First Person (Egocentric) Vision: History and Applications

Francesco Ragusa

First Person Vision@Image Processing Laboratory - http://iplab.dmi.unict.it/fpv

Next Vision - http://www.nextvisionlab.it/

Department of Mathematics and Computer Science - University of Catania

francesco.ragusa@unict.it - https://francescoragusa.github.io/

Agenda

1) Part I: History and motivations [09.00 - 10.30]
    a) Agenda of the tutorial;
    b) Definitions, motivations, history and research trends of First Person (egocentric) Vision;
    c) Seminal works in First Person (Egocentric) Vision;
    d) Differences between Third Person and First Person Vision;
    e) First Person Vision datasets;
    f) Wearable devices to acquire/process first person visual data;
    g) Main research trends in First Person (Egocentric) Vision;

Coffee Break [10.30 – 10.45]

Keynote presentation: Gerhard Rigoll [10.45 – 12.00]

1) **Part II: Fundamental tasks for First Person Vision systems [12.00 – 13.00]**
    a) **Localization;**
    b) **Hand/Object Detection;**
    c) **Action/Activity Recognition;**
    d) **Egocentric Human-Object Interaction;**
    e) **Anticipation;**
    f) **Industrial Applications;**
    g) **Conclusion.**

Before we begin…

The slides of this tutorial are available online at:
https://francescoragusa.github.io/visigrapp2024

Part II

Fundamental Tasks for First Person Vision Systems

Data Acquisition

Four things to pay attention to when collecting first person visual data

Video Quality

Field of View

Wearing Modality

Other Modalities

Data Acquisition – Video Quality

- Try to get a high quality camera to get high quality images!
- Egocentric video is subject to motion blur and exposure issues.

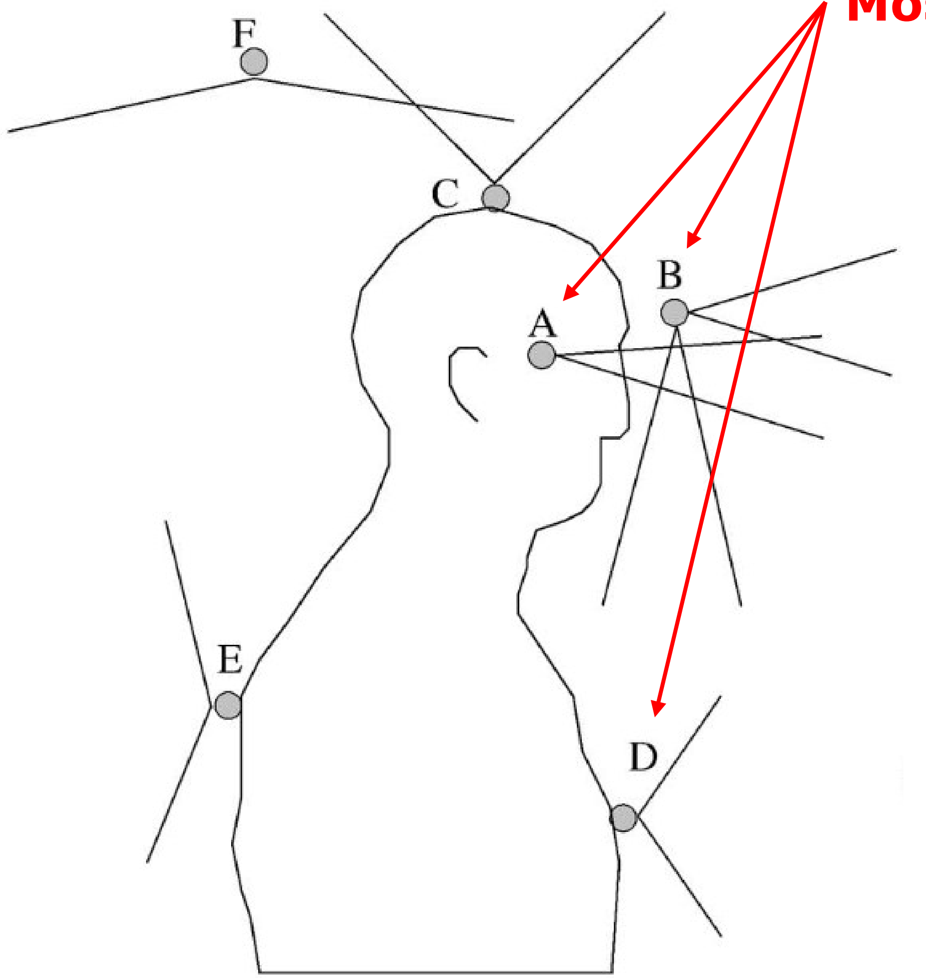
High Quality Video Obtained with a GoPro

Average Quality Video

Data Acquisition – Camera Wearing Modalities

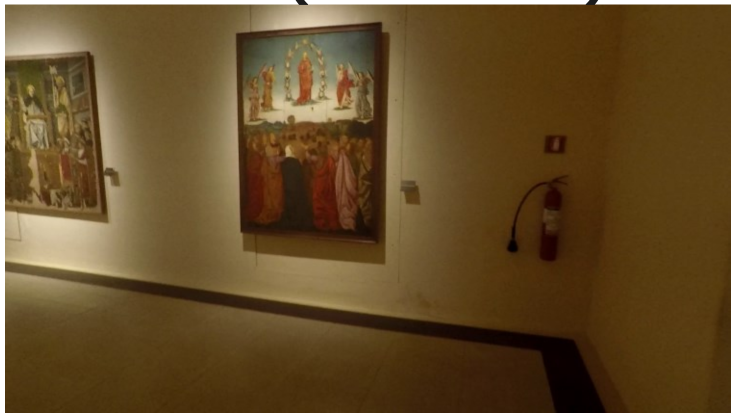
A,B: head mounted, D: chest mounted

Most Common Wearing Modalities



A

B (frontward)

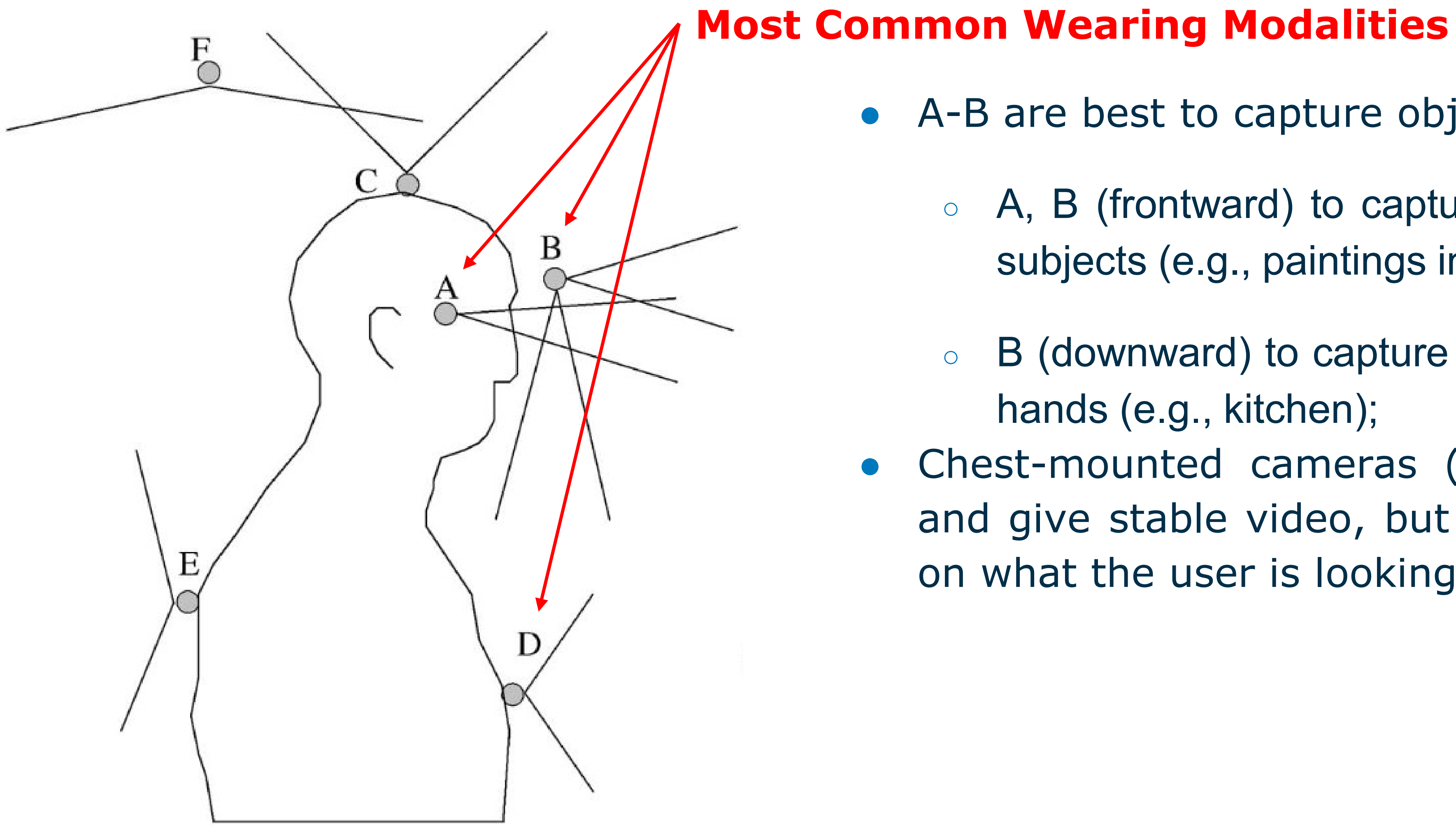


B (downward)

D

Mayol-Cuevas, W. W., Tordoff, B. J., & Murray, D. W. (2009). On the choice and placement of wearable vision sensors. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *39*(2), 414-425.

Data Acquisition – Camera Wearing Modalities



- A-B are best to capture objects:
  - A, B (frontward) to capture objects in front of the subjects (e.g., paintings in a museum);
  - B (downward) to capture objects manipulated with hands (e.g., kitchen);
- Chest-mounted cameras (D) are less obtrusive and give stable video, but they may miss details on what the user is looking at;

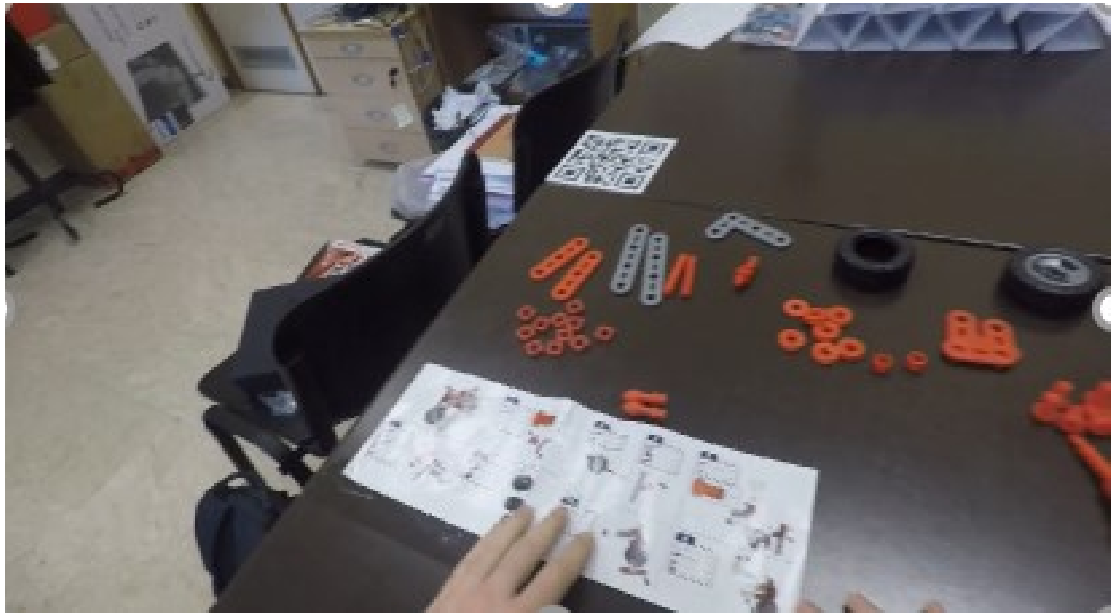
Mayol-Cuevas, W. W., Tordoff, B. J., & Murray, D. W. (2009). On the choice and placement of wearable vision sensors. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *39*(2), 414-425.

Data Acquisition – Field of View (FOV)

A wide FOV allows to capture more scene but it may introduce distortion

Narrow Angle

Wide Angle



Data Acquisition – Other Modalities – Depth

- Depth can improve scene understanding by highlighting the position of objects and hands;

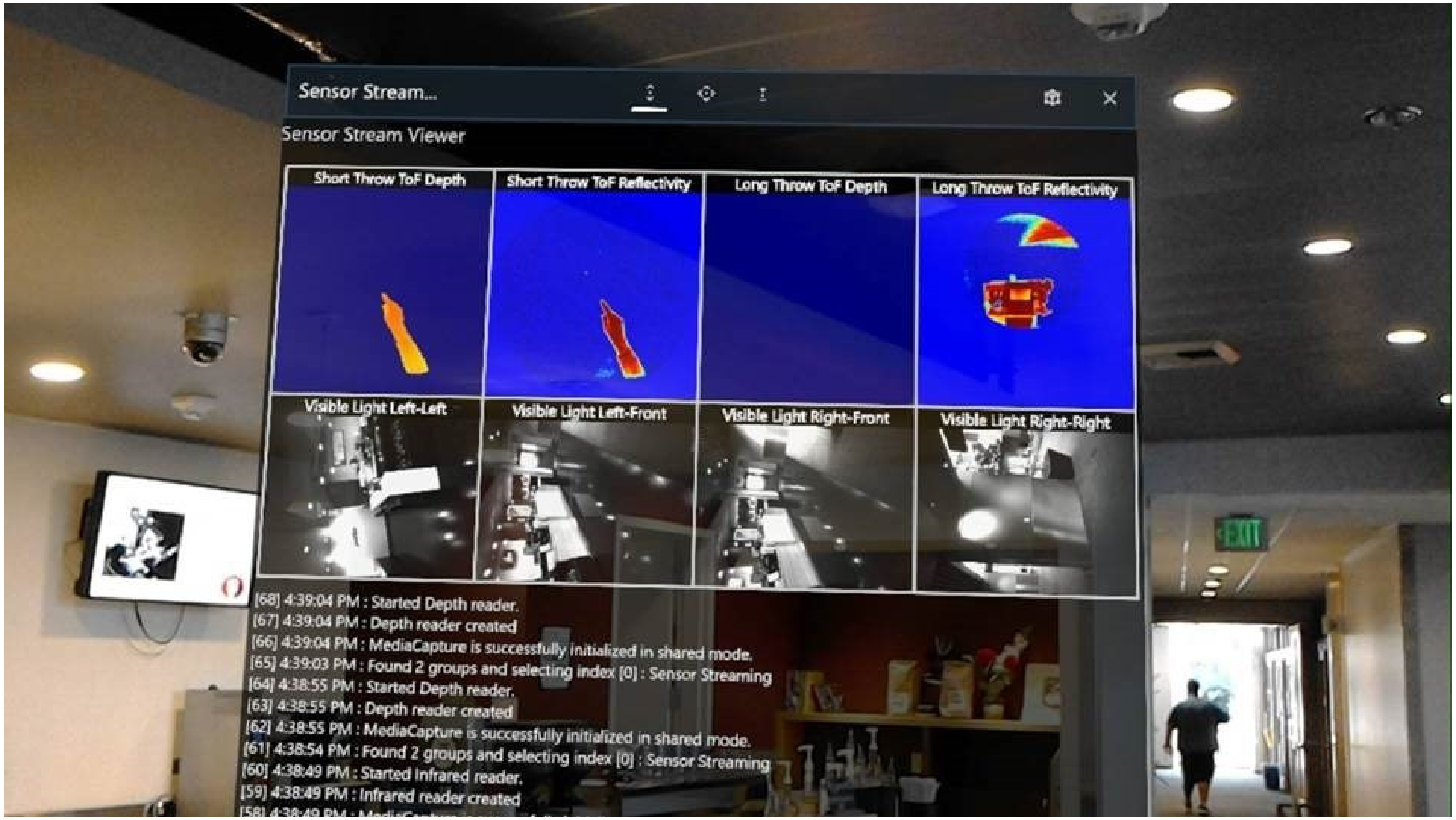
Wan, S., & Aggarwal, J. K. (2015). Mining discriminative states of hands and objects to recognize egocentric actions with a wearable RGBD camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 36-43).

Data Acquisition – Other Modalities – Depth (2)

https://github.com/microsoft/HoloLensForCV

- Microsoft HoloLens has a «Research Mode» which allows to access:
  - short-range depth
  - long-range depth;
  - IR reflectivity;

Microsoft HoloLens Research Mode



https://docs.microsoft.com/en-us/windows/mixed-reality/research-mode

Data Acquisition – Other Modalities – Gaze

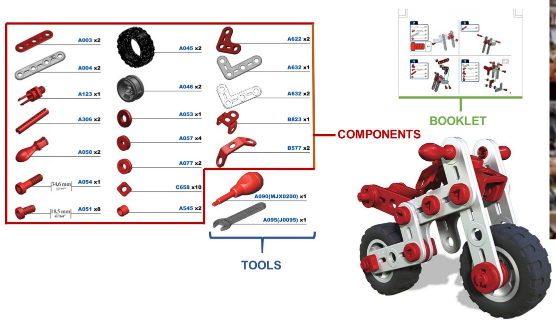
Gaze can give information on what the user is paying attention to.

However, gaze trackers generally require a calibration process (and some expertise).

F. Ragusa, A. Furnari, S. Livatino, G. M. Farinella. The MECCANO Dataset: Understanding Human-Object Interactions from Egocentric Videos in an Industrial-like Domain. WACV 2021 (ORAL) (https://arxiv.org/abs/2010.05654).

Datasets

MECCANO



Plizzari, C., Goletto, G., Furnari, A., Bansal, S., Ragusa, F., Farinella, G. M., Damen., D. & Tommasi, T. (2023). An Outlook into the Future of Egocentric Vision. *arXiv preprint arXiv:2308.07123*.

And More

Datasets (non-exhaustive)

Dataset	URL	Settings	Annotations	Goal
EGO-EXO4D	https://ego-exo4d-data.org/	839 participants performing procedural and physical activities.	Natural language descriptions, segmentation masks, temporal segments of keysteps, task-graphs, proficency labels, 3D human pose	Keystep Recognition, Proficiency Estimation, Relation, Pose Estimation
EGO4D	https://ego4d-data.org/	931 participants performing different activities in different domains.	Different temporal and spatial annotations related to 5 benchmarks	Episodic Memory, Hand-Object Interaction, Audio-Visual Diarization, Social Interactions, Forecasting
EPIC-KITCHENS-100	https://epic-kitchens.github.io/2020-100	Subjects performing unscripted actions in their native kitchens.	Temporal segments	Action recognition, detection, anticipation, retrieval.
MECCANO	https://iplab.dmi.unict.it/MECCANO/	20 subjects assembling a toy motorbike.	Temporal segments, active objects, human-object interactions	Action recognition, Active object detection, Egocentric Human-Object Interaction Detection
ASSEMBLY101	https://assembly-101.github.io/	53 subjects assembling in a cage settings 101 children's toys.	Temporal segments, 3D hand poses	Action recognition, Action Anticipation, Temporal Segmentation

Datasets (non-exhaustive)

Dataset	URL	Settings	Annotations	Goal
ENIGMA-51	https://iplab.dmi.unict.it/ENIGMA-51/	Participants performing procedural activities in the industrial domain.	Textual procedures, Hand and Object annotations, human-object interactions, next-object interactions	Untrimmed temporal annotations of human-object interactions, Egocentric Human-object interactions, short-term object interaction anticipation, NLU of intents and entities
HOLOASSIST	https://holoassist.github.io/	350 instructor-performer pairs which collaboratively complete physical manipulation tasks.	Action and conversational annotations	Action recognition and anticipation, mistake detection, inervention type prediction, 3D hand pose forecasting
ARIA Digital Twin	https://www.projectaria.com/datasets/adt/			

Datasets (non-exhaustive)

Dataset	URL	Settings	Annotations	Goal
EPIC-KITCHENS 2018	https://epic-kitchens.github.io/2018	32 subjects performing unscripted actions in their native environments	action segments, object annotations	Action recognition, Action Anticipation, Object Detection
Charade-Ego	https://allenai.org/plato/charades/	paired first-third person videos	action classes	Action recognition
EGTEA Gaze+	http://ai.stanford.edu/~alireza/GTEA/	32 subjects, 86 sessions, 28 hours	action segments, gaze, hand masks	Understading daily activities, action recognition
ADL	https://www.csee.umbc.edu/~hpirsiav/papers/ADLdataset/	20 subjects performing daily activities in their native environments	activity segments, objects	Detecting activities of daily living
CMU kitchen	http://www.cs.cmu.edu/~espriggs/cmu-mmac/annotations/	multimodal, 18 subjects cooking 5 different recipes: brownies, eggs, pizza, salad, sandwiche	action segments	Understading daily activities
EgoSeg	http://www.vision.huji.ac.il/egoseg/	Long term actions (walking, running, driving, etc.)	long term activity	Temporal Segmentation, Indexing

Datasets (non-exhaustive)

Dataset	URL	Settings	Annotations	Goal
First-Person Social Interactions	http://ai.stanford.edu/~alireza/Disney/	8 subjects at disneyworld	Activities: walking, waiting, gathering, sitting, buying something, eating, etc.	Recognizing social interactions
UEC Dataset	http://www.cs.cmu.edu/~kkitani/datasets/	two choreographed datasets with different egoactions (walk, jump, climb, etc.) + 6 youtube sports videos	activities	Unsupervised activity recognition
JPL	http://michaelryoo.com/jpl-interaction.html	interaction with a robot	activities performed on the robot + pose	Interaction recognition/prediction
Multimodal Egocentric Activity Dataset	http://people.sutd.edu.sg/~1000892/dataset	15 seconds clips of 20 activities	activity (walking, elevator, etc.)	Life-logging
LENA: An egocentric video database of visual lifelog	http://people.sutd.edu.sg/~1000892/dataset	13 activities performed by 10 subjects (Google Glass)	activity (walking, elevator, etc.)	Life-logging

Datasets (non-exhaustive)

Dataset	URL	Settings	Annotations	Goal
FPPA	http://tamaraberg.com/prediction/Prediction.html	Five subjects performing 5 daily actions	activity (drinking water, putting on clothes, etc.)	Temporal prediction
UT Egocentric	http://vision.cs.utexas.edu/projects/egocentric/index.html	3-5 hours long videos capturing a person's day	important regions	Summarization
VINST/ Visual Diaries	http://www.csc.kth.se/cvap/vinst/NovEgoMotion.html	31 videos capturing the visual experience of a subject walkin from metro station to work	location id, novel egomotion	Novelty detection
Bristol Egocentric Object Interaction (BEOID)	https://www.cs.bris.ac.uk/~damen/BEOID/	8 subjects, six locations. Interaction with objects and environment	gaze, objects, mode of interaction (pick, plug, etc.)	Provide assistance on object usage
Object Search Dataset	https://github.com/Mengmi/deepfuturegaze_gan	57 sequences of 55 subjects on search and retrieval tasks	gaze	gaze prediction

Datasets (non-exhaustive)

Dataset	URL	Settings	Annotations	Goal
UNICT-VEDI	http://iplab.dmi.unict.it/VEDI/	different subjects visiting a museum	location, observed objects	localizing visitors of a museum and estimating their attention
UNICT-VEDI-POI	http://iplab.dmi.unict.it/VEDI_POIs/	different subjects visiting a museum	object bounding boxes annotations, observed objects	recognizing points of interest observed by the visitors
Simulated Egocentric Navigations	http://iplab.dmi.unict.it/SimulatedEgocentricNavigations/	simulated navigations of a virtual agent within a large building	3-DOF pose of the agent in each image	egocentric localization
EgoCart	http://iplab.dmi.unict.it/EgocentricShoppingCartLocalization/	egocentric images collected by a shopping cart in a retail store	3-DOF pose of the shopping cart in each image	egocentric localization
Unsupervised Segmentation of Daily Livign Activities	http://iplab.dmi.unict.it/dailylivingactivities	egocentric videos of daily activities	activities	unsupervised segmentation with respect to the activities

Datasets (non-exhaustive)

Dataset	URL	Settings	Annotations	Goal
Visual Market Basket Analysis	http://iplab.dmi.unict.it/vmba/	egocentric images colelcted by a shopping cart in a retail store	class-location of each image	egocentric localization
Location Based Segmentation of Egocentric Videos	http://iplab.dmi.unict.it/PersonalLocationSegmentation/	egocentric videos of daily activities	location classes	egocentric localization, video indexing
Recognition of Personal Locations from Egocentric Videos	http://iplab.dmi.unict.it/PersonalLocations/	egocentric videos clips of daily activities	location classes	recognizing personal locations
EgoGesture	http://www.nlpr.ia.ac.cn/iva/yfzhang/datasets/egogesture.html	2k videos from 50 subjects performing 83 gestures	Gesture labels, depth	Gesture recognition
EgoHands	http://vision.soic.indiana.edu/projects/egohands/	48 videos of interactions between two people	Hand segmentation masks	Egocentric hand segmentation
DoMSEV	http://www.verlab.dcc.ufmg.br/semantic-hyperlapse/cvpr2018-dataset/	80 hours/different activities	Scene/Action labels with IMU, GPS mad depth	Summarization

Datasets (non-exhaustive)

Dataset	URL	Settings	Annotations	Goal
EGO-HPE	http://imagelab.ing.unimore.it/imagelab2015/researchactivity.asp?idAttivita=23	Egocentric videos for head pose estimation	Head pose of the subjects	Head-pose estimation
EGO-GROUP	http://imagelab.ing.unimore.it/imagelab2015/researchactivity.asp?idAttivita=23	18 videos of people engaging social relationships	Social relationships	Understanding social relationships
DR(eye)VE	http://aimagelab.ing.unimore.it/dreyeve	74 videos of people driving	Eye fixations	Autonomous and assisted driving
THU-READ	http://ivg.au.tsinghua.edu.cn/dataset/THU_READ.php	8 subjects performing 40 actions with a head-mounted RGBD camera	Action segments	RGBD egocentric action recognition
EGO-CH	https://iplab.dmi.unict.it/EGO-CH/	70 subjects visiting two cultural sites in Sicily, Italy.	Temporal segments, room-based localization, objects	Room-basd localization, Object detection, Behavioral analysis

Fundamental Tasks

12 Egocentric Vision Research Tasks

1. Localisation
2. 3D Scene Understanding
3. Anticipation
4. Action Recognition
5. Gaze Understanding and Prediction
6. Social Behaviour Understanding
7. Full Body Pose Estimation
8. Hand and Hand-Object Interactions
9. Person Identification
10. Privacy
11. Summarisation
12. Visual Question Answering

Plizzari, C., Goletto, G., Furnari, A., Bansal, S., Ragusa, F., Farinella, G. M., Damen., D. & Tommasi, T. (2023). An Outlook into the Future of Egocentric Vision. *arXiv preprint arXiv:2308.07123*.

Fundamental Tasks

12 Egocentric Vision Research Tasks

1. **Localisation**
2. 3D Scene Understanding
3. **Anticipation**
4. **Action Recognition**
5. Gaze Understanding and Prediction
6. Social Behaviour Understanding
7. Full Body Pose Estimation
8. **Hand and Hand-Object Interactions**
9. Person Identification
10. Privacy
11. Summarisation
12. Visual Question Answering

Plizzari, C., Goletto, G., Furnari, A., Bansal, S., Ragusa, F., Farinella, G. M., Damen., D. & Tommasi, T. (2023). An Outlook into the Future of Egocentric Vision. *arXiv preprint arXiv:2308.07123*.

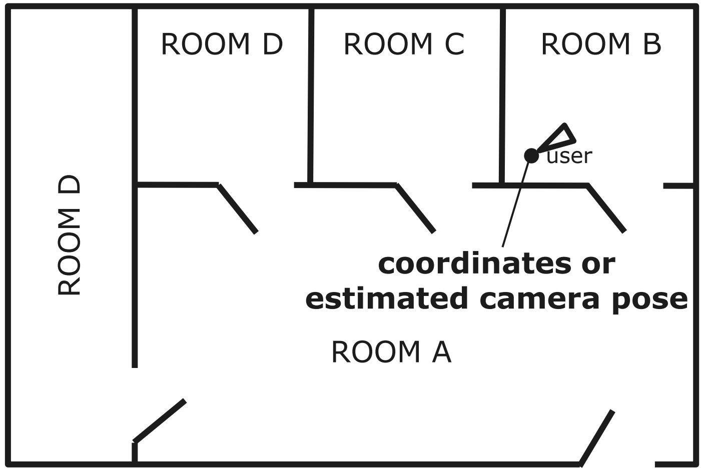
Localization

Localization – Levels of Granularity

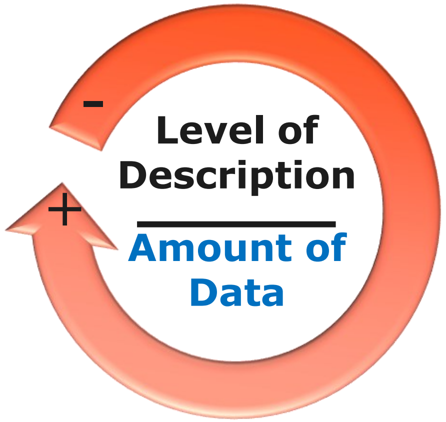
SCENE RECOGNITION

off-the-shelf detectors

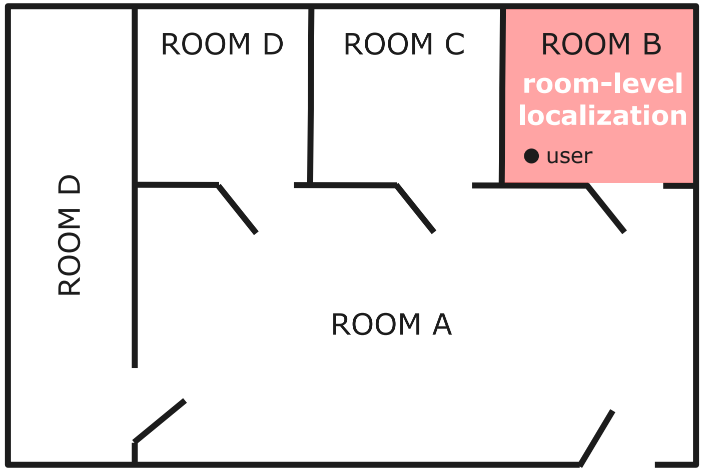
CAMERA POSE-ESTIMATION



3D reconstruction of the building



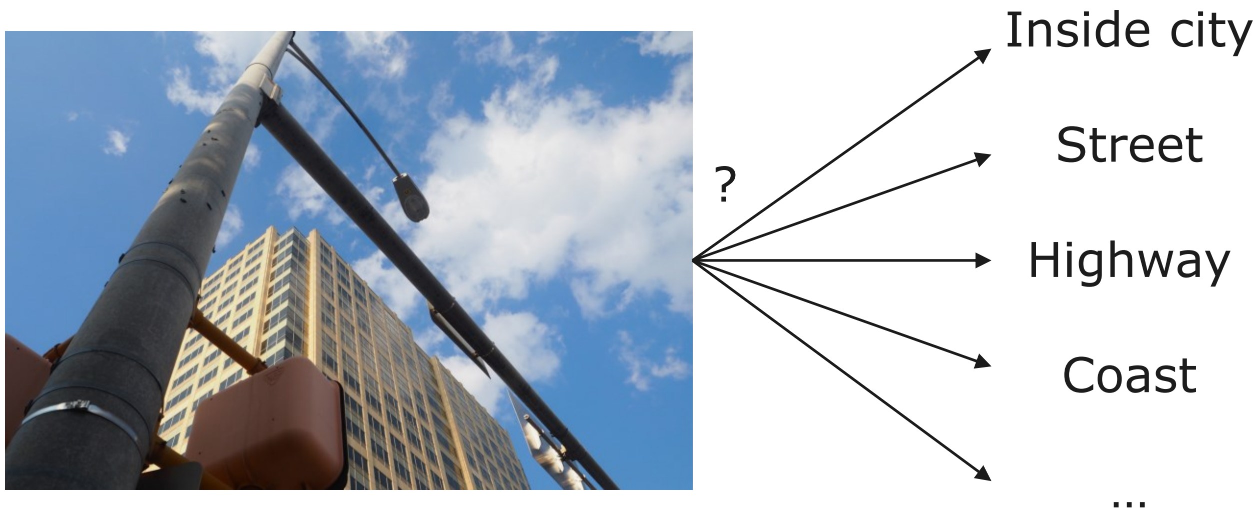
ROOM-LEVEL RECOGNITION



moderate amount of training data

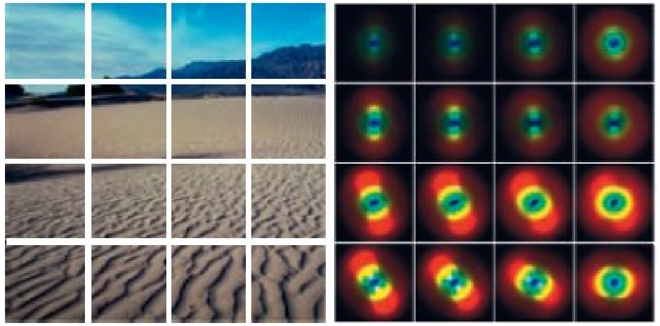
Scene Recognition

- The most basic form of localization;
- Tells what kind of scene the user is in;
- Useful to distinguish between (even for unseen places) :
  - indoor/outdoor
  - natural/artificial
  - conf. room
  - Office
- Can use off-the-shelf detections.



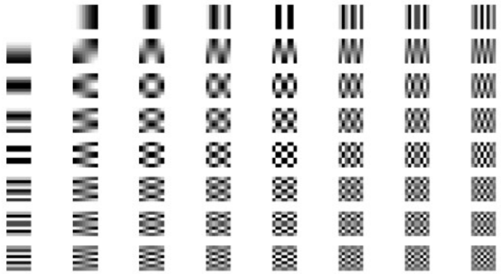
COMPUTATIONALLY INEXPENSIVE ALGORITHMS

GIST Descriptor



Oliva, Aude, and Antonio Torralba. "Modeling the shape of the scene: A holistic representation of the spatial envelope." International journal of computer vision 42.3 (2001): 145-175.

DCT-GIST (runs on the IGP pipeline)



G. M. Farinella, D. Ravì, V. Tomaselli, M. Guarnera, S. Battiato, *"Representing scenes for real-time context classification on mobile devices"*, Pattern Recognition, Elsevier, ISSN 0031-3203, Vol. 48, N. 4, pp. 1082-1096, doi: 10.1016/j.patcog.2014.05.014, 2015

Scene Recognition – Places

DATA & CODE HERE -> http://places2.csail.mit.edu/

GT: cafeteria
top-1: cafeteria (0.179)
top-2: restaurant (0.167)
top-3: dining hall (0.091)
top-4: coffee shop (0.086)
top-5: restaurant patio (0.080)

- Places is a large (10M images – 400+ classes) dataset for scene recognition;
- CNN models trained to recognize 365 scene classes available for download;
- Can be used off-the-shelf!

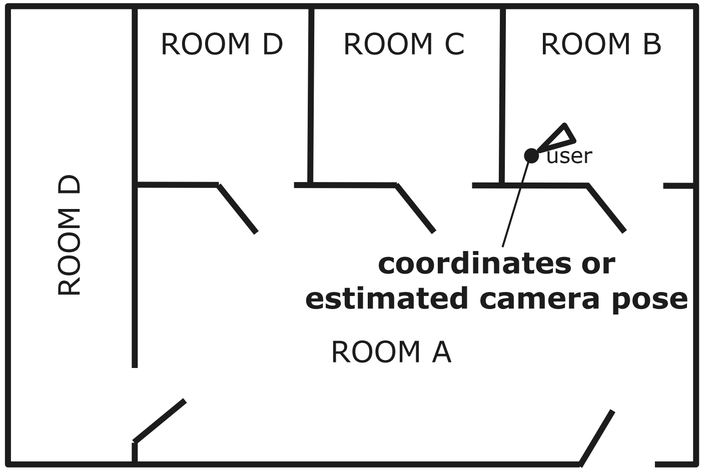
A 10 million Image Database for Scene Recognition B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba *IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017*

Localization – Levels of Granularity

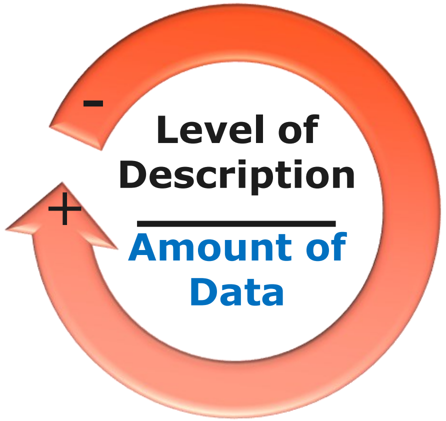
SCENE RECOGNITION

off-the-shelf detectors

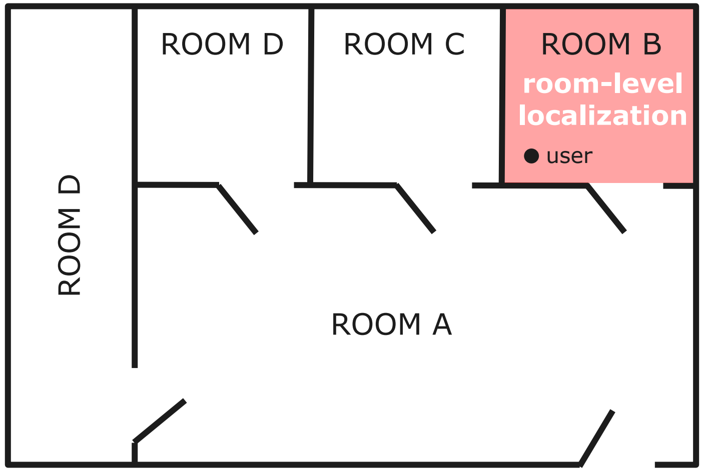
CAMERA POSE-ESTIMATION



3D reconstruction of the building



ROOM-LEVEL RECOGNITION



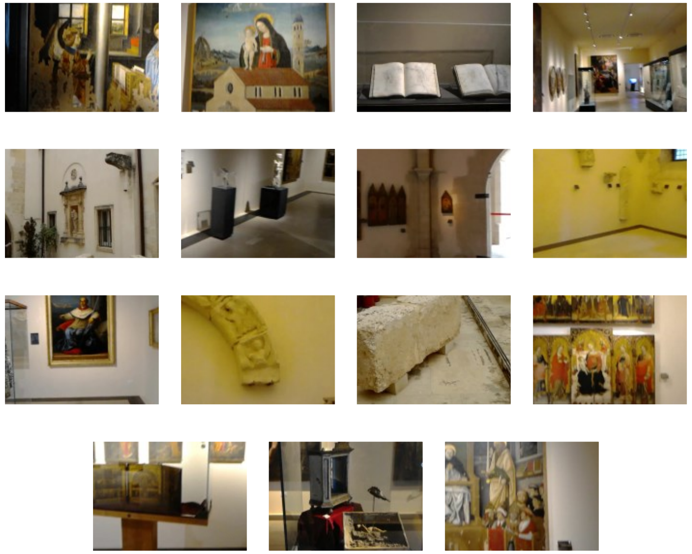
moderate amount of training data

Room-Level Localization – Museums

Cultural Site (e.g., museum) divided into contexts (e.g., rooms)

(videos acquired in the different contexts)

Training Set



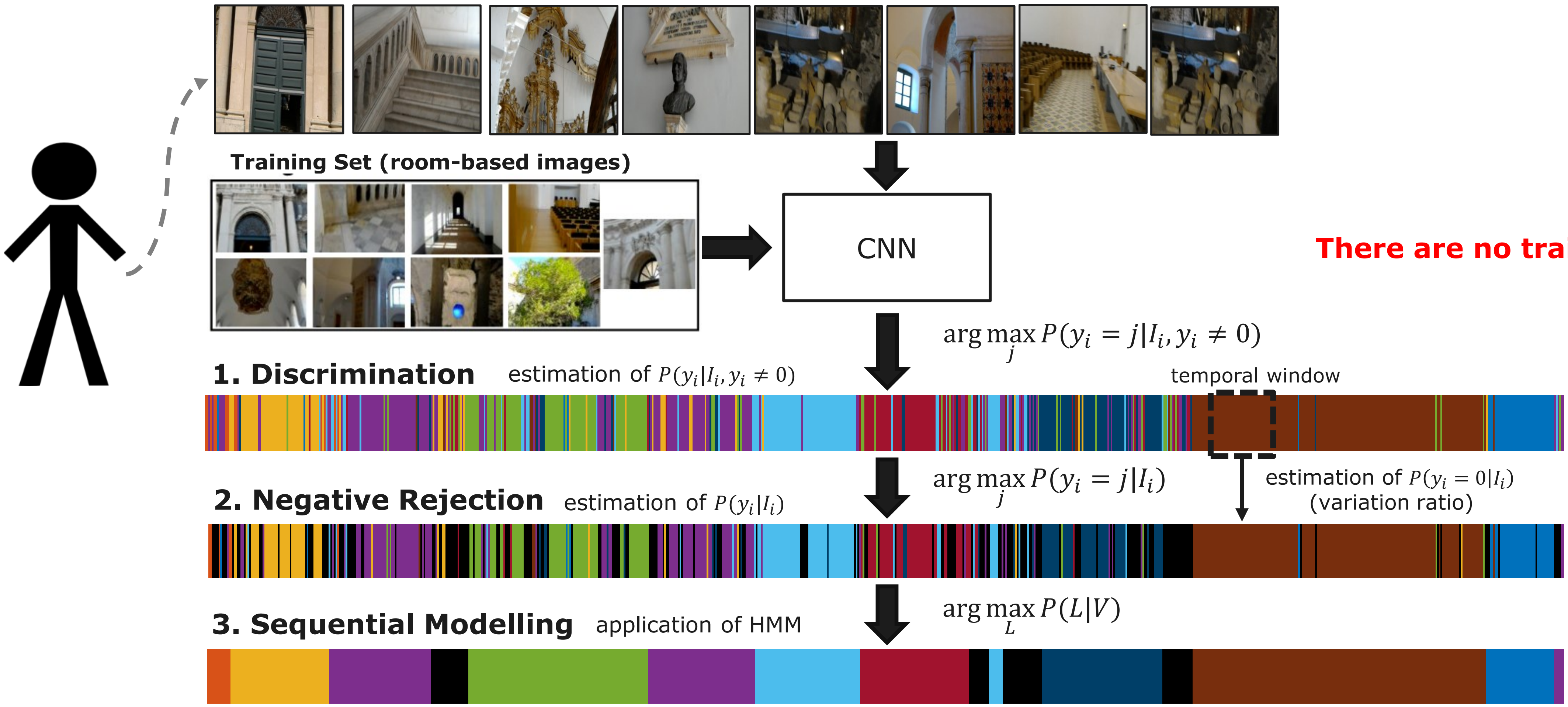
(frames extracted from videos acquired in the different contexts)

F. Ragusa, A. Furnari, S. Battiato, G. Signorello, G. M. Farinella (2020). EGO-CH: Dataset and Fundamental Tasks for Visitors Behavioral Understanding using Egocentric Vision . Pattern Recognition Letters.

Room-Level Localization – Full Model

CODE HERE -> https://iplab.dmi.unict.it/VEDI/

https://iplab.dmi.unict.it/PersonalLocationSegmentation/



There are no training negatives!

A. Furnari, G. M. Farinella, S. Battiato, Personal-Location-Based Temporal Segmentation of Egocentric Video for Lifelogging Applications, Journal of Visual Communication and Image Representation, 2017.

F. Ragusa, A. Furnari, S. Battiato, G. Signorello, G. M. Farinella. Egocentric Visitors Localization in Cultural Sites. In Journal on Computing and Cultural Heritage (JOCCH), 2019.

Room-Level Localization – Demo

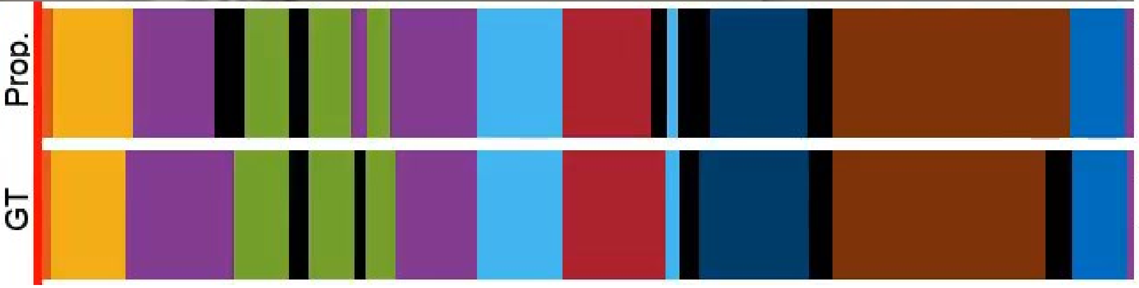
Cutgana
CENTRO UNIVERSITARIO PER LA TUTELA E LA GESTIONE DEGLI AMBIENTI NATURALI E DEGLI AGRO-ECOSISTEMI

VEDI – Vision Exploitation for Data Interpration, PON MISE Horizon 2020

F. Ragusa, A. Furnari, S. Battiato, G. Signorello, G. M. Farinella

Time Spent at Location

LOC	EST	GT
G. Novizi	00:00	00:00
Cortile	00:03	00:03
Scalone	00:00	00:00
Corridoi	00:00	00:00
C. Notte	00:00	00:00
Antiref.	00:00	00:00
S. Mazz.	00:00	00:00
Cucina	00:00	00:00
Ventre	00:00	00:00
Negative	00:00	00:00



Detected Shots for Storyboard Summary

Estimated Probabilities	Predicted Class	GT Class
Giardino dei Novizi		
Cortile	●	●
Scalone Monumentale		
Corridoi		
Coro di Notte		
Antirefettorio		
Aula Santo Mazzarino		
Cucina		
Ventre		
Negative		

Localization-based analysis

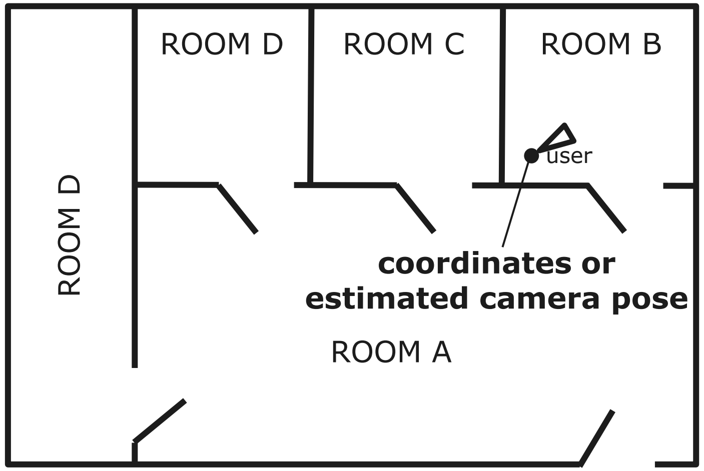
G. M. Farinella, G. Signorello, S. Battiato, A. Furnari, F. Ragusa, R. Leonardi, E. Ragusa, E. Scuderi, A. Lopes, L. Santo, M. Samarotto. VEDI: Vision Exploitation for Data Interpretation. In 20th International Conference on Image Analysis and Processing (ICIAP), 2019

Localization – Levels of Granularity

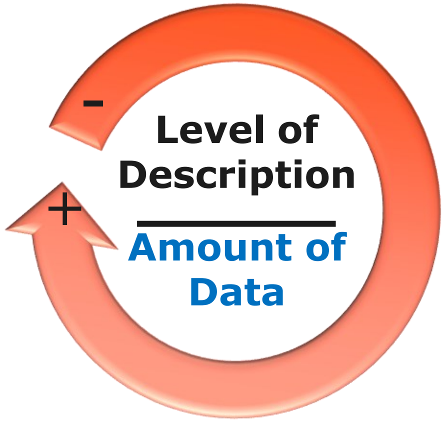
SCENE RECOGNITION

off-the-shelf detectors

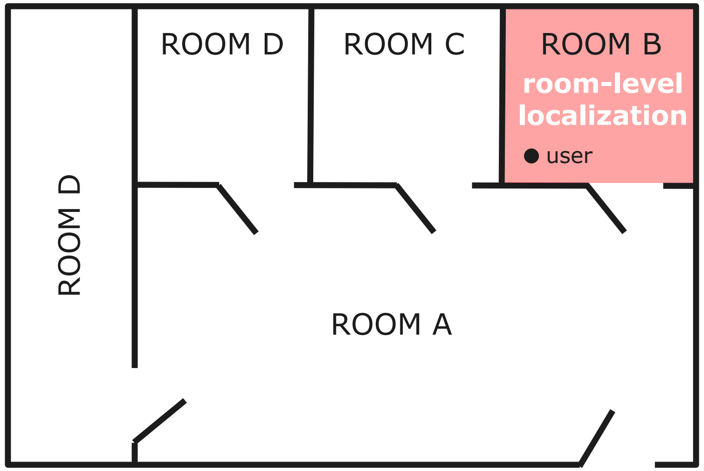
CAMERA POSE-ESTIMATION



3D reconstruction of the building



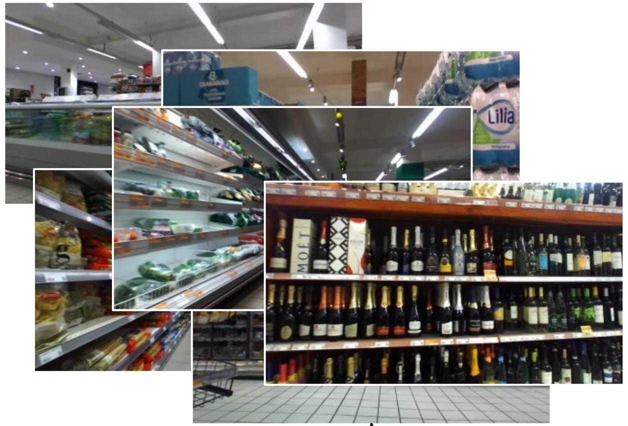
ROOM-LEVEL RECOGNITION



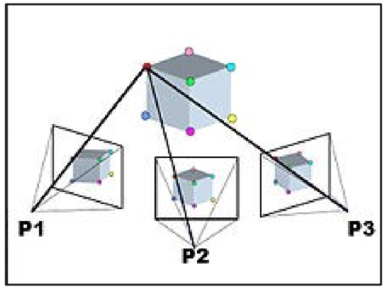
moderate amount of training data

Camera Pose Estimation – Dataset Creation

Images



Structure from Motion (SfM)

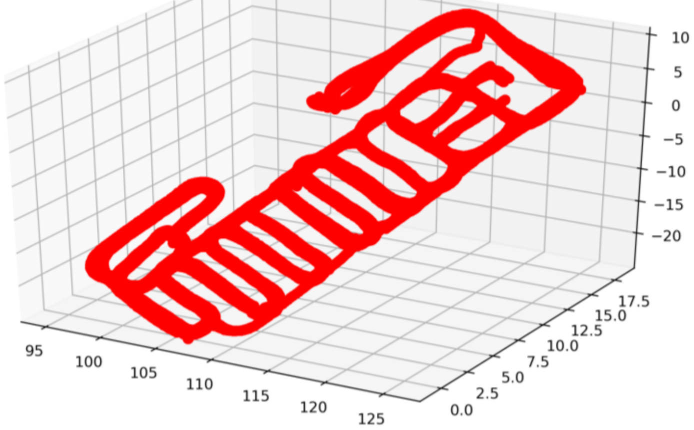


(P,Q)

Attach estimated 6DOF pose to each image

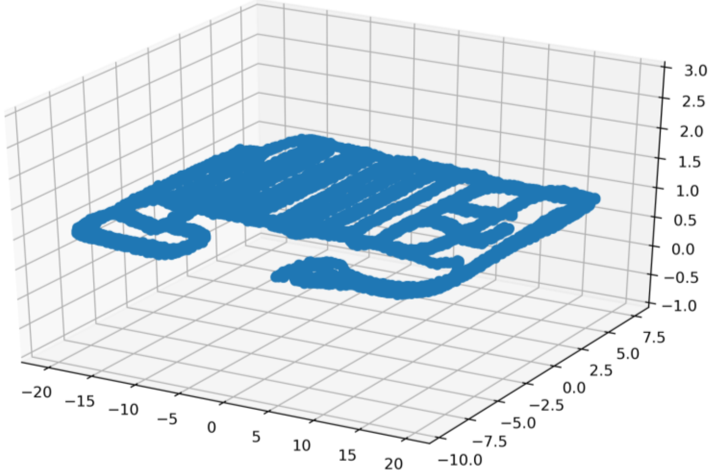
3D Model

Arbitrary Coordinate System (pose/scale)

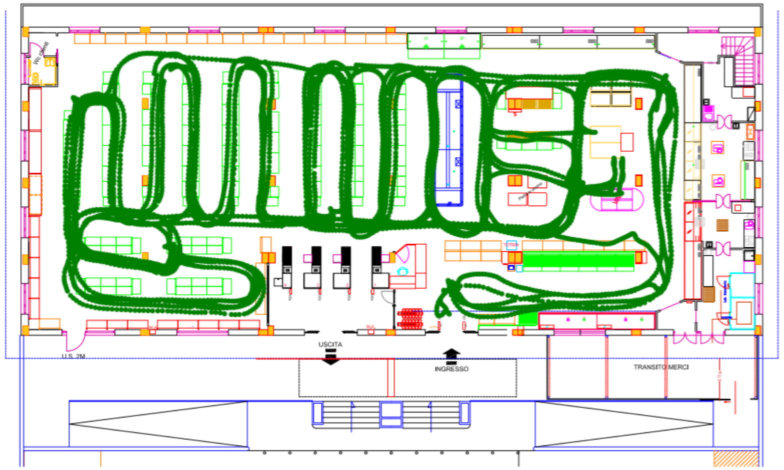


camera poses

PCA



rotated poses



scaled/aligned poses

Camera Pose Estimation – Labeling

Structure from Motion attaches every input image to a 3D model.

Many options available:

COLMAP (free)

https://colmap.github.io/

Visual SFM (free)

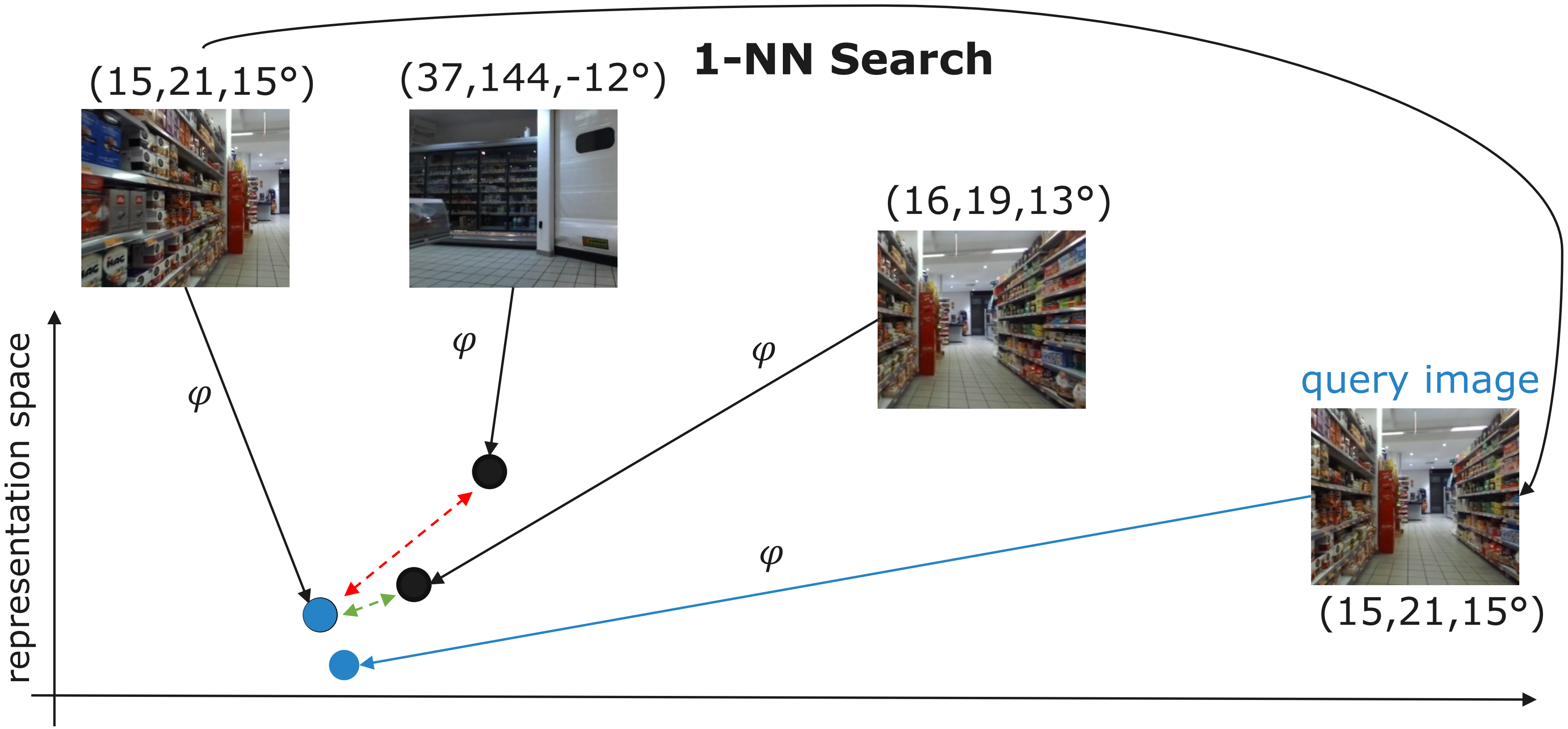
http://ccwu.me/vsfm/

3D Zephir (paid)

https://www.3dflow.net/it/3df-zephyr-pro-3d-models-from-photos/

Camera Pose Estimation – Retrieval Approach

Use deep metric learning to learn a representation function φ which maps close to each other images of nearby locations



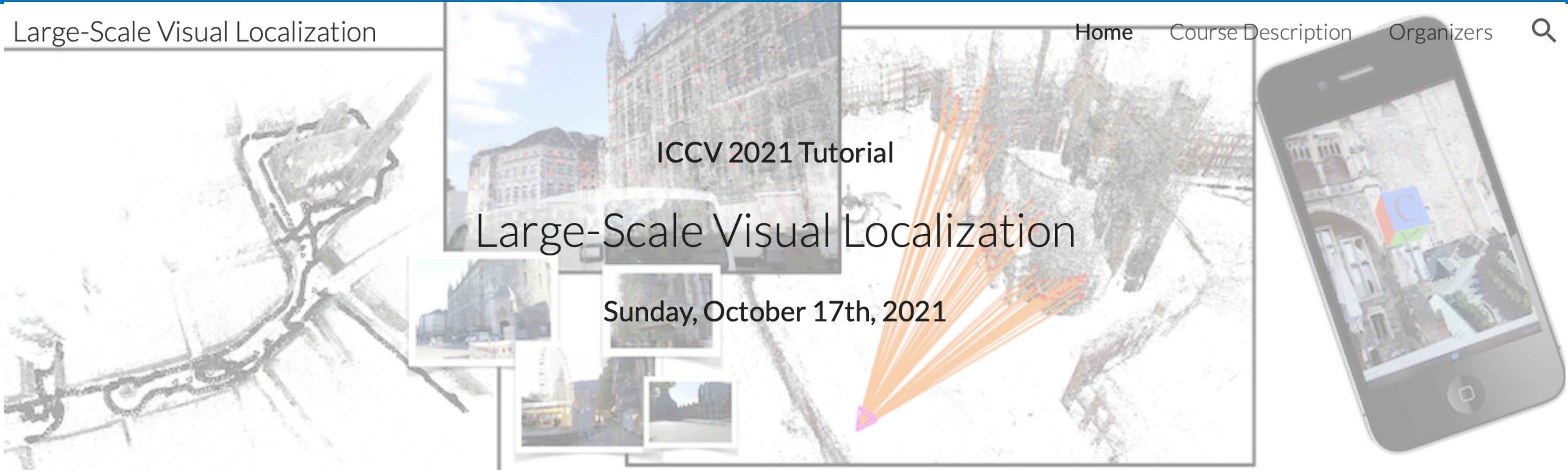
E. Spera, A. Furnari, S. Battiato, G. M. Farinella, Egocentric Shopping Cart Localization, International Conference on Pattern Recognition (ICPR), 2018
S. A. Orlando, A. Furnari, S. Battiato, G. M. Farinella. Image-Based Localization with Simulated Egocentric Navigations. VISAPP 2019

Università di Catania

Visual Localization Literature is Rich!

Large-Scale Visual Localization

Home Course Description Organizers



Course Information

- **When:** Sunday, October 17th, 2021
- **Where:** Online at https://youtu.be/RaVPiIGhdWk
- **Time:** half-day tutorial - starts at 2:30 pm CEST (ics)
- **Preliminary Schedule**
  - Part I: Image Retrieval for Coarse Localization (Giorgos, Yannis)
    - Image Retrieval & Visual Representation [50 min] (Giorgos) [slides]
    - Metric learning: knowledge transfer, data augmentation, and attention [20 min] (Yannis) [slides]

https://sites.google.com/view/lsvpr2021/home

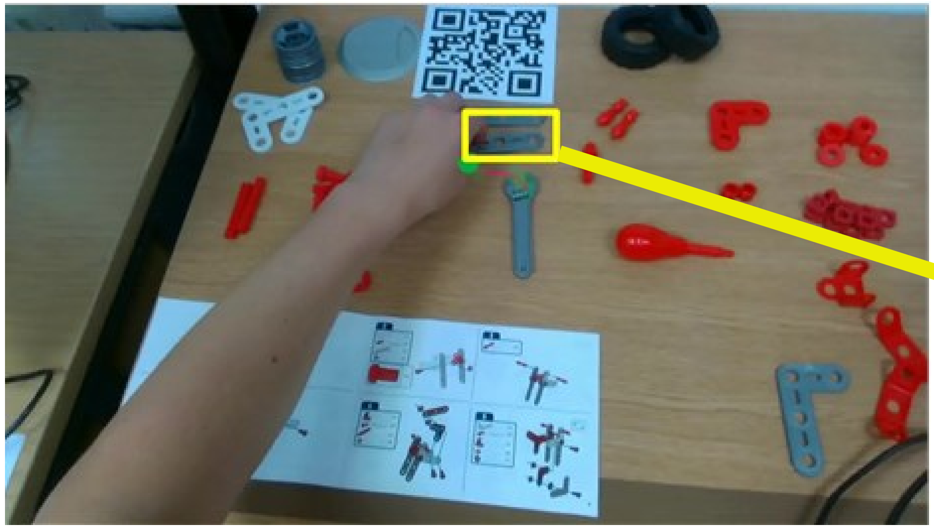
Object Detection/Interaction

Object Detection

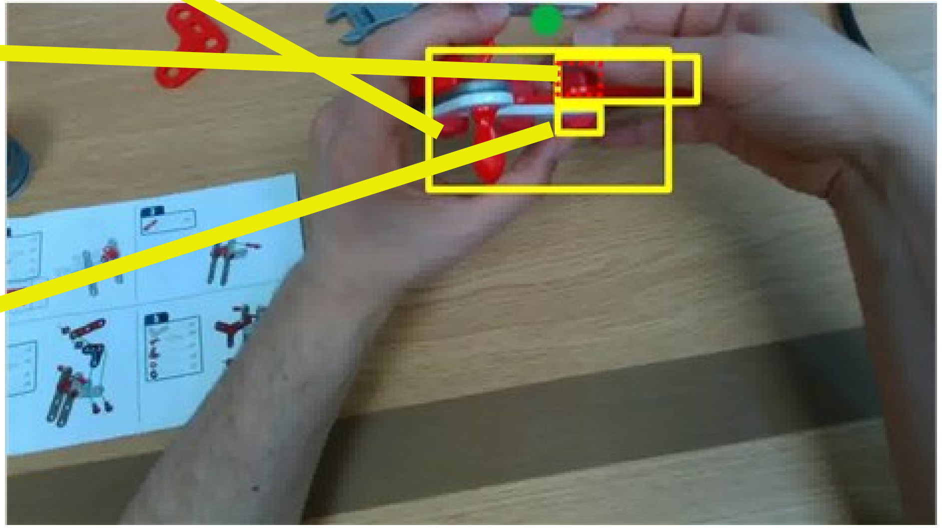
Objects and Actions are tight!

Useful to know what is in the scene

Useful to know what actions can be performed

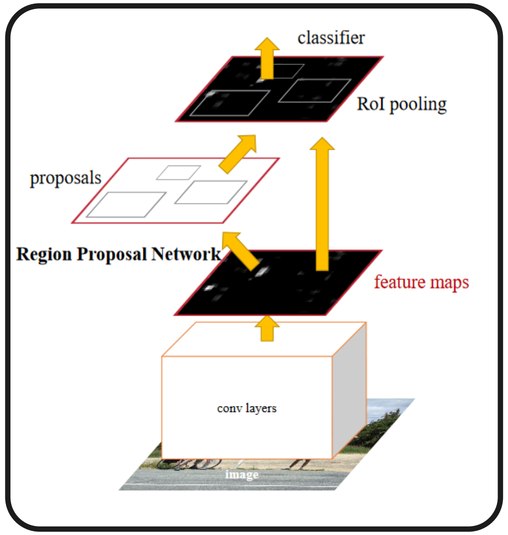


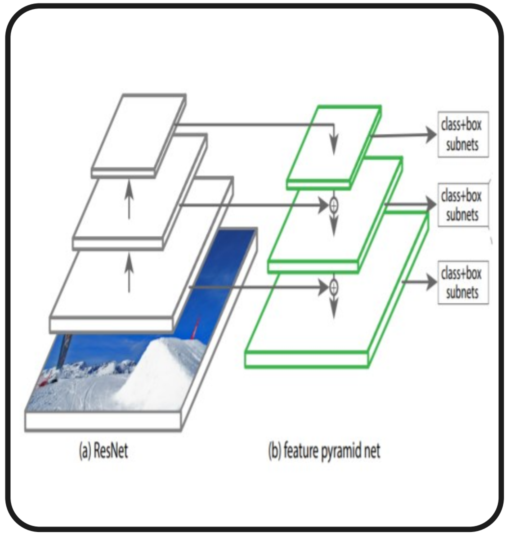
ID	Class
0	instruction booklet
1	gray_angled_perforated_bar
2	partial_model
3	white_angled_perforated_bar
4	wrench
5	screwdriver
6	gray_perforated_bar
7	wheels_axle
8	red_angled_perforated_bar
9	red_perforated_bar
10	rod
11	handlebar
12	screw
13	tire
14	rim
15	washer
16	red_perforated_junction_bar
17	red_4_perforated_junction_bar
18	bolt
19	roller

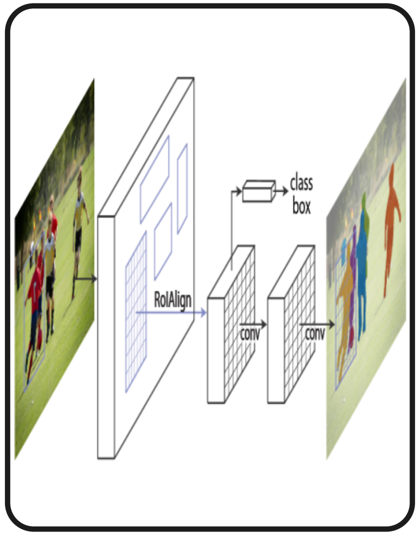


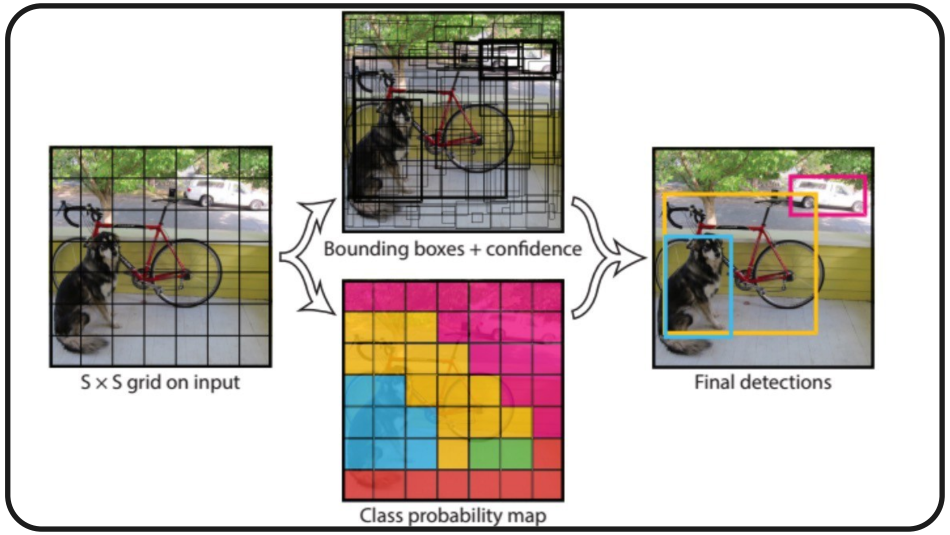
F. Ragusa, A. Furnari, G. M. Farinella. MECCANO: A Multimodal Egocentric Dataset for Humans Behavior Understanding in the Industrial-like Domain. Computer Vision and Image Understanding (CVIU), 2023 (https://arxiv.org/abs/2209.08691).

Off-the-shelf object detectors









Faster-RCNN
(bounding boxes)

RetinaNet
(bounding boxes - faster)

Mask-RCNN
(boxes + segments)

YOLO
(much faster, but less accurate)

https://github.com/facebookresearch/detectron2

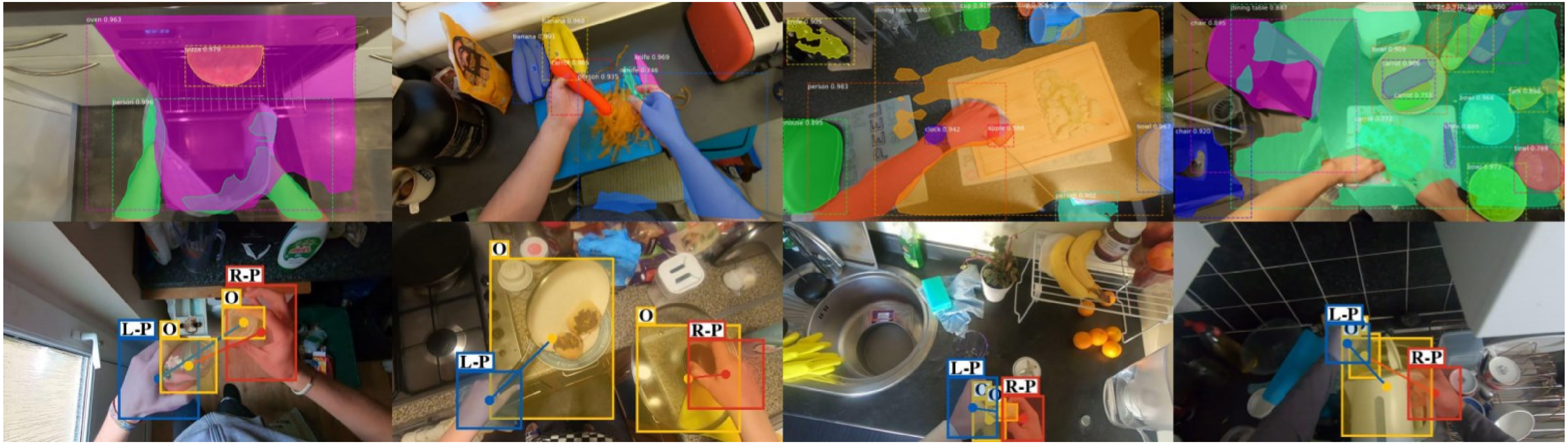
https://pjreddie.com/darknet/yolo/

Transformer-Based Detectors: https://github.com/IDEA-Research/awesome-detection-transformer

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*.
Joseph Redmon, Ali Farhadi, YOLO9000: Better, Faster, Stronger, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017
He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017, October). Mask r-cnn. In *Computer Vision (ICCV), 2017* (pp. 2980-2988). IEEE.

Off-the-shelf detectors on EPIC-KITCHENS

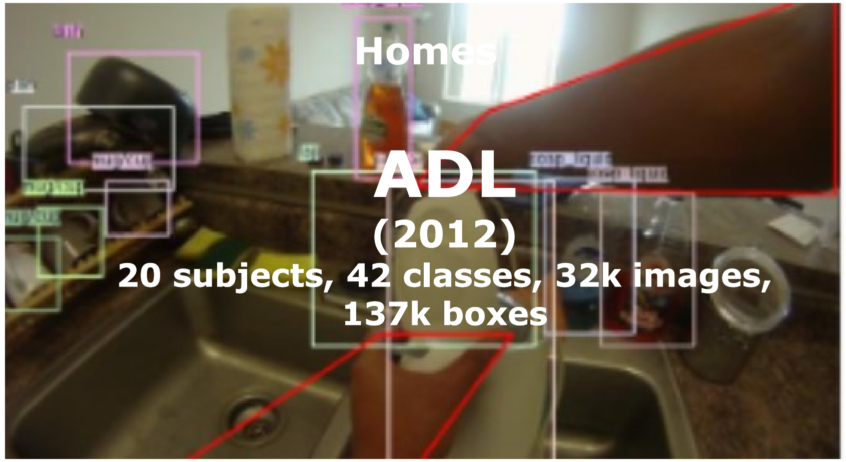
Depending on the scenario, off-the-shelf detectors can be a starting point, but they are not always accurate.



Damen, Doughty, Farinella, Furnari, Kazakos, Moltisanti, Munro, Price, Wray (2020). Rescaling Egocentric Vision. *arXiv preprint arXiv:2006.13256* (2020).

Università di Catania

Train/Finetune your own object detector



https://www.csee.umbc.edu/~hpirsiav/papers/ADLdataset/

http://epic-kitchens.github.io/

https://iplab.dmi.unict.it/EGO-CH/

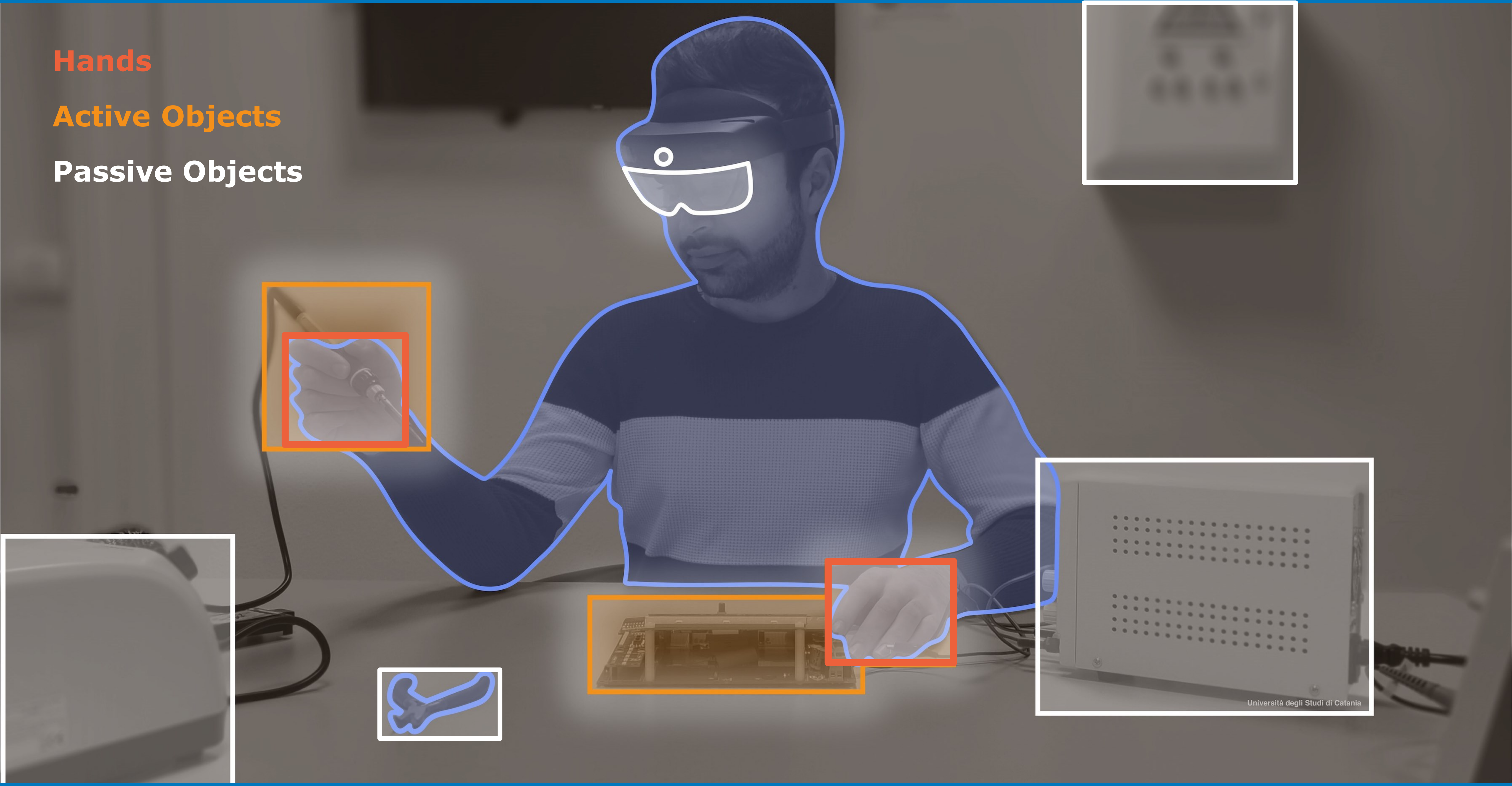
https://iplab.dmi.unict.it/MECCANO/

- In some scenarios, it could be necessary to fine-tune an object-detector with application-specific data.
- Main egocentric datasets providing bounding box annotations.
- EGO4D is multi- domain annotated with 295K bounding boxes.

NEW
EgoObjects!
114K annotated
frames
https://github.com/facebookresearch/EgoObjects

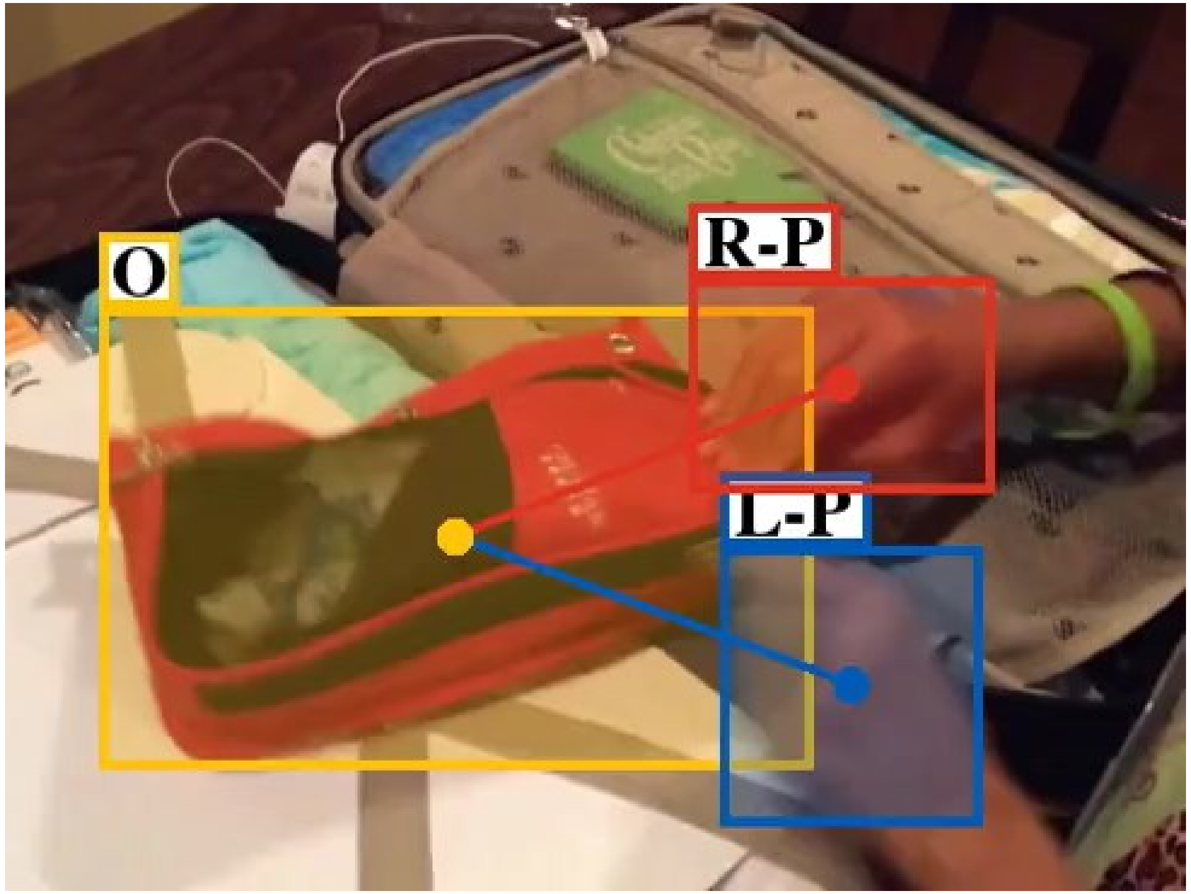
Università di Catania

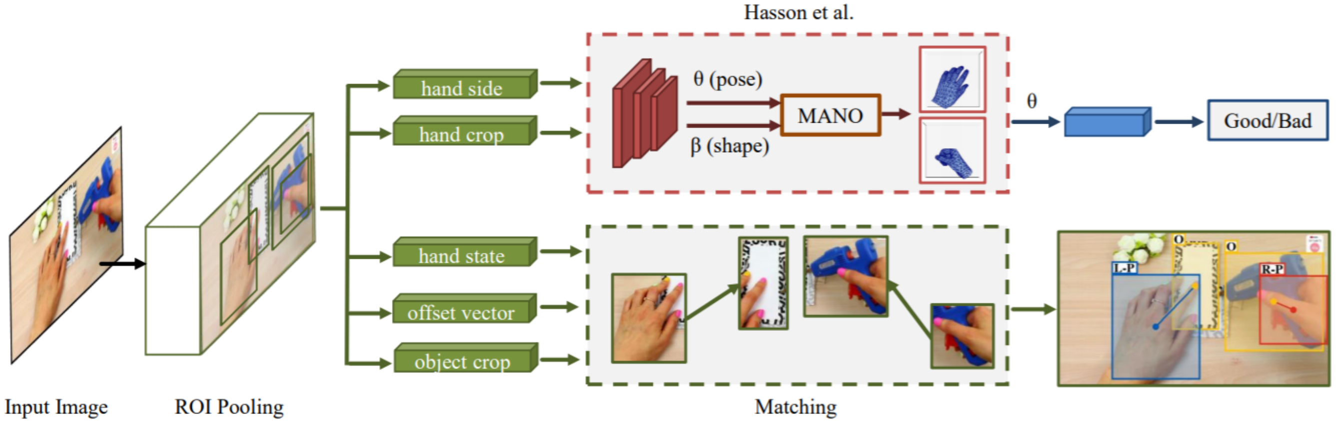
Understanding human-object interactions (HOI)



Hands in Contact – Hands + Objects

CODE & DATA HERE -> https://fouheylab.eecs.umich.edu/~dandans/projects/100DOH/





An «augmented» detector which recognizes:

- The left hand;
- The right hand;
- The interacted object.

VISOR DATASET

Darkhalil, Ahmad, et al. "Epic-kitchens visor benchmark: Video segmentations and object relations." *Advances in Neural Information Processing Systems* 35 (2022): 13745-13758.

Shan, D., Geng, J., Shu, M., & Fouhey, D. F. (2020). Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9869-9878).

Standard approach – relies on huge quantities of real data

Standard approach:

- Collect a lot of images and videos of construction sites;
- Label the data with domain-specific annotations;
- Train and test deep learning algorithms.

Learning in simulation

What if we could learn the «real thing» in simulation?

DATA HERE -> https://iplab.dmi.unict.it/EHOI_SYNTH/

Can simulated data help?

ENIGMA Laboratory

19 objects categories

Rosario Leonardi, Francesco Ragusa, Antonino Furnari, Giovanni Maria Farinella (2022). Egocentric Human-Object Interaction Detection Exploiting Synthetic Data . In International Conference on Image Analysis and Processing (ICIAP)

User-Object Interaction Synthetic Data Generation

Università di Catania

DATA GENERATION

Virtual Replica of Real Environment

Interaction Simulation

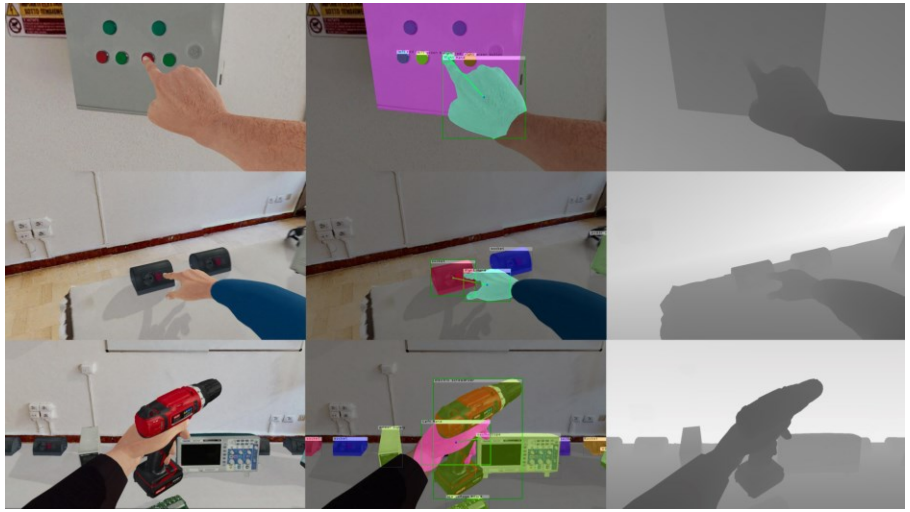
DATASET

Images
39304

Hand annotations
59860

Object annotations
237985

EHOI annotations
35416

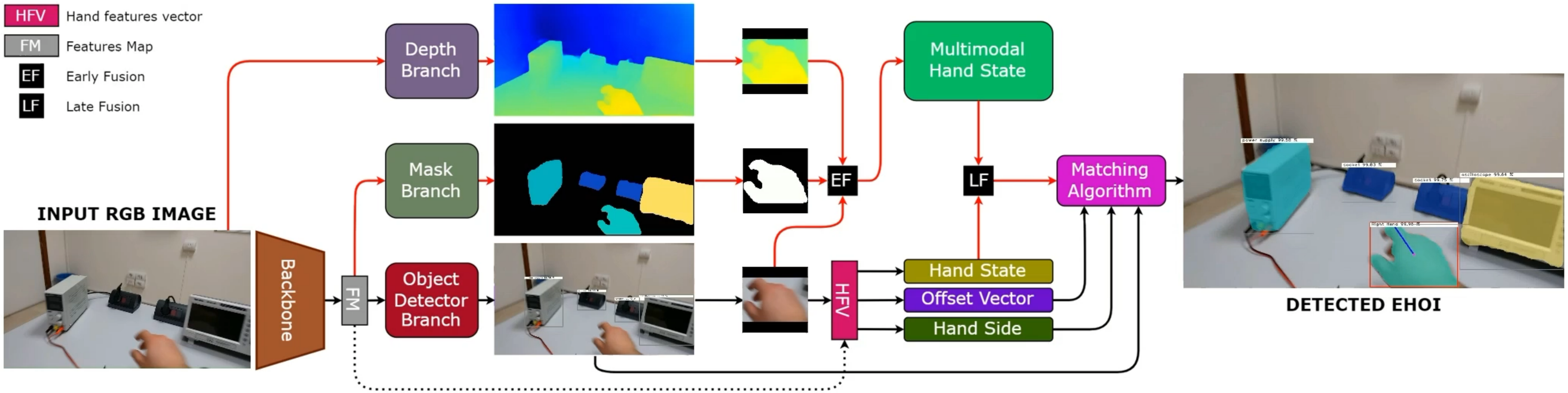


Rosario Leonardi, Francesco Ragusa, Antonino Furnari, Giovanni Maria Farinella (2022). Egocentric Human-Object Interaction Detection Exploiting Synthetic Data . In International Conference on Image Analysis and Processing (ICIAP)

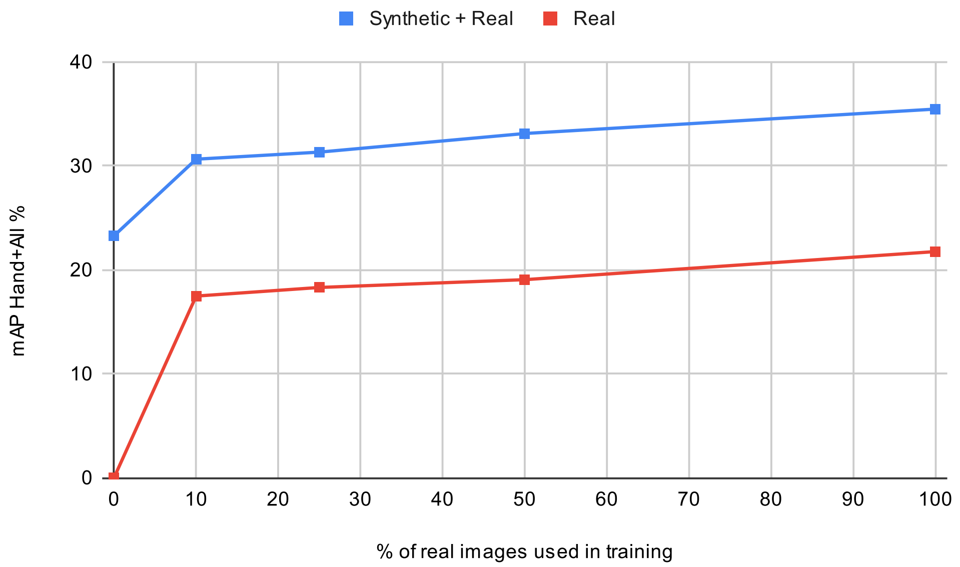
Università di Catania

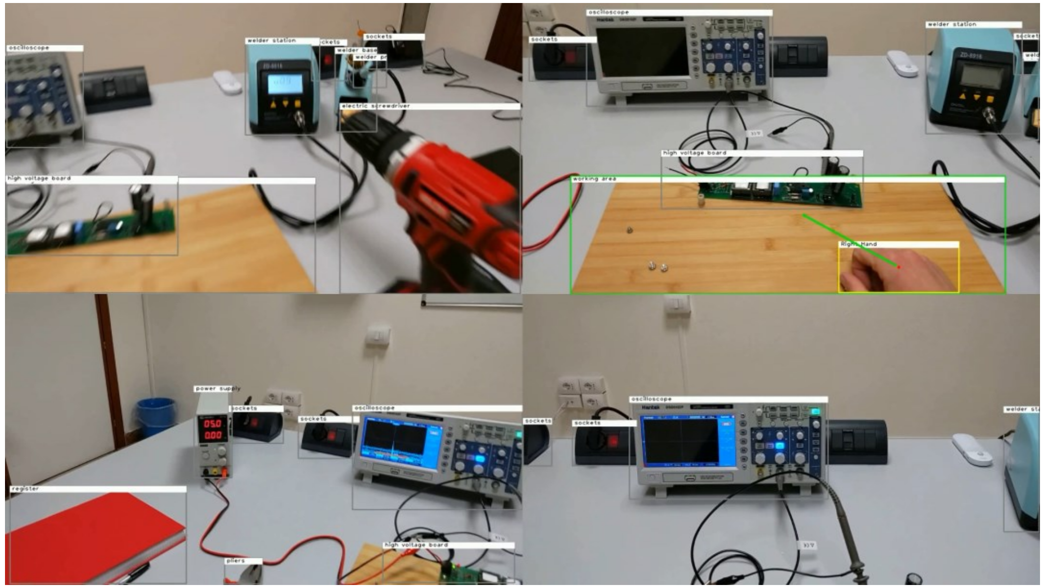
User-Object Interaction Synthetic Data Generation

APPROACH



RESULTS

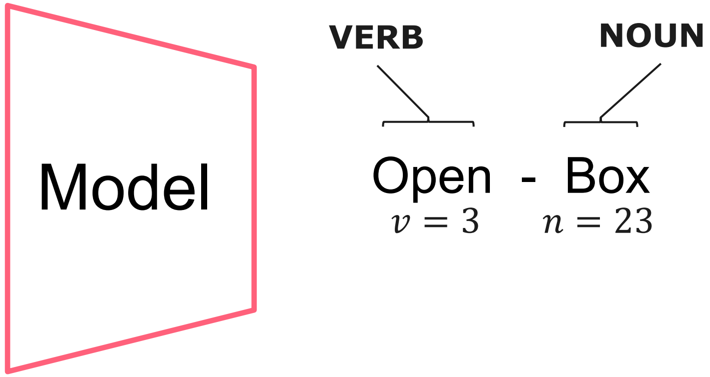


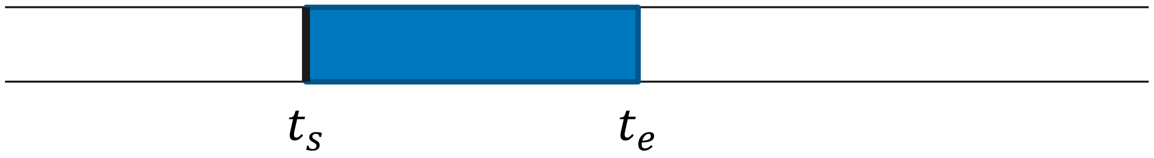


Rosario Leonardi, Francesco Ragusa, Antonino Furnari, Giovanni Maria Farinella (2022). Egocentric Human-Object Interaction Detection Exploiting Synthetic Data . In International Conference on Image Analysis and Processing (ICIAP)

Action Recognition

Action Recognition: Task





"observe a trimmed segment denoted by start and end time and classify the action present in the clip"

As defined in EPIC-KITCHENS-2020

Relation between Action and Interaction

TAKE SCREWDRIVER

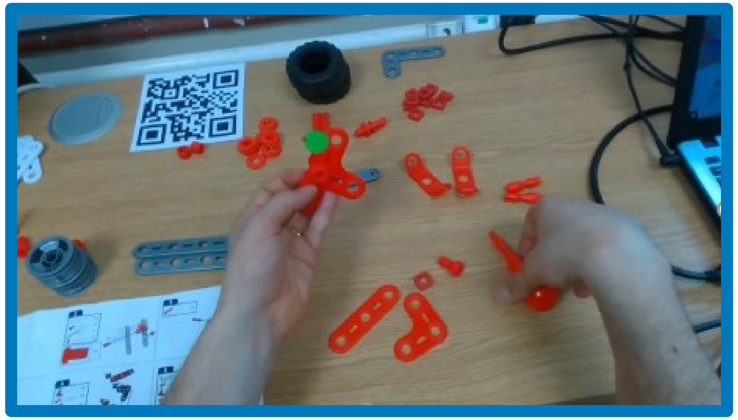
F. Ragusa, A. Furnari, G. M. Farinella. MECCANO: A Multimodal Egocentric Dataset for Humans Behavior Understanding in the Industrial-like Domain. Computer Vision and Image Understanding (CVIU), 2023.

Relation between Action and Interaction

TAKE SCREWDRIVER

Start Action

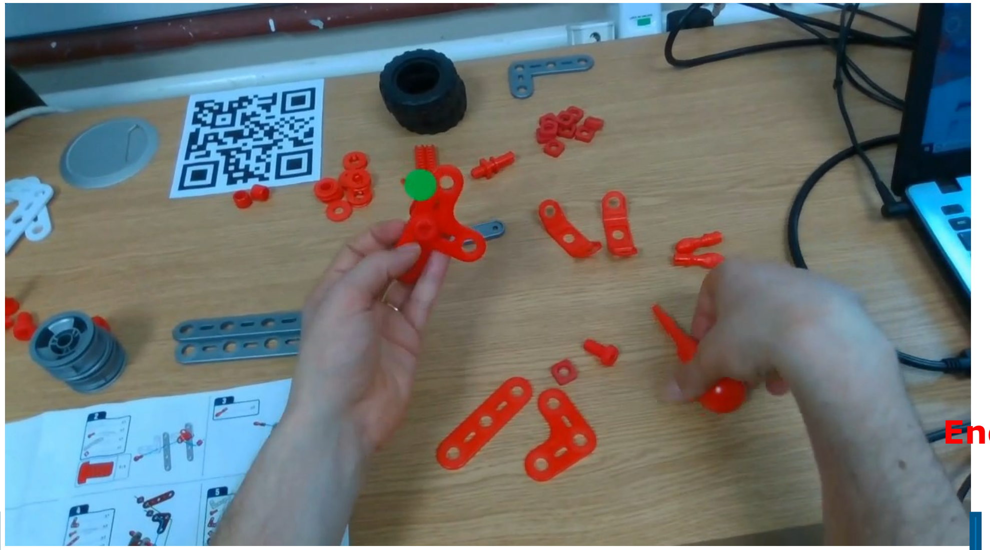
Start Interaction (H-O)



Frame of Contact

Relation between Action and Interaction

TAKE SCREWDRIVER

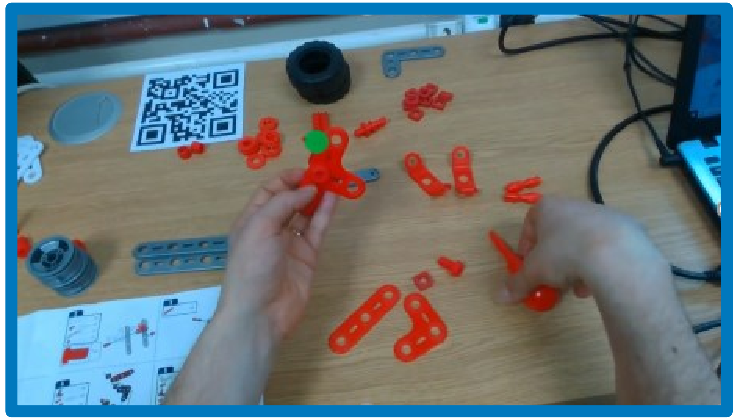


End Interaction

Start Action

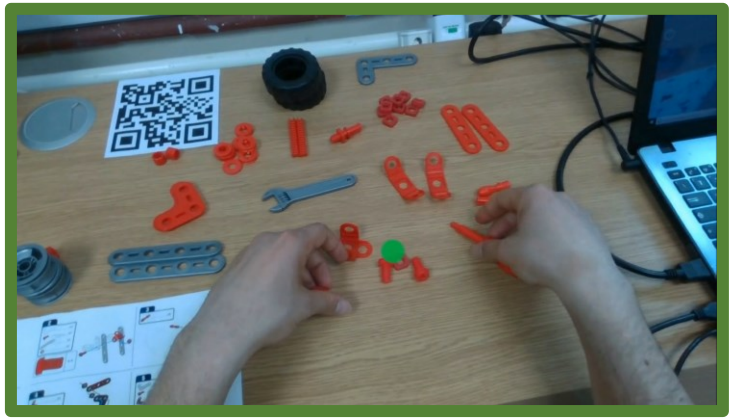
Start Interaction (H-O)

End Action



Frame of Contact

Frame of Decontact



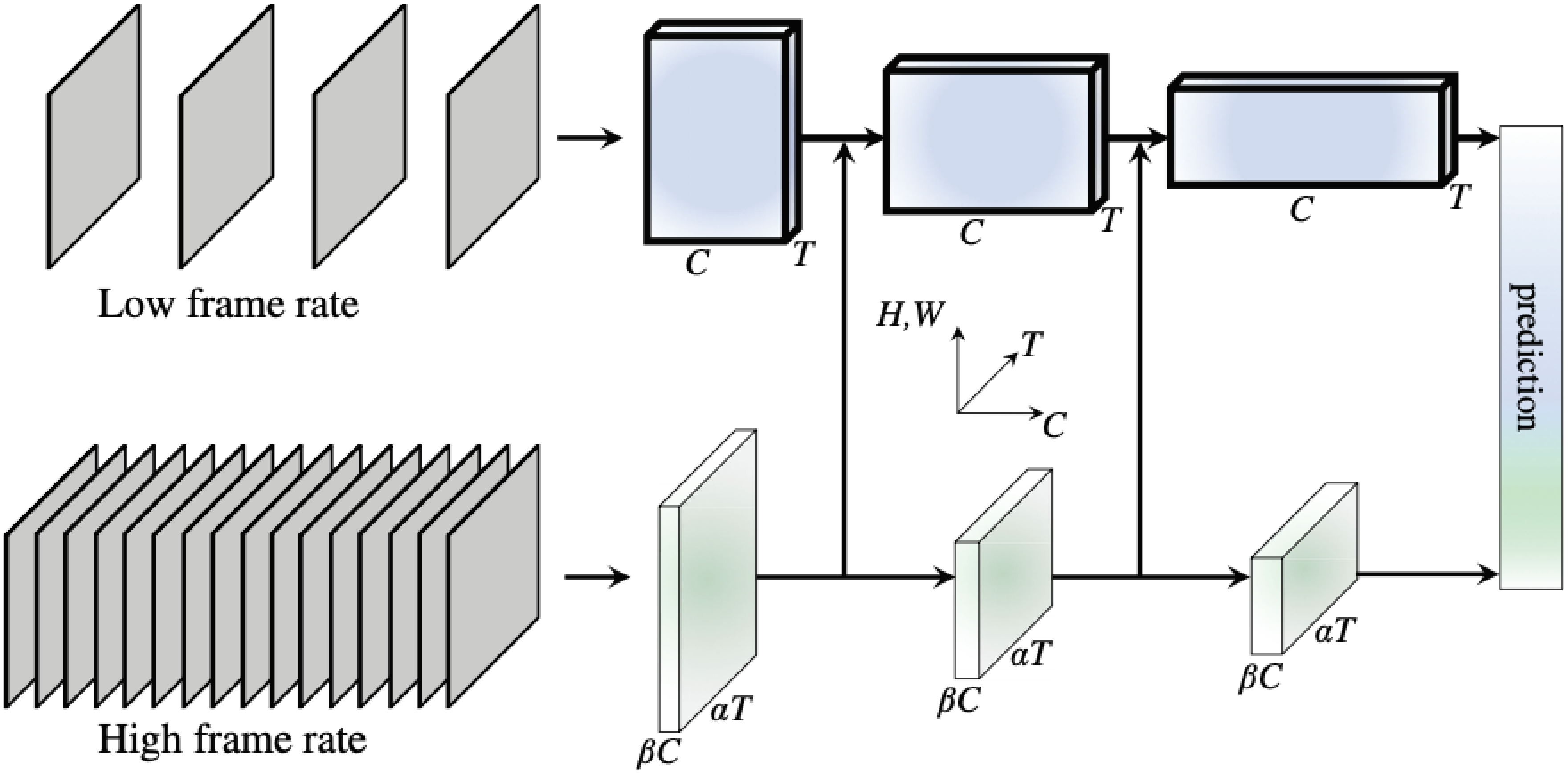
Relation between Action and Interaction

F. Ragusa, A. Furnari, G. M. Farinella. MECCANO: A Multimodal Egocentric Dataset for Humans Behavior Understanding in the Industrial-like Domain. Computer Vision and Image Understanding (CVIU), 2023.

Relation	Verbs	MECCANO verbs
A_s, I_s, I_e, A_e	pat, hit, kick	//
A_s, I_s, A_e, I_e	pick up	take, fit, align, plug, pull
A_s, I_s, A_e, I_e	close, open, turn on, press, push	browse
A_s, A_e	walk, jump, run	//
I_s, A_s, A_e, I_e	wring out, wash, cut, mix	pull
I_s, A_s, I_e, A_e	throw, leave, place	put
I_s, A_s, I_e, A_e	move	browse
I_s, A_s, A_e, I_e	twist, rip	screw, unscrew, tighten, loosen
I_s,A_s, I_e, A_e	stretch, knead, write, watch	check

SlowFast Networks for Video Recognition

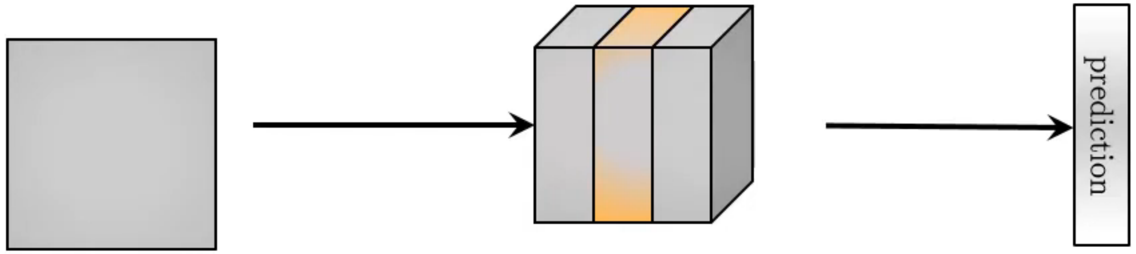
CODE HERE -> https://github.com/facebookresearch/SlowFast



Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6202-6211).

X3D: Expanding Architectures for Efficient Video Recognition

CODE HERE -> https://github.com/facebookresearch/SlowFast



- X-Fast
- X-Temporal
- X-Spatial
- X-Depth
- X-Width
- X-Bottleneck

Feichtenhofer, C. (2020). X3D: Expanding Architectures for Efficient Video Recognition. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 200-210.

PySlowFast

README.md

PySlowFast

PySlowFast is an open source video understanding codebase from FAIR that provides state-of-the-art video classification models with efficient training. This repository includes implementations of the following methods:

- SlowFast Networks for Video Recognition
- Non-local Neural Networks
- A Multigrid Method for Efficiently Training Video Models
- X3D: Progressive Network Expansion for Efficient Video Recognition
- Multiscale Vision Transformers
- A Large-Scale Study on Unsupervised Spatiotemporal Representation Learning
- MViTv2: Improved Multiscale Vision Transformers for Classification and Detection
- Masked Feature Prediction for Self-Supervised Visual Pre-Training
- Masked Autoencoders As Spatiotemporal Learners
- Reversible Vision Transformers

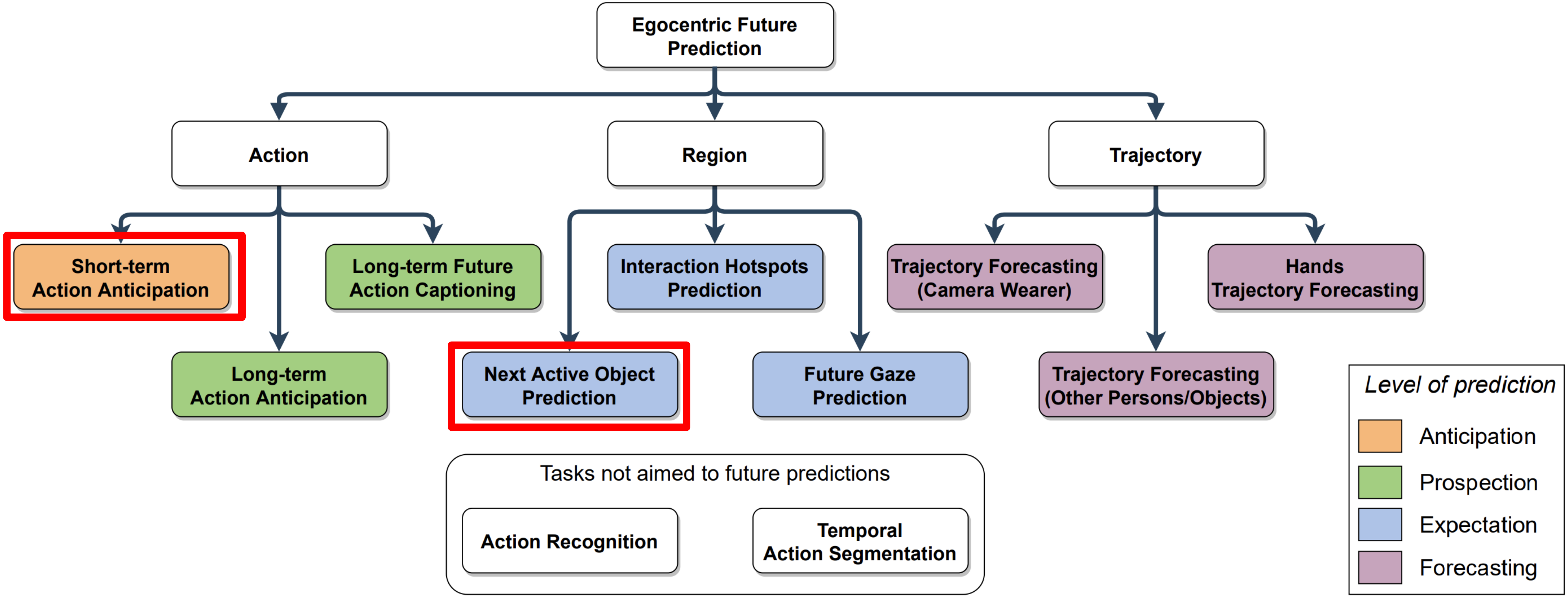
https://github.com/facebookresearch/SlowFast

Anticipation

Personal assistants and Future Predictions

Intelligent assistants should be able to understand what are the user’s goals and what is going to happen in the future.

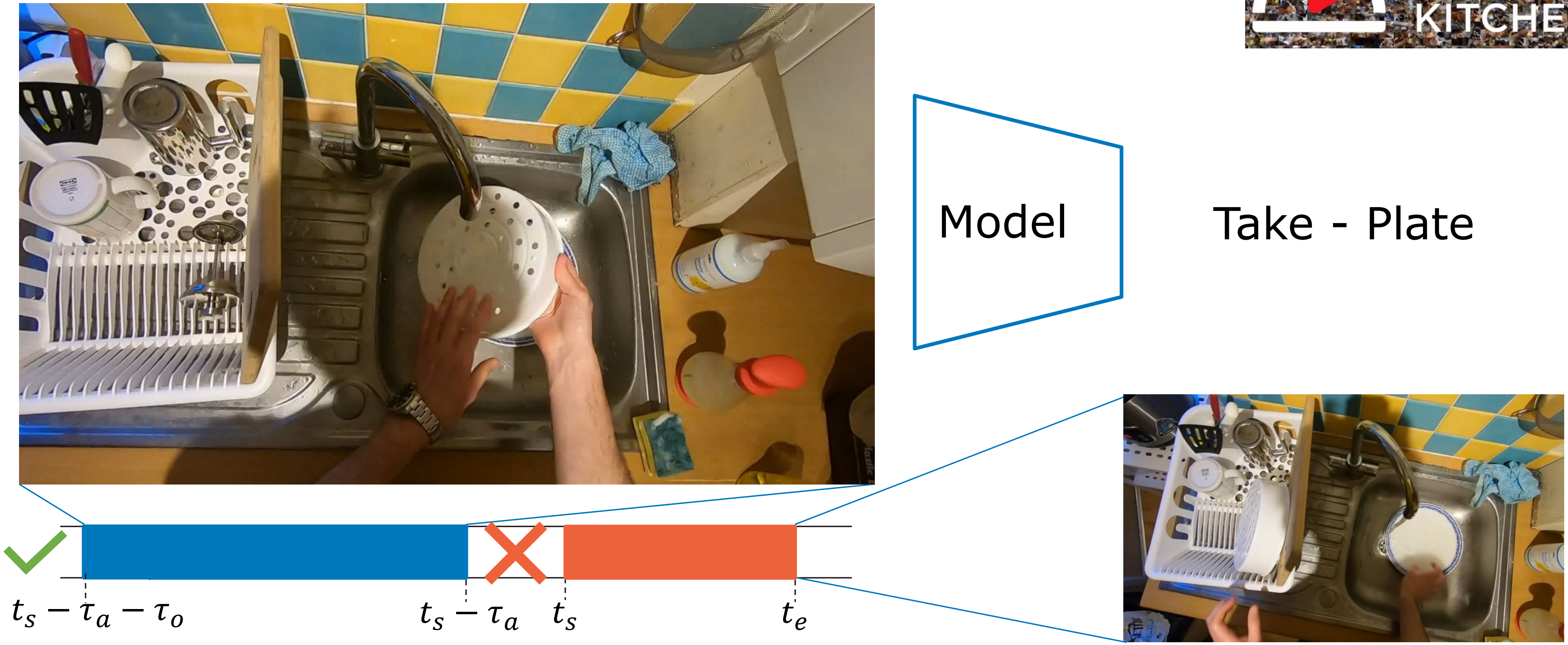
Future Predictions in Egocentric Vision



Ivan Rodin, Antonino Furnari, Dimitrios Mavroedis, Giovanni Maria Farinella (2021). Predicting the Future from First Person (Egocentric) Vision: A Survey. Computer Vision and Image Understanding, 211, pp. 103252.

Egocentric Action Anticipation Task

(observed video)



τ_o arbitrary

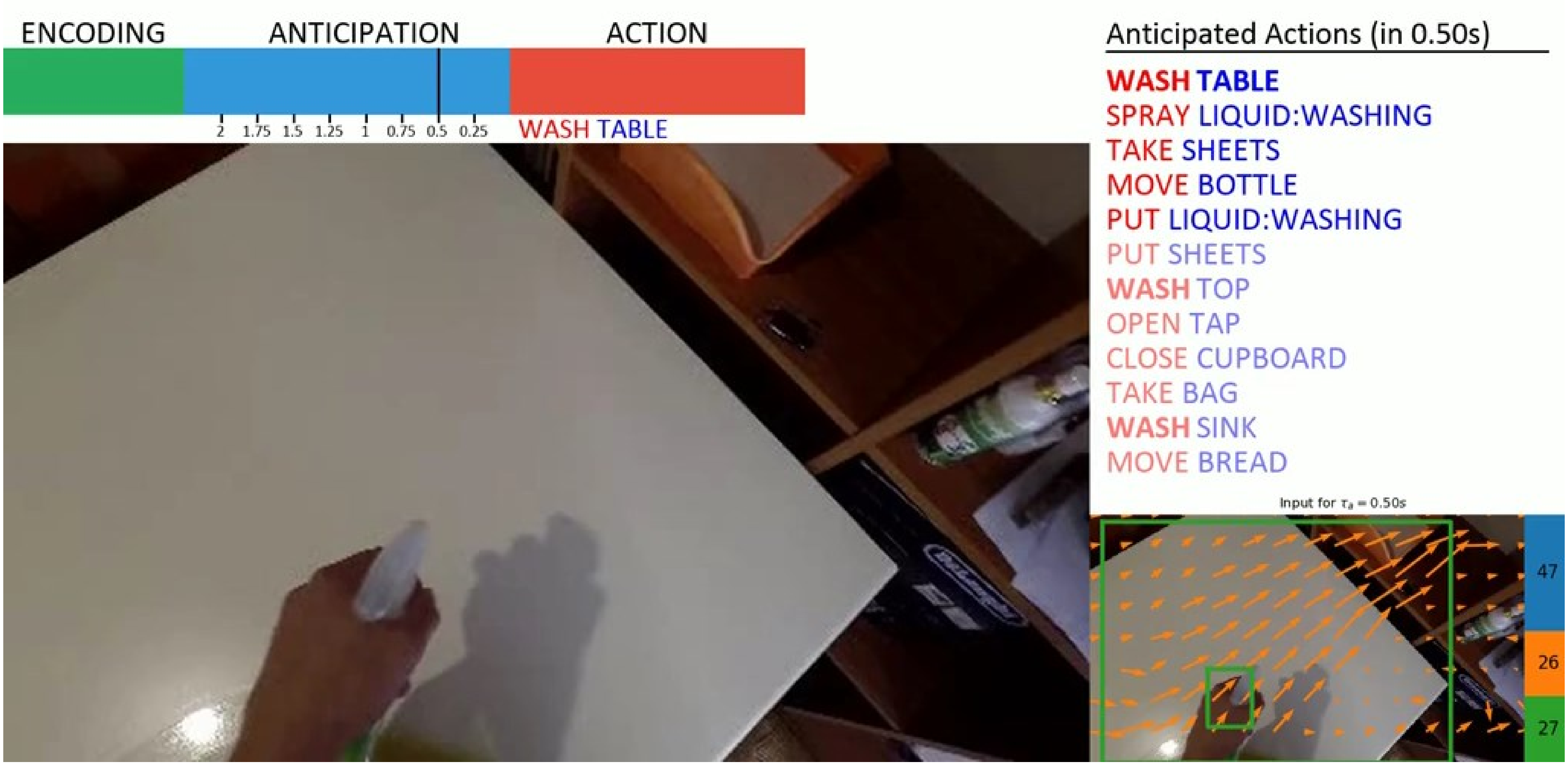
$\tau_a = 1s;$

(unobserved)

Damen, Dima, et al. "Scaling egocentric vision: The epic-kitchens dataset." *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

Dima Damen et al. Rescaling Egocentric Vision . International Journal on Computer Vision (IJCV). 2021

Demo Video: Egocentric Action Anticipation



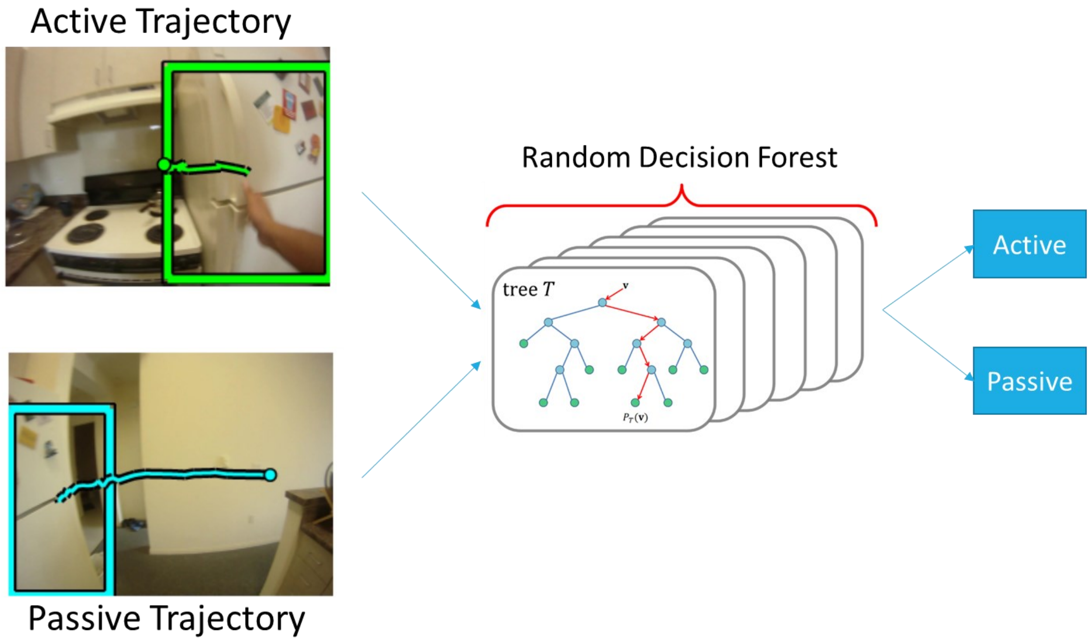
A. Furnari, G. M. Farinella, What Would You Expect? Anticipating Egocentric Actions with Rolling-Unrolling LSTMs and Modality Attention. ICCV 2019 (ORAL).

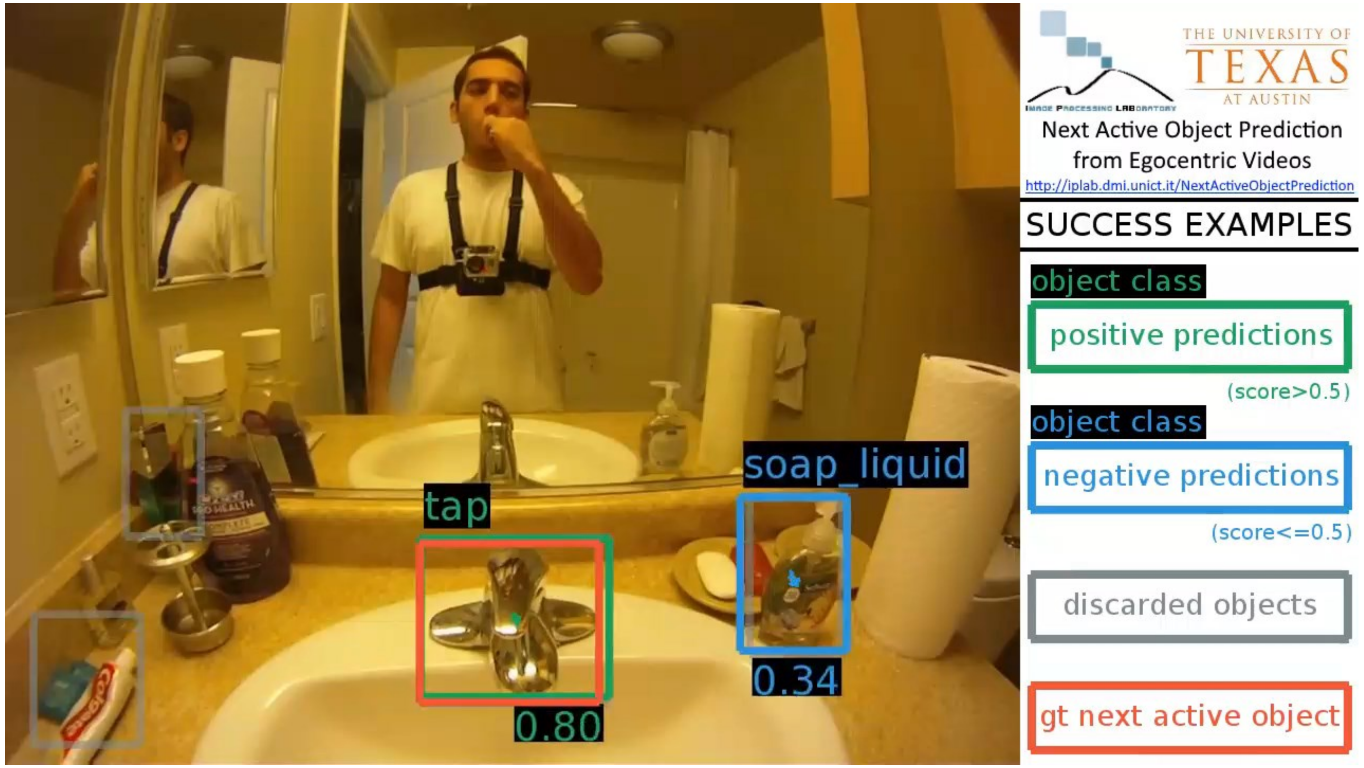
A. Furnari, G. M. Farinella. Rolling-Unrolling LSTMs for Action Anticipation from First-Person Video. TPAMI 2020. http://iplab.dmi.unict.it/rulstm

Anticipation – Next-Active-Objects

http://iplab.dmi.unict.it/NextActiveObjectPrediction/

Use egocentric object trajectories to distinguish passive from next-active-objects (i.e., those which will be used soon by the user).



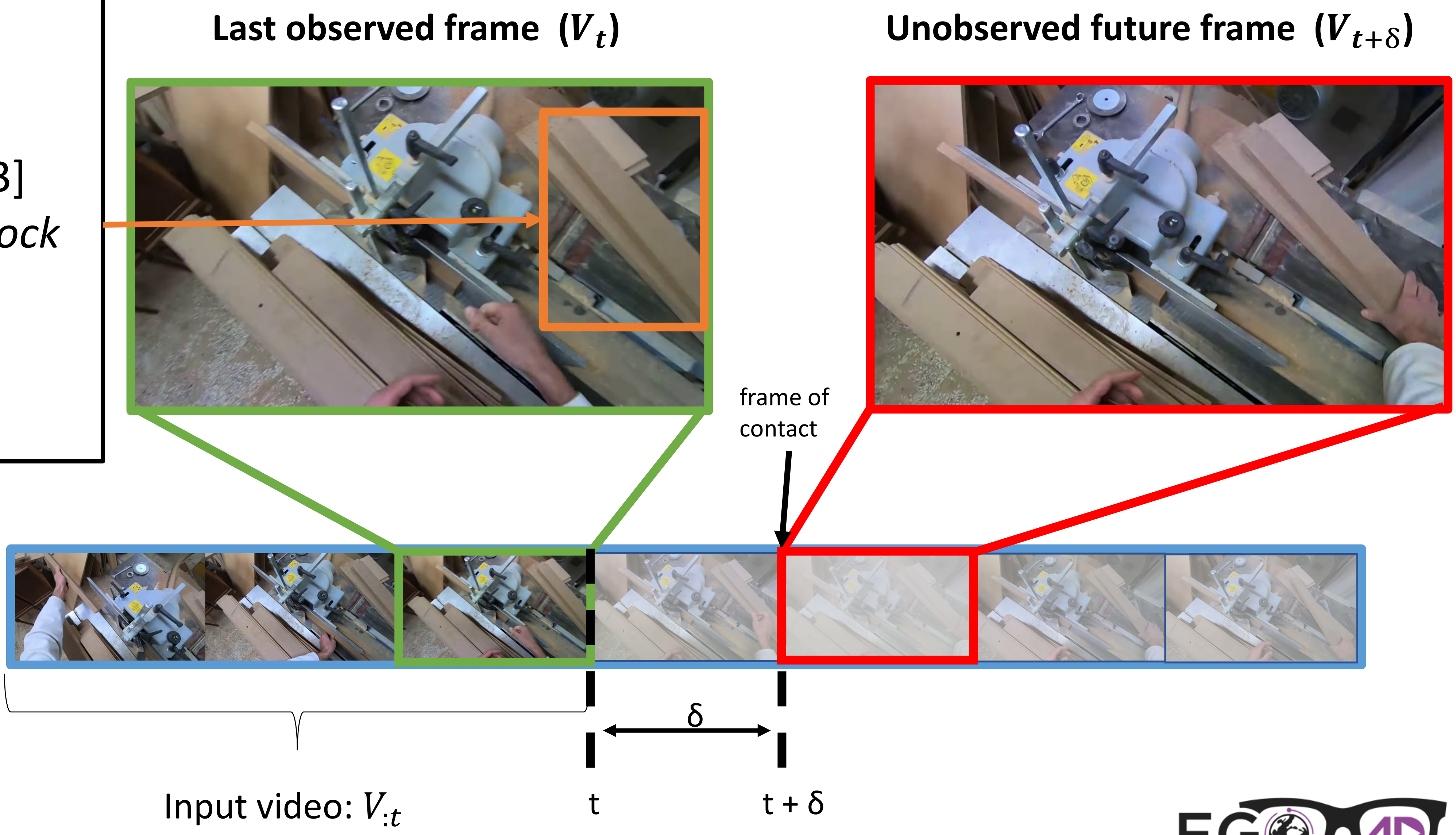


A. Furnari, S. Battiato, K. Grauman, G. M. Farinella, Next-Active-Object Prediction from Egocentric Videos, Journal of Visual Communication and Image Representation, 2017

Short-Term Object Interaction Anticipation

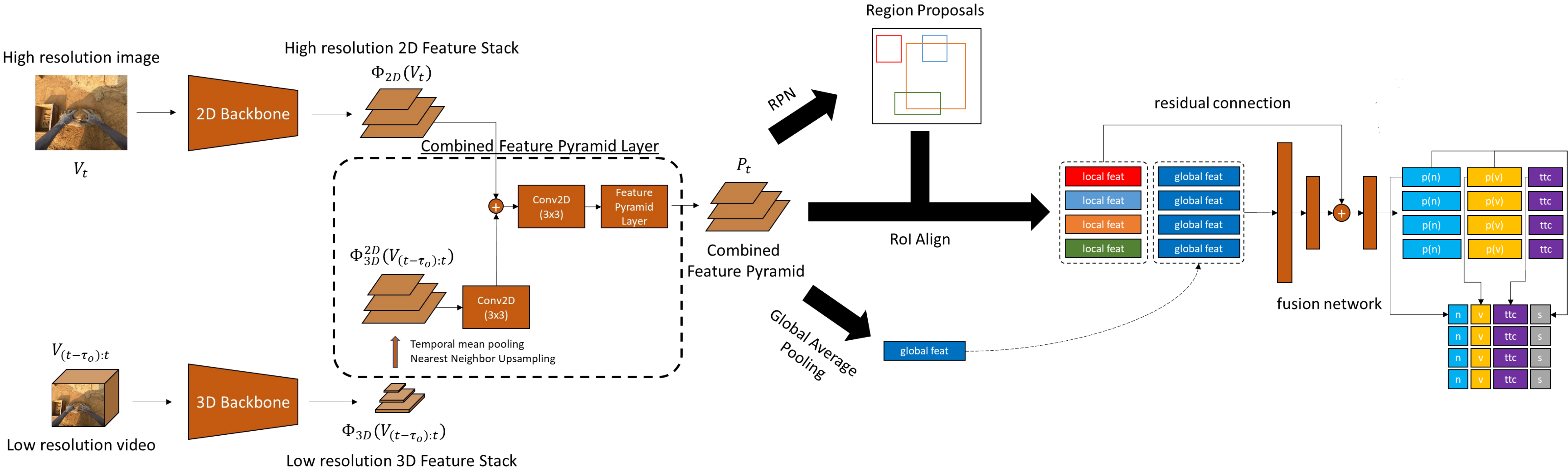
prediction

bbox = [1391,101,531,713]
noun = *wooden block*
verb = *take*
ttc = 0.75s
score = 0.83



StillFast Networks

An end-to-end approach for predicting next-active-objects based on an 2D-3D backbone taking as input a high resolution image and a video clip.



Francesco Ragusa, Giovanni Maria Farinella, Antonino Furnari (2023). StillFast: An End-to-End Approach for Short-Term Object Interaction Anticipation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.

Università di Catania

StillFast Qualitative Results

Can we bring egocentric vision to industry?

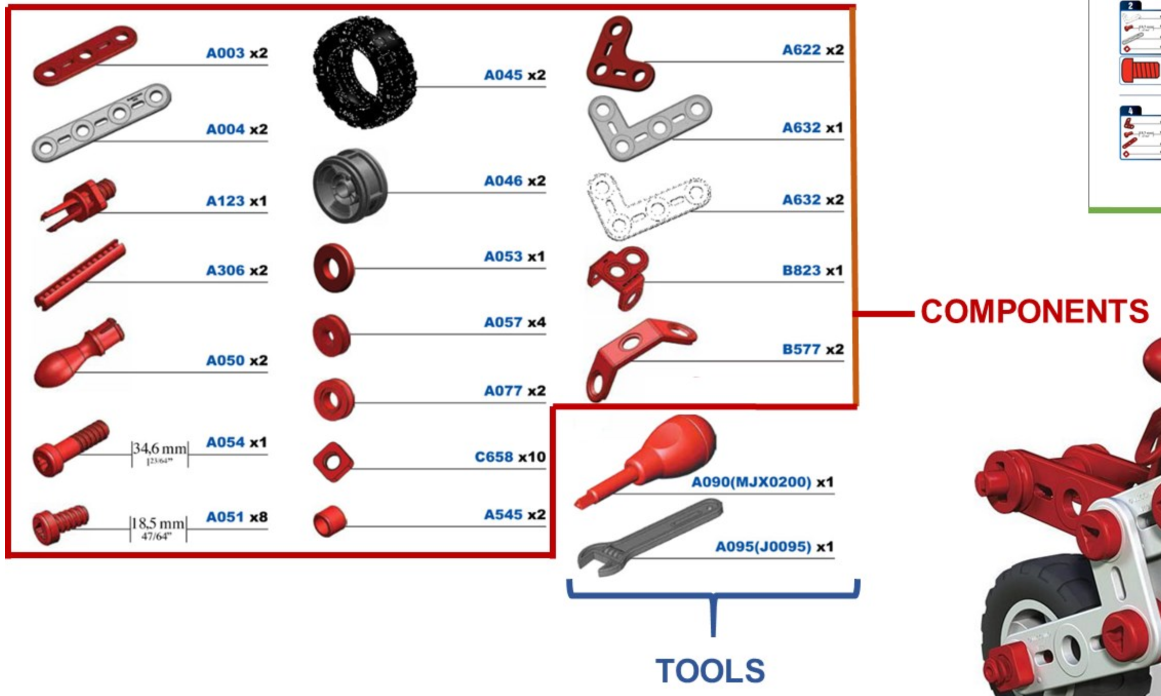
- The factory is a natural place for a wearable assistant;
- Closed-world assumption;
- Current research has considered different scenarios;
- No datasets in industrial-like scenarios;

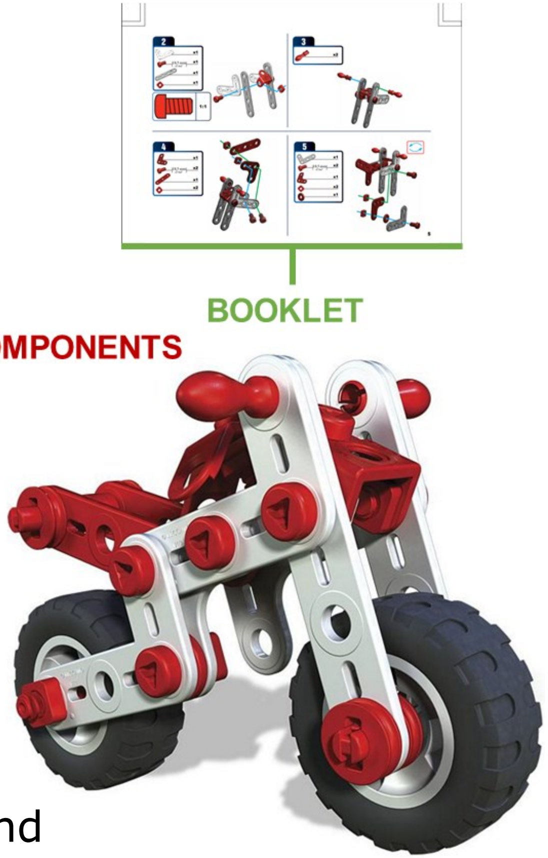
The MECCANO Dataset

Data HERE -> https://iplab.dmi.unict.it/MECCANO/

We asked subjects to record egocentric videos while assembling a toy motorbike.

The assembly required to interact with several parts and two tools.





The scenario is industrial-like, with subjects undertaking interactions with tiny objects and tools in a sequential fashion to reach a goal.

Francesco Ragusa, Antonino Furnari, Salvatore Livatino, Giovanni Maria Farinella (2021). The MECCANO Dataset: Understanding Human-Object Interactions from Egocentric Videos in an Industrial-like Domain. In IEEE Winter Conference on Application of Computer Vision (WACV).

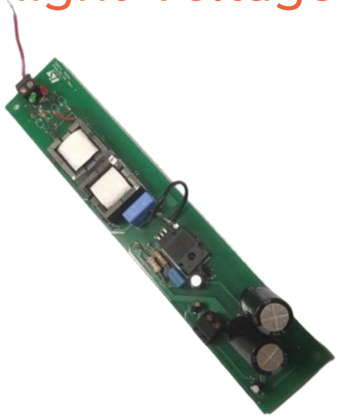
Università di Catania

The ENIGMA-51 Dataset

We designed two procedures consisting of instructions that involve humans interacting with the objects present in the laboratory to achieve the goal of repairing two electrical boards

Low-Voltage

Hight-Voltage



ENIGMA-51: Towards a Fine-Grained Understanding of Human Behavior in Industrial Scenarios. F. Ragusa R. Leonardi, M. Mazzamuto, C. Bonanno, R. Scavo, A. Furnari, G. M. Farinella. WACV (2024).

Industrial Applications

NEXT VISION

Spin-off of the University of Catania

SICILIAE STVDIVM GENERALE 1434

IMAGE PROCESSING LABORATORY

Intelligent Navigation

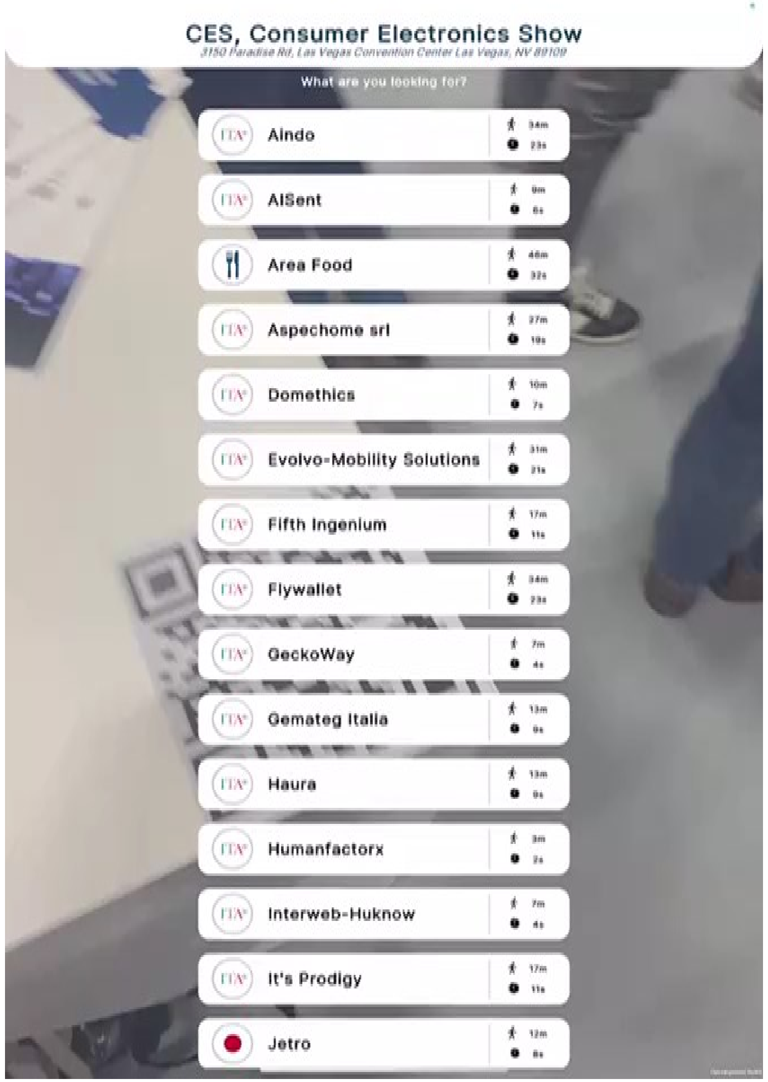
Image-based Localization

Augmented Reality

Multi-platform

Founders of Next Vision are authors of patents related to the developed technologies

https://drive.google.com/file/d/1lle4yF6b1kLp9P3ywqKOi77koTvn5OuE/view?usp=share_link



https://drive.google.com/file/d/1FAkLceBz_wCkDCsAJFq-nYBwFPZVciQV/view?usp=drive_link

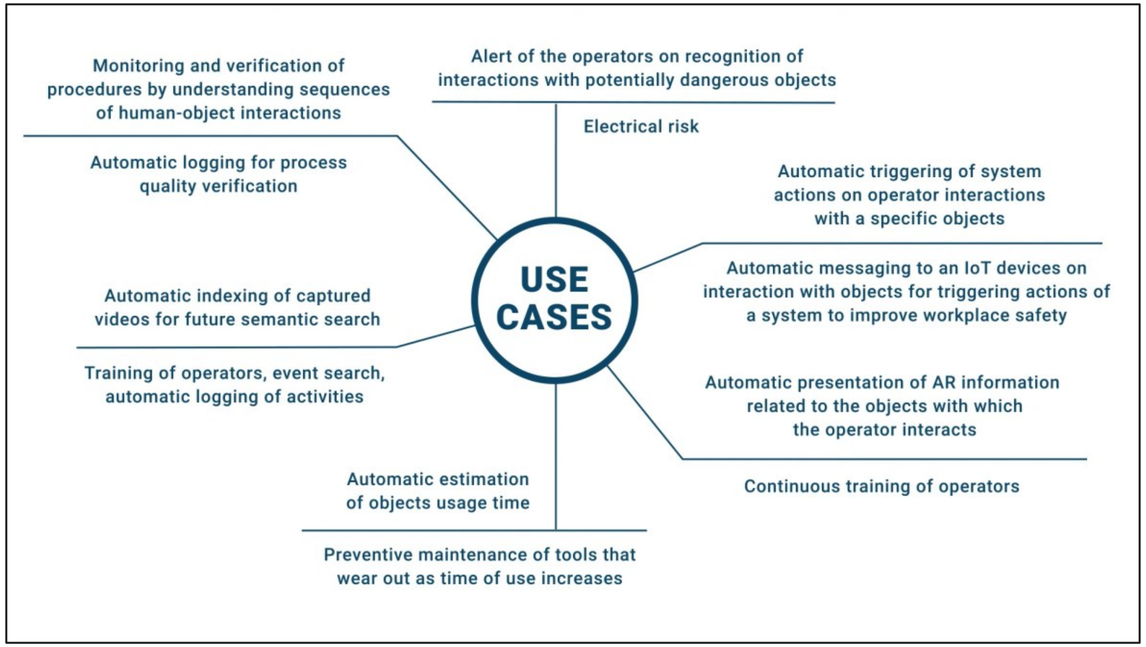
NOKIA

Università di Catania

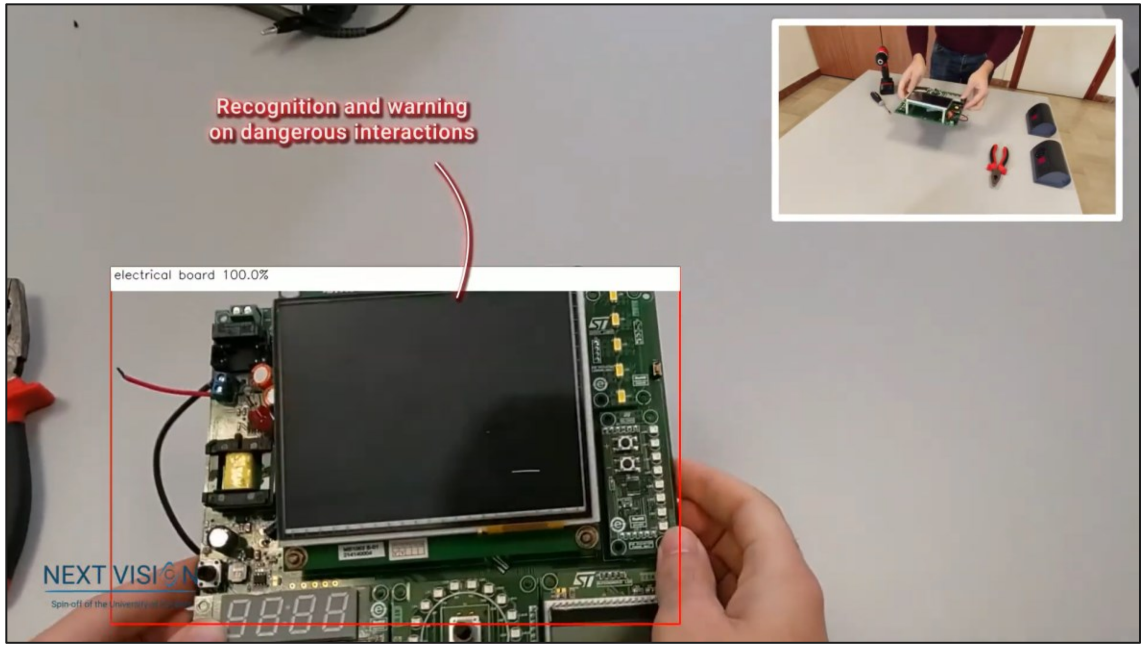
CES

UNIONE EUROPEA · REPUBBLICA ITALIANA · REGIONE SICILIANA · PO FESR SICILIA 2014-2020

- **NAOMI** is an AI Assistant able to support humans to monitor interactions, predict/anticipate next interactions, verify correctness in a sequence of interactions.



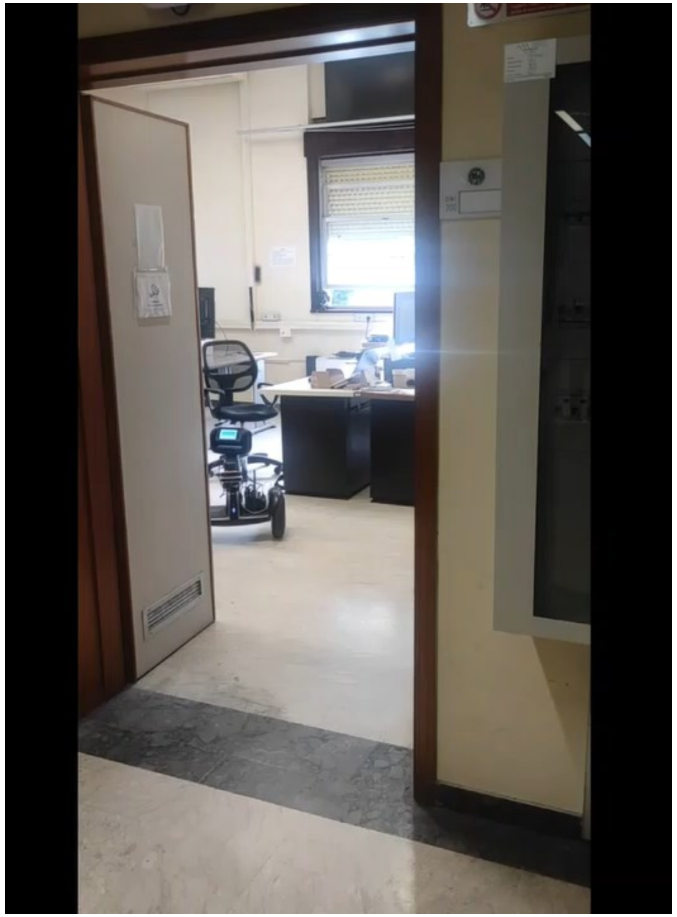
Use cases

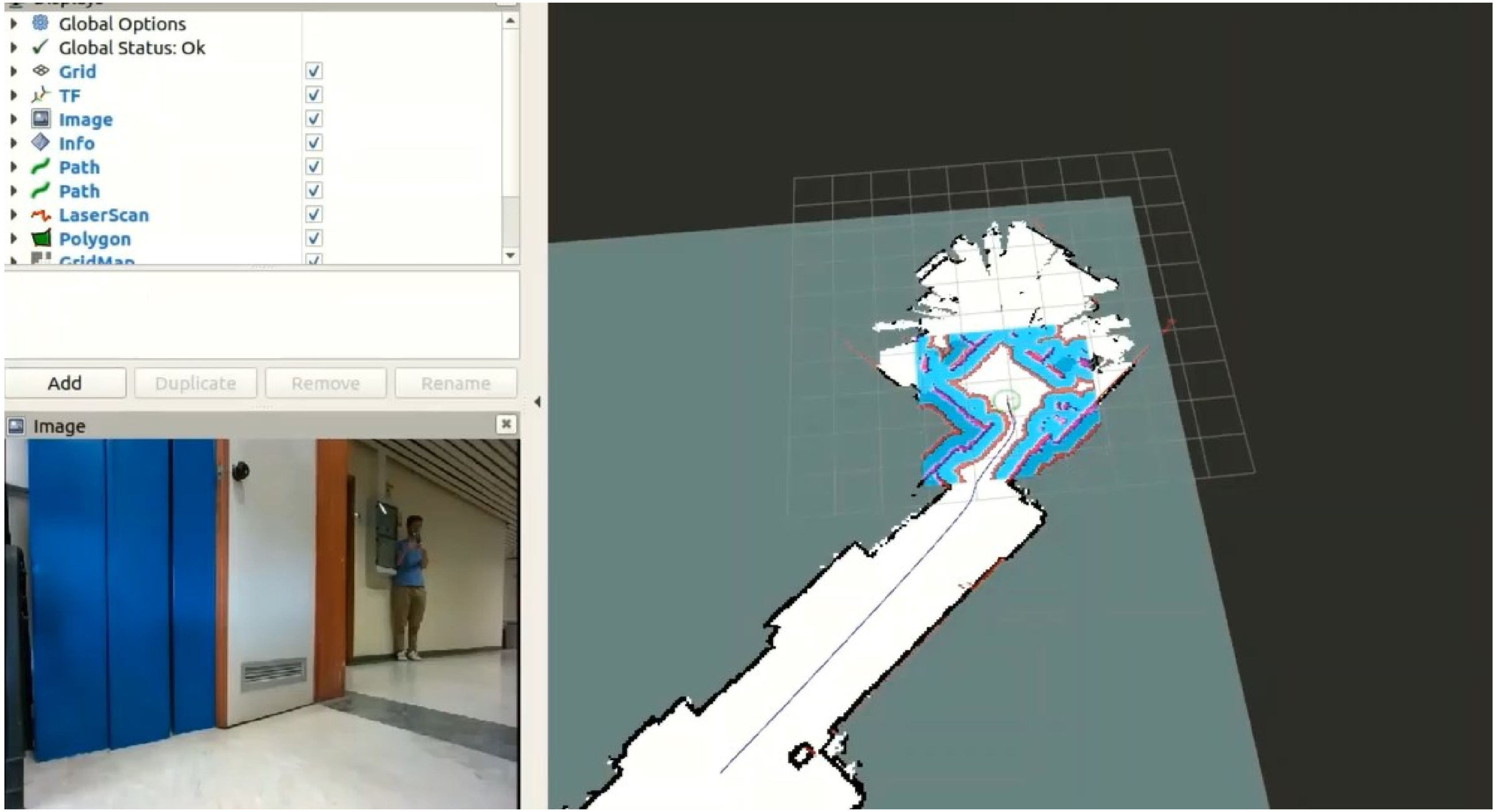


The video shows an example of object interaction monitoring.
The operator is notified on an interaction with a dangerous object.

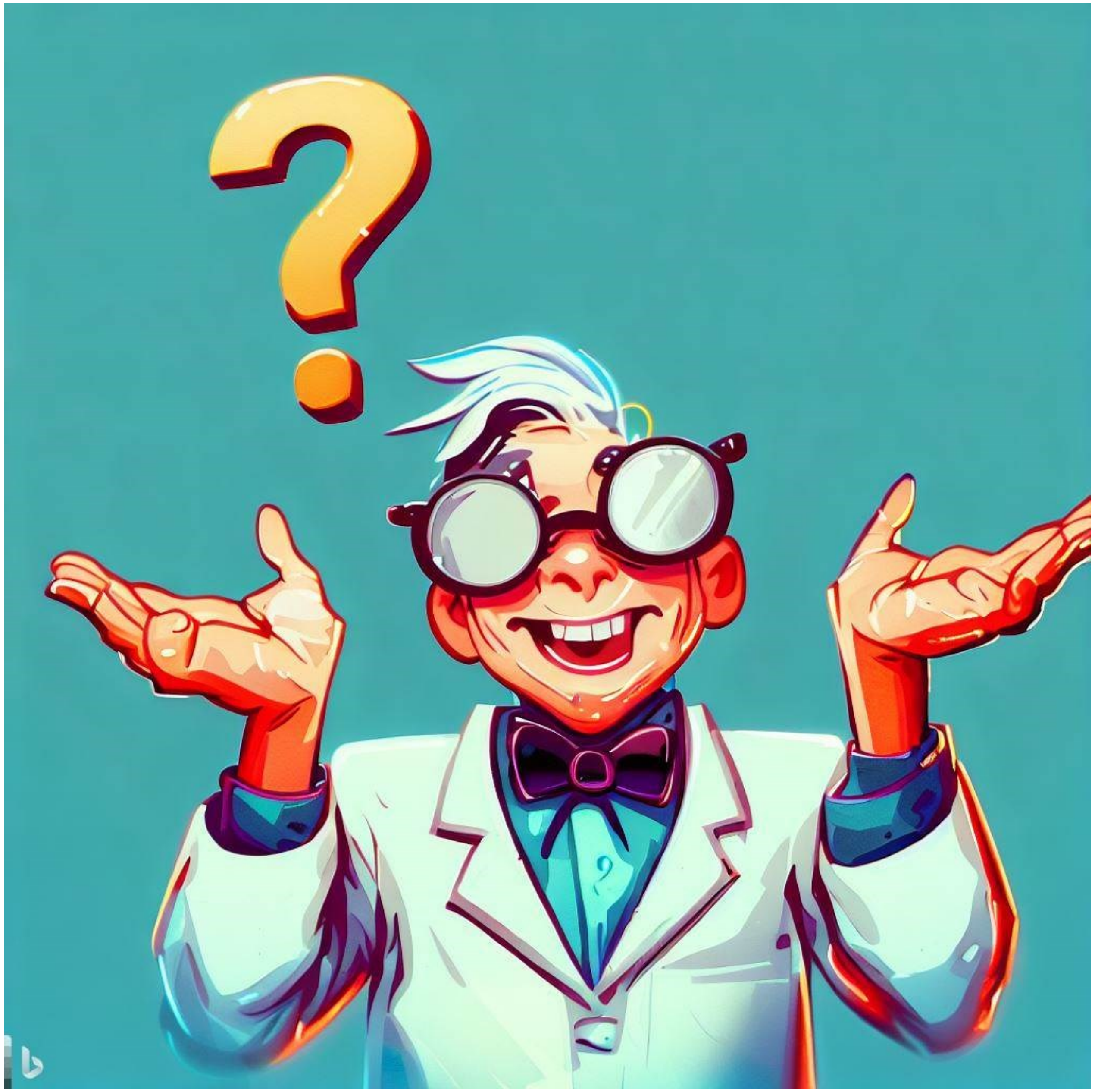
https://drive.google.com/file/d/1oOvhVbyyR7AZ35I-V90Zy7RyRTR7lkD4/view?usp=drive_link

MOVING





https://drive.google.com/file/d/17XrD-syy7pUm5MO4WYm7ZZxgRQsbSl2M/view?usp=drive_link



Doing Research in Egocentric Vision: Where to start?

Accessing Modern Datasets – License

Università di Catania

Data nowadays carries a lot of privacy/social/economic implications, so modern datasets are usually licensed.

! pay attention to which uses are permitted!

EPIC KITCHENS — ABOUT | STATS | DOWNLOADS | CHALLENGES | TEAM

Disclaimer

EPIC-KITCHENS-55 and EPIC-KITCHENS-100 were collected as a tool for research in computer vision. The dataset may have unintended biases (including those of a societal, gender or racial nature).

Copyright

All datasets and benchmarks on this page are copyright by us and published under the Creative Commons Attribution-NonCommercial 4.0 International License. This means that you must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use. You may not use the material for commercial purposes.

For commercial licenses of EPIC-KITCHENS and any of its annotations, email us at uob-epic-kitchens@bristol.ac.uk

EGO4D License Agreement

Obtaining the dataset or any annotations requires you first review our license agreement and accept the terms. Go here (ego4ddataset.com) to review and execute this agreement, and you will be emailed a set of AWS access credentials when your license agreement is approved, which will take ~48hrs. In the meantime, you can check out data overview & sample notebooks here to get familiar with the dataset, and can download the CLI & dataloaders to get setup in advance.

Note that licenses have the option to execute our license agreements as either an individual or on behalf of your institution. You will likely sign the license as an individual. Typically, only institutional signatories at a director or executive level can agree to license terms on behalf of an entire organization.

Also note that once approved your access credentials will expire in 14 days - you're expected to download the data locally, not to consume it from AWS. You can easily renew your license once it expires though: license renewal FAQ

Ego4D Dataset

This information you enter below will be used to generate a data usage agreement. You will receive an email from HelloSign which will step you through the process of signing all the agreements. You can review the data usage agreement at —

http://ego4d.github.io/pdfs/Ego4D-Licenses-Draft.pdf

Note: Only official signatories can sign as organisation

- Individual
- Organization

First name | Last name

Email

Home Address

City | State / Province / County | Country

Submit

Accessing Modern Datasets – Command Line Interfaces

Università di Catania

https://github.com/epic-kitchens/epic-kitchens-download-scripts

EPIC KITCHENS

Download only certain data types

We provide videos, RGB/optical flow frames, GoPro's metadata (for the extension only) and object detection frames (for EPIC KITCHENS-55's videos only). You can also download the consent form templates.

If you want to download only one (or a subset) of the above, you can do so with the following self-explanatory arguments:

- `--videos`
- `--rgb-frames`
- `--flow-frames`
- `--object-detection-images`
- `--masks`
- `--metadata`
- `--consent-forms`

If you want to download only videos, then:

```
python epic_downloader.py --videos
```

Note that these arguments can be **combined** to download multiple things. For example:

```
python epic_downloader.py --rgb-frames --flow-frames
```

Will download both RGB and optical flow frames.

Specifying participants

You can use the argument `--participants` if you want to download data for only a subset of the participants. Participants can be specified with their numerical *or* string ID.

You can specify a single participant, e.g. `--participants 1` or `--participants P01` for participant `P01`, or a comma-separated list of them, e.g. `--participants 1,2,3` or `--participants P01,P02,P03` for participants `P01`, `P02` and `P03`

This argument can also be combined with the aforementioned arguments. For example:

```
python epic_downloader.py --videos --participants 1,2,3
```

Will download only videos from `P01`, `P02` and `P03`.

Modern datasets are HUGE!

- ➢ EPIC-KITCHENS ~ 796 GB
- ➢ EGO4D ~ 30+ TB

https://github.com/facebookresearch/Ego4d/tree/main/ego4d/cli

Data download

Canonical videos and annotations can be downloaded using the following command:

```
python -m ego4d.cli.cli --output_directory="~/ego4d_data" --datasets full_scale annotations --benchmarks FHO
```

v2.0 annotations can be downloaded with:

```
python -m ego4d.cli.cli --output_directory="~/ego4d_data" --datasets annotations --version v2
```

Detailed Flags

Flag Name	Description
`--dataset`	[Required] A list of identifiers to download: [annotations, full_scale, clips] Each dataset will be stored in folders in the output directory with the name of the dataset (e.g. output_dir/v2/full_scale/) and manifest.
`--output_directory`	[Required]A local path where the downloaded files and metadata will be stored
`--metadata`	[Optional] Download the primary `ego4d.json` metadata at the top level (Default: True)
`--benchmarks`	[Optional] A list of benchmarks to filter dataset downloads by - e.g. Narrations/EM/FHO/AV
`-y` `--yes`	[Optional] If this flag is set, then the CLI will not show a prompt asking the user to confirm the download. This is so that the tool can be used as part of shell scripts.
`--aws_profile_name`	[Optional] Defaults to "default". Specifies the AWS profile name from ~/.aws/credentials to use for the download
`--video_uids`	[Optional] List of video or clip UIDs to be downloaded. If not specified, all relevant UIDs will be downloaded.
`--video_uid_file`	[Optional] Path to a whitespace delimited file that contains a list of UIDs. Mutually exclusive with the `video_uids` flag.
`--universities`	[Optional] List of university IDs. If specified, only UIDs from the S3 buckets belonging to the listed universities will be downloaded.
`--version`	[Optional] A version identifier - e.g. "v1" or "v2" (default)
`--no-metadata`	[Optional] Bypass the `ego4d.json` metadata download
`--config`	[Optional] Local path to a config JSON file. If specified, the flags will be read from this file instead of the command line

Datasets

The following datasets are available (not exhaustive):

Dataset	Description
annotations	The full set of annotations for the majority of benchmarks.
full_scale	The full scale version of all videos. (Provide `benchmarks` or `video_uids` filters to reduce the 5TB download size.)
clips	Clips available for benchmark training tasks. (Provide `benchmarks` or `video_uids` filters to reduce the download size.)
video_540ss	The downscaled version of all videos - rescaled to 540px on the short side. (Provide `benchmarks` or `video_uids` filters to reduce the 5TB download size.)
annotations_540ss	The annotations corresponding to the downscaled `video_540ss` videos - primarily differing only in spatial annotations (e.g. bounding boxes).
3d	Annotations for the 3D VQ benchmark.
3d_scans	3D location scans for the 3D VQ benchmark.
3d_scan_keypoints	3D location scan keypoints for the 3D VQ benchmark.
imu	IMU data for the subset of videos available
slowfast8x8_r101_k400	Precomputed action features for the Slowfast 8x8 (R101) model
omnivore_video_swinl	Precomputed action features for the Omnivore Video model
omnivore_image_swinl	Precomputed action features for the Omnivore Image model
fut_loc	Images and annotations for the future locomotion benchmark.
av_models	Model checkpoints for the AV/Social benchmark.
lta_models	Model checkpoints for the Long Term Anticipation benchmark.
moments_models	Model checkpoints for the Moments benchmark.
nlq_models	Model checkpoints for the NLQ benchmark.
sta_models	Model checkpoints for the Short Term Anticipation benchmark.
vq2d_models	Model checkpoints for the 2D VQ benchmark.

EGO4D

Command Line Interfaces Provided to Simplify Download

Lots of Tasks Out There – Challenges!

EPIC KITCHENS

ABOUT STATS DOWNLOADS CHALLENGES TEAM

EPIC-KITCHENS-100 2023 CHALLENGES

Challenge Details with links to ★NEW★ Codalab Leaderboards

New leaderboards are now open for the **challenge phase from Mon Jan 2023**. Check the results of the 2022 chalenge results below

In 2023, we have 9 open challenges. These are

- **New** Semi-Supervised Video Object Segmentation Challenge
- **New** Hand-Object Segmentation Challenge
- **New** TREK-150 Object Tracking Challenge
- **New** EPIC-SOUNDS Audio-Based Interaction Recognition
- Action Recognition
- Action Detection
- Action Anticipation
- UDA for Action Recognition
- Multi-Instance Retrieval

EPIC-Kitchens 2023 Challenges

Jan 23rd 2023,	All leaderboards are open (note new challenges for 2023)
June 1st 2023,	Server Submission Deadline at 23:00:00 UTC
June 6th 2023,	Deadline for Submission of Technical Reports on CMT
Mon June 19 2023,	Results announced at 11th EPIC@CVPR2023 workshop in Vancouver 11th EPIC@CVPR2023 workshop in Vancouver

Challenges Guidelines

The **nine** challenges below and their test sets and evaluation servers are available via CodaLab. The leaderboards will decide the winners for each individual challenge. For each challenge, the CodaLab server page details submission format and evaluation metrics.

This year, we offer **four** new challenges in: Semi-Supervised Video Object Segmentation using the VISOR annotations, Hand-object-segmentations using the VISOR annotations, single-object tracking and audio-based action recognition using the epic-sounds dataset.

https://epic-kitchens.github.io/2023#challenges

Ego4D Challenge 2023

Episodic memory:

- Visual queries with 2D localization (VQ2D) and Visual Queries 3D localization (VQ3D): Given an egocentric video clip and an image crop depicting the query object, return the most recent occurrence of the object in the input video, in terms of contiguous bounding boxes (2D + temporal localization) or the 3D displacement vector from the camera to the object in the environment.
  - Quickstart: Open in Colab
- Natural language queries (NLQ): Given a video clip and a query expressed in natural language, localize the temporal window within all the video history where the answer to the question is evident.
  - Quickstart: Open in Colab
- Moments queries (MQ): Given an egocentric video and an activity name (e.g., a "moment"), localize all instances of that activity in the past video
- EgoTracks: Given an egocentric video and a visual template of an object, localize the bounding box containing the object in each frame of the video along with a confidence score representing the presence of the object. **[NEW for 2023]**
- PACO Zero-Shot: Retrieve the bounding box of a specific object instance from a dataset, based on a textual query describing the instance. Query is composed using object and part attributes describing the object of interest. **[NEW for 2023]**

Hands and Objects:

- Temporal localization: Given an egocentric video clip, localize temporally the key frames that indicate an object state change.
- Object state change classification: Given an egocentric video clip, indicate the presence or absence of an object state change.

Audio-Visual Diarization:

- Audio-visual speaker diarization: Given an egocentric video clip, identify which person spoke and when they spoke.
- Speech transcription: Given an egocentric video clip, transcribe the speech of each person.

Social Understanding:

- Talking to me: Given an egocentric video clip, identify whether someone in the scene is talking to the camera wearer.
- Looking at me: Given an egocentric video clip, identify whether someone in the scene is looking at the camera wearer.

Forecasting:

- Short-term hand object prediction: Given a video clip, predict the next active objects, and, for each of them, predict the next action, and the time to contact.
  - Quickstart: Open in Colab
- Long-term activity prediction: Given a video clip, the goal is to predict what sequence of activities will happen in the future. For example, after kneading dough, list the actions that the baker will do next.

https://ego4d-data.org/docs/challenge/

Challenges – Train/Val/Test scheme

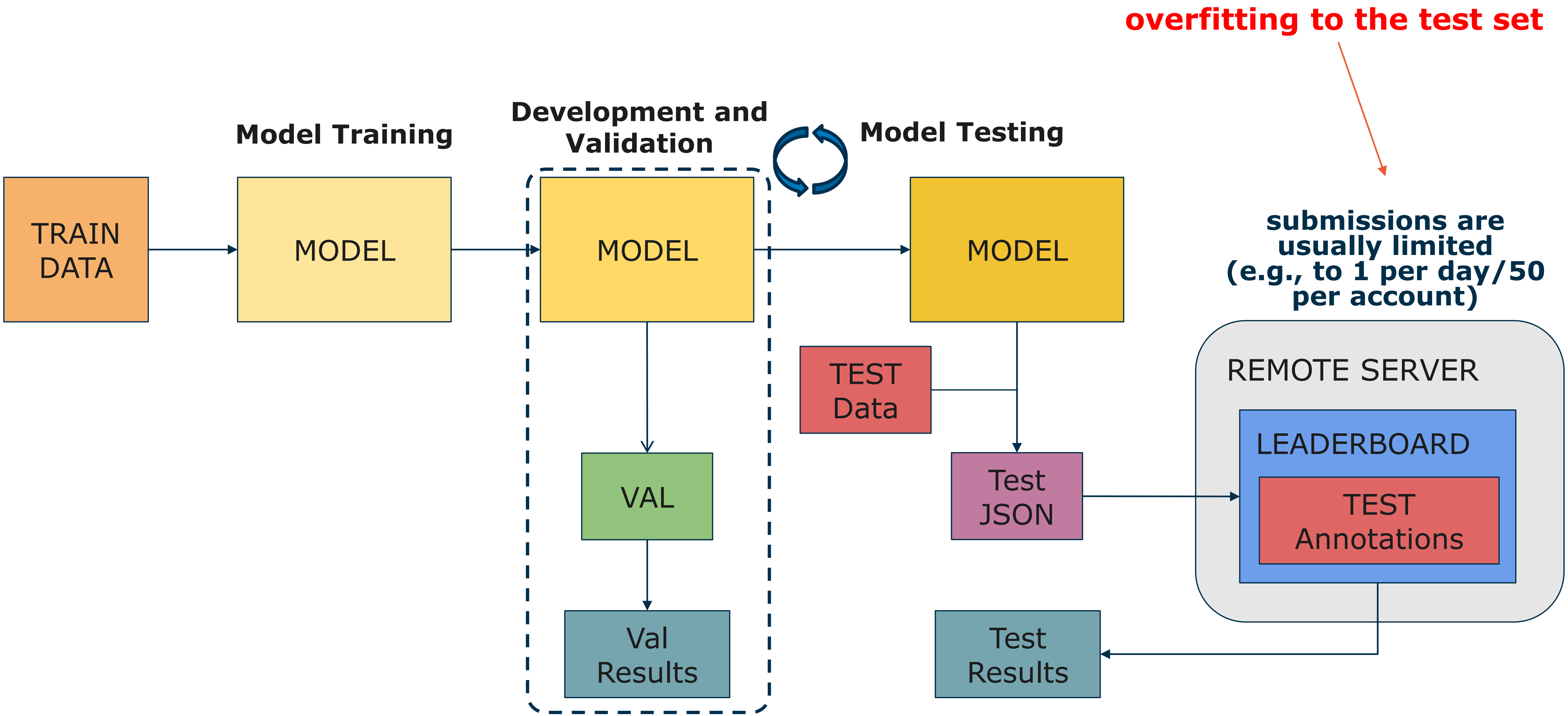
TRAIN

VAL

TEST

- Datasets are usually divided into train/val/test splits;
- All videos are publicly released;
- Train annotations are publicly released and meant for training models for the different challenges;
- Val annotations are publicly released and meant for model development and hyperparameter search;
- Test annotations are private and meant for assessing the performance of models avoiding bias in model design and optimization;
- Hence, the only way to obtain results on the test set is to send model predictions to an evaluation server.

Challenges – Evaluation Server



Università di Catania

Challenges – Evaluation Server

EPIC-KITCHENS-100 Action Anticipation

Organized by antonino - Current server time: Aug. 22, 2023, 9:44 a.m. UTC

Current	End
2023 Open Testing Phase	Competition Ends
June 27, 2023, 8 a.m. UTC	Nov. 25, 2023, 11 p.m. UTC

Test Set (Mean Top-5 Recall)

#	User	Entries	Date of Last Entry	Team Name	SLS			Overall (%)			Unseen (%)			Tail (%)		
					PT	TL	TD	Verb	Noun	Action	Verb	Noun	Action	Verb	Noun	Action
1	latent	29	10/18/22	InAViT IHPC-AISG-LAHA	1.0 (2)	3.0 (2)	3.0 (2)	49.14 (1)	49.97 (1)	23.75 (1)	44.36 (1)	49.28 (1)	23.49 (1)	43.17 (1)	39.91 (1)	18.11 (1)
2	hrgdscs	7	06/01/22		2.0 (1)	3.0 (2)	3.0 (2)	37.91 (4)	41.71 (2)	20.43 (2)	27.94 (4)	37.07 (2)	18.27 (2)	32.43 (4)	36.09 (2)	17.11 (2)
3	corcovadoming	28	06/01/22	NVIDIA-UNIBZ	1.0 (2)	3.0 (2)	4.0 (1)	29.67 (10)	38.46 (4)	19.61 (3)	23.47 (8)	35.25 (4)	16.41 (3)	23.48 (10)	31.11 (6)	16.63 (4)
4	shawn0822	22	06/01/22	ICL-SJTU	2.0 (1)	4.0 (1)	4.0 (1)	41.96 (3)	35.74 (5)	19.53 (4)	33.35 (3)	26.80 (13)	15.85 (5)	41.01 (3)	33.22 (4)	16.87 (3)
5	PCO-PSNRD	7	05/30/22	PCO-PSNRD	2.0 (1)	4.0 (1)	3.0 (2)	30.85 (6)	41.32 (3)	18.68 (5)	25.65 (6)	35.39 (3)	16.32 (4)	24.99 (6)	35.40 (3)	16.14 (5)
6	allenxuuu	1	12/20/21	2021 Open Testing Phase	2.0 (1)	4.0 (1)	4.0 (1)	29.88 (9)	30.40 (15)	17.35 (6)	25.08 (7)	26.08 (14)	14.14 (6)	24.60 (7)	23.68 (12)	14.30 (7)
7	Shawn0822-ICL-SJTU	1	12/20/21	2021 Open Testing Phase	1.0 (2)	4.0 (1)	3.0 (2)	42.32 (2)	34.60 (6)	17.02 (7)	33.36 (2)	25.94 (16)	12.84 (8)	42.47 (2)	31.37 (5)	15.56 (6)
8	shef-AVT-FB-UT	1	12/20/21	2021 Open Testing Phase	2.0 (1)	4.0 (1)	4.0 (1)	26.69 (13)	32.33 (10)	16.74 (8)	21.03 (12)	27.64 (7)	12.89 (7)	19.28 (13)	24.03 (10)	13.81 (8)
9	richard61	8	05/31/22		2.0 (1)	4.0 (1)	4.0 (1)	27.60 (11)	32.45 (9)	16.68 (9)	20.10 (14)	28.13 (5)	12.42 (11)	20.12 (12)	23.89 (11)	13.80 (10)
10	Zeyun-Zhong	12	06/01/22	KIT-IAR-IOSB	1.0 (2)	4.0 (1)	3.0 (2)	30.03 (8)	33.45 (8)	16.65 (10)	23.16 (9)	27.20 (8)	12.63 (10)	23.65 (9)	26.86 (9)	13.80 (9)
11	AVT-FB-UT	1	12/15/21	CVPR 2021 Challenges	2.0 (1)	4.0 (1)	4.0 (1)	25.25 (16)	32.04 (12)	16.53 (11)	20.41 (13)	27.90 (6)	12.79 (9)	17.63 (15)	23.47 (13)	13.62 (11)

https://codalab.lisn.upsaclay.fr/competitions/702

Ego4D Short Term Object Interaction Anticipation Challenge

★ 11

Toggle Participation

Organized by: Ego4D

Published

Starts on: Oct 25, 2022 2:00:00 AM CET (GMT + 2:00)

Ends on: May 20, 2024 2:00:59 AM CET (GMT + 2:00)

Discuss

Leaderboard

Overall Top-5 mAP

Phase: **Test Phase**, Split: **Test Split**

Order by metric

B - Baseline   * - Private   V - Verified   Include private submissions

Rank	Participant team	Noun (↑)	Noun_Verb (↑)	Noun_TTC (↑)	Overall (↑)	Last submission at	Meta Attributes
1	PAVIS (GANO_v2)	25.67	13.60	9.02	5.16	3 months ago	View
2	Host_47324_Team (V2 StilFast Baseline) B	25.06	13.29	9.14	5.12	5 months ago	View
3	Host_47324_Team (V2 Faster RCNN + SlowFast Base) B	26.15	9.45	8.69	3.61	5 months ago	View
4	FPV_UNICT (StillFast)	19.51	9.95	6.45	3.49	11 months ago	View
5	Red Panda (fusion-1)	24.60	9.19	7.64	3.40	11 months ago	View
6	Host_47324_Team (Faster RCNN + SlowFast Baselin) B	20.45	6.78	6.17	2.45	1 year ago	View

https://eval.ai/web/challenges/challenge-page/1623/leaderboard/3910

Conclusion

- First Person Vision paves the way to a variety of user-centric applications;
- However, we are still missing solid building blocks related to fundamental problems of First Person Vision such as action recognition, object detection, action anticipation and human-object interaction detection;
- Consumer devices are starting to appear, but the near future of First Person Vision is in focused applications such as the ones in industrial scenarios.

We Are Hiring!

francesco.ragusa@unict.it – fragusa@nextvisionab.it

THANK YOU!

First Person (Egocentric) Vision: History and Applications

Francesco Ragusa

First Person Vision@Image Processing Laboratory - http://iplab.dmi.unict.it/fpv

Next Vision - http://www.nextvisionlab.it/

Department of Mathematics and Computer Science - University of Catania

francesco.ragusa@unict.it - https://francescoragusa.github.io/