2D Hand Parsing for Egocentric Gesture Recognition

Akanksha Saran
Carnegie Mellon University
Pittsburgh, PA, USA
asaran@andrew.cmu.edu

Kris M. Kitani
Carnegie Mellon University
Pittsburgh, PA, USA
kkitani@cs.cmu.edu

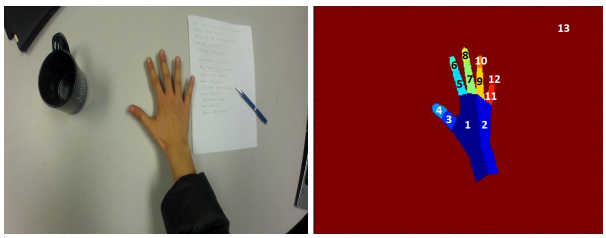


Figure 1. Ground truth labeling scheme of the twelve hand parts and background

The egocentric viewing paradigm presents a novel sensing perspective to observe hand grasps and gestures up-close. In particular, it provides an opportunity to study the finer details of finger placement during hand use. In this work, we explore the task of 2D hand parsing – recognizing the parts of a hand – from a first person point of view and analyze its impact on the task of micro-scale gesture recognition. We show that along with using macro-level hand shape features, which provide a coarse differentiation of hand grasps and gestures, the use of micro-level hand part details can help in the task of gesture and grasp recognition. Specifically, hand parsing information can provide fine finger placement detail which can help distinguish between similarly shaped gestures and grasps. To capture contextual information of hand parts (the possible finger configurations), we propose the use of a structured output regression model to predict pixel-wise hand parts and compute probabilistic hand part features from 2D color images. We observe the impact of using these hand part features along with shape features on two datasets : basic grasping tasks for automated neuromusclar rehabilitation, and alphabets from the American Sign Language. The experiments show that on average, grasps are detected better after merging hand part features with shape features.

Hand Shape Features. Macro-level hand shape features provide coarse differentiation between hand grasps and gestures. As a pre-processing step to hand shape analysis, we detect hands in our egocentric datasets using the per-pixel hand detection approach proposed by Li and Kitani [3]. Once the hand is reliably detected, the shape of the hand is represented by the seven Flusser Suk moments [2]

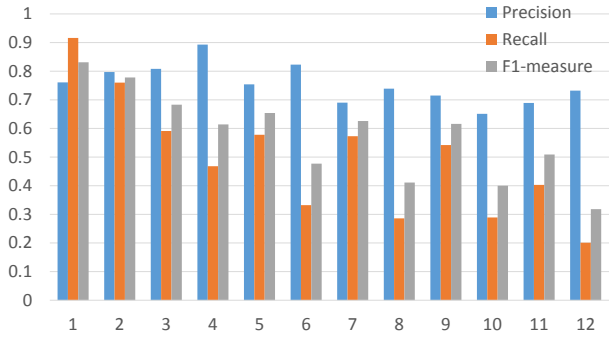


Figure 2. Precision, Recall and F1-metrics for all the 12 hand parts on the entire data set of 245 images after training the Structured Output Random Forest with about 70 images from the data set. The finger tips are the hardest to detect among all hand parts.

and 225 Haar wavelet responses as described by Minnen and Zafrulla [4]. These features can handle rotation invariance(+/-45 degrees). Although the features in [4] use depth-related features, we do not use any 3-D features but only the 232 2-D features in our approach.

Hand Part Features. Hand parsing on a per-pixel level of 2D hand images can be looked at from an object-detection perspective. Inspired by Dollar and Zitnick [1] we use the structured output random forest to learn the hand part labels in a data-driven manner. The most important aspect of this approach is the contextual information it captures while learning the segmentation labels for the hand parts. The input to the regressor is a feature vector with 7228 pixel lookup and pairwise difference features [1] corresponding to a 32x32 patch. The corresponding 16x16 ground truth segmentation mask (the output) is reduced to a single dimension using PCA on a 256 dimensional binary vector describing random pixel pair segmentation memberships to similar or different segments. Output from different trees and different patches are used to determine pixel-wise probabilistic hand part estimates. For every pixel there are several votes for hand part labels. The most likely hand part label is assigned to each pixel as shown in Figure 5. The hand parsing features extracted are the weighted mean coordinates for each of the 12 hand parts (Figure 1).

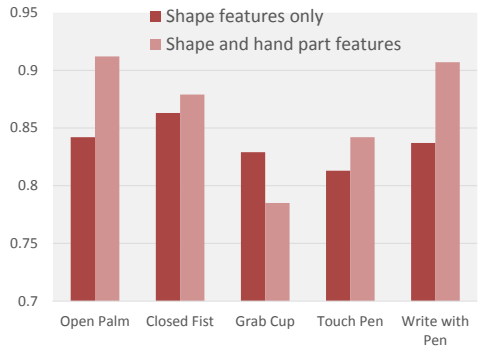


Figure 3. Comparison of F1-scores with and without hand parsing features for nuero-muscular rehabilitation grasps. Apart from the grab cup grasp, we see a boost in detection results for the other grasps after including hand part features. The grab cup grasp seems to obtains more noisy features than informative ones probably due to the small area of the detected hand.

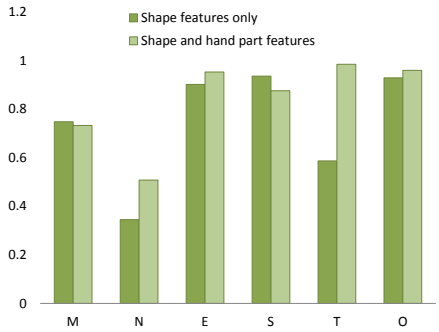
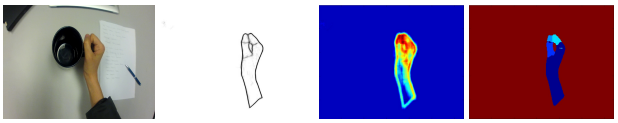


Figure 4. Comparison of F1-scores with and without hand parsing features for American Sign Language gestures. The gestures N and T have similar hand shape but differ just in the placement of the thumb. The hand part information boosts performance for these two gestures in particular. The noise in the hand part features might have also caused the performance of well-detected gestures by shape such as S and M to decrease slightly. Lack of training data comparable with the feature dimensionality for the random forest regressor might also be a reason.

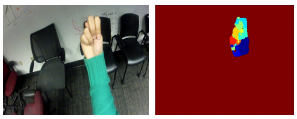
Gesture Recognition. We perform gesture and grasp recognition using hand shape and hand part features. We train a random forest regressor for each class of gestures using hand shape features alone as well as with hand shape and hand part features together. The 232 shape features alone serve as the baseline approach in our experiments. After including the 24 hand part features, the cumulative 256-dimensional feature vector is used as input to the gesture recognition random forest regressor. The output of each regressor is the probability of being detected as a particular gesture.

Experiments and Results

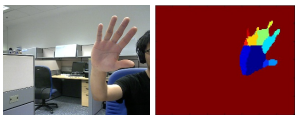
To capture the wide domains of interactions possible with a wearable ego-centric camera, we organize our experiments to look at both the natural hand-object interactions on a flat table surface as well as free hand gestures. First, to analyze how effective the approach is at detecting different hand parts, we use a data set of 245 images consisting of simple open and close hand gestures (Figure 5(c)). Figure 2 shows the precision, recall and F1-score for each of



(a) Neuro-muscular rehabilitation grasp and corresponding edge map, entropy map and per-pixel hand part predictions



(b) American Sign Language



(c) Sample output for simple gestures

Figure 5. Example output of the per-pixel hand part predictions on our data sets

the hand part labels detected. We perform our experiments using one structured output decision tree.

Then we employ two datasets for the task of gesture recognition : (1) automated neuro-muscular rehabilitation (Figure 5a) and (2) American Sign Language (Figure 5b). Egocentric videos have been used in the past to monitor rehabilitation progress with hand contour information [5]. Common neuromuscular grasps capture the interaction with objects on a flat table surface. The F1 scores for a dataset of five such grasps are shown in Figure 3 with and without hand part features. Similarly Figure 4 shows the improvement in F1-scores for the six American sign language gestures after including the hand part features to the shape features.

Conclusion and Future Work

To quantify the differences between highly deformable and similarly appearing hand parts, we propose the use of a per-pixel prediction technique for hand part detection using structured output regressor. This approach aids the process of hand grasp and gesture recognition. Our data-driven approach captures fine differences based on finger placements and aids the process of recognizing similarly shaped gestures effectively. For future work we will incorporate variants of hand parsing features for gesture recognition and cross-dataset evaluations.

References

[1] P. Dollár and C. L. Zitnick. Structured forests for fast edge detection. In *International Conference on Computer Vision (ICCV)*, 2013. 1

[2] J. Flusser and T. Suk. Rotation moment invariants for recognition of symmetric objects. *IEEE Transactions on Image Processing*, 15(12):3784–3790, 2006. 1

[3] C. Li and K. M. Kitani. Pixel-level hand detection in ego-centric videos. In *Computer Vision and Pattern Recognition (CVPR)*, 2013. 1

[4] D. Minnen and Z. Zafrulla. Towards robust cross-user hand tracking and shape recognition. In *Computer Vision Workshops (ICCV Workshops)*, 2011. 1

[5] J. Zariffa and M. R. Popovic. Hand contour detection in wearable camera video using an adaptive histogram region of interest. *Journal of neuroengineering and rehabilitation*, 10(1):114, 2013. 2