Egocentric Video Understanding across Modalities and Domains

Chiara Plizzari

Politecnico di Torino

With the growing popularity of wearable cameras, egocentric vision has become an increasingly researched area. This perspective offers a direct view from the wearer's perspective, enabling a more direct study of human behavior. However, the introduction of these devices also presents unique challenges not encountered with traditional stationary cameras.

The goal of this thesis is to explore how multi-sensory information can address the complexities of egocentric videos. Wearable devices face significant changes in illumination, perspective, and environment, causing action recognition models to depend heavily on their training environments and struggle with generalization to new ones. In the first part of the thesis, we explore solving auxiliary tasks across various information channels from videos to enhance robustness across domains. By integrating RGB data with audio and motion information from optical flow via an auxiliary loss to align feature norms, we demonstrate that the resulting models are more generalizable and perform reliably in unseen environments. We then introduce a method using cross-instance video reconstruction through language to learn robust features against a *scenario shift*, where the same action occurs in different activities, and a *location shift*, where videos are from varied geographical locations. For this, we curated ARGO1M, the largest dataset for action recognition generalization, containing over 1 million video clips. Our findings indicate that textual guidance significantly enhances model performance in unseen scenarios and locations.

In the second part of the thesis, we analyze previously unexplored modalities within egocentric vision. Event cameras, with their high pixel bandwidth, dynamic range, low latency, and low power consumption, effectively address challenges like fast camera motion and background clutter. We introduce N-EPIC-Kitchens, the first dataset for studying event-based data in this domain. Results demonstrate that event data perform competitively compared to traditional RGB and optical flow modalities. Finally, we integrate 3D scene information with appearance-based models to overcome the limitations of 2D images' narrow field of view and incomplete scene views. We introduce the task "Out of Sight, Not Out of Mind", which involves tracking object locations around the user over time, even when not visible, using both frame-based images and 3D object positioning. Our findings show that 3D information significantly enhances

the capability of egocentric vision systems to fully capture and understand the surrounding context.

Throughout this thesis, we highlight the importance of utilizing information from multiple channels and demonstrate that focusing on these aspects can significantly improve egocentric video understanding.