Multimodal Egocentric Action Recognition

Advanced Machine Learning / Data Analysis and Artificial Intelligence

2023/2024

Teaching Assistant:
Gabriele Goletto (gabriele.goletto@polito.it)
Simone Alberto Peirone (simone.peirone@polito.it)

Egocentric Vision

Damen, Dima, et al. "Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100." *International Journal of Computer Vision* (2022): 1-23.

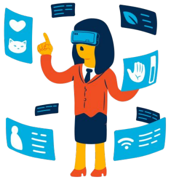
Why Egocentric Action Recognition?

Learn **how humans interact with world** and improve human-robot cooperation

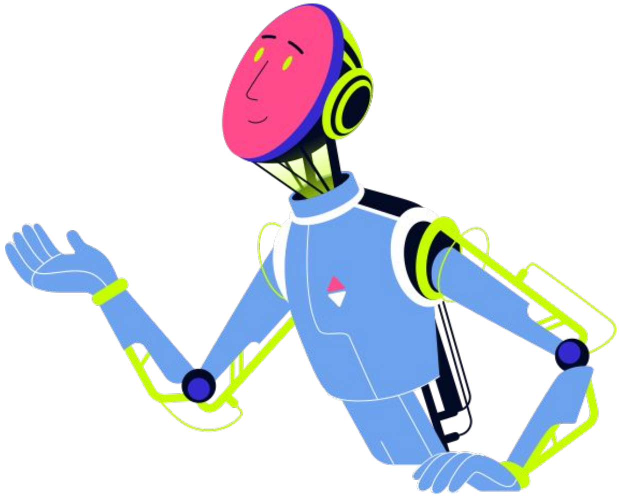
Assistive robotics

Autonomous driving

Industrial applications



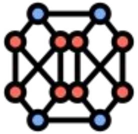
Augmented reality



[2] Núñez-Marcos *et al*,"Egocentric Vision-based Action Recognition: A survey", Neurocomputing 2022

Egocentric Action Recognition (EAR)

A classification problem that aims to assign **labels** to **videos**

Action video →  → **Label**

Neural network

Examples

Open fridge

Close oil

Place pan on hob

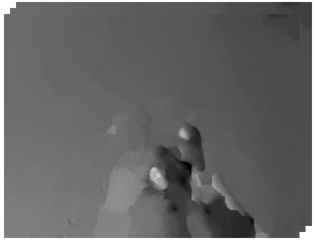
Egocentric Action Recognition (EAR)

Videos provide **multiple sources of information** (a.k.a *modalities*) that we can use for action recognition

RGB frames

✔ Visual appearance (objects and scenes)

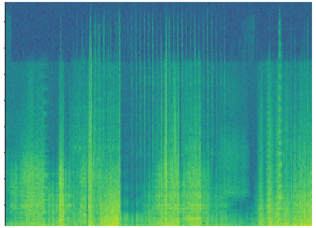
✘ Suffers from occlusions and blur



Optical flow

✔ Focused on motion

✘ Computationally expensive



Audio

✔ Lightweight

✘ Not all activities are recognizable from audio only

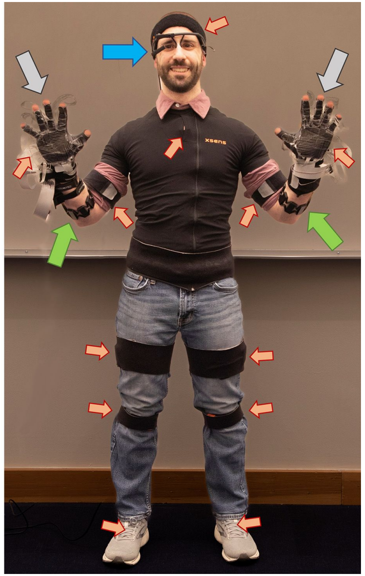
Each modality has its own **strengths** and **weaknesses**

Egocentric vision + wearable sensors

Combine **egocentric vision** with **wearable sensors**

EMG sensors
measures the electrical activity of the muscles

Body and finger tracking
senses the acceleration and velocity of your body



Eye tracking
detects where the eyes are looking at

Tactile sensors
measure the force exerted on the fingers through touch

[3] DelPreto, Joseph, et al. "ActionNet: A Multimodal Dataset for Human Activities Using Wearable Sensors in a Kitchen Environment." NeurIPS22

The EPIC-Kitchens dataset (2018)

https://epic-kitchens.github.io

32 participants

55+ recorded hours

39k action segments in a kitchen

The Action-Net dataset (2022)

https://action-net.csail.mit.edu

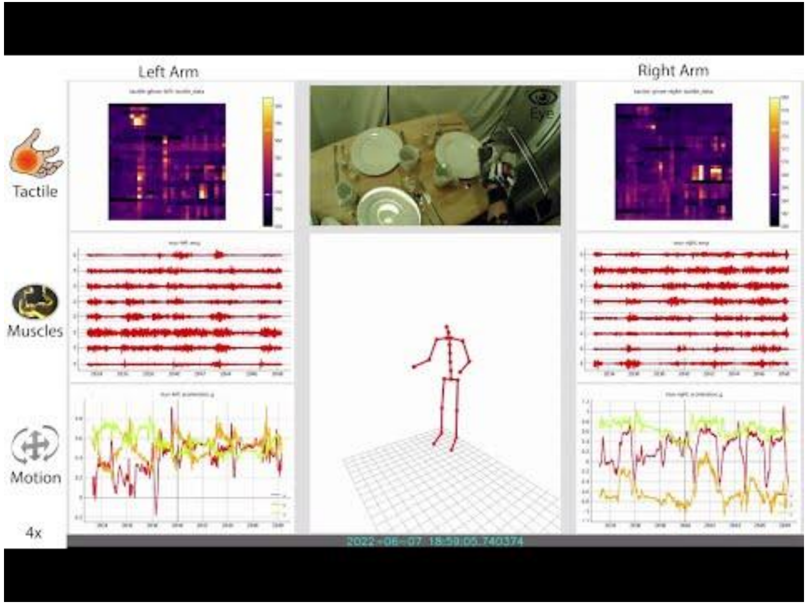
10 subjects

12+ recorded hours

20 unique activities

- Eye tracking
- Microphones
- Body tracking
- **RGB/RGB-D cameras**
- Tactile sensors
- **EMG sensors**

A large number of **cameras** (head-mounted and external) and **wearable sensors**



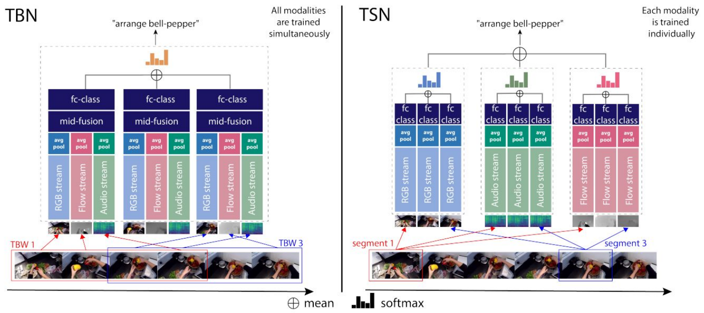
[3] DelPreto, Joseph, et al. "ActionNet: A Multimodal Dataset for Human Activities Using Wearable Sensors in a Kitchen Environment." NeurIPS22

Project steps

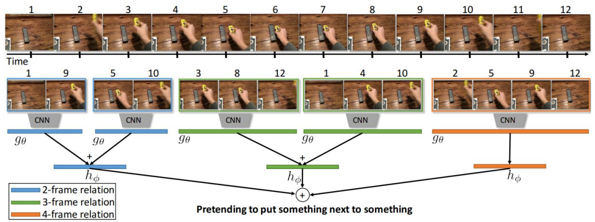
1. **Get familiar with egocentric vision and multimodal learning**
2. Implement common baselines using RGB and optical flow on the EK-55 dataset.
3. Implement one of the following variations:
    a. Multimodal training with RGB and EMG on Action-Net.
    b. A variational autoencoder to translate from RGB to EMG and viceversa on Action-Net. Reconstruct the missing EMG signal in EK and use it to implement a classifier.

Project steps

1. Get familiar with egocentric vision and multimodal learning
2. **Implement common baselines using RGB and optical flow on the EK dataset.**
3. Implement one of the following variations:
   a. Multimodal training with RGB and EMG on Action-Net.
   b. A variational autoencoder to translate from RGB to EMG and viceversa on Action-Net. Reconstruct the missing EMG signal in EK and use it to implement a classifier.



Temporal Binding Network [3] and Temporal Segment Network [4]

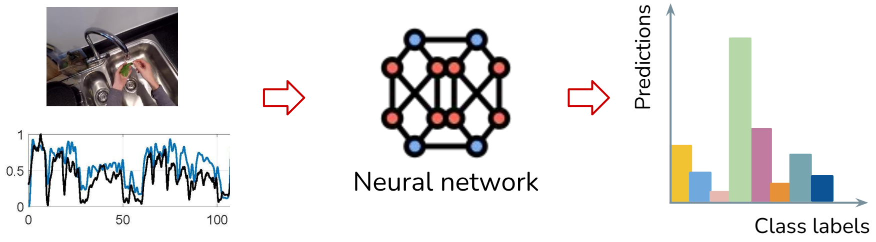


Temporal Relational Reasoning in Videos [5]

[4] Kazakos et al. “Epic-fusion: Audio-visual temporal binding for egocentric action recognition”. ICCV 2019
[5] Wang *et al.* "Temporal segment networks for action recognition in videos." TPAMI 2018
[6] Zhou *et al.* "Temporal relational reasoning in videos." ECCV 2018

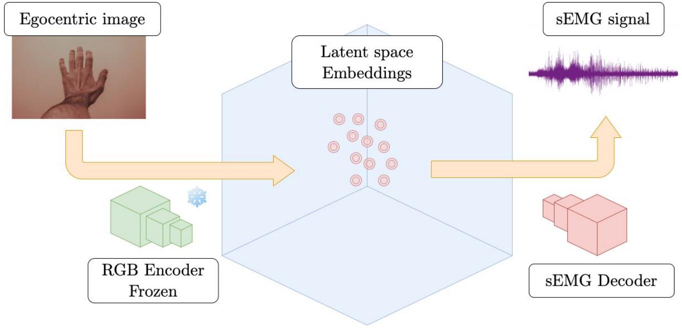
Project steps

1. Get familiar with egocentric vision and multimodal learning
2. Implement common baselines using RGB and optical flow on the EK dataset.
3. **Implement one of the following variations:**
   a. **Multimodal training with RGB and EMG on Action-Net.**
   b. A variational autoencoder to translate from RGB to EMG and viceversa on Action-Net. Reconstruct the missing EMG signal in EK and use it to implement a classifier.



Project steps

1. Get familiar with egocentric vision and multimodal learning
2. Implement common baselines using RGB and optical flow on the EK dataset.
3. **Implement one of the following variations:**
   a. Multimodal training with RGB and EMG on Action-Net.
   b. **A variational autoencoder to translate from RGB to EMG and viceversa. Reconstruct the missing EMG signal on EK-55 and use it to implement a classifier.**



Bibliography

[1] Damen, Dima, et al. "Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100." International Journal of Computer Vision (2022): 1-23.

[1] Núñez-Marcos, Adrián, Gorka Azkune, and Ignacio Arganda-Carreras. "Egocentric vision-based action recognition: a survey." Neurocomputing 472 (2022): 175-197.

[2] DelPreto, Joseph, et al. "ActionNet: A Multimodal Dataset for Human Activities Using Wearable Sensors in a Kitchen Environment." Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track.

[3] Kazakos, Evangelos, et al. "Epic-fusion: Audio-visual temporal binding for egocentric action recognition." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

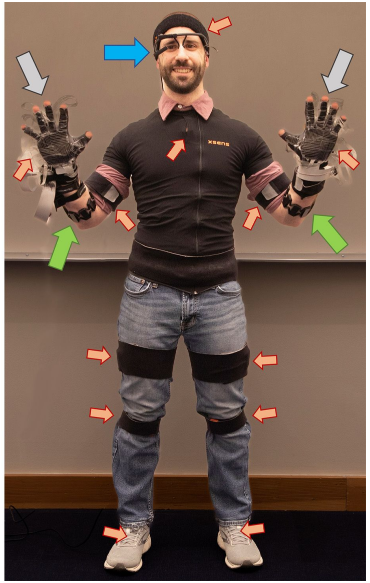
[4] Wang, Limin, et al. "Temporal segment networks for action recognition in videos." IEEE transactions on pattern analysis and machine intelligence 41.11 (2018): 2740-2755.

[5] Zhou, Bolei, et al. "Temporal relational reasoning in videos." Proceedings of the European conference on computer vision (ECCV). 2018.

Multimodal Egocentric Action Recognition

Teaching Assistant:
Gabriele Goletto (gabriele.goletto@polito.it)
Simone Alberto Peirone (simone.peirone@polito.it)

The objective of this project is to explore multimodal **Egocentric Action Recognition (EAR)**, starting from traditional approaches based on RGB, optical flow and audio and moving towards new modalities enabled by wearable sensors. The student should understand typical approaches used in egocentric vision and how multiple modalities can be combined together to perform action recognition or to compensate for the absence of other modalities.



The project is organized in the following steps:

1. Get familiar with egocentric vision and multimodal learning, the task of egocentric action recognition and the most common datasets in the field.
2. Implement EAR models using RGB and optical flow on the EPIC-KITCHENS dataset.
3. Implement one of the following variations:
    a. **Train a multimodal model** on the Action-Net [1] dataset using RGB frames and EMG signals measuring the wearer's muscle activity.
    b. **Train a variational autoencoder** on the Action-Net dataset to translate from RGB to EMG and viceversa. Reconstruct the missing EMG signal on EPIC-KITCHENS [2] and use it to implement a classifier.

Link to project details

[1] DelPreto, Joseph, et al. "ActionNet: A Multimodal Dataset for Human Activities Using Wearable Sensors in a Kitchen Environment." NeurIPS22
[2] Damen, Dima, et al. "Scaling egocentric vision: The epic-kitchens dataset." ECCV 2018.