

Aayush Jung Bahadur Rana

1992	Born in Kathmandu, Nepal
2011-2015	B.Sc.E., Asian Institute of Technology, Thailand
2015-2017	Junior Research Assistant, VGL Lab, Asian Institute of Technology
2020	Research Internship, SRI International, Princeton, NJ
2022	Research Internship, Qualcomm Technologies, San Diego, CA
2017-2023	M.S., University of Central Florida, Orlando, FL
2017-2023	Ph.D., University of Central Florida, Orlando, FL
2023	Senior Engineer, Qualcomm Technologies, San Diego, CA

UCF | Center for Research in Computer Vision

UNIVERSITY OF CENTRAL FLORIDA

FINAL ORAL EXAMINATION

OF

Aayush Jung Bahadur Rana

B.Sc.E., Asian Institute of Technology, 2015
M.S., University of Central Florida, 2023

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

(COMPUTER SCIENCE)

5 July 2023, 1:00 P.M.
Engineering II Room 103

DISSERTATION COMMITTEE

Professor Yogesh Singh Rawat, *Chair*, *yogesh@crcv.ucf.edu*
Professor Mubarak Shah, *shah@crcv.ucf.edu*
Professor Chen Chen, *chen.chen@crcv.ucf.edu*
Professor Shaurya Agarwal, *shaurya.agarwal@ucf.edu*

DISSERTATION RESEARCH IMPACT

Videos capture the inherently sequential nature of the real world, making automatic video understanding an essential need for automatic understanding of the real world. Due to major advancements in camera, communication, and storage hardware, videos have become a widely used data format for crucial applications such as home automation, security, analysis, robotics, and autonomous driving. Existing methods for video understanding require heavy computation and large training data for good performance, this limits how quickly the videos can be processed and how much data can be labeled for training. Real-world video understanding requires analyzing dense scenes and sequential information, which increases the processing time and labeling cost as the video increases in scene density and video length. Therefore, it is crucial to develop video understanding methods that reduce the processing time and labeling cost.

In this dissertation, we first propose a method to improve network efficiency for video understanding task and then provide methods to improve annotation efficiency for video understanding task using end-to-end models, active learning based selection framework and temporal pseudo-labeling strategies. Through these works, we aim to improve network efficiency as well as data annotation efficiency, as an effort to encourage wider development and adaptation of large-scale video understanding methods.

SELECTED PUBLICATIONS

1. **Hybrid Active Learning via Deep Clustering for Video Action Detection,** AJ Rana and YS Rawat, *in CVPR*, 2023.
2. **Are all Frames Equal? Active Sparse Labeling for Video Action Detection,** AJ Rana and YS Rawat, *in NeurIPS*, 2022.
3. **We don't need thousand proposals: Single Shot Actor-Action Detection in Videos,** AJ Rana and YS Rawat, *in WACV*, 2021.
4. **OmViD: Omni-supervised active learning for video action detection,** AJ Rana and YS Rawat, Under Submission in ICCV, 2023.
5. **Gabriella: An online system for real-time activity detection in untrimmed security videos,** MN Rizve, U Demir, P Tirupattur, AJ Rana, K Duarte, IR Dave, YS Rawat and M Shah, *in ICPR*, 2020. *Best Scientific Paper Award.*
6. **SSA2D: Single Shot Actor-Action Detection in Videos (Student Abstract),** AJ Rana and YS Rawat, *in AAAI*, 2021.
7. **LARNet: Latent Action Representation for Human Action Synthesis,** N Biyani, AJ Rana, S Vyas and YS Rawat, *in BMVC*, 2021.
8. **Methods of Real-time Spatio-temporal Activity Detection and Categorization from Untrimmed Video Segments,** Y Rawat, M Shah, AJ Rana, P Tirupattur and MN Rizve, *U.S. Patent 11,468,676*.

DISSERTATION

DEEP VIDEO UNDERSTANDING WITH MODEL EFFICIENCY AND SPARSE ACTIVE LABELING

Dense video understanding is a challenging problem with a wide range of applications in automation, analysis, security, autonomous driving, and robotics. Real-world videos have varying object density due to crowded scene, and they have varying video length based on the underlying application. This creates a challenge for video understanding as the processing time increases with video length and density. Better video understanding methods require more annotated data for training, which also gets costlier to label as the video gets longer and more crowded. These challenges need to be addressed to overcome different video understanding tasks such as classification, detection, segmentation, tracking, summarization and more. As complex tasks such as detection and segmentation need more processing time and labeled data, it is important to improve time and label efficiency for dense video understanding. While deep neural networks for image understanding have improved a lot over the years, it is not the same for video understanding tasks. This comes down to two factors: video models have extra temporal information which increases the computation cost of the model during training and inference, and video datasets with full annotations are costly to label. To improve video understanding further, we need to optimize the models for efficient training and inference by streamlining the entire process. Furthermore, we need to design methods that enable labeling large video dataset efficiently by reducing unnecessary annotation cost, along with methods that enable training video understanding models using sparse annotations.

First, we propose a simple yet effective end-to-end deep network for actor-action detection in videos. Existing methods take a memory intensive top-down approach based on region-proposal networks (RPN) to generate thousands of proposals per frame. We propose to solve this problem from a different perspective where we don't need any proposals by performing pixel level joint actor-action detection, where every pixel of the detected actor is assigned an action label. This makes it time and memory efficient as well as scalable to dense video scenes.

Next, we focus on efficient labeling of videos for action detection to minimize the annotation cost for video action detection training tasks. We propose a novel active learning strategy for sparse labeling for video action detection which estimates the utility of each frame for action detection and uses these sparse labels for training action detection models. The selection is done at frame level and select the most informative frame for each video to annotate. The sparse annotation is used with our novel loss formulation which enables training of action detection model using pseudo-labels.

Then, we extend it further by reducing annotation cost by selecting at video level for annotation which will have more impact on video action detection training. We propose a hybrid active learning strategy which performs efficient labeling by first selecting videos (inter-sample) for labeling and then selecting a few frames from these videos (intra-sample) to be annotated. This strategy reduces the annotation cost from two different aspects leading to significant labeling cost reduction. We also further improve the loss formulation to better utilize sparse annotations.

Finally, we analyze the types of annotation appropriate for each sample and how it affects video action detection. We study several annotation types including i) video level tags, ii) points iii) scribbles, iv) bounding box, and v) pixel level masks and propose a simple active learning strategy to which estimates appropriate types of annotations needed for each video sample based on the usefulness of that video sample. This is combined with a superpixel based pseudo-labeling technique to improve learning of video action detection from a mixed set of annotation.