Video Understanding
for Human Behavior Analysis

Francois.Bremond@inria.fr

INRIA Sophia Antipolis – **STARS team**

Nice University Hospital - **CoBTeK**,

3iA Côte d'Azur
Interdisciplinary Institute for Artificial Intelligence

Video Understanding for Human Behavior Analysis

Objectives:

- to measure objectively human behaviors by recognizing their everyday activities, their emotion, eating habits and lifestyle,
- to improve and optimize the quality of life of people suffering from behavior disorders.

Method:

- Designing vision systems for the recognition of human activities
- Human behavior can be modeled by learning from a large number of data from a variety of sensors.

Video Understanding
for Human Behavior Analysis

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**
  - 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

Video Understanding
for Human Behavior Analysis

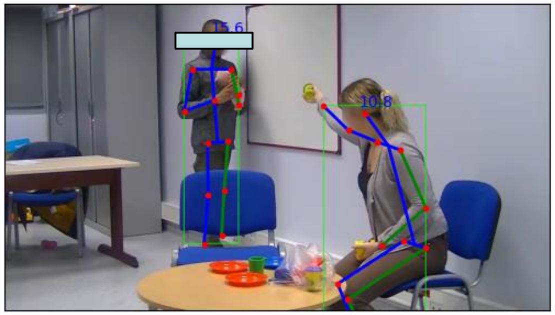
Collaboration with Nice Hospital:

- Mental health from birth to the end of life: clinical trials
  - Children: autism,
  - Adults: schizophrenia, depression,
  - Older adults: dementia, Alzheimer, frailty

Find biomarkers in videos of patient-clinician interaction

Datasets

- with sufficient annotation but with general scene
  - Kinetics, Toyota Smart Home, NTU, iMiGUE, Eyediap
- with appropriate scene but without sufficient annotation
  - MOTAP, CHU Nice, INRIA Nancy, DeepSPA
- with appropriate scene and with specific annotation
  - MPII Group Interaction, ACTIVIS, Mephesto?



3iA Côte d'Azur
Interdisciplinary Institute
for Artificial Intelligence

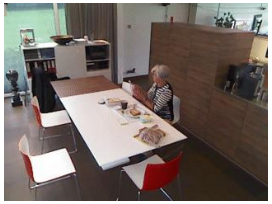
Toyota Smart-Home
Large scale daily living dataset

COMPOSITE & ELEMENTARY ACTIVITIES

Breakfast

Cut bread

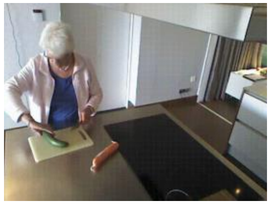
Spread butter



Take ham

Eat at table

Cook



Cut
(vegetable/meat)

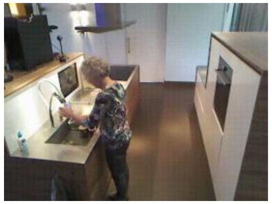
Stir

Use oven

Use stove

Clean dishes

Put sth. in sink



Clean with water

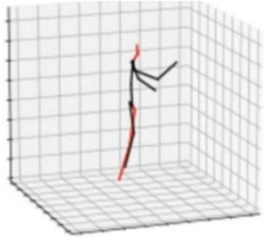
Dry up

Privileged Modalities

RGB

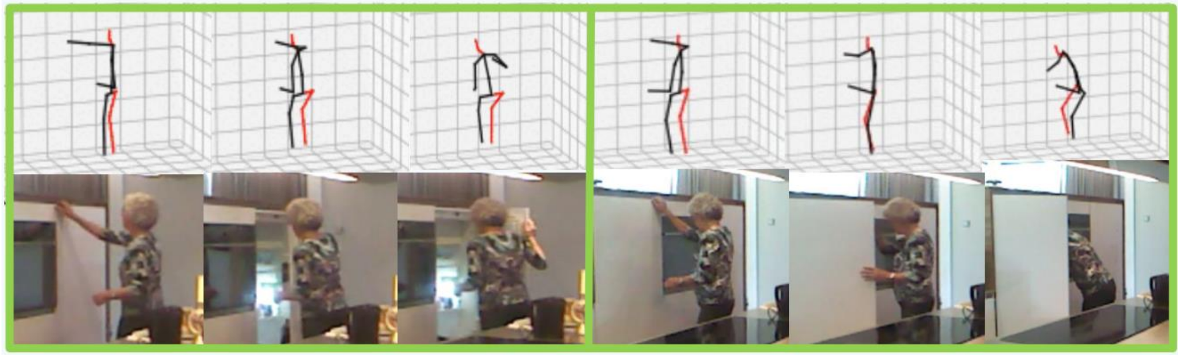
Depth

Optical Flow



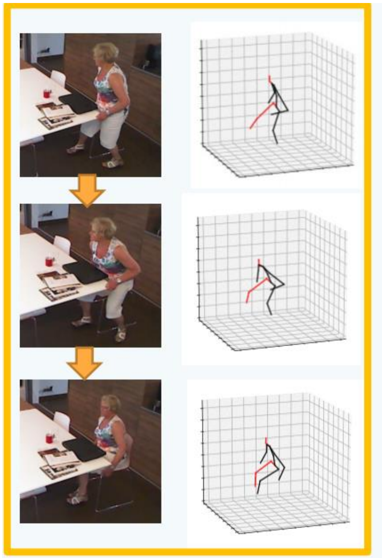
3D skeleton

Complementary Nature



Open fridge

Open cupboard



Sit down

Filtering the noisy appearance patterns

Help capturing the body motion

[ICCV 2021] Learning an Augmented RGB Representation with Cross-Modal Knowledge Distillation for Action Detection

Video Understanding
for Human Behavior Analysis

Find biomarkers in videos of patient-clinician interaction

Methods

- For detecting, tracking people, skeleton and face

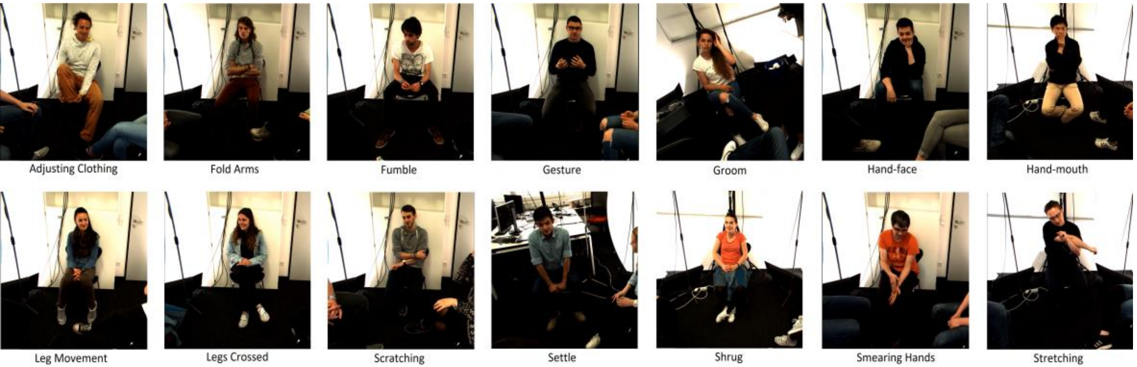
  SSD+DeepSort, YOLOvX+ByteTrack, OpenPose, OpenFace,
- for estimating gaze, head rotation, eye contact and 17 action units

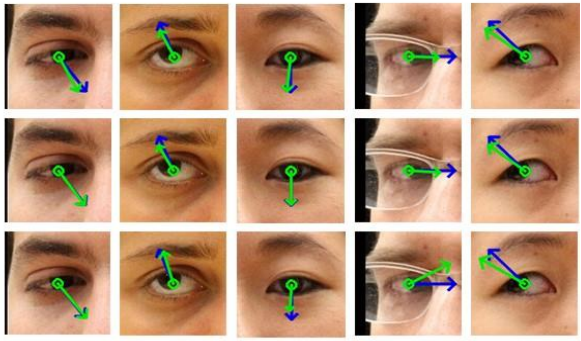
  OpenFace, FLAME
- for recognizing emotions and personality through multi-modalities (e.g. Bio-signals)

  DeepFace, MultiModalMAE, FAt Transformer
- for detecting actions and gestures

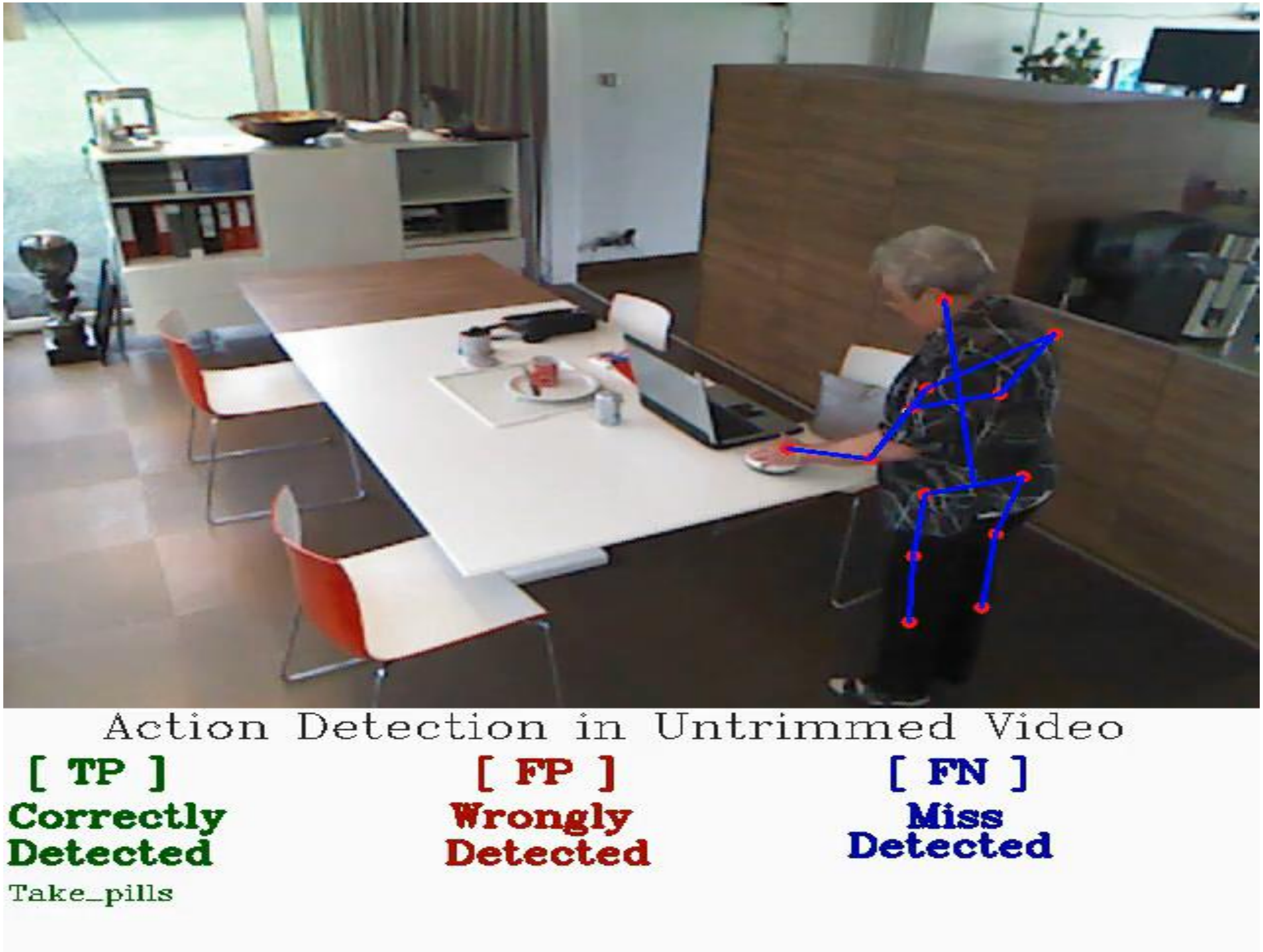
  UNIK, PDAN, VPN++, THORN, MS-TCT
- For data augmentation, anonymization, Video Generation

  G3AN, ImaGINator, LIA



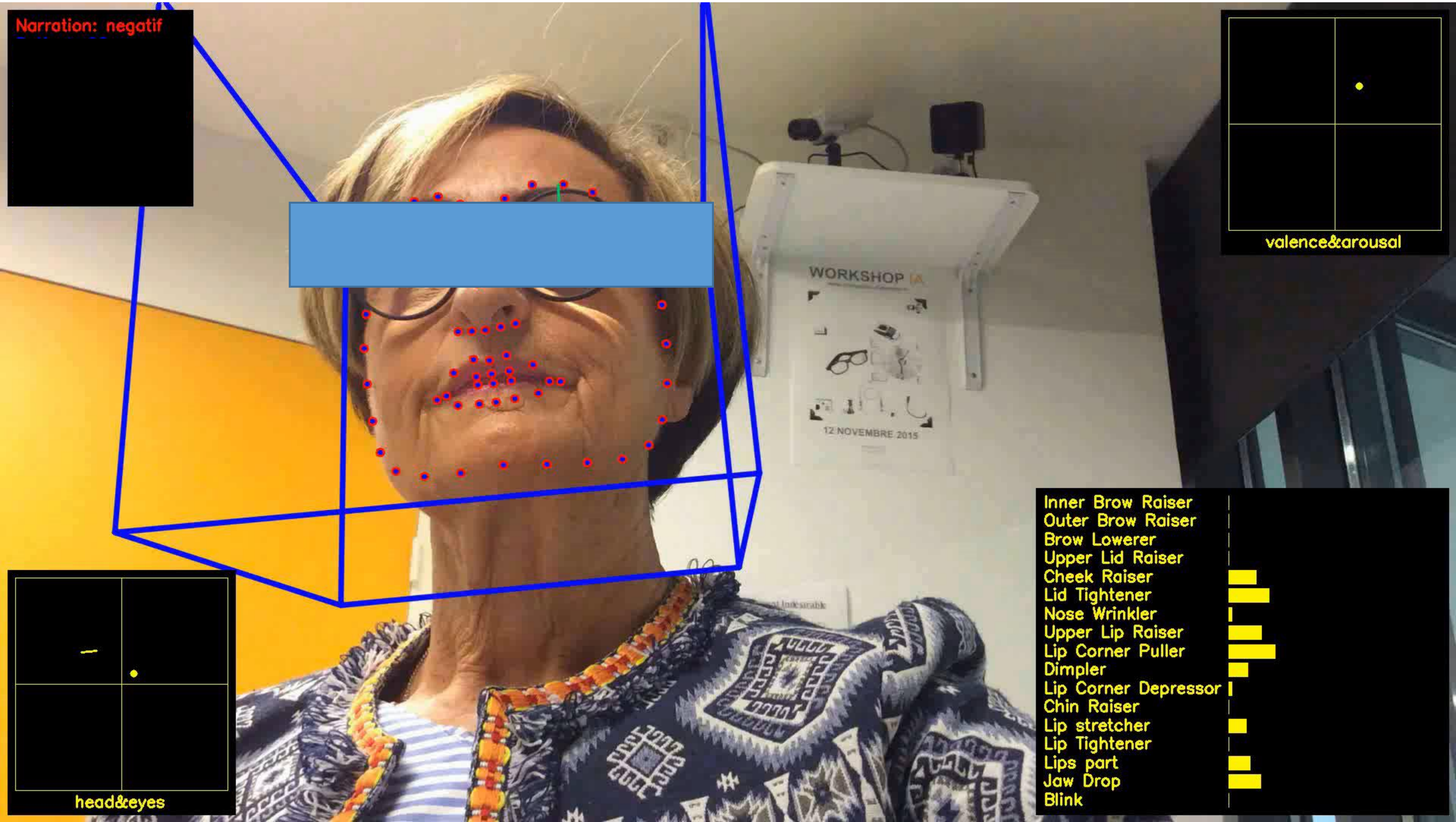


Toyota Smart-Home
Large scale daily living dataset

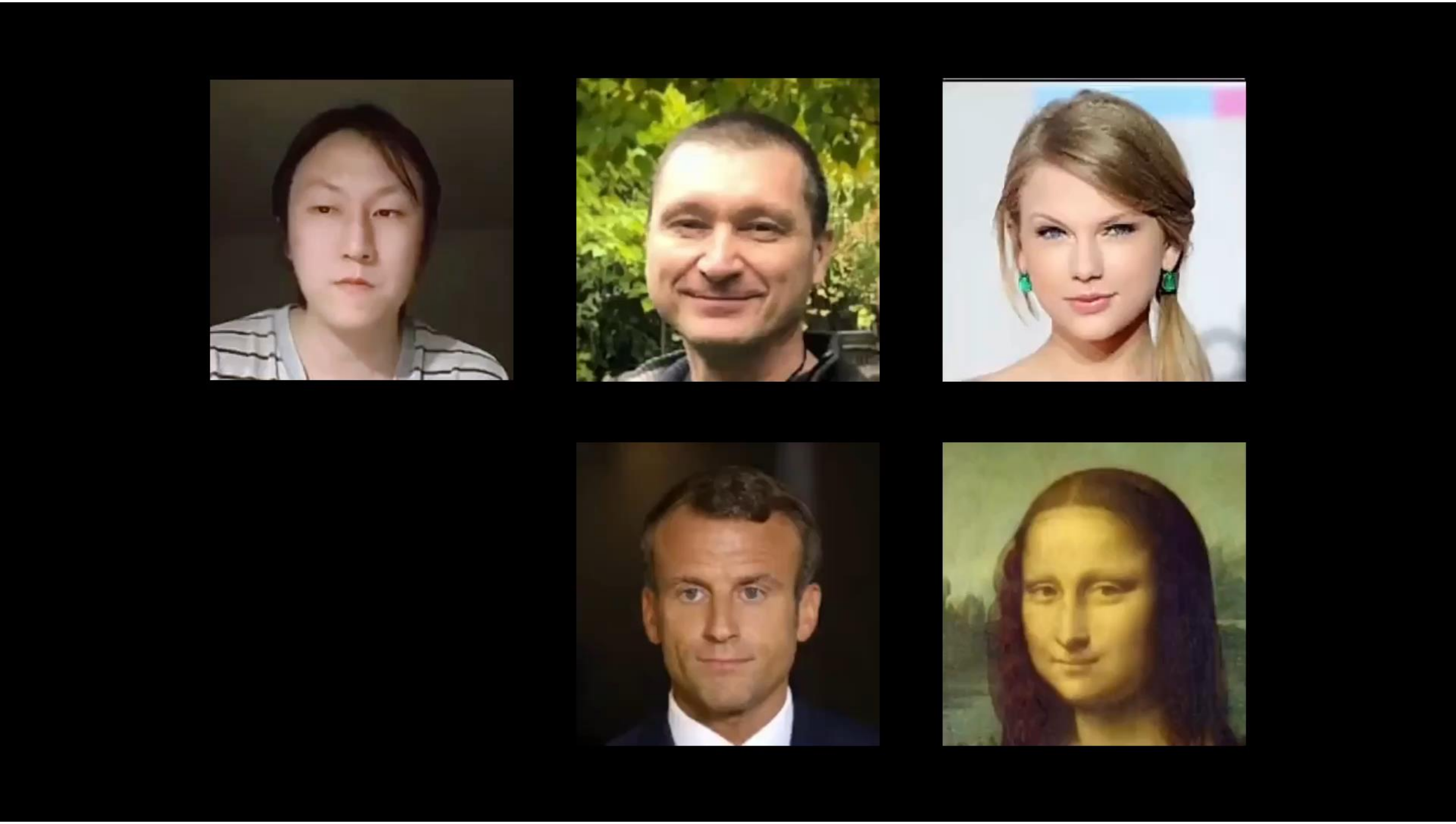


Emotion Recognition : Facial Expression Recognition

Characterizing the state of Apathy using Facial Motion and Emotion



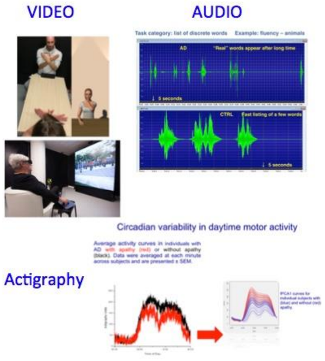
Data Augmentation : Video Generation



Conclusion – People Monitoring

A **global framework** for building real-time video understanding systems:

- **Activity Monitoring** Systems to measure levels of everyday activities: from handcrafted to (un)supervised learned models of activity
- Robust for **long term** video monitoring
- Online and real-time recognition with limited user interaction during training


VIDEO
AUDIO
Task category: list of discrete words
Example: fluency – animals
AD
"Real" words appear after long time
5 seconds
CTRL
Fast listing of a few words
5 seconds
Circadian variability in daytime motor activity
Average activity curves in individuals with AD with apathy (red) or without apathy (black). Data were averaged at each minute across subjects and are presented ± SEM.
Actigraphy
fPCA1 curves for individual subjects with (blue) and without (red) apathy

Perspectives:

- View-point invariant - Real-world settings
- Generate totally unsupervised models
- Generic semantic activity models (cross scenes), Adaptive learning
- Use finer features as input for the algorithm (head, posture, facial, hand, gesture…)
- More semantics, emotion, mental states.
- Multi-modalities (e.g. speech)
- Reaction to Stimulation : Serious Games