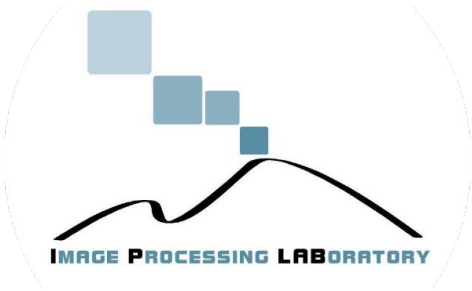
Università
di Catania



Machine Learning

Giovanni Maria Farinella

Department of Mathematics and Computer Science - University of Catania

gfarinella@dmi.unict.it

http://www.dmi.unict.it/farinella/

FPV@IPLAB - http://iplab.dmi.unict.it/fpv

Lecture 1

- Docente e Tutor
- ML @ DMI
- Struttura del Corso di ML
- Informazioni su esame e progetto
- Perchè ML è interessante?
- Applicazioni ed esempi
- Concetti Base

www.dmi.unict.it/farinella

GIOVANNI MARIA FARINELLA

Department of Mathematics and Computer Science - University of Catania

Curriculum Vitae et Studiorum
Request an Updated CV

E-Mail
giovanni.farinella@unict.it

Mobile
Request Mobile Number

Office
+39 095 738 3205

Fax
+39 095 330094
(mark fax ATTN. G.M. Farinella)

BOOKS

Computer Vision - ECCV 2022

Pattern Recognition. ICPR International Workshops and Challenges (link)

Book on CV for Assistive Healthcare (link)

Book on Registration and Recognition (link)

Giovanni Maria Farinella obtained the degree in Computer Science (egregia cum laude) from the University of Catania, Italy, in 2004. He is Founder Member of the IPLAB Research Group at University of Catania since 2005. He also became an Associate Member of the Computer Vision and Robotics Research Group at University of Cambridge in 2006, and Associate Member of the Italian National Research Council since 2018. He was awarded a Doctor of Philosophy (Computer Vision) from the University of Catania in 2008. He is Full Professor at the Department of Mathematics and Computer Science, University of Catania, Italy. His research interests lie in the fields of Computer Vision, Pattern Recognition and Machine Learning, with focus on First Person (Egocentric) Vision. He is Associate Editor of the international journals IEEE Transactions on Pattern Analysis and Machine Intelligence, Pattern Recognition - Elsevier , International Journal of Computer Vision. He has been serving as Area Chair for CVPR 2020/21/22/23, ICCV 2017/19/21/23, ECCV 2020, BMVC 2020, WACV 2019/22, ICPR 2018, and as Program Chair of ECCV 2022, ICIAP 2021 and VISAPP 2019/20/21/22/23. Giovanni Maria Farinella founded (in 2006) and currently directs the International Computer Vision Summer School. He also founded (in 2014) and currently directs the Medical Imaging Summer School. He is member of the European Laboratory for Learning and Intelligent Systems (ELLIS), Senior Member of the IEEE Computer Society, Scientific Advisor of the NVIDIA AI Technology Centre (NVAITC), and board member of the CINI Laboratory of Artificial Intelligence and Intelligent Systems (lead of the area AI for Industry - since 2021). He was awarded the PAMI Mark Everingham Prize 2017 and the Intel's 2022 Outstanding Researcher Award. In addition to academic work, Giovanni's industrial experience includes scientific advisorship to different national and international companies and startups, as well as the leadership as Founder and Chief Scientific Officer of Next Vision - Spinoff of the University of Catania.

Scopus Profile, **Google Scholar Profile**

Giovanni Maria Farinella has got the National Scientific Qualification to Associate Professor (ASN - Seconda Fascia) in the areas of Information Engineering (SSD 09/H1) and Computer Science (SSD 01/B1) on 03/12/2013 and 29/01/2014 respectively.

He has got the National Scientific Qualification to Full Professor (ASN - Prima Fascia), in the areas of Computer Science (SSD 01/B1) and Information Engineering (SSD 09/H1) on 28/03/2018 and 26/07/2018 respectively.

Image Processing Laboratory

First Person (Egocentric) Vision @ IPLAB

NEXT VISION

Spin-off of the University of Catania


COURSES


SEMINARS & PRESENTATIONS


PUBLICATIONS


RESEARCH VISITS


PROJECTS


PROFESSIONAL ACTIVITIES


PICTURES


LINKS & EVENTS

NEWS

- **Are you looking for my help? Send me an email to fix a meeting. Do not hesitate!**
- **Research Scholarship Opportunities**

SCHOLARSHIPS

- **Ph.D. Opportunities**

PhD

- **PostDoctoral Opportunities**

POSTDOCTORAL

- **Reading Group @ IPLAB**

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

FIRST PERSON (EGOCENTRIC) VISION @ IPLAB

ABOUT IPLAB

The Image Processing Laboratory (IPLAB) is part of the Department of Mathematics and Computer Science of the University of Catania, Italy. IPLAB's research focuses in the areas of Image Processing, Computer Vision, Machine Learning and Computer Graphics.

LEARN MORE ABOUT IPLAB

CONTACT PERSON FOR RESEARCH ON FIRST PERSON (EGOCENTRIC) VISION

Prof. Giovanni Maria Farinella - gfarinella@dmi.unict.it

HOME

LAST UPDATE: OCT 5, 2020

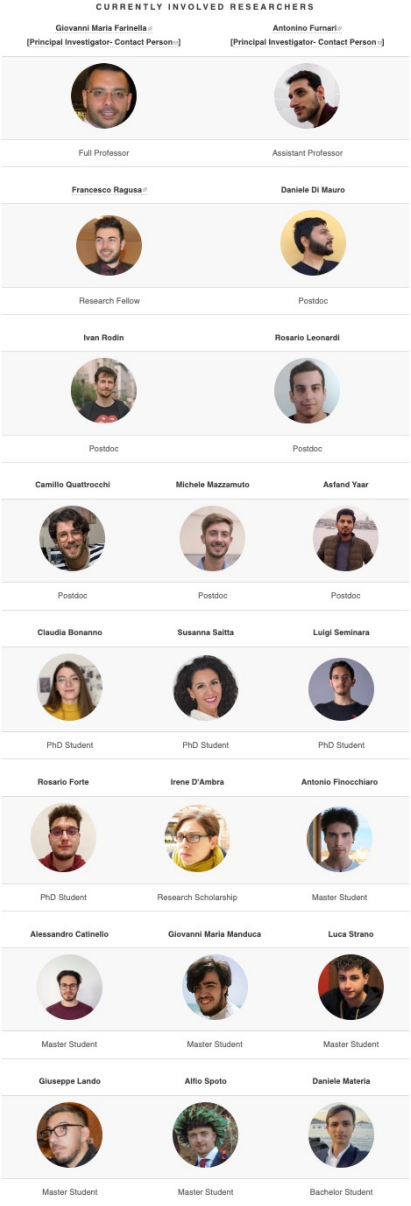
FIRST PERSON (EGOCENTRIC) VISION @ IPLAB

The terms *First Person Vision* or *Egocentric Vision* refer to the study and development of Computer Vision techniques in the scenario in which images and video are acquired from the user's point of view. This is generally done employing *wearable cameras* such as Google Glass, Microsoft HoloLens and GoPro. This acquisition paradigm is in contrast with standard *Third Person Vision* applications which assume that images are acquired by fixed cameras.

Visual content acquired according to the *First Person Vision* paradigm is inherently different from standard *Third Person Vision Content*. Indeed, while a fixed camera observes event from a *neutral* point of view, egocentric visual content captures the personal visual experience of the camera wearer. While Third Person Vision content is often edited and pre-segmented (e.g., movies or collections of YouTube videos), egocentric video is generally acquired in a continuous fashion and hence it tends to be unstructured and difficult to index. While strong assumptions can be generally made on Third Person Visual content, First Person data is inherently characterized by a continuously changing context which must be dealt with.

With its intrinsic mobility, *First Person Vision* poses some new challenges (e.g., changing context, motion blur, unstructured content), and offers unique opportunity to develop true intelligent systems able to assist the user and augment his abilities.

The aim of this page is to present research done by the Image Processing LABoratory(IPLAB) in the scope of First Person Vision.



Rosario Leonardi

FOLLOW

University of Catania
Verified email at unict.it

Computer Vision Machine Learning Egocentric Vision

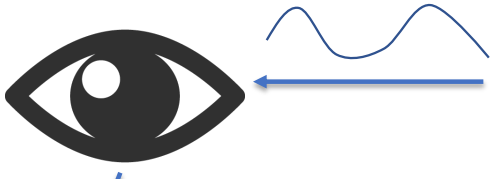
TITLE	CITED BY	YEAR
Egocentric Human-Object Interaction Detection Exploiting Synthetic Data R Leonardi, F Ragusa, A Furnari, GM Farinella Image Analysis and Processing–ICIAP 2022: 21st International Conference …	14	2022
Vedi: Vision exploitation for data interpretation GM Farinella, G Signorello, S Battiato, A Furnari, F Ragusa, R Leonardi, ... Image Analysis and Processing–ICIAP 2019: 20th International Conference …	14	2019
Exploiting multimodal synthetic data for egocentric human-object interaction detection in an industrial scenario R Leonardi, F Ragusa, A Furnari, GM Farinella Computer Vision and Image Understanding 242, 103984	13	2024
Enigma-51: Towards a fine-grained understanding of human behavior in industrial scenarios F Ragusa, R Leonardi, M Mazzamuto, C Bonanno, R Scavo, A Furnari, ... Proceedings of the IEEE/CVF Winter Conference on Applications of Computer …	9 *	2024
Are Synthetic Data Useful for Egocentric Hand-Object Interaction Detection? R Leonardi, A Furnari, F Ragusa, GM Farinella European Conference on Computer Vision, 36-54	3 *	2025
HERO: An Artificial Conversational Assistant to Support Humans in Industrial Scenarios C Bonanno, F Ragusa, R Leonardi, A Furnari, GM Farinella International Conference on Signal Processing and Multimedia Applications …	2	2022
An AR-Based Tool for Acquisition and Automatic Labeling of Human-Object Interactions From First Person Vision L Seminara, F Ragusa, R Leonardi, GM Farinella, A Furnari 2023 IEEE International Conference on Metrology for eXtended Reality …		2023
The UNICT-TEAM Vision Modules for the Mohamed Bin Zayed International Robotics Challenge 2020 S Battiato, L Cantelli, F D'Urso, GM Farinella, L Guarnera, DC Guastella, ... International Conference on Computer Vision and Image Processing, 707-719		2022
Supplementary Material ENIGMA-51: Towards a Fine-Grained Understanding of Human Behavior in Industrial Scenarios F Ragusa, R Leonardi, M Mazzamuto, C Bonanno, R Scavo, A Furnari, ...		

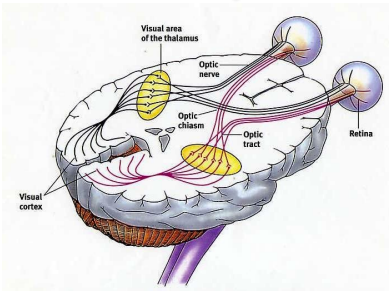
ML @ DMI

Visual Intelligence

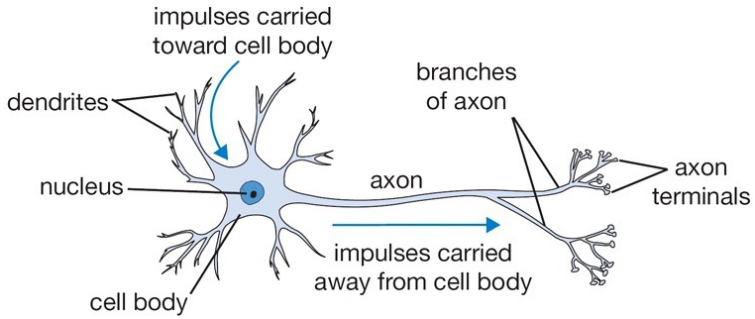
“learning to answer questions about images and videos”

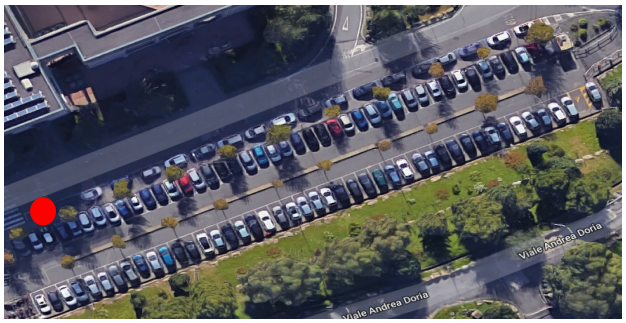
Human Visual System





Human vision involves over 60 billion neurons.





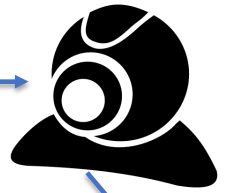
T A S K

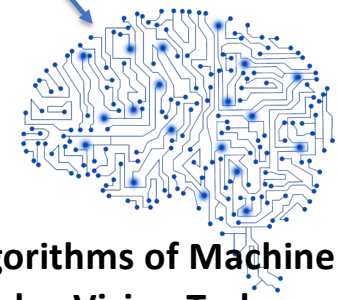
How many cars are in the parking area?

103

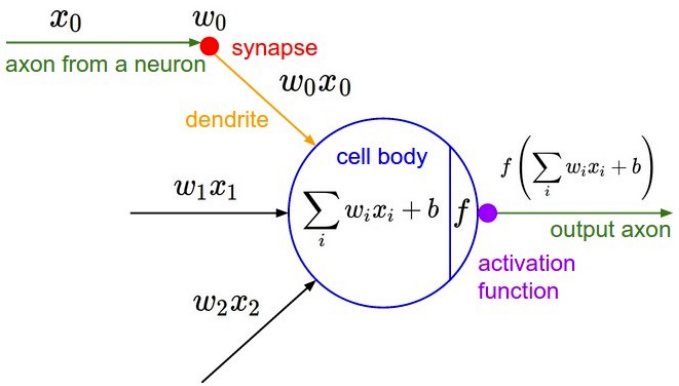
Visual Artificial System

101011001

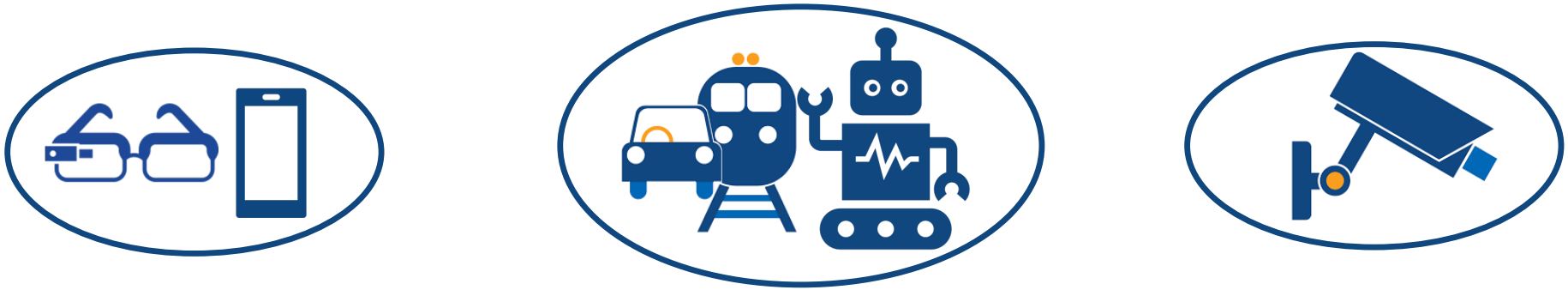




Advanced Algorithms of Machine Learning to Solve Vision Tasks



Visual Intelligence - “Points of Views”



www.dmi.unict.it/farinella - gfarinella@dmi.unict.it

First Person Vision

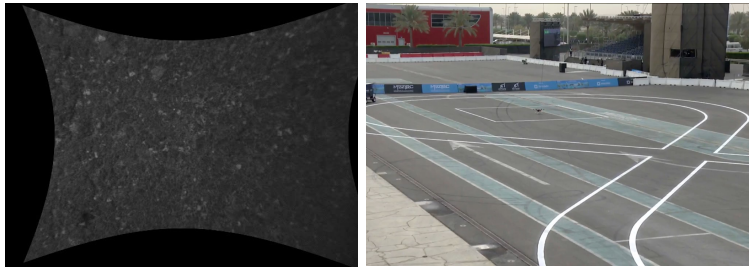
- It refers to systems (hw + sw) that consider **images from the user (or agent) viewpoint** and understand them.
- When available from sensors, information about the **user's head, motion and gaze** (through eye tracking) are also **considered during the interpretation of images**.
- Models and techniques to **understand what a person sees, want to see** (e.g., extra info with AR) or **would like to see** (e.g. in case of visual impairments), from the first person's point of view and **centered on the human perceptual needs (most of the time personal needs)**.
- **Egocentric Vision, Wearable Vision**

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

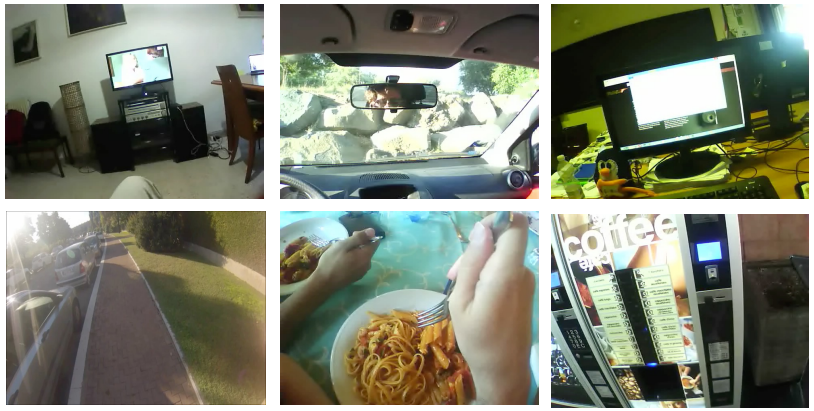
Egocentric (First-Person) Vision

Autonomous — The Human Role and the Point of View — Human

Egocentric Vision for Autonomous Robots



Egocentric Vision for Market Basket Analysis

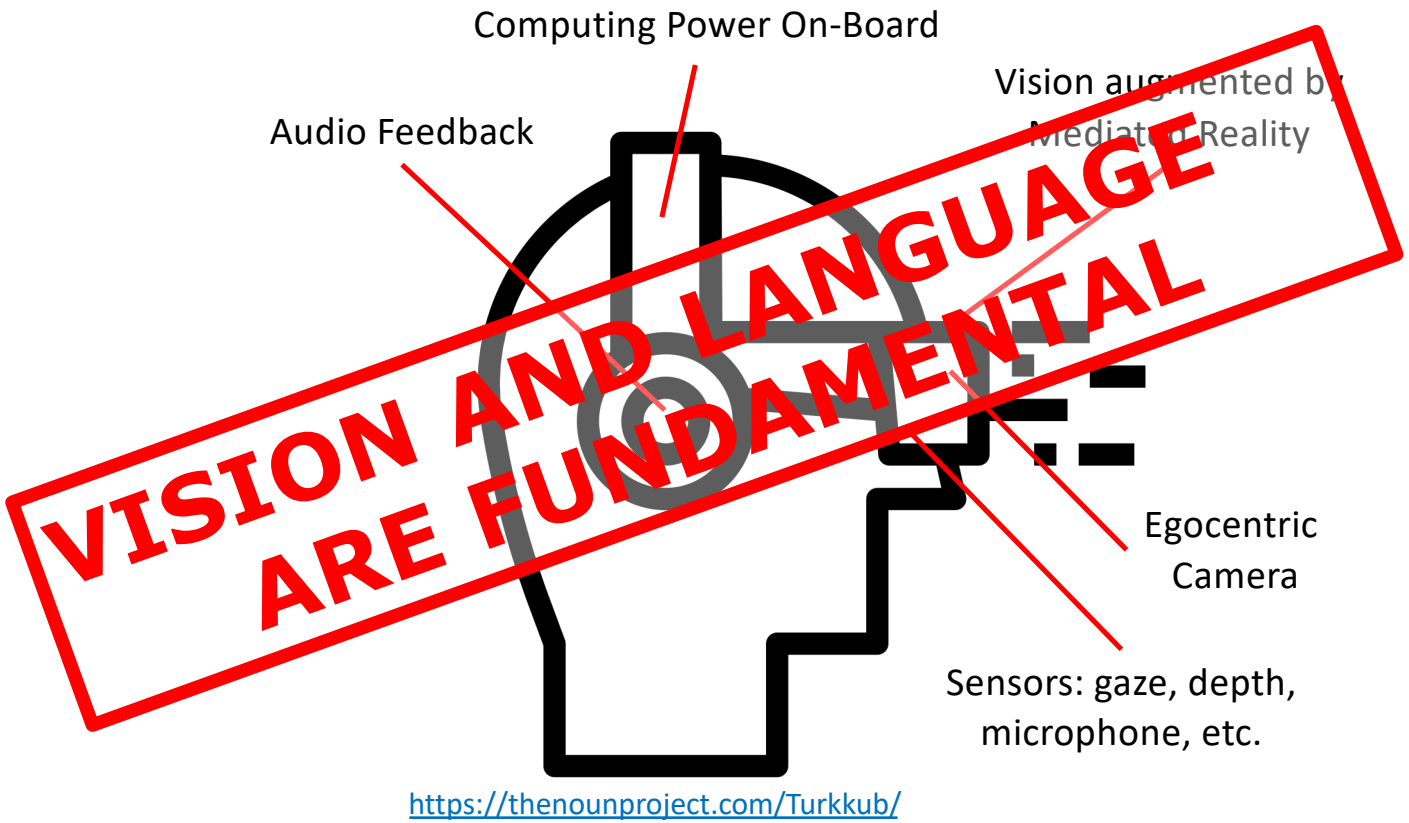


Egocentric Vision to Support the User

- Stress Monitoring and Memory Augmentation
- Summarization and Lifelogging
- Personalized Imaging and Personal Big Data
- Assistive and Quality of Life Applications
- Virtual Assistant
- Object Interaction Prediction
- Action Anticipation
- Behavioural Analysis

www.dmi.unict.it/farinella
gfarinella@dmi.unict.it

An AI-Powered Wearable Virtual Assistant



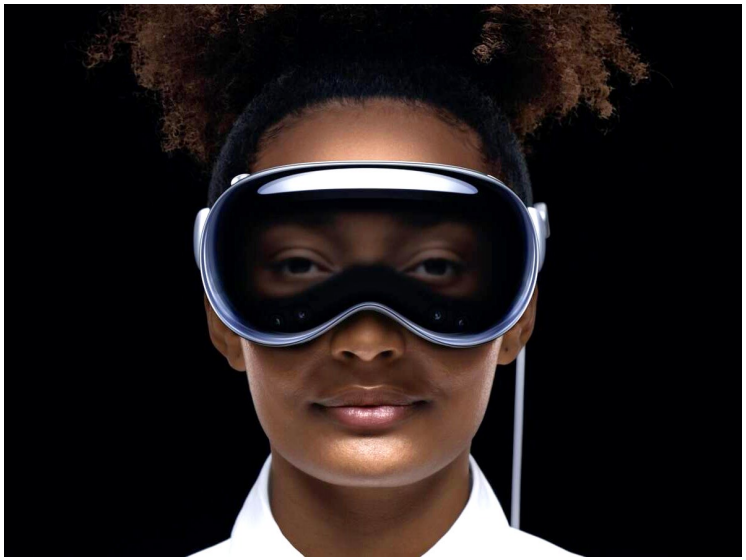
https://thenounproject.com/Turkkub/

A wearable device which perceives the world from our "egocentric" point of view is ideal for implementing an AI-Powered Virtual Assistant

Wearable Devices – Success Examples on the Market

OrCam MyEye
Available since 2015
Health, assistive technologies

Microsoft HoloLens
Available since 2016
Mixed Reality



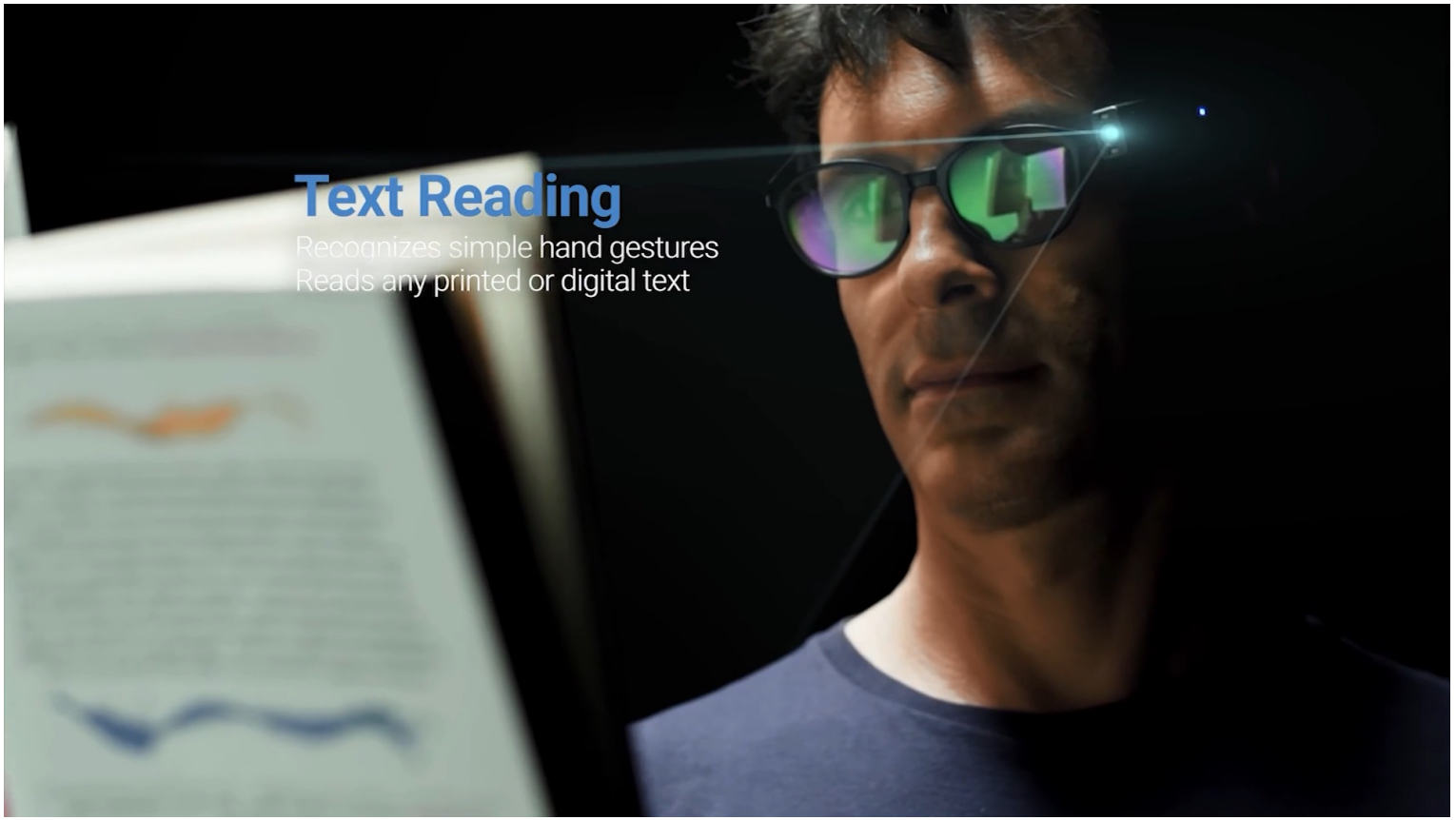
Apple Vision Pro
Pre-order since 2024
Mixed Reality

OrCam MyEye, since 2015

Health, assistive technologies

https://www.orcam.com/

OrCam MyEye, since 2015



https://www.orcam.com/

Project ARIA

Meta

Università di Catania — Egocentric Vision Research – Before EPIC-Kitchens Era

Egocentric Vision has a long research history…

Augmented Reality Through Wearable Computing — 1997
Thad Starner, Steve Mann, Bradley Rhodes, Jeffrey Levine, Jennifer Healey, Dana Kirsch, Roz Picard, and Alex Pentland
The Media Laboratory, Massachusetts Institute of Technology
(augmented reality)

Visual Contextual Awareness in Wearable Computing — 1998
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 — 1999
Bernt Schiele, Nuria Oliver, Tony Jebara, and Alex Pentland
Vision and Modeling Group, MIT Media Laboratory, Cambridge, MA 02139, USA
(object recognition, media memories)

Wearable Visual Robots — 2000
W.W. Mayol, B. Tordoff and D.W. Murray
University of Oxford, Parks Road, Oxford OX1 3PJ, UK
(active vision)

Context-based vision system for place and object recognition — 2003
Antonio Torralba, Kevin P. Murphy, William T. Freeman, Mark A. Rubin
MIT AI lab, Cambridge, MA 02139 / Lincoln Labs, Lexington, MA 02420
(location/object recognition)

Real-Time Localisation and Mapping with Wearable Active Vision — 2003
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)

Do Life-Logging Technologies Support Memory for the Past? An Experimental Study Using SenseCam — 2007
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)

MyPlaces: Detecting Important Settings in a Visual Diary — 2008
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 — 2008
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 retrie

Fast Unsupervised Ego-action Learning for First-Person Sports Videos — 2011
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)

Learning to Recognize Daily Actions using Gaze — 2012
Alireza Fathi, Yin Li, James M. Rehg
(detection and recognition of social interactions)

Learning Hand Activity Models — 2005
W.W. Mayol, Department of Computer Science, University of Bristol
D.W. Murray, Department of Engineering Science, University of Oxford
(hand activity recognition)

Figure-Ground Segmentation Improves Handled Object Recognition in Egocentric Video — 2010
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)

Story-Driven Summarization for Egocentric Video — 2013
Zheng Lu and Kristen Grauman, University of Texas at Austin
(egocentric video sumarization)

Learning to Predict Gaze in Egocentric Video — 2012
Yin Li, Alireza Fathi, James M. Rehg
(gaze prediciton, action recognition)

You-Do, I-Learn: Egocentric unsupervised discovery of objects and their modes of interaction towards video-based guidance — 2016
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 — 2023
Francesco Ragusa*, Antonino Furnari, Giovanni Maria Farinella
(gaze prediciton, procedural video)

Temporal Segmentation of Egocentric Videos — 2014
Yair Poleg, Chetan Arora*, Shmuel Peleg
(egocentric video indexing)

Recognizing Personal Locations from Egocentric Videos — 2016
Antonino Furnari, Giovanni Maria Farinella, *Senior Member, IEEE*, and Sebastiano Battiato, *Senior Member, IEEE*
(localization, indexing, context-aware computing)

Egocentric Future Localization — 2016
Hyun Soo Park, Jyh-Jing Hwang, Yedong Niu, Jianbo Shi
(future localization, navigation)

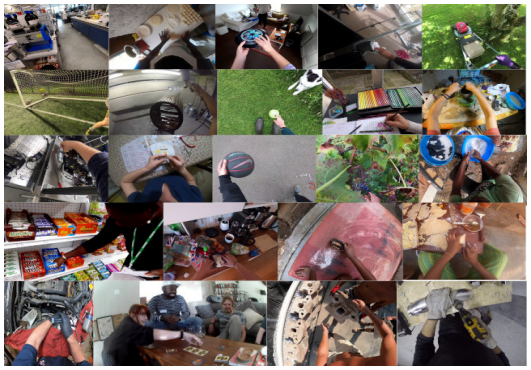
Multi-face tracking by extended bag-of-tracklets in egocentric photo-streams — 2016
Maedeh Aghaei[a,*], Mariella Dimiccoli[a,b], Petia Radeva[a,b]
(lifelogging, face tracking)

SR-clustering: Semantic regularized clustering for egocentric photo streams segmentation — 2017
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)

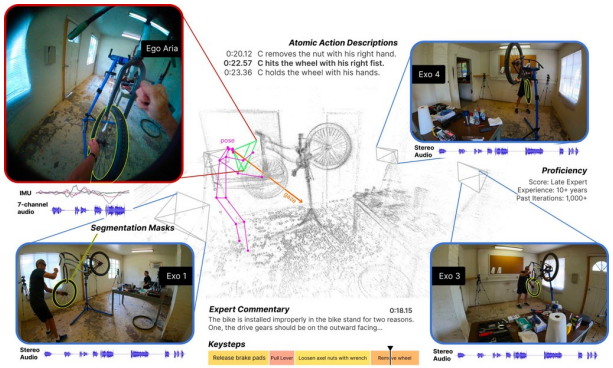
Toward Storytelling From Visual Lifelogging: An Overview — 2017
Marc Bolaños, Mariella Dimiccoli, and Petia Radeva
(lifelogging, survey)

Game Changer - Datasets and Benchmark Suites

EPIC-KITCHENS 55/100
Available Since 2018



EGO4D
Available Since 2021



Ego-Exo4D
Available Since 2023

- Damen, Doughty, Farinella, Fidler, Furnari, Kazakos, Moltisanti, Munro, Perrett, Price, Wray, Scaling Egocentric Vision: The EPIC-KITCHENS Dataset, European Conference on Computer Vision, 2018
- Damen, Doughty, Farinella, Fidler, Furnari, Kazakos, Moltisanti, Munro, Perrett, Price, Wray, The EPIC-KITCHENS Dataset: Collection, Challenges and Baselines, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 2021
- Damen, Doughty, Farinella, Furnari, Kazakos, Ma, Moltisanti, Munro, Perrett, Price, Wray, Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100, International Journal on Computer Vision (IJCV), 2022
- Grauman et al., Around the World in 3,000 Hours of Egocentric Video, IEEE/CVF International Conference on Computer Vision and Pattern Recognition, 2022
- Grauman et al., Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives, CVPR 2024

Some goals of my goup

- Algorithms to understand users' intent and actions performed with hands and feet
  - Localization of the agent and recognition of environment
  - Recognition of the objects observed by the users
  - Recognition/Anticipation of the human-object Interactions
  - Recognition/Anticipation of the human actions
  - Exploiting Synthetic Data for Learning
  - Formalise representations that can be useful for long-term understanding
  - Knowledge Transfer
  - :
  - :
- Application Domains – Binding with Industry
  - Home Assisted Living
  - Retail
  - Cultural Heritage
  - Industrial Environments
- Lack of benchmarks → need to push the community working on the field
  - no common goals and tasks/no common platform, standard/Datasets

www.dmi.unict.it/farinella - gfarinella@dmi.unict.it

Problem

How to represent egocentric videos for long-term understanding?

- **Understand sequences of activities performed by the camera wearer** in different physical locations
- **Egocentric video is by its own nature long-form**
- Egocentric vision systems require algorithms able to **represent and process video over temporal spans that last in the order of minutes or hours**
- Examples of **applications** are **action anticipation, video summarization, and episodic memory retrieval**
- **Lack of a comprehensive and long-form representation of videos** that algorithms can rely on
- **Popular highlevel human-gathered representations** being in the form of textual narrations, verb-noun action labels, temporal bounds for action segments, object bounding boxes, object state changes, and hand-object interaction states, **are all short-range representations describing temporal spans lasting few seconds.**

Egocentric Action Scene Graphs (EASGs)

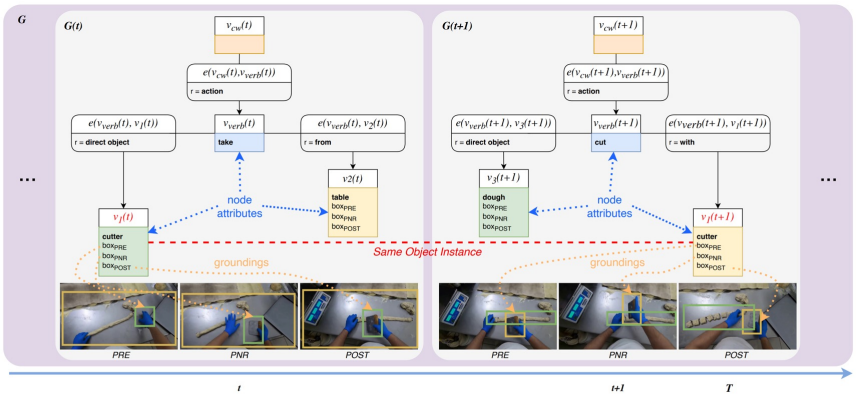


Figure 2. An Egocentric Action Scene Graph (EASG) is a time-varying directed graph $G(t) = (V(t), E(t))$, where nodes $V(t)$ represent either the camera wearer ($v_{cw}(t)$), the action verb ($v_{verb}(t)$), or the involved objects. Edges $E(t)$ represent relationships $e(v_i(t), v_j(t))$ between node pairs. Each node, except for the CW node, can have one or more attributes $att(v_j(t))$ (indicated in blue). Each object has three grounding bounding boxes in the PRE, PNR and $POST$ frames (highlighted in orange). Nodes v_j representing the same object instance maintain the same index across different timesteps (e.g., $v_1(t)$ and $v_1(t+1)$ highlighted in red).

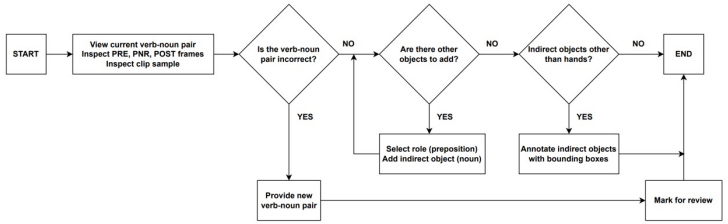


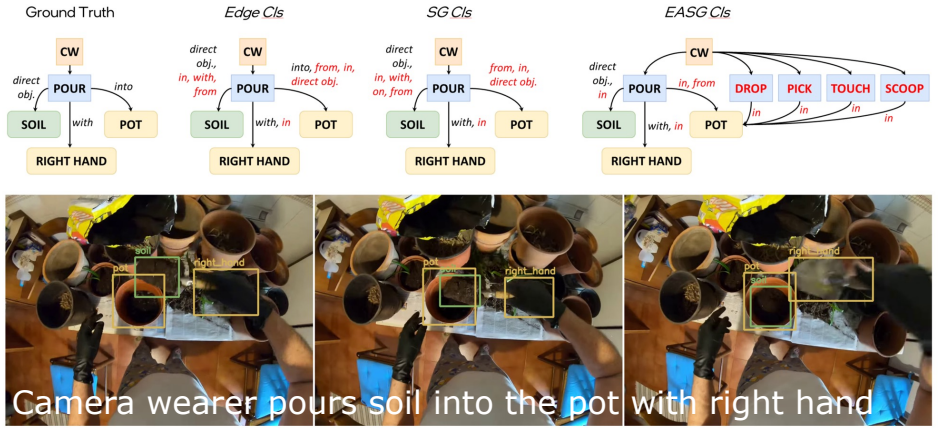
Figure 3. The Ego4D-EASG annotation pipeline. The annotators first review the provided verb-noun pair, the PRE, PNR, $POST$ frames and a clip sampled around PNR. They then check the existing narration and add indirect objects and related groundings, if necessary.

Validation procedure	Example Questions (answers in red)
1. Filtering verb-noun pair	Does CW *take bowl* or *press dough*? take bowl
2. Selecting proper preposition in case of multiple edges between two nodes	Select the preposition which is more appropriate: • CW takes bowl **with** left hand ✓ • CW takes bowl **on** left hand
3. Selecting hand(s) if there are different hands with the same preposition	Does CW take bowl with right hand, with left hand or with both hands? left hand
4. Identifying spatial relations	Is the following statement correct: • The bowl is with flour [Y/N] Y • The bowl is from scale [Y/N] N

Example Frame

Table 1. Examples of questions (with correct answers in red) asked to the annotators in the validation stage to resolve ambiguities between the labels provided in the annotation stage.

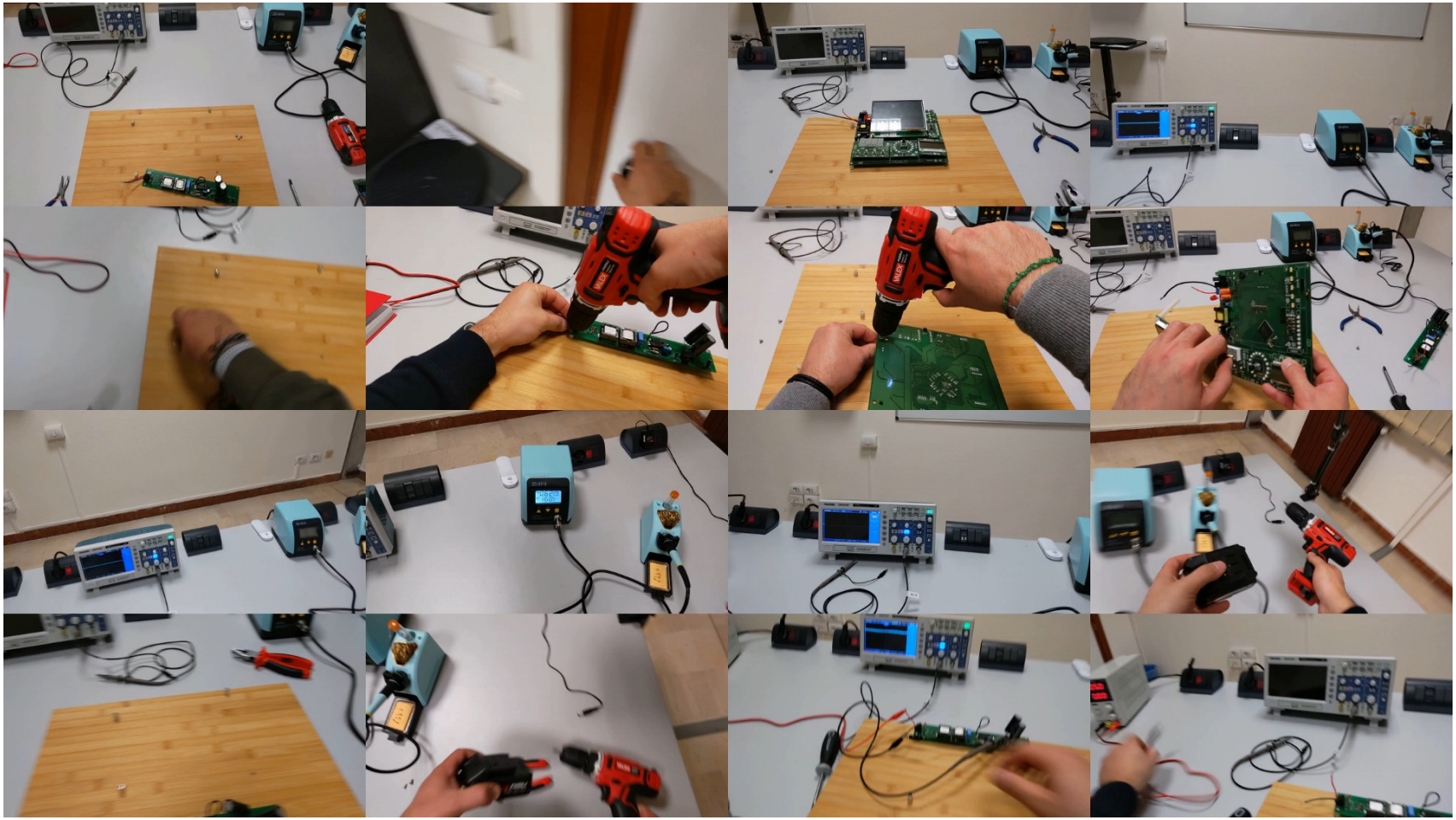
- EASGs Formal Definition
- Novel annotation procedure for Egocentric Videos
- Extension of Ego4D dataset by adding manually labeled Egocentric Action Scene Graphs offering a rich set of annotations designed for long-from egocentric video understanding
- Benchmark for the task of EASG Generation
- Proof of effectiveness of EASGs on two downstream tasks: egocentric action anticipation and egocentric activity summarization
- We will release the dataset to the community as well as the code to replicate experiments: https://github.com/fpv-iplab/EASG



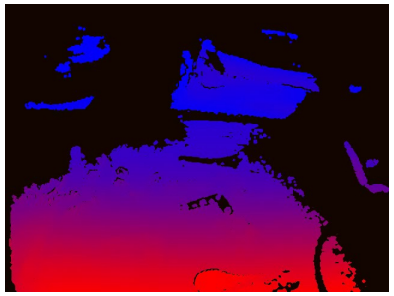
Ivan Rodin, Antonino Furnari, Kyle Min, Subarna Tripathi, Giovanni Maria Farinella, Action Scene Graphs for Long-Form Understanding of Egocentric Videos, CVPR 2024, Arxiv: https://arxiv.org/abs/2312.03391

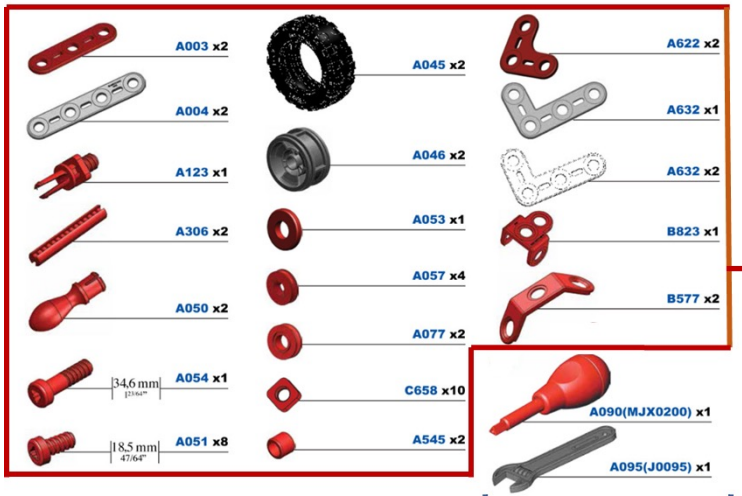
Problem

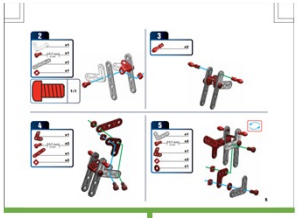
Given an egocentric video the goal is to perform detection, spatial localization and recognition of the objects to which the user is interacting with over time



Università di Catania MECCANO Multimodal Dataset and Benchmark







COMPONENTS


20 Subjects


3 Modalities


20 min. avg. Video length


5 Tasks


8858 Segments


64349 Boxes


20 Objects

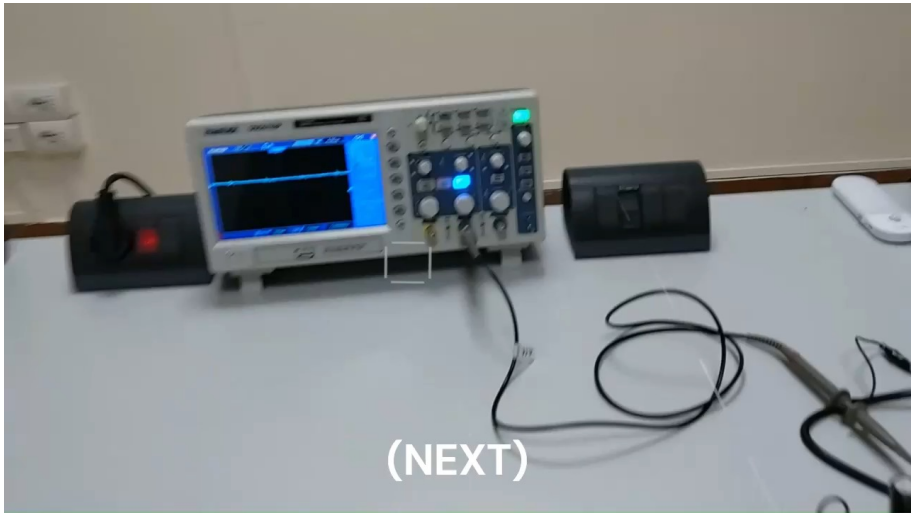

12 Verbs

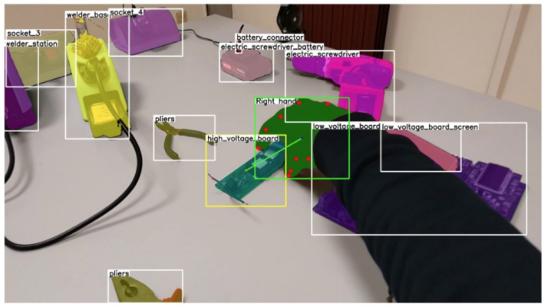

61 Actions

MECCANO Multimodal Dataset, Challenges and Baselines are available at the following link: https://iplab.dmi.unict.it/MECCANO

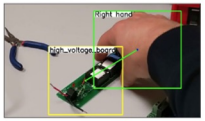
F. Ragusa, A. Furnari, G. M. Farinella. MECCANO: A Multimodal Egocentric Dataset for Humans Behavior Understanding in the Industrial-like Domain. Computer Vision and Image Understanding, 2023

Università di Catania ENIGMA-51 Dataset and Benchmark






Hand-Object boxes

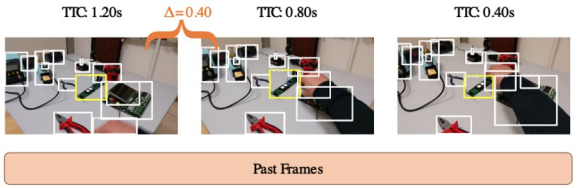

Human-Object Interactions

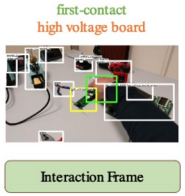

Hand-Object Masks


Hand Keypoints


Environment 3D Model


Object 3D Models





Procedure :

.....

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

.....

22. Turn on the welder using the switch on the corresponding socket (second from right)
23. Set the temperature of the welder to 480 °C using the yellow "UP" button

.....

- **51 egocentric videos** with a resolution of 2272x1278 pixels at 30 fps
- **19 subjects**
- **22 hours** of video recordings
- **45505 RGB frames**
- **2 procedures** consisting of text instructions that involve humans interacting with the objects
- **14036 interactions** temporally annotated indicating the verbs which describe the actions performed
- **275135 objects** and **56473 hands** annotated with bounding boxes
- **12597 interaction frames** annotated with **14036 interactions** and **9342 active objects**
- **37314 next-object interactions** annotated in past frames
- **4 basic verb classes**, and **25 objects classes**
- **3D models** of the environment and the objects

Untrimmed temporal detection of human-object interactions

Egocentric human-object interaction detection

Short-term object interaction anticipation

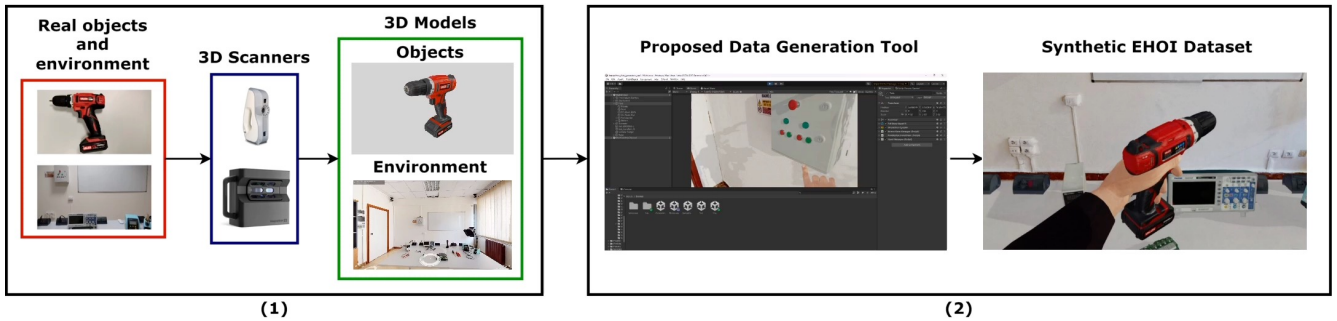
Natural language understanding of intents and entities

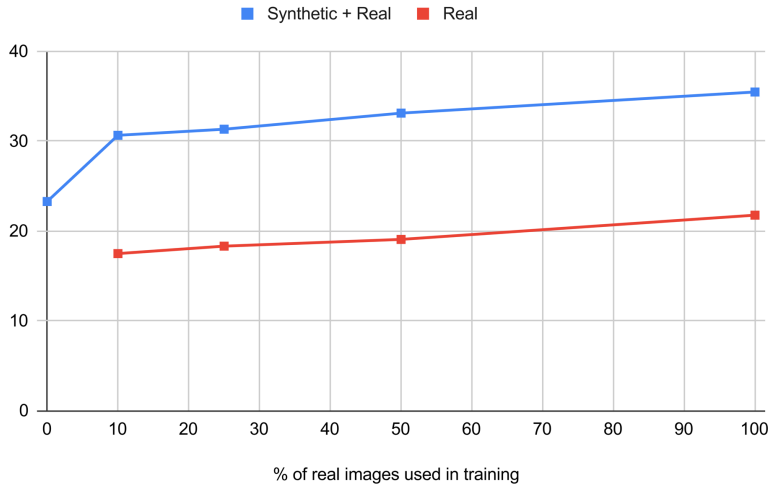
ENIGMA-51 Dataset is available at the following link: https://iplab.dmi.unict.it/ENIGMA-51/

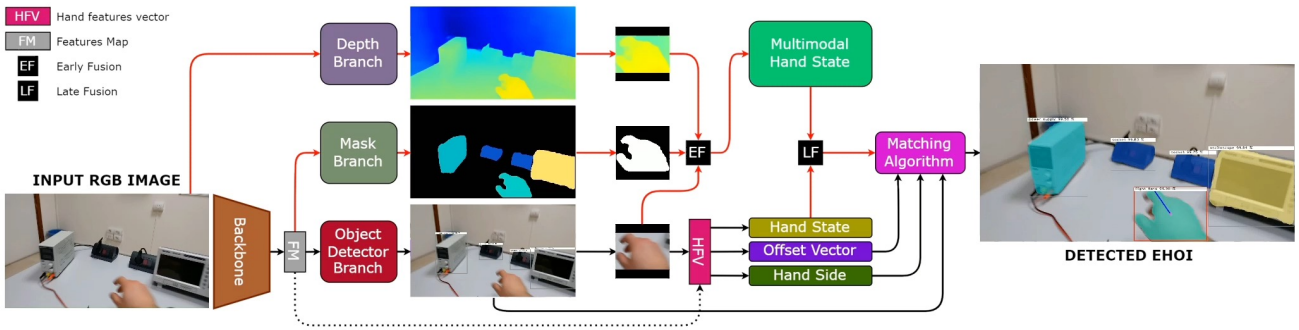
F. Ragusa, R. Leonardi, M. Mazzamuto, C. Bonanno, R. Scavo, A. Furnari, G. M. Farinella. ENIGMA-51: Towards a Fine-Grained Understanding of Human-Object Interactions in Industrial Scenarios, IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024

Are Synthetic Data Useful for Egocentric HOI Detection?

How to provide robust models cutting down labelling time and costs?







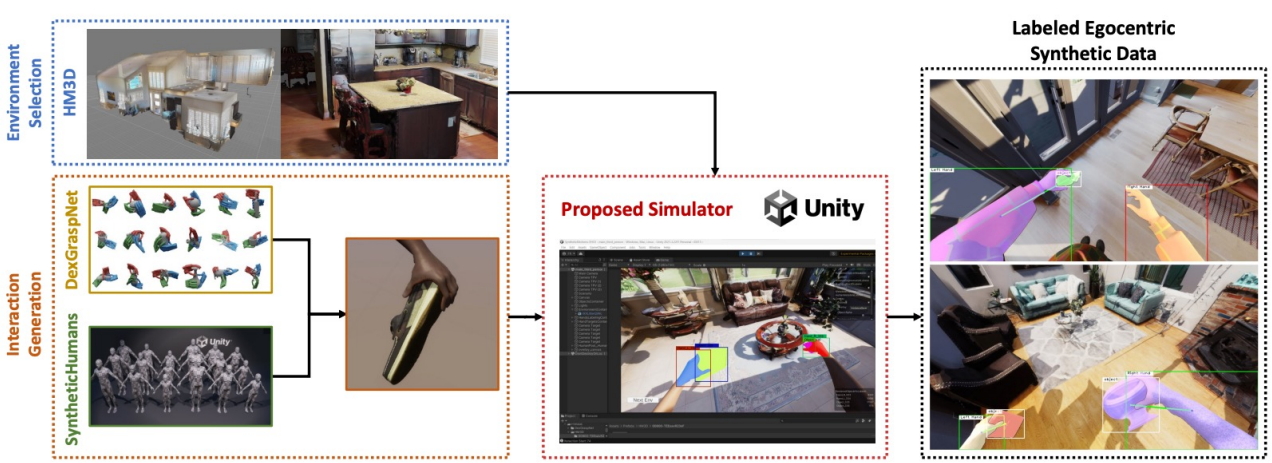
Data to replicate experiments are publicly available:
https://iplab.dmi.unict.it/egoism-hoi/

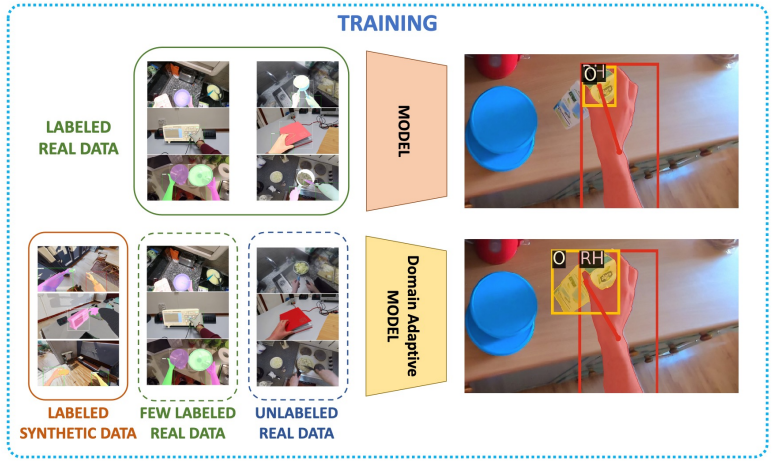
R. Leonardi, F. Ragusa, A. Furnari, G. M. Farinella, Exploiting Multimodal Synthetic Data for Egocentric Human-Object Interaction Detection in an Industrial Scenario, arXiv preprint arXiv:2306.12152, 2023

Università di Catania

Are Synthetic Data Useful for Egocentric HOI Detection?

How to provide robust models cutting down labelling time and costs?





EPIC-KITCHENS VISOR

% Real Labeled Data	Approach	Hand	Hand+Side	Hand+Contact	Object	Hand+Object
0%	Synthetic-Only	28.41	24.89	08.64	01.23	09.88
	UDA	**80.16**	**65.98**	**33.47**	**8.35**	**33.33**
Improvement with respect to Synthetic-Only		**+51.75**	**+41.09**	**+24.83**	**+7.12**	**+23.45**
10% (3,286 images)	Real-Only	87.45	**83.27**	51.98	19.47	38.55
	Synthetic+Real	86.39	82.85	**52.25**	**23.03**	37.62
	SSDA	**89.05**	80.77	46.83	20.41	**44.22**
Improvement with respect to 10% Real-Only		**+1.60**	−0.42	**+0.27**	**+3.56**	**+5.67**
25% (8,215 images)	Real-Only	90.14	**85.66**	53.99	17.85	37.90
	Synthetic+Real	89.98	84.67	**55.88**	18.49	38.19
	SSDA	**90.37**	84.42	52.59	**22.15**	**45.55**
Improvement with respect to 25% Real-Only		**+0.23**	−0.99	**+1.89**	**+4.30**	**+7.65**
50% (16,429 images)	Real-Only	91.16	**86.05**	52.28	17.92	38.15
	Synthetic+Real	**91.34**	85.85	54.09	19.06	43.52
	SSDA	90.94	85.73	**58.02**	**23.49**	**46.47**
Improvement with respect to 50% Real-Only		**+0.18**	−0.20	**+5.74**	**+5.57**	**+8.32**
100% (32,857 images)	Real-Only	**92.25**	88.54	**59.24**	24.23	45.33
	Synthetic+Real	91.45	**88.94**	56.55	**27.77**	44.52
	SSDA	91.83	87.65	57.63	24.03	**46.48**
Improvement with respect to 100% Real-Only		−0.42	**+0.40**	−1.61	**+3.54**	**+1.15**

Ego-HOS

% Real Labeled Data	Approach	Hand	Hand+Side	Hand+Contact	Object	Hand+Object
0%	Synthetic-Only	18.25	15.93	05.33	01.24	07.16
	UDA	**70.30**	**59.21**	**20.84**	**09.65**	**28.16**
Improvement with respect to Synthetic-Only		**+52.05**	**+43.28**	**+15.51**	**+8.41**	**+21.00**
10% (857 images)	Real-Only	76.28	68.92	35.84	16.59	28.44
	Synthetic+Real	77.15	71.64	39.25	17.33	28.74
	SSDA	**83.25**	**73.72**	**47.20**	**22.40**	**36.68**
Improvement with respect to 10% Real-Only		**+6.97**	**+4.80**	**+11.36**	**+5.81**	**+8.24**
25% (2,026 images)	Real-Only	78.94	70.62	41.67	21.83	33.73
	Synthetic+Real	79.60	71.61	46.11	19.87	33.78
	SSDA	**83.79**	**74.28**	**49.00**	**23.82**	**37.16**
Improvement with respect to 25% Real-Only		**+4.85**	**+3.66**	**+7.33**	**+1.99**	**+3.43**
50% (4,379 images)	Real-Only	81.82	73.63	47.27	25.73	36.30
	Synthetic+Real	82.54	74.03	47.92	23.47	34.30
	SSDA	**85.17**	**76.80**	**52.58**	**26.90**	**39.85**
Improvement with respect to 50% Real-Only		**+3.97**	**+3.17**	**+5.31**	**+1.17**	**+3.55**
100% (8,758 images)	Real-Only	84.39	76.24	51.81	26.46	36.16
	Synthetic+Real	84.56	71.56	49.72	23.16	34.68
	SSDA	**85.58**	**76.80**	**51.99**	**27.05**	**39.61**
Improvement with respect to 100% Real-Only		**+1.19**	**+0.56**	**+0.18**	**+0.59**	**+3.45**

ENIGMA-51

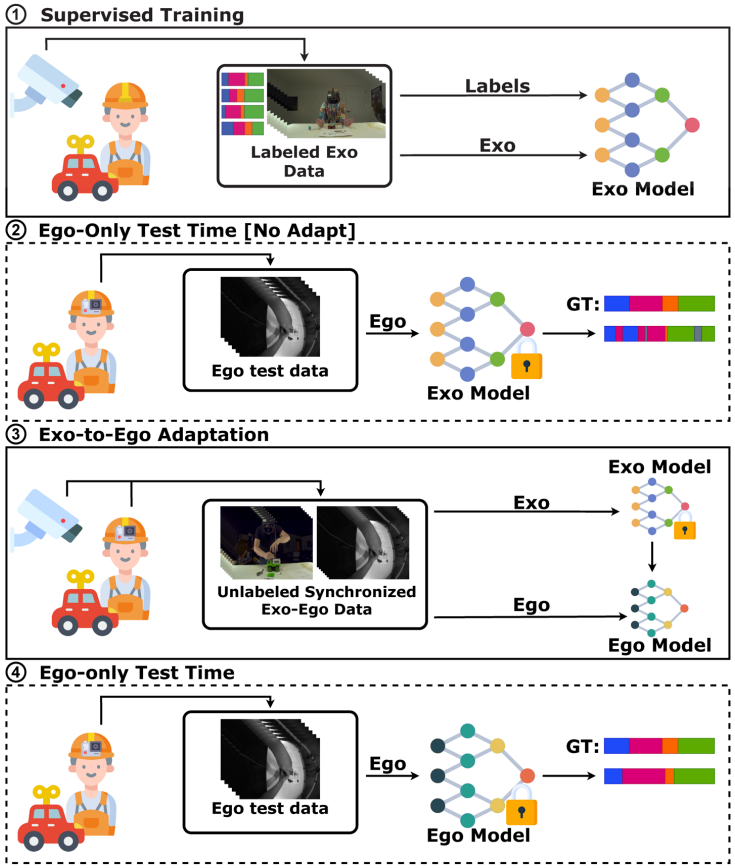
% Real Labeled Data	Approach	In-domain	Hand	Hand+Side	Hand+Contact	Object	Hand+Object
0%	Synthetic-Only		01.07	00.11	00.03	00.99	00.21
	Synthetic-Only	✓	56.05	35.14	15.24	4.79	12.85
	UDA		42.81	14.52	7.97	3.29	6.87
	UDA	✓	**78.83**	**70.91**	**28.14**	**25.84**	**34.78**
Improvement with respect to Synthetic-Only In-domain			**+22.78**	**+35.77**	**+12.90**	**+21.05**	**+21.93**
10% (347 images)	Real-Only		81.25	76.22	37.96	39.53	45.39
	Synthetic+Real		82.02	77.67	42.59	40.45	47.17
	Synthetic+Real	✓	82.19	78.31	42.11	42.23	48.24
	SSDA		**85.40**	**78.62**	**43.56**	**46.97**	**57.08**
	SSDA	✓	84.58	78.42	41.17	46.50	56.69
Improvement with respect to 10% Real-Only			**+4.15**	**+2.40**	**+5.60**	**+7.44**	**+11.69**
25% (870 images)	Real-Only		82.95	78.70	43.52	45.25	51.83
	Synthetic+Real		83.57	80.47	**48.32**	45.64	52.70
	Synthetic+Real	✓	83.31	**80.77**	46.39	47.57	54.57
	SSDA		**84.99**	80.41	46.31	49.34	58.17
	SSDA	✓	84.85	80.30	44.24	**49.37**	**59.48**
Improvement with respect to 25% Real-Only			**+2.04**	**+2.07**	**+4.80**	**+4.12**	**+7.65**
50% (1,739 images)	Real-Only		84.65	80.43	47.41	48.79	57.62
	Synthetic+Real		84.87	81.92	**53.97**	50.07	59.71
	Synthetic+Real	✓	84.81	81.28	50.16	51.42	58.10
	SSDA		**85.67**	82.00	52.20	**52.56**	**63.25**
	SSDA	✓	85.12	**82.01**	48.96	51.94	61.93
Improvement with respect to 50% Real-Only			**+1.02**	**+1.58**	**+6.56**	**+3.77**	**+5.63**
100% (3,479 images)	Real-Only		85.01	81.05	52.32	51.35	63.84
	Synthetic+Real		85.64	82.21	**57.25**	50.91	64.59
	Synthetic+Real	✓	**86.02**	82.50	55.81	53.26	**66.17**
	SSDA		85.94	**82.91**	54.13	52.50	64.41
	SSDA	✓	85.37	82.45	51.60	**53.30**	64.20
Improvement with respect to 100% Real-Only			**+1.01**	**+1.86**	**+4.93**	**+1.95**	**+2.33**

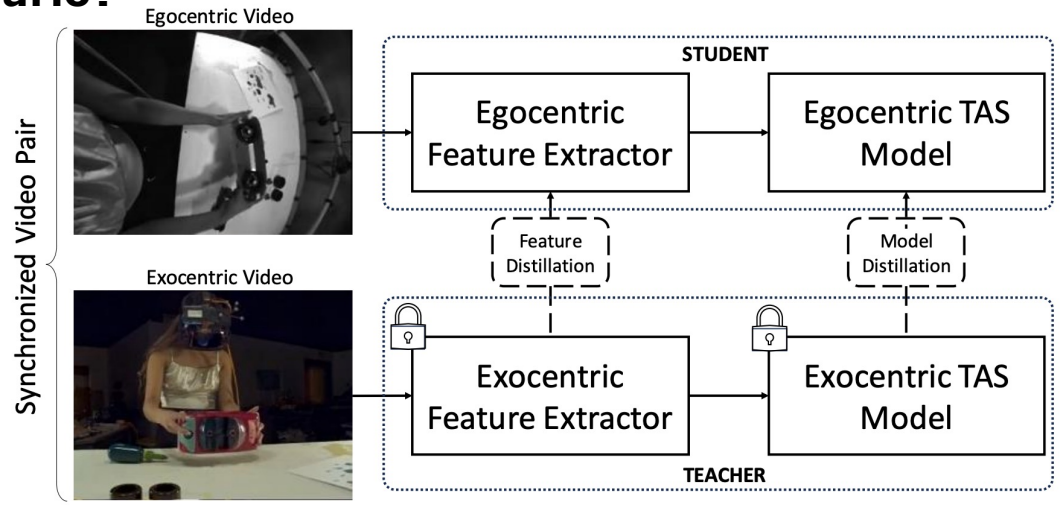
The Tool for Synthetic Data Generation, Code and Data to replicate experiments will be publicly released here: https://iplab.dmi.unict.it/HOI-Synth/

R. Leonardi, A. Furnari, F. Ragusa, G. M. Farinella, Are Synthetic Data Useful for Egocentric Hand-Object Interaction Detection? An Investigation and the HOI-Synth Domain Adaptation Benchmark, arXiv, 2023

Exocentric-to-Egocentric Model Transfer

How to transfer a temporal action segmentation system initially designed for exocentric (fixed) cameras to an egocentric scenario?

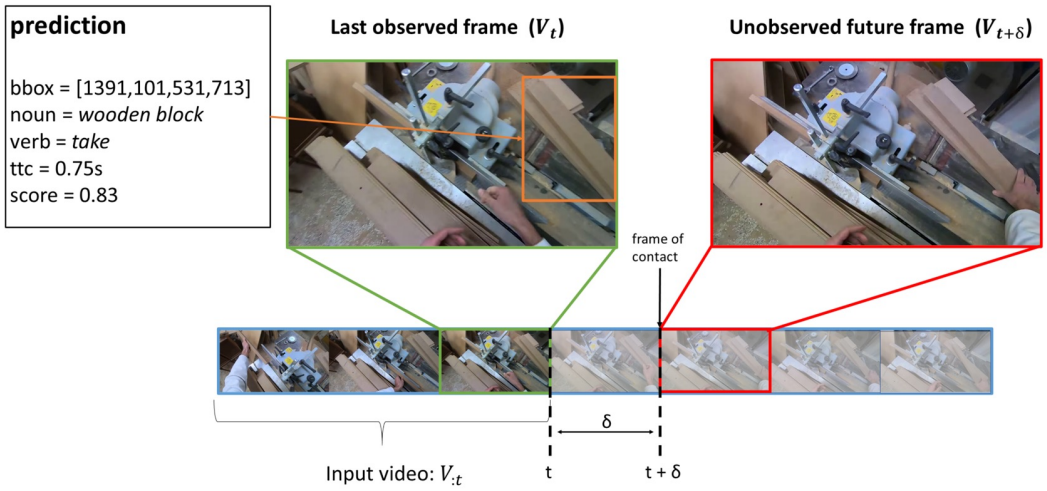


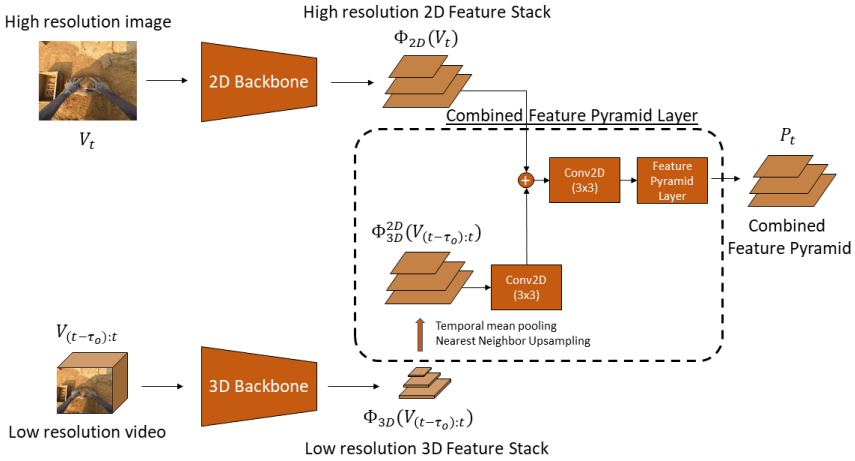


	Adaptation Task	Backbone	Feature Dist.	Model Dist.	Edit	F1{10}	F1{25}	F1{50}	MoF
1	EXO → EXO	TSM			**31.45**	**34.24**	**29.92**	**20.88**	**37.39**
2	(exo-oracle)	DINOv2			28.55	30.13	25.84	17.79	36.03
3	EGO → EGO	TSM			**29.25**	**31.06**	**25.40**	**17.47**	**36.60**
4	(ego-oracle)	DINOv2			26.42	26.50	22.16	14.84	33.47
5					10.53	6.47	2.44	0.56	5.42
6		TSM	✓		12.38	8.73	5.64	1.99	8.21
7				✓	20.17	21.11	18.23	13.12	17.77
8			✓	✓	**25.39**	**26.79**	**22.37**	**15.52**	**30.30**
9	EXO → EGO	Improvement w.r.t. baseline (line 5)			+14.86	+20.32	+19.93	+14.96	+24.88
10					12.60	10.18	7.21	2.40	14.15
11		DINOv2	✓		14.38	10.80	6.53	2.33	10.94
12				✓	**28.59**	**29.58**	**24.84**	**16.38**	**31.36**
13			✓	✓	22.64	23.12	20.22	13.51	27.31
14		Improvement w.r.t. baseline (line 10)			+15.99	+19.40	+17.63	+13.98	+17.21

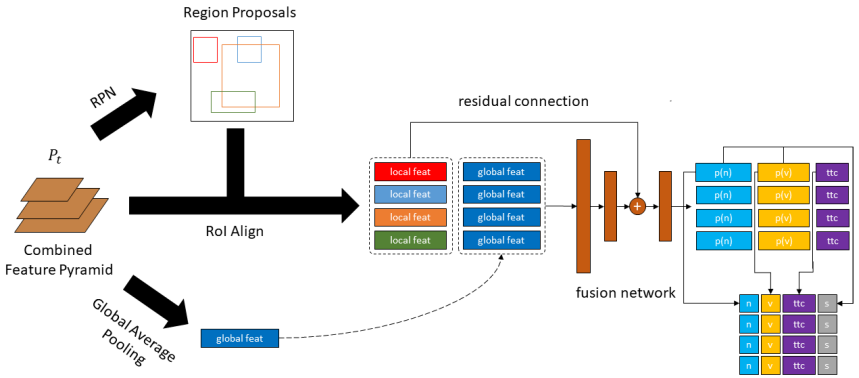
C. Quattrocchi, A. Furnari, D. Di Mauro, M. V. Giuffrida, G. M. Farinella, Synchronization is All You Need: Exocentric-to-Egocentric Transfer for Temporal Action Segmentation with Unlabeled Synchronized Video Pairs, https://arxiv.org/abs/2312.02638, 2023

Short Term Object Interaction – StillFast Architecture



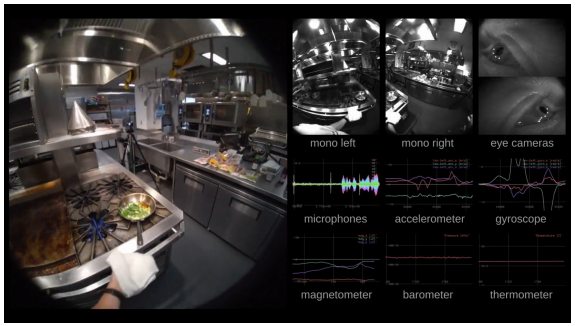


StillFast is ranked first in the public leaderboard of the EGO4D Short Term Object Interaction anticipation challenge 2022. Code available at https://iplab.dmi.unict.it/stillfast/



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

Ego-Exo4D Dataset and Benchmark



A **diverse**, **large-scale multi-modal**, **multi-view**, video dataset and benchmark collected across 13 cities worldwide by 839 camera wearers, capturing **1422 hours** of video of **skilled human activities annotated by experts**.

Procedure understanding

Given a video segment s_i and its segment history $S_{:i-1} = \{s_1, \ldots, s_{i-1}\}$, models have to 1) determine previous keysteps (to be performed before s_i); infer if s_i is 2) optional or 3) a procedural mistake; 4) predict missing keysteps (which should have been performed before s_i); and 5) forecast next keysteps (for which dependencies are satisfied and hence can be executed next).

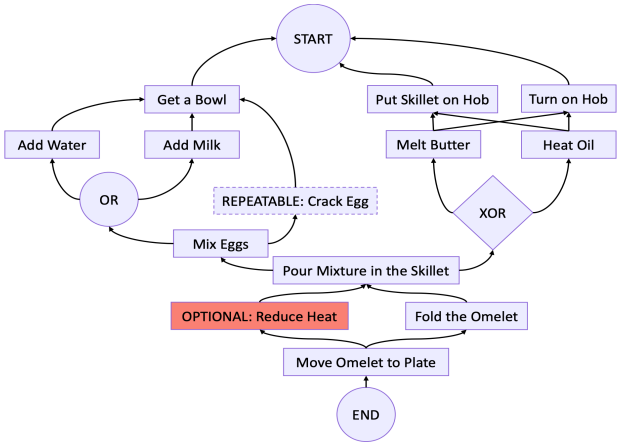


Figure 40. Example task-graph of a "Cooking Omelet" procedure.

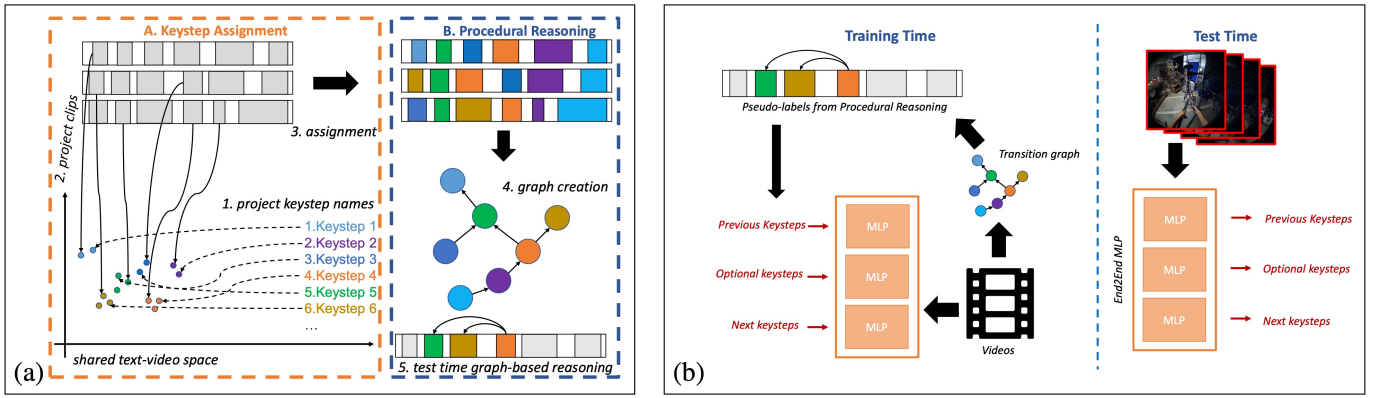


Figure 41. Overview of the two procedure understanding approaches considered in our evaluation: (a) graph-based baselines for procedure understanding rely on a Keystep Assignment and a Procedural Reasoning component; (b) the architecture of our end-to-end baseline.

K. Grauman et al., Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives, arXiv,. https://arxiv.org/abs/2110.07058, 2023

Università di Catania — Streaming Egocentric Action Anticipation

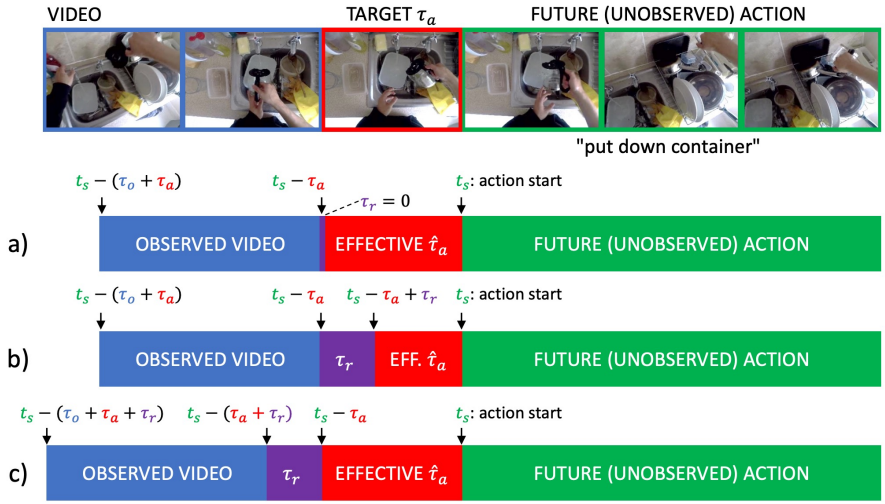


Figure 1: Different schemes to evaluate egocentric action anticipation methods. (a) The ideal commonly used scenario in which the runtime is assumed to be zero. (b) The real-world case in which the runtime is non-negligible. In this case the effective anticipation time is smaller than the target one. (c) A fairer evaluation scheme in which the observation is sampled ahead of time to counterbalance the non-negligible runtime and obtain an effective anticipation time equal to the target one. In the figure, t_s denotes the action start timestamp, τ_o denotes the observation time, τ_a denotes the anticipation time, τ_r denotes the runtime, and $\hat{\tau}_a$ denotes the effective anticipation time.

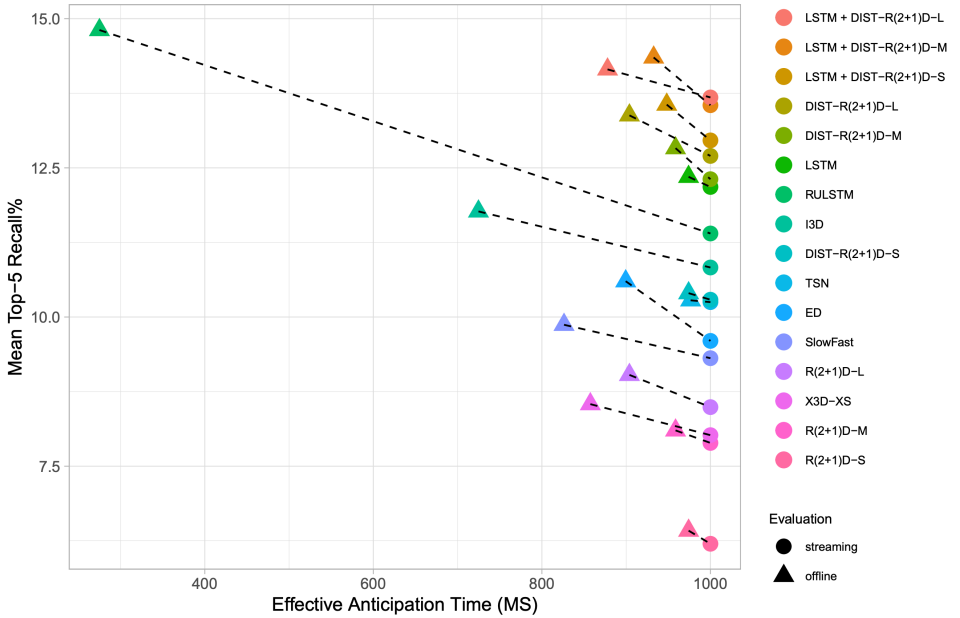


Figure 8: Effect on performance of the streaming egocentric action anticipation evaluation scheme on EPIC-KITCHENS-55. The legend is sorted according to streaming performance. The performance of computationally expensive methods such as RULSTM and I3D tends to be more affected by the streaming scenario, whereas lightweight models such as the proposed DIST-R(2+1)D-L and DIST-R(2+1)D-M are more robust to the streaming evaluation scheme.

Antonino Furnari, Giovanni Maria Farinella, Streaming egocentric action anticipation: an evaluation scheme and approach, Computer Vision and Image Understanding (CVIU), 2023

An Outlook into the Future - What's Relevant in Egovision?

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.

https://arxiv.org/abs/2308.07123

OpenReview.net

An Outlook into the Future of Egocentric Vision

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

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

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

Add Comment

Reply Type: all Author: everybody Visible To: all readers Hidden From: nobody **6 Replies**

Related work on modeling social interactions, especially multimodal dialogue agents

Jaewoo Ahn

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

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

In our recent paper [1], we explored the benefits of a multimodal approach to dialogue personalization. Our findings showed that incorporating both text and images in defining a persona greatly enriched the dialogue agent's understanding and personality 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 integrating 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 text capabilities 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.

We asked feedback to the research community before submission to the IJCV

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

Università di Catania

An Outlook into the Future – Scenarios and Stories

EGO-HOME

Sam is finally home after a long day. EgoAI kept track of Sam's food intake and a tomato soup sounds like the best complementary nutrition

This way the tomato will cook evenly

A 3D projection of Remy helps with cooking

Audible 3D projection

Sam is impressed by how fun it is to cook with his 3D friend

Toaster reminder

EgoAI recommends some more spice

Waves hitting the shore look and sou...

While getting ready for bed, Sam feels an itch on the wrist that has annoyed him the whole day. EgoAI stores a picture of the injury and sends it to Sam's doctor for advice

EgoAI proposes a short clip from his day, but Sam decides not to share it

EGO-WORKER

EgoAI localises Marco and provides ruote instructions to reach his workstation for the day

EgoAI verifies if Marco is properly wearing the Personal Protection Equipment (PPE)

Where should I go today in the factory?

In the past, EgoAI guided Marco to the closest fire extinguisher during a fire

EgoAI passes a mes... from the m... to...

Since the me... devi...

...d turns off the IoT electrical socket while promptly alerting Marco

For the rest of the day, EgoAI validates Marco's work making sure all the procedures are properly and safely completed

By the end of the day, EgoAI checks Marco's feedback for improving future sessions

EGO-TOURIST

EgoAI prepares Claire a personalised and exciting one-day itinerary in Turin

EgoAI suggests an half-day visit to the Egyptian museum

EgoAI activates the 3D projection of Cleopatra to guide and interact with Claire

Claire feels transported to ancient Egypt

Claire asks Cleopatra for a good place for a pizza

...pizza place for ...ening Claire about the ...d various Italian monuments

...e next ... scheduled to arrive in 20 minutes

EgoAI offers a egocentric view from the chef who prepared her that delicacy

EgoAI suggests Claire a proper Italian coffee at a nearby cafè, sided by a slice of bunet, Turin-based dessert

EgoAI actively saved snapshots and videos of the day

EgoAI retrieves the closest souvenir shop based on Claire's taste and budget

EGO-POLICE

EgoAI is constantly pinpointing Judy's position and would send an alert to the headquarters if she encounters unusual events or dangerous situations

EgoAI helps Judy navigate through the shortest safe path to target places

Judy was able to swiftly arrest him

EgoAI has both thermal and multi-spectral sensors

Thanks to its sensors, EgoAI calculates a low risk for explosive content

Judy also appreciated the help of EgoAI when she had to manage an abandoned backpack

EgoAI projects a clear red circle around the backpack with the minimal stand-off distance

EgoAI connects Judy with the bomb squad and live-shares the observed scene

Thanks to EgoAI, all the relevant events are saved and transformed into a document with related images and video recordings

EgoAI guides Judy with exact instructions to grasp the backpack and open it

The sensitive information is properly identified and secured under admin rights to protect citizens' privacy

EGO-DESIGNER

EgoAI also suggests adding velvet couches on the right and a carved wooden table on the left

...access ...the database of the equipment warehouse; Stanley can search for the available pieces of furniture

EgoAI also allows Stanley to visualise how the actors should move in the space considering that there will be musicians in the middle of the room

EgoAI shares the scene with the actors. Through their own EgoAI, they are immersed inside the changing and moving 3D computer-generated environment

EgoAI assists make-up artists with advanced 3D modelling techniques to project guidelines on the actor's face while applying make-up

EgoAI also assists the director. He is able to preview the planned scene and light effects in real-time while shooting the scene

EGO-AI

From Narratives to Multimodal EgoAI Research Tasks

EGO-Home

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

EGO-Worker

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

EGO-Tourist

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

EGO-Police

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

EGO-Designer

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

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

Industrial Partners on Projects and and PhD Programs (since 2006)

Xenia SOFTWARE SOLUTIONS
2014 -

Meta facebook
2019 -

NVIDIA
2020 -

PHILIPS
2018 -

intel
2021 -

OSRAM
2016 -2018

TOYOTA
2024 -

life.augmented
2006 -

JOINTOPENLAB

TIM
2013 - 2018

linkverse
2019 - 2022

Centro Studi
Process Development & Applied Research
gruppo orizzonti holding
2015 - 2018

orangedev
2018 -

morpheos
2018 -

PARK SMART
2015 - 2019

NEXT VISION
2021 -

Projects

2020-2023 (XENIA)

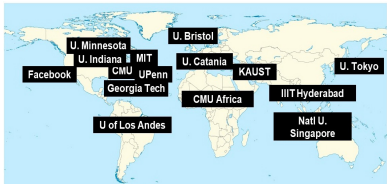
Localization, monitoring, and anticipation in industrial environment from First Person Vision

2020-2023 (XENIA)

Localization and attention estimation in cultural sites from First Person Vision

2019-2023 (Philips)

Personal Healthcare Interface Leveraging Human-Machine Natural Interactions from First Person Vision



2019-Today

(Meta - Facebook)

+3000 hrs of Egocentric Videos for Ego Tasks Benchmark

Other Projects supported by Industry

- Assistive Wearable Technologies and AR for Rehabilitation (AIAS 2021-2024)
- Long Term Egocentic Vision (INTEL – from Oct 2021)
- Procedural Understanding (Toyota – from 2024)

Projects can be found at https://iplab.dmi.unict.it/fpv/projects

Università di Catania ENIGMA Wearable Assistant Demonstrator @ UNICT

Patent Accepted

Number: 102020000027759

Acceptance date: 30/11/2022

Enhance Training

Improve Safety

Energy Saving

Spin-off of the University of Catania

Founders



Xenia® SOFTWARE SOLUTIONS

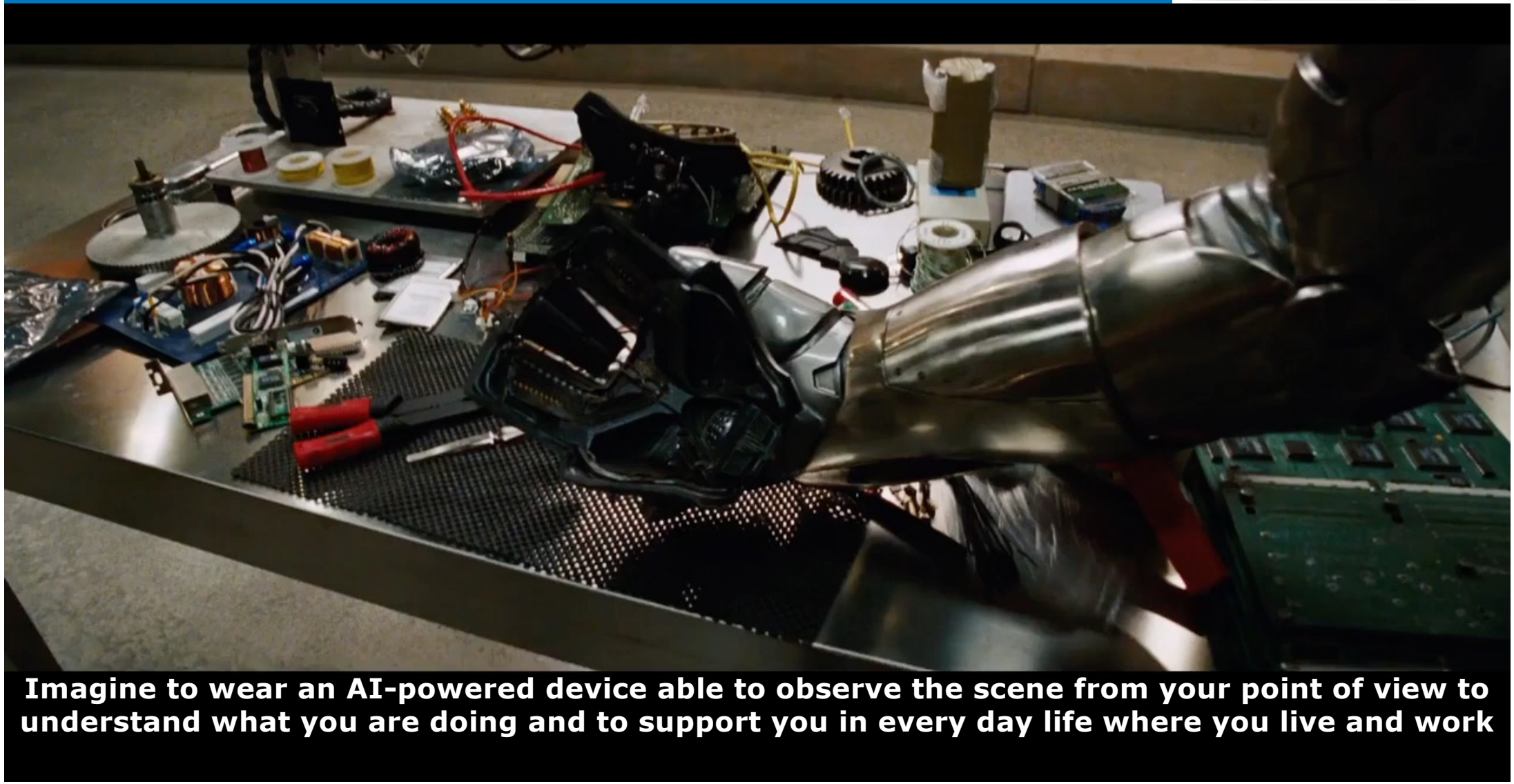
Founders

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

Università di Catania

What will the future AI be? We wonder!

NEXT VISION



Imagine to wear an AI-powered device able to observe the scene from your point of view to understand what you are doing and to support you in every day life where you live and work

Dispositivi indossabili

CONSUMER

Microsoft HoloLens 2

VUZIX BLADE

Magic Leap

EPSON MOVERIO

INDUSTRIAL

Trimble XR10 (based on Holo2)

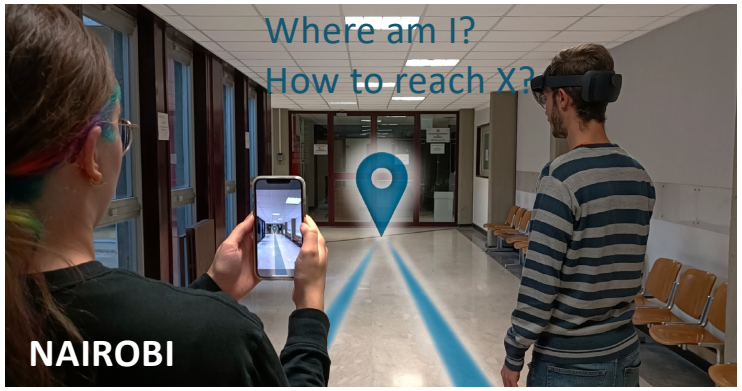
Vuzix M4000

EPSON BT2200

REALWEAR HMT-1Z1

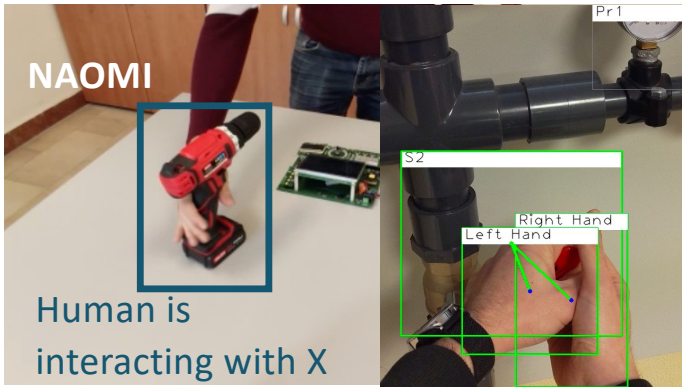
Four Solutions

1



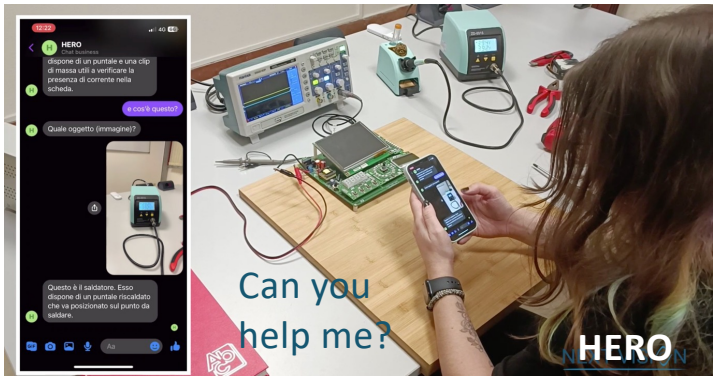
Localization of Humans and Navigation| Aware of Context

2



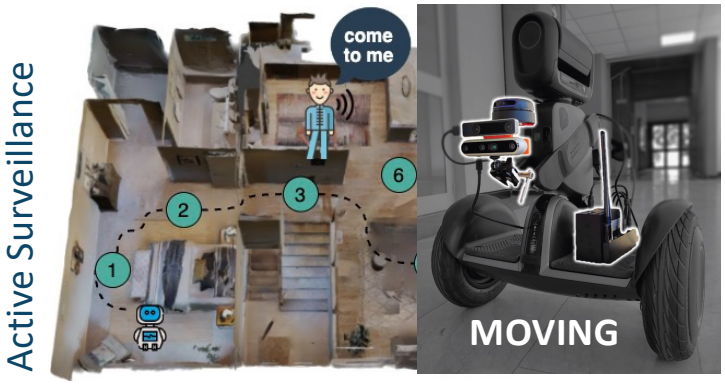
Human-Object Interaction | Aware of Objects

3



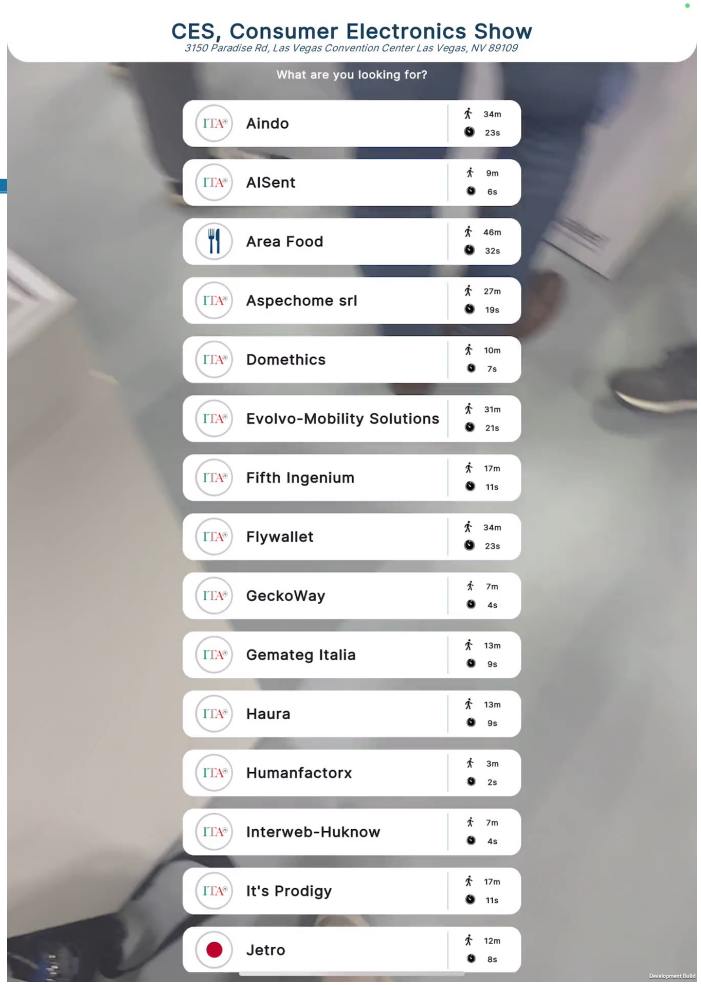
Conversational AI Agents | Aware of Human Needs and Intents

4



Autonomous AI Agents | Aware of Cooperation

Demo Video - NAIROBI



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

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

NOKIA

Università di Catania

UNIONE EUROPEA

REPUBBLICA ITALIANA

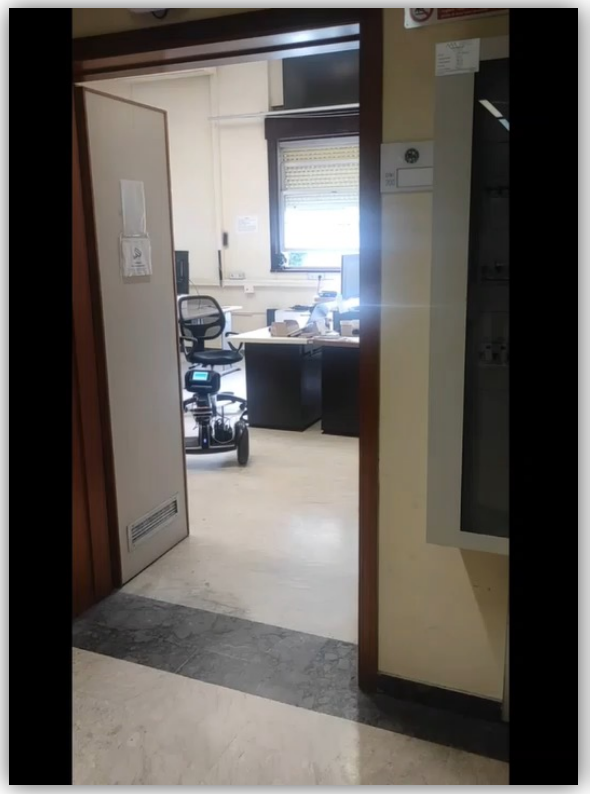
REGIONE SICILIANA

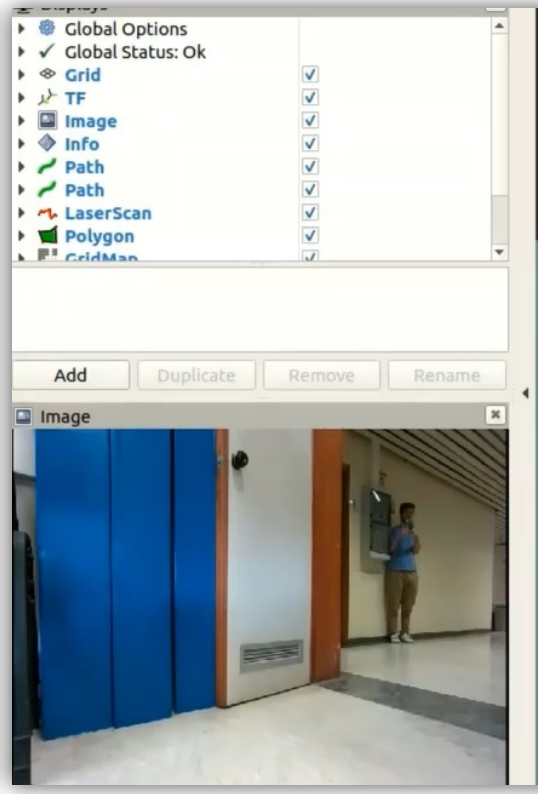
FESR SICILIA 2014-2020

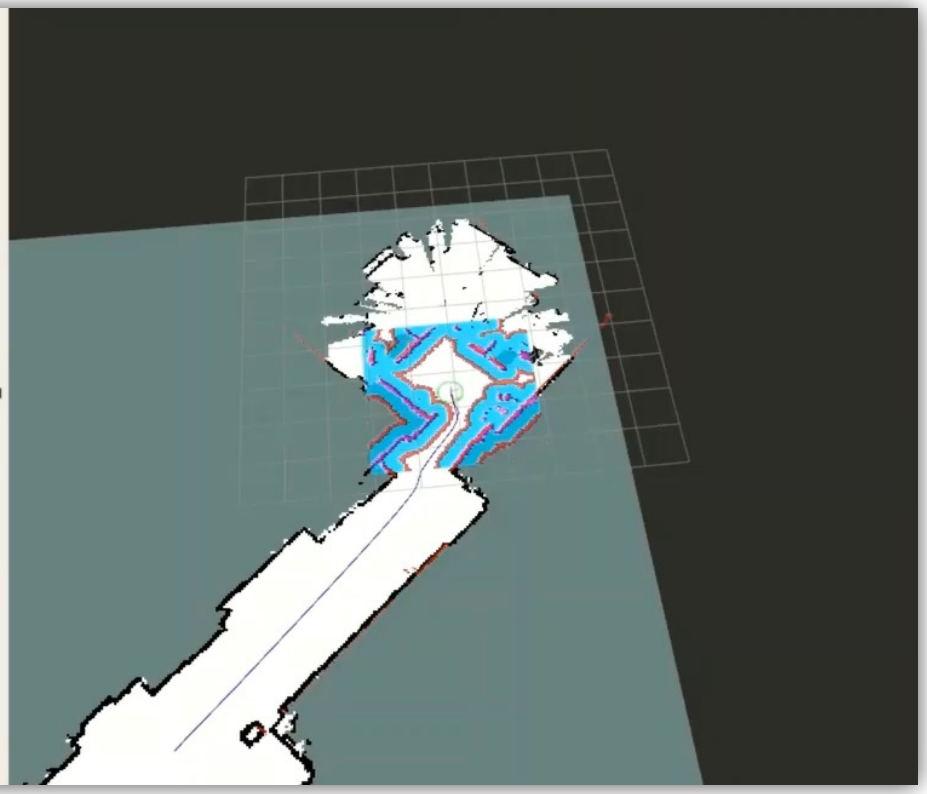
CES

HERO is a Vision & Language Conversational Agent
Wearable version in coming soon...

MOVING




Global Options
Global Status: Ok
Grid
TF
Image
Info
Path
Path
LaserScan
Polygon
Add
Duplicate
Remove
Rename
Image



ML Course 2023-2024

Struttura del corso

11 Lezioni Teoriche

6 Laboratori (Flipped Classroom)

4 Lezioni Industriali (in collaborazione con Luxottica, Xenia Progetti, STMicroelectronics, AI datascience)

4 Lezioni di Ricerca (in collaborazione con University of Bath, University of Zaragoza, University of Rome La Sapienza, Esperti di Data Law)

Sito: http://www.dmi.unict.it/farinella/ML

- Credenziali
  - studente
  - corsoml

Canale Teams – Codice: j8rls3j

Libri: la biblioteca del DMI è provvista di libri utili per approfondimenti. Si veda anche il sito del corso per riferimenti.

Modalità di Esame

- Esami
  - Scritto + Progetto (in team max 3 persone)
  - 50% Scritto, 50% Progetto
- Vedi note online

Stage e Tesi

- Sperimentali
- In collaborazione con Aziende (Next Vision, Xenia, STMicroelectronics, ecc..)
- E’ consigliato iniziare la tesi all’inizio del secondo anno (6+ mesi)

Why is interesting to study ML

Application Domains and Examples

ML Application Domains

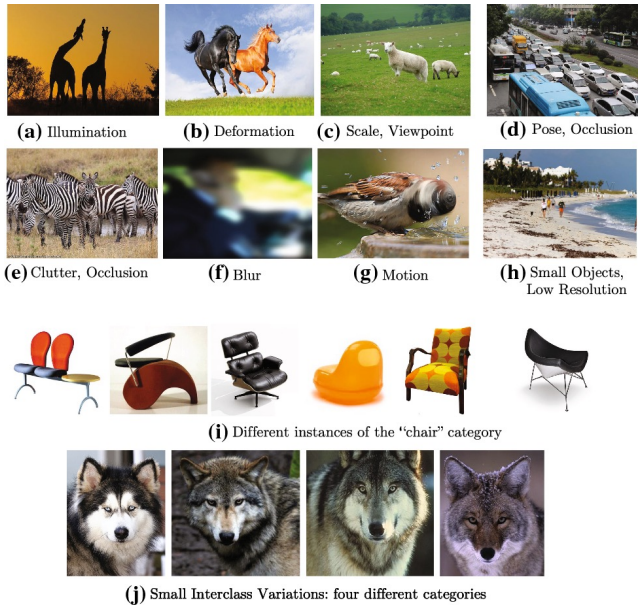
- Computer Vision
- Robotics
- Gaming
- Natural Language Processing and Speech
- Art
- …

Machine Learning for Computer Vision

- Object Recognition and Segmentation
- Image Captioning
- Human pose Estimation
- Summarization and Indexing
- Localization
- First Person Vision for Anticipation

Why is interesting

- Vision is challenging!
  - For a small patch of resolution 256x256 with 256 pixel values
  - A total of $2^{8x256x256} = 2^{524288}$ of possible images
  - In comparison there are about ~10^{24} stars in the universe
- Image variability
  - Different viewpoint, scales, deformations, occlusions…
- Semantic Variation
  - Intra-class Variation (think to faces)
  - Inter-class overlaps (thinks twins)



Machine Learning for Robotics / Artificial Agents

- Self-Driving Cars
- Activity Recognition
- Tracking and Location Estimation
- Gaming

Why is interesting

- Agents (e.g., Robots) are usually considered in controlled environments
  - Laboratory settings, Predictable positions, Standardized tasks (like in factory robots)
- Real Life
  - Environments constantly change, new tasks need to be learnt without guidance, unexpected factors must be dealt with

Machine Learning for Natural Language Processing, Speech Recognition, Audio Recognition

- Language Understanding
- Speech Understanding
- Audio recognition
- Conversational Agents (e.g., ChatGPT)

Why is interesting

- NLP is a complex task
  - synonymy (“chair”, “stool” or “beautiful”, “handsome”)
  - ambiguity (“Cut to the chase”)
- NLP is very high dimensional
  - assuming 150K English words, we need to learn 150K classifiers
  - with quite sparse data for most of them

Machine Learning for Art

- Imitating Famous Painters
- Image to Image Translation
- Generating Music
- Generating Films

Machine Learning for Computer Graphics

- Generating Realistic Images
- Generating Avatars
- Generating Video

Why is interesting

- Music, painting, etc. are tasks that are uniquely human
  - Difficult to model
  - Even more difficult to evaluate (if not impossible)
- If machines can generate novel pieces even remotely resembling art, they must have understood something about “beauty”, “harmony”, etc.
- Have they really learned to generate new art, however?
  - Or do they just fool us with their tricks?

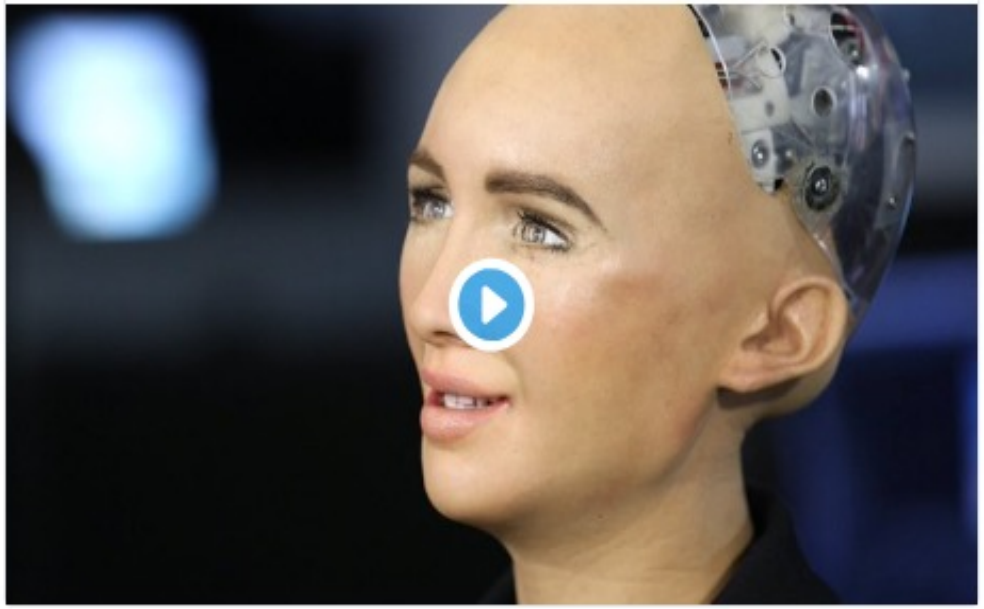
We can continue giving examples on security, chemistry, industry, etc, etc…
but you have got the gist!

Nowadays, ML is basically everywhere

Take care to fake AI!

Tech Insider @techinsider

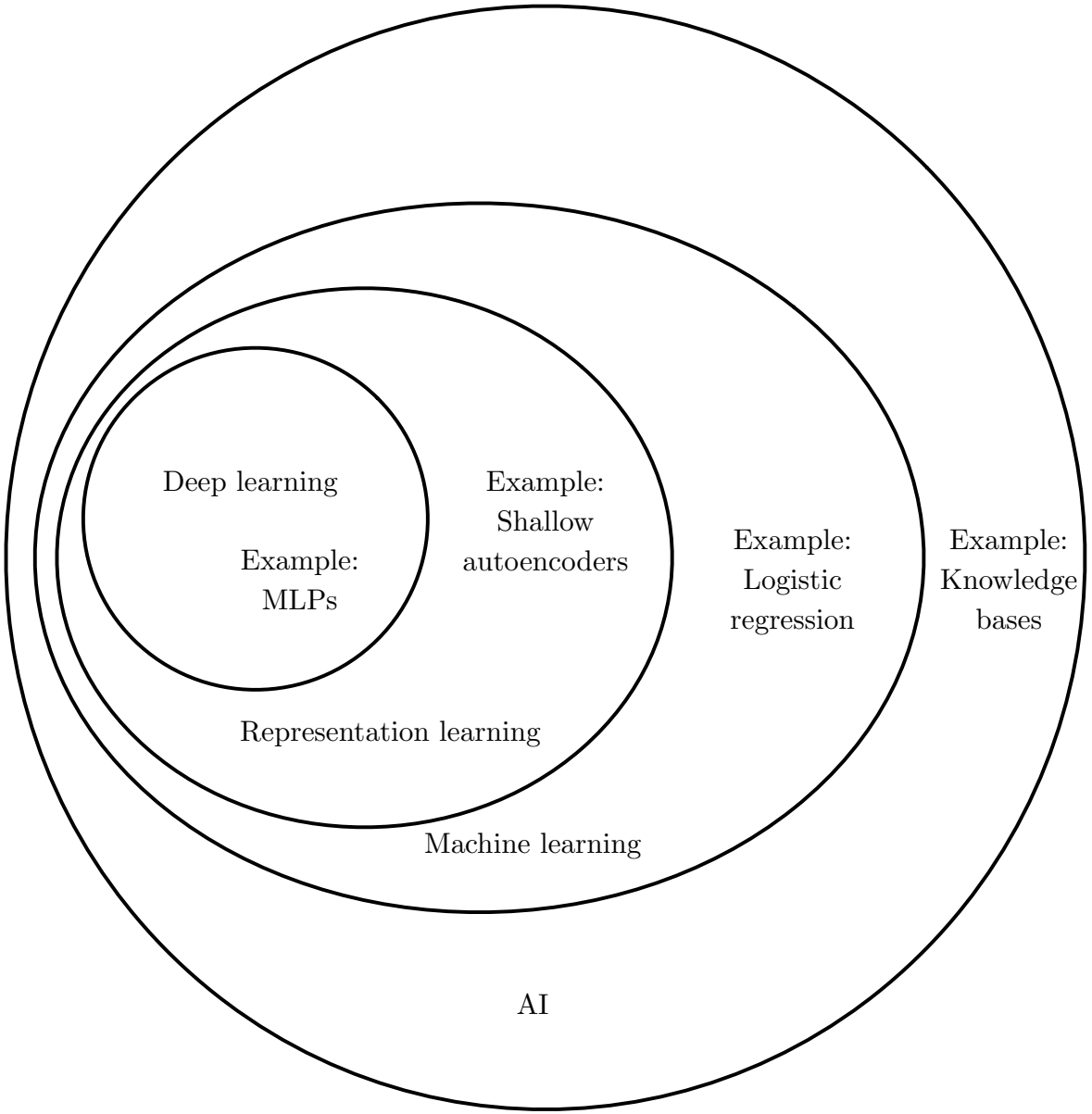
We talked to Sophia — the first-ever robot citizen that once said it would 'destroy humans'



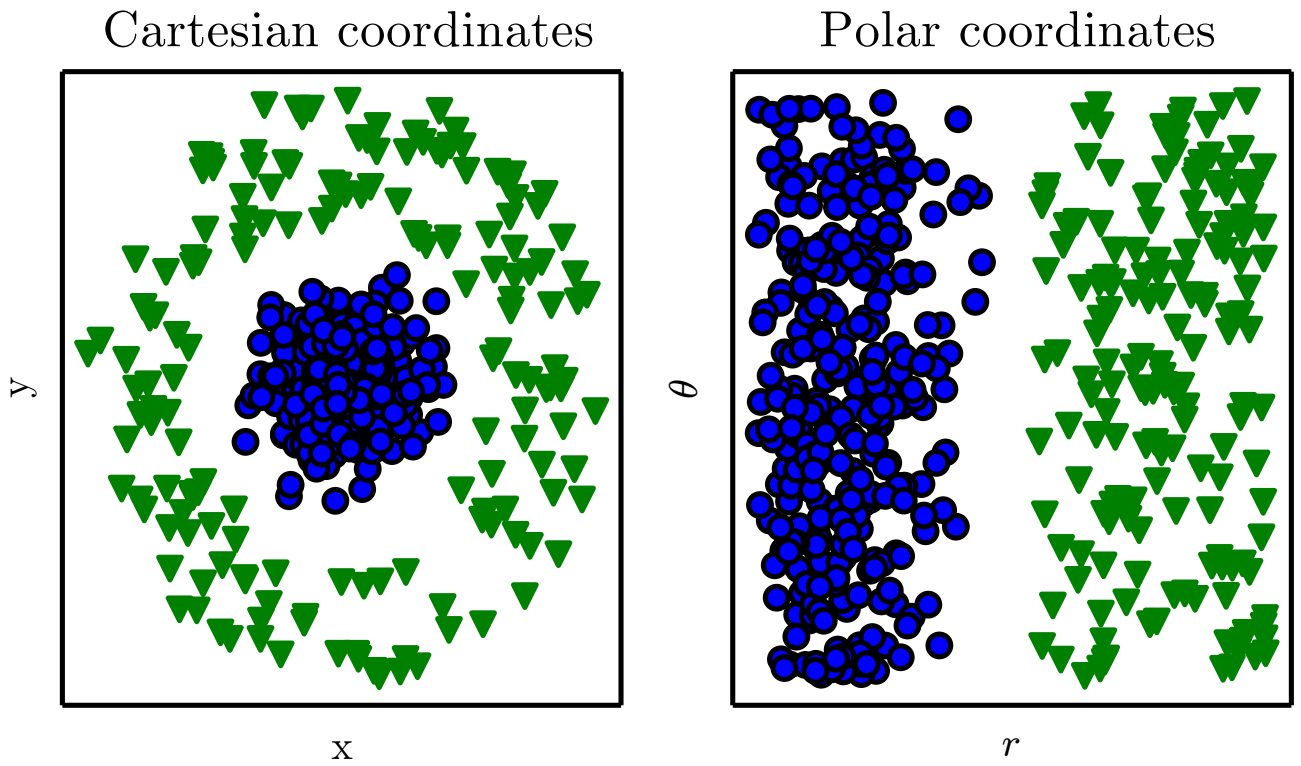
♡ 1,415 10:29 PM - Jan 4, 2018

Yann LeCun
@ylecun

This is to AI as prestidigitation is to real magic.
Perhaps we should call this "Cargo Cult AI" or "Potemkin AI" or "Wizard-of-Oz AI".
In other words, it's complete bullsh*t (pardon my French).
Tech Insider: you are complicit in this scam.

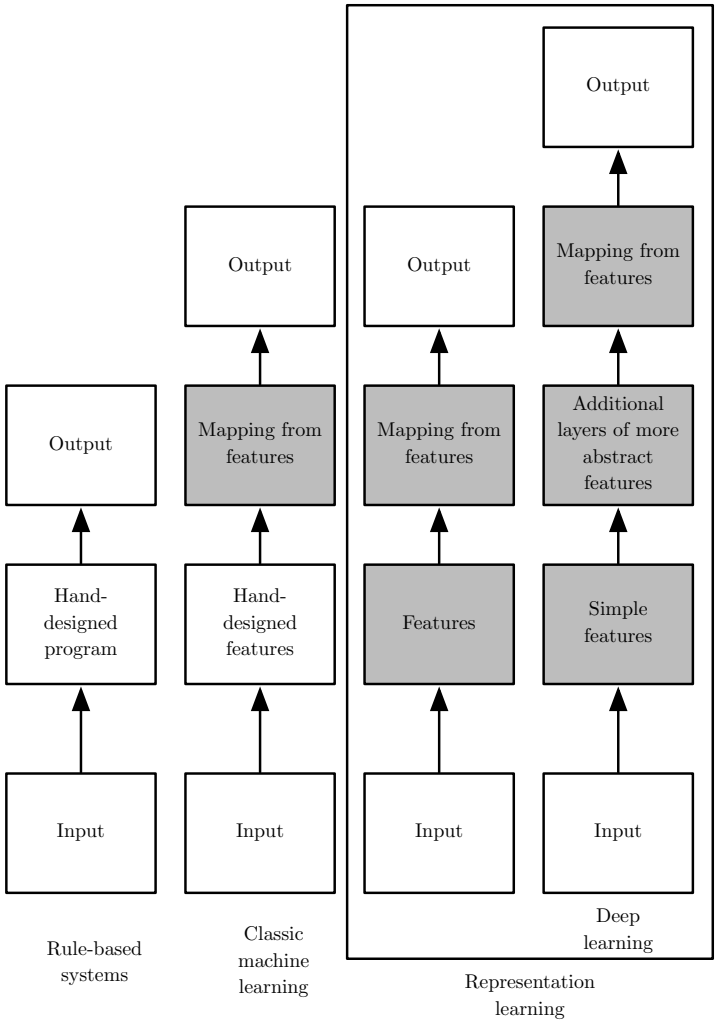

Deep learning
Example:
MLPs
Example:
Shallow
autoencoders
Representation learning
Example:
Logistic
regression
Machine learning
Example:
Knowledge
bases
AI

Representation of Data Matter



*Deep Learning Book (*www.deeplearningbook.org)

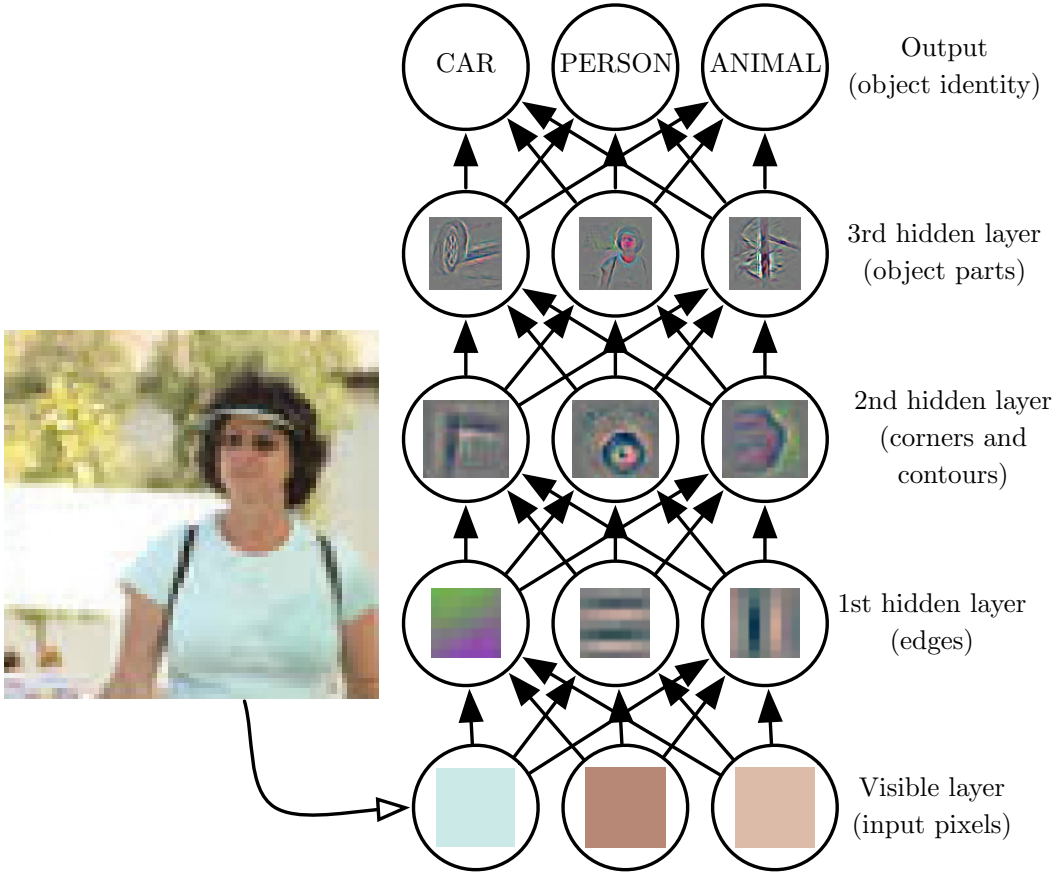
Learning Multiple Components



Learning Procedure

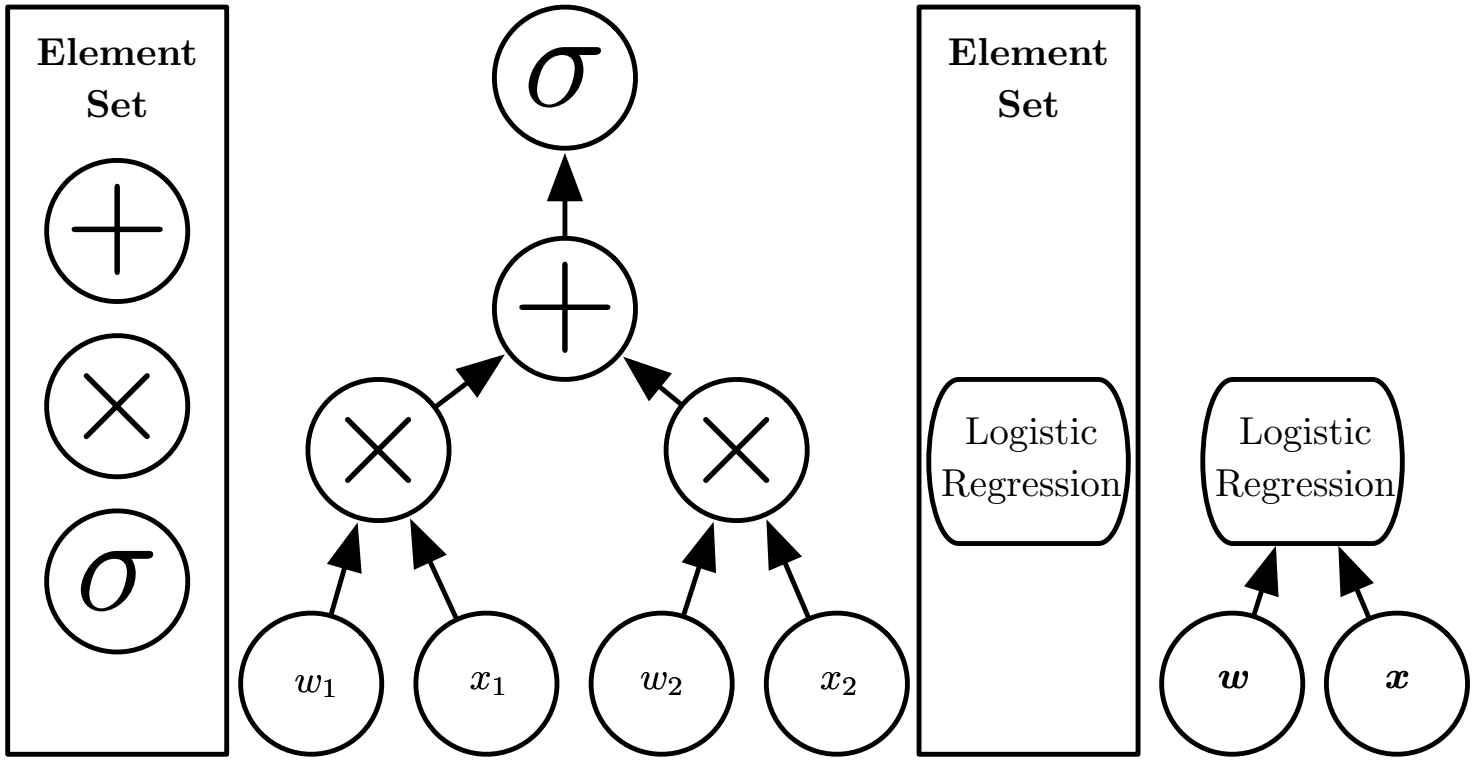
*Deep Learning Book (*www.deeplearningbook.org)

Repeated Composition



Deep Learning Book (www.deeplearningbook.org)

Computational Graphs
Mapping an Input to an Output



*Deep Learning Book (*www.deeplearningbook.org)