Egocentric Vision: Emerging Trends and Human-Centric Applications

Francesco Ragusa

LIVE Group @ UNICT - https://iplab.dmi.unict.it/live/

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

Department of Mathematics and Computer Science - University of Catania

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

Università di Catania

IMAGE PROCESSING LABORATORY

LIVE

LIVE Group @ UNICT

Università di Catania The LIVE Group @ UNICT

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

http://www.nextvisionlab.it/

19 Members
1 Full Professor
1 Assistant Professor
3 Post Docs
2 PhD Students
7 Master Students
1 Lab Assistant
4 Visiting PhD Students

Before we begin...

The slides of this tutorial are available online at:
https://francescoragusa.github.io/iciap2025

Agenda

1) Part I: History and motivations [14.30 - 15.30]
    a) Agenda of the tutorial;
    b) Perception and Egocentric Vision;
    c) Seminal works in 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;
    h) What’s next?
    i) Industrial Applications

Agenda

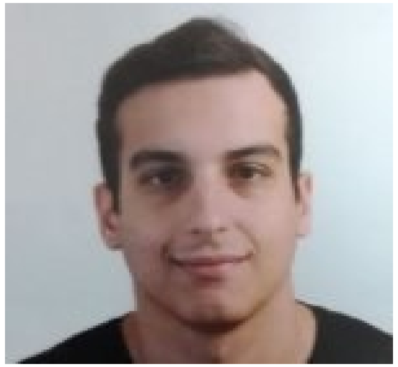
1) Part I: History and motivations [14.30 - 15.30]
   a) Agenda of the tutorial;
   b) Perception and Egocentric Vision;
   c) Seminal works in Egocentric Vision;
   d) Differences between Third Person and First Person Vision;
   e) First Person Vision datasets;
   f) Wearable devices to acquire/process first person visual data;
   g) Main research trends in First Person (Egocentric) Vision;
   h) What’s next?
   i) Industrial Applications

Coffee Break [15.30 – 15.50]

Coffee Breaks are organized autonomously by each workshop to best suit their schedule and format. Participants will receive coupons to enjoy coffee and snacks at the campus café

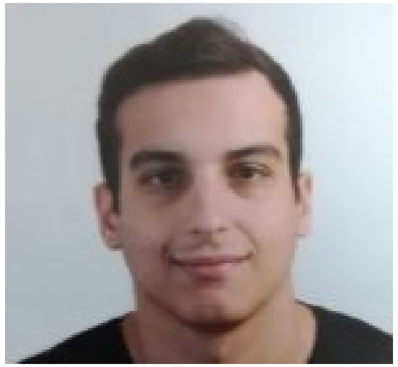
Agenda

2) Part II: Hand-Object Interactions in Egocentric Vision [15.50 – 16.50]
    a) Introduction to Hand-Object Interactions Detection
    b) Datasets and Benchmarks for Hand-Object Interactions in Egocentric Vision
    c) Models and Architectures for Hand-Object Interactions Detection
    d) Open Challenges



Agenda

2) Part II: Hand-Object Interactions in Egocentric Vision [15.50 – 16.50]
    a) Introduction to Hand-Object Interactions Detection
    b) Datasets and Benchmarks for Hand-Object Interactions in Egocentric Vision
    c) Models and Architectures for Hand-Object Interactions Detection
    d) Open Challenges



Short Break [16.50 – 17.00]

Agenda

3) Part III: Gaze Understanding and Visual-Language Benchmarks [17.00 – 18.00]
    a) Gaze Signal Fundamentals
    b) Gaze-Based Dataset
    c) Gaze signal in computer vision
    d) Building procedural assistant with VLLM
    e) Open Challenges and Future Directions

Part I

History and Motivations

Perception and Egocentric Vision

Perception

Perception is the process of **receiving**, **organize** and **interpret** information in order to give meaning to the surrounding world.

Perception is the process of **receiving**, **organize** and **interpret** information in order to give meaning to the surrounding world.

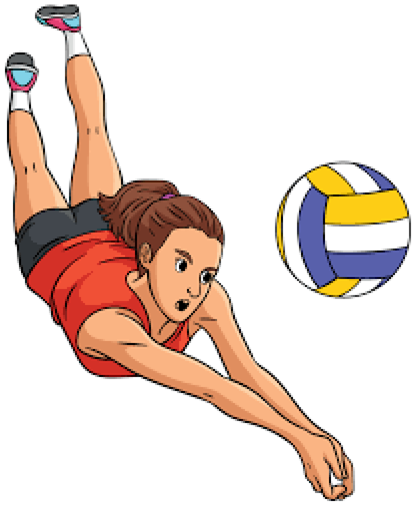
Perception is the process of **receiving**, **organize** and **interpret** information in order to give meaning to the surrounding world.

Perception is the process of **receiving**, **organize** and **interpret** information in order to give meaning to the surrounding world.

E. Corona, A. Pumarola, G. Aleny, M. N. Francesc, R. Gregory. GanHand: Predicting Human Grasp Affordances in Multi-Object Scenes, CVPR, 2020.

Perception

Perception is the process of **receiving**, **organize** and **interpret** information in order to give meaning to the surrounding world.

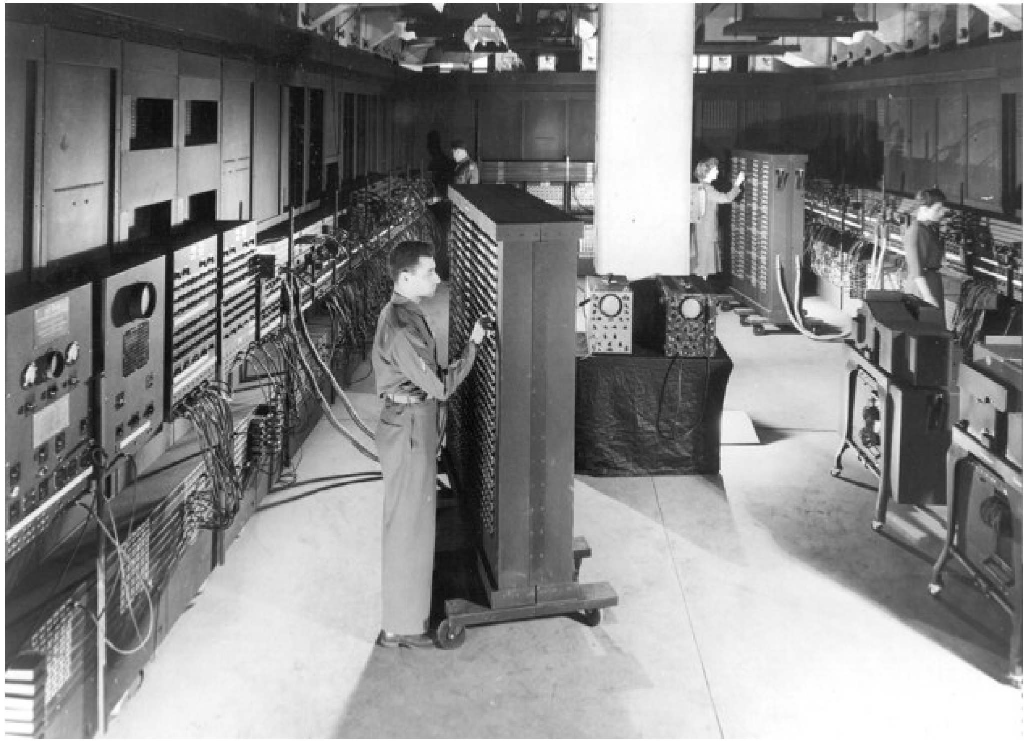


Computer vision enables computers to **acquire**, **process**, **analyze** and **understand** digital images, and extract of high-dimensional data from the real world in order to produce numerical or symbolic information

Computer vision enables computers to **acquire**, **process**, **analyze** and **understand** digital images, and extract of high-dimensional data from the real world in order to produce numerical or symbolic information, e.g. in the forms of decisions

Perception is the process of **receiving**, **organize** and **interpret** information in order to give meaning to the surrounding world.

The Revolution of Personal Computing



Mainframe Era (1950s–1970s)

Centralized, inaccessible, institutional

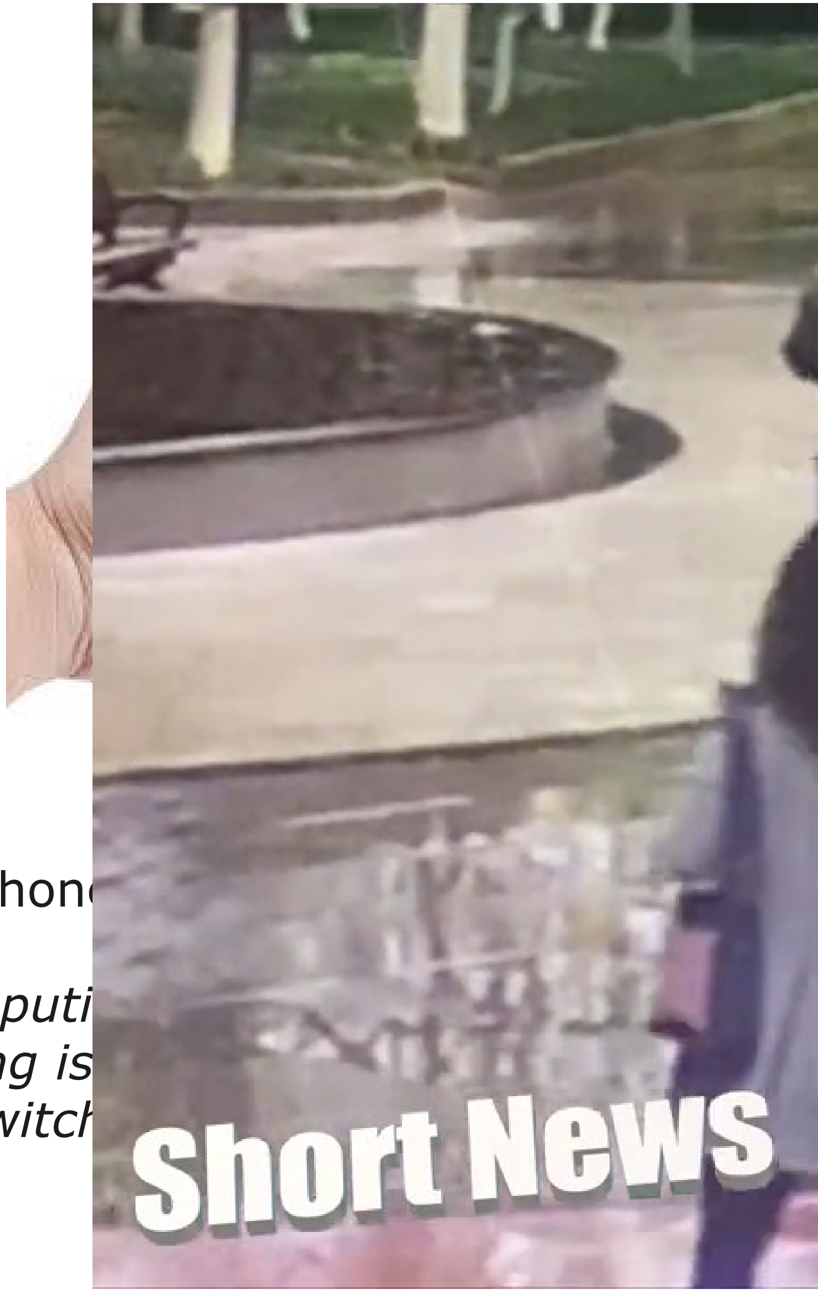
Personal Computer Era (1980s–1990s)

Desktop computing enters homes and offices

The Revolution of Personal Computing

Laptop Era (1990s–2000s)

Computing for the mass, but not mobile and not context aware - dedicated access to computing



Smartphon

Computi
Computing is
forces to switch

The Revolution of Personal Computing

Smartphone Era (2007–present)

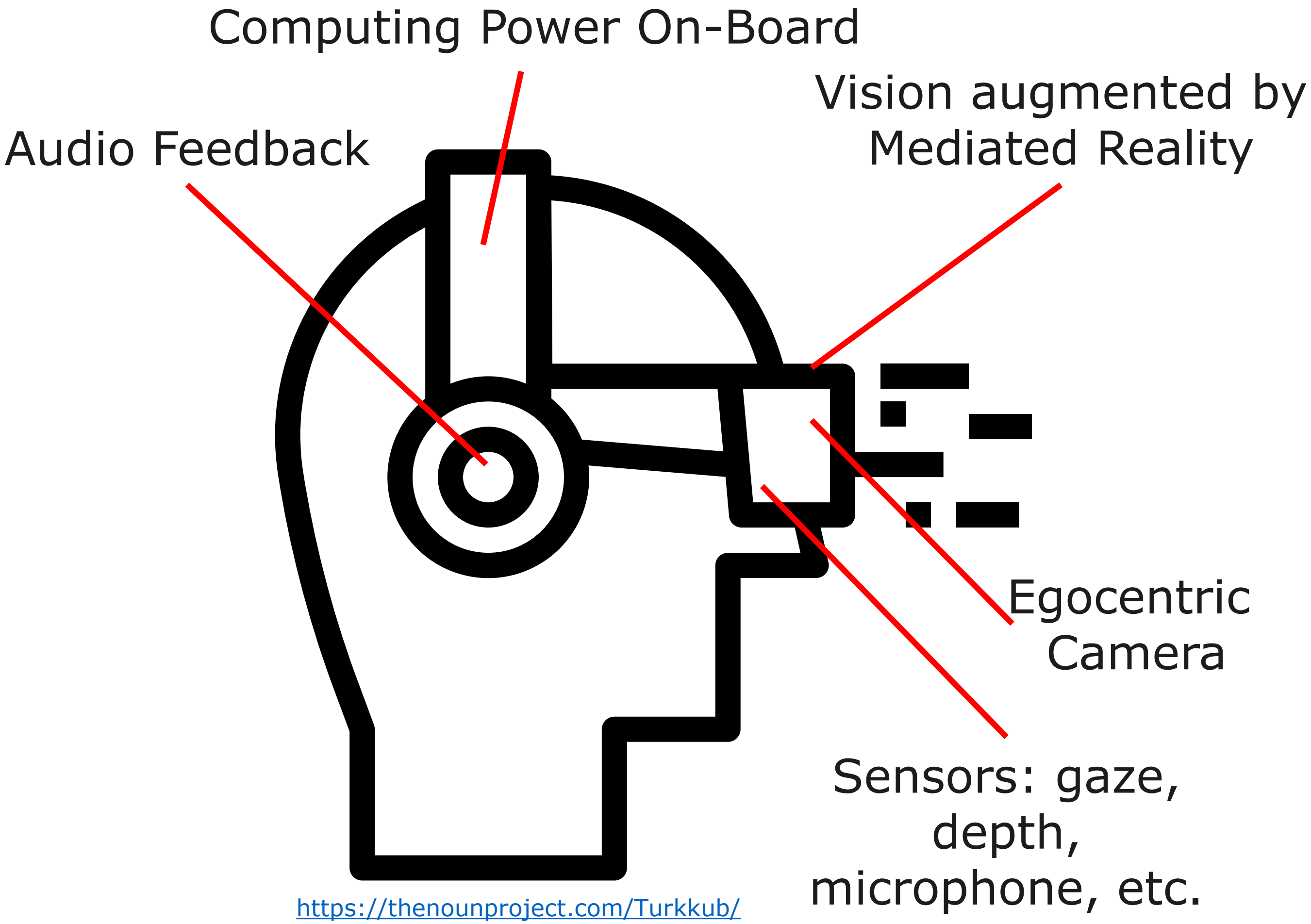
Computing in your pocket.
Computing is always accessible, but forces to switch between the digital and real world

Smartglasses Era (Now and Future)

Hands-free, always-on, egocentric vision.
Computing everywere with minimal switch between real and digital worlds

An AI-Powered Virtual Assistant

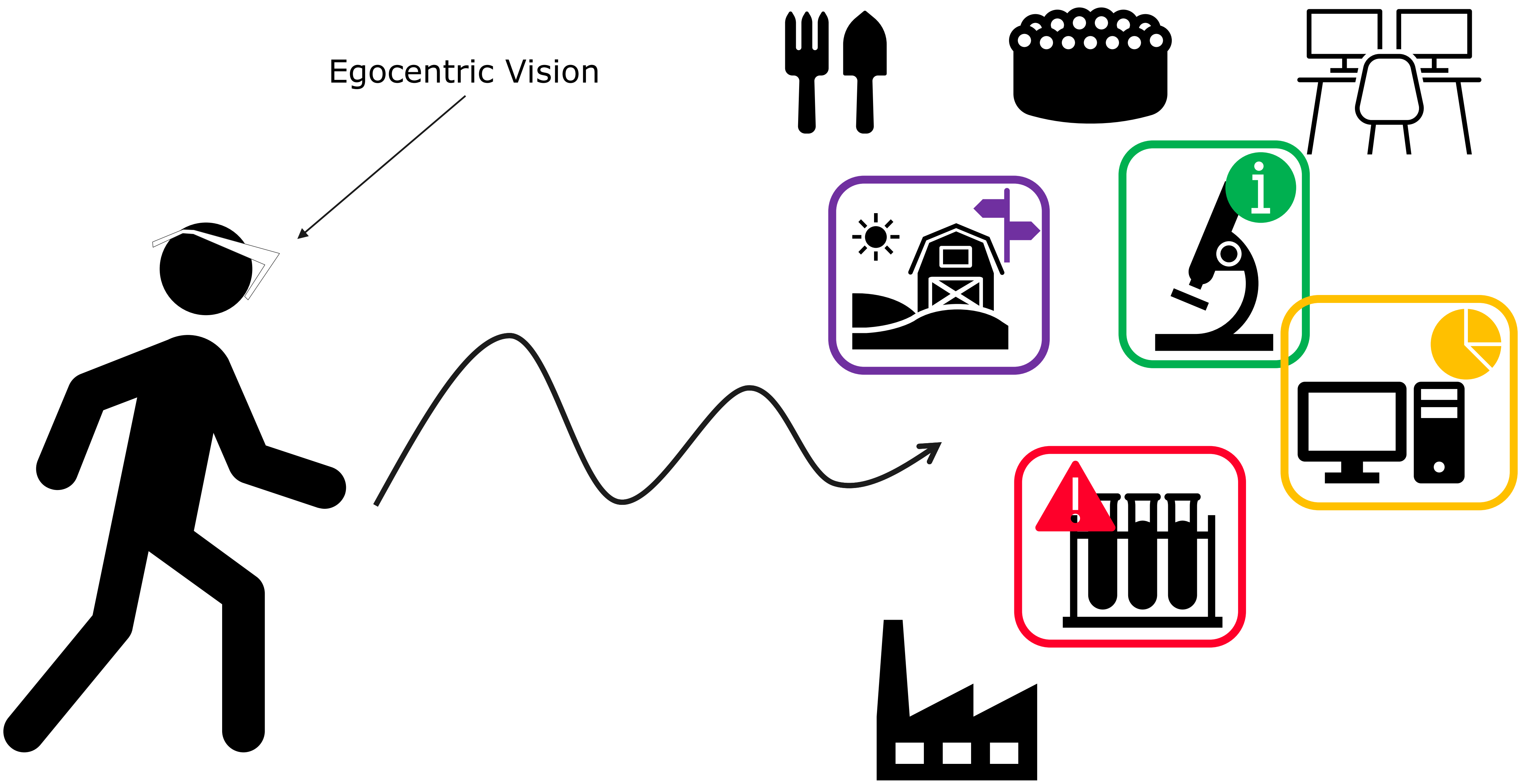
"her" 2013 movie



https://thenounproject.com/Turkkub/

A wearable device which perceives the world from our "egocentric" point of view is perfect for implementing a virtual assistant

A Virtual Personal Assistant



(Egocentric) Computer Vision is Fundamental!

Ego-Exo centric Vision

Exocentric

- ✓ Easy to setup
- ✓ Controlled Field of View
- × Doesn't always see everything
- × Not really portable

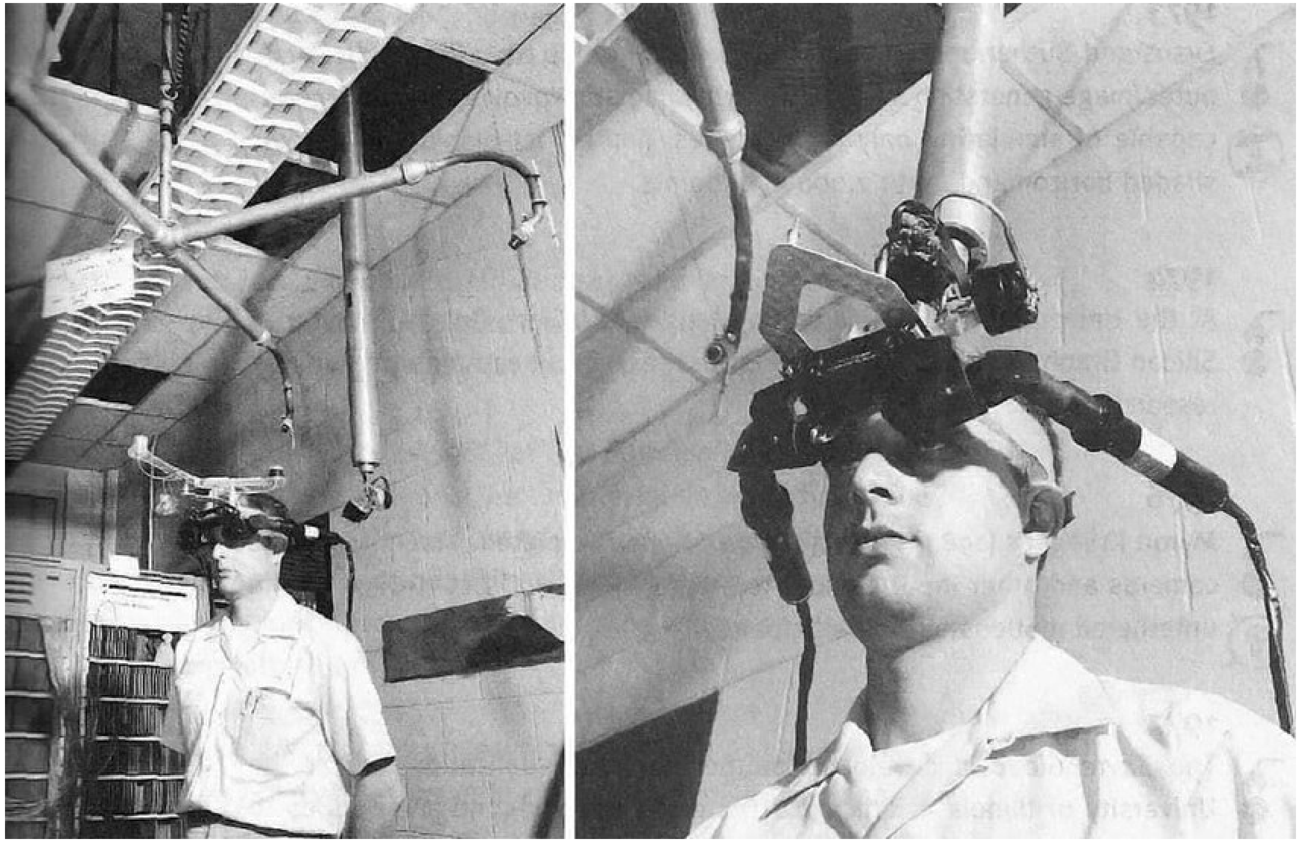
Egocentric

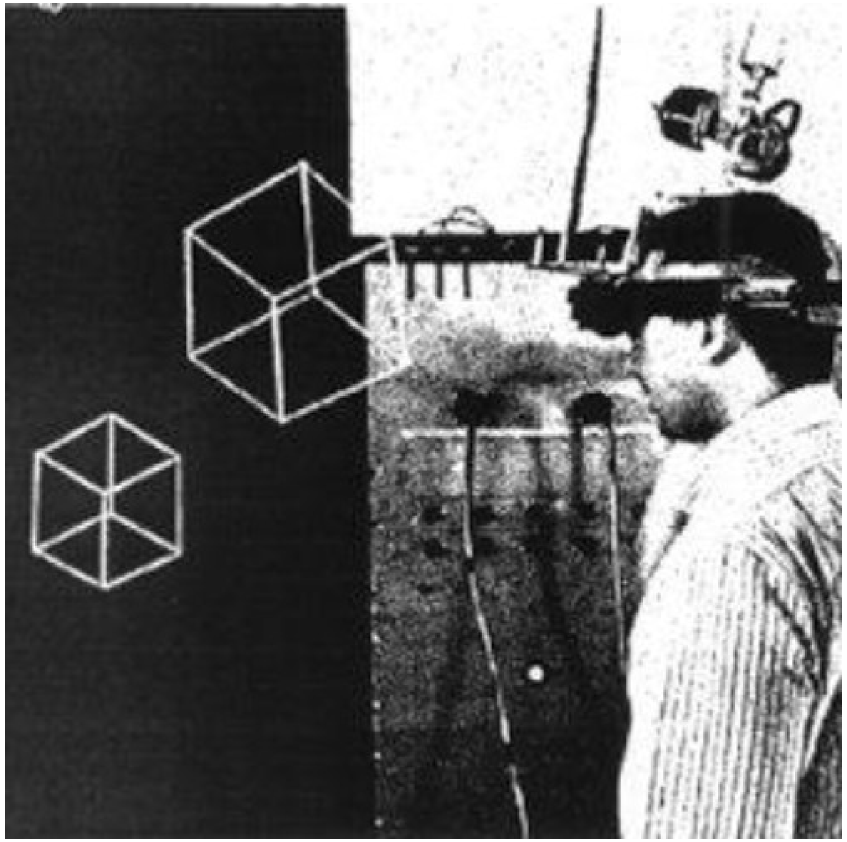
- ✓ Content is always relevant
- ✓ Intrinsically mobile
- × High variability
- × Operational constraints

Receive/Acquire Information

Head Mounted Display (1968)

In 1968 Ivan Sutherland invented the first "head mounted display" (HMD), a stereoscopic display mounted on the head of the user which allowed to show wireframe rooms.



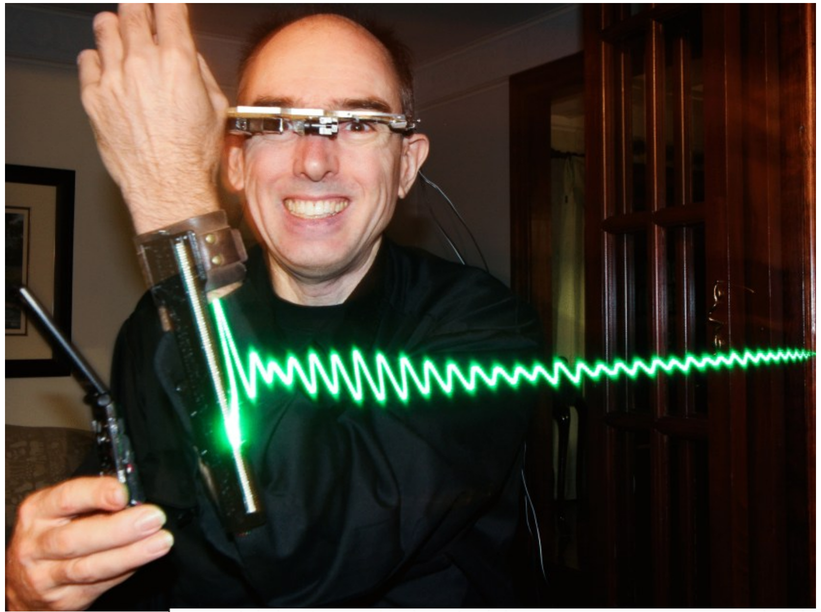


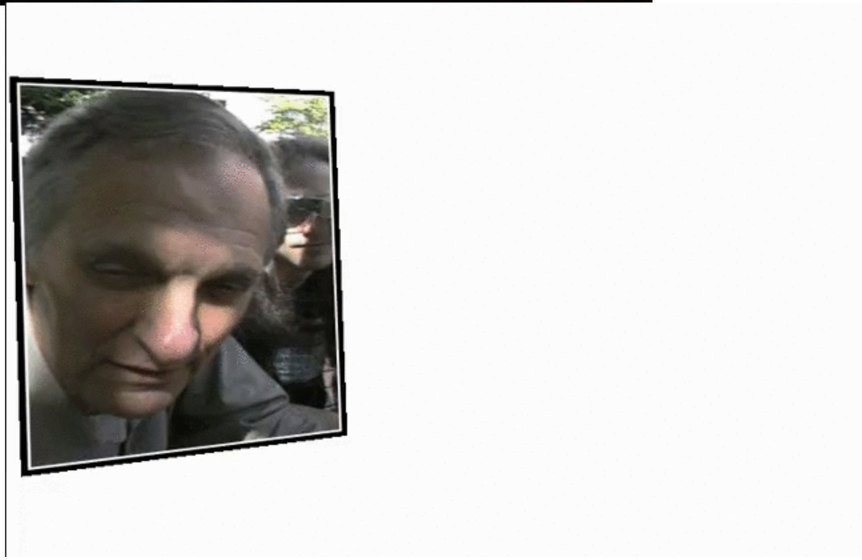
Due to its weight, the display was fixed to the ceiling with a pipe, for which it was called «sword of Damocles».

The Birth of Wearable Computing

In the 80s and 90s Steve Mann (PhD in Media Arts and Sciences at MIT, 1997) invented a number of wearable computers featuring video capabilities, computing capabilities, and a wearable screen for feedback. **Steve Mann is often referred to as «the father of wearable computing»**

The Birth of Wearable Computing





- EyeTap Digital Eye Glass
- SWIM (Sequential Wave Imprinting Machine)
- High-dynamic range imaging (HDR)
- Smartwatch
- Visual Orbits

Steve Mann. "Compositing multiple pictures of the same scene." *Proc. IS&T Annual Meeting, 1993.*
Steve Mann, "Wearable computing: a first step toward personal imaging," in *Computer*, vol. 30, no. 2, pp. 25-32, Feb. 1997.

MIT Media Lab in 1997

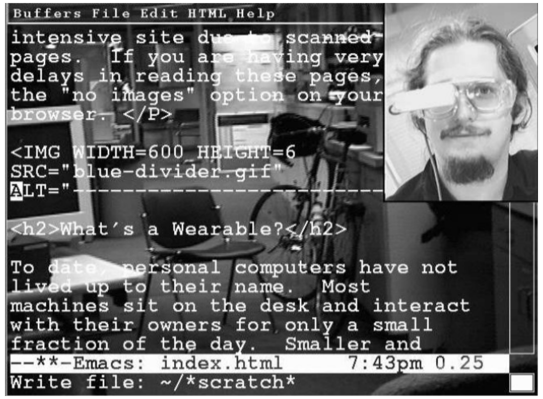
MIT Media Lab Seminal Works, late 1990s

Augmented Reality Through Wearable Computing

Thad Starner, Steve Mann, Bradley Rhodes, Jeffrey Levine
Jennifer Healey, Dana Kirsch, Roz Picard, and Alex Pentland

The Media Laboratory
Massachusetts Institute of Technology

(augmented reality)



1997

1998

Visual Contextual Awareness in Wearable Computing

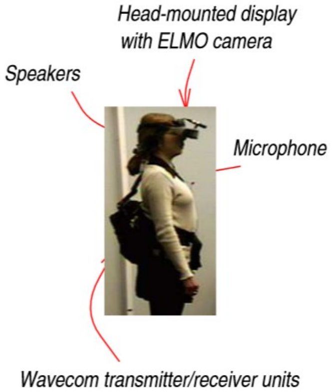
Thad Starner Bernt Schiele Alex Pentland
Media Laboratory, Massachusetts Institute of Technology

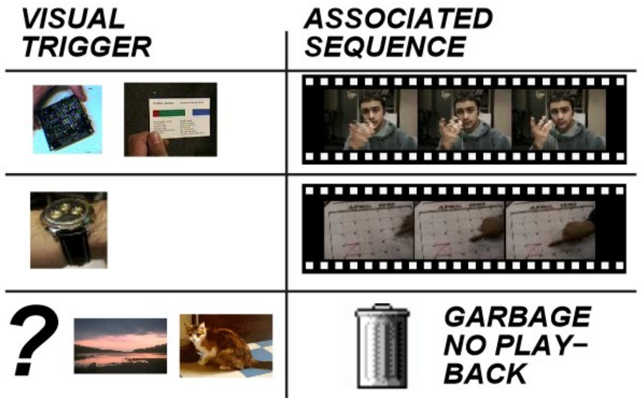
(location and task recognition)

An Interactive Computer Vision System DyPERS: Dynamic Personal Enhanced Reality System

Bernt Schiele, Nuria Oliver, Tony Jebara, and Alex Pentland
Vision and Modeling Group
MIT Media Laboratory, Cambridge, MA 02139, USA

(object recognition, media memories)





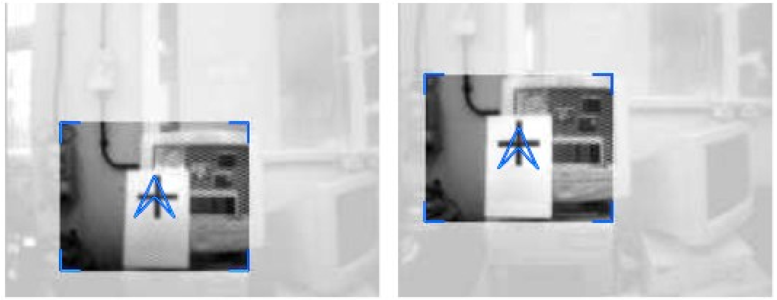
1999

Early 2000s

Wearable Visual Robots

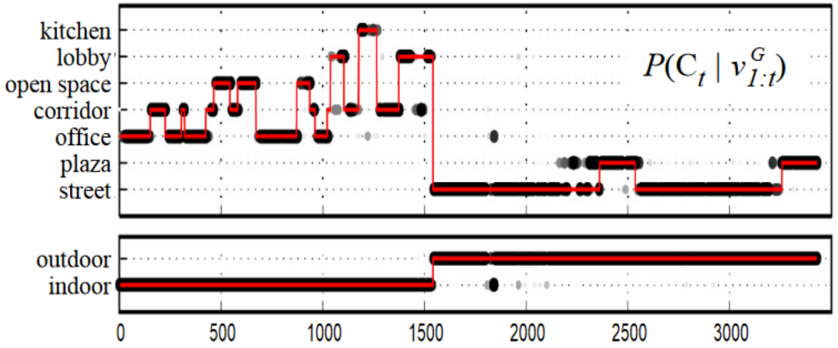
W.W. Mayol, B. Tordoff and D.W. Murray
University of Oxford, Parks Road, Oxford OX1 3PJ, UK

(active vision)



2000

2003



Context-based vision system for place and object recognition

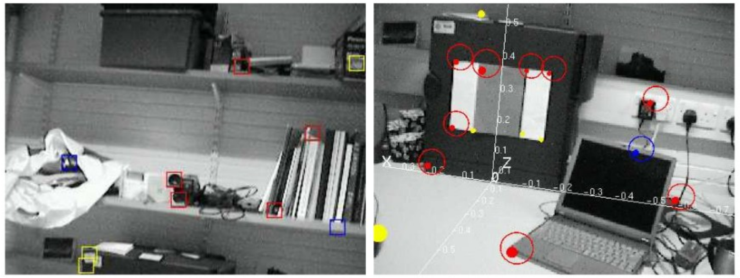
Antonio Torralba, MIT AI lab, Cambridge, MA 02139
Kevin P. Murphy, MIT AI lab, Cambridge, MA 02139
William T. Freeman, MIT AI lab, Cambridge, MA 02139
Mark A. Rubin, Lincoln Labs, Lexington, MA 02420

(location/object recognition)

Real-Time Localisation and Mapping with Wearable Active Vision *

Andrew J. Davison, Walterio W. Mayol and David W. Murray
Robotics Research Group
Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, UK

(active vision, SLAM)



2003

Late 2000s

Wearable Hand *Activity* Recognition for Event Summarization

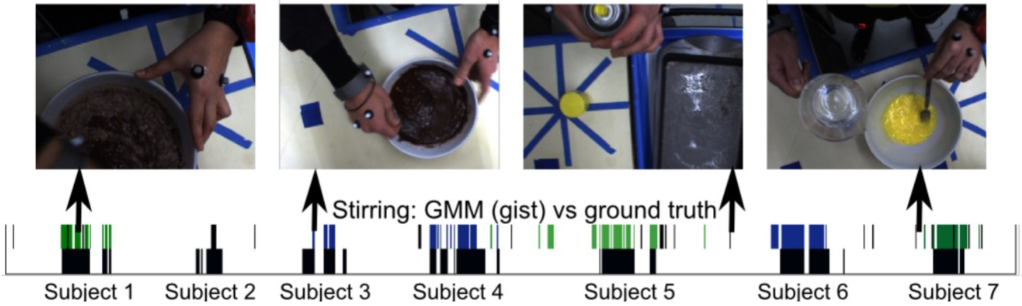
W.W. Mayol
Department of Computer Science
University of Bristol

D.W. Murray
Department of Engineering Science
University of Oxford

(hand activity recognition)

2005

2009



Temporal Segmentation and Activity Classification from First-person Sensing

Ekaterina H. Spriggs, Fernando De La Torre, Martial Hebert
Carnegie Mellon University.

(activity classification)

Figure-Ground Segmentation Improves Handled Object Recognition in Egocentric Video

Xiaofeng Ren
Intel Labs Seattle
1100 NE 45th Street, Seattle, WA 98105

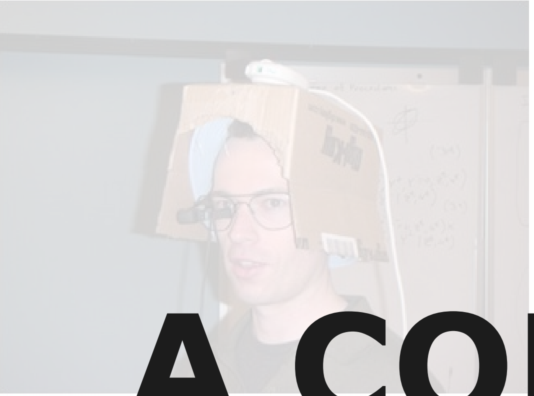
Chunhui Gu
University of California at Berkeley
Berkeley, CA 94720

(handheld object recognition)

2010

Università di Catania

Hardware, 1990s – 2000s



A COMMON HARDWARE PLATFORM WAS MISSING!

Microsoft SenseCam, 2004

"A day in Rome"

- SenseCam is a wearable camera that takes photos automatically;
- Originally conceived as a «personal blackbox» accident recorder;
- Used in the MyLifeBits project, inspired by Bush’s Memex;
- Inspired a series of conferences and many research papers.

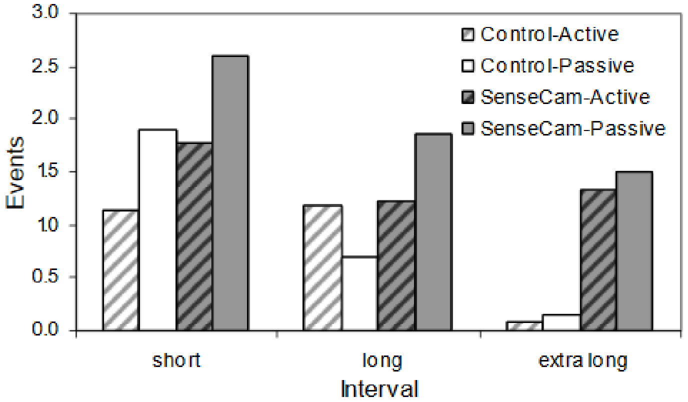
https://www.microsoft.com/en-us/research/project/sensecam/

Bell, Gordon, and Jim Gemmell. *Your life, uploaded: The digital way to better memory, health, and productivity*. Penguin, 2010.

Do Life-Logging Technologies Support Memory for the Past? An Experimental Study Using SenseCam

Abigail Sellen, Andrew Fogg, Mike Aitken*, Steve Hodges, Carsten Rother and Ken Wood
Microsoft Research Cambridge
7 JJ Thomson Ave, Cambridge, UK, CB3 0FB
*Behavioural & Clinical Neuroscience Institute
Dept. of Psychology, University of Cambridge

(health, memory augmentation)



2007

2008

(a) Reading in bed

(b) Having dinner

MyPlaces: Detecting Important Settings in a Visual Diary

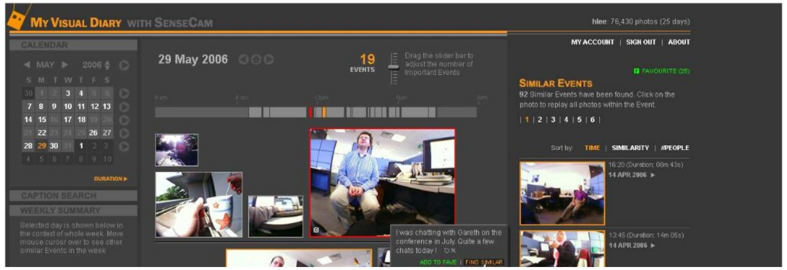
Michael Blighe and Noel E. O'Connor
Centre for Digital Video Processing, Adaptive Information Cluster
Dublin City University, Ireland
{blighem, oconnorn}@eeng.dcu.ie

(lifelogging, place recognition)

Constructing a SenseCam Visual Diary as a Media Process

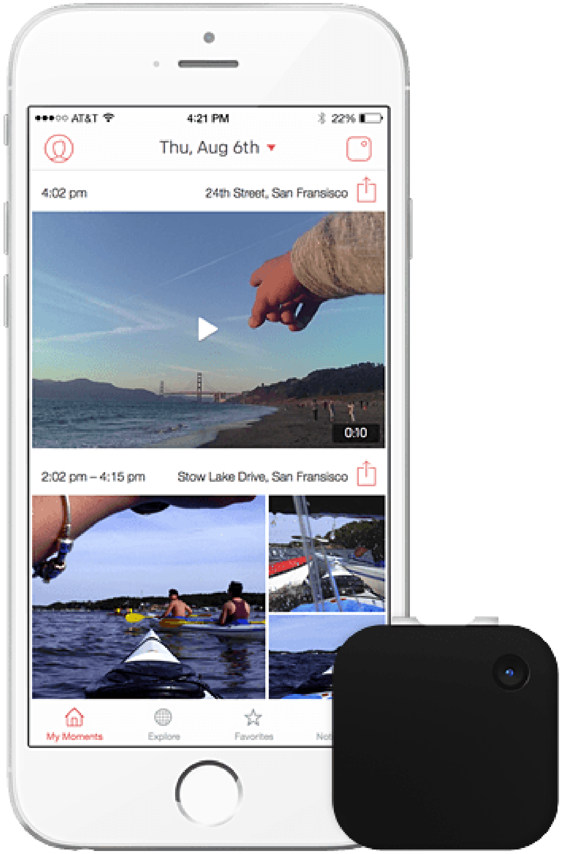
Hyowon Lee, Alan F. Smeaton, Noel O'Connor, Gareth Jones, Michael Blighe, Daragh Byrne, Aiden Doherty, and Cathal Gurrin
Centre for Digital Video Processing & Adaptive Information Cluster,
Dublin City University

(lifelogging, multimedia retrieval)



2008

Narrative Clip, 2012



http://getnarrative.com/

Research Using Narrative Clip

Multi-face tracking by extended bag-of-tracklets in egocentric photo-streams

Maedeh Aghaei[a,*], Mariella Dimiccoli[a,b], Petia Radeva[a,b]

(lifelogging, face tracking)

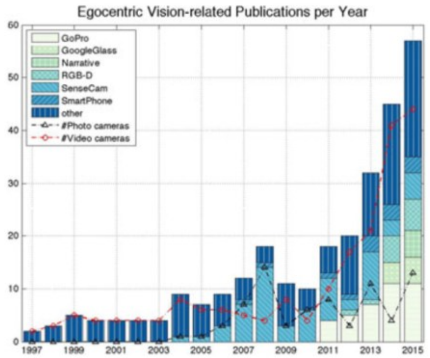
2016

2017

SR-clustering: Semantic regularized clustering for egocentric photo streams segmentation

Mariella Dimiccoli[a,c,1,*], Marc Bolaños[a,1,*], Estefania Talavera[a,b], Maedeh Aghaei[a], Stavri G. Nikolov[d], Petia Radeva[a,c,*]

(lifelogging, event segmentation)



Toward Storytelling From Visual Lifelogging: An Overview

Marc Bolaños, Mariella Dimiccoli, and Petia Radeva

(lifelogging, survey)

2017

What About Video?

Università di Catania

GoPro HD Hero, 2010

different wearing modalities

head-mounted chest-mounted

wrist-mounted helmet-mounted

https://www.youtube.com/watch?v=D4iU-EOJYK8

Early Research On Egocentric Video

Fast Unsupervised Ego-Action Learning for First-Person Sports Videos

Kris M. Kitani
UEC Tokyo
Tokyo, Japan

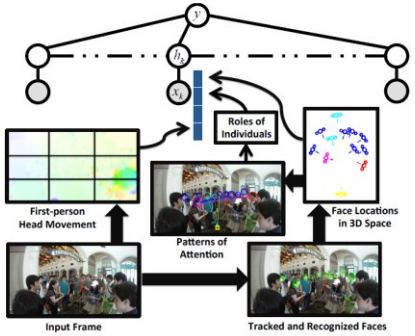
Takahiro Okabe, Yoichi Sato
University of Tokyo
Tokyo, Japan

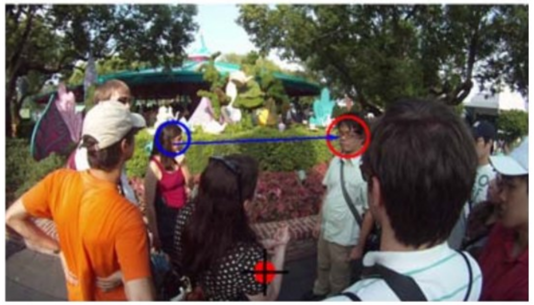
Akihiro Sugimoto
National Institute of Informatics
Tokyo, Japan

(unsupervised action recognition, video indexing)

2011

2012





Social Interactions: A First-Person Perspective

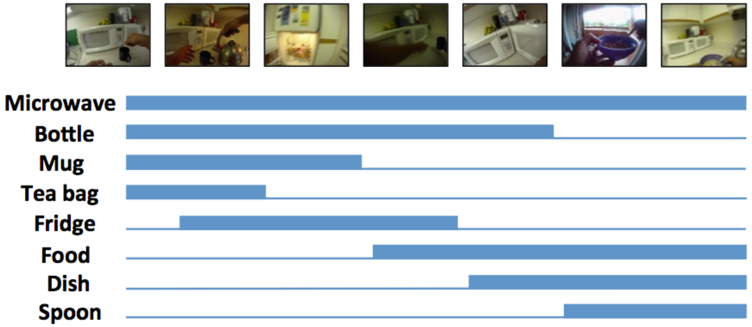
Alireza Fathi[1], Jessica K. Hodgins[2,3], James M. Rehg[1]

(detection and recognition of social interactions)

Story-Driven Summarization for Egocentric Video

Zheng Lu and Kristen Grauman
University of Texas at Austin

(egocentric video sumarization)



Our method

2013

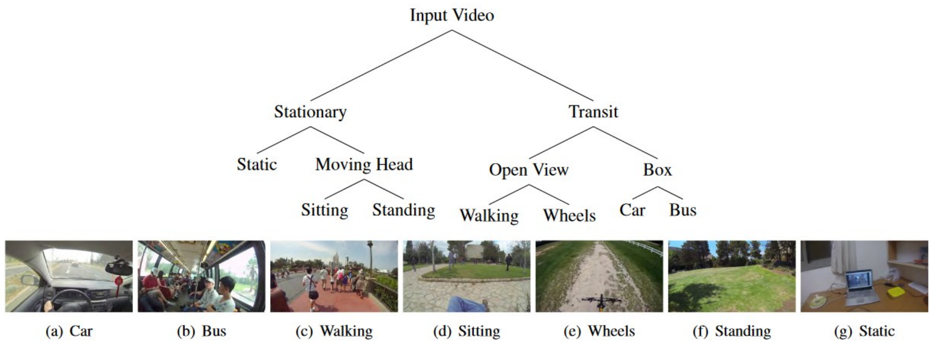
Università di Catania

Later Research On Egocentric Video

Temporal Segmentation of Egocentric Videos

Yair Poleg Chetan Arora* Shmuel Peleg

(egocentric video indexing)

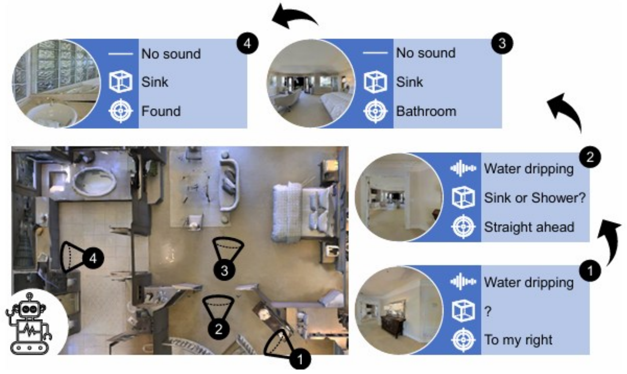


2014

Semantic Audio-Visual Navigation

Changan Chen[1,2] Ziad Al-Halah[1] Kristen Grauman[1,2]
[1]UT Austin [2]Facebook AI Research

2021

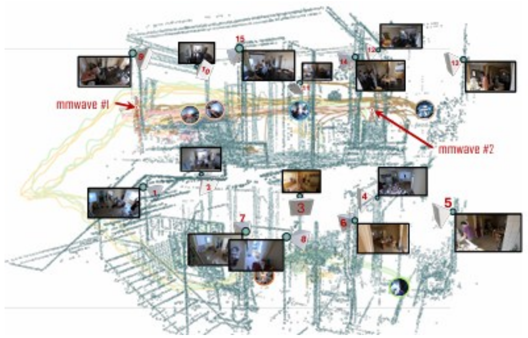


EgoLife: Towards Egocentric Life Assistant

The EgoLife Team

https://egolife-ai.github.io/

(video understanding, egocentric assistant)

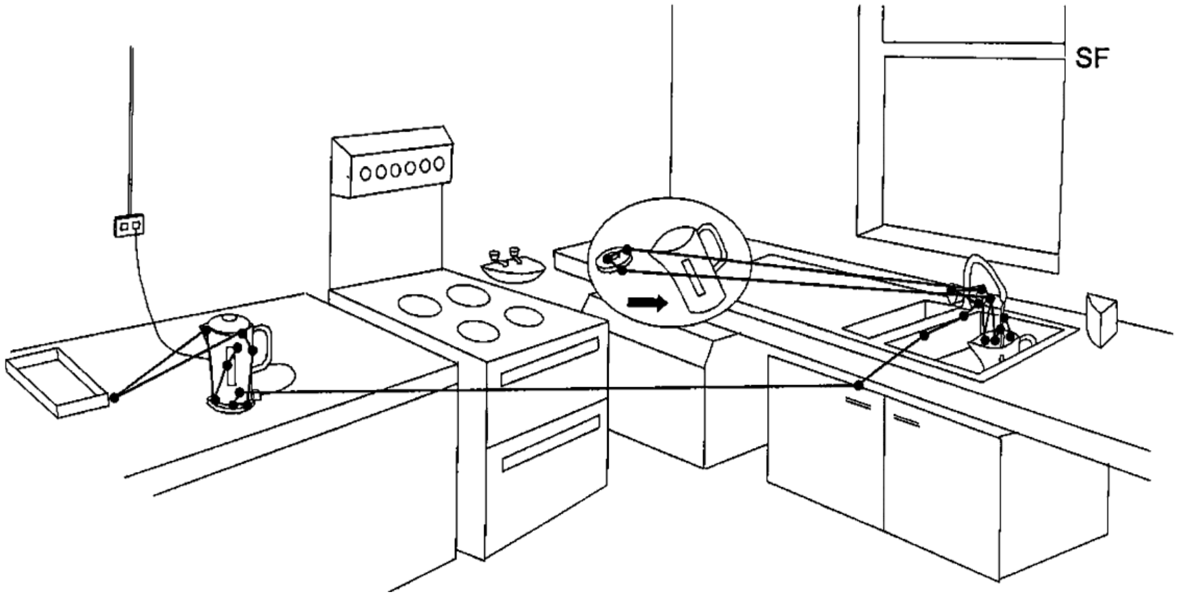


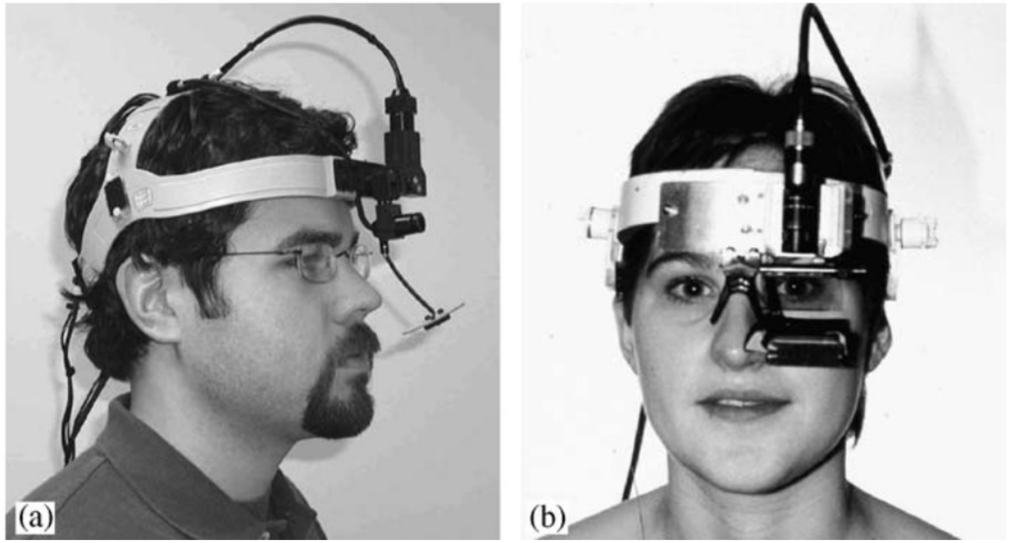
2025

Gaze Trackers

Eye movements and the control of actions in everyday life

Michael F. Land





Prototype by Land (1993)

Gaze is important in Egocentric Vision!

Tobii Pro Glasses 2 (2014)

Microsoft HoloLens 2 (2016)

Mobile Eye-XG (2013)

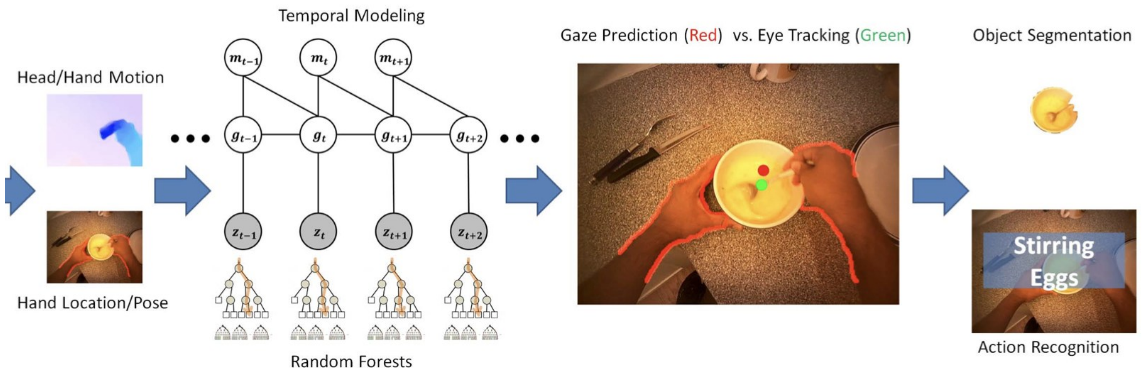
Pupil Eye Trackers (2014 -)

Research On Gaze

Learning to Predict Gaze in Egocentric Video

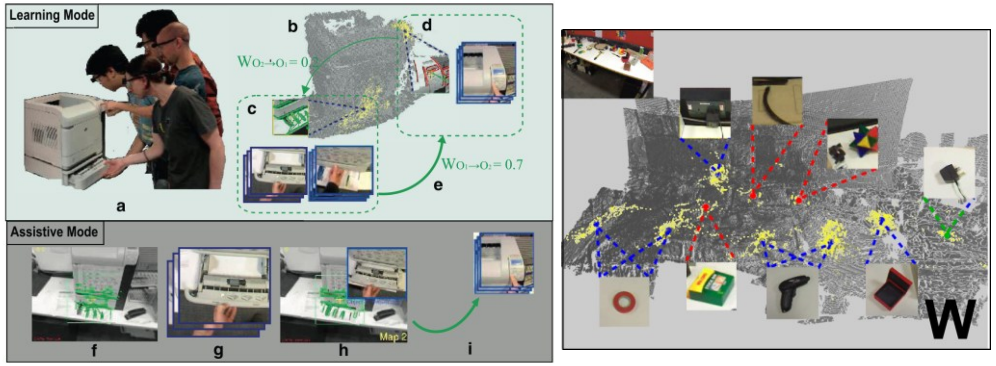
Yin Li, Alireza Fathi, James M. Rehg

(gaze prediciton, action recognition)



2012

2016



You-Do, I-Learn: Egocentric unsupervised discovery of objects and their modes of interaction towards video-based guidance

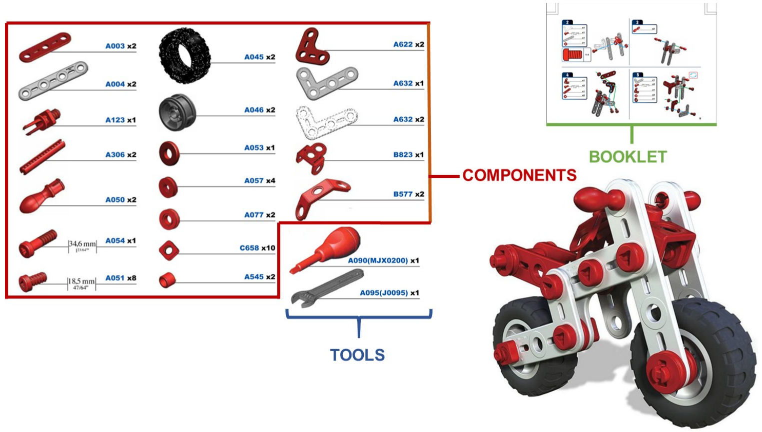
Dima Damen*, Teesid Leelasawassuk, Walterio Mayol-Cuevas

(object usage discovery, assistance)

MECCANO: A multimodal egocentric dataset for humans behavior understanding in the industrial-like domain

Francesco Ragusa *, Antonino Furnari, Giovanni Maria Farinella

(gaze prediciton, procedural video)



2023

OrCam MyEye, since 2015

Health, assistive technologies

https://www.orcam.com/

OrCam MyEye, since 2015

https://www.orcam.com/

Microsoft HoloLens, since 2016 – HoloLens2 in 2020

Mixed Reality

https://www.microsoft.com/hololens

https://youtu.be/eqFqtAJMtYE

Università di Catania

Microsoft HoloLens2 – Towards Industrial Applications

HoloLens 2

An ergonomic, untethered self-contained holographic device with enterprise-ready applications to increase user accuracy and output.

$3,500

HoloLens 2 Industrial Edition

A HoloLens 2 that is designed and tested to support regulated environments such as clean rooms and hazardous locations.

$4,950

Trimble XR10 with HoloLens 2

A hardhat-integrated HoloLens 2 that is purpose-built for personnel in dirty, loud, and safety-controlled work site environments.

$5,199

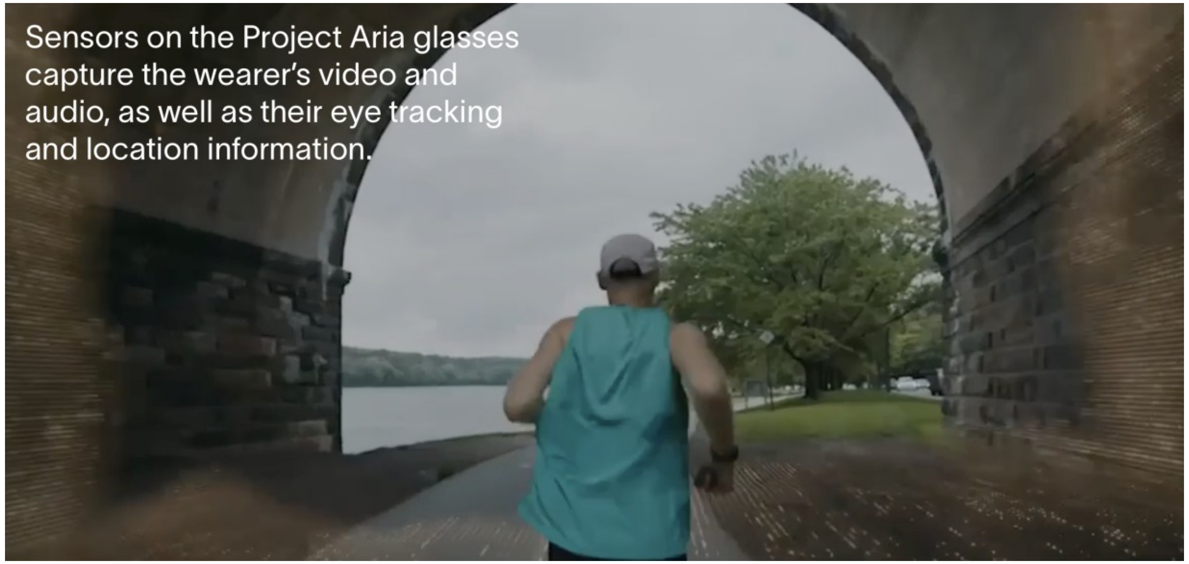
https://www.microsoft.com/en-us/hololens/buy

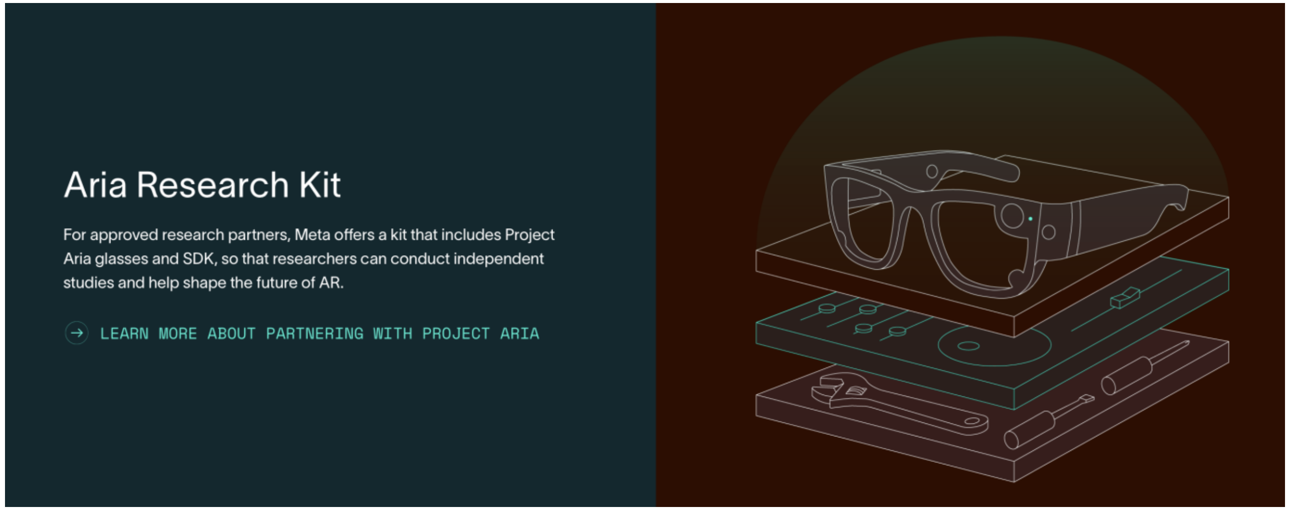
Università di Catania

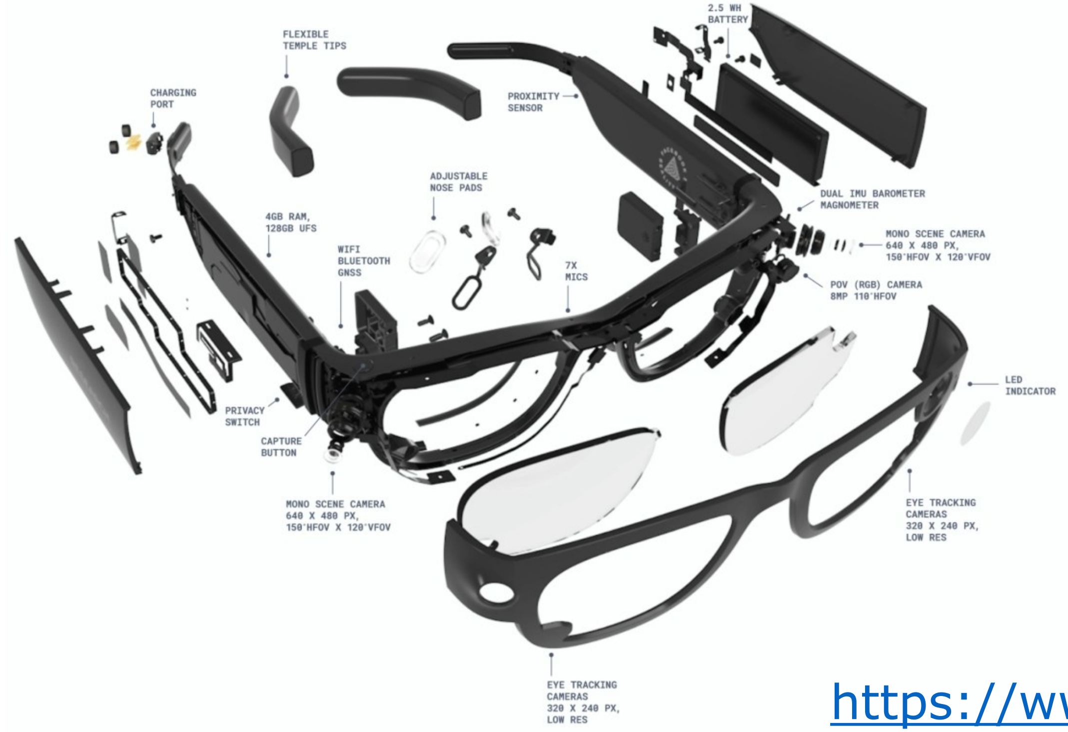
Magic Leap, since 2018 - Magic Leap 2 in 2022

https://www.magicleap.com/magic-leap-2

Meta’s Project Aria

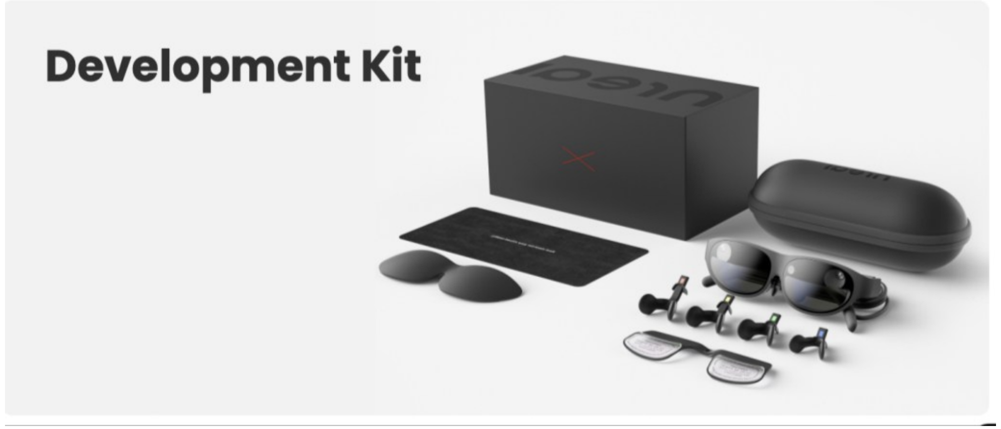


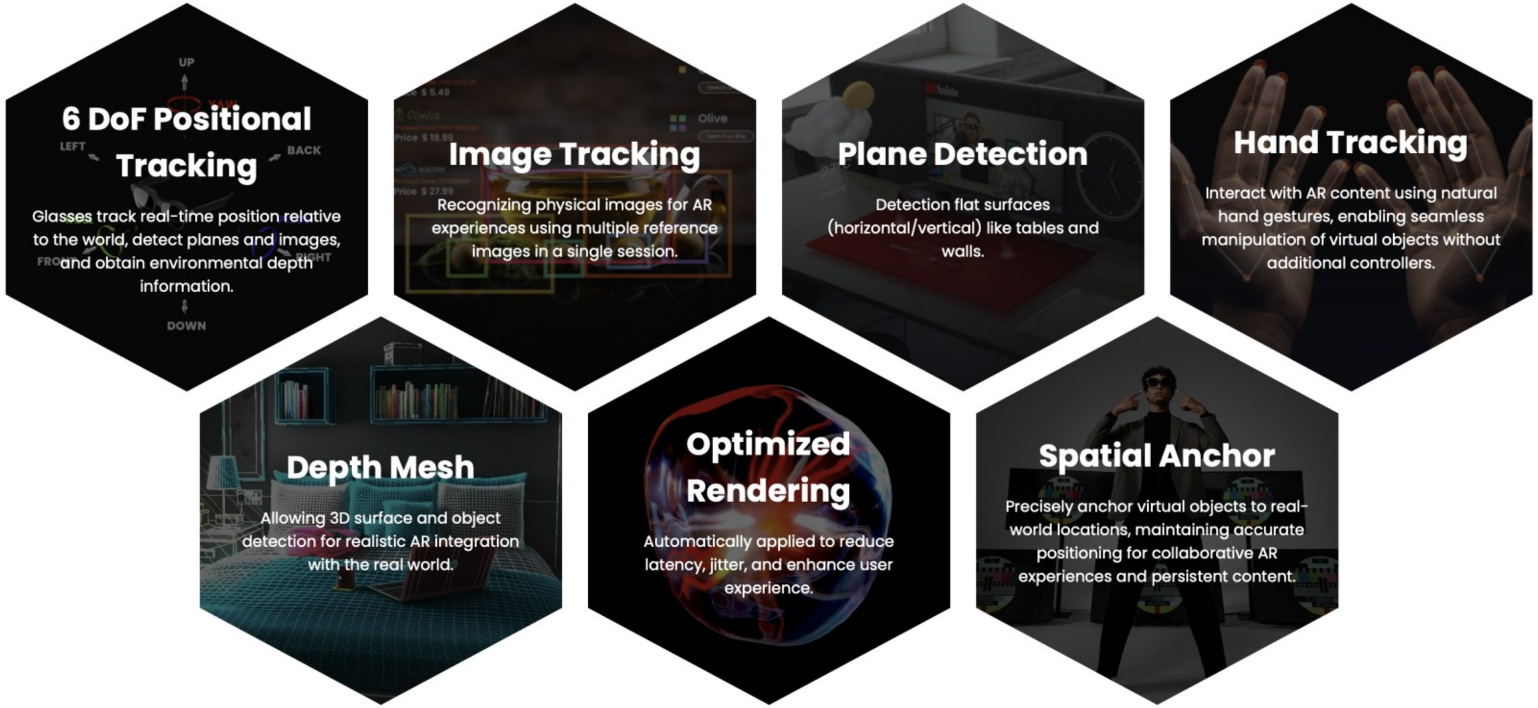




https://www.projectaria.com

XREAL

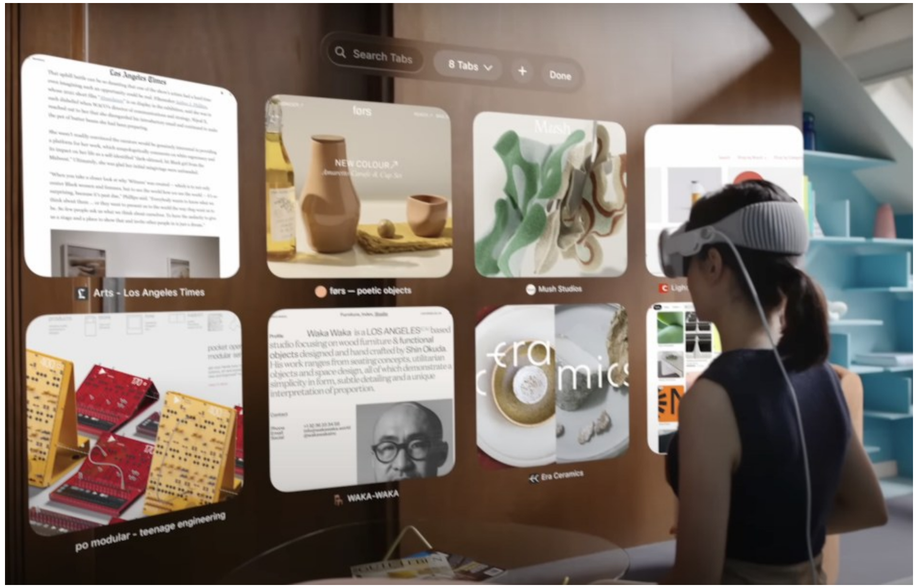


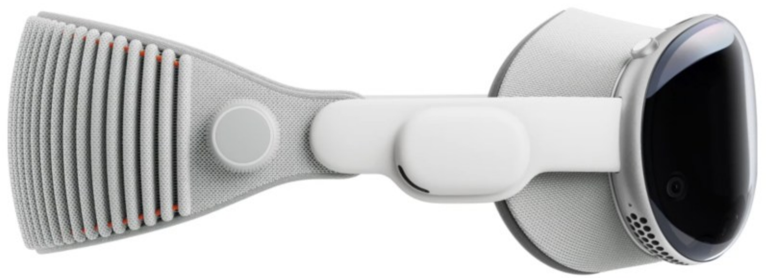


https://www.xreal.com/

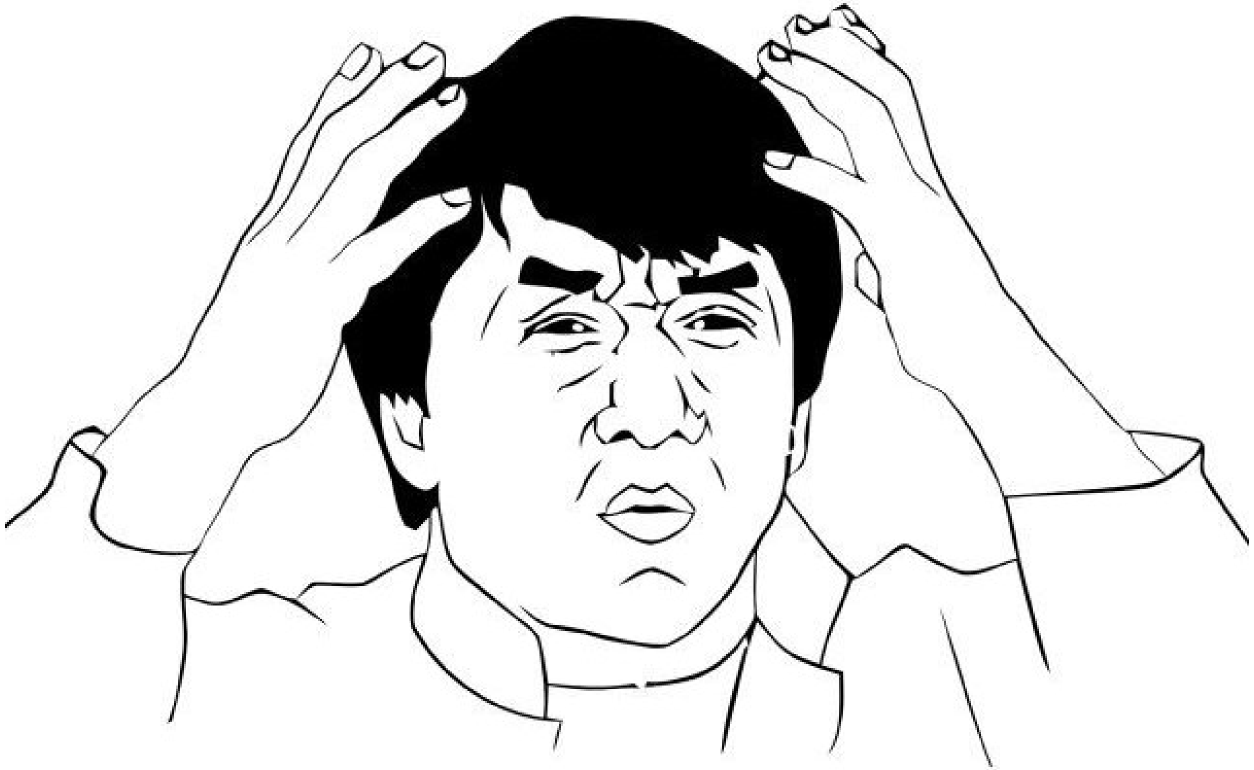
Università di Catania

Apple Vision Pro





https://www.apple.com/apple-vision-pro/



Too Many Devices?

towards standardization…

OpenXR

Unified API supported by many AR and VR devices

https://www.khronos.org/openxr/

Università di Catania

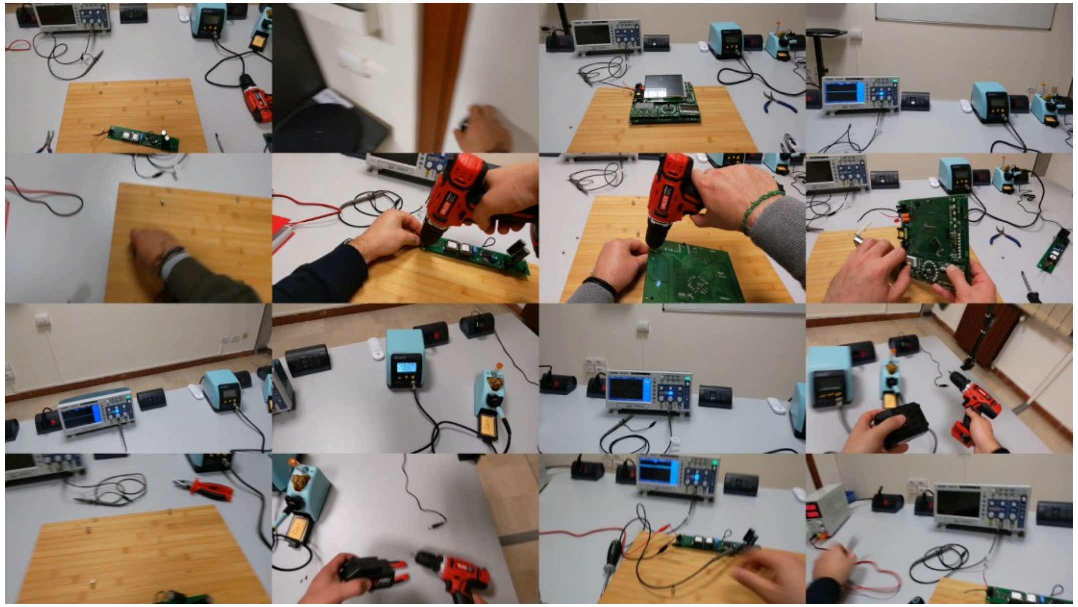
Circa 2017 – most of the discussion still in workshops

Workshop on Egocentric (First Person) Vision

ACVR

EPIC

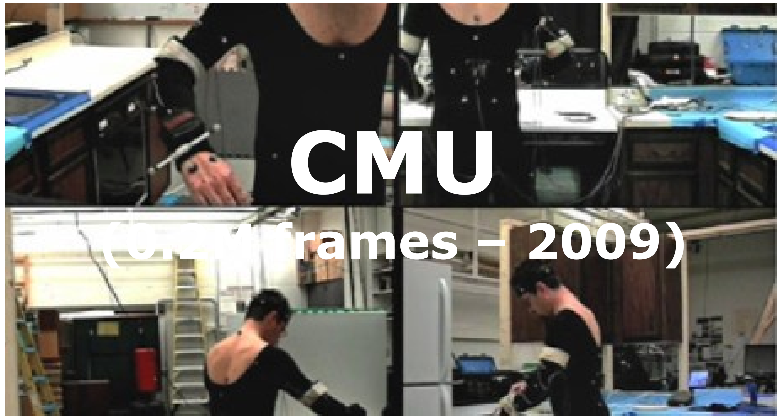
EGO4D



Digital Information

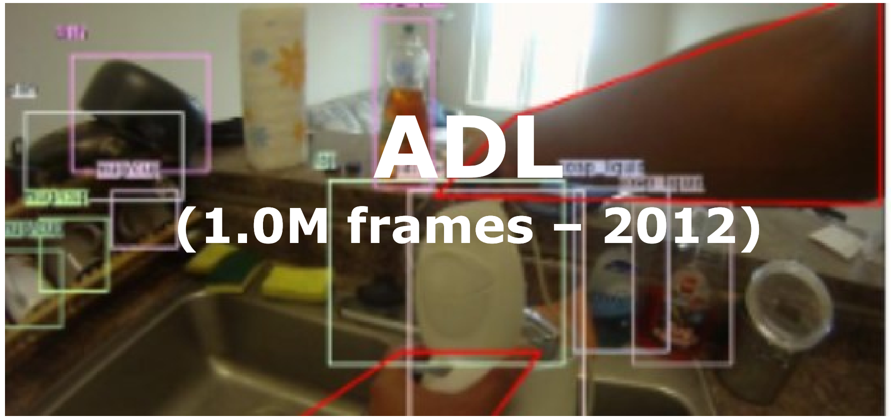
Università di Catania

First Person Vision Research – Datasets (up to 2018)



http://www.cs.cmu.edu/~espriggs/cmu-mmac/annotations/

http://www.cbi.gatech.edu/fpv/



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

https://allenai.org/plato/charades/

http://www.cbi.gatech.edu/fpv/

Università di Catania

The EPIC series

Scaling Egocentric Vision: The EPIC-KITCHENS Dataset

Dima Damen[1][0000−0001−8804−6238], Hazel Doughty[1], Giovanni Maria Farinella[2], Sanja Fidler[3], Antonino Furnari[2], Evangelos Kazakos[1], Davide Moltisanti[1], Jonathan Munro[1], Toby Perrett[1], Will Price[1], and Michael Wray[1]

[1]Uni. of Bristol, UK [2]Uni. of Catania, Italy, [3]Uni. of Toronto, Canada

Abstract. First-person vision is gaining interest as it offers a unique viewpoint on people's interaction with objects, their attention, and even intention. However, progress in this challenging domain has been relatively slow due to the lack of sufficiently large datasets. In this paper, we introduce EPIC-KITCHENS, a large-scale egocentric video benchmark recorded by 32 participants in their native kitchen environments. Our videos depict **non-scripted** daily activities: we simply asked each participant to start recording every time they entered their kitchen. Recording took place in 4 cities (in North America and Europe) by participants belonging to 10 different nationalities, resulting in highly diverse cooking styles. Our dataset features 55 hours of video consisting of 11.5M frames, which we densely labelled for a total of 39.6K action segments and 454.3K object bounding boxes. Our annotation is unique in that we had the participants narrate their own videos (after recording), thus reflecting true intention, and we crowd-sourced ground-truths based on these. We describe our object, action and anticipation challenges, and evaluate several baselines over two test splits, *seen* and *unseen* kitchens.

Keywords: Egocentric Vision, Dataset, Benchmarks, First-Person Vision, Egocentric Object Detection, Action Recognition and Anticipation

EPIC-Kitchens 55

Rescaling Egocentric Vision: Collection Pipeline and Challenges for EPIC-KITCHENS-100

Dima Damen · Hazel Doughty · Giovanni Maria Farinella · Antonino Furnari · Evangelos Kazakos · Jian Ma · Davide Moltisanti · Jonathan Munro · Toby Perrett · Will Price · Michael Wray

Received: 18 Jan 2021, Revised: 23 Aug 2021, Accepted: 17 Sep 2021

Abstract This paper introduces the pipeline to extend the largest dataset in egocentric vision, EPIC-KITCHENS. The effort culminates in EPIC-KITCHENS-100, a collection of 100 hours, 20M frames, 90K actions in 700 variable-length videos, capturing long-term unscripted activities in 45 environments, using head-mounted cameras. Compared to its previous version [1], EPIC-KITCHENS-100 has been annotated using a novel pipeline that allows denser (54% more actions per minute) and more complete annotations of fine-grained actions (+128% more action segments). This collection enables new challenges such as action detection and evaluating the "test of time" — i.e. whether models trained on data collected in 2018 can generalise to new footage collected two years later. recognition. For each challenge, we define the task, provide baselines and evaluation metrics[1].

Keywords Video Dataset, Egocentric Vision, First-Person Vision, Action Understanding, Multi-Benchmark Large-Scale Dataset, Annotation Quality

1 Introduction and Related Datasets

Since the dawn of machine learning for computer vision, datasets have been curated to train models, for single tasks from classification [2,3] to detection [4,5], captioning [6,7] and segmentation [8,9]. Increasingly, datasets have been used for novel tasks, through pre-training [10, 11], self-supervision [12,13] or additional annotations [14, 15]. However, task adaptation demonstrates that models

EPIC-Kitchens 100

EPIC-KITCHENS VISOR Benchmark VIdeo Segmentations and Object Relations

Ahmad Darkhalil** Dandan Shan** Bin Zhu** Jian Ma** Amlan Kar Richard Higgins Sanja Fidler David Fouhey Dima Damen

Uni. of Bristol, UK Uni. of Michigan, US Uni. of Toronto, CA *: Co-First Authors

Abstract

We introduce VISOR, a new dataset of pixel annotations and a benchmark suite for segmenting hands and active objects in egocentric video. VISOR annotates videos from EPIC-KITCHENS, which comes with a new set of challenges not encountered in current video segmentation datasets. Specifically, we need to ensure both short- and long-term consistency of pixel-level annotations as objects undergo transformative interactions, e.g. an onion is peeled, diced and cooked - where we aim to obtain accurate pixel-level annotations of the peel, onion pieces, chopping board, knife, pan, as well as the acting hands. VISOR introduces an annotation pipeline, AI-powered in parts, for scalability and quality. In total, we publicly release 272K manual semantic masks of 257 object classes, 9.9M interpolated dense masks, 67K hand-object relations, covering 36 hours of 179 untrimmed videos. Along with the annotations, we introduce three challenges in video object segmentation, interaction understanding and long-term reasoning.

For data, code and leaderboards: http://epic-kitchens.github.io/VISOR

EPIC-Kitchens VISOR

EPIC-SOUNDS: A Large-Scale Dataset of Actions that Sound

Jaesung Huh[1*], Jacob Chalk[2*], Evangelos Kazakos[3], Dima Damen[2], Andrew Zisserman[1]

[1]Visual Geometry Group, Department of Engineering Science, University of Oxford, UK
[2]Department of Computer Science, University of Bristol, UK
[3] CIIRC, Czech Technical University in Prague, Czech Republic

https://epic-kitchens.github.io/epic-sounds/

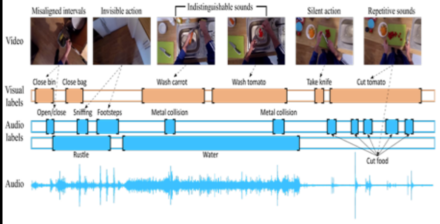


Fig. 1: Sample video with corresponding audio from EPIC-KITCHENS-100 [1]. We compare the already published **visual labels** with our collected EPIC-SOUNDS **audio labels**. We demonstrate the differences between the modality annotations, both in temporal extent and class labels, highlighting: **Misaligned intervals:** temporal boundaries are distinct; **Invisible action:** action not seen in the video, but which produces distinct sounds (0-to-1 matching); **Indistinguishable sounds:** sounds from two distinct visual actions, but are audibly inseparable; **Silent action:** visual action that does not have audible sounds (1-to-0); and visual actions containing multiple **repetitive sounds** (1-to-N).

Abstract—We introduce EPIC-SOUNDS, a large-scale dataset of audio annotations capturing temporal extents and class labels within the audio stream of the egocentric videos. We propose

Index Terms—audio recognition, action recognition, audio event detection, audio dataset, data collection, dataset

EPIC-SOUNDS

EPIC Fields Marrying 3D Geometry and Video Understanding

Vadim Tschernezki** Ahmad Darkhalil** Zhifan Zhu** David Fouhey Iro Laina Diane Larlus Dima Damen Andrea Vedaldi

VGG, University of Oxford University of Bristol New York University NAVER LABS Europe *: Equal Contribution

Abstract

Neural rendering is fuelling a unification of learning, 3D geometry and video understanding that has been waiting for more than two decades. Progress, however, is still hampered by a lack of suitable datasets and benchmarks. To address this gap, we introduce EPIC Fields, an augmentation of EPIC-KITCHENS with 3D camera information. Like other datasets for neural rendering, EPIC Fields removes the complex and expensive step of reconstructing cameras using photogrammetry, and allows researchers to focus on modelling problems. We illustrate the challenge of photogrammetry in egocentric videos of dynamic actions and propose innovations to address them. Compared to other neural rendering datasets, EPIC Fields is better tailored to video understanding because it is paired with labelled action segments and the recent VISOR segment annotations. To further motivate the community, we also evaluate three benchmark tasks in neural rendering and segmenting dynamic objects, with strong baselines that showcase what is not possible today. We also highlight the advantage of geometry in semi-supervised video object segmentations on the VISOR annotations. EPIC Fields reconstructs 96% of videos in EPIC-KITCHENS, registering 19M frames in 99 hours recorded in 45 kitchens, and is available from: http://epic-kitchens.github.io/epic-fields

EPIC-FIELDS

HD-EPIC: A Highly-Detailed Egocentric Video Dataset

Toby Perrett* Ahmad Darkhalil* Saptarshi Sinha* Omar Emara* Sam Pollard* Kranti Parida* Kaiting Liu* Prajwal Gatti* Siddhant Bansal* Kevin Flanagan* Jacob Chalk* Zhifan Zhu* Rhodri Guerrier* Fahd Abdelazim* Bin Zhu Davide Moltisanti Michael Wray Hazel Doughty Dima Damen

Uni. of Bristol Leiden Uni. Singapore Management Uni. Uni. of Bath *: Equal Contribution

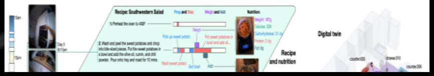
http://hd-epic.github.io

Abstract

We present a validation dataset of newly-collected kitchen-based egocentric videos, manually annotated with highly detailed and interconnected ground-truth labels covering: recipe steps, fine-grained actions, ingredients with nutritional values, moving objects, and audio annotations. Importantly, all annotations are grounded in 3D through digital twinning of the scene, fixtures, object locations, and primed with gaze. Footage is collected from unscripted recordings in diverse home environments, making HD-EPIC the first dataset collected in-the-wild but with detailed annotations matching those in controlled lab environments.

We show the potential of our highly-detailed annotations through a challenging VQA benchmark of 26K questions assessing the capability to recognise recipes, ingredients, nutrition, fine-grained actions, 3D perception, object motion, and gaze direction. The powerful long-context Gemini Pro only achieves 38.5% on this benchmark, showcasing its difficulty and highlighting shortcomings in current VLMs. We additionally assess action recognition, sound recognition, and long-term video-object segmentation on HD-EPIC.

HD-EPIC is 41 hours of video in 9 kitchens with digital twins of 413 kitchen fixtures, capturing 69 recipes, 59K fine-grained actions, 51K audio events, 20K object movements and 37K object masks lifted to 3D. On average, we have 263 annotations per minute of our unscripted videos.



HD-EPIC

EPIC-KITCHENS

TEAM

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro and Toby Perrett, Will Price, Michael Wray (2021). The EPIC-KITCHENS Dataset: Collection, Challenges and Baselines. PAMI, 43(11), pp. 4125-4141.

University of Bristol

University of Toronto

Università degli Studi di Catania

Dima Damen
Principal Investigator
University of Bristol
United Kingom

Sanja Fidler
Co-Investigator
University of Toronto
Canada

Giovanni Maria Farinella
Co-Investigator
University of Catania
Italy

Davide Moltisanti
(Apr 2017 -)
University of Bristol

Michael Wray
(Apr 2017 -)
University of Bristol

Hazel Doughty
(Apr 2017 -)
University of Bristol

Toby Perrett
(Apr 2017 -)
University of Bristol

Antonino Furnari
(Jul 2017 -)
University of Catania

Jonathan Munro
(Sep 2017 -)
University of Bristol

Evangelos Kazakos
(Sep 2017 -)
University of Bristol

Will Price
(Oct 2017 -)
University of Bristol

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro and Toby Perrett, Will Price, Michael Wray (2021). The EPIC-KITCHENS Dataset: Collection, Challenges and Baselines. PAMI, 43(11), pp. 4125-4141.

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

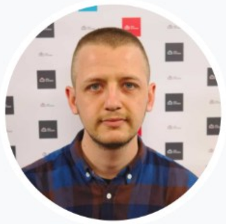
EPIC-KITCHENS-100

Dima Damen
University of Bristol

Hazel Doughty
University of Bristol

Giovanni M. Farinella
University of Catania

Antonino Furnari
University of Catania



Evangelos Kazakos
University of Bristol

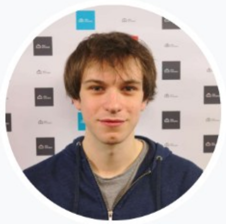
Jian Ma
University of Bristol

Davide Moltisanti
University of Bristol

Jonathan Munro
University of Bristol

Toby Perrett
University of Bristol

Will Price
University of Bristol



Michael Wray
University of Bristol

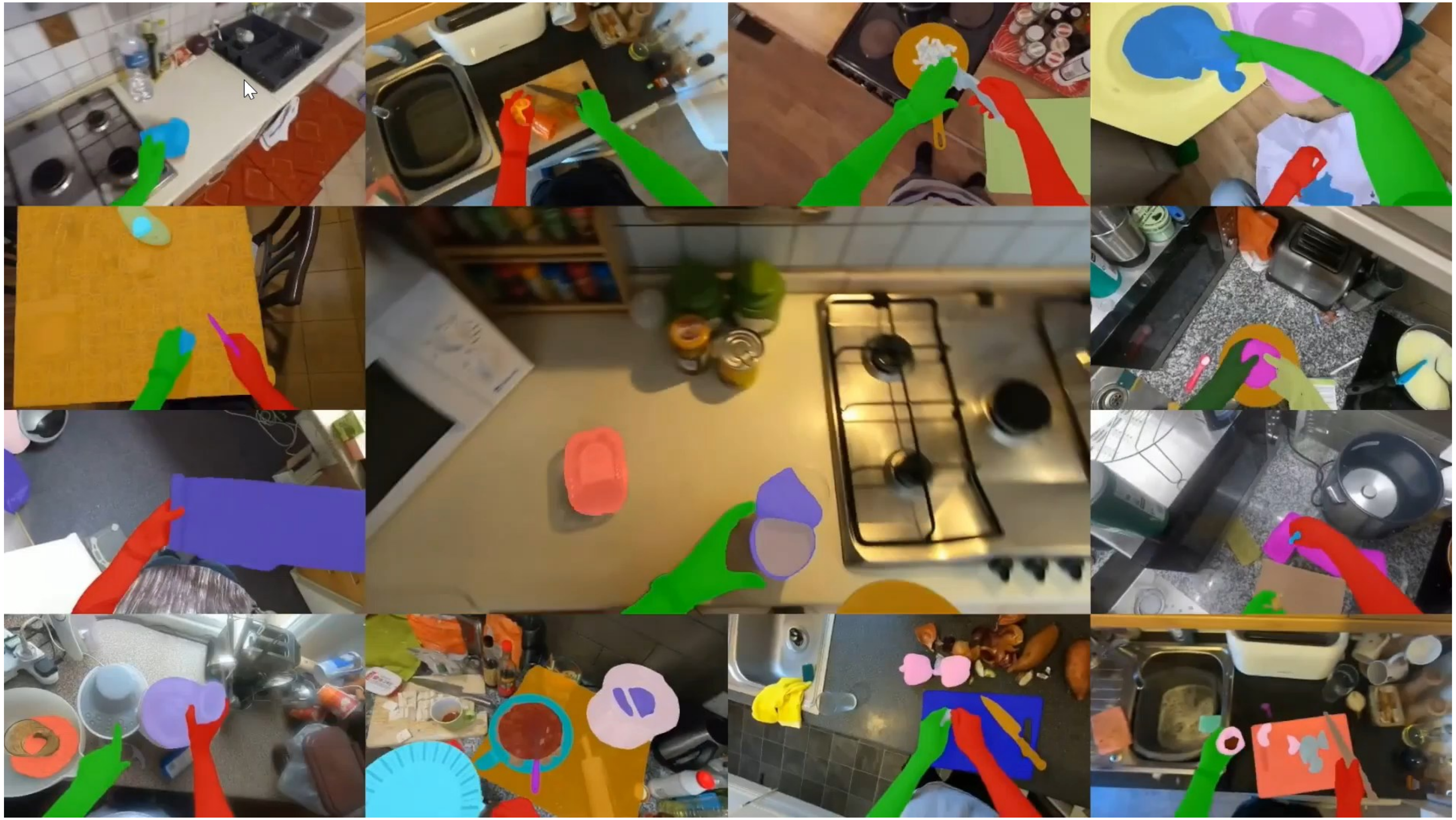
Dima Damen, Hazel Doughty, Giovanni M. Farinella, Antonino Furnari, Evengelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, Michael Wray (2021). Rescaling Egocentric Vision . International Journal on Computer Vision (IJCV) , abs/2006.13256

Bigger…. Better… Denser…

	EPIC-KITCHENS-55	EPIC-KITCHENS-100
No. of Hours	55	100
No. of Kitchens	32	45
No. of Videos	432	700
No. of Action Segments	39,432	89,979
Action Classes	2,747	4,025
Verb Classes	125	97
Noun Classes	331	300
Splits	Train/Test	Train/Val/Test
No. of Challenges	3	6 (4 new challenges)

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

EPIC-Kitchens VISOR



- 272K manual sparse masks for hands and active objects;
- Hand-object contact relations;
- 1477 unique entities;
- 22 categories.

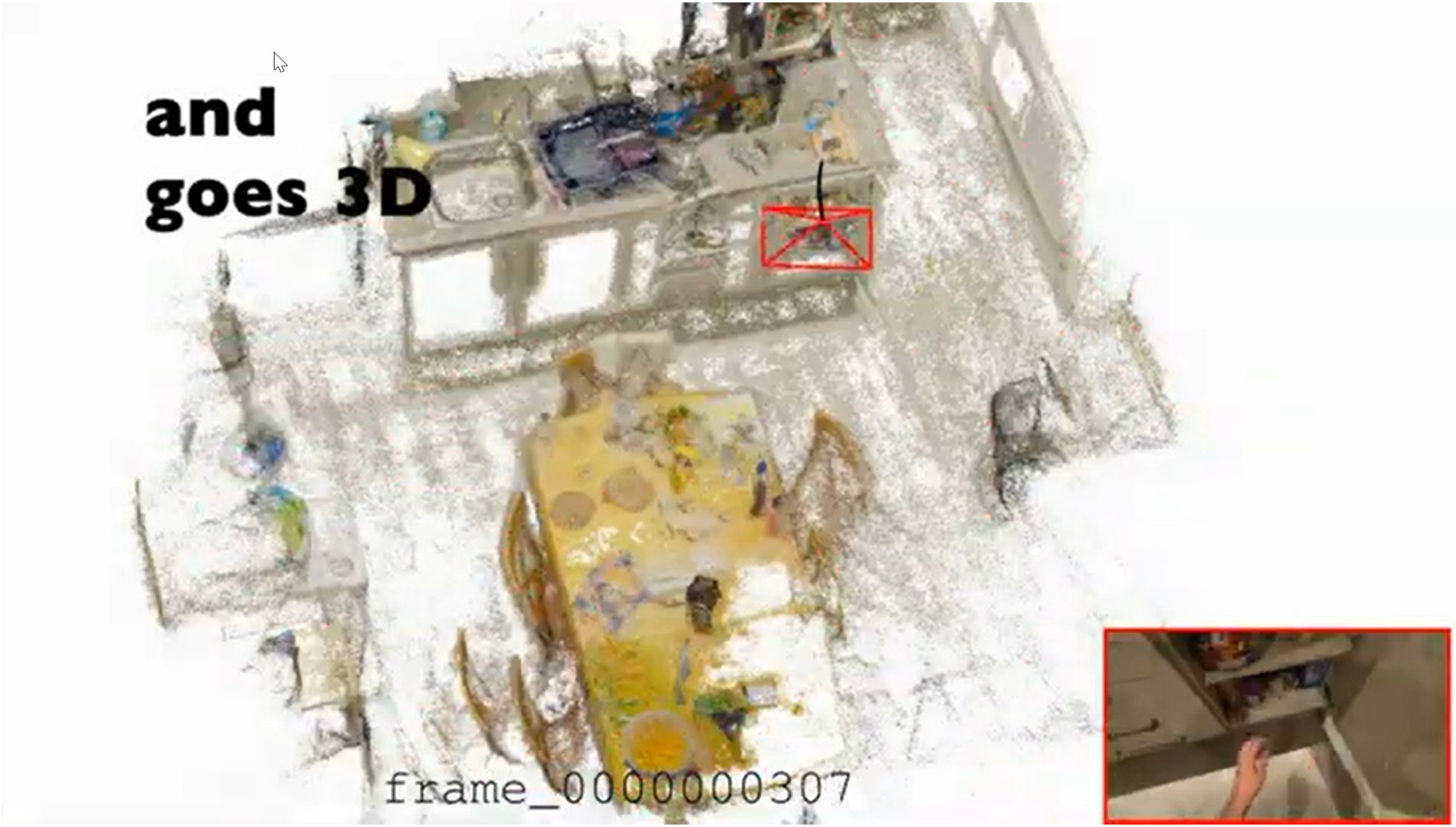
https://epic-kitchens.github.io/VISOR

EPIC-SOUNDS

- 74.8K categorised audio segments;
- Material-based collision sounds;
- Repetitive sounds;
- 44 classes.

https://epic-kitchens.github.io/epic-sounds

EPIC-FIELDS



- 19M registered frames;
- Camera poses;
- 3D reconstruction;
- Paired with VISOR annotations.

https://epic-kitchens.github.io/epic-fields

- Recipe and Nutrition;
- Preparation and Step;
- Narrations;
- Audio Annotations;
- Digital Twin;
- Gaze Priming;

https://hd-epic.github.io

2025 Challenges

- Semi-Supervised Video Object Segmentation Challenge
- EPIC-SOUNDS Audio-Based Interaction Recognition
- EPIC-SOUNDS Audio-Based Interaction Recognition
- Action Recognition
- Action Detection
- UDA for Action Recognition
- Multi-Instance Retrieval

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

EPIC-KITCHENS Workshops & Challenges

EPIC-KITCHENS-100- 2022 Challenges Report

RESULTS - 2024 CHALLENGES (JUNE 2024)

EPIC-Kitchens Challenges @CVPR2024, Seattle, US

2024 CHALLENGE WINNERS

		Team	Member	Affiliations
Action Recognition	1	KAUST-4Paradigm -MoonshotAI-Nvidia	Shuming Liu	King Abdullah University of Science and Technology
			Lin Sui	4Paradigm Inc
			Chen-Lin Zhang	Moonshot AI
			Fangzhou Mu	NVIDIA
			Chen Zhao	King Abdullah University of Science and Technology
			Bernard Ghanem	King Abdullah University of Science and Technology
	2	DeepGlint (dg_team)	Yingxin Xia	DeepGlint and Harbin Institute of Technology
			Ninghua Yang	DeepGlint
			Kaicheng Yang	DeepGlint
			Xiang An	DeepGlint
			Xiangzi Dai	DeepGlint
			Weimo Deng	DeepGlint
			Ziyong Feng	DeepGlint
	3	Shanghai AI Laboratory (Ayiyayi)	Baoqi Pei	Shanghai AI Laboratory and Zhejiang University
			Yifei Huang	Shanghai AI Laboratory
			Guo Chen	Shanghai AI Laboratory and Nanjing University
			Jilan Xu	Shanghai AI Laboratory and Fudan University
			Yicheng Liu	Nanjing University
			Yuping He	Nanjing University
			Kanghua Pan	Nanjing University
			Tong Lu	Nanjing University
			Limin Wang	Shanghai AI Laboratory
			Yali Wang	Shanghai AI Laboratory
			Yu Qiao	Shanghai AI Laboratory

First Joint Egocentric Vision (EgoVis) Workshop

Held in Conjunction with CVPR 2024

17 June 2024 - Seattle, USA

Room: Summit 428


THAT'S JUST IMPRESSIVE

Can We Scale?

EGO 4D

Consortium

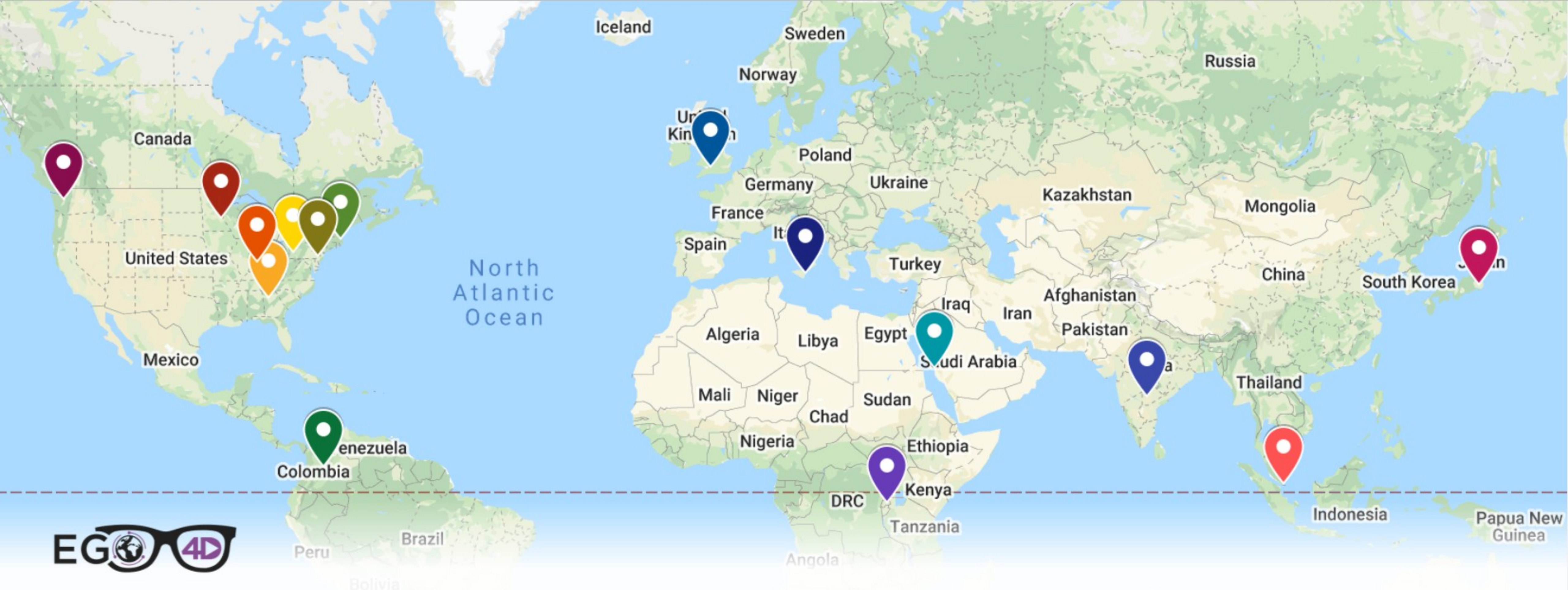
Carnegie Mellon University · Università di Catania · NUS National University of Singapore · King Abdullah University of Science and Technology · 東京大学 The University of Tokyo · University of Bristol · Indiana University Bloomington · University of Minnesota · MIT · Penn University of Pennsylvania · Georgia Institute of Technology · International Institute of Information Technology Hyderabad · Carnegie Mellon University Africa · Universidad de los Andes Colombia · Facebook AI

Ego4D: Around the World in 3,000 Hours of Egocentric Video

84 authors

Kristen Grauman[1,2], Andrew Westbury[1], Eugene Byrne*[1], Zachary Chavis*[3], Antonino Furnari*[4], Rohit Girdhar*[1], Jackson Hamburger*[1], Hao Jiang*[5], Miao Liu*[6], Xingyu Liu*[7], Miguel Martin*[1], Tushar Nagarajan*[1,2], Ilija Radosavovic*[8], Santhosh Kumar Ramakrishnan*[1,2], Fiona Ryan*[6], Jayant Sharma*[3], Michael Wray*[9], Mengmeng Xu*[10], Eric Zhongcong Xu*[11], Chen Zhao*[10], Siddhant Bansal[17], Dhruv Batra[1], Vincent Cartillier[1,6], Sean Crane[7], Tien Do[3], Morrie Doulaty[13], Akshay Erapalli[13], Christoph Feichtenhofer[1], Adriano Fragomeni[9], Qichen Fu[7], Christian Fuegen[13], Abrham Gebreselasie[12], Cristina González[14], James Hillis[5], Xuhua Huang[7], Yifei Huang[15], Wenqi Jia[6], Weslie Khoo[16], Jachym Kolar[13], Satwik Kottur[13], Anurag Kumar[5], Federico Landini[13], Chao Li[5], Zhenqiang Li[15], Karttikeya Mangalam[1,8], Raghava Modhugu[17], Jonathan Munro[9], Tullie Murrell[1], Takumi Nishiyasu[15], Will Price[9], Paola Ruiz Puentes[14], Merey Ramazanova[10], Leda Sari[5], Kiran Somasundaram[5], Audrey Southerland[6], Yusuke Sugano[15], Ruijie Tao[11], Minh Vo[5], Yuchen Wang[16], Xindi Wu[7], Takuma Yagi[15], Yunyi Zhu[11], Pablo Arbeláez†[14], David Crandall†[16], Dima Damen†[9], Giovanni Maria Farinella†[4], Bernard Ghanem†[10], Vamsi Krishna Ithapu†[5], C. V. Jawahar†[17], Hanbyul Joo†[1], Kris Kitani†[7], Haizhou Li†[11], Richard Newcombe†[5], Aude Oliva†[18], Hyun Soo Park†[3], James M. Rehg†[6], Yoichi Sato†[15], Jianbo Shi†[19], Mike Zheng Shou†[11], Antonio Torralba†[18], Lorenzo Torresani†[1,20], Mingfei Yan†[5], Jitendra Malik[1,8]

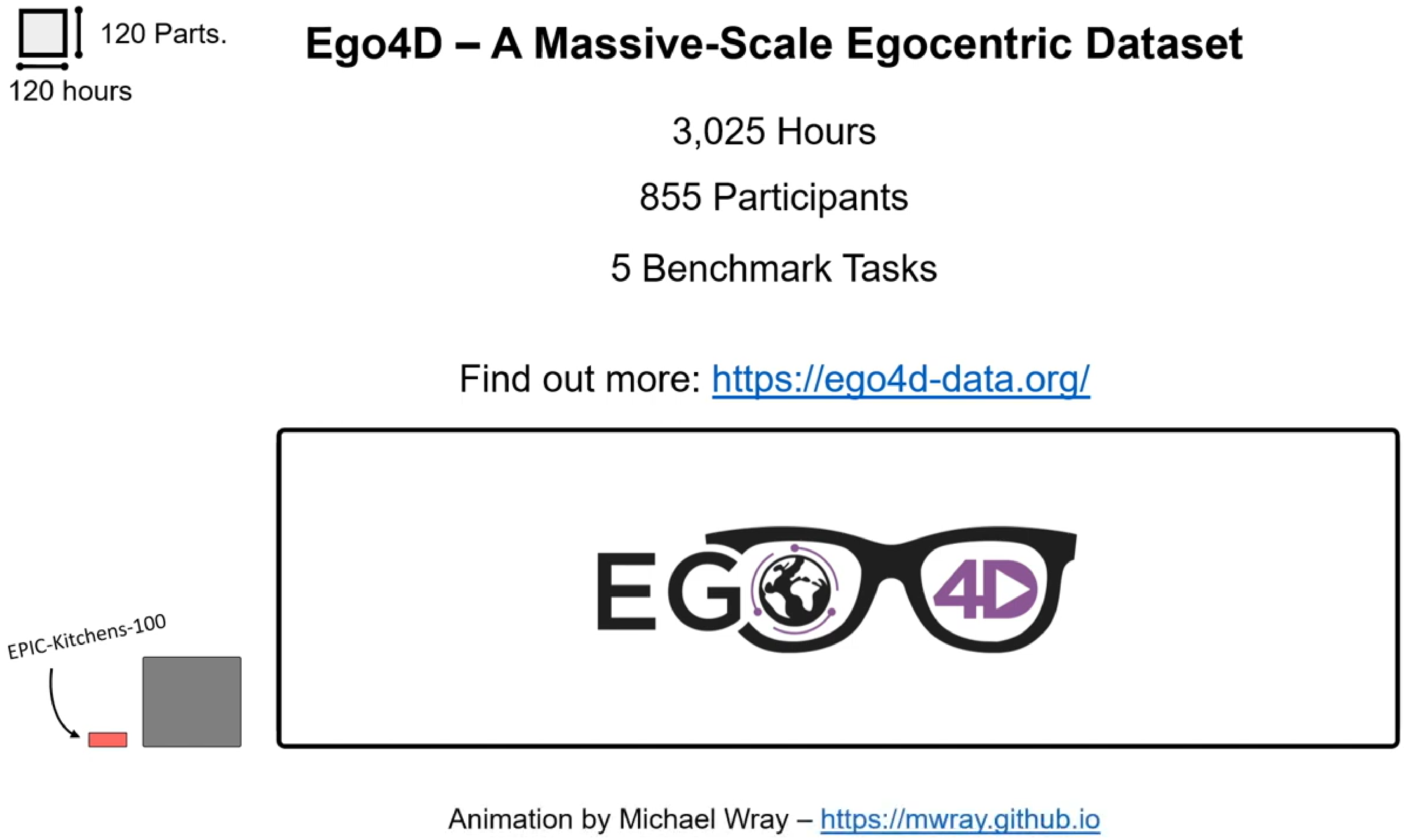
[1]Facebook AI Research (FAIR), [2]University of Texas at Austin, [3]University of Minnesota, [4]University of Catania, [5]Facebook Reality Labs, [6]Georgia Tech, [7]Carnegie Mellon University, [8]UC Berkeley, [9]University of Bristol, [10]King Abdullah University of Science and Technology, [11]National University of Singapore, [12]Carnegie Mellon University Africa, [13]Facebook, [14]Universidad de los Andes, [15]University of Tokyo, [16]Indiana University, [17]International Institute of Information Technology, Hyderabad, [18]MIT, [19]University of Pennsylvania, [20]Dartmouth



A massive-scale, egocentric dataset and benchmark suite collected across 74 worldwide locations and 9 countries, with over 3,025 hours of daily-life activity video.

855 Subjects | 74 Locations | 9 Countries | 3025 Hours | 3D Scans | Audio | Gaze

EGO4D – Massive Scale



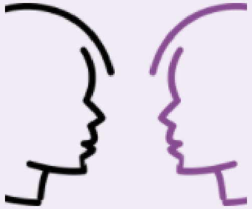
Animation by Michael Wray - https://www.youtube.com/watch?v=p78-V2RiKo

Benchmarks and Challenges

Episodic Memory

Hand-Object Interactions

AV Diarization



Social

Forecasting

1st Ego4D Workshop @ CVPR 2022

Held in conjunction with 10th EPIC Workshop

19 and 20 June 2022

2nd International Ego4D Workshop @ ECCV 2022

24 October 2022

3rd International Ego4D Workshop @ CVPR 2023

Held in conjunction with 11th EPIC Workshop

19 June 2023

First Joint Egocentric Vision (EgoVis) Workshop

Held in Conjunction with CVPR 2024

17 June 2024 - Seattle, USA

Room: Summit 428

Happy Ending?

EGO-EXO4D

EGO-EXO4D Team

THE UNIVERSITY OF TEXAS AT AUSTIN

Universidad de los Andes Colombia

SFU SIMON FRASER UNIVERSITY

Carnegie Mellon University

Meta

Università di Catania

Carnegie Mellon University Africa

NUS National University of Singapore

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

東京大学 THE UNIVERSITY OF TOKYO

جامعة الملك عبدالله للعلوم والتقنية

King Abdullah University of Science and Technology

UNIVERSITY OF MINNESOTA

Penn UNIVERSITY of PENNSYLVANIA

University of BRISTOL

INDIANA UNIVERSITY BLOOMINGTON

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

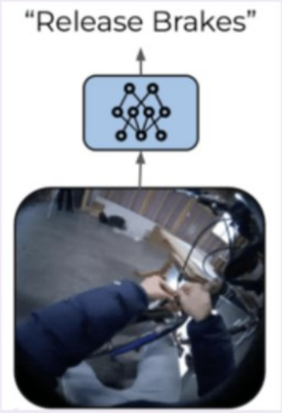
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

https://ego-exo4d-data.org/

Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives

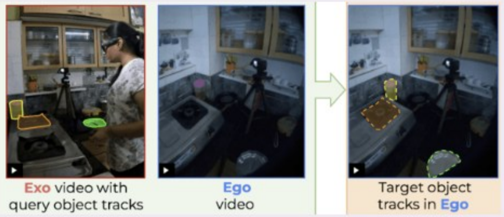
Kristen Grauman[1,2], Andrew Westbury[1], Lorenzo Torresani[1], Kris Kitani[1,3], Jitendra Malik[1,4], Triantafyllos Afouras[*1], Kumar Ashutosh[*1,2], Vijay Baiyya[*5], Siddhant Bansal[*6,7], Bikram Boote[*8], Eugene Byrne[*1,9], Zach Chavis[*10], Joya Chen[*11], Feng Cheng[*1], Fu-Jen Chu[*1], Sean Crane[*9], Avijit Dasgupta[*7], Jing Dong[*5], Maria Escobar[*12], Cristhian Forigua[*12], Abrham Gebreselasie[*9], Sanjay Haresh[*13], Jing Huang[*1], Md Mohaiminul Islam[*14], Suyog Jain[*1], Rawal Khirodkar[*9], Devansh Kukreja[*1], Kevin J Liang[*1], Jia-Wei Liu[*11], Sagnik Majumder[*1,2], Yongsen Mao[*13], Miguel Martin[*1], Effrosyni Mavroudi[*1], Tushar Nagarajan[*1], Francesco Ragusa[*15], Santhosh Kumar Ramakrishnan[*2], Luigi Seminara[*15], Arjun Somayazulu[*2], Yale Song[*1], Shan Su[*16], Zihui Xue[*1,2], Edward Zhang[*16], Jinxu Zhang[*16], Angela Castillo[12], Changan Chen[2], Xinzhu Fu[11], Ryosuke Furuta[17], Cristina González[12], Prince Gupta[5], Jiabo Hu[18], Yifei Huang[17], Yiming Huang[16], Weslie Khoo[19], Anush Kumar[10], Robert Kuo[18], Sach Lakhavani[5], Miao Liu[18], Mi Luo[2], Zhengyi Luo[3], Brighid Meredith[18], Austin Miller[18], Oluwatumininu Oguntola[14], Xiaqing Pan[5], Penny Peng[18], Shraman Pramanick[20], Merey Ramazanova[21], Fiona Ryan[22], Wei Shan[14], Kiran Somasundaram[5], Chenan Song[11], Audrey Southerland[22], Masatoshi Tateno[17], Huiyu Wang[1], Yuchen Wang[19], Takuma Yagi[17], Mingfei Yan[5], Xitong Yang[1], Zecheng Yu[17], Shengxin Cindy Zha[18], Chen Zhao[21], Ziwei Zhao[19], Zhifan Zhu[6], Jeff Zhuo[14], Pablo Arbeláez[†12], Gedas Bertasius[†14], David Crandall[†19], Dima Damen[†6], Jakob Engel[†5], Giovanni Maria Farinella[†15], Antonino Furnari[†15], Bernard Ghanem[†21], Judy Hoffman[†22], C. V. Jawahar[†7], Richard Newcombe[†5], Hyun Soo Park[†10], James M. Rehg[†8], Yoichi Sato[†17], Manolis Savva[†13], Jianbo Shi[†16], Mike Zheng Shou[†11], and Michael Wray[†6]

Challenges

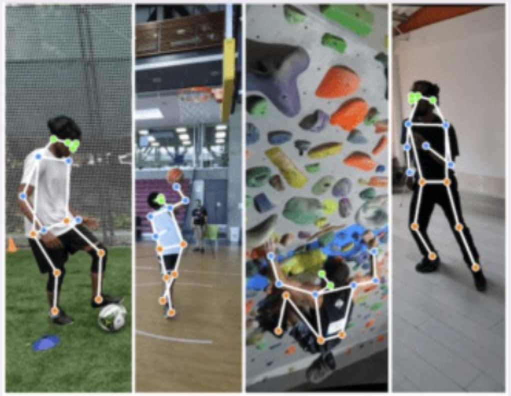


Keystep Recognition

Proficiency Estimation



Relation



Pose Estimation

Second Joint Egocentric Vision (EgoVis) Workshop

Held in Conjunction with CVPR 2025

11 or 12 June 2025 - Nashville, USA

Ego-Exo4D

Ego4D

EPIC-Kitchens

What about Industrial Domain?

UNIVERSITÀ degli STUDI di CATANIA | University of Hertfordshire UH | xenia DOCUMENT MANAGEMENT | Abstract | Dataset | Code | Tasks | Paper | Supp.Material | Acknow. | Related Work | People

The MECCANO Dataset: Understanding Human-Object Interactions from Egocentric Videos in an Industrial-like Domain

F. Ragusa[1,3], A. Furnari[1], S. Livatino[2], G. M. Farinella[1]

[1]IPLab, Department of Mathematics and Computer Science - University of Catania, IT
[2]University of Hertfordshire, Hatfield, Hertfordshire, U.K.
[3]Xenia Gestione Documentale s.r.l. - Xenia Progetti s.r.l., Acicastello, Catania, IT

The new version of MECCANO is available here!

Assembly101: A Large-Scale Multi-View Video Dataset for Understanding Procedural Activities

Fadime Sener[1] Dibyadip Chatterjee[2] Daniel Shelepov[1] Kun He[1]
Dipika Singhania[2] Robert Wang[1] Angela Yao[2]

[1]Reality Labs at Meta
[2]National University of Singapore

CVPR 2022

Paper | Dataset | Code | Sample | Codalab Challenge

IndustReal: A Dataset for Procedure Step Recognition Handling Execution Errors in Egocentric Videos in an Industrial-Like Setting

Tim J. Schoonbeek[1], Tim Houben[1], Hans Onvlee[2], Peter H.N. de With[1], Fons van der Sommen[1],
[1]Eindhoven University of Technology, [2]ASML Research

Published in: WACV 2024

Paper | arXiv | Video | Code | Data | Poster

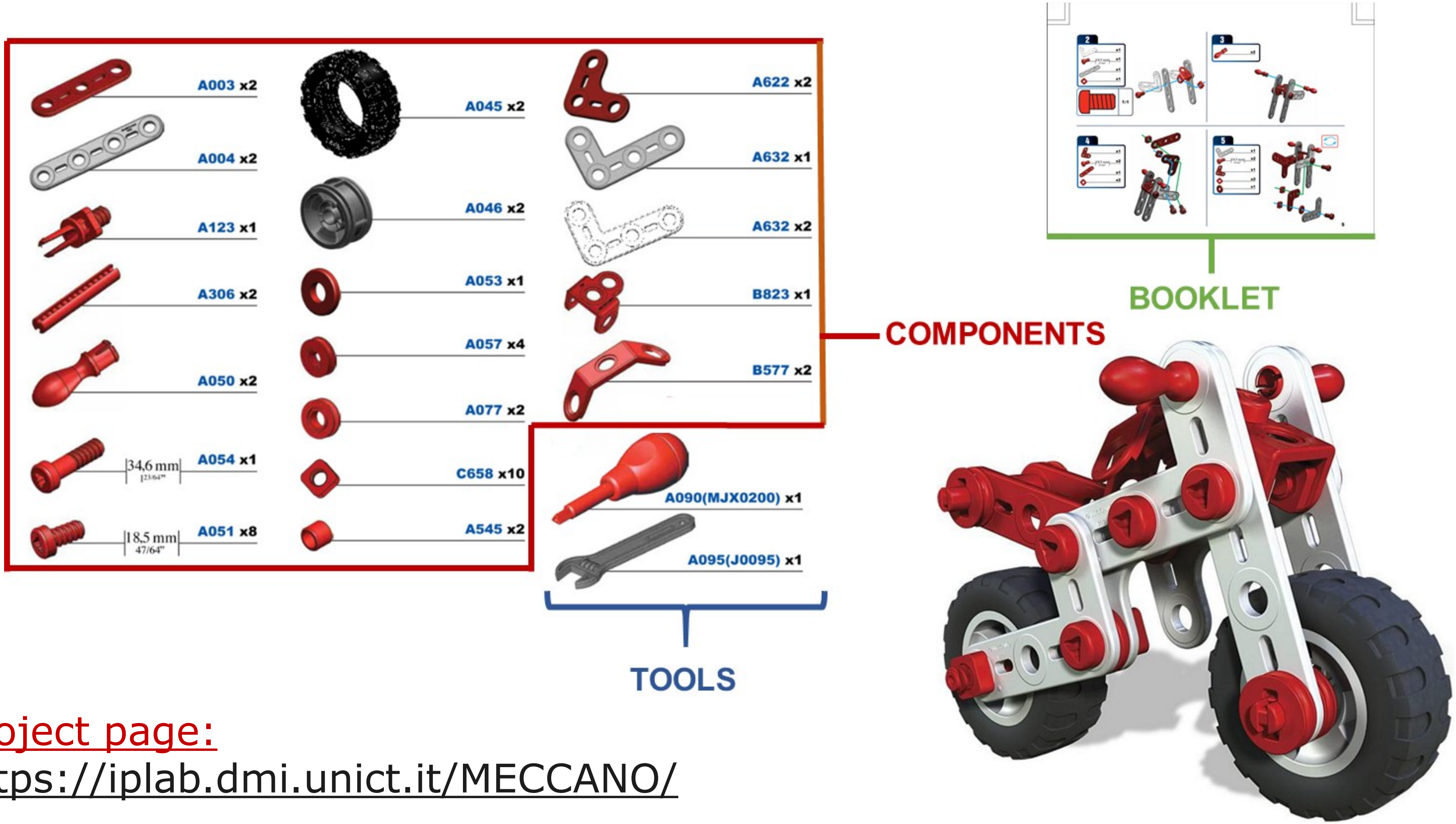
Abstract

ENIGMA-51: Towards a Fine-Grained Understanding of Human Behavior in Industrial Scenarios

ENIGMA-51 is a new egocentric dataset acquired in an industrial scenario by 19 subjects who followed instructions to complete the repair of electrical boards using industrial tools (e.g., electric screwdriver) and equipments (e.g., oscilloscope). The 51 egocentric video sequences are densely annotated with a rich set of labels that enable the systematic study of human behavior in the industrial domain. We provide benchmarks on four tasks related to human behavior: 1) untrimmed temporal detection of human-object interactions, 2) egocentric human-object interaction detection, 3) short-term object interaction anticipation and 4) natural language understanding of intents and entities. Baseline results show that the ENIGMA-51 dataset poses a challenging benchmark to study human behavior in industrial scenarios.

Code | Data

The MECCANO Dataset

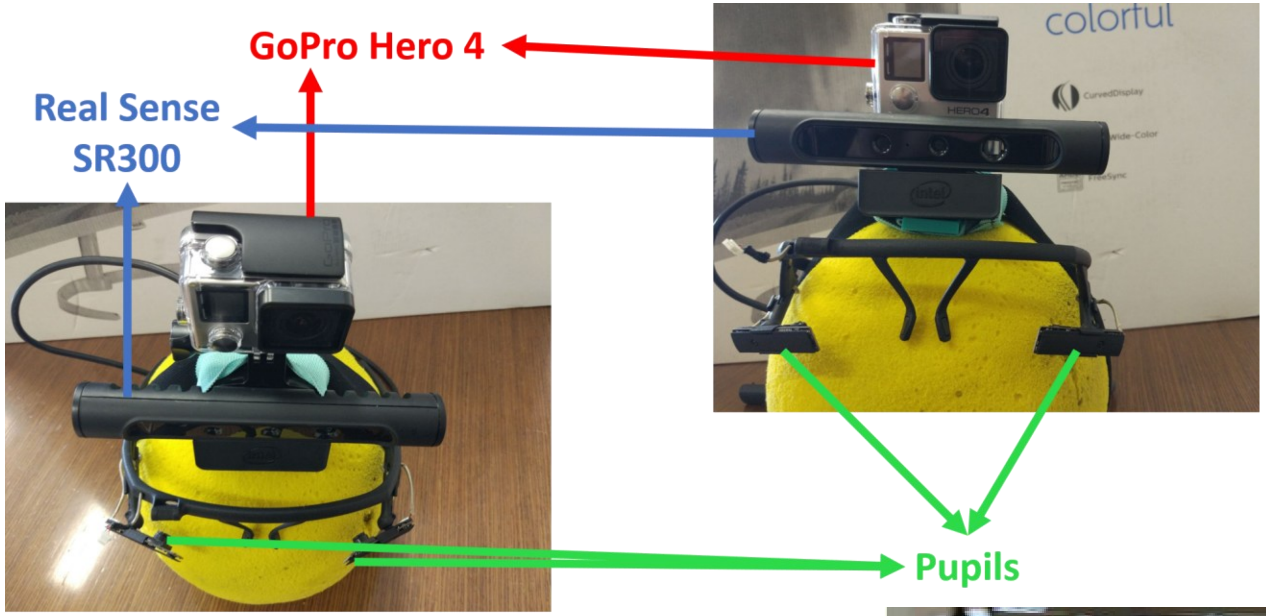


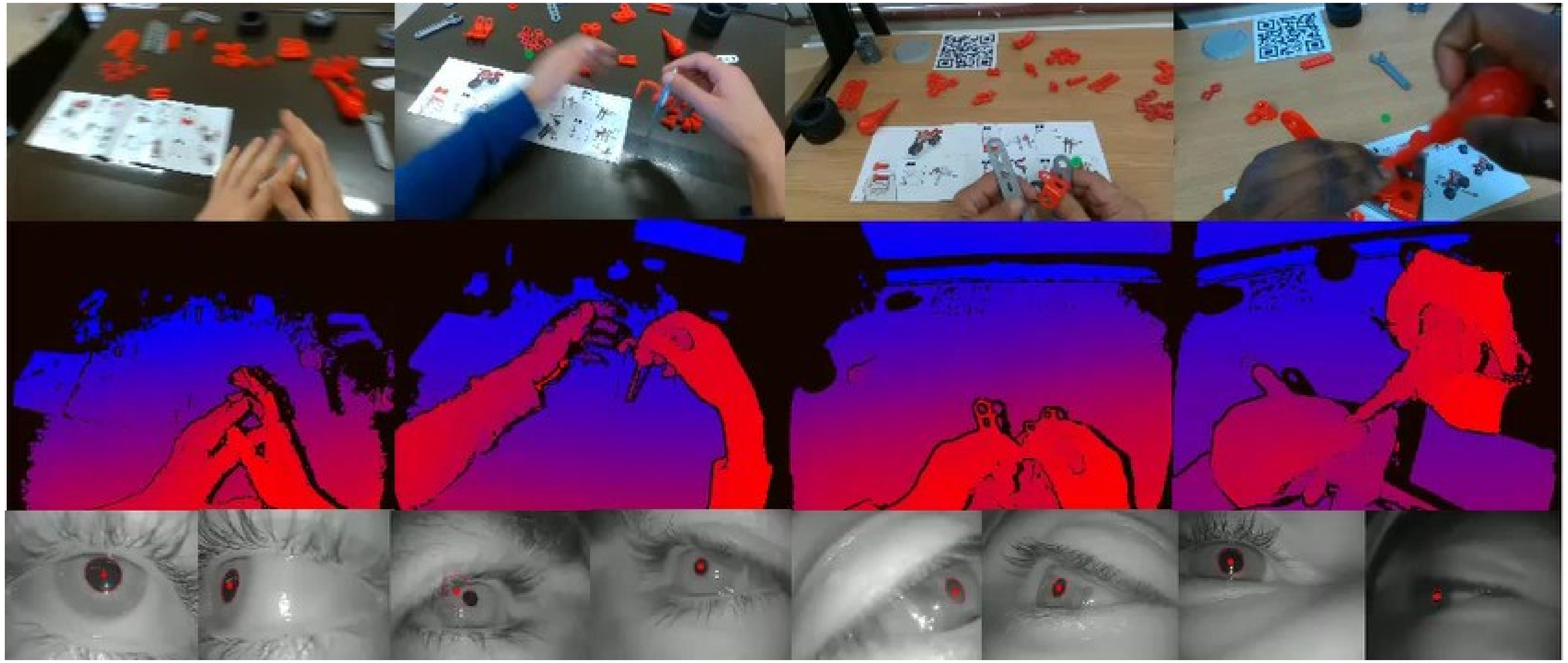
Project page:
https://iplab.dmi.unict.it/MECCANO/

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).

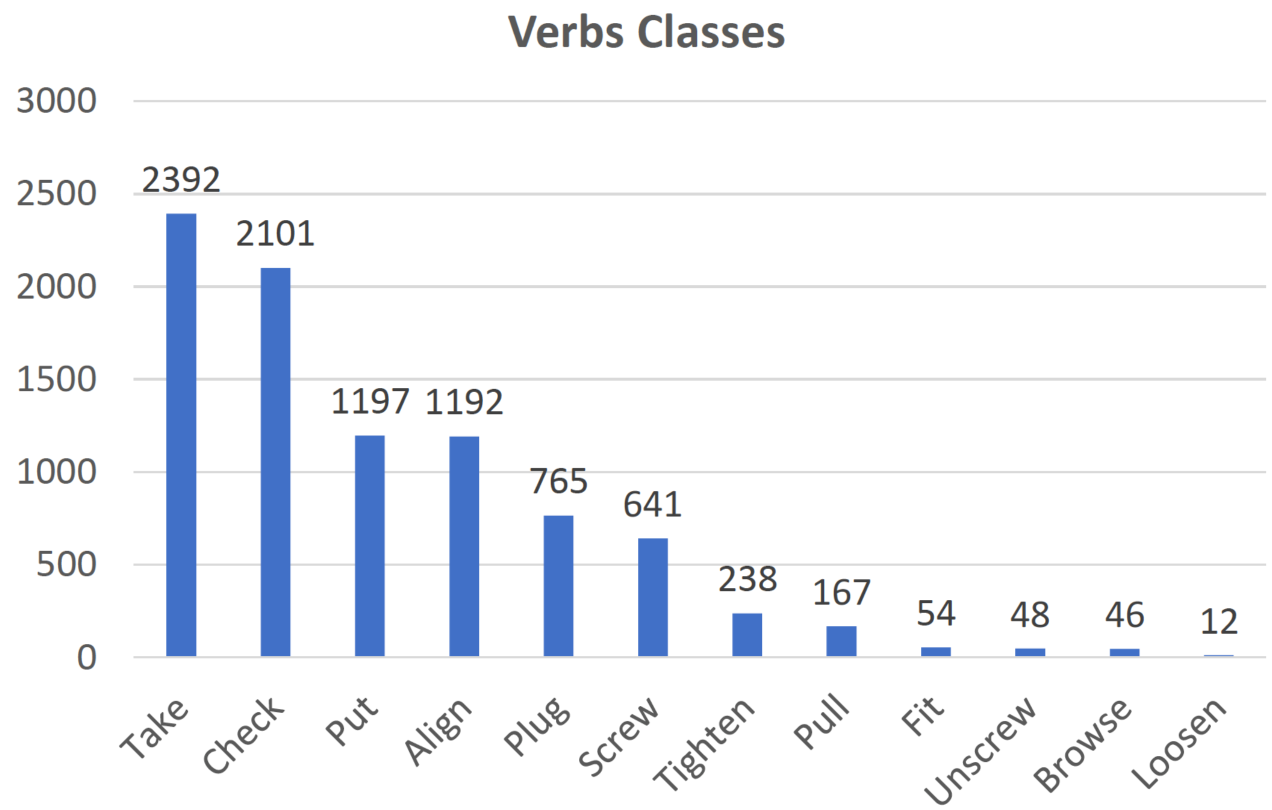
Università di Catania

Data Acquisition



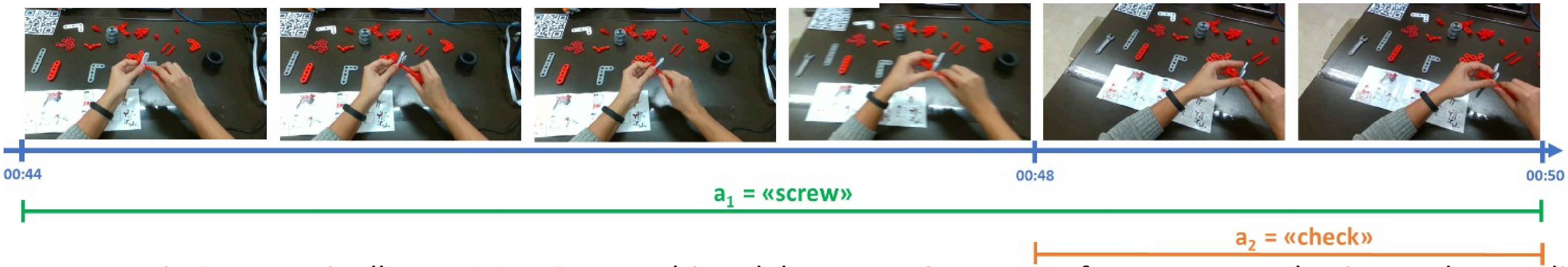


Data Annotation: Temporal Verb Annotations



8857 video segments

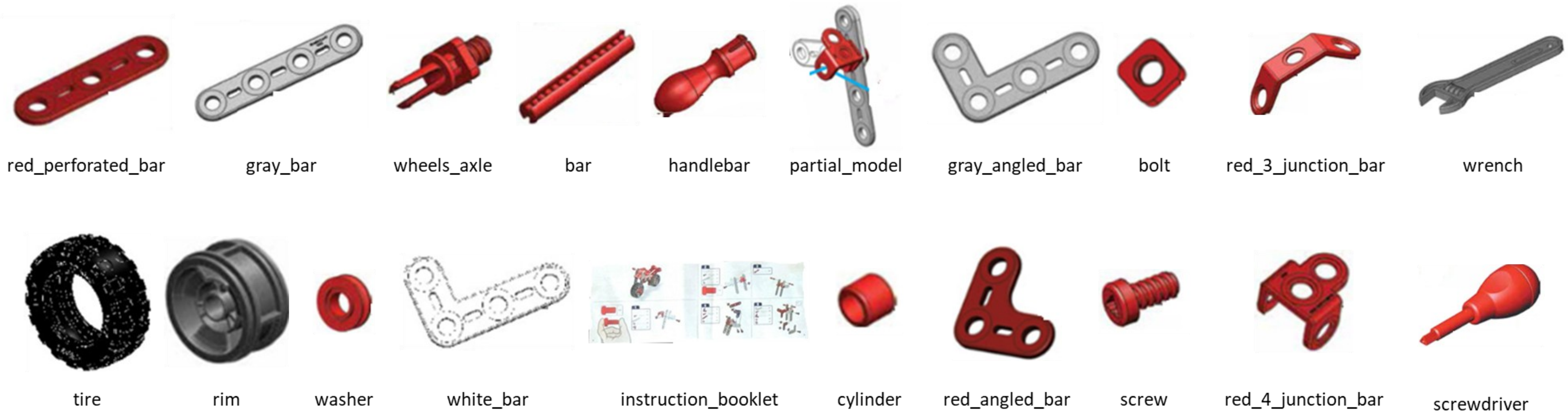
1401 overlap segments (15.82%)

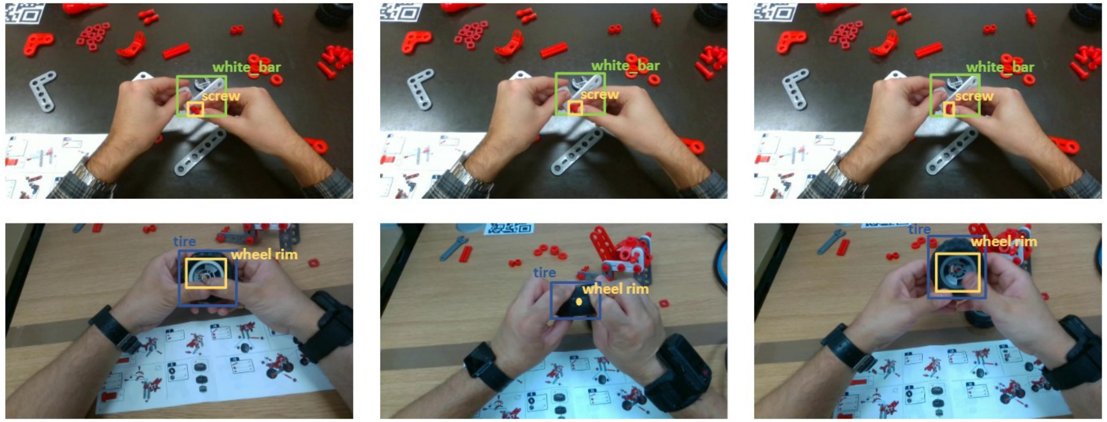


F. Ragusa, A. Furnari, G. M. Farinella. MECCANO: A Multimodal Egocentric Dataset for Humans Behavior Understanding in the Industrial-like Domain. Computer Vision and Image Understanding (CVIU), 2023 (https://arxiv.org/abs/2209.08691).

Università di Catania

Data Annotation: Active Object Bounding Boxes

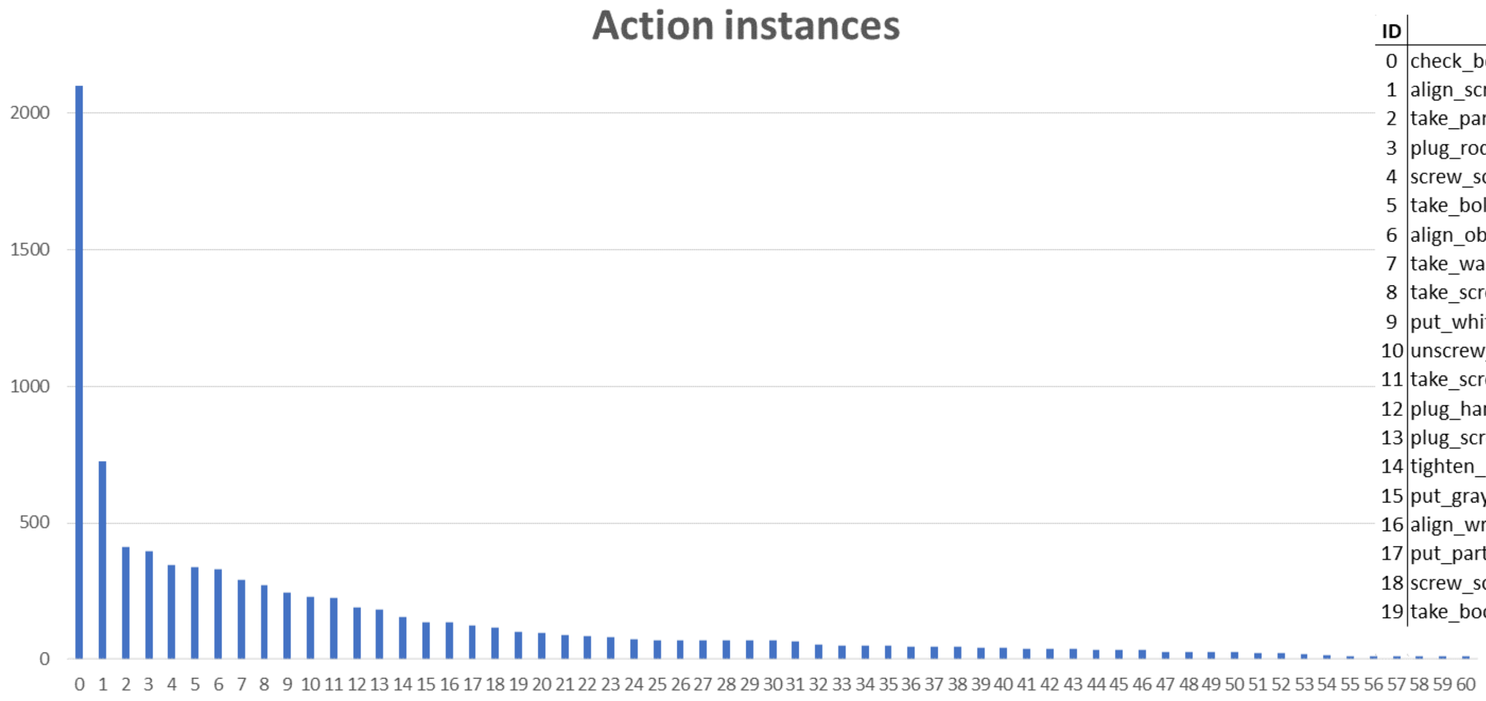




64439 frames

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).

Data Annotation: Action Annotations



ID	Action
0	check_booklet
1	align_screwdriver_to_screw
2	take_partial_model
3	plug_rod
4	screw_screw_with_screwdriver
5	take_bolt
6	align_objects
7	take_washer
8	take_screw
9	put_white_angled_perforated_bar
10	unscrew_screw_with_hands
11	take_screwdriver
12	plug_handlebar
13	plug_screw
14	tighten_nut_with_wrench
15	put_gray_perforated_bar
16	align_wrench_to_bolt
17	put_partial_model
18	screw_screw_with_hands
19	take_booklet

ID	Action
20	put_screwdriver
21	put_red_perforated_junction_bar
22	put_gray_angled_perforated_bar
23	take_red_perforated_bar
24	take_gray_perforated_bar
25	take_red_angled_perforated_bar
26	tighten_nut_with_hands
27	take_white_angled_perforated_bar
28	take_rod
29	put_tire
30	put_roller
31	pull_partial_model
32	pull_screw
33	take_gray_angled_perforated_bar
34	take_tire
35	pull_rod
36	take_wrench
37	browse_booklet
38	take_roller
39	take_handlebar

ID	Action
40	take_red_perforated_junction_bar
41	fit_rim_tire
42	take_rim
43	take_red_4_perforated_junction_bar
44	put_screw
45	put_rod
46	put_washer
47	unscrew_screw_with_screwdriver
48	put_red_perforated_bar
49	put_wrench
50	put_bolt
51	take_wheels_axle
52	put_wheels_axle
53	put_red_angled_perforated_bar
54	put_red_4_perforated_junction_bar
55	take_objects
56	put_objects
57	loosen_bolt_with_hands
58	put_booklet
59	put_rim
60	put_handlebar

align screadriver to screw

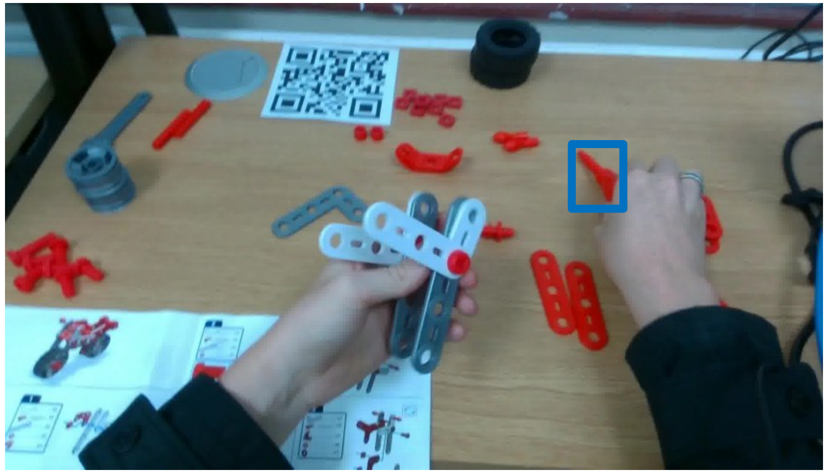
F. Ragusa, A. Furnari, G. M. Farinella. MECCANO: A Multimodal Egocentric Dataset for Humans Behavior Understanding in the Industrial-like Domain. Computer Vision and Image Understanding (CVIU), 2023 (https://arxiv.org/abs/2209.08691).

Egocentric Human-Object Interaction

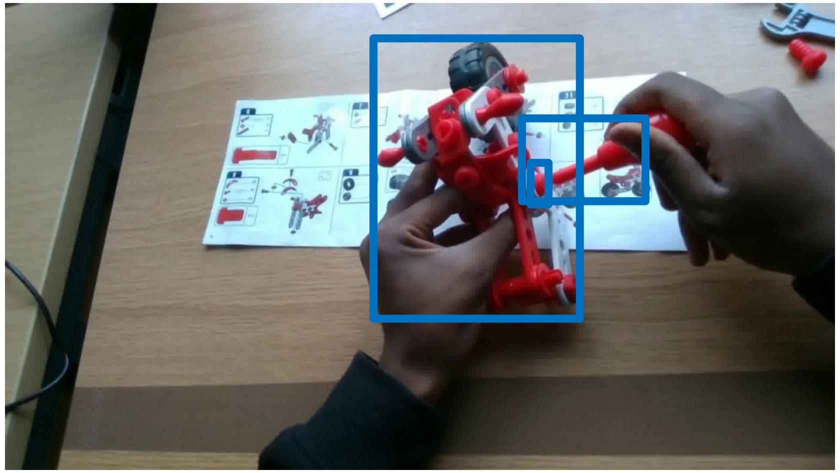
$O = \{o_1, o_2, \ldots, o_n\}$

$V = \{v_1, v_2, \ldots, v_m\}$

$e = (v_h, \{o_1, o_2, \ldots, o_i\})$

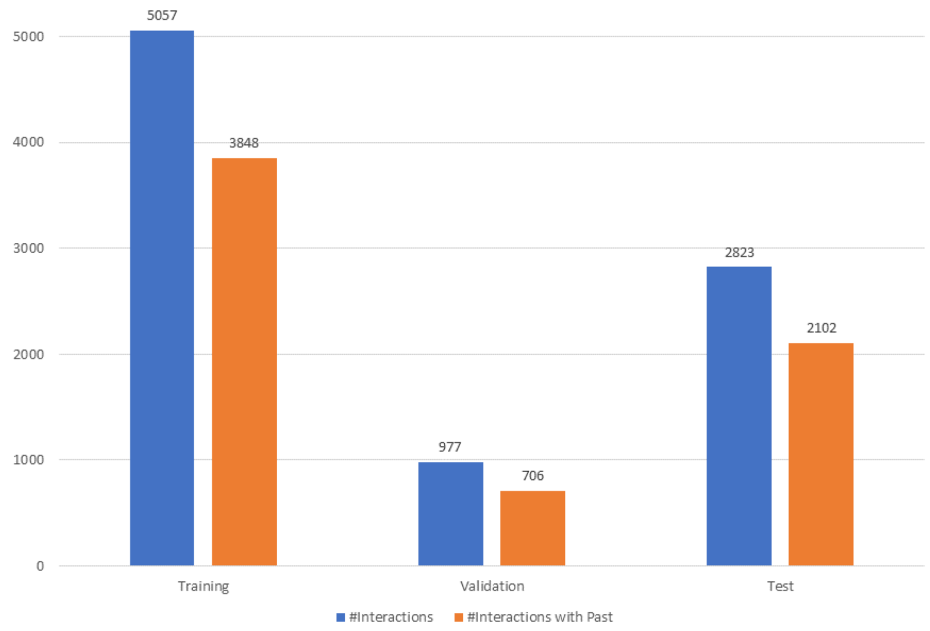


<take, screwdriver>



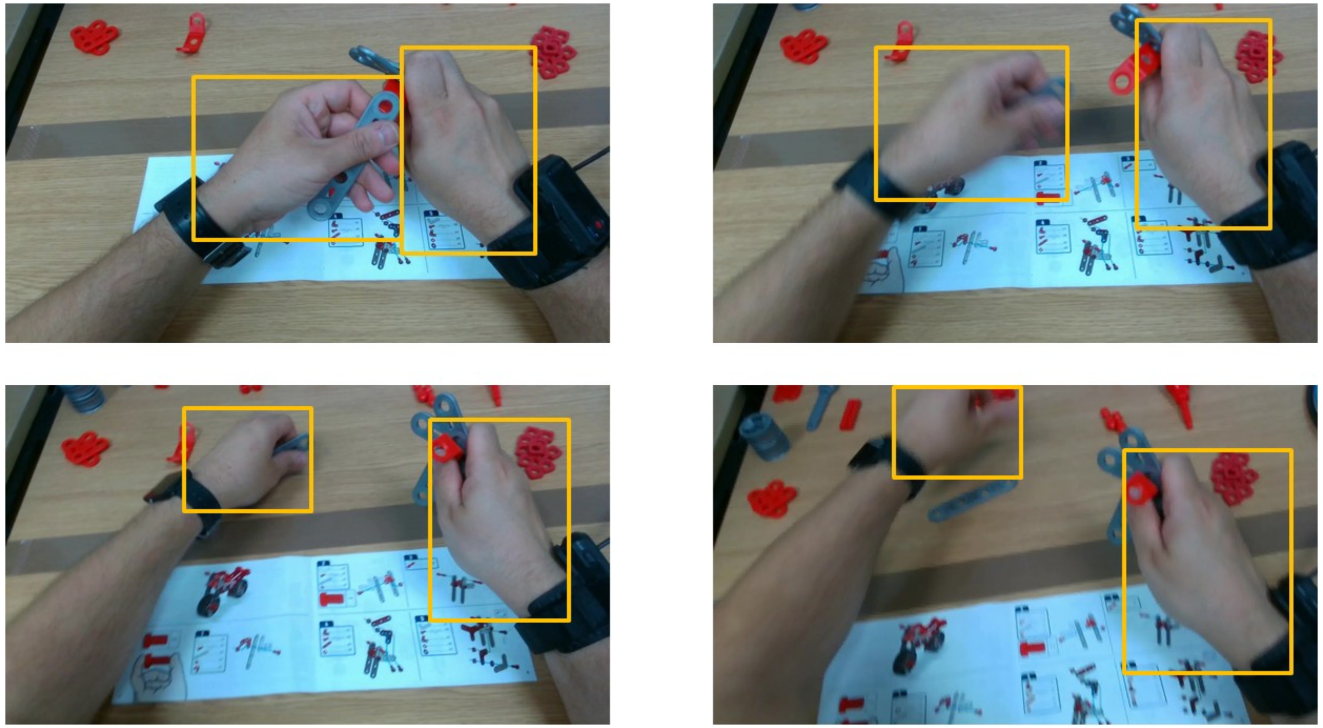
<screw, {screwdriver, screw, partial_model}>

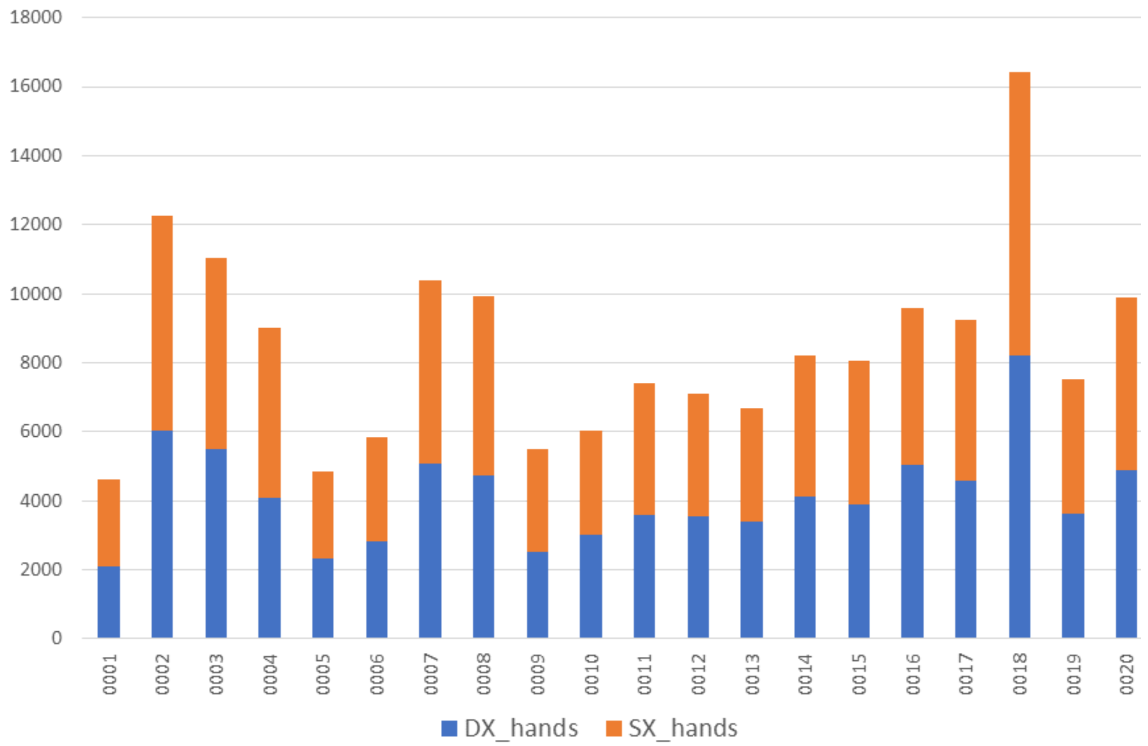
Data Annotation: Next Active Object Annotations



Video	Interactions	Interactions with past
0001	319	257
0002	586	452
0003	573	429
0004	485	372
0005	251	200
0006	307	234
0007	493	367
0008	550	384
0009	289	289
0010	304	194
0011	400	310
0012	384	258
0013	313	244
0014	434	297
0015	425	324
0016	576	436
0017	484	339
0018	788	603
0019	400	294
0020	496	373
Total	8857	6656

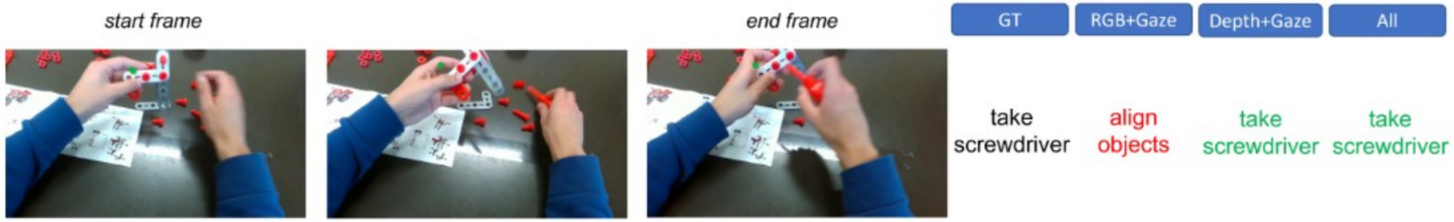
Data Annotation: Hands Annotations



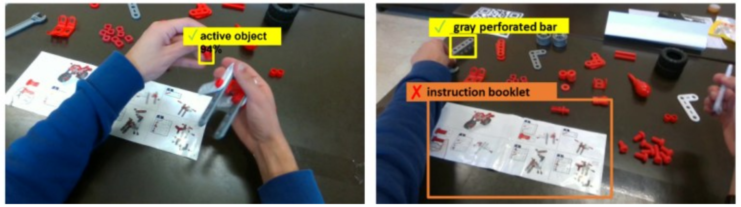


MECCANO Benchmark

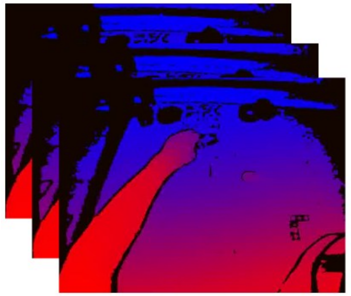
1) Action Recognition

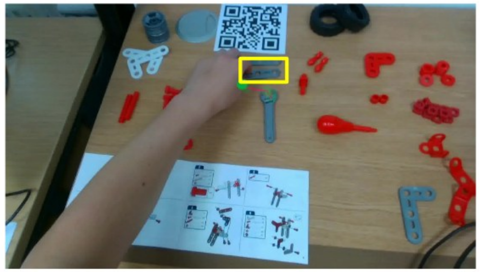


2) Active Object Detection and Recognition



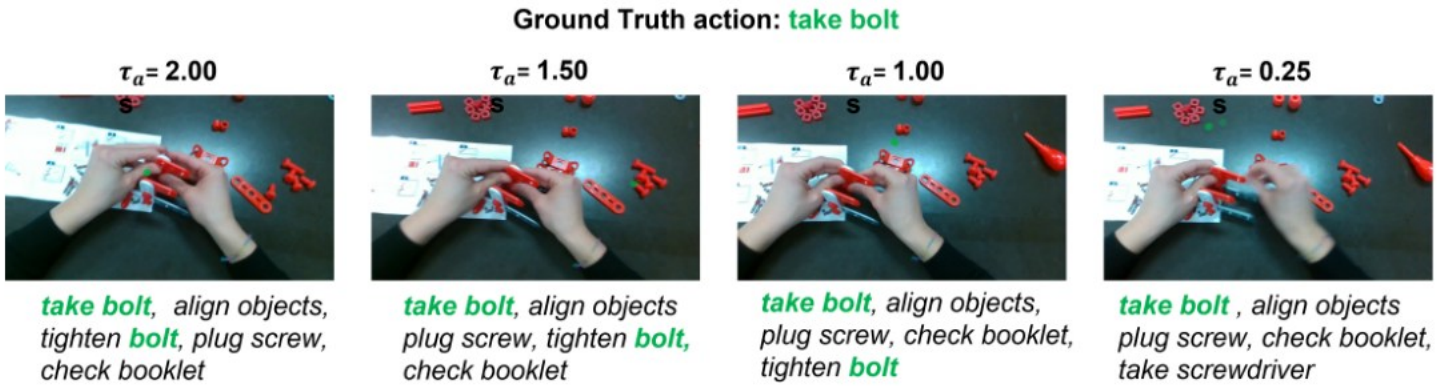
3) EHOI Detection


<take>

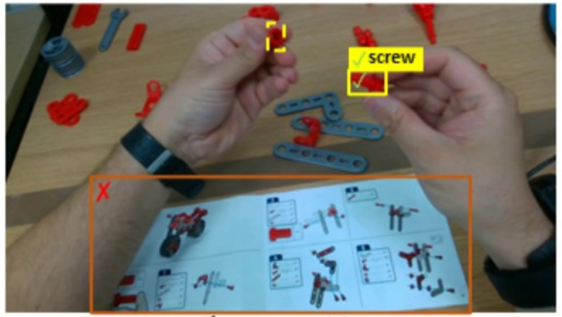

<gray perforated bar>

MECCANO Benchmark

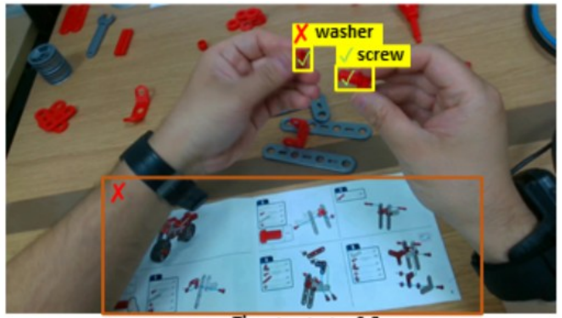
4) Action Anticipation



5) Next-Active Object (NAO) Detection



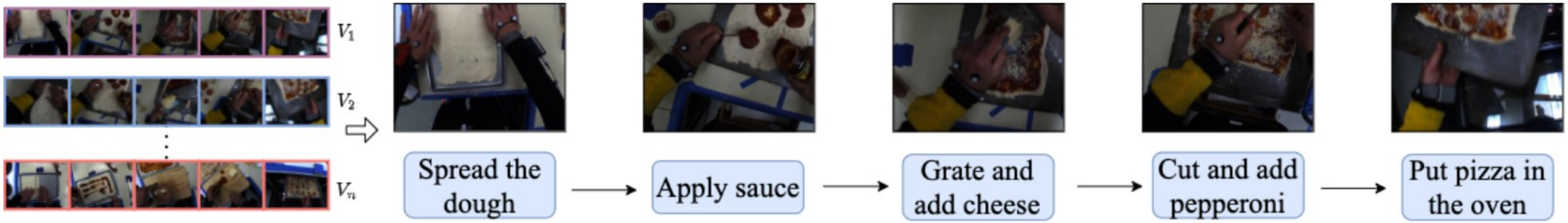
Time to start = 1.6s



Time to start = 0.8s

Procedural Learning

Given multiple videos of a task, the goal is to identify the key-steps and their order to perform the task.



1) EgoProceL (proposed)
2) CMU-MMAC
3) EGTEA Gaze+
4) MECCANO
5) EPIC-Tent

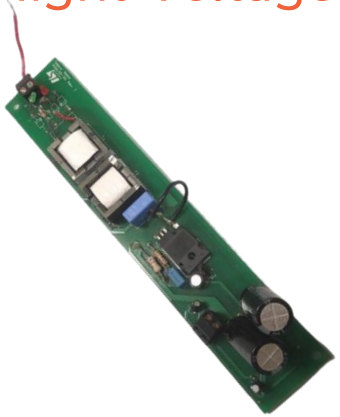
B. Siddhant, A. Chetan, C. V. Jawahar, My View is the Best View: Procedure Learning from Egocentric Videos. In European Conference on Computer Vision (ECCV), 2022.

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

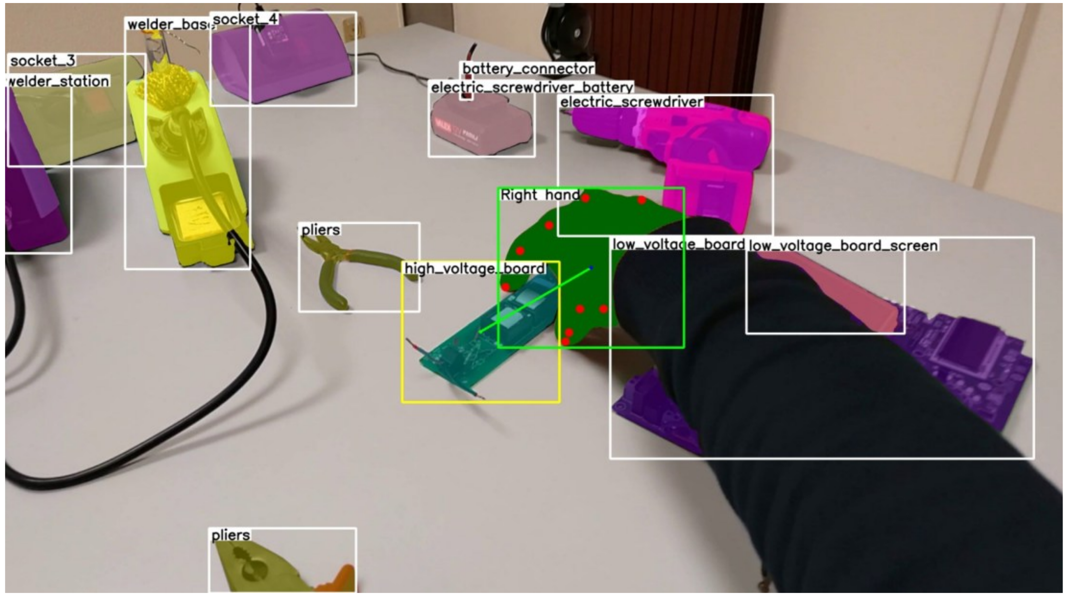
Low-Voltage

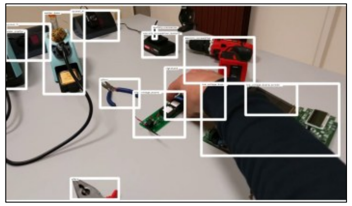
Hight-Voltage



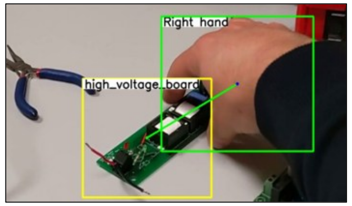
ENIGMA-51: Towards a Fine-Grained Understanding of Human Behavior in Industrial Scenarios. F. Ragusa R. Leonardi, M. Mazzamuto, C. Bonanno, R. Scavo, A. Furnari, G. M. Farinella. WACV (2024).

ENIGMA-51: Annotations

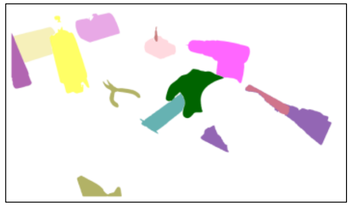




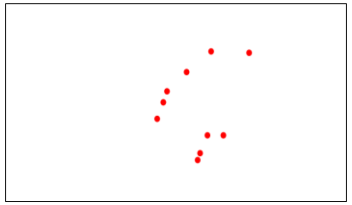
Hand-Object boxes



Human-Object Interactions



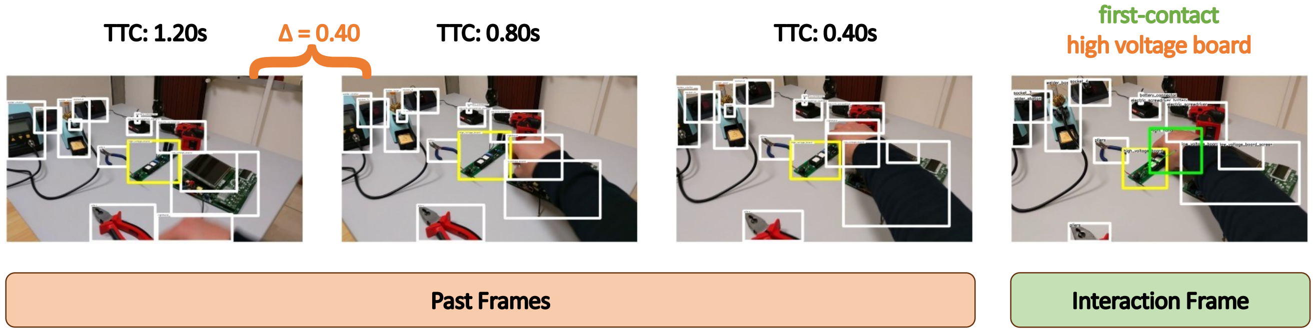
Hand-Object Masks



Hand Keypoints

Environment 3D Model

Object 3D Models



Procedure :

……

4. Take the high voltage board and put it on the working area

5. Take the screwdriver

……

22. Turn on the welder using the switch on the corresponding socket (second from right)

23. Set the temperature of the welder to 480 °C using the yellow “UP” button

……

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).

Untrimmed temporal detection of human-object interactions

Egocentric human-object interaction detection

Short-term object interaction anticipation

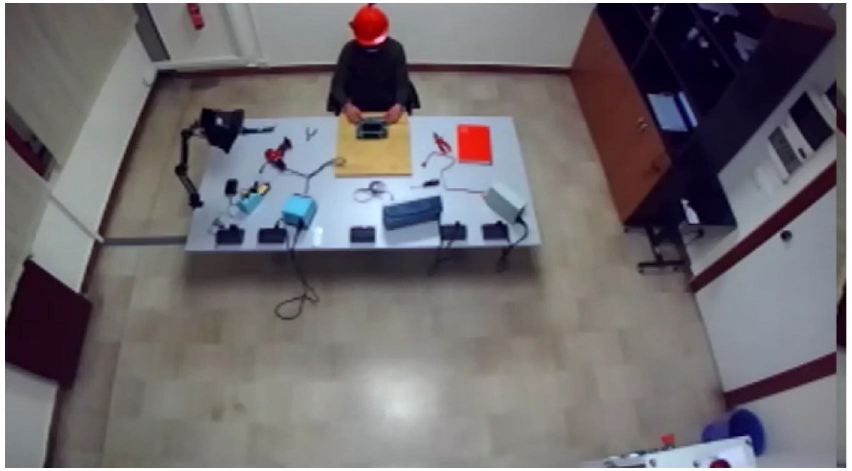
Natural language understanding of intents and entities

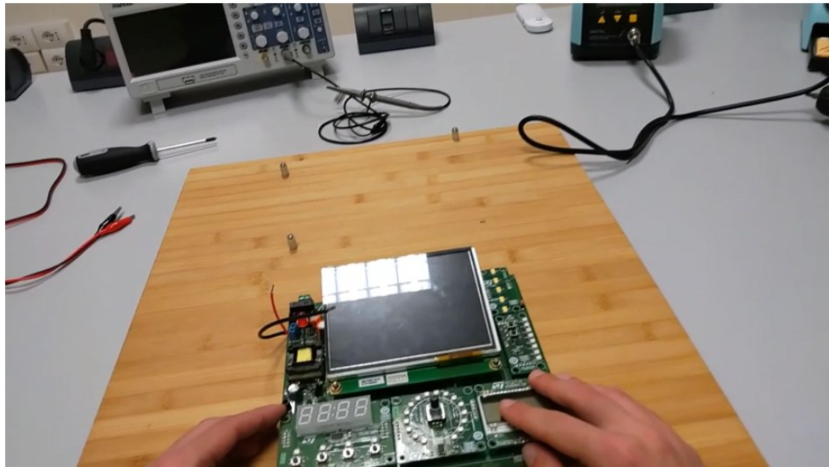
Università di Catania

ENIGMA-360 (Extension)

Exocentric

Egocentric





35 subjects

7 tasks

- Temporal Action Segmentation
- Keystep Recognition
- Hand Object Interaction Segmentation

Conversations are missing

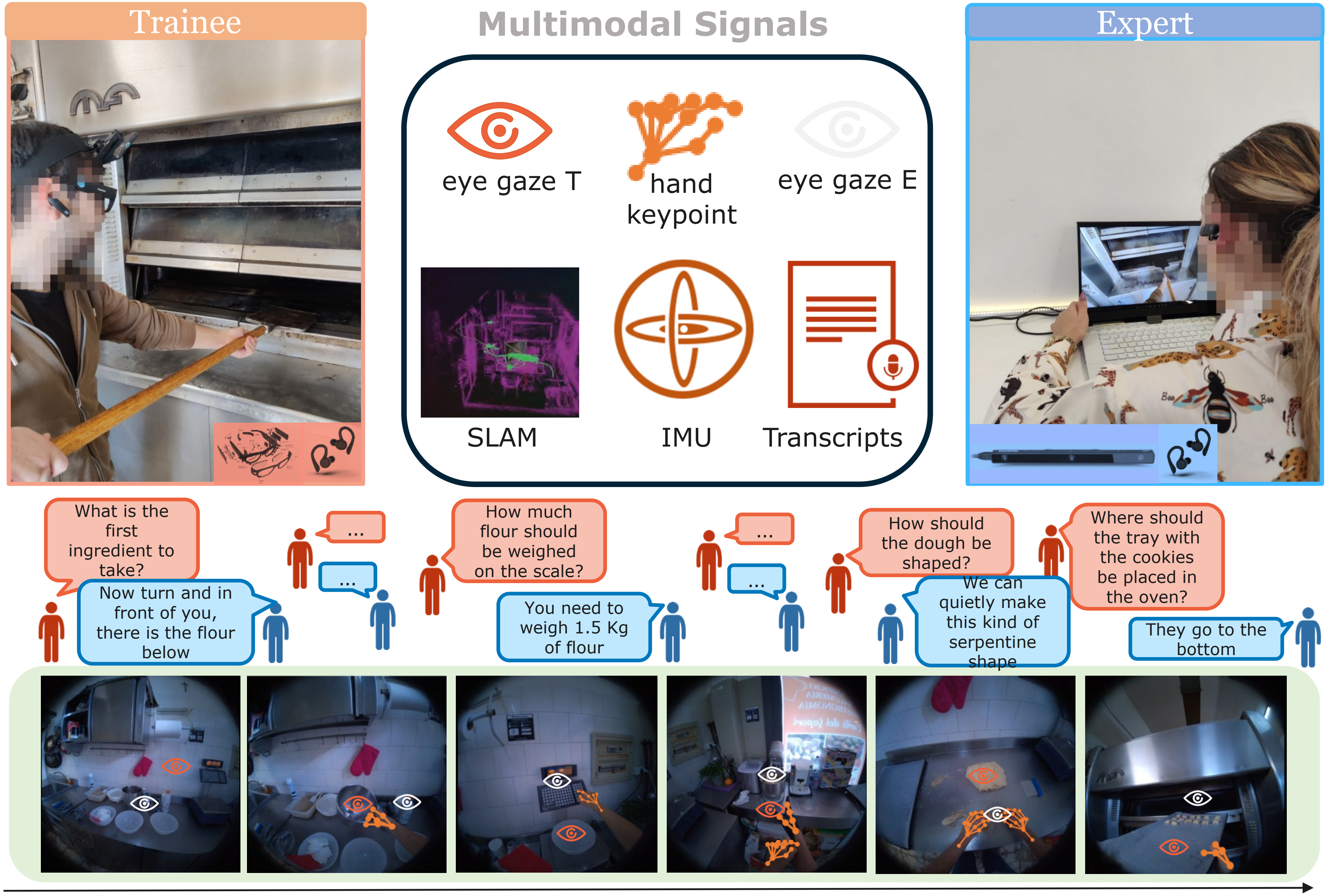
F. Ragusa, M. Mazzamuto, R. Forte, I. D'Ambra, J. Fort, J. Engel, A. Furnari, G. M. Farinella (2026). Ego-EXTRA: video-language Egocentric Dataset for EXpert-TRAinee assistance. In IEEE Winter Conference on Application of Computer Vision (WACV)

The Ideal Personal Assistant

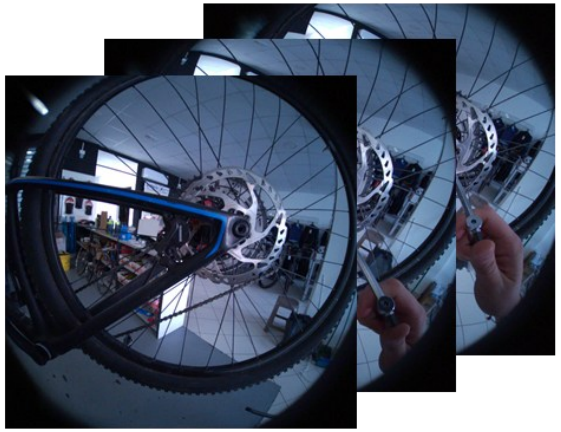
F. Ragusa, M. Mazzamuto, R. Forte, I. D'Ambra, J. Fort, J. Engel, A. Furnari, G. M. Farinella (2026). Ego-EXTRA: video-language Egocentric Dataset for EXpert-TRAinee assistance. In IEEE Winter Conference on Application of Computer Vision (WACV)


Ego-EXTRA: Egocentric dataset of EXpert-TRAinee assistance

Ego-Extra


Trainee
Multimodal Signals
Expert
eye gaze T
hand keypoint
eye gaze E
SLAM
IMU
Transcripts
What is the first ingredient to take?
Now turn and in front of you, there is the flour below
...
...
How much flour should be weighed on the scale?
You need to weigh 1.5 Kg of flour
...
...
How should the dough be shaped?
We can quietly make this kind of serpentine shape
Where should the tray with the cookies be placed in the oven?
They go to the bottom

Multiple-Choice Question Answering



Video Clip

"Do I need to worry that the wheel might fall?"

Trainee's question

Input

→ MiniGPT4-video → **C**

→ Video-LLama → **A**

→ Expert (GT) → **A**

A) "No, not at this moment. Now, hold it like that. "

B) "Maybe we should stop and secure everything again to be absolutely sure."

C) "No, but it's better to use additional supports or have someone assist you just in case."

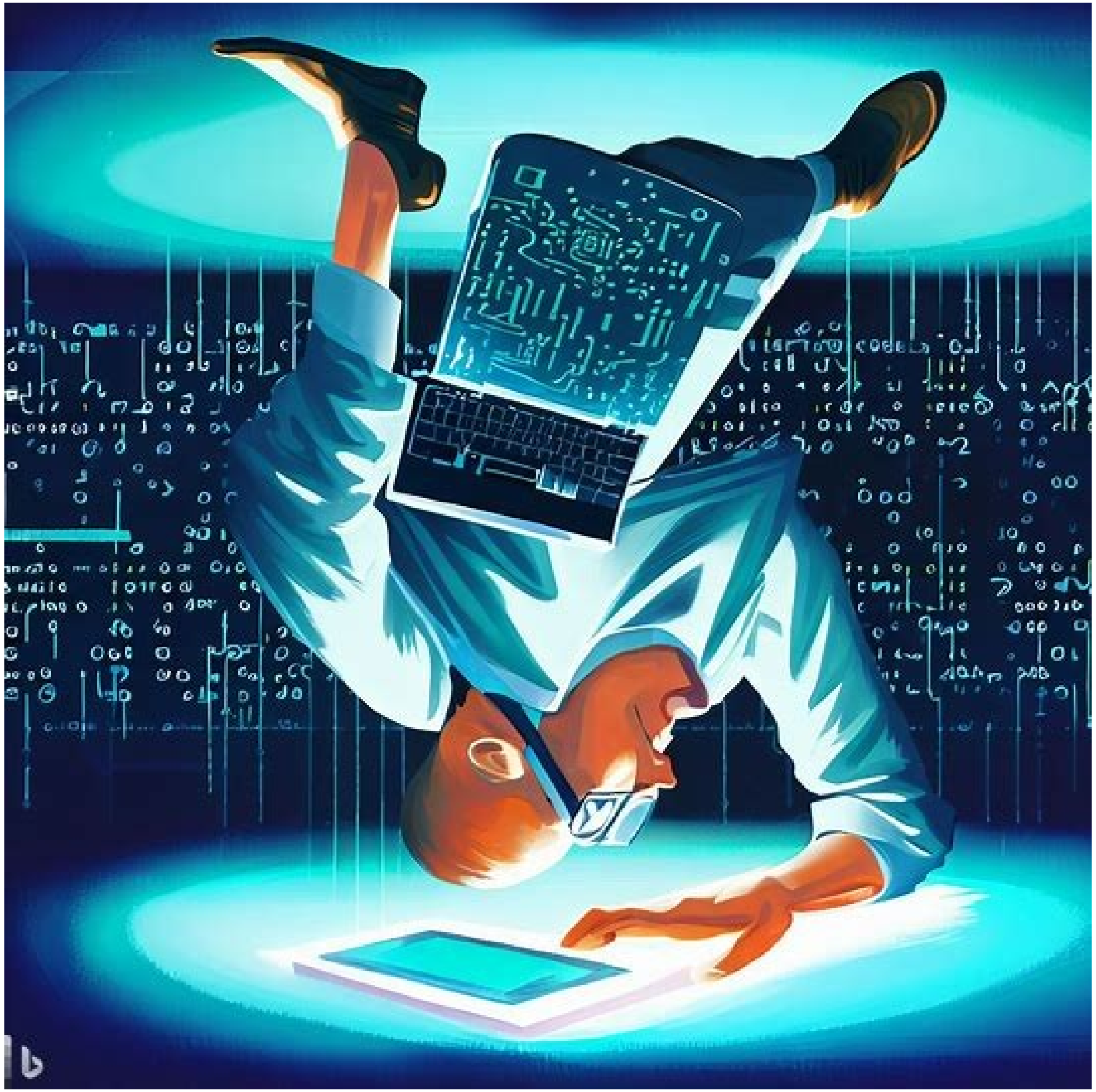
D) "No, just let go and see if it stays in place."

F. Ragusa, M. Mazzamuto, R. Forte, I. D'Ambra, J. Fort, J. Engel, A. Furnari, G. M. Farinella (2026). Ego-EXTRA: video-language Egocentric Dataset for EXpert-TRAinee assistance. In IEEE Winter Conference on Application of Computer Vision (WACV)

What’s Next?

An Outlook into the Future

What's Relevant in Egovision? A top-down approach



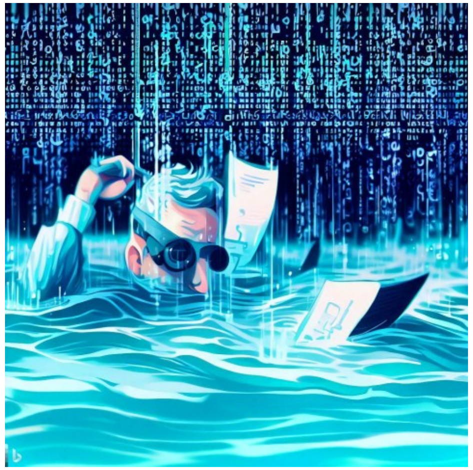
Imagine the Future

Write Stories in Different Scenarios

Extract Important Tasks from the Stories

Go in-depth with Tasks and Datasets

A lot of data!



Rather than being extensive, we considered **seminal** and **state-of-the-art** works

An Outlook into the Future of Egocentric Vision

An Outlook into the Future of Egocentric Vision

Chiara Plizzari* · Gabriele Goletto* · Antonino Furnari* · Siddhant Bansal* · Francesco Ragusa* · Giovanni Maria Farinella† · Dima Damen† · Tatiana Tommasi†

Politecnico di Torino · University of BRISTOL · Università di Catania

Received: date / Accepted: date

Abstract *What will the future be? We wonder!* In this survey, we explore the gap between current research in egocentric vision and the ever-anticipated future, where wearable computing, with outward facing cameras and digital overlays, is expected to be integrated in our every day lives. To understand this gap, the article starts by envisaging the future through character-based stories, showcasing through examples the limitations of current technology. We then provide a mapping between this future and previously defined research tasks. For each task, we survey its seminal works, current state-of-the-art methodologies and available datasets, then reflect on shortcomings that limit its applicability to future research. Note that this survey focuses on software models for egocentric vision, independent of any specific hardware. The paper concludes with recommendations for areas of immediate explorations so as to unlock our path to the future always-on, personalised and life-enhancing egocentric vision.

Keywords Egocentric Vision, Future, Survey, Localisation, Scene Understanding, Recognition, Anticipation, Gaze Prediction, Social Understanding, Body Pose Estimation, Hand and Hand-Object Interaction, Person Identification, Summarisation, Dialogue, Privacy

Contents

1 Introduction

Designing and building tools able to support human activities, improve quality of life, and enhance individuals' abilities to achieve their goals is the ever-lasting aspiration of our species. Among all inventions, digital computing has already had a revolutionary effect on human history. Of particular note is mobile technology, currently integrated in our lives through hand-held devices, i.e. *mobile smart phones*. These are nowadays the de facto for outdoor navigation, capturing static and moving footage of our everyday and connecting us to both familiar and novel connections and experiences.

However, humans have been dreaming about the next-version of such mobile technology — wearable computing, for a considerable amount of time. Imaginations

*: Equal Contribution/First Author
†: Equal Senior Author
C. Plizzari, G. Goletto and T. Tommasi, Politecnico di Torino, Italy · A. Furnari, F. Ragusa and G. M. Farinella, University of Catania, Italy · S. Bansal and D. Damen, University of Bristol, UK. E-mail: Tatiana.Tommasi@polito.it

https://arxiv.org/abs/2308.07123

OpenReview.net

An Outlook into the Future of Egocentric Vision

Chiara Plizzari, Gabriele Goletto, Antonino Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Dima Damen, Tatiana Tommasi

14 Aug 2023 OpenReview Archive Direct Upload Readers: Everyone Show Revisions

Abstract: What will the future be? We wonder! In this survey, we explore the gap between current research in egocentric vision and the ever-anticipated future, where wearable computing, with outward facing cameras and digital overlays, is expected to be integrated in our every day lives. To understand this gap, the article starts by envisaging the future through character-based stories, showcasing through examples the limitations of current technology. We then provide a mapping between this future and previously defined research tasks. For each task, we survey its seminal works, current state-of-the-art methodologies and available datasets, then reflect on shortcomings that limit its applicability to future research. Note that this survey focuses on software models for egocentric vision, independent of any specific hardware. The paper concludes with recommendations for areas of immediate explorations so as to unlock our path to the future always-on, personalised and life-enhancing egocentric vision.

Add Comment

Reply Type: all Author: everybody Visible To: all readers Hidden From: nobody

6 Replies

Related work on modeling social interactions, especially multimodal dialogue agents

Jaewoo Ahn

18 Aug 2023 OpenReview Archive Paper22166 Comment Readers: Everyone Show Revisions

Comment:

I've been reading your fascinating work and wanted to contribute a suggestion based on my recent research in multimodal dialogue agents.

In our recent paper [1], we explored the benefits of a multimodal approach to dialogue personalization. Our study showed that incorporating both text and images in defining a persona greatly enriched the dialogue agent's understanding and personalization capabilities. Specifically, the image modality (i.e., egocentric vision) allowed the dialogue agents to access and better understand their personal characteristics and experiences based on their "episodic memory".

Drawing from this, I propose that there is a strong case to be made for the integration of egocentric vision into the domain of personalized dialogue agent responses. Egocentric vision, being intrinsically tied to personal perspective and experience, can serve as a valuable addition to a persona's episodic memory. This integration can enable chatbots to generate more contextually aware, and personalized responses based on the visual experiences of a user. The fusion of such vision-based episodic memory with textual modalities can be also a promising avenue for future research in personalized dialogue agents.

[1] Ahn et al. MPCHAT: Towards Multimodal Persona-Grounded Conversation, ACL 2023 (https://aclanthology.org/2023.acl-long.189/)

Add Comment

Related work on egocentric full-body pose estimation

Jiaxi Jiang

17 Aug 2023 (modified: 17 Aug 2023) OpenReview Archive Paper22166 Comment Readers: Everyone Show Revisions

Comment:

Thanks for the nice paper, that's awesome!

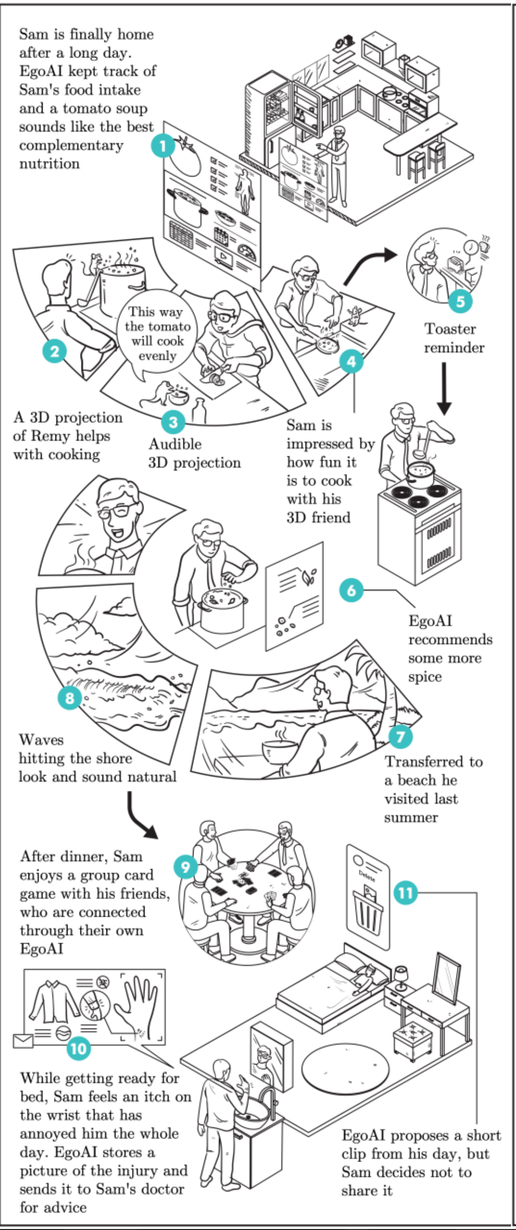
I would really appreciate if our work (AvatarPoser [1] and EgoPoser [2]) on the topic of egocentric full-body pose estimation can also be presented in this review paper.

https://openreview.net/forum?id=V3974SUk1w

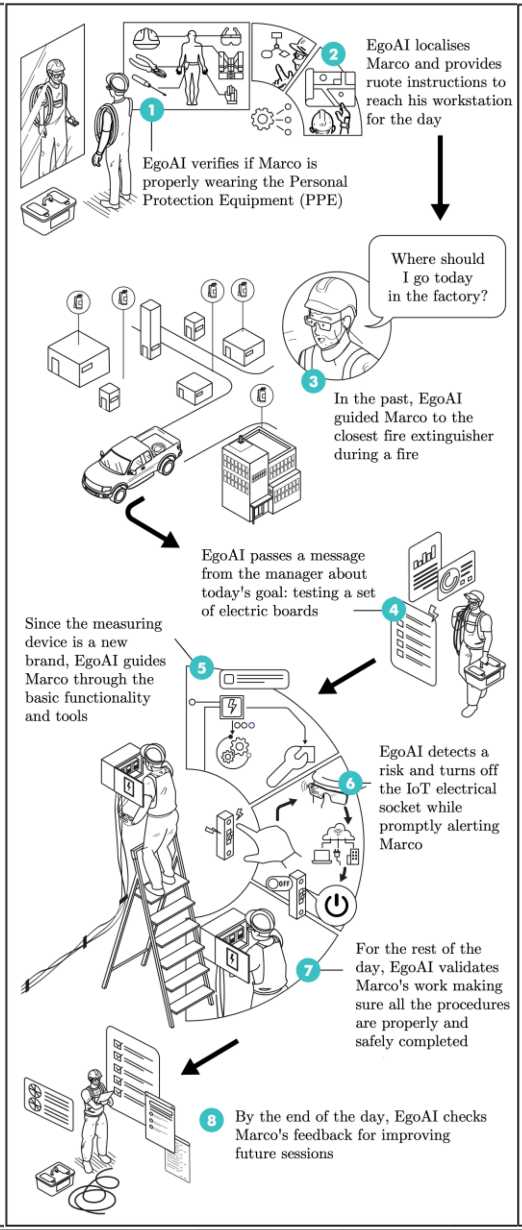
Università di Catania

An Outlook into the Future – Futuristic Stories

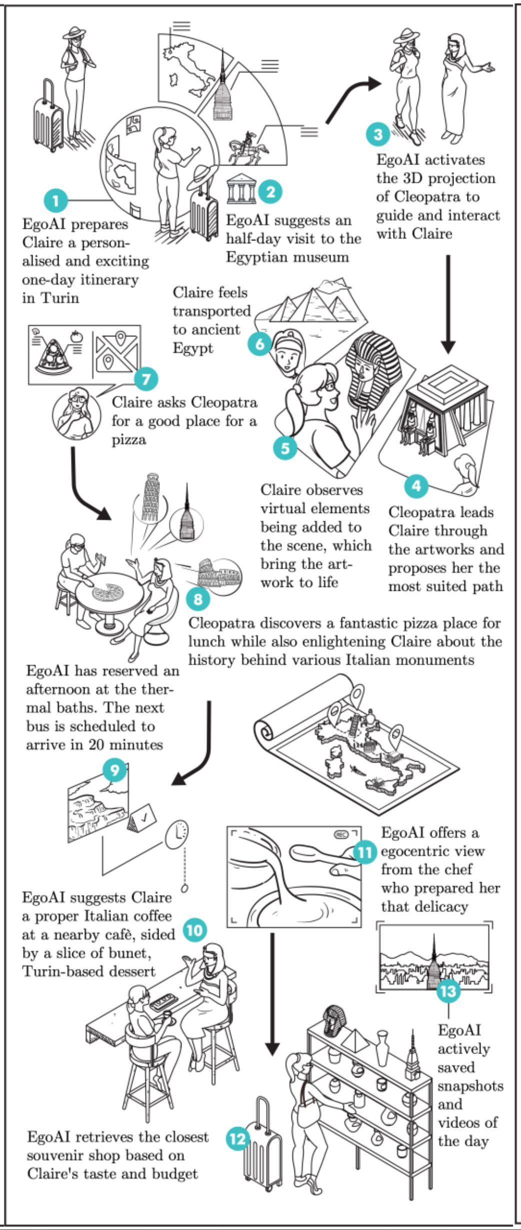
EGO-HOME



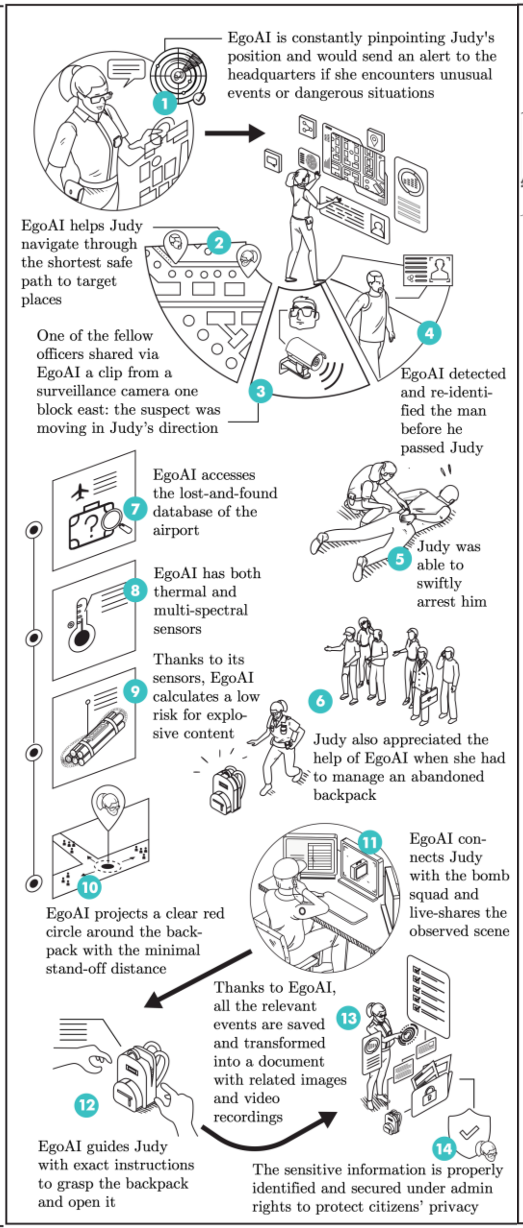
EGO-WORKER



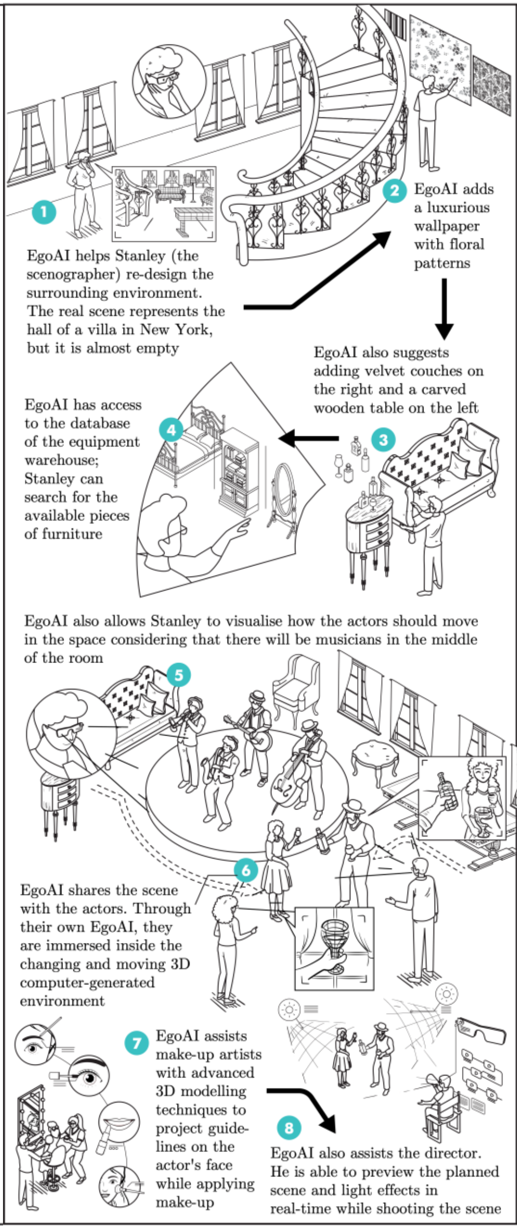
EGO-TOURIST



EGO-POLICE



EGO-DESIGNER



From Narratives to Research Tasks

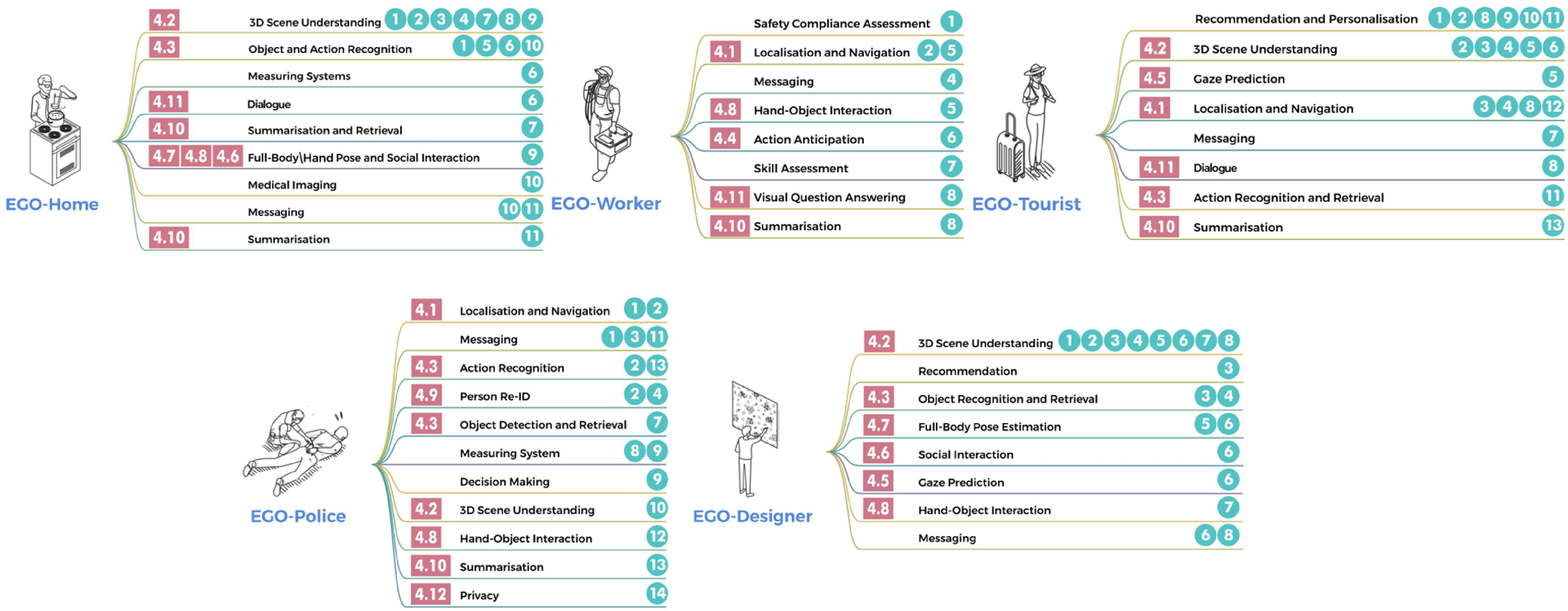
12 Egocentric Vision Research Tasks

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

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. *IJCV, 2024*.

Università di Catania

Links between Stories and Tasks



perspective and provides ego-based assistance. We associate story parts with research tasks (marked by section number) and later revisit the link between these

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. *IJCV, 2024*.

General Datasets

Table 1 General Egocentric Datasets - Collection Characteristics. †: For EGTEA, Audio was collected but not made public. ⋆: For Ego4D, apart from RGB, the other modalities are present for subsets of the data.

Dataset	Settings	Signals	Hours	Sequences	AVG. video duration	Participants
MECCANO (Ragusa et al 2023b)	Industrial	RGB, depth, gaze	6.9	20	20.79 min	20
ADL (Pirsiavash and Ramanan 2012)	Daily activities	RGB	10.0	20	30.00 min	20
HOI4D (Liu et al 2022c)	Table-Top	RGB, depth	22.2	4000	0.33 min	9
EGTEA Gaze+† (Li et al 2021a)	Kitchen	RGB, gaze	27.9	86	19.53 min	32
UTE (Lee et al 2012)	Daily Activities	RGB	37.0	10	222.00 min	4
EGO-CH (Ragusa et al 2020a)	Cultural Sites	RGB	37.1	180	12.37 min	70
FPSI (Fathi et al 2012a)	Recreational Site	RGB	42.0	8	315.00 min	8
KrishnaCam (Singh et al 2016a)	Daily Routine	RGB, GPS, acc	69.9	460	9.13 min	1
EPIC-KITCHENS-100 (Damen et al 2022)	Kitchens	RGB, audio	100.0	700	8.57 min	37
Assembly101 (Sener et al 2022)	Industrial	RGB, multi-view	167.0	1425	7.10 min	53
Ego4D⋆ (Grauman et al 2022)	Multi Domain	RGB, Audio, 3D, gaze, IMU, multi	3670.0	9650	24.11 min	931

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. *IJCV, 2024.*

General Datasets

Table 2 General Egocentric Datasets - Current set of annotations. *: For Ego4D, apart from narrations, the remaining annotations are only available for subsets of the dataset depending on the benchmark

Dataset	Annotations
MECCANO (Ragusa et al 2023b)	Temporal action segments, hand & object bounding boxes, hand-object interactions, next-active object
ADL (Pirsiavash and Ramanan 2012)	Temporal action segments, objects bounding boxes, hand-object interactions
HOI4D (Liu et al 2022c)	Temporal action segments, 3D hand poses and object poses, panoptic and motion segmentation, object meshes, scene point clouds
EGTEA Gaze+ (Li et al 2021a)	Temporal action segments, hand masks, gaze
UTE (Lee et al 2012)	Text descriptions, object segmentations
EGO-CH (Ragusa et al 2020a)	Temporal locations, object bounding boxes, surveys, object masks
FPSI (Fathi et al 2012a)	Temporal social interaction segments
KrishnaCam (Singh et al 2016a)	Motion classes, virtual webcams, popular locations
EPIC-KITCHENS-100 (Damen et al 2022)	Temporal action video segments, Temporal audio segments, narrations, hand and objects masks, hand-object interactions, camera poses
Assembly101 (Sener et al 2022)	Temporal action segments, 3D hand poses
Ego4D* (Grauman et al 2022)	Narrations, Temporal action segments, moment queries, speaker labels, diarisation, hand bounding boxes, time to contact, active objects bounding boxes, trajectories, next-active objects bounding boxes

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. *IJCV, 2024.*

General Datasets

Table 3 General Egocentric Datasets - Current set of tasks: **4.1** Localisation, **4.2** 3D Scene Understanding, **4.3** Recognition, **4.4** Anticipation, **4.5** Gaze Understanding and Prediction, **4.6** Social Behaviour Understanding, **4.7** Full-body Pose Estimation, **4.8** Hand and Hand-Object Interactions, **4.9** Person Identification, **4.10** Summarisation, **4.11** Dialogue, **4.12** Privacy.

Dataset \ Task	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	4.10	4.11	4.12
MECCANO (Ragusa et al 2023b)			✓	✓	✓			✓				
ADL (Pirsiavash and Ramanan 2012)			✓	✓						✓		
HOI4D (Liu et al 2022c)								✓				
EGTEA Gaze+ (Li et al 2021a)			✓	✓	✓			✓				
UTE (Lee et al 2012)								✓		✓		
EGO-CH (Ragusa et al 2020a)	✓											
FPSI (Fathi et al 2012a)						✓				✓		✓
KrishnaCam (Singh et al 2016a)				✓								
EPIC-KITCHENS-100 (Damen et al 2022)		✓	✓	✓				✓			✓	✓
Assembly101 (Sener et al 2022)			✓					✓				
Ego4D (Grauman et al 2022)			✓	✓	✓	✓		✓		✓	✓	

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. *IJCV, 2024.*

Industrial Applications

NEXT VISION

Spin-off of the University of Catania

https://www.nextvisionlab.it/

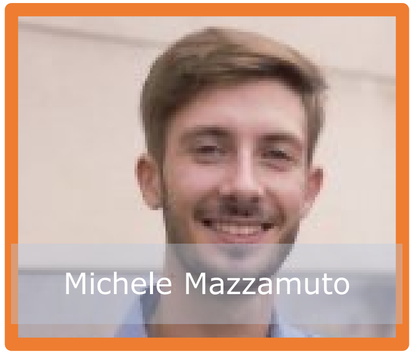
LIVE

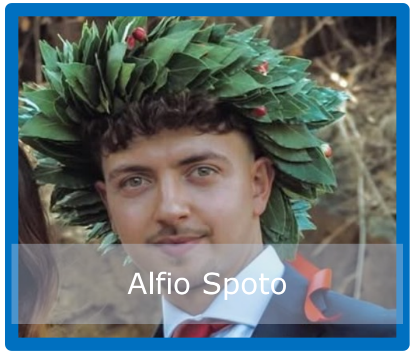
SICILIAE STVDIVM GENERALE 1434

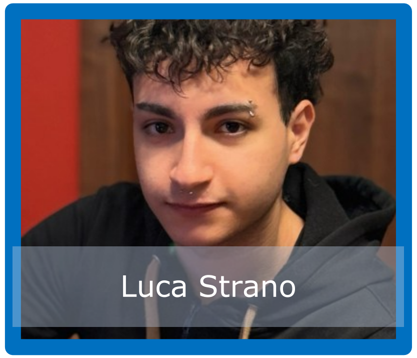
IMAGE PROCESSING LABORATORY

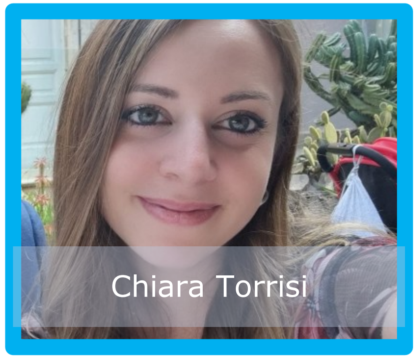
Università di Catania

Next Vision Team









10 Members

Chief Scientific Officier

Chief Executive Officier

Chief R&D Officier

Senior R&D Engineer

R&D Engineer

HR and Project Manager

xenia® SOFTWARE SOLUTIONS

An academic Spinoff of UNICT based in Sicily with an International Vision

Intelligent Navigation

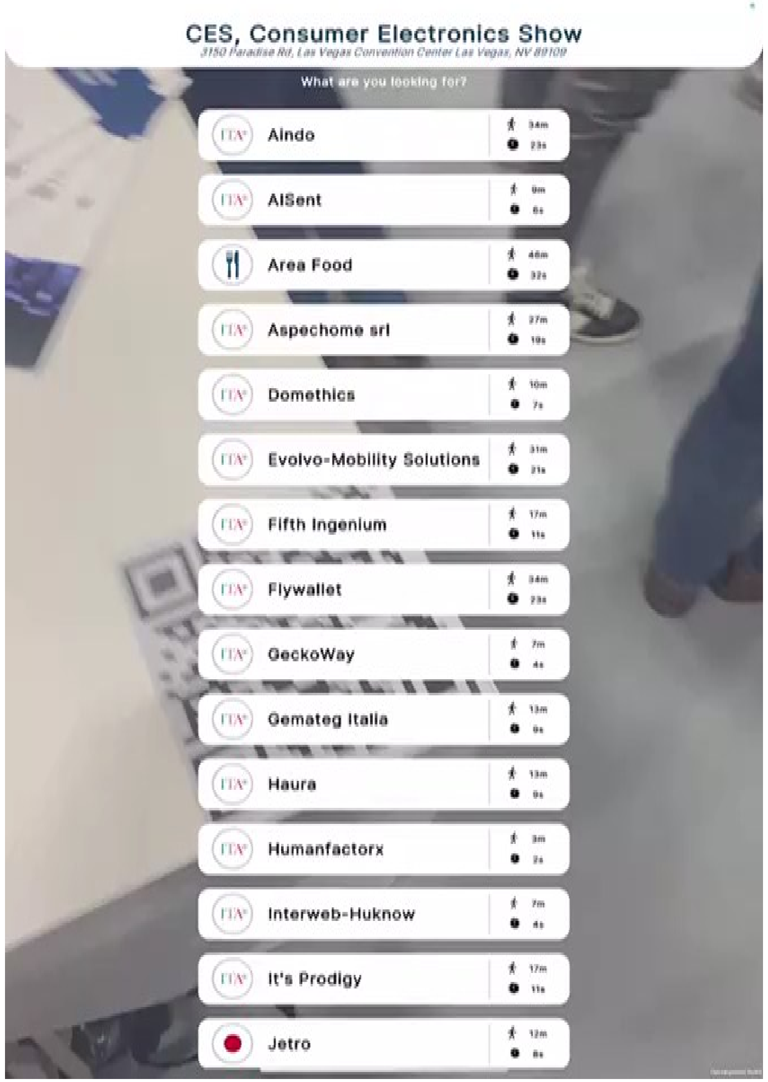
Image-based Localization

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



NOKIA

Università di Catania

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

CES

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

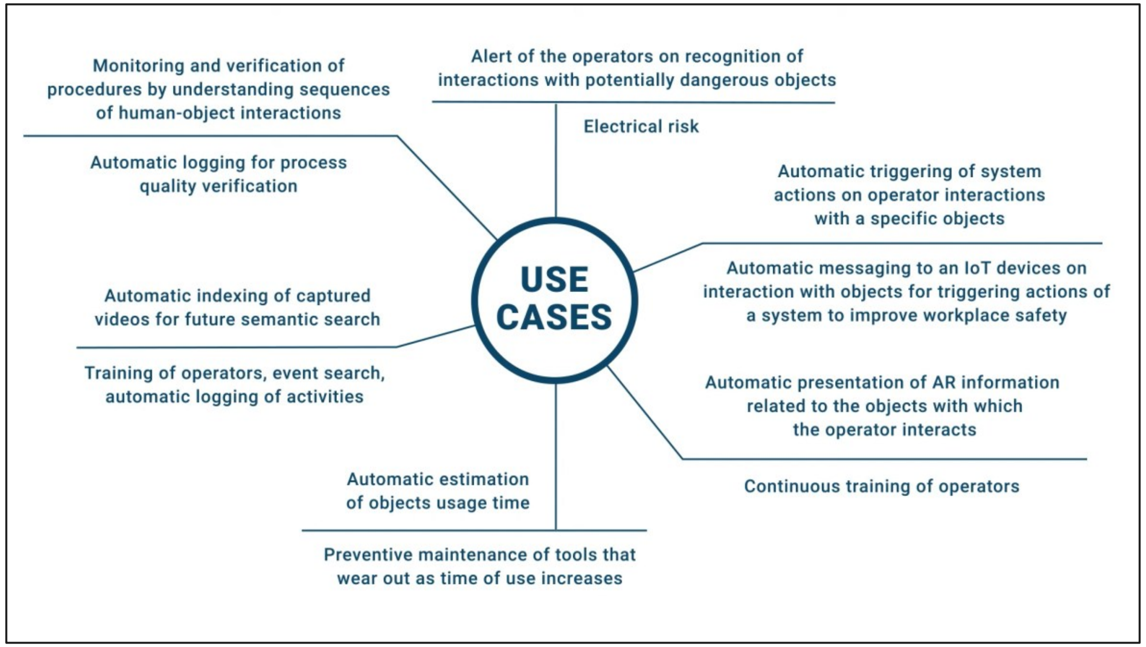
NAIROBI for Rehabilitation

Michele Mazzamuto, Francesco Ragusa, Antonino Furnari, Irene D'Ambra, Antonia Guarriera, Armando Sorbello, Giovanni Maria Farinella (2024). A Mixed Reality Application to Help Impaired People Rehabilitate Outside Clinical Environments. In IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE).

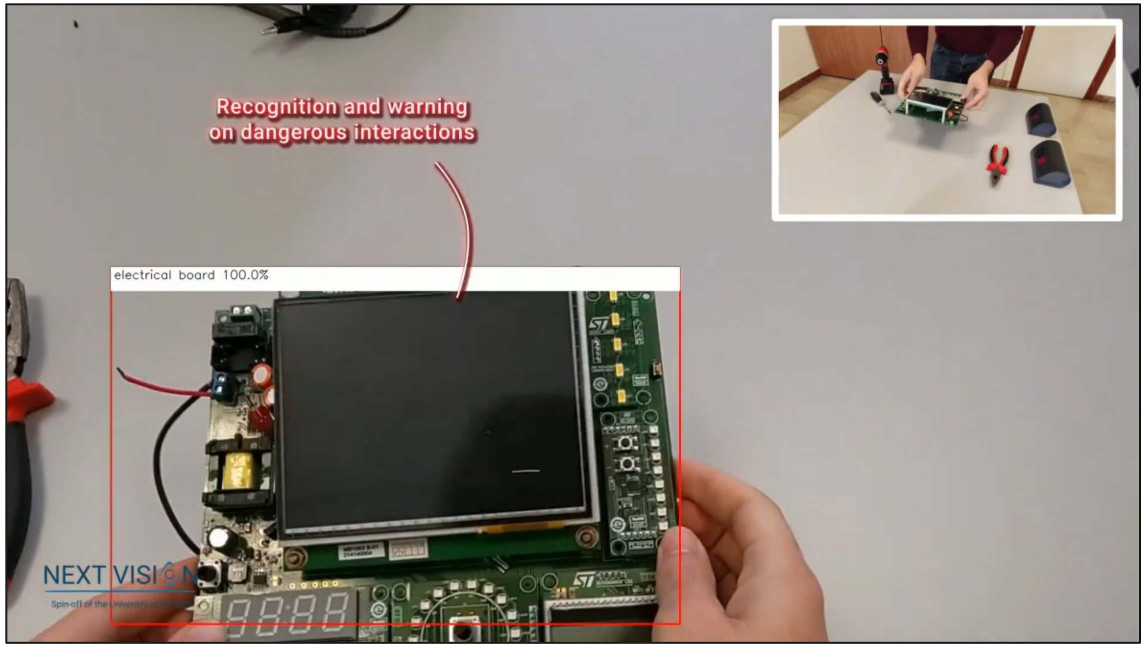
NAIROBI for Rehabilitation

Michele Mazzamuto, Francesco Ragusa, Antonino Furnari, Irene D'Ambra, Antonia Guarriera, Armando Sorbello, Giovanni Maria Farinella (2024). A Mixed Reality Application to Help Impaired People Rehabilitate Outside Clinical Environments. In IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE).

- **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

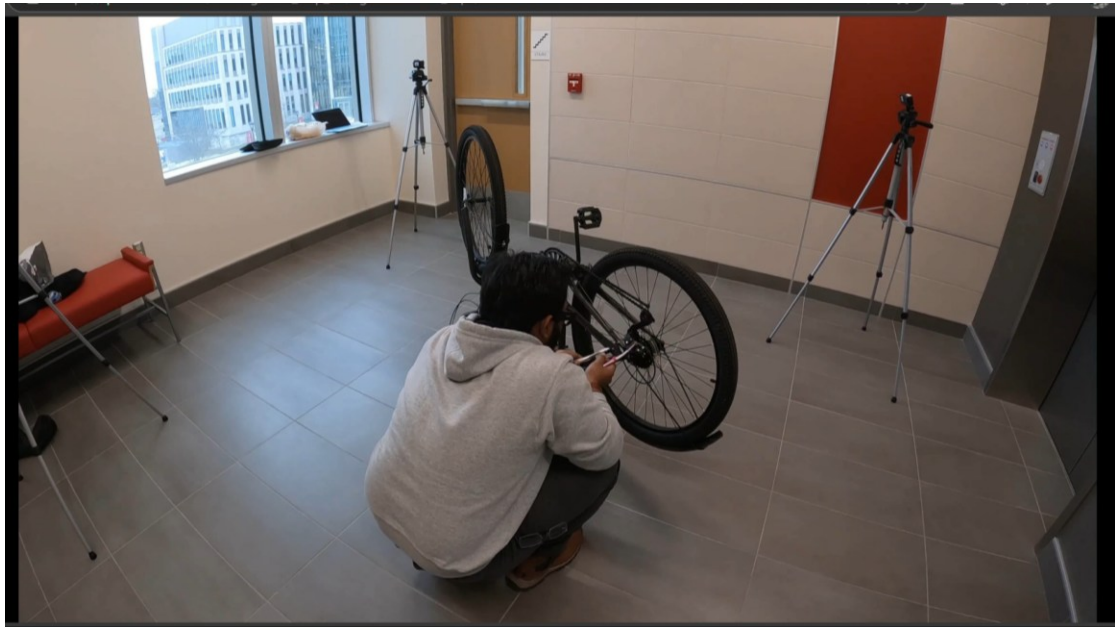


The video shows an example of object interaction monitoring. The operator is notified on an interaction with a dangerous object.

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


NEXT VISION
NEXT VISION
Q 0025
224 225 225
115 500

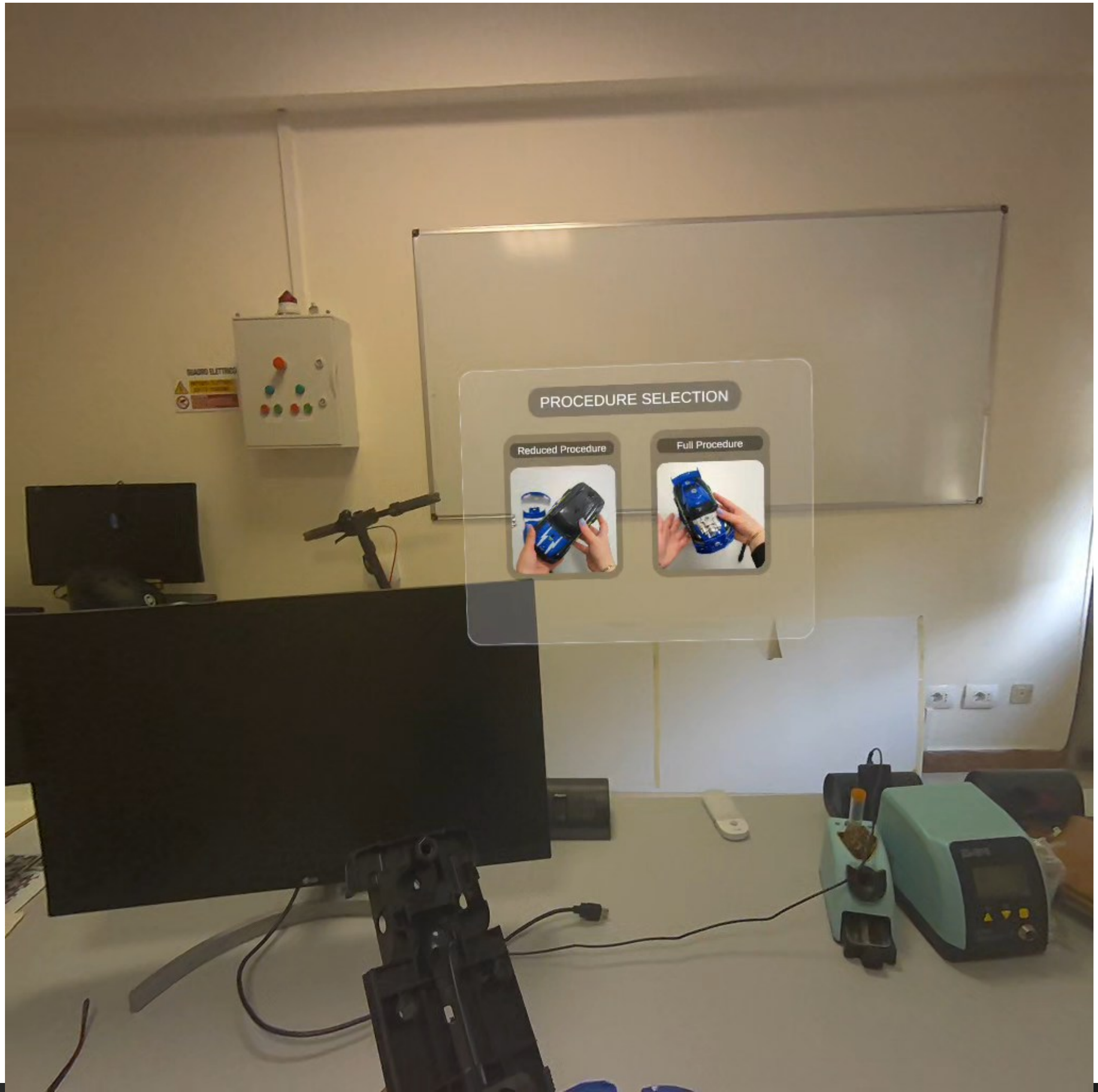
Skill Assessment



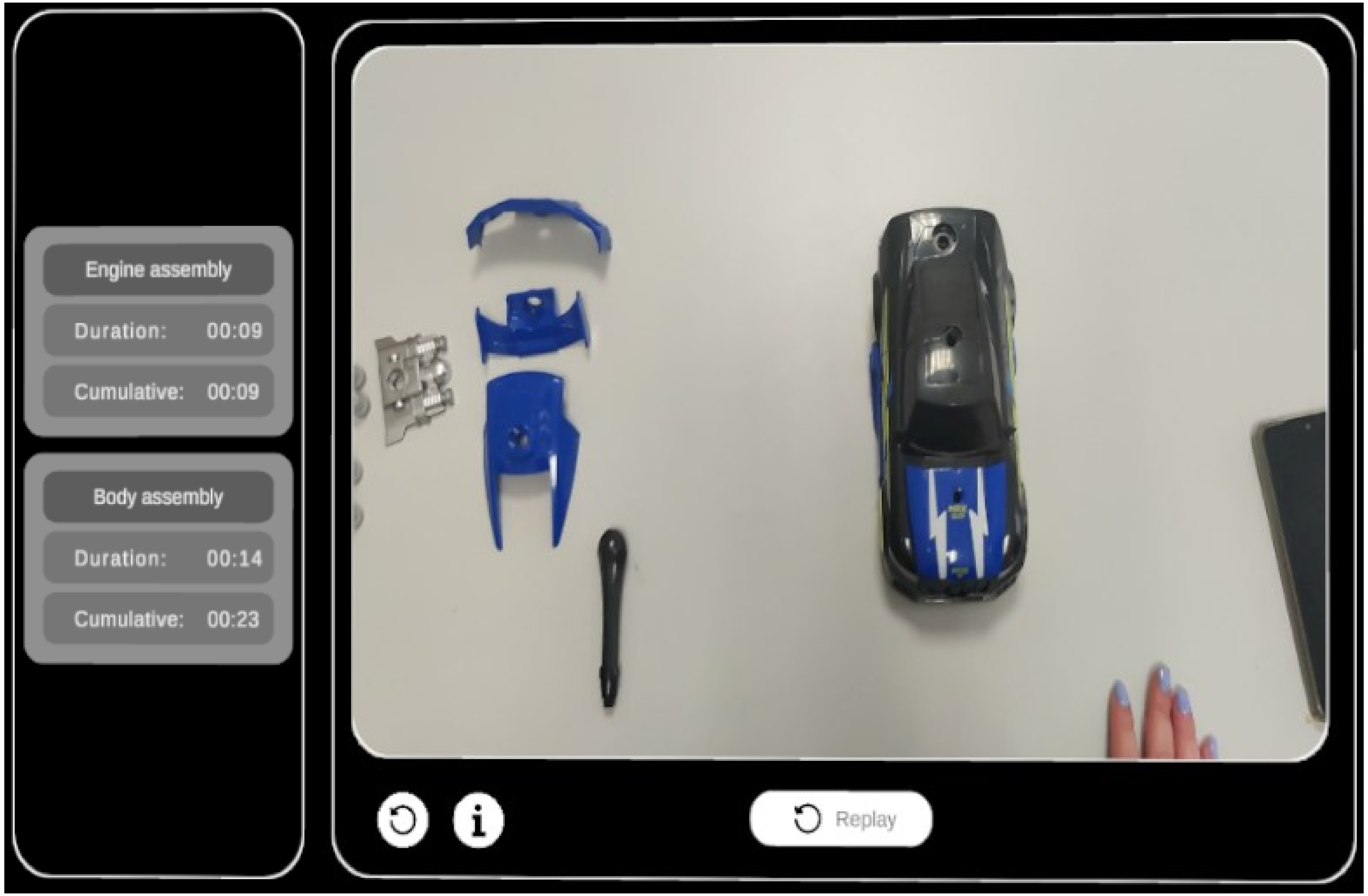
Beginner

Expert

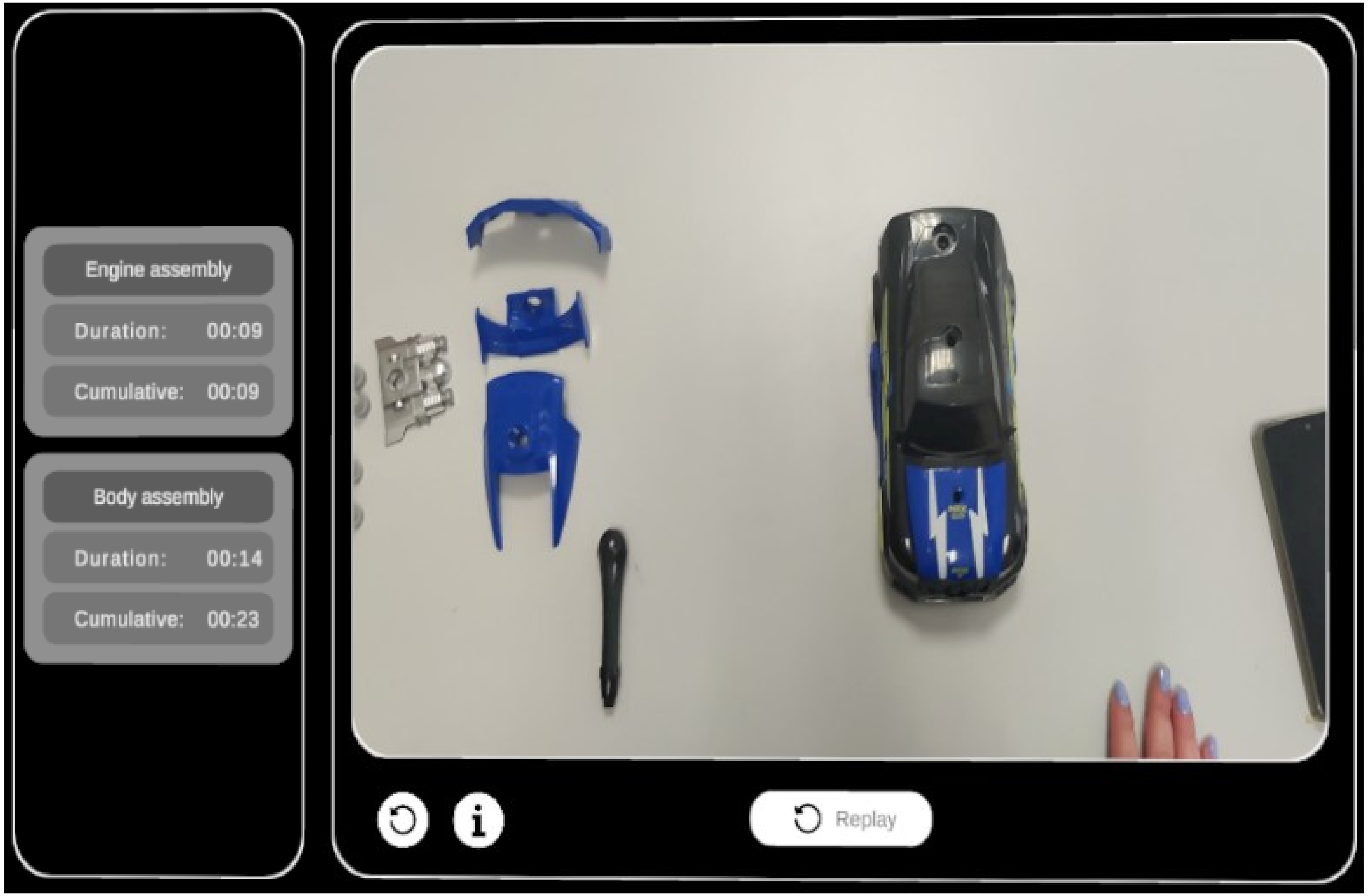
Skill Assessment



Skill Assessment



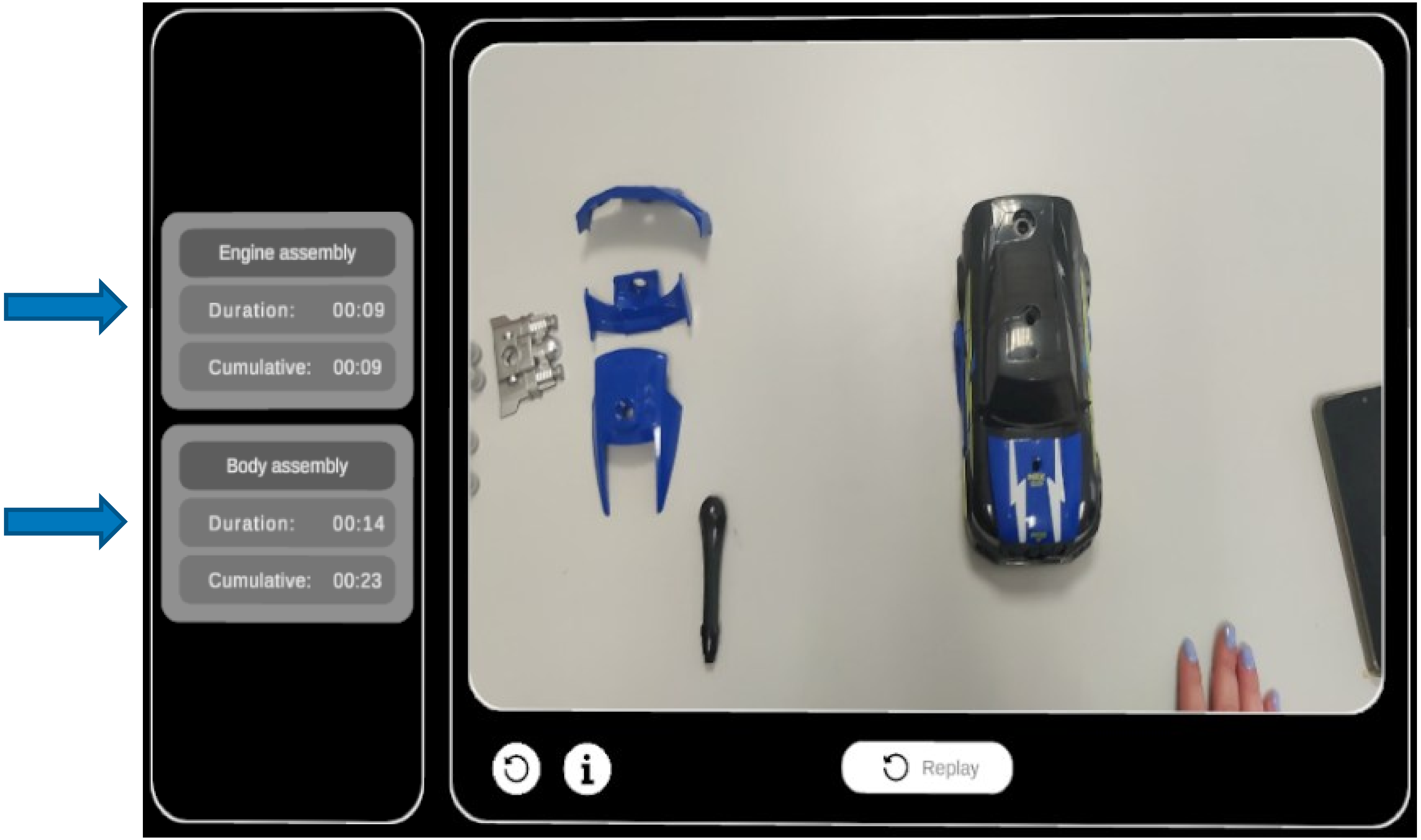
Step Info **Video Tutorial**



Step Info

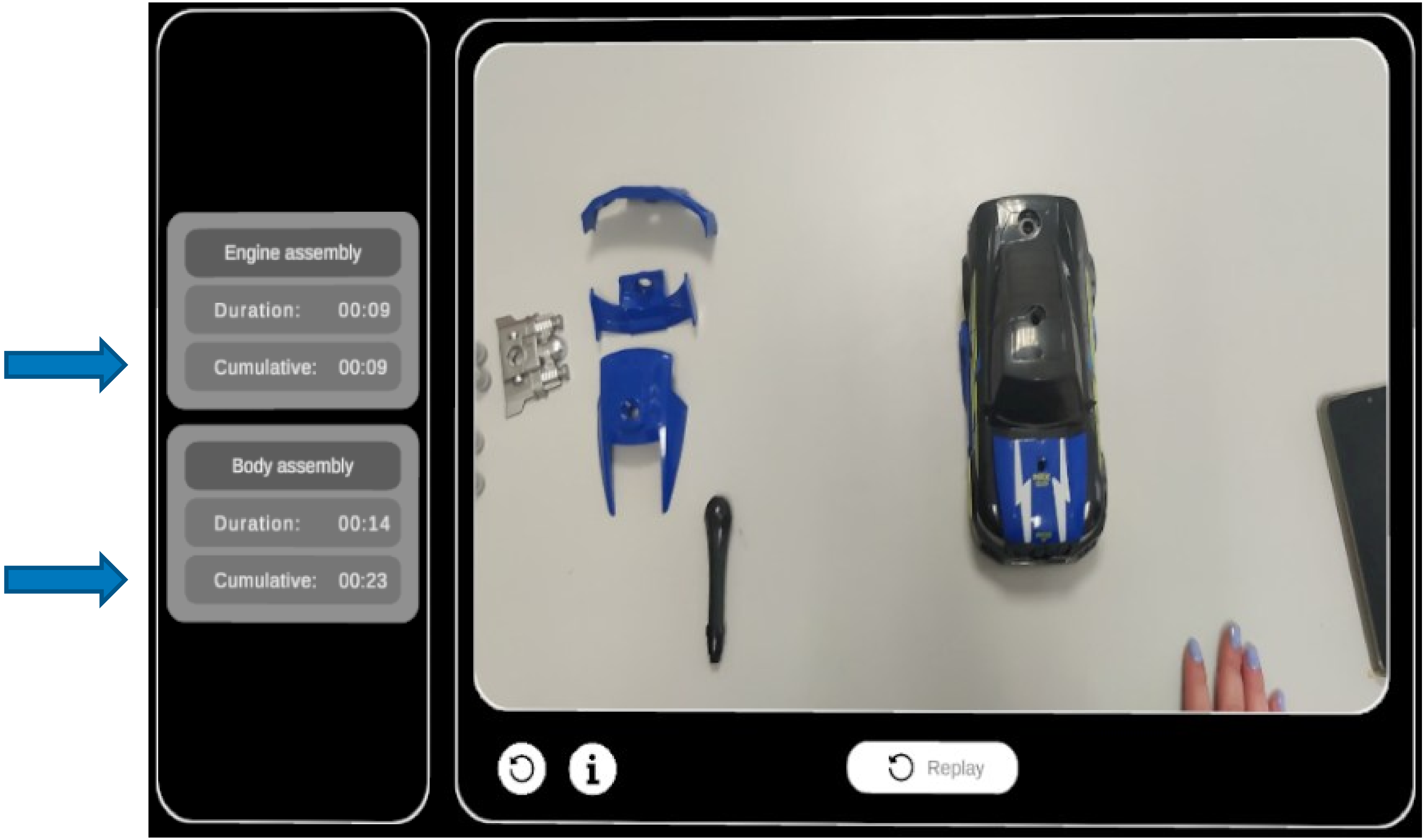
Video Tutorial

Skill Assessment



Step Info

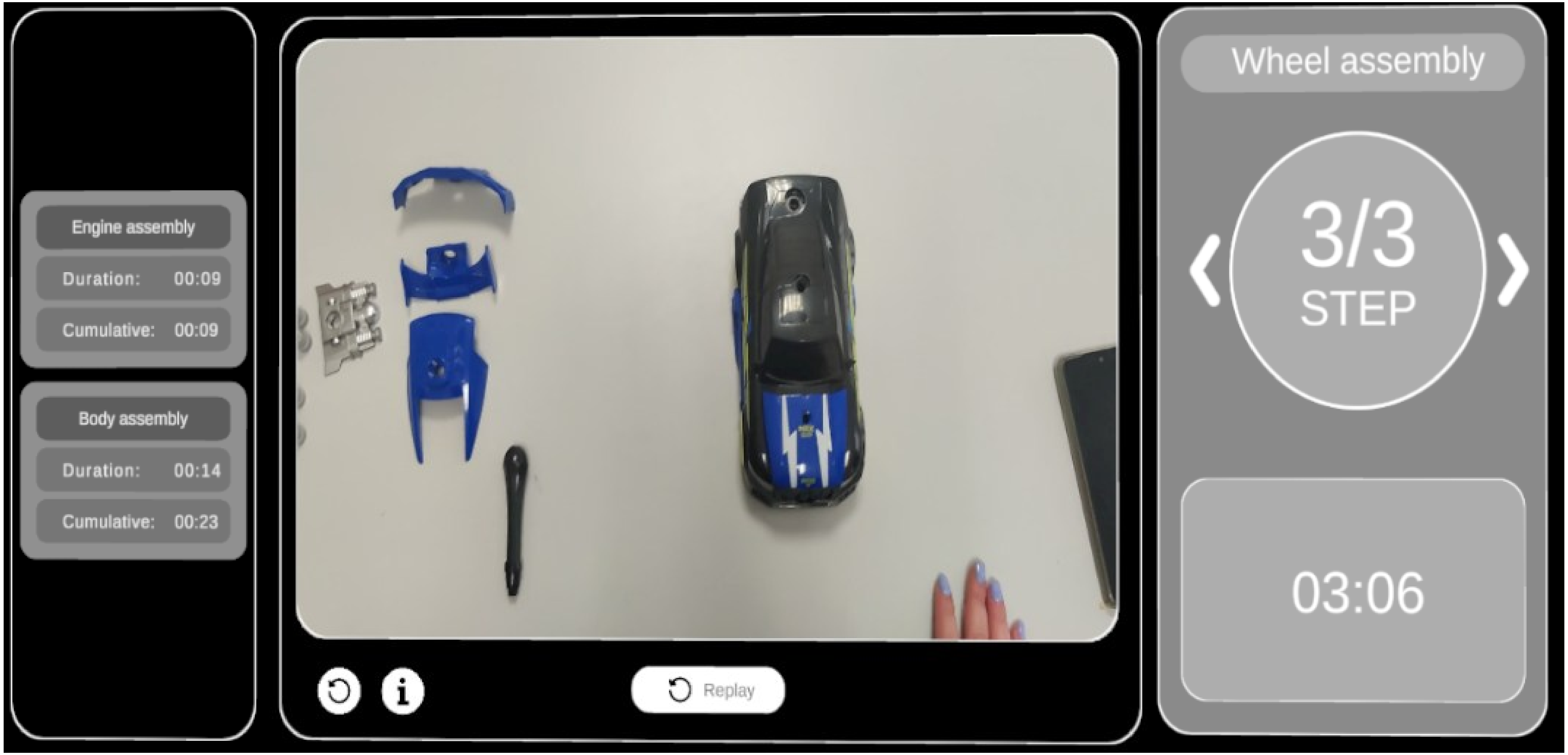
Video Tutorial



Step Info

Video Tutorial

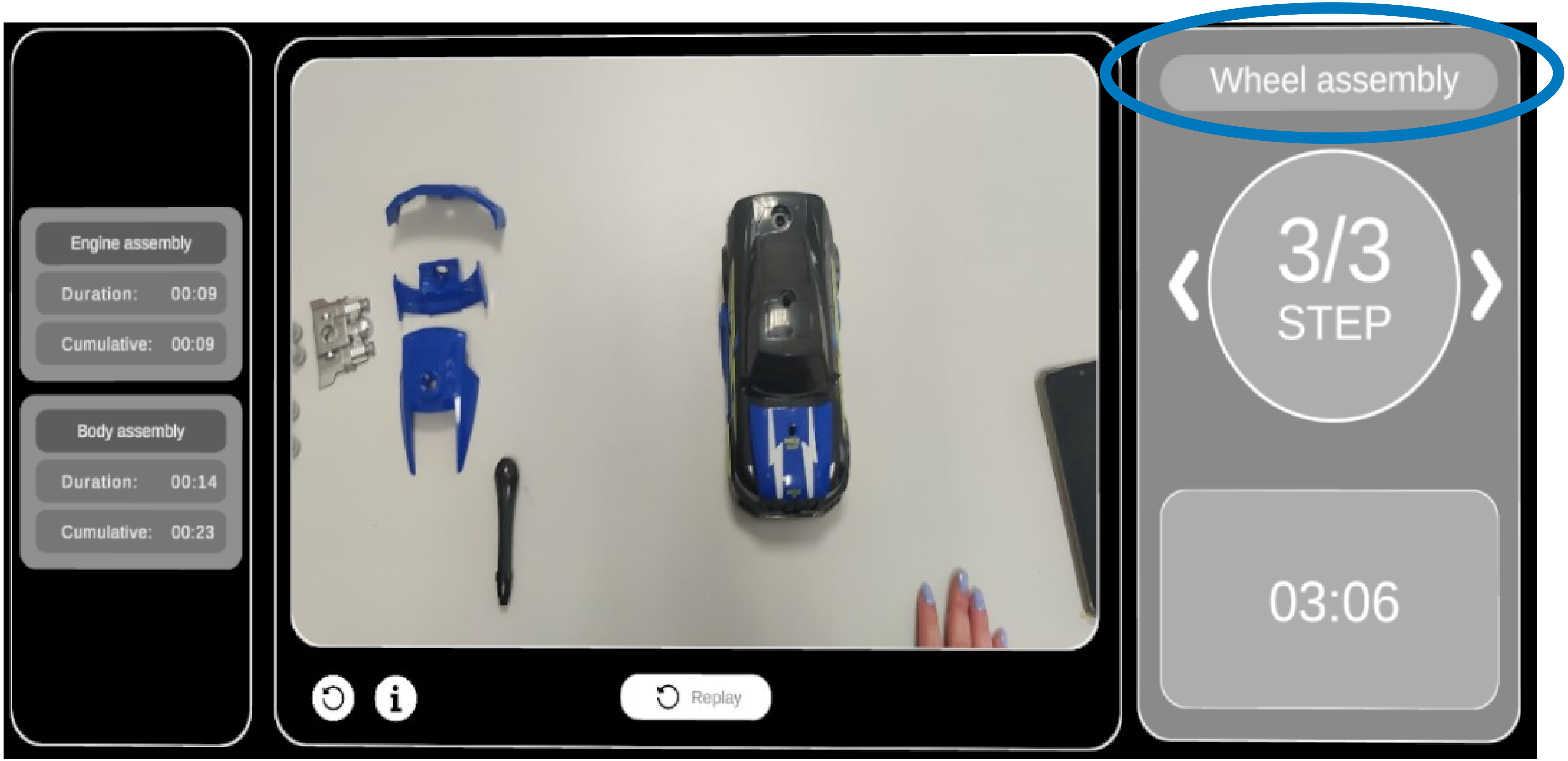
Skill Assessment



Step

Video

Procedure

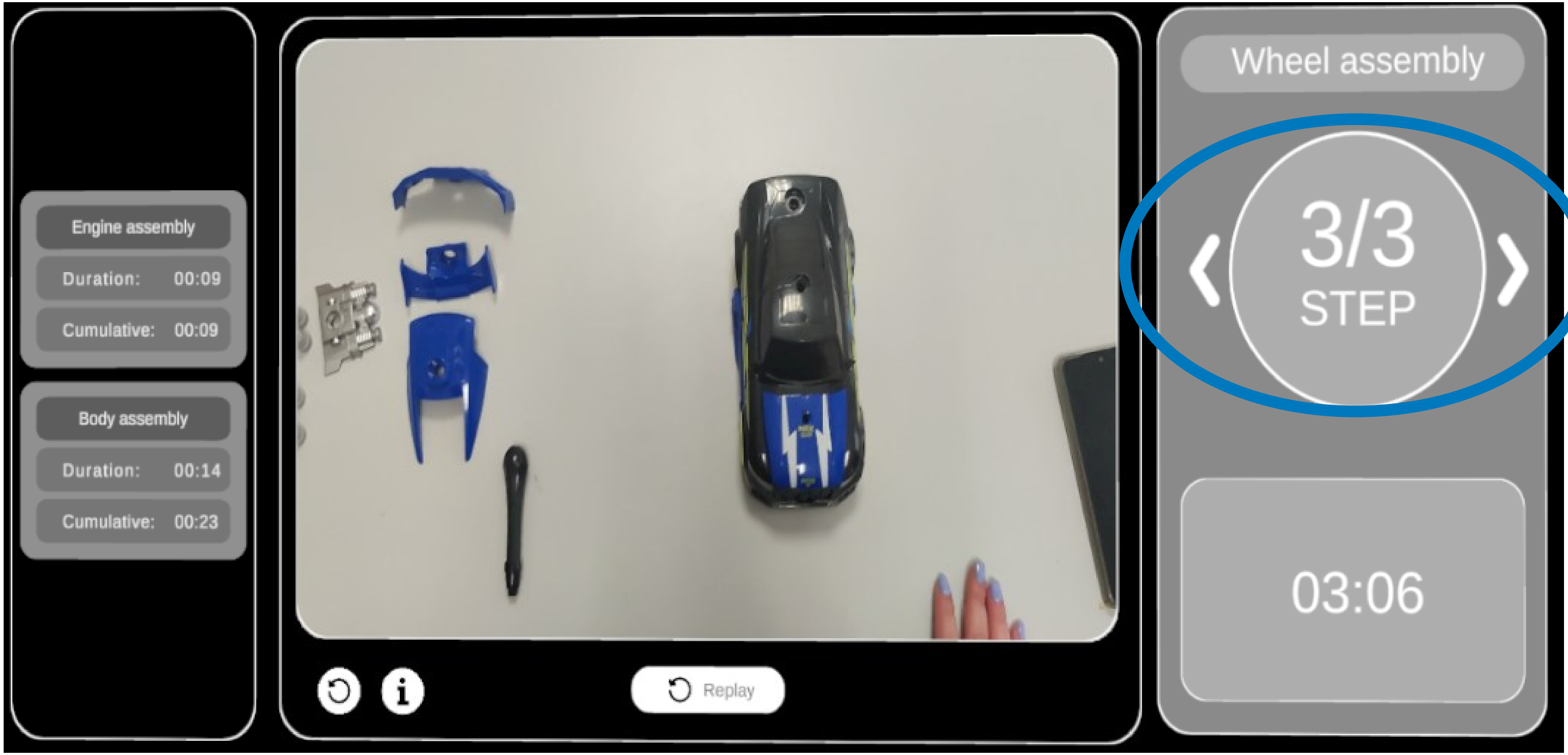


Step Info

Video Tutorial

Procedure Info

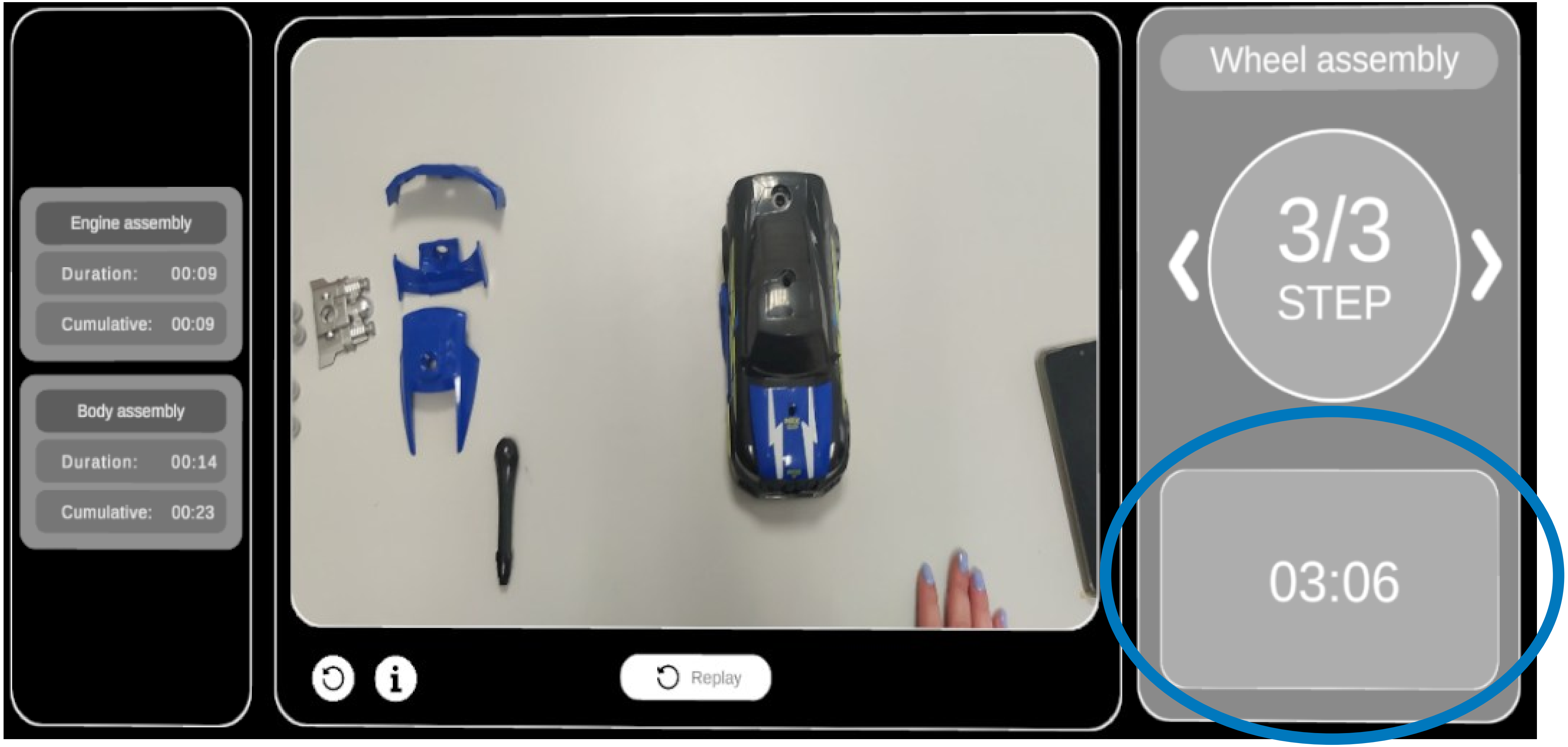
Skill Assessment



Step

Video

Procedure



Step Info **Video Tutorial** **Procedure Info**

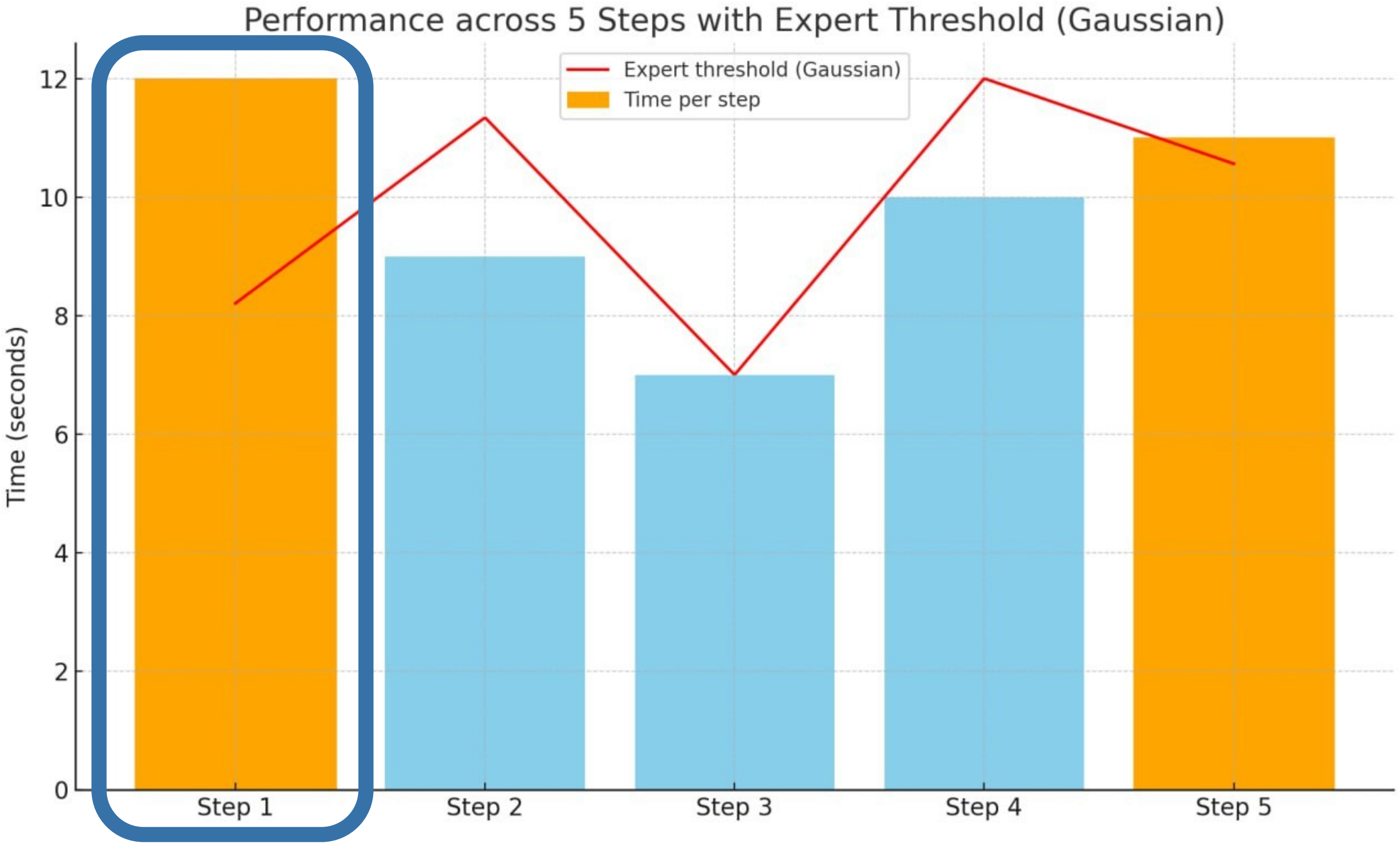
Skill Assessment

Step Info

Procedure Info

Università di Catania

Skill Assessment



Step Info

Procedure Info

Skill Assessment

Step Info

Procedure Info

Skill Assessment

Step Info

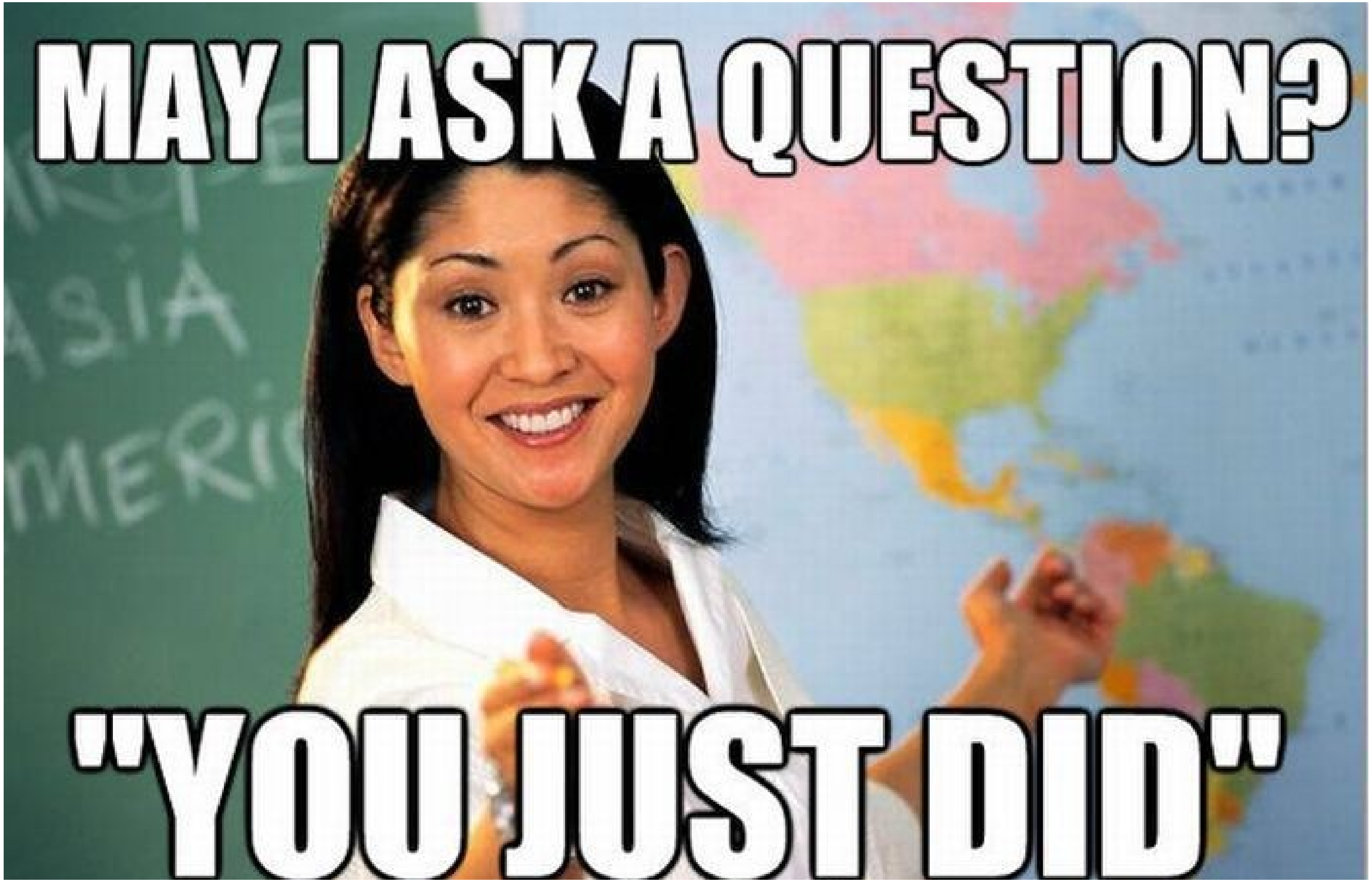
Procedure Info

Conclusion of Part I

It's an exciting time for wearable devices & egocentric vision!

Hardware is increasingly available as big tech gests interested.

Large datasets and pre-defined challenges can help get started to explore the field


MAY I ASK A QUESTION?
"YOU JUST DID"

THANK YOU!

Egocentric Vision:
Emerging Trends and Human-Centric Applications

Francesco Ragusa

LIVE Group @ UNICT - https://iplab.dmi.unict.it/live/

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

Department of Mathematics and Computer Science - University of Catania

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

Agenda

2) **Part II: Hand-Object Interactions in Egocentric Vision [15.50 – 16.50]**
    a) **Introduction to Hand-Object Interactions Detection**
    b) **Datasets and Benchmarks for Hand-Object Interactions in Egocentric Vision**
    c) **Models and Architectures for Hand-Object Interactions Detection**
    d) **Open Challenges**