# Abstract

Understanding human behavior by identifying hand gestures and grasps from images/videos captured by consumer-grade cameras has recently gained a lot of research interest. However, in most cases, research works either focus on detecting and localizing hands or directly uses images or video frames for recognition of gestures and grasps without considering the hand location information. In this thesis, we initially address the task of hand localization and subsequently tackle hand gesture as well as grasp recognition.

In the first part of our work, we focus on detecting and segmenting hands in various scenarios including unconstrained environment, cluttered background, egocentric vision, interaction with another person, computer, and various objects. For this purpose, we first propose four deep learning-based skin segmentation algorithms which aid in removal of false-positives from hand detection results.

For the hand detection problem, we propose four algorithms based on RCNN, Faster-RCNN, multiple-scale Faster-RCNN, and RetinaNet. We initially investigated state-of-the-art object detectors such as RCNN and Faster-RCNN for hand detection and find that performance can be boosted if we use a multiple-scale architecture. We also investigate the recently proposed RetinaNet which is inherently multi-scale and utilizes focal loss.

At the end of the first part of the thesis, we also investigate a problem similar to hand detection, namely, hand segmentation. Hand segmentation provides finer location information than hand detection. For this task, we propose a U-Net-based hand segmentation algorithm utilizing dense layers and dilated layers in both supervised and semi-supervised modalities.

In the second phase of this thesis, we use prior information about hand location which we determined in the first part of our work to develop more complex applications such as static gesture recognition, grasp, and dynamic gesture classification. We devised a method to classify grasps from images using bilinear label-dependent low rank sparse feature learning.

We also investigate the static hand gesture recognition problem where subjects interact with the computer or another person. Towards this end, we propose two algorithms which take segmented hands as input. First, we propose a static hand gesture recognition method based on structurally incoherent non-negative matrix factorization. Next, we present a supervised dictionary training algorithm which is utilized to acquire a low rank, non-negative, sparse, discriminative image feature representation for classification.

Finally, in the thesis we consider dynamic hand gestures where motion plays an important role and we exploit prior information about hand pose and body posture to aid more accurate classification. We experimented with publicly available benchmark datasets for grasp understanding, static gesture recognition, dynamic gesture recognition and achieve state-of-the-art results.

***Index terms***— Multiple-scale Faster-RCNN, adversarial learning, dilated convolution, structurally incoherent NMF, pose-based CNN feature, long short term memory, bilinear feature, low rank, sparsity, non-negative matrix factorization, KL-divergence