Grasp Type Revisited: A Modern Perspective on A Classical Feature for Vision

Yezhou Yang[1], Cornelia Fermüller[1], Yi Li[2], Yiannis Aloimonos[1]
[1]Computer Vision Lab, University of Maryland, College Park. [2]NICTA and ANU.

The grasp type provides crucial information about human action. In this paper we present a study centered around human grasp type recognition and its applications in computer vision. The goal of this research is to provide intelligent systems with the capability to recognize the human grasp type in unconstrained static or dynamic scenes. To be specific, our system takes in an unconstrained image patch around the human hand, and outputs which category of grasp type is used . In the rest of the paper, we show that this capability 1) is very useful for predicting human action intention and 2) helps to further understand human action by introducing a finer layer of granularity. Further experiments on two publicly available dataset empirically support that we can 1) infer human action intention in static scenes and 2) segment videos of human manipulation actions into finer segments based on the grasp type evolution. Additionally, we provide a labeled grasp type image data set and a human intention data set for further research.

Human Grasp Types: We use a categorization into seven grasp types. First we distinguish, according to the most commonly used classification (based on functionality), into power and precision grasps. We then further distinguish among the power grasps, whether they are cylindrical, spherical, or hook. Similarly, we distinguish the precision grasps into pinch, tripodal and lumbrical. Additionally, we also consider a Rest or Extension position (no grasping performed). Fig. 1 illustrates the grasp categories.

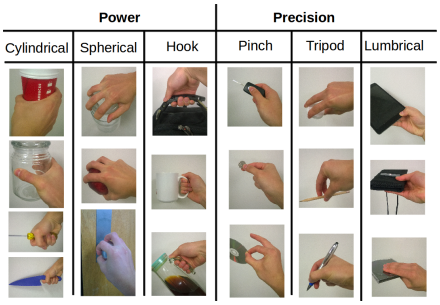

Figure 1: The grasp types considered. Grasps which cannot be categorized into the six types here are considered as the "Rest and Extension" (no grasping performed).

CNN for Grasp Type Recognition: We used a five layer CNN (including the input layer and one fully-connected perception layer for regression output) for grasp type classification. We achieved an average of 59% classification accuracy using the CNN based method, and showed that it outperforms hand-crafted feature based baseline methods. Fig. 2 shows some correct grasp type predictions (denoted by black boxes), and some failure examples (denoted by red and blue bounding boxes). Blue boxes denote a correct prediction of the underlying high-level grasp type in either the "Power" or "Precision" category, but incorrect recognition in finer categories. Red boxes denote a confusion between"Power" and "Precision".

From Grasp Type to Action Intention: Our hypothesis is that the grasp type is a strong indicator of human action intention. In order to validate this, we train an additional classifier layer for recognizing human action intention. We choose here a categorization into three human action intentions ("Force-oriented", "Skill-oriented" and "Casual"), closely related to the functional classification discussed above (Fig. 1). A subset of images from the Oxford hand dataset serves as testing bed for action intention classification and we achieved an average 65% prediction accuracy. Fig. 3 shows some interesting correct cases.

Finer segment action using grasp type evolution: In manipulation actions involving tools and objects, the dynamic changes of grasp type characterize the start and end of these finer actions. We labeled each hand with the grasp type of highest belief score in each frame. After applying a one dimensional mode filtering for temporal smoothing, we segmented the action whenever one hand changes grasp type. Fig. 4 shows two examples of intermediate grasp type recognition and the detected segmentation. Using the grasp type temporal evolution, we achieved 78% recall and 80% precision in fine grain manipulation action segmentation tasks.

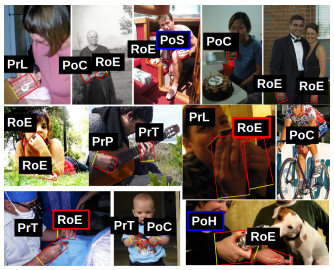

Figure 2: Examples of correct and false grasp type classification. PoC: Power Cylindrical; PoS: Power Spherical; PoH: Power Hook; PrP: Precision Pinch; PrT: Precision Tripod; PrL: Precision Lumbrical; RoE: Rest or Extension.

Figure 3: Examples of predicting action intention.

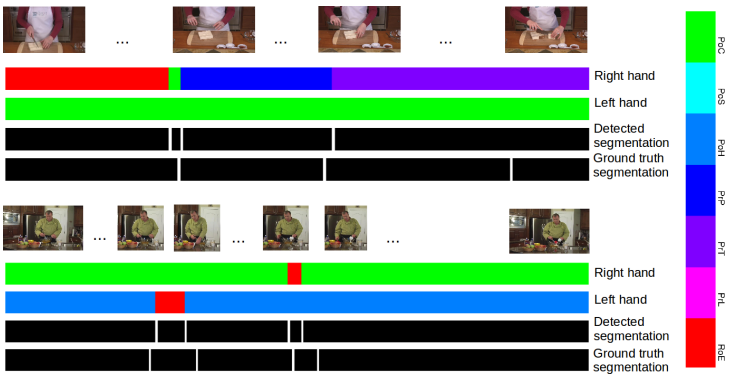

Figure 4: Grasp type recognition along timeline and video segmentation results compared with ground truth segments.

Conclusion: Recognizing grasp type and its use in inference for human action intention and fine level segmentation of human manipulation actions, are novel problems in computer vision. We have proposed a CNN based learning framework to address these problems with decent success. We hope our contributions can help advance the field of static scene understanding and human action fine level analysis, and we hope that they can be useful to other researchers in other applications. Additionally, we augmented a currently available hand data set and a cooking data set with grasp type labels, and provided human action intention labels for a subset of them. We will make this augmented data sets available for future research.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.