論 文 の 内 容 の 要 旨

論文題目　Human Activity Understanding by Multilateral Relation Mining
（多面的関係性マイニングによる人物行動理解）

氏　名　楊 麗錦

Understanding human activity from videos is the key to the next-generation human-oriented assistive AI technology. It is also the core technique to various real-world applications such as surveillance, home-assistance robot, autonomous driving and VR/AR systems. In recent years, with the rapid development of deep learning techniques, remarkable progress has been made on this topic, accompanied by various strong deep backbone models that can extract powerful features to describe the undergoing action in the video. However, most previous work only focus on increasing the representation ability by designing sophisticated network architectures, without fully leveraging the high-level relations hidden in the videos, which could largely enhance the human activity understanding performance.

This thesis focuses on mining multilateral relations for human activity understanding. To be more specific, we aim to find important high-level relations, for example, local-global relations and relation between multiple modalities, and then leverage these relations on human activity understanding task. Based on the degree of supervision, human activity understanding can be roughly divided into three setting categories: the fully-supervised setting, the unsupervised setting, and the open-set setting. Since different types of relations should be used in different settings, in this thesis, we explore the multilateral relation mining in all of these three settings.

Supervised human action understanding can be performed on videos taken from multiple perspectives, among which, first-person videos taken by wearable cameras record human behaviors from the same perspective as humans daily observation. This unique perspective enables a wide range of applications such as VR/AR, human computer interaction and home assistance techniques. Regarding the problem that most methods are designed for third-person videos and perform sub-optimally on the first-person perspective, this thesis first focus on supervised activity recognition in first-person videos. One of the major reasons that previous

works do not perform well on first-person action recognition is that the unique field of view makes actions sometimes happen outside the video viewing range. Thus, this thesis mines the relation between local and global deep features, leveraging a global knowledge of all the local clips to identify which clip is the most discriminative one and suppress the less important clip feature. In order to effectively leverage the local and global feature relations, we introduce a novel stacked temporal attention module, enabling a progressive relation mining and refinement.

To enhance the application of activity understanding algorithms in the real-world application, one major obstacles is that there does not exist sufficient labels to enable the adaptation of trained deep neural networks in the numerous in-the-wild scenarios. While unsupervised domain adaptation techniques are applicable to address this issue by minimizing the domain gaps between seen and unseen scenarios, most previous works only focus on the appearance, without fully exploiting the characteristic of videos. Videos add one additional temporal dimension compared to images, which naturally provides multiple modalities such as optical flow and audio. In this thesis we mine the relation among these multiple modalities within videos to perform activity understanding across domains. We found that each modality has its strength in a certain aspect, and the interaction of these modalities can provide useful information for understanding activities in unseen environments.

To further escalate the practicability of human activity understanding in-the-wild, the ability of recognizing a pre-defined set of actions is far from sufficient. It is essential that any kind of activities described in natural languages can be effectively modeled. In this thesis, we also focus on the activity understanding in this open-set setting, where we aim to find the location of action in a video given a natural language sentence description as input. We further focus on a more challenging weakly-supervised setting where we do not have access to ground truth action location, but only video-level video-sentence correspondence. We formulate the relation in this problem as the ranking between positive location proposals and the negative ones, and learn to output correct proposals in a self-supervised manner.