DONG ZHANG

1984	Born in Datong, China
2007	B.E., Zhejiang University, Hangzhou, China
2013	Research Associate Intern, SRI International, Princeton, NJ
2014	Computer Vision Intern, Siemens Corporation, Princeton, NJ
2015	Interim Engineering Intern, Qualcomm, Vienna, Austria
2011-16	Ph.D., University of Central Florida, Orlando, Florida.

UNIVERSITY OF CENTRAL FLORIDA
CENTER FOR RESEARCH IN COMPUTER VISION

FINAL ORAL EXAMINATION

OF

DONG ZHANG

B.E., ZHEJIANG UNIVERSITY, 2007

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

(COMPUTER SCIENCE)

11 July, 2016, 10:00 A.M.
HEC 103

DISSERTATION COMMITTEE

Professor Mubarak Shah, *Chairman*, *shah@crcv.ucf.edu*
Professor Ulas Bagci, *bagci@crcv.ucf.edu*
Professor Guo Jun Qi, *guojun.qi@ucf.edu*
Professor Hae-Bum Yun, *hae-bum.yun@ucf.edu*

DISSERTATION RESEARCH IMPACT

This dissertation contributes to video object segmentation and human pose estimation. Video object segmentation is crucial for many real-world applications, such as video editing, movie production, and vision guided surgery; which typically have been performed manually. Automatic video object segmentation can save the editors dramatic amounts of time and ensure fast media production; it can also be applied to organ segmentation for vision guided surgery. Similarly, human pose estimation is essential for Human-Computer Interaction (HCI), and its applications include video games, virtual/augmented reality, and healthcare. Non-invasive HCI approaches are desperately needed in the video gaming industry. With the recent availability of affordable virtual reality headsets, automatic human pose estimation can offer users seamless virtual interaction. Also, in healthcare applications, accurate human body state and human pose estimation can provide very useful data.

SELECTED PUBLICATIONS (total citation: 122)

1. **Human Pose Estimation in Videos,** D. Zhang and M. Shah, *in International Conference on Computer Vision (**ICCV**)*, 2015.
2. **Video Object Co-Segmentation by Regulated Maximum Weight Cliques,** D. Zhang, O. Javed, and M. Shah, *in European Conference on Computer Vision (**ECCV**)*, 2014.
3. **Video Object Segmentation through Spatially Accurate and Temporally Dense Extraction of Primary Object Regions,** D. Zhang, O. Javed, and M. Shah, *in IEEE International Conference on Computer Vision and Pattern Recognition (**CVPR**)*, 2013.
4. **Guidewire Tracking Using a Novel Sequential Segment Optimization Method in Interventional X-Ray Videos,** B. Chen, Z. Wu, S. Sun, D. Zhang, and T. Chen, *in IEEE International Symposium on Biomedical Imaging (**ISBI**)*, 2016.
5. **Visual Odometry in Dynamical Scenes,** D. Zhang, and P. Li, *in Sensors and Transducers*, 2012.
6. **Motion Detection for Rapidly Moving Cameras in Fully 3D Scenes,** D. Zhang and P. Li, *in Pacific-Rim Symposium on Image and Video Technology (PSIVT)*, 2010.

PATENTS

1. D. Zhang, and M. Shah, **Human Pose Estimation in Unconstrained Videos**, University of Central Florida, 2016. (U.S. Patent Application filed)
2. D. Zhang, S. Sun, Z. Wu, B.-J. Chen, A. Meyer, and T. Chen, **Vessel Tree Tracking in Angiography Videos**, Siemens Corporate Research, 2015. (U.S. Patent Application filed)

DISSERTATION

SPATIOTEMPORAL GRAPHS FOR OBJECT SEGMENTATION AND HUMAN POSE ESTIMATION IN VIDEOS

Images and videos can be naturally represented by graphs, with spatial graphs for images and spatiotemporal graphs for videos. However, for different applications, there are usually different formulations of the graphs, and algorithms for each formulation have different complexities. Therefore, wisely formulating the problem to ensure an accurate and efficient solution is one of the core issues in Computer Vision research. We explore three problems in this domain to demonstrate how to formulate all of these problems in terms of spatiotemporal graphs and obtain good and efficient solutions.

The first problem we explore is video object segmentation. The goal is to segment the primary moving objects in the videos. This problem is important for many applications, such as content based video retrieval, video summarization, activity understanding and targeted content replacement. In our framework, we use object proposals, which are object-like regions obtained by low-level visual cues. Each object proposal has an object-ness score associated with it, which indicates how likely this object proposal corresponds to an object. The problem is formulated as a directed acyclic graph, for which nodes represent the object proposals and edges represent the spatiotemporal relationship between nodes. A dynamic programming solution is employed to select one object proposal from each video frame, while ensuring their consistency throughout the video frames. Gaussian mixture models (GMMs) are used for modeling the background and foreground, and Markov Random Fields (MRFs) are employed to smooth the pixel-level segmentation.

In the above spatiotemporal graph formulation, we consider the object segmentation in only single video. Next, we consider multiple videos and model the video co-segmentation problem as a spatiotemporal graph. The goal here is to simultaneously segment the moving objects from multiple videos and assign common objects the same labels. The problem is formulated as a regulated maximum clique problem using object proposals. The object proposals are tracked in adjacent frames to generate a pool of candidate tracklets. Then an undirected graph is built with the nodes corresponding to the tracklets from all the videos and edges representing the similarities between the tracklets. A modified Bron-Kerbosch Algorithm is applied to the graph in order to select the prominent objects contained in these videos, hence relate the segmentation of each object in different videos.

In online and surveillance videos, the most important object class is the human. In contrast to generic video object segmentation and co-segmentation, specific knowledge about humans, which is defined by a pose (i.e. human skeleton) can be employed to help the segmentation and tracking of people in the videos. We formulate the problem of human pose estimation in videos using the spatiotemporal graph. In the formulation, the nodes represent different body parts in the video frames and edges represent the spatiotemporal relationship between body parts in adjacent frames. The graph is carefully designed to ensure an exact and efficient solution. The overall objective for the new formulation is to remove the simple cycles from the traditional graph-based formulations. Dynamic programming is employed in different stages in the method to select the best tracklets and human pose configurations.