

Center for Research in Computer Vision

UNIVERSITY OF CENTRAL FLORIDA

FINAL ORAL EXAMINATION

OF

Praveen Tirupattur

M.S., Technical University of Kaiserslautern, 2016
B.Tech., Jawaharlal Nehru Technological University, 2010

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY
(COMPUTER SCIENCE)

27 June 2024, 9:00 A.M.
Research I - Room 101

DISSERTATION COMMITTEE

Professor Mubarak Shah, *Chair, shah@crcv.ucf.edu*
Professor Yogesh Singh Rawat, *yogesh@crcv.ucf.edu*
Professor Chen Chen, *chen.chen@crcv.ucf.edu*
Professor Shuo Pang, *pang@creol.ucf.edu*

DISSERTATION RESEARCH IMPACT

Video action understanding involves comprehending human actions in videos, a challenging task for computer vision systems due to inter-class variations and complex scenes. Central to this task are four fundamental questions: What, When, Where, and Who. This dissertation addresses these by proposing methods for multi-view action classification (What?), modeling relationships for precise temporal localization (When?) of actions, introducing Gabriella for real-time action detection (Where?) in security videos, and developing joint action and actor recognition (Who?) for robotic care.

Enhanced multi-view action classification significantly advances surveillance and sports analysis by integrating multiple camera perspectives for a comprehensive understanding of activities. Precise temporal localization within videos enhances video analytics and content moderation by accurately identifying specific moments of interest. Gabriella, an advanced video surveillance system, excels in monitoring large fields-of-view and detect diverse activities, enabling detection of multiple subjects and recognition of the activities performed with high precision. Joint action and actor recognition capabilities empowers robots to provide personalized care, thereby improving the quality of life for users through tailored support and interactions.

SELECTED PUBLICATIONS & PATENT (h-index: 6, citation: 203)

1. **DVANet: Disentangling View and Action Features for Multi View Action Recognition,** N Siddiqui, P Tirupattur, and M Shah, *in AAAI*, 2024.
2. **Modeling Multi-Label Action Dependencies for Temporal Action Localization,** P Tirupattur, K Duarte, YS Rawat, and M Shah, *in CVPR*, 2021. Oral.
3. **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.*
4. **Thoughtviz: Visualizing Human Thoughts Using Generative Adversarial Network,** P Tirupattur, YS Rawat, C Spampinato, and M Shah, *in ACM-MM*, 2018.
5. **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

VIDEO ACTION UNDERSTANDING: ADVANCING ACTION RECOGNITION, TEMPORAL LOCALIZATION AND DETECTION

Video action understanding delves into comprehending human actions depicted in videos, presenting a formidable challenge for computer vision systems. The multifaceted nature of video content, characterized by inter-class variations and complex scenes, poses hurdles that necessitate innovative solutions. Addressing these challenges is pivotal to various domains, including but not limited to video surveillance, activity recognition, human-computer interaction, and the development of autonomous systems. Central to the task of video action understanding are four fundamental questions: What, When, Where, and Who, each encapsulating a distinct aspect of action analysis. Action classification, the "What?", involves the categorization of human actions into predefined classes, enabling the discernment of specific actions within video segments. Temporal action localization, focusing on the "When?", goes beyond mere classification and delves into predicting the temporal extent of actions, encountering challenges arising from variable durations, occlusions, and cluttered backgrounds. Action detection, concerned with the "Where?", amalgamates classification and localization to identify and pinpoint multiple instances of actions across spatial and temporal domains. The aspect of "Who?" remains largely unexplored within the realm of video action understanding, though it finds application in vision-based biometrics and the emerging field of Action Biometrics, which seeks to address both action classification and individual identification. Despite notable progress in research related to these tasks, numerous challenges persist, underscoring the need for ongoing innovation and the development of robust methodologies to tackle them effectively. In this dissertation, we propose novel approaches aimed at overcoming these challenges, thereby advancing the state-of-the-art in video action understanding.

First, we propose a novel method for multi-view action classification ("What?") to enhance action classification accuracy by disentangling action representations from view information. Our approach utilizes learnable transformer decoder queries and supervised contrastive losses to foster robust feature learning resilient to viewpoint shifts. Our method outperforms all uni-modal models on multiple multi-view action recognition datasets, showcasing its effectiveness.

Next, we tackle multi-label temporal action localization ("When?") and propose a method to leverage inherent relationships between actions in real-world videos to improve localization performance. We distinguish the relationships as co-occurrence (actions that occur at the same time) and temporal dependencies (actions that precede or follow each other). Our proposed attention-based architecture models these relationships using a novel Multi-Label Action Dependency (MLAD) layer, enhancing localization performance. We introduce evaluation metrics to consider these dependencies and show superior performance over existing multi-label temporal action localization methods.

Subsequently, we introduce Gabriella, a real-time online system for activity detection ("Where?") in security videos, addressing challenges such as a large field-of-view and multiple activities. Our three-stage approach includes tubelet extraction, activity classification, and online tubelet merging, significantly reducing computation time. We perform experiments on the VIRAT and MEVA datasets and demonstrate the effectiveness of the proposed approach in terms of speed (~100 fps) and performance, achieving state-of-the-art results.

Finally, we introduce an approach to solve the novel task of joint action and actor recognition ("Who?"). Our transformer-based model leverages disentangled representation learning to separate actor and action features for improved performance. We evaluate our approach on multiple real-world datasets, such as the NTU-RGBD 120, PKU-MMD, and Toyota Smart Home datasets, and show improved performance over the baselines.



Praveen Tirupattur

1989	Born in Hyderabad, India
2006-2010	B.Tech., Jawaharlal Nehru Technological University, India
2010-2013	Software Engineer, Progress Software
2013-2016	M.S., Technical University of Kaiserslautern, Germany
2016-2017	Backend Software Developer, Xyrality GmbH
2022	Research Internship, Pinterest
2023	Research Internship, Amazon
2017-2024	Ph.D., University of Central Florida, Orlando, FL