
Politecnico di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Computer and Control Engineering ($37^{th}$cycle)

# Towards Computationally-Efficient Solutions for Real-World Challenges in Egocentric Vision

By

**Gabriele Goletto**

******

**Supervisor(s):**
Prof. Barbara Caputo, Supervisor

Politecnico di Torino
2025

# Towards Computationally-Efficient Solutions for Real-World Challenges in Egocentric Vision

Gabriele Goletto

In recent years, egocentric vision has emerged as a compelling paradigm for understanding human behavior through wearable cameras. By capturing the world from a first-person perspective, this approach provides rich contextual information about users' actions, interactions, and environments - opening up exciting possibilities for applications such as assistive robotics, augmented reality, and human-computer interaction. However, turning this potential into real-world impact requires addressing several applicability challenges. Unlike typical commercial applications, egocentric setups are tightly coupled with users' experiences, meaning they should ideally run on edge devices to avoid constant streaming and enable real-time feedback, all while coping with long-duration recordings and the high variability of daily life, including different users and evolving scenes. These constraints entail both technical challenges - requiring models to generalize across diverse domains, subjects, and conditions - as well as computational requirements, as they must be capable of real-time processing within memory and energy budgets.

This thesis tackles these real-world challenges by proposing scalable and efficient solutions across three complementary directions. First, we focus on the feasibility of deploying egocentric action recognition models on edge devices. We evaluate efficiency metrics such as inference speed and power consumption, and we introduce new action recognition evaluation protocols that explicitly consider practical deployment constraints, such as the lack of prior knowledge of action boundaries at inference time. To this end, we propose a lightweight, model-agnostic technique that adapts existing architectures for real-time processing of egocentric videos, enabling online action recognition without requiring knowledge of action boundaries.

In the second part of the thesis, we shift our focus to the challenge of long video understanding. Wearable devices often record videos in an untrimmed fashion, capturing prolonged activities. This is particularly relevant as many human procedures naturally unfold over long temporal horizons. However, repeatedly processing such lengthy recordings is computationally expensive and inefficient. Inspired by the human's ability to maintain information from a single watching, we propose a novel approach that constructs a persistent representation of the interactions in the

recording from a single pass through the video. This representation allows querying action sequences efficiently, without reprocessing the raw video, enabling reasoning over how interactions evolve over time. To systematically evaluate this approach, we introduce a benchmark designed to assess fine-grained video understanding capabilities through a diverse set of challenging queries.

Finally, we showcase the limits of egocentric action recognition systems in unseen environments and demonstrate how exploiting good inductive scene biases and novel sensing modalities can improve the robustness and efficiency of egocentric models. We investigate how scene-level knowledge - particularly the notion that actions are tied to "activity-centric zones" that can be discovered in an unsupervised fashion - can guide action recognition across domains by decoupling these zones from their environment-based appearance. Furthermore, we highlight the potential of event cameras, which offer high temporal resolution and ultra-low power consumption, as a promising alternative to traditional vision sensors. We introduce the first event-based extension of a large-scale egocentric dataset and demonstrate that event data can match or even exceed the performance of RGB-based models, with significantly lower power consumption. By improving generalization, we aim to reduce the need for retraining across different deployment settings, thereby enhancing the scalability and long-term efficiency of egocentric vision models.

Throughout the thesis, we emphasize solutions that move egocentric vision closer to real-world deployment by addressing efficiency, scalability, and generalization. By rethinking how egocentric videos are processed, represented, and sensed, we contribute toward building vision systems that are not only smarter but also more practical and sustainable.