VISMAC 23

International Summer School on Machine Vision

Università di Catania

NEXT VISION

IMAGE PROCESSING LABORATORY

Tutorial on Egocentric Vision

Antonino Furnari

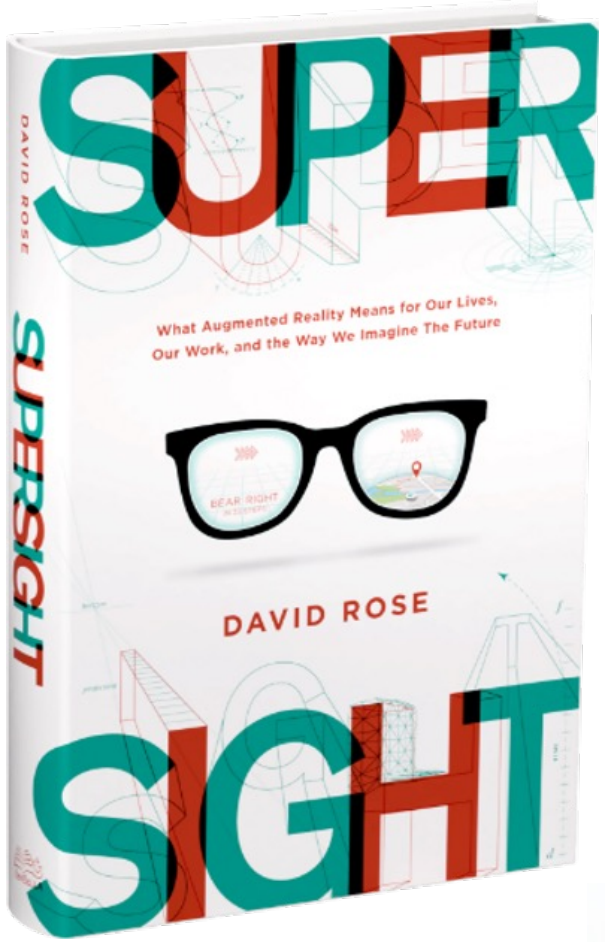
First Person Vision@Image Processing Laboratory - http://iplab.dmi.unict.it/fpv

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

Department of Mathematics and Computer Science - University of Catania

antonino.furnari@unict.it - http://www.antoninofurnari.it/

The evolution of men



«The human eye has not significantly evolved in millennia»

«Although we've invented glasses to correct our vision, and microscopes and telescopes for specialized tasks, our ancestors perceived the world much as we do. But thanks to a set of exponentially advancing technologies over the next decade, that's about to change radically.»

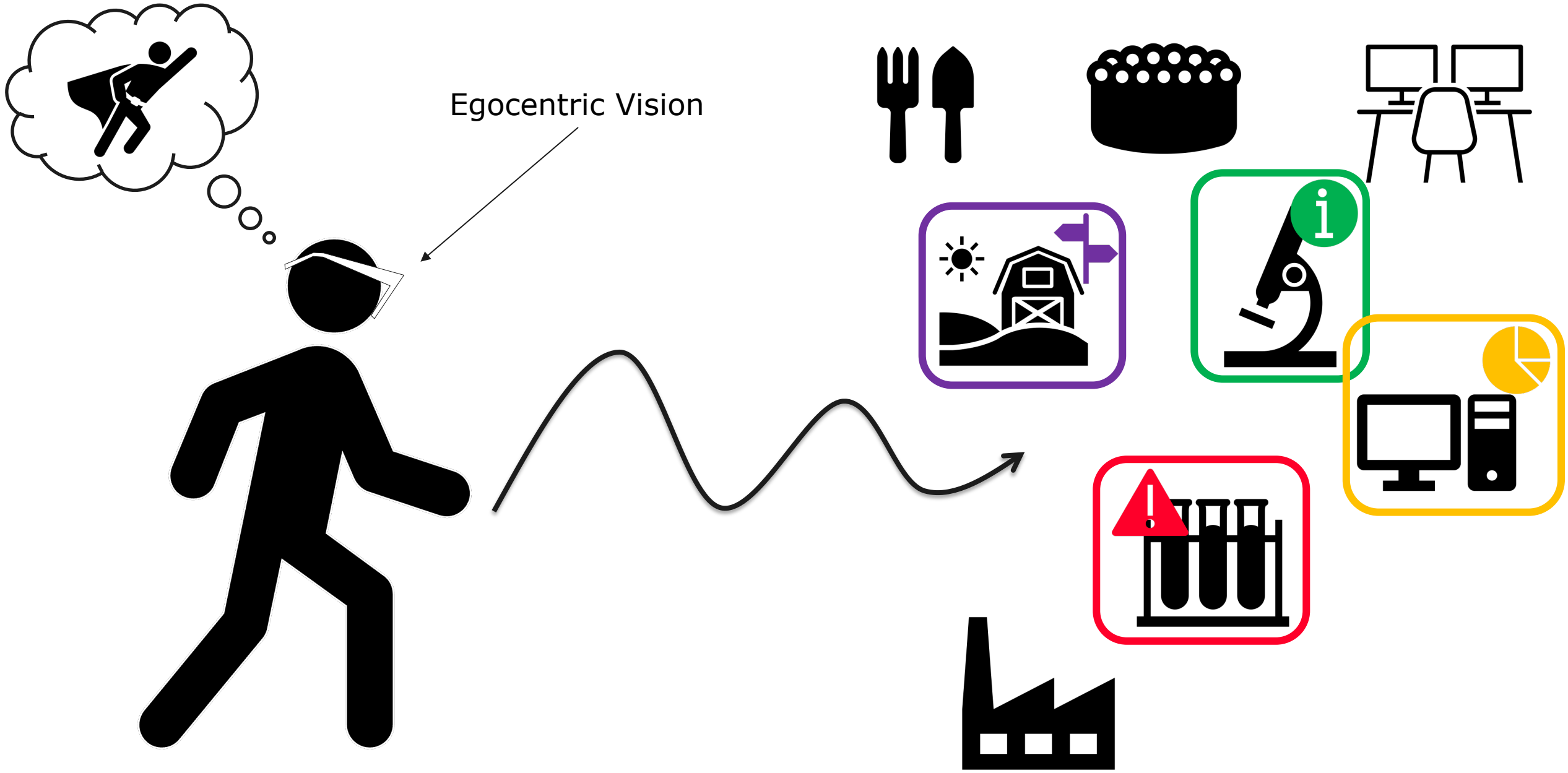
A better eye?

Clip from the Terminator 2-Judgment day movie: https://youtu.be/9MeaaCwBW28

Ref: https://www.redsharknews.com/vr_and_ar/item/3539-terminator-2-vision-the-augmented-reality-standard-for-25-years

A better human!

A Virtual Personal ~~Super Hero~~ 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 contraints

Fixed Camera

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

Roadmap

1

Egocentric Vision: A Retrospective

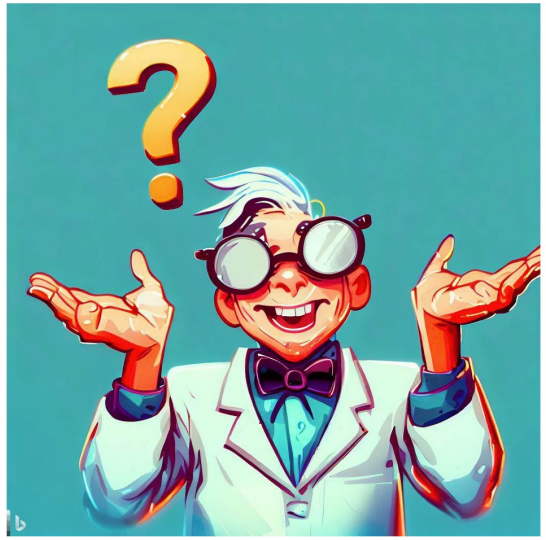
2

The Cyborg Dream

3

An Outlook into the Future

4

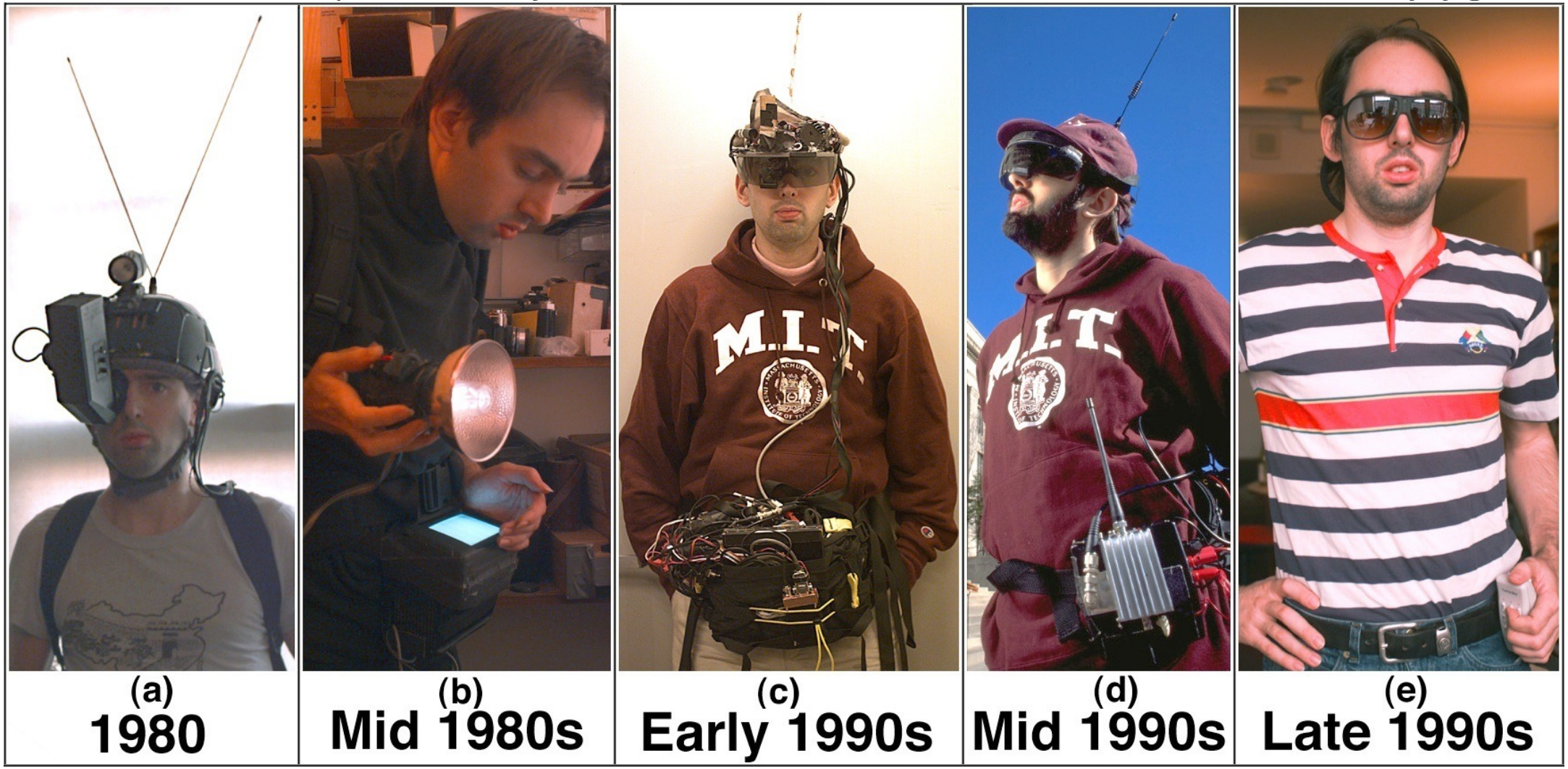


Doing Research in Egocentric Vision: Where to start?

Egocentric Vision: A Retrospective

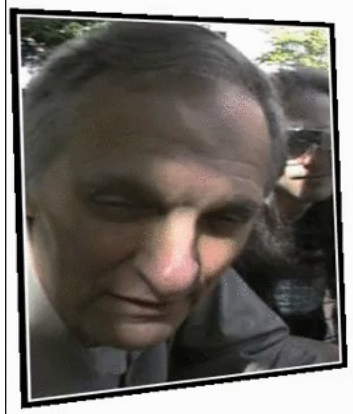
The Birth of Wearable Computing

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

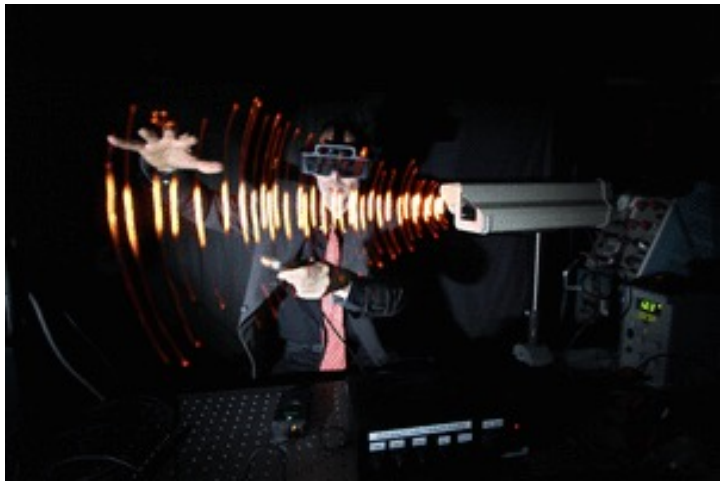


In the 80s and 90s Steve Mann (PhD in Media Arts and Sciences at MIT, 1997) invented a number of wearable computers featuring video capabilities, computing capabilities, and a werable screen for feedback. **Steve Mann is often referred to as «the father of wearable computing»**

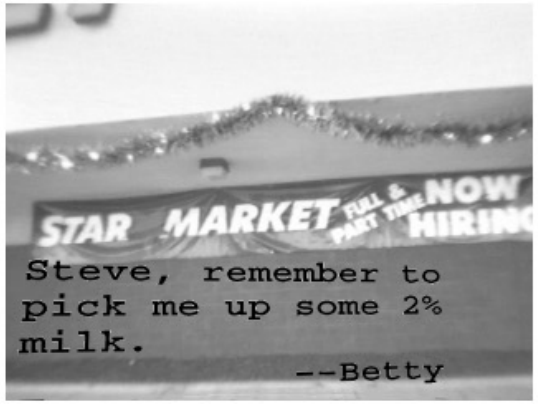
First Wearable Computing Applications



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

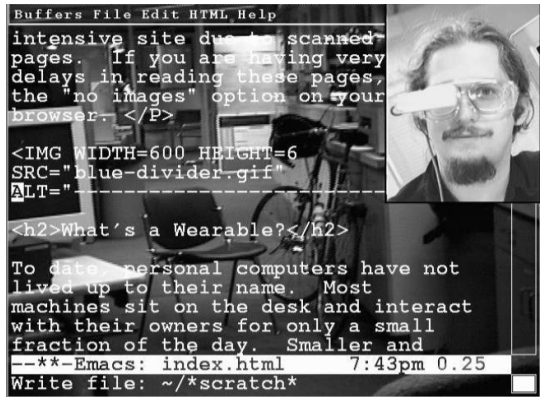
MIT Media Lab Seminal Works, late 1990s

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

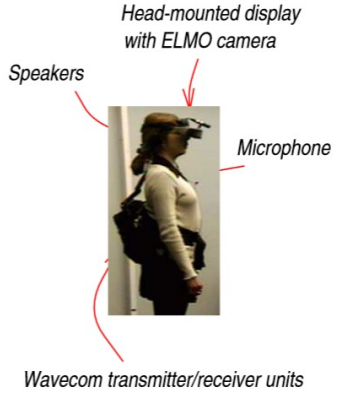
Thad Starner Bernt Schiele Alex Pentland
Media Laboratory, Massachusetts Institute of Technology

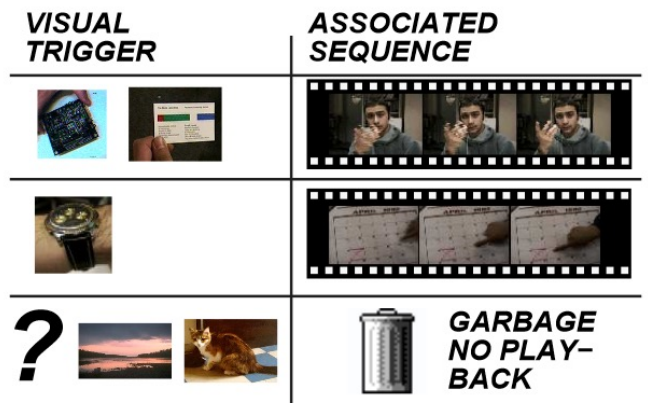
(location and task recognition)

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)





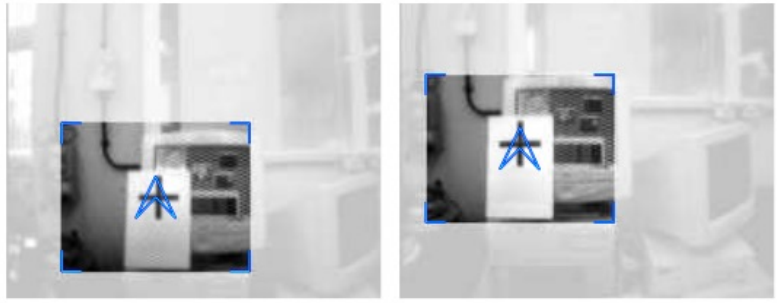
1999

Early 2000s

Wearable Visual Robots

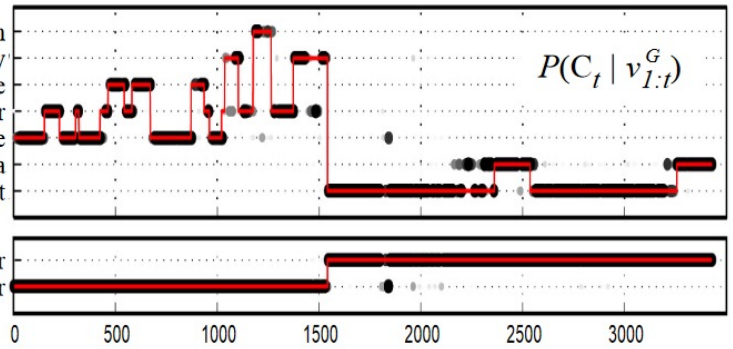
W.W. Mayol, B. Tordoff and D.W. Murray
University of Oxford, Parks Road, Oxford OX1 3PJ, UK

(active vision)



2000

2003



Context-based vision system for place and object recognition

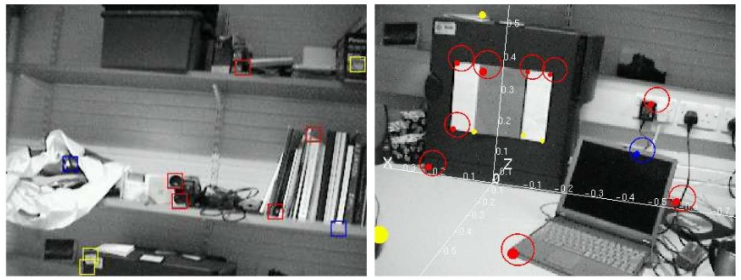
Antonio Torralba, MIT AI lab, Cambridge, MA 02139
Kevin P. Murphy, MIT AI lab, Cambridge, MA 02139
William T. Freeman, MIT AI lab, Cambridge, MA 02139
Mark A. Rubin, Lincoln Labs, Lexington, MA 02420

(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

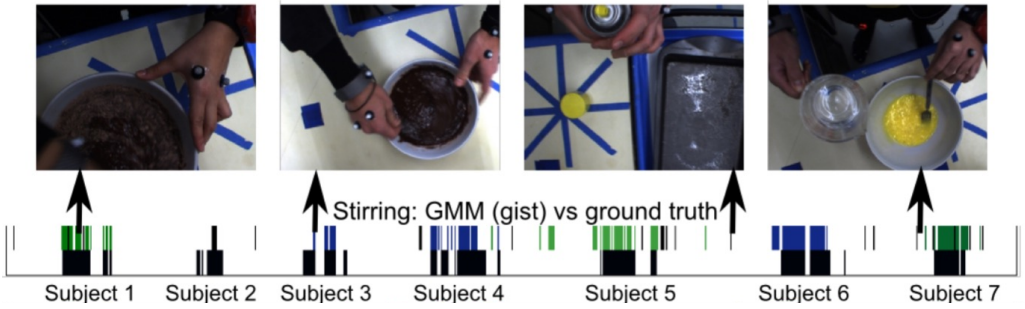
W.W. Mayol
Department of Computer Science
University of Bristol

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.

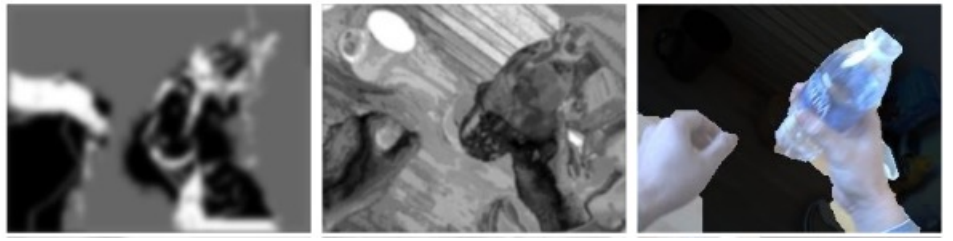
(activity classification)

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)

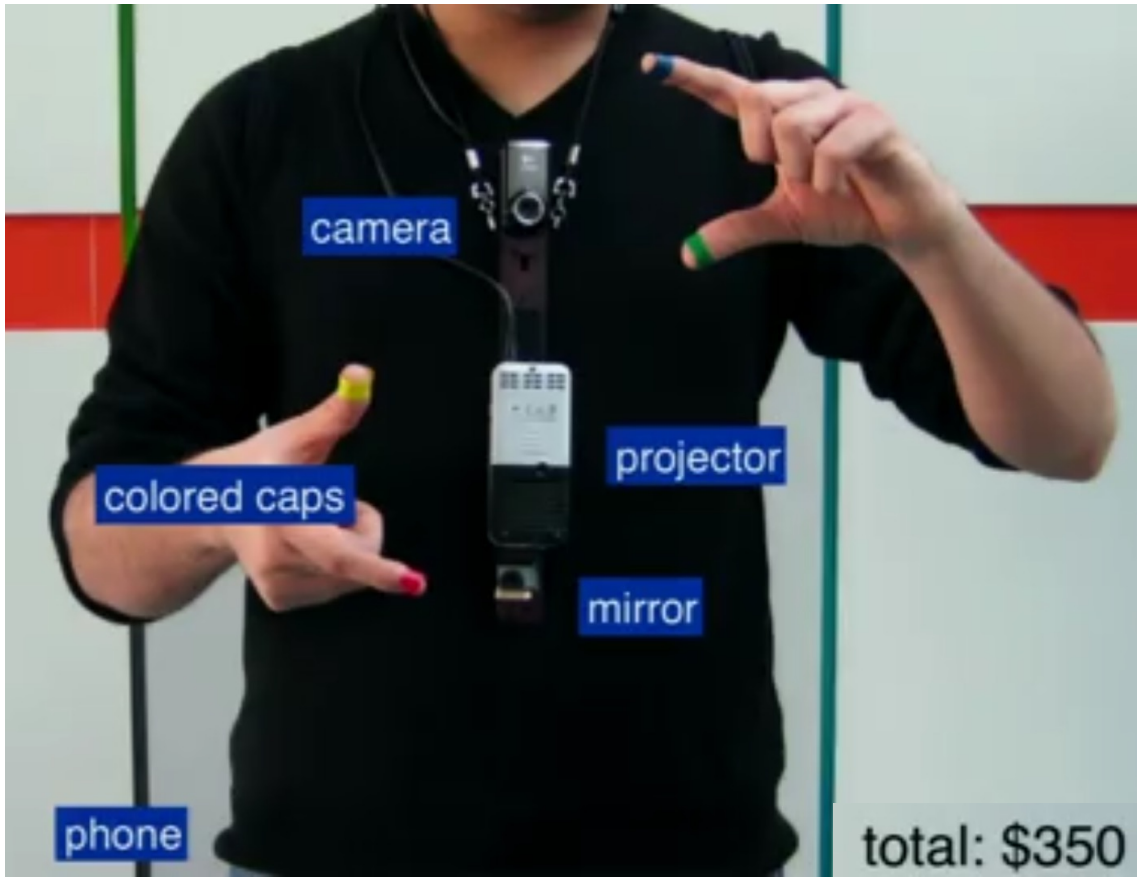


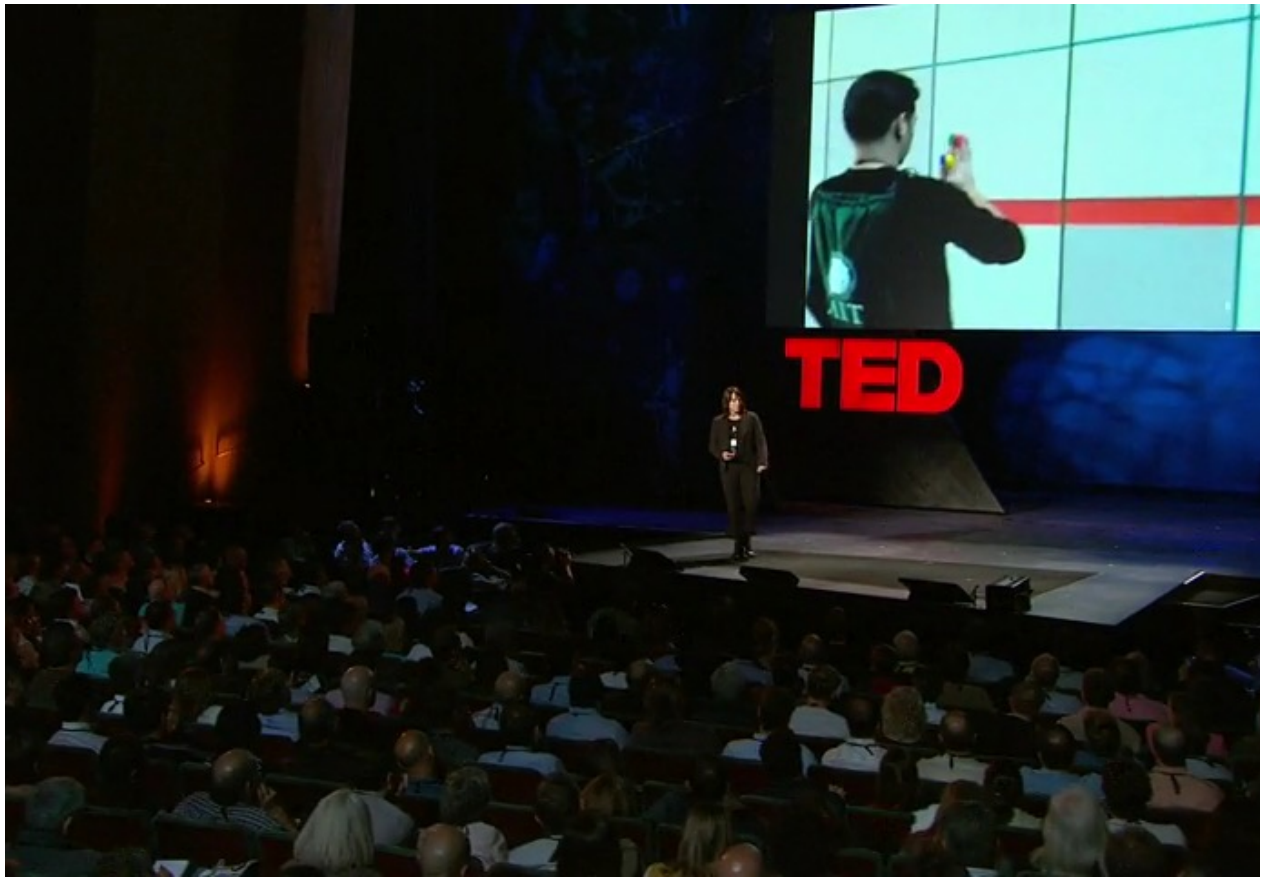
2010

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»



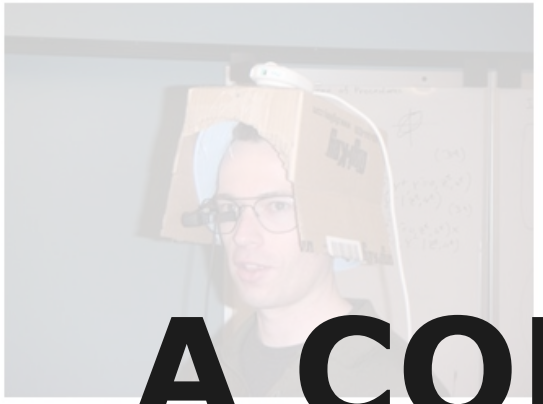


Pattie Maes & Pranav Mistry (MIT) @ TED

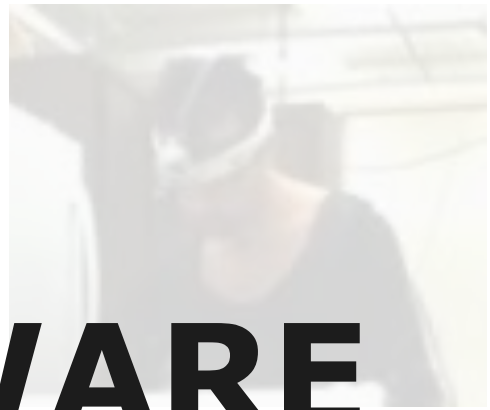
https://www.ted.com/talks/pattie_maes_demos_the_sixth_sense

RADIO SILENCE

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/

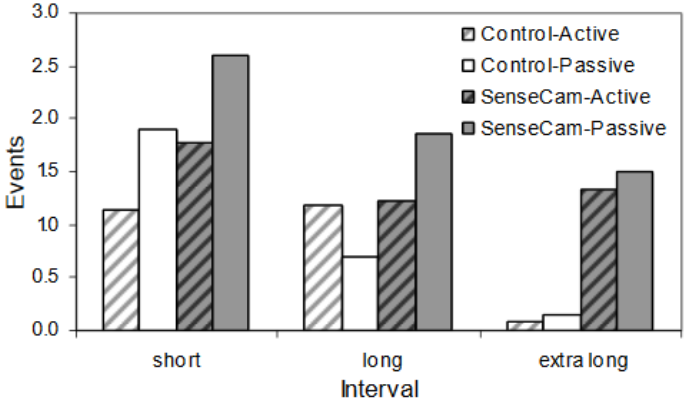
Bell, Gordon, and Jim Gemmell. *Your life, uploaded: The digital way to better memory, health, and productivity*. Penguin, 2010.

Research using Microsoft SenseCam

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

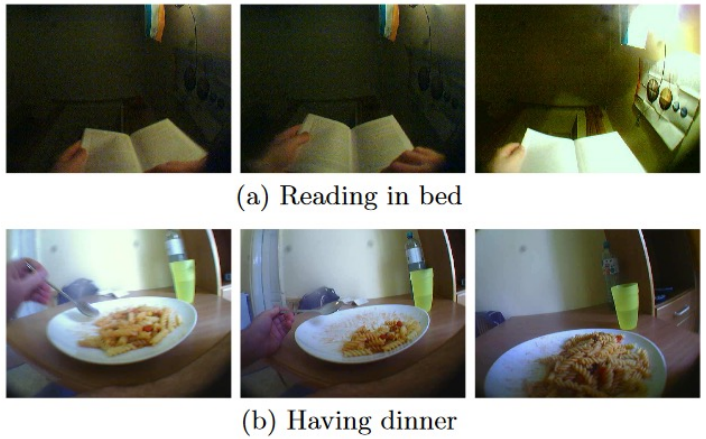
Abigail Sellen, Andrew Fogg, Mike Aitken*, Steve Hodges, Carsten Rother and Ken Wood
Microsoft Research Cambridge
7 JJ Thomson Ave, Cambridge, UK, CB3 0FB
*Behavioural & Clinical Neuroscience Institute
Dept. of Psychology, University of Cambridge

(health, memory augmentation)



2007

2008



(a) Reading in bed

(b) Having dinner

MyPlaces: Detecting Important Settings in a Visual Diary

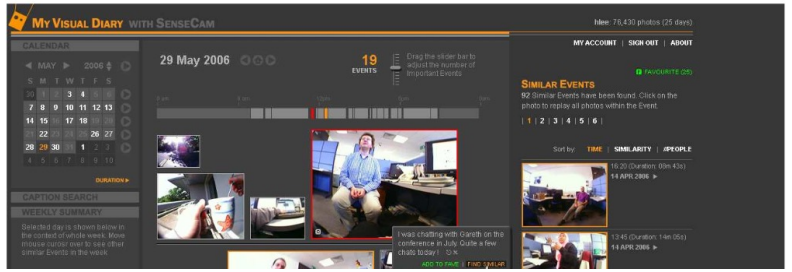
Michael Blighe and Noel E. O'Connor
Centre for Digital Video Processing, Adaptive Information Cluster
Dublin City University, Ireland
{blighem, oconnorn}@eeng.dcu.ie

(lifelogging, place recognition)

Constructing a SenseCam Visual Diary as a Media Process

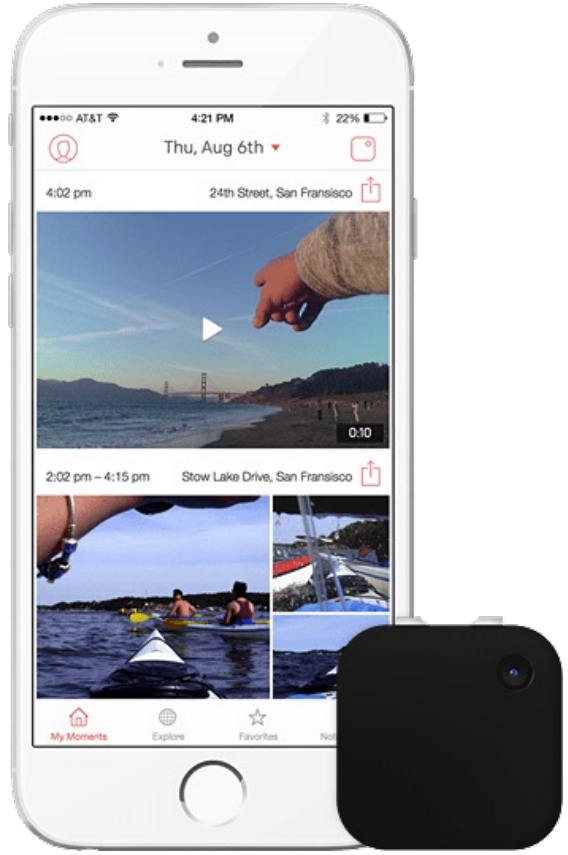
Hyowon Lee, Alan F. Smeaton, Noel O'Connor, Gareth Jones, Michael Blighe, Daragh Byrne, Aiden Doherty, and Cathal Gurrin
Centre for Digital Video Processing & Adaptive Information Cluster,
Dublin City University

(lifelogging, multimedia retrieval)



2008

Narrative Clip, 2012



http://getnarrative.com/

Research Using Narrative Clip

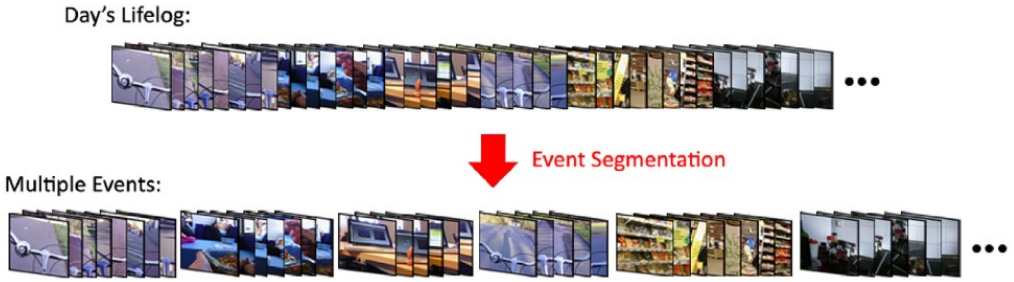
Multi-face tracking by extended bag-of-tracklets in egocentric photo-streams

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

(lifelogging, face tracking)

2016

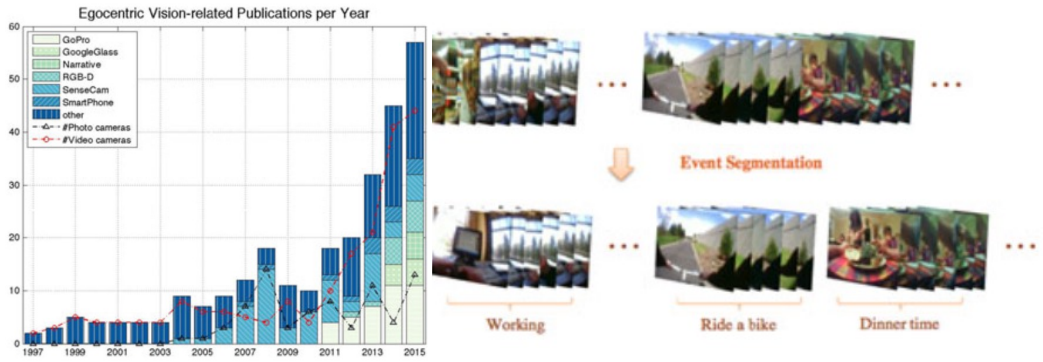
2017



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)



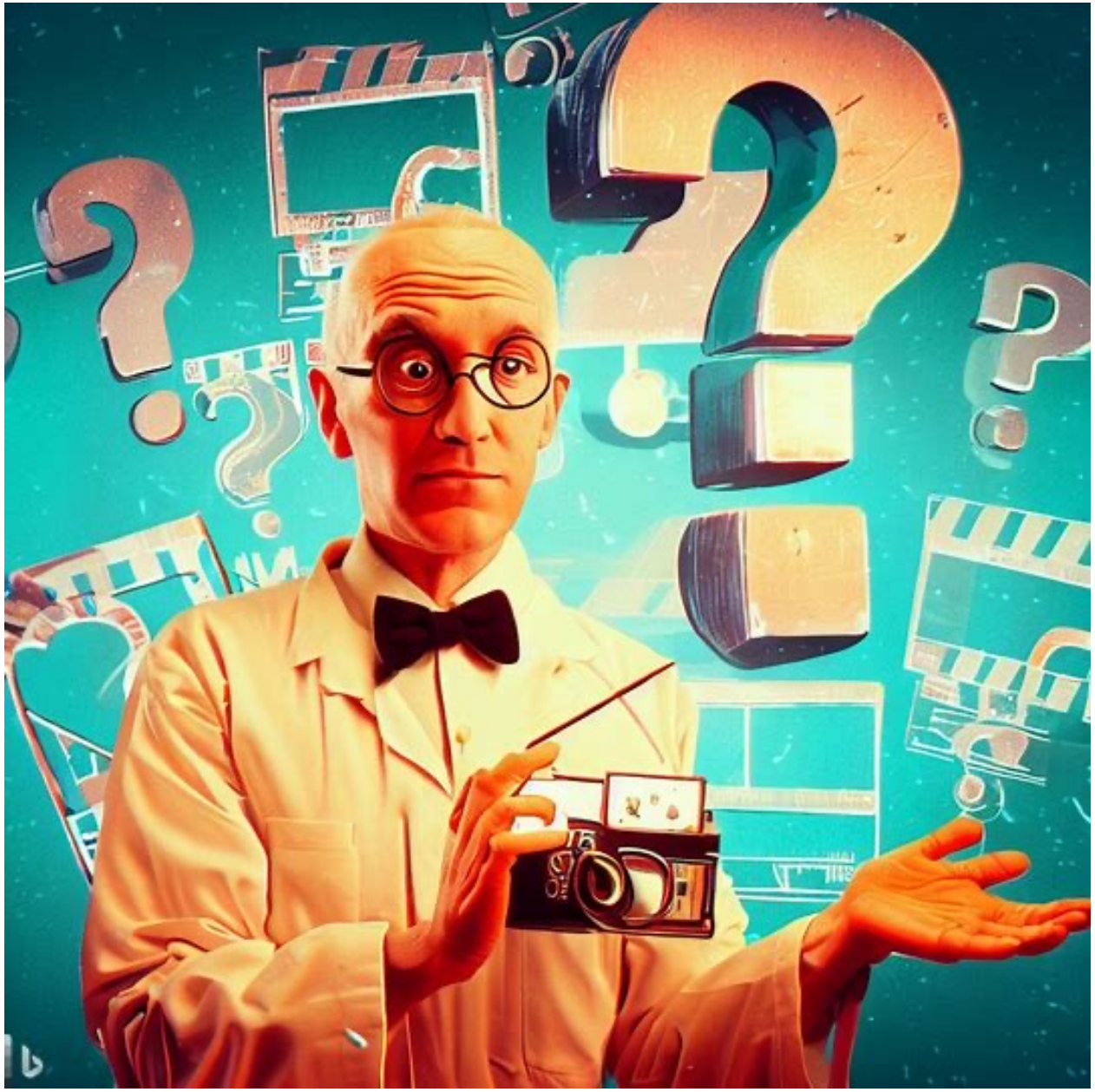
Toward Storytelling From Visual Lifelogging: An Overview

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

Early Research On Egocentric Video

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

Kris M. Kitani
UEC Tokyo
Tokyo, Japan

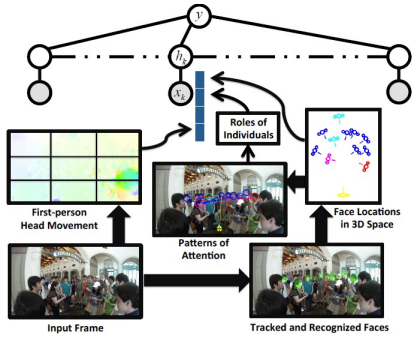
Takahiro Okabe, Yoichi Sato
University of Tokyo
Tokyo, Japan

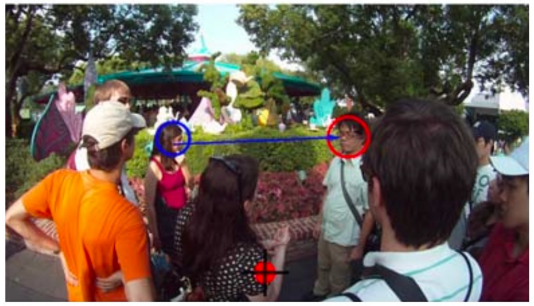
Akihiro Sugimoto
National Institute of Informatics
Tokyo, Japan

(unsupervised action recognition, video indexing)

2011

2012





Social Interactions: A First-Person Perspective

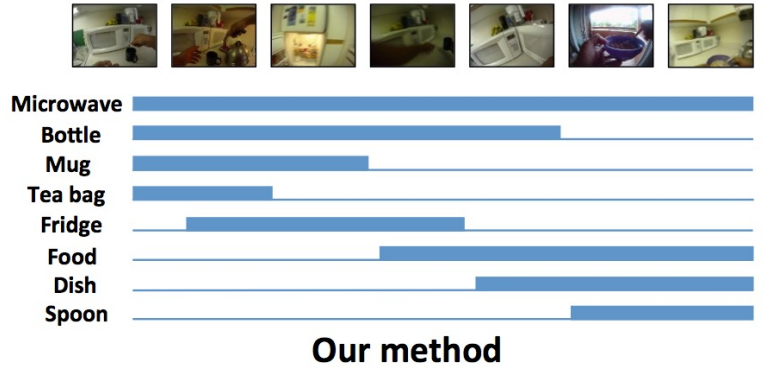
Alireza Fathi[1], Jessica K. Hodgins[2,3], James M. Rehg[1]

(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)



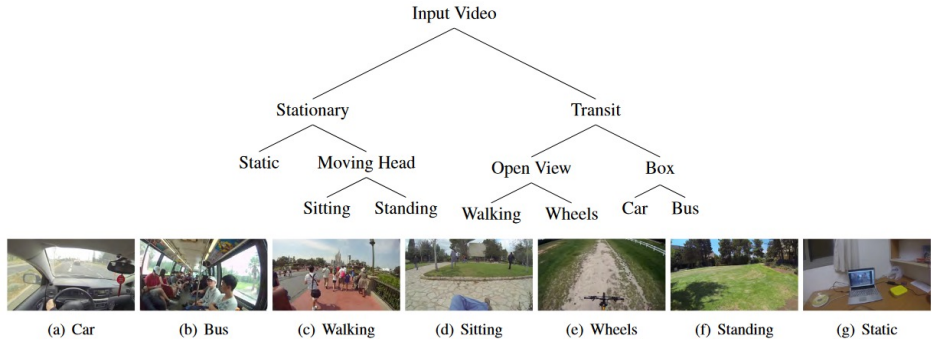
2013

Later Research On Egocentric Video

Temporal Segmentation of Egocentric Videos

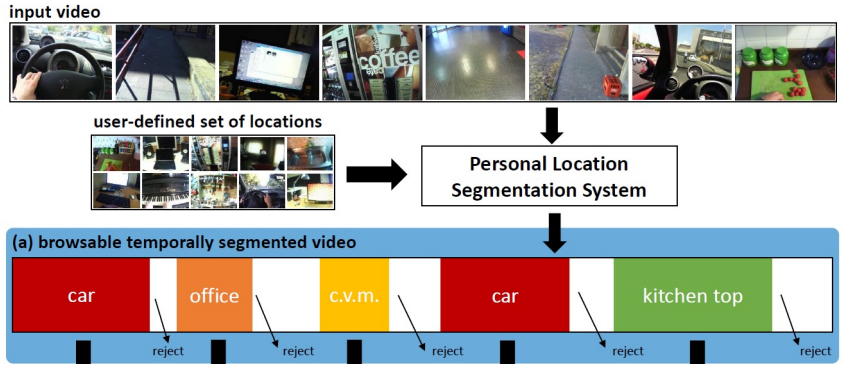
Yair Poleg Chetan Arora* Shmuel Peleg

(egocentric video indexing)



2014

2016



Recognizing Personal Locations from Egocentric Videos

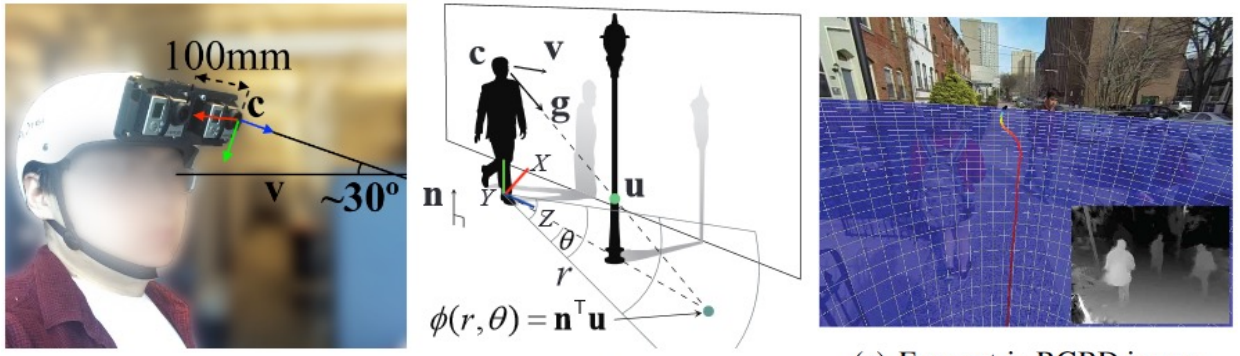
Antonino Furnari, Giovanni Maria Farinella, *Senior Member, IEEE,* and Sebastiano Battiato, *Senior Member, IEEE*

(localization, indexing, context-aware computing)

Egocentric Future Localization

Hyun Soo Park Jyh-Jing Hwang Yedong Niu Jianbo Shi

(future localization, navigation)



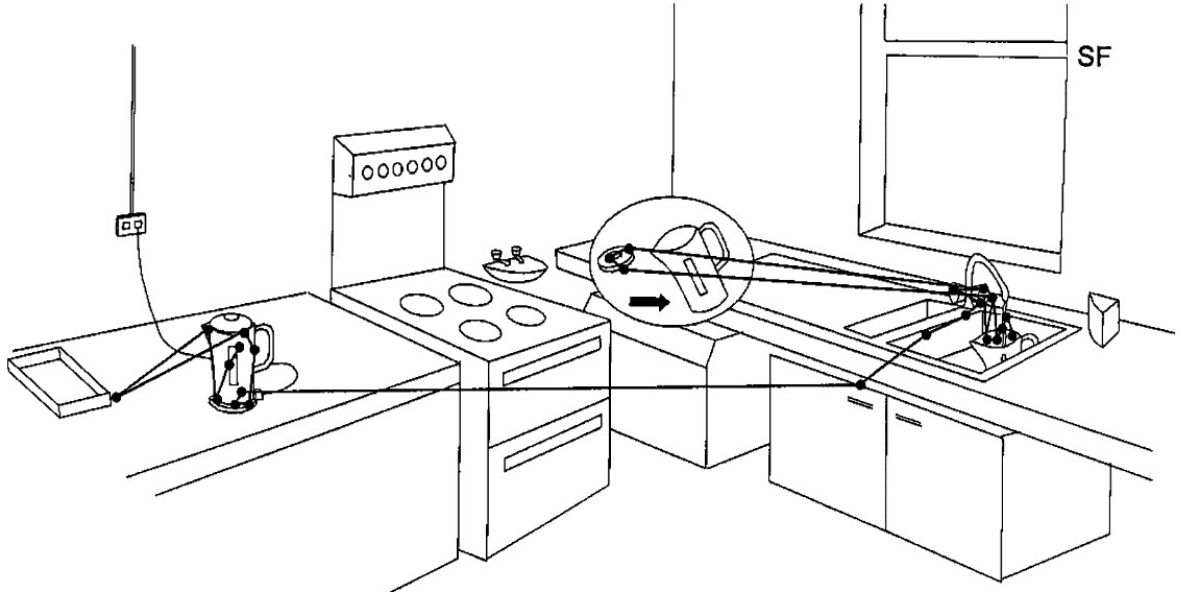
(a) Ego-stereo cameras (b) Geometry (c) Egocentric RGBD image

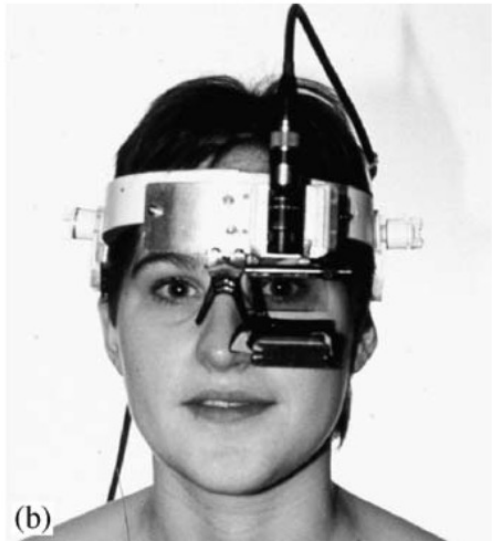
2016

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 -)

Inward- and outward-looking cameras

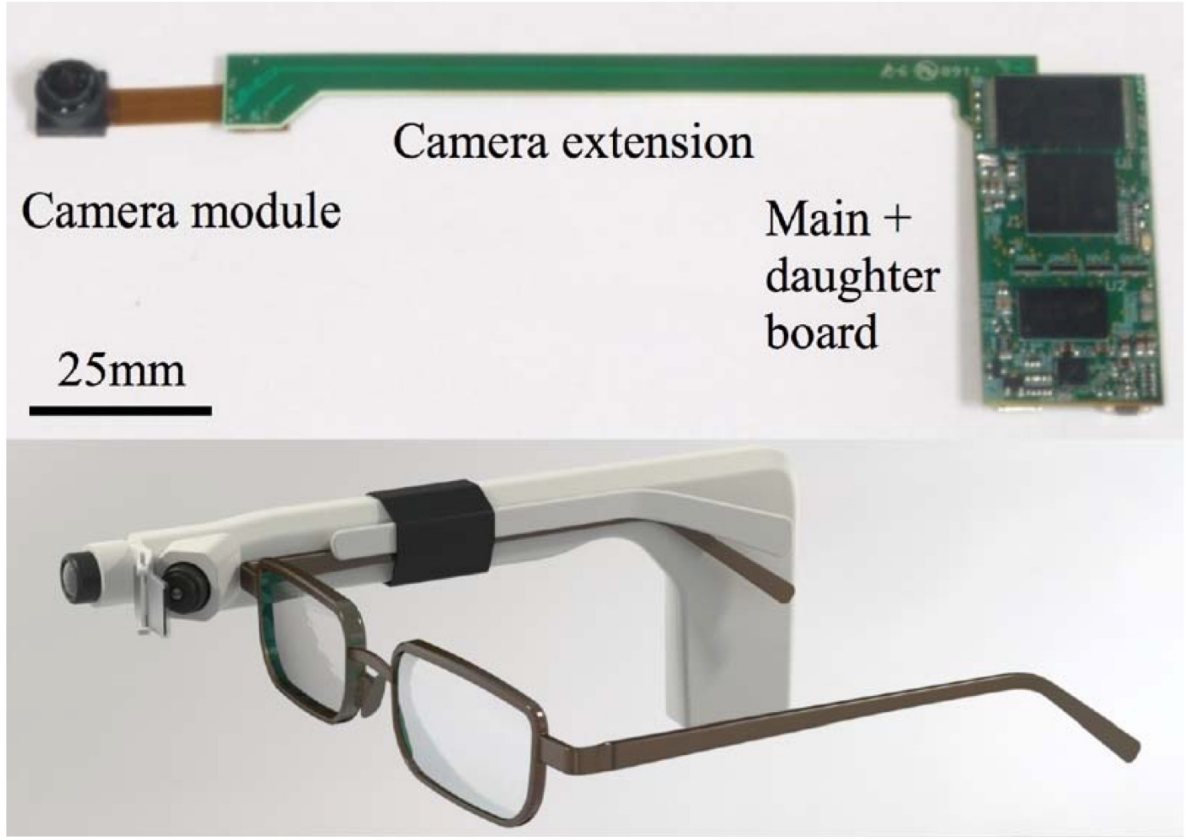


Fig. 2. ***A version of the FPV system. Top: Electronics for on-board image capture and recording; Bottom: Casing attached to a pair of eyeglasses showing both the inward- and outward-looking cameras.***

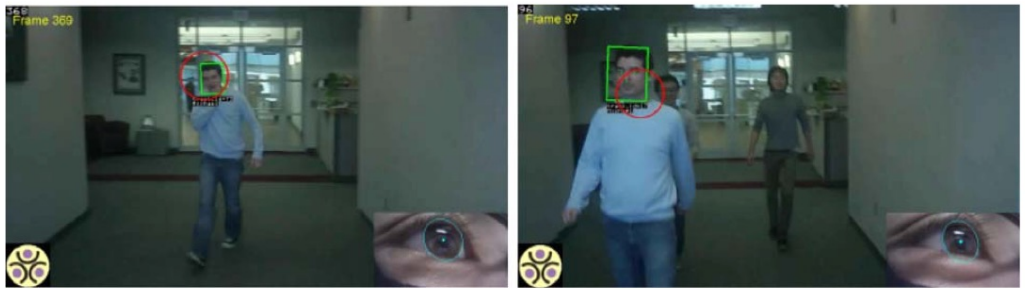


Fig. 12. ***Recognizing people from FPV: The location of the detected face is shown in green and the gaze direction estimated by the eye tracker is shown in red. The name of the recognized person is displayed.***

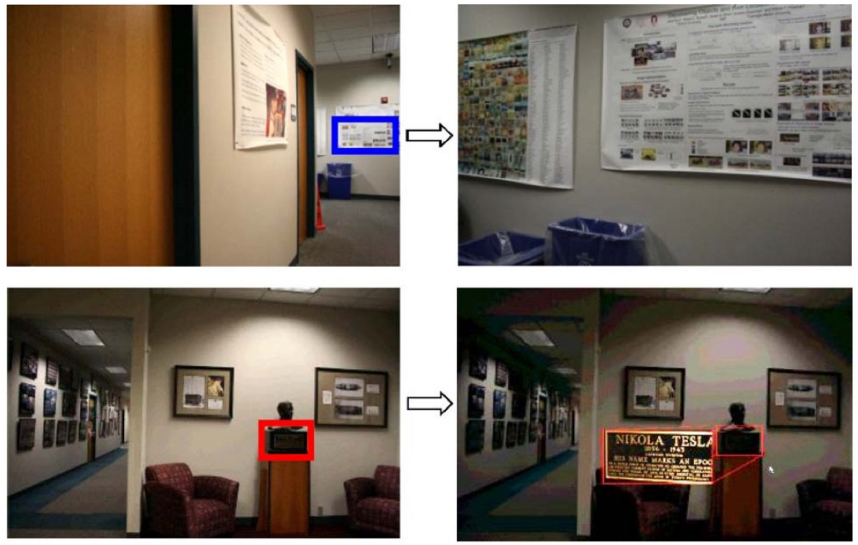


Fig. 7. ***Intelligent zoom concept: Images with legible signage (right) are generated by matching images from FPV (left) with a large database of images.***

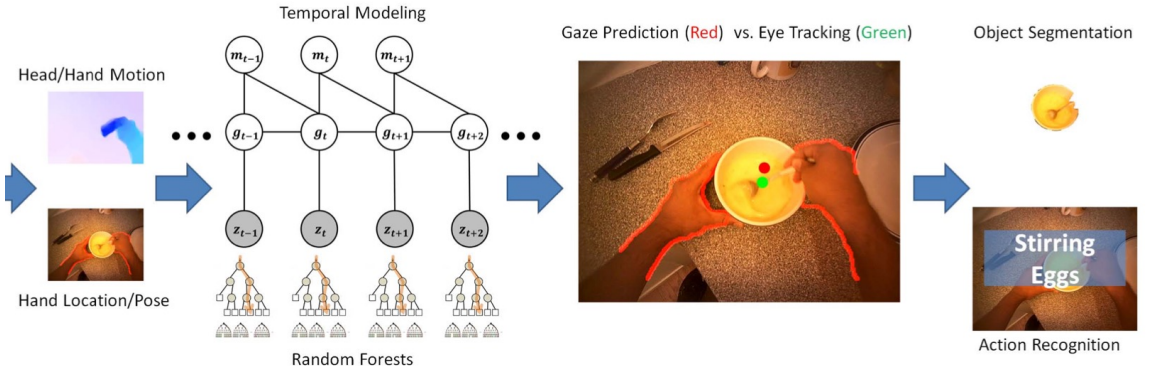
Kanade, T., & Hebert, M. (2012). First-person vision. *Proceedings of the IEEE*, *100*(8), 2442-2453.

Research On Gaze

Learning to Predict Gaze in Egocentric Video

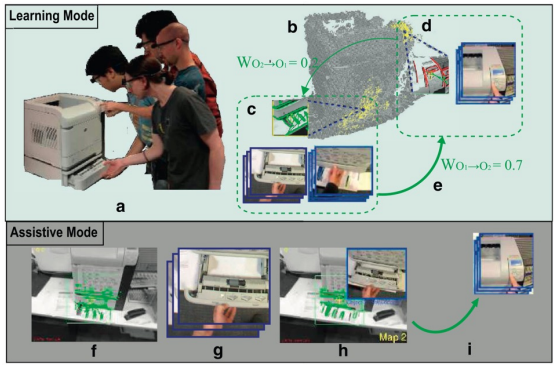
Yin Li, Alireza Fathi, James M. Rehg

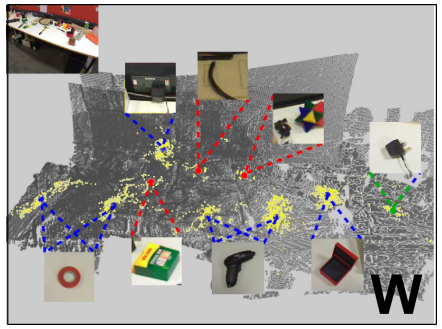
(gaze prediciton, action recognition)



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

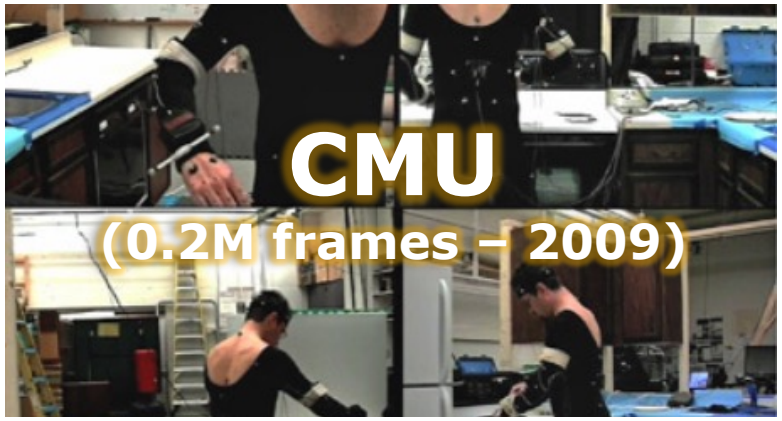
Workshop on Egocentric (First Person) Vision

ACVR

EPIC

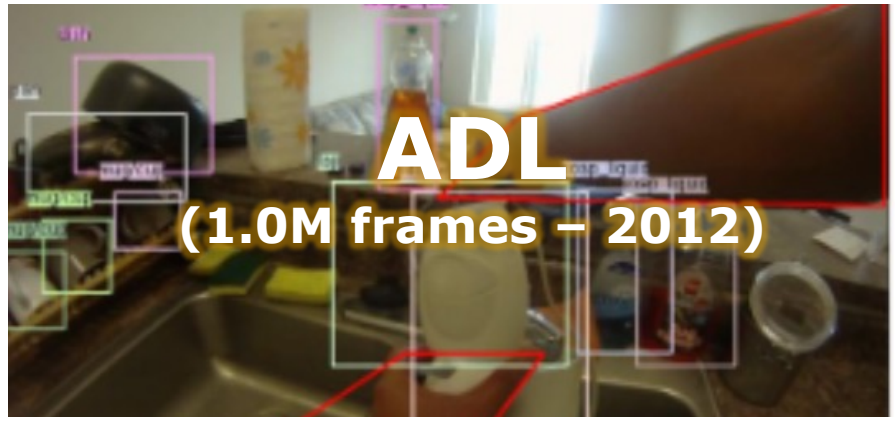
EGOAPP **LTA**

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.

EPIC KITCHENS

University of BRISTOL

UNIVERSITÀ degli STUDI di CATANIA

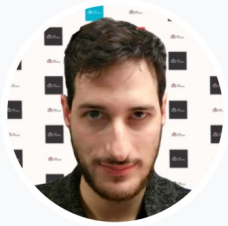
https://epic-kitchens.github.io/

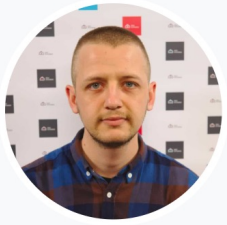
EPIC-KITCHENS-100


Dima Damen
University of Bristol


Hazel Doughty
University of Bristol


Giovanni M. Farinella
University of Catania

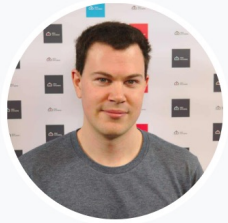

Antonino Furnari
University of Catania


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

Improved annotation pipeline

EPIC-KITCHENS-55

VIDEO ACQUISITION

AUDIO COMMENTARY

ACTION SEGMENTS

SEMANTIC PARSING

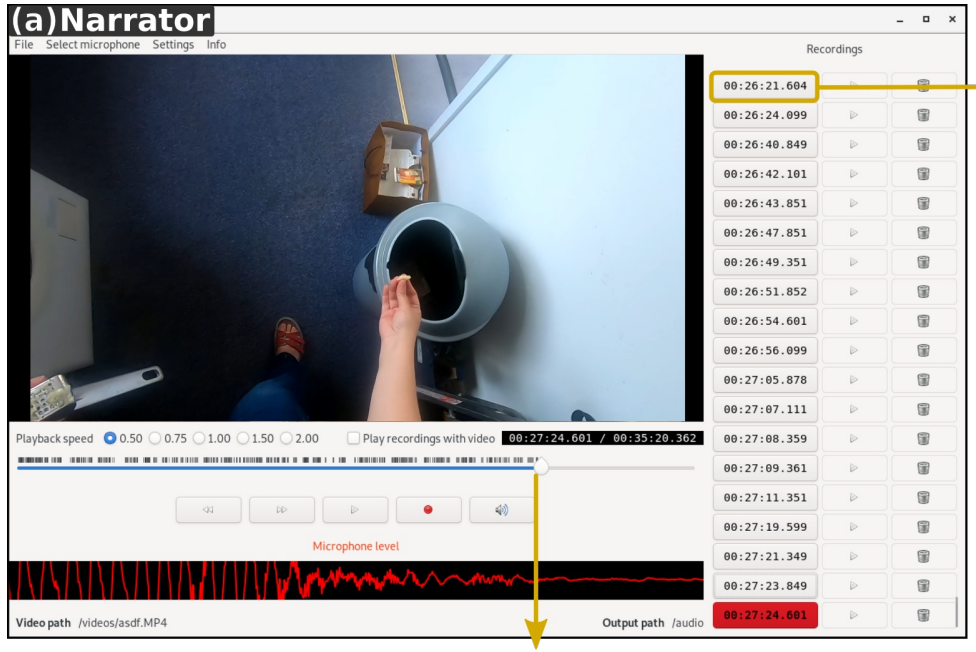
VERB-NOUN REPRESENTATION

CUT ONION

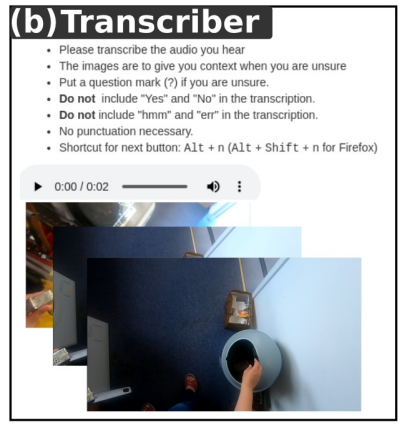
TURN-OFF TAP

DRY CUP

EPIC-KITCHENS-100



narration



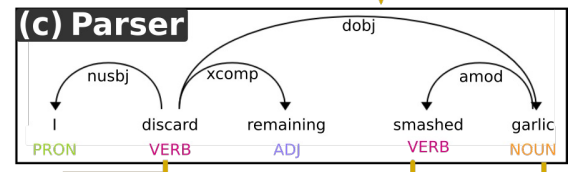
清理桌上多余的蒜泥

caption

Discard remaining smashed garlic

timestamp

action segment



<discard>

13: throw

verb class

<garlic:smashed>

51: garlic

noun class

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/

2023 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

Can We Scale?

Consortium

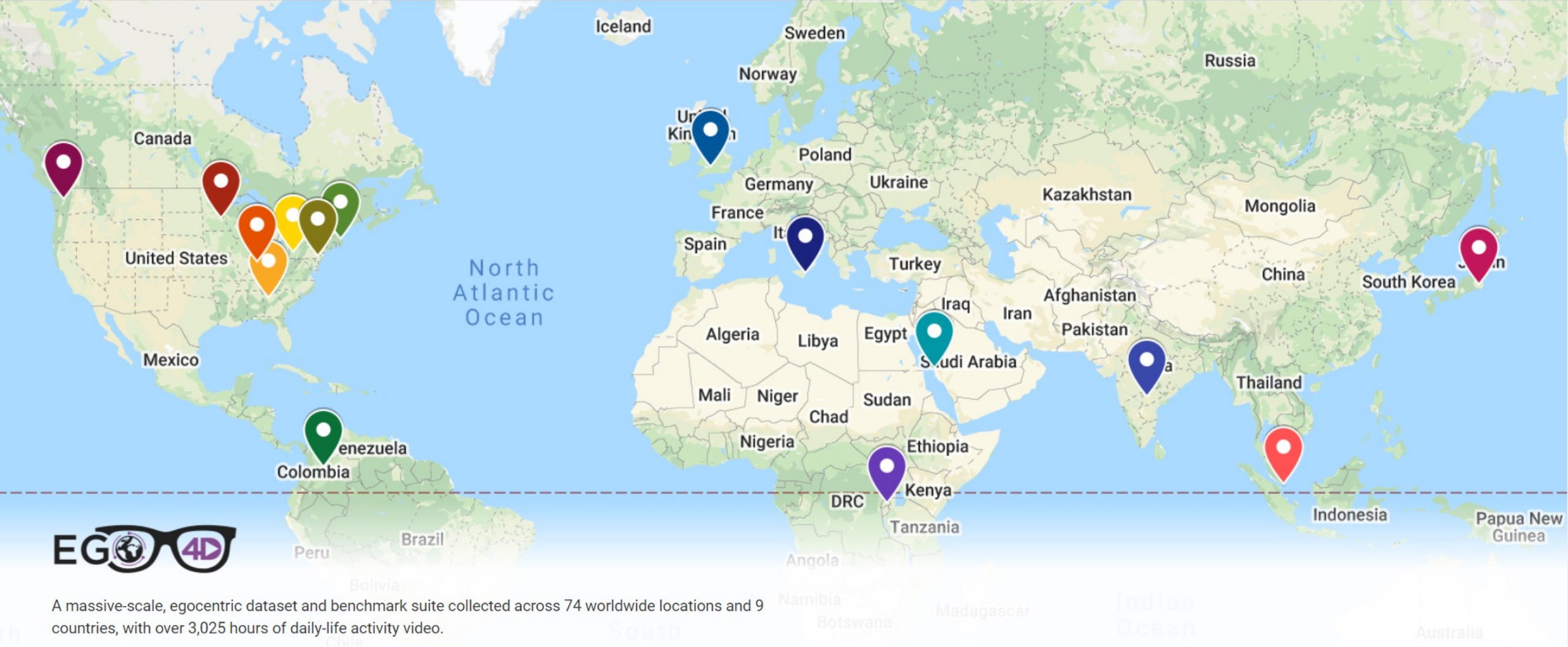
Carnegie Mellon University
Università di Catania
NUS National University of Singapore
King Abdullah University of Science and Technology
東京大学 THE UNIVERSITY OF TOKYO
University of BRISTOL
INDIANA UNIVERSITY BLOOMINGTON
UNIVERSITY OF MINNESOTA
MIT
Penn University of Pennsylvania
Georgia Institute of Technology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



EGO 4D

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

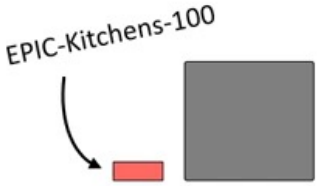
Ego4D – A Massive-Scale Egocentric Dataset

3,025 Hours

855 Participants

5 Benchmark Tasks

Find out more: https://ego4d-data.org/



Animation by Michael Wray – https://mwray.github.io

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

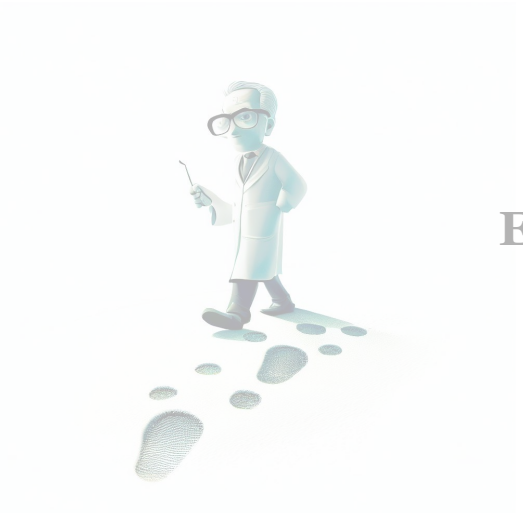
24 October 2022

3rd International Ego4D Workshop @ CVPR 2023

Held in conjunction with 11th EPIC Workshop

19 June 2023

1



Egocentric Vision: A Retrospective

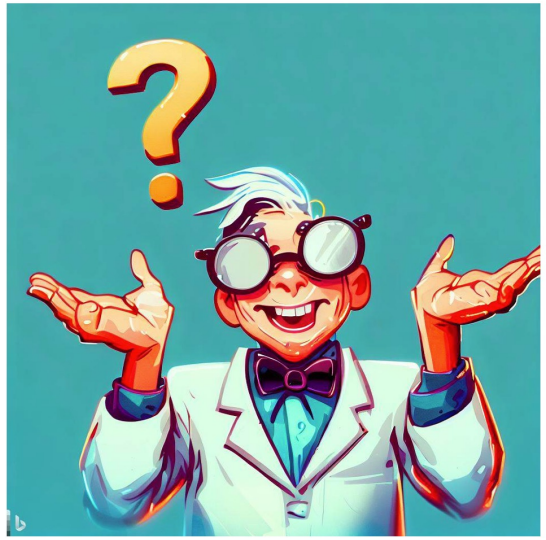
2

The Cyborg Dream

3

An Outlook into the Future

4



Doing Research in Egocentric Vision: Where to start?

The Cyborg Dream

Google Glass, 2012

- 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.

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

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)

Manufacturing Solutions

LEARN MORE

Warehouse Solutions

LEARN MORE

Field Service & Remote Assist Solutions

LEARN MORE

Tele-Medicine Solutions

LEARN MORE

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

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.

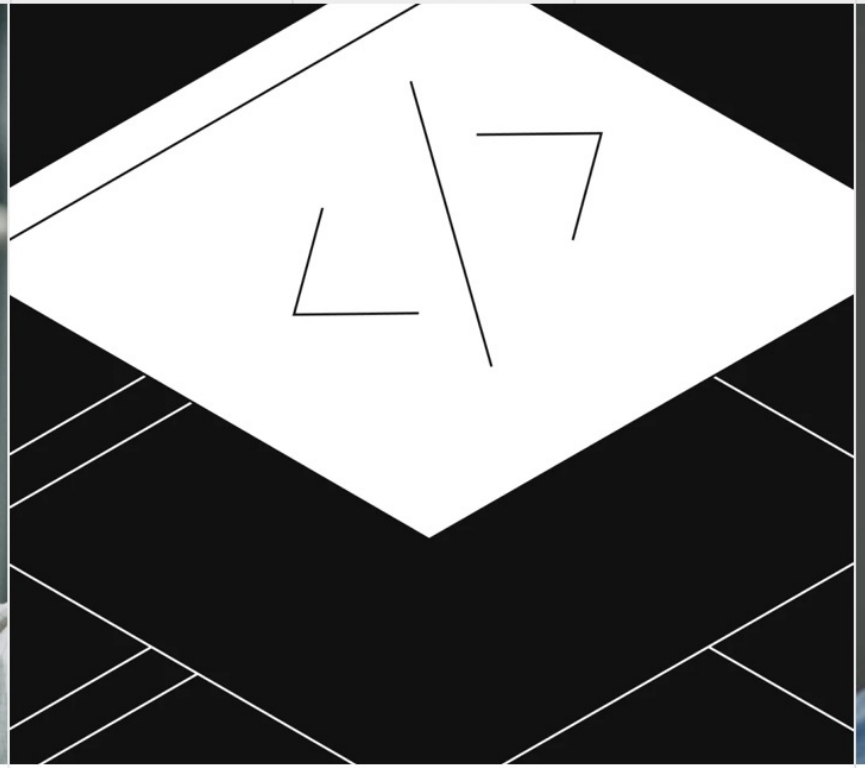
$5,199

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

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

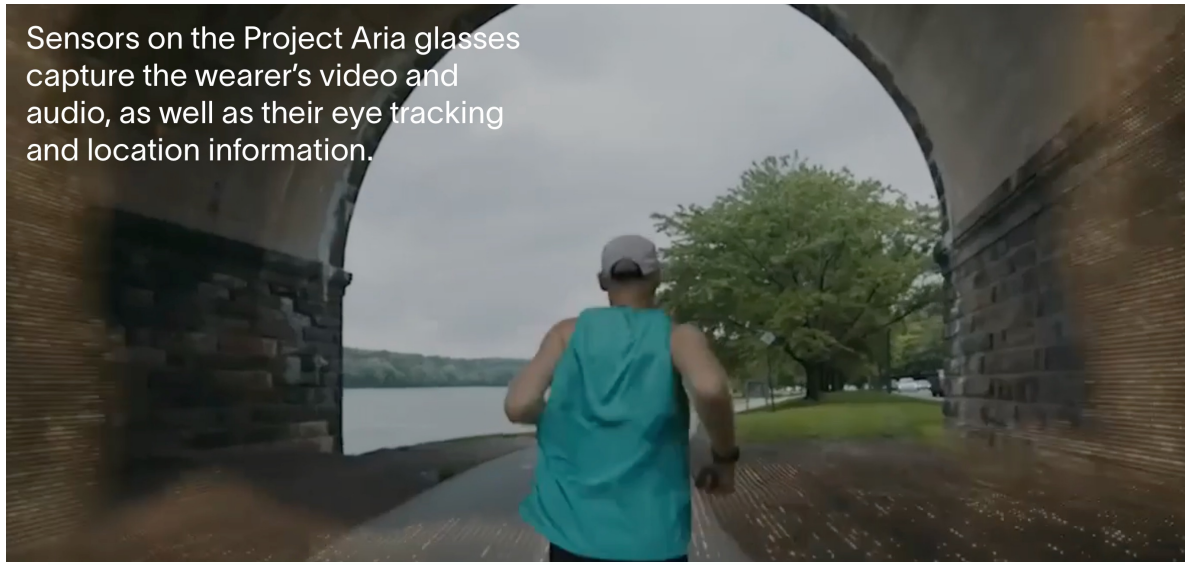
Magic Leap 2 is purpose-built on an open platform to integrate with leading enterprise multi-device management (MDM) systems.

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/

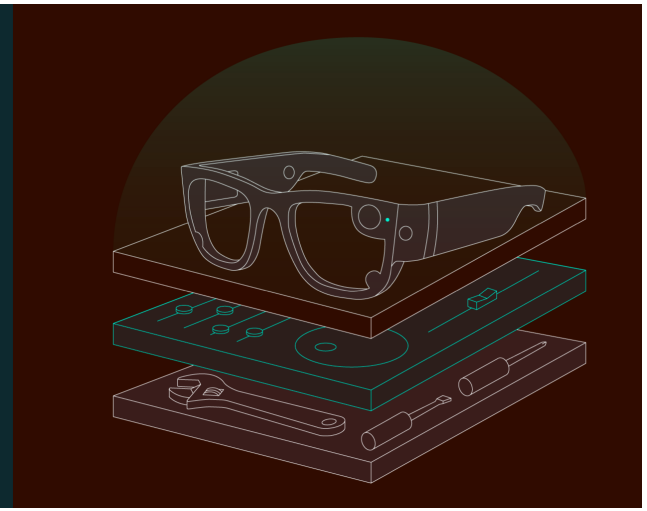
Università di Catania — Meta's Project Aria

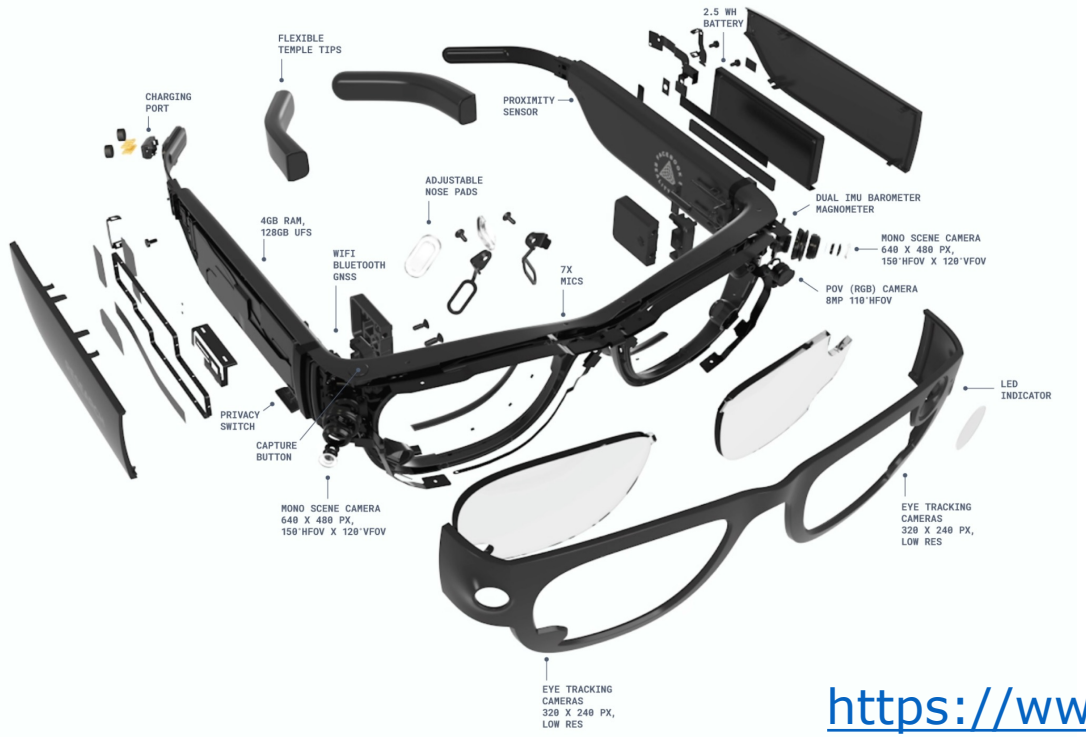


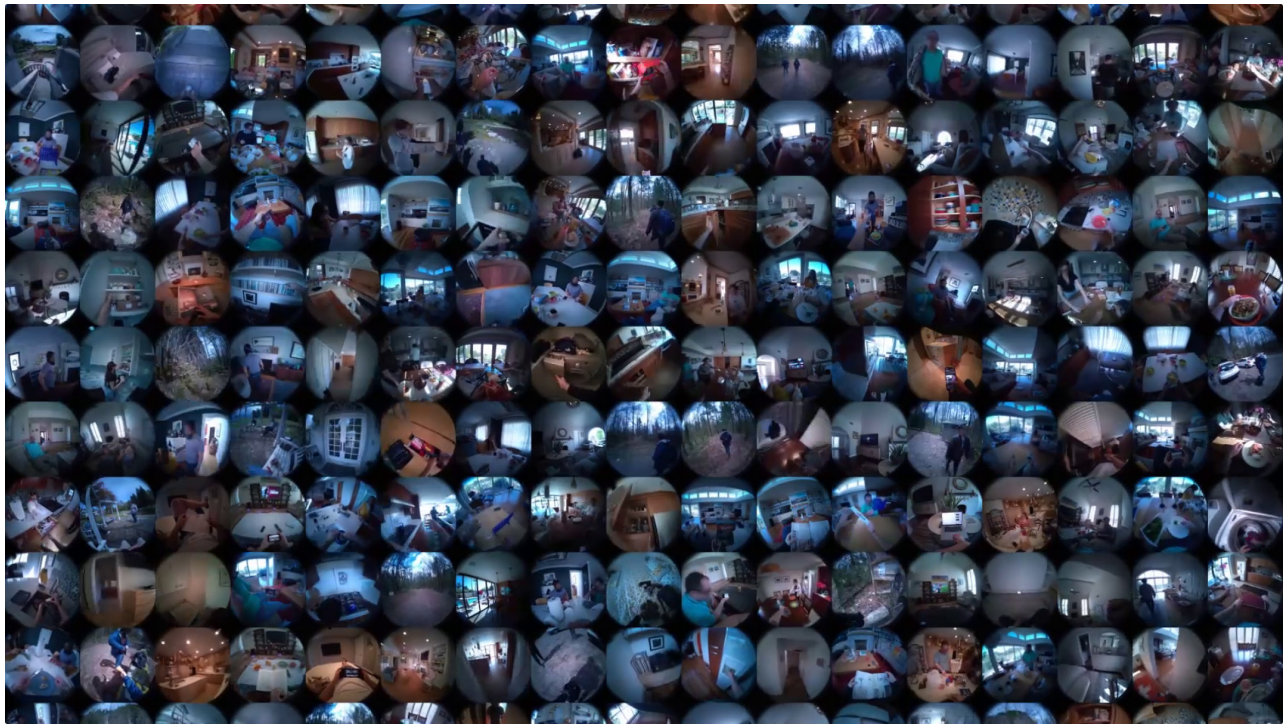
Aria Research Kit

For approved research partners, Meta offers a kit that includes Project Aria glasses and SDK, so that researchers can conduct independent studies and help shape the future of AR.

LEARN MORE ABOUT PARTNERING WITH PROJECT ARIA

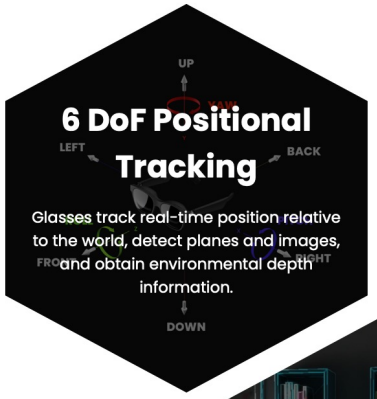


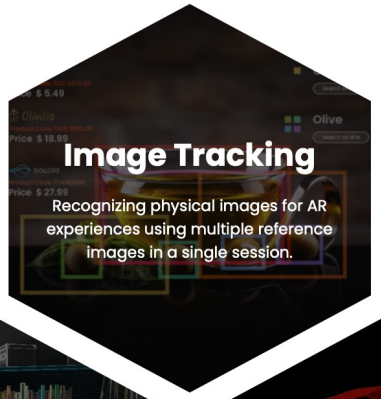


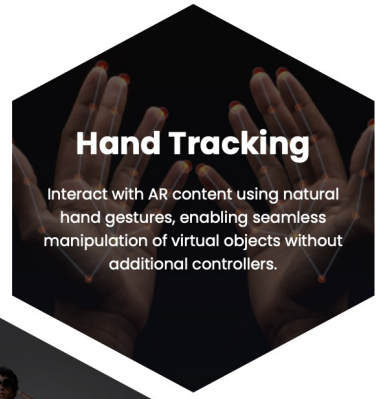


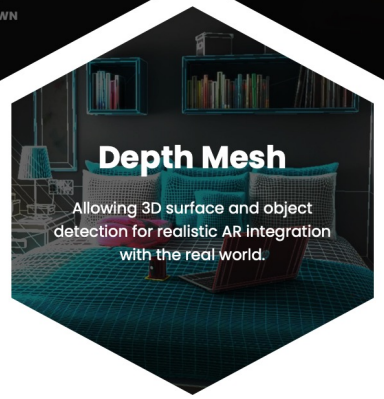
https://www.projectaria.com

Development Kit



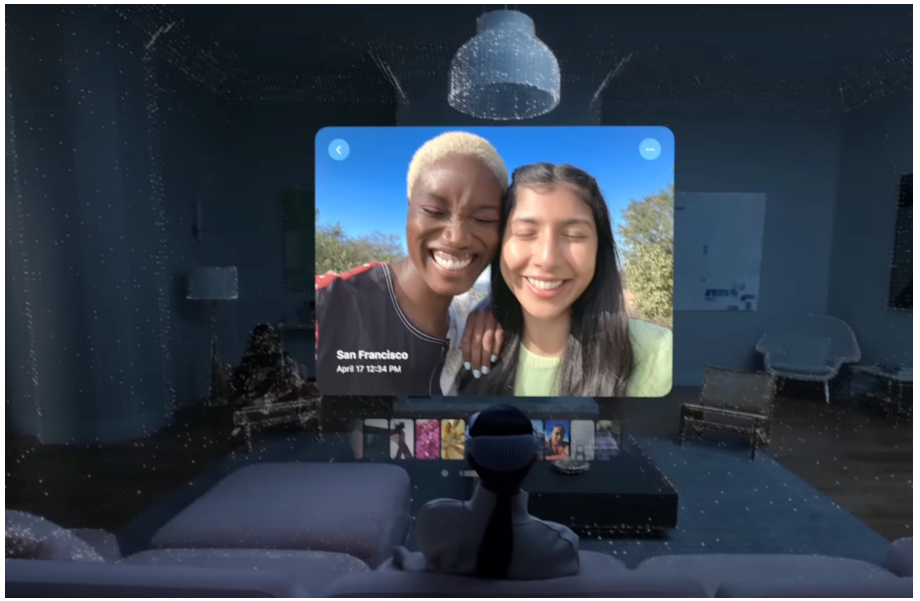


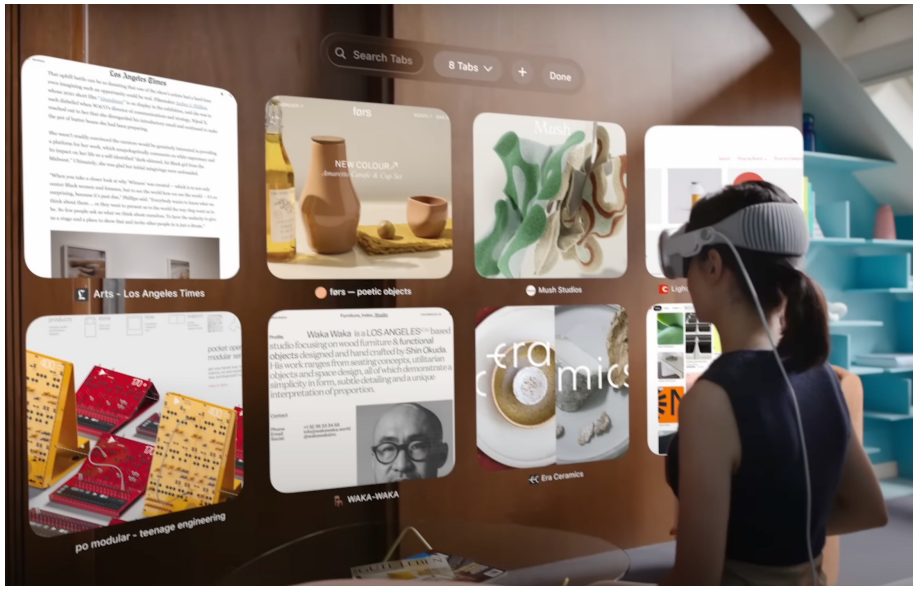


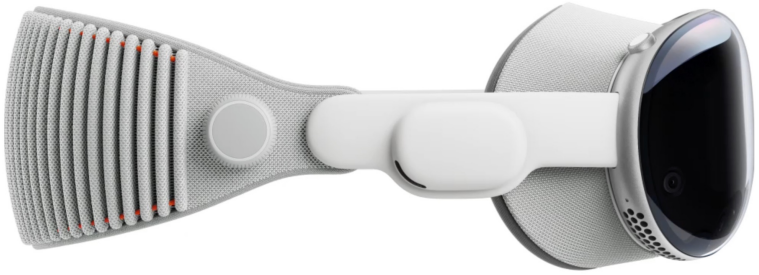


https://www.xreal.com/

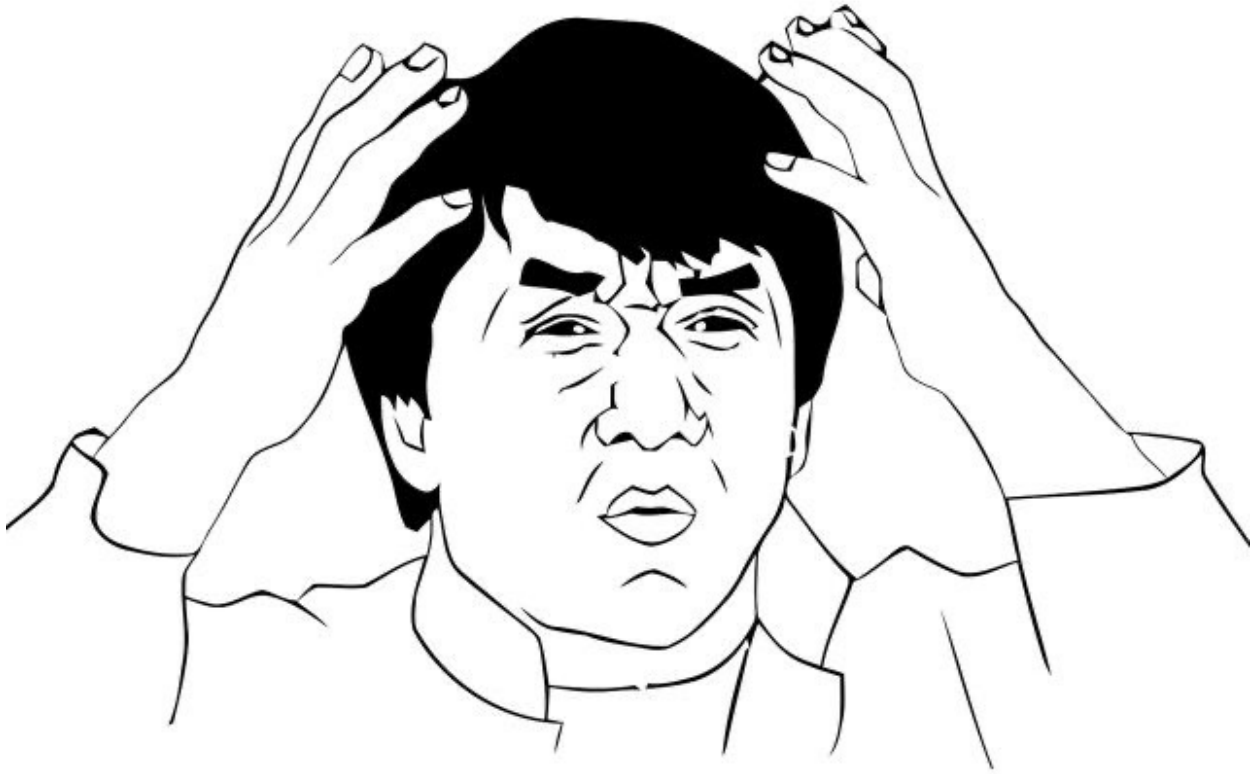
Apple Vision Pro







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



Too Many Devices?

towards standardization…

OpenXR

Unified API supported by many AR and VR devices

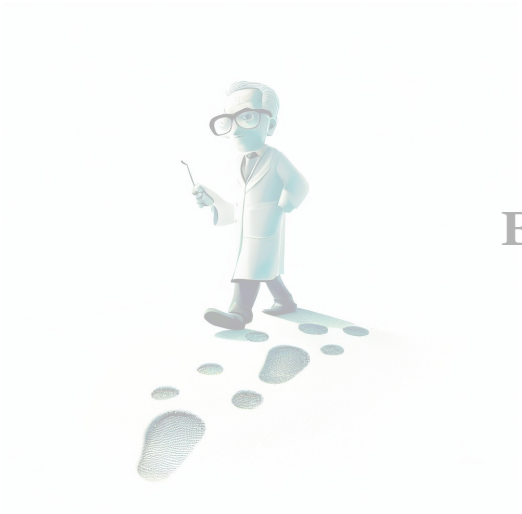
https://www.khronos.org/openxr/

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

1



Egocentric Vision: A Retrospective

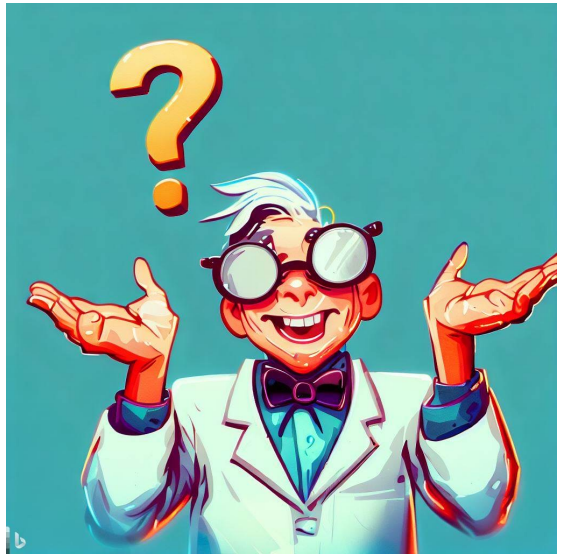
2

The Cyborg Dream

3

An Outlook into the Future

4



Doing Research in Egocentric Vision: Where to start?

An Outlook into the Future

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

Immagine the Future

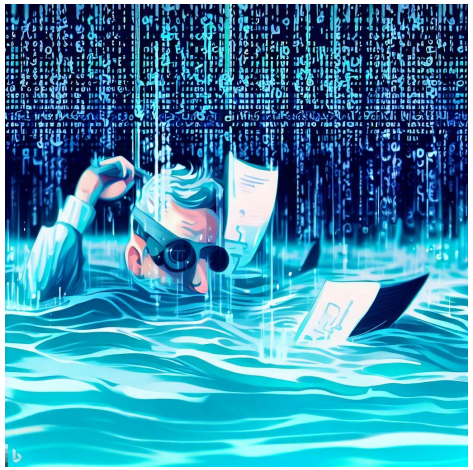
Write Stories in Different Scenarios

Extract Important Tasks from the Stories

Go in-depth with Tasks and Datasets

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

A lot of data!



An Outlook into the Future

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†

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, Anticipation, Recognition, Gaze Prediction, Social Understanding, Body Pose Estimation, Hand and Hand-Object Interaction, Person Identification, Privacy, Summarisation, VQA

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 are present in movies, fictional novels and pop culture[1]. Notwithstanding the fast progress of Artificial Intelligence, and the hardware advances of the last ten years, our ability to fulfil this dream is lagging behind.

In computer vision, research papers on egocentric vision have instead limited their focus to a handful of applications, where current technology can already make a difference. These are: training or monitoring in industrial settings, performing adhoc and infrequent tasks such as assembling a piece of furniture, preparing a new recipe, or playing a group game in a social setting. These showcase egocentric wearables as niche devices very distant from everyone's everyday needs. This perspective has not only limited our chances to convince others that egocentric vision is a key technology of our future, but it also restricted our ability to push the boundaries and remove obstacles to the integration of egocentric devices as the ultimate replacement of the *mobile phone* with unlocking of additional capabilities.

*: 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] Few examples: (1) Molly's Vision-Enhancing Lenses from the *Neuromancer* novel, William Gibson, 1984. (2) JVC Personal Video Glasses from the *Back to the Future II* movie, 1989. (3) Iron Man Suits with J.A.R.V.I.S. AI system from Marvel movies 2008-2015. (4) AI Earbuds and smartphone in shirt pocket from the *Her* movie, 2013. (5) E.D.I.T.H. smart glasses from the *Spider-Man: Far From Home* movie, 2019.

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.

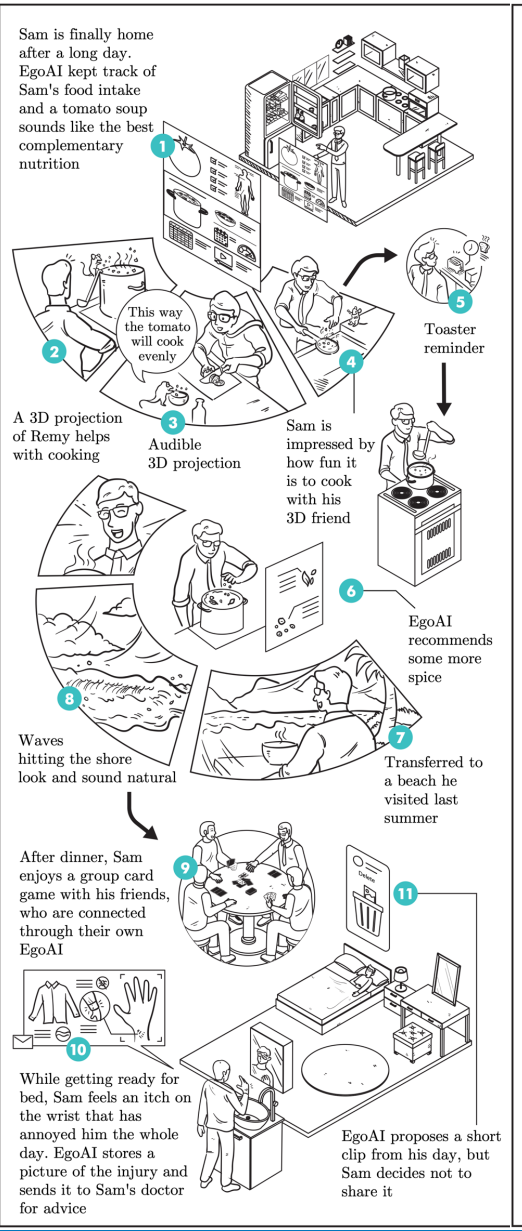
https://arxiv.org/abs/2308.07123

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

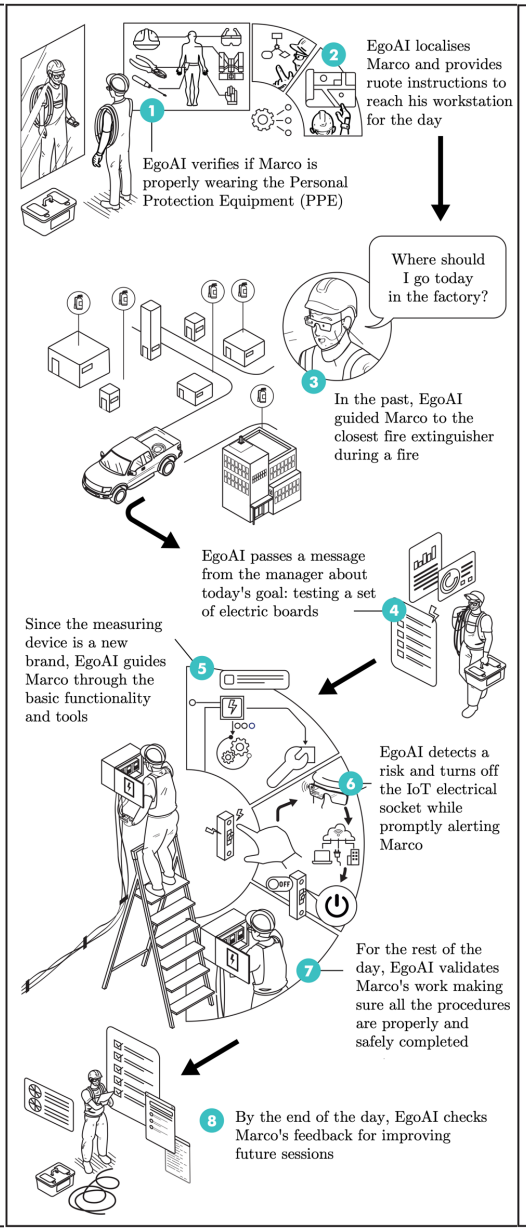
Università di Catania

An Outlook into the Future – Futuristic Stories

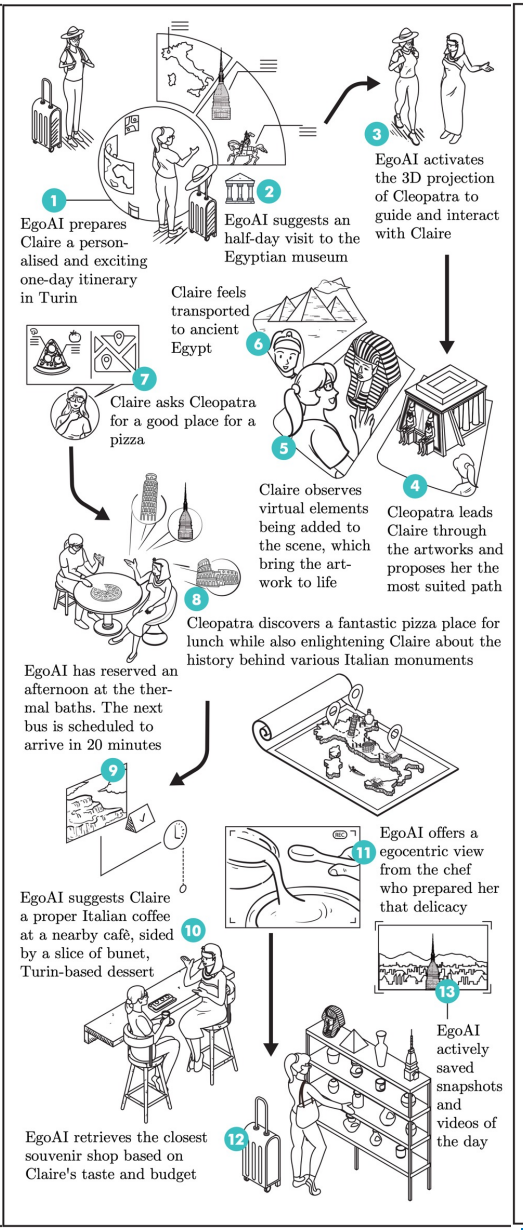
EGO-HOME



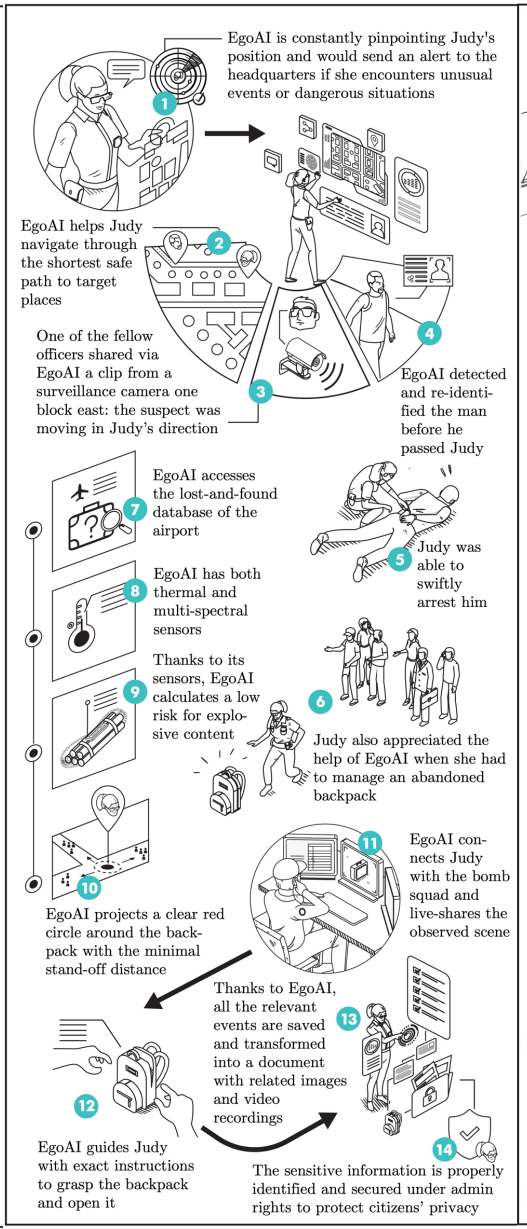
EGO-WORKER



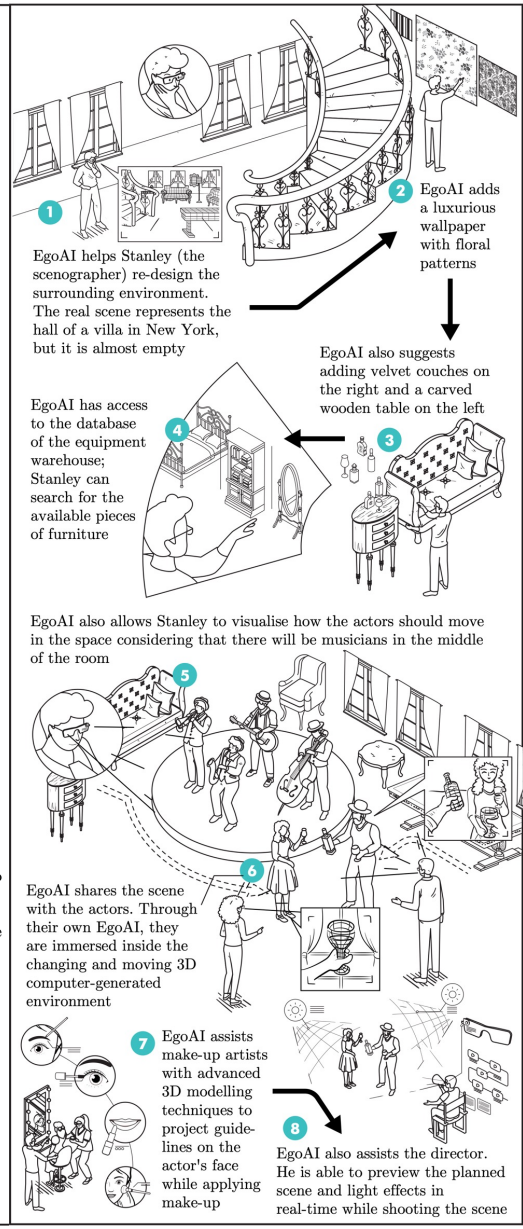
EGO-TOURIST



EGO-POLICE



EGO-DESIGNER



From Narratives to Research Tasks

12 Egocentric Vision Research Tasks

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

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

Links between Stories and Tasks

EGO-Home

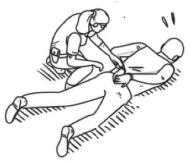
- 3D Scene Understanding 1 2 3 4 7 8 9
- Object and Action Recognition 1 5 6 10
- Measuring Systems 6
- Visual Question Answering 6
- Summarisation and Retrieval 7
- Full-Body Pose and Social Interaction 9
- Medical Imaging 10
- Messaging 10 11
- Summarisation 11

EGO-Worker

- Safety Compliance Assessment 1
- Localisation and Navigation 2 5
- Messaging 4
- Hand-Object Interaction 5
- Action Anticipation 6
- Skill Assessment 7
- Visual Question Answering 8
- Summarisation 8

EGO-Tourist

- Recommendation and Personalisation 1 2 8 9 10 11
- 3D Scene Understanding 2 3 4 5 6
- Gaze Prediction 5
- Localisation and Navigation 3 4 8 12
- Messaging 7
- Visual Question Answering 8
- Action Recognition and Retrieval 11
- Summarisation 13



EGO-Police

- Localisation and Navigation 1 2
- Messaging 1 3 11
- Action Recognition 2 13
- Person Re-ID 2 4
- Object Detection and Retrieval 7
- Measuring System 8 9
- Decision Making 9
- 3D Scene Understanding 10
- Hand-Object Interaction 12
- Summarisation 13
- Privacy 14

EGO-Designer

- 3D Scene Understanding 1 2 3 4 5 6 7 8
- Recommendation 3
- Object Recognition and Retrieval 3 4
- Full-Body Pose Estimation 5 6
- Social Interaction 6
- Gaze Prediction 6
- Hand-Object Interaction 7
- Messaging 6 8

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 Dataset - 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 2022b)	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

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 2022b)	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

Table 3 General Egocentric Datasets - Current set of tasks: **4.1** Localisation, **4.2** 3D Scene Understanding, **4.3** Anticipation, **4.4** Action Recognition, **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** Privacy, **4.11** Summarisation, **4.12** Visual Question Answering.

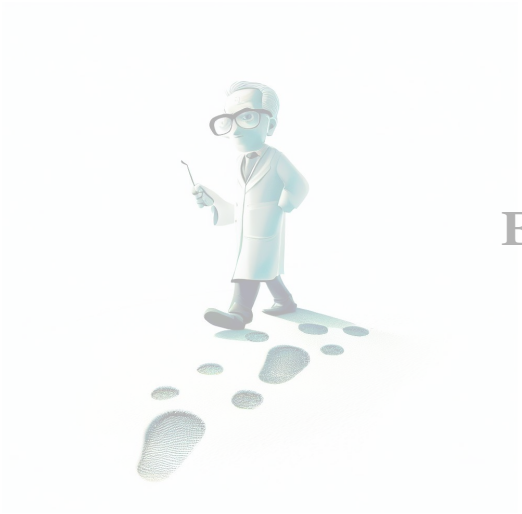
4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	4.10	4.11	4.12
		✓	✓	✓			✓				
		✓	✓							✓	
							✓				
		✓	✓	✓			✓				
							✓			✓	
✓											
					✓				✓	✓	
		✓									
	✓	✓	✓				✓		✓		
			✓				✓				
		✓	✓	✓	✓		✓			✓	✓

Mapping to tasks

Data Statistics

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*.

1



Egocentric Vision: A Retrospective

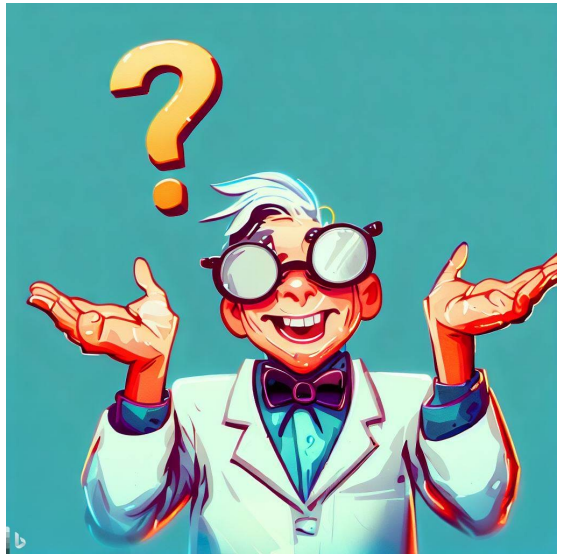
2

The Cyborg Dream

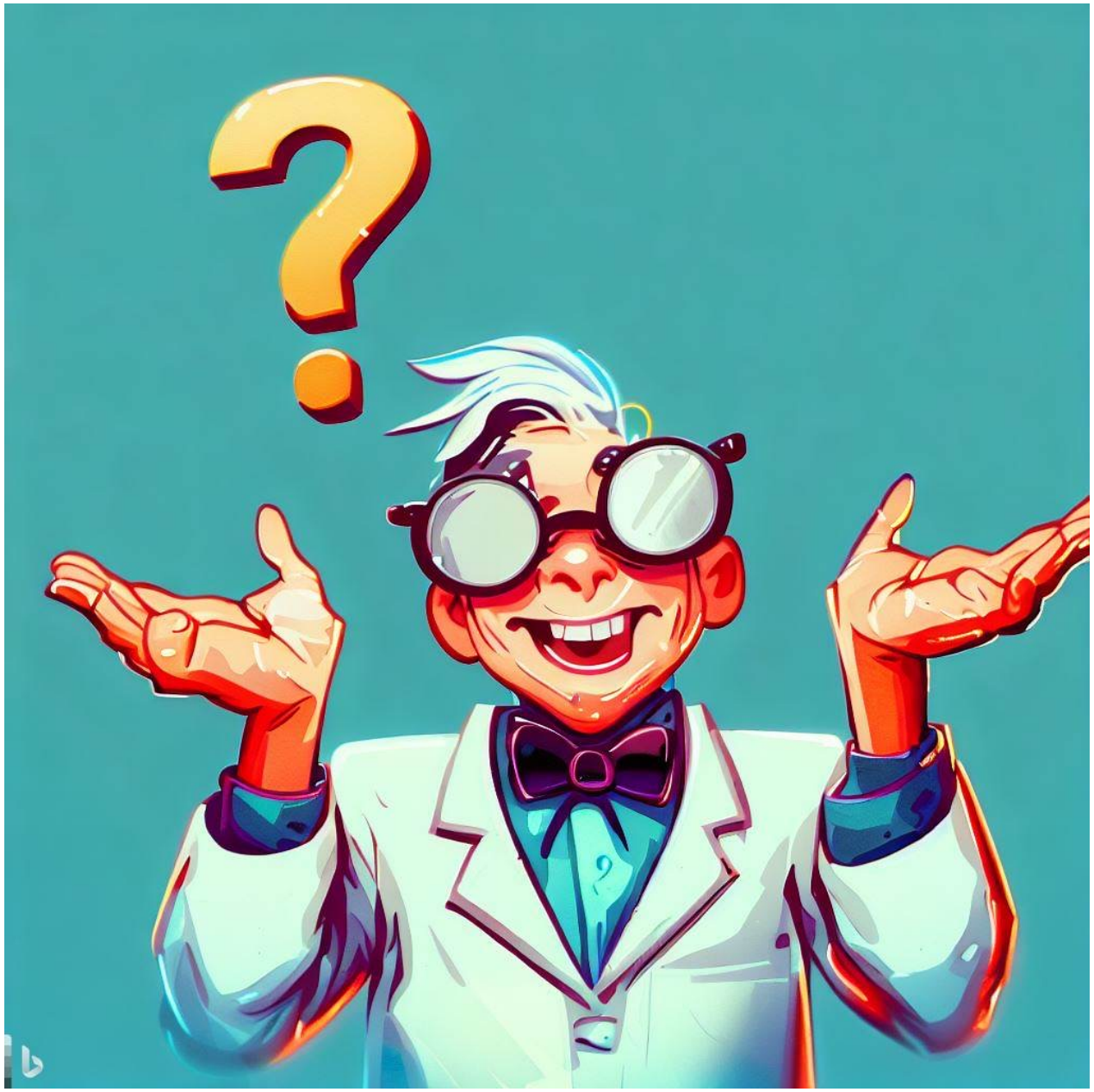
3

An Outlook into the Future

4

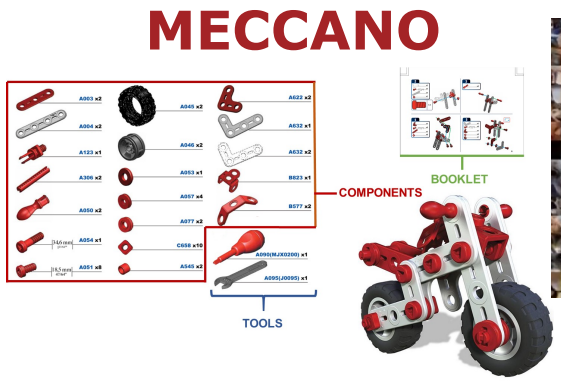


Doing Research in Egocentric Vision: Where to start?



Doing Research in Egocentric Vision: Where to start?

Lots of Data Out There

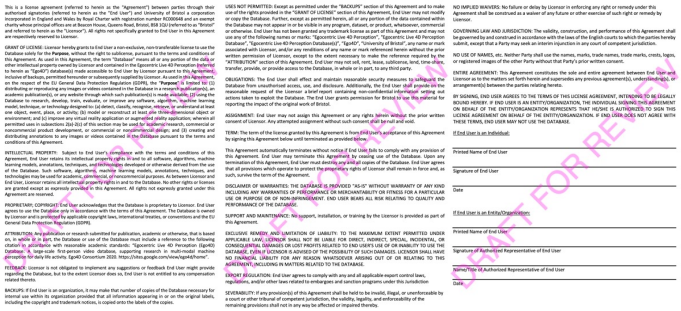


Plizzari, C., Goletto, G., Furnari, A., Bansal, S., Ragusa, F., Farinella, G. M., Damen., D. & Tommasi, T. (2023). An Outlook into the Future of Egocentric Vision. *arXiv preprint arXiv:2308.07123*. **And More**

Accessing Modern Datasets – License

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

! pay attention to which uses are permitted!



EPIC KITCHENS

ABOUT STATS DOWNLOADS CHALLENGES TEAM

Disclaimer

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

Copyright

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

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

EGO4D License Agreement

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

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

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

Ego4D Dataset

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

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

Note: Only official signatories can sign as organisation

Individual Organization

First name | Last name

Email

Home Address

City | State / Province / County | Country

Submit

Accessing Modern Datasets – Command Line Interfaces

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

Download only certain data types

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

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

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

If you want to download only videos, then:

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

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

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

Will download both RGB and optical flow frames.

Specifying participants

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

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

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

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

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

Modern datasets are HUGE!

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

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

Data download

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

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

v2.0 annotations can be downloaded with:

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

Detailed Flags

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

Datasets

The following datasets are available (not exhaustive):

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

EGO 4D

Command Line Interfaces Provided to Simplify Download

Lots of Tasks Out There – Challenges!

EPIC-KITCHENS-100 2023 CHALLENGES

Challenge Details with links to ★NEW★ Codalab Leaderboards

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

In 2023, we have 9 open challenges. These are

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

EPIC-Kitchens 2023 Challenges

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

Challenges Guidelines

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

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

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

Ego4D Challenge 2023

Episodic memory:

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

Hands and Objects:

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

Audio-Visual Diarization:

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

Social Understanding:

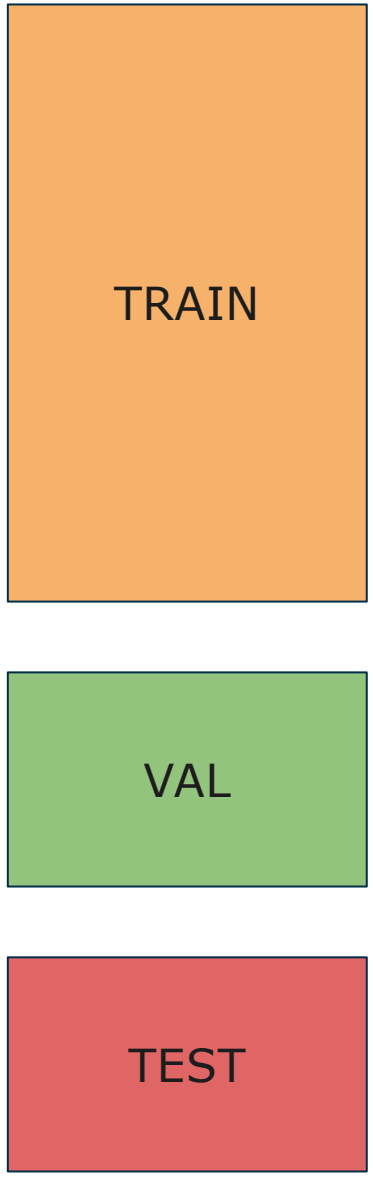
- Talking to me: Given an egocentric video clip, identify whether someone in the scene is talking to the camera wearer.
- Looking at me: Given an egocentric video clip, identify whether someone in the scene is looking at the camera wearer.

Forecasting:

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

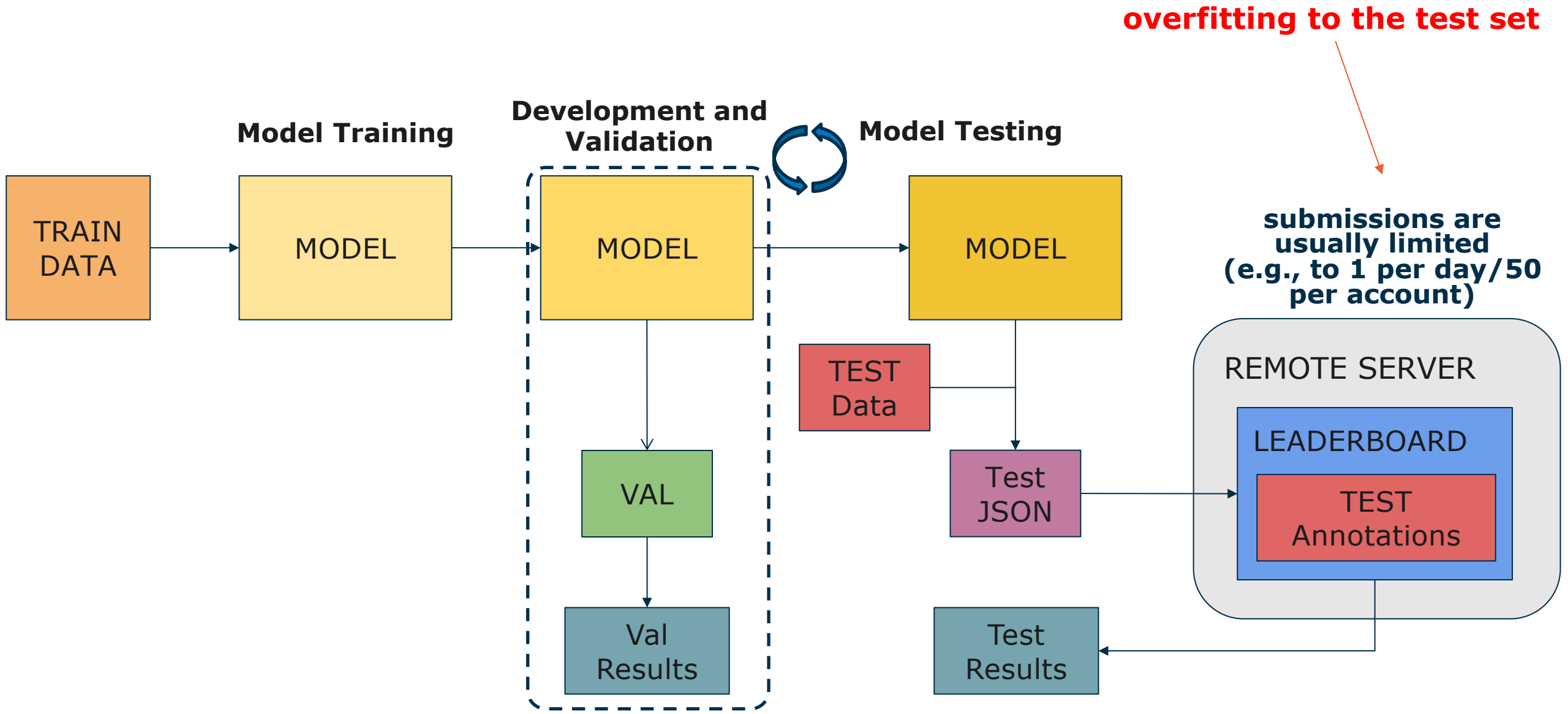
https://ego4d-data.org/docs/challenge/

Challenges – Train/Val/Test scheme



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

Challenges – Evaluation Server



Challenges – Evaluation Server

EPIC-KITCHENS-100 Action Anticipation

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

▶ Current: 2023 Open Testing Phase — June 27, 2023, 8 a.m. UTC

End: Competition Ends — Nov. 25, 2023, 11 p.m. UTC

Test Set (Mean Top-5 Recall)

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

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

Ego4D Short Term Object Interaction Anticipation Challenge

★ 11

Toggle Participation

Organized by: Ego4D
Published
Starts on: Oct 25, 2022 2:00:00 AM CET (GMT + 2:00)
Ends on: May 20, 2024 2:00:59 AM CET (GMT + 2:00)

Discuss

Leaderboard

Overall Top-5 mAP

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

Order by metric

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

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

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

Challenges – A Good Way to Start!

C3-Action-Anticipation

Challenge

To submit and participate to this challenge, register at the Action Anticipation Codalab Challenge

Evaluation Code

https://github.com/epic-kitchens/C3-Action-Anticipation

This repository contains the official code to evaluate egocentric action anticipation methods on the EPIC-KITCHENS-100 validation set.

Requirements

In order to use the evaluation code, you will need to install a few packages. You can install these requirements with:

```
pip install -r requirements.txt
```

Usage

You can use this evaluation code to evaluate submissions on the valuation set in the official JSON format. To do so, you will need to first download the public EPIC-KITCHENS-100 annotations with:

```
git clone https://github.com/epic-kitchens/epic-kitchens-100-annotations.git
```

You can then evaluate your json file with:

```
python evaluate_anticipation_json_ek100.py path_to_json path_to_annotations
```

Example json file

We provide an example json file which has been generated using our "chance" action anticipation baseline. To evaluate this json, you first need to unzip its archive with:

```
unzip action_anticipation_chance_baseline_validation.zip
```

After that, you can evaluate the json file with:

Short-Term Object Interaction Anticipation

https://github.com/EGO4D/forecasting/blob/main/SHORT_TERM_ANTICIPATION.md

Please note that this code refers to the old baseline. The code for the new baseline is available here: https://github.com/fpv-iplab/stillfast

This README reports information on how to train and test the baseline model for the Short-Term Object Interaction Anticipation task part of the forecasting benchmark of the Ego4D dataset. The following sections discuss how to download and prepare the data, download the pre-trained models and train and test the different components of the baseline.

This code has been tested both with v1.0 and v2.0 data. See here for more information on the v2.0 update.

Data

The first step is to download the data using the CLI avaiable at https://github.com/facebookresearch/Ego4d.

Data download

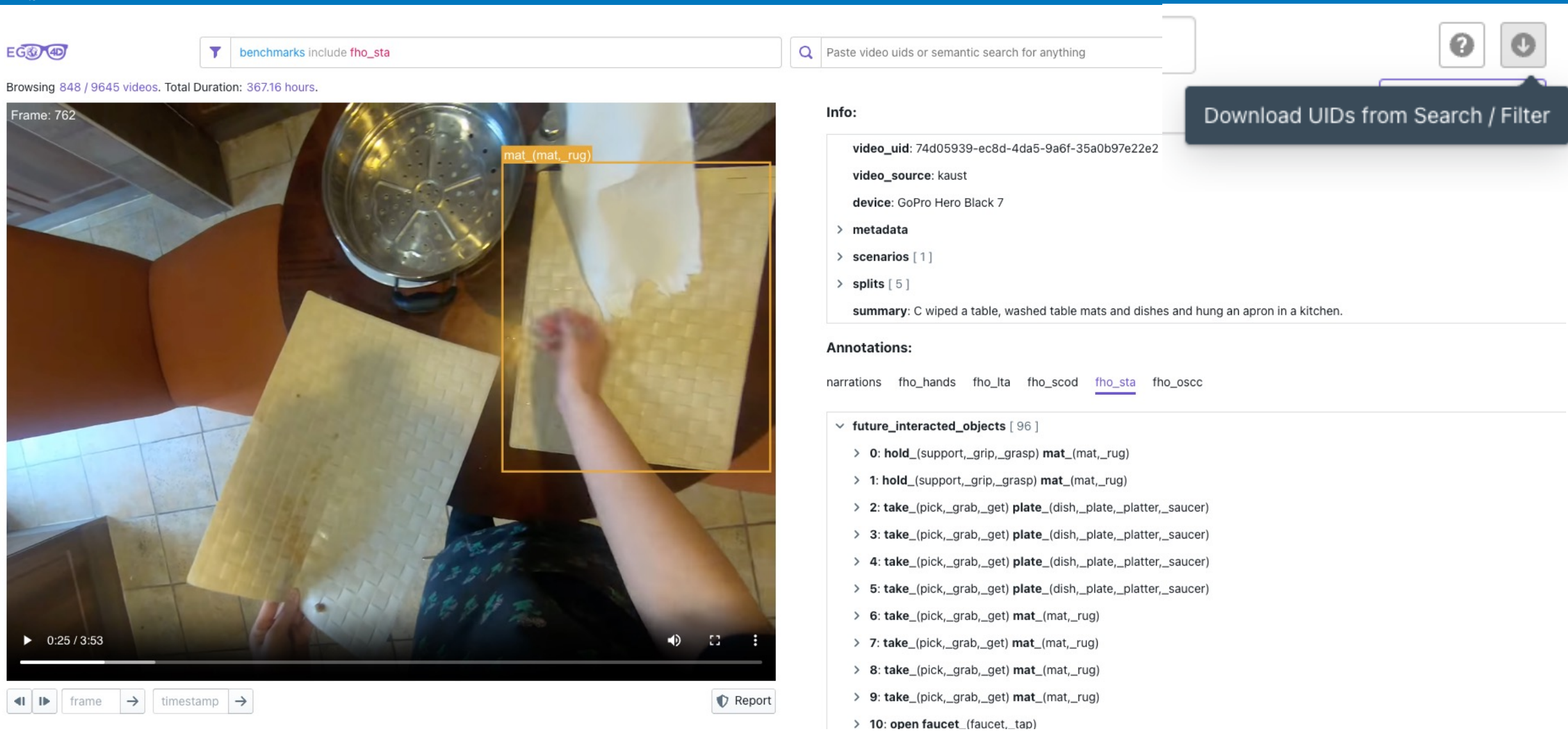
Canonical videos and annotations can be downloaded using the following command:

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

How to teach an old dog new tricks

Use existing data to investigate new tasks.

Università di Catania EGO4D Visualizer



https://visualize.ego4d-data.org/

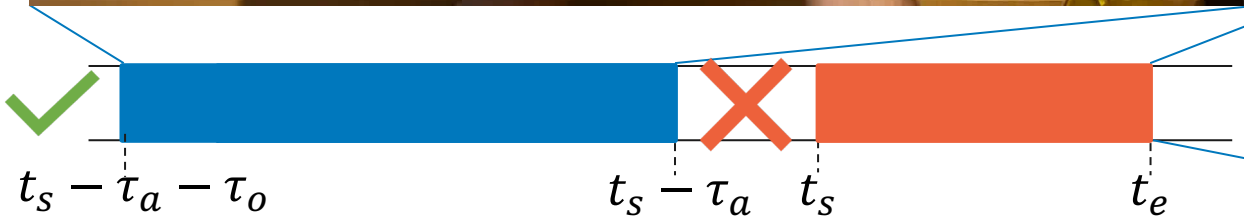
Practical: Rolling-Unrolling LSTMs

Egocentric Action Anticipation Task

(observed video)

Model

Take - Plate



τ_o arbitrary

$\tau_a = 1s$;

(unobserved)

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

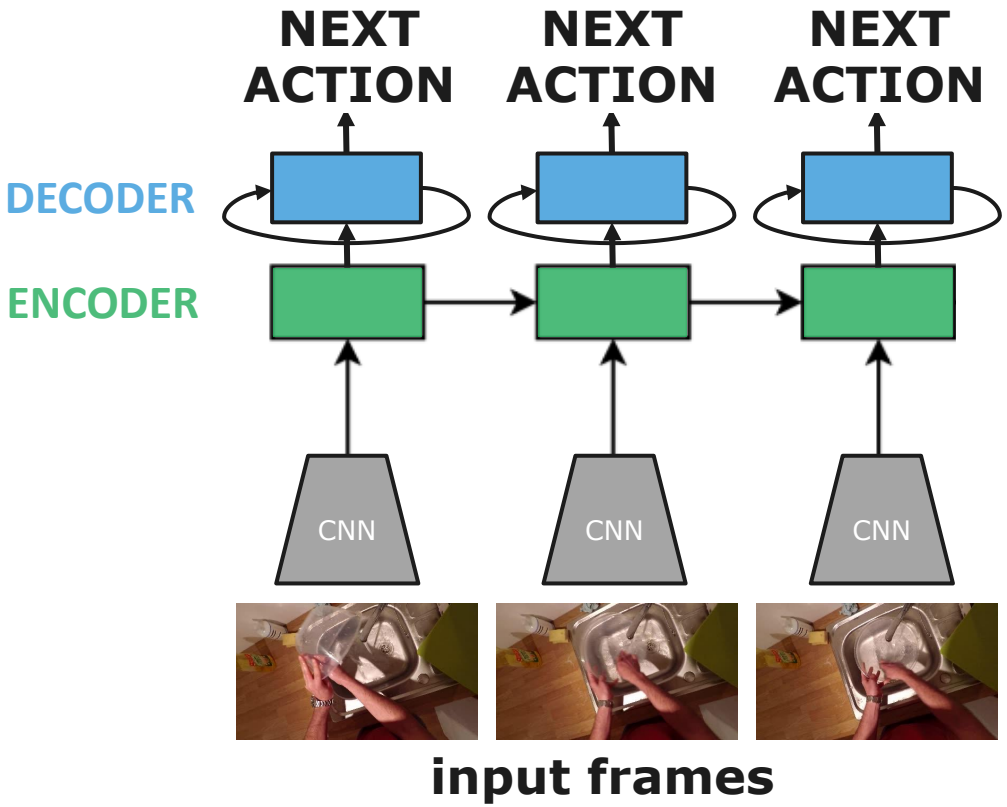
Dima Damen et al. Rescaling Egocentric Vision . International Journal on Computer Vision (IJCV). 2021

Sequence to Sequence Models

Rolling-LSTM **ENCODING**

Unrolling-LSTM **INFERENCE**

We take inspiration from sequence to sequence models.



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

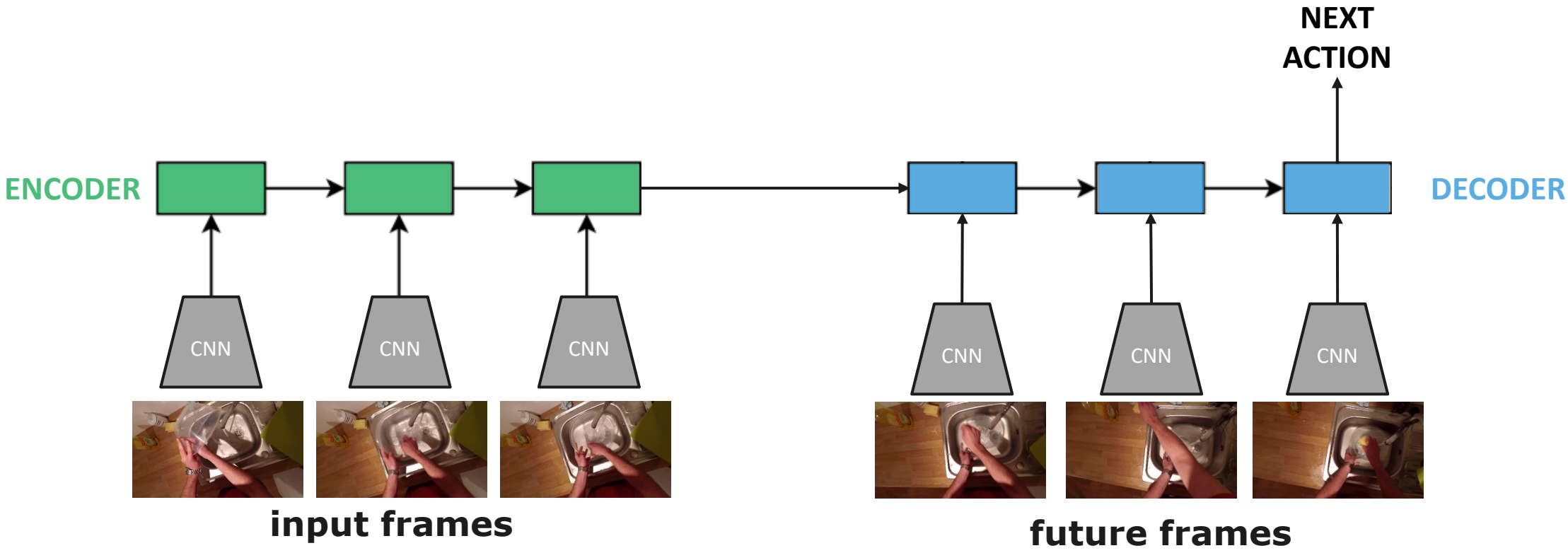
A. Furnari, G. M. Farinella. Rolling-Unrolling LSTMs for Action Anticipation from First-Person Video. TPAMI 2020. http://iplab.dmi.unict.it/rulstm

Sequence Completion Pre-Training

Rolling-LSTM **ENCODING**

Unrolling-LSTM **INFERENCE**

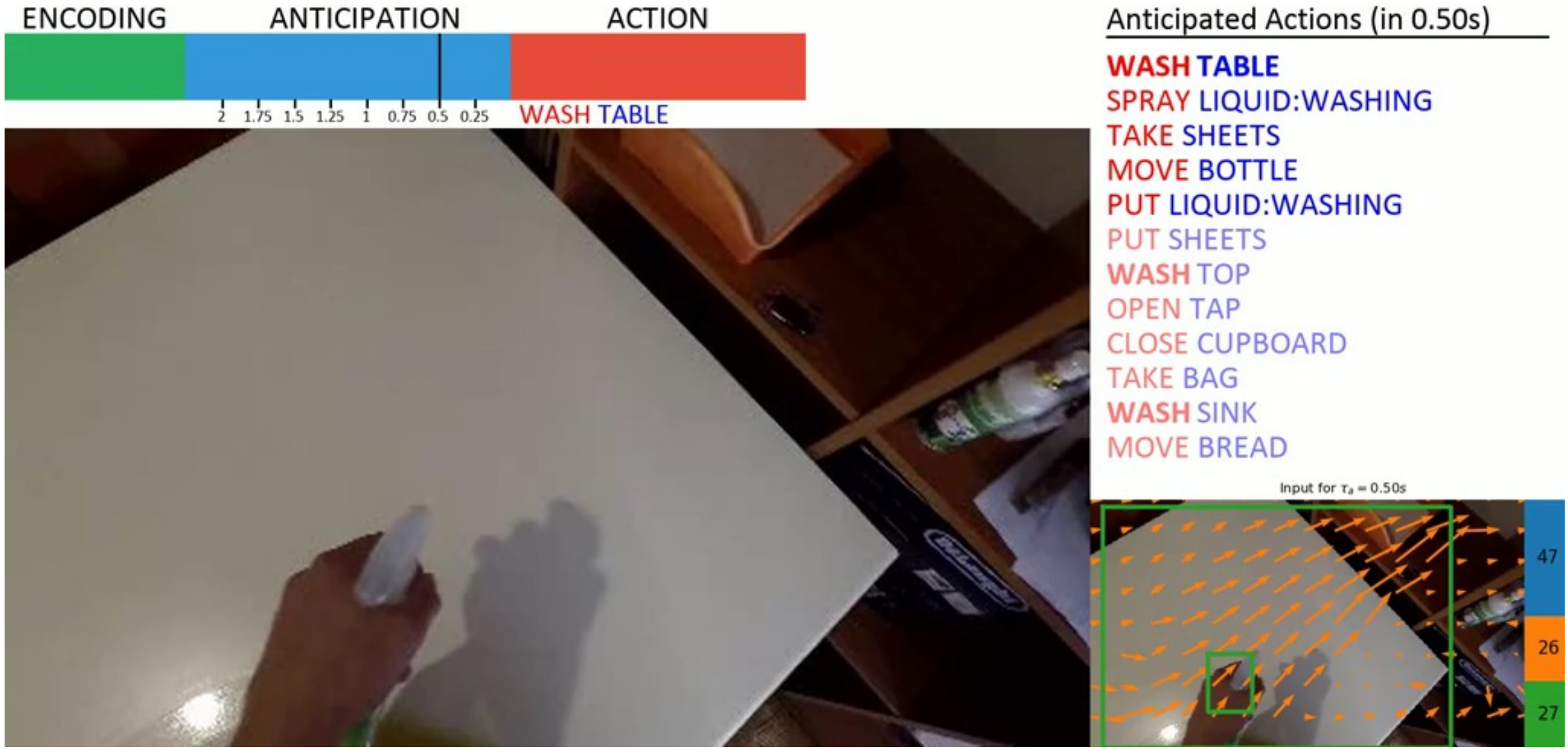
To encourage the Rolling-LSTM to only perform encoding and not anticipation, we pre-train the model feeding future frames to the Unrolling-LSTM.



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

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

Demo Video: Egocentric Action Anticipation



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

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

Rolling-Unrolling LSTM Quickstart

1. Go to: https://github.com/fpv-iplab/rulstm
2. Then click on «Open in Colab»

README.md

What Would You Expect? Anticipating Egocentric Actions with Rolling-Unrolling LSTMs and Modality Attention

See the quickstart here 👉 Open in Colab

This repository hosts the code related to the following papers:

Antonino Furnari and Giovanni Maria Farinella, Rolling-Unrolling LSTMs for Action Anticipation from First-Person Video. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). 2020. Download

Antonino Furnari and Giovanni Maria Farinella, What Would You Expect? Anticipating Egocentric Actions with Rolling-Unrolling LSTMs and Modality Attention. International Conference on Computer Vision, 2019. Download

Please also see the project web page at http://iplab.dmi.unict.it/rulstm.

If you use the code/models hosted in this repository, please cite the following papers:

```
@article{furnari2020rulstm,
  author = {Antonino Furnari and Giovanni Maria Farinella},
  journal = {IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)},
  title = {Rolling-Unrolling LSTMs for Action Anticipation from First-Person Video},
  year = {2020}
}
```

```
@inproceedings{furnari2019rulstm,
  title = { What Would You Expect? Anticipating Egocentric Actions with Rolling-Unrolling LSTMs and
  author = { Antonino Furnari and Giovanni Maria Farinella },
  year = { 2019 },
  booktitle = { International Conference on Computer Vision (ICCV) },
}
```

3. Follow the instructions

Rolling-Unrolling LSTM Quickstart

Antonino Furnari - antonino.furnari@unict.it - https://www.antoninofurnari.it/

Introduction

This quickstart will guide you through a simplified training loop for the Rolling-Unrolling LSTM model. Please refer to the official repository for more information: https://github.com/fpv-iplab/rulstm and all options.

Preliminaries

For this quickstart, we need to install the `lmdb` library, which allows to access the LMDB dataset containing our data.

Let's install our library with the following command:

```
[1] !pip install lmdb

    Collecting lmdb
      Downloading lmdb-1.4.1-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (299 kB)
         ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 299.2/299.2 kB 5.4 MB/s eta 0:00:00
    Installing collected packages: lmdb
    Successfully installed lmdb-1.4.1
```

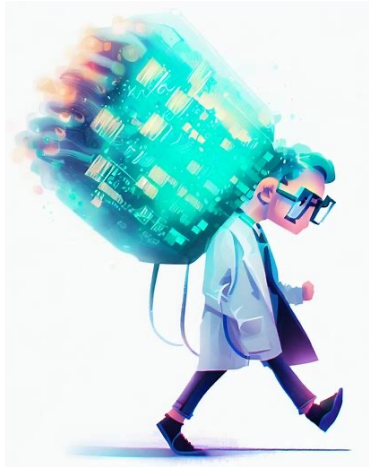
4. Answer the questions along the way

> **Question 1**
>
> Have a look at the `main.py` file in the repository and answer the following questions:
>
> - Where is the training loop located?
> - Where is the model loaded?
> - Where does the logging happen?

Conclusion

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

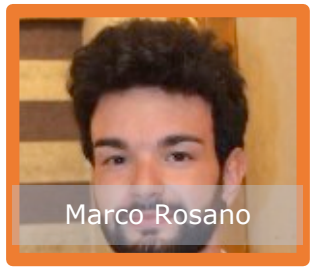
The FPV@IPLAB Group

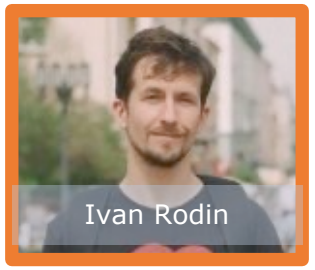

Giovanni Maria Farinella


Antonino Furnari


Francesco Ragusa

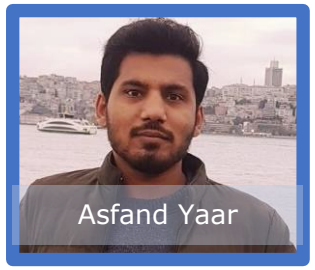

Daniele Di Mauro


Marco Rosano


Ivan Rodin

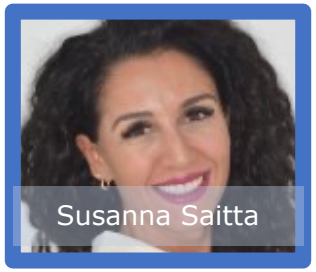

Rosario Leonardi

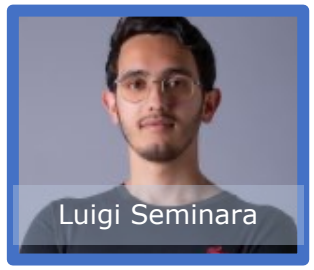

Camillo Quattrocchi


Asfand Yaar

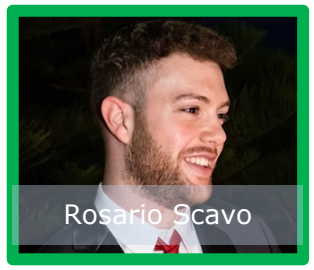

Michele Mazzamuto


Claudia Bonanno


Susanna Saitta


Luigi Seminara


Adriana Maccarrone


Rosario Scavo


Rosario Forte


Francesca Del Zoppo

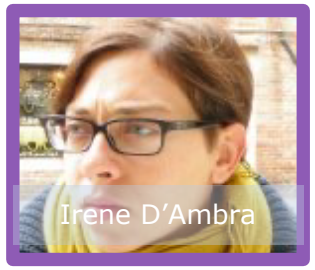

Irene D'Ambra

IMAGE PROCESSING LABORATORY

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

NEXT VISION

http://www.nextvisionlab.it/

17 Members

1 Full Professor

1 Assistant Professor

1 Researcher

3 Post Docs

7 PhD Students

3 Master Students

1 Intern

1 Lab Assistant






Università di Catania

IMAGE PROCESSING LABORATORY

FPV @ IPLAB Group

Università di Catania

We Are Hiring!

VISMAC 23

International Summer School on Machine Vision

Università di Catania

NEXT VISION

IMAGE PROCESSING LABORATORY

Tutorial on Egocentric Vision

Thank You!

Antonino Furnari

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

antonino.furnari@unict.it - http://www.antoninofurnari.it/