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@cmu.edu dteney@andrew.cmu.edu kkitani@cs.cmu.edu

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. 3), 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.

In order to facilitate proper discrimination between fine-grained grasp categories, we propose a visual classification algorithm to extract finger locations - a 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. To validate our approach, we introduce a grasp dataset recorded with a wearable camera (945 image labels), where the hand and its parts have been manually segmented with pixel-wise accuracy (over 25 million labeled pixels). 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.

II. PROPOSED GRASP ANALYSIS PIPELINE

Our proposed two stage grasp analysis pipeline is visualized in Fig. 2. In the first stage, texture and shape-based predictors are used to obtain rough segmentations of the hand. The output of these two modalities are then merged with a third predictor to generate a final segmentation (hand parse). In the second stage, high-order features are extracted using hand parsing results and are processed by a classifier to predict the grasp category.

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 [9], [10]. 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. Our texture feature is formed by computing color histograms in LUV space, and histogram of gradients (HOG) features over a 16 × 16 image patch. We train a structured output random forest following [11] 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...
then used as an input to a random forest to classify the grasp type. We describe a set of useful high-level features below.

1) Global Image Representation: The global image representation (GIR) \cite{15} 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

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

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,

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

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

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) \cite{15} to capture the relative spatial relationships between different hand parts. This feature computes a statistical approximation of the distance/orientation by using a probabilistically weighted value over a set of sparse keypoints. The pairwise (weighted) distance between each pair of points is computed using the following equation,

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

where $i$ and $j$ are the indices 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 $D(n,m)$
Fig. 5. 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.

<table>
<thead>
<tr>
<th>TABLE I</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HAND PARSING PERFORMANCE ON THE CMU GRASPING DATASET (F-MEASURE)</strong></td>
</tr>
<tr>
<td>Texture only</td>
</tr>
<tr>
<td>-------------</td>
</tr>
<tr>
<td>Index finger</td>
</tr>
<tr>
<td>Middle finger</td>
</tr>
<tr>
<td>Ring finger</td>
</tr>
<tr>
<td>Pinky finger</td>
</tr>
<tr>
<td>Thumb</td>
</tr>
<tr>
<td>Palm</td>
</tr>
<tr>
<td>Average</td>
</tr>
</tbody>
</table>

now represents the angle between the line joining $n$ and $m$, and the $x$ axis of the image.

### III. EXPERIMENTAL VALIDATION

#### A. Evaluation of Hand Parsing

To show the importance of using our multi-modal hand parsing algorithm, 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 train 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.

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 are shown in Fig. 6. 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.

#### B. 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 Cai et al. [7] which uses a HOG descriptor masked with the results of a hand detection algorithm [12]. The recognition performance over each grasp type on the CMU dataset is given in Table II. We observe that our proposed approach using all features in Section II-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 has a very beneficial impact on the grasp recognition performance.

### IV. CONCLUSION

We have proposed a grasp analysis pipeline for recognizing human grasps in monocular videos recorded by a wearable camera. A two stage algorithm detects hand parts in the first stage and aggregates that data in the second stage to determine the grasp type. Experiments show 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 proposed approach using such high-level features outperforms the state-of-the-art by over 30 percentage points. We evaluated our approach on a novel dataset and showed that can discriminate between visually similar, yet functionally different grasp types.

### REFERENCES


Fig. 6. 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.


