically similar from their definition (finger articulation and object geometry), such as *Thumb-2 Finger* and *Thumb-3 Finger*. Another reason, which is important to be noted here, is annotation confidence. In real-world setting, a subject is doing natural manipulation tasks without performing specific grasp type from any order, therefore some recorded hand poses are not corresponding exactly to any grasp types in existing grasp taxonomies. While human raters are inclined to annotate unknown hand poses to any close grasp types in their mind, the annotation becomes inconsistent for such unknown poses, of which the close grasp types are interrelated in training. By inspecting data samples of the interrelated grasp types based on the learned visual structures, it can help researchers to refine grasp annotations of low confidence or even to define a set of new distinct grasp types.

In present work, the visual structures are learned by iteratively clustering predefined grasp types based on a supervised learning process. However, it is insufficient to deal with undefined hand-object interactions often appeared in new scenarios. This can be addressed by integrating an unsupervised clustering method for discovering unknown grasp types. As done in the work of Huang *et al.* [11], an unsupervised clustering method is utilized to obtain a diverse set of hand-object interactions based on hand appearance, from which new distinct grasp types can be discovered. By adding newly discovered grasp types into existing grasp taxonomy, the grasp analysis system would be more adaptable to new scenarios.

VI. Conclusions

We proposed an egocentric vision-based system to automate the hand grasp analysis in large amounts of video data recorded with a wearable camera. Given an egocentric video, hands are automatically detected, and grasp classifiers are trained to recognize different grasp types based on state-of-the-art computer vision techniques. Furthermore, intuitive visual structures of hand grasps are learned by an iterative grasp clustering method.

The system performance is evaluated in both laboratory and real-world scenarios. In laboratory scenario, the system achieves high performance grasp recognition (92% accuracy) for specific users, and shows its potential for generalizing across different tasks (76% accuracy) and users (73% accuracy). Although the recognition performance degrades a lot (59% accuracy with the proposed feature) in real-world scenario, our work shows considerable potential for developing automatic systems for analyzing everyday hand grasp usage with large scale of data. Moreover, the automatically learned visual structures of hand grasps give researchers the flexibility of finding a good balance between more robust classification and more detailed grasp analysis.

In future work, we plan to expand the current dataset to include hand grasp data from more subjects that cover different ages and races, in order to validate and further improve the system reliability to generalize to large population. Besides, we also plan to extend the system to deal with both RGB and depth data so as to make the system more stable in real-world environments as wearable RGB-D cameras may become available in the near future.

References

[1] M. R. Cutkosky, "On grasp choice, grasp models, and the design of hands for manufacturing tasks," *IEEE Transactions on Robotics and Automation*, vol. 5, no. 3, pp. 269–279, 1989.

[2] A. D. Keller, *Studies to Determine the Functional Requirements for Hand and Arm prosthesis*. Department of Engineering University of California, 1947.

[3] 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*, vol. 32, no. 7, pp. 1635–1639, 2001.

[4] J. Case-Smith, C. Pehoski, A. O. T. Association *et al.*, *Development of Hand Skills in Children*. American Occupational Therapy Association, 1992.

[5] M. Cai, K. M. Kitani, and Y. Sato, "A scalable approach for understanding the visual structures of hand grasps," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 1360–1366.

[6] G. Schlesinger, "Der mechanische aufbau der kunstlichen glieder," *Ersatzglieder und Arbeitshilfen fur Kriegsbeschadigte und Unfallverletzte*, pp. 321–661, 1919.

[7] J. R. Napier, "The prehensile movements of the human hand," *Journal of Bone and Joint Surgery*, vol. 38, no. 4, pp. 902–913, 1956.

[8] T. Iberall, G. Bingham, and M. Arbib, "Opposition space as a structuring concept for the analysis of skilled hand movements," *Experimental brain research series*, vol. 15, pp. 158–173, 1986.

[9] S. B. Kang and K. Ikeuchi, "Toward automatic robot instruction from perception-recognizing a grasp from observation," *IEEE Transactions on Robotics and Automation*, vol. 9, no. 4, pp. 432–443, 1993.

[10] T. Feix, R. Pawlik, H.-B. Schmiedmayer, J. Romero, and D. Kragic, "A comprehensive grasp taxonomy," in *Proceedings of the Robotics: Science and Systems Conference Workshop on Understanding the Human Hand for Advancing Robotic Manipulation*, 2009, pp. 2–3.

[11] D.-A. Huang, M. Ma, W.-C. Ma, and K. M. Kitani, "How do we use our hands? discovering a diverse set of common grasps," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 666–675.

[12] J. Romero, T. Feix, H. Kjellstrom, and D. Kragic, "Spatio-temporal modeling of grasping actions," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2010, pp. 2103–2108.

[13] I. M. Bullock, J. Z. Zheng, S. Rosa, C. Guertler, and A. M. Dollar, "Grasp frequency and usage in daily household and machine shop tasks," *IEEE Transactions on Haptics*, vol. 6, no. 3, pp. 296–308, 2013.

[14] R. Deimel and O. Brock, "A novel type of compliant, underactuated robotic hand for dexterous grasping," *Proceedings of the Robotics: Science and Systems Conference (RSS)*, pp. 1687–1692, 2014.

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

[16] T. Feix, I. Bullock, and A. Dollar, "Analysis of human grasping behavior: Object characteristics and grasp type," *IEEE Transactions on Haptics*, vol. 7, no. 3, pp. 311–323, 2014.

[17] M. Santello, M. Flanders, and J. F. Soechting, "Postural hand synergies for tool use," *The Journal of Neuroscience*, vol. 18, no. 23, pp. 10 105–10 115, 1998.

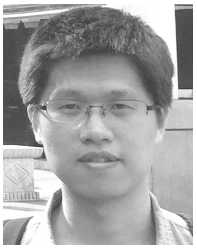
[18] H. Friedrich, V. Grossmann, M. Ehrenmann, O. Rogalla, R. Zöllner, and R. Dillmann, "Towards cognitive elementary operators: grasp classification using neural network classifiers," in *Proceedings of the IASTED International Conference on Intelligent Systems and Control (ISC)*, vol. 1, 1999, pp. 88–93.

[19] 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*, vol. 21, no. 1, pp. 47–57, 2005.

[20] S. Ekvall and D. Kragic, "Grasp recognition for programming by demonstration," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2005, pp. 748–753.

[21] H. Kjellstrom, J. Romero, and D. Kragic, "Visual recognition of grasps for human-to-robot mapping," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2008, pp. 3192–3199.

[22] H. Hamer, K. Schindler, E. Koller-Meier, and L. Van Gool, "Tracking a hand manipulating an object," in *Proceedings of the IEEE International Conference On Computer Vision (ICCV)*. IEEE, 2009, pp. 1475–1482.
[23] I. Oikonomidis, N. Kyriazis, and A. A. Argyros, "Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2011, pp. 2088–2095.
[24] J. Romero, H. Kjellström, C. H. Ek, and D. Kragic, "Non-parametric hand pose estimation with object context," *Image and Vision Computing*, vol. 31, no. 8, pp. 555–564, 2013.
[25] M. Cai, K. Kitani, and Y. Sato, "Understanding hand-object manipulation with grasp types and object attributes," in *Robotics: Science and Systems Conference (RSS)*, 2016.
[26] Y. Yang, Y. Li, C. Fermuller, and Y. Aloimonos, "Grasp type revisited: A modern perspective of a classical feature for vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pp. 400–408.
[27] C. Li and K. M. Kitani, "Pixel-level hand detection in ego-centric videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013, pp. 3570–3577.
[28] Y. Li, A. Fathi, and J. M. Rehg, "Learning to predict gaze in egocentric video," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2013, pp. 3216–3223.
[29] L. Baraldi, F. Paci, G. Serra, L. Benini, and R. Cucchiara, "Gesture recognition in ego-centric videos using dense trajectories and hand segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2014, pp. 702–707.
[30] G. Rogez, J. S. S. III, and D. Ramanan, "First-person pose recognition using egocentric workspaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pp. 4325–4333.
[31] A. Saran, D. Teney, and K. M. Kitani, "Hand parsing for fine-grained recognition of human grasps in monocular images," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 5052–5058.
[32] A. A. Argyros and M. I. Lourakis, "Real-time tracking of multiple skin-colored objects with a possibly moving camera," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2004, pp. 368–379.
[33] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1. IEEE, 2005, pp. 886–893.
[34] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of Advances in neural information processing systems (NIPS)*, 2012, pp. 1097–1105.
[36] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014, pp. 1725–1732.
[37] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.
[38] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014, pp. 580–587.
[39] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 3169–3176.
[40] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Image Analysis*. Springer, 2003, pp. 363–370.
[41] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2013, pp. 3551–3558.
[42] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2010, pp. 143–156.
[43] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*. Citeseer, 2000, pp. 61–74.
[44] I. M. Bullock, T. Feix, and A. M. Dollar, "The yale human grasping dataset: Grasp, object, and task data in household and machine shop environments," *The International Journal of Robotics Research*, vol. 34, no. 3, pp. 251–255, 2015.
[45] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014, pp. 675–678.
[46] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.
[47] H. N. Djidjev, G. E. Pantziou, and C. D. Zaroliagis, "Computing shortest paths and distances in planar graphs," in *Automata, Languages and Programming*. Springer, 1991, pp. 327–338.



Minjie Cai received the B.S. and M.S. degrees in electronics and information engineering from Northwestern Polytechnical University, Xi'an, China, in 2008 and 2011 respectively, and the Ph.D. degree in information science and technology from The University of Tokyo, Tokyo, Japan, in 2016.

He is currently a Postdoctoral Researcher with the Institute of Industrial Science, The University of Tokyo. His research interests include hand manipulation analysis, first-person vision and its applications.

Kris M. Kitani received the B.S. degree in electrical engineering from University of Southern California, CA, USA, in 2000, the M.S. and Ph.D. degrees from The University of Tokyo, Tokyo, Japan, in 2005 and 2008, respectively.

He is currently an Assistant Research Professor with the Robotics Institute, Carnegie Mellon University. His research interests include first person vision, action modeling, hand detection and gesture analysis.

Yoichi Sato received the B.S. degree from The University of Tokyo, Tokyo, Japan, in 1990, the M.S. and Ph.D. degrees in robotics from the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA, in 1993 and 1997, respectively.

He is currently a Professor with the Institute of Industrial Science, The University of Tokyo. His research interests include physics-based vision, reflectance analysis, image-based modeling and rendering, and gaze and gesture analysis.