The Department of Electrical and Computer Engineering

Announces the

Final Defense of Dissertation

Tahmida Mahmud

Doctor of Philosophy, Graduate Program in Electrical Engineering
University of California, Riverside

Date: June 26, 2019
Time: 3:00 PM – 5:00 PM
Location: Winston Chung Hall 315

Title: Near-Future Prediction in Videos: Applications in Video Annotation and Frame Reconstruction

Abstract: Near-future prediction in videos has crucial impact on a wide range of practical applications which require anticipatory response. In videos, prediction can be performed in different spaces such as labels, captions and frames. Labels can be predicted for a longer horizon in future but are less informative than frames. Video frames are much richer in content than labels but only a few frames can be predicted ahead. Captions lie in between these two extremes: they can describe changes in activities for a longer prediction horizon and provide a much richer description than labels. In this thesis, we provide three distinct prediction frameworks leveraged upon different computer vision and machine learning techniques.

Most of the existing works on labeling human activities focus on the recognition or early recognition problem where complete or partial observations of the activity are available. However, in the prediction problem we are addressing, no observation of the future activity is available beforehand. We propose a system that can infer about the labels and the starting time of a sequence of future unobserved activities combining different context attributes from the observed portion of the video. Next, we propose a sequence-to-sequence learning-based approach using an encoder-decoder LSTM pair for captioning the near-future unobserved activity sequences. We propose an adversarial approach using conditional Generative Adversarial Network (cGAN) for multi-sensor frame reconstruction where we learn a mapping between inter-camera and intra-camera frames and for multi-modal frame reconstruction where we learn a mapping between 3D LIDAR point clouds and RGB images. However, these solution methods require lots of labeled data. State-of-the-art video annotation approaches assume that there is no latency for looking up the correct category of label and the annotator is required to watch the whole video segment. We propose a novel early prediction framework so that video annotation becomes scalable.