National Technical University of Athens (NTUA)
Intelligent Robotics & Automation Lab (IRAL) και CVSP
Athena Research Center / Institute of Robotics
https://robotics.ntua.gr

# Multimodal Robot Perception and Interaction

Petros Maragos

Keynote Talk, PETRA 2023, Kerkyra, 05 July 2023

# Human Actions & Activities



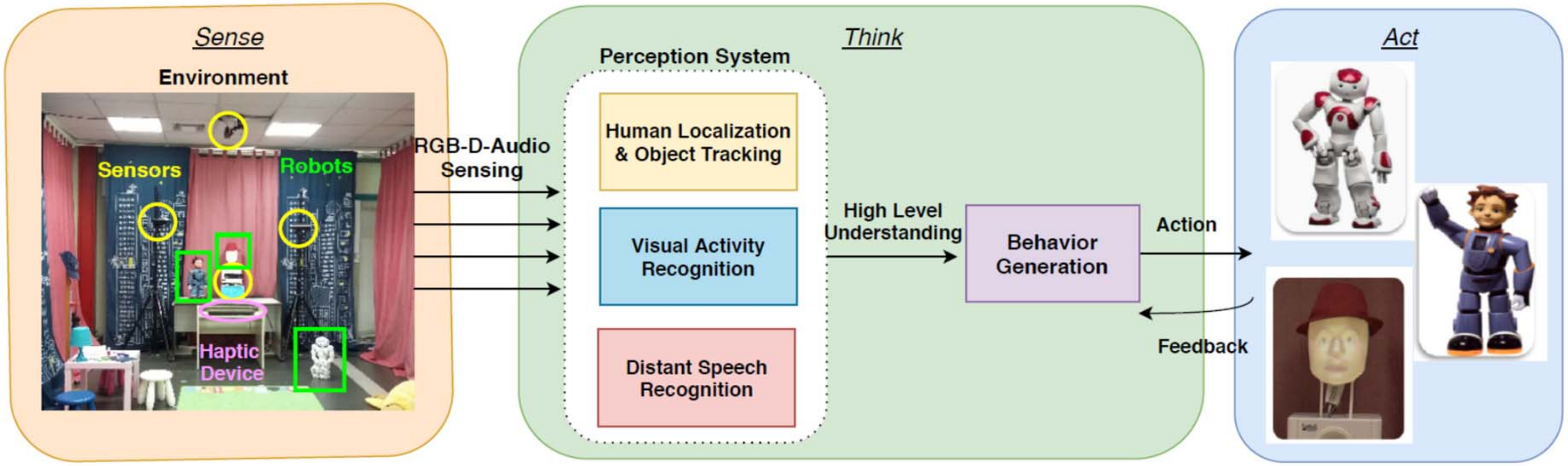
# Human Activity Recognition(HAR)

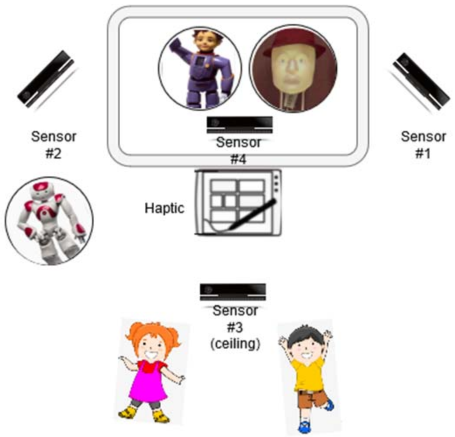
## Vision-based
- Robotics
- Surveillance

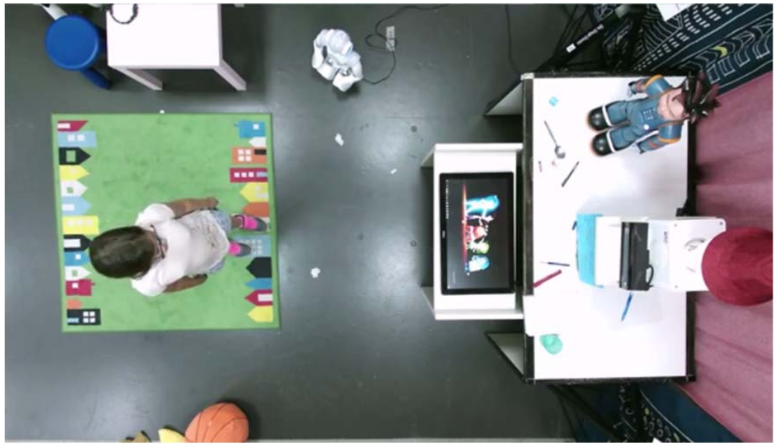
## Speech-based
- Smart Home
- Navigation

## Sensor-based
- Healthcare
- Sports

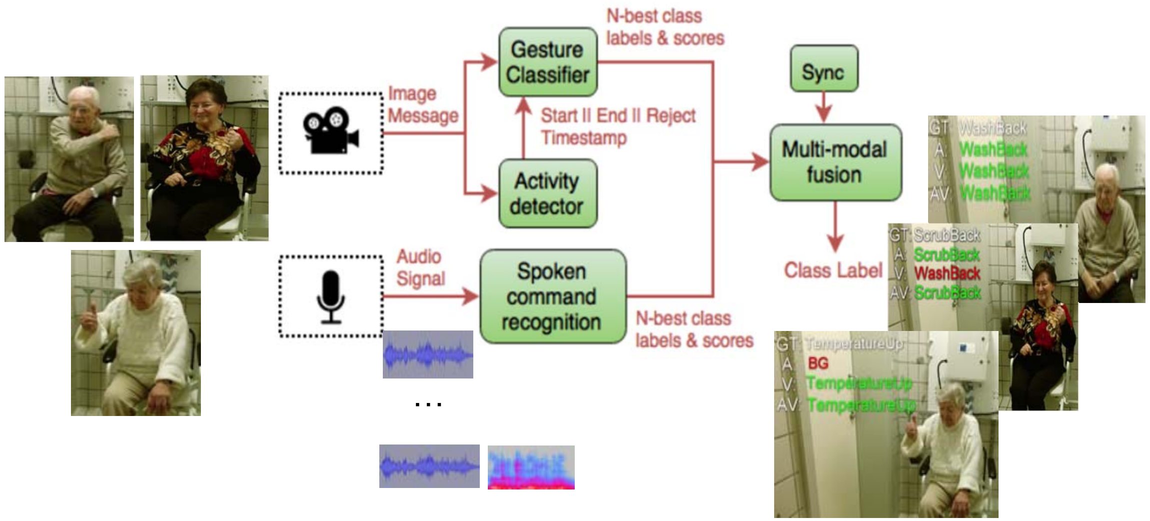
# Area 1. Audio-Visual Child-Robot Communication & Interaction



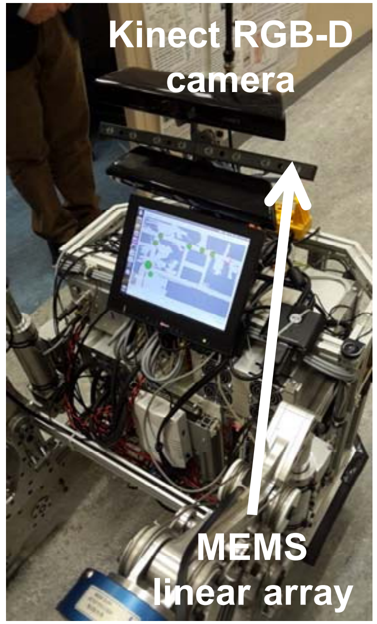



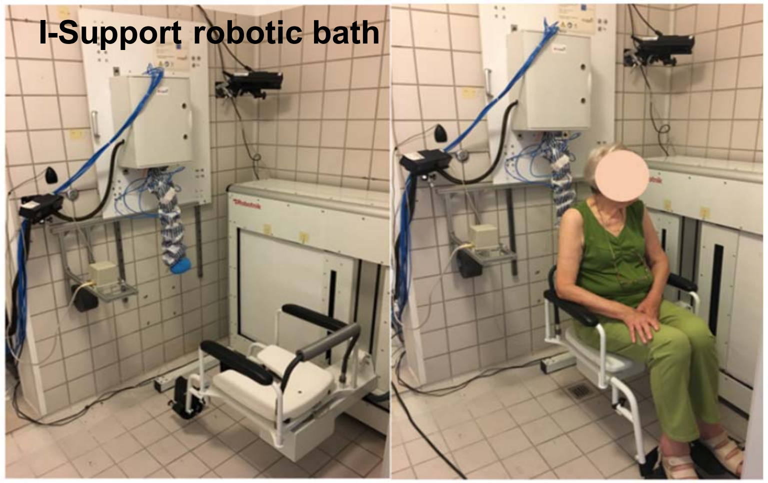
# Area 2. Audio-Visual Human-Robot Interaction in Assistive Robotics



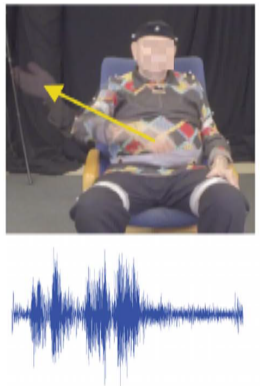

i-Walk

MOBOT robotic platform



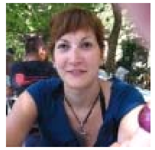



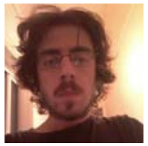

Audio-Gestural Commands



# IRAL+CVSP: members & collaborators

- **Faculty:** Petros Maragos  Costas Tzafestas  Alexandros Potamianos 

- **Post-Doc Researchers:**
      
  - Panagiotis Filntisis
  - George Moustris
  - George Retsinas
  - Antigoni Tsiami
  - Athanasia Zlatintsi

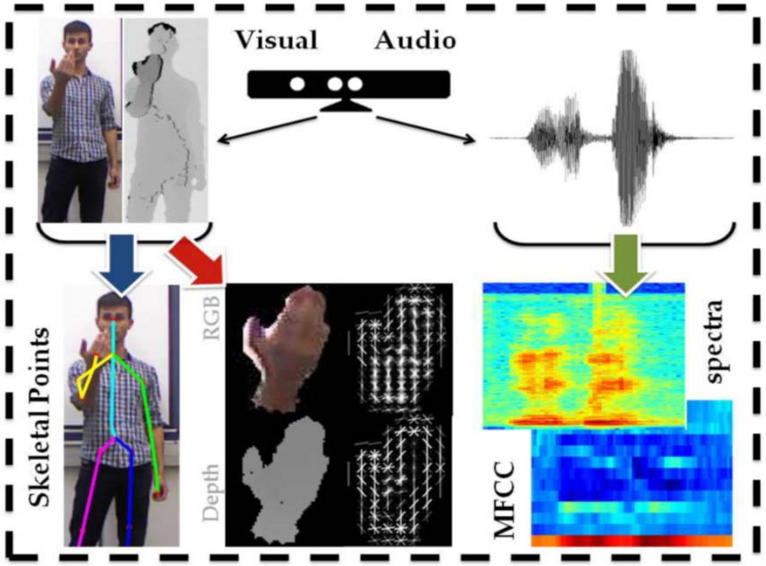
- **PhD/MEng GRAs:**
        
  - Dafni Anagnastopoulou
  - Niki Efthymiou
  - Christos Garoufis
  - Nikos Kardaris
  - Panagiotis Mermigas
  - Paraskevas Oikonomou
  - George Paraskevopoulos

- **Technical/Management Support:**
  - Despina Kassianidi, Vicky Platitsa, Fotini Stamelou

- **Collaborators:**
  - A. Dometios, G. Chalvatzaki, P. Koutras, A. Katsamanis, V. Pitsikalis, G. Potamianos

# Early Work

# Multimodal Gesture Recognition

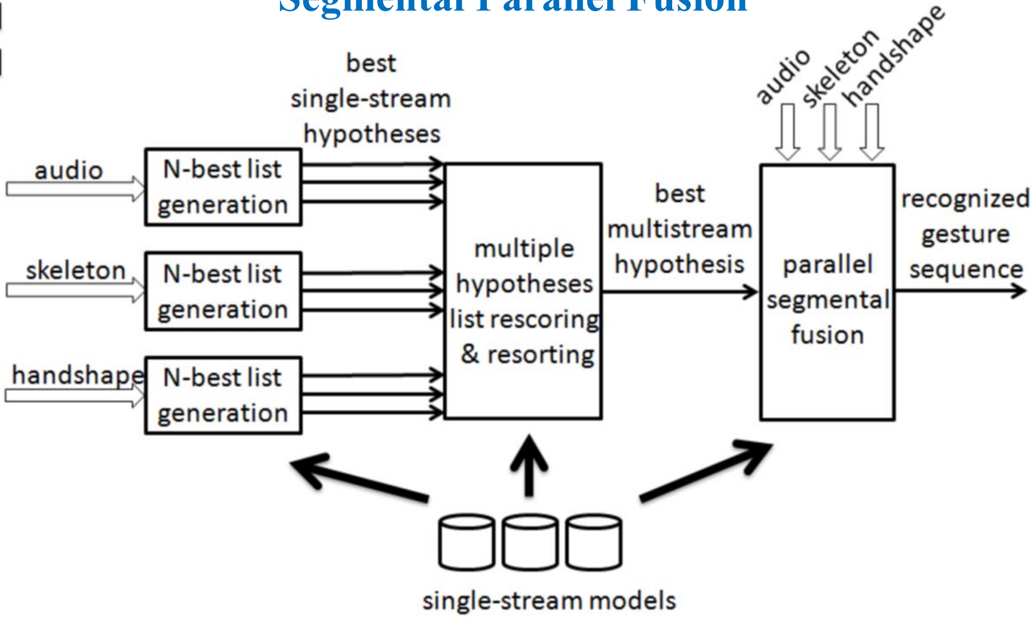



- ChaLearn Corpus: 20 signs of Italian language, 22 different users, 20 repetitions/user
  [S. Escalera et al., , "*Multimodal Gesture Recognition Challenge 2013: Dataset and results*", ACM Int'l Conf. Multimodal Interaction, 2013.]

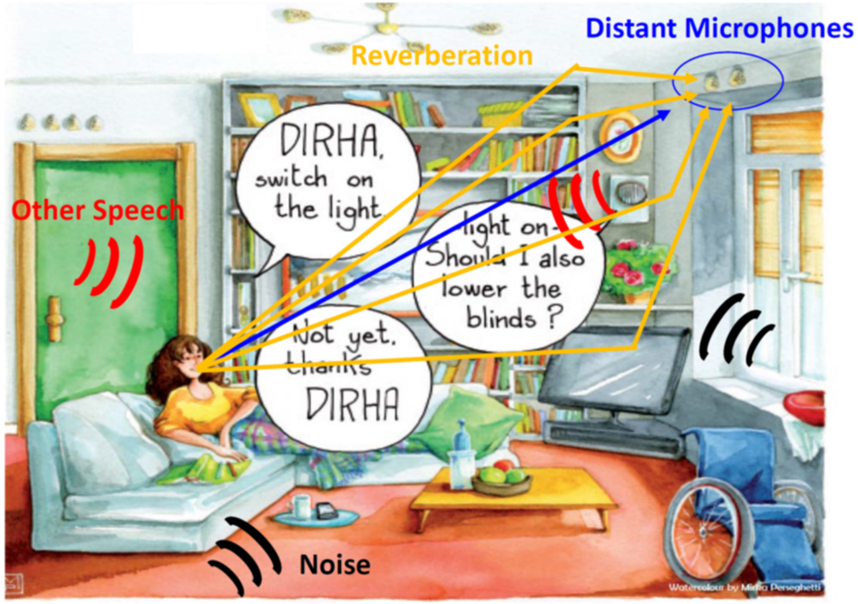
- Accuracy 93.3%: top rank in ChaLearn (ACM 2013 Gesture Challenge – 50 teams)

  [ V. Pitsikalis, A. Katsamanis, S. Theodorakis and P. Maragos, "*Multimodal Gesture Recognition via Multiple Hypotheses Rescoring*", JMLR 2015. ]

## Multimodal Hypothesis Rescoring + Segmental Parallel Fusion



## Distant Speech Recognition in Voice-enabled Interfaces



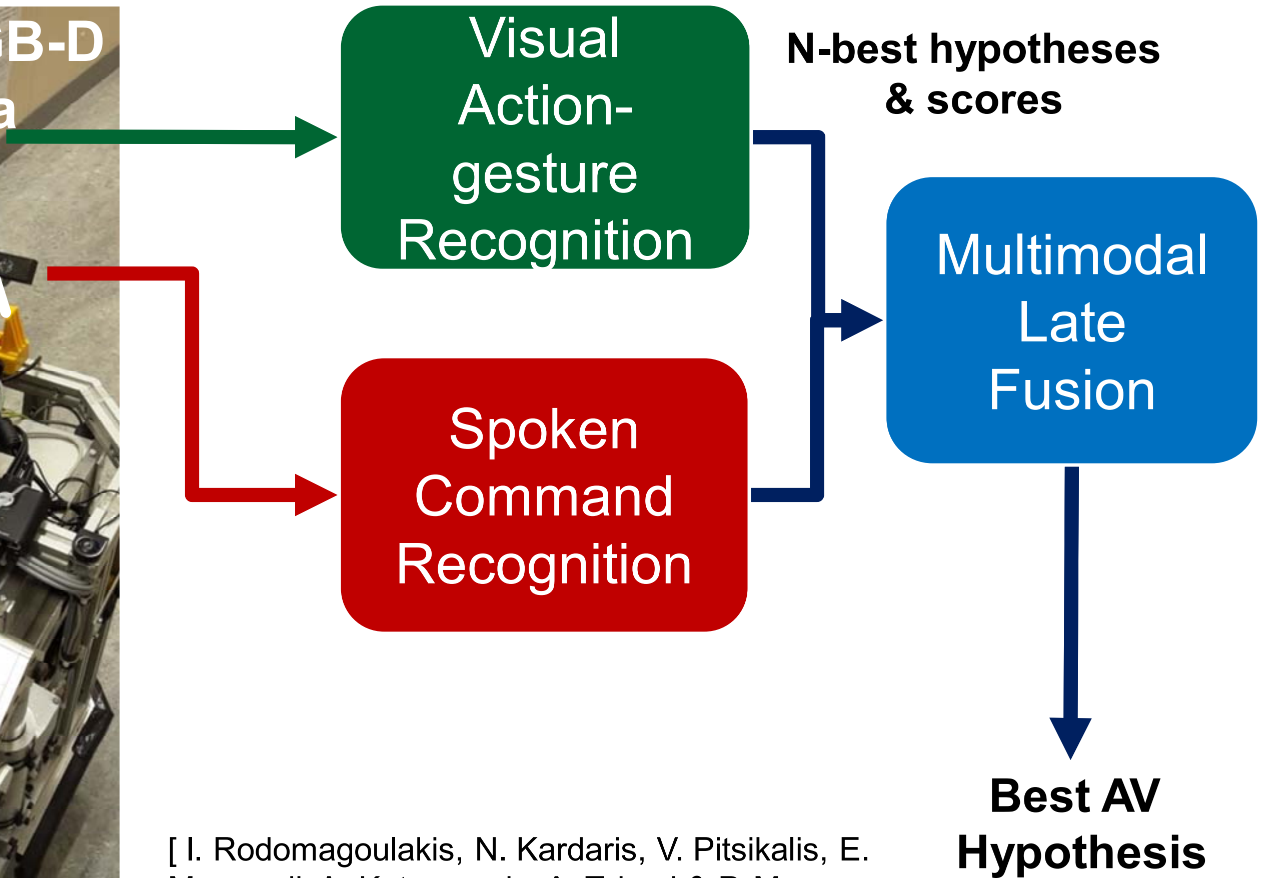

I. Rodomagoulakis, A. Katsamanis, G. Potamianos, P. Giannoulis, A. Tsiami, P. Maragos, "Room-localized Spoken Command Recognition in Multi-room, Multi-microphone Environments", *Computer Speech & Language*, 2017.

## Smart Office Demo
## ("Σπιτάκι μου")


https://www.youtube.com/watch?v=zf5wSKv9wKs

# Audio-Gestural Command Recognition in Human-Robot Interaction (HRI)

**MOBOT robotic platform**

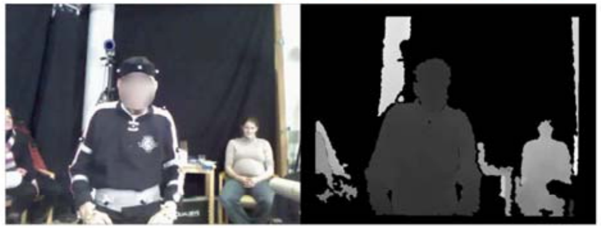

[ I. Rodomagoulakis, N. Kardaris, V. Pitsikalis, E. Mavroudi, A. Katsamanis, A. Tsiami & P. Maragos, ICASSP 2016 ]

# MOBOT: Multi-Sensor Data for Assistive Robotics

Kinect1 RGB Data

Kinect Depth Data

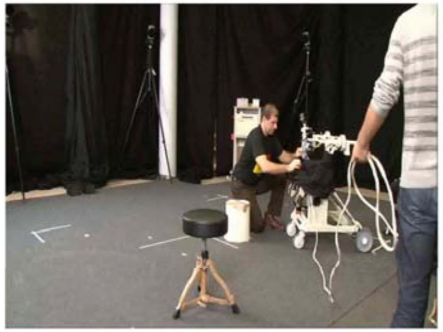
Kinect1 RGB
Kinect1 Depth
MEMS Audio Data



Go Pro RGB Data

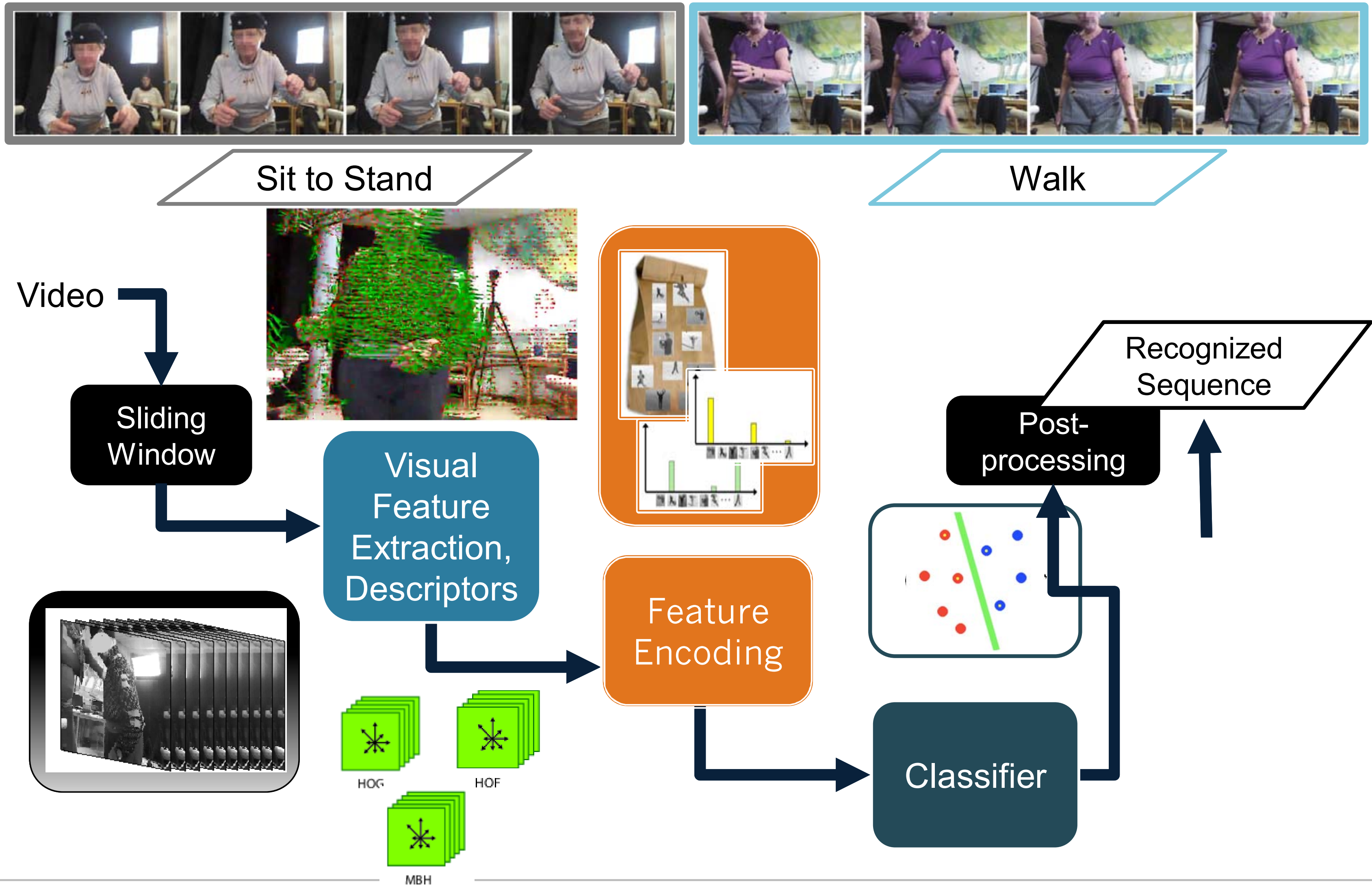
HD1 Camera Data

HD2 Camera Data

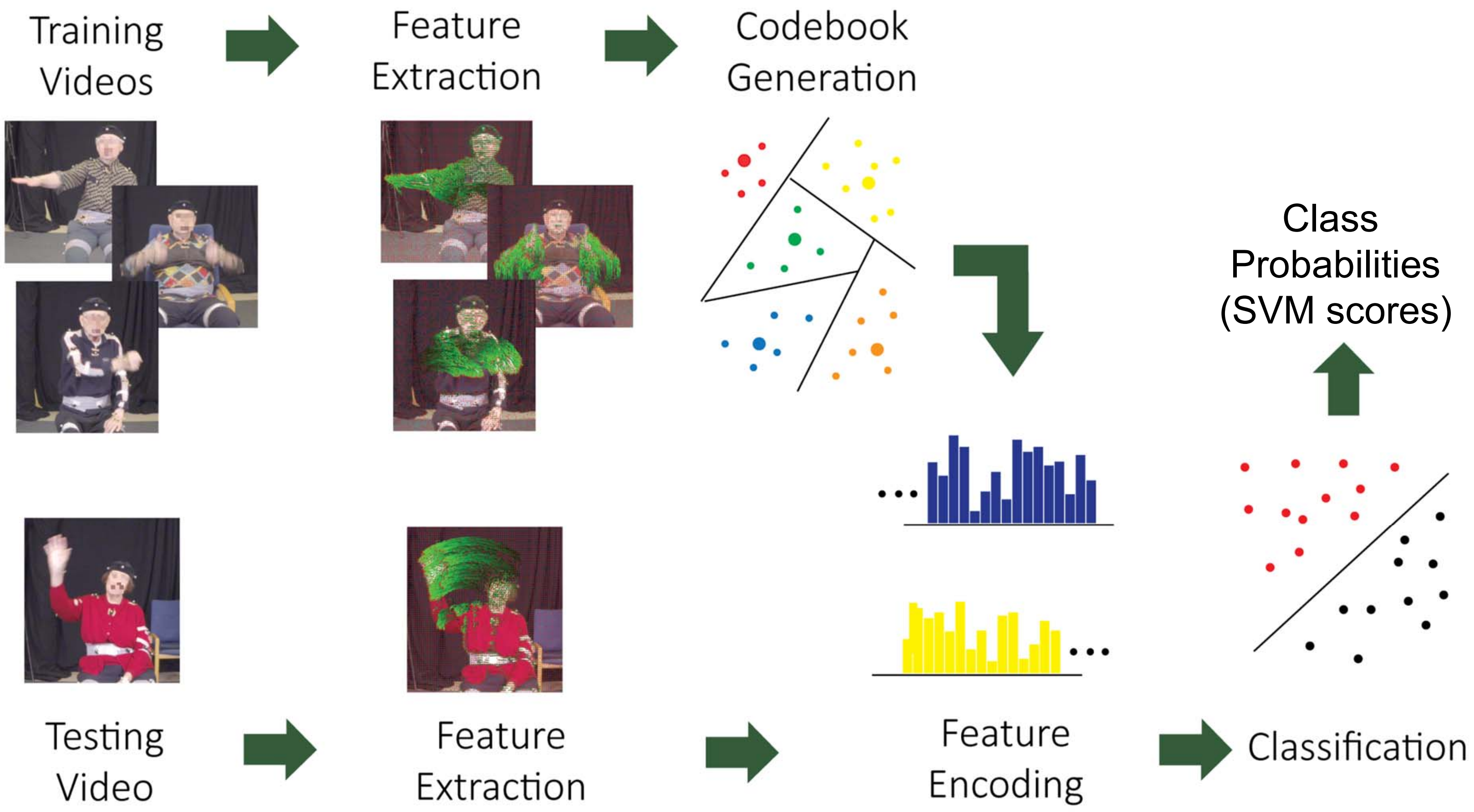


- Visual noise by intruders. Noisy acoustics (background, speakers overlap, distance)
- Multiple subjects in the scene, even at same depth level
- Frequent and extreme occlusions, missing body parts (e.g. face)
- Significant variation in subjects pose, actions, visibility, background

# Visual action recognition pipeline

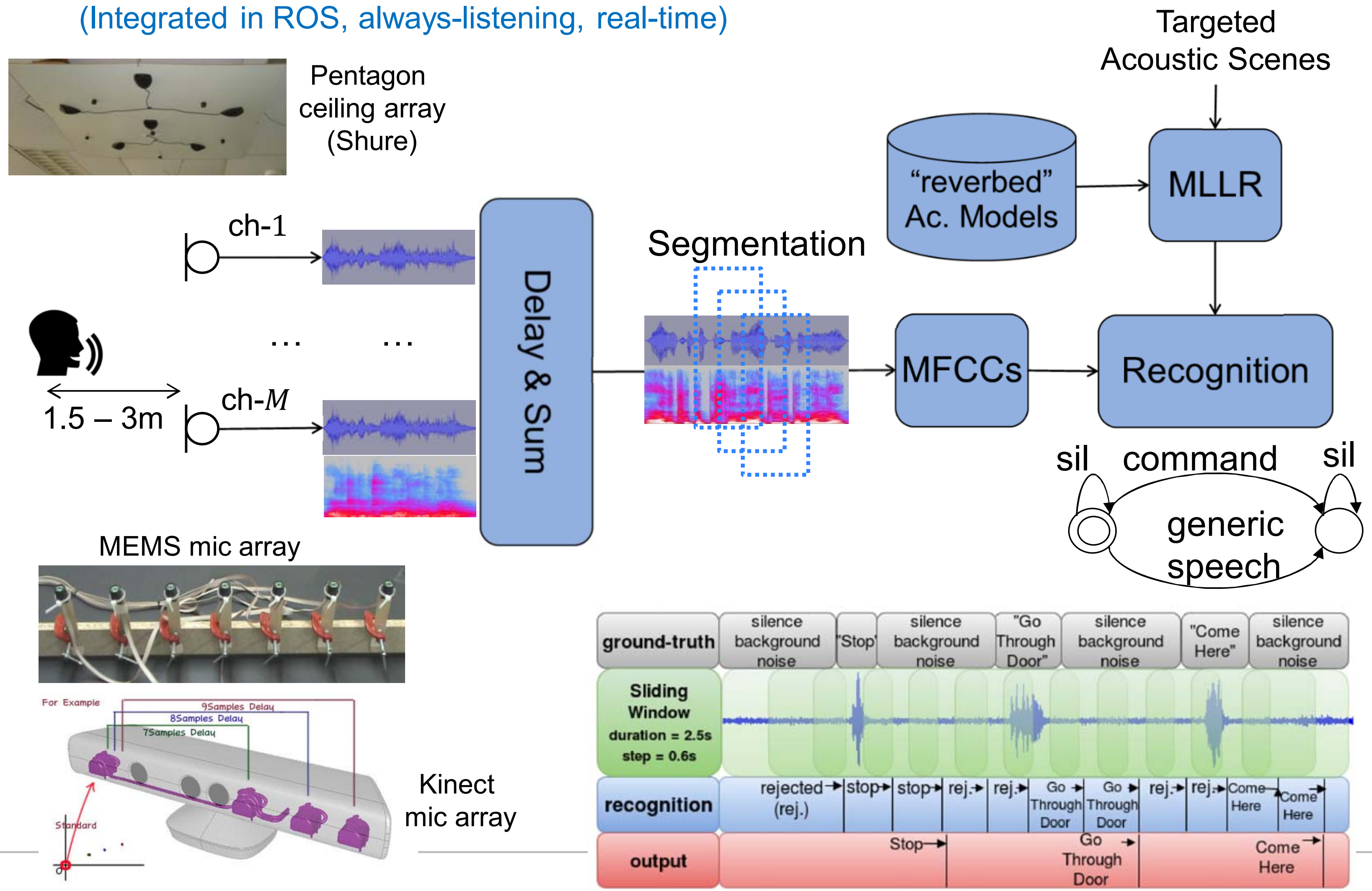


# Visual Gesture Classification Pipeline



# Online Spoken Command Recognition for HRI

- **Greek, German, Italian, English**
  (Integrated in ROS, always-listening, real-time)



# Audio-Visual Fusion: Hypotheses Rescoring

speech & gesture recognition

(Mobot-I.6a data: 8p,8g)

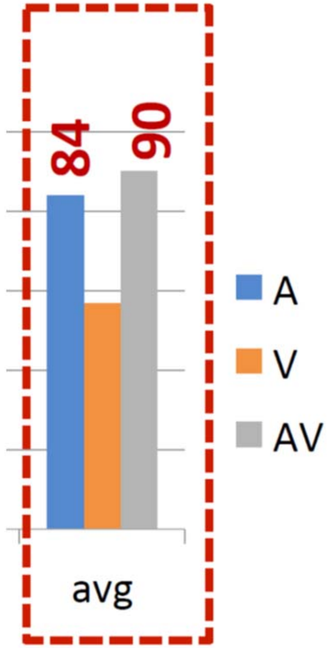
**N-best**

**spoken commands hypotheses**

| | hypothesis | normalized score |
|---|---|---|
| A1 | Help | 0.2 |
| A2 | Stop | 0.19 |
| A3 | park | 0.12 |
| | … | |
| A19 | go straight | 0.01 |

**visual gesture hypotheses**

| | Hypothesis | normalized score |
|---|---|---|
| V1 | Stop | 0.5 |
| V2 | go away | 0.15 |
| V3 | help | 0.12 |
| | … | |
| V19 | go straight | 0.01 |




**fusion hypotheses**

| | hypothesis | combined score |
|---|---|---|
| F1 | Stop | 0.205 |
| F2 | help | 0.196 |

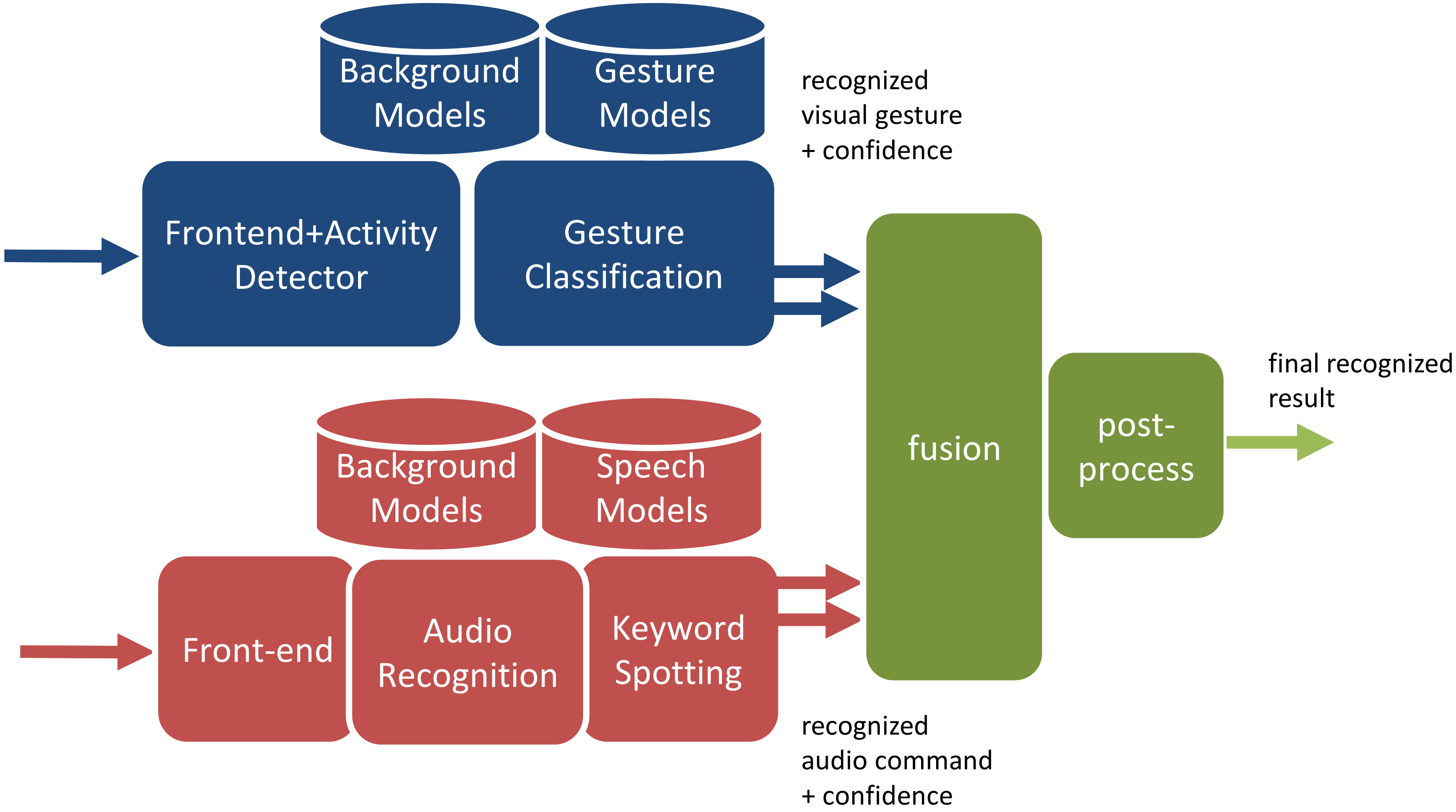
$w_a, w_v$ : modality weights

$$\mathrm{MAX}(w_a \times score(A_1) + w_v \times score(V_3), w_a \times score(A_2) + w_v \times score(V_1))$$

# Audio-Gestural Command Recognition

## Online processing system – Open Source Software

http://robotics.ntua.gr/projects/building-multimodal-interfaces

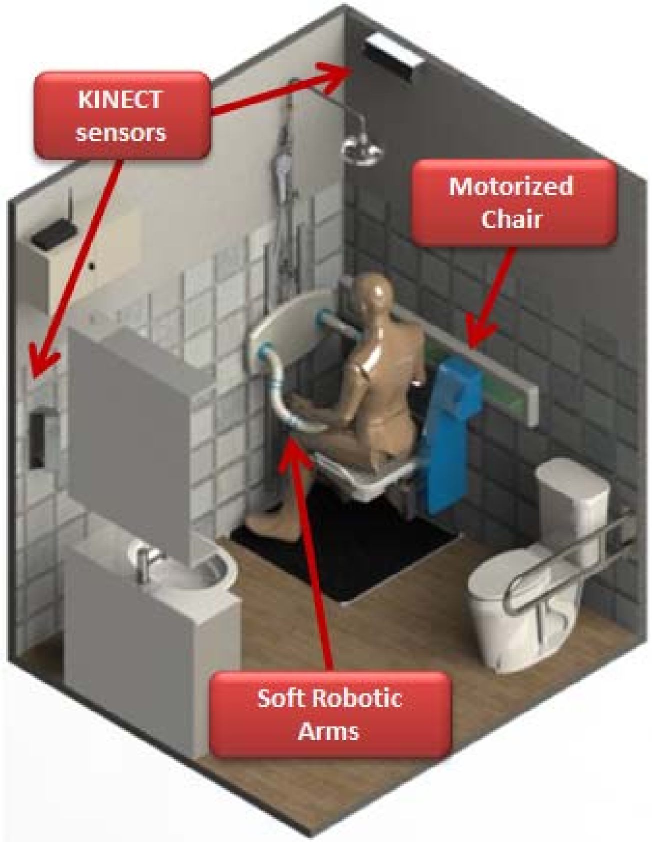



N. Kardaris, I. Rodomagoulakis, V. Pitsikalis, A. Arvanitakis and P. Maragos, *A platform for building new human-computer interface systems that support online automatic recognition of audio-gestural commands*, Proc. ACM Multimedia 2016.

# i-Support project

I-Support project goals:

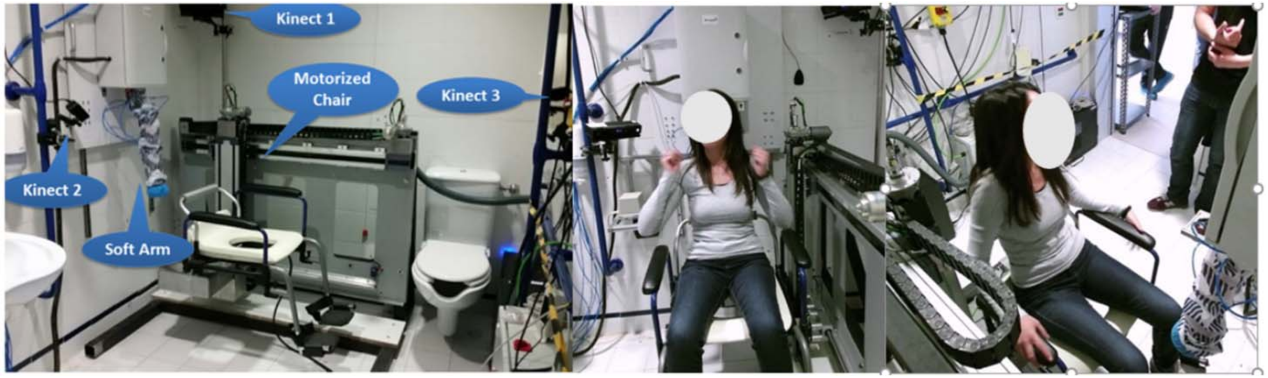
- Assist elderly people with their bathing activities in a **safe**, **effective** and **independent** manner.
- Ensure safe entry and exit from the bathing area.
- **Support** and **reinforce** elderly users’ **motor capabilities** and **strength.**
- Adapt, integrate and effectively control soft arms.
- **Continuous**, **natural** and **intuitive human – machine interaction** using audio and/or gestural commands**.**



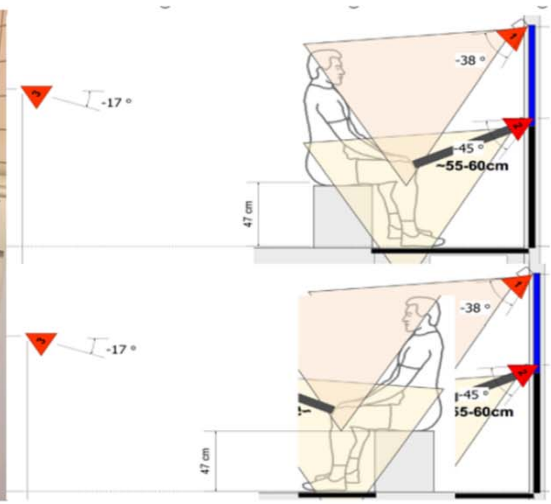

[ Zlatintsi et al., “I-Support: A robotic platform for an assistive bathing robot for the elderly population”, *Robotics & Autonomous Systems* 2020.]

# i-Support: Validation Setup

FSL, Rome




Bethanien, Heidelberg



# I-SUPPORT system video

[ Zlatintsi et al., “I-Support: A robotic platform for an assistive bathing robot for the elderly population”, *Robotics & Autonomous Systems* 2020.]

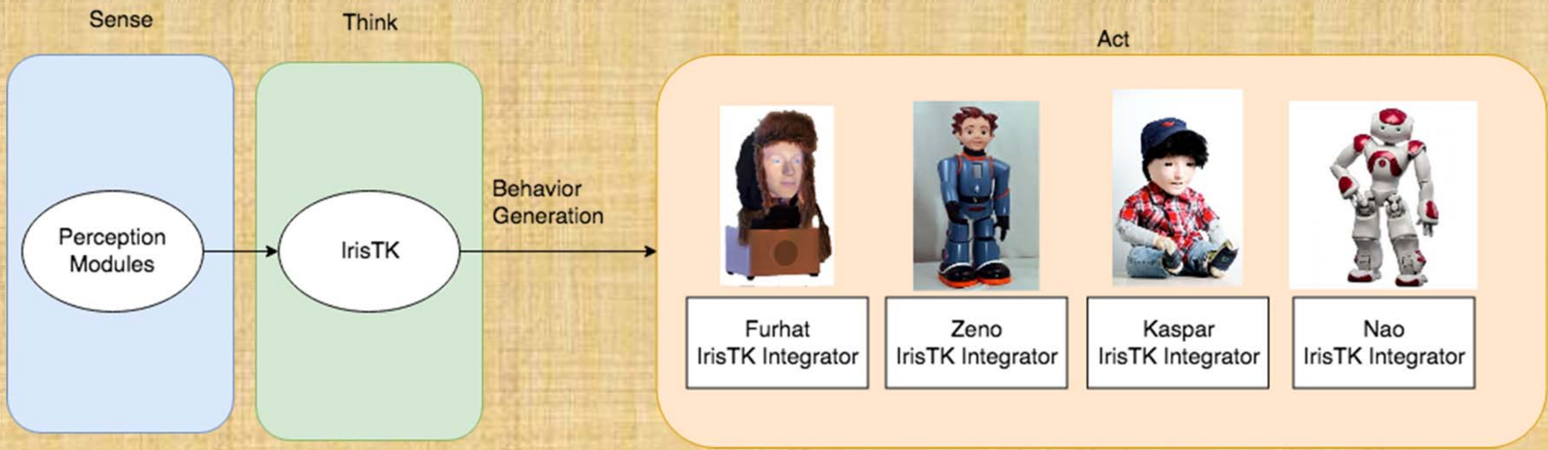
# BabyRobot: Child-Robot Communication & Collaboration

- Create robots that analyze and track human behavior over time in the context of their surroundings using audio-visual monitoring to establish common ground and intention-reading capabilities

- Focus on typically developing and autistic spectrum children.

- Define, implement and evaluate child-robot interaction application scenarios for developing specific socio-affective, communication and collaboration skills.

http://www.babyrobot.eu/



# Child-Robot Interaction: Demo

- Develop core audio-visual processing technology to extract low-, mid-, & high-level HRI information from AV sensors.


https://www.youtube.com/watch?v=DWIm9zCK9dk&ab_channel=IRALNTUALaboratory

# Child-Robot Interaction nowadays



- Attracts **interdisciplinary scientific interest**

→ focuses on children's mental and cognitive development

- Wide range of applications with social robots for **education** and **edutainment**
- In classrooms, social robots create **more pleasant learning** and motivate children to participate more
- Numerous studies focus on topics related to the conditions of Child Robot Interaction (CRI)
- Robotic agents in such studies are mostly semi-autonomous or tele-operated
- Advancements in machine learning lead to create more **intelligent perception systems** → encouraging robotic use by non-experts

# ChildBot: Multi-robot perception and interaction with children



- **ChildBot**: Integrated modular robotic system with
  - multiple perception modules
  - multiple robotic agents
  - wide range of tasks

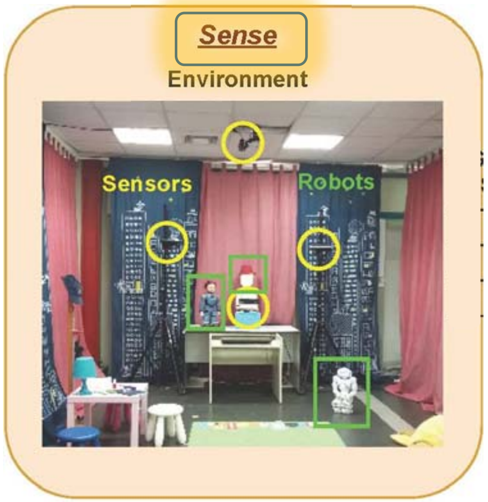
- **Contributions:**
  - An integrated system for AudioVisual Human-Robot Interaction
  - Perception modules for multimodal scene understanding
  - Evaluated on Spontaneous children data during CRI
  - Used for child-robot interactions with Typical Development children or children with Autism Spectrum Disorders

[ A. Tsiami, P. Filntisis, N. Efthymiou, P. Koutras, G. Potamianos, P. Maragos, "Multi3: Multi-sensory Perception System for Multi-modal Child Interaction with Multiple Robots", *Proc. Int'l Conf. Robotics & Automation*, 2018. ]

[ N. Efthymiou, P. P. Filntisis, P. Koutras, A. Tsiami, J. Hadfield, G. Potamianos and P. Maragos, "**ChildBot**: Multi-robot perception and interaction with children", *Robotics and Autonomous Systems*, 2022. ]

# ChildBot: three-layer architecture: Sense – Think - Act

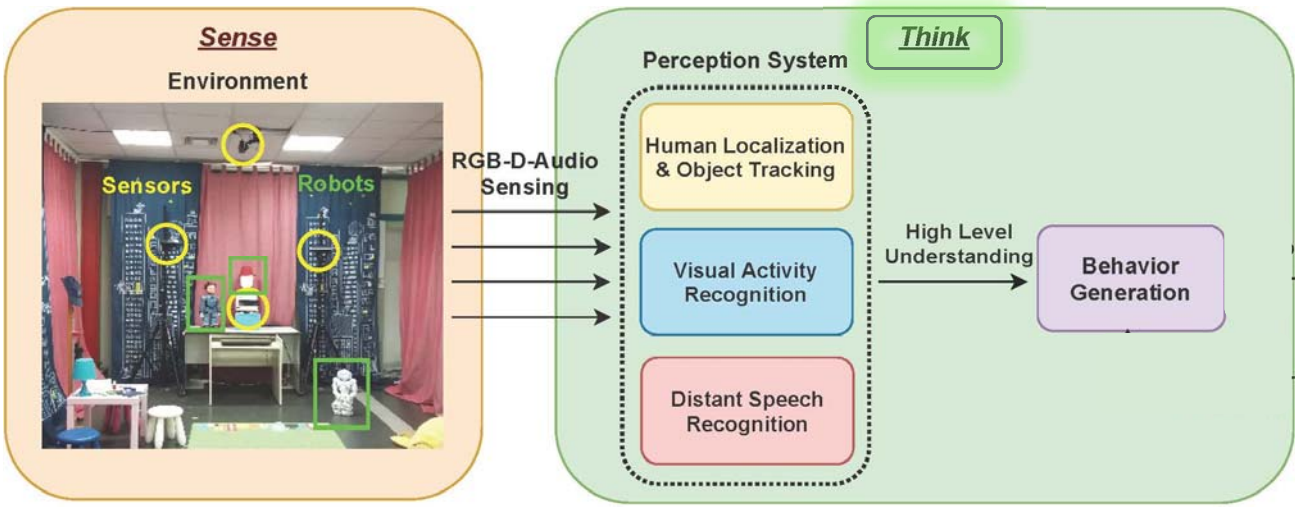





- **Network of sensors**: 4 compact sensors (RGB cameras + Depth, microphone arrays)
  - Avoiding occlusions
  - Fusion of different data streams
  - Bypassing thesensing limitations of individual robotic systems
- **Multiple Robots**: NAO, Zeno, Furhat

# ChildBot: three-layer architecture: Sense – Think - Act

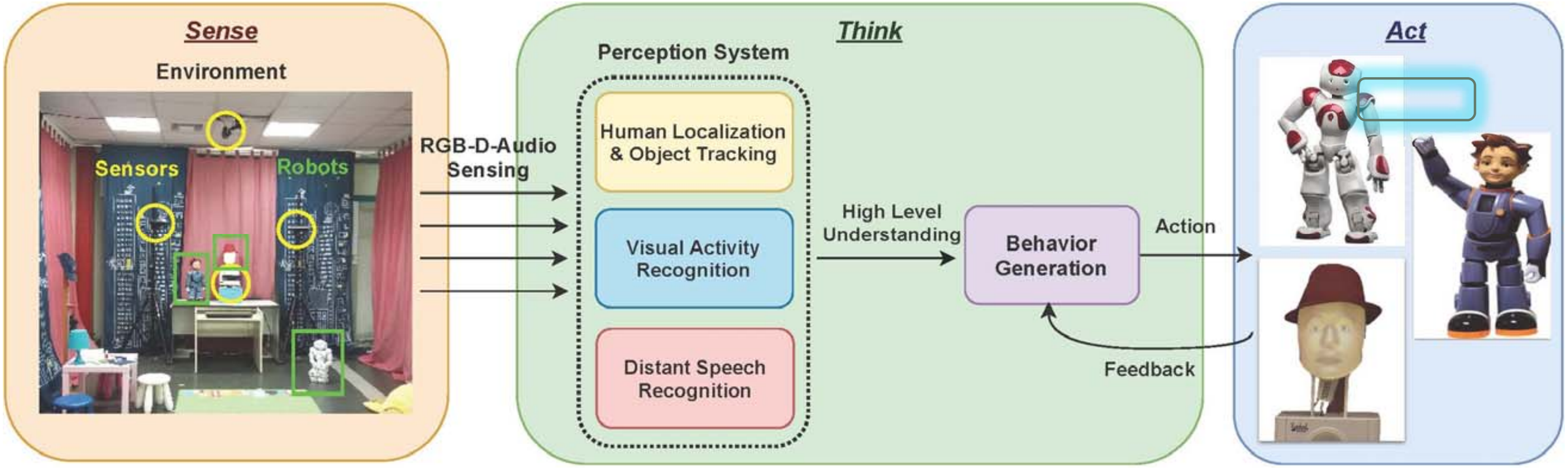





- **Perception systems**:
  - Audio-Visual Active Speaker Localization and 6-DoF ObjectTracking
  - Visual Activity Recognition
  - Distant Speech Recognition
- **Continuous/ High-level understanding** of children's actions
- **Behavior generation** module decides and controls robotic agents

# ChildBot: Αρχιτεκτονική 3 επιπέδων: Sense – Think - Act





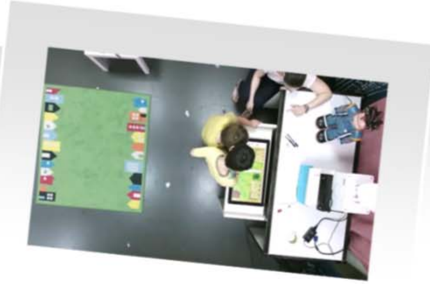

- Act:
  - Robots' movements
  - Robots' speech
  - Other multimodal actions (e.g. touch screen)
- Feedback
- System is ready to detect new events:
  - Sense
  - Act
  - Monitor (feedback etc.)

# ChildBot: Train & Evaluate on children data



Show me the gesture
Kinect #1

Pantomime
Kinect #2



Form a Farm
Kinect #3

Express the Feeling
Kinect #4

- 20 adults
- 31 typical development children 6-10 years old
- 15 children with Autism Spectrum Disorder
- 2 types of collected data:
  - Development data
  - Use-case related data
- 6 individual games
- 3 cooperative games

Statistics of the most important child activities during the data collection.

| Collected data | Event type | Number of events |
|---|---|---|
| Development data | Utterances | 977 |
| | Gestures | 196 |
| | Pantomimes | 336 |
| Use-case related data | Utterances | 641 |
| | Gestures | 143 |
| | Pantomimes | 109 |

# Children vs Adults models



Different training schemes (on Development data):

- Adults models
- Children models
- Mixed model

need for children specific models

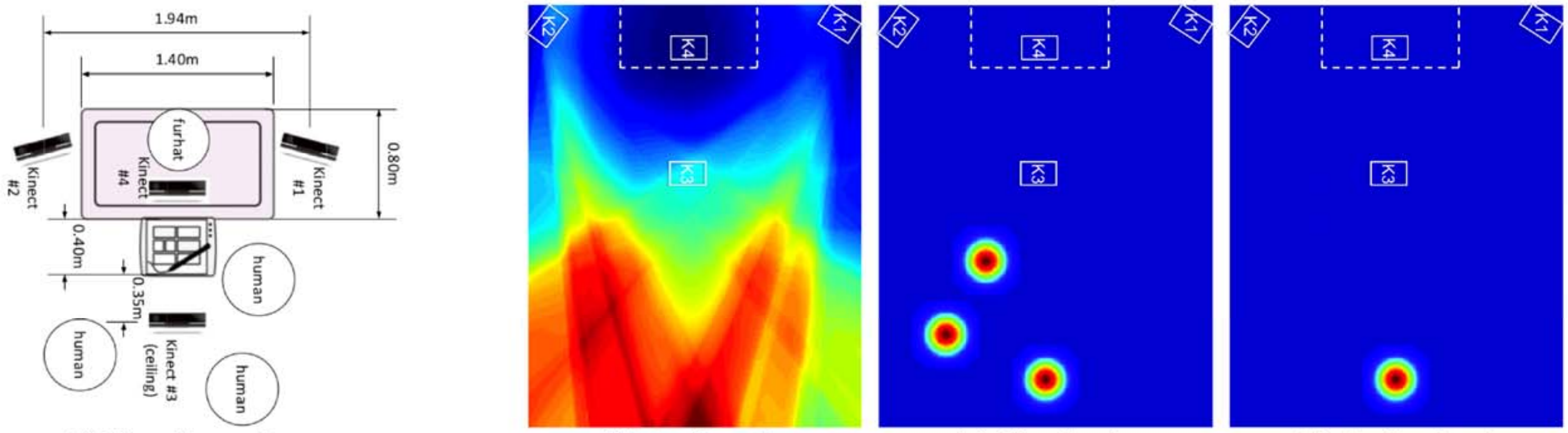
| | | Gesture Recognition Training scheme | | |
|---|---|---|---|---|
| Test | | Adults | Children | Mixed |
| Adults | Avg | 86.49 | 56.32 | 87.36 |
| | Fuse | 92.19 | 62.08 | **95.10** |
| Children | Avg | 49.92 | 70.99 | 72.30 |
| | Fuse | 56.25 | **83.80** | 80.09 |

| | | Action Recognition Training scheme | | |
|---|---|---|---|---|
| Test | | Adults | Children | Mixed |
| Adults | Avg | 78.39 | 63.00 | 78.39 |
| | Fuse | **87.36** | 72.53 | 86.26 |
| Children | Avg | 46.55 | 65.74 | 65.88 |
| | Fuse | 56.51 | **74.46** | 74.26 |

| | DSR-Adaptation scheme | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | No-adapt | | Adults | | Children | | Mixed | |
| Test | WCOR | SCOR | WCOR | SCOR | WCOR | SCOR | WCOR | SCOR |
| Adults | 97.54 | 91.25 | 99.58 | **98.87** | 96.73 | 93.20 | 99.50 | 98.43 |
| Children | 79.06 | 69.95 | 75.31 | 71.20 | 97.81 | **95.50** | 90.71 | 82.06 |

[ N. Efthymiou, P. P. Filntisis, P. Koutras, A. Tsiami, J. Hadfield, G. Potamianos and P. Maragos, "**ChildBot**: Multi-robot perception and interaction with children", *Robotics and Autonomous Systems*, 2022. ]

# Audio - Visual Active Speaker Localization



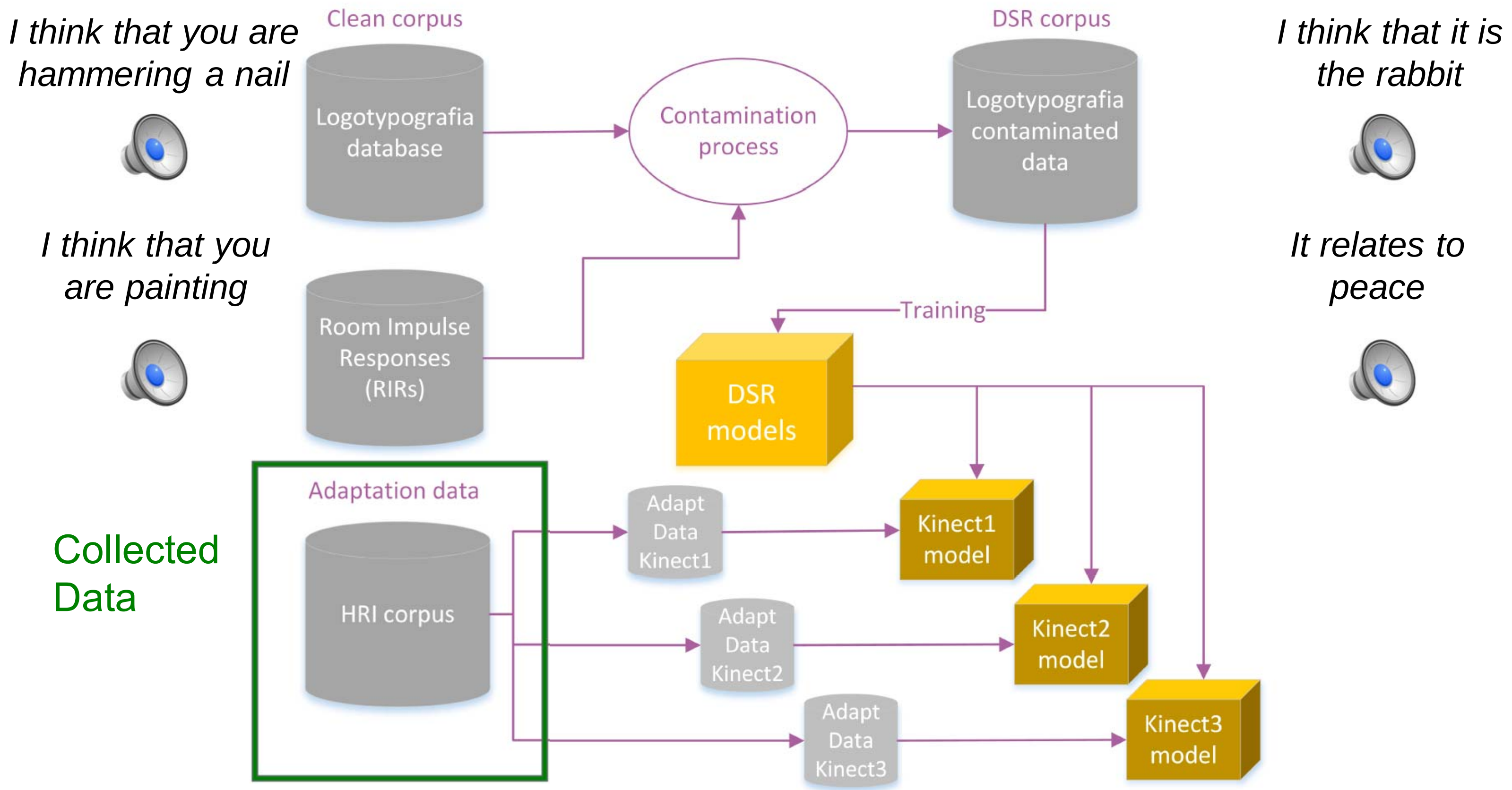
(a) Experimental setup (b) Audio-only (c) Visual-only (d) Audio-visual

- person tracking using 3D skeleton
- choosing the person closest to the auditory source position
- Rcor: percentage of correct estimations (deviation from ground truth less than 0.5m)
  - Audio Source Localization: 45.51%
  - Audio-Visual Localization: 85.58%

[ Tsiami et al., Proc. ICASSP 2018. ]

# Distant Speech Recognition System



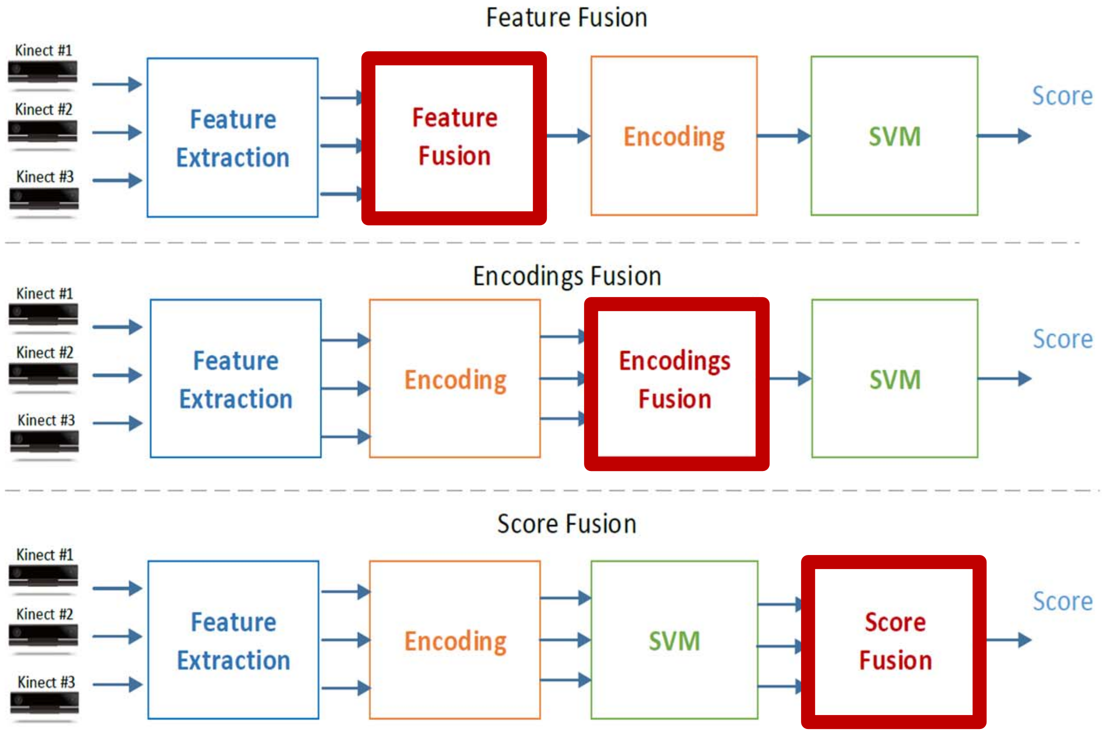



➢DSR model training and adaptation per Kinect (Greek models)

# ChildBot: Activity Recognition Systems

- Multiple views of children activities
- 2 set of activities:
  - Gestures (communicative gestures)
  - More general movements (pantomimes)
- Single-view (only RGB, no depth) and Multi-view fusion activity recognition
- Ablation Studies:
  - Different dense trajectories features (Traj., HOG, HOF, MBH) & combination of them
  - Encoding methods (BoW, VLAD)
  - Different fusion schemes

# ChildBot: Multi-view Fusion for Activity Recognition




- Feature Fusion: Early fusion of low-level descriptors
- Encodings Fusion: Middle fusion of encodings
- Score Fusion: Late fusion deploying the resulted probabilities for the recognition, from each sensor

[ Efthymiou et al., Proc. ICIP 2018. ]

# Evaluation of activity recognition system



| Recognition System (VLAD – Comb. Feats.) | | Single-View Accuracy (%) | Multi-View Accuracy(%) |
|---|---|---|---|
| Gestures | Development Data | 71 - 81 | 83 - 85 |
| | Use-case related Data | 61 - 72 | 69 - 74 |
| Actions | Development Data | 59 - 76 | 77 - 79 |
| | Use-case related Data | 38 - 59 | 62 - 69 |

- Best system architecture:
  - use of combined features and VLAD encoding
- **85% gesture** & **79% action recognition accuracy** on development data
- Decrease ~10% accuracy on use-case related data
- Best fusion results:
  - **score** fusion on development data (gestures & pantomimes)
  - **encodings** fusion for gestures & **features** fusion for pantomimes on use-case related data

[ N. Efthymiou et al., "ChildBot", *Robotics and Autonomous Systems*, 2022. ]

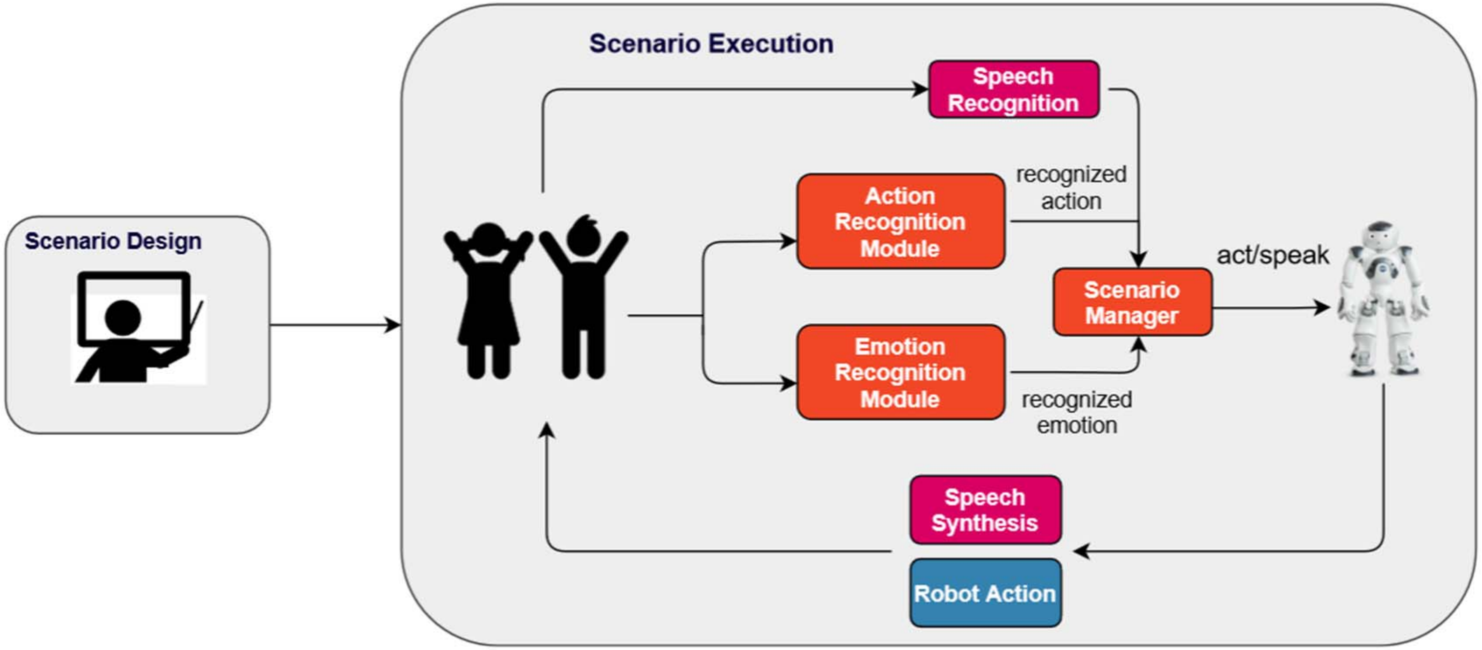
# Child-Robot Interaction: Multiple Children (RPS)


[ A. Tsiami, P. Filntisis, N. Efthymiou, P. Koutras, G. Potamianos, P. Maragos, "Multi3: Multi-sensory Perception System for Multi-modal Child Interaction with Multiple Robots", *Proc. Int'l Conf. Robotics & Automation*, 2018. ]

[ N. Efthymiou, P. P. Filntisis, P. Koutras, A. Tsiami, J. Hadfield, G. Potamianos and P. Maragos, "ChildBot: Multi-robot perception and interaction with children", *Robotics and Autonomous Systems*, 2022. ]

# TeachBot: Child-Robot Interaction System for Edutainment






- Focus on visual information: Action & Emotion Recognition
- Lightweight and low-cost system (NAO, compact cam+mic, computational system)
- Use in classrooms: Design and execute edutainment scenarios

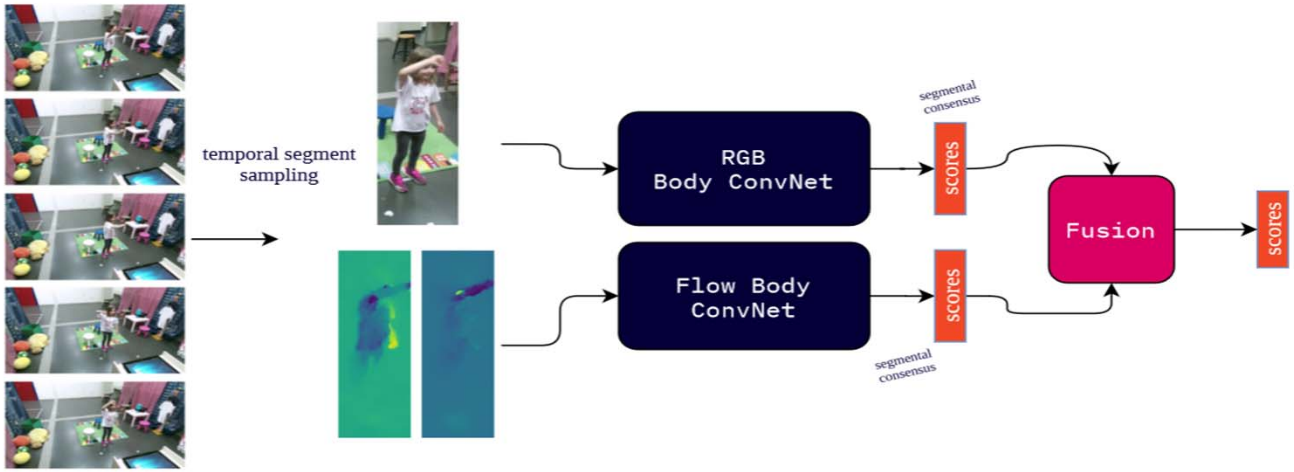
[ Efthymiou et al., PETRA 2021. ]

# TeachBot: Visual Perception



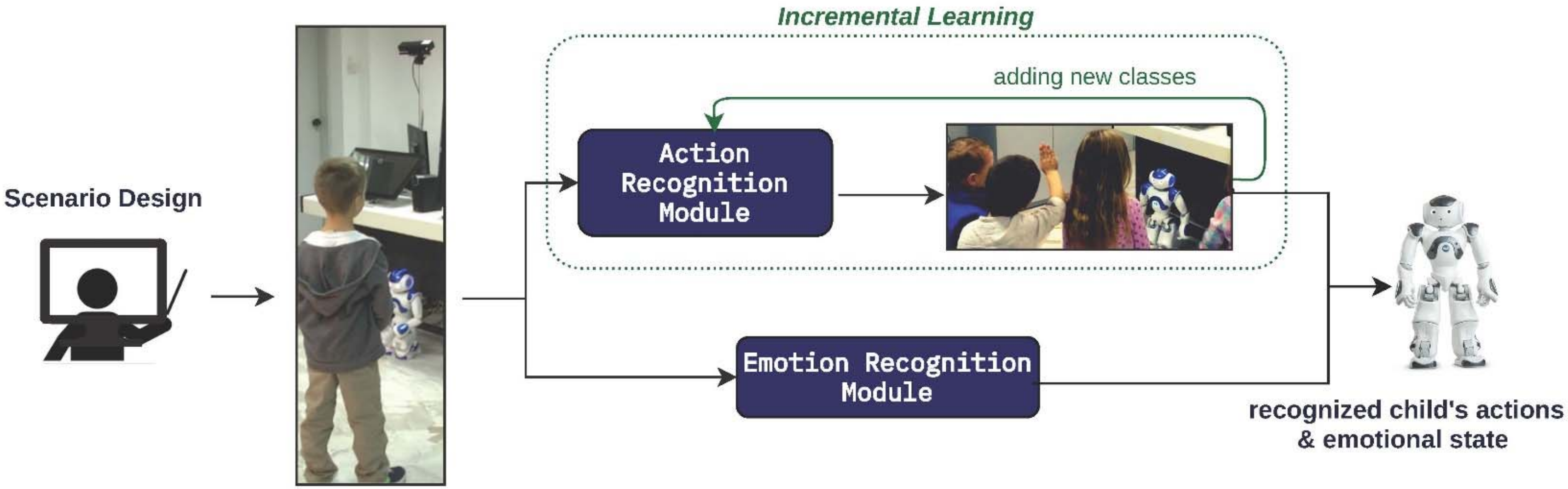
Both perception modules based on **Temporal Segment Networks - TSN**

- **Random selection of *K* short segments** from each video clip
- **Crop** the video around the area of interest (face or body)
- Two streams are used: **RGB** and **Optical Flow**.




- Helps the generalization
- Reduces the computational cost
- Reduces the redundant information

# TeachBot: Employing Incremental Learning




- Learning new classes
- No need for training using new and old classes
- Without significant decrease on the recognition of the old classes
- First work considering Incremental Learning for action recognition in CRI
- Outperforms the state-of-the-art at a low-computational cost

[ N. Efthymiou, P. P. Filntisis, G. Potamianos, and P. Maragos, "Visual Robotic Perception System with Incremental Learning for Child–Robot Interaction Scenarios," *Technologies*, 2021. ]

# Experimental Results – Single-view: TSN – Pantomimes (BabyRobot data)



*Ablation study*

| Segments | Accuracy (%) | Time/Training Epoch (s) | Time/Validation Epoch (s) |
|---|---|---|---|
| **RGB** | | | |
| 1 | 36.74 | 5.2 | 0.4 |
| 3 | 40.95 | 6.0 | 0.8 |
| 5 | 47.43 | 8.8 | 1.0 |
| 10 | 49.56 | 14.6 | 1.4 |
| **Flow** | | | |
| 1 | 58.75 | 5.4 | 0.6 |
| 3 | 71.77 | 10.3 | 1.2 |
| 5 | 75.96 | 16.3 | 1.8 |
| 10 | $\boxed{76.82}$ | 31.3 | 3.2 |

- K=5 segments for each of RGB & Optical Flow
- Low computational cost
- Pre-trained model is very important

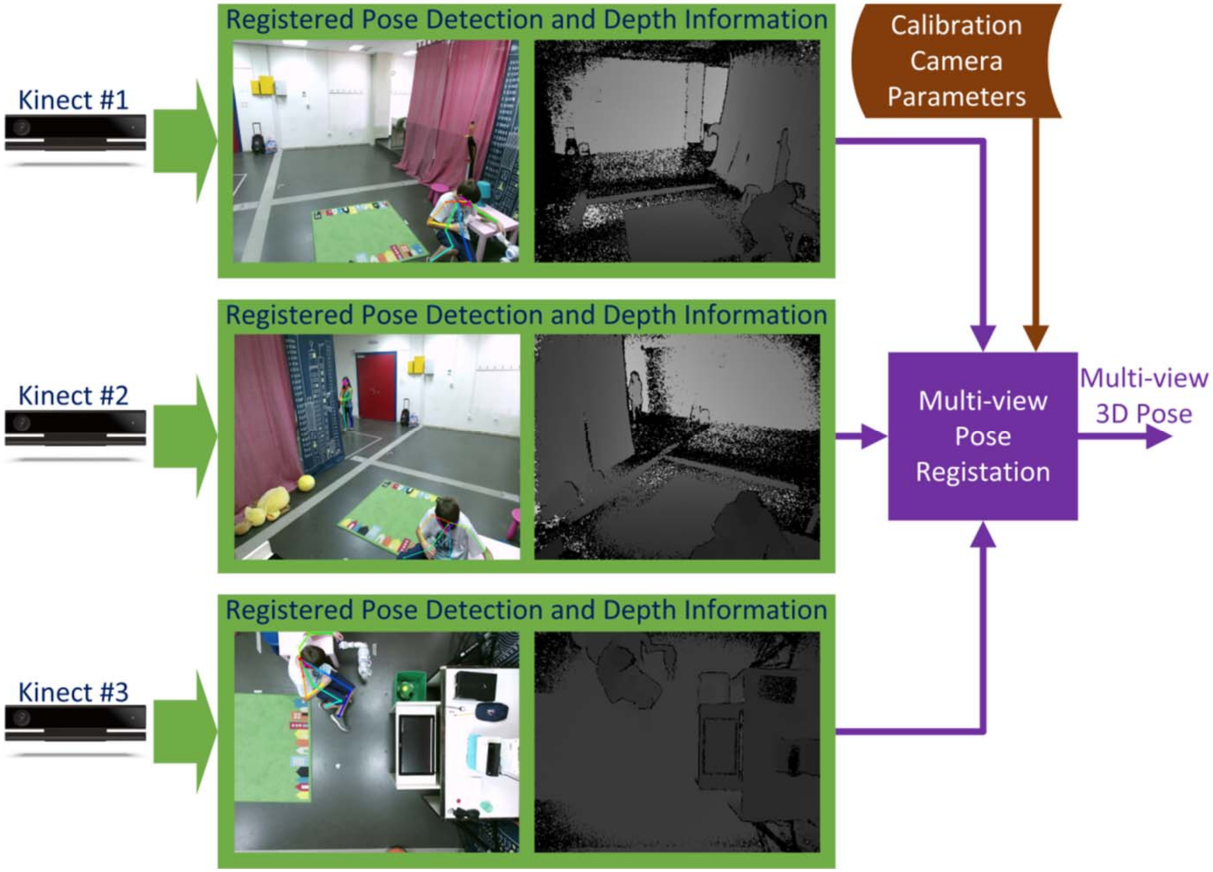
*Recognition accuracy (%)*
*for 13 children pantomimes*

| model | Accuracy (%) |
|---|---|
| RGB-Kinetics | 47.14 |
| RGB-ImageNet | 42.75 |
| Flow-Kinetics | 74.75 |
| Flow-ImageNet | 63.49 |
| **RGB-Kinetics + Flow-Kinetics** | $\boxed{76.23}$ |
| RGB-ImageNet + Flow-ImageNet | 64.10 |
| Dense Traj. Ensemble [14] | 74.15 |
| C3D [14] | 59.38 |

- Engagement estimation is an important factor for **improving Child Robot Interactions quality.**
  - ✓ Robots recognize children's engagement level.
  - ✓ Robots adapt their behavior according to children's cognitive state.

**Engagement:** the level at which the child is both attentive and cooperative with their partner towards their common goal.

# Multiple view 3D Pose




Use 3D pose as indicative of the engagement level.

- Feed OpenPose 2D body skeleton key pts and their Depths into a NN to predict 3D Pose [Zimmermann et al. ICRA 2018].
- [Hadfield et al. IROS 2019] TD Children, Joint Attention: Detect 3D pose from multiple cameras and detect robot's head to estimate i) Distance between child and partner (robot or mother) and ii) Orientation of child's body wrt partner. Fuse poses.

# Engagement Estimation in Child Robot Interactions



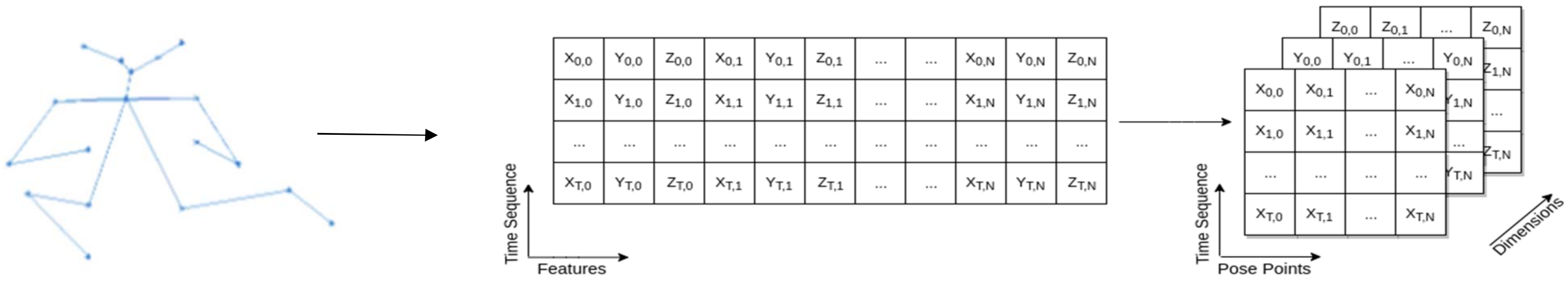
Engagement 0

Engagement 1

Engagement 2

- **Goal:** Develop a reliable method of engagement estimation in diverse CRIs.
- Experiments with 5 different datasets:
  - **TD Joint Attention** (Typically developing children in simple interactions with robots)
  - **ASD Joint Attention** (Children with autism spectrum disorders (ASD) in simple interactions with robots)
  - **ASD Games** (Children with ASD in 4 different games with robots)
  - **BabyAffect** (Children with ASD in interactions with their mothers at home)
  - **ASD School** (Children with ASD in interactions with robots at their school)

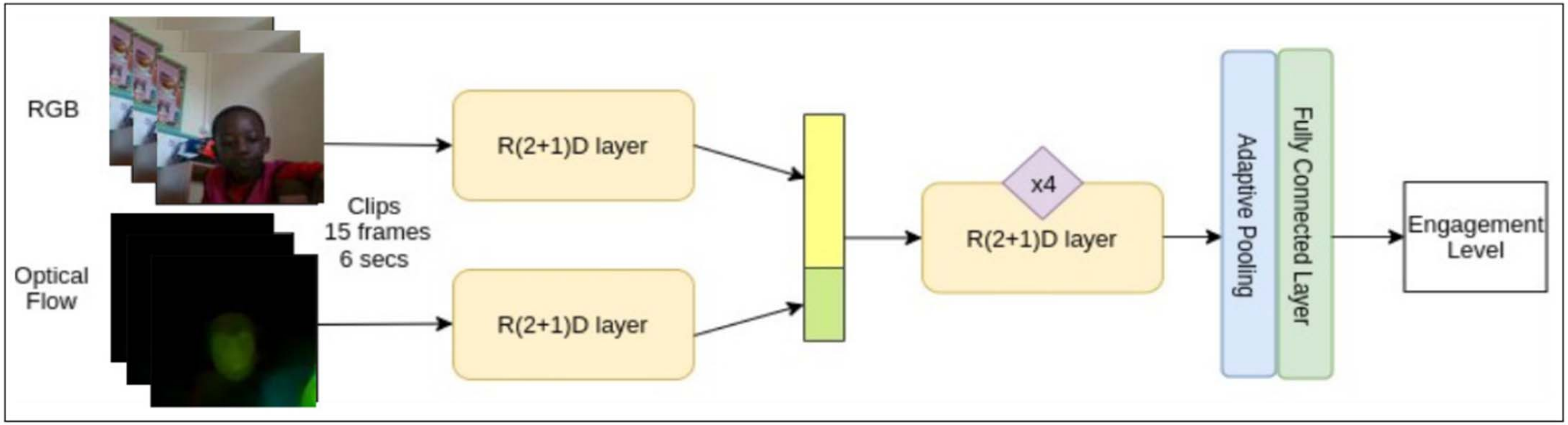
# Child Engagement Estimation Using Pose




- Training of **CNNs** using **pose keypoints.**
- **Rearrange** feature vectors to **resemble images** in order to **retain the temporal information**.
- Use of 2D pose or construction of 3D pose when multiple views and depth are available.

[ Anagnostopoulou et al., Proc. ICRA 2021 ]

# Child Engagement Estimation Using R(2+1)D Networks

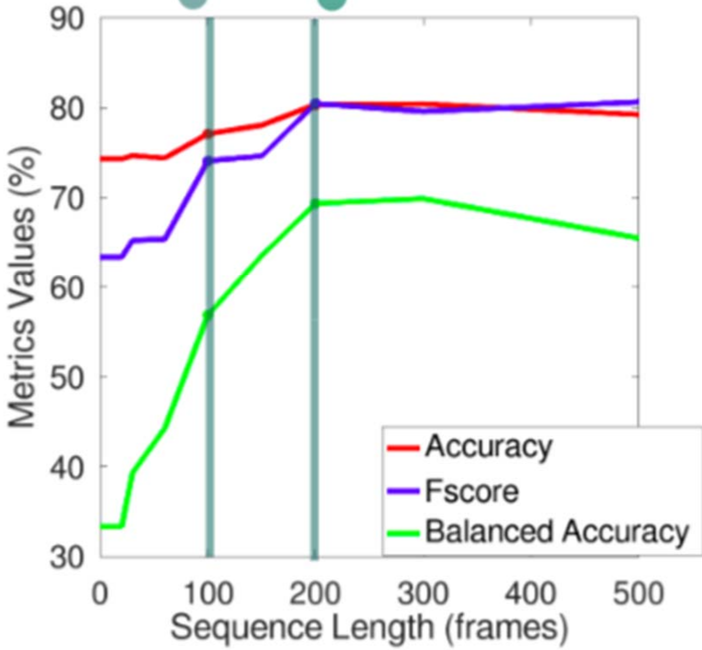



- Engagement estimation mostly in **static** interactions (i.e. a child sitting in front of a desk with a smart monitor and a robot).
- Training of **R(2+1)D** in **6 seconds** intervals.
- Use of **RGB information** in combination with **optical flow** information.

[ D. Anagnostopoulou, N. Efthymiou, C. Papailiou, P. Maragos, "Child Engagement Estimation in Heterogeneous Child-Robot Interactions Using Spatiotemporal Visual Cues", Proc. IROS 2022 ]

# Child Engagement Estimation- Conclusions

Sequences **larger than 100 frames** (approximately 3 seconds) allow the network to train and estimate engagement.

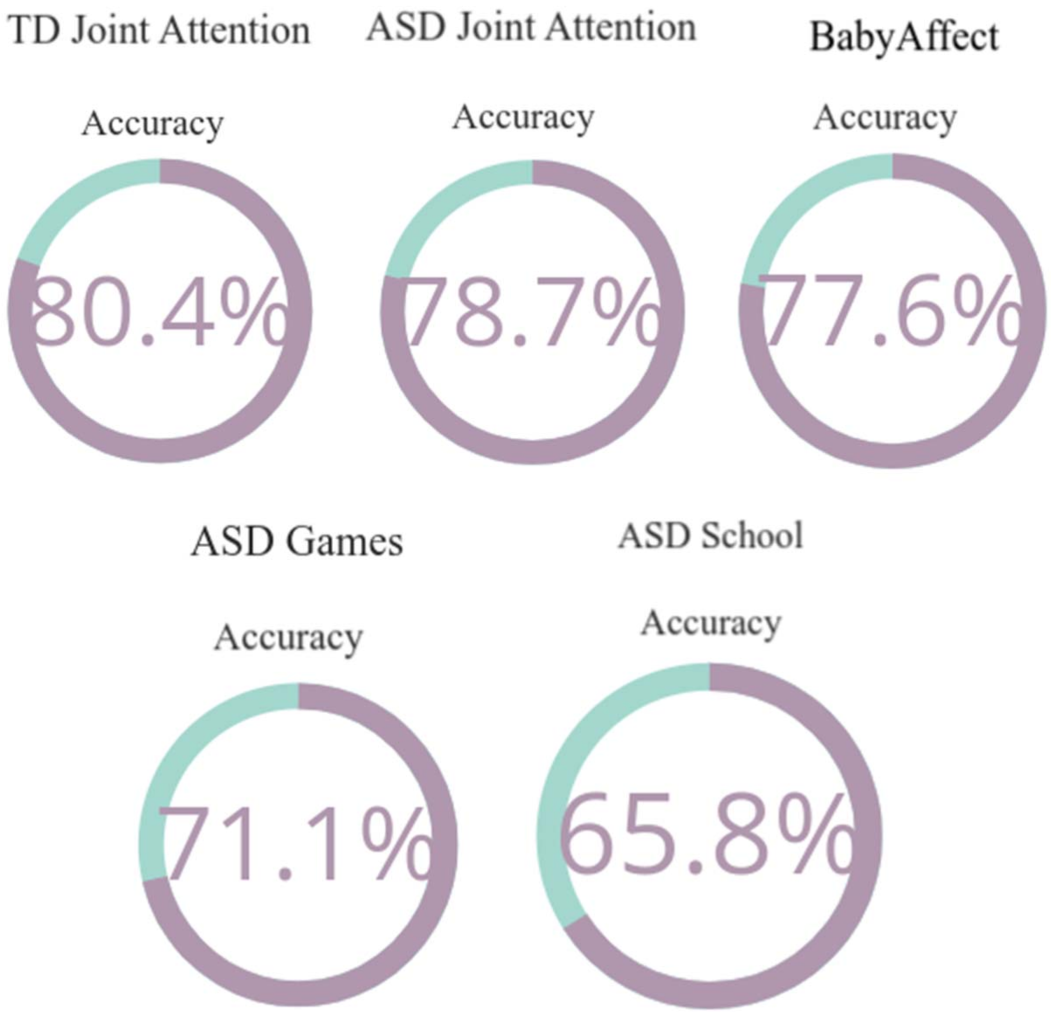



**Best results:** sequences of **200 frames**, i.e. approximately 6 to 7 seconds

**Psychologists' conclusions:** Humans express feelings and intentions in a shared social space through movements organized in a time frame ranging from 3 to 7 seconds. **The time frame 3 to 7 seconds is considered fundamental to human motoric and perceptual functions.**

# Child Engagement Estimation - Results







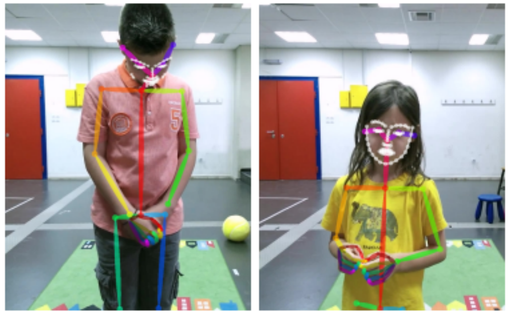
**Levels of Engagement**
0: unengaged
1: partially engaged
2: fully engaged

# Visual Emotion Recognition - Patterns

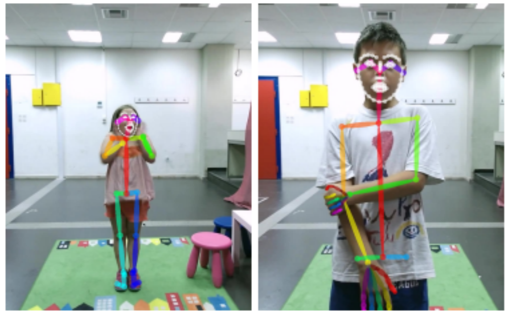


**happiness**
mainly facial, rare jumping and/or open raised hands, body erect, upright head

**sadness**
crying (with hands in front on face), motionless, head looking down, contracted chest



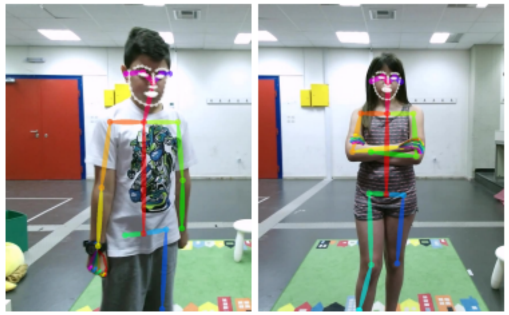
**surprise**
expanded chest, hand movement without specific patterns, either positive or negative surprise



**fear**
quick eye gaze, weak facial expressions, arms crossed in front of body, head sink

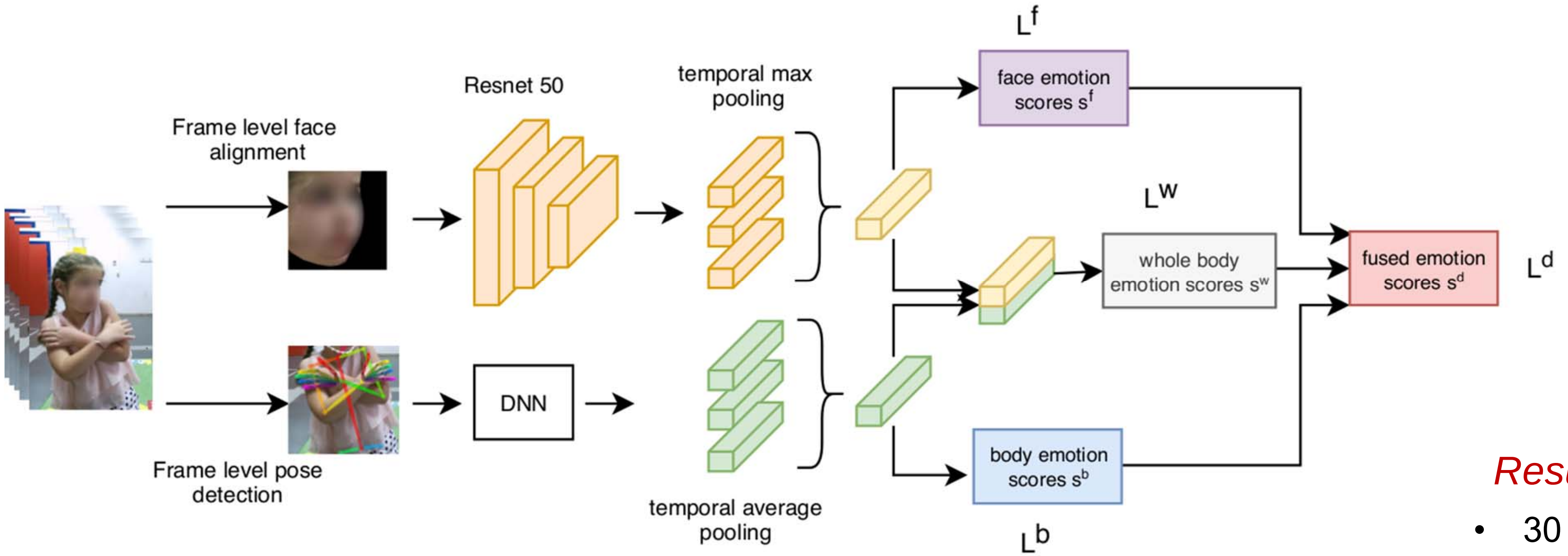
**disgust**
mainly with facial expression (tongue out), movement away from/hands against robot

**anger**
clenched fists, arms crossed, squared shoulders



# Visual Emotion Recognition – HMT network






Hierarchical Multi-label Training (**HMT**) for recognition of affect from multiple visual cues.

$$\mathcal{L} = \mathcal{L}^f(y^f, \tilde{s}^f) + \mathcal{L}^b(y^b, \tilde{s}^b) + \mathcal{L}^w(y, \tilde{s}^w) + \mathcal{L}^d(y, \tilde{s}^d)$$

[ Filntisis et al., IEEE Robotics and Automation Letters, 2019. ]

*Results of HMT network*

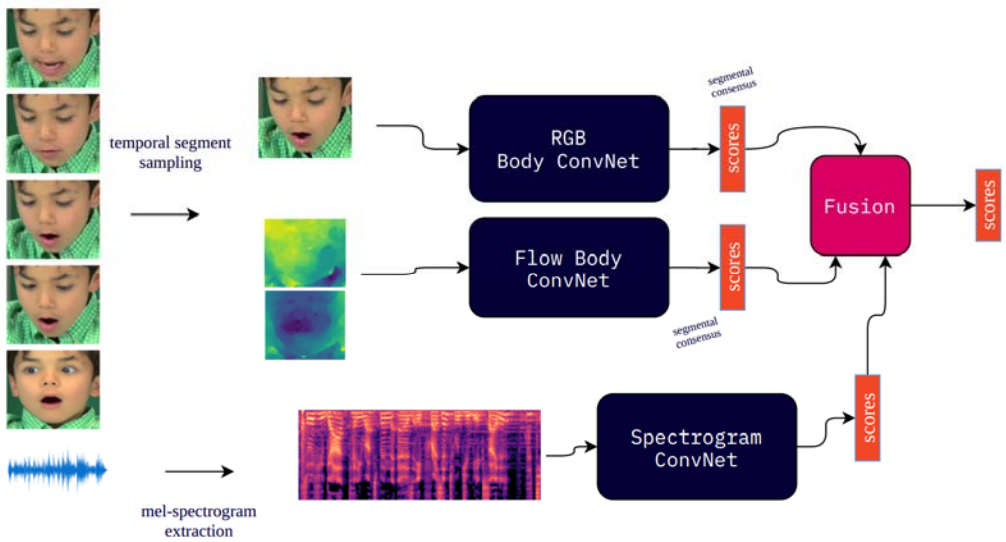
- 30 children × 6 emotions for two sessions: Acted and Spontaneous
- multi-label annotations

| | Acc |
|---|---|
| Body branch | 0.36 |
| Face branch | 0.58 |
| **Fusion** | **0.71** |

# Visual Emotion Recognition: Demo Video


[ Filntisis et al., IEEE Robotics and Automation Letters, 2019. ]

- Use a Temporal Segments Network approach for the visual modality:
  - Ignore redundant information of consecutive frames
  - Lightweight – can run in real time (great for HRI scenarios)
- CNN-based speech emotion recognition applied on spectrograms




*Results on EmoReact dataset*

| | ROC AUC |
|---|---|
| RGB | 0.786 |
| Flow | 0.757 |
| Audio | 0.750 |
| Fusion | 0.799 |

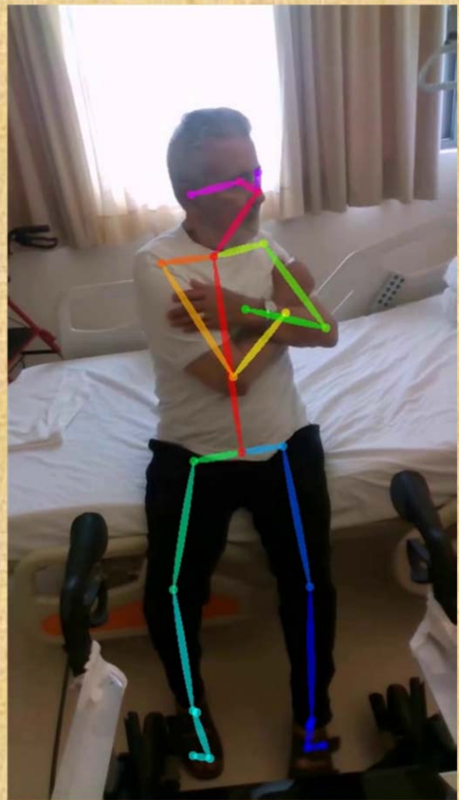
Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., and Van Gool, L. *Temporal Segment Networks: Towards good practices for deep action recognition*. ICCV 2016

Nojavanasghari, B., Baltrušaitis, T., Hughes, C. E., & Morency, L. P. *EmoReact: a multimodal approach and dataset for recognizing emotional responses in children*. ICMI 2016.

[P.P. Filntisis, N. Efthymiou, G. Potamianos and P. Maragos. An AudioVisual Child Emotion Recognition System for Child-Robot Interaction Applications. Proc. EUSIPCO 2021]

# i-Walk

## Intelligent Robotic Walker for mobility and cognitive assistance of elderly and motor-impaired people



website: www.i-walk.gr

G. Moustris et al., "The i-Walk Lightweight Assistive Rollator: First Evaluation Study", *Frontiers in Robotics and AI*, 2021.

# Motivation for Healthcare Robotics

- **Constant growth of elderly population**

Projected global population over 60 by WHO

| 2015 | 2030 | 2050 |
| --- | --- | --- |
| **0,9 Billion** | **1,4 Billion** | **2,1 Billion** |

- **Difficulties in performing Personal Care Activities (e.g. showering, dressing, indoor or outdoor transferring)**

Augmented Nursing Care

**In – House**
Family financial burden

**Clinical**
Nursing staff
Increased requirements

# i-Walk: rehabilitation and social platform

**microphone array**
(eMeet)

**RGB-D cameras**
(RealSense)

**lasers**
(Hokuyo)

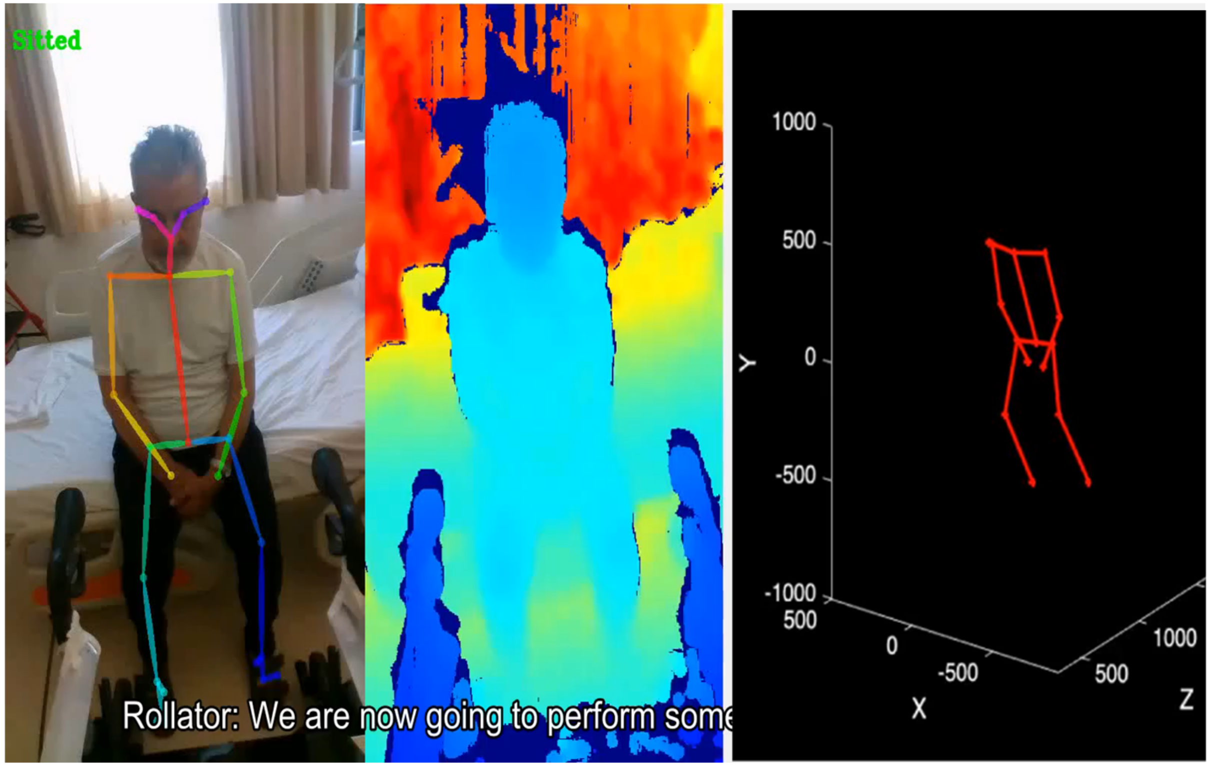
**i-Walk rollator features**:

- One 4-microphone array + speaker
- 2 RGB-D RealSense cameras
- 2 Hokuyo lasers
- 1 touch screen
- 2 motorized wheels
- 2 selection buttons

# i-Walk demo: Rehabilitation Exercises

RealSense camera

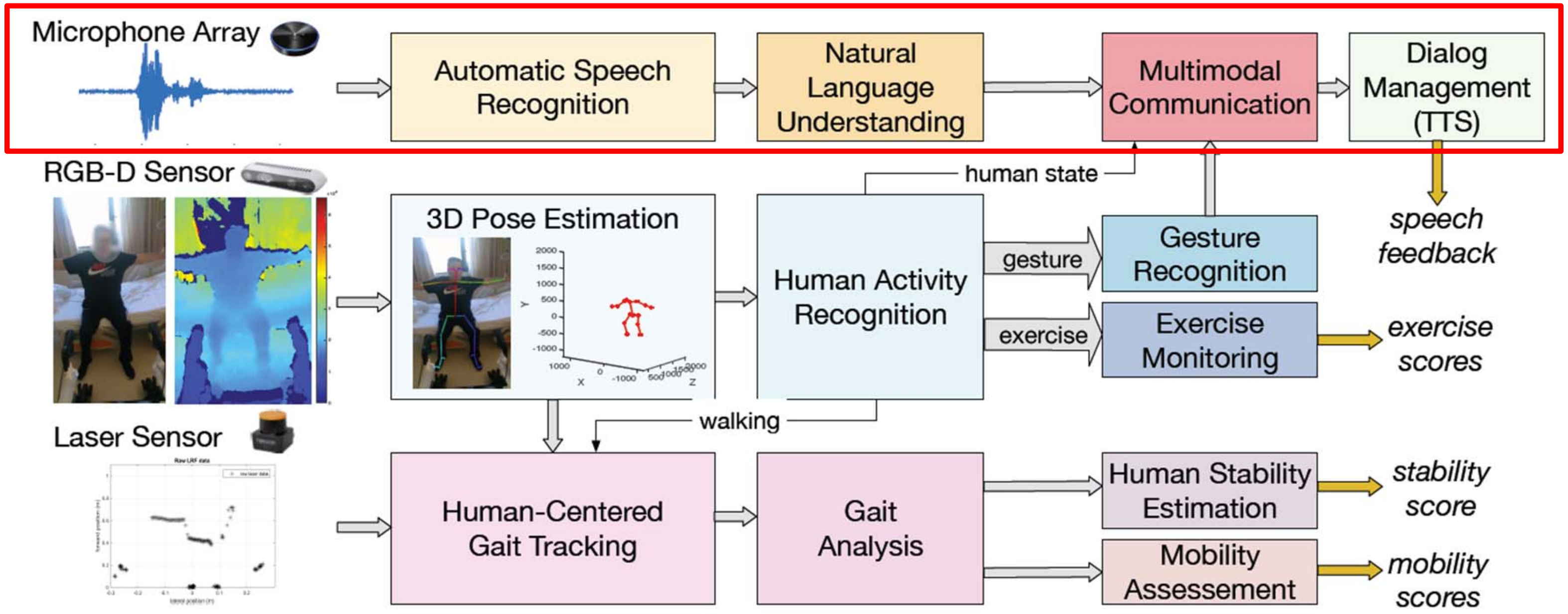
website: www.i-walk.gr




Inverse perspective mapping

RGB 2D Pose + Depth → 3D Pose

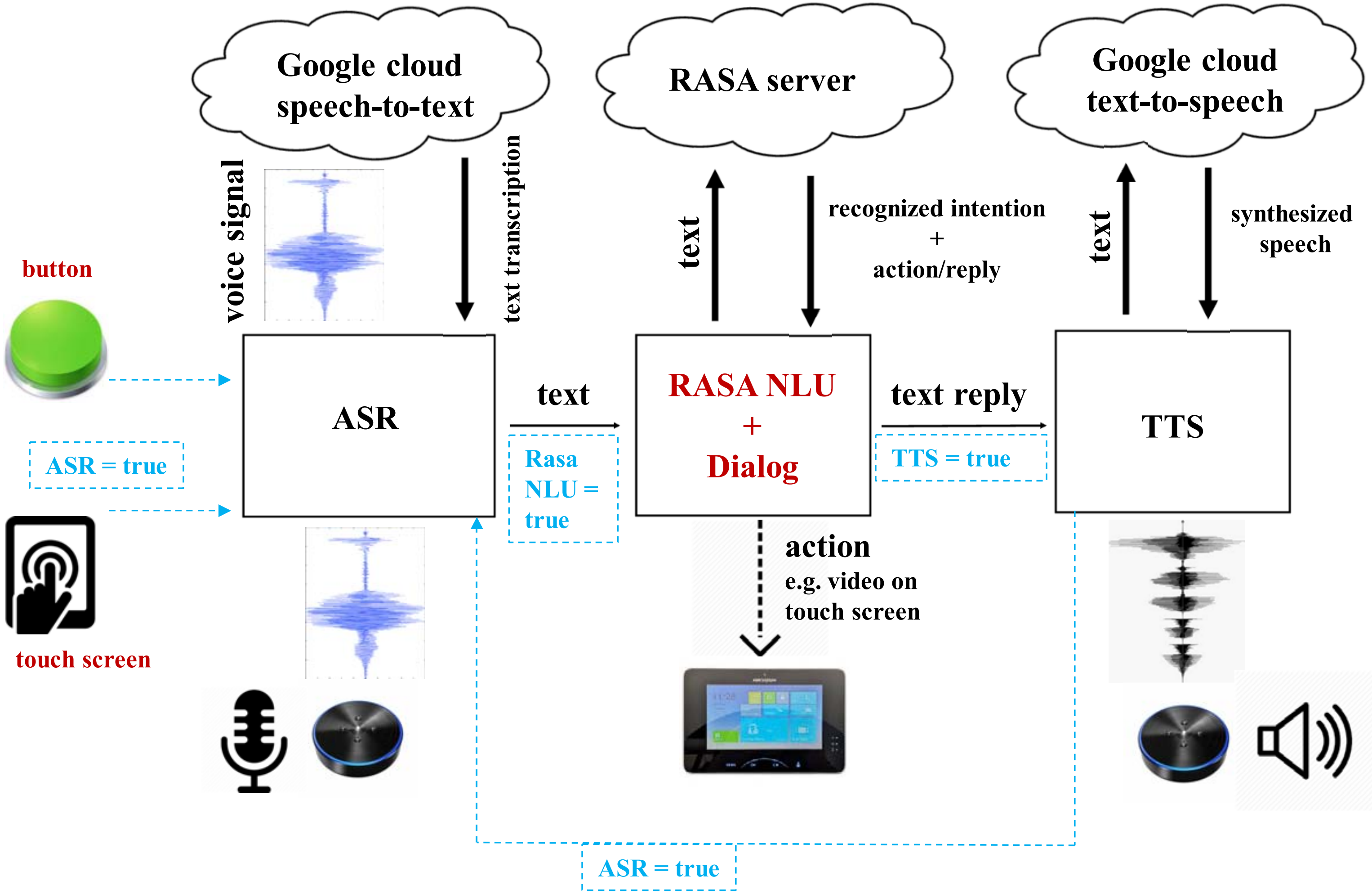
# i-Walk: Multimodal assistance for elderly people (Speech)




**Speech system goals:**

- patients use their voice to express their **intentions**
- the system gives proper feedback and performs designated actions
- continuous speech recognition is required for naturalness in expression
- variability for intention expression, i.e. the patient should express one intention with many possible ways
- speech interaction and user experience should be as easy, simple, natural and pleasant as possible

# i-Walk modular speech system in-depth



# i-Walk speech system technical assessment - I

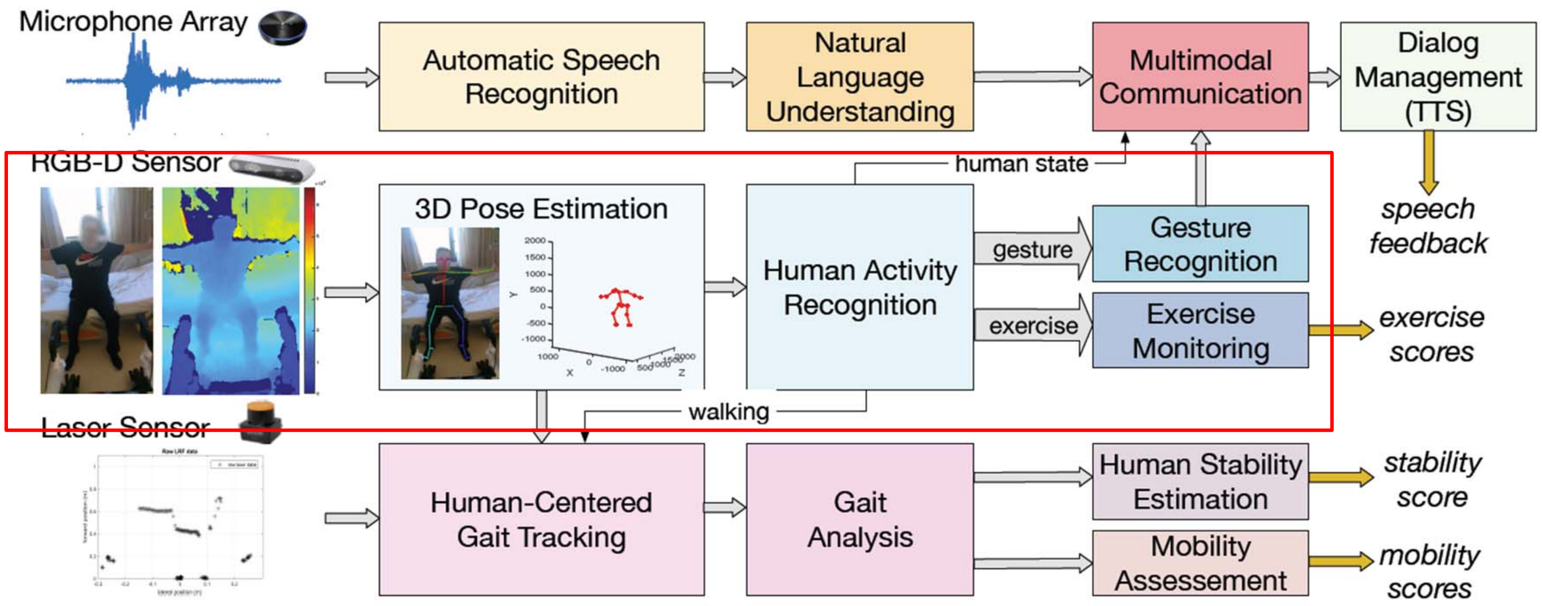
- performs very **fast**, although it gets information from servers and not locally
  - speed is not always desirable as in TTS case
- **modular** architecture
  - ASR, Dialog and TTS do not have any dependency among them; can be substituted by any other corresponding system
  - communication is achieved through ROS signals, components are standby waiting for an activation signal: flexibility
- **internet**-dependent
  - since it uses Google APIs, an internet connection is required

- **evaluation results** (Diaplasi, Kalamata, July 2021)
  - 23 patients: SCOR recognition = 43% WCOR = 49% INTENT = 64%
  - 12 carers: SCOR recognition = 90% WCOR = 93% INTENT = 96%

# i-Walk speech system technical assessment - II

- **advantages**:
  - recognizes even unseen paraphrases for intent expression
  - accepts multimodal inputs → robustness
  - real-time performance with almost no delay
- **main issues**:
  - noise and low volume of speech – voice pathologies
  - long system feedbacks → lack of attention
  - patients need some training on how to use properly
- **Future goals**:
  - more robust ASR for pathological speech
  - richer dialog flows
  - more human characteristics (sense of humor?)
  - short introductory video for patient training

# i-Walk: Multimodal assistance for elderly people (Vision)



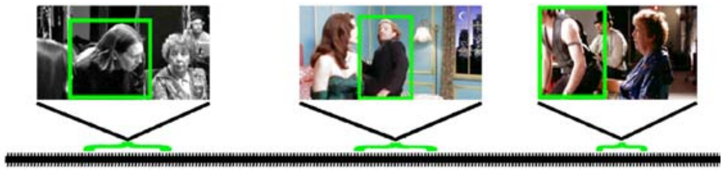

**Action Recognition:**

- Recognize user's state
- Evaluate user's performance on rehabilitation exercises
- Communicate using manual gestures

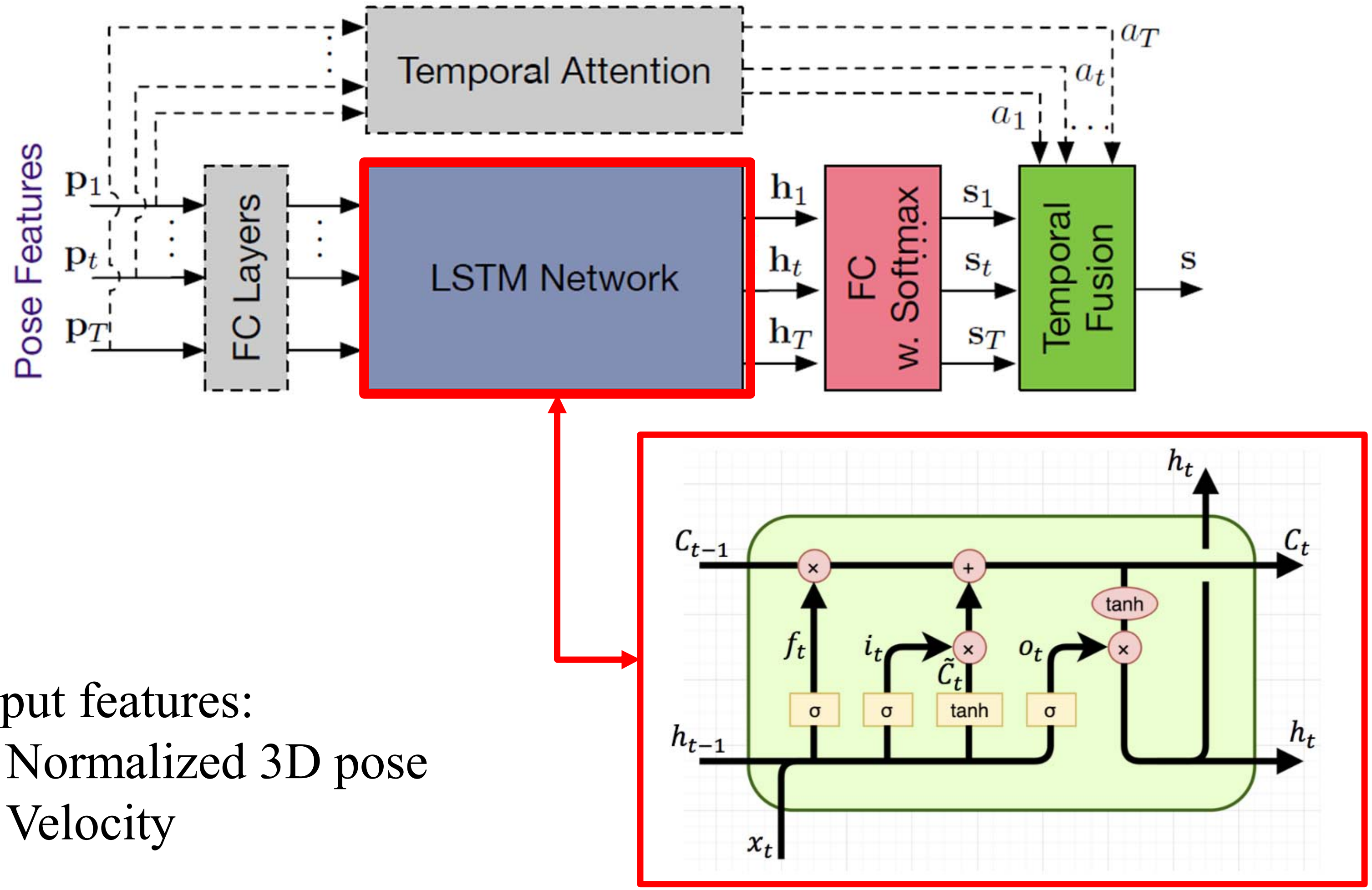
# iWalk: Vision Challenges

- Real-time action detection and classification is required
- Limited computational resources
- Limited training data
- Large performance variability among users





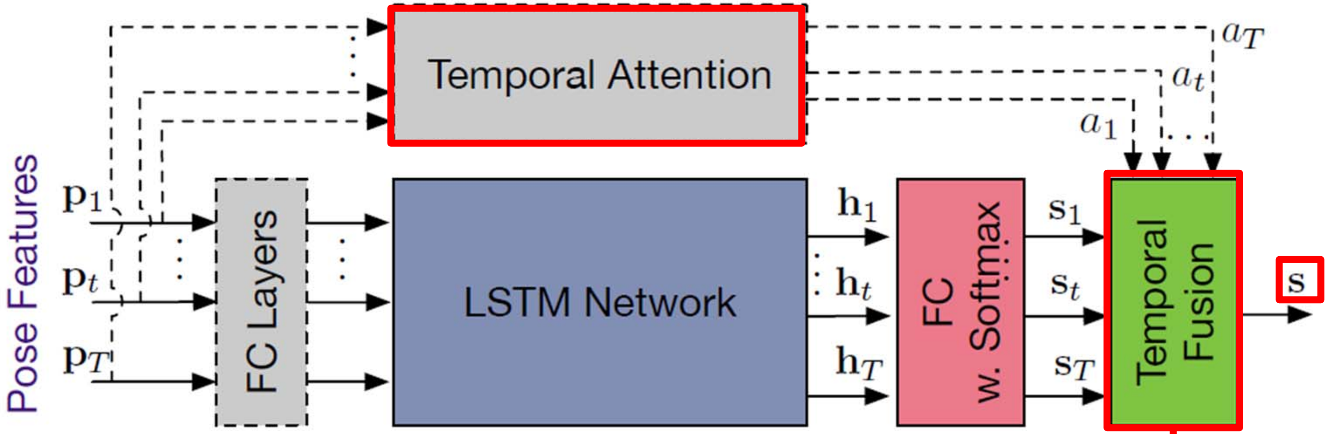
# LSTMs for Action Recognition




Input features:

- Normalized 3D pose
- Velocity

# Temporal Pooling on LSTM's output




- **Average**: $\boldsymbol{s} = \frac{1}{T}\sum_1^T \boldsymbol{s}_t$
- **Max**: $\boldsymbol{s} = \max_{t=1\ldots T} \boldsymbol{s}_t$
- **Weighted average**: $\boldsymbol{s} = \frac{1}{T}\sum_1^T \boldsymbol{\gamma}_t \boldsymbol{s}_t$
- **Attention**:
  - $\boldsymbol{b}_t = FC(\boldsymbol{p}_t)$
  - $\boldsymbol{s} = \frac{1}{T}\sum_1^T \boldsymbol{\alpha}_t \boldsymbol{s}_t$
  - $\boldsymbol{\alpha}_t = \frac{e^{\boldsymbol{b}_t}}{\sum_1^T e^{\boldsymbol{b}_t}}$

# The i-Walk Multimodal Dataset (Y1)

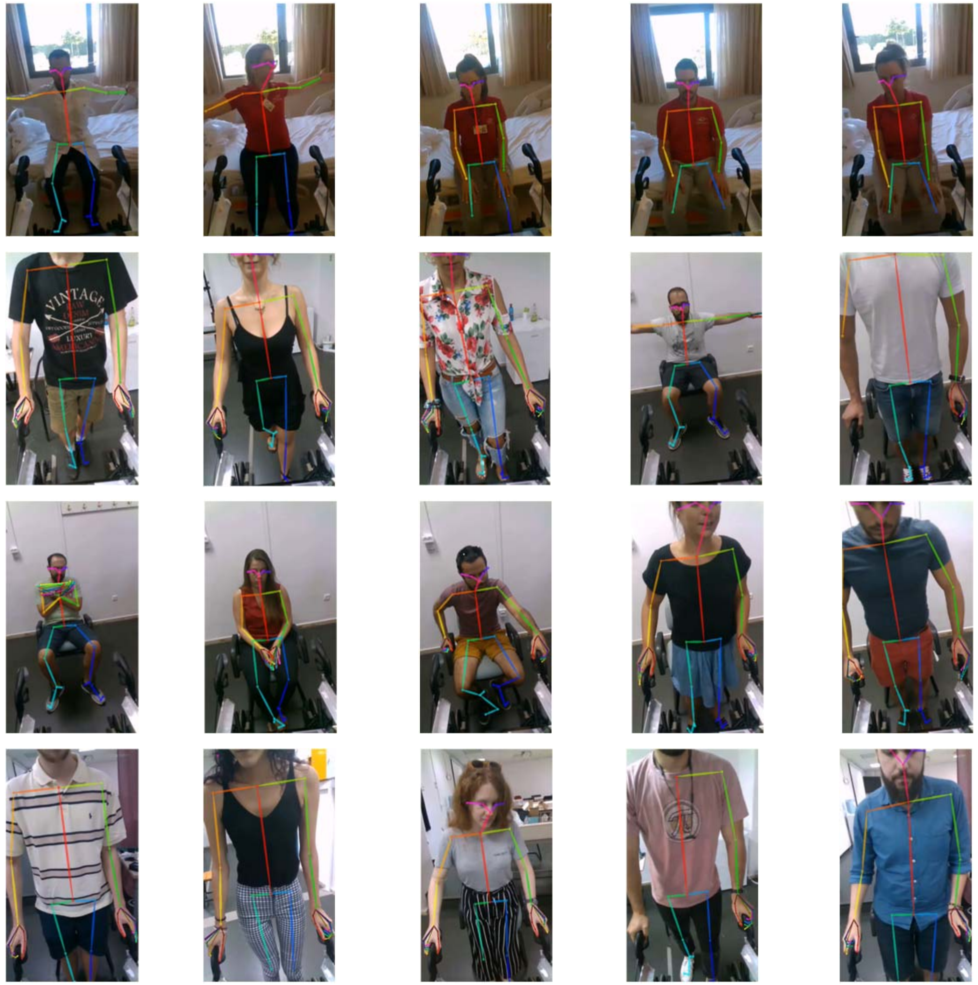
**Actions**

| | |
|---|---|
| 1 | Sitted |
| 2 | StandUpPrep |
| 3 | StandUp |
| 4 | SitDown |
| 5 | Walking |
| 6 | Standing-still |
| 7 | HandCross |
| 8 | HandCrossTurn |
| 9 | HandOpenTurn |
| 10 | HandOpen |
| 11 | WeightMoves |
| 12 | StepsHigh |
| 13 | TurnStanding |
| 14 | Gesture |

**Gestures**

| | |
|---|---|
| a | ComeCloser |
| b | WantStandUp |
| c | WantSitDown |
| d | Stop |
| e | End |

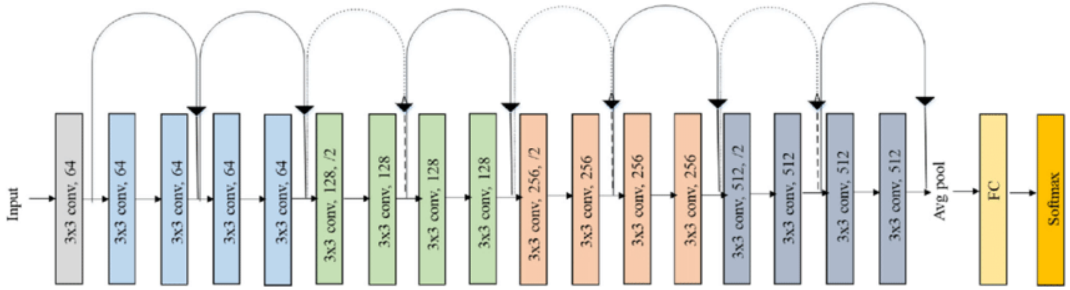
## 20 healthy users

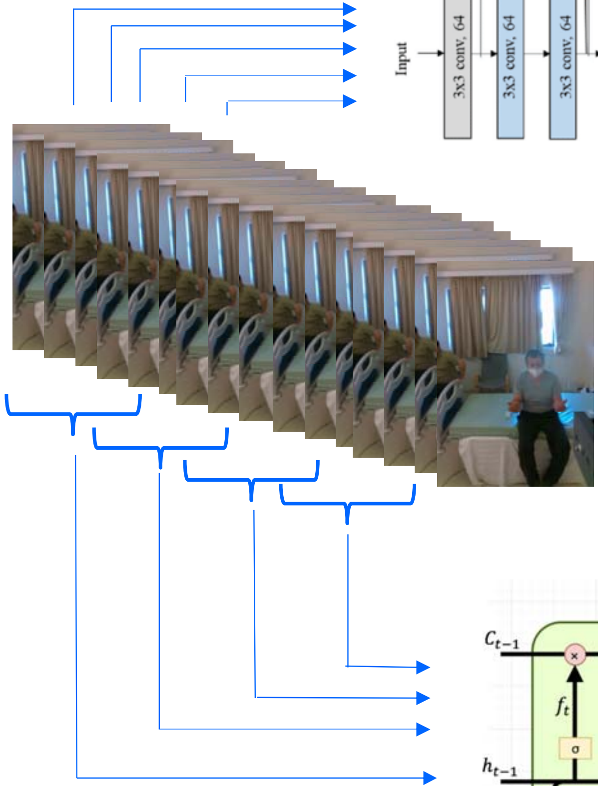


# The i-Walk Multimodal Dataset (Y1)

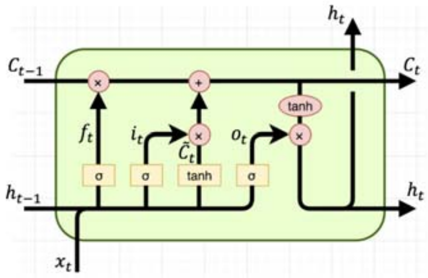
**13 patients**

# Overall Gesture Recognition System (Y3)



**Classification (frame-by-frame)**

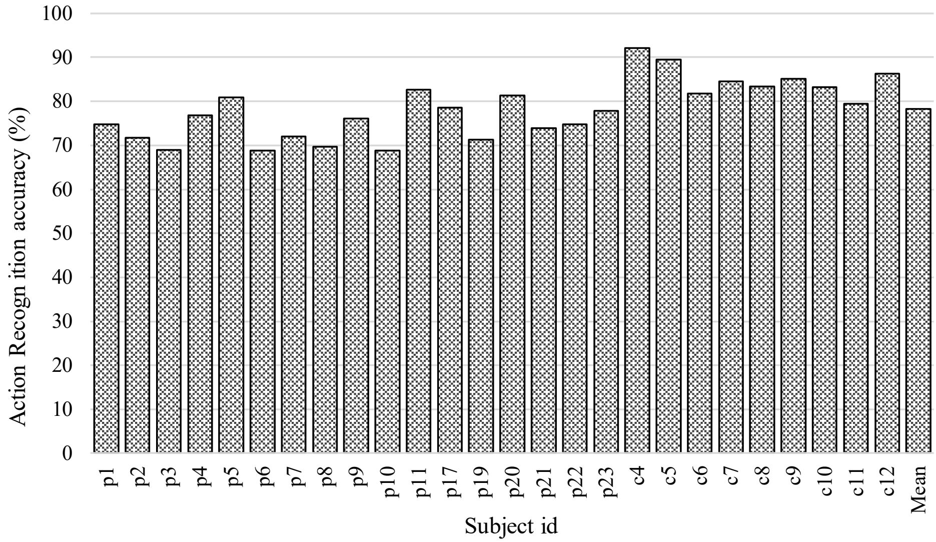


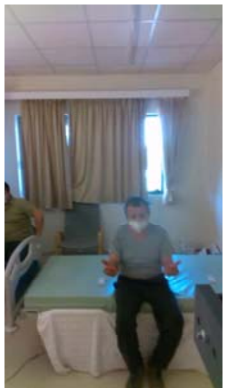

**Temporal Localization (sliding window)**

# i-Walk Validation Study 2: Results (Y3, July 2021)

- **Action Recognition: 78.1%** (10 classes)
  Patients: 75.6%
  Carers: 85.0%

- **Gesture Recognition: 72.1%**
  (4 gestures)



| | HandCross | HandCrossTurn | HandOpen | HandOpenTurn | Seated | StandUp | StandUpPrep | StandingStill | TurnStanding | Walking |
|---|---|---|---|---|---|---|---|---|---|---|
| HandCross | 0.53 | 0.19 | 0 | 0 | 0.04 | 0 | 0.01 | 0 | 0 | 0 |
| HandCrossTurn | 0.05 | 0.96 | 0 | 0 | 0 | 0 | 0 | 0 | 0.01 | 0 |
| HandOpen | 0.05 | 0.02 | 0.42 | 0.04 | 0.13 | 0 | 0.35 | 0 | 0 | 0 |
| HandOpenTurn | 0.04 | 0.02 | 0.36 | 0.49 | 0.02 | 0.01 | 0.18 | 0 | 0.03 | 0 |
| Seated | 0.07 | 0.08 | 0.06 | 0.01 | 0.88 | 0.07 | 0.18 | 0 | 0.01 | 0 |
| StandUp | 0.01 | 0 | 0 | 0 | 0.02 | 0.25 | 0.09 | 0 | 0 | 0 |
| StandUpPrep | 0.01 | 0 | 0 | 0.01 | 0.04 | 0.01 | 0.57 | 0 | 0 | 0 |
| StandingStill | 0.07 | 0 | 0.01 | 0 | 0.03 | 0.07 | 0.09 | 0.67 | 0.26 | 0.01 |
| TurnStanding | 0.08 | 0 | 0 | 0.05 | 0.02 | 0.1 | 0.05 | 0 | 0.8 | 0 |
| Walking | 0.03 | 0 | 0 | 0 | 0.01 | 0.01 | 0.13 | 0.19 | 0.09 | 0.77 |

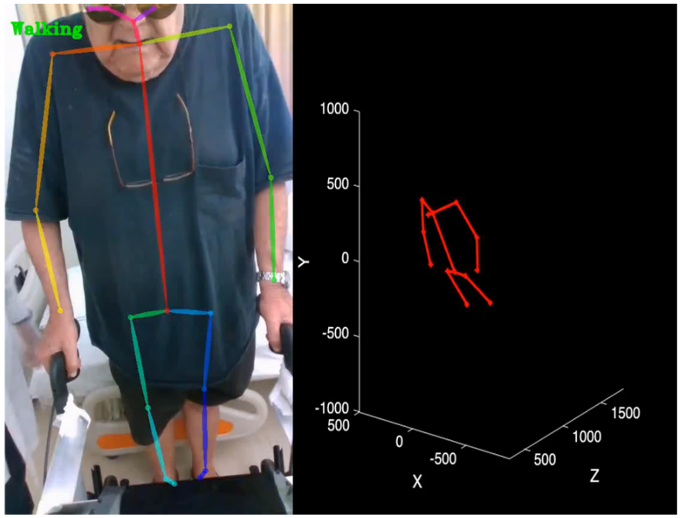
# i-Walk: Future Work on Action Recognition

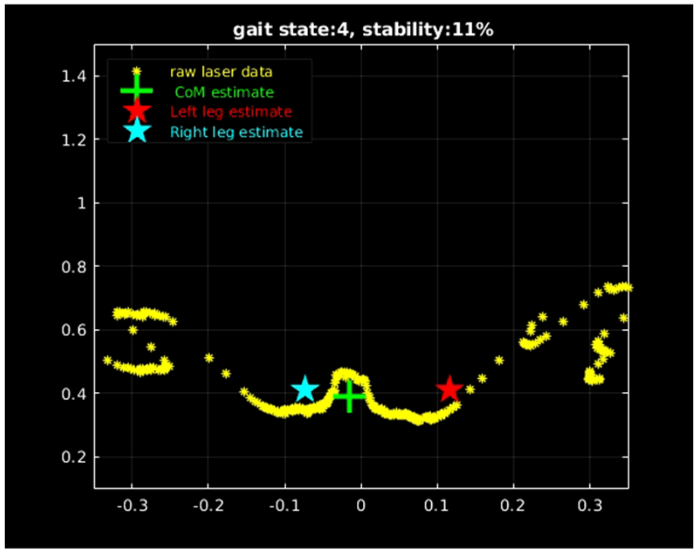
- **Combine multiple complementary channels/modalities efficiently** (pose, static appearance, depth).
  - Different actions/gestures may be more easily recognized using different modalities → focus on the most discriminative modality.
  - Different modalities are more discriminative in different scales (e.g. whole arm movement with a characteristic handshape) → multiscale-multimodal approach.

- **On-line temporal action detection** is a challenging task.
  - Most work on temporal action detection is formulated as an offline problem, in which the start and end times of actions are determined after the entire video is fully observed.
  - Action duration is variable and usually the evolution of an action cannot be predicted in advance by a few frames.
  - Encoder-decoder architectures (input = features sequence, output = sequence of per-class weights).
  - Use multiple temporal window suggestions and choose the one that classifies an action clip with the greatest certainty.
  - Estimate the optimal sliding window length each moment.

## Experimental results in Scenario 1: Rehabilitation exercises & Action Recognition

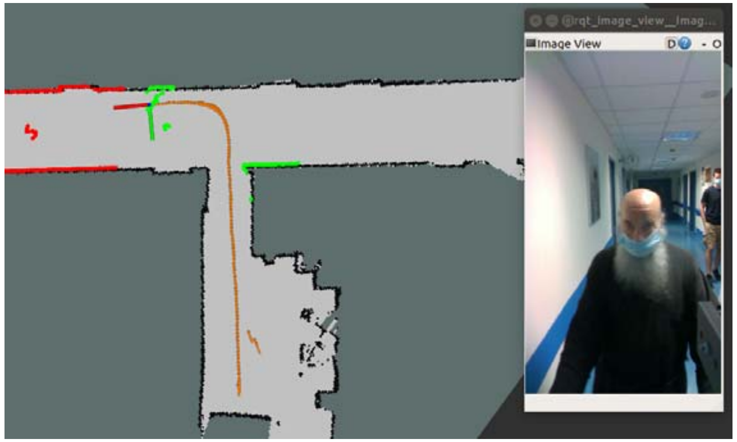
# Experimental results in Scenario 2: Transfer to bathroom



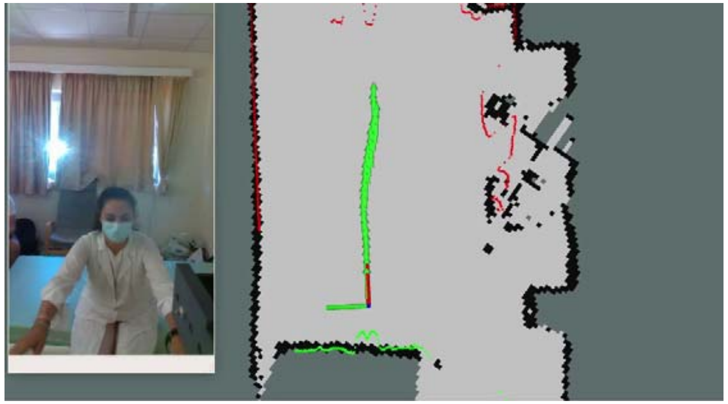
# i-Walk: Control and Motion Planning

- Localization & Mapping

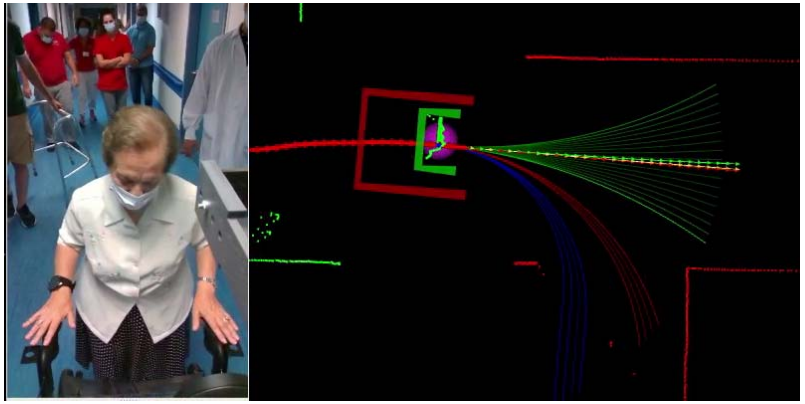



- Assisted navigation

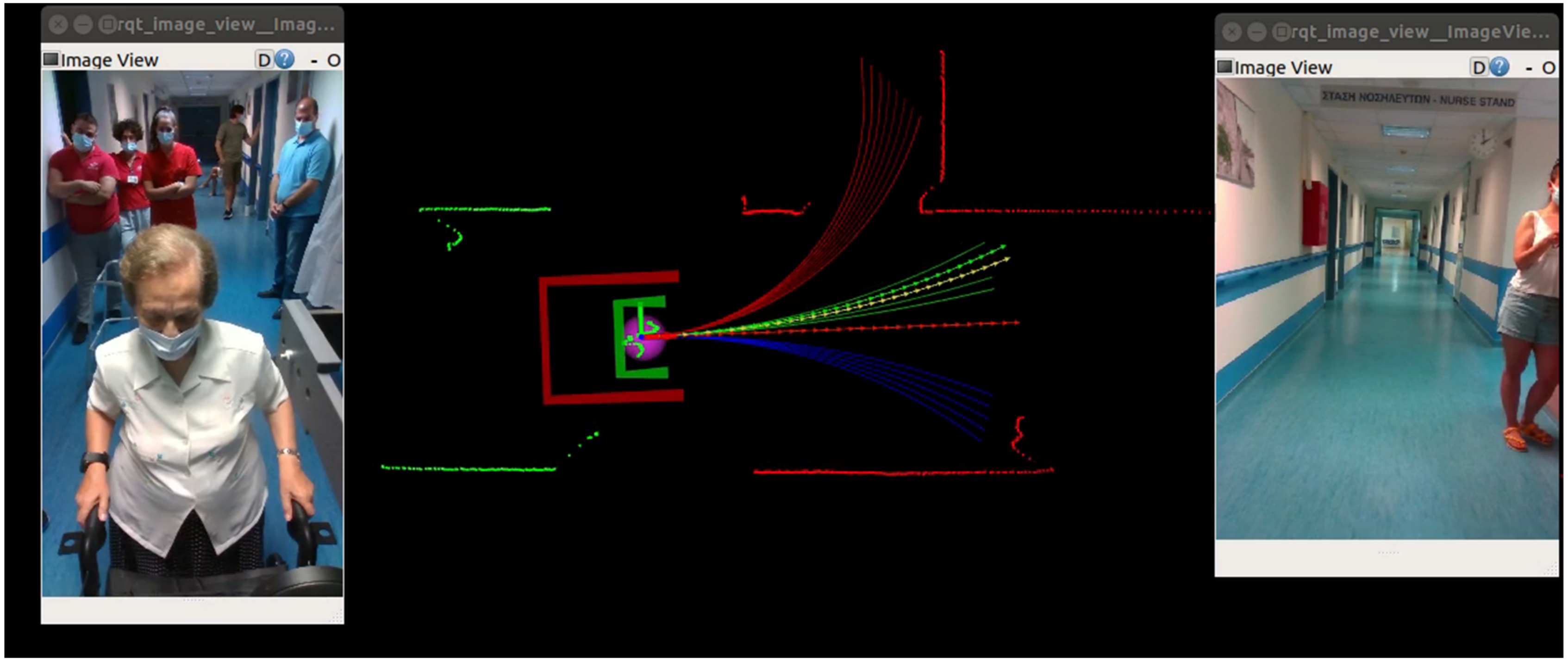

- User approach



- User front-following



# Front-Following



# Conclusions and Future Perspectives

- **Systems**: Developed integrated Mutimodal-Perception HRI systems that can work in a variety of environments (Adults, Children, Patients, multiple Robots) in Real-time to serve a variety of scenaria. Can be Trained with satisfactory performance with small datasets which can be collected with suitable developed tools.
- **Methods:** Used multiple cameras and multiple microphones to improve recognition in complex action spaces (bath, hospitals, children playrooms) to increase robustness and overall performance. Can benefit from pose estimation.
- **Datasets**: Collected many datasets from various HRI scenaria which offer the potential of investigating better deep learning methods.
- **Interdisciplinary**: collaborate with researchers from medicine and psychology to develop systems and datasets meaningful in healthcare and social robotics.
- Ongoing work:
  - **Social Robotics**: Child-Robot Interaction.

  **Future**: incremental learning, comp-light networks, engagement / attention, behavioral state estimation
  - **Healthcare Assistive Robotics**: Robot Assistants/Companions for mobility and cognitive assistance of elderly and motor-impaired people.

  **Future**: robust ASR for pathological speech and richer dialog, combine multiple cues for action recognition, multiscale-multimodal, online action detection.

For more information, demos, and current results: http://robotics.ntua.gr