Tutorial 3: Human Pose Estimation and Action Recognition

Human pose estimation is a fundamental problem in computer vision which can enable many applications. With the rapid development of convolutional neural network, there has been exciting progress in human pose estimation, and impressive performance has been achieved. Human action understanding is of high-level semantics, playing a critical role in many applications like human-computer interaction and visual surveillance. In this past few years, there has been a large progress in action and gesture recognition thanks partly to the progress in the body and hand pose estimation.. Effective techniques have been developed to address many challenging issues in real world environments such as dynamic and cluttered background as well as occlusions. In this tutorial, we first introduce the basics of the tutorial and then state-of-art algorithms for the human pose estimation problem will be discussed. We then present in-depth studies on the recently developed action recognition algorithms. The topics cover the human pose estimation and action recognition using regular and depth cameras, techniques that leverage body and hand poses, as well as action analysis in user-generated consumer videos, such as movies and Youtube videos.

Human Pose Estimation and Keypoint Detection

For human pose estimation, we will roughly categorize the recent approaches into two types: top-down and bottom-up. Top-down approaches rely on the human detection first to localize the human bounding-box and then a single person skeleton estimation algorithm is utilized to localize the human keypoints. Bottom-up approaches, on the other hand, try to localize the human joints first and then assemble the joints into human instances. We will give a comprehensive overview of the recently developed techniques on both approaches, and discuss the challenges and potential applications.

Video-based Action Recognition and Space-Time Localization

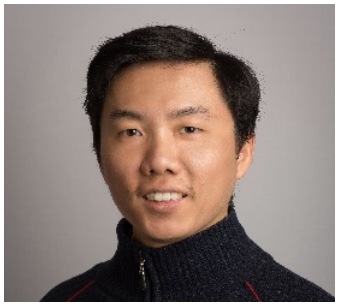
For action recognition, so far, majority of the research has been focusing on the categorization of a whole video sequence where one action label is assigned to the entire video sequence. However, the granularity of the resulting concept labels is usually too coarse to be useful in real world applications. It is not uncommon that a video sequence contains multiple activities that may occur simultaneously at different locations. Therefore, instead of annotating the entire video sequence as an individual event, efficient and powerful tools are required to “locate” the actions, despite the cluttered and dynamic scenes. Action localization in videos involves finding both the spatial and temporal extent of an action, that is, where and when an action/activity occurs. It is inherently more difficult than classification. For video surveillance applications, space-time localization of both normal and abnormal actions (e.g., abandoning luggage, going off normal path) is of great importance. An overview of action recognition with emphasis on location will be introduced and we will highlight the recently developed techniques in this direction.

Depth-camera based Action and Gesture Recognition

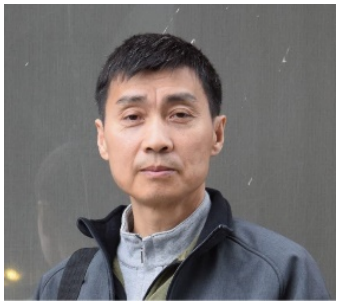
For depth-camera based action recognition, many researchers use the 3D skeletons as the basic features. Different techniques have been developed to leverage the 3D skeletons for action recognition. Some of them use recursive neural networks while others do not. We will cover both types of techniques. For depth-camera based gesture recognition, one popular approach is to use the hand pose. Effective techniques have been developed to estimate the 3D hand pose from depth maps. In this tutorial, we will cover the 3D hand pose estimation and its applications to hand gesture recognition.

Speakers:

Gang Yu
Megvii (Face++), China



Junsong Yuan
University at Buffalo, USA



Zicheng Liu
Microsoft Research, USA