Carnegie Mellon University

THE ROBOTICS INSTITUTE

# THESIS DEFENSE

## Exploiting Point Motion, Shape Deformation, and Semantic Priors for Dynamic 3D Reconstruction in the Wild

Tuesday, August 20, 2019

3002 Newell Simon Hall

10:00 a.m.

### Thesis Committee:

Srinivasa Narasimhan
Co-chair

Yaser Sheikh
Co-chair

Michael Kaess

Marc Pollefeys
ETH/Microsoft

## Minh Phuoc Vo

### Abstract

With the advent of affordable and high-quality smartphone cameras, any significant events will be massively captured both actively and passively from multiple perspectives. This opens up exciting opportunities for low-cost high-end VFX effects and large scale media analytics. However, automatically organizing large scale visual data and creating a comprehensive 3D scene model is still an unsolved problem. State of the art 3D reconstruction algorithms are mostly applicable to static scenes, mainly due to the lack of triangulation constraints for dynamic objects observed by unsynchronized cameras and the difficulties in finding reliable correspondences across cameras in diverse and dynamic settings.

This thesis aims to provide a computational pipeline for high-quality 3D reconstruction of the dynamic scene captured by multiple unsynchronized video cameras in the wild. The key is to exploit the physics of motion dynamics, shape deformation, scene semantics, and the interplay between them. Toward this end, this thesis makes four enabling technical contributions.

First, this thesis introduces a spatiotemporal bundle adjustment algorithm to accurately estimate a sparse set of 3D trajectories of dynamic objects from multiple unsynchronized mobile video cameras. The lack of triangulation constraint on dynamic points is solved by carefully integrating physics-based motion prior describing how points move over time. This algorithm takes advantage of the unsynchronized video streams to estimate 3D motion reconstruction in the wild at much higher temporal resolution than the input videos.

Second, this thesis presents a simple but powerful self-supervised framework to adapt a generic person appearance descriptor to the unlabeled videos by exploiting motion tracking, mutual exclusion constraints, and multi-view geometry without any manual annotations. The adapted descriptor is strongly discriminative and enables a tracking-by-clustering formulation. This advantage enables a first-of-a-kind accurate and consistent markerless motion tracking of multiple people participating in a complex group activity from mobile cameras in the wild with further application to multi-angle video cutting for intuitive tracking visualization.

Third, this thesis creates a framework for 3D tracking of the rigidly moving objects even in severe occlusions by fusing single-view unstructured tracklets and multi-view semantic structured keypoints reconstruction. No spatial correspondences are needed for the unstructured points. No temporal correspondences are needed for the structured points. The imprecise but accurate 3D structured keypoint is compensated by the sparse but precise 3D unstructured tracks, leading to improvements in both structured keypoints localization and motion tracking of the entire object.

Fourth, this thesis presents a single-shot illumination decomposition method for dense dynamic shape capture of highly textured surfaces illuminated by multiple projectors. The decomposition scheme assumes smooth shape deformation and can accurately recover the illumination image of different projectors and the texture images of the scene from their mixed appearances.