# A Scalable Approach for Understanding the Visual Structures of Hand Grasps

Minjie Cai[1], Kris M. Kitani[2] and Yoichi Sato[1]

***Abstract*— Our goal is to automatically recognize hand grasps and to discover the visual structures (relationships) between hand grasps using wearable cameras. Wearable cameras provide a first-person perspective which enables continuous visual hand grasp analysis of everyday activities. In contrast to previous work focused on manual analysis of first-person videos of hand grasps, we propose a fully automatic vision-based approach for grasp analysis. A set of grasp classifiers are trained for discriminating between different grasp types based on large margin visual predictors. Building on the output of these grasp classifiers, visual structures among hand grasps are learned based on an iterative discriminative clustering procedure. We first evaluated our classifiers on a controlled indoor grasp dataset and then validated the analytic power of our approach on real-world data taken from a machinist. The average F1 score of our grasp classifiers achieves over 0.80 for the indoor grasp dataset. Analysis of real-world video shows that it is possible to automatically learn intuitive visual grasp structures that are consistent with expert-designed grasp taxonomies.**

## I. Introduction

This work aims to provide a scalable computer vision-based framework for understanding and analyzing the use of human hands. In particular, we propose a fully automatic vision-based approach for recognizing hand grasps and learning visual grasp structures using a wearable camera.

For over a century, analysis of hands and their interactions with the physical world has attracted great focus from researchers across different domains such as neuromuscular rehabilitation [1], robotic arm design [2] and motor control analysis [3]. In robotics, the study of hand function has provided critical information regarding design of robotic and prosthetic hands [4][2]. However, traditional approaches to grasp analysis have been developed primarily in a controlled laboratory setting, which often includes intrusive hand-contact sensors or calibrated cameras.

Wearable cameras overcome the constraints of other modes of direct sensing by allowing for continuous recording of natural hand interactions. Furthermore, the first-person view is an ideal viewing perspective for grasp analysis since a hand and an object being grasped are naturally located in the center of the visual field. However, the most significant benefit of a wearable camera is that it enables the study of hand grasps at a large scale. It is now possible to effortlessly record hours of video for analyzing hand grasps.

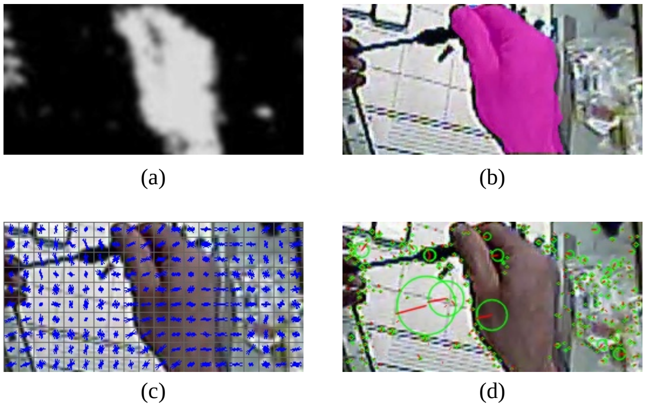


Fig. 1. Examples of our automated vision-based framework which enables large scale analysis of hand use. (a) Skin detection (b) Hand segmentation (c) Global gradient histograms (d) Salient interest points

In this work, we take a departure from classical prehensile techniques (i.e., manual annotation) and develop automatic computer vision techniques that can be used as a tool to advance studies in prehensile analysis. Examples of computer vision techniques used in our work are shown in Fig. 1. 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 backgrounds with extreme ego-motion. Based on robust hand detection, grasp-related features are extracted which encode the hand shape and object context, and grasp classifiers are trained for discriminating between different grasp types. Finally, the grasp classifiers are used to learn the visual similarities between grasps to automatically build an appearance based grasp hierarchy, which we call *the visual structure of hand grasps*. In our experiments, the analysis of real-world video shows that it is possible to automatically learn intuitive visual grasp structures that are consistent with expert-designed grasp taxonomies.

The contributions of this work are as follows: 1) We propose a fully automatic vision-based approach which can achieve robust grasp recognition performance only with a single wearable camera. 2) We propose a method for learning visual structures of hand grasps using a visual clustering approach which enables the system to automatically learn task-based grasp taxonomies.

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

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

## II. Related Work

### A. Hand Grasp Taxonomies

Grasp taxonomies have been studied for almost a century to better understand the use of human hands [5][4][6][2][7][8]. Early work by Schlesinger [5] classified hand grasps into 6 major categories based on hand shape and object properties. In 1956, Napier proposed a scheme [6] that divides grasps into power and precision grasps based on requirements of the manipulation task. The categorizations of power and precision grasps was widely adopted by researchers in the medical, biomechanical and robotic fields. In studying grasps in manufacturing tasks, Cutkosky provided a comprehensive hand grasp taxonomy [2] which played an important role in guiding robotic hand design. In the early 1990's, Kang and Ikeuchi [7] presented a computational framework for grasp identification, allowing automatic grasp planning of a robotic system from a demonstrated human grasp.

Recently with the advances in wearable camera technology, research has focused on prehensile analysis from large datasets of first-person point-of-view (POV) video. Work from Yale University [9][10][11] used several hours of first-person POV video to observe human grasping behavior. In the tradition of previous work on prehensile analysis, the process required many hours of visual inspection by skilled annotators. However, as it becomes easier to acquire large amounts of visual data, it is clear that manual approaches will not scale to larger datasets. Therefore it is the aim of this work to propose a scalable automatic vision-based framework that will help to support next generation research in the area of prehensile analysis using a large amount of video data.

### B. Automated Grasp Analysis

Automated data-driven approaches for prehensile analysis have been developed primarily in a controlled laboratory setting. Hand tracking devices such as data gloves or inertial sensors have been used to obtain detailed measurements of joint angles and positions on the hand [12][13][14][15]. Since sensors are directly embedded on the hand, hand movements can be measured with very high accuracy. However, the main limitation is that they must be worn and can sometimes inhibit hand interactions. Vision-based hand pose estimation systems [16][17][18][19][20] allow a completely non-contact form of hand interactions. However, most hand pose estimation systems also require a controlled environment of calibrated cameras and require that hand interactions are recorded in a laboratory setting. In order to understand the natural statistics over hand use, it is critical that hand interactions can be observed in normal activities of daily living – outside of the laboratory. In this work, we develop techniques targeted to analyze videos of everyday hand interactions recorded with a wearable first-person POV camera. [21] has recently presented results on hand pose estimation from an egocentric RGB-D camera. Instead our work focuses on hand grasp recognition based on only RGB images.

## III. GRASP LEARNING FRAMEWORK

We desire to have a scalable grasp analysis framework which can learn discriminative classifiers and visual structures of hand grasps automatically from videos. To this end, we adopt state-of-the-art hand detection techniques to segment hand regions from egocentric videos, we extract grasp-related features for training discriminative grasp classifiers and we use supervised clustering method to learn visual structures of hand grasps.

### A. Hand Segmentation

Robustly identifying hand regions with a wearable camera is a challenging yet essential pre-processing needed to automate hand grasp analysis. As the camera is mobile, the background is rapidly changing, hands are moving without constraint and the camera can move with extreme ego-motion. Recent work on detecting hand regions using a wearable camera has shown that robust hand detection performance can be achieved if the hand model is rapidly adapted to changes in imaging conditions [22]. Following [22], we train a multi-model hand detector composed by a collection of hand pixel classifiers indexed by global appearance models. Given a test image, the global appearance modeled 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, we generate a probability map for each image as illustrated in Figure 2(b). The value of each pixel represents the likelihood of being a hand pixel in the original image.

Once the hand probability map has been detected, hand region, which contains most of the grasp information, is then segmented with a bounding box. Candidate hand regions with arms are first selected by binarizing the probability map with a threshold. Regions under a certain area proportion are discarded and at most two regions are retained. Fig. 2(c) shows two candidate hand regions painted with green and orange contours. In this paper we only consider the right hand grasp. The left hand is suppressed by simply selecting the candidate hand region which is right-most. If no hand region is detected, that is when no hands are visible, the image is discarded. Each hand region is extracted with a fixed size bounding box which is shown as the white rectangle in Fig. 2(c). In detail, ellipse parameters (length of long/short axis, angle) are fitted to the original 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 across various manipulation tasks.

### B. Feature Representation

In expert-defined grasp taxonomies, different grasp types are often identified by hand postures, object properties and types of hand-object interactions. Therefore, we extracted

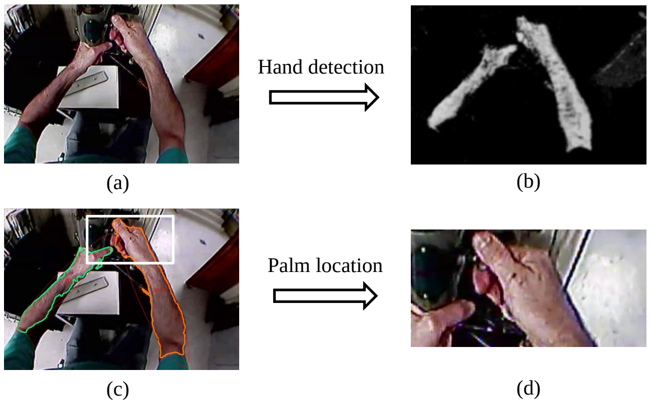


Fig. 2. Example of hand segmentation. (a) Image from first-person POV video (b) Skin probability map (c) Candidate hand regions (d) Palm region within a bounding box

grasp-related features for palm regions which encodes the shape of different hand postures and visual context of manipulated objects.

*1) Hand Shape:* Hand shapes are represented with Histogram of Oriented Gradient (HOG) [23] computed from a palm region. HOG features are an image descriptor based on collected local distributions of intensity gradients and have been widely used in object detection. The HOG features are computed by first dividing a palm 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 used a cell size of $8 \times 8$ pixels, block size of $16 \times 16$ pixels, and window size of $160 \times 80$ pixels with 9 orientation bins. A visualization of example HOG features is shown in the bottom-left of Fig. 3.

In our experiments we examined three variants of the HOG feature descriptor. The first is the global HOG feature described above. The second is a dimension-reduced version of HOG using Principle Component Analysis (HOG-PCA) to reduce the dimension of feature descriptor from 6156 to 100. The third is HOG features weighted by a skin probability map (HandHOG). HandHOG effectively suppresses gradients due to object being manipulated or background regions. As shown in Fig. 3, HOG features corresponding to non-hand regions are removed by weighting each block histogram by squared hand probability at the center of the block.

*2) Object Context:* We extract features based on local keypoints in order to capture the visual context of the object and hand-object interaction. In particular, we extract the following two local gradient descriptors.

We extract SIFT features [24] as a representation of the visual context of manipulated objects. Example keypoints are visualized in Fig. 4 where the scale and orientation of each keypoint are illustrated with a circle and a red radius. Histogram of gradients around each keypoint is computed

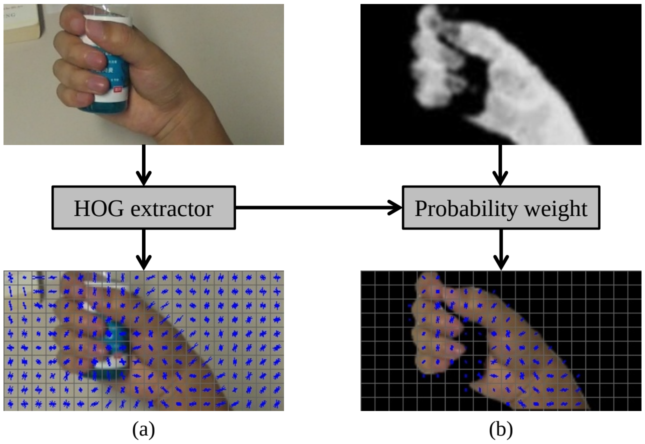


Fig. 3. Visualization of HOG features. (a) HOG (b) HandHOG

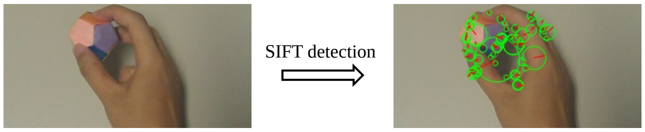


Fig. 4. Visualization of SIFT keypoints.

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 an image descriptor which contains the frequency of keypoint patterns. A total of 100 keypoint patterns are generated using k-means clustering over all keypoint descriptors.

In addition to the SIFT BOW, we also used the same approach to obtain a 100-dimensional image descriptor counting frequency of block-based HOG features which are generated using k-means clustering over all block HOG descriptors. The two 100-dimensional feature vectors are then concatenated together to generate a new feature (BlockHOG-SIFT).

### *C. Grasp Recognition and Abstraction*

We trained one-versus-all multi-class grasp classifiers for the grasp types defined in Feix's taxonomy [8]. We use this taxonomy since it is the most complete one in existence and has previously been applied to grasp analysis in [10][11]. We performed probability calibration [25] for each classifier in order to produce comparable scores. During testing, each frame is classified to the grasp type of the classifier with the highest score.

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

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

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

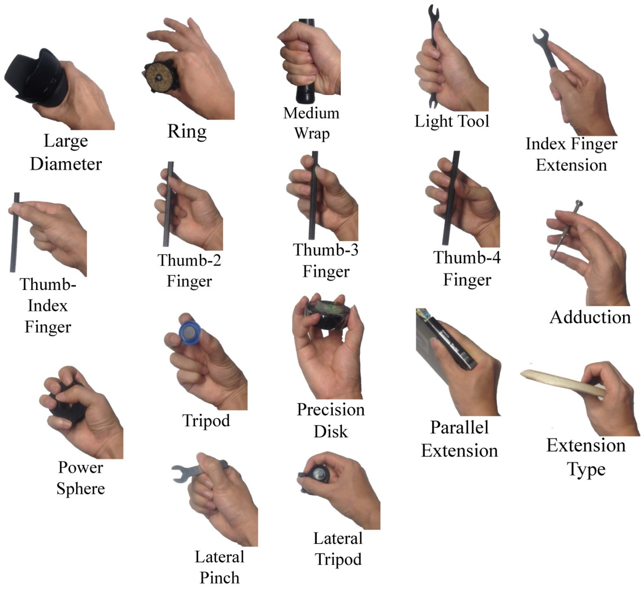


Fig. 5. 17 different grasp types from [8][10].

are the number of samples from grasp type $i$ and grasp type $j$, respectively.

Based on the correlation index, we implemented a supervised grasp clustering algorithm by iteratively clustering two most similar grasp types. The algorithm is described in Algorithm 1. This process defines a visual structure between grasp types – a grasp dendrogram.

**Algorithm 1** Supervised 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 explore the effectiveness of our visual features for recognizing grasp types, we created a new dataset under controlled environment (we call it “UT Grasp Dataset”). Only a subset of grasp types in Feix’s taxonomy are considered in our dataset, since not all the grasp types are commonly used in everyday activities. We selected 17 grasp types as shown in Fig. 5 based on the statistical result of grasp prevalence provided by Bullock et al. [10]. Four subjects were asked to grasp a set of objects placed on a desktop after brief demonstration of how to perform each type of grasps. Each subject performed hand grasps with a unique set of objects (*e.g.*, different objects with a cylindrical shape are used by different subjects in the *medium wrap* grasp type). Video was recoded by a HD head mounted camera (GoPro Hero2) at 30 fps while subjects performed each grasp type with varying hand poses. The recorded video was then downsized to 960 × 540 pixels.

To examine our approach in more natural environments, we use a real-world grasp dataset [26], which is composed of 20 video sequences recording a machinist’s daily work (we call it “Machinist Grasp Dataset”). The Machinist Grasp Dataset is part of a larger human grasping dataset provided by Yale University and is manually labeled with grasp types. The video quality of the Machinist Grasp Dataset is relatively low with the image resolution of 640x480 pixels. In our experiments on Machinist Grasp Dataset, we removed rare grasp types and focused on 17 remaining ones which at least take place three times through out all sequences. The 17 grasp types in Machinist Grasp Dataset as shown in Fig. 9 are slightly different from that in UT Grasp Dataset since grasp usage varies in different tasks.

Hand regions were segmented with a bounding box with the size of 320 × 160 for UT Grasp Dataset and 256 × 128 for Machinist Grasp Dataset. Then four feature descriptors (HOG, HOG-PCA, HandHOG, and BlockHOG-SIFT) were extracted for each of the segmented hand regions as explained in Section III-B. Finally, three types of classifiers were trained by using the obtained feature descriptors: (1) Linear Support Vector Machine (SVM-linear), (2) SVM with Radial Basis Function kernel (SVM-rbf), and (3) Exemplar SVM (ESVM). The average F1 score computed from a weighted average of the F1 score of each grasp type is used for evaluating the grasp recognition performance. Value ranges from 0 to 1, where 1 represents perfect performance.

### A. *Performance of Grasp Recognition*

We applied our approach to UT Grasp Dataset and Machinist Grasp Dataset to see how visual features can discriminate between different grasp types in both controlled and natural environments.

First we present grasp recognition results for a single user on UT Grasp Dataset. We trained and tested grasp classifiers for each user using 5-fold cross validation. The average F1 scores of the 17 grasp classifiersare shown in Table I for different feature descriptors and different machine learning algorithms. From Table I, we can see global features (HOG, HOG-PCA, HandHOG) outperform local feature histograms (BlockHOG-SIFT). While different hand grasps may share similar statistics of local gradient patterns, we observe that global gradient information is important for robust classification. Although the separation between hand and object in HandHOG seems intuitive and well-motivated, HandHOG performs slightly worse than HOG in nearly all cases. This is in part because of the hand segmentation noises, but also because HOG encodes additional information about the appearance of the object being held. The big performance gap between SVM-linear and SVM-rbf, especially when using HOG-PCA, indicates that hand grasps have wide variance in pose and are therefore not linearly separable. More importantly, the experimental results show that it is possible to construct high performance vision-based task-specific classifiers for a single user.

TABLE I

PERFORMANCE OF SINGLE USER ON UT GRASP DATASET

| | SVM-linear | SVM-rbf | ESVM |
|---|---|---|---|
| HOG | 0.85 | 0.86 | **0.89** |
| HOG-PCA | 0.79 | 0.88 | **0.89** |
| HandHOG | 0.8 | 0.85 | **0.88** |
| BlockHOG-SIFT | 0.79 | **0.8** | 0.79 |

The grasp recognition performance on Machinist Grasp Dataset using 5-fold cross validation is shown in Table II. Note that the dataset contains nearly eight hours of video data recording a machinist's daily work, thus it provides a good platform to evaluate how our vision-based approach works under real-world conditions. The combination of HOG-PCA and SVM-rbf achieves the best average F1 of 0.42, the average F1 for classification of 17 classes is 0.06 at the chance level. Although the absolute performance is still low, we believe that the result demonstrates the potential of automatic visual classification of grasp types in a realistic setting.

Some examples of true positives and false positives are shown in Fig. 6. Two columns to the left of the dashed line show true positives of a grasp type of which the prototype is illustrated in the left-most column. The false positives are shown in the right side of Fig. 6. From these examples, we see that 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. Differentiating between fine-grained categories such as these will require more advanced vision-techniques for extracting exact finger positions. This is left to our future work.

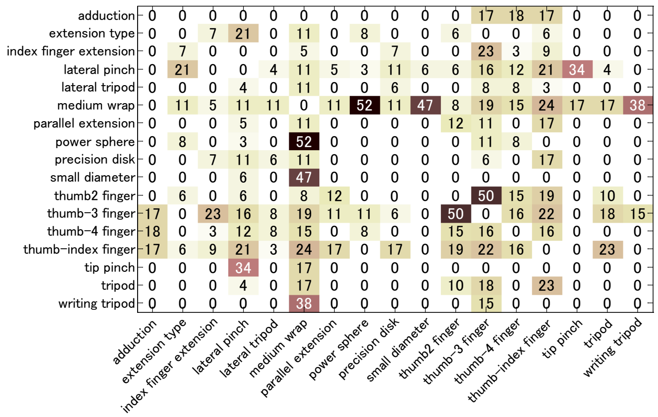
TABLE II

PERFORMANCE ON MACHINIST GRASP DATASET

| | SVM-linear | SVM-rbf | ESVM |
|---|---|---|---|
| HOG | 0.31 | 0.37 | **0.39** |
| HOG-PCA | 0.18 | **0.42** | 0.38 |
| HandHOG | 0.32 | **0.38** | 0.34 |
| BlockHOG-SIFT | 0.29 | **0.39** | 0.37 |

### B. Learning the Visual Structure between Grasps

Here we show how the correlation between visually trained grasp classifiers can be used to discover the visual structures of hand grasps. We computed the correlation index between all pairs of grasp types for Machinist Grasp Dataset based on classification results using combination of HOG-PCA and SVM-rbf. The correlation matrix of 17 grasp types is shown in Fig. 7, where each element indicates the correlation index (scaled by 100 for visualization) between a pair of grasp types indexed by rows and columns. Top 5 pairs of grasp types with highest correlation index are shown in Fig. 8.



| | adduction | extension type | index finger extension | lateral pinch | lateral tripod | medium wrap | parallel extension | power sphere | precision disk | small diameter | thumb2 finger | thumb−3 finger | thumb−4 finger | thumb−index finger | tip pinch | tripod | writing tripod |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| adduction | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 17 | 18 | 17 | 0 | 0 | 0 |
| extension type | 0 | 0 | 7 | 21 | 0 | 11 | 0 | 8 | 0 | 0 | 6 | 0 | 0 | 6 | 0 | 0 | 0 |
| index finger extension | 0 | 7 | 0 | 0 | 0 | 5 | 0 | 0 | 7 | 0 | 0 | 23 | 3 | 9 | 0 | 0 | 0 |
| lateral pinch | 0 | 21 | 0 | 0 | 4 | 11 | 5 | 3 | 11 | 6 | 6 | 16 | 12 | 21 | 34 | 4 | 0 |
| lateral tripod | 0 | 0 | 0 | 4 | 0 | 11 | 0 | 0 | 6 | 0 | 0 | 8 | 8 | 3 | 0 | 0 | 0 |
| medium wrap | 0 | 11 | 5 | 11 | 11 | 0 | 11 | 52 | 11 | 47 | 8 | 19 | 15 | 24 | 17 | 17 | 38 |
| parallel extension | 0 | 0 | 0 | 5 | 0 | 11 | 0 | 0 | 0 | 0 | 12 | 11 | 0 | 17 | 0 | 0 | 0 |
| power sphere | 0 | 8 | 0 | 3 | 0 | 52 | 0 | 0 | 0 | 0 | 0 | 11 | 8 | 0 | 0 | 0 | 0 |
| precision disk | 0 | 0 | 7 | 11 | 6 | 11 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 17 | 0 | 0 | 0 |
| small diameter | 0 | 0 | 0 | 6 | 0 | 47 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| thumb2 finger | 0 | 6 | 0 | 6 | 0 | 8 | 12 | 0 | 0 | 0 | 0 | 50 | 15 | 19 | 0 | 10 | 0 |
| thumb−3 finger | 17 | 0 | 23 | 16 | 8 | 19 | 11 | 11 | 6 | 0 | 50 | 0 | 16 | 22 | 0 | 18 | 15 |
| thumb−4 finger | 18 | 0 | 3 | 12 | 8 | 15 | 0 | 8 | 0 | 0 | 15 | 16 | 0 | 16 | 0 | 0 | 0 |
| thumb−index finger | 17 | 6 | 9 | 21 | 3 | 24 | 17 | 0 | 17 | 0 | 19 | 22 | 16 | 0 | 0 | 23 | 0 |
| tip pinch | 0 | 0 | 0 | 34 | 0 | 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| tripod | 0 | 0 | 0 | 4 | 0 | 17 | 0 | 0 | 0 | 0 | 10 | 18 | 0 | 23 | 0 | 0 | 0 |
| writing tripod | 0 | 0 | 0 | 0 | 0 | 38 | 0 | 0 | 0 | 0 | 0 | 15 | 0 | 0 | 0 | 0 | 0 |

Fig. 7. Correlation matrix of 17 grasp types.

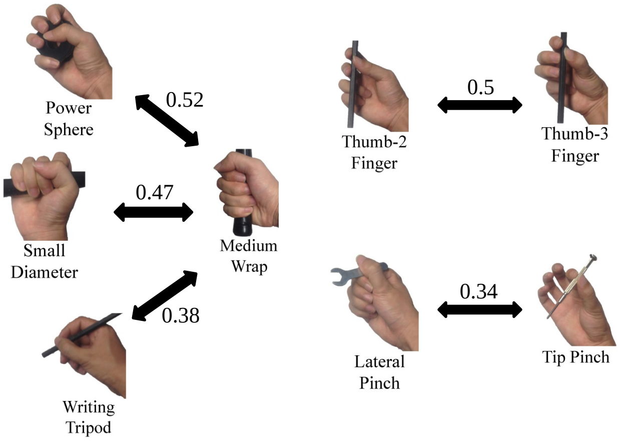


Fig. 8. Top 5 grasp correlations.

Following the iterative supervised clustering algorithm described in Algorithm 1, we constructed a dendrogram of grasp types 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 is shown in Fig. 9. Grasp types with the highest classifier correlation are clustered first at lower level nodes, while those dissimilar with each other are clustered later at higher levels in the tree. The original grasp types from Feix's taxonomy are located at the leaf nodes (level-0). We observe that for the first six iterations, grasps are clustered in a manner consistent with known divisions of power and precision grasps in expert-designed grasp taxonomies[2][8]. With the exception of *Writing Tripod* and *Extension Type*, the division between power and precision grasps are preserved until level-12 (the 12-th iteration) of

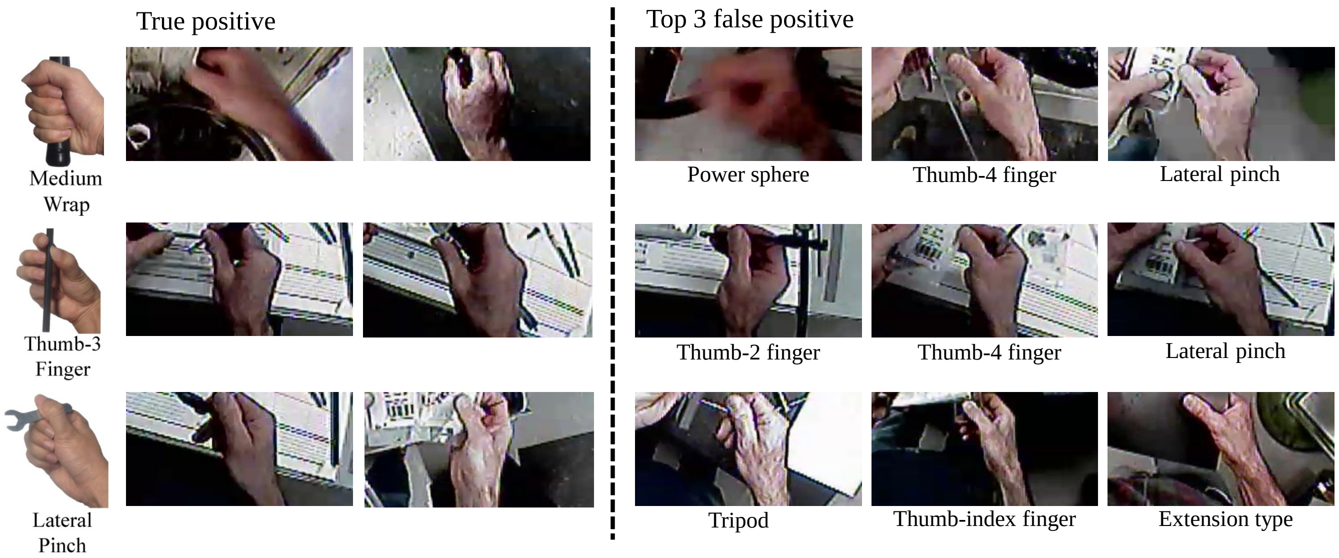


Fig. 6. Examples of true positives and false positives on Machinist Grasp Dataset.

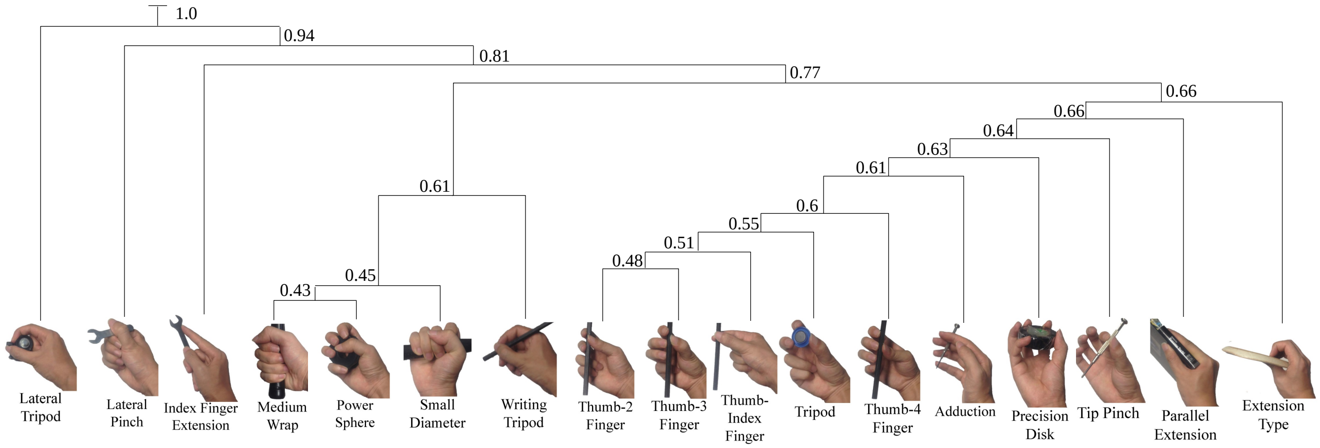


Fig. 9. Dendrogram of hand grasp types. Average F1 scores computed at different abstraction levels are added near each clustering node.

the grasp hierarchy.

The more important observation however is that the visual structures of grasps for the machinist has been learned automatically in a data-driven manner. While classical grasp taxonomies have been created through deep introspection, the shared uncertainty between visual classifiers can also be used to learn intuitive hierarchies over human grasps.

### C. Recognition using Grasp Abstractions

Based on the dendrogram in Fig. 9 it is possible to 'cut' the tree at different levels to obtain different set of grasp clusters. Furthermore, each slice (abstraction) level can be interpreted as a new grasp taxonomy. By learning new grasp classifiers for each category of the new taxonomies, we can achieve a trade-off between more detailed classification and more robust classification. Average F1 scores are computed for grasp recognition at each level of grasp abstractions in Fig. 9. If we utilize a higher level of the tree to define grasp categories, we obtain more reliable grasp classification. For example, for level-12 of the tree, we will be able to differentiate between 5 grasps with an average F1 score of 0.66. On the other hand, choosing level-5 will allows us to differentiate between 12 grasps with an average F1 score of 0.55.

The changes of grasp recognition performance at different levels of the grasp dendrogram is shown in Fig. 10. The average F1 grows up steadily until level-6 since at initial six iterations similar grasp types are being clustered together. From level-7 to level-12, average F1 increases relatively slowly compared to previous steps. For example, average F1 of level-11 and level-12 are almost the same (0.66). This can be explained as newly clustered grasp types become more dissimilar and thus only limited improvement of recognition performance is achieved. Average F1 increases dramatically from level-13 since big grasp clusters are merged together and chance of misclassification is low.

This learned visual structure gives researchers the flexibility of finding a good balance between better performance

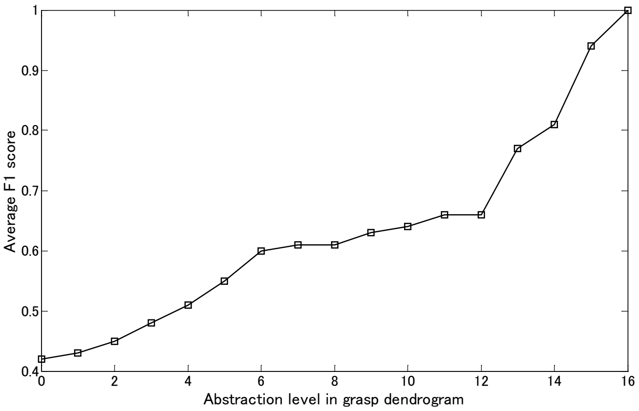


Fig. 10. Grasp recognition performance at different levels of grasp abstractions.

and more detailed grasps analysis.

## V. Conclusion

We proposed a vision-based approach to automate grasp analysis for large amounts of video data. Discriminative classifiers were trained to recognize different grasp types based on computer vision techniques. Visual structures of hand grasps were learned by a supervised grasp clustering method. Our work shows the potential for using computer vision techniques for analyzing hand grasps with large scale of data in real-life settings.

There still exists a lot of work to do to improve grasp recognition performance. Fine-grained grasp recognition is lacking in this paper while it is crucial information for a scientist studying human behavior. The temporal aspect of grasping is obviated in this paper and it would be helpful to impose temporal coherence to improve classification performance. Moreover, explicit object attributes such as weight, shape and size are important factors affecting human grasp selection. We believe a reliable detection framework of object attributes would be very useful in inferring grasp usage. These problems will be addressed in our future work.

## References

[1] 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.

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

[3] J. Case-Smith, C. Pehoski, A. O. T. Association, *et al.*, *Development of hand skills in children.* American Occupational Therapy Association, 1992.

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

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

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

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

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

[9] J. Z. Zheng, S. De La Rosa, and A. M. Dollar, "An investigation of grasp type and frequency in daily household and machine shop tasks," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on.* IEEE, 2011, pp. 4169–4175.

[10] 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.* IEEE, 2013, pp. 1068–1075.

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

[12] 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.

[13] 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.

[14] K. Bernardin, K. Ogawara, K. Ikeuchi, and R. Dillmann, "A sensor fusion approach for recognizing continuous human grasping sequences using hidden markov models," *Robotics, IEEE Transactions on*, vol. 21, no. 1, pp. 47–57, 2005.

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

[16] V. Athitsos and S. Sclaroff, "Estimating 3d hand pose from a cluttered image," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 2. IEEE, 2003, pp. II–432.

[17] H. Hamer, K. Schindler, E. Koller-Meier, and L. Van Gool, "Tracking a hand manipulating an object," in *Computer Vision, 2009 IEEE 12th International Conference On.* IEEE, 2009, pp. 1475–1482.

[18] 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 *Computer Vision (ICCV), 2011 IEEE International Conference on.* IEEE, 2011, pp. 2088–2095.

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

[20] 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.

[21] G. Rogez, J. S. Supancic III, M. Khademi, J. M. M. Montiel, and D. Ramanan, "3d hand pose detection in egocentric rgb-d images," *arXiv preprint arXiv:1412.0065*, 2014.

[22] C. Li and K. M. Kitani, "Pixel-level hand detection in ego-centric videos," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on.* IEEE, 2013, pp. 3570–3577.

[23] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[24] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[25] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in large margin classifiers.* Citeseer, 1999.

[26] I. Bullock, T. Feix, and A. Dollar, "The yale human grasping data set: Grasp, object and task data in household and machine shop environments," 2014.