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

Xuewen Yang

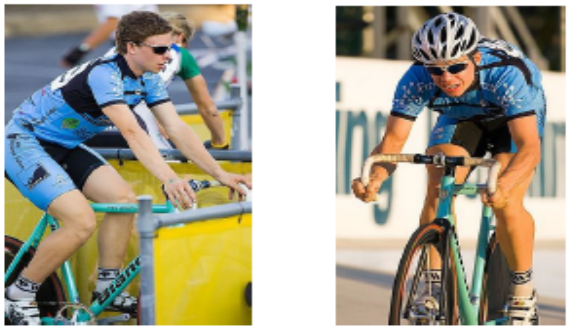
July 16 2018

Abstract

The grasp type provides crucial information about human action. However, recognizing the grasp type from unconstrained scenes is challenging because of the large variations in appearance, occlusions and geometric distortions.In this paper, first the author present a convolutional neural network to classify functional hand grasp types. Experiments on a public static scene hand data set validate good performance of the presented method. Then they present two applications utilizing grasp type classification: (a) inference of human action intention and (b) fine level manipulation action segmentation. Experiments on both tasks demonstrate the usefulness of grasp type as a cognitive feature for computer vision. This study shows that the grasp type is a powerful symbolic representation for action understanding, and thus opens new avenues for future research.

1. Introduction

The grasp type contains fine-grain information about human action. Consider the two scenes in Figure 1 from the VOC challenge. Current computer vision systems can easily detect that there is one bicycle and one cyclist (human being) in the image. Through human pose estimation, the system can further confirm that these two cyclists are riding the bike. But humans can tell that the cyclist on the left side literally is not riding the bicycle since his hands are posing in a Rest or Extension grasp next to the handlebar while the cyclist on the right side is racing because his hands firmly hold the handlebar with a Power Cylindrical grasp. In other words, the recognition of grasp type is essential for a more detailed analysis of human action, beyond the processes of current state-of-the-art vision systems. Here they 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,this system takes in an unconstrained image patch around the human hand, and outputs which category of grasp type is used.



(a) (b)

Figure 1. (a) Rest or Extension on the handlebar vs. (b) Firmly power cylindrical grasping the handlebar.

2. Approach

The author briefly summarize the basic concepts of Convolutional Neural Networks (CNN), and then they present their implementations for grasp type recognition, human action intention prediction and fine level manipulation action segmentation using the change of grasp type over time.

2.1. Human Grasp Types

Humans, when looking at a photograph, can more or less tell what kind of grasp the person in the picture is using.The question becomes, whether using the current state-of-the-art computer vision technique, whether we can develop a system that learns the pattern from human labeled data and recognizes grasp type from a patch around each hand? In the following section,They present our take and show that a grasp type recognition model with decent robustness can be learned using Convolutional Neural Network (CNN) techniques.

2.2. CNN for Grasp Type Recognition

Convolutional Neural Network (CNN) is a multilayer learning framework, which may consist of an input lay-

er,a few convolutional layers and an output layer. The goal of CNN is to learn a hierarchy of feature representations.Response maps in each layer are convolved with a number of filters and further down-sampled by pooling operations.These pooling operations aggregate values in a smaller region by down-sampling functions including max, min, and average sampling. In this work the author adopt the softmax loss function which is given by Eq 1:

$$L(t,y) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{K=1}^{C}(t_k^n \log \frac{e^{y_k^n}}{\sum_{n=1}^{C} e^{y_m^n}}) \qquad (1)$$

2.3. Human Action Intention

Our ability to interpret other peoples actions hinges crucially on predicting their intentionality. Even 18-month-old infants behave altruistically when they observe an adult accidentally dropping a marker on the floor but out of his reach, and they can predict his intention to pick up the marker [2]. From the point view of machine learning for intelligent systems and human-robot collaboration, due to the differences in the embodiment of humans and robots,a direct mapping of action signals is problematic.This solution is that the robot predicts the intent of the observed human activity and implements the same intention using its own sensorimotor apparatus [1].

References

[1] D. Song, N. Kyriazis, I. Oikonomidis, and C. Papazov. Predicting human intention in visual observations of hand/object interactions. In *IEEE International Conference on Robotics and Automation*, 2013. 2

[2] F. Warneken and M. Tomasello. Altruistic helping in human infants and young chimpanzees. *Science*, 2006. 2