Hand Parsing for Fine-Grained Recognition of Human Grasps in Monocular Images

Akanksha Saran, Damien Teney and Kris M. Kitani
The Robotics Institute, Carnegie Mellon University
Pittsburgh, PA 15213, USA
asaran@alumni.cmu.edu dteney@andrew.cmu.edu kkitani@cs.cmu.edu

***Abstract*— We propose a novel method for performing fine-grained recognition of human hand grasp types using a single monocular image to allow computational systems to better understand human hand use. In particular, we focus on recognizing challenging grasp categories which differ only by subtle variations in finger configurations. While much of the prior work on understanding human hand grasps has been based on manual detection of grasps in video, this is the first work to automate the analysis process for fine-grained grasp classification. Instead of attempting to utilize a parametric model of the hand, we propose a hand parsing framework which leverages a data-driven learning to generate a pixel-wise segmentation of a hand into finger and palm regions. The proposed approach makes use of appearance-based cues such as finger texture and hand shape to accurately determine hand parts. We then build on the hand parsing result to compute high-level grasp features to learn a supervised fine-grained grasp classifier. To validate our approach, we introduce a grasp dataset recorded with a wearable camera, where the hand and its parts have been manually segmented with pixel-wise accuracy. Our results show that our proposed automatic hand parsing technique can improve grasp classification accuracy by over 30 percentage points over a state-of-the-art grasp recognition technique.**

I. Introduction

The study of human hand usage has been a topic of longstanding interest in the robotics community [1], [2], [3], [4], [5], [6] where research results are typically obtained through many hours of visual observation and thoughtful introspection. Recently, supervised [7] and unsupervised [8] computer vision-based approaches have been proposed in an effort to automate the process of gathering hand use statistics. However, these works are only able to categorize grasps using rough hand shape and cannot resolve differences between many similar looking grasps. In this work, we focus on a subset of grasps with similar appearance, yet differing by subtle finger placement or in the number of fingers involved during manipulation. In contrast to previous work, we show that we can automatically differentiate between grasp types with high accuracy, which was not possible with previous approaches.

Different types of grasps may be functionally different yet visually very similar. For example, we cannot simply use rough hand appearance to differentiate the "lateral tripod" and "medium wrap" [4] (Fig. 4), since they only differ in the detailed placement of the thumb relative to the fingers and object. This example shows the importance of accurately localizing hand parts and identifying their relative placement to disambiguate between certain grasp types. We call this task of differentiating between similar looking yet functionally distinct grasps as *fine-grained* grasp recognition.

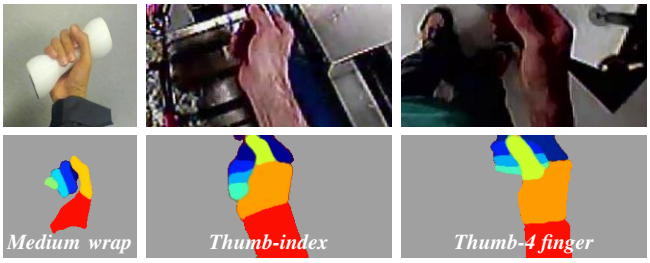


Fig. 1. Our grasps analysis framework infers hand part segmentations to recognize grasps types.

In order to facilitate proper discrimination between fine-grained grasp categories, a visual classification algorithm needs to possess the ability to extract finger location. This requirement leads to a proposed two-stage approach, where fingers are localized in the first stage and features based on relative finger locations are used to classify the grasp in the second stage.

The first stage of our pipeline performs *hand parsing*, which we define as the pixel-level localization and segmentation of the palm and individual fingers of the hand (Fig. 1). We take a data-driven approach similar to the body part estimation of the Kinect [9] to directly learn the mapping from appearance to segmentations (as opposed to defining a parametric hand model). In particular, we use a collection of data-driven predictors which make use of both local hand texture and local shape information to accurately segment hand parts.

The second stage of our pipeline is used to process the output of the hand parsing stage to classify the grasp types. In particular, we build high-order hand features from the results of the hand parsing stage, and use them to classify the observed grasp type (see Fig. 2 for an overview).

In order to capture the true statistics over human hand use, an automated visual grasp classification algorithm will need to be able to observe people engaging in everyday manipulation tasks. This requirement has led to our use of video recorded with wearable cameras. As with previous work [5], [8], [7], wearable cameras can be mounted directly

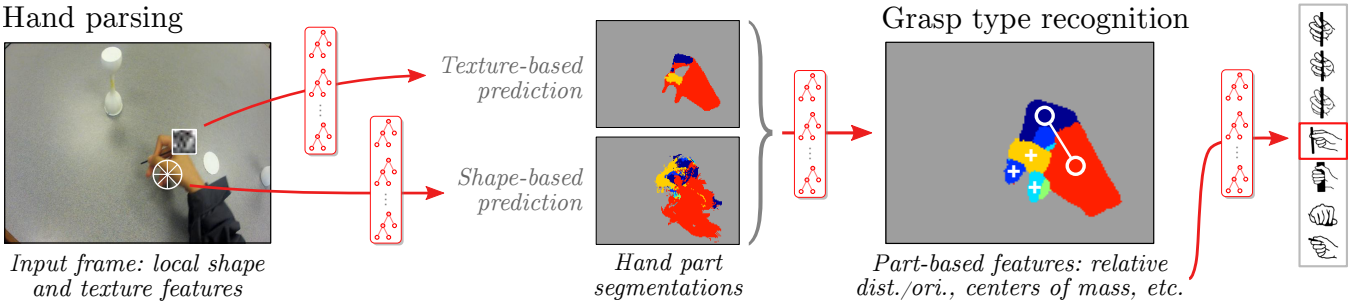


Fig. 2. We use texture and shape cues as input to independently-trained predictors, implemented as random forests. The pixel-level segmentations of hand parts are then fused with another predictor. The resulting segmentation is much more stable and robust to noise than the individual predictions. In the final stage, we extract high-order features (e.g., pairwise relationships between segments) from the hand parse to classify the grasp into learned categories.

on a person and can be used to monitor everyday activities. Moreover, a head mounted camera is a passive sensing technology and does not require sensing elements to be fixed on the hands like data-gloves or IMUs. Furthermore, the egocentric paradigm favorably limits the variability of hand-camera configurations making the visible size and pose of the hand relatively stable. This stability in viewpoint makes egocentric videos particularly amenable to many non-parametric data-driven vision algorithms.

The contributions of this work are threefold:

1) A data-driven image-based hand parsing technique using texture and shape features,
2) Pair-wise hand part features effective for fine-grained grasp recognition,
3) A novel dataset of egocentric images featuring various types of grasps, annotated with hand part locations and grasp types (see Fig. 3 and 7).

We evaluate the performance of our proposed contributions using our new CMU grasp dataset along with the publicly available Yale Human Grasping dataset [5]. Our experiments demonstrate that our proposed approach outperforms the state-of-the-art method for grasp recognition by over 30 percentage points (a significant 50% improvement in recognition accuracy).

II. Related Work

A. Hand Grasp Analysis

The study of hand usage and their interactions with the physical world has attracted much attention across various fields, from robotic arm design [1], [10], to neuromuscular rehabilitation or motor control analysis [11], [12], [13]. The earliest work explored the space of hand manipulation through discrete grasp categories or taxonomies [10], [14]. In the context of robotic manipulation, domain constrained categorical representation of grasps such as Cutkosky and Wright's hand grasp taxonomy [1] played an important role in guiding robotic hand design. In the 1990s, Kang and Ikeuchi [15], [16], [17], [2], [3] presented an important paradigm of using the classification of human grasps (power and precision grasps) to help automate robotic manipulation. More recent work has utilized large amounts of video/image data to understand the scope of grasps [5] and the complexity of everyday object interactions [6].

B. Sensing techniques

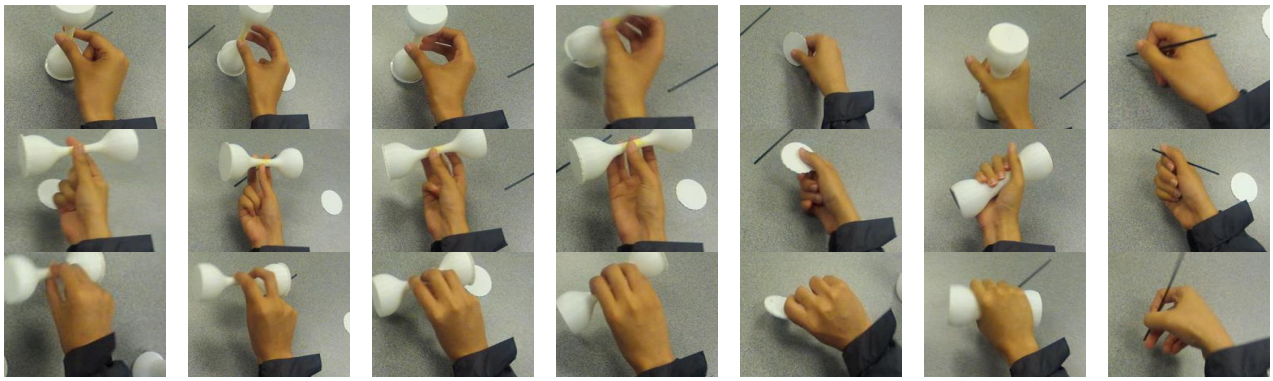
Automated grasp recognition was first studied with data gloves and inertial sensors [18], [13], [2], [3]. Although such sensors can provide detailed measurements of joint angles and finger positions, they must be worn over the hand. Hand-worn devices often inhibit natural hand interactions and can restrict data capture to laboratory settings. Marker-based motion capture [19], [20] allows hands to be in contact with objects but also requires that markers be attached to the hands. An alternative is to use techniques to estimate the pose of the hand directly from appearance. These approaches however often require some form of 3D input data [21], [22], [23], [24], [25]. While 3D sensing technologies such as IR projection-based RGBD cameras or multi-camera systems can aid in estimating hand pose, the footprint of multi-camera setups or the power consumption of IR projectors can be inhibiting for a wearable sensing scenario.

C. Vision-based hand analysis

Much of the work in computer vision has been motivated by gestural interfaces with an emphasis on tracking hand motion and hand gesture recognition. Many classical gesture recognition approaches were designed to differentiate between hand configurations with distinct shape features. In such cases, contour, shape or gradient information was sufficient to recognize various gestures [26], [22], [27].

Generic visual hand trackers have been successfully proposed previously to recover articulated hand pose from images [28], [21], [24], [29], [30]. Oikonomidis *et al.* [31] use hand-object interactions as constraints to fit a high degree-

Fig. 4. Challenges of the visual recognition of grasps: the lateral tripod (**left**) and medium wrap (**right**) have different functions yet look visually similar (Yale human grasping dataset [5]).



Thumb-index | Thumb-2 finger | Thumb-3 finger | Thumb-4 finger | Lateral pinch | Medium wrap | Lateral tripod

Fig. 3. The proposed CMU grasp dataset consists of 7 highly challenging fine-grained grasp categories (frames cropped to show hand close-ups).

of-freedom 3D hand model to the input of a calibrated multi-camera system.

Recently, a number of methods have been designed to detect hands from wearable cameras under changing lighting conditions [32], [33], [34], [35], [36]. While these approaches are effective at detecting and segmenting the hands in egocentric videos, they do not provide the detailed hand part locations which are needed for discriminating between fine-grained grasp categories.

III. Proposed Grasp Analysis Pipeline

Our proposed two stage grasp analysis pipeline is visualized in Fig. 2. In the first stage, a texture-based predictor and a shape-based predictor are used to obtain rough segmentations of the hand image. The output of these two modalities are then merged with a third predictor to generate a segmentation (hand parse) which takes into account both types of appearance cues. In the second stage, high-order features (e.g., pairwise relationships between segments) are extracted from the hand parsing results and are processed by a classifier to predict the grasp category. We describe the details of our proposed approach below.

A. Stage 1: Hand Parsing

We take a non-parametric data-driven approach by learning a direct mapping from an image patch to a segmentation mask – a technique commonly used in semantic scene segmentation [37], [38]. In this work, we use variants of the random forest regressor for two modes of appearance information: texture and shape.

1) Textural Cues for Hand Parsing: We use local texture cues to capture subtle differences of appearance caused by the wrinkling of skin, or the bending and crossing of fingers. For example, in the case of the *medium wrap* grasp (Fig. 6) where all fingers wrap around an object, the aligned fingers generate a distinct line pattern (a textural cue) which can be learned in a data-driven fashion.

Our texture feature is formed by computing color histograms in LUV space and gradients over a 16×16 image patch. Empirical tests showed that color features have a higher contribution for classifying pixels into hand parts, however gradient features also provide a significant boost in output performance. We train a structured output random forest using these input features following [39] to learn a direct mapping from a feature patch to a corresponding segmentation patch. The structured random forest is essentially acting as an efficient nearest neighbor classifier, where the known segmentation masks corresponding to various feature patches are stored at the leaves of the random forest. When the structured random forest is given a 16×16 image patch as input, it returns a corresponding pixel output – a 16×16 hand part segmentation patch.

2) Shape Cues for Hand Parsing: We use shape cues to identify hand parts, such as the fingertips, and the contours of the hand. For example, a small patch centered on a small semi-circular skin region provides strong evidence that we may be observing the tip of a finger. To extract local shape information, we first use the skin detector of Li and Kitani [33]. Their method obtains highly robust hand detection results by training hundreds of skin appearance models and adaptively applying appropriate models depending on the illumination conditions of the scene.

From a binary skin detection map extracted using [33], we build our local shape context descriptor as a spatial indicator feature which encodes the presence or absence of skin over circular perimeter around a particular pixel, similar to [40], [41]. Empirical tests showed that a fixed circular perimeter of radius 20 (pixels) works best using 360 orientation bins. A random forest is used as the predictor which takes as input the shape features over a 16×16 image patch and outputs probabilities of the input pixel belonging to each hand part.

3) Fusing Multi-modal Parsing Results: A final fusion step is critical for assuring a well-balanced hand parsing result, as the segmentation output of each feature modality has different strengths (and weaknesses). The segmentation output of the texture-based classifier tends to be conservative (i.e., high precision), only segmenting hand parts that are easy to identify. On the other hand, the shape-based classifier is more inclusive (i.e., high recall) of all hand parts but at the cost of more noisy segmentation. A third classifier is used to merge the best results of the hand parsing output for each modality (Fig. 7). This usage of a classifier output to the

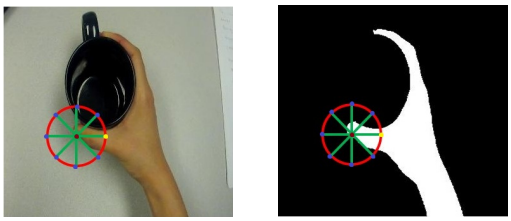


Fig. 5. Shape descriptor. Shape information is encoded by a binary vector, where each element represents the presence of skin-colored pixels along a circular path around a pixel.

next stage is reminiscent of the auto context idea [42] and it's applications to segmentation [38].

To train the fusion model, the probabilistic output of each modality is used as the training data and the ground truth hand part label is used as the training label. The input is a small image patch to a random forest regressor, where each pixel of the patch is represented by a *distribution* over hand parts labels. If we use a 16×16 image patch as input, and there are 6 different hand part labels, the dimensionality of the input data for a single image patch will be 1536 ($16 \times 16 \times 6$). Furthermore, since we are merging the hand parsing results for two modalities, the dimensions of the input features is doubled. The regressor outputs probabilities for the input pixel belonging to each hand part.

B. Stage 2: Fine-grained Grasp Recognition

As the final stage of our proposed approach, we use the output of the hand parsing stage to compute features necessary for predicting the grasp type. Since grasp types are defined by relative configuration of fingers, we introduce a set of high-level features which capture the relative placement of fingers during a grasp. These high-level features are then used as an input to a random forest regressor which outputs 7 class probabilities to classify the grasp type. We describe a set of useful high-level features below.

1) Global Image Representation: The global image representation (GIR) [43] is a histogram counting the number of pixels for each hand part. Each bin of the histogram essentially encodes the (weighted) size of the hand part. The histogram value of the i^{th} hand part $h(i)$ is defined as

$$h(i) = \frac{1}{N} \sum_{n=1}^{N} p_n(i), \tag{1}$$

where $p_n(i)$ is the probability of the n^{th} pixel belonging to the i^{th} hand part, and N is the total number of non-background pixels (high probability background pixels are removed before computing this feature).

2) Center of Mass: The center of mass (CoM) of each hand part (e.g., index finger, thumb, palm) is computed as a weighted average using the output of the hand parsing stage. In particular, for a single hand part i, the first order moment is computed as,

$$c_x(i) = \frac{1}{N} \sum_{n=1}^{N} p_n(i) \cdot x_n, \tag{2}$$

$$c_y(i) = \frac{1}{N} \sum_{n=1}^{N} p_n(i) \cdot y_n, \tag{3}$$

where n is the index of the pixels in the image, N is the total number of pixels in the image and $p_n(i)$ is the probability that the n^{th} pixel belongs to the i^{th} hand part. When there are N hand parts, the CoM feature is a vector of dimension $2N$.

3) Pairwise part distance and orientation: We use the pairwise part distance (PPD) and pairwise part orientation (PPO) [43] to capture the relative spatial relationships between different hand parts. Instead of computing an exact distance between hand parts, this feature computes a statistical approximation of the distance/orientation by computing a probabilistically weighted value over a set of sparse keypoints.

First, a collection of 500 sparse keypoint locations (drawn from a uniform distribution) are used to represent the distribution of hand parts. Second, the pairwise (weighted) distance between each pair of points is computed using the following equation,

$$PPD(i,j) = \sum_{n,m} p_n(i) \cdot p_m(j) \cdot D(n,m) \tag{4}$$

where i and j are the indicies of two hand parts, $D(n,m)$ is the Euclidean distance between the n^{th} and m^{th} keypoints. The PPO is computed in a similar fashion, where $D(n,m)$ now represents the angle between the line joining n and m, and the x axis of the image.

In summary, we partition our entire pipeline into four different random forests - the first three perform segmentation and the last one classifies grasp types. Even though other partitions are plausible approaches for training the objective end-to-end, simpler pipelines would loose the flexibility of using different forests (classic vs structured), tailoring the intermediate features and the auto context effect [42]. These abilities of our pipeline prove advantageous in capturing subtle differences in hand grasps.

IV. Experimental Validation

A. Datasets

To evaluate the accuracy and robustness of our proposed approach we utilize two datasets. The first dataset, the CMU grasp dataset, is a densely labeled grasp dataset recorded in a moderately controlled office environment. This dataset is used to evaluate our various design choices and to carefully measure the accuracy of our proposed approach. The second dataset, the Yale Human Grasping dataset [5], is a temporally labeled dataset recorded in real-world situations. This dataset is used to quantify the robustness of our approach on unconstrained videos.

Thumb index | Thumb 2 finger | Thumb 3 finger | Thumb 4 finger | Lateral pinch | Medium wrap | Lateral tripod

Fig. 6. Grasp categories used in this work. Labels are taken from from Feix *et al.*'s taxonomy [4]. These grasps cover a wide range of object-hand interactions, yet differ only by small differences in finger placement.

We introduce the densely annotated CMU grasp dataset for detailed quantitative analysis of hand parsing and fine-grained grasp recognition. The dataset contains 7 different fine-grained grasp types recorded in a controlled environment: (1) lateral pinch, (2) lateral tripod, (3) medium wrap, (4) thumb-index finger, (5) thumb-2 finger, (6) thumb-3 finger, and (7) thumb-4 finger (Fig. 6). We selected these grasps as they share similar appearance but differ by slight changes in finger positions. The video is captured with a head-mounted GoPro Hero 2 camera at a HD resolution of 1920×1080 by one user. In each grasp sequence, the user transports a set of abstract objects designed to allow for several grasping strategies. The user also moves during the entire grasp sequence inducing moderate camera ego-motion and variations in viewpoint. Image pixels of the dataset are labeled with a hand part (or background label) and the entire image is labeled with a grasp type. The CMU grasp dataset consists of over 945 image labels and over 25 million labeled pixels.

We also use the Yale Human Grasping Dataset [5] to validate our proposed approach. We use a subset of videos of this dataset for one user and test our method for seven grasp types (Fig. 6) out of the total 17 grasps annotated in this dataset. The Yale dataset is recorded with a wearable camera (VGA resolution) in real-world scenarios captured by a machinist in a workshop and by a housekeeping staff member in a hotel (Fig. 4). Since the video is recorded 'in the wild', the recognition task using computer vision is extremely challenging, as illumination conditions change and background is cluttered with real-world objects. It is also the only publicly available video dataset with grasp annotations defined over short temporal windows. Hand locations and hand parts are not labeled as part of this dataset. Other hand gesture datasets used for vision-based hand analysis do not contain traditional grasp categories [44], [45] or do not have grasp labels since they were designed primarily for object and action recognition [46], [44].

B. Evaluation of Hand Parsing

To show the importance of using our multi-modal hand parsing algorithm (the first stage of our proposed grasp recognition pipeline), we perform ablative analysis to contrast our proposed approach against models which only use a single modality (texture or shape). More specifically, we perform tests on the CMU grasp dataset using three hand parsing models: (1) texture-only, (2) shape-only and (3) our proposed multi-modal feature model. We use a 80/20 (training/testing) split to evaluate each hand parsing model for our experiments. Since the output of each model is a probabilistic distribution over labels at each pixel, we associate a pixel to the label (hand part) that has the highest probability to compute the F measure.

TABLE I

HAND PARSING PERFORMANCE ON THE CMU GRASPING DATASET (F-MEASURE)

	Texture only	Shape only	Proposed
Index finger	0.14	0.25	**0.61**
Middle finger	0.10	0.01	**0.44**
Ring finger	0.32	0.15	**0.61**
Pinky finger	0.64	0.26	**0.65**
Thumb	0.43	0.30	**0.72**
Palm	0.79	0.65	**0.85**
Average	0.40	0.27	**0.64**

The F-measure are shown in Table I. The texture-based predictor has an average F-measure of 0.40 and the local shape predictor has an F-measure of 0.27. By combining the weak information from these two modes, performance increases significantly to 0.64. This result shows that our proposed multi-modal fusion framework is necessary for better hand parsing performance.

Qualitative results for hand parsing are shown in Fig. 7. The results illustrate the smoothing effect of the fusion step when compared to the noisy output of the individual texture and shape-based hand parsing results. Robustness of the segmentation results to varying illumination conditions rely on the training data. Increased diversity of global illumination in the training data qualitatively improves performance.

C. Evaluation of Grasp Recognition

We evaluate the ability of our grasp recognition technique to differentiate between fine-grained grasp categories. We compare our results to a state-of-the-art approach by Cai et al. [7] which uses a HOG descriptor masked with the results of a hand detection algorithm [33]. We also perform ablative analysis over the high-level features introduced in Section III-B to show how each feature contributes to overall recognition performance. For each experiment a random forest is used to estimate the grasp category label using the respective feature representation as input. The data is partitioned into a 80/20 (training/testing) split and average accuracies are computed using cross-validation.

The average recognition performance over both the CMU and Yale datasets for 7 grasp types (Fig. 6) are included in Table II. The pairwise part orientation feature performs the best with 81% accuracy when used in isolation. The GIR feature, which simply encodes the size of each hand part performs the worst with an accuracy of 75%. A similar trend was observed on the Yale grasp dataset. As expected, the absolute performance on the Yale dataset is lower than the CMU dataset because the Yale dataset contains challenging real-world human activities.

The recognition performance over each grasp type on the CMU dataset is given in Table III. We observe that our

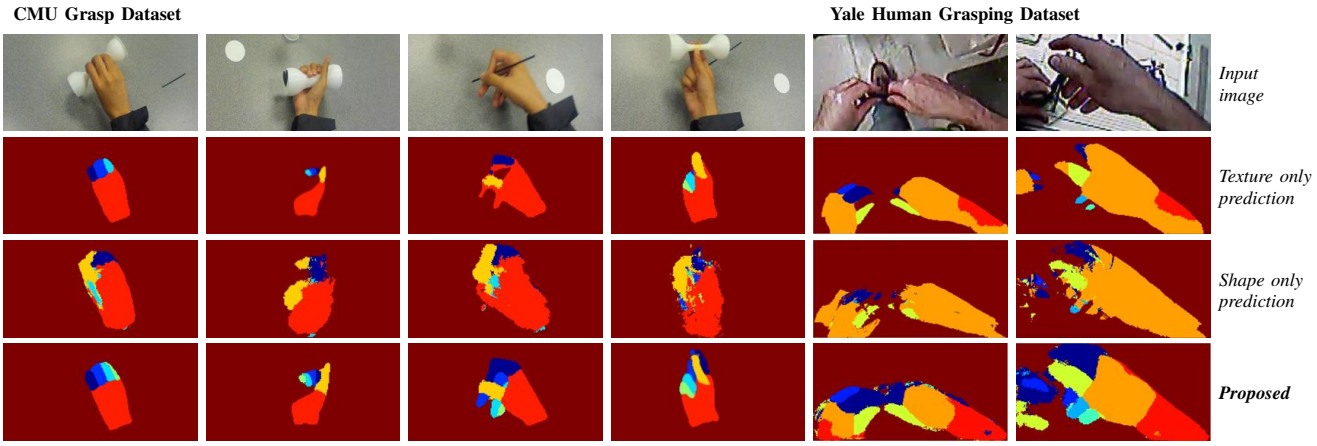


Fig. 7. Qualitative comparison of our ablative analysis. Our proposed multi-modal model combines the strengths of texture and shape predictions to generate an improved hand parsing result.

TABLE II
ABLATIVE ANALYSIS OVER FEATURE REPRESENTATIONS (ACCURACY)

	CMU	Yale
Global Image Representation	75%	41%
Probabilistic Center of Mass	80%	45%
Pairwise Part Distance	80%	46%
Pairwise Part Orientation	81%	48%

TABLE III
RECOGNITION PERFORMANCE PER GRASP TYPE ON **CMU** DATASET (ACCURACY)

	Masked HOG [7]	Proposed
Lateral Pinch	**100**%	92%
Lateral Tripod	80%	**93**%
Medium Wrap	45%	**95**%
Thumb-2 finger	78%	**91**%
Thumb-3 finger	11%	**77**%
Thumb-4 finger	91%	**93**%
Thumb-index finger	23%	**93**%
Average	61%	**91**%

proposed approach using *all features* in Section III-B yields significant improvements for certain grasps such as 'Thumb-3 finger' which increases accuracy by 66 percentage points or 'Thumb-index finger' where accuracy improves by 70 percentage points. This result shows that our use of hand parsing results as an intermediate representation of grasp type has a very beneficial impact on the grasp recognition performance.

V. CONCLUSION

We have proposed a grasp analysis pipeline for recognizing human grasps in monocular videos recorded by a wearable camera. We described a two stage algorithm that detects hand parts (hand parsing) in the first stage and aggregates that data in the second stage to determine the grasp type. We showed through experiments that the first-stage hand parsing technique is able to accurately segment individual hand parts. Furthermore, we showed that multi-modal inputs (both texture and shape) are needed for robust performance. It was also shown that high-level features based on reliable hand parsing results are critical for the success of fine-grained grasp recognition. Our experiments showed that our proposed approach can improve over the state-of-the-art by over 30 percentage points by using such high-level features. We evaluated our approach on two grasping datasets and showed that our approach is able to discriminate between visually similar, yet functionally different grasp types.

VI. DISCUSSION

Our proposed approach shows improvement in performance over a state of the art technique for grasp recognition, though it is limited in terms of: (1) egocentric views of monocular cameras, (2) single user grasp recognition and (3) evaluation of 7 grasp categories only (Fig. 6).

The motivation to choose the egocentric viewing perspective was the suitability of wearable cameras to record activities of daily living and capture stable size and pose of hands. Wearable cameras are therefore favorable to study human grasps. In general, the proposed approach is applicable to all monocular views but will require further evaluation.

Cross-user generalization of grasp recognition is a challenge due to the difference in shape, color and sizes of hands across people. Our work reports high grasp recognition performance for a single user. In future work, we will evaluate if our approach accounts for all possible variabilities in hand characteristics and quantitatively generalizes to multi-user grasp recognition.

The 7 grasps chosen for evaluation in this paper (Fig. 6) are visually similar with subtle differences, which makes it challenging for manual as well algorithmic inspection. These grasps have also been shown to have a high grasp span (versatility to handle a wide range of objects) by Bullock *et al.*[5]. However the 17 grasp categories used in other works [5], [7] include more visually distinct grasp types and evaluating the scalability of this approach to more grasp types

will be a direction for future work.

ACKNOWLEDGEMENT

This research was supported in part by a JST CREST grant.

REFERENCES

[1] M. R. Cutkosky, "On grasp choice, grasp models, and the design of hands for manufacturing tasks," *IEEE Transactions on Robotics and Automation*, 1989.

[2] S. B. Kang and K. Ikeuchi, "A robot system that observes and replicates grasping tasks," in *International Conference on Computer Vision (ICCV)*, 1995.

[3] S. B. Kang and K. Ikeuchi, "Toward automatic robot instruction from perception-mapping human grasps to manipulator grasps," *IEEE Transactions on Robotics and Automation*, 1997.

[4] 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.

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

[6] Y. N. J. Liu, F. Feng and N. S. Pollard, "A taxonomy of everyday grasps in action," in *IEEE International Conference on Humanoid Robots*, 2014.

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

[8] D.-A. Huang, W.-C. Ma, M. Ma, and K. M. Kitani, "How do we use our hands? discovering a diverse set of common grasps," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[9] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

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

[11] J. M. Elliott and K. Connolly, "A classification of manipulative hand movements," *Developmental Medicine & Child Neurology*, no. 3, 1984.

[12] C.-S. J. and P. C., "Development of hand skills in children," in *American Occupational Therapy Association*, 1992.

[13] I. Dejmal and M. Zacksenhouse, "Coordinative structure of manipulative hand-movements facilitates their recognition," *IEEE Transactions on Biomedical Engineering*, 2006.

[14] J. R. Napier, "The prehensile movements of the human hand," *Journal of bone and joint surgery*, 1956.

[15] S. B. Kang and K. Ikeuchi, "A framework for recognizing grasps," Tech. Rep., 1991.

[16] S. B. Kang and K. Ikeuchi, "Grasp recognition using the contact web," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1992.

[17] S. B. Kang and K. Ikeuchi, "A grasp abstraction hierarchy for recognition of grasping tasks from observation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1993.

[18] S. Ekvall and D. Kragic, "Grasp recognition for programming by demonstration," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2005.

[19] P. G. Kry and D. K. Pai, "Interaction capture and synthesis," *ACM Transactions on Graphics (TOG)*, 2006.

[20] N. S. Pollard and V. B. Zordan, "Physically based grasping control from example," in *Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation*, 2005.

[21] V. Athitsos and S. Sclaroff, "Estimating 3D hand pose from a cluttered image," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2003.

[22] C. Schwarz and N. da Vitoria Lobo, "Segment-based hand pose estimation," in *Canadian Conference on Computer and Robot Vision (CRV)*, 2005.

[23] H. Hamer, K. Schindler, E. Koller-Meier, and L. J. V. Gool, "Tracking a hand manipulating an object," in *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[24] C. Keskin, F. Kıraç, Y. E. Kara, and L. Akarun, "Hand pose estimation and hand shape classification using multi-layered randomized decision forests," in *European Conference on Computer Vision (ECCV)*, 2012.

[25] P. Doliotis, V. Athitsos, D. Kosmopoulos, and S. Perantonis, "Hand shape and 3D pose estimation using depth data from a single cluttered frame," in *Advances in Visual Computing*, 2012.

[26] E.-J. Ong and R. Bowden, "A boosted classifier tree for hand shape detection," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2004.

[27] T.-T.-H. Tran and T.-T.-M. Nguyen, "Invariant lighting hand posture classification," in *IEEE International Conference on Progress in Informatics and Computing (PIC)*, 2010.

[28] J. M. Rehg and T. Kanade, "Digiteyes: Vision-based hand tracking for human-computer interaction," in *IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, 1994.

[29] J. Romero, H. Kjellström, C. H. Ek, and D. Kragic, "Non-parametric hand pose estimation with object context," *Image and Vision Computing*, 2013.

[30] B. Stenger, A. Thayananthan, P. H. Torr, and R. Cipolla, "Model-based hand tracking using a hierarchical bayesian filter," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2006.

[31] I. Oikonomidis, N. Kyriazis, and A. Argyros, "Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints," in *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[32] M. Kölsch and M. Turk, "Robust hand detection." in *IEEE International Conference on Automatic Face and Gesture Recognition (FGR)*, 2004.

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

[34] A. Fathi, A. Farhadi, and J. M. Rehg, "Understanding egocentric activities," in *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[35] 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 (CVPR) Workshops*, 2014.

[36] C. Li and K. M. Kitani, "Model recommendation with virtual probes for egocentric hand detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[37] P. Kontschieder, S. Rota Bulo, H. Bischof, and M. Pelillo, "Structured class-labels in random forests for semantic image labelling," in *, IEEE International Conference on Computer Vision (ICCV)*, 2011.

[38] D. Munoz, J. A. Bagnell, and M. Hebert, "Stacked hierarchical labeling," in *European Conference on Computer Vision (ECCV)*, 2010.

[39] P. Dollár and C. L. Zitnick, "Structured forests for fast edge detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[40] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2002.

[41] G. Mori, S. Belongie, and J. Malik, "Shape contexts enable efficient retrieval of similar shapes," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001.

[42] Z. Tu, "Auto-context and its application to high-level vision tasks," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008.*, 2008.

[43] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar, "Food recognition using statistics of pairwise local features," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[44] T.-K. Kim, K.-Y. K. Wong, and R. Cipolla, "Tensor canonical correlation analysis for action classification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[45] Y. Hsiao, J. Sanchez-Riera, T. Lim, K. Hua, and W. Cheng, "Lared: a large RGB-D extensible hand gesture dataset," in *Multimedia Systems Conference*, 2014.

[46] X. Ren and M. Philipose, "Egocentric recognition of handled objects: Benchmark and analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.