First Person (Egocentric) Vision for Human-Centric Assistance: History, Building Blocks, and Applications

Antonino Furnari, Francesco Ragusa

Image Processing Laboratory - http://iplab.dmi.unict.it/

Department of Mathematics and Computer Science - University of Catania

Next Vision s.r.l., Italy

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

francesco.ragusa@unict.it - https://iplab.dmi.unict.it/ragusa/

http://iplab.dmi.unict.it/fpv - https://www.nextvisionlab.it/

IMAGE PROCESSING LABORATORY

iplab.dmi.unict.it

- Located in Catania, Sicily, Italy
- More than 25 people:
  - 2 Full Professors;
  - 2 Associate Professors;
  - 2 Assistant Professors
  - 3 Research Fellows;
  - 3 Postdocs;
  - 16 PhD Students;
  - Students and consultants;
- Collaborations with local industries;
- Research interests:
  - First Person (Egocentric) Vision;
  - Multimedia Security and Forensics;
  - Cultural Heritage;
  - Social Media Mining;

Antonino Furnari

Francesco Ragusa

Before we begin…

The slides of this tutorial are available online at:

http://www.antoninofurnari.it/talks/iciap2022

Agenda

ANTONINO

Part I: Definitions, motivations, history and research trends [14.00 - 15.45]

- **What is first person vision? What is it for?**
- **What makes it different from third person vision?**
- **History of First Person Vision: visions, ideas, research, devices;**
- **Where do we go from here? Research trends, datasets and challenges.**

FRANCESCO

Part II: Building Blocks for First Person Vision Systems [16.15 – 18.00]

- Data Acquisition & Datasets;
- Fundamental Taks in First Person Vision:
  - Localization;
  - Hand/Object Detection;
  - Attention;
  - Action/Activities;
  - Anticipation
- Conclusion

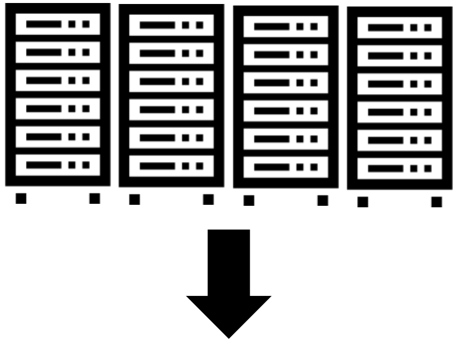
Part I

Definitions, Motivations, History, Research Trends and Applications

The Revolution of Personal Computing

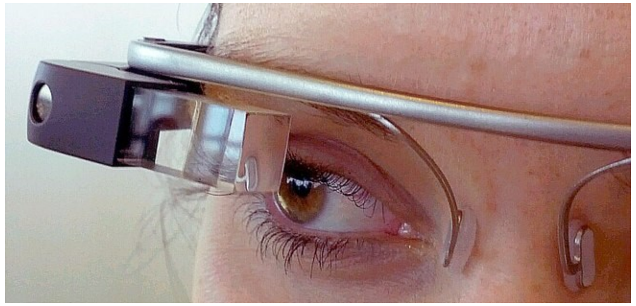
After personal computers and smartphones, mixed reality is the third wave of computing

– 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

What Shall We Expect from Wearable Computers?

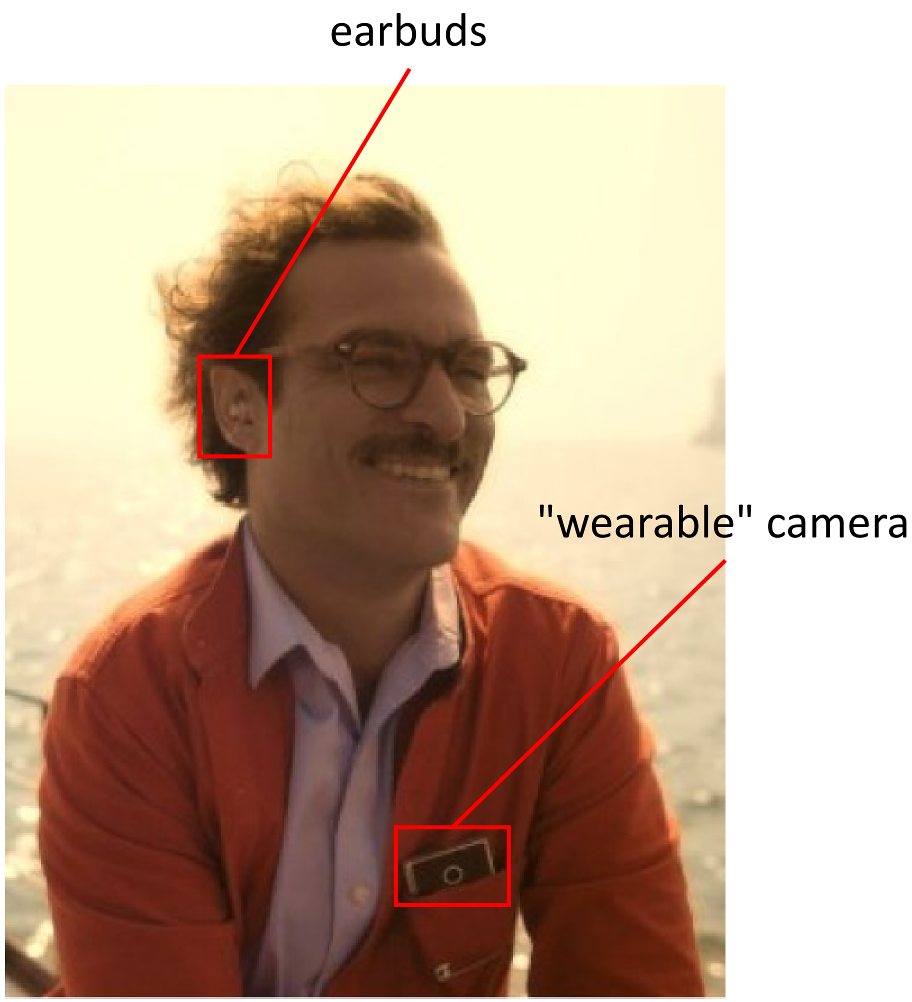
Wearable Computers define a world in which the user is central and access to computation is simplified and **blended** in the real world

Wearable Computers are the perfect interface to build **personal assistants** capable of automating computation to augment our abilities.

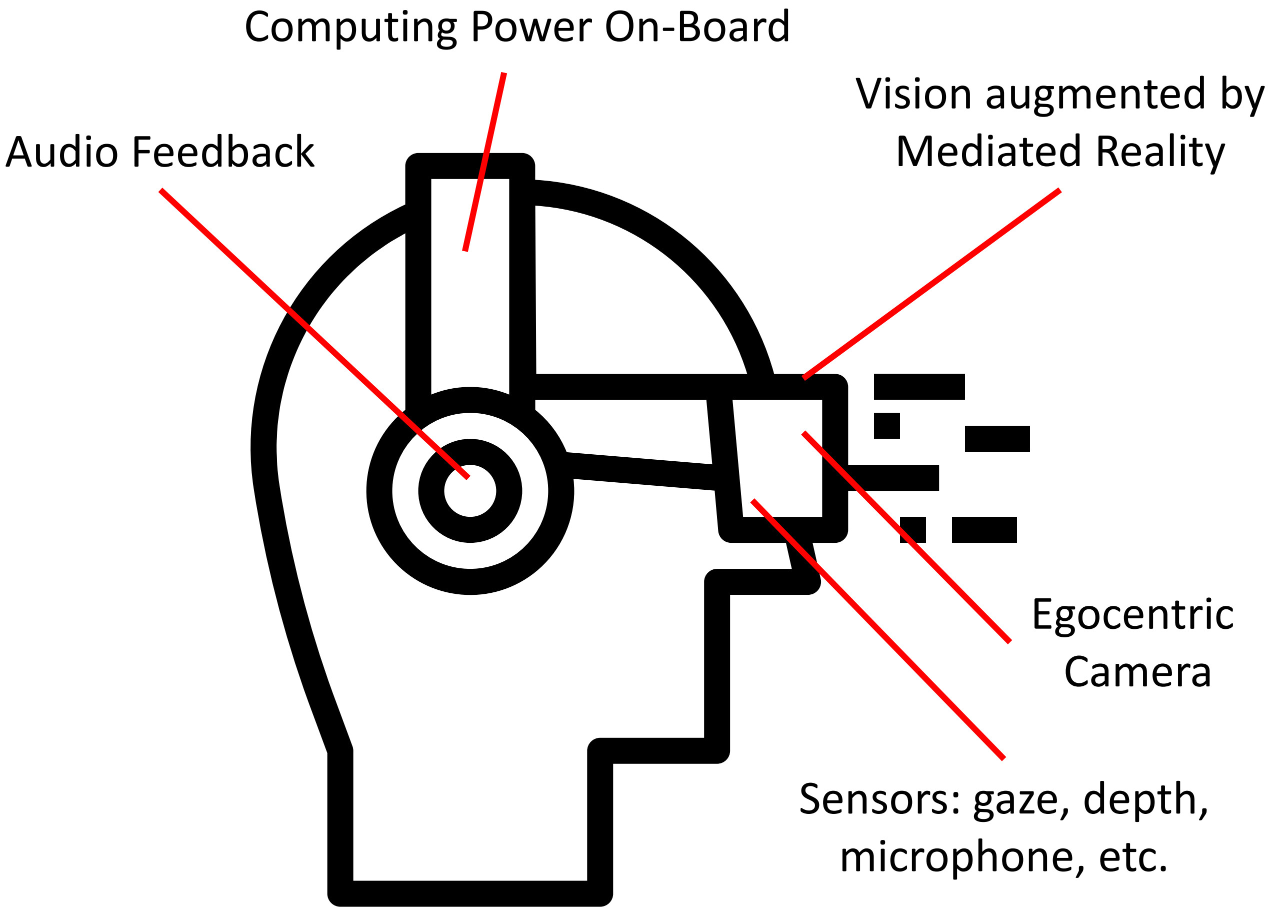
Hence, they need to **understand** where we are, how the physical world around us is made, and what are our objectives.

Vision is fundamental!

An AI-Powered Virtual Assistant



"her" 2013 movie

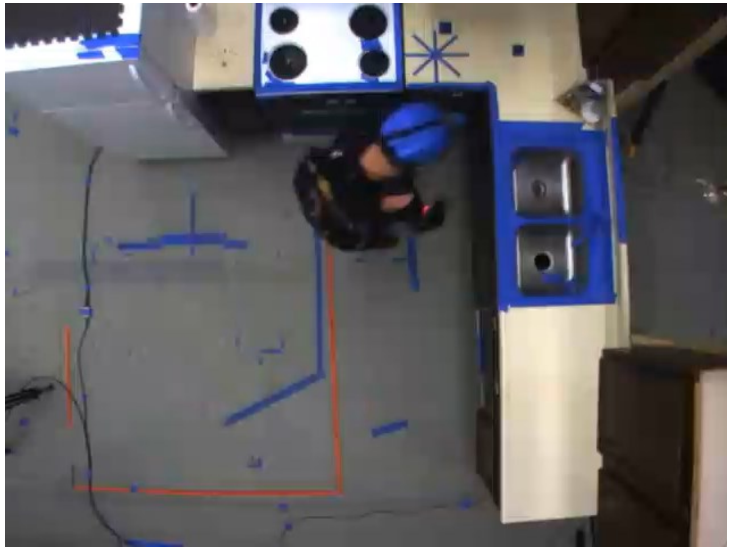


https://thenounproject.com/Turkkub/

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

Can’t we just apply standard CV?

Third Person Camera

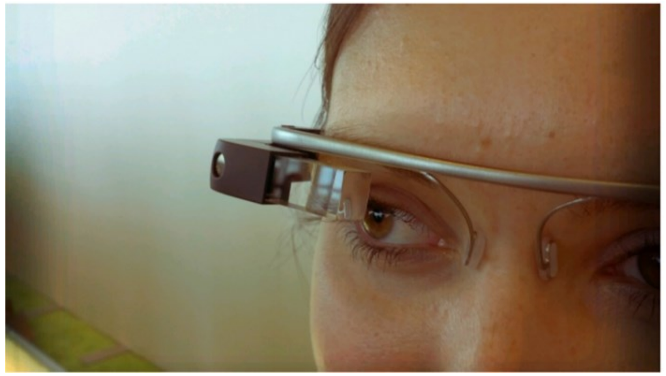


First Person Camera

Fixed Camera

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

Wearable Camera



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

Features of First Person Vision

- **Sees «what the user sees»**
  - The acquired video always «tells something» about the user;
    - Behavior understanding, Embodied perception;
- **Naturally mobile**
  - FPV can be used to build intelligent systems able to assist the user and augment their abilities;
    - Third wave of personal computing;
- **Exposed to huge amounts of personal data**
  - FPV can be used to build AIs which learn from personal data;
    - Can learn to predict the user's goal.

Virtual Personal Assistant

“All about the user”

“First Person Vision, which senses the environment and the subject’s activities from his/her view point, is advantageous for understanding the behavior, intent, and environment of a person.”

Takeo Kanade and Martial Herbert. "First-person vision." *Proceedings of the IEEE* 100.8 (2012): 2442-2453.

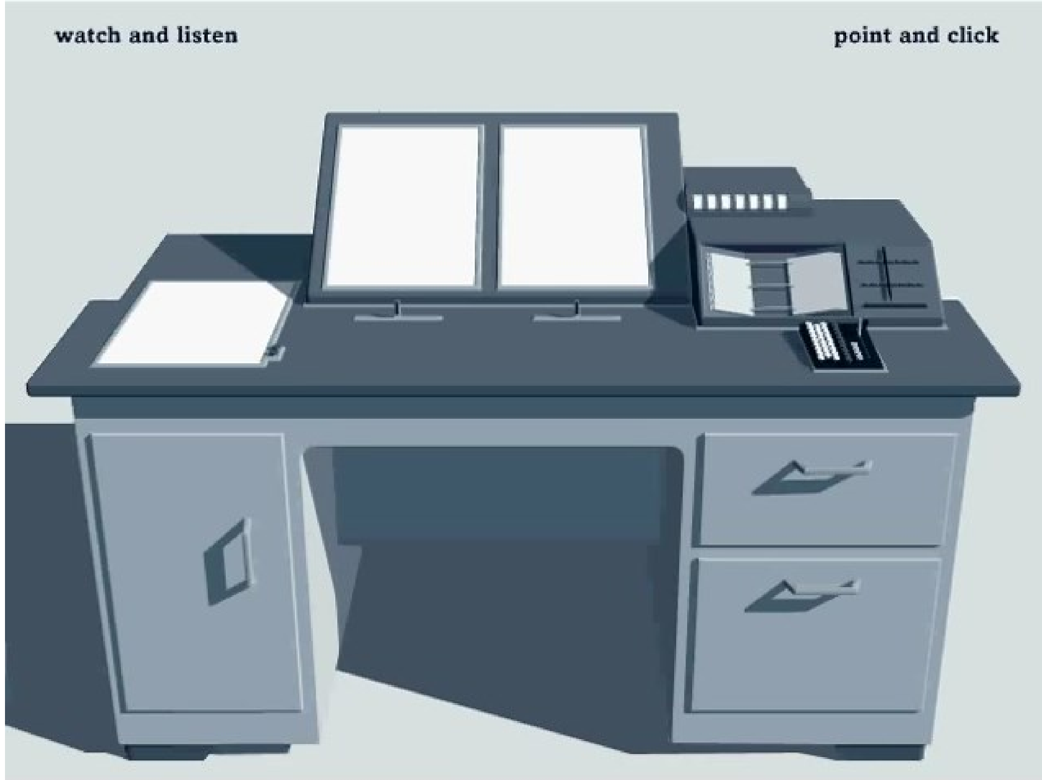
Note on Terminology

Different terms have been used to refer to very similar concepts. The most common ones are as follows:

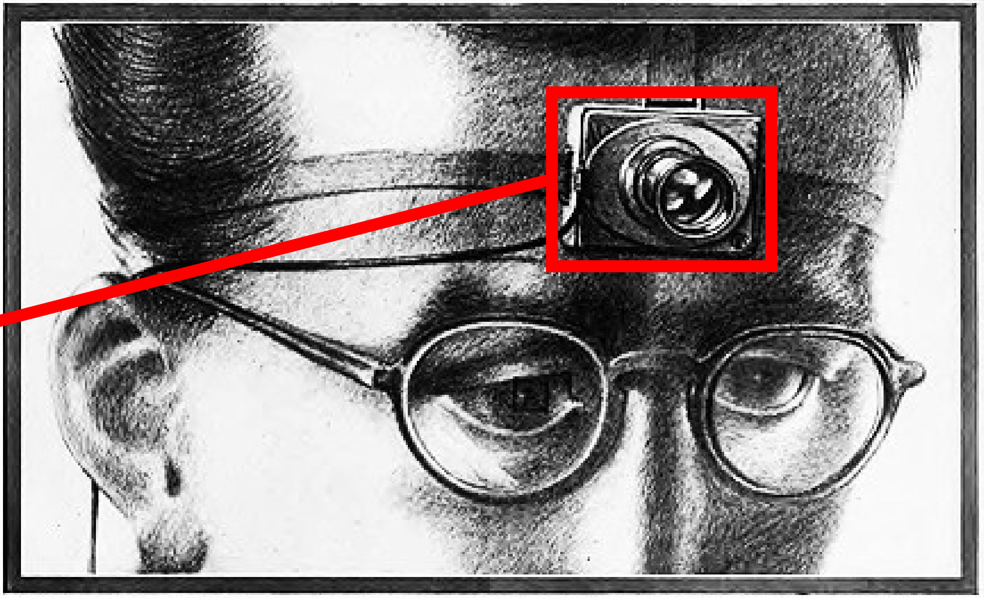
Term	Etymology
First Person Vision (FPV)	Computer Vision for images and videos acquired from a First Person View, as opposed to the classic Third Person View
Egocentric Vision (Ego-Vision)	Computer Vision for visual data «about me» (from Greek/Latin «ego»=«I»)
Wearable Vision	Computer Vision for wearable devices

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

The Birth of Wearable Computing

http://wearcam.org

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.

http://wearcam.org/previous_experiences/

WearCam, Steve Mann, 1994-1996

- In 1994 Steve Mann invented the first wearable camera;
- WearCam streamed images to Mann’s personal page from 1994 to 1996;

Steve Mann, "Wearable computing: a first step toward personal imaging," in *Computer*, vol. 30, no. 2, pp. 25-32, Feb. 1997.

Cybersquare

Steve Mann
MIT Media Lab

Wearable Computing: A First Step Toward Personal Imaging

Miniaturization of components has enabled systems that are wearable and nearly invisible, so that individuals can move about and interact freely, supported by their personal information domain.

Wearable Computer Vision: The Goal

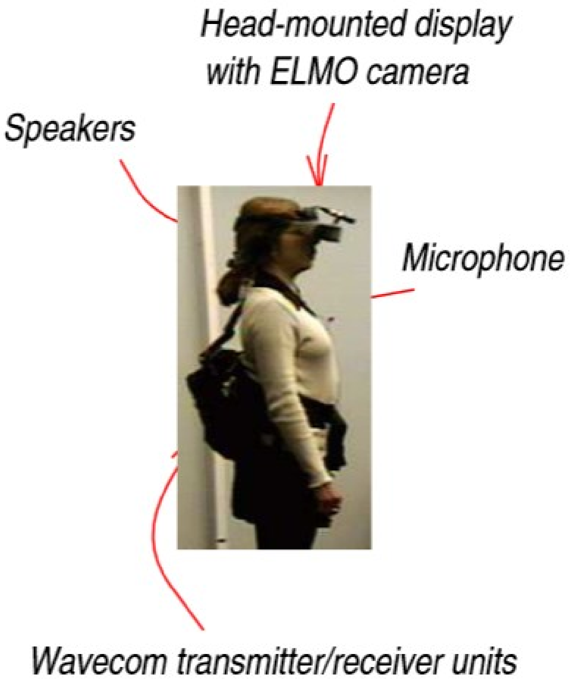
Clip from movie Termintor 2-Judgment day: https://youtu.be/9MeaaCwBW28

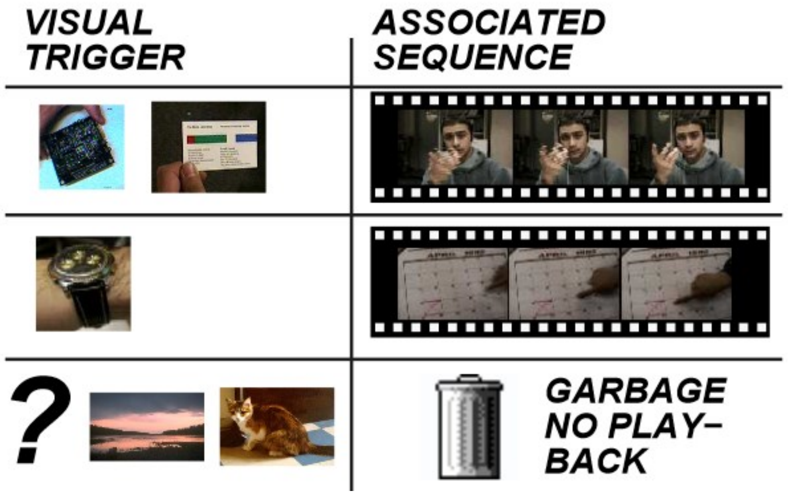
Ref: https://www.redsharknews.com/vr_and_ar/item/3539-terminator-2-vision-the-augmented-reality-standard-for-25-years

MIT Media Lab in 1997

DyPERS, 1998

www.nuriaoliver.com/dypers/

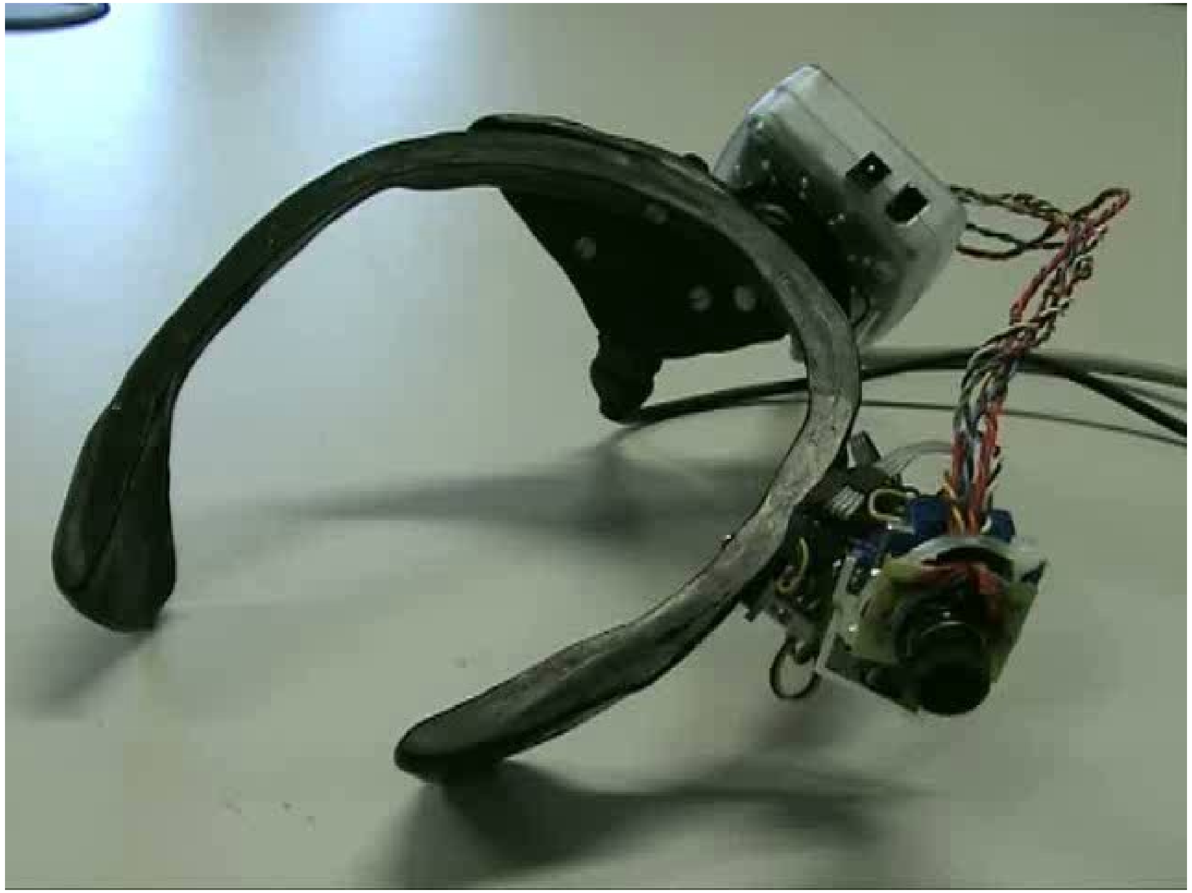


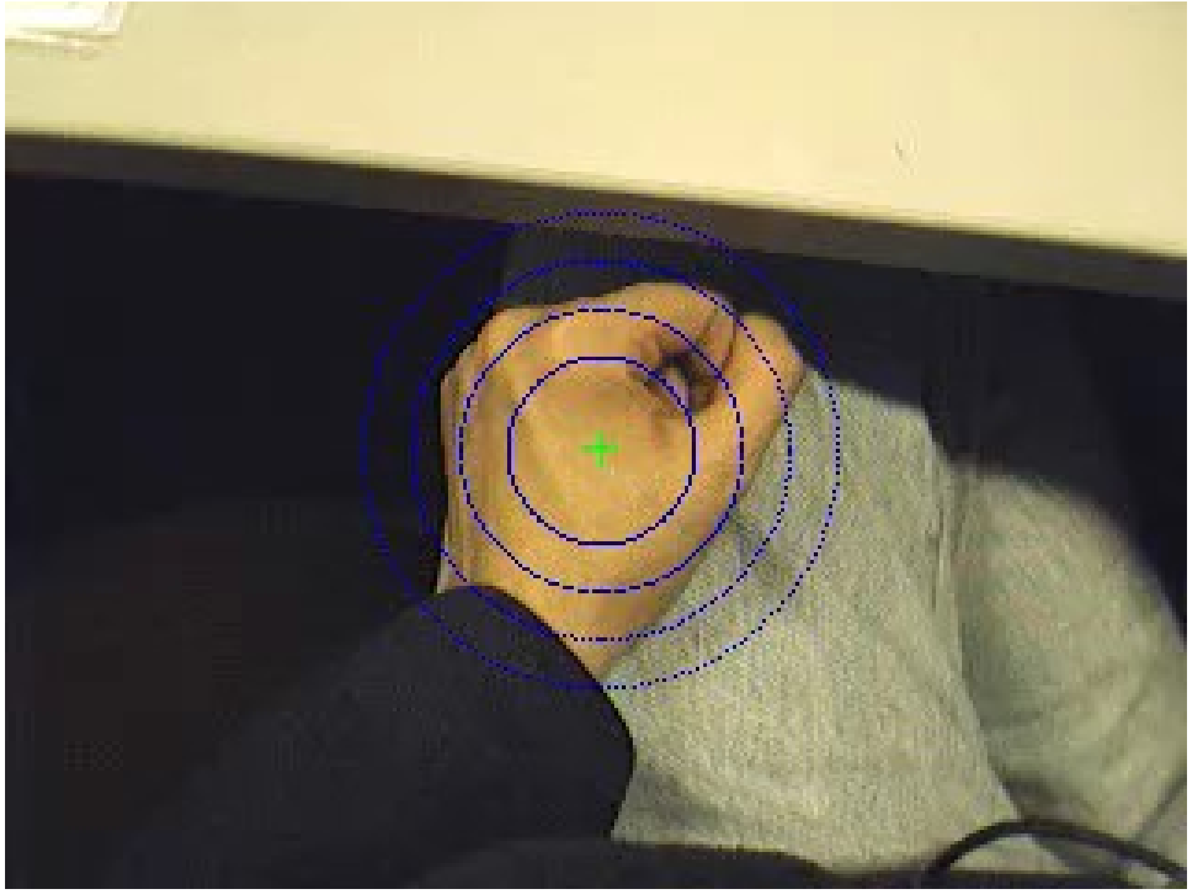


Jebara, T., Schiele, B., Oliver, N., & Pentland, A. (1998). DyPERS: Dynamic personal enhanced reality system. In *In Proc. 1998 Image Understanding Workshop*.

Wearable Visual Robots, 2002 - 2004

http://people.cs.bris.ac.uk/~wmayol/research/





W.W. Mayol, B. Tordoff and D.W. Murray. Wearable Visual Robots. Selected papers from ISWC00, Personal And Ubiquitous Computing Journal. Springer-Verlag. Volume 6 pp37-48. 2002.
W.W. Mayol, A.J. Davison, B.J. Tordoff, N.D. Molton, and D.W. Murray. Interaction between hand and wearable camera in 2D and 3D environments. Proc. British Machine Vision Conference 2004.

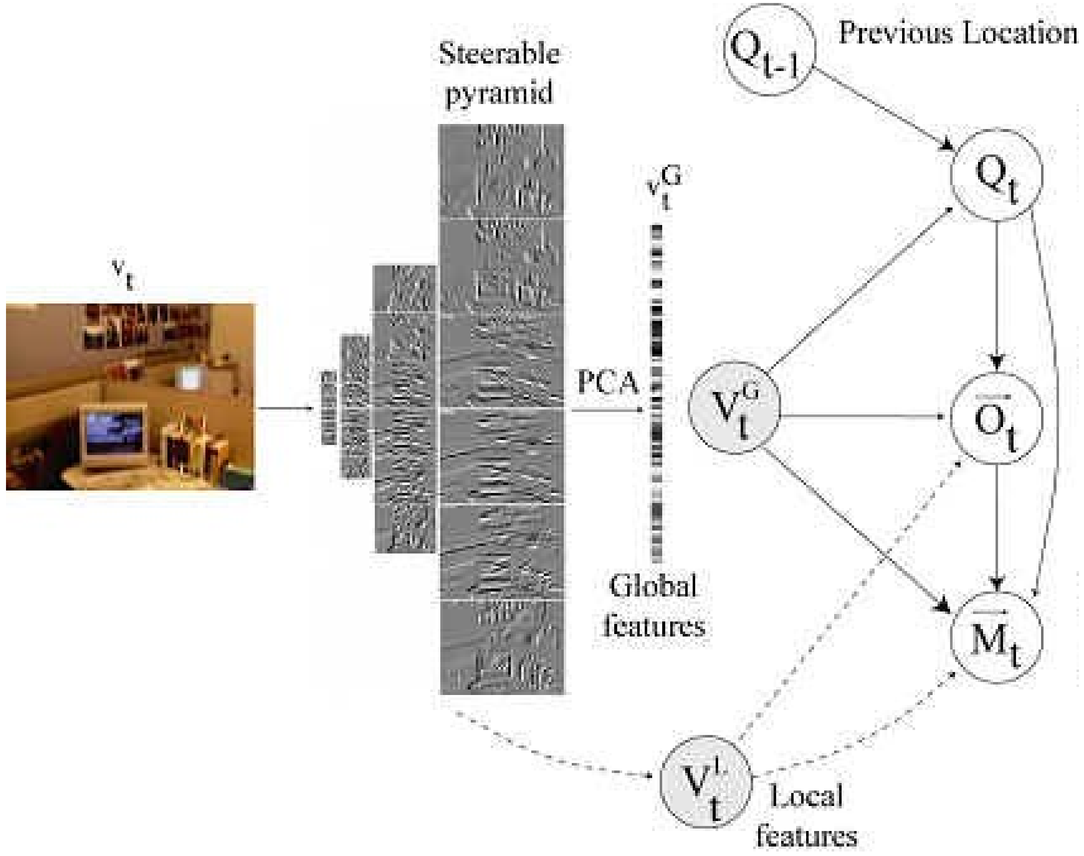
SLAM and Augmented Reality, 2004-2008

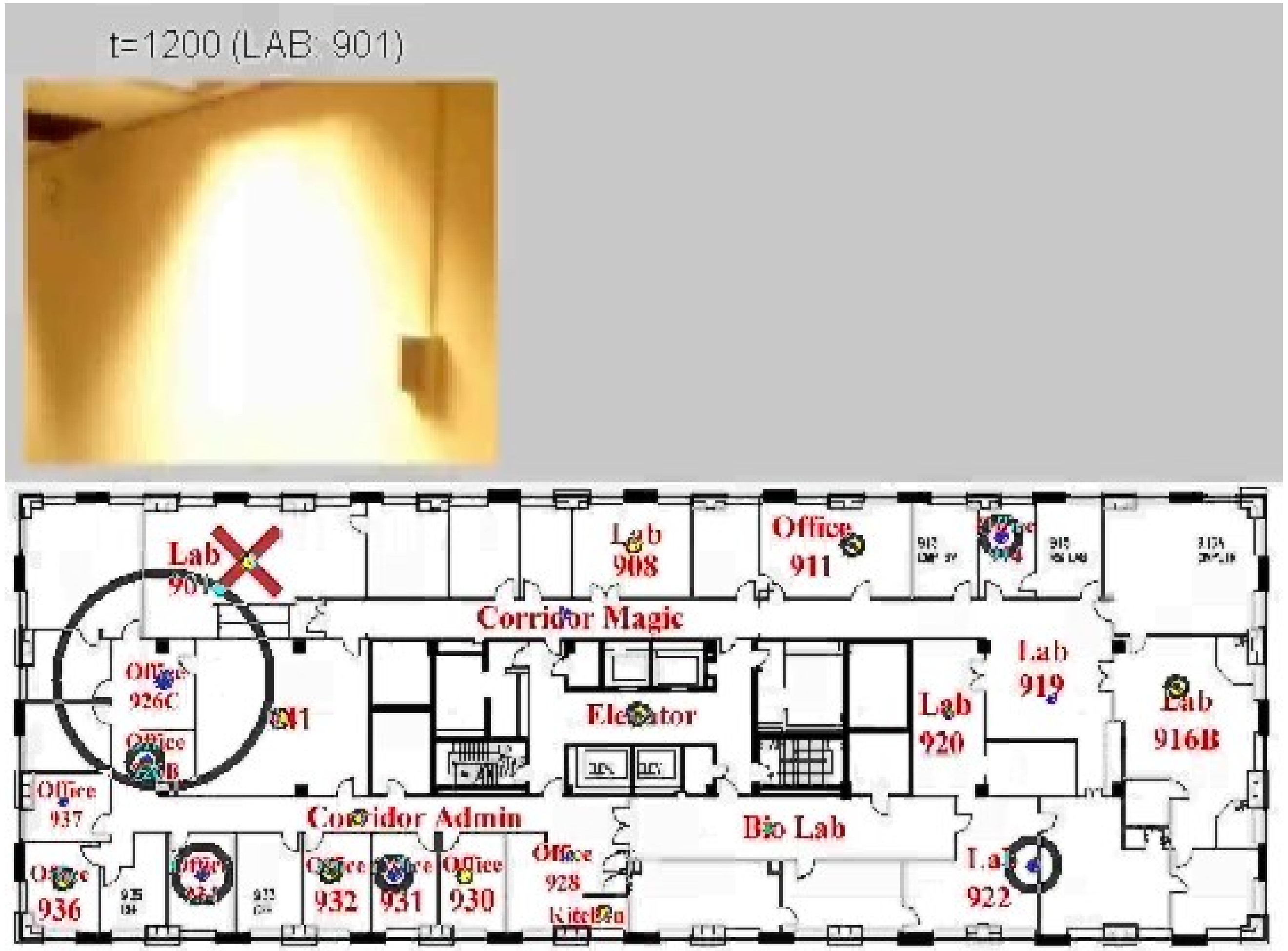
http://people.cs.bris.ac.uk/~wmayol/research/

- W.W. Mayol, A.J. Davison, B.J. Tordoff, N.D. Molton, and D.W. Murray. Interaction between hand and wearable camera in 2D and 3D environments. Proc. British Machine Vision Conference 2004. London, UK, September. 2004.
- Pished Bunnun, Walterio Mayol-Cuevas, OutlinAR: an assisted interactive model building system with reduced computational effort. 7th IEEE and ACM International Symposium on Mixed and Augmented Reality. September 2008.

Place and scene recognition from FPV, 2003

https://www.cs.ubc.ca/~murphyk/Vision/placeRecognition.html



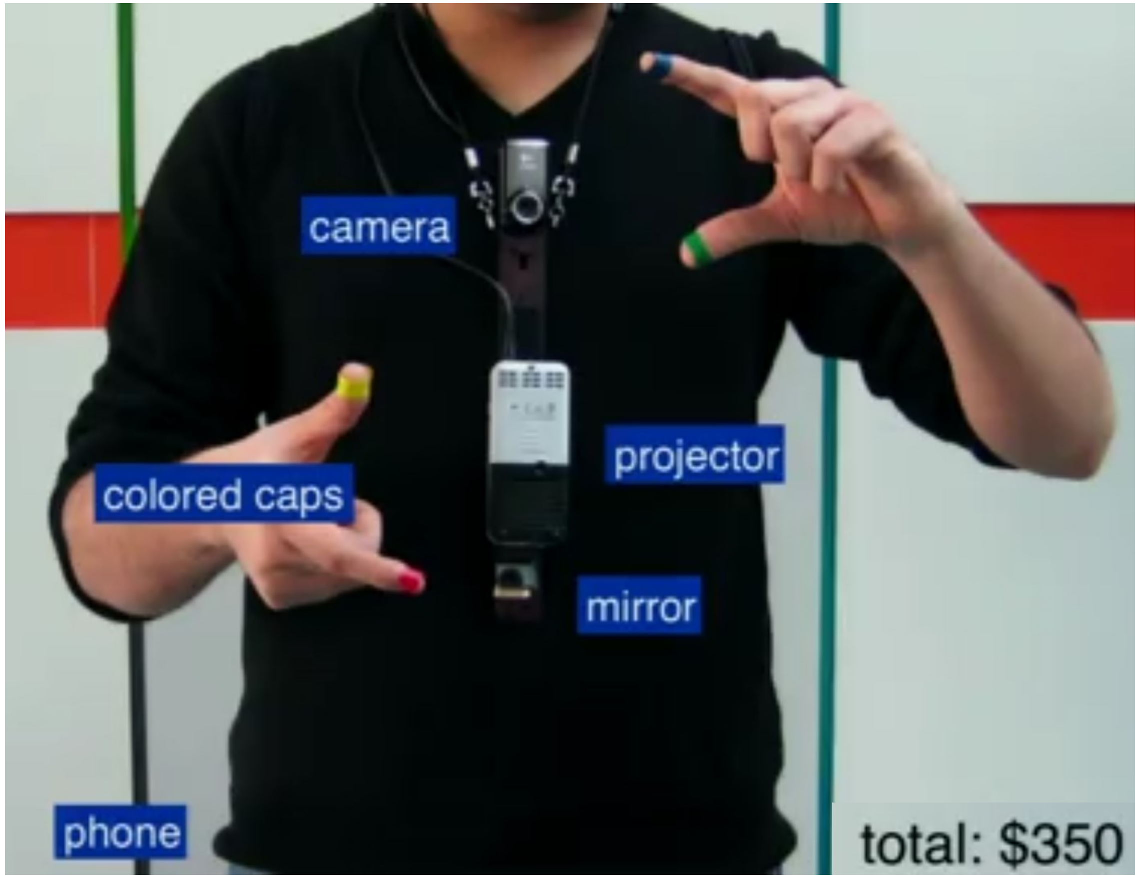


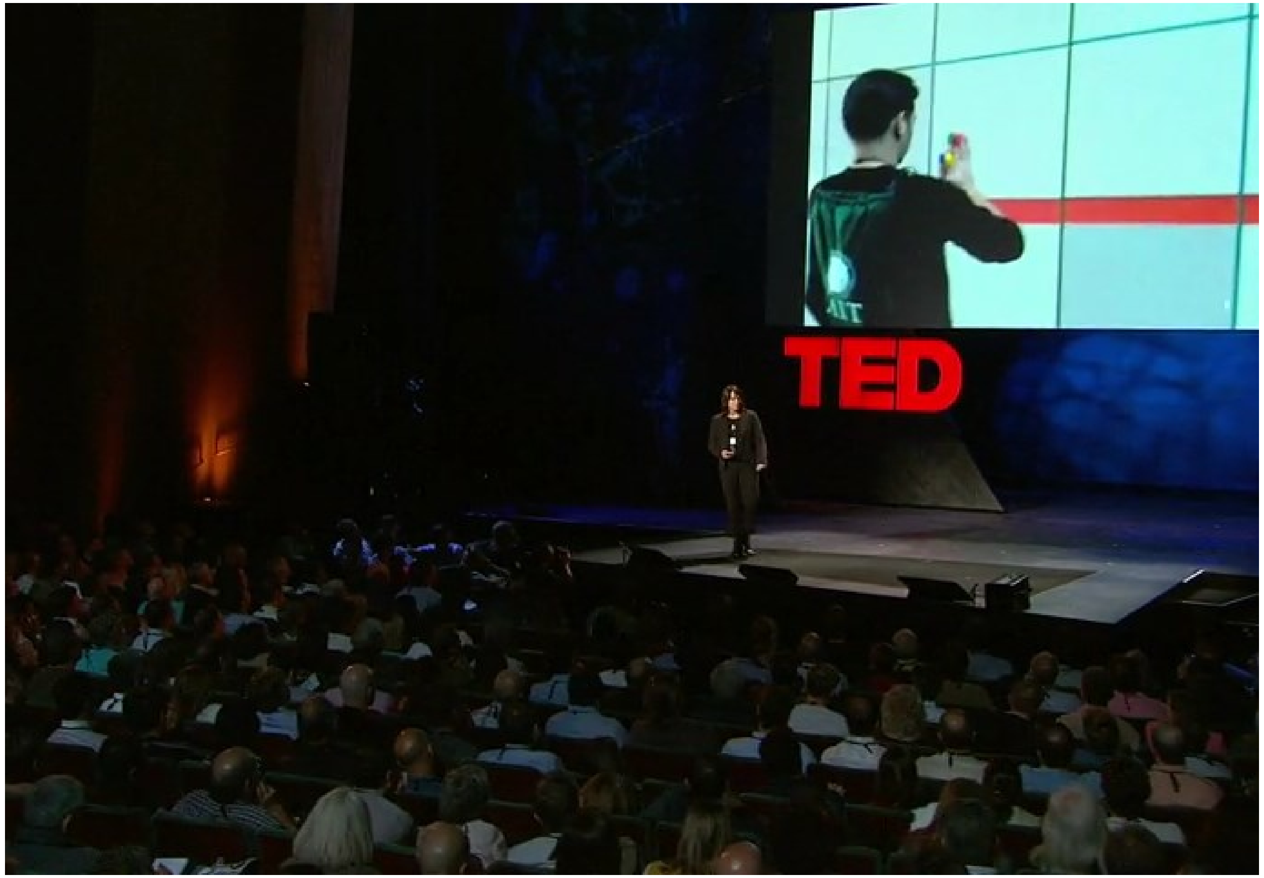
Torralba, A., Murphy, K. P., Freeman, W. T., & Rubin, M. A. (2003). Context-based vision system for place and object recognition. In *ICCV 2003*

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»

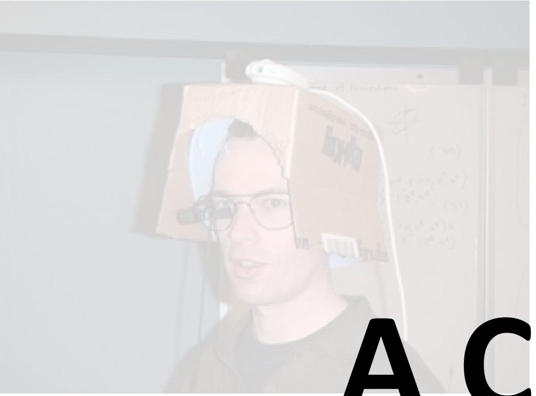




Pattie Maes & Pranav Mistry (MIT) @ TED https://www.ted.com/talks/pattie_maes_demos_the_sixth_sense

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/

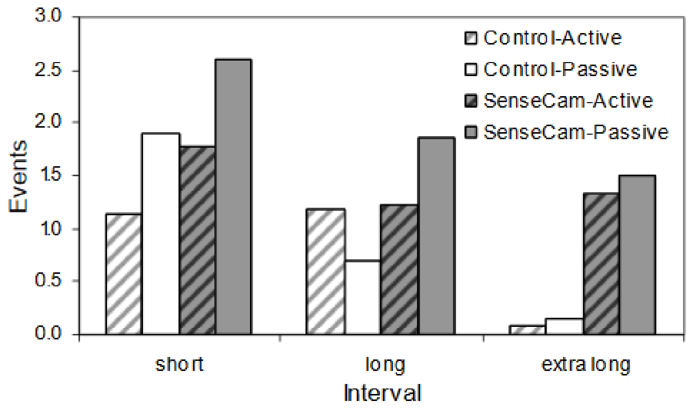
Bell, Gordon, and Jim Gemmell. *Your life, uploaded: The digital way to better memory, health, and productivity*. Penguin, 2010.

Research using Microsoft SenseCam

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

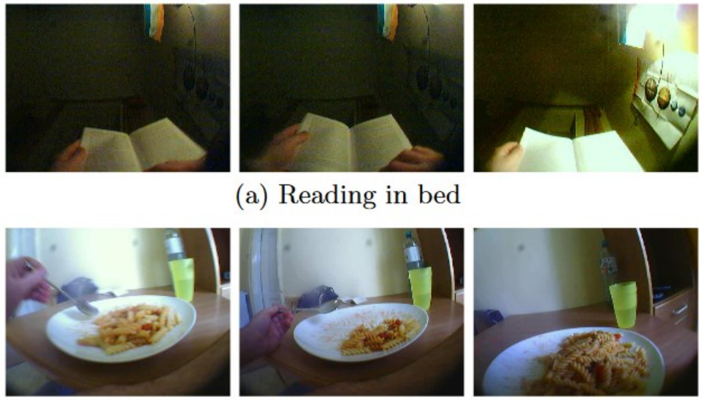
Abigail Sellen, Andrew Fogg, Mike Aitken*, Steve Hodges, Carsten Rother and Ken Wood
Microsoft Research Cambridge
7 JJ Thomson Ave, Cambridge, UK, CB3 0FB
*Behavioural & Clinical Neuroscience Institute
Dept. of Psychology, University of Cambridge

(health, memory augmentation)



2007

2008



(a) Reading in bed

(b) Having dinner

MyPlaces: Detecting Important Settings in a Visual Diary

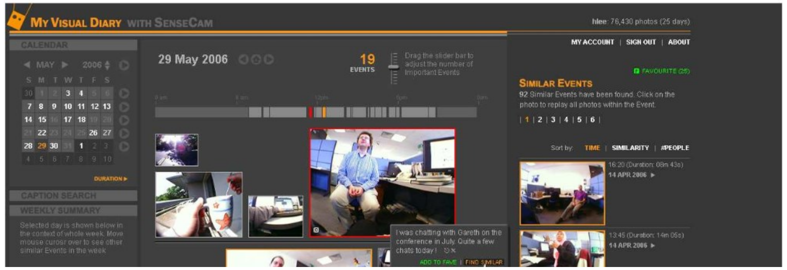
Michael Blighe and Noel E. O'Connor
Centre for Digital Video Processing, Adaptive Information Cluster
Dublin City University, Ireland
{blighem, oconnorn}@eeng.dcu.ie

(lifelogging, place recognition)

Constructing a SenseCam Visual Diary as a Media Process

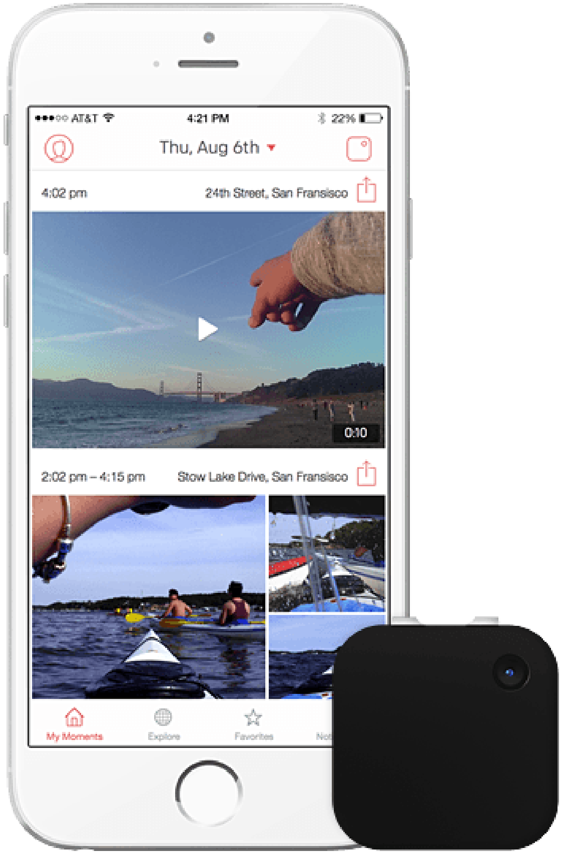
Hyowon Lee, Alan F. Smeaton, Noel O'Connor, Gareth Jones, Michael Blighe, Daragh Byrne, Aiden Doherty, and Cathal Gurrin
Centre for Digital Video Processing & Adaptive Information Cluster,
Dublin City University

(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)

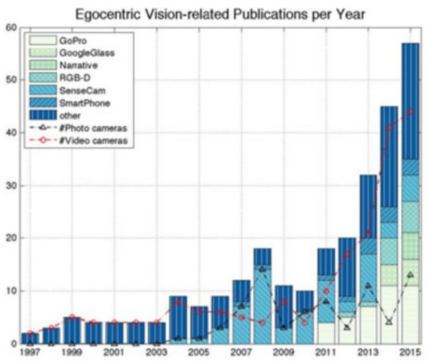
2016

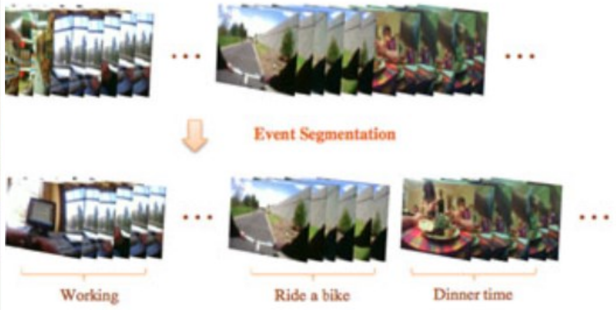
2017

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

Looxcie, 2010


«mobile, connected, hands free, streaming video camera»

(unsupervised action recognition, video indexing)

Unsupervised Ego-Action Learning, 2011

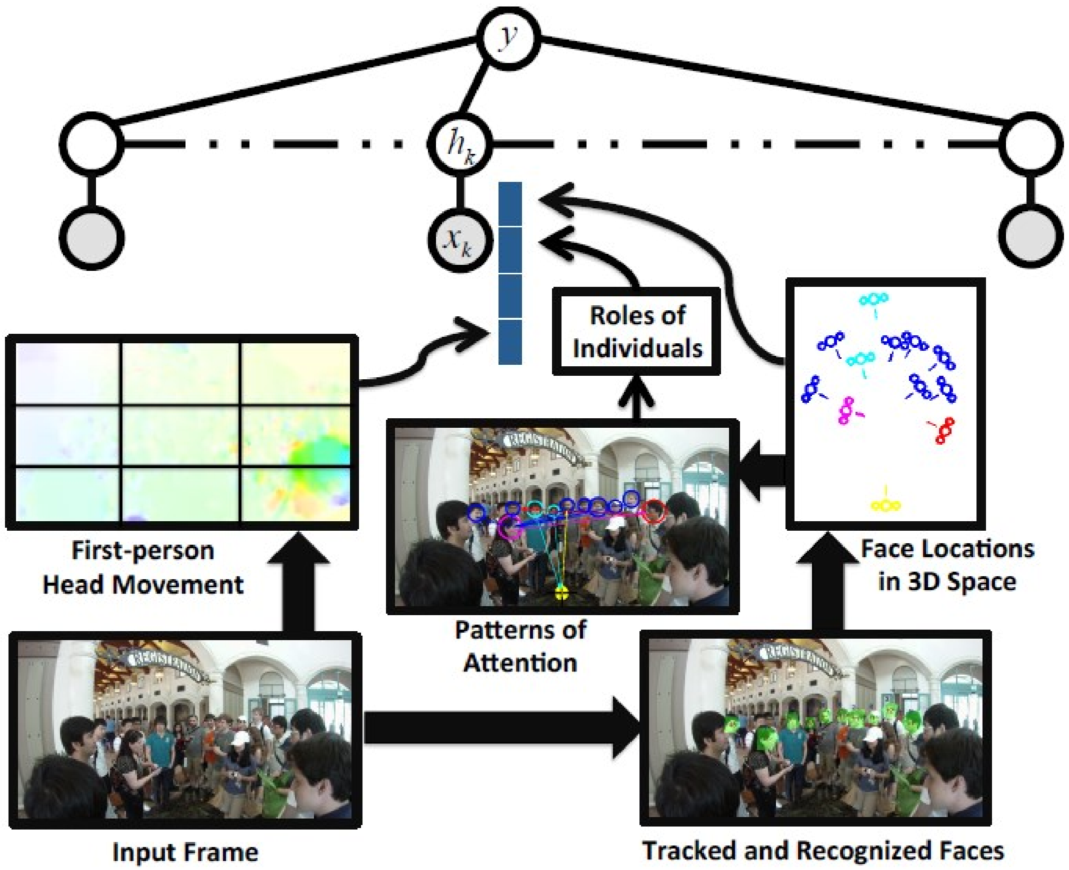
https://www.youtube.com/watch?v=12CZu4Xlb_U

Kitani, K. M., Okabe, T., Sato, Y., & Sugimoto, A. (2011, June). Fast unsupervised ego-action learning for first-person sports videos. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* (pp. 3241-3248). IEEE.

(detection and recognition of social interactions)

Social Interaction Recognition, 2012

https://player.vimeo.com/video/37507972

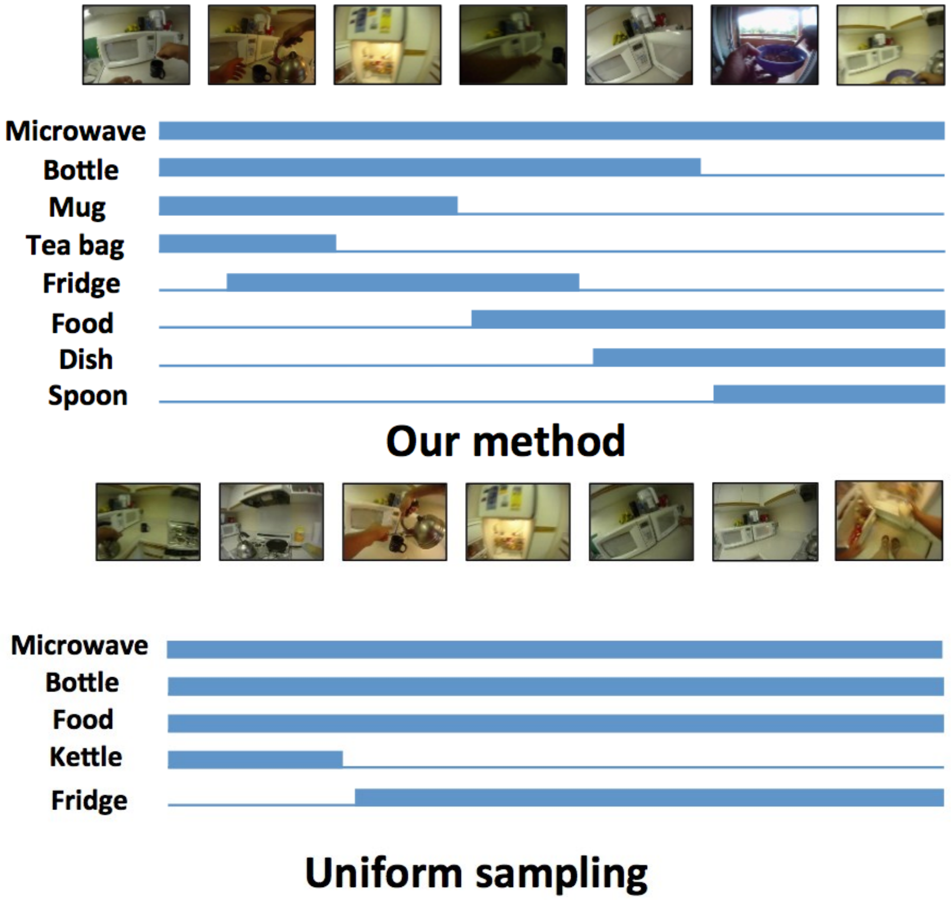


Fathi, A., Hodgins, J. K., & Rehg, J. M. (2012, June). Social interactions: A first-person perspective. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (pp. 1226-1233). IEEE.

(egocentric video sumarization)

Egocentric Video Summarization, 2013

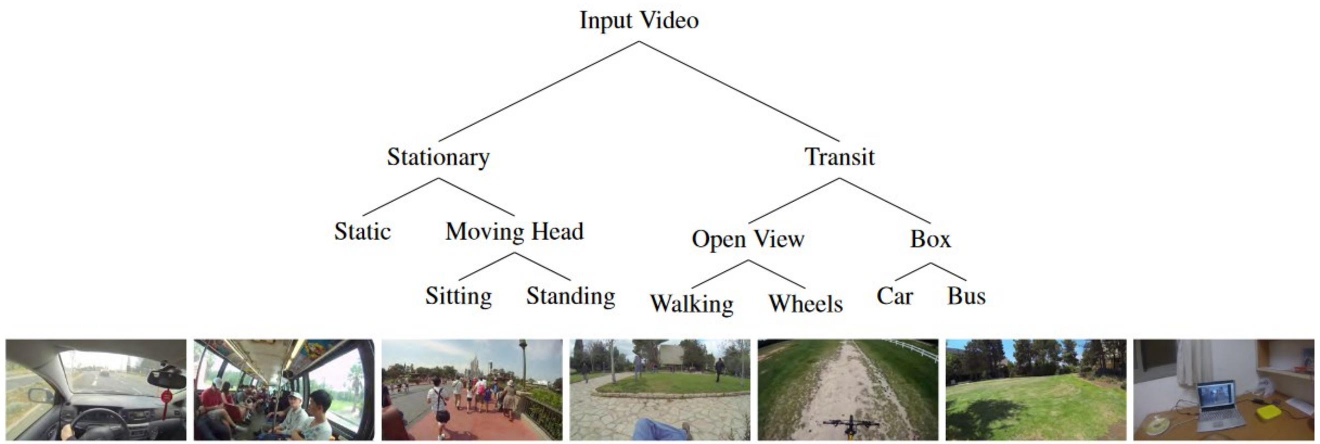
http://vision.cs.utexas.edu/projects/egocentric/storydriven.html



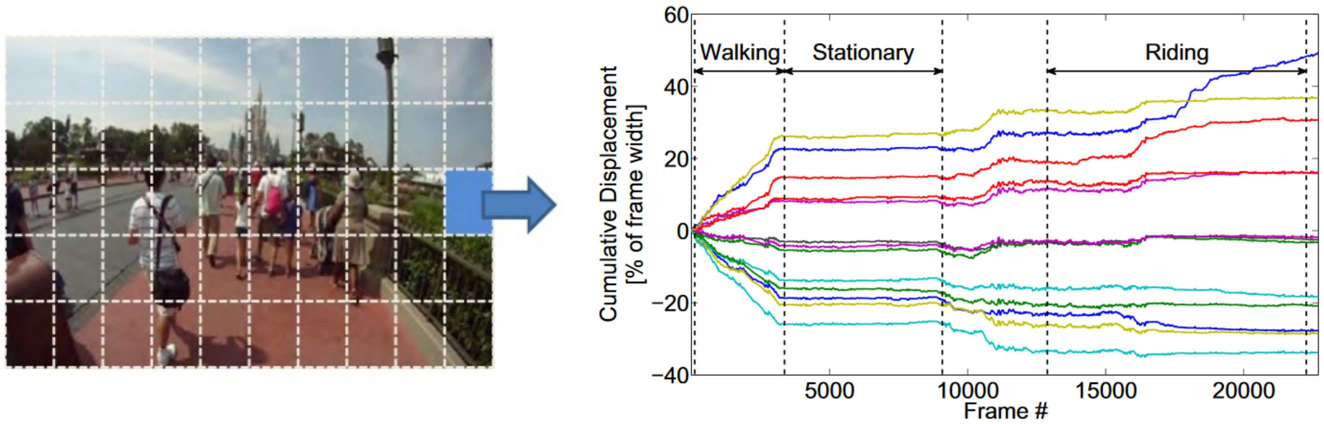
- Story-Driven Summarization for Egocentric Video. Zheng Lu and Kristen Grauman. Computer Vision and Pattern Recognition (CVPR), 2013
- Discovering Important People and Objects for Egocentric Video Summarization. Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. CVPR 2012

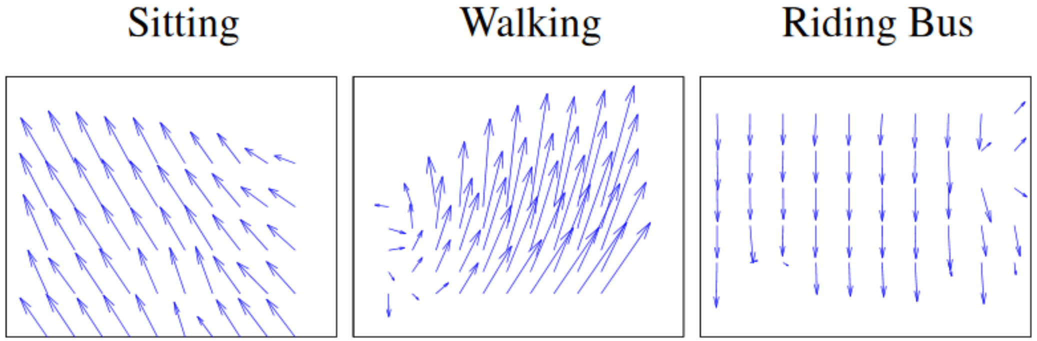
(egocentric video indexing)

Temporal Segmentation of Egocentric Video, 2014

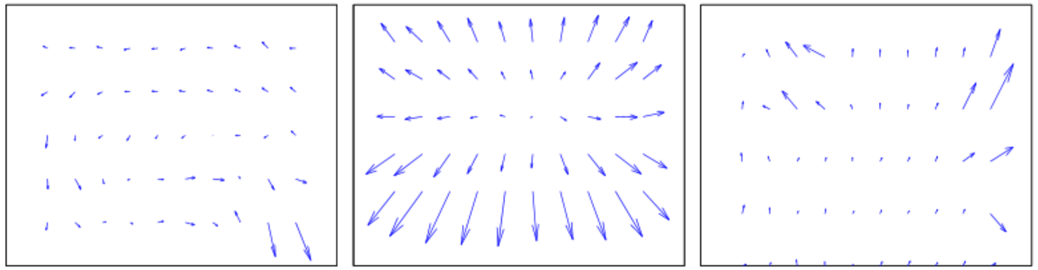


(a) Car (b) Bus (c) Walking (d) Sitting (e) Wheels (f) Standing (g) Static





(a) Instantaneous (x, y) displacement vectors are dominated by the head rotation, and the effects of the activity, e.g. sitting, walking, or riding, is too small to observe.



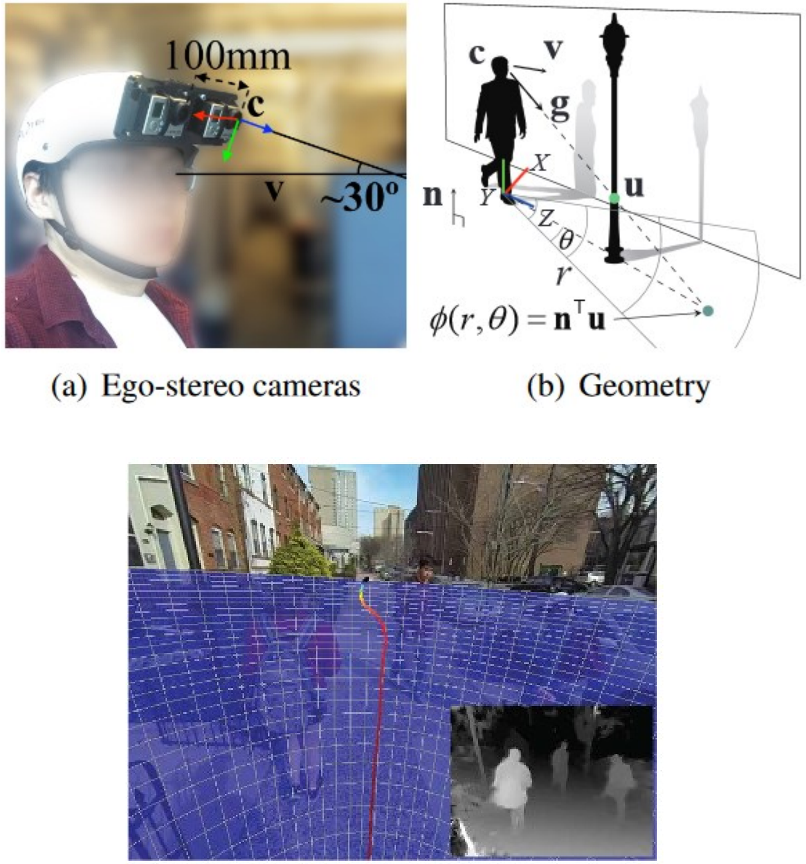
(b) Motion vectors obtained from the cumulative displacement curves as given in Eq. 3. Effects of head rotations are removed, and the direction of vectors are now noiseless. For 'walking' the vectors are large and have radial direction. In the 'sitting' case, the magnitude mostly zero. Riding ('car'/'bus') has a mixed pattern.

Poleg, Y., Arora, C., & Peleg, S. (2014). Temporal segmentation of egocentric videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2537-2544).

(future localization, navigation)

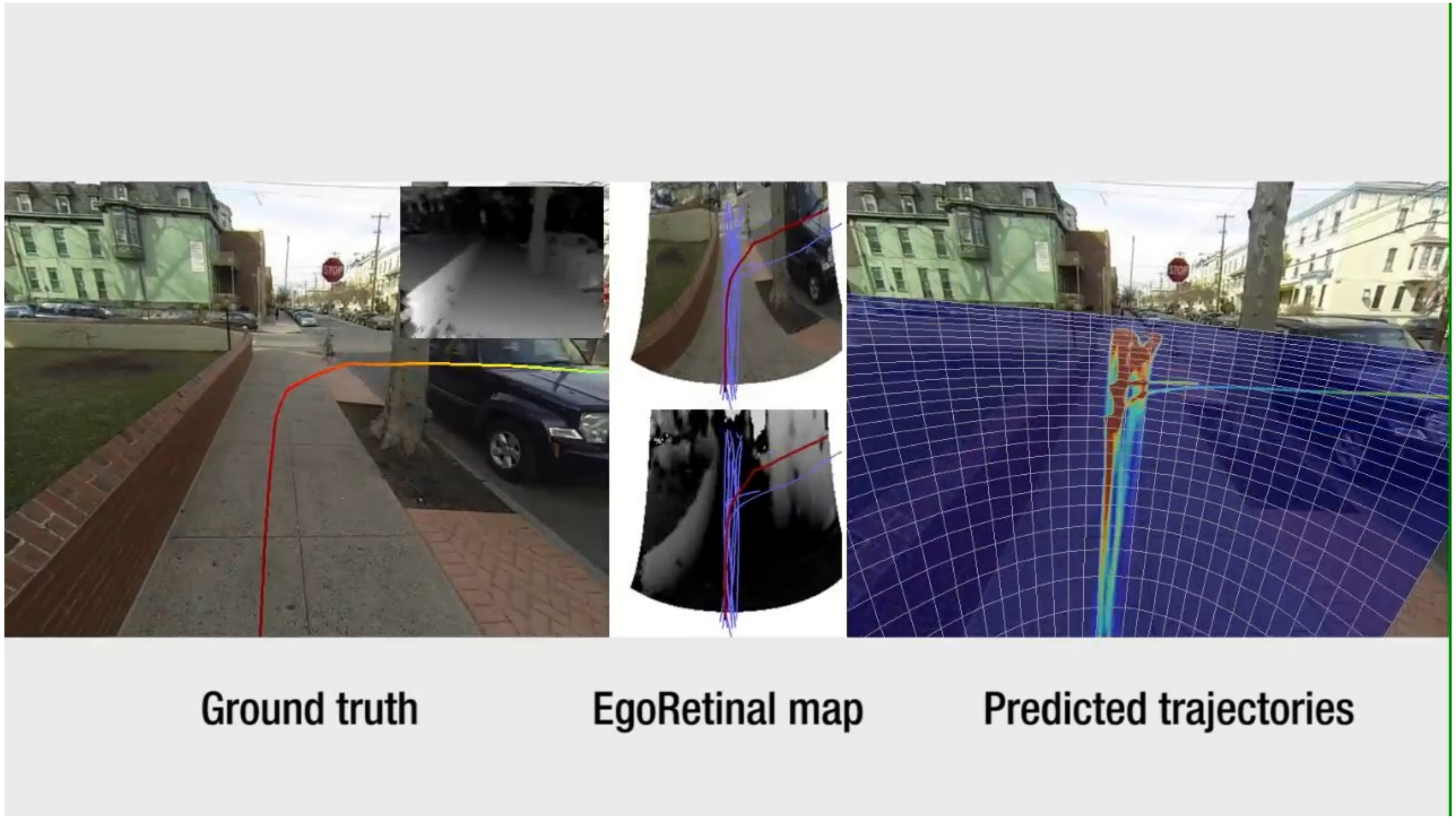
Future Localization in Egocentric Video, 2016

https://www-users.cs.umn.edu/~hspark/future_loc.html



(a) Ego-stereo cameras (b) Geometry

(c) Egocentric RGBD image

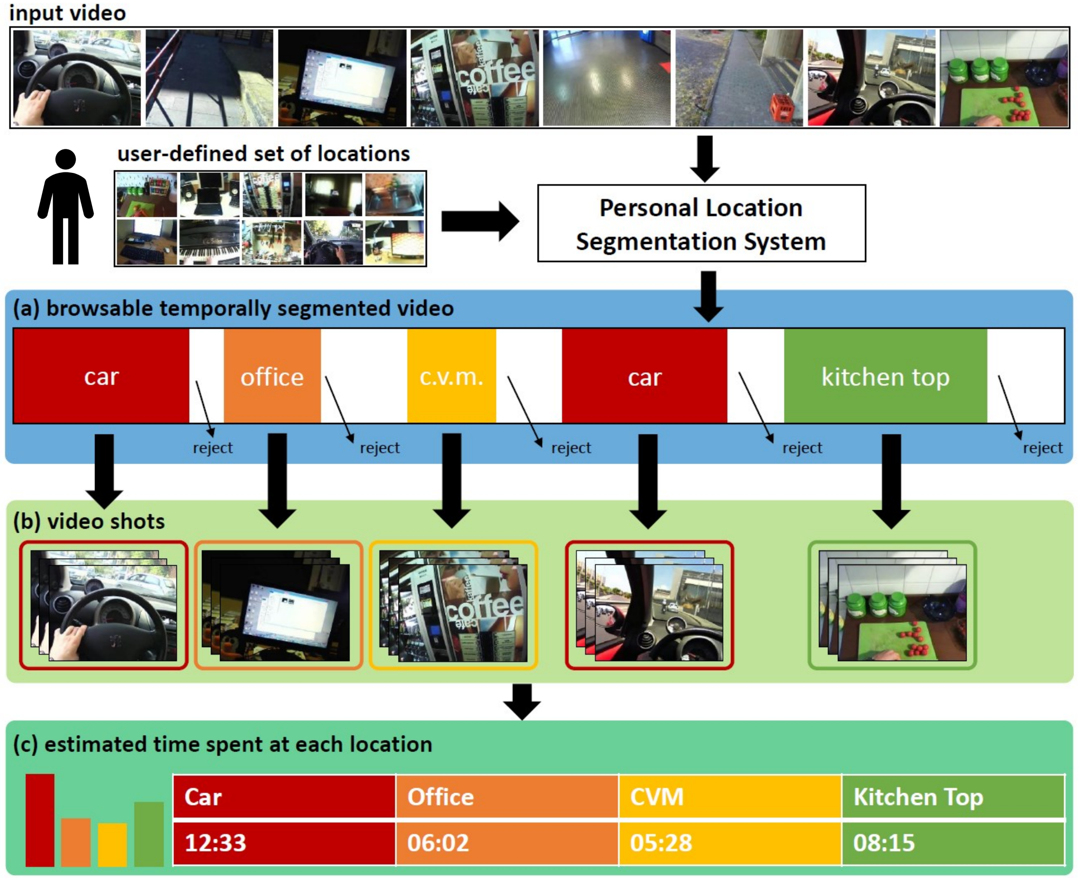


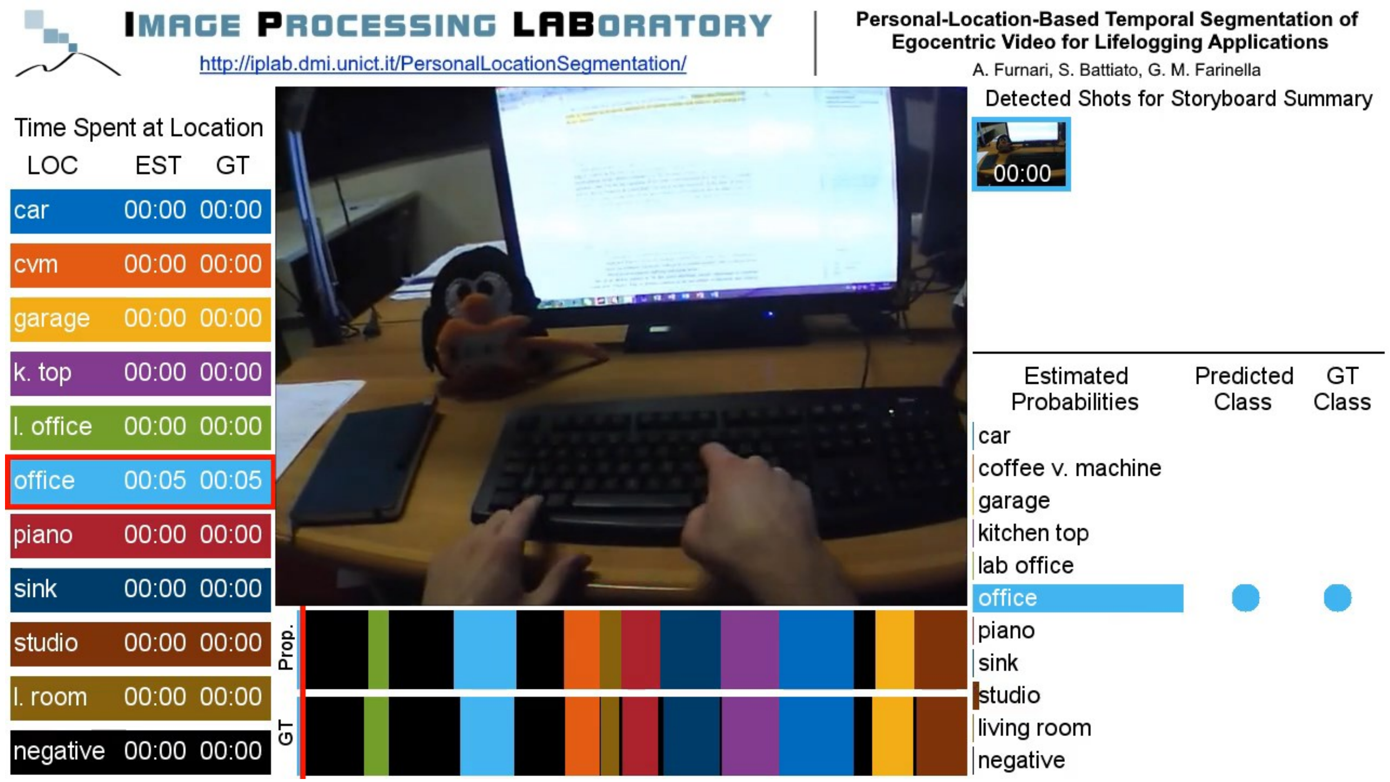
Hyun Soo Park, Jyh-Jing Hwang, Yedong Niu, and Jianbo Shi "Egocentric Future Localization" Conference on Computer Vision and Pattern Recognition (CVPR), 2016,

(localization, indexing, context-aware computing)

Egocentric Location Recognition, 2018

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

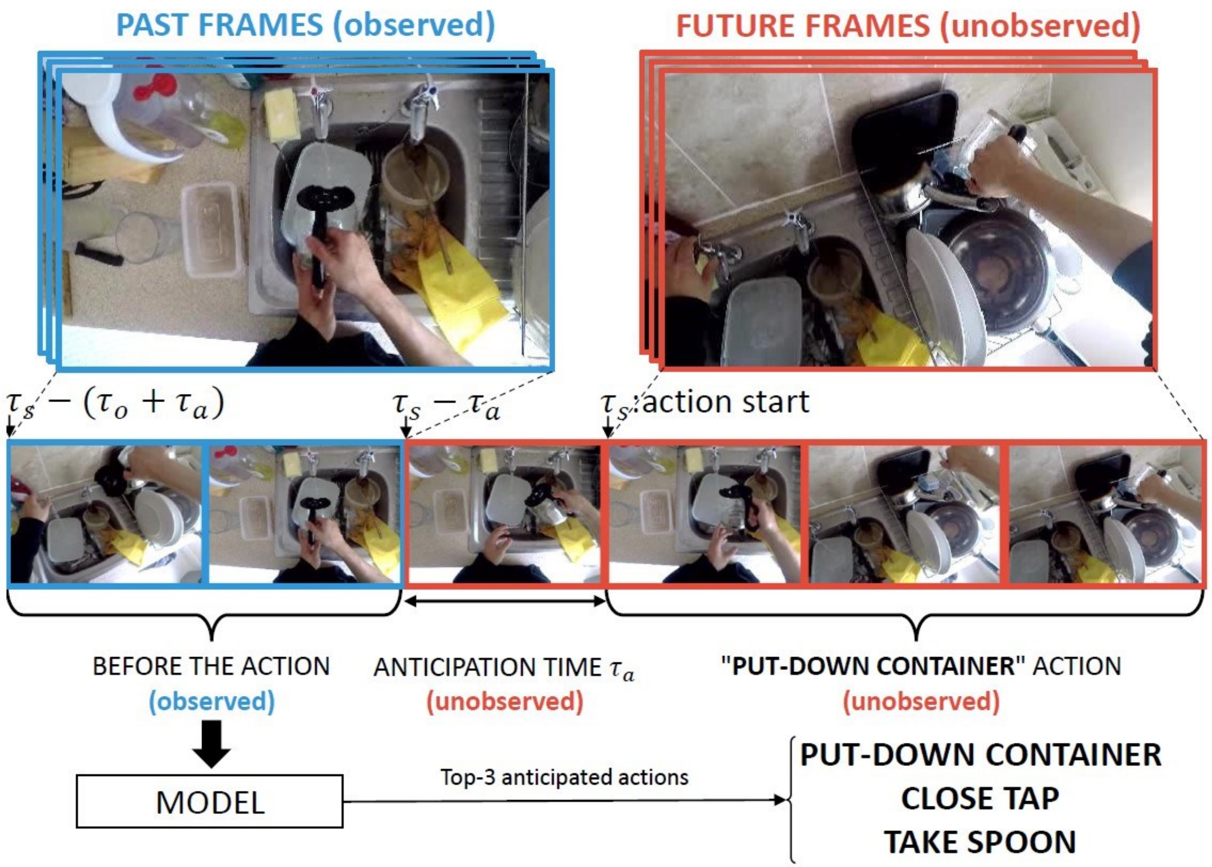


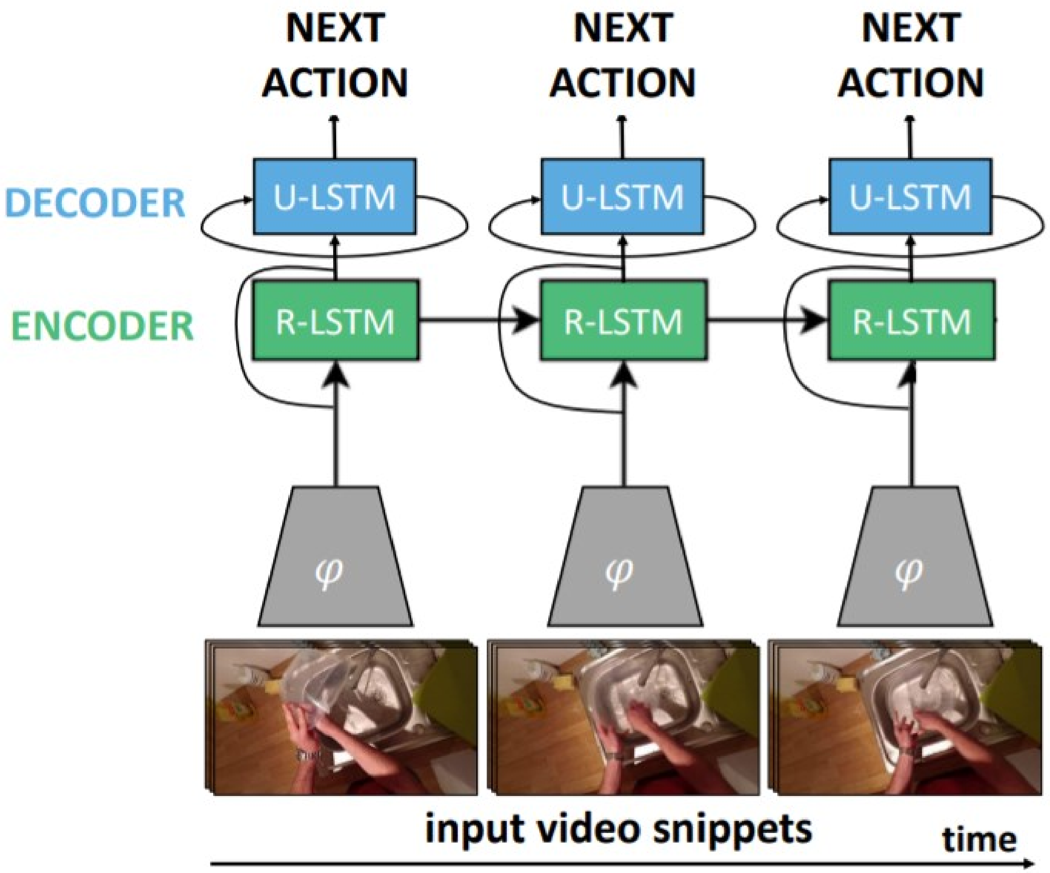


- A. Furnari, G. M. Farinella, S. Battiato, Recognition of Personal Locations from Egocentric Videos, IEEE Transactions on Human-Machine Systems, 2016.
- A. Furnari, S. Battiato, G. M. Farinella, Personal-Location-Based Temporal Segmentation of Egocentric Video for Lifelogging Applications . Journal of Visual Communication and Image Representation , 52 , pp. 1-12, 2018.

(future predictions)

Egocentric Action Anticipation, 2020



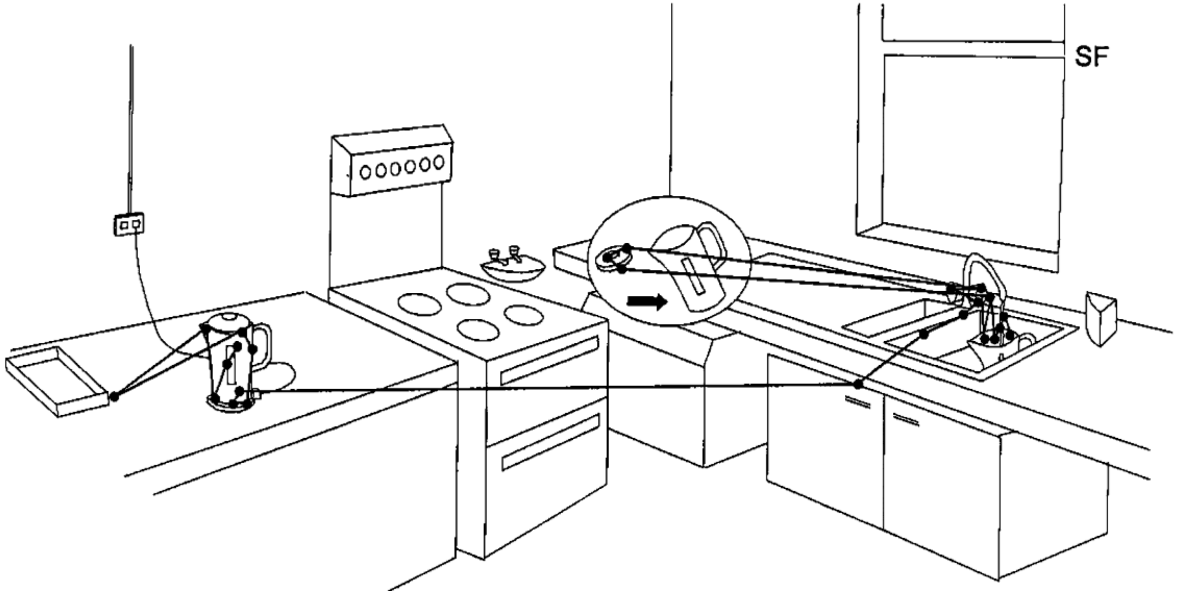


A. Furnari, G. M. Farinella, Rolling-Unrolling LSTMs for Action Anticipation from First-Person Video. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). 2020.

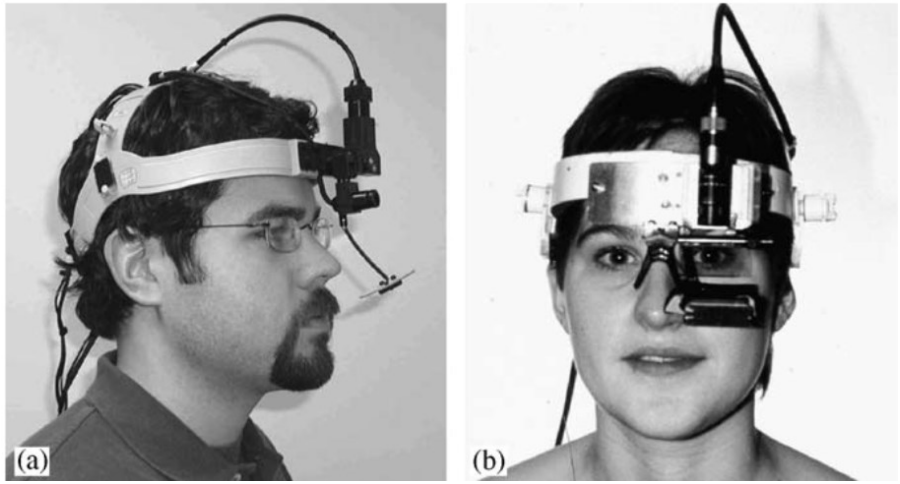
Gaze Trackers

Eye movements and the control of actions in everyday life

Michael F. Land



Gaze is important for First person Vision!



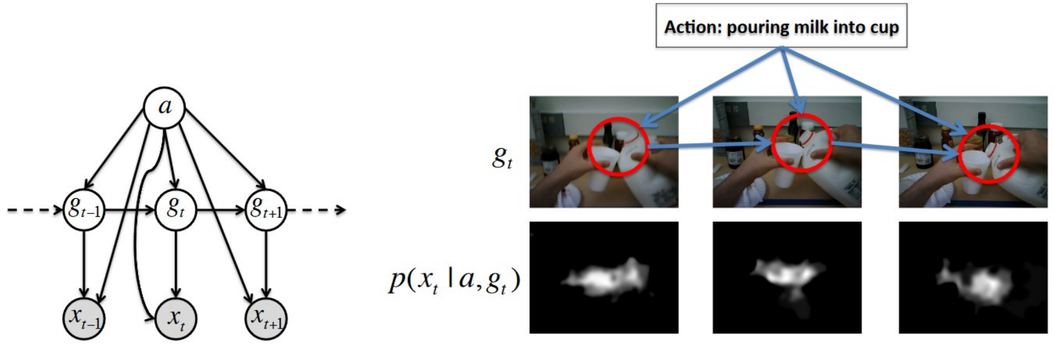
Prototype by Land (1993)

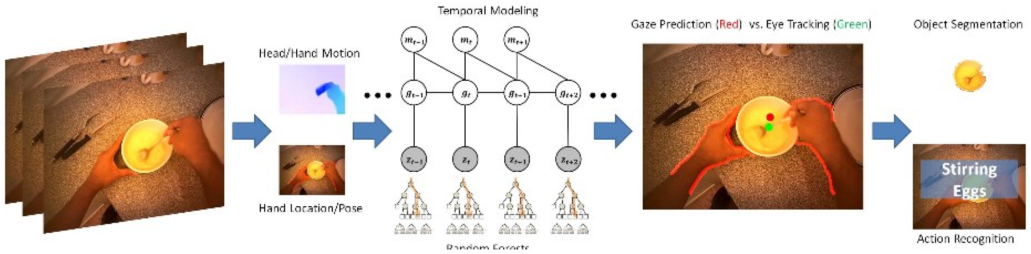
Mobile Eye-XG (2013)

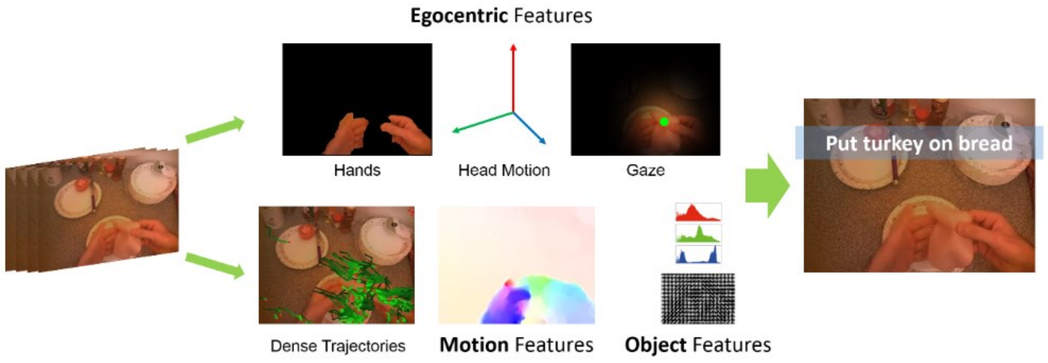
Pupil Eye Tracker (2014)

Tobii Pro Glasses 2 (2014)

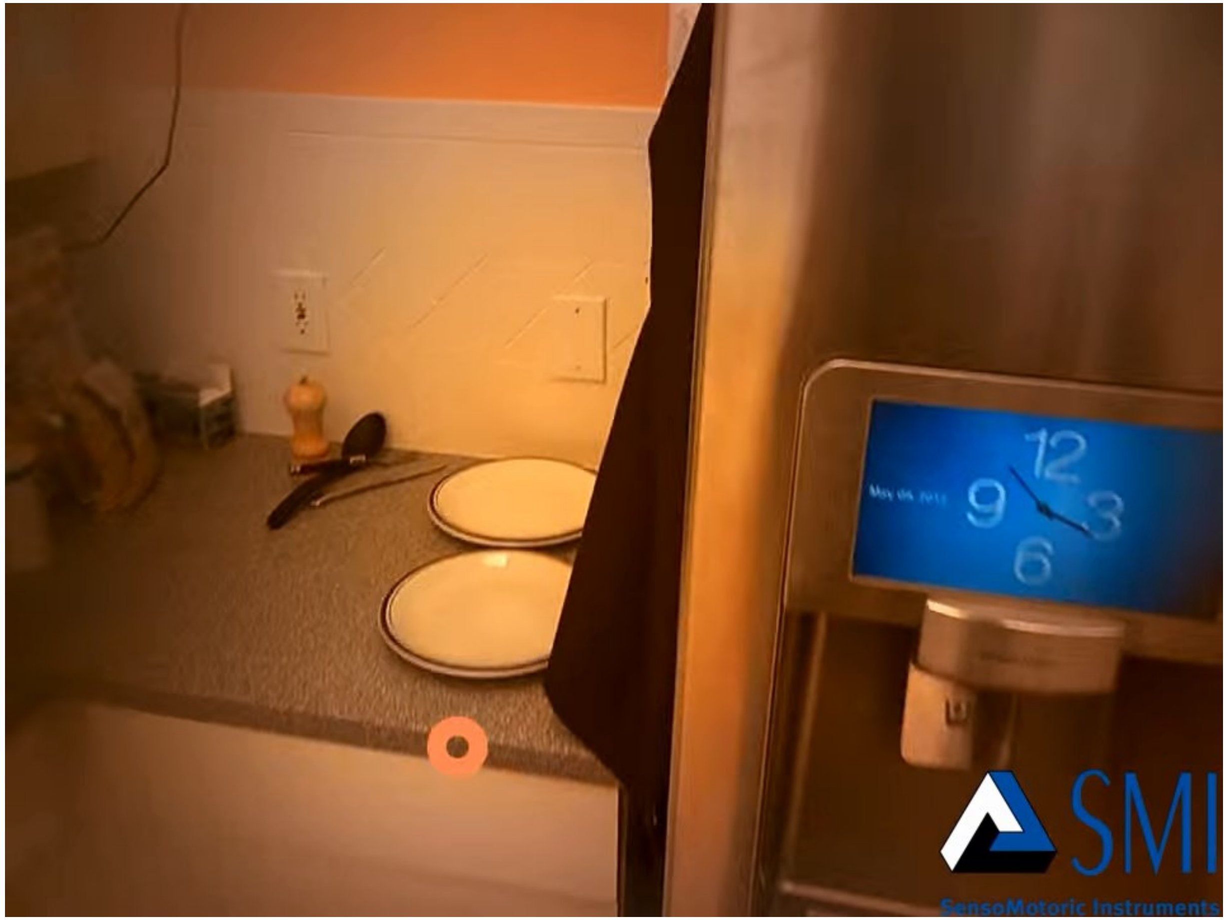
Gaze & Actions Using Gaze, 2012 - 2015







http://ai.stanford.edu/~alireza/GTEA_Gaze_Website/



- Fathi, A., Li, Y., & Rehg, J. M. (2012, October). Learning to recognize daily actions using gaze. In *European Conference on Computer Vision* (pp. 314-327). Springer, Berlin, Heidelberg.
- Li, Yin, Alireza Fathi, and James M. Rehg. "Learning to predict gaze in egocentric video." *Proceedings of the IEEE International Conference on Computer Vision*. 2013.
- Li, Y., Ye, Z., & Rehg, J. M. (2015). Delving into egocentric actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 287-295).

(object usage discovery, assistance)

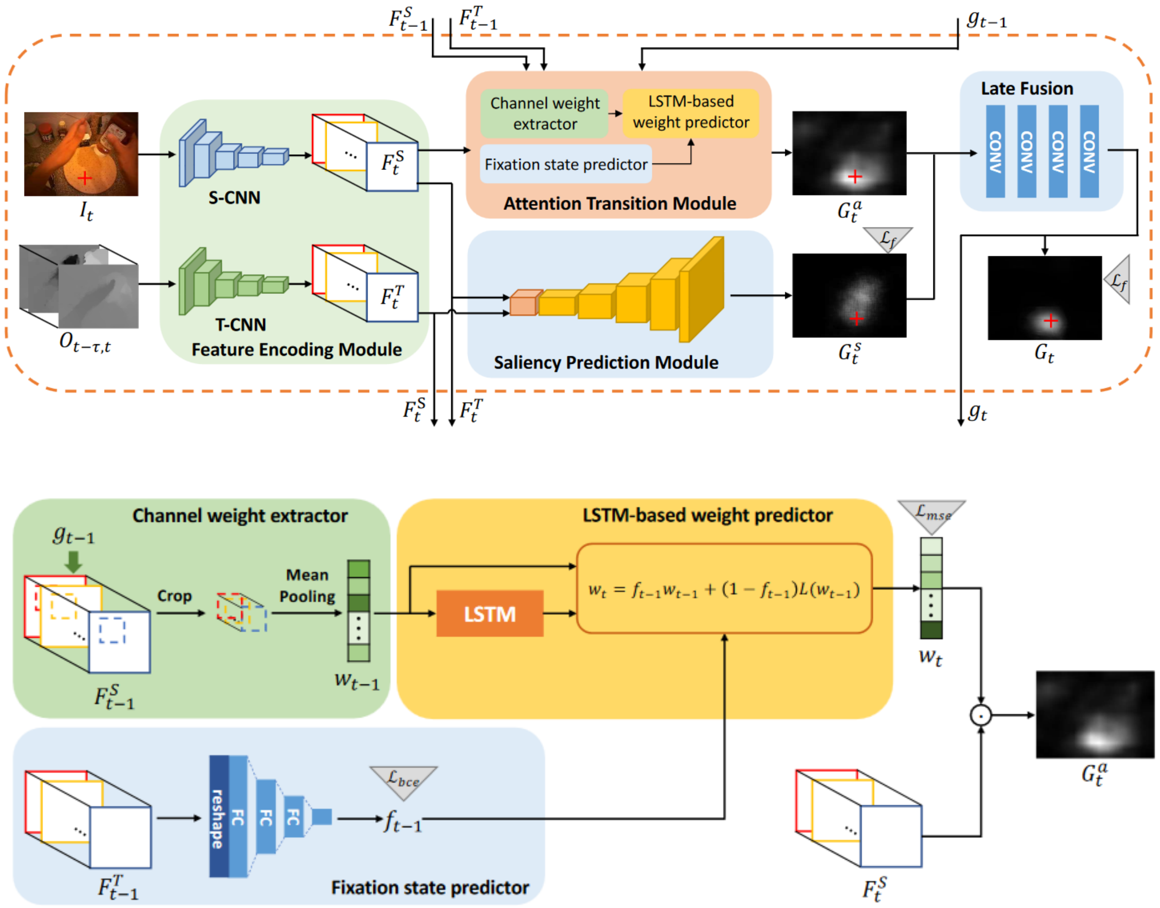
You-Do, I-Learn, 2016

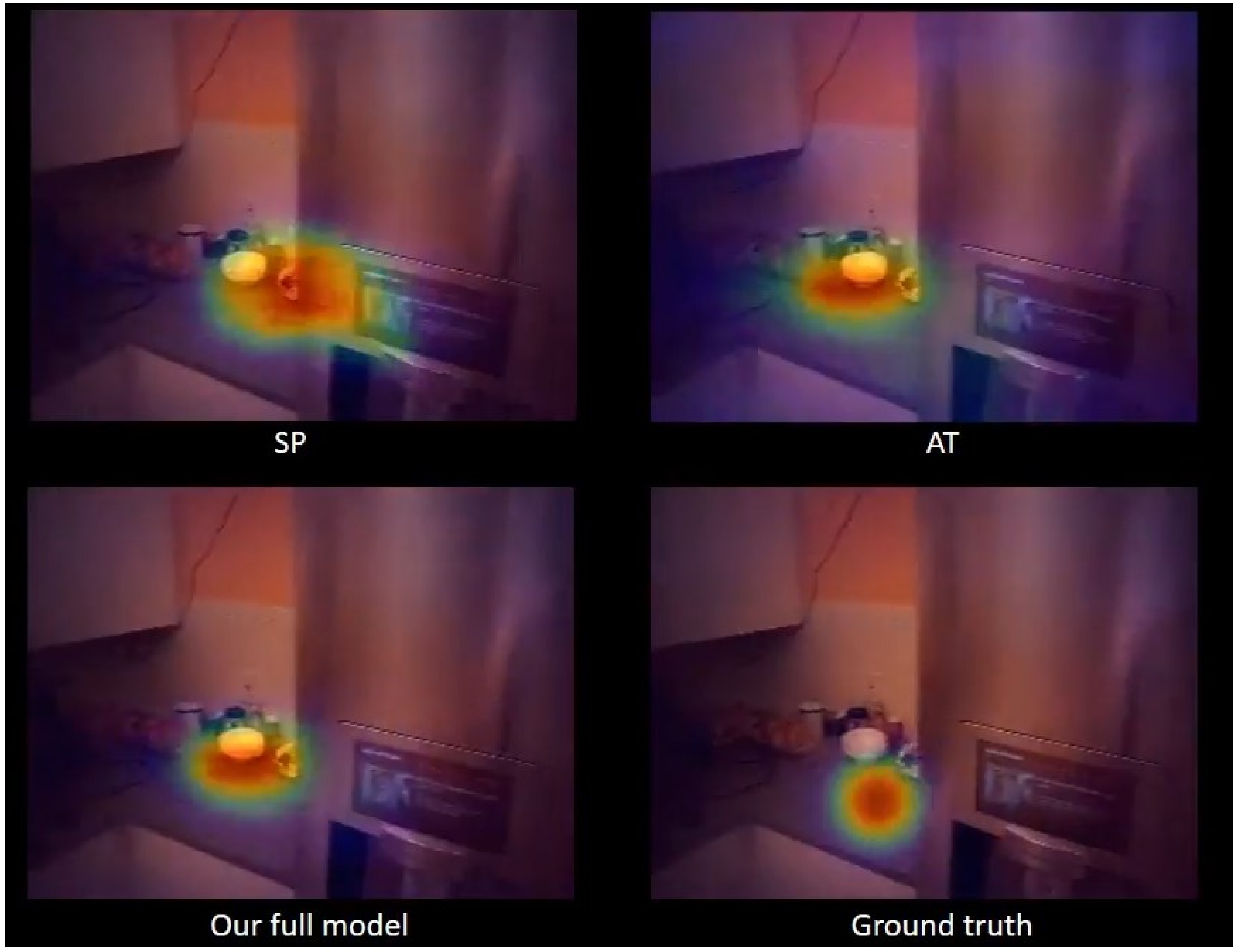
Learning Mode

Assistive Mode

Damen, D., Leelasawassuk, T., Haines, O., Calway, A., & Mayol-Cuevas, W. W. (2014, September). You-Do, I-Learn: Discovering Task Relevant Objects and their Modes of Interaction from Multi-User Egocentric Video. In *BMVC* (Vol. 2, p. 3). https://www.youtube.com/watch?v=vUeRJmwm7DA

Gaze Prediction, 2018

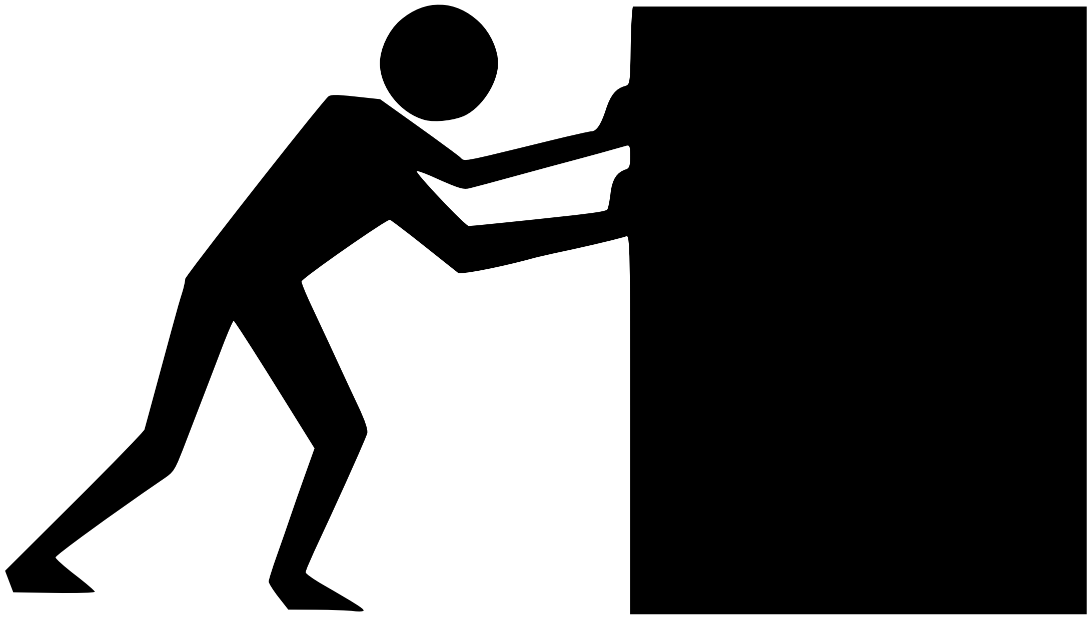




Huang, Y., Cai, M., Li, Z., & Sato, Y. (2018). Predicting Gaze in Egocentric Video by Learning Task-dependent Attention Transition. *ECCV 2018.*

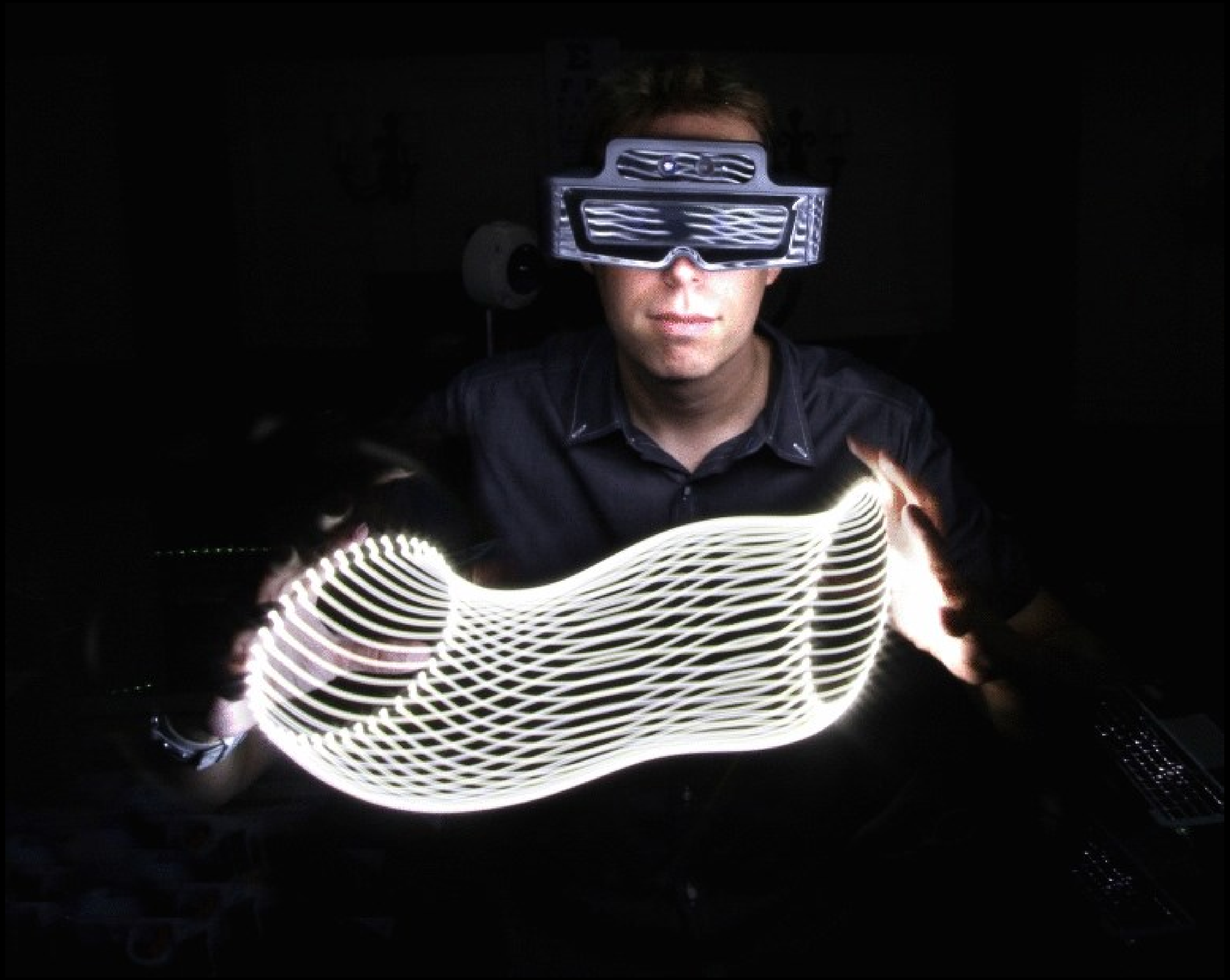
Acquisition devices helped research

however, they moved the focus from action to analysis



ACTION → ANALYSIS

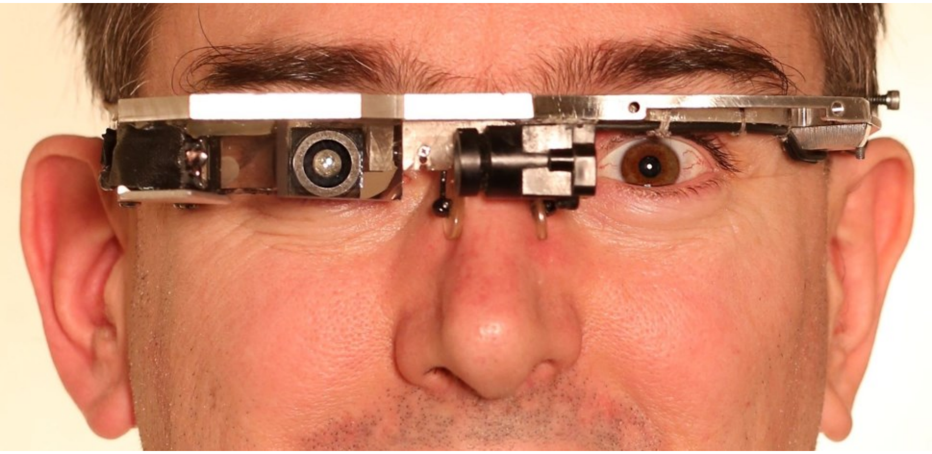
Is the Cyborg dream still possible?



wearcam.org/

EyeTap, Steve Mann, 2000s

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



Mann, S., Fung, J., Aimone, C., Sehgal, A., & Chen, D. (2005). Designing EyeTap digital eyeglasses for continuous lifelong capture and sharing of personal experiences. *Alt. Chi, Proc. CHI 2005*.

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.

Brief Timeline of Google Glass (up to 2014)

2009
- Google creates a research lab to work on a prototype of smart glasses. The first prototype looked like a scuba mask attached to a laptop you carried around with a backpack.

2011
- Early designs to make Google Glass light and portable. The product is there, but there is not clear use for it.

2012
- Google Glass is officially announced to the public with a futuristic video of what it will allow to do (but still doesn't). People love the concept of Google Glass.

2013
- The Glass Explorer program is launched as an early access campaign to gain insight into how people would use Google Glass.

2014
- Google Glass turns out to be an immature product with no clear use case. Its video recording capabilities raise serious privacy issued and the product is ridiculized by the media. The launch of a consumer version is suspended.

Reference: https://medium.com/swlh/the-unexpected-rebirth-of-google-glass-96f6060a62f2

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 with See-Through Display for Augmented Reality, since 2012

https://www.epson.co.uk/products/see-through-mobile-viewer/moverio-bt-300

Vuzix (Since 2012)

Manufacturing Solutions

LEARN MORE

Warehouse Solutions

LEARN MORE

Field Service & Remote Assist Solutions

LEARN MORE

Tele-Medicine Solutions

LEARN MORE

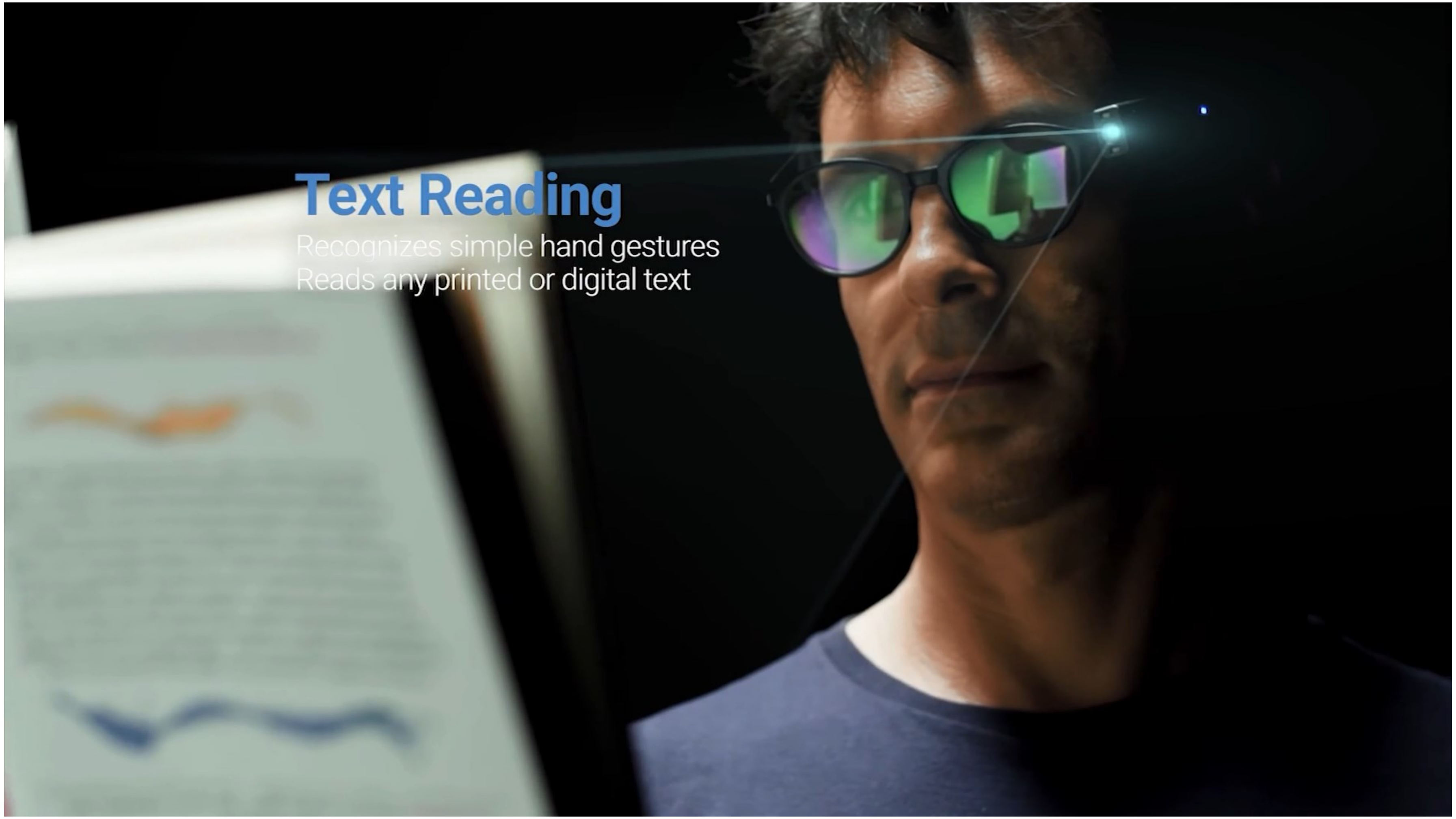
https://www.vuzix.com/

OrCam MyEye, since 2015

Health, assistive technologies

https://www.orcam.com/

OrCam MyEye, since 2015



https://www.orcam.com/

https://www.microsoft.com/hololens

Mixed Reality

Microsoft HoloLens, since 2016 – HoloLens2 in 2020

https://youtu.be/eqFqtAJMtYE

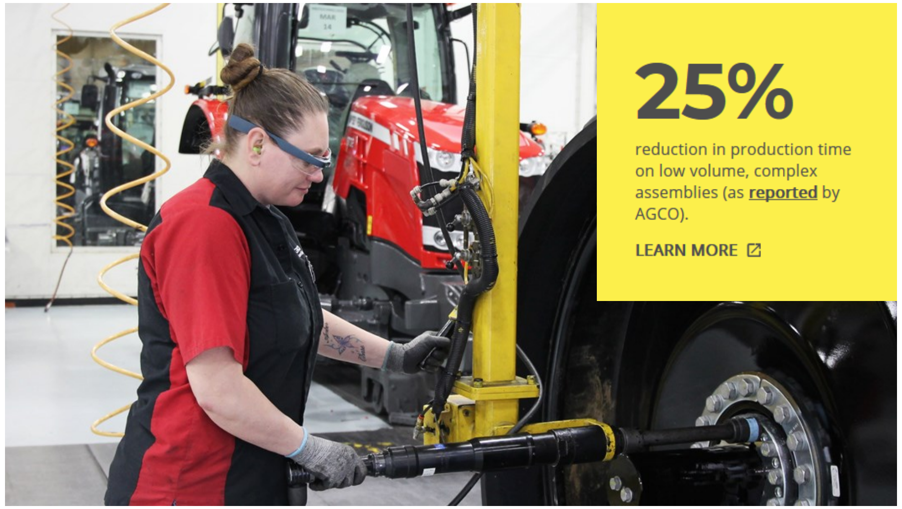
real use cases in industrial scenarios, where privacy is not a issue

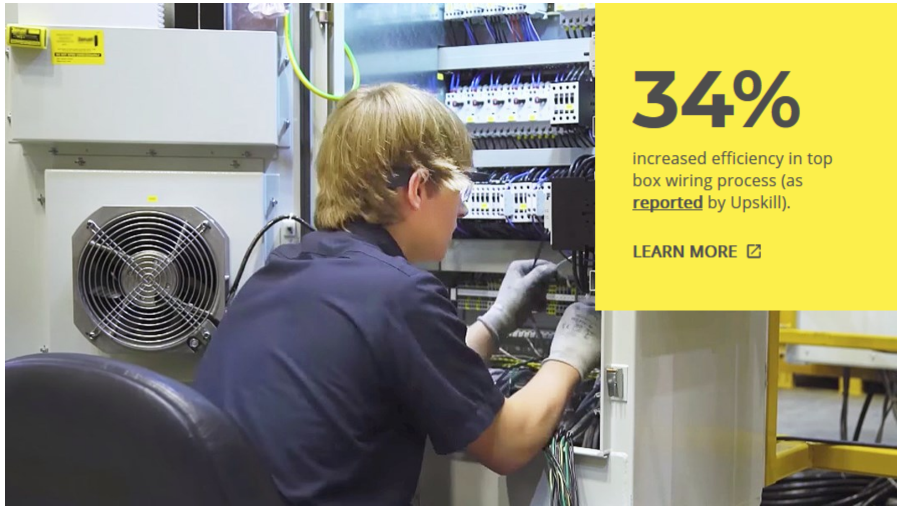
Google Glass Enterprise Edition, since 2017

Stay hands-on

Work smarter

Instant expertise





https://www.x.company/glass/

Magic Leap, since 2018

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

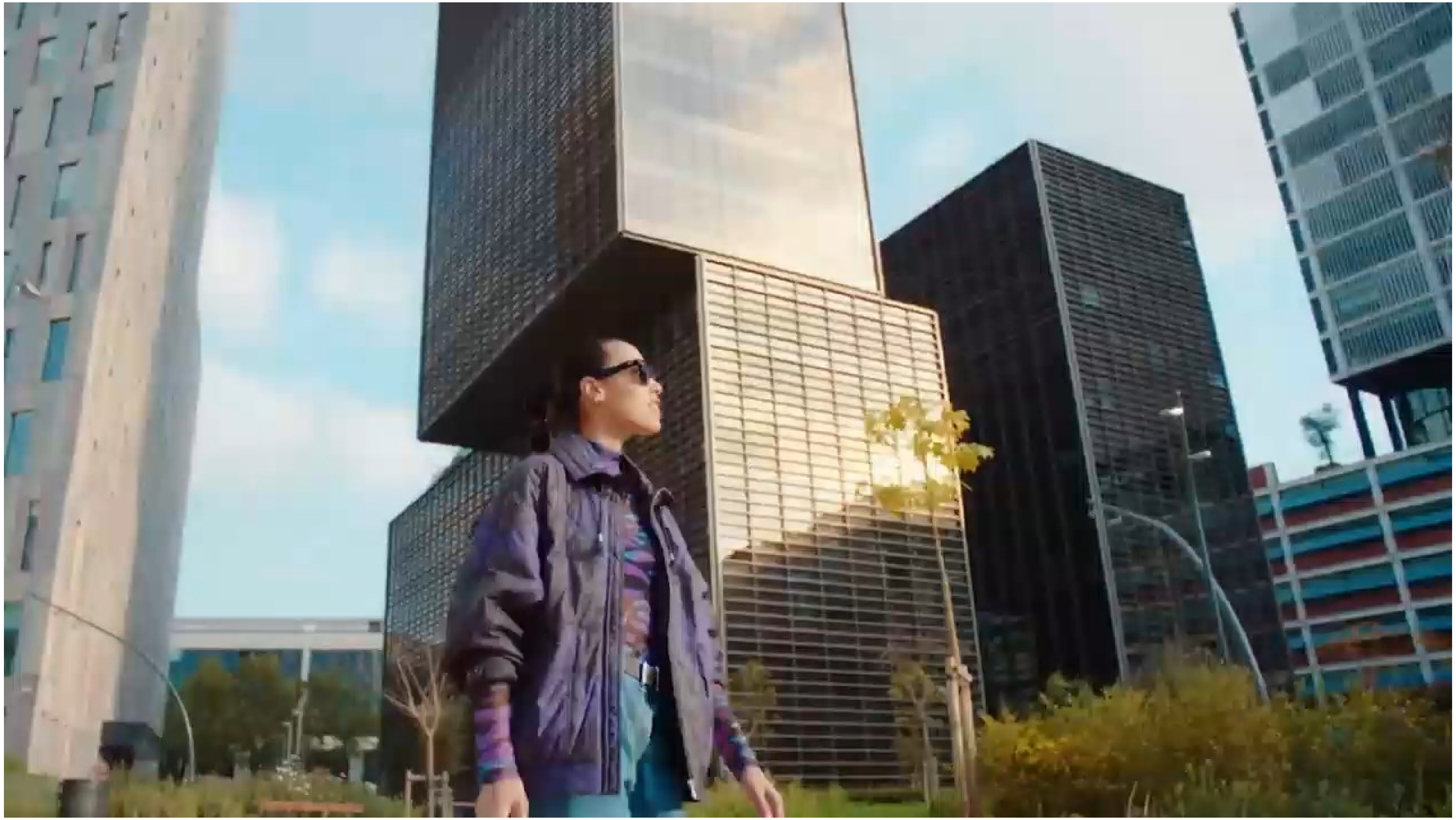
Magic Leap 2 Announced (March 2022)

Magic Leap 2. The most immersive AR headset for enterprise.

Meta’s Project Aria

https://about.facebook.com/realitylabs/projectaria/

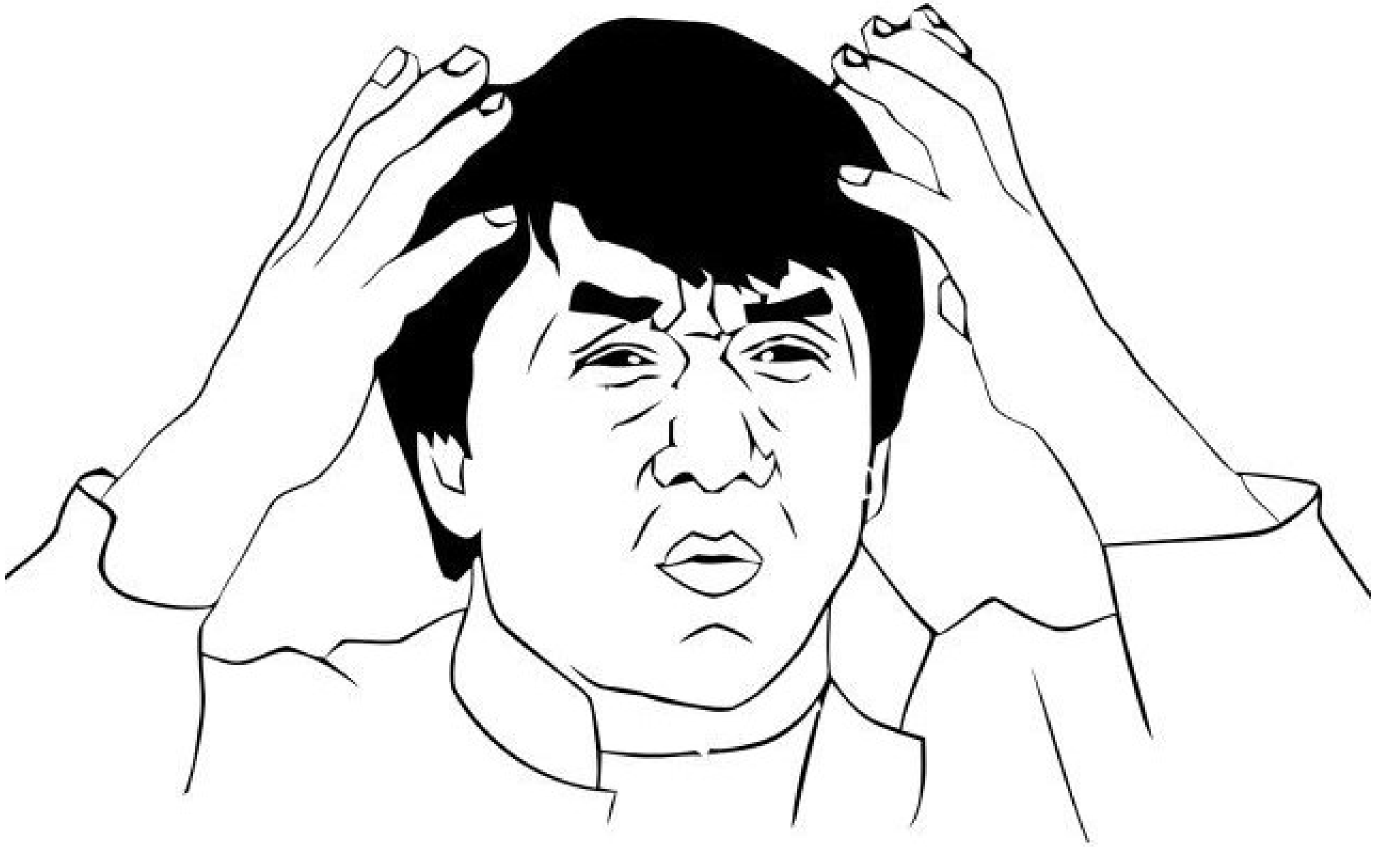
Facebook Rayban Stories



https://www.ray-ban.com/italy/ray-ban-stories

nreal

https://www.nreal.ai/



Too Many Devices?

towards standardization…

OpenXR

Unified API supported by many AR and VR devices

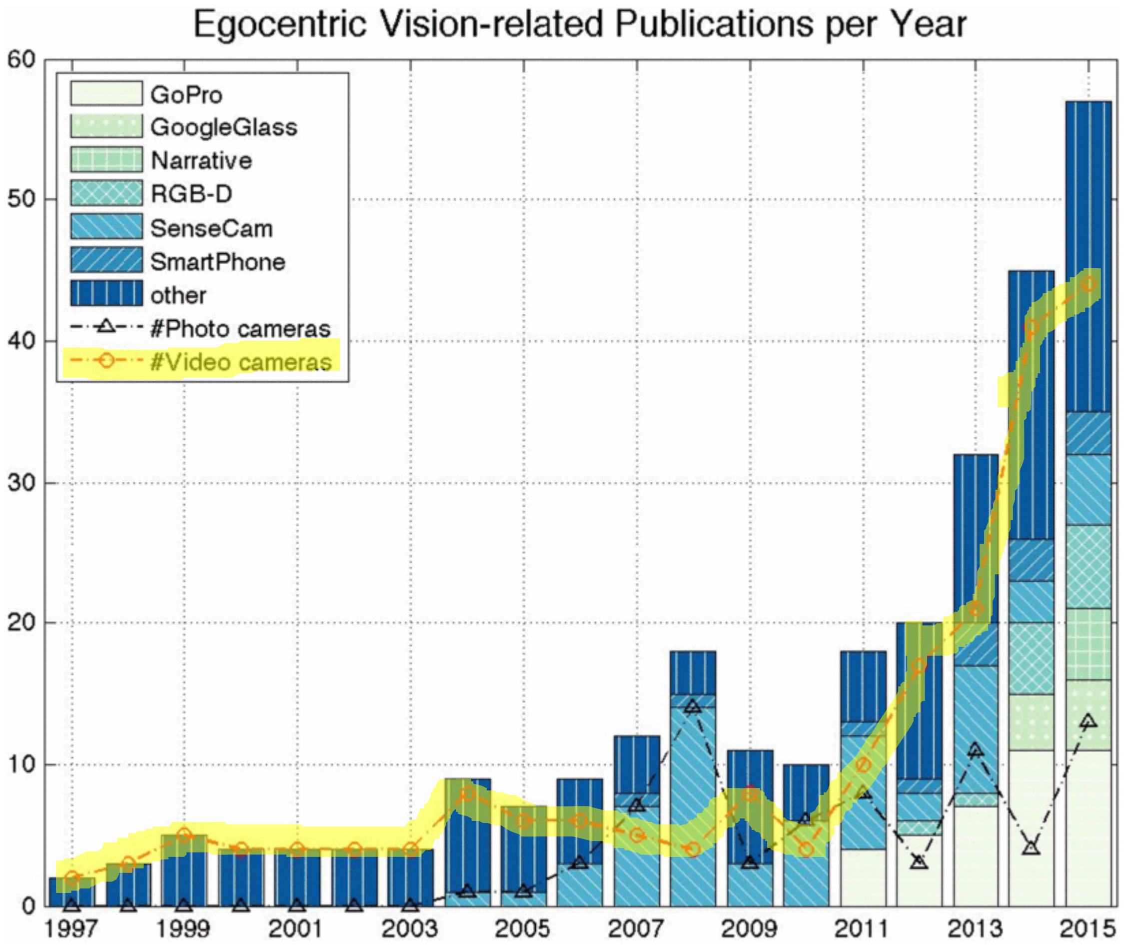
https://www.khronos.org/openxr/

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

First Person Vision Research – Trends



M. Bolaños, M. Dimiccoli and P. Radeva, "Toward Storytelling From Visual Lifelogging: An Overview," in *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 1, pp. 77-90, Feb. 2017.

Growing number of research papers on First Person Vision, expecially with video

…and more

Consumer Devices

Narrative Clip
(http://getnarrative.com/)

€ 229

GoPro
(https://shop.gopro.com)

From € 220

Pupil Eye Tracker
(https://pupil-labs.com/store/)

From € 2850

Microsoft HoloLens 2
(https://www.microsoft.com/en-us/hololens/buy)

From $ 3500

Magic Leap
(https://www.magicleap.com/magic-leap-one)

$ 2295

nreal
(https://shop.nreal.ai/cart)

From $1,199

First Person Vision Research – Conferences

Many conferences/workshops/symposia on wearable computing/first person vision:

Past:

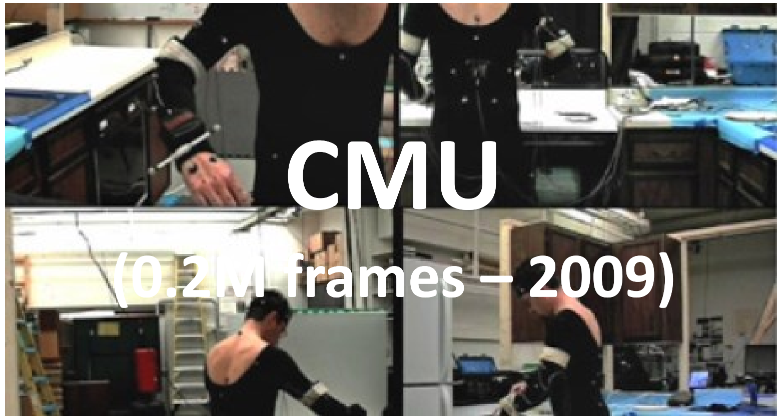
- **SenseCam** series – 2009, 2010, 2012, 2013;
- Workshop on Lifelogging Tools and Applications (**LTA**) – 2016;
- Workshops on Egocentric Vision @ CVPR – 2009, 2012, 2014, 2016

Current:

- International Symposium on Wearable Computers (**ISWC**) – yearly since 1997;
- **UbiComp**/Pervasive/HUC – yearly since 1999;
- ECCV/ICCV Workshop on Assistive Computer Vision and Robotics (**ACVR**) – yearly since 2013;
- **EPIC@X** Workshop Series – yearly since 2016.
- **EGO4D** Workshops – since 2022.
- Special issues in top journals (e.g., TPAMI);
- Many works on First Person Vision appearing in top computer vision conferences (CVPR, ICCV, ECCV) and Journals (TPAMI, TIP, IJCV, PR);

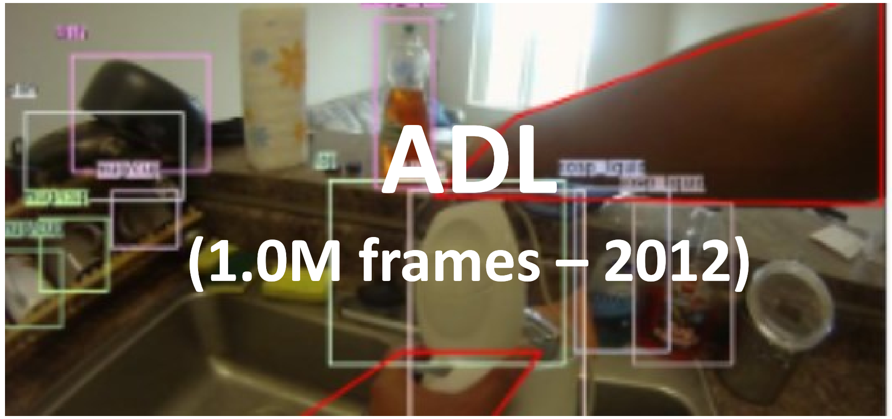
First Person Vision Research – Datasets (up to 2018)

Source: https://arxiv.org/abs/1804.02748



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/

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

EPIC-KITCHENS TEAM

University of Bristol

University of Toronto

Università degli Studi di Catania

Dima Damen
Principal Investigator
University of Bristol
United Kingdom

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


32
KITCHENS

EPIC-KITCHENS
ANNOTATION

NOVEL
AUDIO COMMENTARY APPROACH

VIDEO ACQUISITION
↓
AUDIO COMMENTARY
↓
ACTION SEGMENTS
↓
SEMANTIC PARSING
↓
VERB-NOUN REPRESENTATION

CUT ONION

TURN-OFF TAP

DRY CUP

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

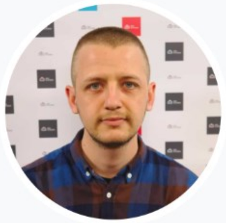
EPIC-KITCHENS-100


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

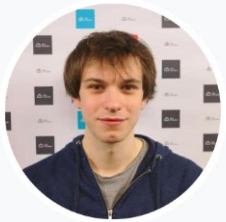

Jian Ma
University of Bristol


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

EPIC KITCHENS

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/

2022 EPIC-KITCHENS Challenges (OPEN!)

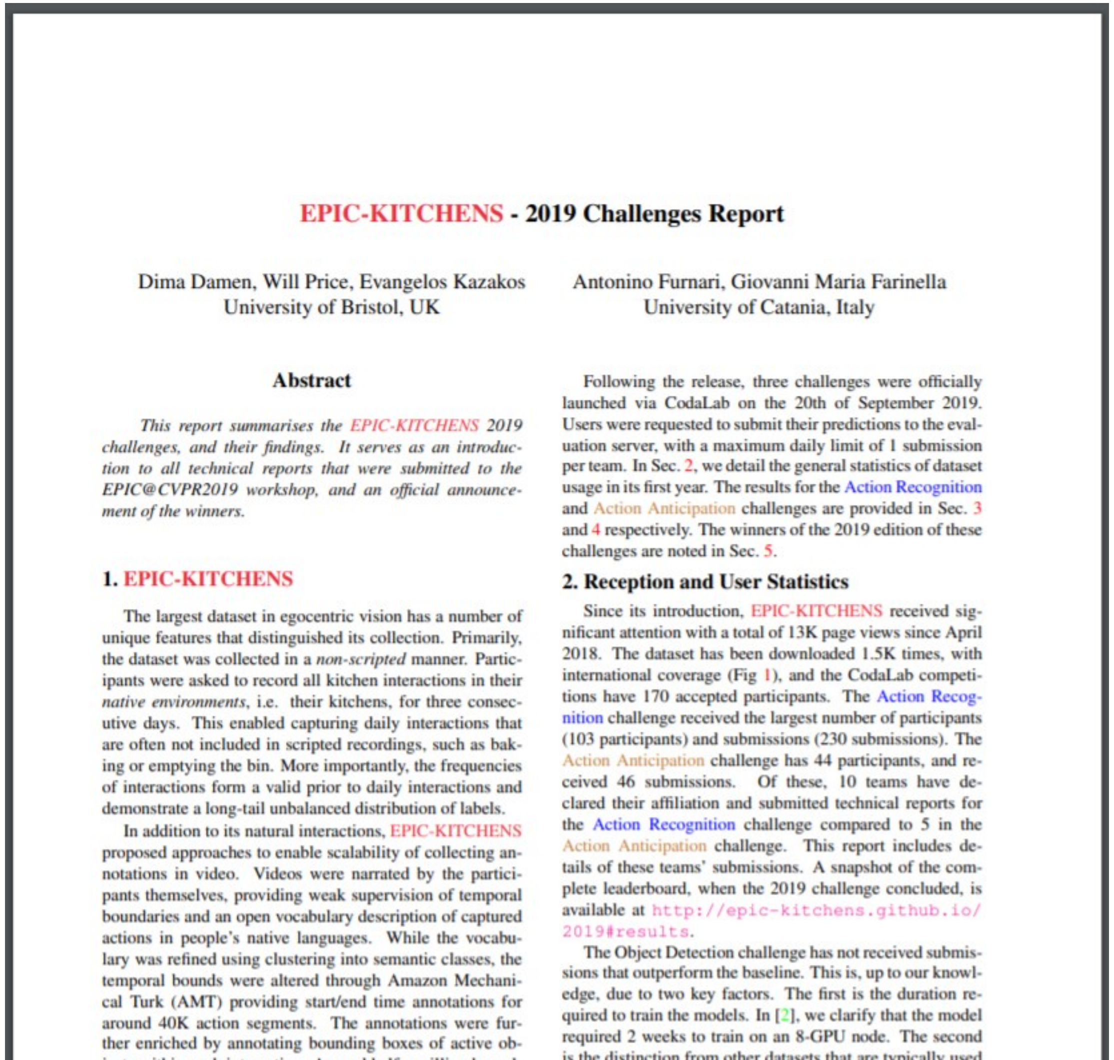
- Action Recognition
- Action Detection
- Action Anticipation
- Unsupervised Domain Adaptation for Action Recognition
- Multi-Instance Retrieval

- Challenges open for submission!
- The winners will be announced at CVPR 2022!
- Goals:
  - Get the community to focus on specific issues
  - Push the state of the art in FPV forward!

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

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

EPIC-KITCHENS – 2019 Challenges Report



EPIC-KITCHENS - 2019 Challenges Report

Dima Damen, Will Price, Evangelos Kazakos
University of Bristol, UK

Antonino Furnari, Giovanni Maria Farinella
University of Catania, Italy

Abstract

This report summarises the EPIC-KITCHENS 2019 challenges, and their findings. It serves as an introduction to all technical reports that were submitted to the EPIC@CVPR2019 workshop, and an official announcement of the winners.

1. EPIC-KITCHENS

The largest dataset in egocentric vision has a number of unique features that distinguished its collection. Primarily, the dataset was collected in a *non-scripted* manner. Participants were asked to record all kitchen interactions in their *native environments*, i.e. their kitchens, for three consecutive days. This enabled capturing daily interactions that are often not included in scripted recordings, such as baking or emptying the bin. More importantly, the frequencies of interactions form a valid prior to daily interactions and demonstrate a long-tail unbalanced distribution of labels.

In addition to its natural interactions, EPIC-KITCHENS proposed approaches to enable scalability of collecting annotations in video. Videos were narrated by the participants themselves, providing weak supervision of temporal boundaries and an open vocabulary description of captured actions in people's native languages. While the vocabulary was refined using clustering into semantic classes, the temporal bounds were altered through Amazon Mechanical Turk (AMT) providing start/end time annotations for around 40K action segments. The annotations were further enriched by annotating bounding boxes of active ob-

Following the release, three challenges were officially launched via CodaLab on the 20th of September 2019. Users were requested to submit their predictions to the evaluation server, with a maximum daily limit of 1 submission per team. In Sec. 2, we detail the general statistics of dataset usage in its first year. The results for the Action Recognition and Action Anticipation challenges are provided in Sec. 3 and 4 respectively. The winners of the 2019 edition of these challenges are noted in Sec. 5.

2. Reception and User Statistics

Since its introduction, EPIC-KITCHENS received significant attention with a total of 13K page views since April 2018. The dataset has been downloaded 1.5K times, with international coverage (Fig 1), and the CodaLab competitions have 170 accepted participants. The Action Recognition challenge received the largest number of participants (103 participants) and submissions (230 submissions). The Action Anticipation challenge has 44 participants, and received 46 submissions. Of these, 10 teams have declared their affiliation and submitted technical reports for the Action Recognition challenge compared to 5 in the Action Anticipation challenge. This report includes details of these teams' submissions. A snapshot of the complete leaderboard, when the 2019 challenge concluded, is available at http://epic-kitchens.github.io/2019#results.

The Object Detection challenge has not received submissions that outperform the baseline. This is, up to our knowledge, due to two key factors. The first is the duration required to train the models. In [2], we clarify that the model required 2 weeks to train on an 8-GPU node. The second is the distinction from other datasets that are typically used

Winners of the 2019 edition

		Team	Member	Affiliations
Action Recognition	1	UTS-Baidu (wasun)	Xiaohan Wang	University of Technology Sydney,Baidu Research
			Yu Wu	University of Technology Sydney,Baidu Research
			Linchao Zhu	University of Technology Sydney
			Yi Yang	University of Technology Sydney
	2	FAIR (deeptigp)	Deepti Ghadiyaram	Facebook AI
			Matt Feiszli	Facebook AI
			Du Tran	Facebook AI
			Xueting Yan	Facebook AI
			Heng Wang	Facebook AI
			Dhruv Mahajan	Facebook AI
	3	FBK-HUPBA (sudhakran)	Swathikiran Sudhakaran	FBK, University of Trento
			Sergio Escalera	CVC, Universitat de Barcelona
			Oswald Lanz	FBK, University of Trento
Action Anticipation	1	RML (Nour)	Nour Eldin Elmadany	Ryerson University
			Yifeng He	Ryerson University
			Ling Guan	Ryerson University
	2	Inria-Facebook (masterchef)	Antoine Miech	Inria, Ecole Normale Superieure
			Ivan Laptev	Inria, Ecole Normale Superieure
			Josef Sivic	Inria, Ecole Normale Superieure, CIRC
			Heng Wang	Facebook AI
			Lorenzo Torresani	Facebook AI
			Du Train	Facebook AI
	3	NTU (zhe2325138)	Zhe-Yu Liu	National Taiwan University
			Ya-Liang Chung	National Taiwan University
			Chih-Hung Liang	National Taiwan University
			Yun-Hsuan Liu	National Taiwan University
			Ke-Jyun Wang	National Taiwan University
			Winston Hsu	National Taiwan University
	3	Bonn (yassersouri)	Yaser Souri	University of Bonn
			Tridivraj Bhattacharyya	University of Bonn
			Juergen Gall	University of Bonn
			Luca Minciullo	Toyota Motor Europe

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

EPIC-KITCHENS – 2020 Challenges Report

EPIC-KITCHENS-55 - 2020 Challenges Report

Dima Damen, Evangelos Kazakos, Will Price, Jian Ma, Hazel Doughty
University of Bristol, UK

Antonino Furnari, Giovanni Maria Farinella
University of Catania, Italy

Abstract

This report summarises the EPIC-KITCHENS-55 2020 challenges, and their findings. It serves as an introduction to all technical reports that were submitted to the EPIC@CVPR2020 workshop, and an official announcement of the winners.

1. EPIC-KITCHENS-55

As the largest dataset in egocentric vision, EPIC-KITCHENS-55 continued to receive significant attention from the research community over the past year. EPIC-KITCHENS-55 has a number of unique features that distinguished its collection, including *non-scripted* and *untrimmed* nature of the footage captured in participants' *native environments*. More details on the dataset's collection and annotation pipeline are available in this year's PAMI extended version [3].

This report details the submissions and winners of the 2020 edition of the three challenges available on CodaLab: Action Recognition, Action Anticipation and object detection. For each challenge, submissions were limited per team to a maximum of 50 submissions in total, as well as a maximum daily limit of 1 submission. In Sec. 2, we detail the general statistics of dataset usage in its first year. The results for the Action Recognition, Action Anticipation and Object Detection in Video challenges are provided in Sec. 3, 4 and 5 respectively. The winners of the 2020 edition of these challenges are noted in Sec. 6.

Details of the 2019 challenges are available from the technical report [4].

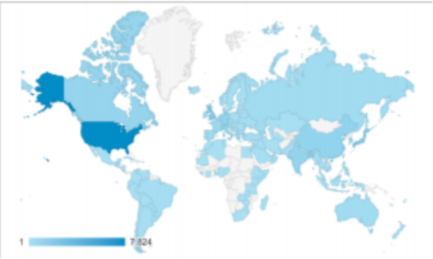


Figure 1: Heatmap of countries based on EPIC-KITCHENS-55 page view statistics.

shows page views of the dataset's website, based on country. The Action Recognition challenge received the largest number of participants (46 teams) and submissions (368 submissions). The Action Anticipation challenge has 20 participating teams, and received 244 submissions. The Object Detection in Video challenge has 16 participating teams and 182 submissions. Of these, 8 teams have declared their affiliation and submitted technical reports for the Action Recognition challenge compared to 5 in the Action Anticipation challenge and 4 in the Object Detection in Video challenge. This report includes details of these teams' submissions. A snapshot of the complete leaderboard, when the 2020 challenge concluded on the 30th of May, is available at `http://epic-kitchens.github.io/2020#results`.

Winners of the 2020 edition

	S1	S2	Team	Member	Affiliations
Action Recognition	1	1	UTS-Baidu (wasun)	Xiaohan Wang	University of Technology Sydney, Baidu Research
				Yu Wu	University of Technology Sydney, Baidu Research
				Linchao Zhu	University of Technology Sydney
				Yi Yang	University of Technology Sydney
				Yueting Zhuang	Zhejiang University
	2	3	NUS-CVML (action-banks)	Fadime Sener	University of Bonn
				Dipika Singhania	National University of Singapore
				Angela Yao	National University of Singapore
	4	2	GT-WISC-MPI (aptx4869lm)	Miao Liu	Georgia Institute of Technology
				Yin Li	University of Wisconsin-Madison
				James M. Rehg	Georgia Institute of Technology
	3	5	FBK-HUPBA (sudhakran)	Swathikiran Sudhakaran	FBK, University of Trento
				Sergio Escalera	CVC, Universitat de Barcelona
				Oswald Lanz	FBK, University of Trento
	3	6	SAIC-Cambridge (tnet)	Juan-Manuel Perez-Rua	Samsung AI Centre, Cambridge
				Antoine Toisoul	Samsung AI Centre, Cambridge
				Brais Martinez	Samsung AI Centre, Cambridge
				Victor Escorcia	Samsung AI Centre, Cambridge
				Li Zhang	Samsung AI Centre, Cambridge
				Xiatian Zhu	Samsung AI Centre, Cambridge
				Tao Xiang	Samsung AI Centre Cambridge, Univ of Surrey
Action Anticipation	1	3	NUS-CVML (action-banks)	Fadime Sener	University of Bonn
				Dipika Singhania	National University of Singapore
				Angela Yao	National University of Singapore
	2	1	Ego-OMG (edessale)	Eadom Dessalene	University of Maryland, College Park
				Michael Maynord	University of Maryland, College Park
				Chinmaya Devaraj	University of Maryland, College Park
				Cornelia Fermuller	University of Maryland, College Park
				Yiannis Aloimonos	University of Maryland, College Park
	2	2	VI-I2R (chengyi)	Ying Sun	A*STAR, Singapore
				Yi Cheng	A*STAR, Singapore
				Mei Chee Leong	A*STAR, Singapore
				Hui Li Tan	A*STAR, Singapore
				Kenan E. Ak	A*STAR, Singapore
Object Detection in Video	1	2	hutom (killerchef)	Jihun Yoon	hutom
				Seungbum Hong	hutom
				Sanha Jeong	hutom
				Min-Kook Choi	hutom
	3	1	FB AI (gb7)	Gedas Bertasius	Facebook AI
				Lorenzo Torresani	Facebook AI
	2	3	DHARI (kide)	Kaide Li	ZheJiang Dahua Technology
				Bingyan Liao	ZheJiang Dahua Technology
				Laifeng Hu	ZheJiang Dahua Technology
				Yaonong Wang	ZheJiang Dahua Technology

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

EPIC-KITCHENS – 2021 Challenges Report

EPIC-KITCHENS-100- 2021 Challenges Report

Dima Damen, Adriano Fragomeni, Jonathan Munro, Toby Perrett, Daniel Whettam, Michael Wray
University of Bristol, UK

Antonino Furnari, Giovanni Maria Farinella
University of Catania, Italy

Davide Moltisanti
NTU, Singapore

Abstract

This report summarises the EPIC-KITCHENS-100 2021 challenges, and their findings. It serves as an introduction to all technical reports that were submitted to the EPIC@CVPR2021 workshop, and an official announcement of the winners.

1. EPIC-KITCHENS-100

In July 2020, EPIC-KITCHENS-100 was released as the next version of the EPIC-KITCHENS dataset. EPIC-KITCHENS-100, like its previous version EPIC-KITCHENS-55, has a number of unique features that distinguished its collection, including *non-scripted* and *untrimmed* nature of the footage captured in participants' *native environments*. The dataset was extended in footage, up to 100 hours of annotated egocentric footage. More importantly, the pipeline for annotations was revised and improved on every step including the pause-and-talk narrator, which increased the density and correctness of the annotations. More details and statistics on EPIC-KITCHENS-100 can be found at [5]. Notably, each submission is requested to provide their level of supervision following the proposed Supervision Levels Scale (SLS) [9].

This report details the submissions and winners of the 2021 edition of the five challenges available on CodaLab: Action Recognition, Action Anticipation, Action Detection, Unsupervised Domain Adaptation for Recognition and Multi-Instance Retrieval. For each challenge, submissions were limited per team to a maximum of 50 submissions in total, as well as a maximum daily limit of 1 submission. In Sec. 2, we detail the general statistics of dataset usage. The results for all challenges are provided in Sec. 3-7. The winners of the 2021 edition of these challenges are noted in

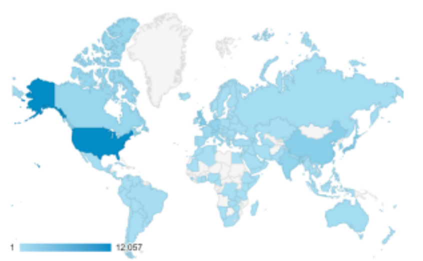


Figure 1: Heatmap of countries based on EPIC-KITCHENS-100 webpage view statistics.

United States	744	China	447	Japan	237
Germany	162	United Kingdom	89	India	82
Unknown	73	France	65	Canada	51
Spain	50	Netherlands	49	Singapore	44
Australia	36	Turkey	33	Italy	33
South Korea	31	Russia	16	Iran	10
Finland	8	Taiwan	7	Austria	7
Malaysia	6	Switzerland	6	Brazil	6
Greece	5	Ireland	4	Thailand	3
Israel	3	Mexico	3	Poland	3
Romania	2	Pakistan	2	Slovakia	2
Vietnam	2	Belgium	2	Croatia	2
Sweden	2	Ukraine	1	Libya	1
Belarus	1	Portugal	1	Algeria	1
Morocco	1	Myanmar	1	Argentina	1

Table 1: Downloads for EPIC-KITCHENS dataset, by country

2021 challenge concluded on the 28th of May, is

Winners of the 2021 edition

		Team	Member	Affiliations
Action Recognition	①	SCUT-JD (hrgdscs)	Zeyu Jiang	South China University of Technology
			Changxing Ding	South China University of Technology
			Canwei Zhang	South China University of Technology
			Dacheng Tao	JD Explore Academy
	②	NUS-HUST-THU-Alibaba (ZiyuanHuang)	Ziyuan Huang	National University of Singapore
			Zhiwu Qing	Huazhong University of Science and Technology
			Xiang Wang	Huazhong University of Science and Technology
			Yutong Feng	Tsinghua University
			Shiwei Zhang	DAMO Academy, Alibaba Group
			Jianwen Jiang	DAMO Academy, Alibaba Group
			Zhurong Xia	DAMO Academy, Alibaba Group
			Mingqian Tang	DAMO Academy, Alibaba Group
			Nong Sang	Huazhong University of Science and Technology
			Marcelo H. Ang Jr	National University of Singapore
	③	SAIC-FBK-UB (Sudhakaran)	Swathikiran Sudhakaran	Samsung AI Center, Cambridge
			Adrian Bulat	Samsung AI Center, Cambridge
			Juan-Manuel Perez-Rua	Samsung AI Center, Cambridge
			Alex Falcon	Fondazione Bruno Kessler - FBK, Trento
			Sergio Escalera	Universitat de Barcelona, Spain
			Oswald Lanz	Fondazione Bruno Kessler - FBK, Trento
			Brais Martinez	Samsung AI Center, Cambridge
			Georgios Tzimiropoulos	Samsung AI Center, Cambridge
Action Anticipation	①	AVT-FB-UT (shef)	Rohit Girdhar	Facebook AI Research
			Kristen Grauman	Facebook AI Research
	②	Panasonic-CNSIC-PSNRD (panasonic)	Yutaro Yamamuro	Panasonic System Networks R&D Lab
			Kazuki Hanazawa	Panasonic System Networks R&D Lab
			Masahiro Shida	Panasonic System Networks R&D Lab
			Tsuyoshi Kodake	Panasonic System Networks R&D Lab
			Shinji Takenaka	Panasonic System Networks R&D Lab
			Yuji Sato	Connected Solutions Company, Panasonic
			Takeshi Fujimatsu	Connected Solutions Company, Panasonic
	③	ICL-SJTU (Shawn0822)	Xiao Gu	Imperial College London
			Jianing Qiu	Imperial College London
			Yao Guo	Shanghai Jiao Tong University
			Benny Lo	Imperial College London
			Guang-Zhong Yang	Shanghai Jiao Tong University
Action Detection	①	HUST-NUS-THU-Alibaba (ZiyuanHuang)	Zhiwu Qing	Huazhong University of Science and Technology
			Ziyuan Huang	National University of Singapore
			Xiang Wang	Huazhong University of Science and Technology
			Yutong Feng	Tsinghua University
			Shiwei Zhang	DAMO Academy, Alibaba Group
			Jianwen Jiang	DAMO Academy, Alibaba Group
			Mingqian Tang	DAMO Academy, Alibaba Group
			Changxin Gao	Huazhong University of Science and Technology
			Marcelo H. Ang Jr	National University of Singapore
			Nong Sang	Huazhong University of Science and Technology
	②	LocTransformer (evangelion)	Chen-Lin Zhang	Nanjing University
			Jianxin Wu	Nanjing University
			Yin Li	University of Wisconsin-Madison
UDA for Recognition	①	A*STAR (chengyi)	Yi Cheng	A*STAR, Singapore
			Fen Fang	A*STAR, Singapore
			Ying Sun	A*STAR, Singapore
	②	Tokyo (M3EM)	Lijin Yang	University of Tokyo, Japan
			Yifei Huang	NUniversity of Tokyo, Japan
			Yusuke Sugano	University of Tokyo, Japan
			Yoichi Sato	University of Tokyo, Japan
	③	Torino (plnet)	Chiara Plizzari	Politecnico di Torino, Italy
			Mirco Planamente	Politecnico di Torino, Italy
			Emanuele Alberti	Politecnico di Torino, Italy
			Barbara Caputo	Politecnico di Torino, Italy
MI Retrieval	①	IIE-MRG (haoxiaoshuai)	Xiaoshuai Hao	Institute of Information Engineering, CAS
			Wanqian Zhang	Institute of Information Engineering, CAS
			Dejie Yang	Institute of Information Engineering, CAS
			Shu Zhao	Institute of Information Engineering, CAS
			Dayan Wu	Institute of Information Engineering, CAS
			Bo Li	Institute of Information Engineering, CAS
			Weiping Wang	Institute of Information Engineering, CAS

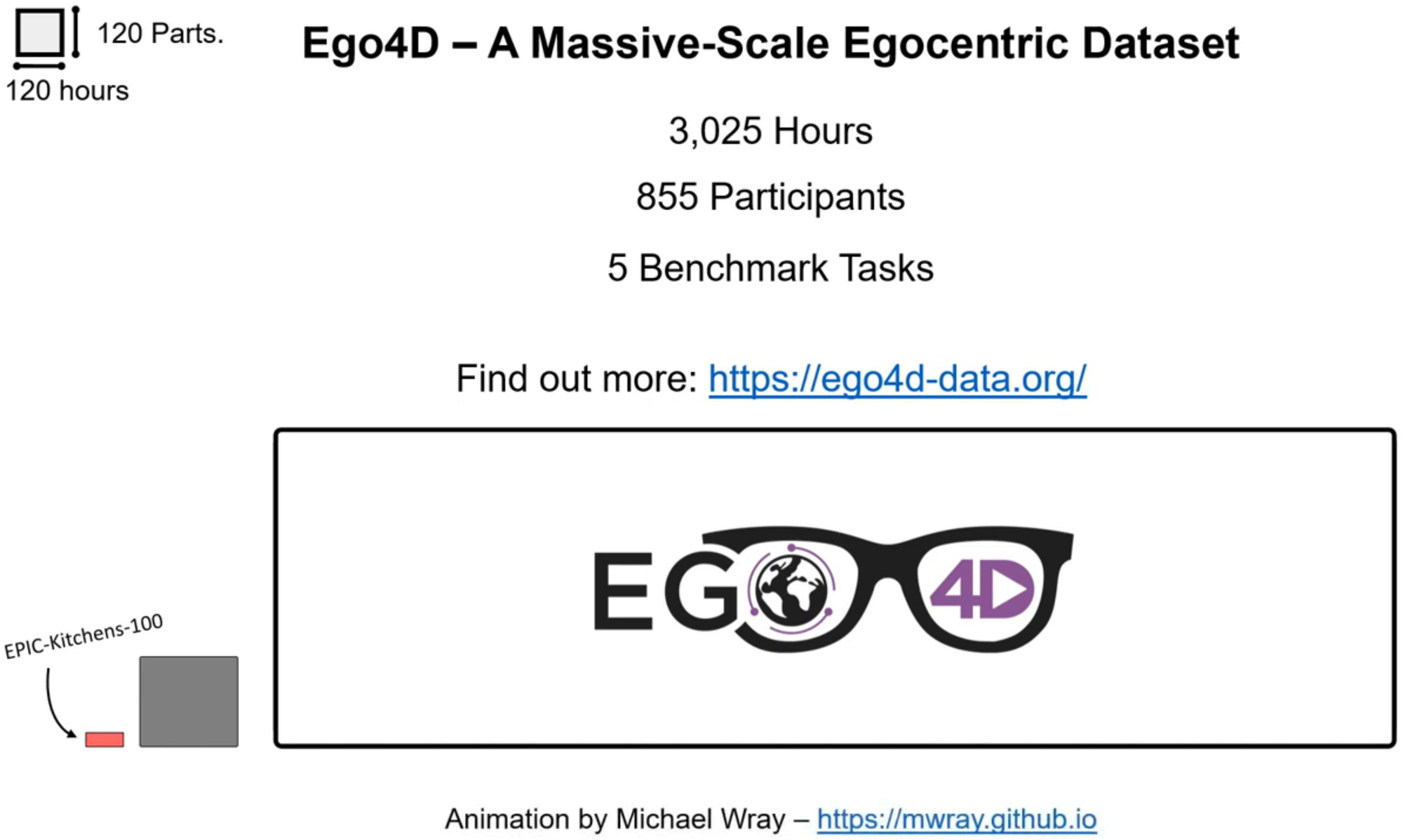
Consortium

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

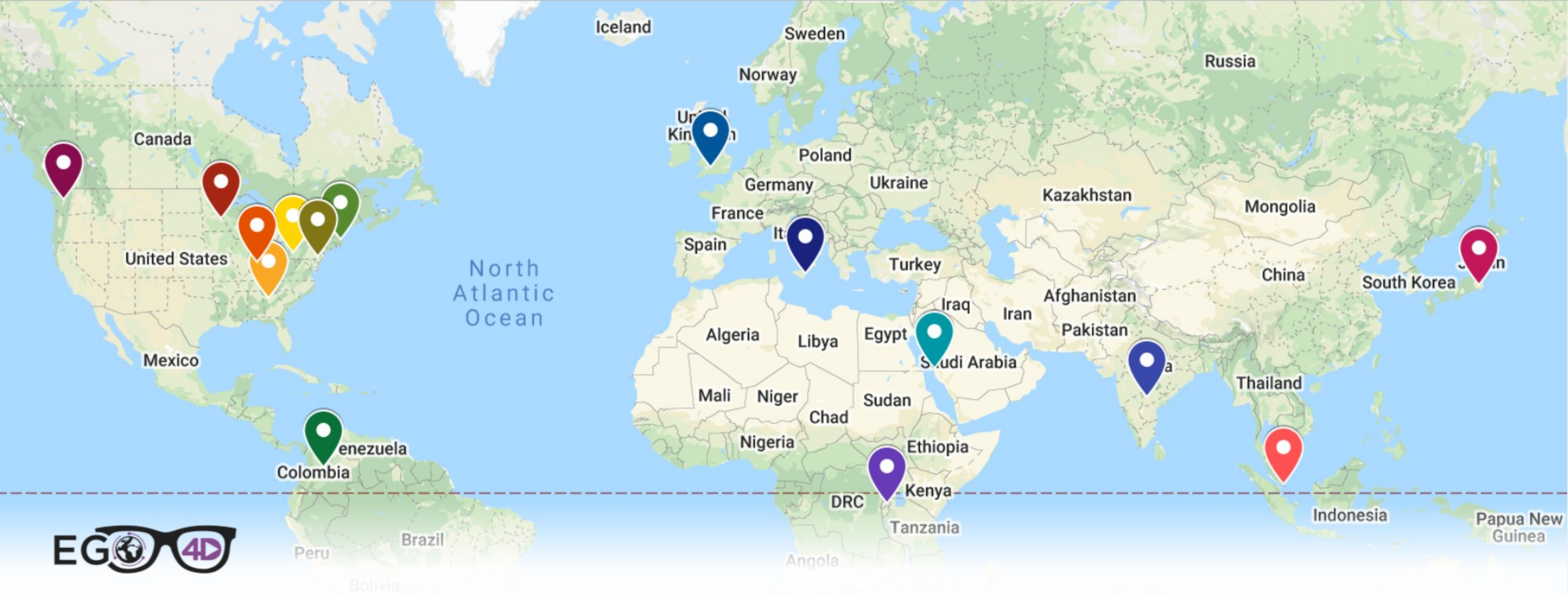
Kristen Grauman[1,2], Andrew Westbury[1], Eugene Byrne[*1], Zachary Chavis[*3], Antonino Furnari[*4], Rohit Girdhar[*1], Jackson Hamburger[*1], Hao Jiang[*5], Miao Liu[*6], Xingyu Liu[*7], Miguel Martin[*1], Tushar Nagarajan[*1,2], Ilija Radosavovic[*8], Santhosh Kumar Ramakrishnan[*1,2], Fiona Ryan[*6], 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

EGO4D – Massive Scale



Animation by Michael Wray - https://www.youtube.com/watch?v=_p78-V2RiKo



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

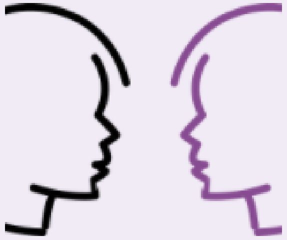

Geographic diversity
ASIA
AFRICA
SOUTH AMERICA
OCEANIA
Baking
Doing laundry
Multi-perspective
IMU
Sports
Shopping
Stereo vision
Reading
Human locomotion
Gardening
Sewing / Knitting
Pets
Playing games
Social interaction
3D
Video + 3D scans

Challenges

Episodic Memory

Hand-Object Interactions

AV Diarization



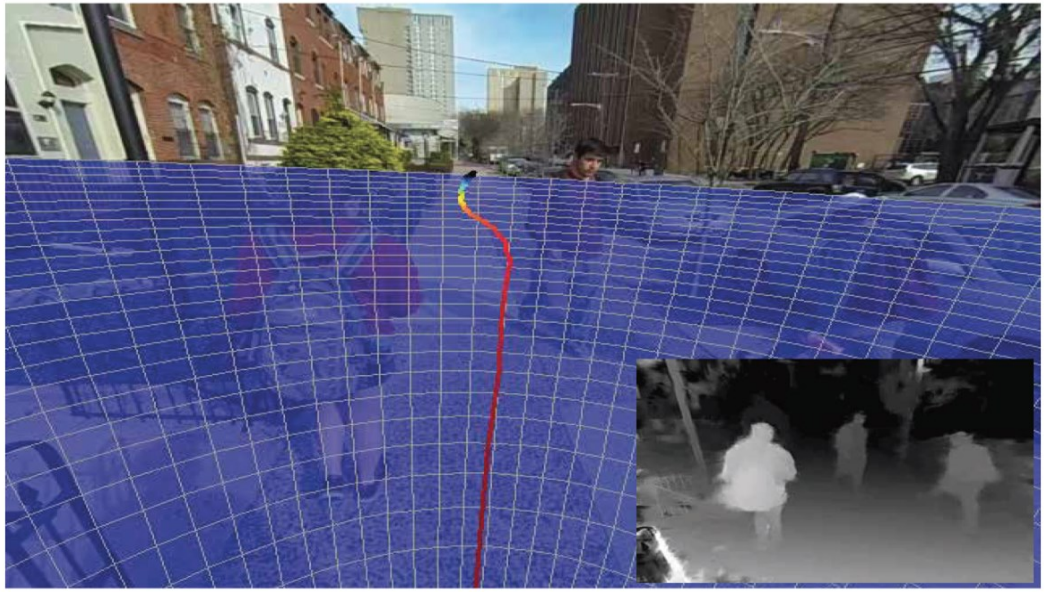
Social

Forecasting

Four Ego4D Forecasting Challenges

Two related Position and Trajectory Prediction

Future Locomotion Movements

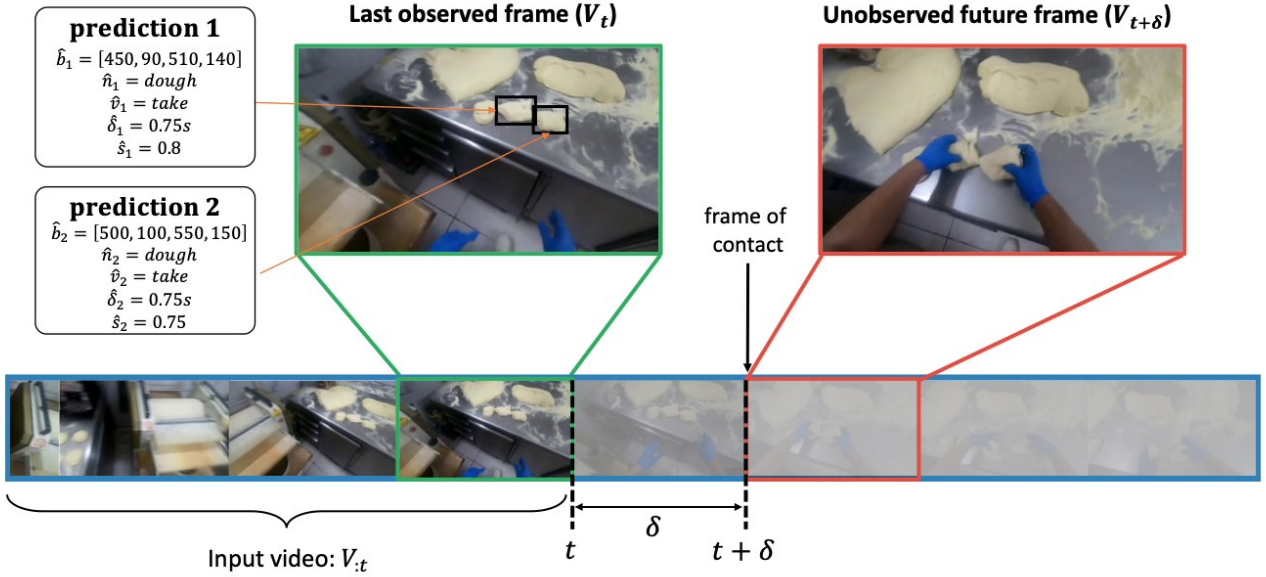


Future Hands Movements

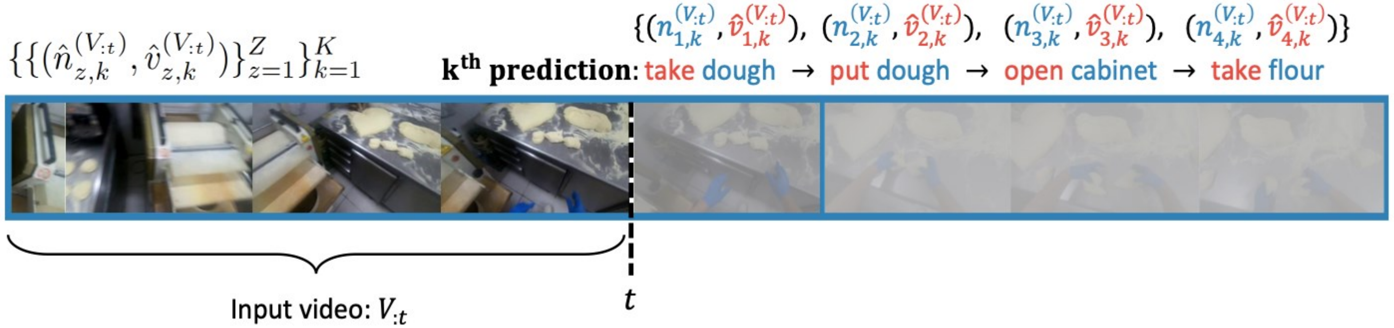
Four Ego4D Forecasting Challenges

Two related Object Interaction Anticipation

Short-Term Anticipation



Long-Term Anticipation



1st Ego4D Workshop @ CVPR 2022

Held in conjunction with 10th EPIC Workshop

19 and 20 June 2022

Overview

In 2022, we will host 16 challenges, representing each of Ego4D's five benchmarks. These are:

Episodic memory:

- Visual queries with 2D localization and VQ 3D localization: Given an egocentric video clip and an image crop depicting the query object, return the last time the object was seen in the input video, in terms of the tracked bounding box (2D + temporal localization) or the 3D displacement vector from the camera to the object in the environment.
- Natural language queries: 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.
- Moments queries: Given an egocentric video and an activity name (e.g., a "moment"), localize all instances of that activity in the past video

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.
- State change object detection: Given an egocentric video clip, identify the objects whose states are changing and outline them with bounding boxes.

Audio-Visual Diarization & Social:

- Audio-visual localization: Given an egocentric video clip, localize the speakers in the visual field of view.
- Audio-visual speaker diarization: Given an egocentric video clip, identify which person spoke and when they spoke.
- Audio-only Diarization Challenge: Given an egocentric video clip, identify which person spoke and when they spoke based on audio alone.
- Speech transcription: Given an egocentric video clip, transcribe the speech of each person.
- 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:

- **Locomotion forecasting**: Given a video frame and the past trajectory, predict the future ego positions of the camera wearer (in the form of a 3D trajectory).
- Hand forecasting: Given a short preceding video clip, predict where the hand will be visible in the future, in terms of a bounding box center in keyframes.
- Short-term hand object prediction: Given a video clip, predict the next active objects, the next action, and the time to contact.
- 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, what will the baker do next?

16 challenges;
Deadline: 1st June;
Winners announced during the workshop.

http://ego4d-data.org/Workshop/CVPR22/

Doing research on First Person Vision now is much easier than in the past!

- Consumer wearable devices;
- Capability to handle huge quantities of data:
  - Hardware (CPUs, GPUs);
  - Deep Learning;
- Industrial interest:
  - Microsoft's HoloLens2;
  - Magic Leap;
  - Google Glass Enterprise Edition;
  - Meta's Project Aria;
- Conferences and workshops on FPV;
  - + many papers on FPV published in top vision conferences (CVPR, ICCV, ECCV);
- Datasets and standard challenges are available.

Take-Home Messages

- Technological advances allowed the creation of efficient platforms for First Person Vision;
- First Person Vision has a great potential for focused application scenarios:
  - Assistive Technologies;
  - Health;
  - Industrial scenarios;
- Big players are moving towards consumer products, with different hardware platform becoming increasingly available;
- It’s a good moment for First Person Vision research, with technology advancing and datasets/challenges attracting the interest of the community.

Question Time

Agenda

ANTONINO

Part I: Definitions, motivations, history and research trends [14.00 - 15.45]

- What is first person vision? What is it for?
- What makes it different from third person vision?
- History of First Person Vision: visions, ideas, research, devices;
- Where do we go from here? Research trends, datasets and challenges.

FRANCESCO

Part II: Building Blocks for First Person Vision Systems [16.15 – 18.00]

- **Data Acquisition & Datasets;**
- **Fundamental Taks in First Person Vision:**
  - **Localization;**
  - **Hand/Object Detection;**
  - **Attention;**
  - **Action/Activities;**
  - **Anticipation**
- **Conclusion**

Non-Exhaustive List of References

- Marr, David. "Vision: A computational investigation into the human representation and processing of visual information. MIT Press." Cambridge, Massachusetts (1982).
- Takeo Kanade and Martial Herbert. "First-person vision." Proceedings of the IEEE 100.8 (2012): 2442-2453.
- Bush, Vannevar. "As we may think." The atlantic monthly 176.1 (1945): 101-108.
- WearCam Website. http://wearcam.org/. Accessed 04/02/2019.
- Steve Mann, "Wearable computing: a first step toward personal imaging," in Computer, vol. 30, no. 2, pp. 25-32, Feb. 1997.

Non-Exhaustive List of References

- Steve Mann. "Compositing multiple pictures of the same scene." Proc. IS&T Annual Meeting, 1993.
- Starner, Thad, et al. "Augmented reality through wearable computing." Presence: Teleoperators &Starner, T., Mann, S., Rhodes, B., Levine, J., Healey, J., Kirsch, D., ... & Pentland, A. (1997). Augmented reality through wearable computing. Presence: Teleoperators & Virtual Environments, 6(4), 386-398.
- Starner, T., Schiele, B., & Pentland, A. (1998, October). Visual contextual awareness in wearable computing. In Wearable Computers, 1998. Digest of Papers. Second International Symposium on (pp. 50-57). IEEE.
- Schiele, B., Oliver, N., Jebara, T., & Pentland, A. (1999, January). An interactive computer vision system dypers: Dynamic personal enhanced reality system. In International Conference on Computer Vision Systems (pp. 51-65). Springer, Berlin, Heidelberg.
- Mayol, W. W., Tordoff, B. J., & Murray, D. W. (2002). Wearable visual robots. Personal and Ubiquitous Computing, 6(1), 37-48.

Non-Exhaustive List of References

- Torralba, A., Murphy, K. P., Freeman, W. T., & Rubin, M. A. (2003, October). Context-based vision system for place and object recognition. In ICCV (Vol. 3, pp. 273-280).
- Davison, A. J., Mayol, W. W., & Murray, D. W. (2003, October). Real-time localization and mapping with wearable active vision. In Mixed and Augmented Reality, 2003. Proceedings. The Second IEEE and ACM International Symposium on (pp. 18-27). IEEE.
- Mayol, W. W., & Murray, D. W. (2005, October). Wearable hand activity recognition for event summarization. In Wearable Computers, 2005. Proceedings. Ninth IEEE International Symposium On (pp. 122-129). IEEE.
- Spriggs, Ekaterina H., Fernando De La Torre, and Martial Hebert. "Temporal segmentation and activity classification from first-person sensing." Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference On. IEEE, 2009.
- Ren, Xiaofeng, and Chunhui Gu. "Figure-ground segmentation improves handled object recognition in egocentric video." Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010.

Non-Exhaustive List of References

- Pattie Maes (MIT) @ TED https://www.ted.com/talks/pattie_maes_demos_the_sixth_sense. Accessed 04/02/2019.
- Bell, Gordon, and Jim Gemmell. Your life, uploaded: The digital way to better memory, health, and productivity. Penguin, 2010.
- Project SenseCam. https://www.microsoft.com/en-us/research/project/sensecam/. Accessed 04/02/2019.
- Sellen, A. J., Fogg, A., Aitken, M., Hodges, S., Rother, C., & Wood, K. (2007, April). Do life-logging technologies support memory for the past?: an experimental study using sensecam. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 81-90). ACM.
- Blighe, M., & O'Connor, N. E. (2008, July). MyPlaces: detecting important settings in a visual diary. In Proceedings of the 2008 international conference on Content-based image and video retrieval (pp. 195-204). ACM.

Non-Exhaustive List of References

- Lee, H., Smeaton, A. F., O'Connor, N. E., Jones, G., Blighe, M., Byrne, D., ... & Gurrin, C. (2008). Constructing a SenseCam visual diary as a media process. Multimedia Systems, 14(6), 341-349.
- Jojic, N., Perina, A., & Murino, V. (2010). Structural epitome: a way to summarize one's visual experience. In Advances in neural information processing systems (pp. 1027-1035).
- Gurrin, C., Smeaton, A. F., & Doherty, A. R. (2014). Lifelogging: Personal big data. Foundations and Trends® in Information Retrieval, 8(1), 1-125.
- Perina, A., Mohammadi, S., Jojic, N., & Murino, V. (2017, October). Summarization and classification of wearable camera streams by learning the distributions over deep features of out-of-sample image sequences. In Proceedings of the IEEE International Conference on Computer Vision (pp. 4326-4334).
- Aghaei, M., Dimiccoli, M., & Radeva, P. (2016). Multi-face tracking by extended bag-of-tracklets in egocentric photo-streams. Computer Vision and Image Understanding, 149, 146-156.

Non-Exhaustive List of References

- Dimiccoli, M., Bolaños, M., Talavera, E., Aghaei, M., Nikolov, S. G., & Radeva, P. (2017). Sr-clustering: Semantic regularized clustering for egocentric photo streams segmentation. Computer Vision and Image Understanding, 155, 55-69.
- Bolanos, M., Dimiccoli, M., & Radeva, P. (2017). Toward storytelling from visual lifelogging: An overview. IEEE Transactions on Human-Machine Systems, 47(1), 77-90.
- Kitani, K. M., Okabe, T., Sato, Y., & Sugimoto, A. (2011, June). Fast unsupervised ego-action learning for first-person sports videos. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on (pp. 3241-3248). IEEE.
- Fathi, A., Ren, X., & Rehg, J. M. (2011, June). Learning to recognize objects in egocentric activities. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference On (pp. 3281-3288). IEEE.
- Pirsiavash, H., & Ramanan, D. (2012, June). Detecting activities of daily living in first-person camera views. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on (pp. 2847-2854). IEEE.

Non-Exhaustive List of References

- Fathi, A., Hodgins, J. K., & Rehg, J. M. (2012, June). Social interactions: A first-person perspective. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on (pp. 1226-1233). IEEE.
- Poleg, Y., Arora, C., & Peleg, S. (2014). Temporal segmentation of egocentric videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2537-2544).
- Lee, Yong Jae, Joydeep Ghosh, and Kristen Grauman. "Discovering important people and objects for egocentric video summarization." Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012.
- Lu, Zheng, and Kristen Grauman. "Story-driven summarization for egocentric video." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013.
- Soran, Bilge, Ali Farhadi, and Linda Shapiro. "Generating notifications for missing actions: Don't forget to turn the lights off!." Proceedings of the IEEE International Conference on Computer Vision. 2015.

Non-Exhaustive List of References

- Furnari, A., Farinella, G. M., & Battiato, S. (2017). Recognizing personal locations from egocentric videos. IEEE Transactions on Human-Machine Systems, 47(1), 6-18.
- Land, Michael F. "Eye movements and the control of actions in everyday life." Progress in retinal and eye research 25.3 (2006): 296-324.
- Ogaki, K., Kitani, K. M., Sugano, Y., & Sato, Y. (2012, June). Coupling eye-motion and ego-motion features for first-person activity recognition. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on (pp. 1-7). IEEE.
- Fathi, A., Li, Y., & Rehg, J. M. (2012, October). Learning to recognize daily actions using gaze. In European Conference on Computer Vision (pp. 314-327). Springer, Berlin, Heidelberg.
- Bulling, A., Roggen, D., & Troester, G. (2011). What's in the Eyes for Context-Awareness?. IEEE Pervasive Computing, 10(2), 48-57.

Non-Exhaustive List of References

- Li, Y., Fathi, A., & Rehg, J. M. (2013). Learning to predict gaze in egocentric video. In Proceedings of the IEEE International Conference on Computer Vision (pp. 3216-3223).
- Li, Y., Ye, Z., & Rehg, J. M. (2015). Delving into egocentric actions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 287-295).
- Damen, D., Leelasawassuk, T., Haines, O., Calway, A., & Mayol-Cuevas, W. W. (2014, September). You-Do, I-Learn: Discovering Task Relevant Objects and their Modes of Interaction from Multi-User Egocentric Video. In BMVC (Vol. 2, p. 3).
- Zhang, M., Teck Ma, K., Hwee Lim, J., Zhao, Q., & Feng, J. (2017). Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4372-4381).
- Huang, Y., Cai, M., Li, Z., & Sato, Y. (2018). Predicting Gaze in Egocentric Video by Learning Task-dependent Attention Transition. ECCV 2018.

Non-Exhaustive List of References

- Li, Y., Liu, M., & Rehg, J. M. (2018). In the eye of beholder: Joint learning of gaze and actions in first person video. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 619-635).
- Lee, S., Bambach, S., Crandall, D. J., Franchak, J. M., & Yu, C. (2014). This hand is my hand: A probabilistic approach to hand disambiguation in egocentric video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 543-550).
- Bambach, S., Lee, S., Crandall, D. J., & Yu, C. (2015). Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1949-1957).
- Leelasawassuk, Teesid, Dima Damen, and Walterio Mayol-Cuevas. "Automated capture and delivery of assistive task guidance with an eyewear computer: the GlaciAR system." Proceedings of the 8th Augmented Human International Conference. ACM, 2017.
- The unexpected rebirth of Google Glass. https://medium.com/swlh/the-unexpected-rebirth-of-google-glass-96f6060a62f2. Accessed 04/02/2019.

Non-Exhaustive List of References

- Why Google Glass broke. https://www.nytimes.com/2015/02/05/style/why-google-glass-broke.html. Accessed 04/02/2019.
- Templeman, R., Korayem, M., Crandall, D. J., & Kapadia, A. (2014, February). PlaceAvoider: Steering First-Person Cameras away from Sensitive Spaces. In NDSS (pp. 23-26).
- Ryoo, M. S., Rothrock, B., Fleming, C., & Yang, H. J. (2017). Privacy-Preserving Human Activity Recognition from Extreme Low Resolution. In AAAI (pp. 4255-4262).
- Dimiccoli, M., Marín, J., & Thomaz, E. (2018). Mitigating Bystander Privacy Concerns in Egocentric Activity Recognition with Deep Learning and Intentional Image Degradation. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 1(4), 132.
- Microsoft secures $480 million HoloLens contract from US Army. https://www.theverge.com/2018/11/28/18116939/microsoft-army-hololens-480-million-contract-magic-leap. Accessed 04/02/2019.

Non-Exhaustive List of References

- Sigurdsson, G. A., Gupta, A., Schmid, C., Farhadi, A., & Alahari, K. (2018). Charades-Ego: A Large-Scale Dataset of Paired Third and First Person Videos. arXiv preprint arXiv:1804.09626.
- Damen, D., Doughty, H., Farinella, G. M., Fidler, S., Furnari, A., Kazakos, E., ... & Wray, M. (2018). Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. ECCV 2018.
- Kazakos, Evangelos and Nagrani, Arsha and Zisserman, Andrew and Damen, Dima, EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition, ICCV 2019.
- Sudhakaran, Swathikiran and Escalera, Sergio and Lanz, Oswald, LSTA: Long Short-Term Attention for Egocentric Action Recognition, CVPR 2019.
- Li, Y., Liu, M., & Rehg, J. M., In the eye of beholder: Joint learning of gaze and actions in first person video, ECCV 2018.

Non-Exhaustive List of References

- Tekin, Bugra and Bogo, Federica and Pollefeys, Marc H+O: Unified Egocentric Recognition of 3D Hand-Object Poses and Interactions, CVPR 2019.
- Miao Liu, Siyu Tang, Yin Li, James M. Rehg, Forecasting Human-Object Interaction: Joint Prediction of Motor Attention and Actions in First Person Video, ECCV 2020.
- Nagarajan, Tushar and Li, Yanghao and Feichtenhofer, Christoph and Grauman, Kristen, EGO-TOPO: Environment Affordances from Egocentric Video, CVPR 2020.
- A. Furnari, G. M. Farinella, Rolling-Unrolling LSTMs for Action Anticipation from First-Person Video. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). 2020.
- Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro and Toby Perrett, Will Price, Michael Wray (2020). The EPIC-KITCHENS Dataset: Collection, Challenges and Baselines. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI).

Non-Exhaustive List of References

- F. Ragusa, A. Furnari, S. Livatino, G. M. Farinella. The MECCANO Dataset: Understanding Human-Object Interactions from Egocentric Videos in an Industrial-like Domain. In IEEE Winter Conference on Applications of Computer Vision (WACV) 2021.
- Giovanni Pasqualino, Antonino Furnari, Giovanni Signorello, Giovanni Maria Farinella (2021). An Unsupervised Domain Adaptation Scheme for Single-Stage Artwork Recognition in Cultural Sites. Image and Vision Computing.
- Siddhant Bansal, Awesome Egocentric Vision GitHub repository, 2021. https://github.com/Sid2697/awesome-egocentric-vision