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/



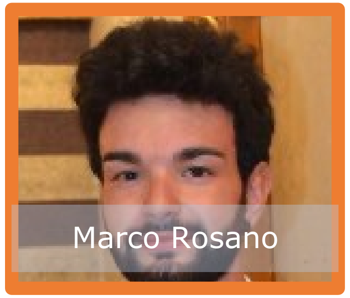


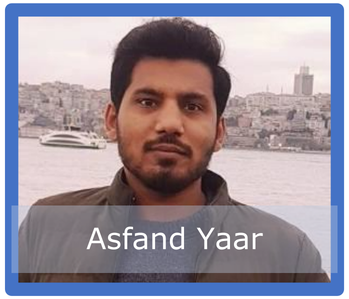

Università di Catania

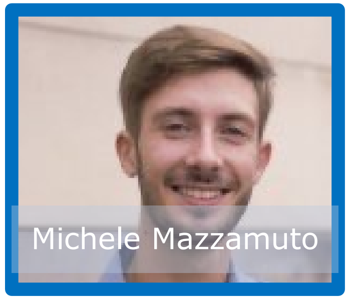
IMAGE PROCESSING LABORATORY

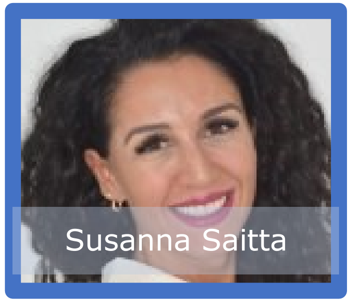
FPV @ IPLAB Group

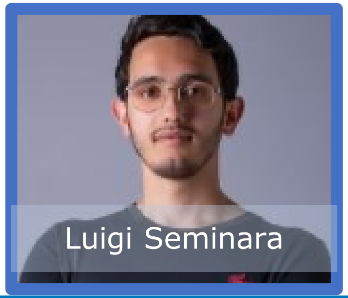
The FPV@IPLAB Group

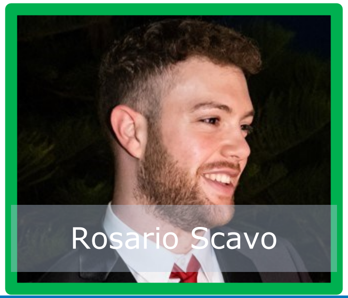












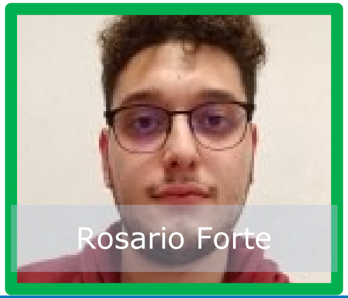


IMAGE PROCESSING LABORATORY

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

NEXT VISION

http://www.nextvisionlab.it/

16 Members
1 Full Professor
1 Assistant Professor
1 Researcher
3 Post Docs
7 PhD Students
2 Master Students
1 Lab Assistant

Before we begin…

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

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.

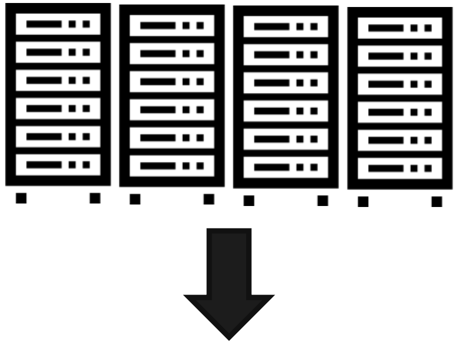
Part I

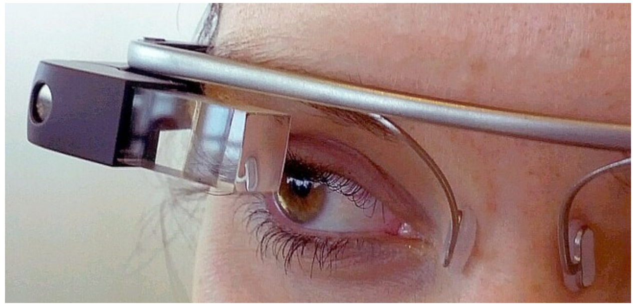
History and Motivations

The Revolution of Personal Computing

After personal computers and smartphones, wearable devices are the <u>third wave of computing</u>

– Marc Pollefeys, Lab Director, Microsoft Mixed Reality and AI Zurich



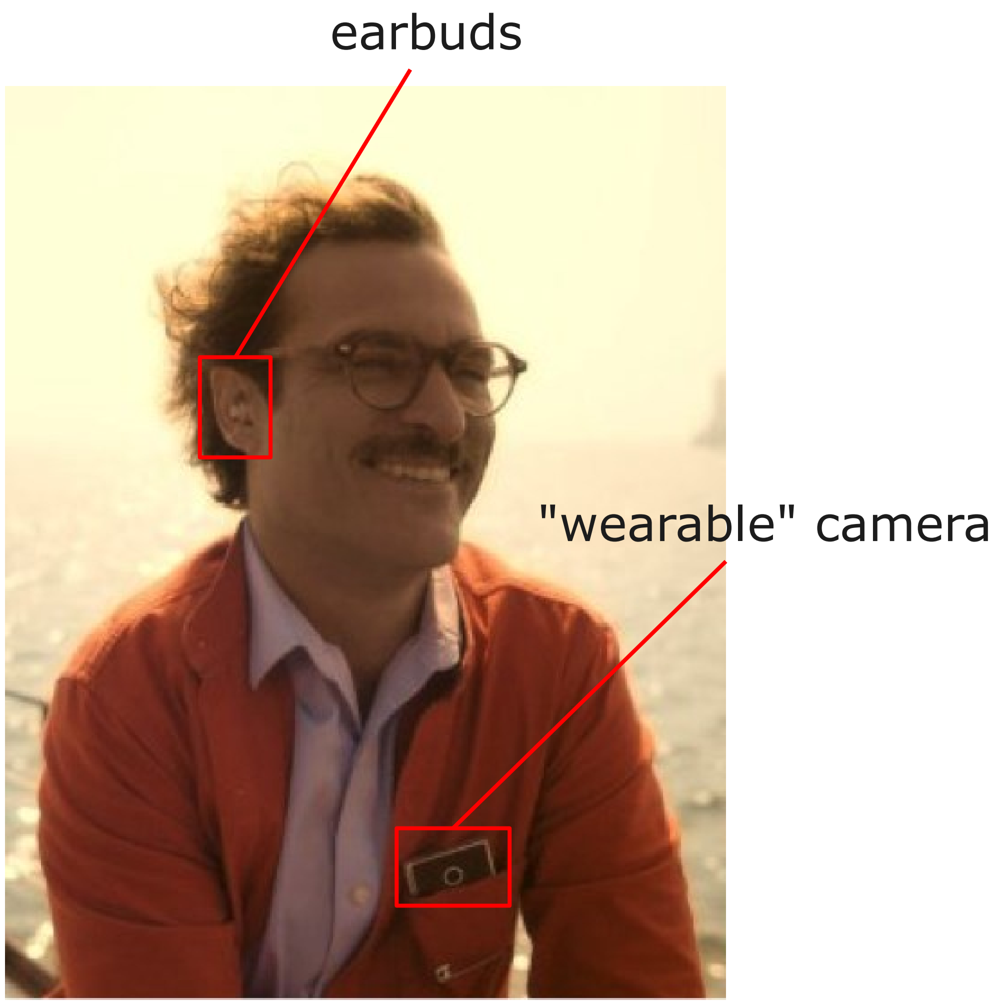


Personal Computers: computing for the mass, but not mobile and not context aware - dedicated access to computing

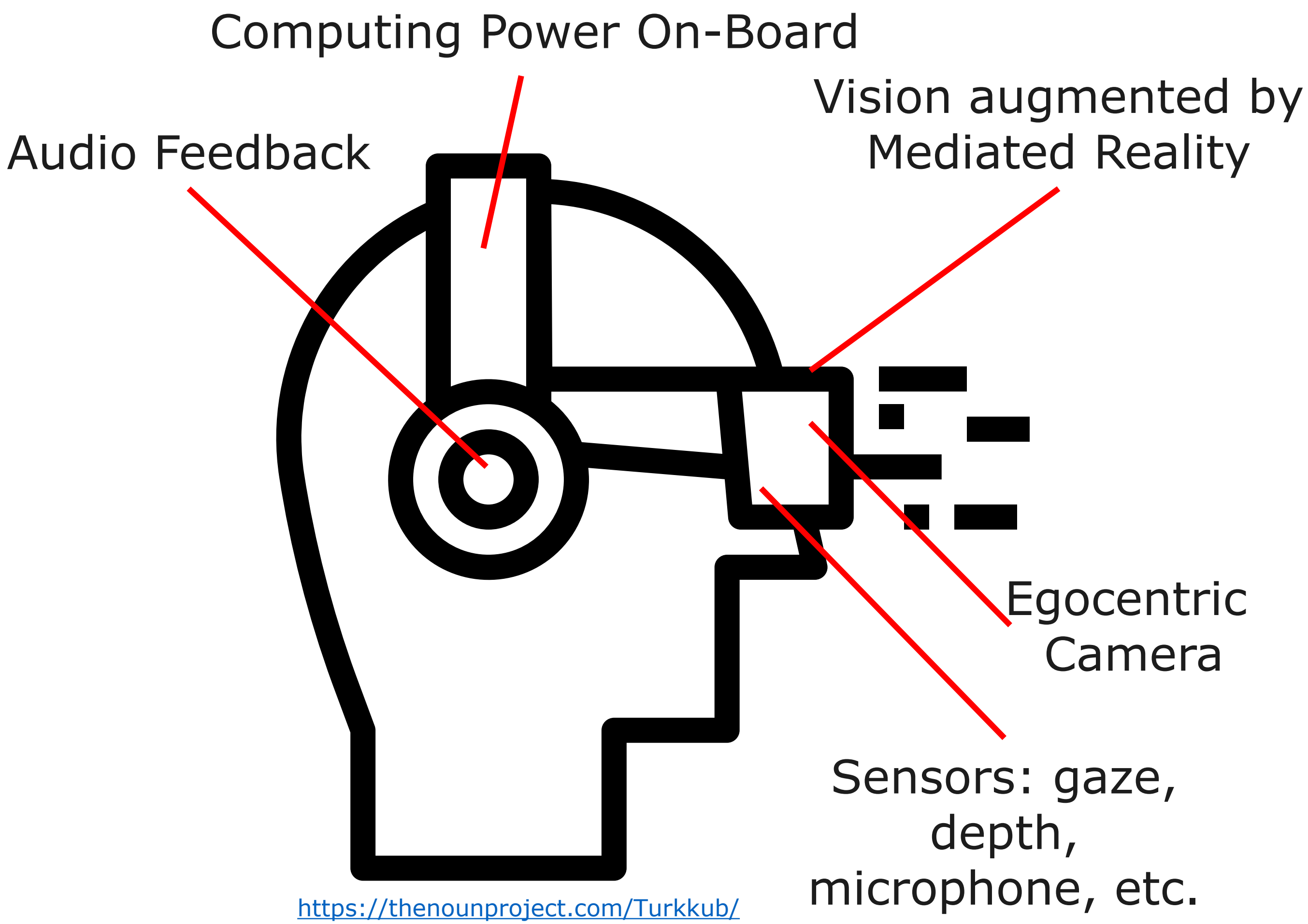
Smartphones: mobile computing is always accessible, but forces to switch between the digital and real world

Eyeworn Devices: computing everywere with minimal switch between real and digital worlds

An AI-Powered Virtual Assistant

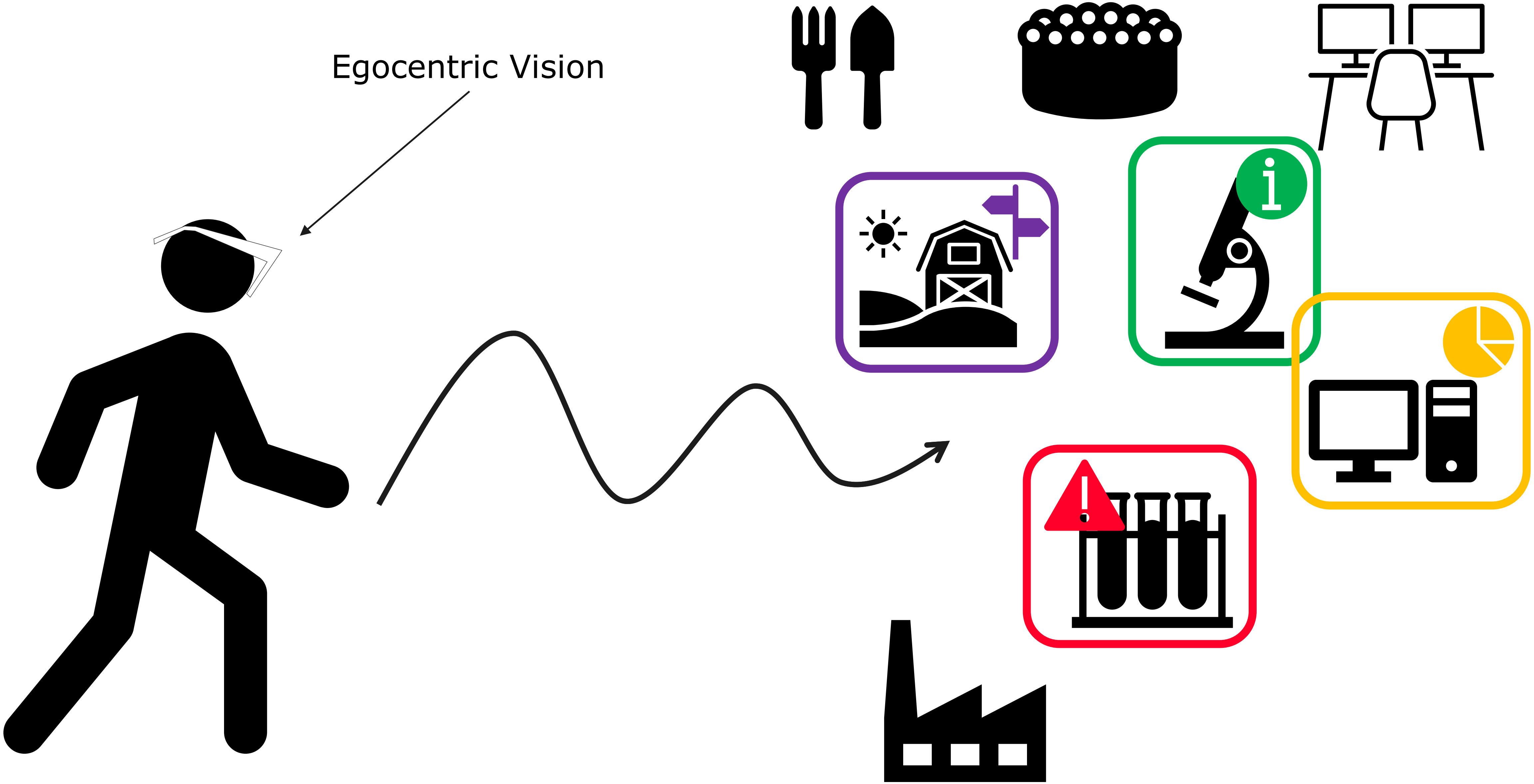


"her" 2013 movie



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

A Virtual Personal Assistant



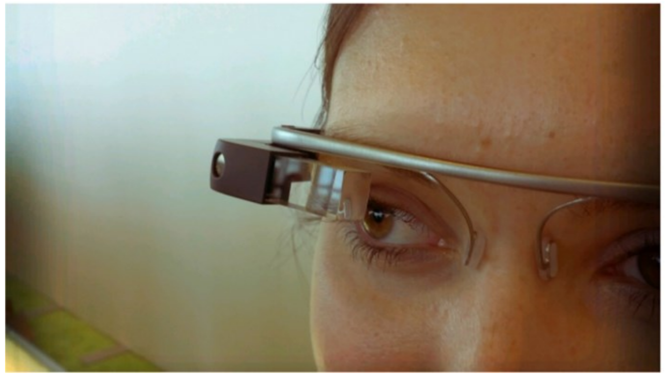
(Egocentric) Computer Vision is Fundamental!

Can’t we just apply standard CV?

First Person Camera

Third Person Camera

Wearable Camera

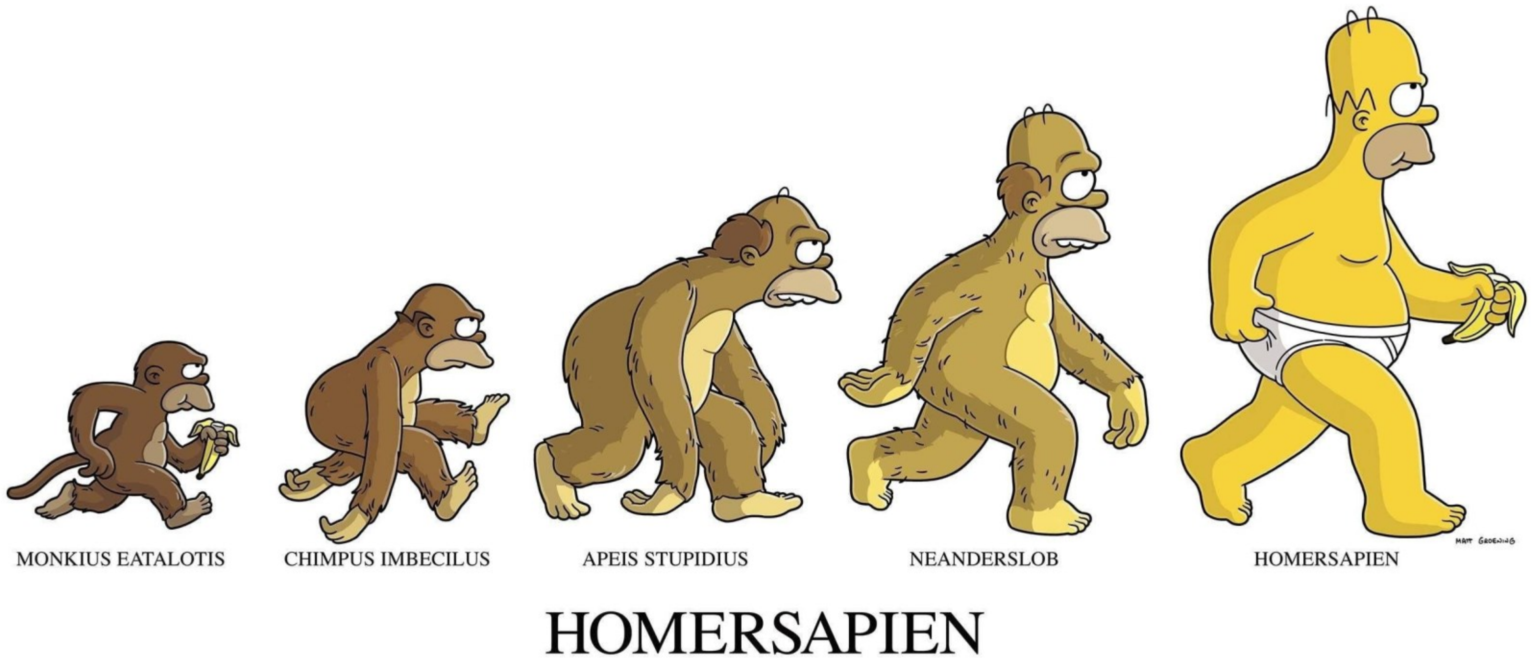


- ✓ Content is always relevant
- ✓ Intrinsically mobile
- × High variability
- × Operational constraints

Fixed Camera

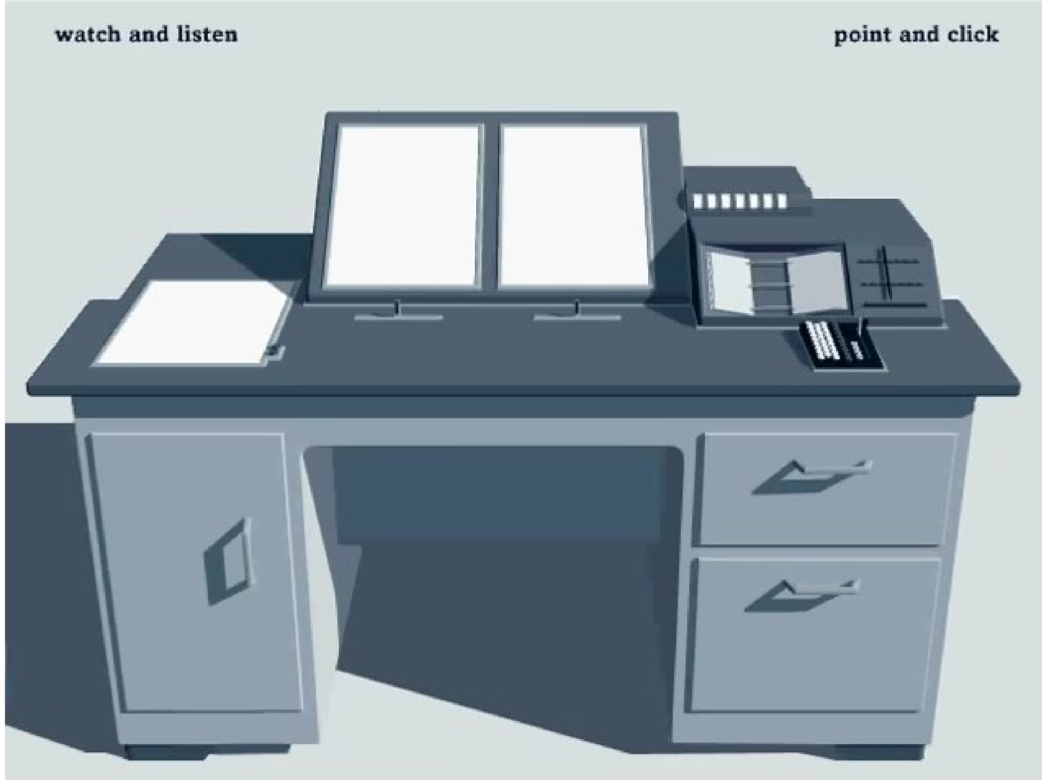
- ✓ Easy to setup
- ✓ Controlled Field of View
- × Doesn’t always see everything
- × Not really portable

When Did it All Begin?

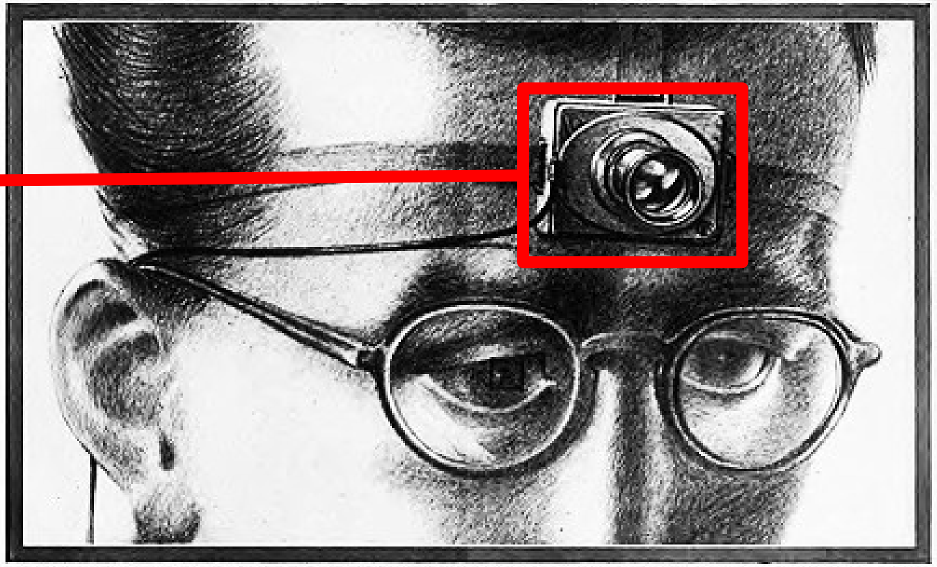


Università di Catania — Bush's Memex, 1945

"Certainly, progress in photography is not going to stop. […] Let us project this trend ahead to a logical, if not inevitable, outcome. The camera hound of the future wears on his forehead a lump a little larger than a walnut."



https://www.youtube.com/watch?v=c539cK58ees



A SCIENTIST OF THE FUTURE RECORDS EXPERIMENTS WITH A TINY CAMERA FITTED WITH UNIVERSAL-FOCUS LENS. THE SMALL SQUARE IN THE EYEGLASS AT THE LEFT SIGHTS THE OBJECT

AS WE MAY THINK

A TOP U. S. SCIENTIST FORESEES A POSSIBLE FUTURE WORLD IN WHICH MAN-MADE MACHINES WILL START TO THINK

by VANNEVAR BUSH

DIRECTOR OF THE OFFICE OF SCIENTIFIC RESEARCH AND DEVELOPMENT

Condensed from the *Atlantic Monthly*, July 1945

This has not been a scientists' war; it has been a war in which all have had a part. The scientists, burying their old professional competition in the demand of a common cause, have shared greatly and learned much. It has been exhilarating to work in effective partnership. What are the scientists to do next?

For the biologists, and particularly for the medical scientists, there can be little indecision, for their war work has hardly required them to leave the old paths. Many indeed have been able to carry on their war research in their familiar peacetime laboratories. Their objectives remain much the same.

It is the physicists who have been thrown most violently off stride, who have left academic pursuits for the making of strange destructive gadgets, who have had to devise new methods for their unanticipated assignments. They have done their part on the devices that made it possible to turn back the enemy. They have worked in combined effort with the physicists of our allies. They have felt within themselves the stir of achievement. They have been part of a great team. Now one asks where they will find objectives worthy of their best.

* * *

There is a growing mountain of research. But there is increased evidence that we are being bogged down today as specialization extends. The investigator is staggered by the findings and conclusions of thousands of other workers—conclusions which he cannot find time to grasp, much less to remember, as they appear. Yet specialization becomes increasingly necessary for progress, and the effort to bridge between disciplines is correspondingly superficial.

Professionally our methods of transmitting and reviewing the results of research are generations old and by now are totally inadequate for their purpose. If the aggregate time spent in writing scholarly works and in reading them could be evaluated, the ratio between these amounts of time might well be startling. Those who conscientiously attempt to keep abreast of current thought, even in restricted fields, by close and continuous reading might well shy away from an examination calculated to show how much of the previous month's efforts could be produced on call.

Mendel's concept of the laws of genetics was lost to the world for a generation because his publication did not reach the few who were capable of grasping and extending it. This sort of catastrophe is undoubtedly being repeated all about us as truly significant attainments become lost in the mass of the inconsequential.

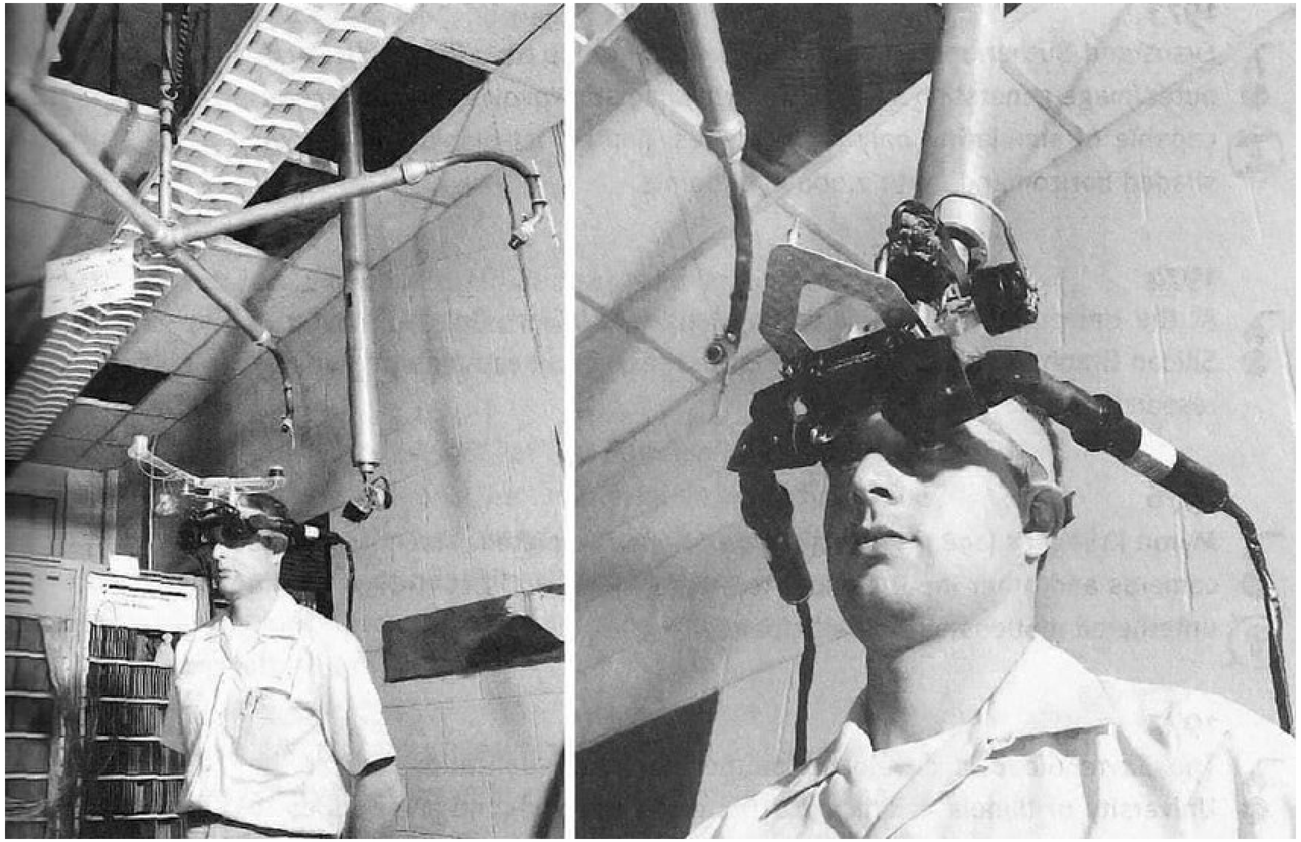
Publication has been extended far beyond our present ability to make real use of the record. The summation of human experience is being expanded at a prodigious rate, and the means we use for threading through the consequent maze to the momentarily important item is the same as was used in the days of square-rigged ships.

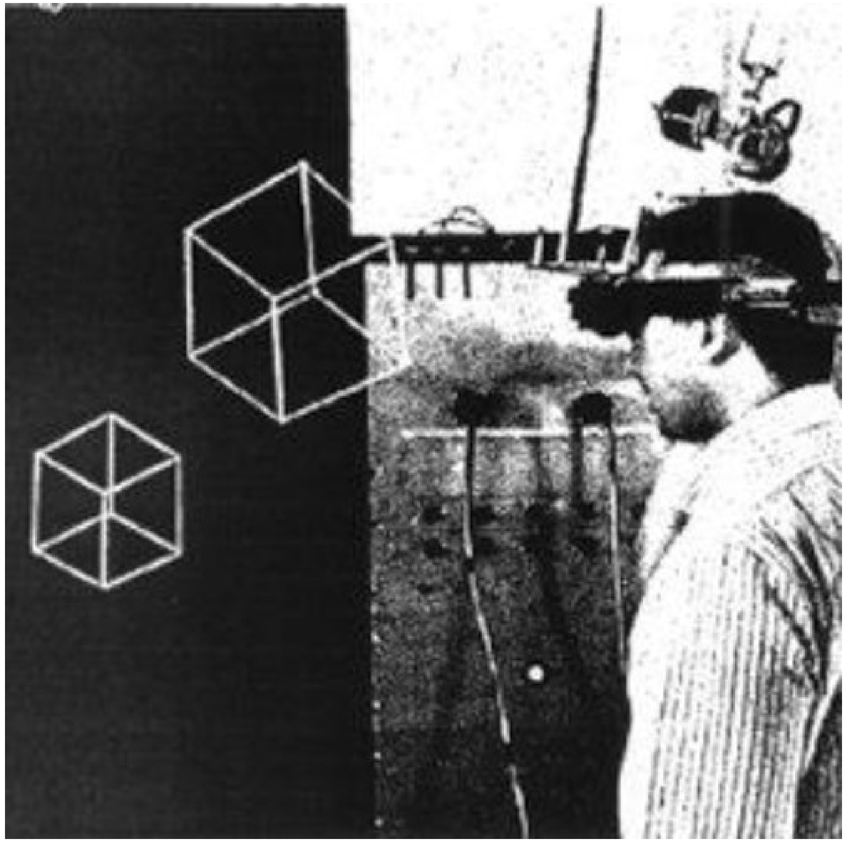
But there are signs of a change as new and powerful instrumentalities come into use. Photocells capable of seeing things in a physical sense, advanced photography which can record what is seen or even what is not, thermionic tubes capable of controlling potent forces under the guidance of

112

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

Steve Mann's "wearable computer" and "reality mediator" inventions of the 1970s have evolved into what looks like ordinary eyeglasses.

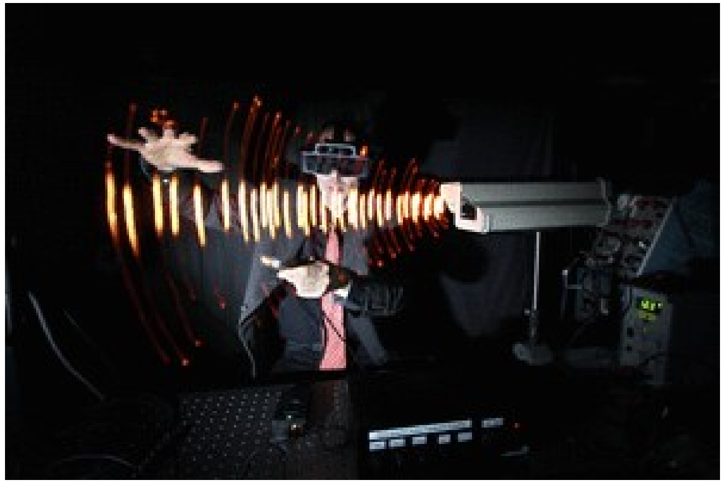
(a)	(b)	(c)	(d)	(e)
1980	Mid 1980s	Early 1990s	Mid 1990s	Late 1990s

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 werable screen for feedback. **Steve Mann is often referred to as «the father of wearable computing»**

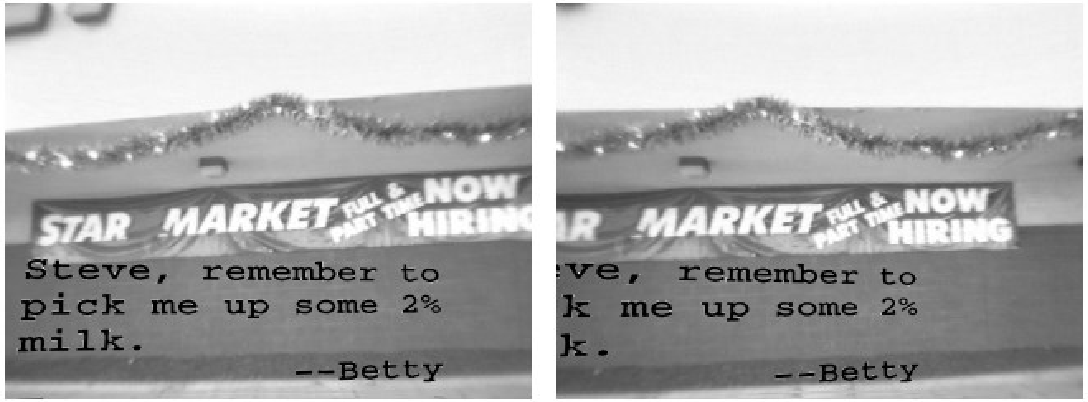
Università di Catania

First Wearable Computing Applications

Visual Orbits

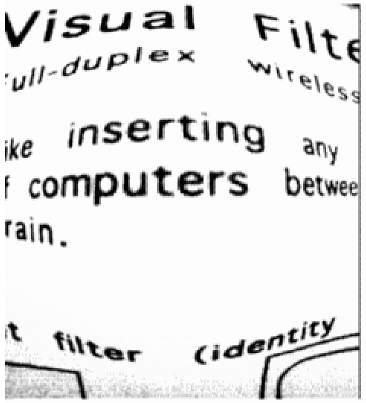


Meta-Vision



Spatialized Reminders

Spatialized Shopping List



Visual Filters

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

Egocentric Computer Vision: The Goal

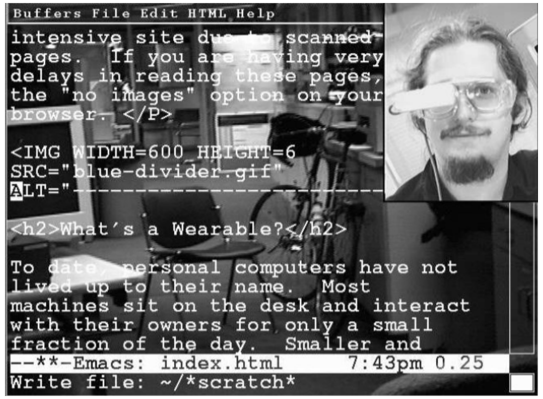
Clip from movie Terminator 2-Judgment day: https://youtu.be/9MeaaCwBW28
Ref: https://www.redsharknews.com/vr_and_ar/item/3539-terminator-2-vision-the-augmented-reality-standard-for-25-years

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

(augumented reality)



1997

1998

Visual Contextual Awareness in Wearable Computing

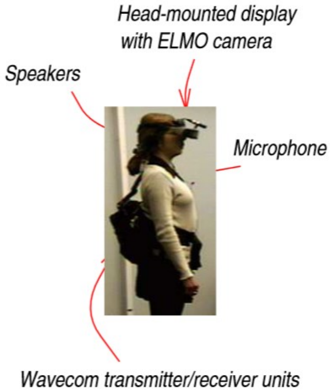
Thad Starner Bernt Schiele Alex Pentland
Media Laboratory, Massachusetts Institute of Technology

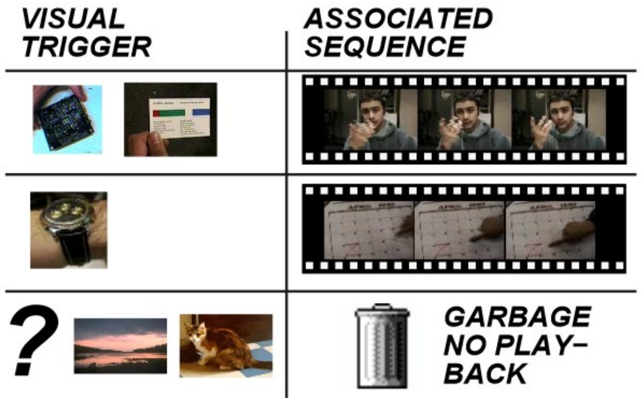
(location and task recognition)

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)





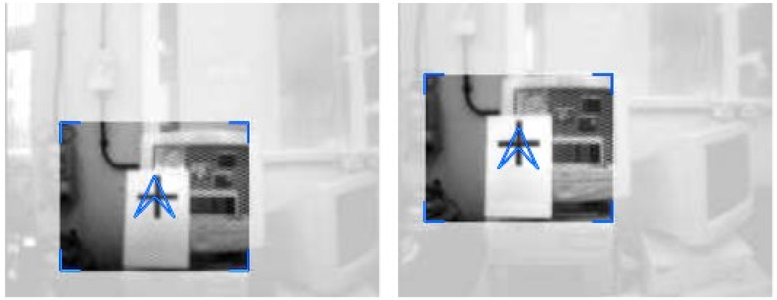
1999

Early 2000s

Wearable Visual Robots

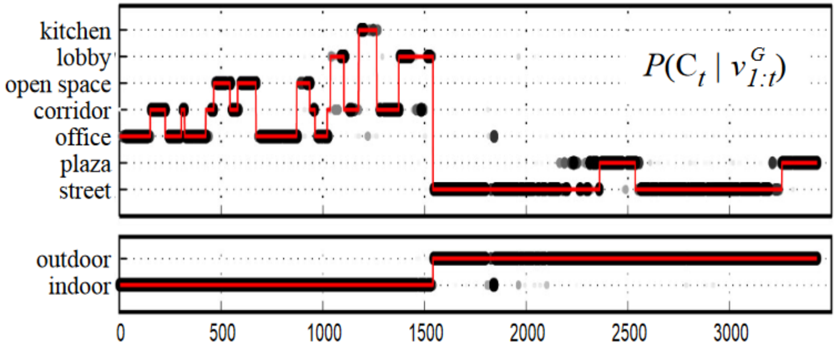
W.W. Mayol, B. Tordoff and D.W. Murray
University of Oxford, Parks Road, Oxford OX1 3PJ, UK

(active vision)



2000

2003



Context-based vision system for place and object recognition

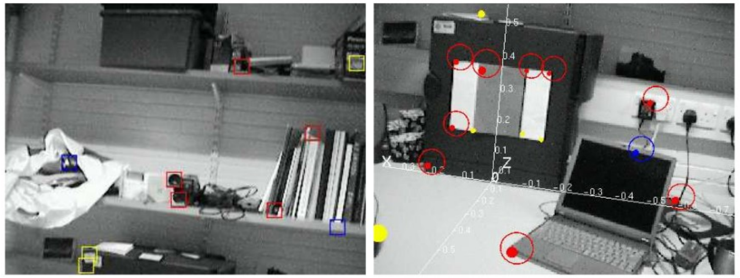
Antonio Torralba, MIT AI lab, Cambridge, MA 02139
Kevin P. Murphy, MIT AI lab, Cambridge, MA 02139
William T. Freeman, MIT AI lab, Cambridge, MA 02139
Mark A. Rubin, Lincoln Labs, Lexington, MA 02420

(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

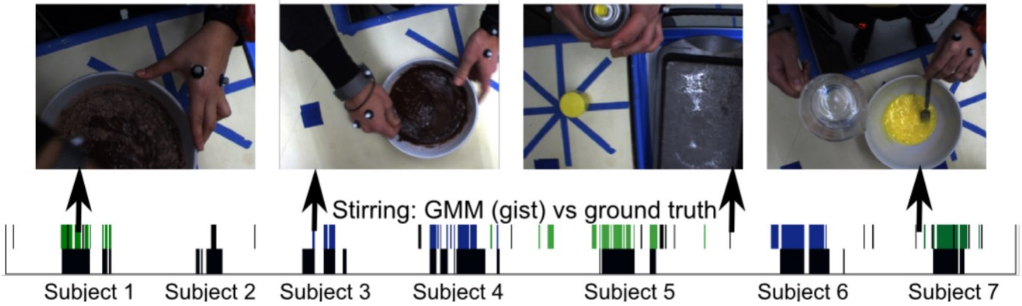
W.W. Mayol
Department of Computer Science
University of Bristol

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

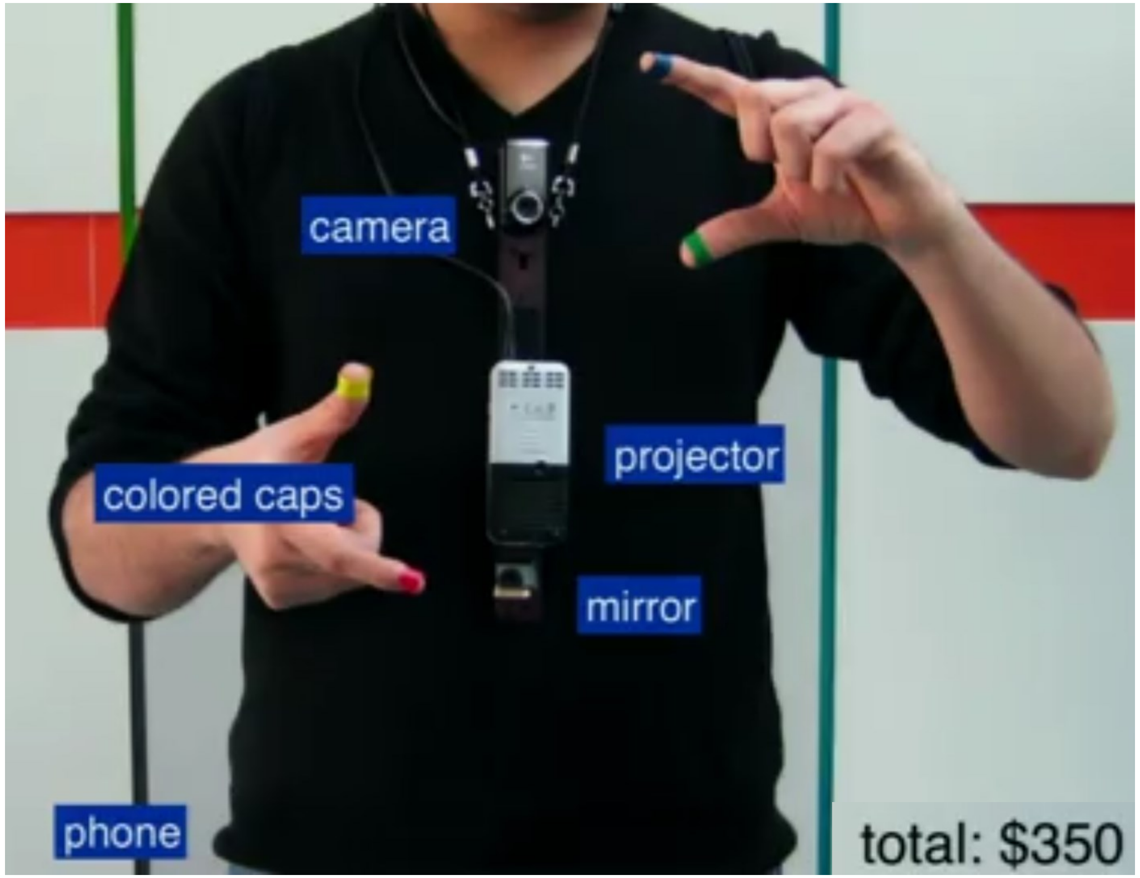
(handheld object recognition)

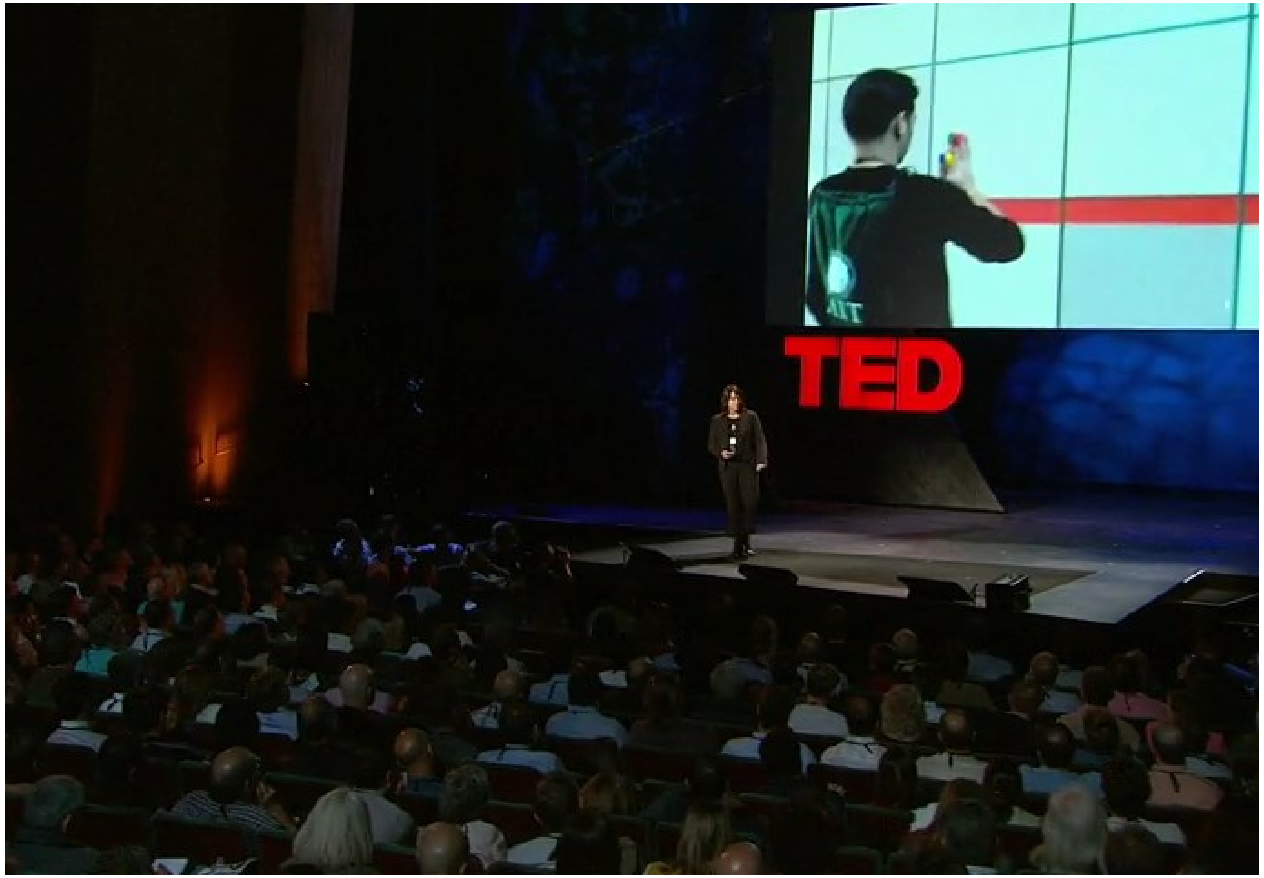
2010

Sixth Sense, 2009

Neck worn camera with a projector and a gesture-based user interface.

«to give people access to information without requiring that the user changes any of their behavior»





Pattie Maes & Pranav Mistry (MIT) @ TED

https://www.ted.com/talks/pattie_maes_demos_the_sixth_sense

Hardware, 1990s – 2000s

A COMMON HARDWARE PLATFORM WAS MISSING!

Microsoft SenseCam, 2004

"A day in Rome"

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

- 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.

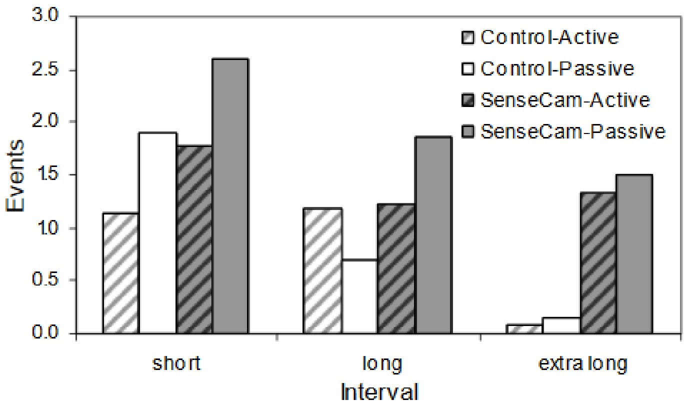
Bell, Gordon, and Jim Gemmell. *Your life, uploaded: The digital way to better memory, health, and productivity*. Penguin, 2010.

Research using Microsoft SenseCam

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

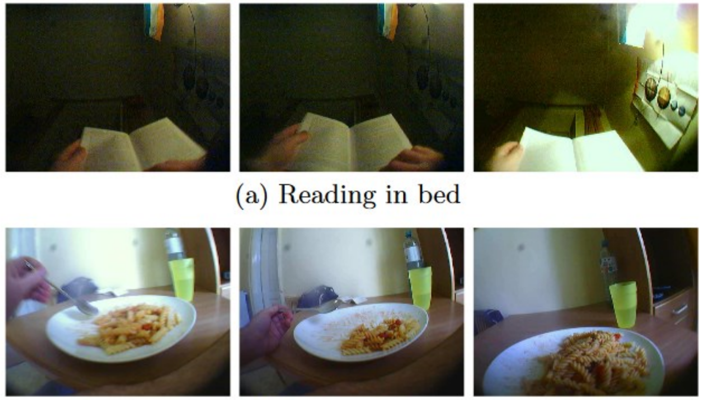
Abigail Sellen, Andrew Fogg, Mike Aitken*, Steve Hodges, Carsten Rother and Ken Wood
Microsoft Research Cambridge
7 JJ Thomson Ave, Cambridge, UK, CB3 0FB
*Behavioural & Clinical Neuroscience Institute
Dept. of Psychology, University of Cambridge

(health, memory augmentation)



2007

2008



(a) Reading in bed

(b) Having dinner

MyPlaces: Detecting Important Settings in a Visual Diary

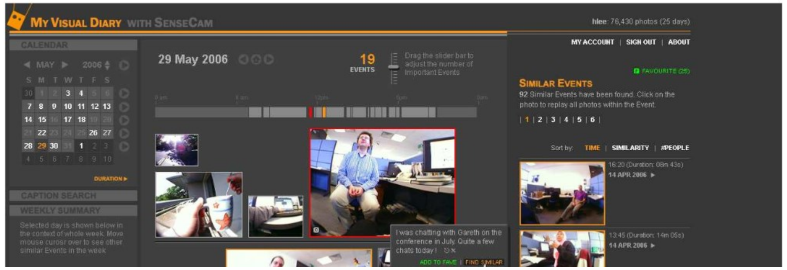
Michael Blighe and Noel E. O'Connor
Centre for Digital Video Processing, Adaptive Information Cluster
Dublin City University, Ireland
{blighem, oconnorn}@eeng.dcu.ie

(lifelogging, place recognition)

Constructing a SenseCam Visual Diary as a Media Process

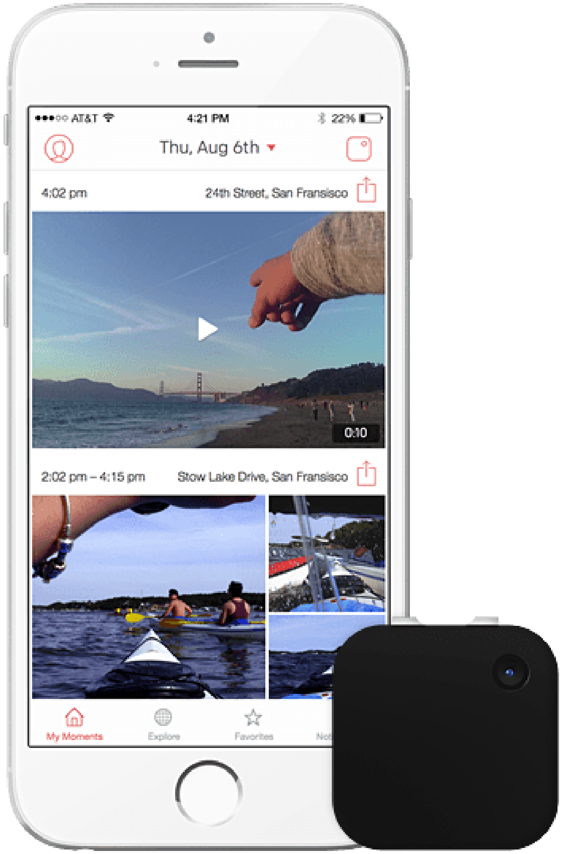
Hyowon Lee, Alan F. Smeaton, Noel O'Connor, Gareth Jones, Michael Blighe, Daragh Byrne, Aiden Doherty, and Cathal Gurrin
Centre for Digital Video Processing & Adaptive Information Cluster,
Dublin City University

(lifelogging, multimedia retrieval)



2008

Narrative Clip, 2012



http://getnarrative.com/

Università di Catania

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)

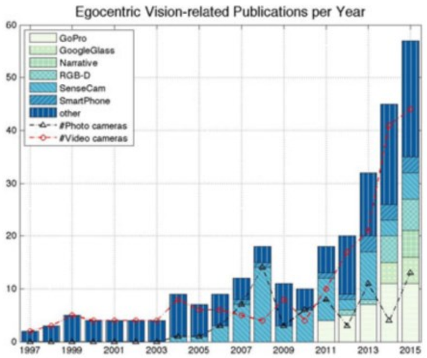
2016

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?

GoPro HD Hero, 2010

different wearing modalities

head-mounted

chest-mounted

wrist-mounted

helmet-mounted

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

Università di Catania

Early Research On Egocentric Video

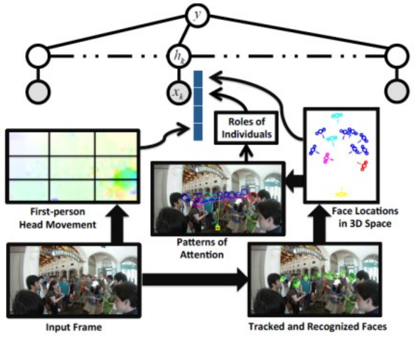
2011

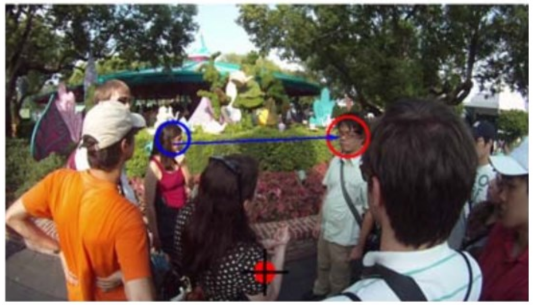
Fast Unsupervised Ego-Action Learning for First-Person Sports Videos

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)

2012





Social Interactions: A First-Person Perspective

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

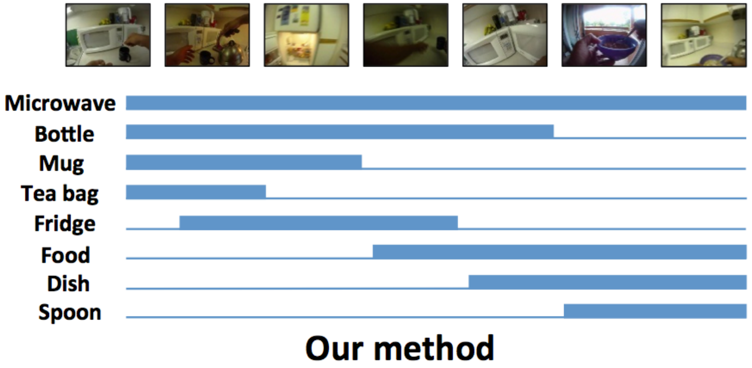
(detection and recognition of social interactions)

2013

Story-Driven Summarization for Egocentric Video

Zheng Lu and Kristen Grauman
University of Texas at Austin

(egocentric video sumarization)

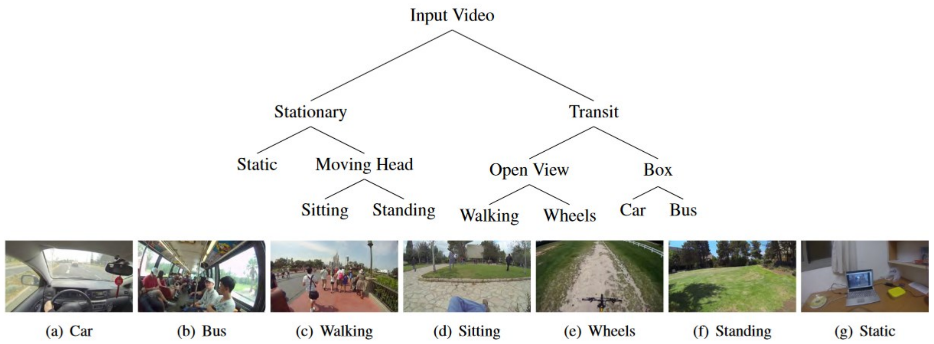


Later Research On Egocentric Video

Temporal Segmentation of Egocentric Videos

Yair Poleg   Chetan Arora*   Shmuel Peleg

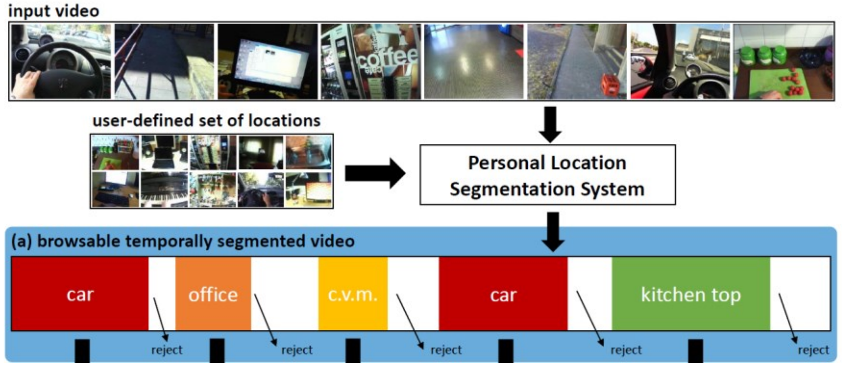
(egocentric video indexing)



(a) Car (b) Bus (c) Walking (d) Sitting (e) Wheels (f) Standing (g) Static

2014

2016



Recognizing Personal Locations from Egocentric Videos

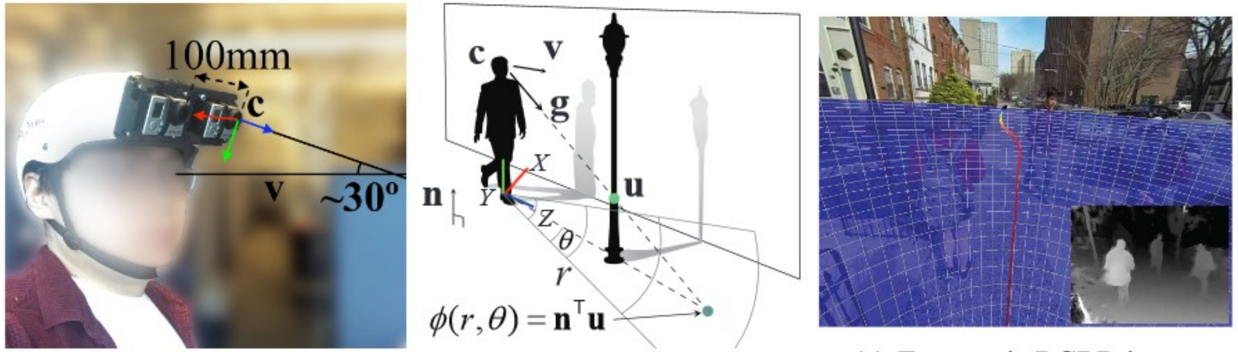
Antonino Furnari, Giovanni Maria Farinella, *Senior Member, IEEE,* and Sebastiano Battiato, *Senior Member, IEEE*

(localization, indexing, context-aware computing)

Egocentric Future Localization

Hyun Soo Park   Jyh-Jing Hwang   Yedong Niu   Jianbo Shi

(future localization, navigation)



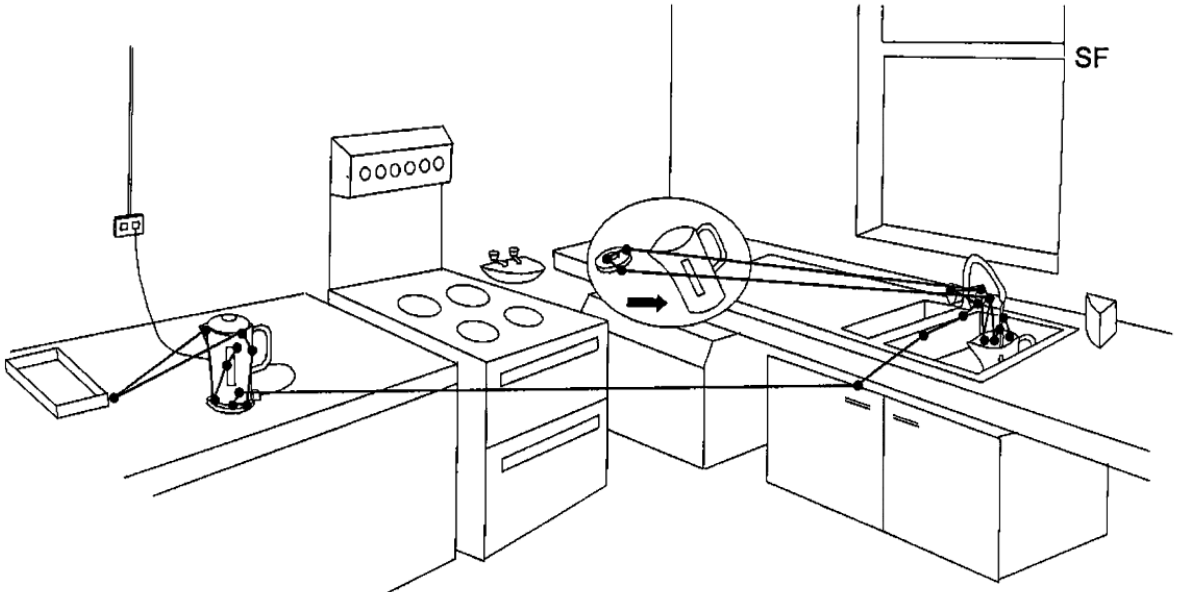
(a) Ego-stereo cameras (b) Geometry (c) Egocentric RGBD image

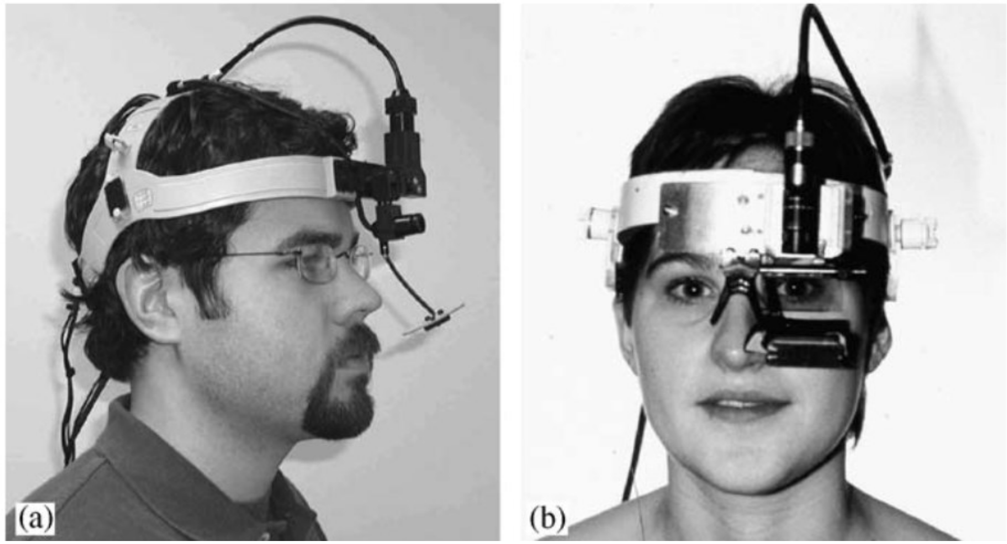
2016

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)

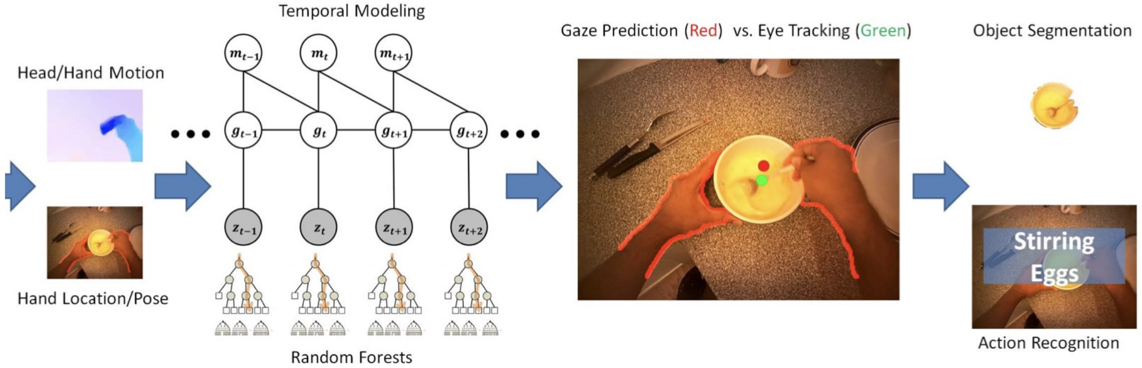
Pupil Eye Trackers (2014 -)

Research On Gaze

Learning to Predict Gaze in Egocentric Video

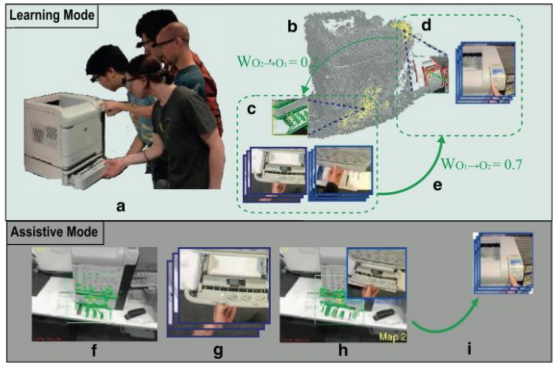
Yin Li, Alireza Fathi, James M. Rehg

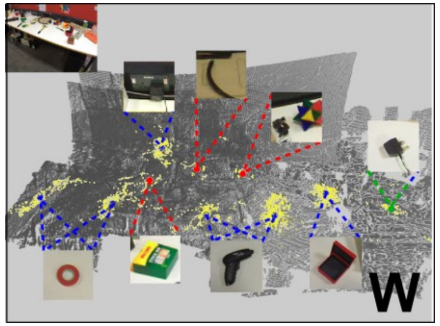
(gaze prediciton, action recognition)



2012

2016





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

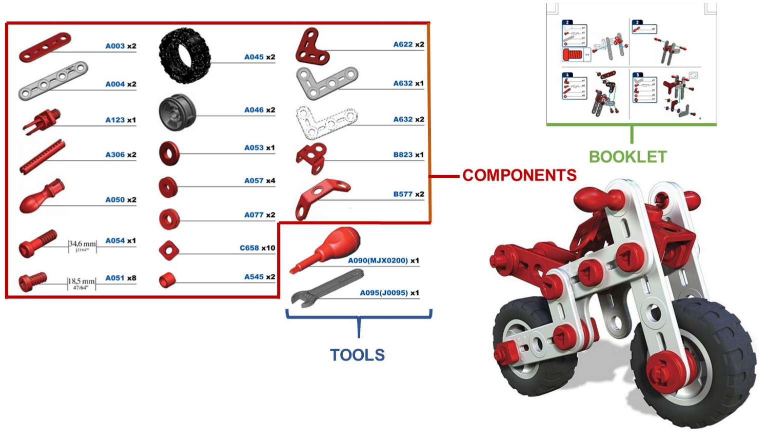
Dima Damen*, Teesid Leelasawassuk, Walterio Mayol-Cuevas

(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

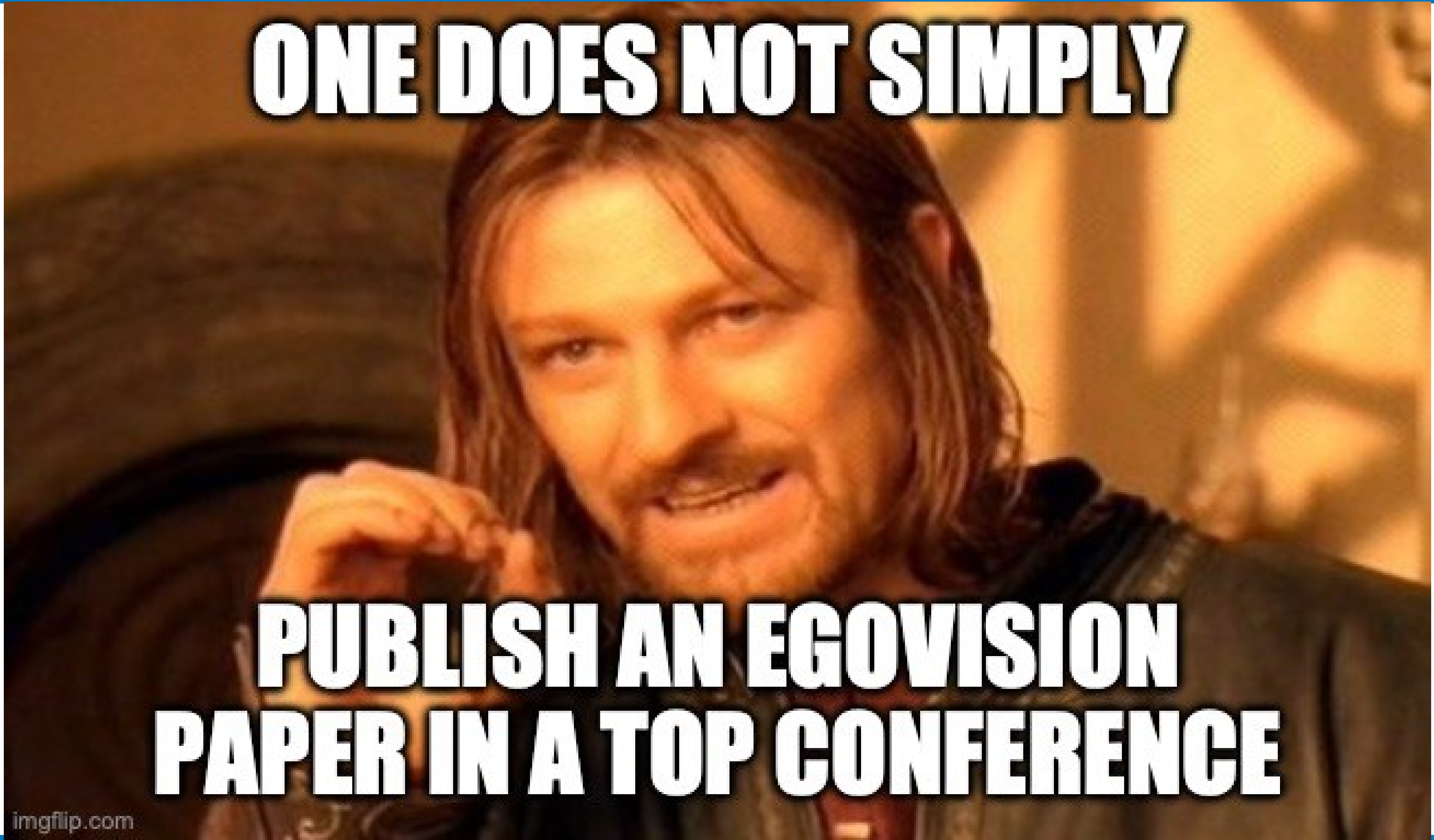
Circa 2017 – most of the discussion still in workshops

Workshop on Egocentric (First Person) Vision

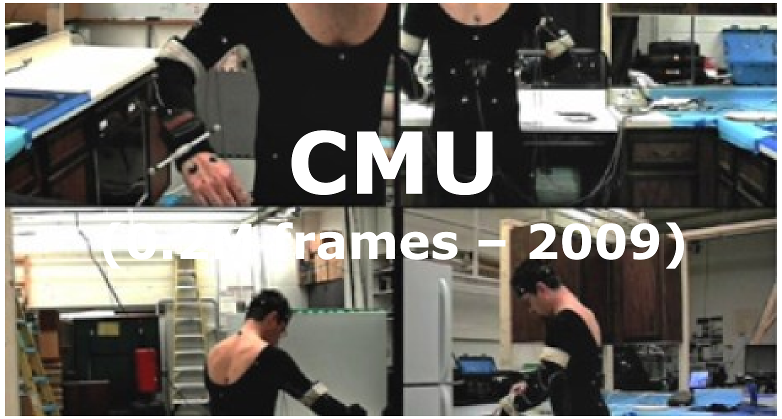
ACVR

EPIC

EGO4D

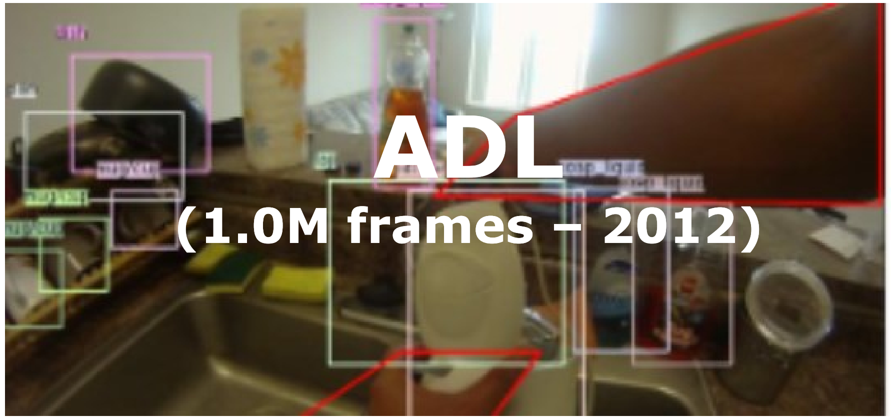

ONE DOES NOT SIMPLY
PUBLISH AN EGOVISION
PAPER IN A TOP CONFERENCE
imgflip.com

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/

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

EPIC-KITCHENS

TEAM

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.

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.

2024 Challenges

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

EPIC-KITCHENS Workshops & Challenges

EPIC-KITCHENS-100- 2022 Challenges Report

EPIC-KITCHENS-100- 2023 Challenges Report

Dima Damen, Jacob Chalk, Ahmad Darkhalil, Toby Perrett, Daniel Whettam, Saptarshi Sinha, Michael Wray, Bin Zhu
University of Bristol, UK

Antonino Furnari, Francesco Ragusa, Giovanni Maria Farinella
University of Catania, Italy

Dandan Shan, David Fouhey
University of Michigan, US

Matteo Dunnhofer, Christian Micheloni
University of Udine, Italy

Jaesung Huh, Andrew Zisserman
University of Oxford, UK

Abstract

This report presents the findings from the 5th EPIC-KITCHENS-100 challenges, opened from Jan 2023 and concluded on the 1st of June 2023. It serves as an introduction to all technical reports that were submitted to the 11th EPIC@CVPR2023 workshop, and an official announcement of the winners.

The report covers 4 new challenges, announced for the first time in the 2023 round as well as 5 recurring challenges

1. Datasets

The challenges cover three datasets publicly available.

The 5 recurring challenges are based on the publicly available EPIC-KITCHENS-100 dataset. In summary, EPIC-KITCHENS-100 provides 20M frames of egocentric footage, captured in an unscripted manner, with carefully collated annotations of 90K fine-grained actions. Details of how the dataset was collected and annnotated are available in our IJCV paper [8]. The challenges are: Action Recognition, Action Anticipation, Action Detection, Unsupervised

The TREK-150 Object Tracking is based on the TREK-150 dataset [20] released in 2021. The challenge focuses on single object tracking in egocentric videos.

Finally, one challenge is based on EPIC-SOUNDS [27]. Released last year as well, EPIC-SOUNDS annotates 78.4k categorised segments of audible events and actions, distributed across 44 classes. A challenge on audio-only action recognition was run this year as: EPIC-SOUNDS Audio-Based Interaction Recognition.

Each challenge we released codebase with pre-trained models, features and evaluation scripts:

- Action Recognition at https://github.com/epic-kitchens/C1-Action-Recognition: Five pre-trained models were made available using the codebases: TSN, TRN, TBN, TSM and SlowFast, as well as evaluation script.
- Action Detection at https://github.com/epic-kitchens/C2-Action-Detection: with pre-extracted features, a baseline using BMN model and evaluation script.
- Action Anticipation at https://github.com/epic-kitchens/C3-Action-Anticipation with pre-extracted features, RULSTM base model and


IMPRESSIVE, MOST IMPRESSIVE
BUT YOU ARE NOT A JEDI YET
makeameme.org

Can We Scale?

EGO4D

Consortium

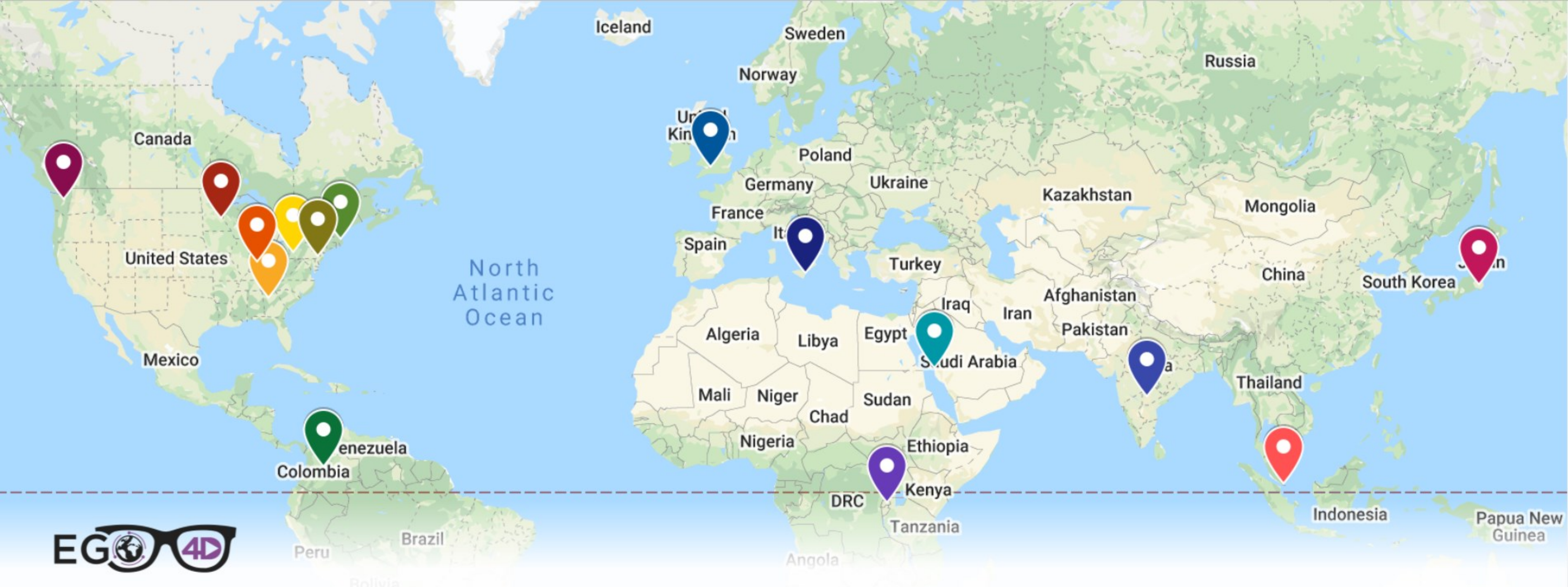
Carnegie Mellon University
Università di Catania
NUS National University of Singapore
King Abdullah University of Science and Technology
東京大学 THE UNIVERSITY OF TOKYO
University of BRISTOL
INDIANA UNIVERSITY BLOOMINGTON
UNIVERSITY OF MINNESOTA
MIT
Penn UNIVERSITY of PENNSYLVANIA
Georgia Institute of Technology
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

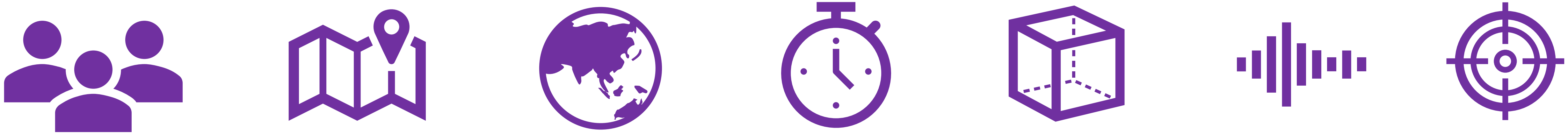
84 authors

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.



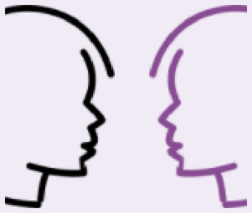
855 Subjects | 74 Locations | 9 Countries | 3025 Hours | 3D Scans | Audio | Gaze

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

Happy Ending?

EGO-EXO4D

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

NUS National University of Singapore

Carnegie Mellon University Africa

International Institute of Information Technology Hyderabad

東京大学 The University of Tokyo

University of Minnesota

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Penn University of Pennsylvania

University of Bristol

Indiana University Bloomington

The University of North Carolina at Chapel Hill

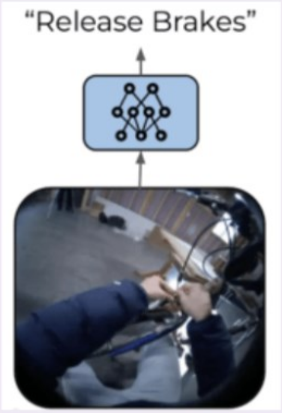
University of Illinois Urbana-Champaign

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

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

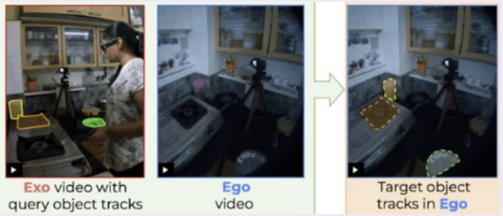
Kristen Grauman[1,2], Andrew Westbury[1], Lorenzo Torresani[1], Kris Kitani[1,3], Jitendra Malik[1,4], Triantafyllos Afouras[*1], Kumar Ashutosh[*1,2], Vijay Baiyya[*5], Siddhant Bansal[*6,7], Bikram Boote[*8], Eugene Byrne[*1,9], Zach Chavis[*10], Joya Chen[*11], Feng Cheng[*1], Fu-Jen Chu[*1], Sean 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

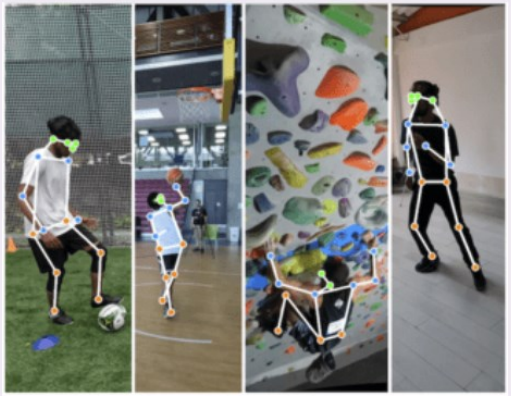


Keystep Recognition

Proficiency Estimation



Relation



Pose Estimation

First Joint Egocentric Vision (EgoVis) Workshop

Held in Conjunction with CVPR 2024

17 June 2024 - Seattle, USA

EGO-EXO4D

EGO4D

EPIC KITCHENS

Ego-Exo4D

Ego4D

EPIC-Kitchens

Università di Catania

What about Industrial Domain?

MECCANO: A Multimodal Egocentric Dataset for Humans Behavior Understanding in the Industrial-like Domain

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

[1]FPV@IPLab, Department of Mathematics and Computer Science - University of Catania, Italy
[2]Next Vision s.r.l., Spin-off of the University of Catania, Italy

Running ICIAP competition with Prize!

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

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

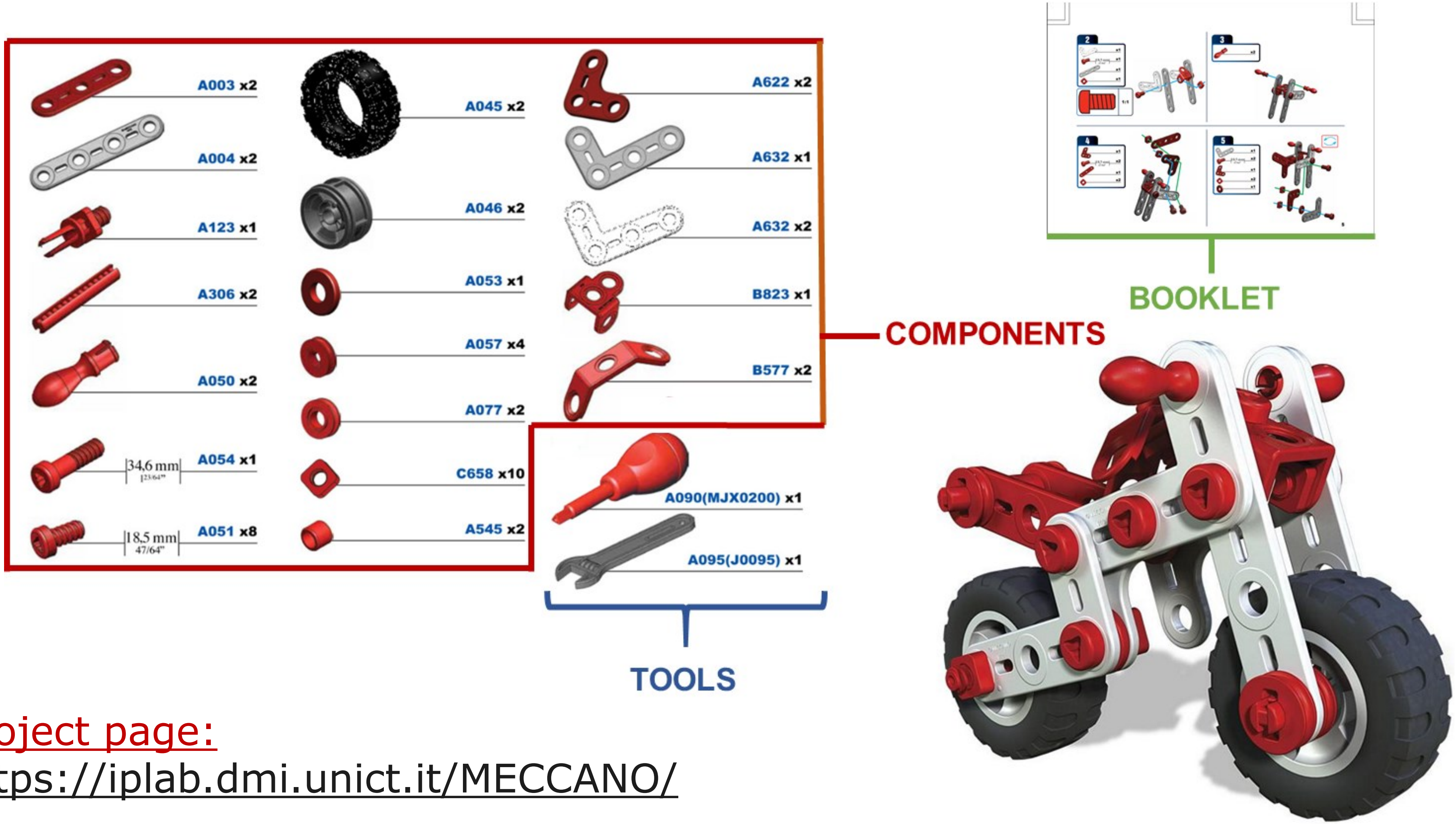
Abstract

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

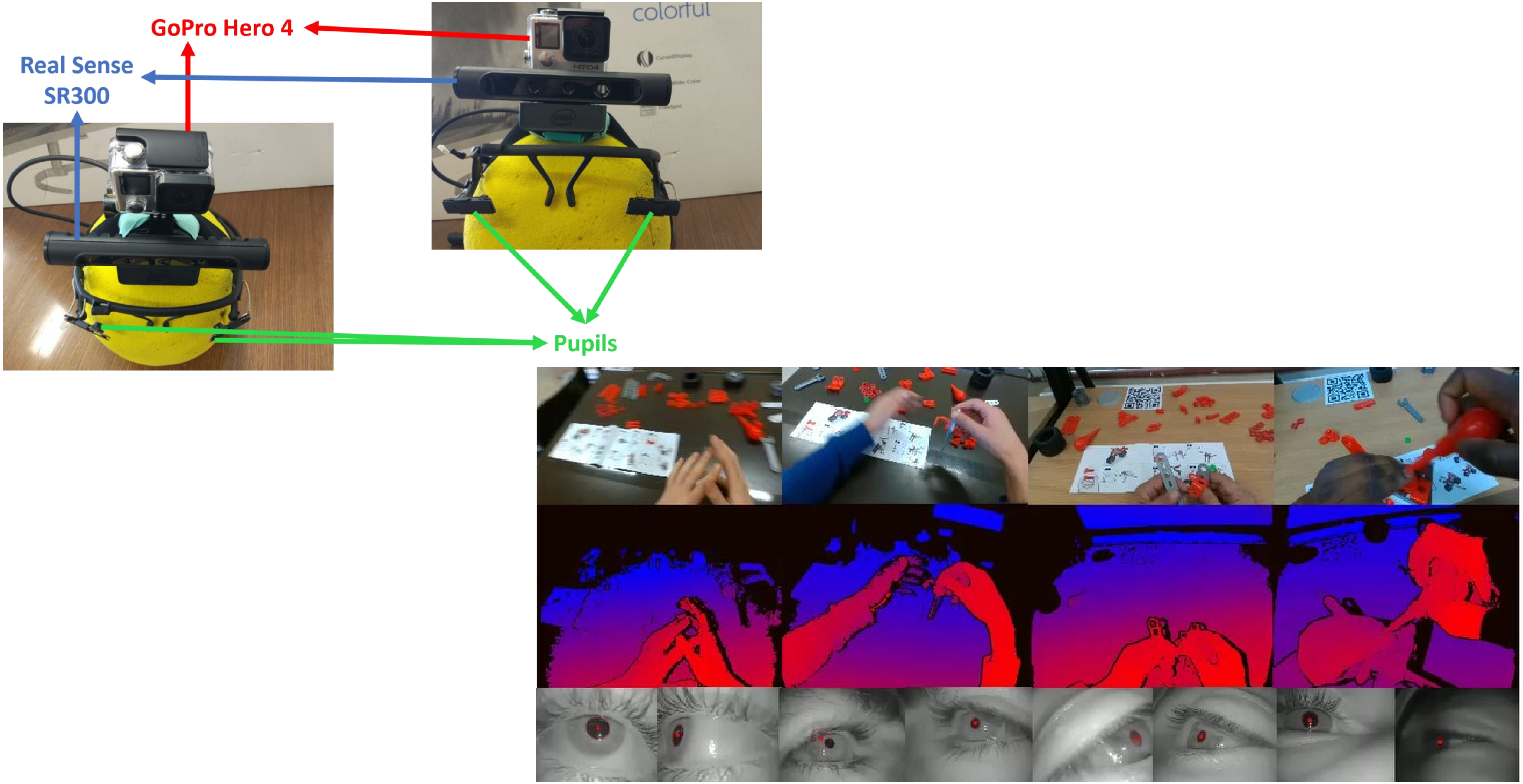
The MECCANO Dataset



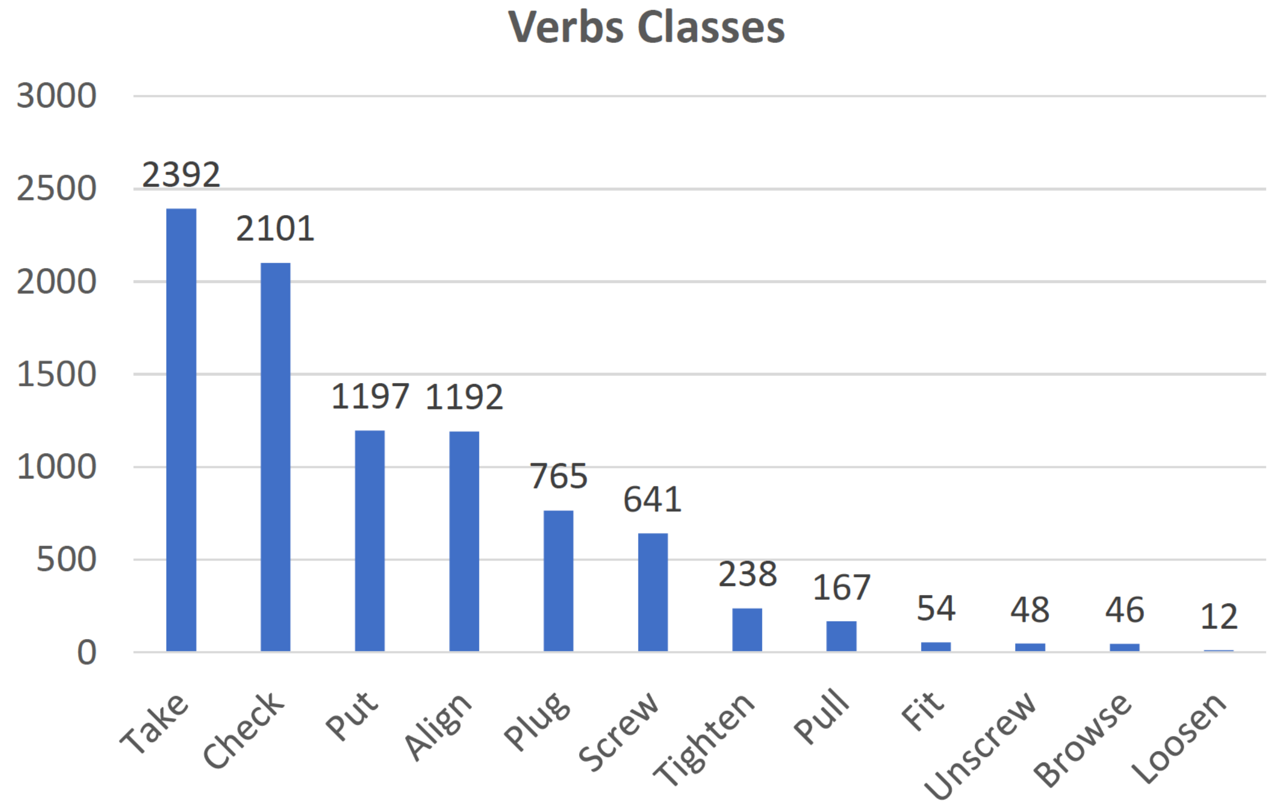
Project page:
https://iplab.dmi.unict.it/MECCANO/

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).

Data Acquisition

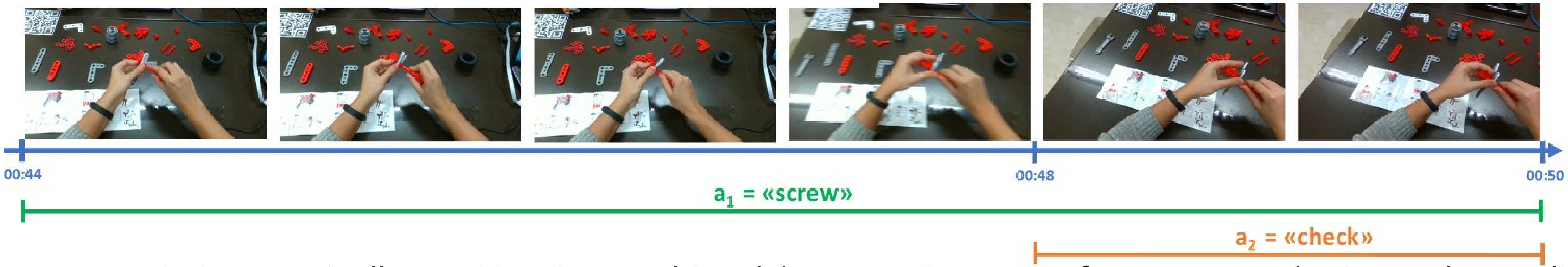


Data Annotation: Temporal Verb Annotations



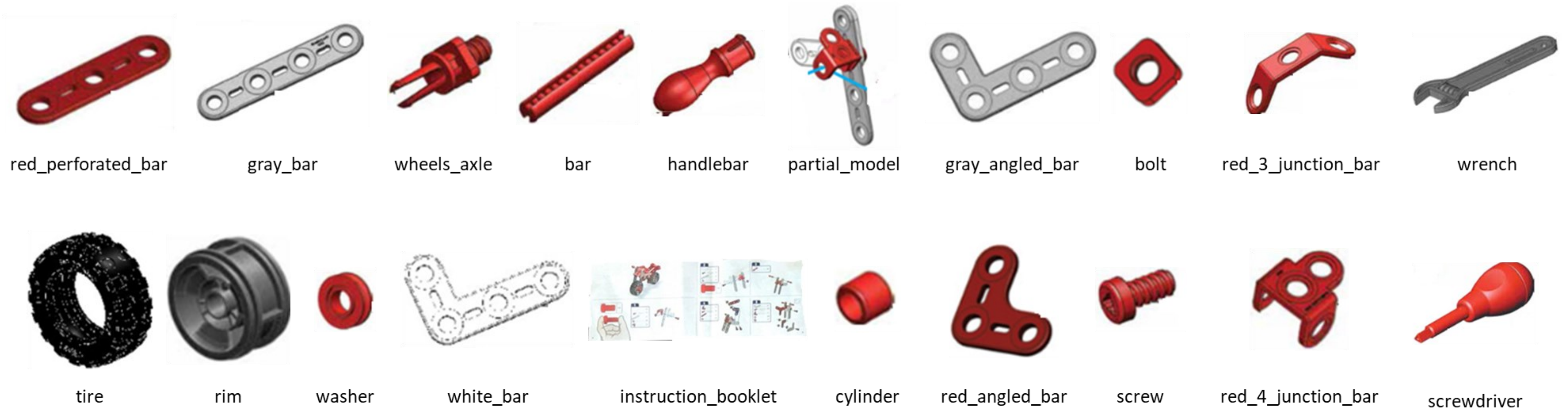
8857 video segments

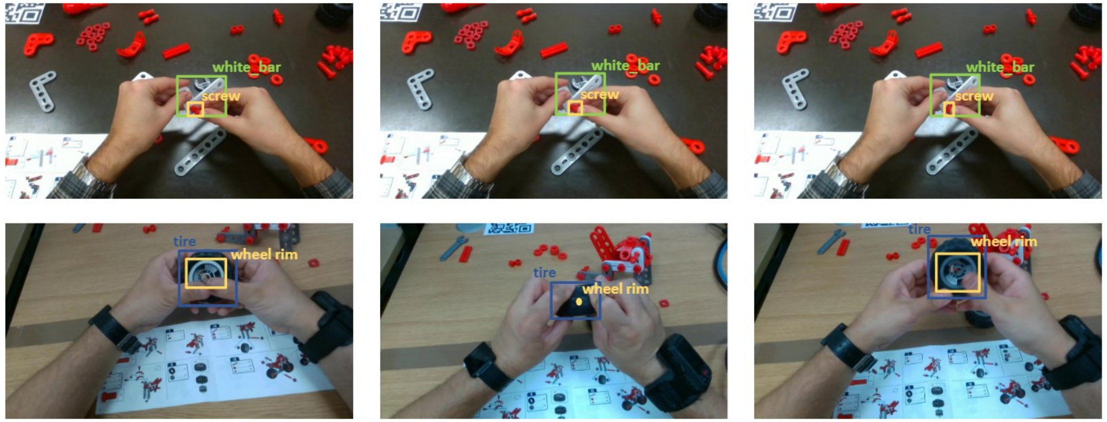
1401 overlap segments (15.82%)



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).

Data Annotation: Active Object Bounding Boxes

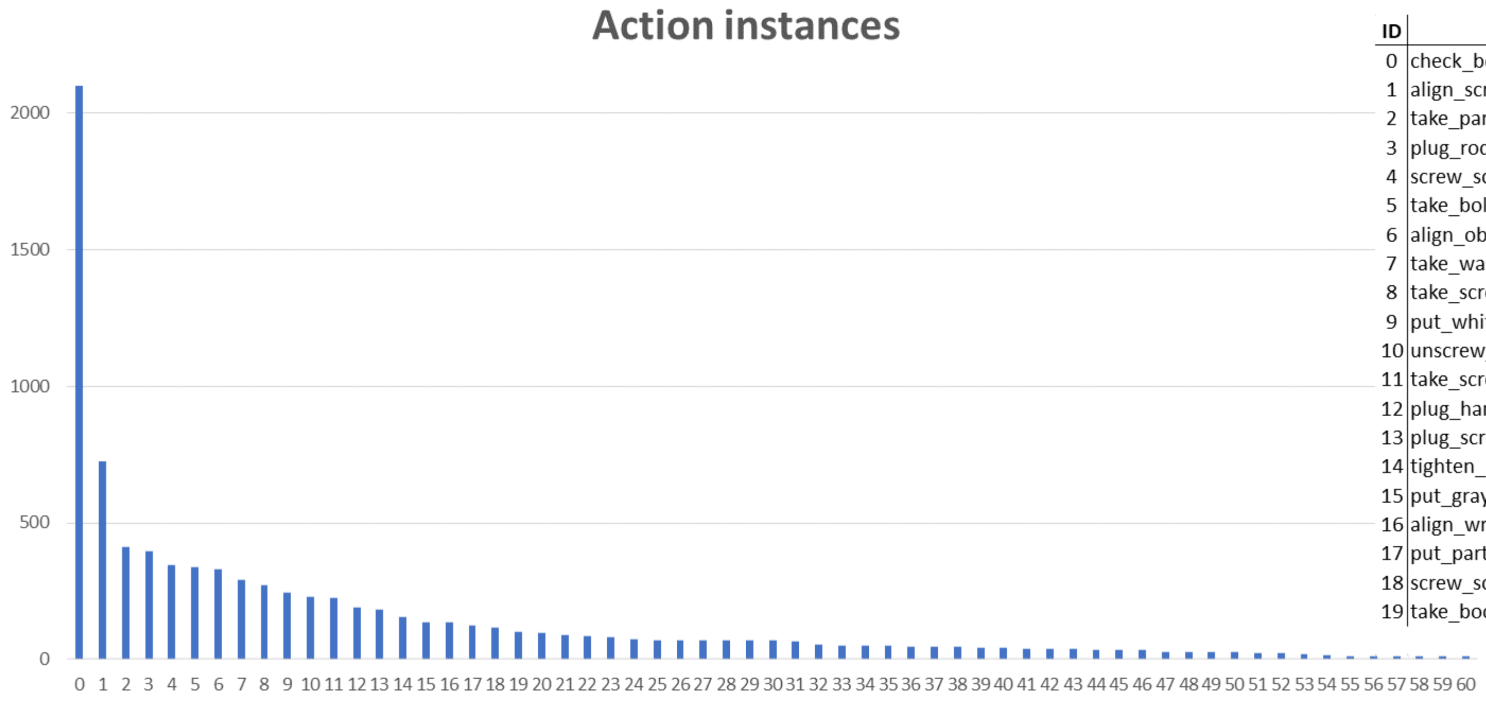




64439 frames

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	ID	Action	ID	Action
0	check_booklet	20	put_screwdriver	40	take_red_perforated_junction_bar
1	align_screwdriver_to_screw	21	put_red_perforated_junction_bar	41	fit_rim_tire
2	take_partial_model	22	put_gray_angled_perforated_bar	42	take_rim
3	plug_rod	23	take_red_perforated_bar	43	take_red_4_perforated_junction_bar
4	screw_screw_with_screwdriver	24	take_gray_perforated_bar	44	put_screw
5	take_bolt	25	take_red_angled_perforated_bar	45	put_rod
6	align_objects	26	tighten_nut_with_hands	46	put_washer
7	take_washer	27	take_white_angled_perforated_bar	47	unscrew_screw_with_screwdriver
8	take_screw	28	take_rod	48	put_red_perforated_bar
9	put_white_angled_perforated_bar	29	put_tire	49	put_wrench
10	unscrew_screw_with_hands	30	put_roller	50	put_bolt
11	take_screwdriver	31	pull_partial_model	51	take_wheels_axle
12	plug_handlebar	32	pull_screw	52	put_wheels_axle
13	plug_screw	33	take_gray_angled_perforated_bar	53	put_red_angled_perforated_bar
14	tighten_nut_with_wrench	34	take_tire	54	put_red_4_perforated_junction_bar
15	put_gray_perforated_bar	35	pull_rod	55	take_objects
16	align_wrench_to_bolt	36	take_wrench	56	put_objects
17	put_partial_model	37	browse_booklet	57	loosen_bolt_with_hands
18	screw_screw_with_hands	38	take_roller	58	put_booklet
19	take_booklet	39	take_handlebar	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).

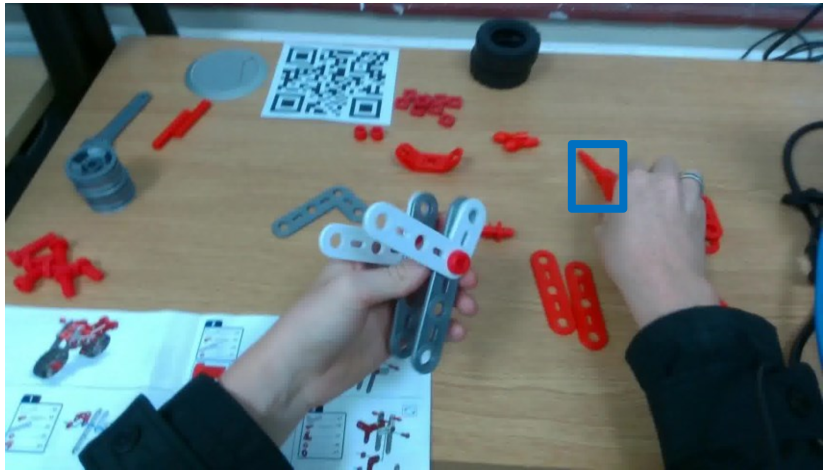
Data Annotation: Egocentric Human-Object Interactions

Egocentric Human-Object Interaction

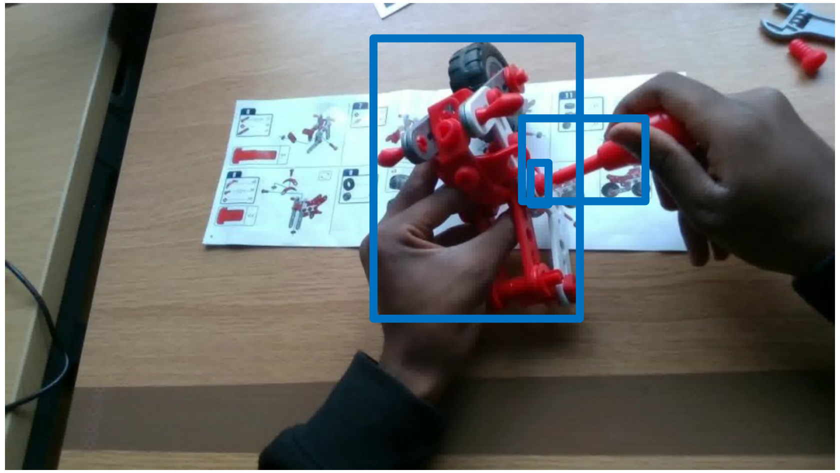
$O = \{o_1, o_2, ..., o_n\}$

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

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

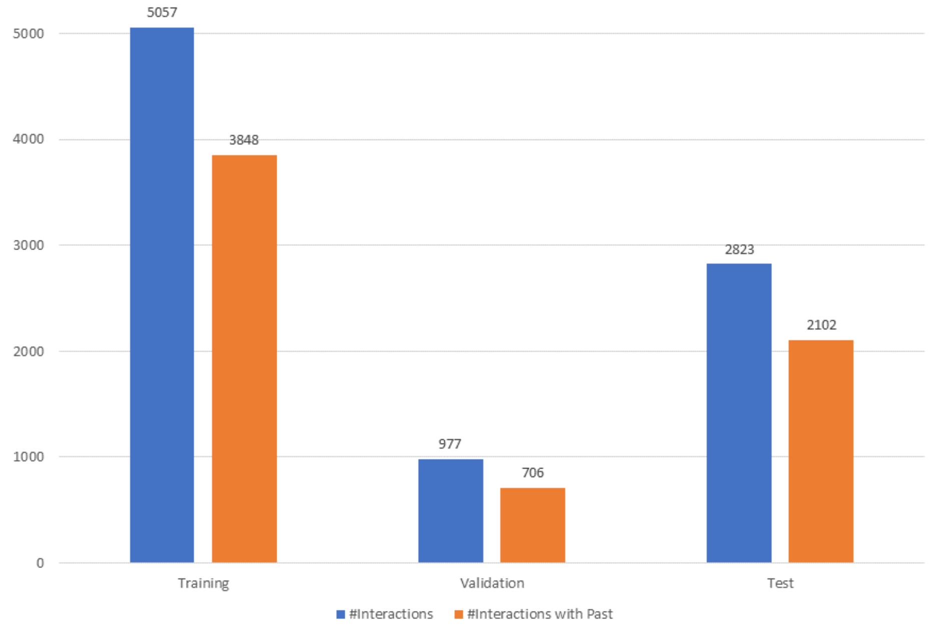


<take, screwdriver>



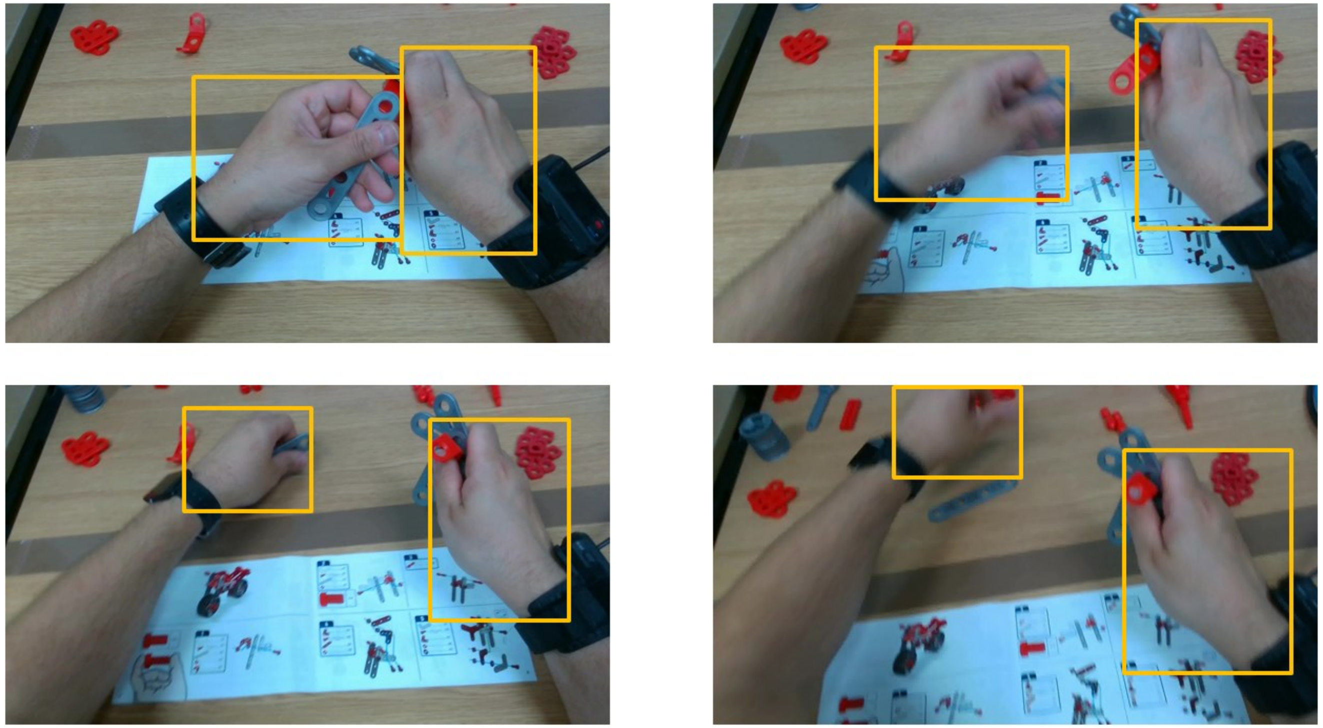
<screw, {screwdriver, screw, partial_model}>

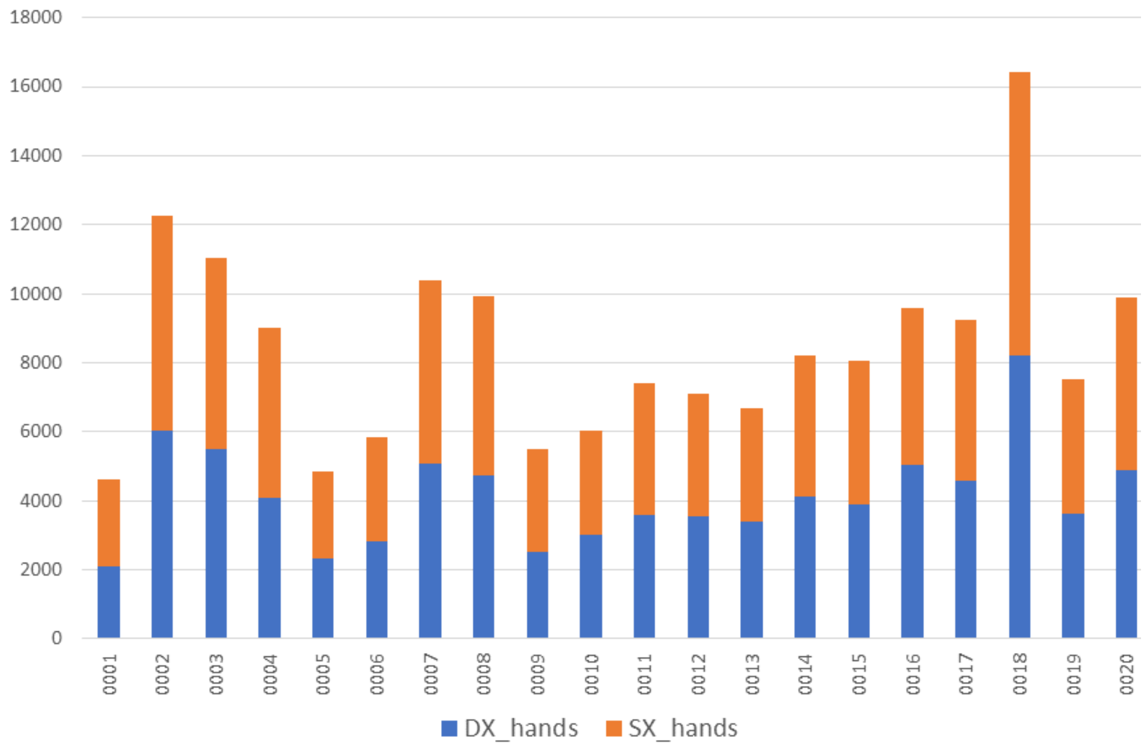
Data Annotation: Next Active Object Annotations



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

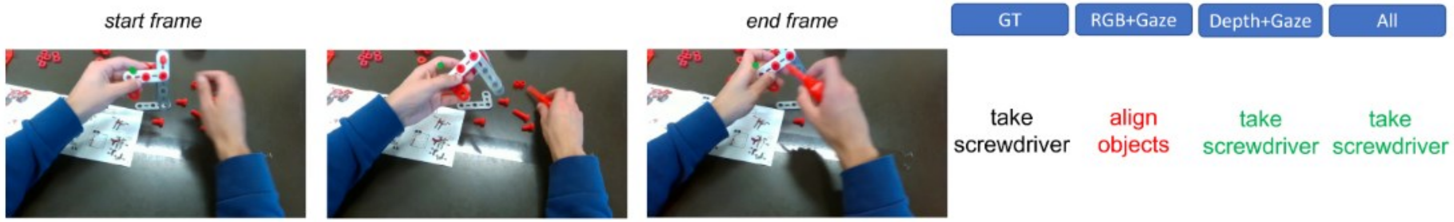
Data Annotation: Hands Annotations



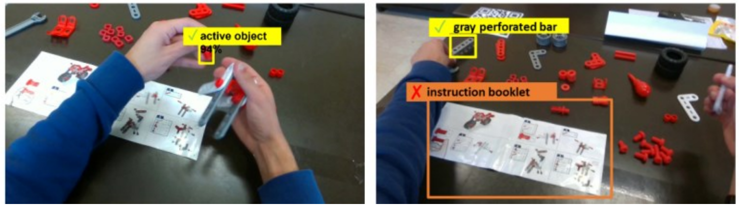


MECCANO Benchmark

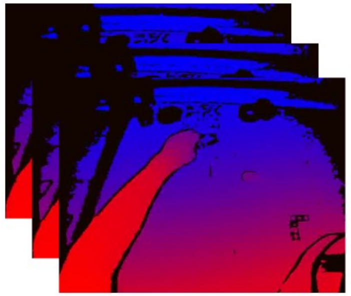
1) Action Recognition

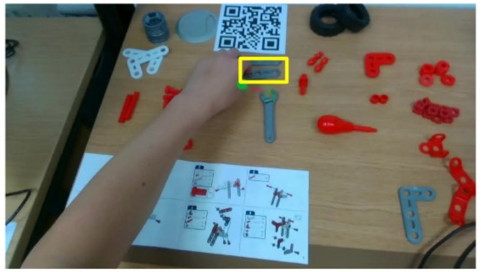


2) Active Object Detection and Recognition

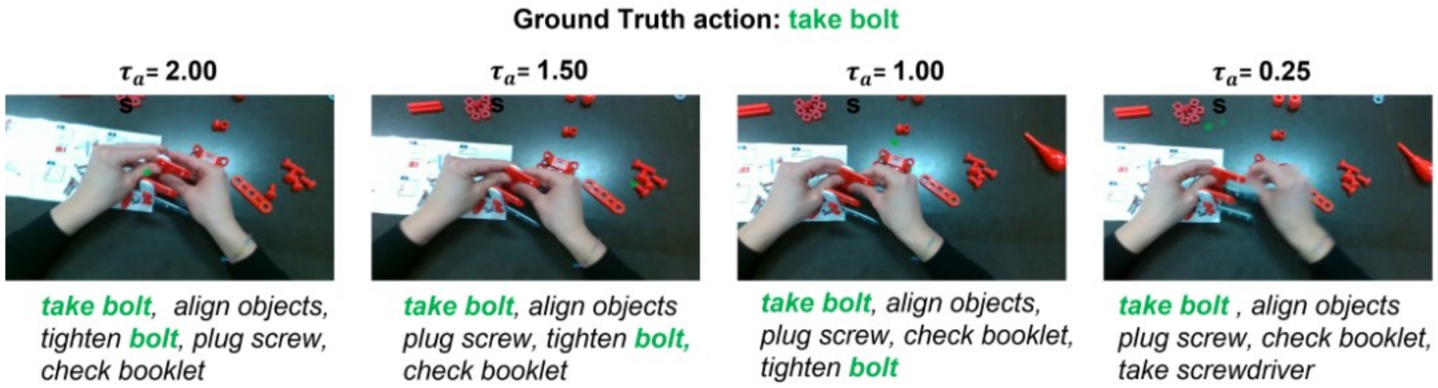


3) EHOI Detection

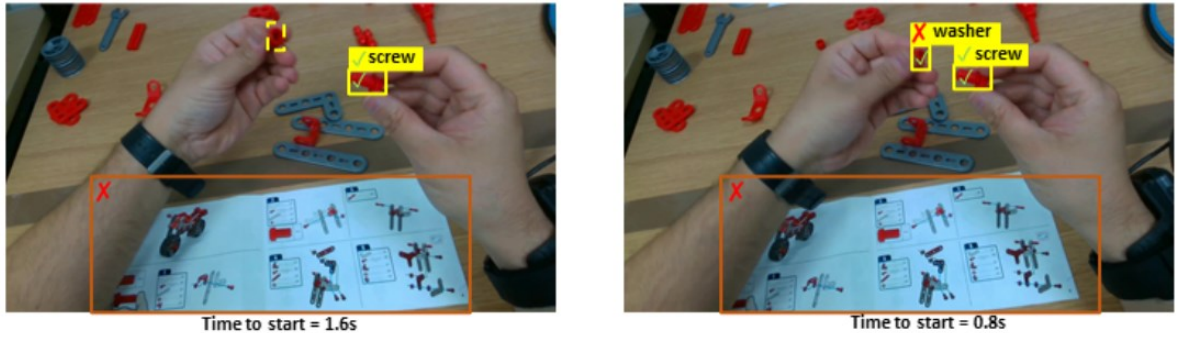

<take>


<gray perforated bar>

4) Action Anticipation

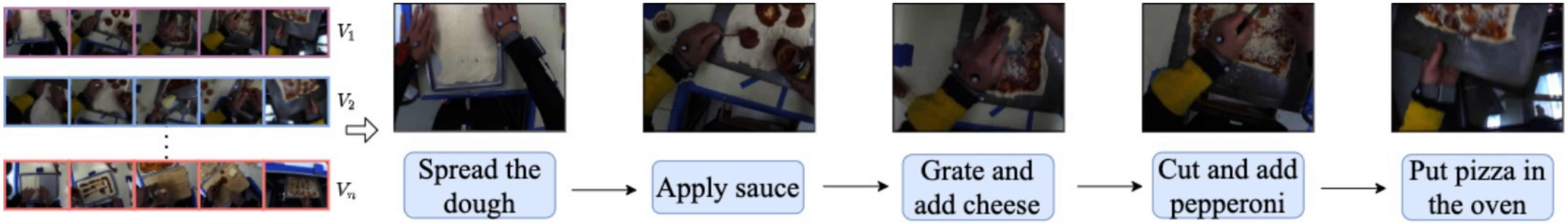


5) Next-Active Object (NAO) Detection



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.

Action Recognition Competition

Università di Catania

ICIAP 2023 UDINE

NEXT VISION
Spin-off of the University of Catania

Leaderboard 2023

Rank	Team	Top-1 Accuracy	Top-5 Accuracy	Technical Report
1	**UCF**	**52.82**	**83.85**	⤓
2	UNIBZ	52.57	81.53	⤓
3	LUBECK	51.82	83.35	⤓
4	MACAU	50.30	78.46	⤓
5	*Baseline (RGB-Depth-Gaze)*	49.66	77.82	
6	TORONTO	49.52	74.21	⤓
7	*Baseline (RGB-Depth)*	49.49	77.61	
8	CUNY	24.69	52.46	⤓

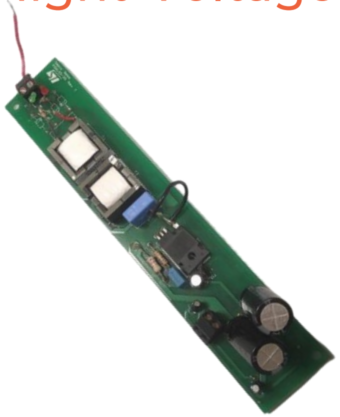
https://iplab.dmi.unict.it/MECCANO/challenge.html

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

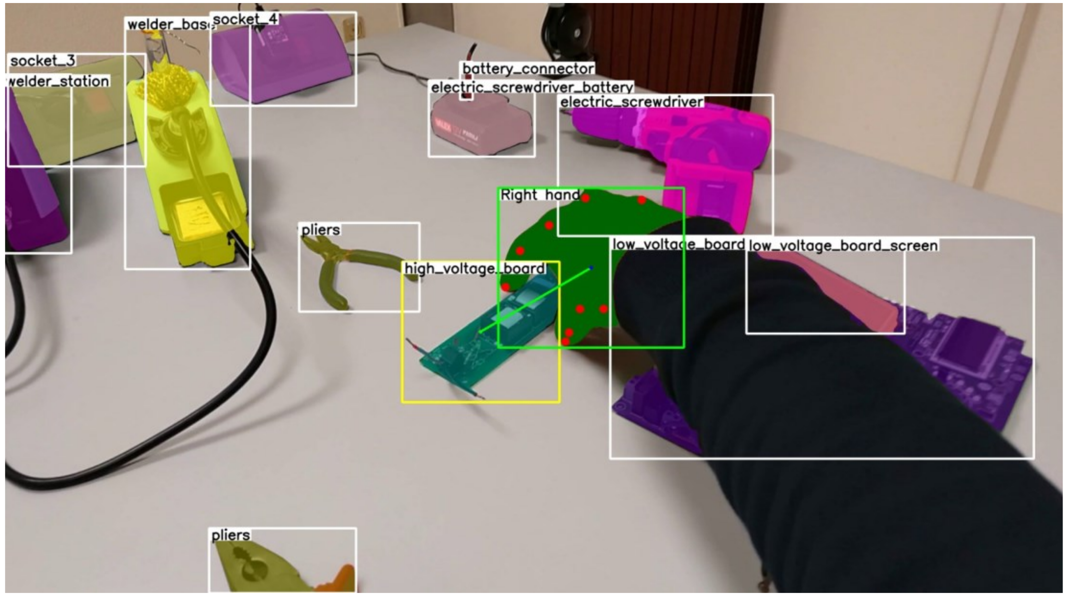
Low-Voltage

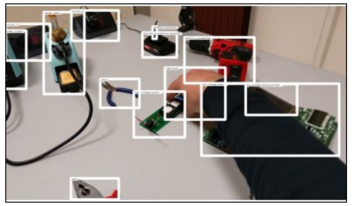
Hight-Voltage



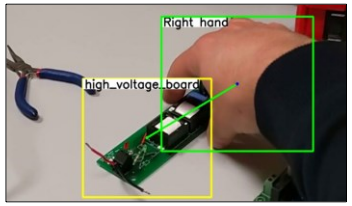
ENIGMA-51: Towards a Fine-Grained Understanding of Human Behavior in Industrial Scenarios. F. Ragusa R. Leonardi, M. Mazzamuto, C. Bonanno, R. Scavo, A. Furnari, G. M. Farinella. WACV (2024).

ENIGMA-51: Annotations

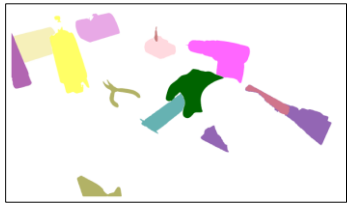




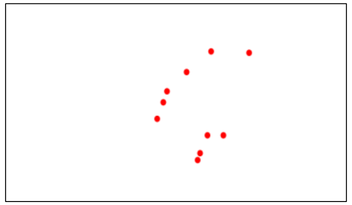
Hand-Object boxes



Human-Object Interactions



Hand-Object Masks



Hand Keypoints

Environment 3D Model

Object 3D Models

TTC: 1.20s Δ = 0.40 TTC: 0.80s TTC: 0.40s first-contact high voltage board

Past Frames

Interaction Frame

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).

Benchmark

Untrimmed temporal detection of human-object interactions

Egocentric human-object interaction detection

Short-term object interaction anticipation

Natural language understanding of intents and entities

Google Glass, 2012

https://www.youtube.com/watch?v=YAXTQL3jPFk

- Google envisioned a future in which smart glasses replace smartphones;
- The goal of Google Glass was to make computation available to the user when they need it and get out of the way when they dont.

The Failure of Google Glass, 2014

https://www.youtube.com/watch?v=ClvI9fZaz6M

Google Glass failed because of the lack of clear use cases + privacy issues.

Consumer Wearable Cameras

Is this it?

SenseCam

2004

Vicon Revue

2010

Autographer

2016

2013

Looxcie

2014

2010

Google Glass

2014

2012

Success Cases

Epson Moverio Smart Glasses for Augmented Reality, since 2012

Moverio BT-40

- USB-C connectivity
- Full HD 1080p
- Second screen privacy

OUR PRICE:

£579.00

incl. VAT (£482.50 ex. VAT)

In Stock

Learn more ▶

Buy Now ▶

FIND A DEALER ▶

REQUEST A CALLBACK ▶

SUPPORT ▶

Moverio BT-40S

- Intelligent Controller
- Full HD 1080p
- Commercial applications

OUR PRICE:

£1,002.00

incl. VAT (£835.00 ex. VAT)

In Stock

Learn more ▶

Buy Now ▶

FIND A DEALER ▶

REQUEST A CALLBACK ▶

SUPPORT ▶

Moverio BT-45CS

- Centred 8MP camera
- Rugged design
- Intelligent Controller

OUR PRICE:

£1,836.00

incl. VAT (£1,530.00 ex. VAT)

In Stock

Learn more ▶

Buy Now ▶

FIND A DEALER ▶

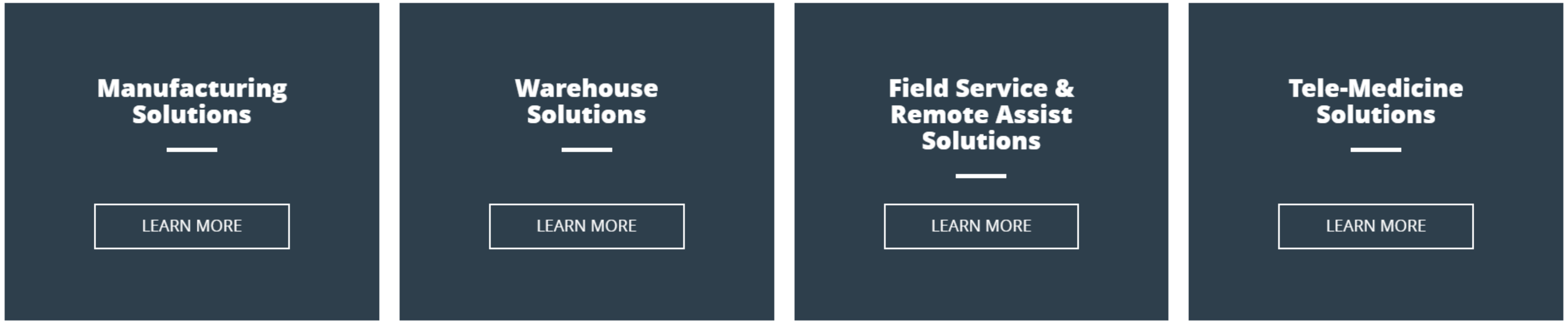
REQUEST A CALLBACK ▶

SUPPORT ▶

focused application scenarios

https://www.epson.co.uk/en_GB/search/allproducts?text=smart+glasses

Vuzix (Since 2012)



https://www.vuzix.com/

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

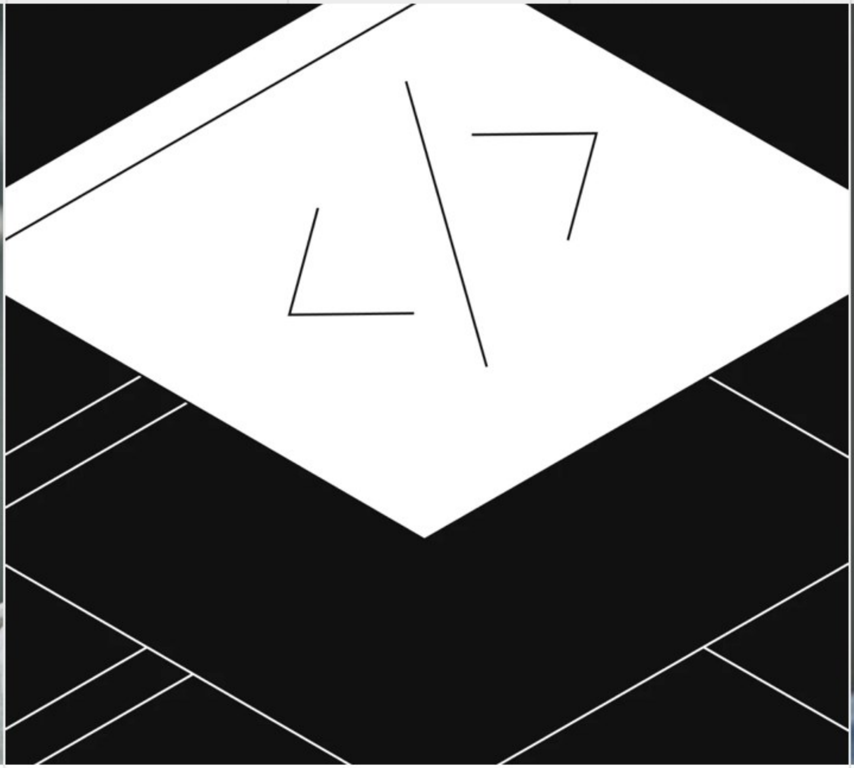
Magic Leap, since 2018 - Magic Leap 2 in 2022

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

Magic Leap 2 – Immersive Enterprise AR Device

Scalable

Magic Leap 2 is built to support scalable augmented reality (AR) solutions necessitating multiple simultaneous users.



Integrative

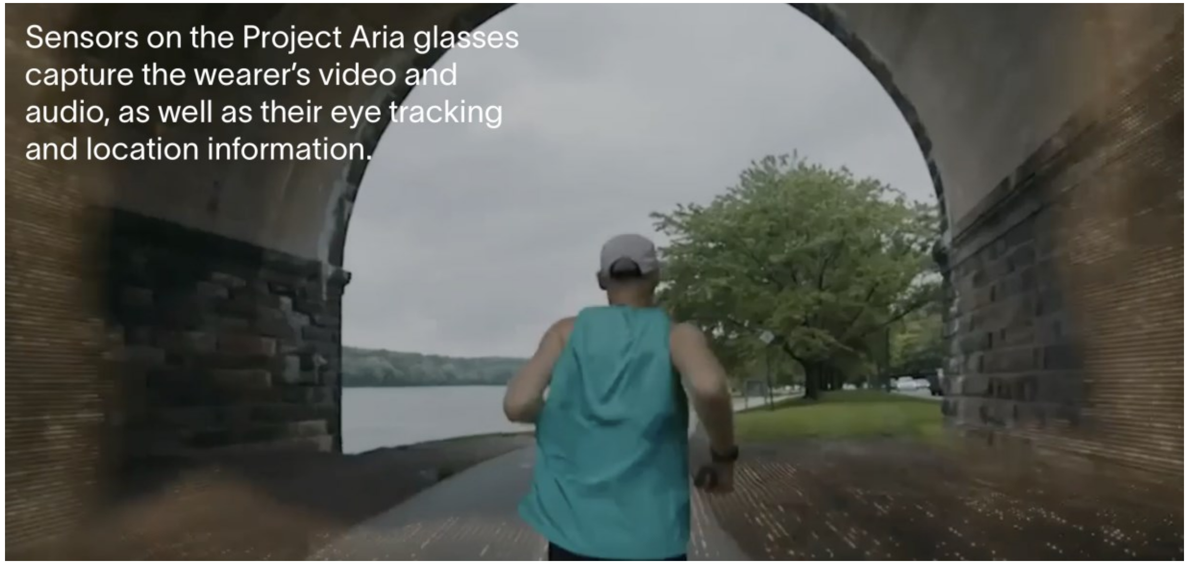
Magic Leap 2 is purpose-built on an open platform to integrate with leading enterprise multi-device management (MDM) systems.

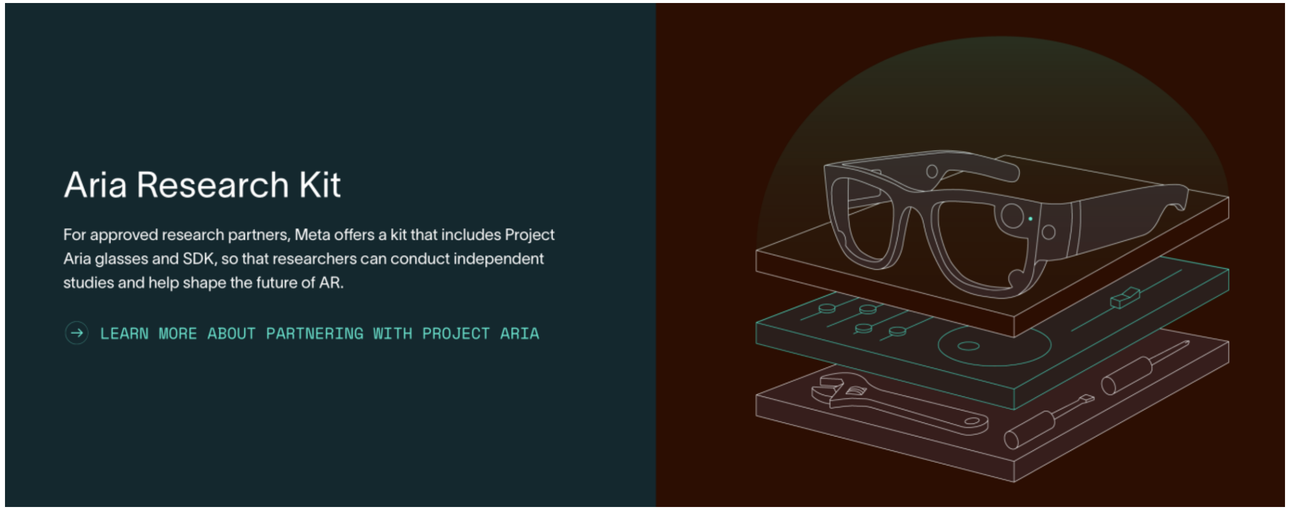
Secure

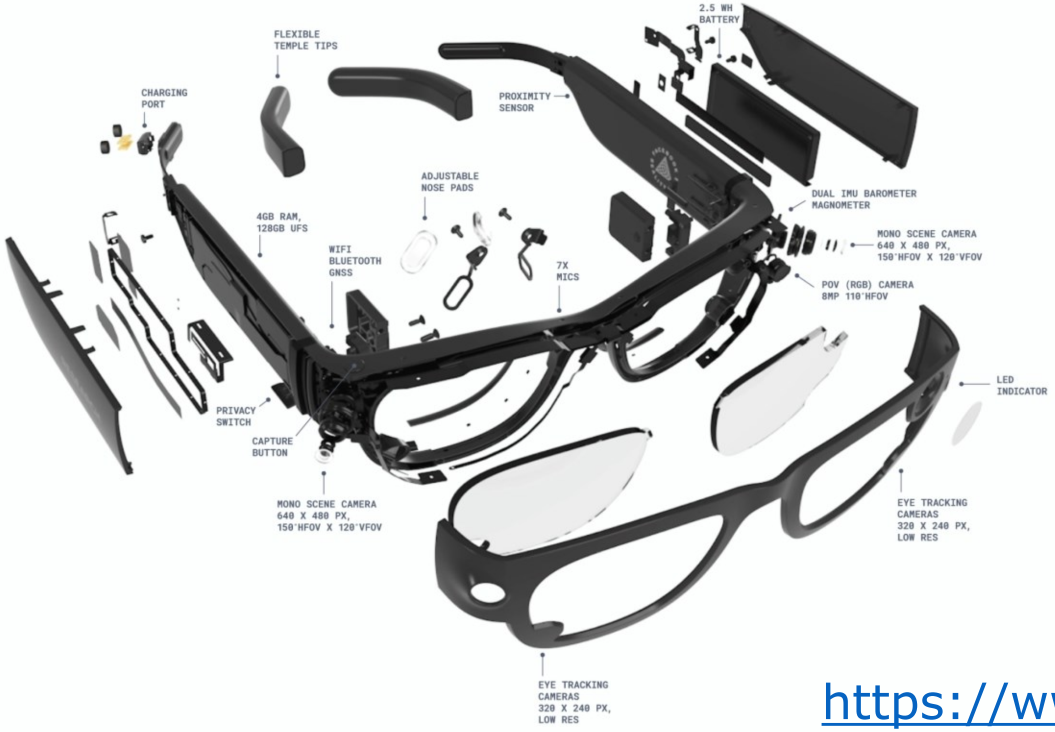
Store your data anywhere and use any preferred cloud setup. Magic Leap 2 lets users retain control of their data and is compatible with leading enterprise security protocols.

https://www.magicleap.com/en-us/

Meta’s Project Aria

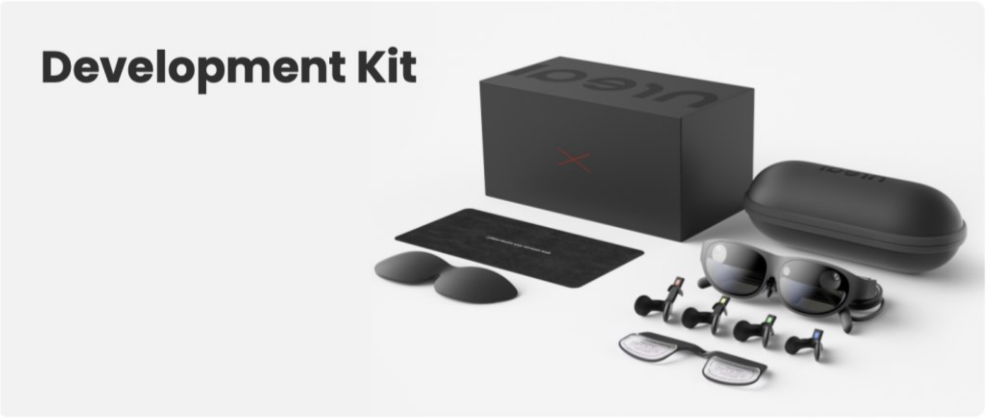


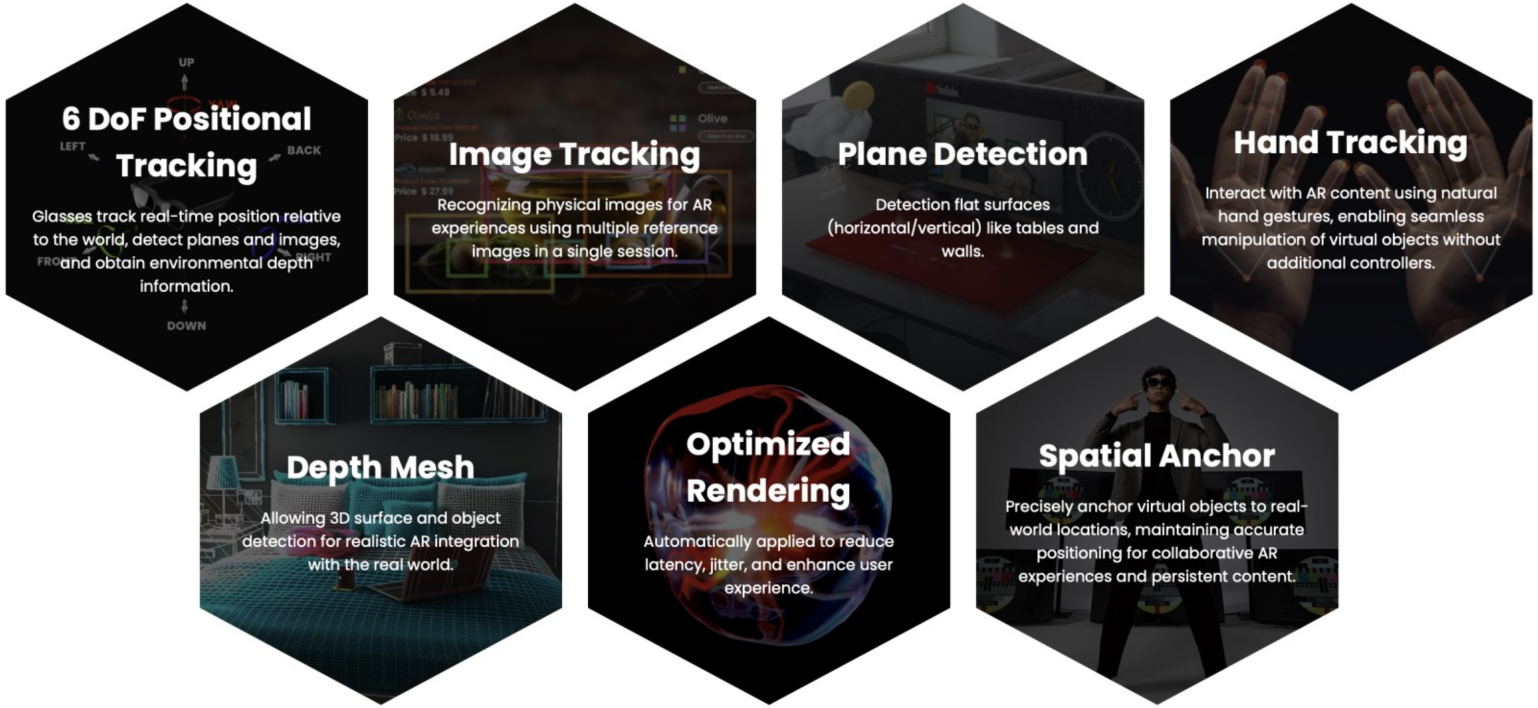




https://www.projectaria.com

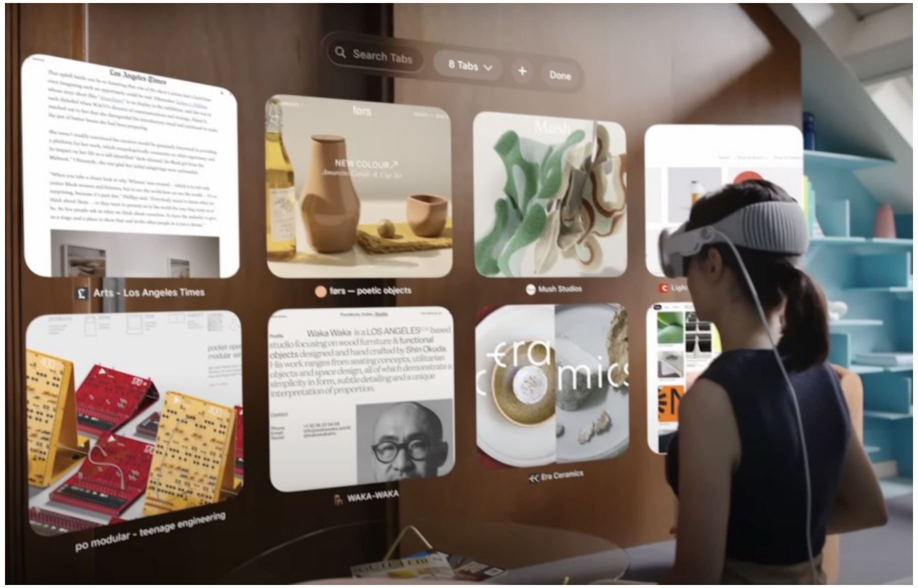
XREAL



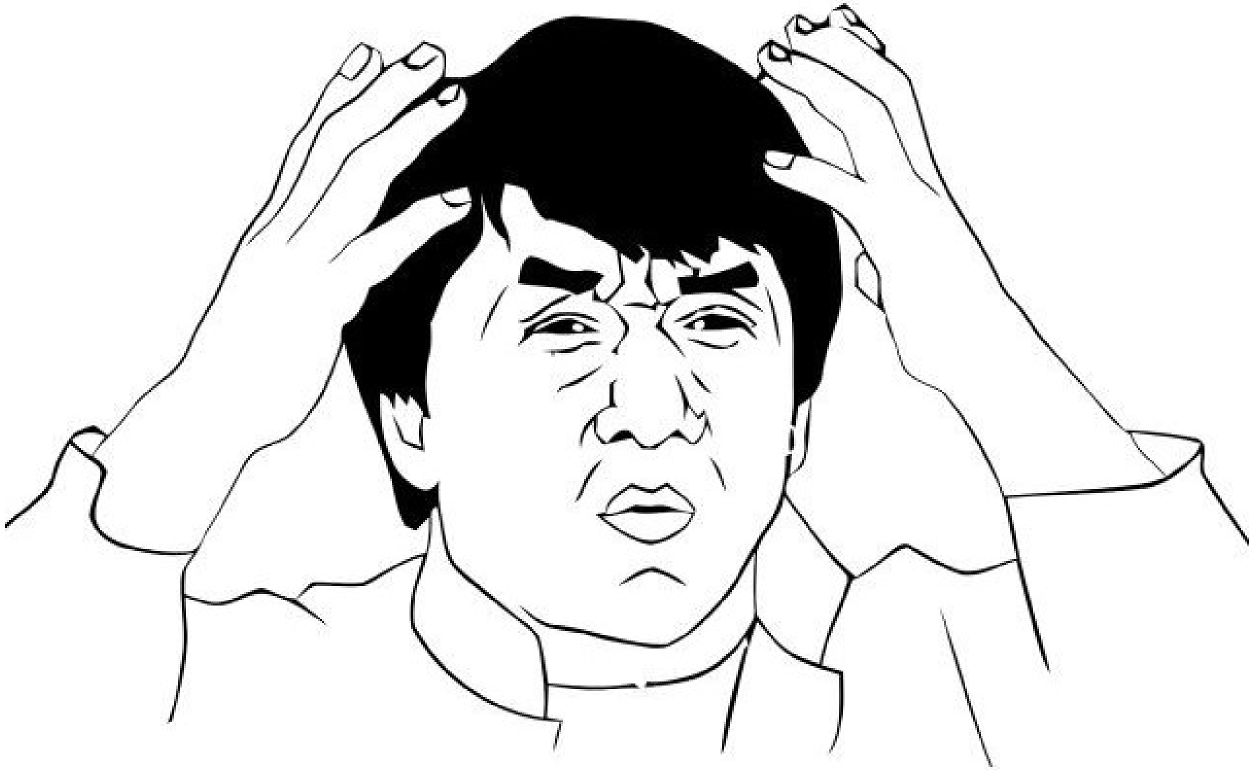


https://www.xreal.com/

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/

Snapdragon Spaces

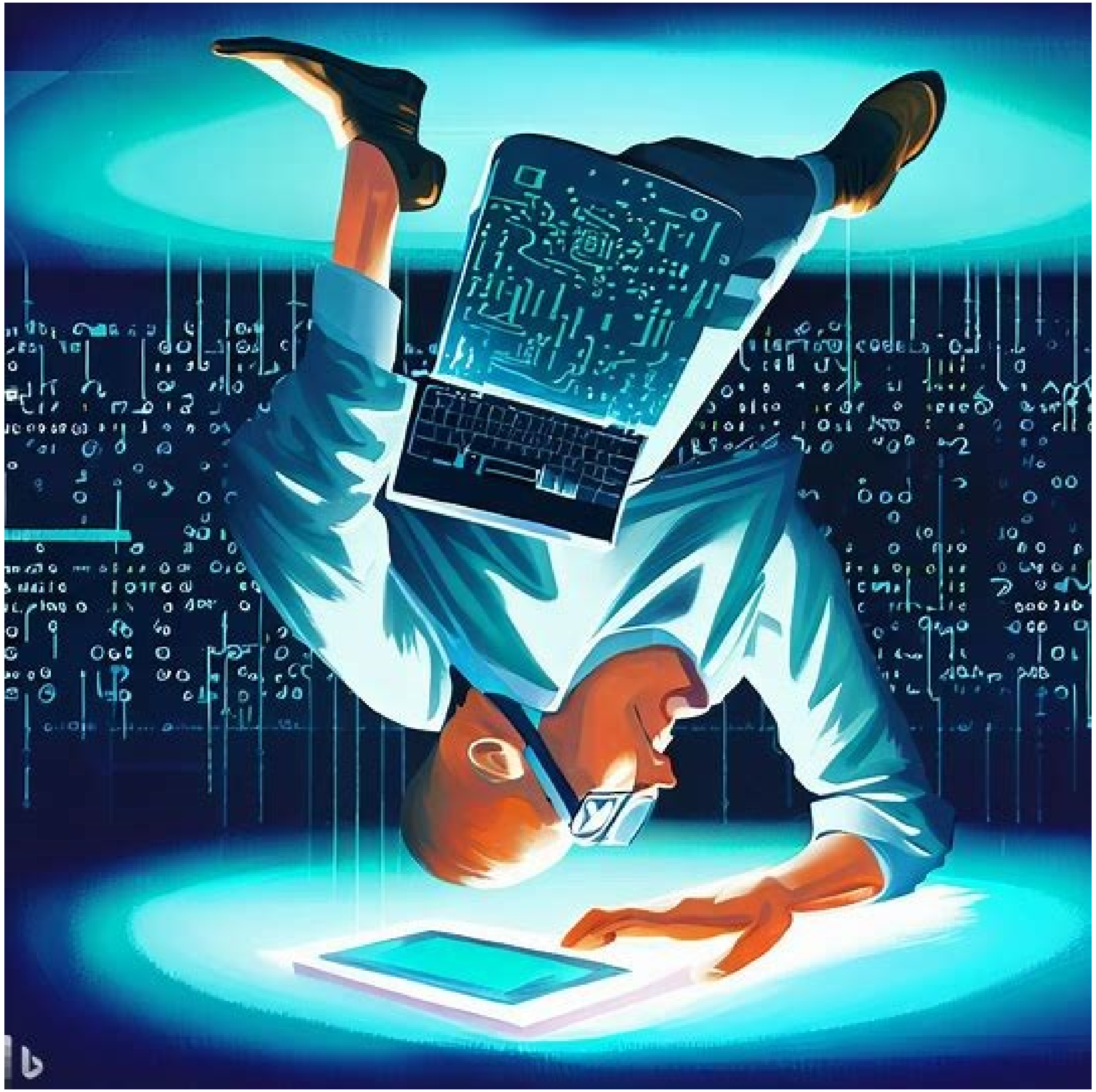
“The Snapdragon Spaces XR Developer Platform reduces developer friction by providing a uniform set of augmented reality features independent of device manufacturers. This allows developers to seamlessly blend the lines between our physical and digital realities and transform the world around us in ways limited only by our imaginations.”

https://www.qualcomm.com/products/features/snapdragon-spaces-xr-platform

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

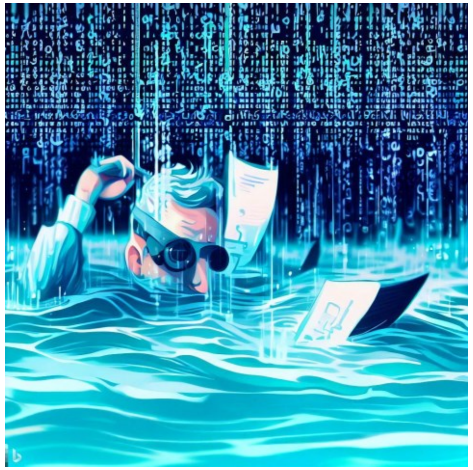
↓

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

*: 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

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

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

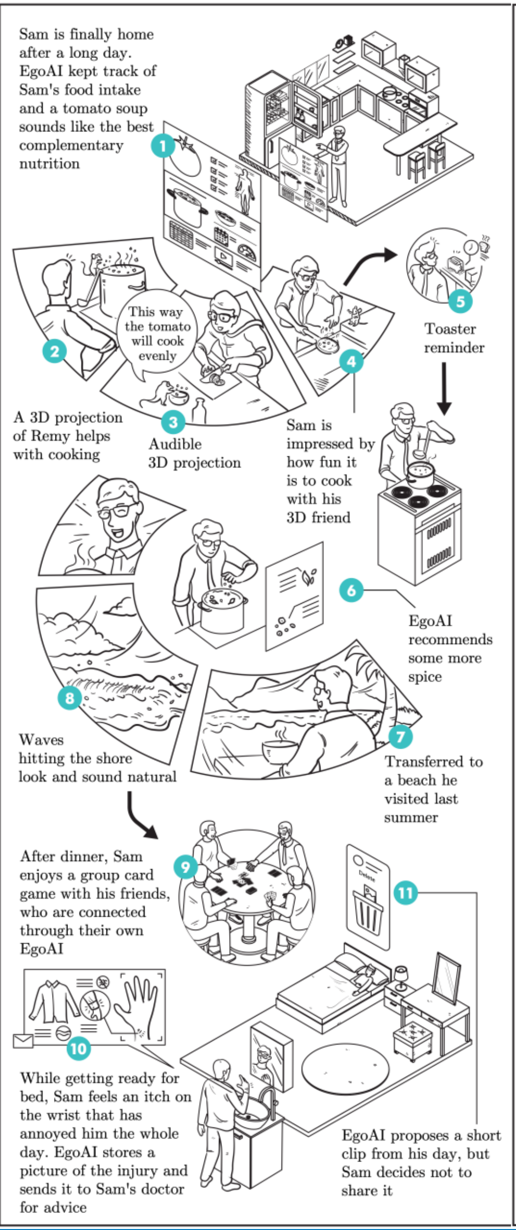
Comment:
Thanks for the nice paper, that's awesome!

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.

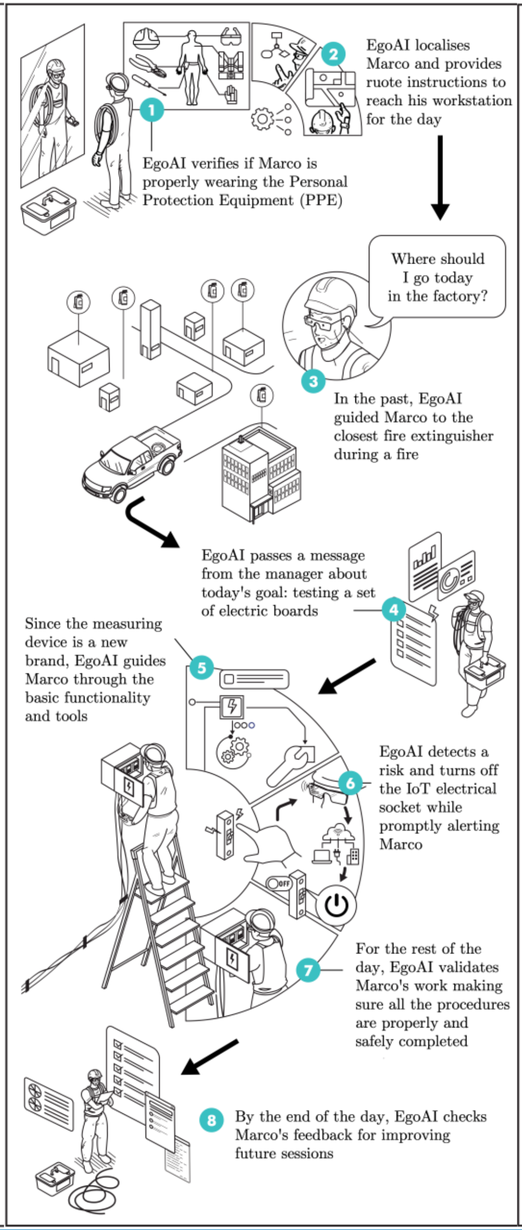
https://openreview.net/forum?id=V3974SUk1w

Università di Catania An Outlook into the Future – Futuristic Stories

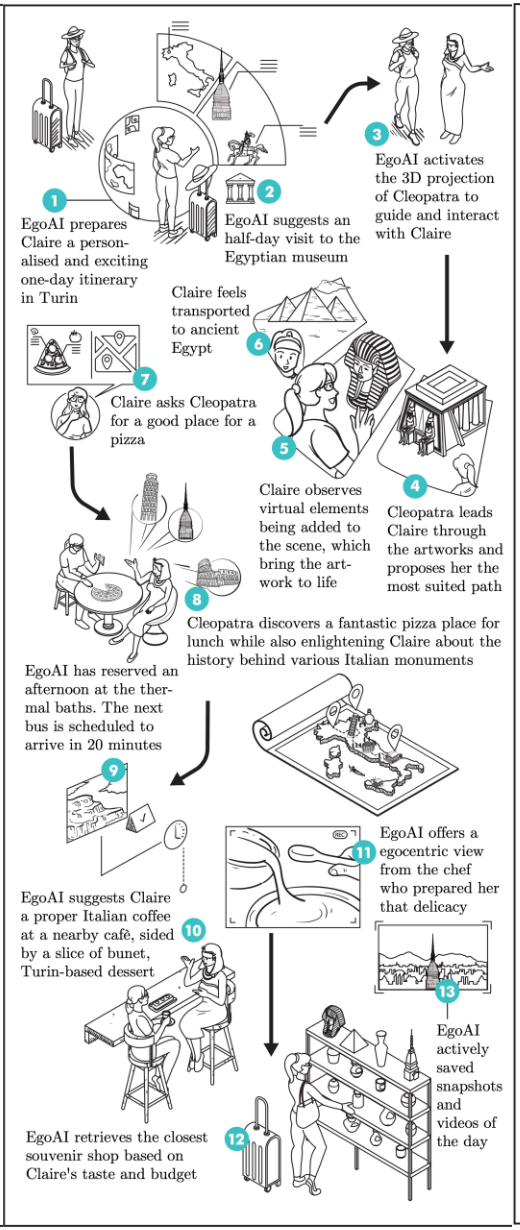
EGO-HOME



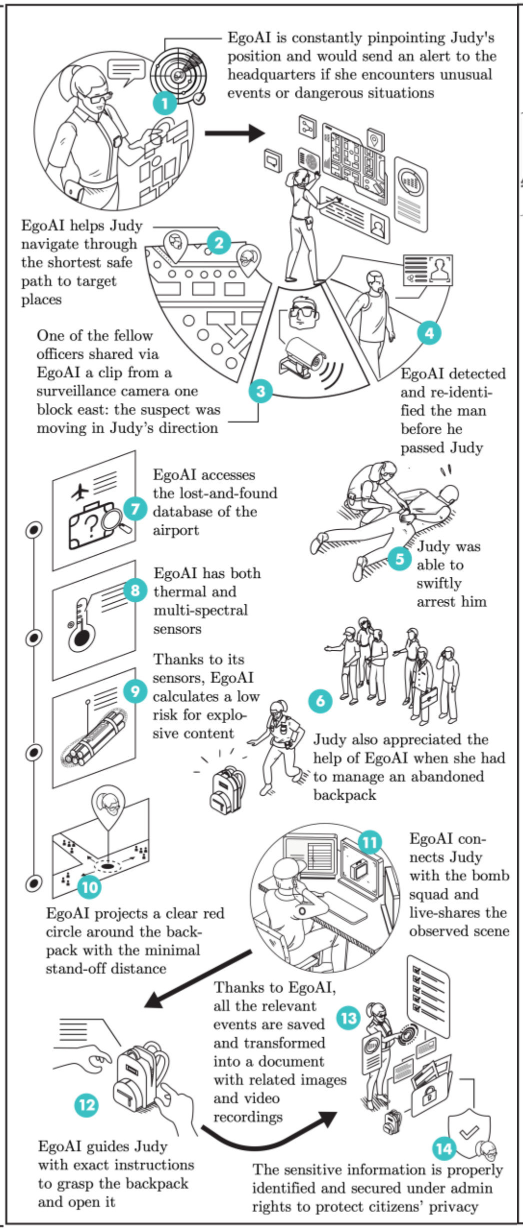
EGO-WORKER



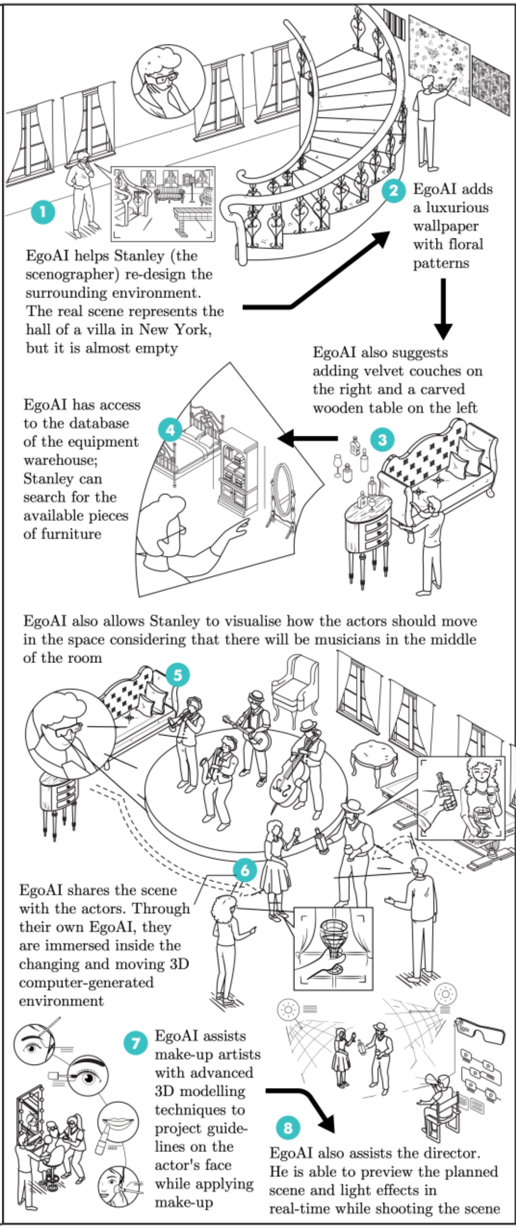
EGO-TOURIST



EGO-POLICE



EGO-DESIGNER



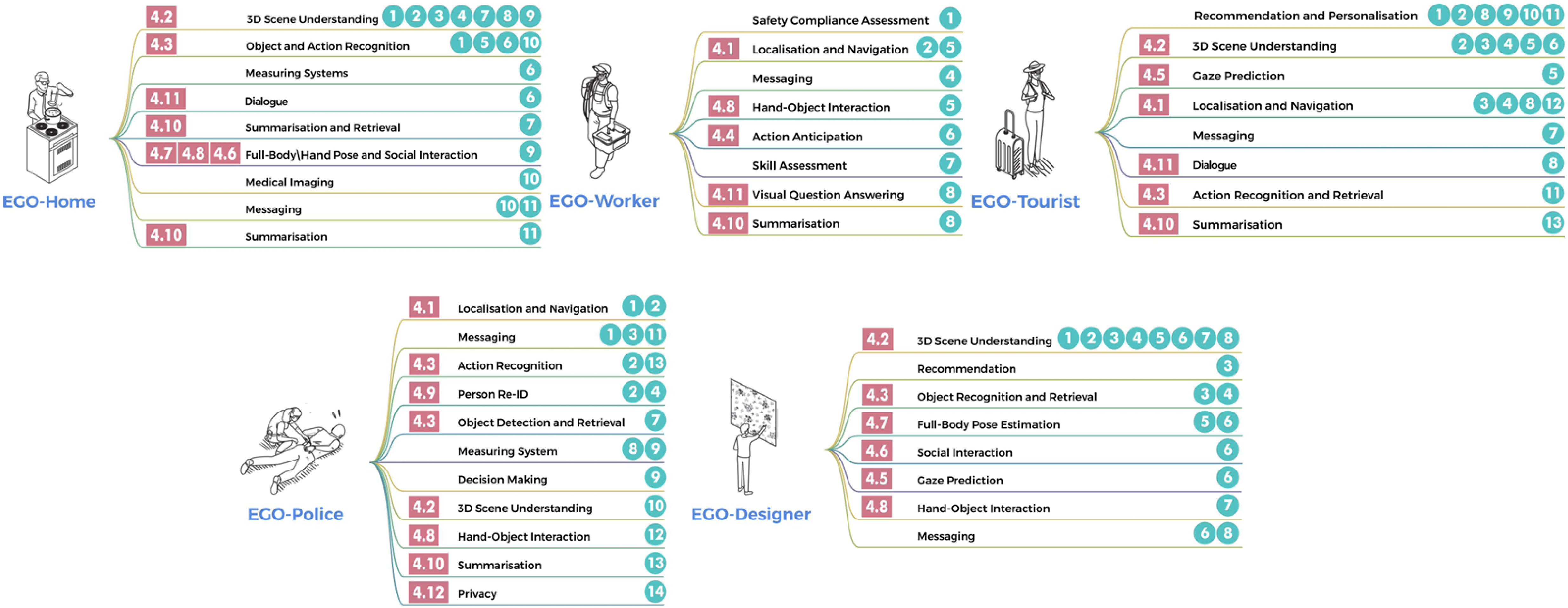
From Narratives to Research Tasks

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. *arXiv preprint arXiv:2308.07123*.

Links between Stories and Tasks



perspective and provides ego-based assistance. We associate story p arts 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. *arXiv preprint arXiv:2308.07123*.

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. *arXiv preprint arXiv:2308.07123*.

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. *arXiv preprint arXiv:2308.07123*.

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. *arXiv preprint arXiv:2308.07123*.

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


TIME FOR...
QUESTIONS

Università di Catania

NEXT VISION

Spin-off of the University of Catania

IMAGE PROCESSING LABORATORY

VISIGRAPP 2024
19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications
Rome, Italy 27 - 29 February, 2024
GRAPP HUCAPP IVAPP VISAPP

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/

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.