論 文 の 内 容 の 要 旨

論文題目　Modeling Human Behaviors from First-person Perspective
（一人称視点映像解析による人物行動のモデリング）

氏　　名　　黄　逸飛

Understanding the behavior of human automatically from daily tasks is important for multiple domains such as robotics, psychology, and motor skill analysis. With the commercial success of wearable cameras, analyzing human behavior using videos taken by first-person cameras is becoming popular, as this human-centric perspective is naturally suited to gathering visual information about everyday observations and interactions, which in turn can reveal the attention, activities, and goals of its wearer. By taking advantage of the first-person point-of-view paradigm, I present models for human behavior understanding based on machine learning that focuses on the two types of human behavior: human gaze and human actions. However, the tasks of modeling human gaze and actions from first-person view video are challenging due to rapidly changing background, cluttered objects in the scene, field-of-view limitations and hand-object occlusions. To tackle the challenges, I propose approaches to reason about the rich high-level semantic information contained in the first-person videos to model human gaze and actions.

The thesis work is composed by three components which address the understanding of different types of human behavior from first-person view videos: (1) An approach for human gaze prediction from first-person videos that uses task-dependent attention transition; (2) A graph-based method for localizing and recognizing human actions from videos using the temporal relation among actions; (3) A unified framework for jointly recognizing human action and predicting human gaze, since human action and gaze are deeply correlated.

The study of human gaze plays a vital role in understanding the attention mechanism of human since gaze is one of the most direct way of human attention. Since human gaze is not always attracted by the salient regions but also dependent on the undergoing task, I propose a hybrid model for gaze prediction based on deep neural networks which integrates task-dependent attention transition with bottom-up saliency estimation. In particular, the task-dependent attention transition is learned with a recurrent neural network to exploit the temporal context of gaze fixations, e.g., looking at a cup after moving gaze away

from a grasped bottle. Analysis on real-world videos show that the proposed model significantly outperforms state-of-the-art gaze prediction methods and is able to learn meaningful transition of human attention.

Understanding human actions has always been one of the fundamental research problems of computer vision. Different from third-person perspective, actions from first-person perspective is more difficult to capture due to the camera motion and limitations on field of view. To alleviate these problems, I aim to learn the relation of multiple action segments in various time spans, to help with better localization and classification of human actions. This task is also known as action segmentation. I design a method to model the relations by using two Graph Convolution Networks (GCNs) where each node represents an action segment. The two graphs have different edge properties to account for boundary regression and classification tasks, respectively. By applying graph convolution, we can update each node's representation based on its relationship with neighboring nodes. The updated representation is then used for improved action segmentation. Through extensive experiments, effectiveness of the proposed method is verified that modeling relations using GCNs can help to better locate the action boundaries and recognize the action categories.

Building on the work of human gaze prediction and action segmentation, a further step is taken to study the mutual influence of the two human behaviors. My assumption is that during the procedure of performing a manipulation task, on the one hand, what a person is doing determines where the person is looking at. On the other hand, the gaze location reveals gaze regions which contain important and information about the undergoing action and also the non-gaze regions that include complimentary clues for differentiating some fine-grained actions. To validate the assumptions, I use a mutual context network (MCN) that jointly learns action-dependent gaze prediction and gaze-guided action recognition in an end-to-end manner. Experiments on multiple public egocentric video datasets demonstrate that our MCN achieves state-of-the-art performance of both gaze prediction and action recognition. Our experiments also show that action-dependent gaze patterns could be learned with our method.