論文の内容の要旨

論文題目　Understanding Hand Manipulation from First-Person View Videos (一人称視点映像からの手操作解析に関する研究)

氏　名　　蔡　敏捷

本文

Understanding the ways how human hands interact with objects (hand manipulation) automatically from daily tasks is important for domains such as robotics, human grasp understanding, and motor skill analysis. To promote the study of daily hand manipulation, I present a recognition framework for hand manipulation under first-person vision paradigm with a wearable camera, which overcomes the constraints of tactile sensors and calibrated cameras used in traditional approaches. However, the tasks of recognizing different types of hand manipulation from first-person view video are challenging due to rapidly changing background, ambiguous hand appearance and mutual hand-object occlusions. To tackle the challenges, I propose approaches to reason about semantic information of hands and objects which are considered critical in understanding hand manipulation.

The thesis work is composed by three components which address different aspects of understanding hand manipulation from first-person view videos: (1) An image-based approach for hand grasp analysis from image appearance is presented, which plays a central role in understanding hand manipulation; (2) A

sequence-based method is proposed for hand grasp analysis from a different perspective of hand dynamics rather than static appearance; (3) An unified framework for recognizing grasp types, object attributes and manipulation actions is proposed, in which semantic relationship between hands, objects, and actions is modeled.

The study of hand grasp plays a central role in understanding hand manipulation since hand grasp characterizes the ways how hand hold an object and implies attribute information of the manipulated object. Therefore, an appearance-based approach for hand grasp analysis under first-person vision (FPV) paradigm is first presented. The proposed approach recognizes the types of hand grasp from image appearance and analyzes visual similarity among different grasp types (visual structures of hand grasp). Experiment results demonstrate the potential of automatic grasp recognition in unstructured environments. Analysis of real-world video shows that it is possible to automatically learn intuitive visual grasp structures that are consistent with expert-designed grasp taxonomies.

Appearance-based method is insufficient to discriminate between different grasp types which are ambiguous from a single image, and is sensitive to unreliable hand detection. To address this problem, I propose a sequence-based method to study hand grasp from perspective of hand dynamics. In particular, a feature representation which encodes dynamical information of hand appearance and motion is proposed based on hand-guided feature tracking from image sequences. In addition, I propose a metric for comparing hierarchical clusters in order to quantitatively evaluate the consistency between different visual structures of hand grasp. Through extensive experiments, effectiveness of the proposed method is verified that hand dynamics can help improve grasp recognition and learn more consistent grasp structures.

Building on the work of hand grasp analysis, a further step is taken to study hand manipulation in a broader scale. I believe that grasp types together with object attributes provide complementary information for characterizing different manipulation actions. Thus, I propose an unified model for recognizing hand grasp types, object attributes and manipulation actions from a single image. Experiments strongly support the hypothesis that: (1) Attribute information of the manipulated object can be extracted without any specific object detectors by exploring spatial hand-object configuration; (2) Contextual information between grasp types and object attributes is important in dealing with mutual hand-object occlusions; (3) Action models that address the semantic relationship with grasp types and object

attributes outperform traditional appearance-based models which are not designed to take into account semantic constraints and are overfit to image appearance.