Computational Research in Boston and Beyond Seminar

Skill learning from video

Kristen Grauman
(University of Texas at Austin)

Abstract

What would it mean for AI to understand skilled human activity? In augmented reality (AR), a person wearing smart glasses could quickly pick up new skills with a virtual AI coach that provides real-time guidance. In robot learning, a robot watching people in its environment could acquire new dexterous manipulation skills with less physical experience. Realizing this vision demands significant advances in video understanding, in terms of the degree of detail, viewpoint flexibility, and proficiency assessment. In this talk I'll present our recent progress tackling these challenges. This includes 4D models to anticipate human activity in long-form video; video-language capabilities for generating fine-grained descriptions and constructive commentary; and cross-view representations able to bridge the exocentric-egocentric divide—from the view of the teacher to the view of the learner. I'll also illustrate the impact of these ideas for AI coaching prototypes that guide users through new skills or provide feedback on their physical performance, transforming how-to videos into personalized AI assistants.

Bio

Kristen Grauman is a Professor in the Department of Computer Science at the University of Texas at Austin. Her research in computer vision and machine learning focuses on video understanding and embodied perception. Before joining UT-Austin in 2007, she received her Ph.D. at MIT. She is a AAAS Fellow, IEEE Fellow, AAAI Fellow, Sloan Fellow, and recipient of the 2026 Hill Prize in AI, the 2025 Huang Prize, and the 2013 Computers and Thought Award. She and her collaborators have been recognized with several Best Paper awards in computer vision, including a 2011 Marr Prize and a 2017 Helmholtz Prize (test of time award). She has served as Associate Editor-in-Chief for PAMI and Program Chair of CVPR 2015, NeurIPS 2018, and ICCV 2023.

MIT
MATHEMATICS

https://math.mit.edu/crib/