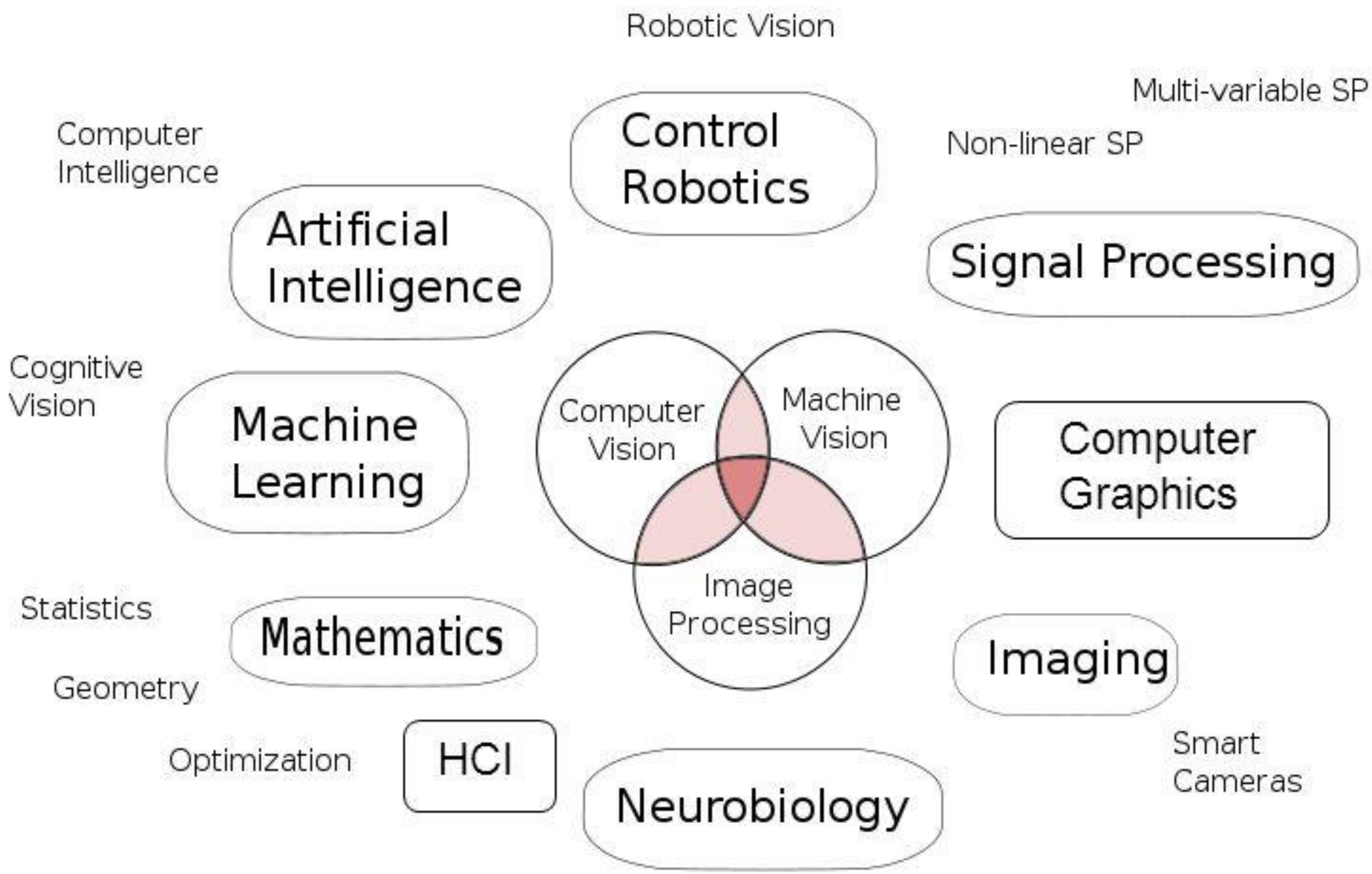
Deep Learning for Computer Vision

UCA Master 2 Data Science

INRIA Sophia Antipolis – **STARS team**

14 January / 25 March 2025

Vision is multidisciplinary



- **Computer Vision** is a subfield of artificial intelligence as machine learning.
- Techniques in machine learning and other subfields of AI (e.g. NLP) can be borrowed and reused in computer vision.

Computer Vision: many Tasks

Computer Vision is an interdisciplinary scientific field that deals with how computers can be made to gain high-level understanding from digital images or videos.

From the perspective of engineering, it seeks to automate tasks that the human visual system can do. [Wikipedia]

Computer Vision Tasks:

- Recognition of Entities : Images, 2/3D Objects, People/Pose/Face/Gaze or Emotions/Events
  - Classification
  - Detection, segmentation
  - Retrieval
- Motion analysis
  - Optical flow
  - Tracking of objects, ReID
- Image/video synthesis, generation
- Image restoration, super resolution, denoising, 3D geometry
- Biometrics, medical image, remote sensing,..
- Multimodalities (text, audio, depth, physiological, etc...)

Video Analytics (or VCA) applies CV & ML algorithms to extract/analysis content from videos

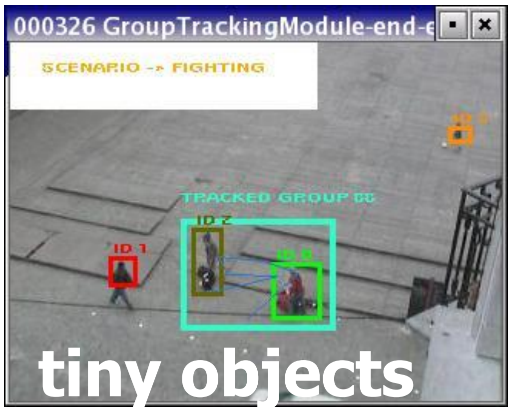
Video Analytics: many research Domains

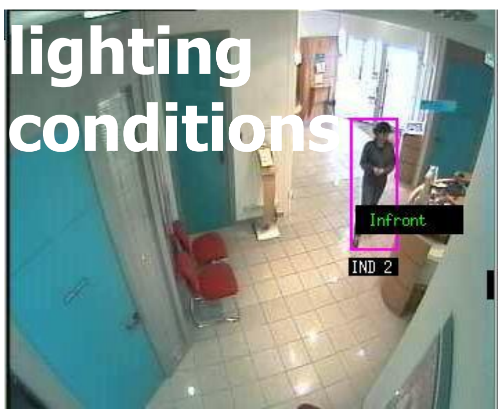
- Smart Sensors: Acquisition (dedicated hardware), thermal, omni-directional, PTZ, cmos, IP, tri CCD, RGBD Kinect, FPGA, DSP, GPU.
- Networking: UDP, scalable compression, secure transmission, indexing and storage.
- Image Processing/**Computer Vision**: feature extraction, Deep CNN, 2D object detection, active vision, tracking of people using 3D geometric approaches
- Event Recognition: Probabilistic approaches HMM, DBN, logics, symbolic constraint networks
- Multi-Sensor Information Fusion: cameras (overlapping, distant) + microphones, contact sensors, physiological sensors, optical cells, RFID
- Reusable Systems: Real-time distributed dependable platform for video surveillance, OSGI, adaptable systems, Machine learning
- System Optimization: complexity reduction (# parameters, Flops) matrix factorization, distillation
- Visualization: 3D animation, ergonomic, video abstraction, annotation, simulation, HCI, interactive surface.

Video Analytics : Issues

Practical issues

Video Understanding systems have poor performances over time, can be hardly modified and do not provide semantics





Video Analytics : Issues

V1) Acquisition information:

- V1.1) Camera configuration: mono or multi cameras,
- V1.2) Camera type: CCD, CMOS, large field of view, colour, event, thermal cameras (infrared), Depth
- V1.3) Compression ratio: no compression up to high compression,
- V1.4) Camera motion: static, oscillations (e.g., camera on a pillar agitated by the wind), relative motion (e.g., camera looking outside a train), vibrations (e.g., camera looking inside a train),
- V1.5) Camera position: top view, side view, close view, far view,
- V1.6) Camera frame rate: from 25 down to 1 frame per second,
- V1.7) Image resolution: from low to high resolution, deformation,

V2) Scene content:

- V2.1) Classes of physical objects of interest: people, vehicles, crowd, mix of people and vehicles,
- V2.2) Scene type: indoor, outdoor or both,
- V2.3) Scene location: parking, tarmac of airport, office, road, bus, a park,
- V2.4) Weather conditions: night, sun, clouds, rain (falling and settled), fog, snow, sunset, sunrise,
- V2.5) Clutter: empty scenes up to scenes containing many contextual objects (e.g., desk, chair),
- V2.6) Illumination conditions: artificial versus natural light, both artificial and natural light,
- V2.7) Illumination strength: from dark to bright scenes,

Video Analytics : Issues

V3) Technical issues:

- V3.1) Illumination changes: none, slow or fast variations,
- V3.2) Reflections: reflections due to windows, reflections in pools of standing water, reflections,
- V3.3) Shadows: scenes containing weak shadows up to scenes containing contrasted shadows (with textured or coloured background),
- V3.4) Moving Contextual objects: displacement of a chair, escalator management, oscillation of trees and bushes, curtains,
- V3.5) Static occlusion: no occlusion up to partial and full occlusion due to contextual objects,
- V3.6) Dynamic occlusion: none, up to one person occluded by a car, by another person,
- V3.7) Crossings of physical objects: none up to high frequency of crossings and high number of implied objects,
- V3.8) Distance between the camera and physical objects of interest: close up to far,
- V3.9) Speed of physical objects of interest: stopped, slow or fast objects,
- V3.10) Posture/orientation of physical objects of interest: lying, crouching, sitting, standing,
- V3.11) Calibration issues: little or large perspective distortion, 3D information

Video Analytics Applications

- Strong impact in transportation (metro station, trains, airports, aircraft, harbors)
- Traffic monitoring (parking, vehicle counting, street monitoring, driver assistance, self-driving car)
- Control access, intrusion detection and Video surveillance in public places, building, biometrics, face recognition
- Store monitoring, Retail, Aware House, Bank agency
- Health (HomeCare) patient monitoring,
- Video communication (Mediaspace, 3D virtual reality, augmented reality)
- Sports monitoring (Tennis coach, Soccer analytics, F1, Swimming pool monitoring), rehabilitation, relapse
- Other application domains : Robotics, Drones, Teaching, Biology, Animal Behaviors, Risk management …

➢ Creation of start-up
  - ➢ Keeneo: http://www.keeneo.com/
  - ➢ Ekinnox: https://www.ekinnox.com/


Video Analytics : Scientific Issues

Performance: robustness of real-time (vision) algorithms

Bridging the gaps at different abstraction levels:

- From sensors to image processing [sensor world]
- From image processing to 4D (3D + time) analysis [physical world]
- From 4D analysis to semantics [end-user world]

Uncertainty management: [how reliable]

- uncertainty management of noisy data (imprecise, incomplete, missing, corrupted)
- formalization of the expertise (fuzzy, subjective, incoherent, implicit knowledge, partial models)

Independence of the models/methods versus: [how generic]

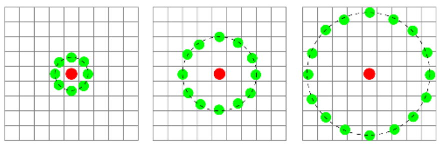
- Sensors (position, type), scenes, low level processing and target applications
- several spatio-temporal scales

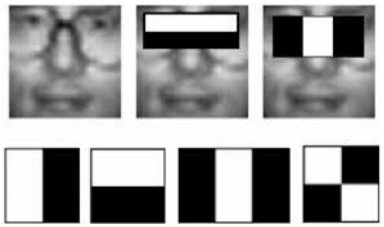
Knowledge management :

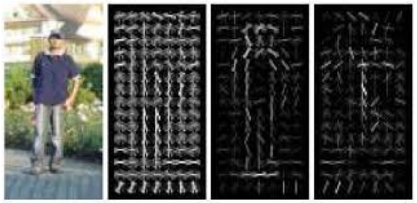
- Bottom-up versus top-down, focus of attention
- Regularities, invariants, models and context awareness
- Knowledge acquisition versus ((none, semi)-supervised, incremental) learning techniques
- Formalization, modeling, ontology, standardization

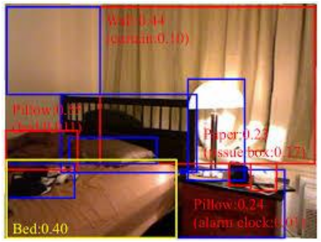
A brief history of Computer Vision

Geometric, Statistics, handcrafted features









LBP, 1994
Local Binary Patterns

Viola & Jones, 2001
Face Detection

Dalal & Triggs, 2005
HOG

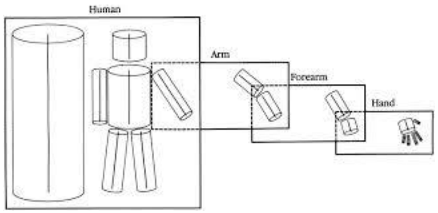
Everingham, 2012
PASCAL Challenge

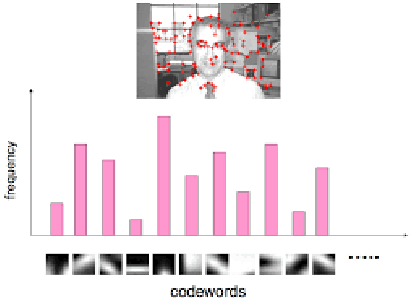
Felzenswalb & Ramanan, 2009
Deformable Part Model

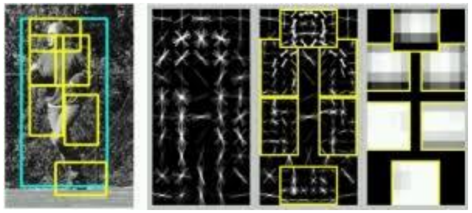
David Lowe, 1999
SIFT

David Marr, 1970s
from images to geometric
blobs, edges, 3-D models

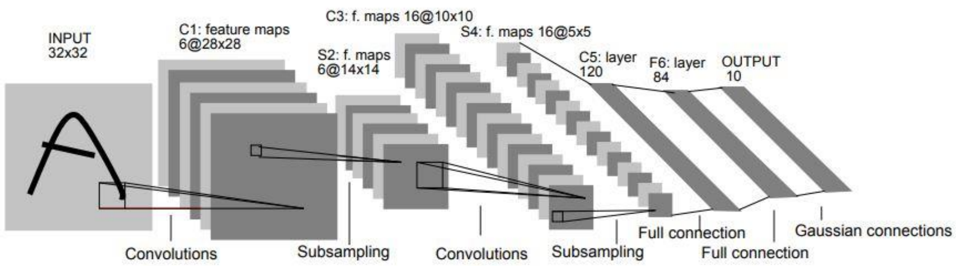
Sivic & Zisserman, 2003
Bags of words

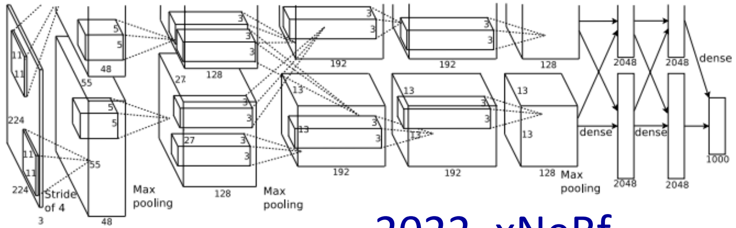






A brief history of Deep Learning





2022, xNeRf
Diffusion M

Minsky & Papert, 1969
perceptron

LeCun, Bengio, 1998
LeNet-5
Gradient-based learning

Krizhevsky, Hinton, 2012
AlexNet

NAS, 2018
3D Conv, GAN

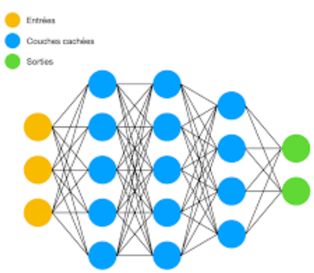
2023
Foundation M
Visual prompt
Adapters

LeCun, 1990
convolutional networks

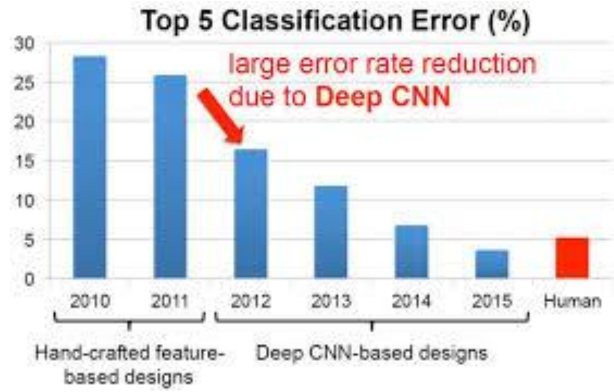
Li Fei-Fei, 2009
Image-net
22K categories and 15M images

Ross Girshick, 2016
Faster RCNN, ResNet

2020
Transformers
DETR



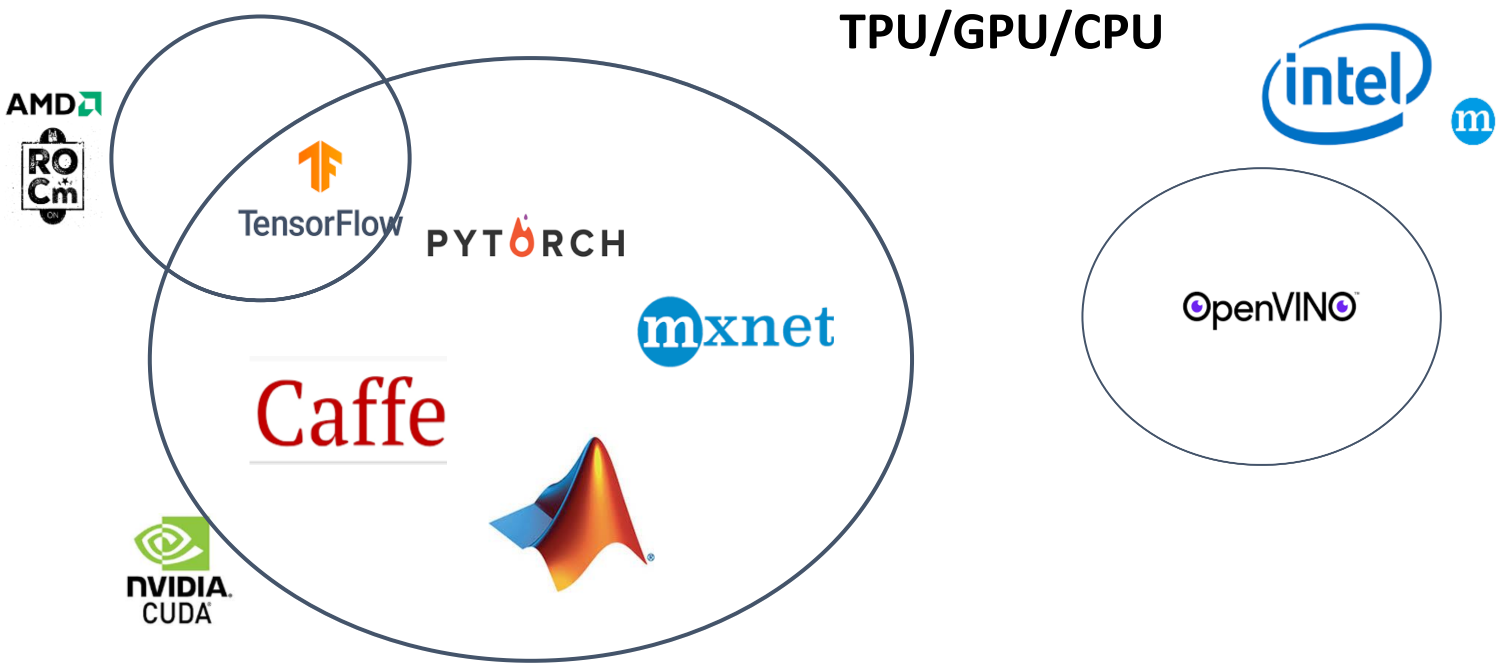
ImageNet Large Scale Visual Recognition Challenge
Russakovsky et al. IJCV 2015



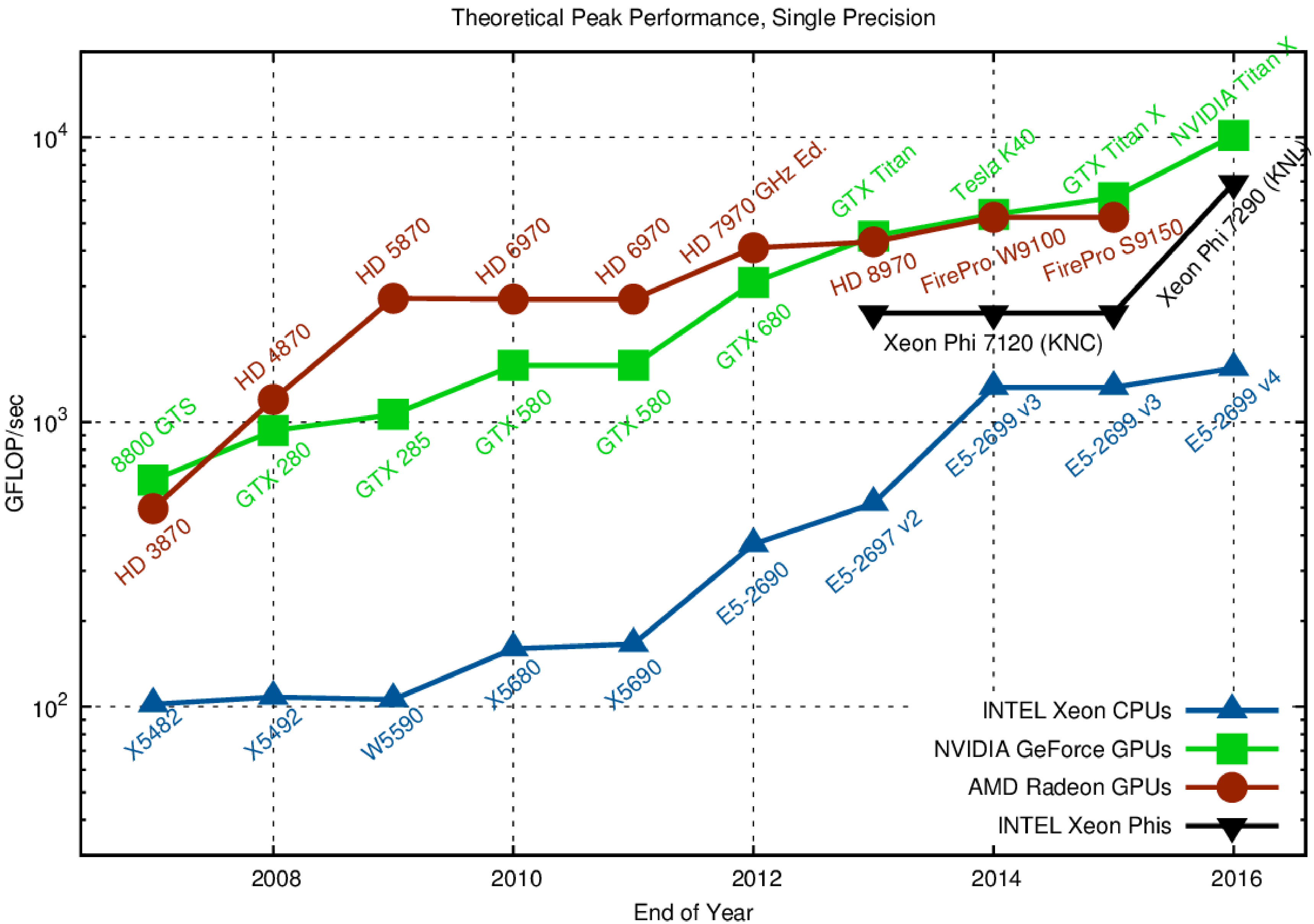
Components for Deep Learning

3 Components for Deep Learning:

- Hardware: High Computation
- Software: Deep Learning Algorithms, Libraries
- Data : Images, Videos, Annotation

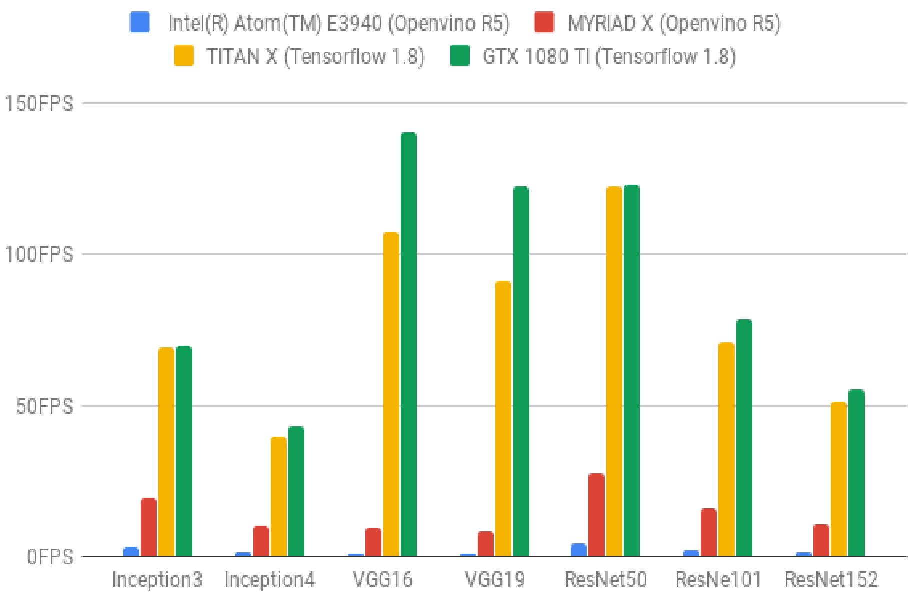


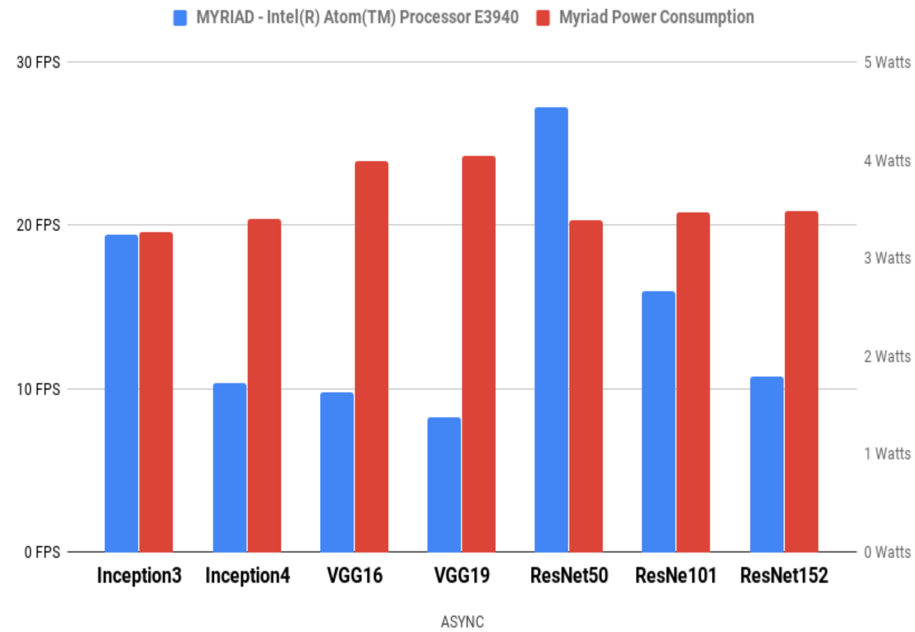
Deep Learning Hardware



Deep Learning Hardware

Large variety of GPUs for various needs:





Limitations on Nvidia Deep learning on Embedded hardware

- Power consumption : GTX 1080: 250 W > Myriad X: 5 W
- Only 3 years of Warranty (at least 8 years needed)

Deep Learning Software

Libraries (high level API)

- Caffe — (Berkeley Vision Lab)
- **TensorFlow** — (Google)
- CNTK — (Microsoft) - discontinued
- Torch — (Facebook) - discontinued
  - **PyTorch** — (Facebook/Meta)
- Theano — (MILA) – discontinued
- MXNet – Apache Software Foundation
- built on top of other libraries:
  - **Keras** — (Individual initiative + Google push)

Networks/Architectures

A neural network consisting of convolutional or recurrent layers or both, which extracts features from an image/video.

- VGG16, Alexnet,
- Siamese, U-Net, HourGlass, VAE, [coupled networks]
- RNN, GRU, LSTM
- ResNet, Inception, Inception-Resnet, DenseNet, [parallel branches, bottleneck, skip conn., residual link]
- I3D, 3DResNet, R(2+1)D, 3D-DenseNet, ResNeXt, [ST separation, channel group]
- Videos: TCN, Slow-Fast, FPN
- NAS: AssembleNet
- GAN, Diffusion Models
- Transformers: ViT, ViViT, Swin

Models/Framework

A complete end-to-end system performing a well-defined vision task

- FRCNN, Mask-RCNN; SSD, YOLO, RetinaNet (detection/segmentation),
- FCNN (Fully Convolutional, segmentation)
- DinoV2, CLIP, GroundedDino, VideoMAE

Data : machine learning

Machine Learning : Data-Driven Approach

- Collect a dataset of images and labels – expansive – to be curated
- Use Machine Learning to train a classifier [training&validation] risk of overfitting
- Evaluate/test the classifier on new unseen images [testing/inference]

Machine Learning : Few Paradigms

- supervised learning
  - Learn to map an input (data) to known labels (ground-truth), which can be discrete (classification) or continuous (regression)
  - Transfer learning: pre-training + finetuning - linear evaluation
- unsupervised learning
  - Learn a compact representation (i.e. distribution) of the data that can be useful for downstream tasks
    - Methods: density estimation, clustering, sampling, dimension reduction,
    - but in some cases, labels can be obtained automatically, transforming an unsupervised task to supervised
  - Domain Adaptation: labels for a source domain, but few or no labels for the target domain
  - Domain Generalization: life-long learning, unknown target domain (runtime)
  - Self-Supervision: a form of unsupervised learning (generic) where the data with a pre-task (reconstruction) provides the supervision, normalization, regularization (add constraints, penalty)
- semi-supervised
  - Semi (partial, zero-one-few-shots) - weakly supervised (vague or ambiguous/noisy labels),
- reinforcement learning
  - learn to predict the next actions, supervised by rewards.

Data : machine learning

Image DataSets - Challenges

- CIFAR10 (CIFAR100, MNIST)
  - 10 classes/ 50,000 training images/ 10,000 testing images [1998 - 2006]
- Pascal VOC
  - 20 object categories, 11.5K images, detection + segmentation [2006 - 2012]
- Image-net - ILSVRC
  - 22K categories and 15M images; (subset) 1K categories and 1.2M images [2009 – 2012]
- MS COCO
  - 90 object categories, 183 K images, detection + segmentation + keypoints [2014]
- OpenImages
  - 600 object categories, 1.7 – 10 M images, detection – weakly annotated [2018-2019]

Video DataSets

- Kinetics
  - 400-600-700 action classes, 325-650K video clips [2017-2019]
- ActivityNet-200
  - 200 action classes, 20K untrimmed videos, 31K action instances [2016]
- MSRDailyActivity3D:
  - 16 action classes, 320 video clips [2012]
- NTU RGB+D
  - 60/120 action classes, 56880/120K videos [2016/2019]
- Toyota Smarthome
  - 31/51 action classes, 16129/536 videos, 41K action instances [2019/20]

STARS Inria Research Team

Objective: designing vision systems for the recognition of human activities

Challenges:

- Perception of Human Activities : **robustness**
  - Long term activities (from sec to months),
  - Real-world scenarios,
  - Real-time processing with high resolution.
- Semantic Activity Recognition : **semantic gap**
  - From pixels to semantics, uncertainty management,
  - Human activities including complex interactions with many agents, vehicles, …
  - Fine grained facial expressions, rich 3D spatio-temporal relationships.
- Learning representation: **effective models**
  - Combining Multi-modalities: RGB, 2D/3D Pose, Flow, bio-signals, voice, …
  - Cross spatial and temporal dimensions : LSTM, TCN, Transformers, …
  - Using learning mechanisms: fusion, multi-tasks, guided-Attention, Self-Attention, Knowledge Distillation, contrastive learning,
  - In various learning modes : supervised, weakly-supervised, cross-datasets, unsupervised, self-learning, life long learning
- **Applications** : Safety & Health (CoBTeK from Nice Hospital : Behavior Disorder)

People Detection in real world situations

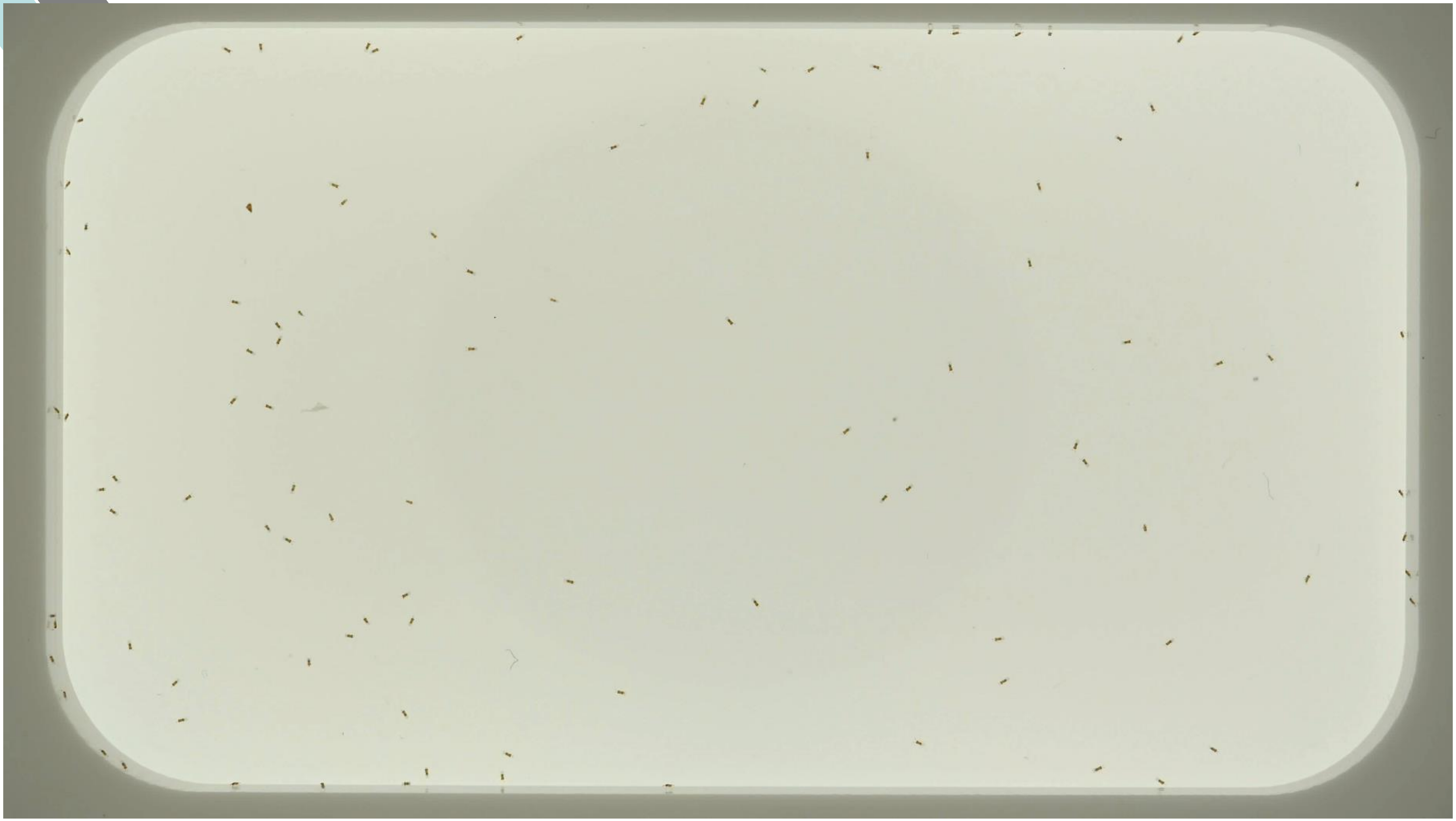
People Tracking in real world situations

People Tracking in real world situations

People Tracking and Segmentation on MOT

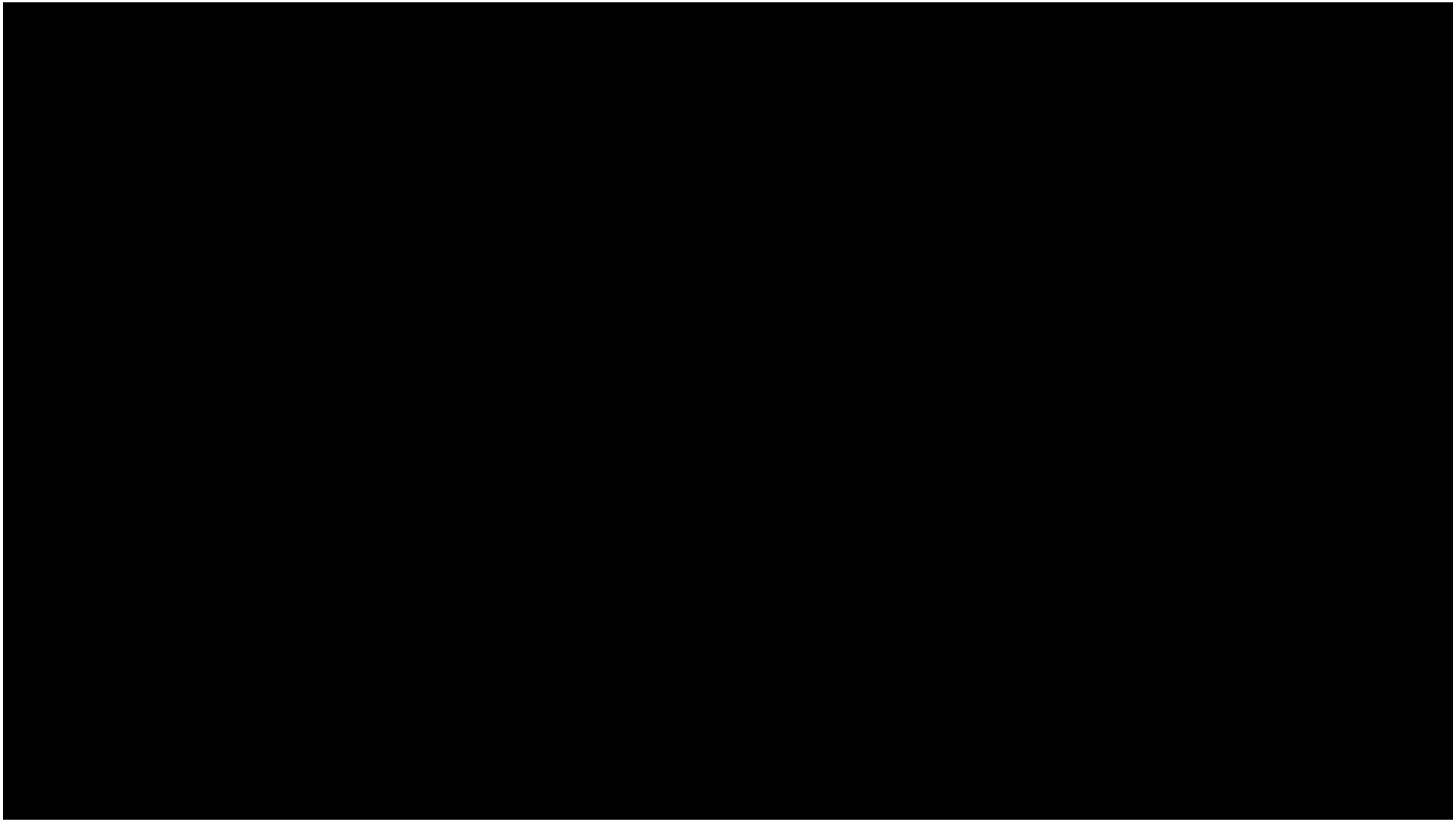
Grounded DINO + Segment Anything (SAM) + Track Anything

Analysis of trichogramma behavior with video tracking



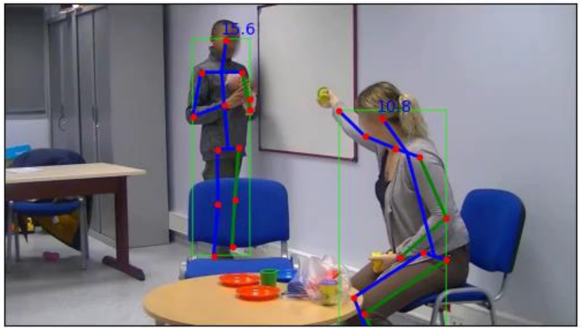
Activity monitoring at ICP with AD patients

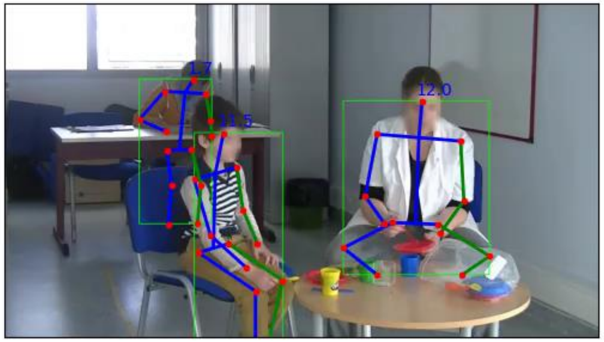
Visualization of older adult performance while accomplishing the semi-guided tasks.



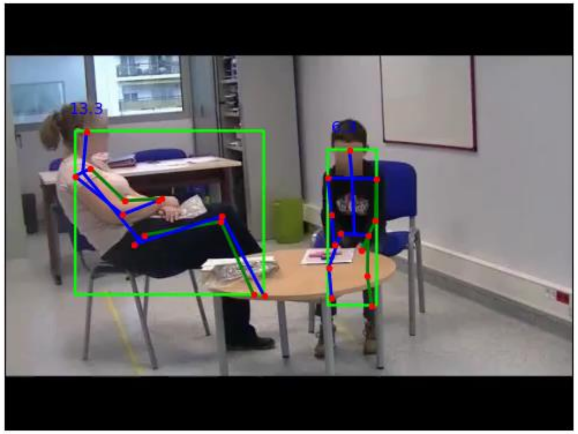
ACt4autism: children behavior

Objective quantification of atypical behaviors (stereotypies) on which the diagnosis of autism (ADOS) is based.





- Analysis of the atypical postures of the child with ASD.
- Global analysis of the movements of the child with ASD with agitation.
- Eye tracker analysis to measure joint attention.



Toyota Smart-Home
Large scale daily living dataset

Example 1

Challenges :

1. Composite Activities
e.g. Cook
3. Low Camera Framing
e.g. Dump in Trash

Person 02

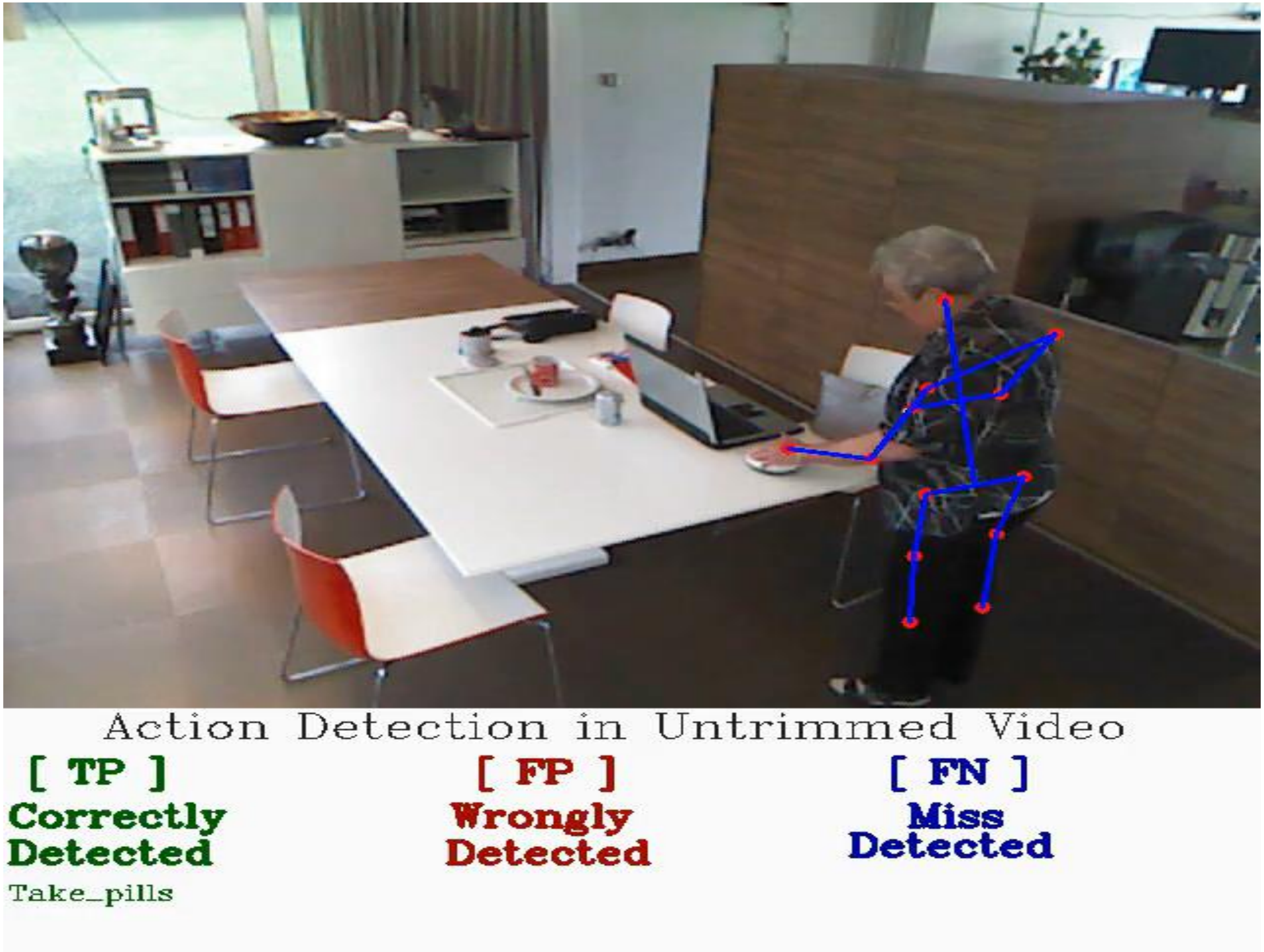
Camera 03

Frame 2379

Annotated Activities By Category

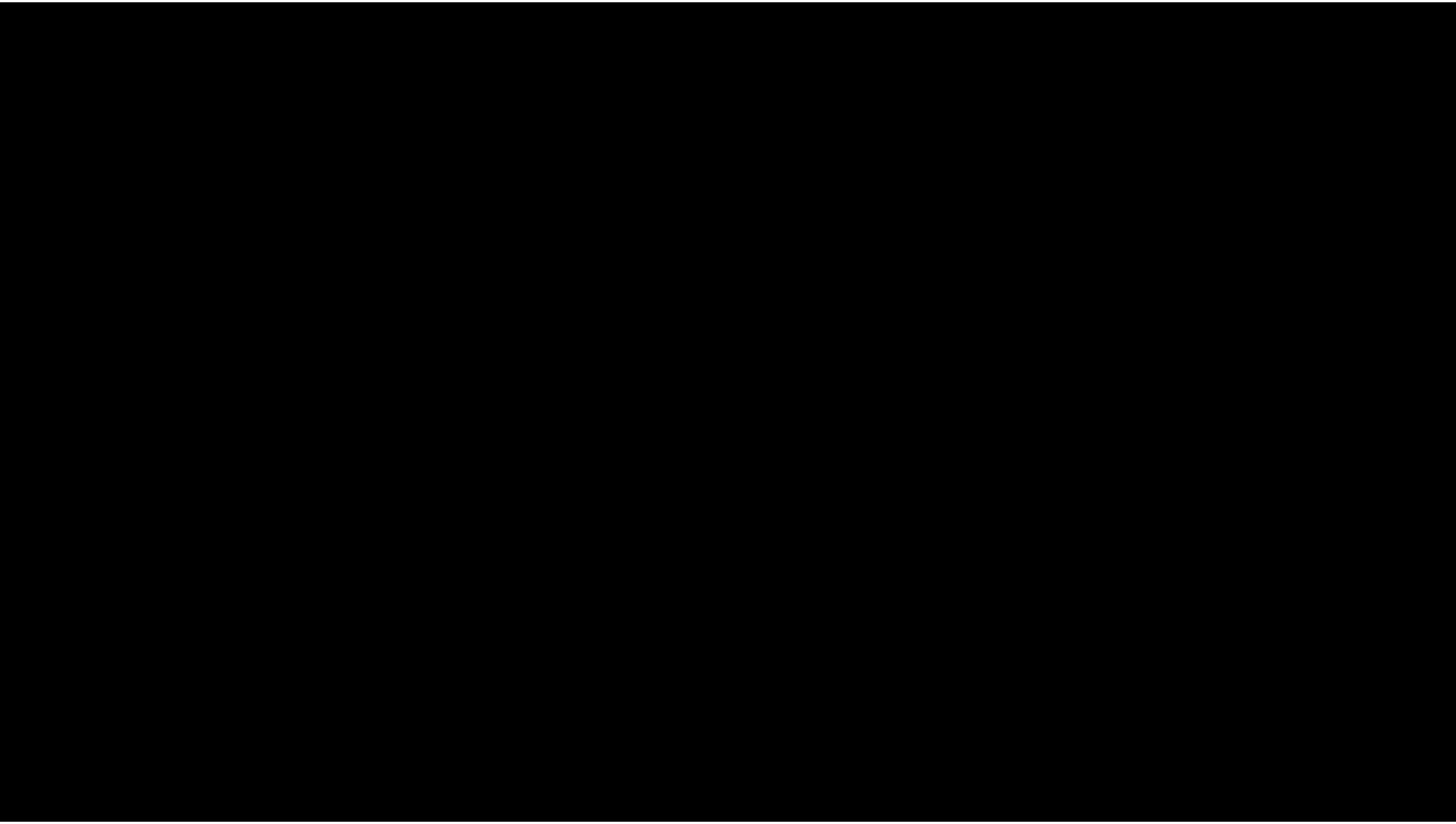
Single	**Composite & Elementary**	**Object−based**
Take_sth._off_table Walk	Cook	

Toyota Smart-Home
Large scale daily living dataset



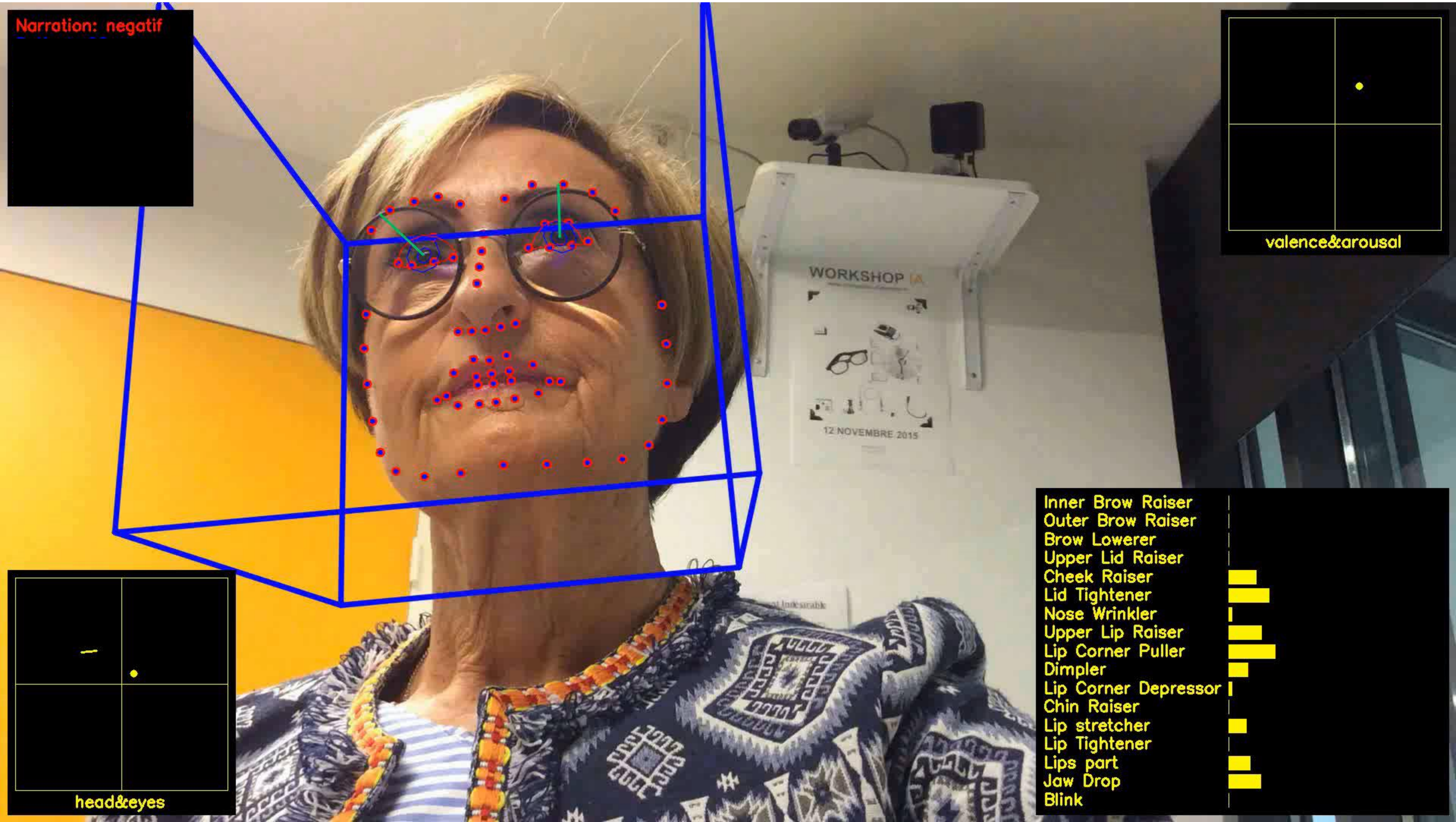
Praxis and Gesture Recognition

(short demo)



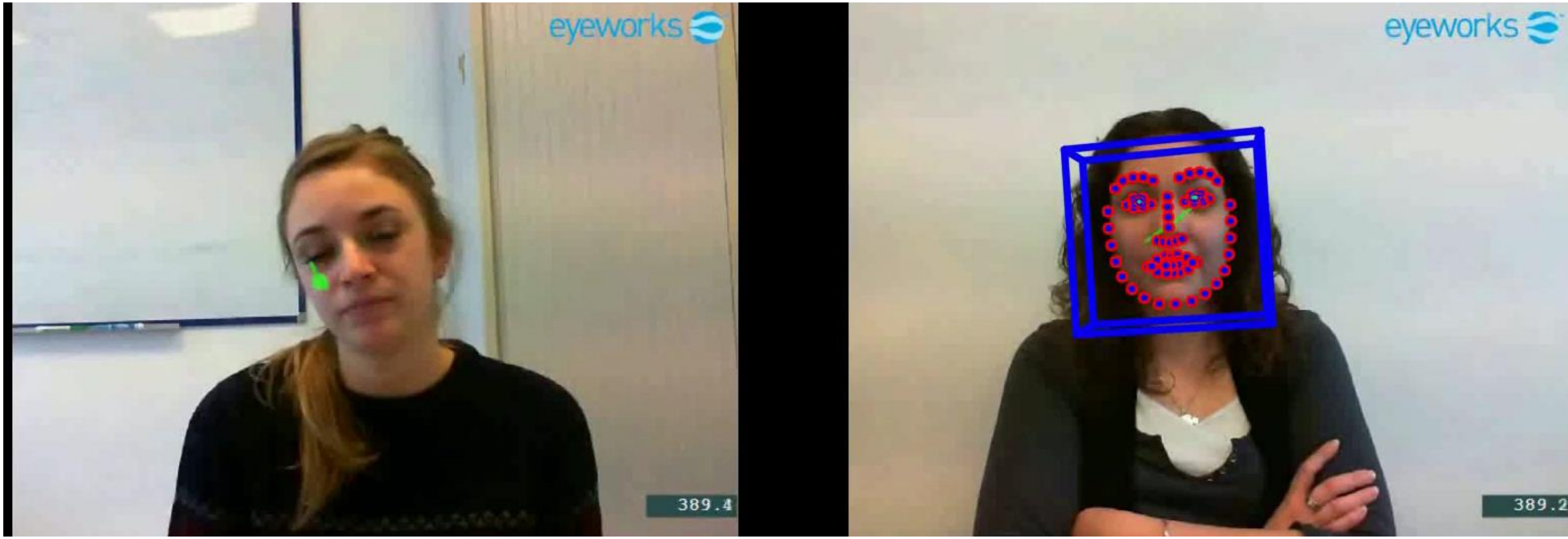
Emotion Recognition : Facial Expression Recognition

Characterizing the state of Apathy using Facial Motion and Emotion



Emotion Recognition : gaze estimation

Characterization of gaze (attention) during speech: case of schizophrenia (rupture of content).



Green dot: eye tracker

Video generation to increase facial expressions

Related Courses @ UCA

MSc Data Science and Artificial Intelligence

http://univ-cotedazur.fr/en/idex/formations-idex/data-science/

Master 1:

- Statistical Learning
- Data visualization
- Machine Learning Algorithms
- Introduction to Deep Learning, more on Deep Learning

Master 2:

- Statistical Learning
- Machine Learning - Bayesian
- **Advanced Deep Learning :** Deep understanding of Deep Neural Networks and the latest deep neural architectures in practice of these recent architectures

Educational Objectives:

- Discuss well-known methods from low-level description to intermediate representation, and their dependence on the end task
  - Focus on recent, state of the art methods and large scale applications
  - Study a data-driven approach where the entire pipeline is optimized end-to-end, jointly in a supervised fashion, according to a task-dependent objective
- Implement them to get insight on the inner deep learning mechanisms
- Implementation issues in DL are crucial:
  - Programming language support
  - Documentation quality
  - Community support
  - Learning curve
  - Stability
  - Speed
  - Scalability (multi-GPU, distributed)

Course Planning

Each session : lecture (theoretical) + practice

- **Lecture 1:** Introduction to CV : Francois + Tomasz
  - Traditional and modern Computer Vision & Artificial Intelligence [FB]
  - Neural Networks for CV : Image Classification [TS]
  - Practice: Image Classification with Pytorch
- **Lecture 2, 3:** Object Detection/Tracking : Tomasz
  - Object detection techniques will include Faster-RCNN, YOLO and ByteTrack, Sushi.
  - Each will be deeply described and compared.
- **Lecture 4:** Video and Action Classification, LSTM, TCN, Transformer : Snehashis
- **Lecture 5:** Action Detection and Anticipation: Snehashis
  - Dense Trajectories, different video aggregation techniques, two-streams, LSTMs for AR, 3D ConvNets
  - Attention Mechanism : spatial attention for image classification, spatio-temporal attention for action recognition, Transformer.
- **Lecture 6:** Image and Video Generation (Diffusion Models) : Seongro Yoon
- **Lecture 7:** Foundation Models : Mahmoud
- **Lecture 8:** Article presentation : all

How to Contact Us

- Course Website:
  - http://www-sop.inria.fr/members/Francois.Bremond/MSclass/deepLearningWinterSchool25/UCA_master/schedule.html
  - Syllabus, lecture slides, schedule, etc

- Emails:
  - Tomasz Stanczyk: tomasz.stanczyk@inria.fr
  - Seongro Yoon: seong-ro.yoon@inria.fr
  - Snehashis Majhi : snehashis.majhi@inria.fr
  - Mahmoud Ali: mahmoud.ali@inria.fr
  - Francois Bremond: francois.bremond@inria.fr





Evaluation Policy

- Engagement while attending class (oral) : 30%
  - Answering questions
  - Practical training, assignments

- Project, Article presentation: 70%
  - 6 groups of 1 or 2 students
  - Select 1 article out of 10
  - Last day: slide presentation : 20 min + 10 min questions
    - Motivation
    - State-of-the-art
    - Proposed approach
    - Performance/limitations
    - Future directions