ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Magic Leap Distributed Computing

Prof. R. Wattenhofer

Egocentric Action Recognition

Egocentric Action Recognition is a specialized field in computer vision and artificial intelligence that focuses on recognizing human actions from a first-person perspective, where the camera is worn by the person performing the actions. This unique viewpoint provides rich visual information that can be leveraged for understanding human activities in various scenarios.

While most research efforts in Egocentric Action Recognition (EAR) typically concentrate on enhancing classification performance, our project diverges by placing a primary focus on optimizing time-effectiveness. We want to focus this project towards Head-Mounted Displays (HMDs) such as MagicLeap 2 (ML2). The headset context shifts the optimization priority from performance to efficiency due to some crucial aspects:

- Need for Online Processing. Head-mounted displays require real-time action recognition to provide immediate feedback and enhance user experience.
- Limited Battery Life. Efficient time processing is crucial for conserving the limited battery life of head-mounted displays, ensuring prolonged usage without frequent recharging.
- Simpler Hardware. By prioritizing time-effectiveness, we aim to develop algorithms that can run efficiently on simpler hardware, such as CPUs, reducing the computational burden and enhancing the feasibility of deployment on head-mounted displays. Moreover, this could allow integration on simpler devices.

Contact

- Ard Kastrati : kard@ethz.ch, ETZ G61.3
- Matteo Macchini : mmacchini@magicleap.com, Magic Leap
- Dushan Vasilevski : dvasilevski@magicleap.com, Magic Leap

Project Outline

We identify the following tasks, however the direction of the project is flexible:

- Perform literature review on EAR models (with focus on multimodal approaches), object tracking, and optimized vision backbones (TinyViT, LeViT).
- Explore different datasets for EAR. Consider creating a detailed comparison chart that highlights the strengths and weaknesses of each dataset.
- Identify baseline models for comparison. Evaluate not only their accuracy, but also their inference speed.
- Experiment with different multimodal models by changing the feature set and/or the backbone. Possible input sources include vision, head pose, eye tracking, hand tracking, and object tracking.
- Analyze trade-off between accuracy and inference speed on cpu for the selected systems.
- Evaluate effectiveness of ROI approaches to reduce image size, extract context, and improve the attention mechanism.
- Perform ablation studies on selected systems.
- Integrate on device and test.