Computer Vision, Speech Communication & Signal Processing Group,
National Technical University of Athens, Greece (NTUA)
Robotic Perception and Interaction Unit,
Athena Research and Innovation Center (Athena RIC)

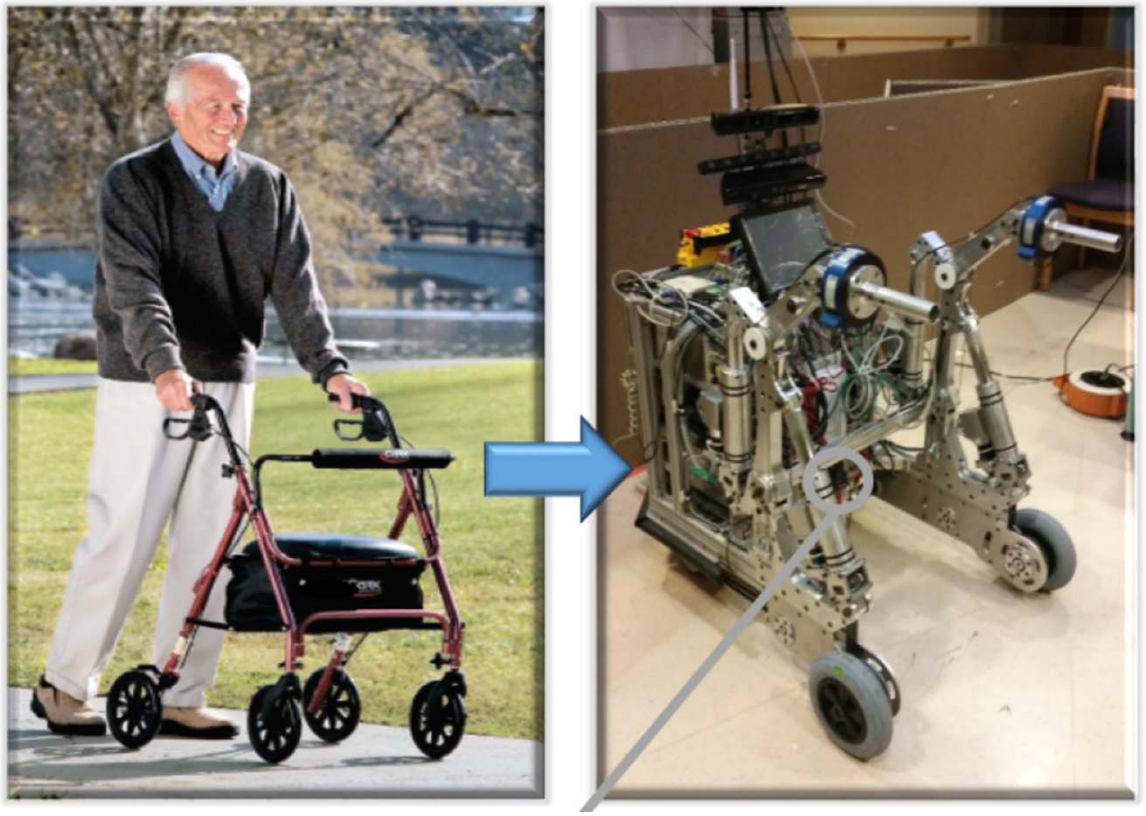
Action, Gesture and Spoken Command Recognition in Human-Robot Interaction

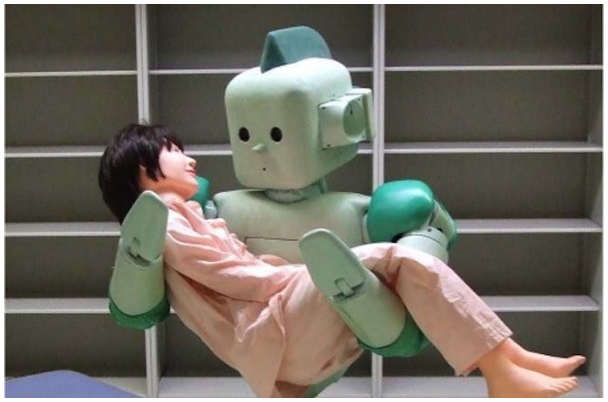
Petros Maragos

Seminar Talk at Tsinghua University, Beijing, China, 22 Sep. 2017

Multimodal HRI: Applications and Challenges

assistive robotics





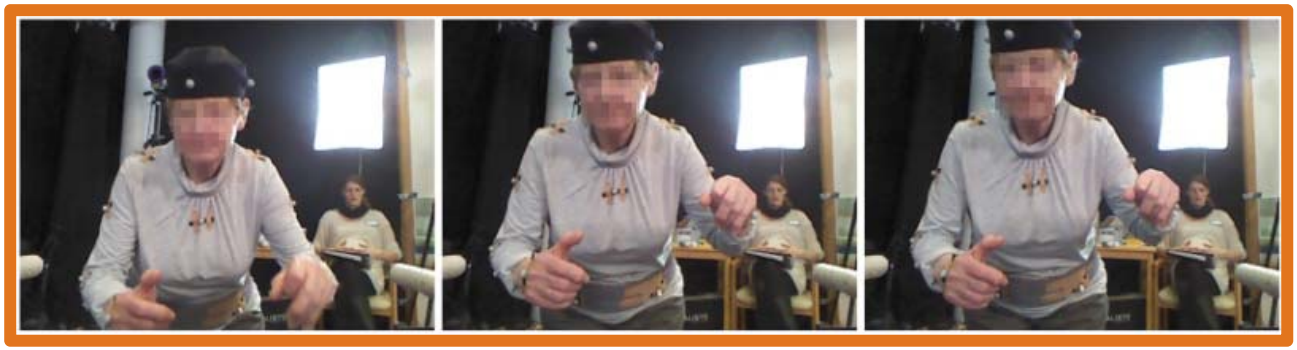
education, entertainment

- Challenges
  - Speech: distance from microphones, noisy acoustic scenes, variabilities
  - Visual recognition: noisy backgrounds, motion, variabilities
  - Multimodal fusion: incorporation of multiple sensors, integration issues
  - Elderly users, Children

Visual Activity Recognition

action → gesture → sign

Action: sit to stand



Gestures: come here, come near

Sign:
(GSL) Europe

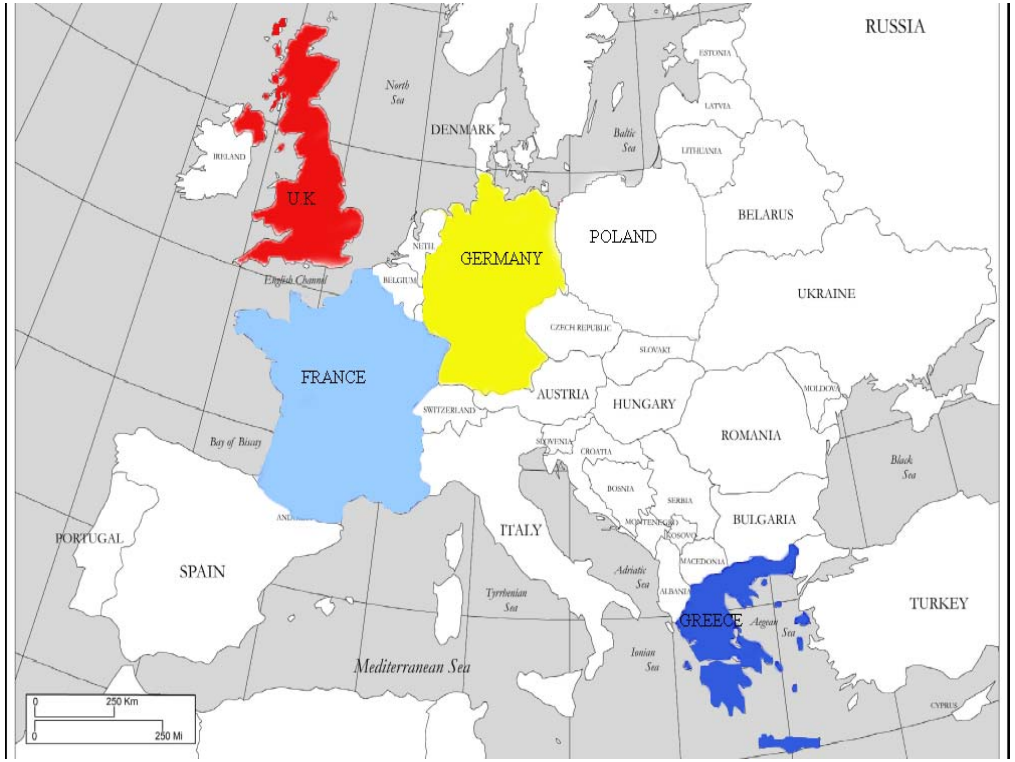
Outline

- Earlier work: Sign Language Recognition
- Action Recognition
- Gesture Recognition
- Spoken Command recognition
- Audio-Visual Fusion for improved gesture recognition
- Gait analysis
- Applications in EU projects:
  - MOBOT
  - I-SUPPORT
  - BabyRobot (if time permits)

EU Project: Dicta-Sign

Dicta-Sign: Sign Language Recognition, Generation and Modelling with Application in Deaf Communication

FP7-ICT-2007.2.2, Grant # 231135, Duration: 2009-2012 , website: http://dicta-sign.eu



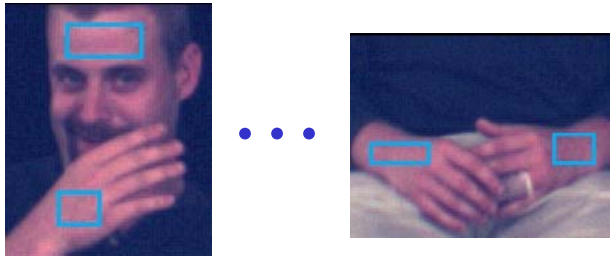
Our Team's Contribution

- Visual Processing Frontend
- Sign Language Recognition
- Data-driven and Phonetic-driven SubUnits
- Co-development of Greek Sign Language Corpus

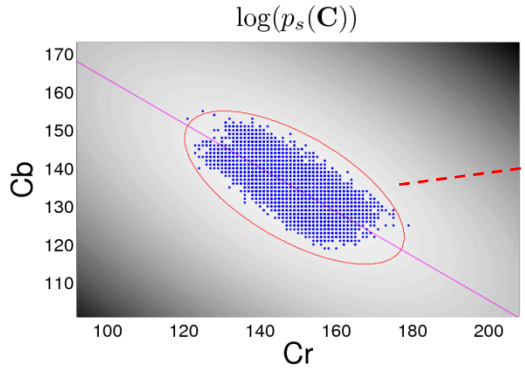
Partners: AthenaRC-ILSP (Greece), NTUA (Greece), UHH (Germany), UEA (U.K.), UniS (U.K.), CNRS (France), UPS (France), Websourd (France).

Initial Head & Hands Tracking (1/2)

- Skin color modeling



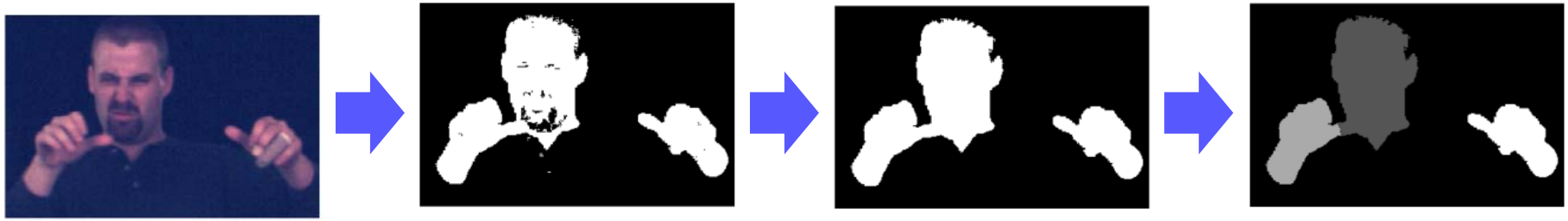
training samples



fitted pdf

color thresholding
for the skin mask S_0

- Morphological processing and segmentation of the skin mask



input

skin mask S_0

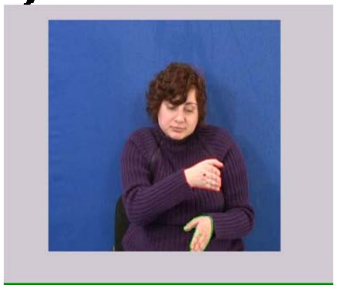
refinement of S_0
- generalized hole filling
- area opening

segmentation
- connected components
- competitive rec. opening

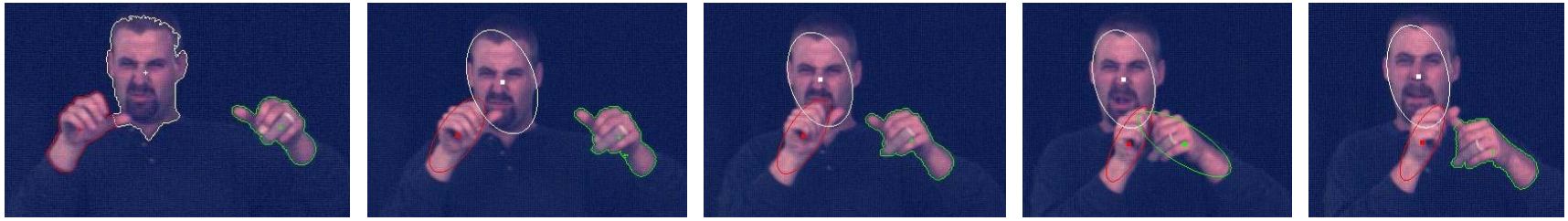
[S. Theodorakis, V. Pitsikalis and P. Maragos, Image & Vision Comp. 2014]

[A. Rousos, S. Theodorakis, V. Pitsikalis and P. Maragos, J. Mach. Learn. Res. 2013]

Initial Head & Hands Tracking (2/2)



- Main parts of tracking:
  - ellipses fitting,
  - fwd-bkwd prediction (ellipse parameters),
  - template matching (ellipse bbox),
  - probabilistic constraints

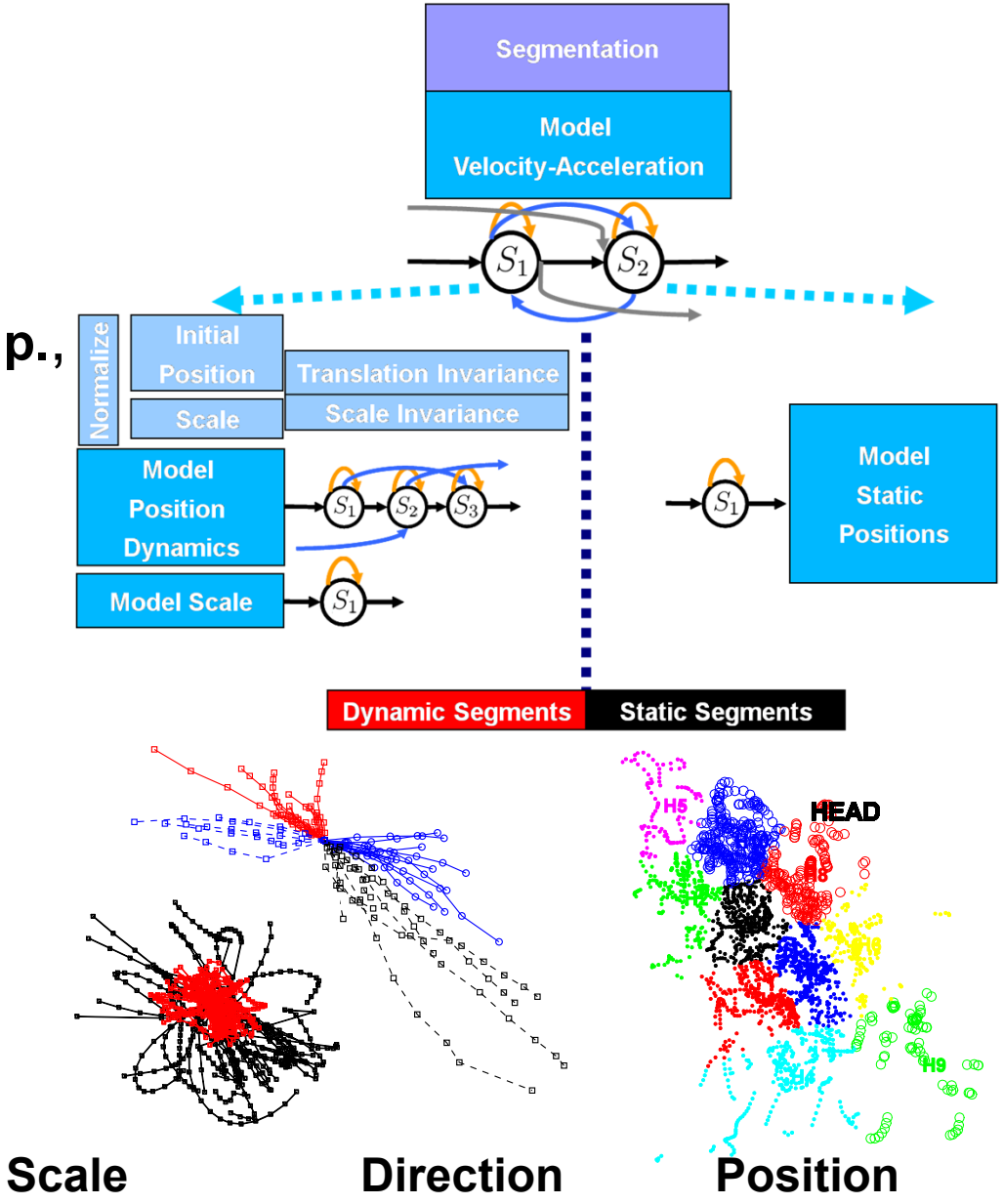


- Output: set of skin region masks together with one or multiple body-part labels

Hand Subunits: Dynamic-Static

- Provide Missing Sequential Structure
- **Dynamic-Static** Segmentation: **Intuitive, Unsup.,** Segments + Labels
- Separate Modeling, SUs, Clustering w.r.t. Feature type, Parameters and Architecture; Normalize features
- Training, HMMs
- Data-Driven Lexicon

[S. Theodorakis, V. Pitsikalis and P. Maragos, IVC 2014]



Handshape Modeling: AAM, Dynamic & Static Priors

- Shape-Appearance (SA) Representation

cropped frame ***I(x)*** hand mask *M*

SA image ***f(x)***

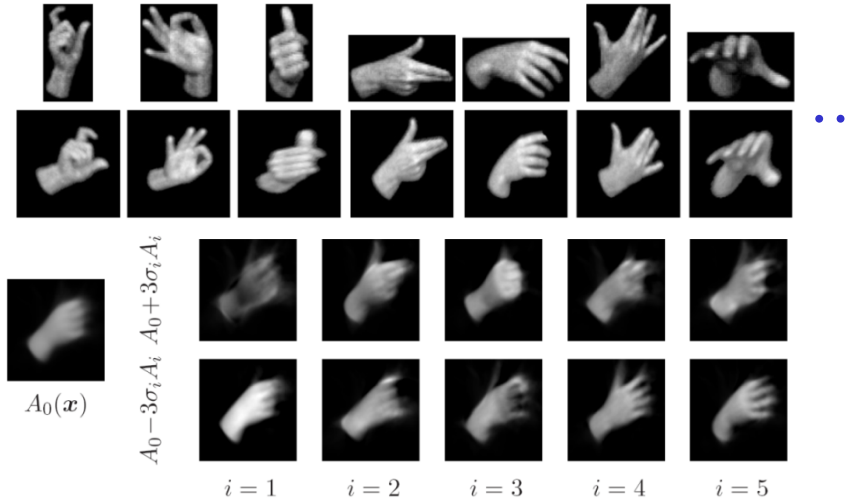
$$f(\boldsymbol{x}) = \begin{cases} g(\boldsymbol{I}(\boldsymbol{x})), & \text{if } \boldsymbol{x} \in M \\ -c_b & \text{else} \end{cases}$$

- Generative model

$$f(W_{\boldsymbol{p}}(\boldsymbol{x})) \approx A_0(\boldsymbol{x}) + \sum_{i=1}^{N_c} \lambda_i A_i(\boldsymbol{x})$$

$W_{\boldsymbol{p}}(\boldsymbol{x})$: *2D affine transform with parameters* $\boldsymbol{p} \in \Re^6$

- Training of the Model
  - Affine alignment of the training set
    - generalization of the procrustes analysis
    - iterative manual feedback
  - PCA to learn $A_i(\boldsymbol{x})$
    - keep only Nc=35 components



- Fitting : Find parameters **λ,p** that minimize:

$$E(\boldsymbol{\lambda}, \boldsymbol{p}) = E_{rec}(\boldsymbol{\lambda}, \boldsymbol{p}) + w_S E_S(\boldsymbol{\lambda}, \boldsymbol{p}) + w_D E_D(\boldsymbol{\lambda}, \boldsymbol{p})$$

Handshape Classification

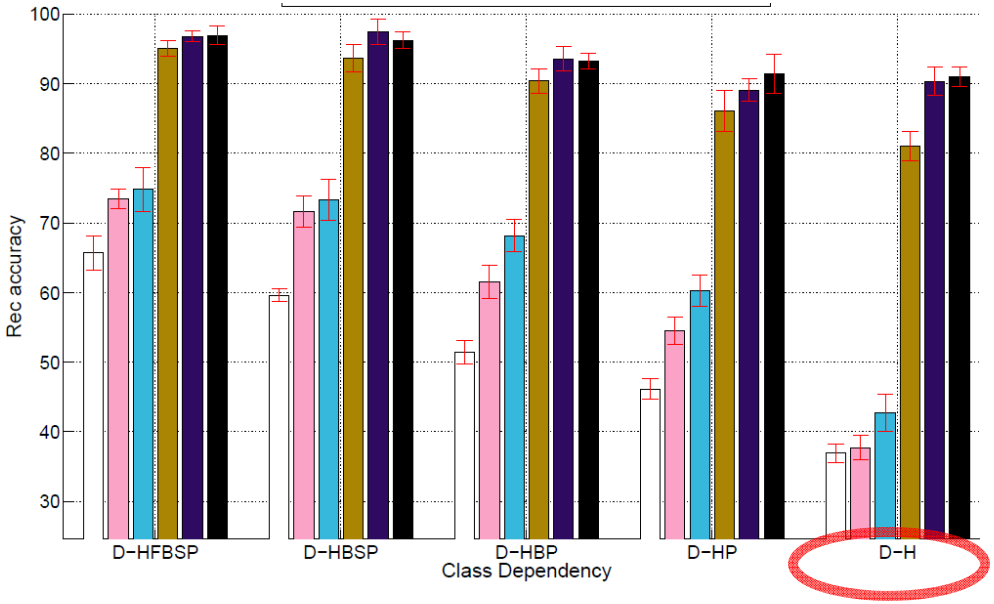
BU400

Proposed: Affine SAM
Direct Similarity SAM
Direct Translation Scale SAM

Aff-SAM
DTS-SAM
DS-SAM

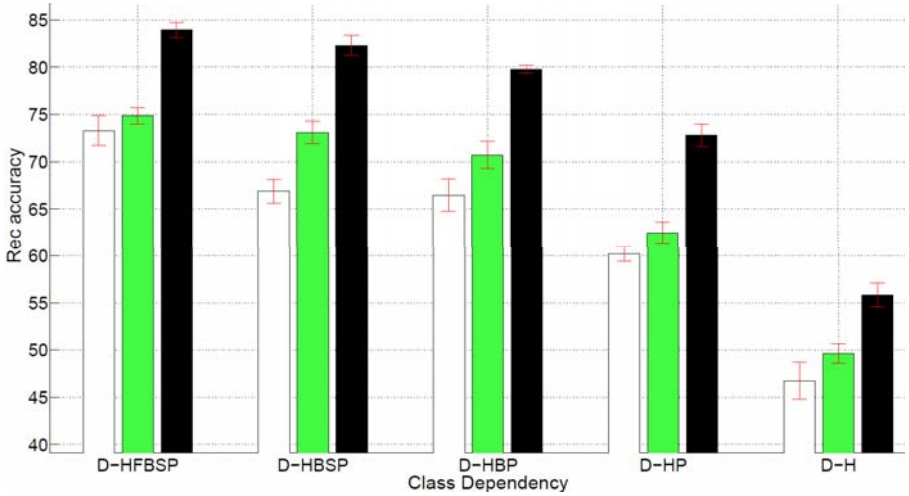
Fourier Descriptors (FD)
Moments (M)
Region Based (RB)

FD
RB
M



No occlusions

With Occlusions

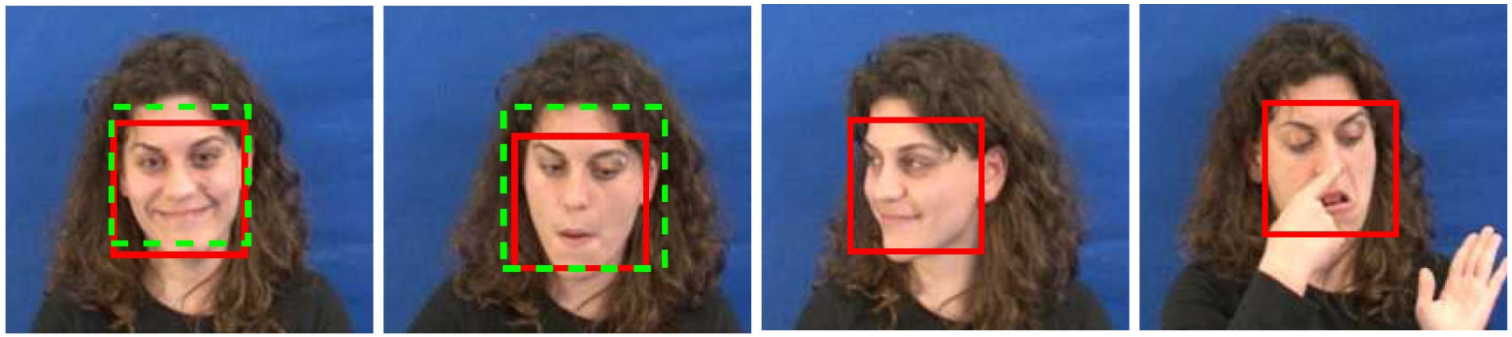


Class Dependency

Class Dependency	Annotation Parameters F	S	B	P	H
D-HFSBP	D	D	D	D	D
D-HSBP	*	D	D	D	D
D-HBP	*	*	D	D	D
D-HP	*	*	*	D	D
D-H	*	*	*	*	D

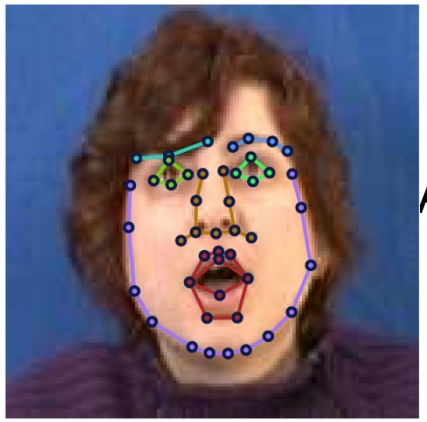
[Rousos, Theodorakis, Pitsikalis and Maragos, JMLR 2013]

Face Modeling and Feature Extraction



- Face Detection

(Viola-Jones with Kalman Filtering)



- Global and Local

Active Appearance Modeling

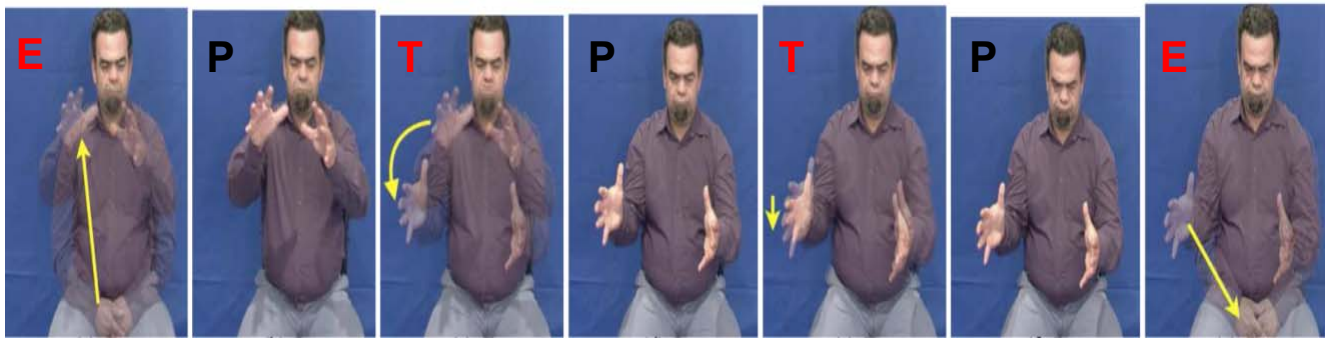
- AAM Tracking

on GSL Continuous Corpus

[E. Antonakos, V. Pitsikalis and P. Maragos, JIVP 2014]

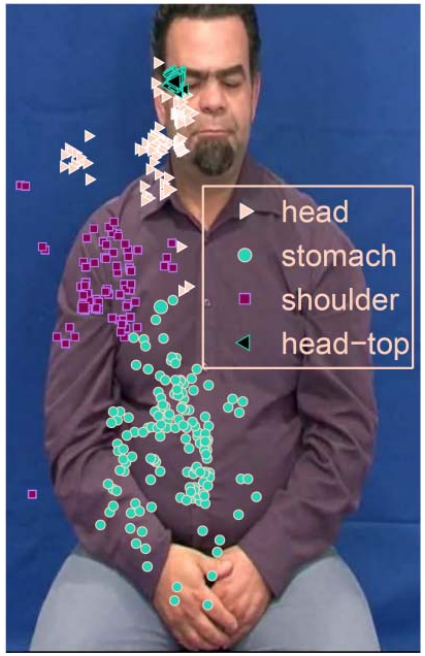
PDTS Phonetic Subunits (Examples)

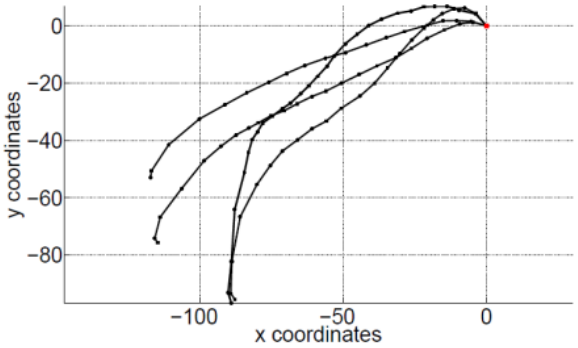
Pitsikalis, Theodorakis, Vogler, Maragos, CVRP-2011
Workshop on Gesture Recognition (Best paper award).



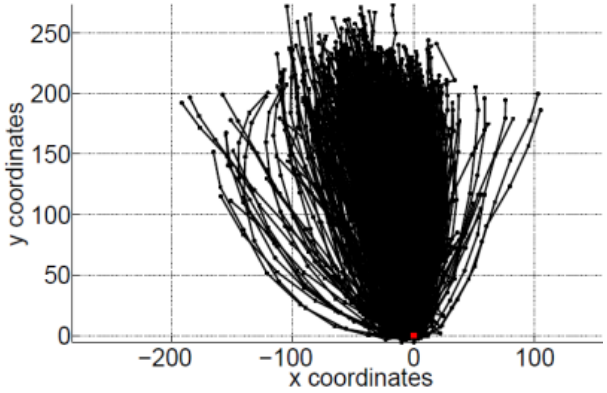
Transition/Epenthesis Segments
Superimposed Initial-End Frames + Arrow

Posture/Detention Segments
Single Frame

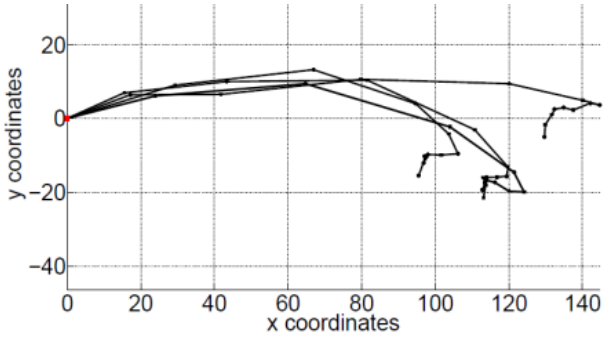




(c) T-down-right



(a) E-to-head

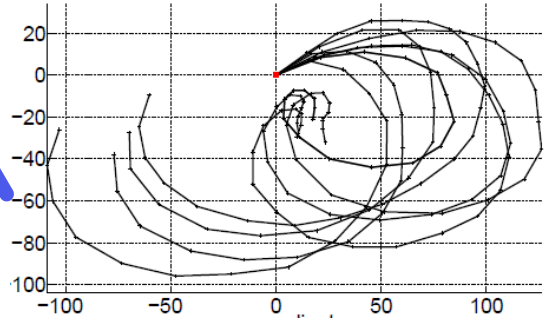


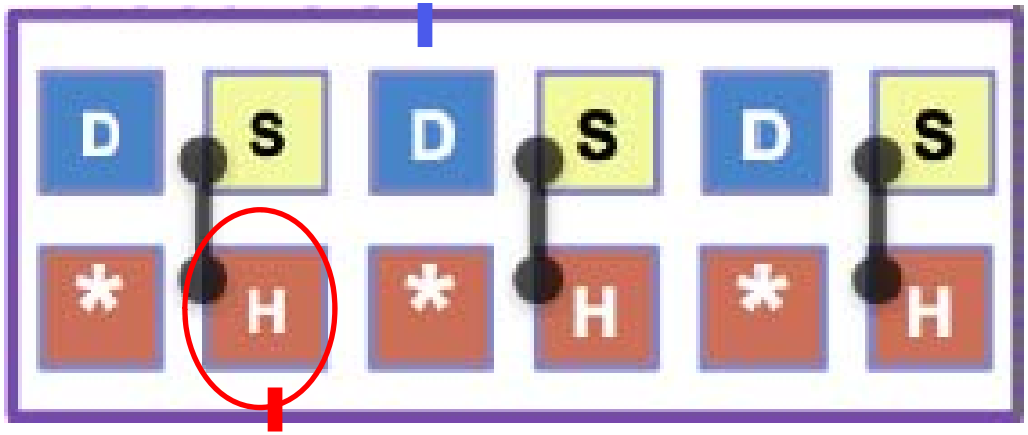
(d) T-in-left

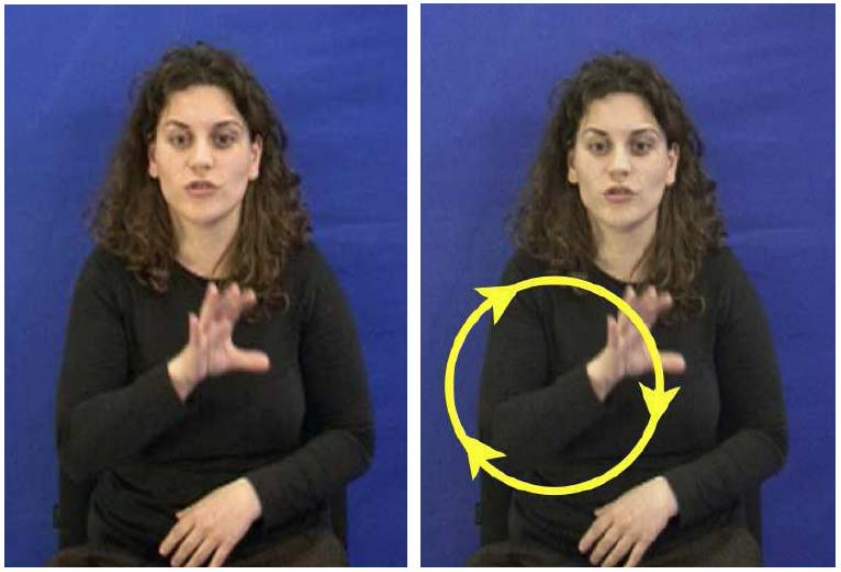
Movement-Position and Handshape Fusion

Transition PDTS Subunit

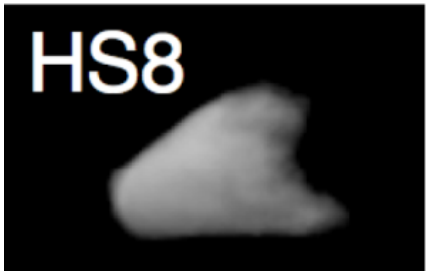
T-circular





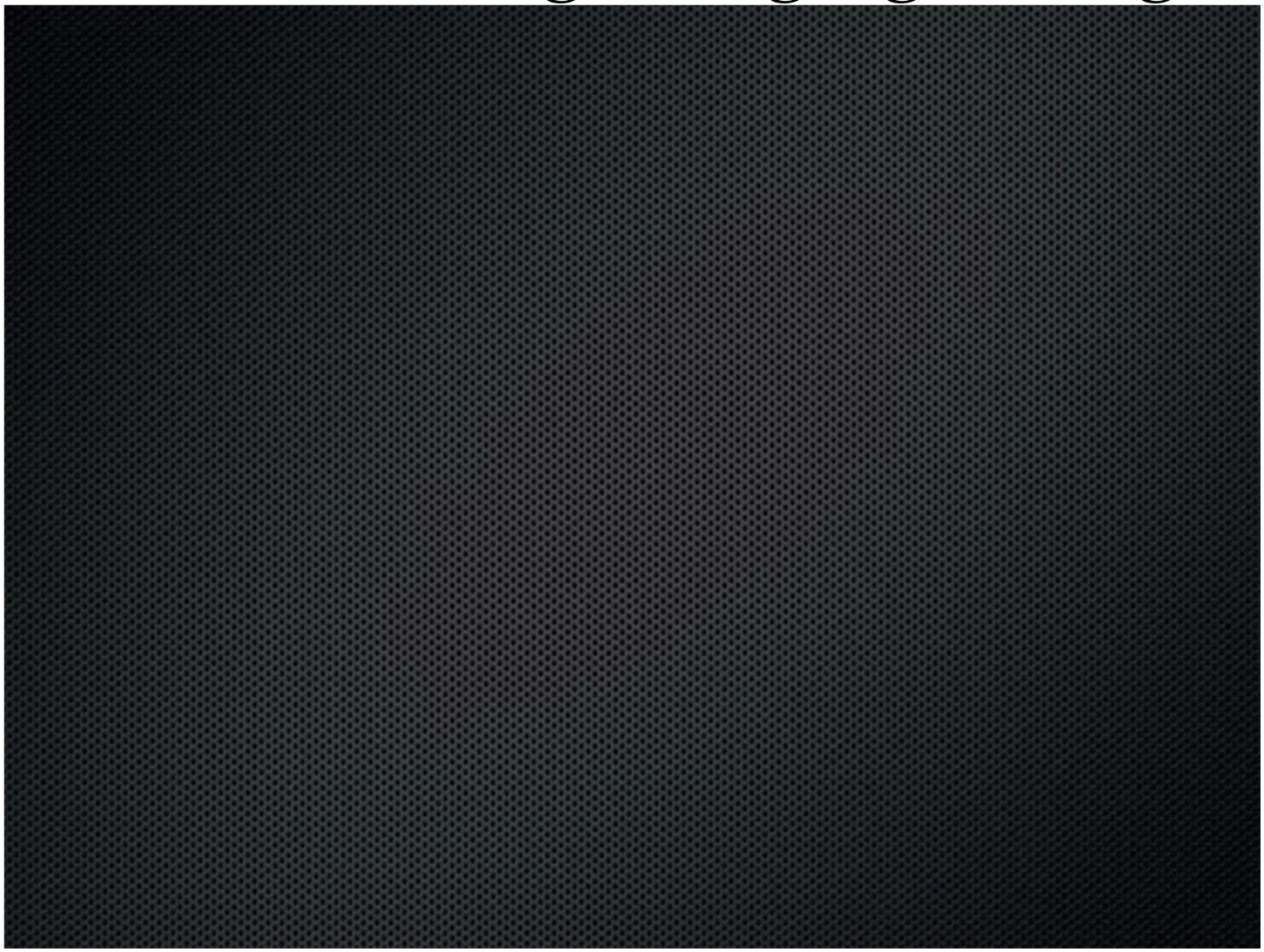


P5-HS8 T-circular



Handshape Subunit

Demo: Continuous Sign Language Recognition

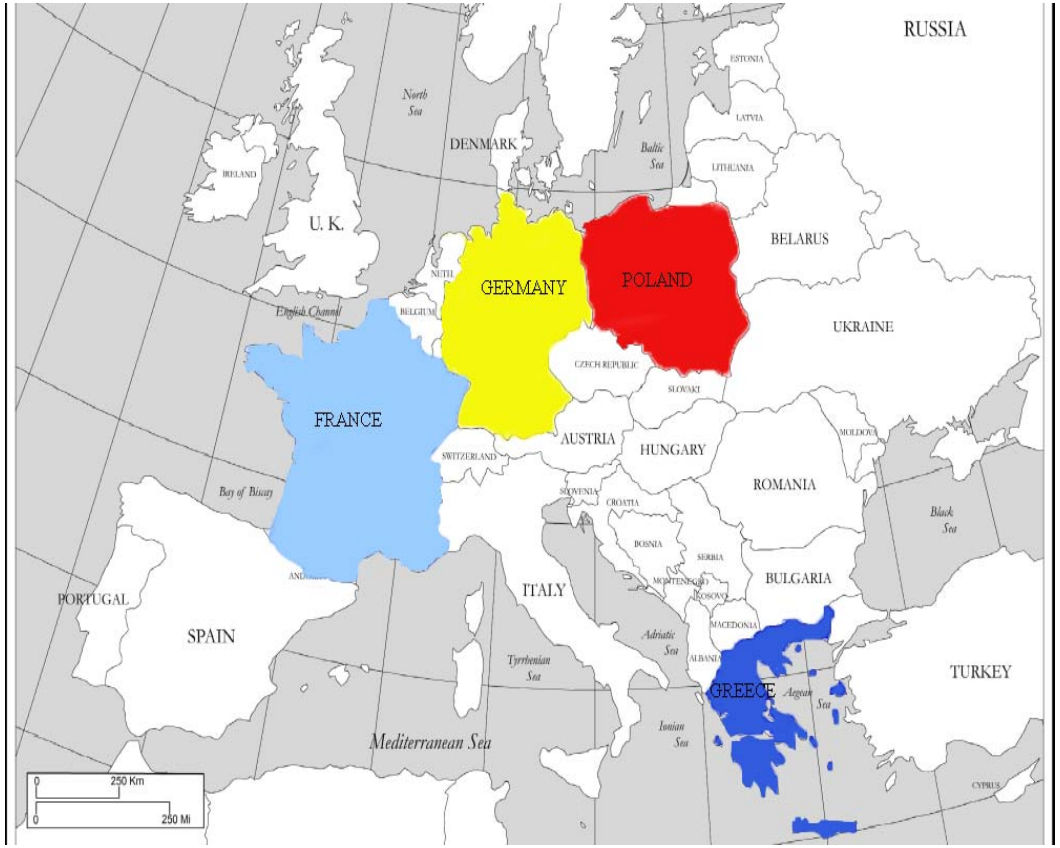


[S. Theodorakis, V. Pitsikalis and P. Maragos, IVC 2014]

EU Project: MOBOT

MOBOT: Intelligent Active MObility Aid RoBOT integrating Multimodal Communication

FP7- ICT-2011.2.1, Grant # 600796, Duration: 2013-2016, website: http:// mobot-project.eu/



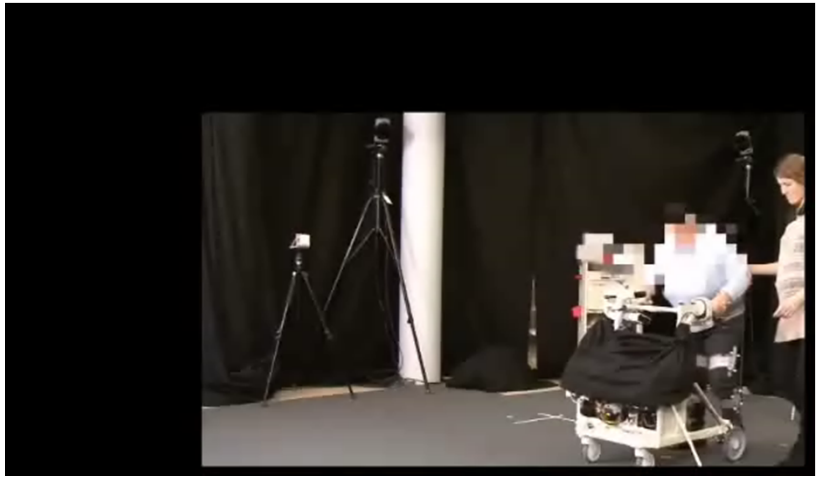
Our Team's Contribution

- Action and Gesture Recognition
- Spoken Command Recognition
- Audio-Visual Fusion
- Multimodal HRI
- Gait analysis and Robotics Control

Partners: TUM / UWE (Germany / UK), ICCS-NTUA (Greece), INRIA (France), AthenaRC (Greece), UHEI (Germany), Bethanien (Germany), DIAPLASIS (Greece), ACCREA (Poland).

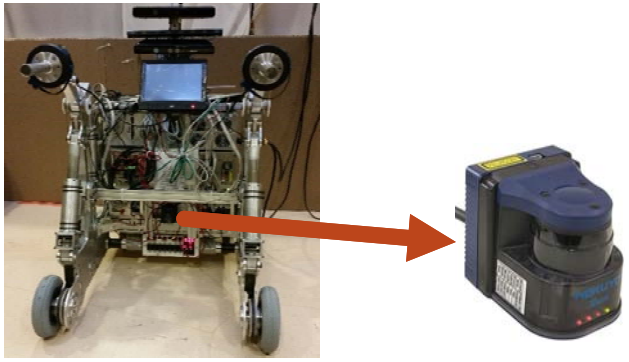
Motivation

Experiments conducted at
Bethanien Geriatric Center Heidelberg



Mobility & **Cognitive** impairments, prevalent in **elderly** population, limiting factors for ***Activities of Daily Living*** **(ADLs)**

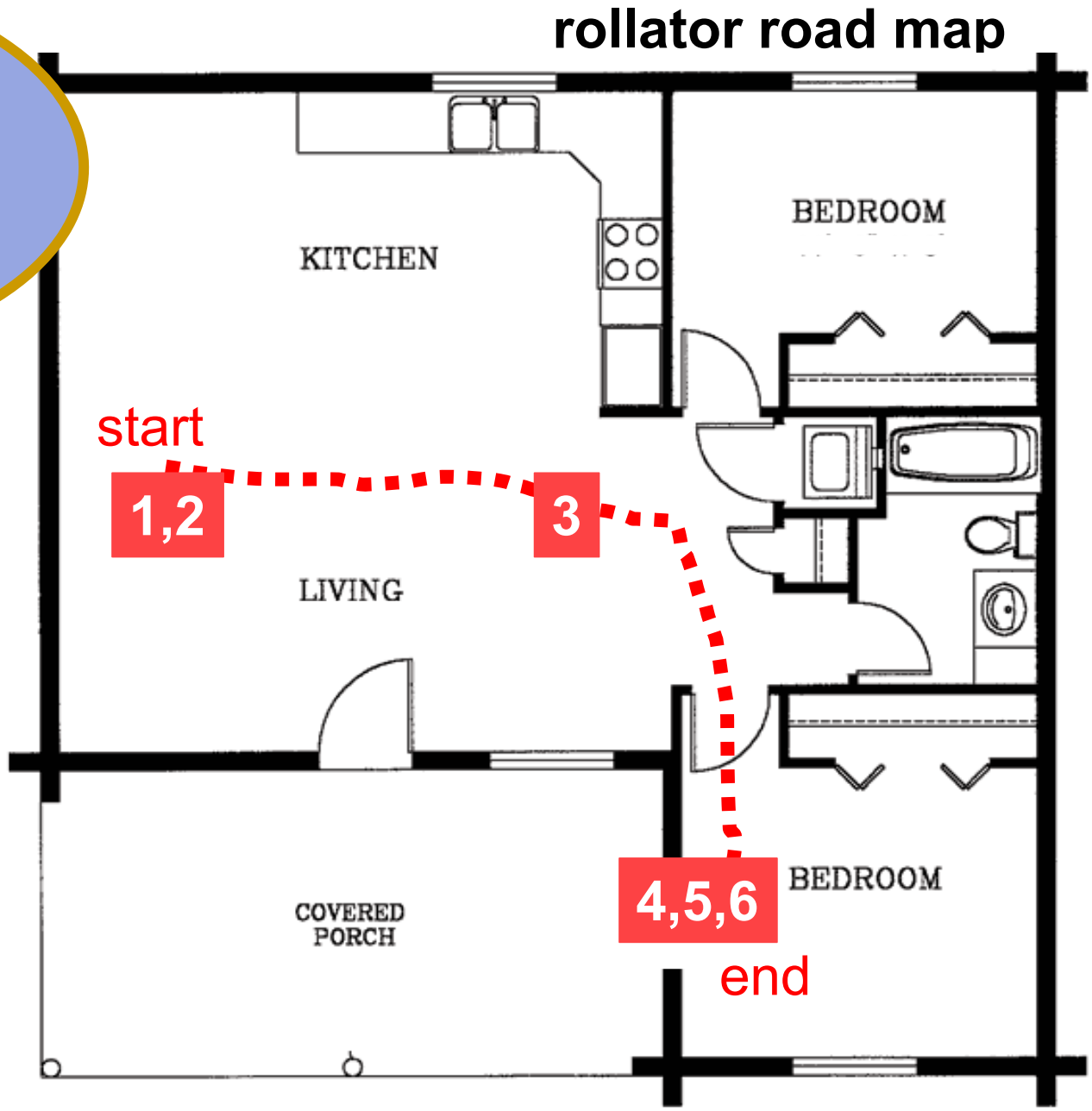
Intelligent assistive devices (robotic Rollator) aiming to provide ***context-aware*** and ***user-adaptive*** mobility (**walking**) assistance



MOBOT rollator

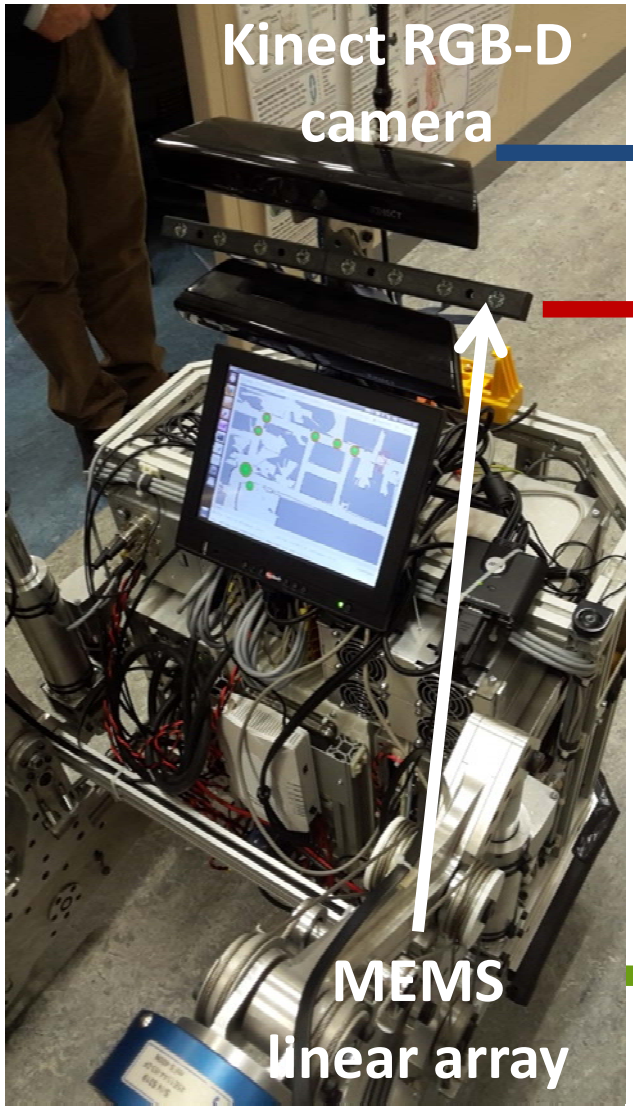
MOBOT robotic rollator and assistive scenario

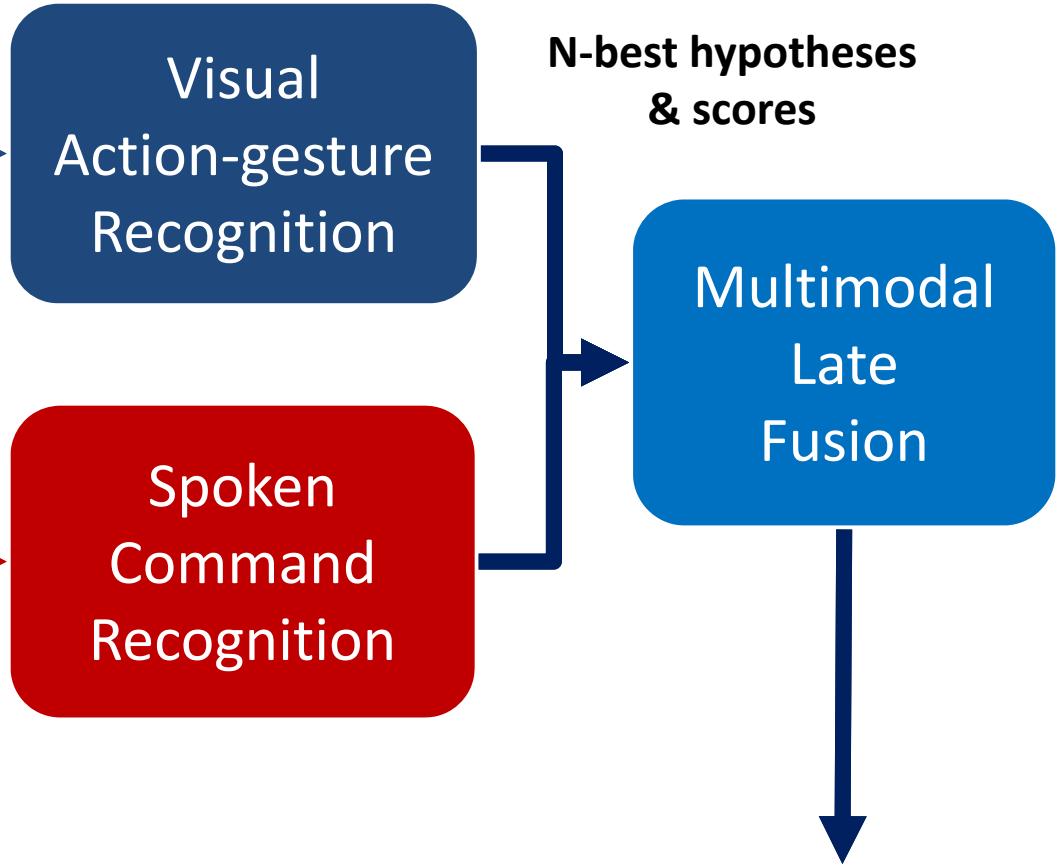
|1| I want to stand up
|2| go to the bedroom
|3] turn right
|4| stop
|5| I want to sit down
|6| what time is it?



Overall working system: audio-gestural command recognition

MOBOT robotic platform

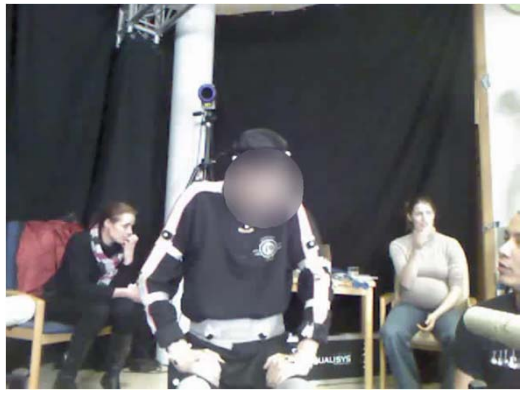




Overview scheme of our multimodal interface

Multi-Sensor Data for HRI

Kinect1 RGB Data

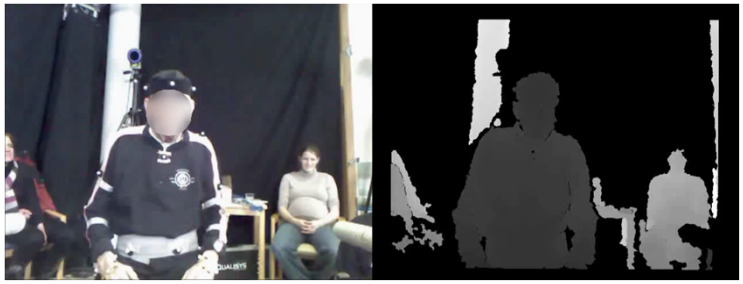


Kinect Depth Data

Kinect1 RGB

Kinect1 Depth

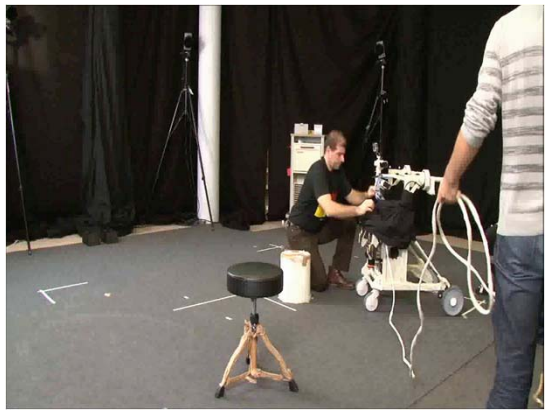
MEMS Audio Data



Go Pro RGB Data

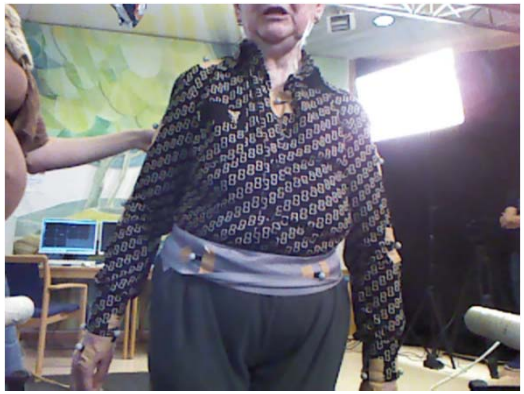
HD1 Camera Data

HD2 Camera Data

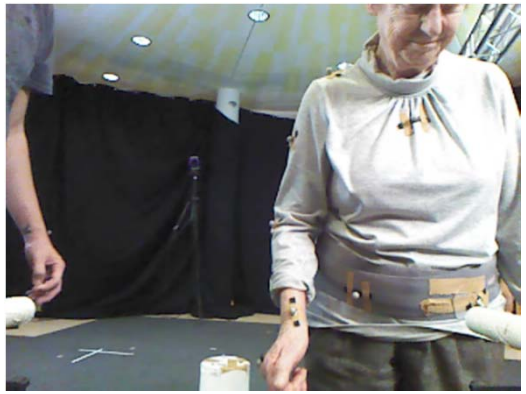


Action Sample Data and Challenges

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



Stand-to-Sit – P1



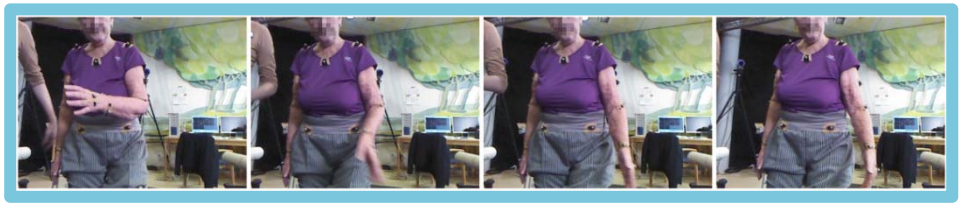
Stand-to-Sit – P3

Stand-to-Sit – P4

ACTION RECOGNITION

Visual action recognition pipeline

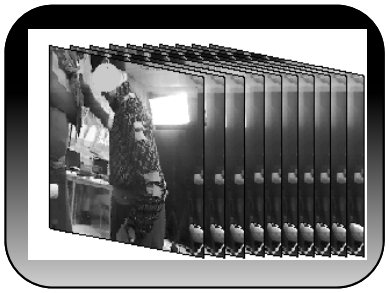
Sit to Stand

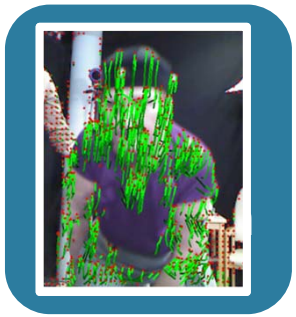


Walk

Video

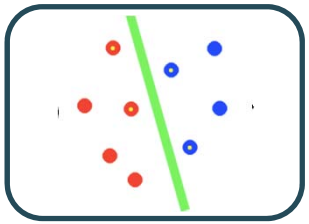
Sliding Window





Visual Feature Extraction

Feature Encoding



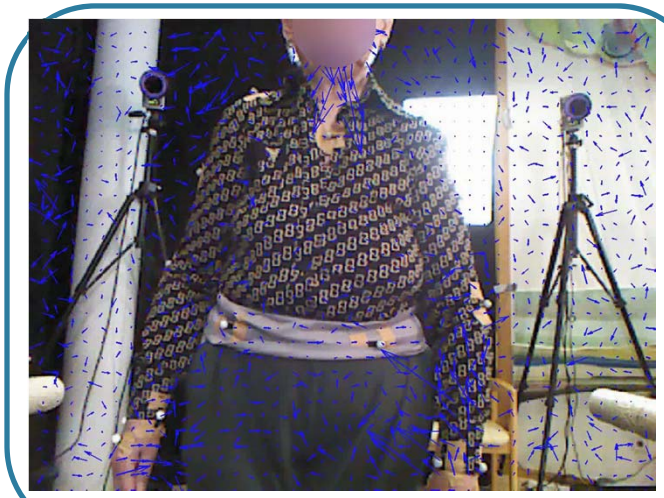
Classifier

Post-processing

Recognized Sequence

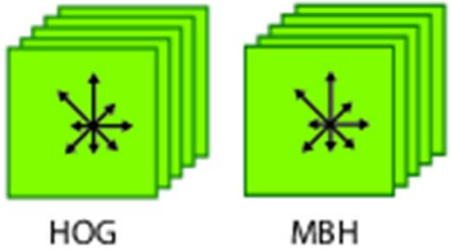
Visual Front-End

Video



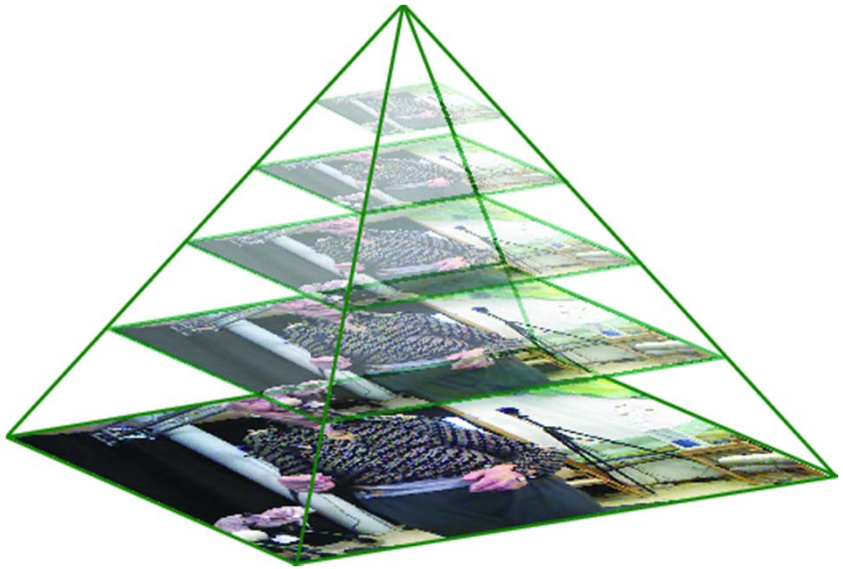
Optical Flow

Dense Trajectories

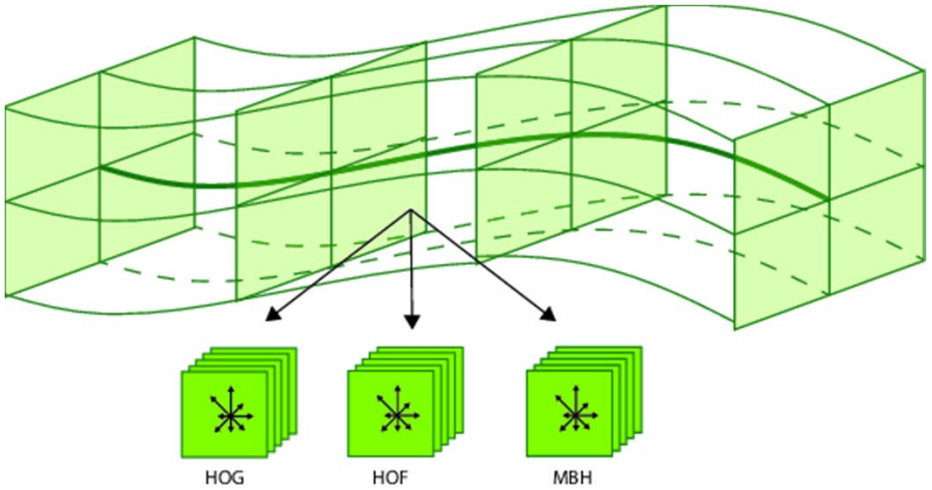


Feature Descriptors

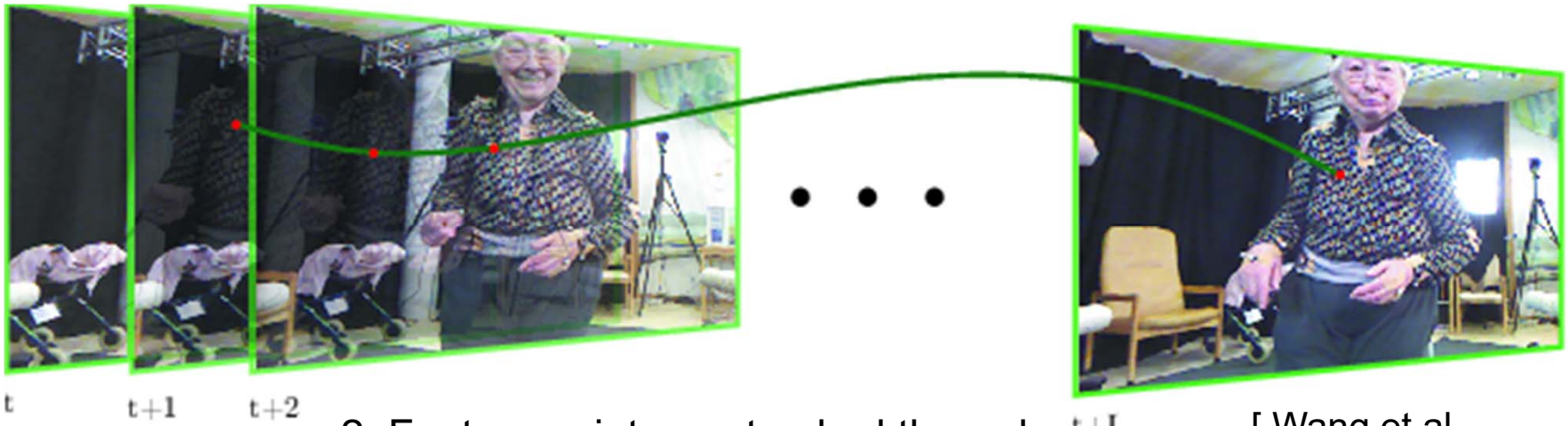
Features: Dense Trajectories



1. Feature points are sampled on a regular grid in multiple scales



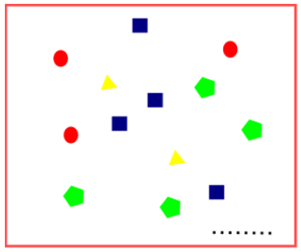
3. Descriptors are computed in space-time volumes along trajectories

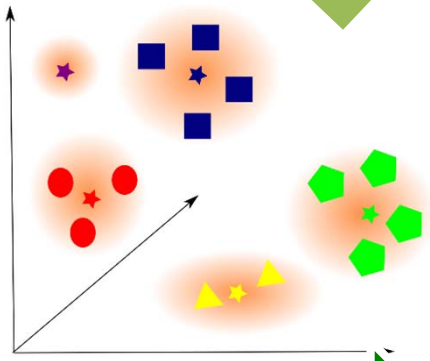


2. Feature points are tracked through consecutive video frames

[Wang et al. IJCV 2013]

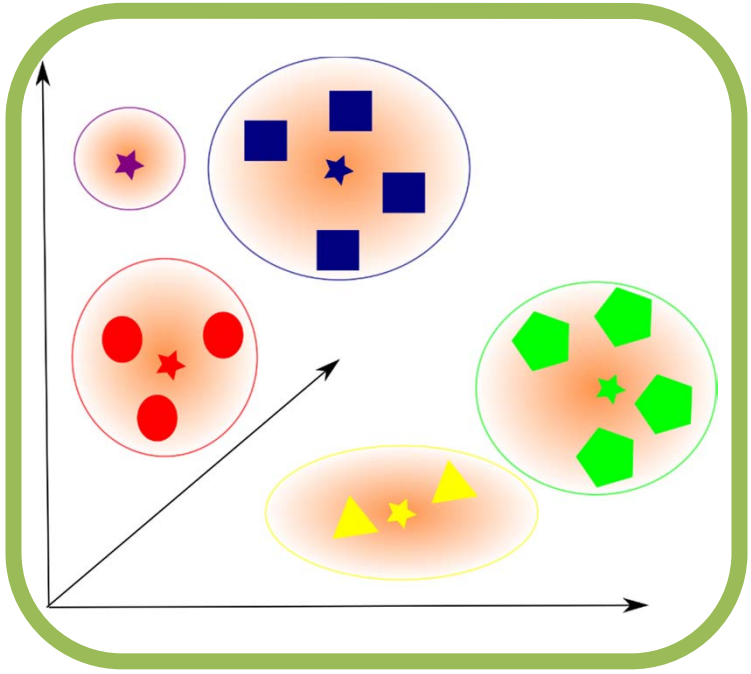
Feature Encoding

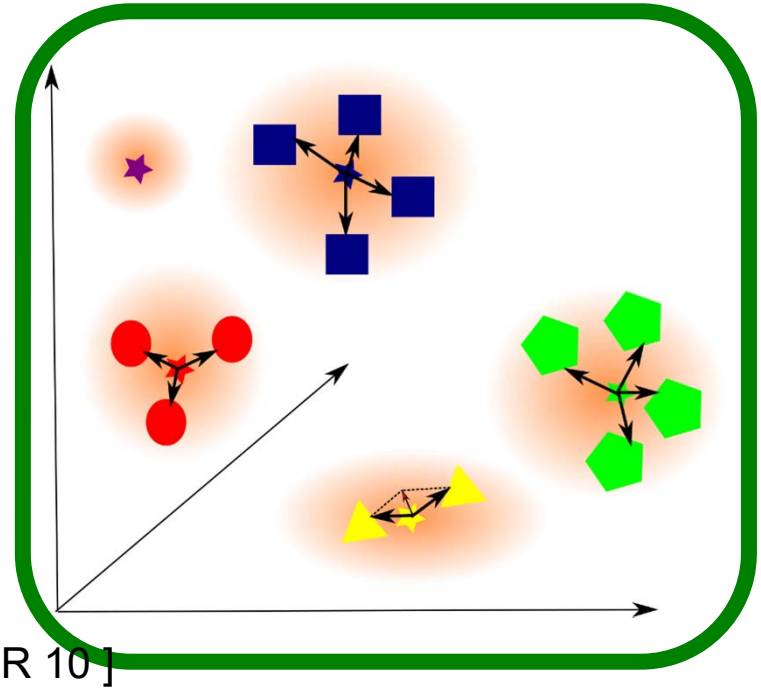


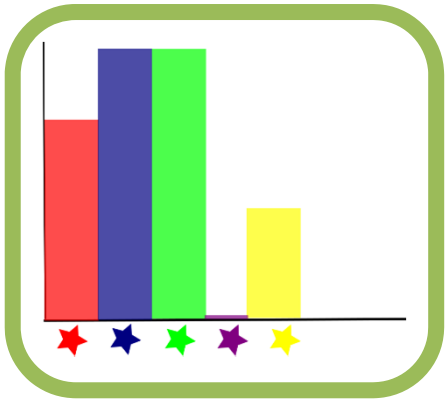


BOF

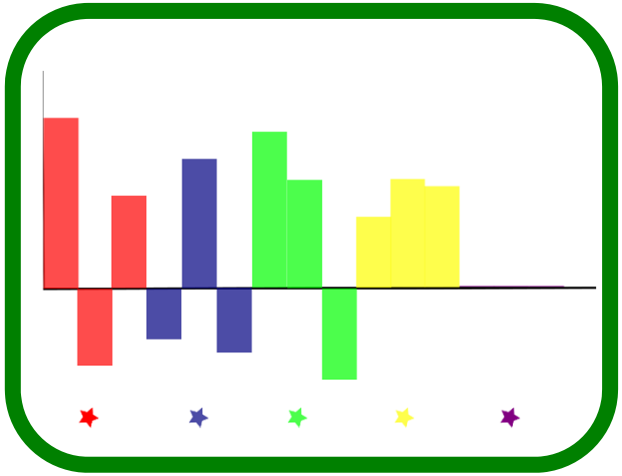
VLAD







BOF - Size: K

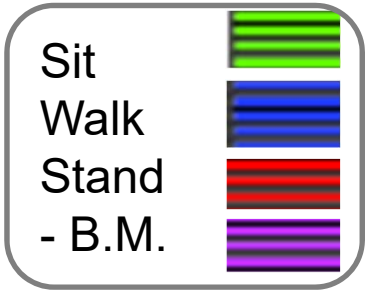


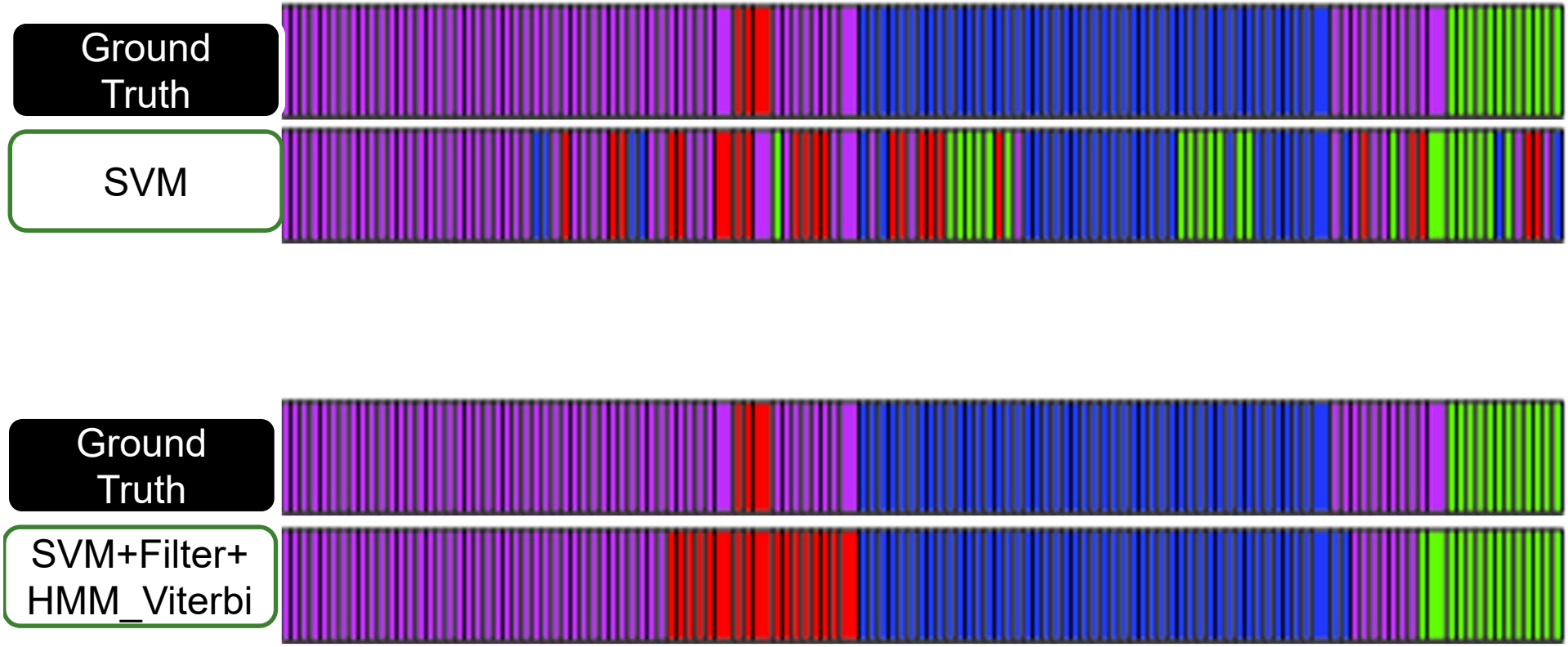
VLAD - Size: K*D

[Jegou et.al CVPR 10]

Action Recognition video and results

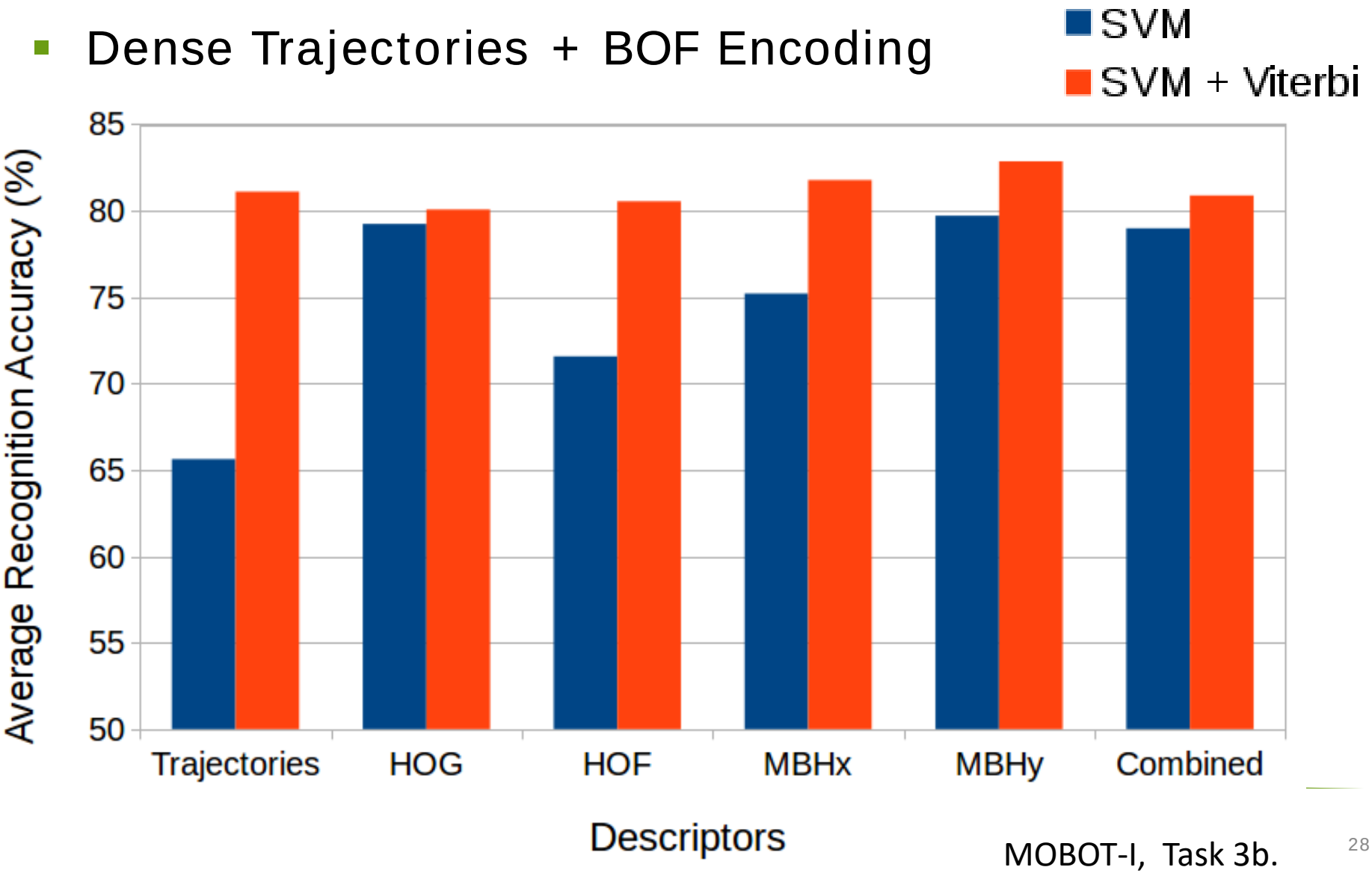
Temporal Segmentation Results





Action Recognition Results (4a, 6p)**: Descriptors + Post-processing Smoothing**

- Dense Trajectories + BOF Encoding



GESTURE RECOGNITION

Gesture Recognition

Challenging task of recognizing human gestural movements:

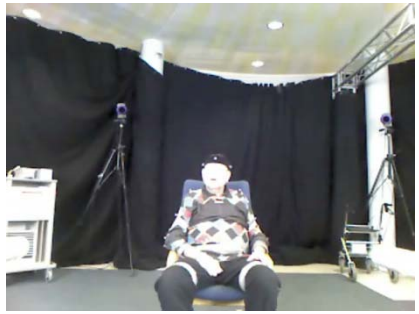
- Large variability in gesture performance.
- Some gestures can be performed with left or right hand.

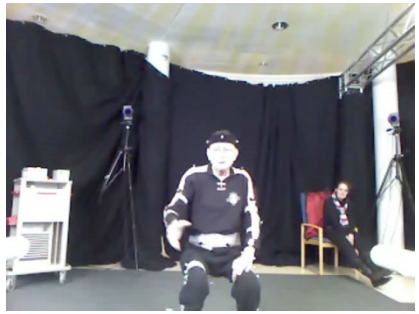
Come Closer



Park

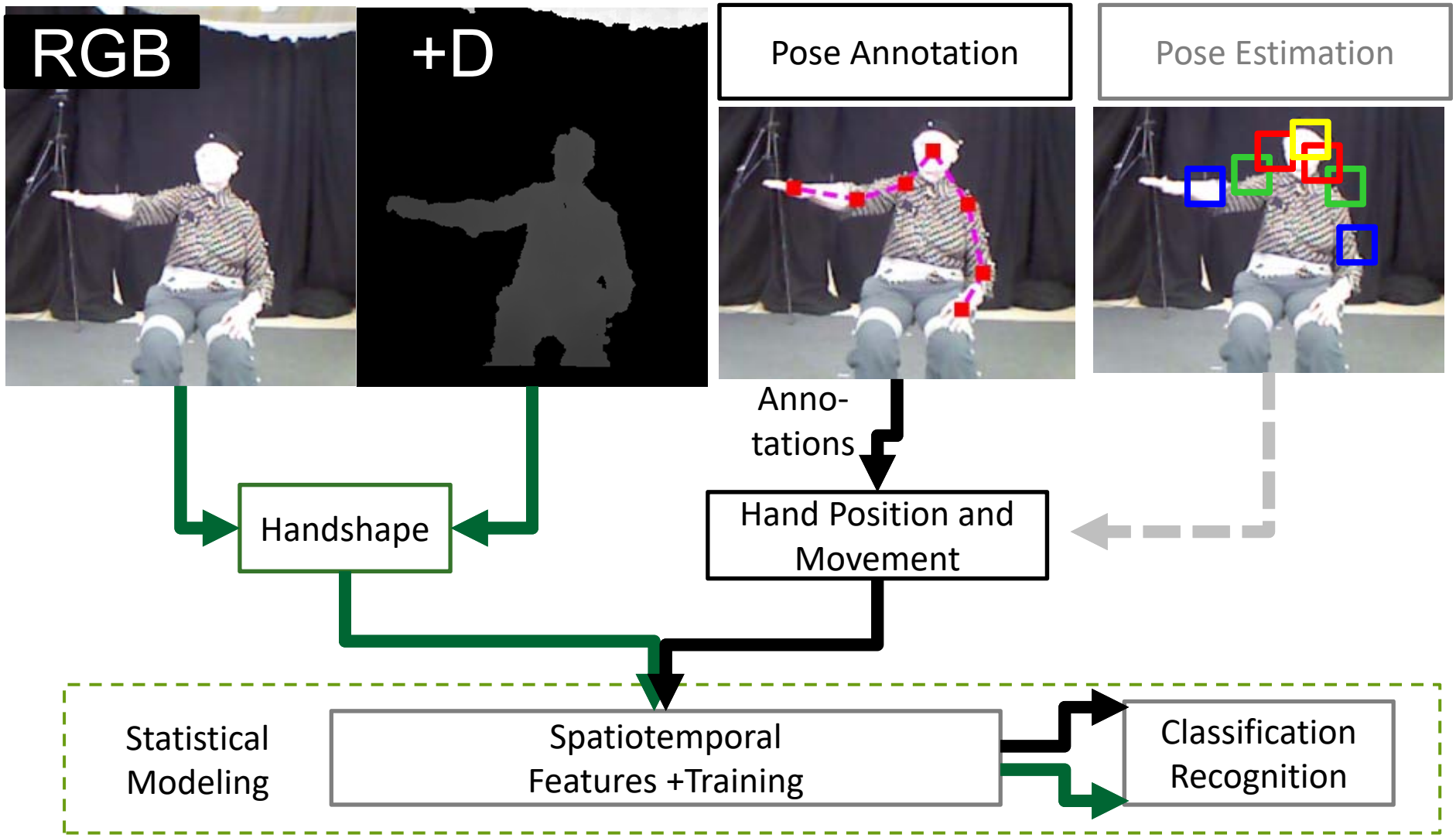
I want to Sit Down

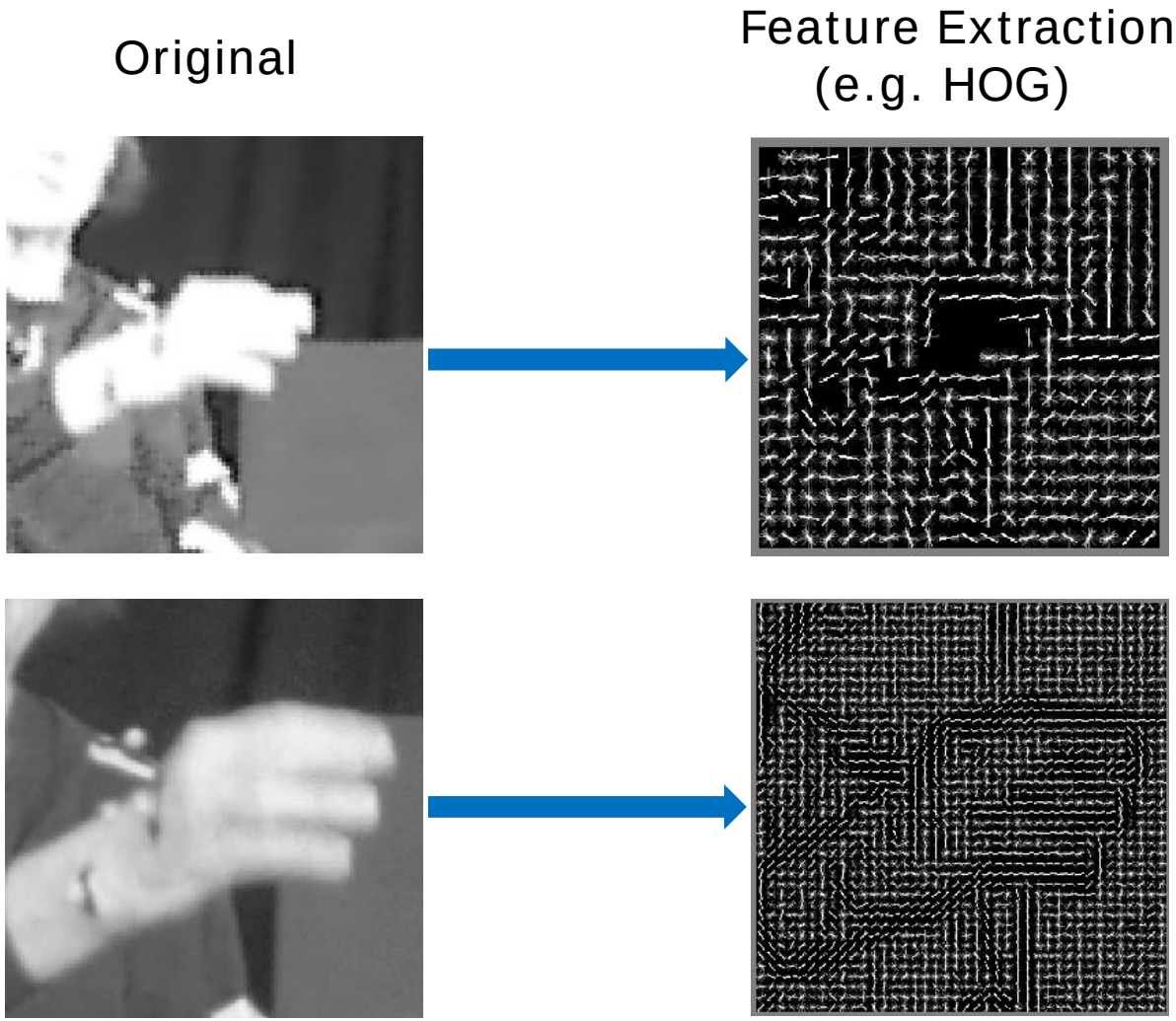
I want to Perform a Task

Overview: Visual Gesture Recognition

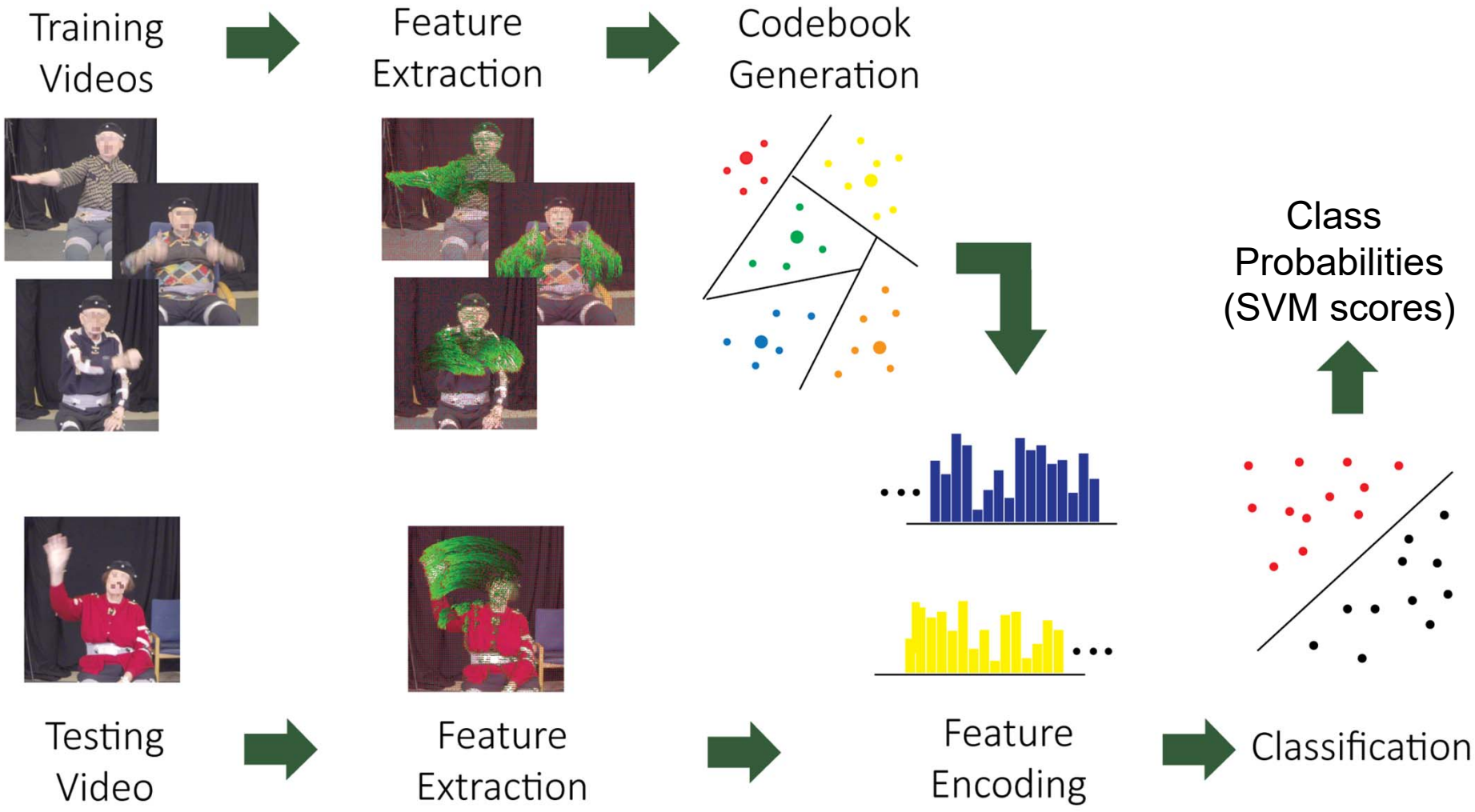


Handshape Feature Extraction

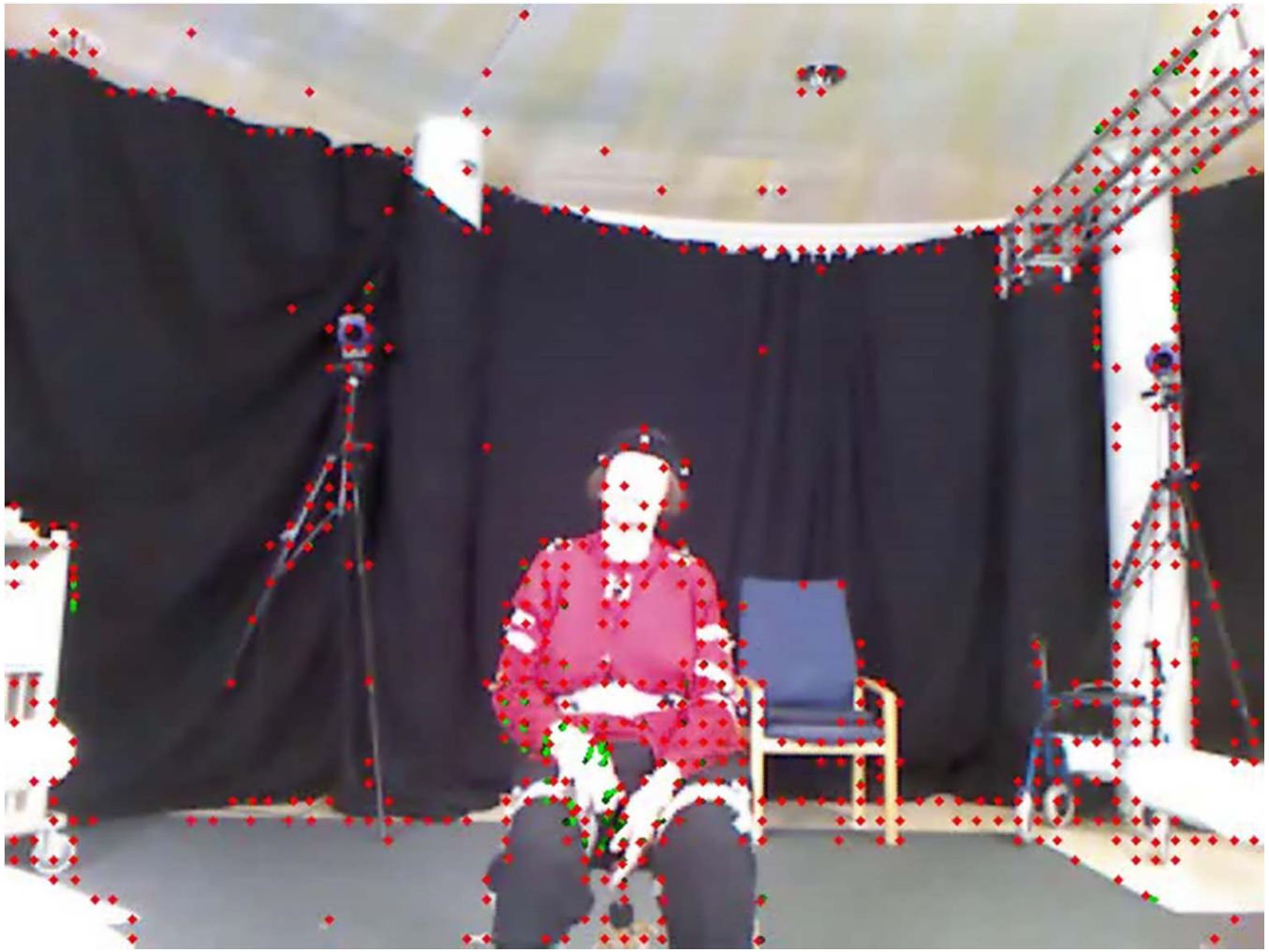
- Kinect
- GoPro



Visual Gesture Classification Pipeline



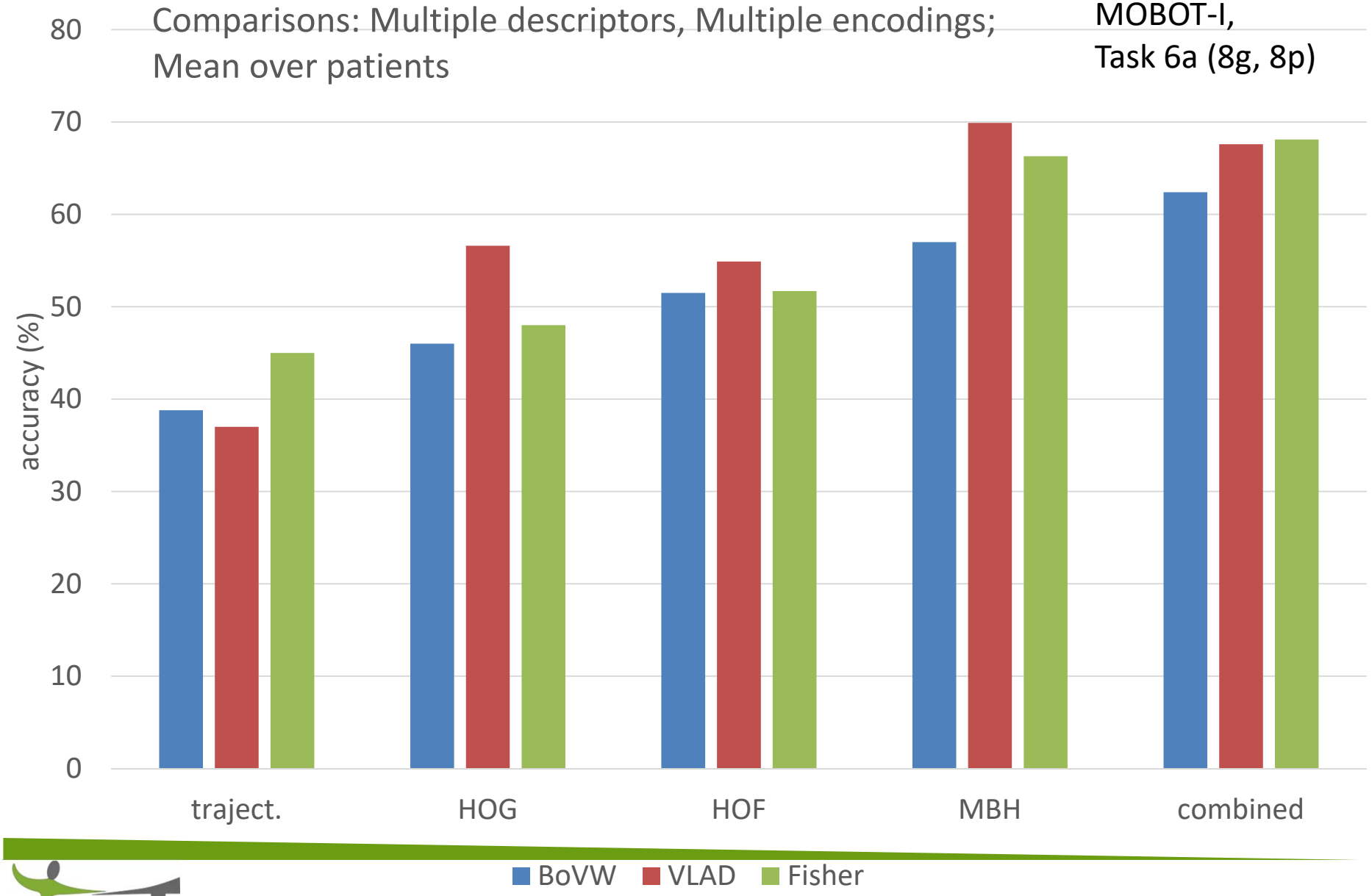
Applying Dense Trajectories on gesture data



Extended results on Gesture Recognition

Comparisons: Multiple descriptors, Multiple encodings;
Mean over patients

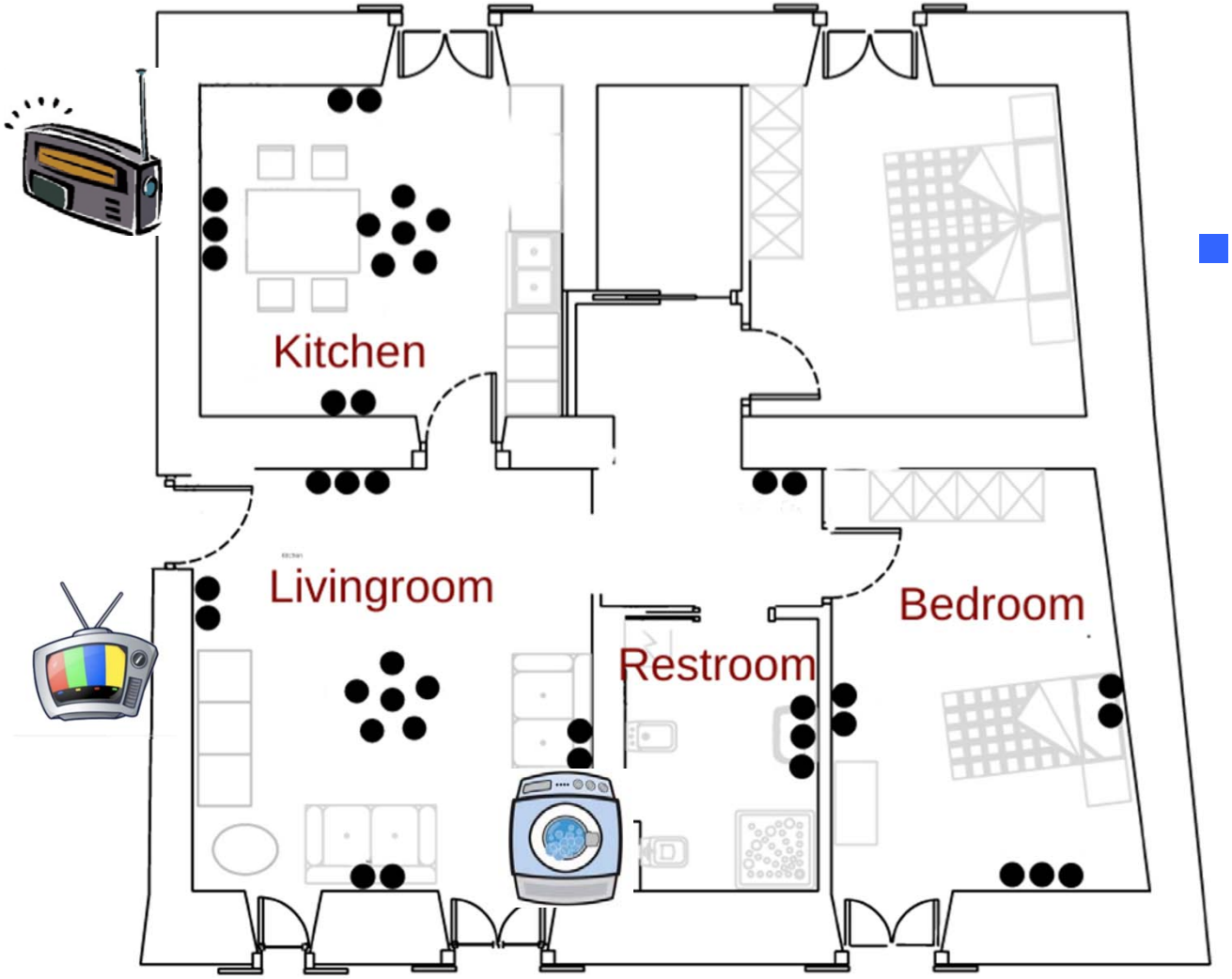
MOBOT-I,
Task 6a (8g, 8p)



SPOKEN COMMAND RECOGNITION

Late afternoon, Anna interacts with the system

Smart home voice interface



Sweet home listen!
Turn on the lights in the living room!

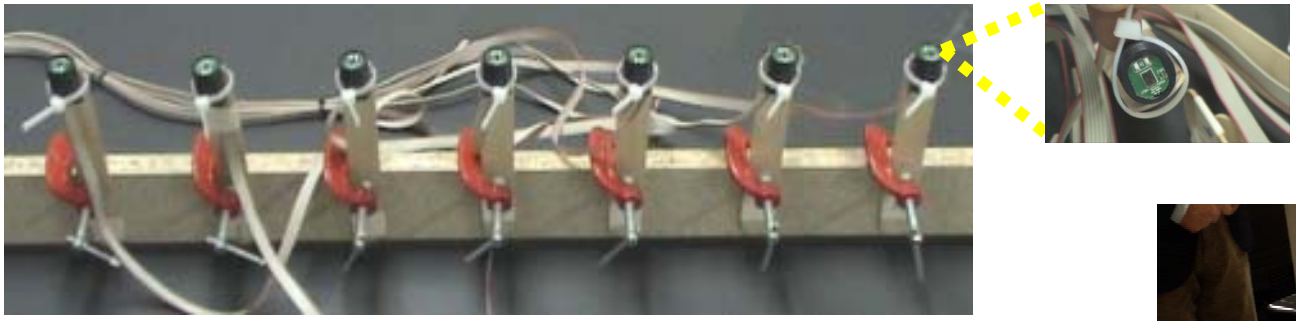
- Main technologies:
  - Voice Activity Detection
  - Acoustic Event Detection
  - Speaker Localization
  - Speech Enhancement
  - Keyword Spotting
  - Far-field command recognition

DIRHA demo (“spitaki mou”)

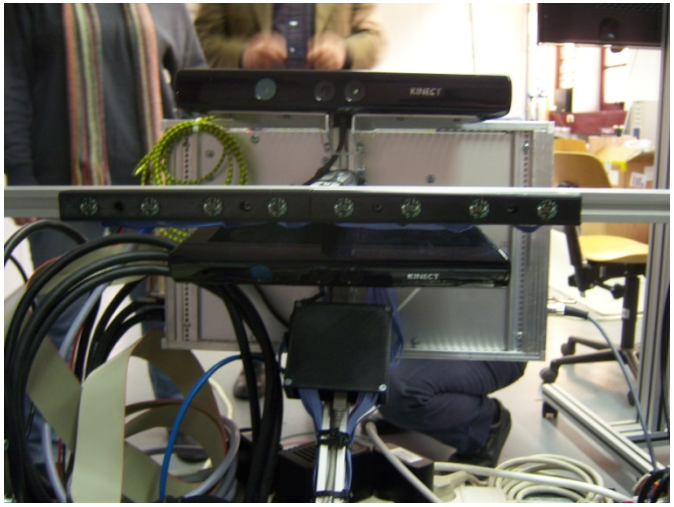
https://www.youtube.com/watch?v=zf5wSKv9wKs

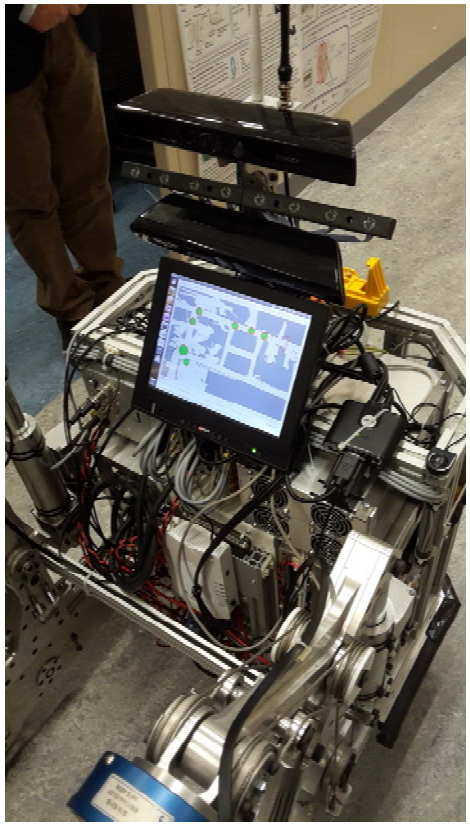
- A.Tsiami, A. Katsamanis, I. Rodomagoulakis, G. Potamianos, and P. Maragos, “Home sweet home… Listen!” , *Show & Tell Demo Presentation,* ICASSP 2016.
- I. Rodomagoulakis, A. Katsamanis, G. Potamianos, P. Giannoulis, A. Tsiami, P. Maragos, “Room-localized spoken command recognition in multi-room, multi-microphone environments”, *Computer Speech & Language*, 2017.

MEMS Audio Sensors Array



- MEMS Microphone Array
  - 8 channels
  - 48 kHz Sampling, 16bit PCM

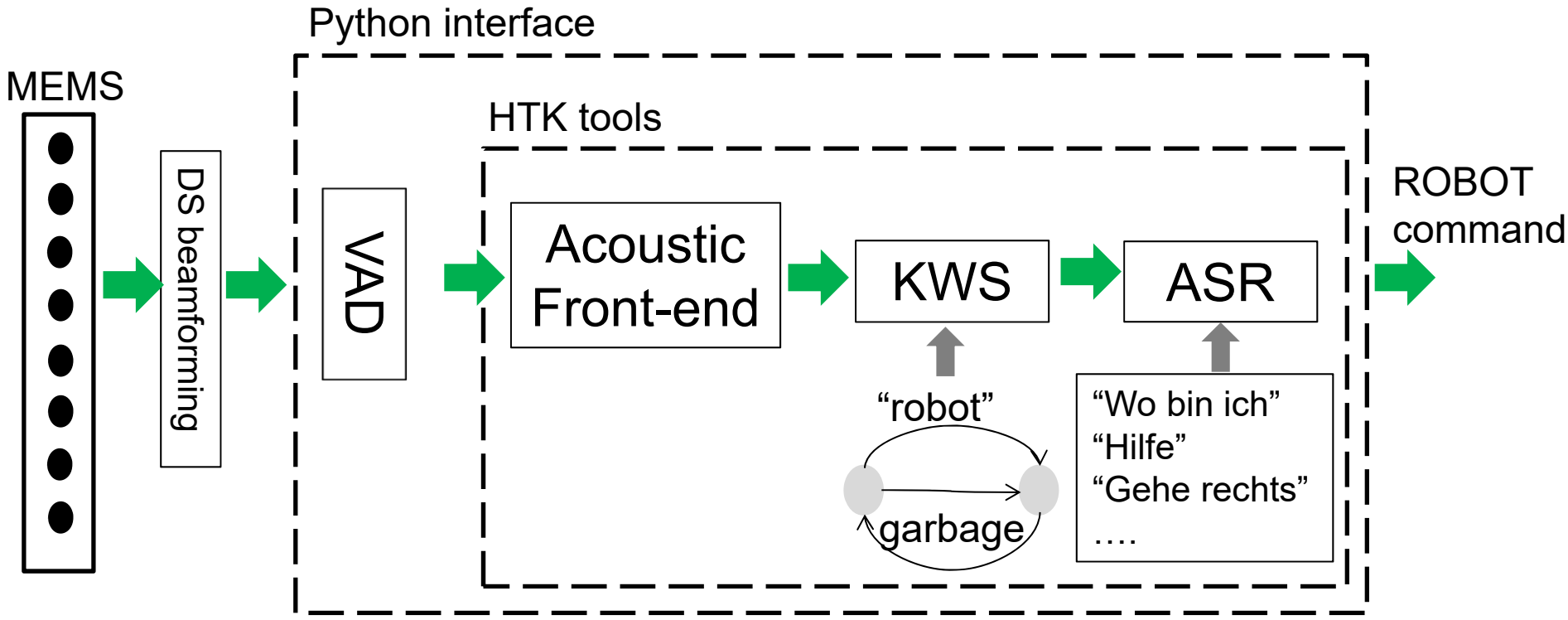




MEMS mounted on robot rollator

Spoken-Command Recognition Module for HRI

- integrated in ROS, always-listening mode, real time performance



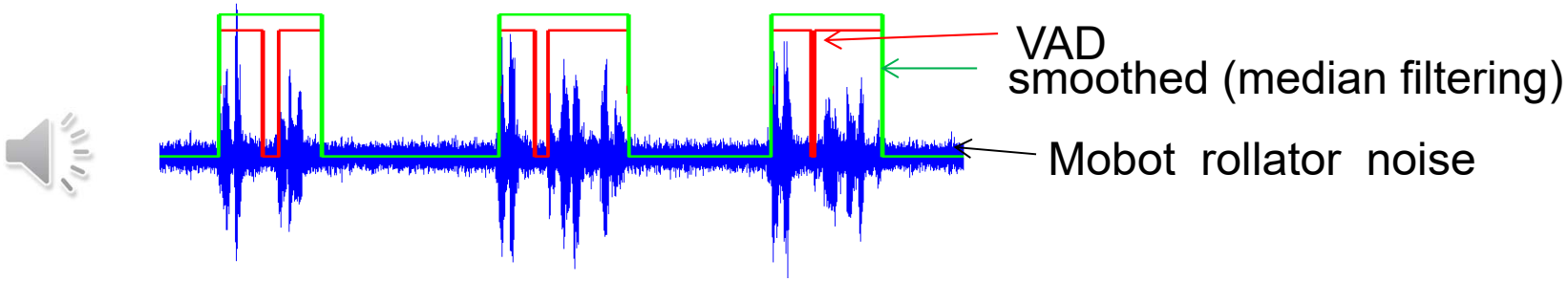
Spoken Command Recognition Results (MOBOT-I data)

- MOBOT Scenario 3.b
  - users: Elderly/patients, native Germans
  - setup: Users are sitting 2-3m in front of the platform
  - conditions: Noisy (overlapping speech, background noises)
  - data: 8 users, 19 audio-gestural commands, 3-4 repetitions per user
- classification results (ground-truth segmentation)
  - leave-one-out adaptation-testing

patients models	p1	p12	p13	p9	p4	p7	p8	p11	**avg**
CW-tri	40.41	61.07	53.00	50.83	36.51	23.26	27.24	21.24	**38.87**
+MLLR	79.79	71.14	78.80	78.75	76.19	76.26	73.42	48.67	**72.67**

+34%

Spoken Command Recognition Results (MOBOT-II data)



- users: Normal (between 20-40 years old)
- setup: Holding and moving the MOBOT platform (following mode)
- conditions: Quiet indoors environment (laboratory)
- data: 10 users, 21 audio-gestural commands, 5 repetitions
- recognition results: Leave-one-out adaptation-testing

segmentation	users										avg
	models	u1	u2	u3	u4	u5	u6	u7	u8	u9	
ground truth	CW-tri	57.37	50.79	64.74	48.94	48.94	61.64	66.84	66.58	56.05	**57.33**
	+MLLR	96.05	95.53	98.95	94.71	94.68	97.35	97.89	98.95	95.79	**96.13**
VAD		94.21	87.77	97.35	75.00	76.82	92.82	85.32	61.48	94.24	**85.13**

AUDIO-VISUAL FUSION

for

MULTIMODAL GESTURE RECOGNITION

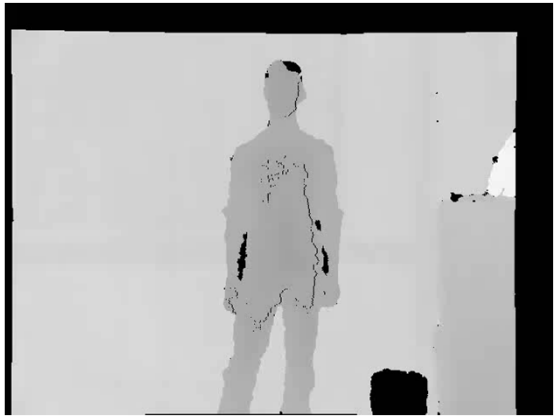
Multimodal Gesture Signals from Kinect-0 Sensor

(from CHALEARN 2013 Database: 20 Italian gesture phrases, 22 users, ~20 repetitions)

RGB Video & Audio

Depth

(vieniqui - *come here*)



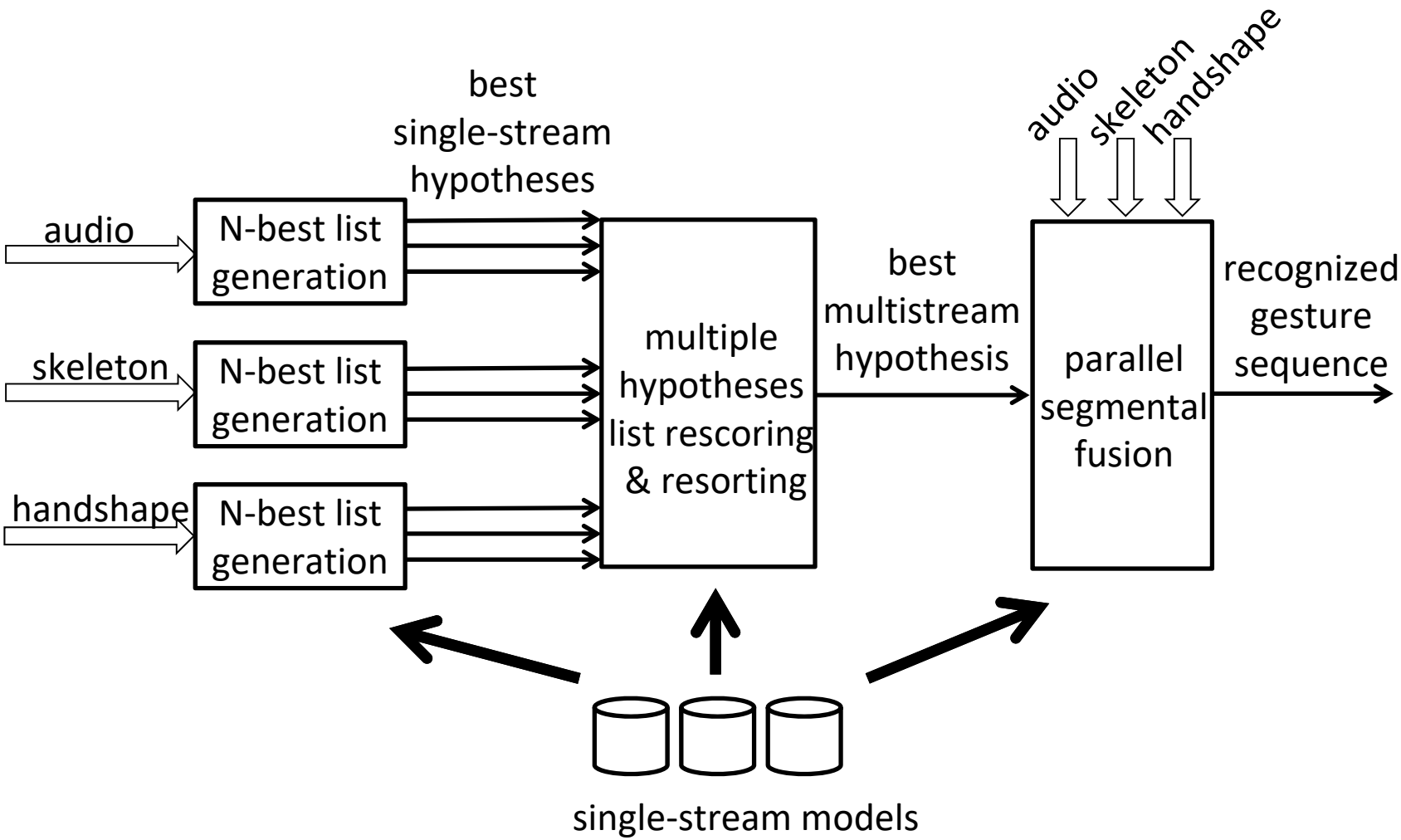
Skeleton

(vieniqui - *come here*)

User Mask

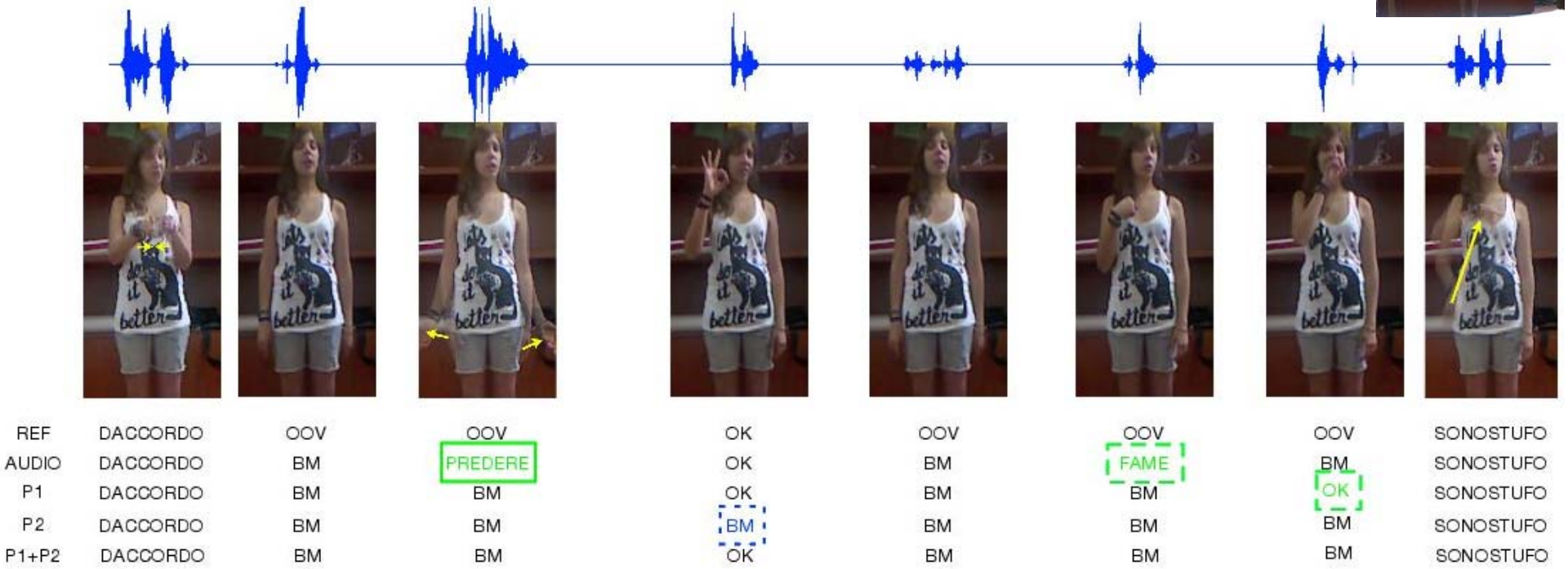
(vieniqui - *come here*)

Overview: Multimodal Hypothesis Rescoring + Segmental Parallel Fusion



[V. Pitsikalis, A. Katsamanis, S. Theodorakis & P. Maragos, JMLR 2015]

Audio-Visual Fusion & Recognition



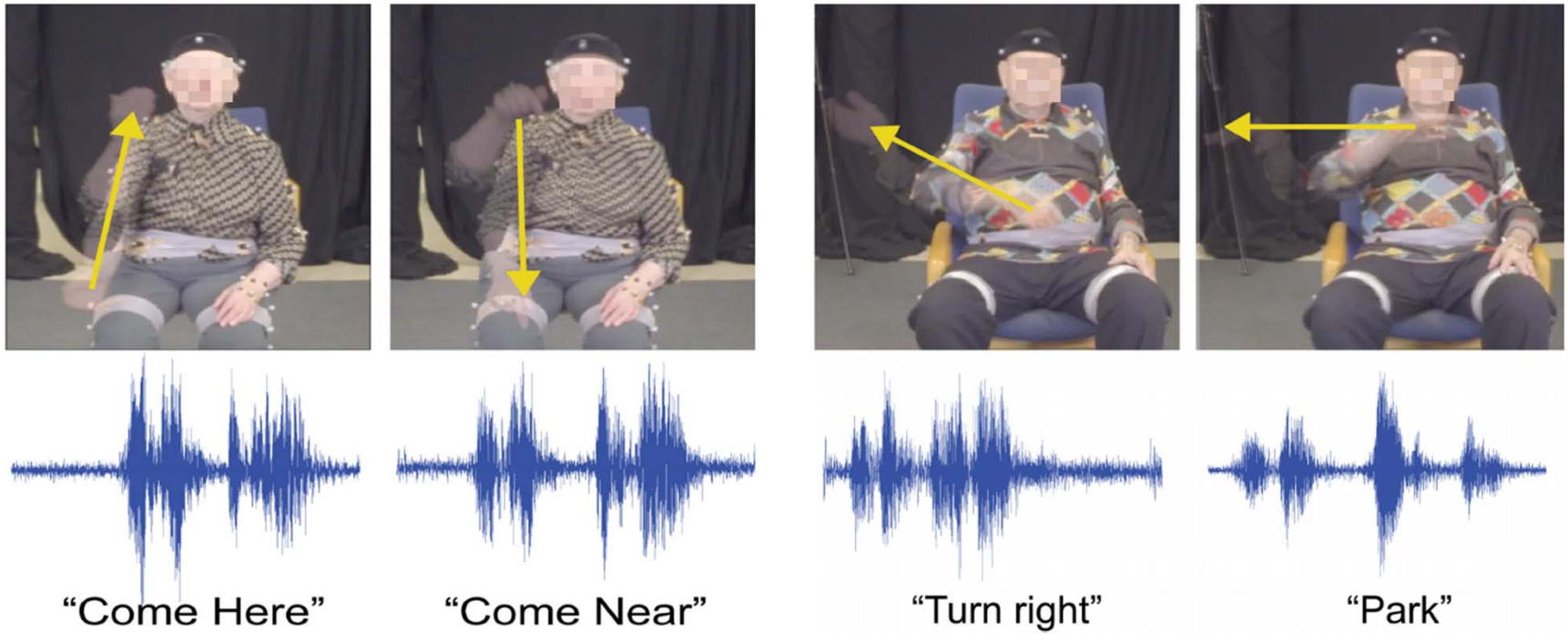
- Audio and visual modalities for A-V gesture word sequence.
- Ground truth transcriptions (“REF”) and decoding results for audio and 3 different fusion schemes.
- Achieved top performance (93.3%) in gesture challenge CHALEARN (ACM ICMI 2013).

[V. Pitsikalis, A. Katsamanis, S. Theodorakis & P. Maragos, JMLR 2015]

Multimodal fusion: Complementarity of visual and audio modalities

Similar audio,
distinguishable gesture

Distinguishable audio,
similar gesture

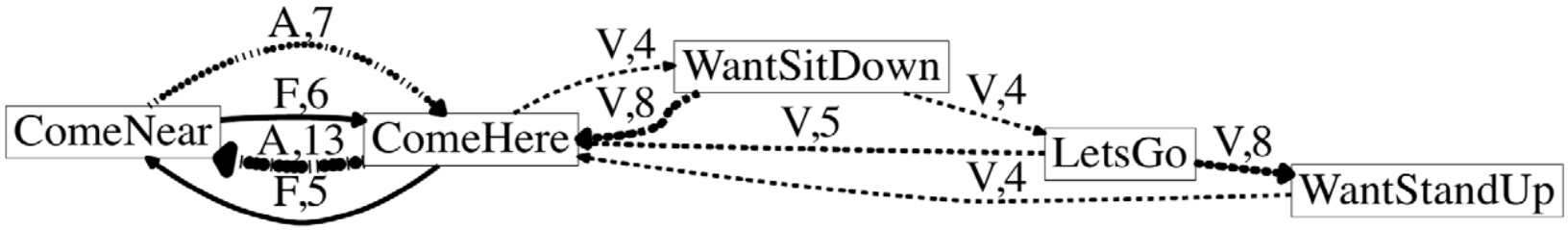


Multimodal gesture classification results

- Leave-one-out experiments (Mobot-I.6a data: 8p,8g)
- Unimodal: audio (A) and visual (V)
- Multimodal (AV): N-best list rescoring

	p1	p11	p12	p13	p4	p7	p8	p9	avg
A	96.87	81.81	79.16	87.50	78.26	79.16	90.32	80.00	84.13
V	40.62	63.63	70.83	34.37	65.21	50.00	64.51	66.66	56.98
AV	87.50	90.90	95.83	84.37	100.0	79.16	96.77	86.66	**90.15**

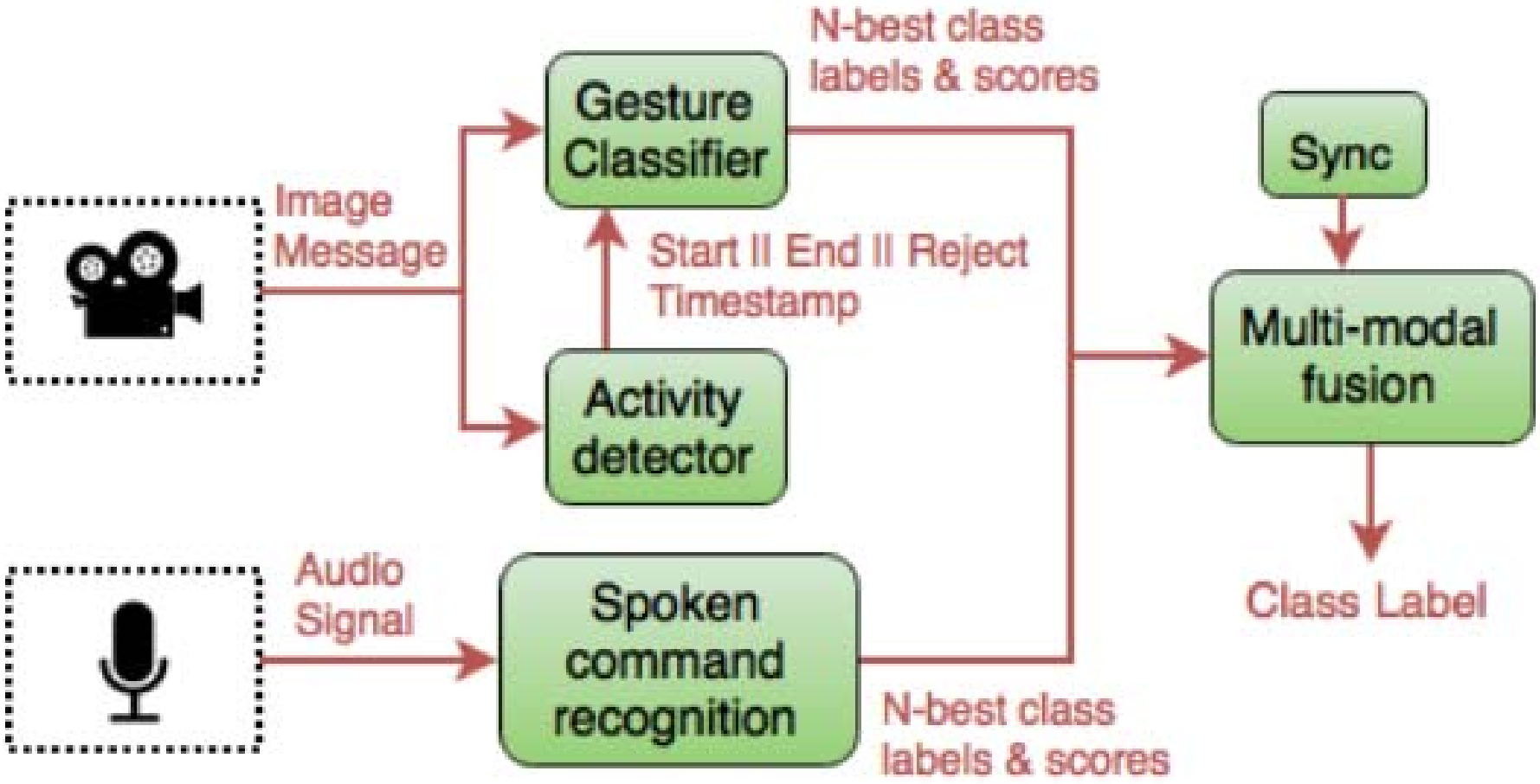
Multimodal confusability graph



[I.Rodomagoulakis, N.Kardaris, V.Pitsikalis, E.Mavroudi, A.Katsamanis, A.Tsiami & P.Maragos, *Multimodal Human Action Recognition in Assistive Human-Robot Interaction*, ICASSP 2016.]

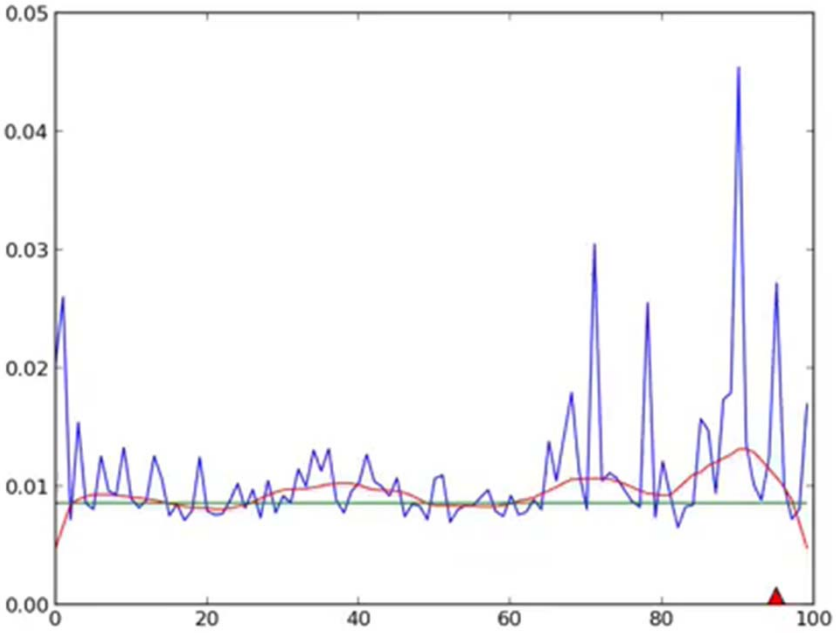
HRI Online Multimodal System Architecture

- ROS based integration
  - Spoken command recognition node
  - Activity detection node
  - Gesture classifier node
  - Multimodal fusion node
- Communication using ROS messages



Online processing details

Activity / Non activity

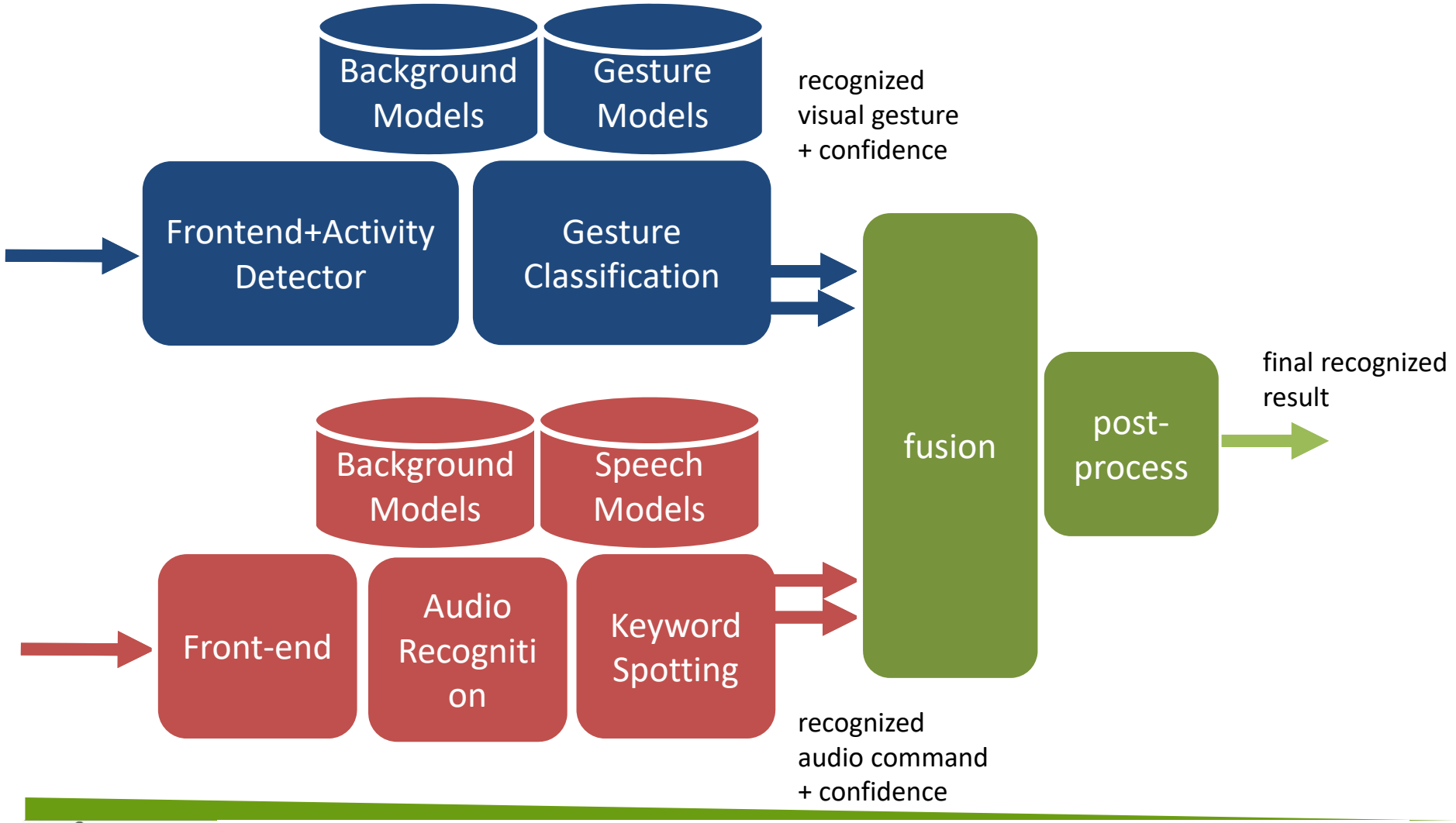


- Online processing; “Always-listening” + Activity detection

Audio-Visual gesture recognition

Online processing system – Open Source Software

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

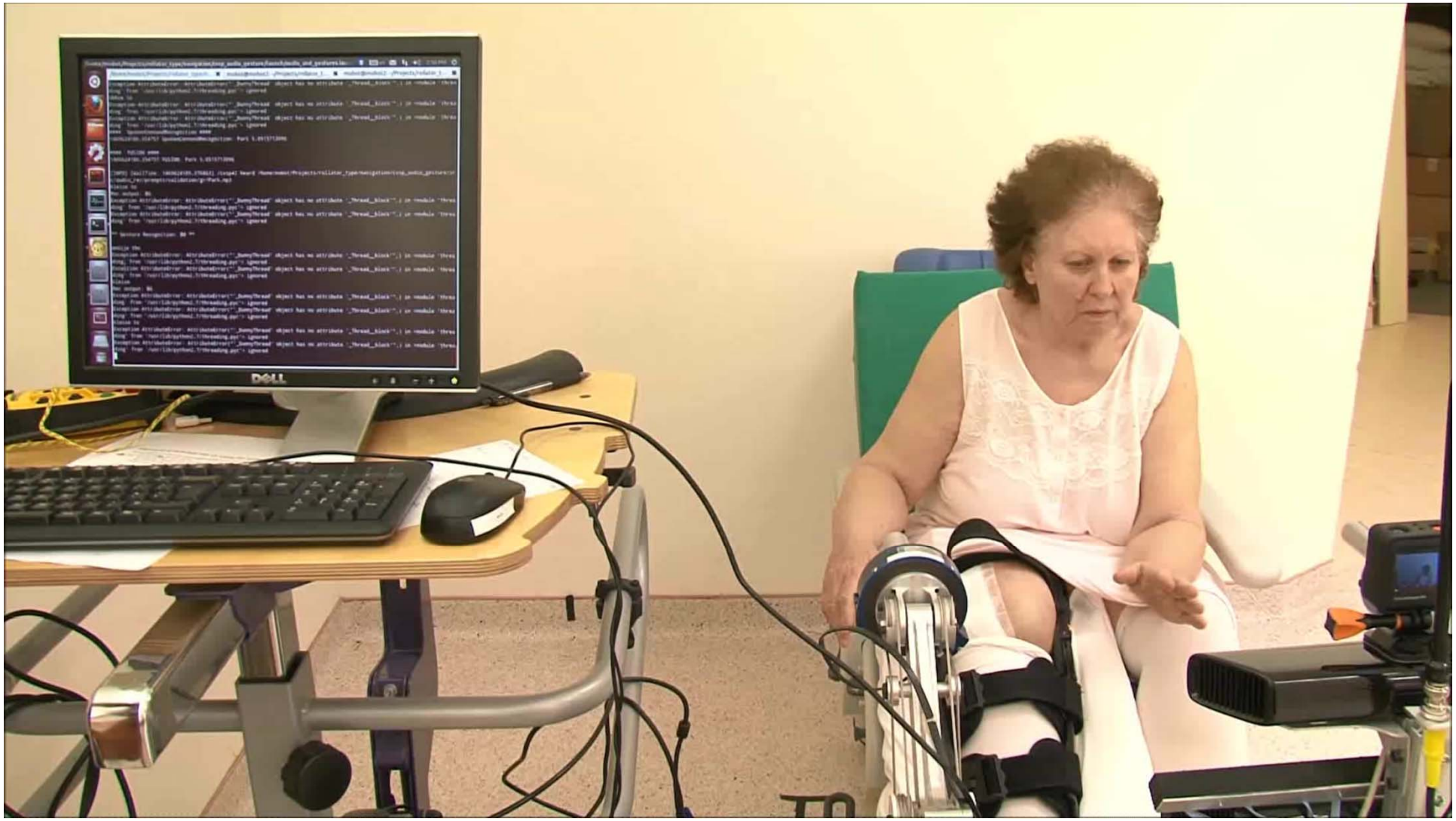


N. Kardaris, I. Rodomagoulakis, V. Pitsikalis, A. Arvanitakis and P. Maragos, *A platform for building new human-computer interface systems that support online automatic recognition of audio-gestural commands*, Proc. ACM Multimedia 2016, Amsterdam, The Netherlands, 15-19 Oct. 2016.

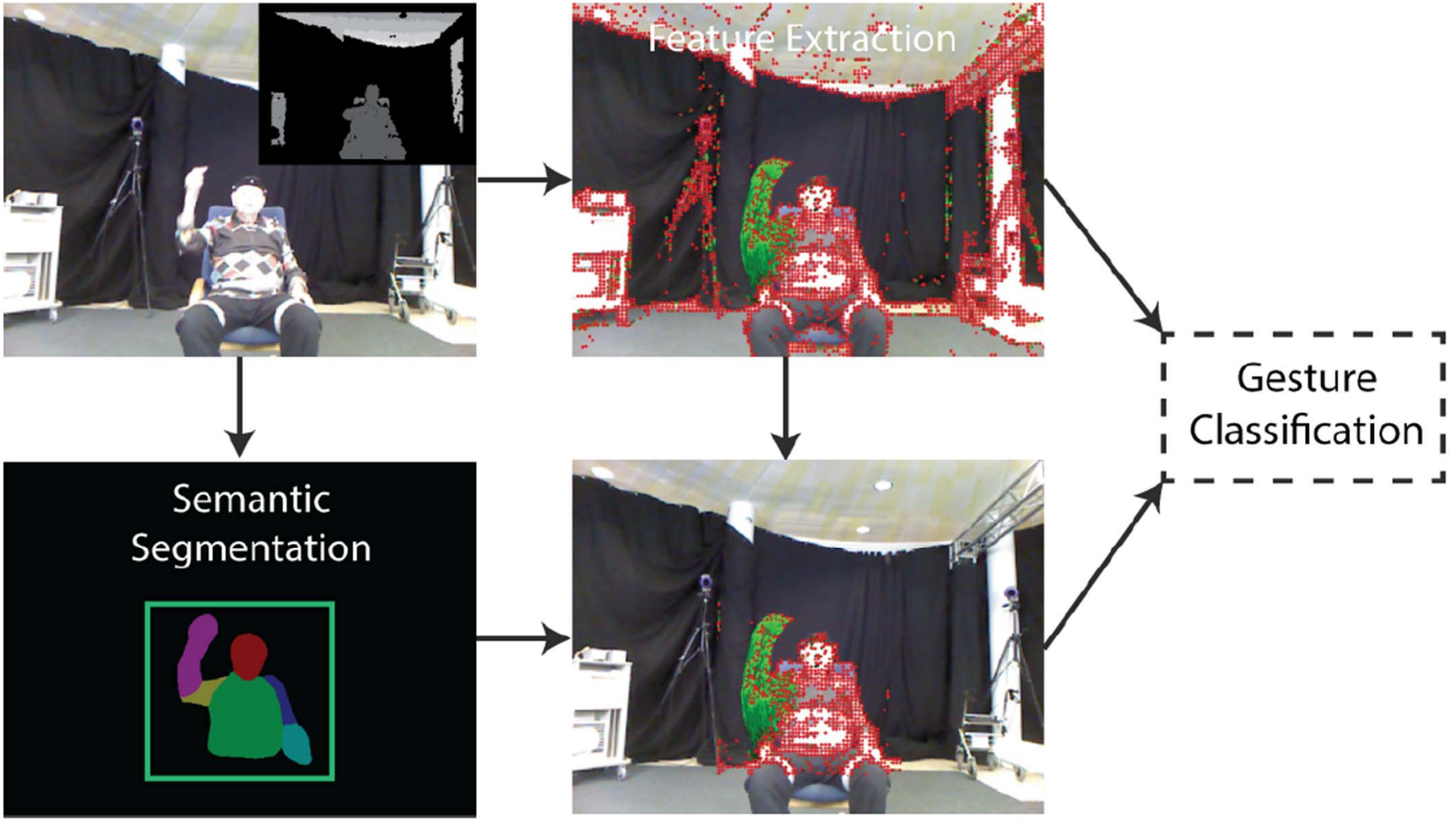
Validation experiments (Bethanien, Heidelberg):

Audio-Gestural recognition in action (1/2)

Validation experiments (Diaplasis, Kalamata)



Visual Synergy: : SS+GR



foreground/background+gesture recognition

Visual Synergy: SS+GR

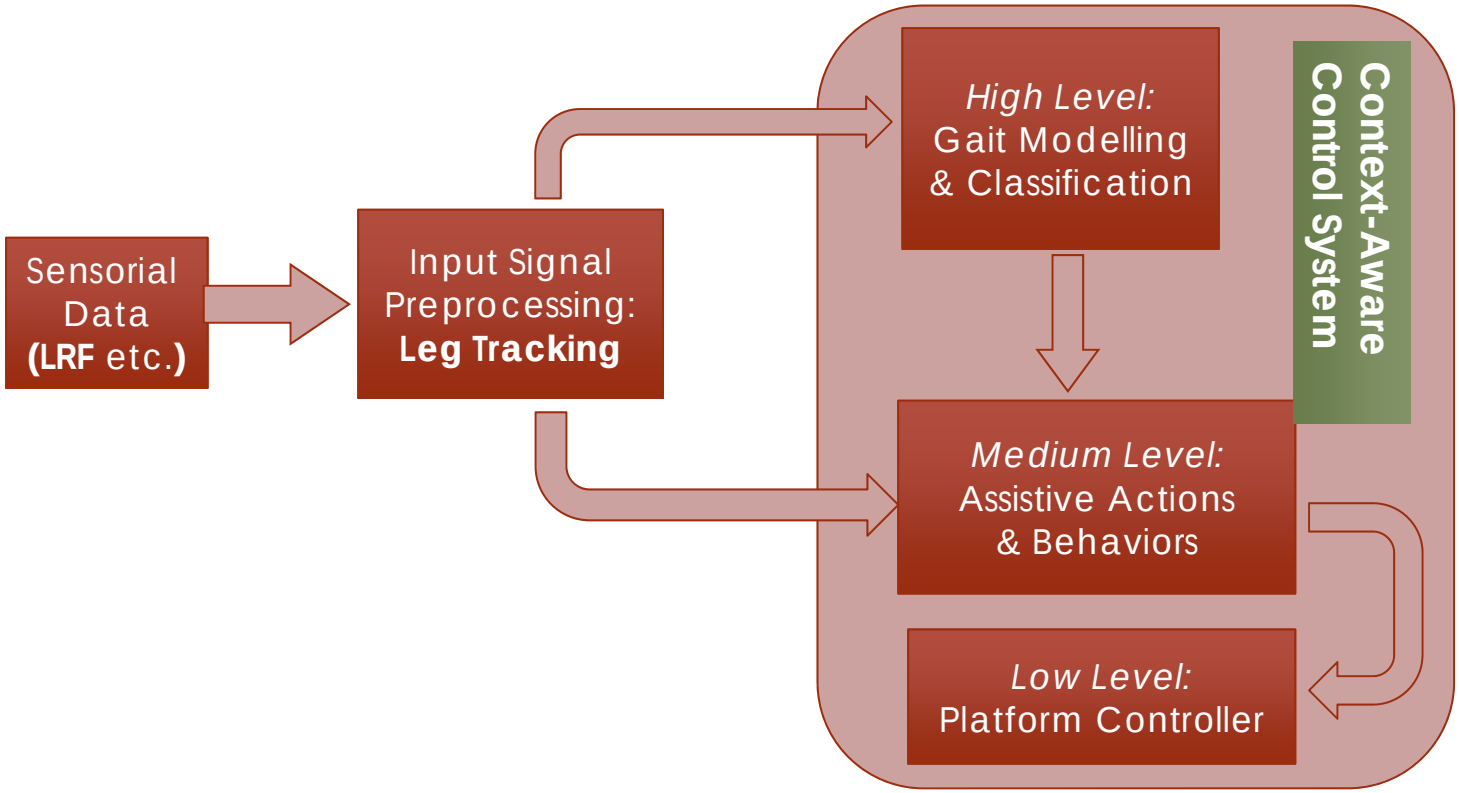
Feat. Descr.	GR	SS+GR	Impr. (%)
traject.	38.8	42.1	8.59
HOG	46.0	46.7	1.47
HOF	51.5	56.3	9.33
MBH	57.0	63.9	12.18
combined	62.4	65.6	5.22

Average classification accuracy (%) over all 8 patients using our baseline method (first column) and employing the foreground-background mask (second column). Results show a consistent improvement (third column) over multiple feature descriptors. Results are obtained using the BoVW encoding.

Ref:

A. Guler, N. Kardaris, S. Chandra, V. Pitsikalis, C. Werner, K. Hauer, C. Tzafestas, P. Maragos and I. Kokkinos, "*Human Joint Angle Estimation and Gesture Recognition for Assistive Robotic Vision*" in ECCV workshop on Assistive Computer Vision and Robotics (ACVR-16), Oct. 2016.

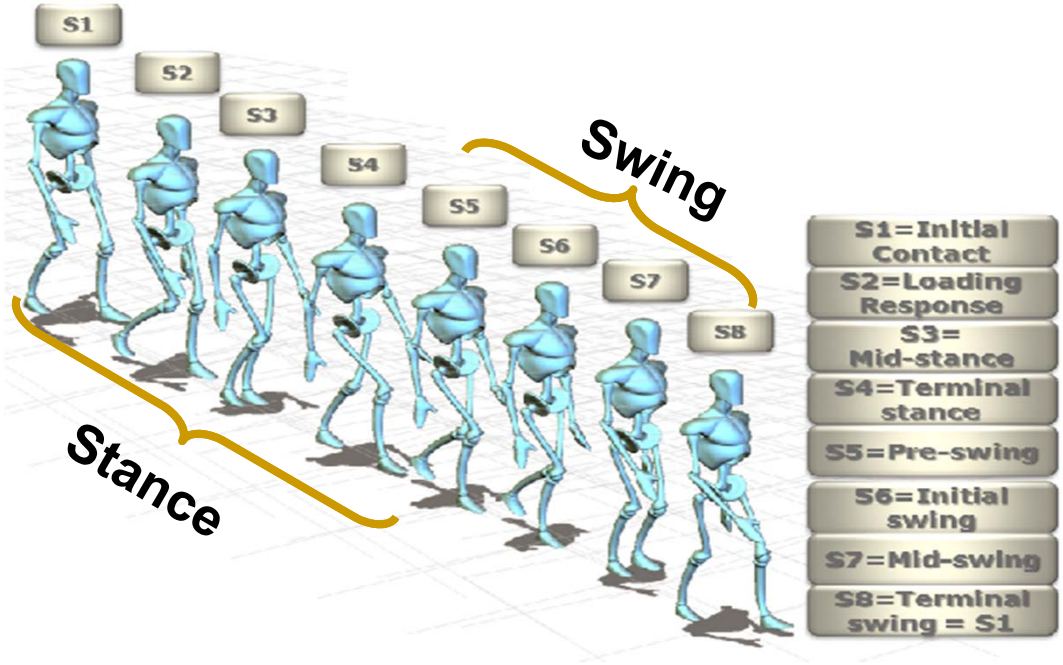
GAIT ANALYSIS



Human Gait Cycle

Two main phases in the gait cycle

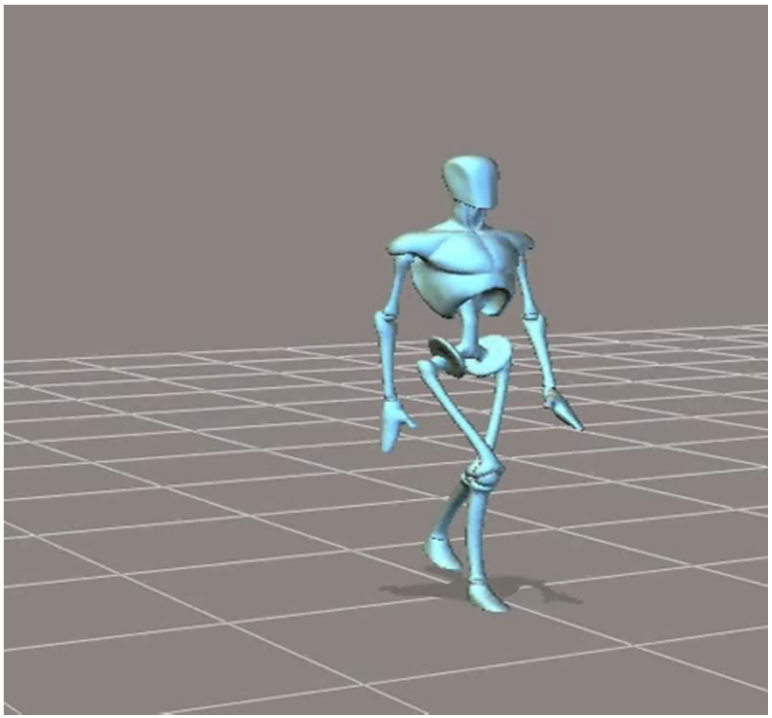
- Stance phase: the foot is on the ground
- Swing phase: the same foot is no longer in contact with the ground and is swinging through in preparation for the next foot strike



Human Gait Cycle

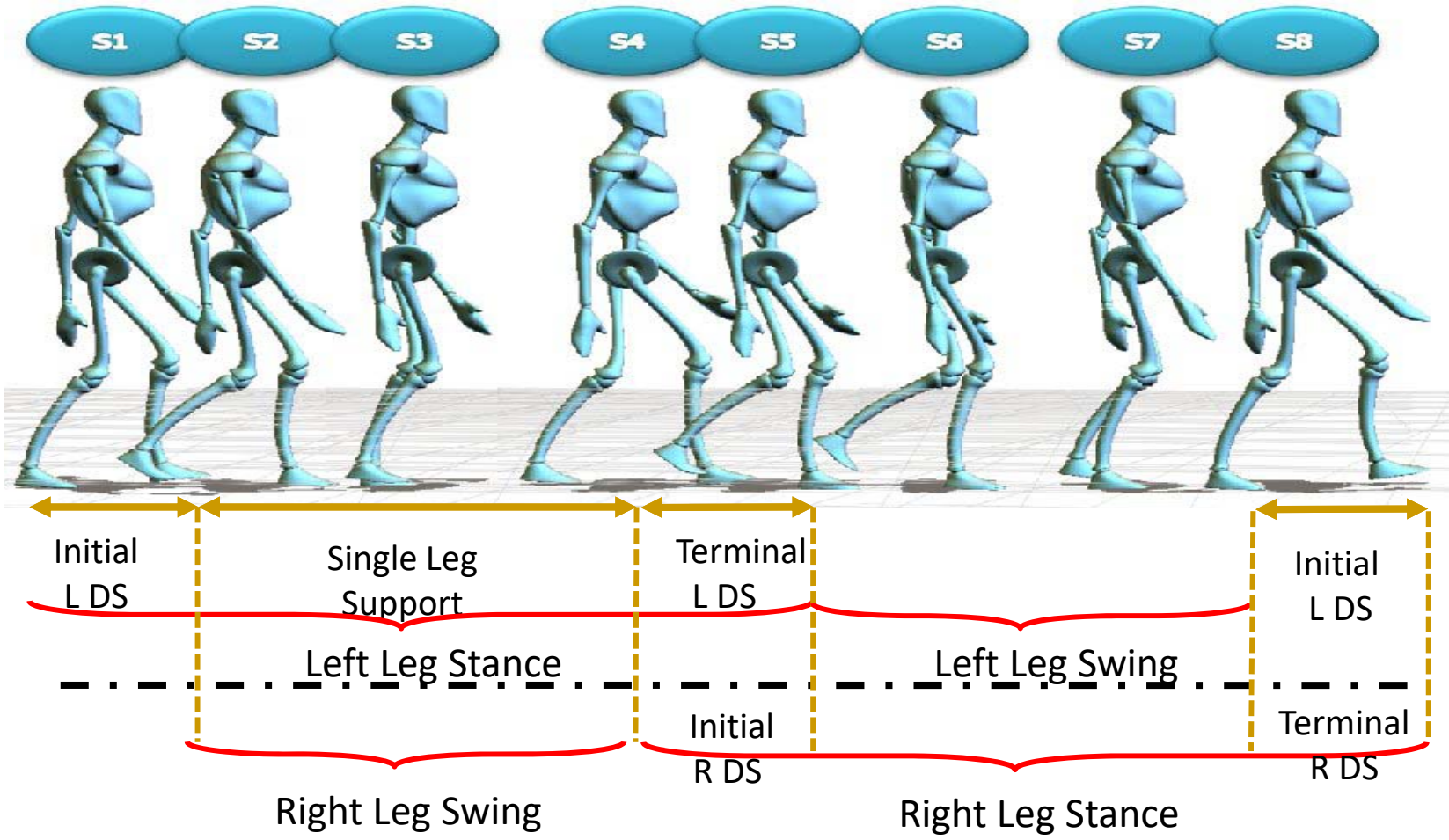
Comprises eight (8) events, which can generally describe any gait:

1. Initial contact (0%) - IC, (**Heel Strike:** initiates the gait cycle)
2. Loading response (0-10%) - LR, (**Foot Flat:** the plantar surface of the foot touches the ground)
3. Mid-stance (10-30%) - MS,
4. Terminal stance (30-50%) - TS, (**Heel Off:** the heel loses contact with the ground)
5. Pre-swing (50-60%) - PW, (**Toe Off:** the foot leaves the ground)
6. Initial Swing (60-70%) - IW,
7. Mid-swing (70-85%) - MW,
8. Terminal swing (85-100%) – TW



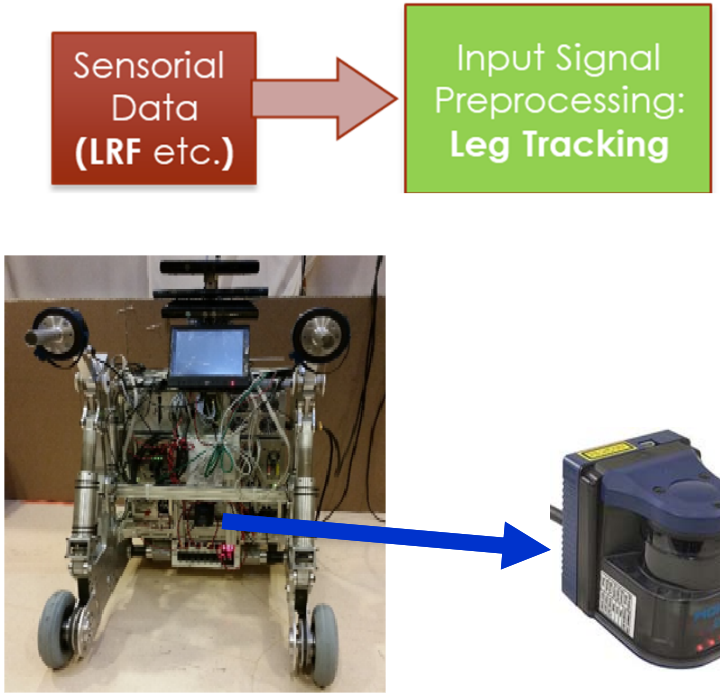
Objective: Develop a computational model to capture the stochastic, dynamic nature of human g.c. and be used for on-line characterization & estimation purposes

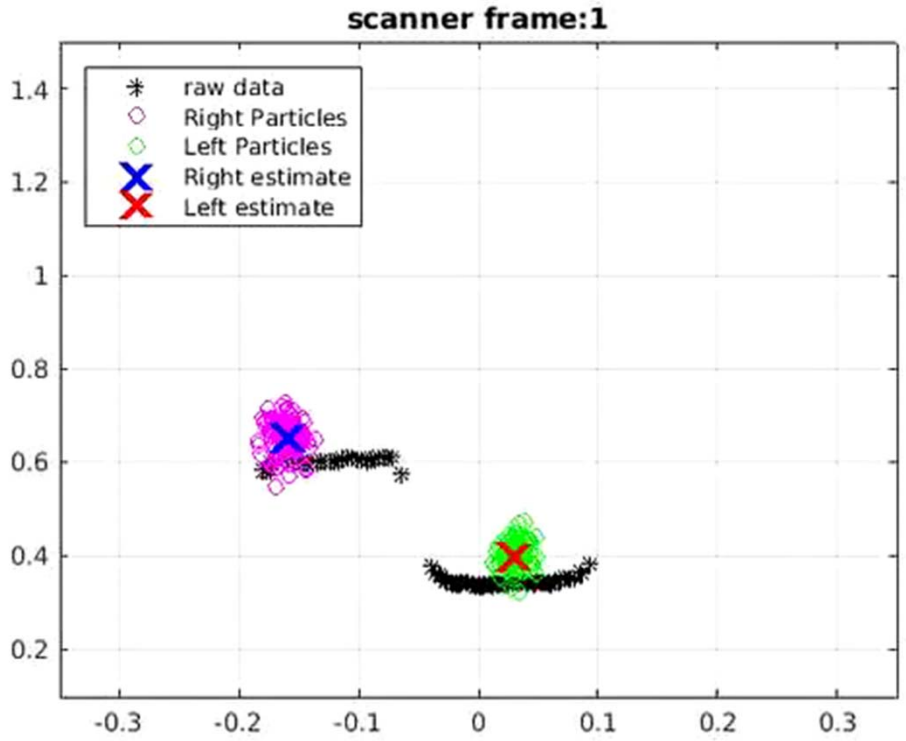
Human Gait Cycle



Leg Tracking: Particle Filters with Probabilistic Data Association

- Probabilistic Data Association Particle Filtering (PDA-PF) system Tracks the user legs, [1]
- PDA-PF sequentially estimates the relative position and velocity of the patients legs w.r.t. the robotic rollator

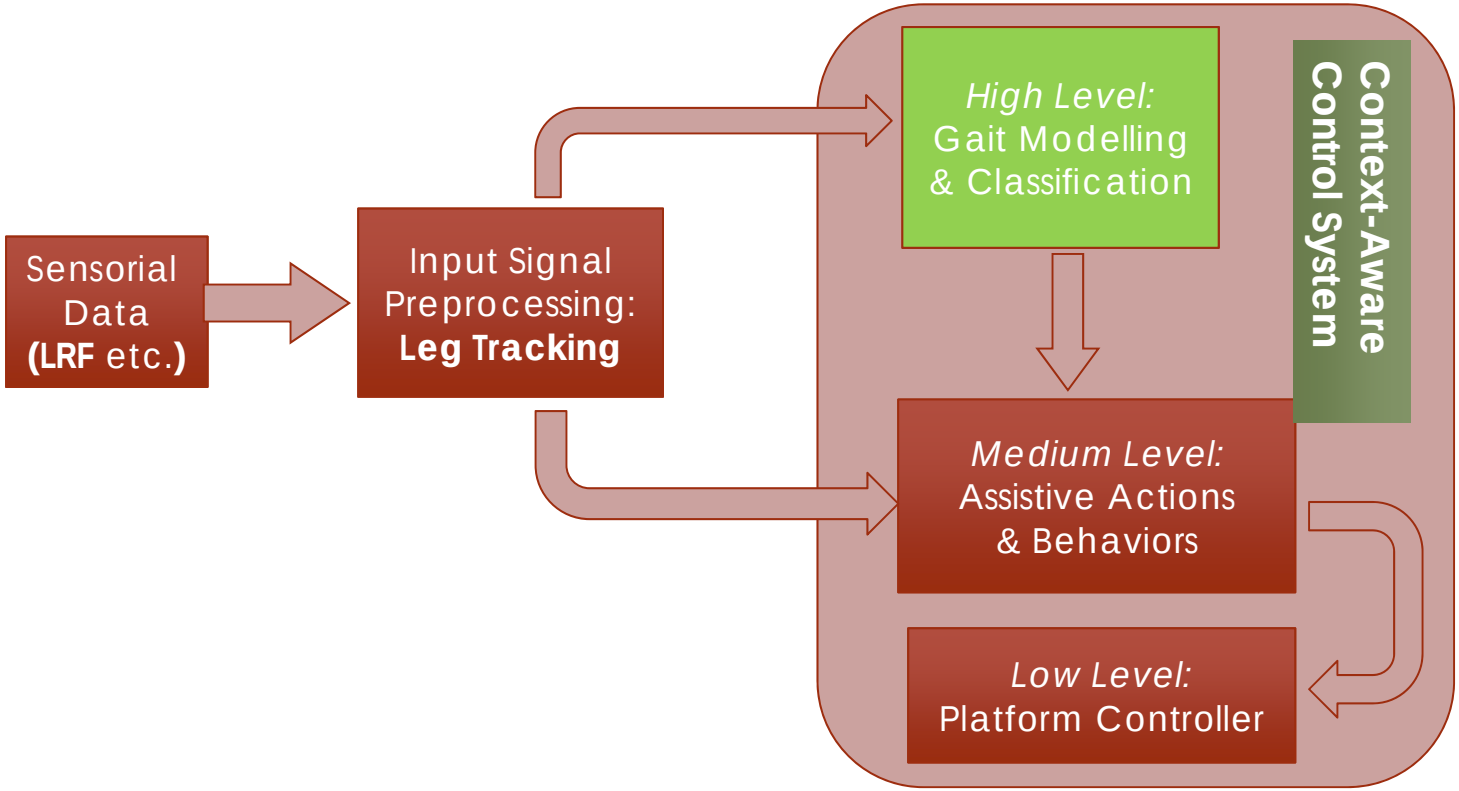




[1] G. Chalvatzaki, X. Papageorgiou, C. Tzafestas and P. Maragos,, "Comparative experimental validation of human gait tracking algorithms for an intelligent robotic rollator," in ICRA 2017.

HMMs-based Gait Phases Recognition System

Posterior estimates of the legs' states are fed into two HMMs that recognize the left and right gait cycles respectively, [2].



[2] G. Chalvatzaki, X. Papageorgiou, C. Tzafestas and P. Maragos, "Estimating double support in pathological gaits using an HMM-based analyzer for an intelligent robotic walker," in RO-MAN 2017.

Gait parameters estimation based on HMMs

Use and **validation** of HMM-based gait analysis for **reliable estimation** of clinically-relevant **gait parameters**

- HMM-based **recognition of gait phases** transitions
- Computation of **gait parameters based on this on-line** gait segmentation
- Application areas: medical diagnosis, rehabilitation progress

Temporal Gait parameters:

- **Stride Time** (the duration of each gait cycle)
- **Stance Time** (the stance phase duration in one cycle)
- **Double Support Time** (the time period when both feet are in contact with the ground)

Spatial Gait parameters

- **Stride Length** (the distance trevelled by both legs in the duration of one stride)
- **Gait Speed** (the ratio of stride length to stride time)

Validation studies

Two experimental setups from the MOBOT project:

- Validation using ground truth data from a Motion Capture System, [2].
- Validation using ground truth from a GAITRite System, [3,4].

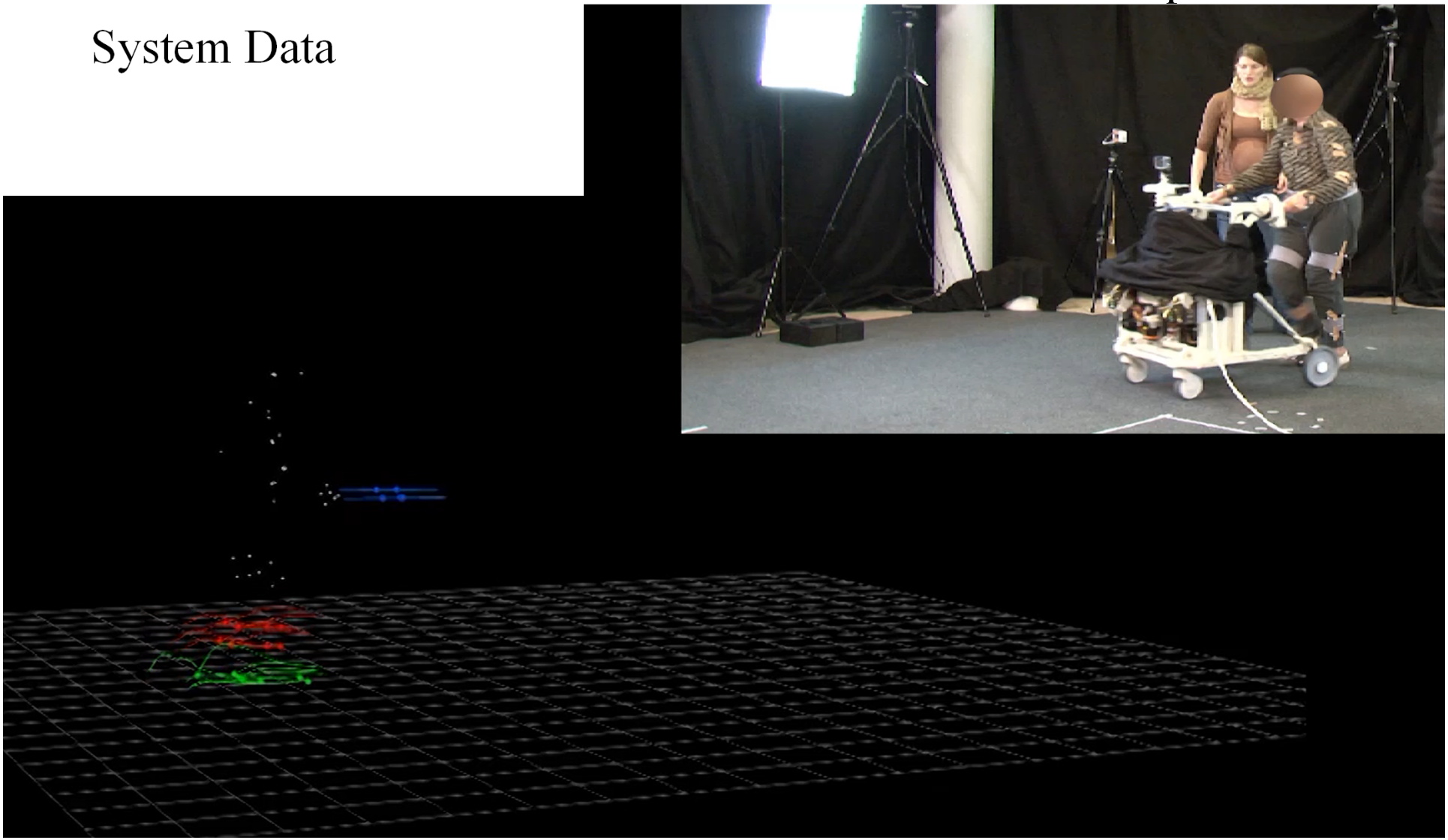
[2] G. Chalvatzaki, X. Papageorgiou, C. Tzafestas and P. Maragos, “Estimating double support in pathological gaits using an HMM-based analyzer for an intelligent robotic walker,” in RO-MAN 2017.
[3] X. Papageorgiou, G. Chalvatzaki, K. Lianos, C. Werner, K. Hauer, C. Tzafestas, P. Maragos,, “Experimental validation of human pathological gait analysis for an assisted living intelligent robotic walker,” in Biorob 2016.
[4] G. Chalvatzaki, X. Papageorgiou, C. Tzafestas and P. Maragos, ‘’HMM-based Pathological Gait Analyzer for a User-Adaptive Intelligent Robotic Walker, in EUSIPCO- Workshop: "MultiLearn 2017.

1st Validation study: Experimental Protocol

- Data were collected using a **HOKUYO rapid laser sensor** (mean sampling frequency of 512 planar points every 28ms) mounted on the robotic rollator
- The subjects walked with **physical support of the rollator** on a straight direction of about 3 m, performed a 180º turn and returned to initial position
- Under appropriate **carer's supervision**
- The patients were wearing their **normal cloths** (no special clothing) and were instructed to walk as normally as possible

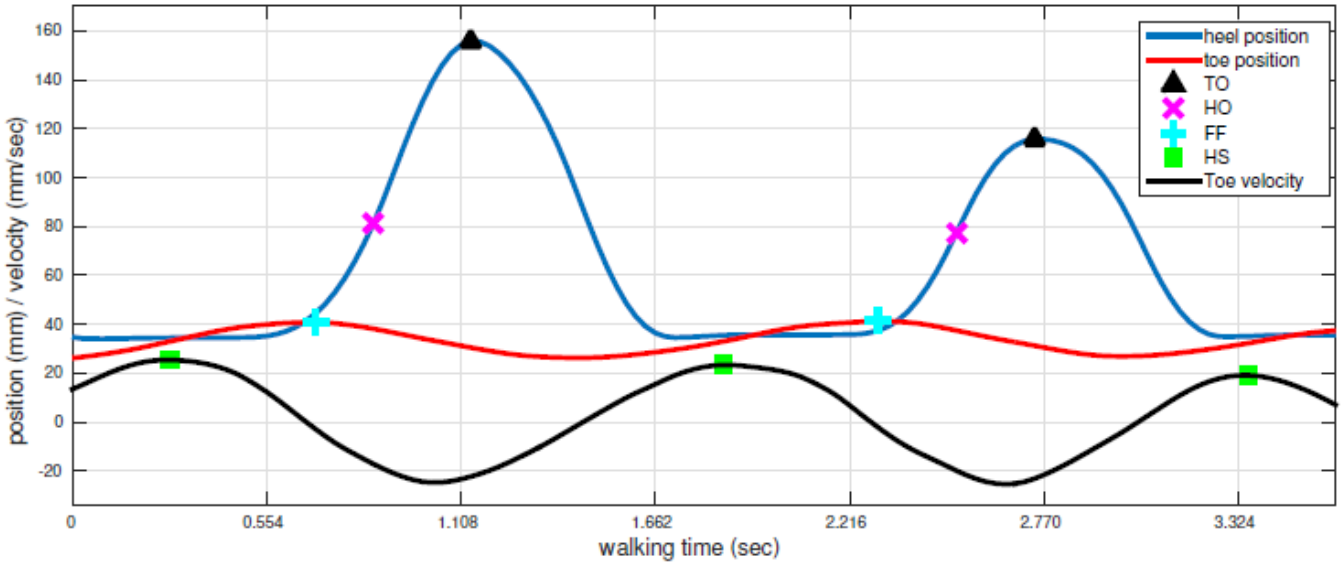
1st Validation study

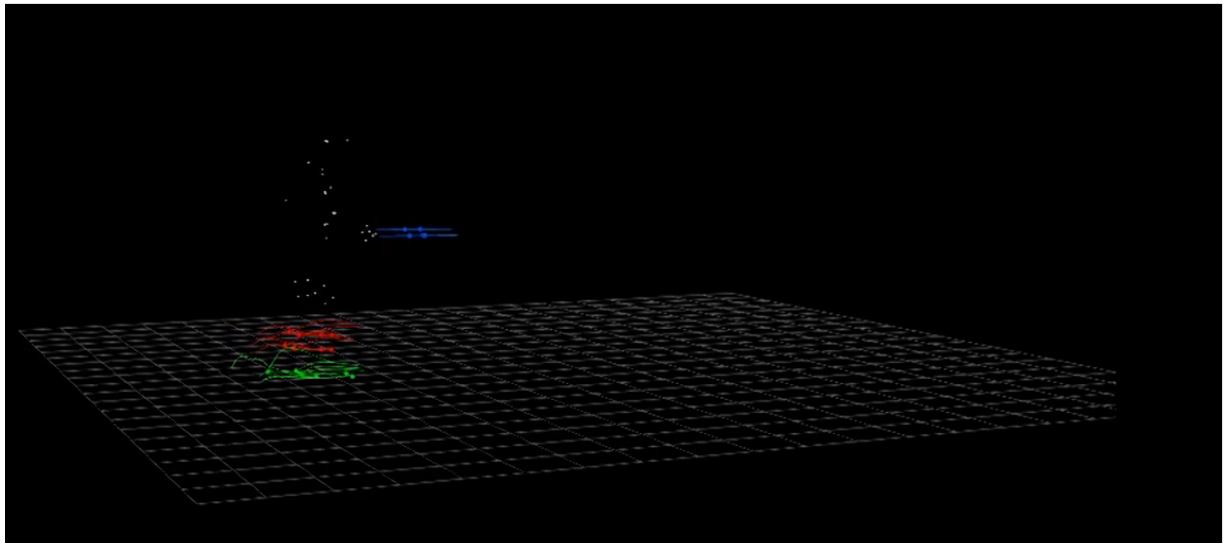
Ground Truth: Gait Phases Detection from Motion Capture System Data



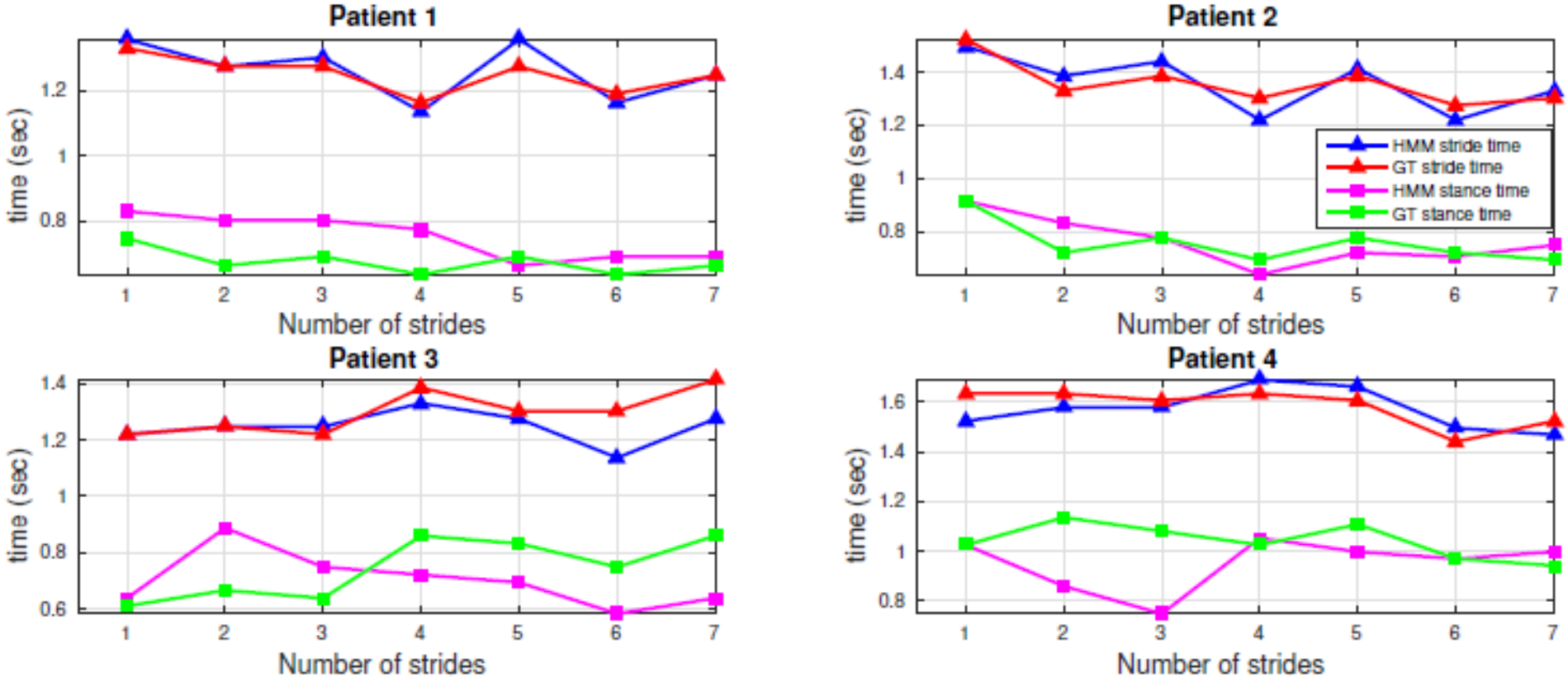
1st Validation study

Gait Phases Detection: Motion Capture Data



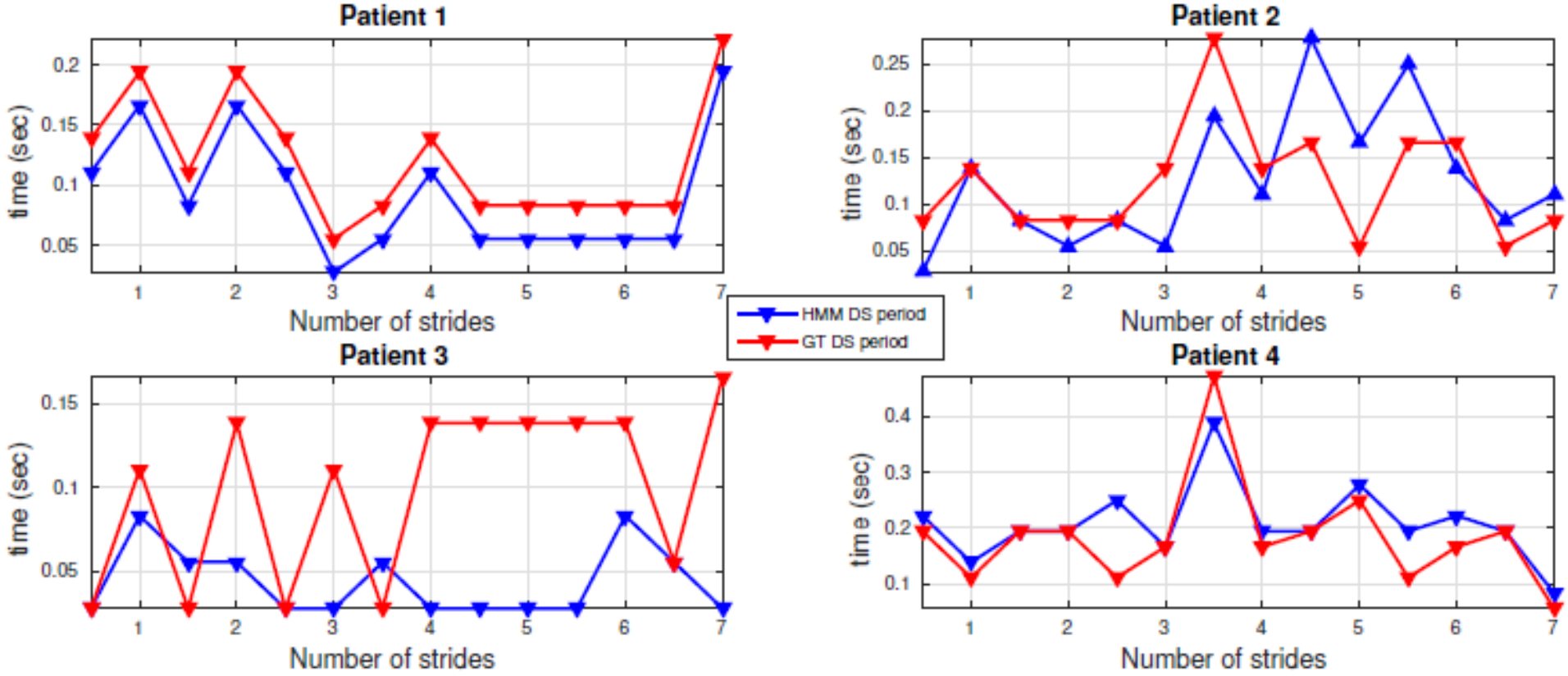


1st Validation Study: Results



Evolution of the estimated **stride time and stance time** over the stride number obtained by the proposed the HMM-based approach and the respective stride and stance times provided by the ground truth data for the patients of the study

1st Validation Study: Results



Evolution over the stride number of the **Double Support (DS) duration** estimates obtained by the proposed HMM approach compared to ground truth data

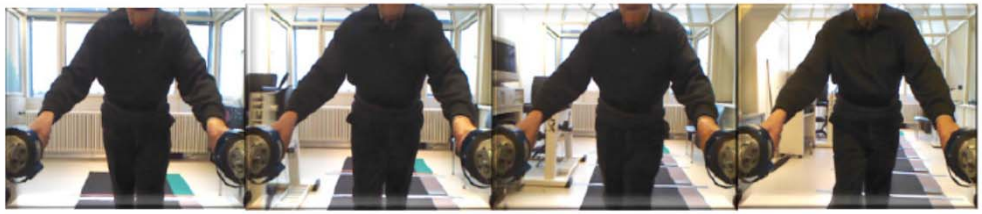
2nd Validation study: Setup

POMA score: Performance Oriented Mobility Assessment

TABLE I: Demographics

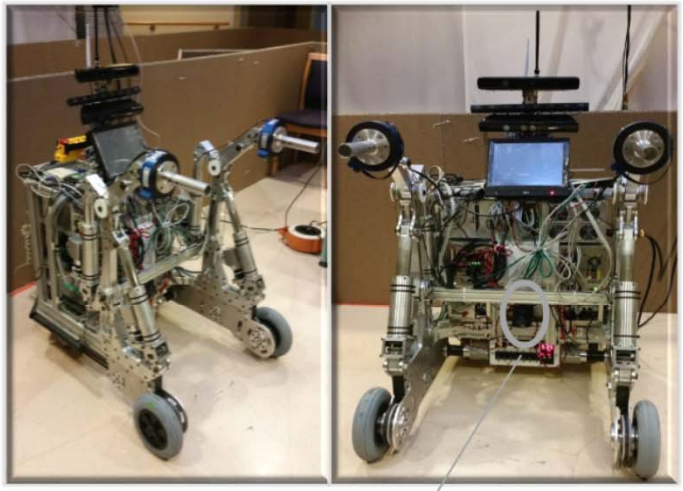
Subject	1	2	3	4	5	6
Age	89	83	83	71	82	82
Sex	F	F	F	F	F	M
POMA	7	11	19	20	26	27
Falls	yes	yes	yes	yes	yes	yes

Demographics for the subjects that participated in the experiments.



Snapshots of a subject walking on the GAITRite walkway assisted by the robotic platform, during one stride.

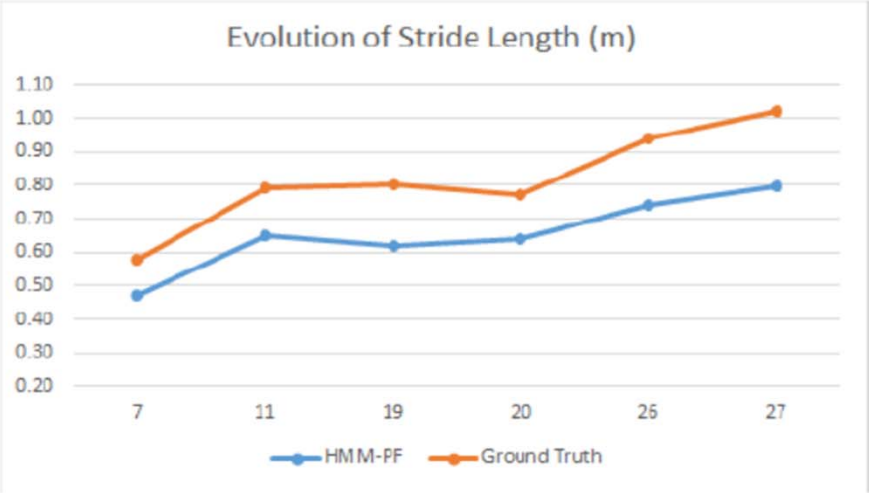
The captured footprints of the subject by the GAITRite System.



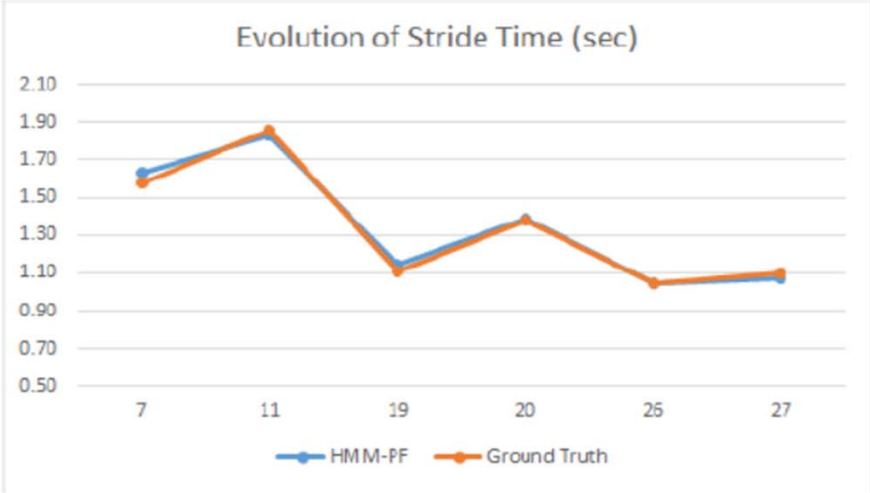
The GAITRite System mat.

2nd Validation study: Results

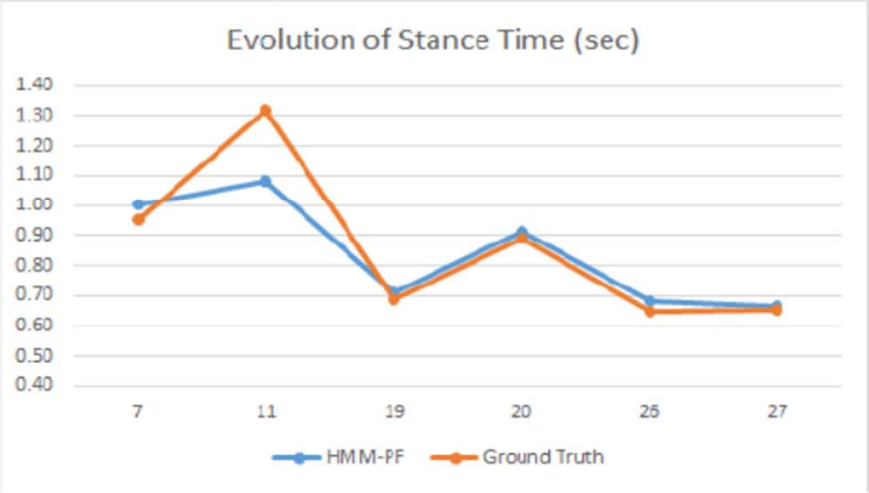
Gait Parameters Extraction



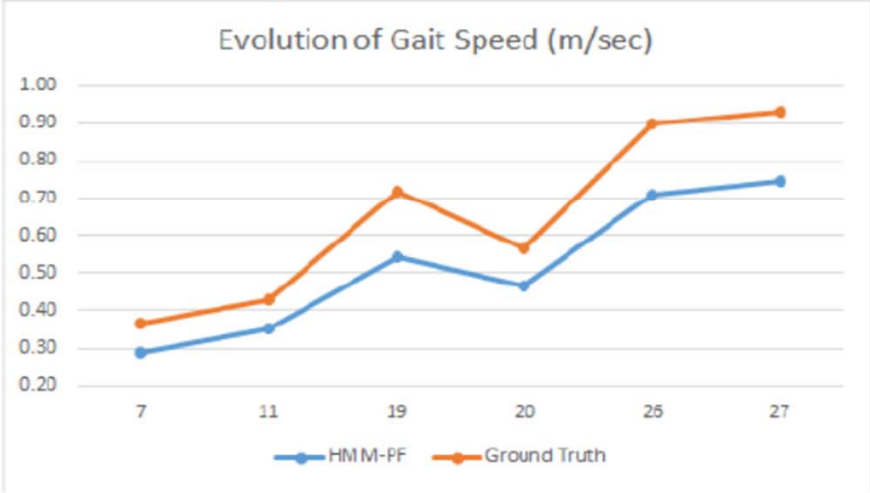
(a) Evolution of the average stride length as it was estimated by the HMM-based system w.r.t. Ground truth data according to the POMA score of the patients.



(b) Evolution of the average stride time as it was estimated by the HMM-based system w.r.t. Ground truth data according to the POMA score of the patients.



(c) Evolution of the average stance time as it was estimated by the HMM-based system w.r.t. Ground truth data according to the POMA score of the patients.

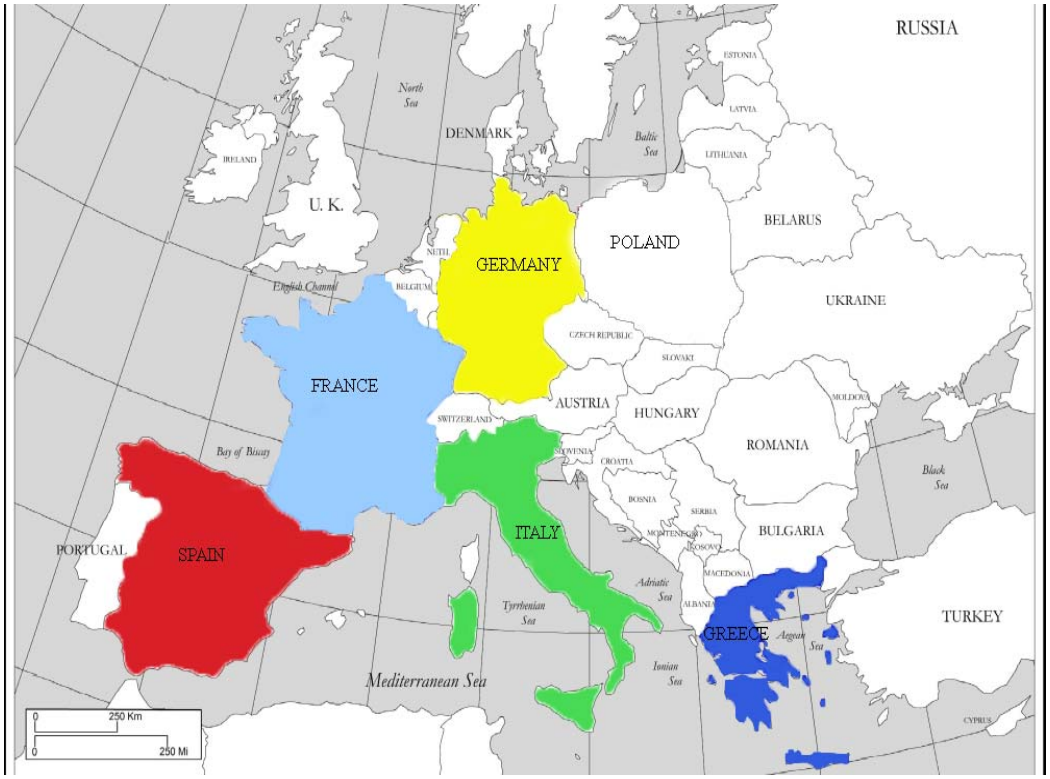


(d) Evolution of the average gait speed as it was estimated by the HMM-based system w.r.t. Ground truth data according to the POMA score of the patients.

EU Project: I-SUPPORT

ICT-Supported Bath Robots

H2020-PHC-2014-single-stage, Grant # 643666, Duration: 2015-2018 , website: www.i-support-project.eu/

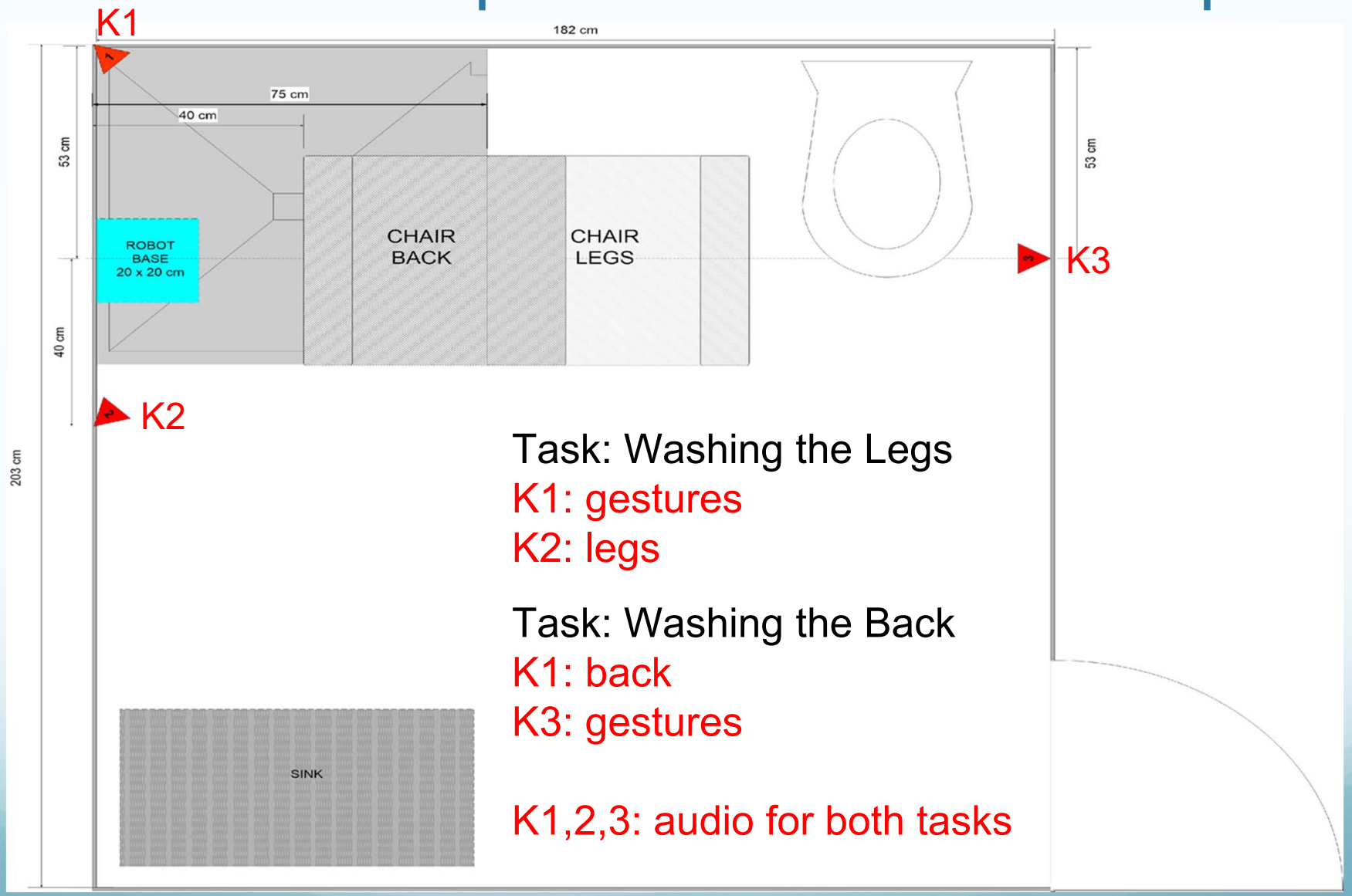


Our Team's Contribution

- Action/Gesture recognition
- Spoken Command Recognition
- Co-development of A-V data corpus
- Robotics control

Partners: ICCS-NTUA (Greece), Robotnik (Spain), SSSA(Italy), KIT (Germany), OMEGATECH (Greece), INRIA (France), Fondazione Santa Lucia (Italy), Bethanien (Germany), University of Applied Sciences (Germany)

Refined Setup of the 3 Kinects - Top



Task: Washing the Legs

K1: gestures

K2: legs

Task: Washing the Back

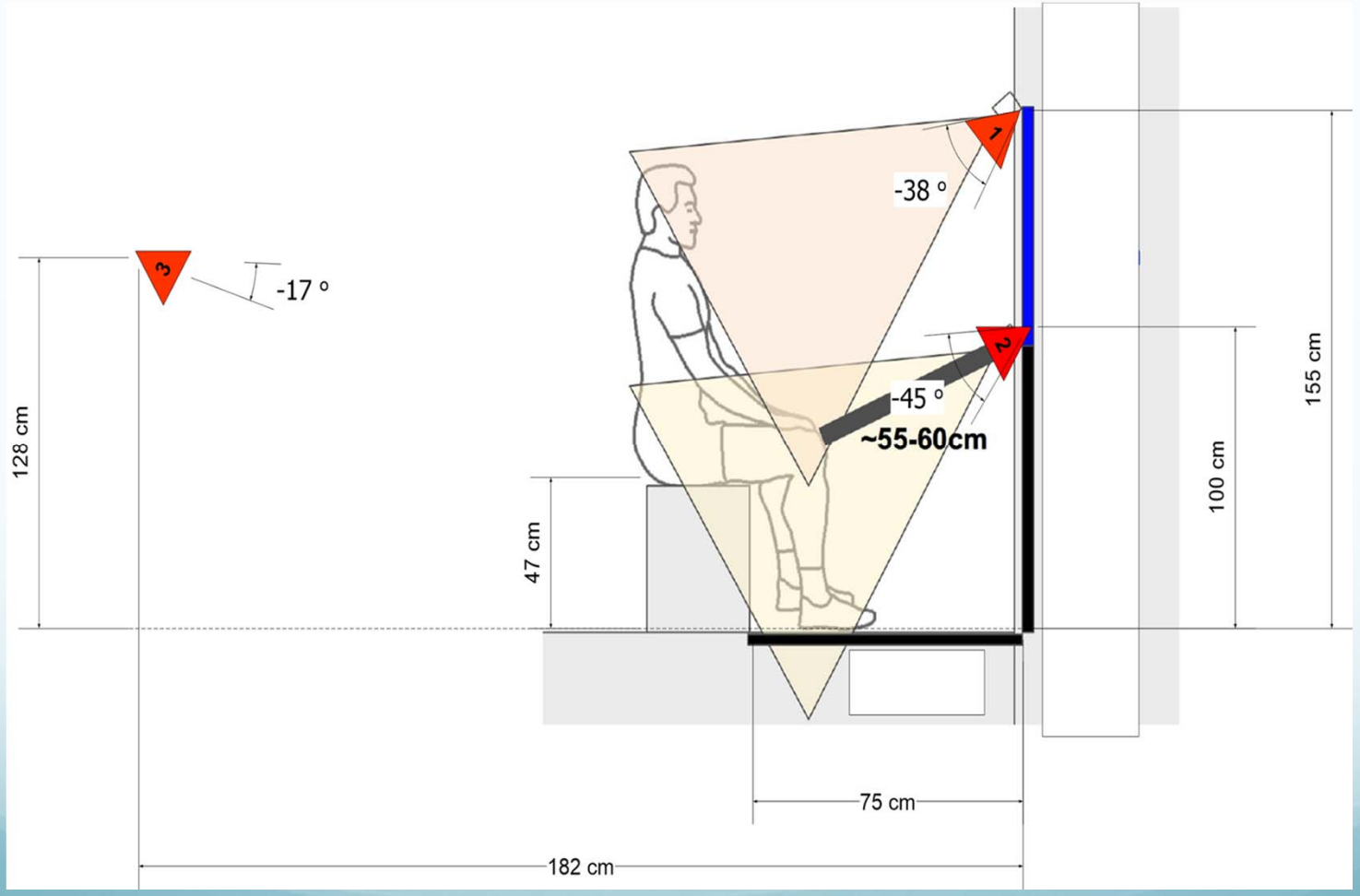
K1: back

K3: gestures

K1,2,3: audio for both tasks

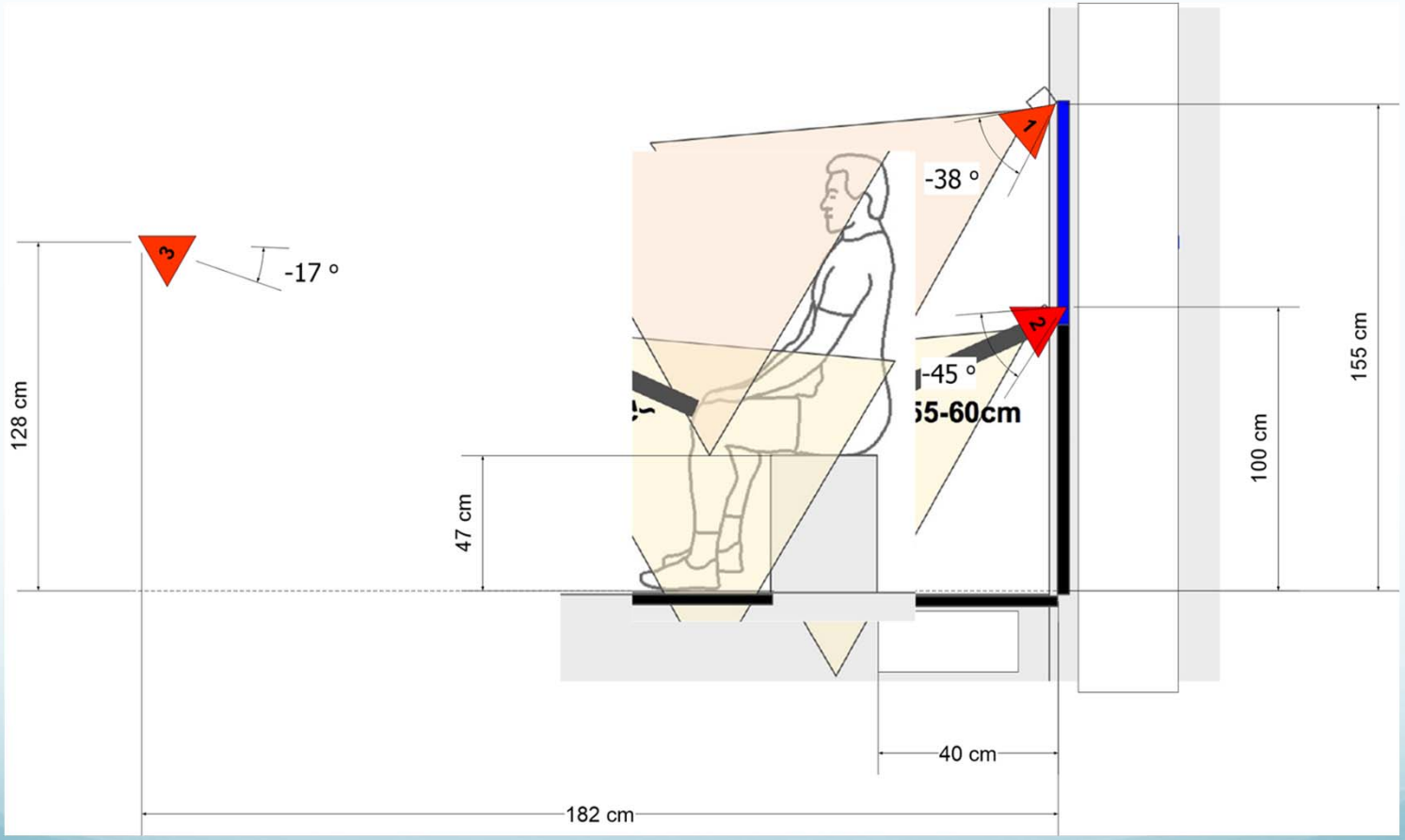
Setup of the 3 Kinects - Legs

Setup for the task “Washing the Legs”

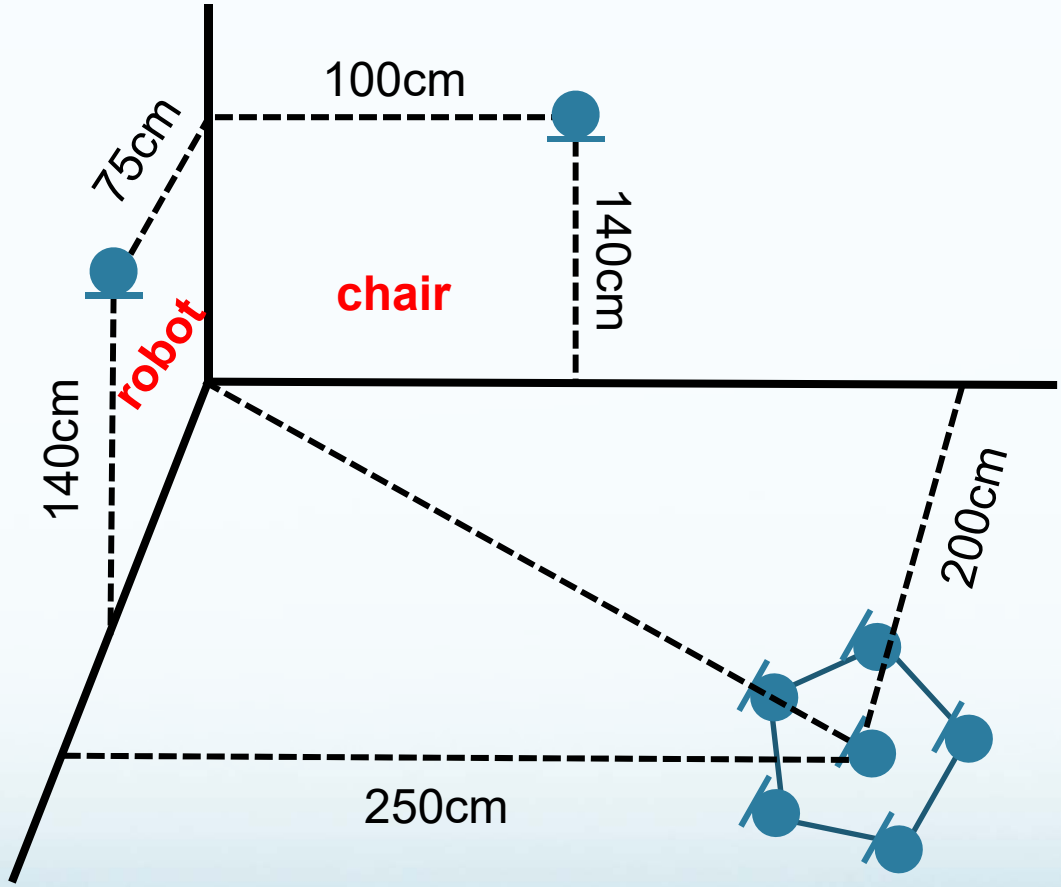


Setup of the 3 Kinects - Back

Setup for the task “Washing the Back”



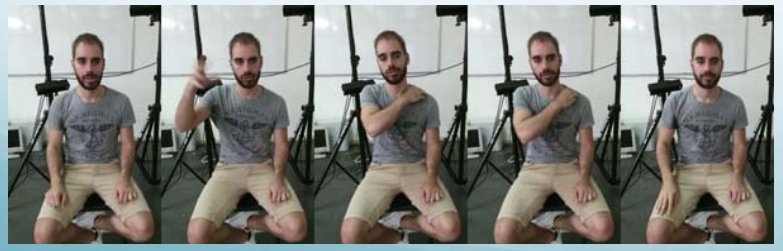
Setup of Audio Sensors (Shure mics)



Data Collection

- ICCS and KIT implemented the system architecture for the audio-visual robot perception:
  - feedback from clinical partners for the A+G commands
- @ KIT: collection of the audio-gestural dataset
- @ ICCS: small audio-gestural development dataset, incl. 28 gestures (24 users) and 23 spoken commands from 8 users
- @ FSL: collection of 10 Italian A+G commands from 7 normal users and 5 patients for pre-validation experiments

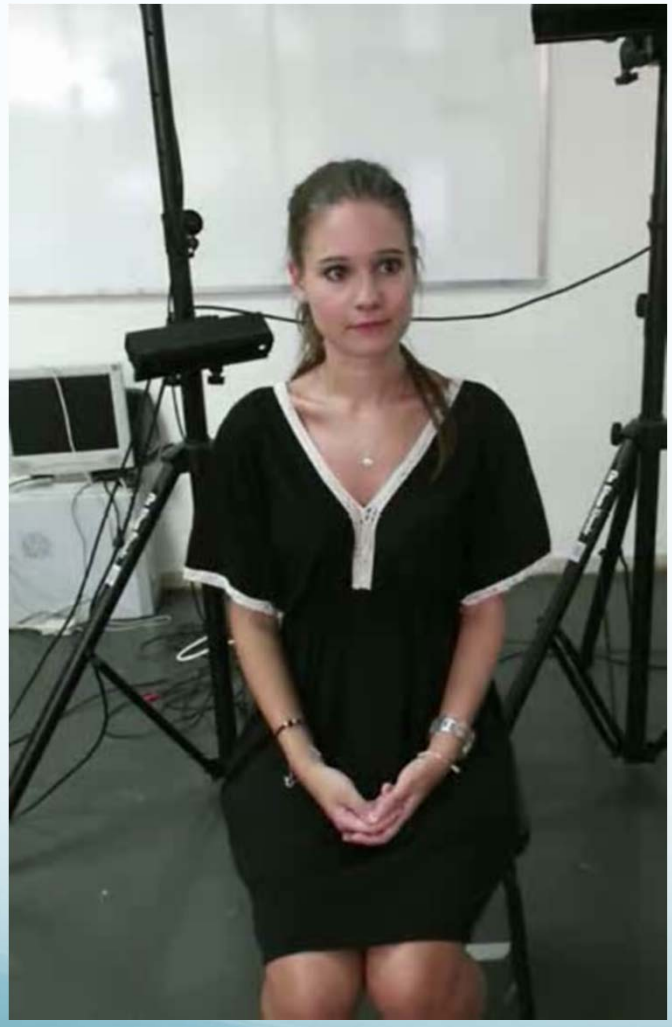
wash back



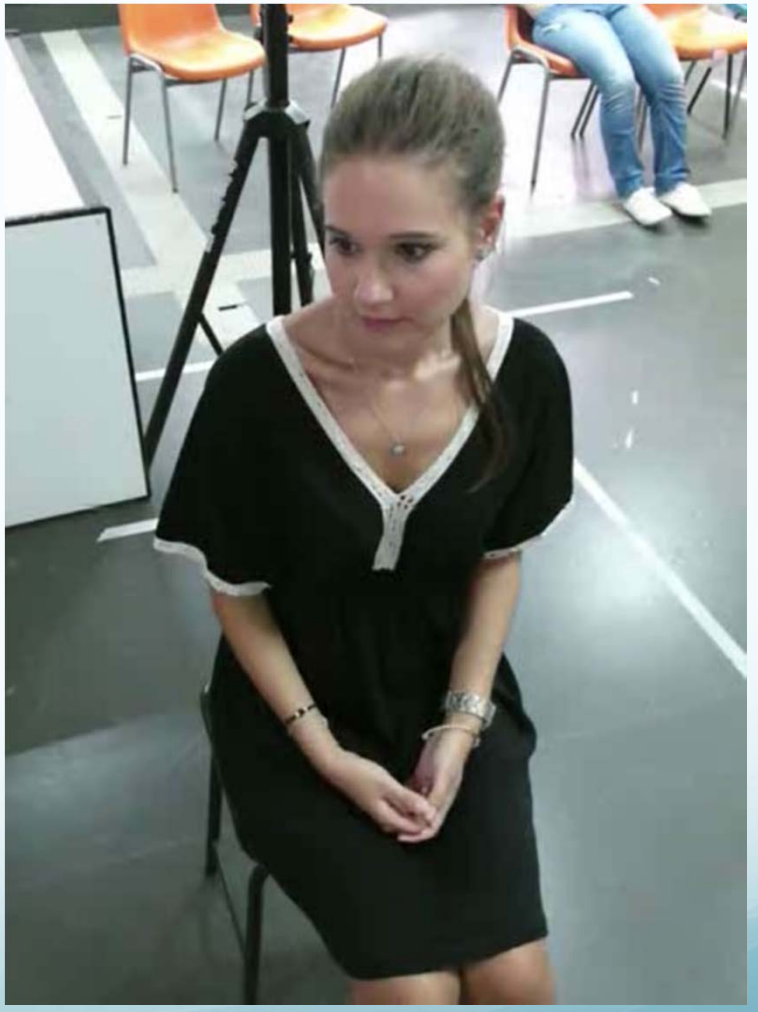
wash legs

ICCS development dataset

Kinect 3 view for "Washing the Back"

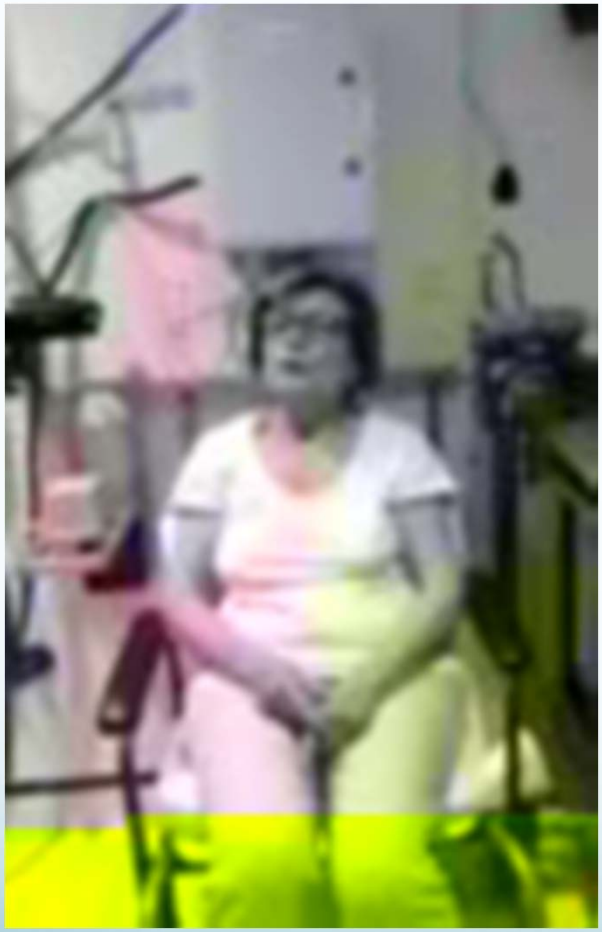


Kinect 1 view for "Washing the Legs"

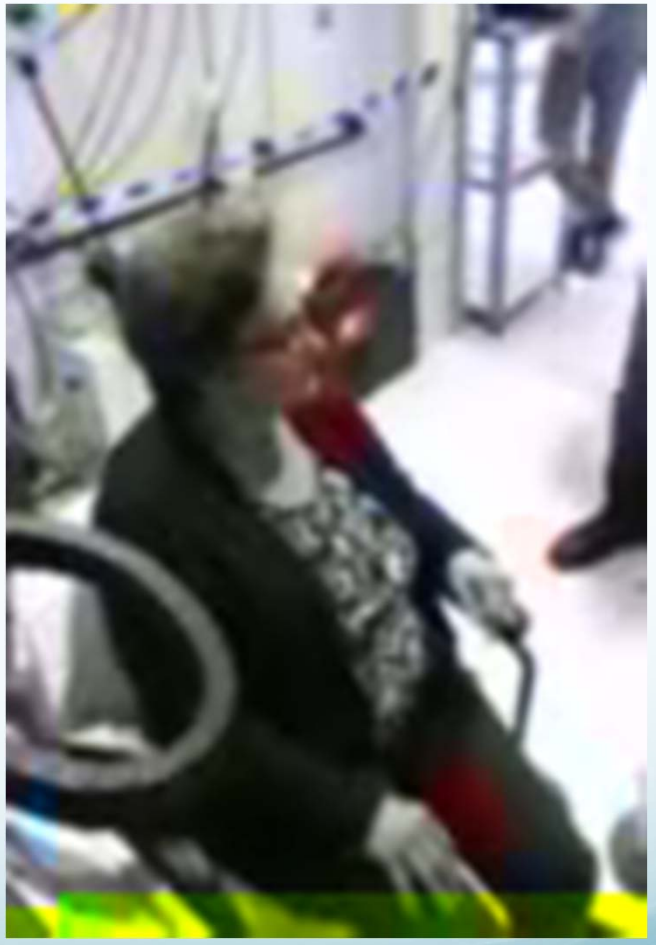


FSL pre-validation dataset

Kinect 3 view for "Washing the Back"



Kinect 1 view for "Washing the Legs"



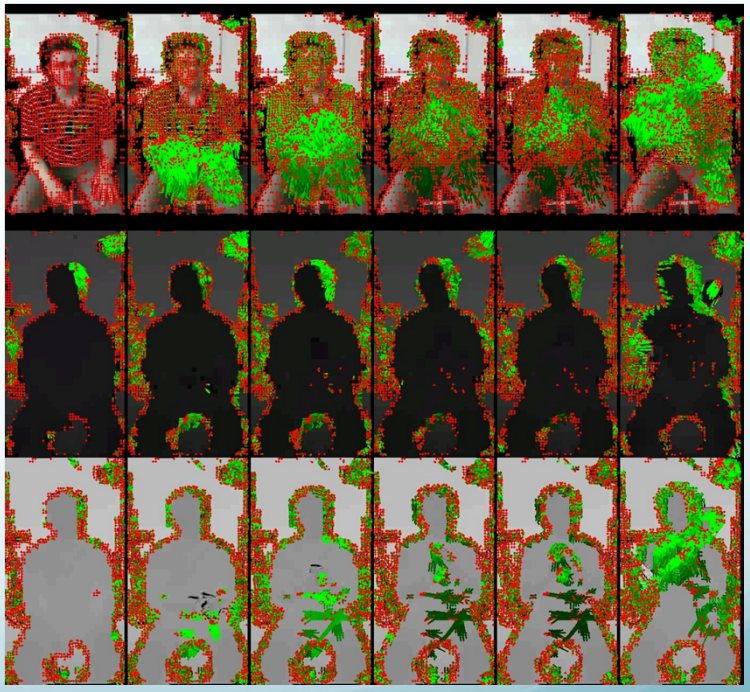
Gesture Recognition – Depth Modality

- Experiments with Depth and Log-Depth streams
- Extraction of Dense Trajectories performs better on the Log-Depth stream

RGB stream

Log-Depth stream

Dense Trajectories



Gesture Classification – Results

- **ICCS Dataset**
  - Two different setups
  - Two different streams
  - Different Encoding Methods
  - Different Features

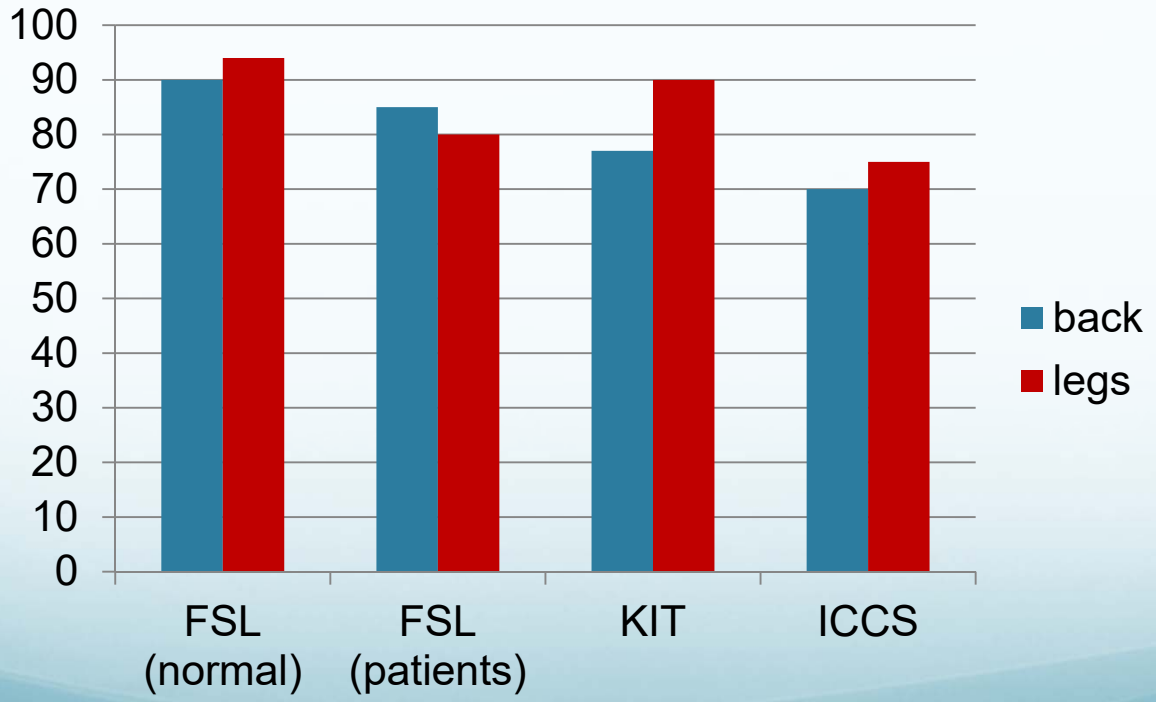
		Task: Legs		Task: Back	
Feat.	Encoding	RGB	D	RGB	D
Traj.	BoVW	69.64	60.52	77.84	60.87
HOG		41.01	53.34	58.51	57.14
HOF		74.15	66.26	82.92	71.58
MBH		77.36	65.31	80.81	65.73
Comb.		**80.88**	**74.41**	**83.92**	**75.70**
Traj.	VLAD	69.22	52.66	74.34	54.14
HOG		49.86	65.99	61.23	65.63
HOF		76.54	72.88	83.17	78.07
MBH		78.35	75.12	82.54	73.09
Comb.		**83.00**	**78.49**	**84.54**	**81.18**

- **KIT Dataset**
  - Two different setups
  - Average gesture recognition accuracy:
    - Legs (8 gestures): 83%
    - Back (10 gestures): 75%

- **FSL Dataset**
  - Train/Fine-tuning the models for audio-visual gesture recognition
  - Average gesture recognition accuracy for the 5 gestures used in validation:
    - Legs: 85% , Back: 75%

Spoken Command Classification Results (4 languages)

Dataset	Microphones	# Users	# Commands	Language
FSL	Kinect	7 normal, 5 patients	10	Italian
KIT	Shure	7 normal	10	German
ICCS	Kinect	8 normal	19	German



FSL Preliminary first Validation Results

- 25 naïve patients
- 5 audio-gestural commands
- Task: Always-listening Online A-G Command Recognition
- Challenges: Acoustic noise, low voice, weak motions
- Multimodal Command Recognition Rate:

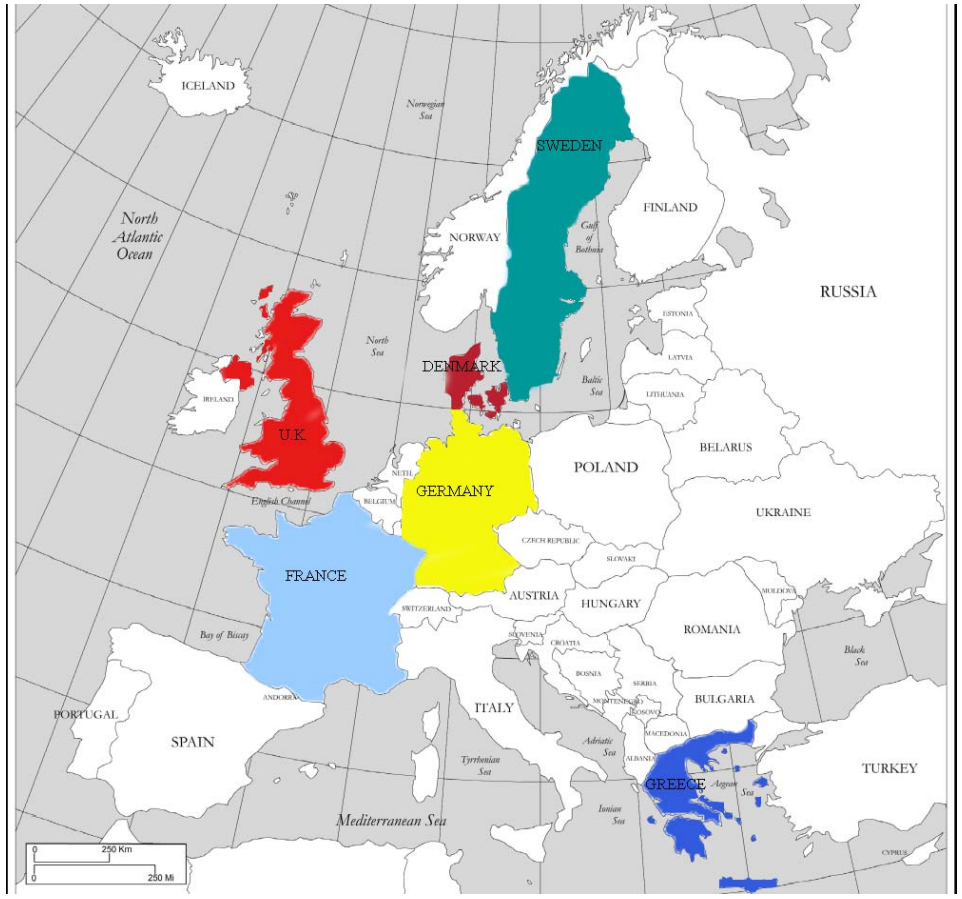
$$\text{MCRR} = \frac{\#\ commands\ correctly\ recognized\ by\ system}{\#\ commands\ correctly\ performed\ by\ user}$$

position	System's MCRR%	User performance (speech)	User performance (gestures)	# false alarms
Back	87.16%	99.16%	78.15%	3
Legs	79.52%	98.29%	81.14%	3

EU Project: Baby Robot

BabyRobot: Child-Robot Communication and Collaboration: Edutainment, Behavioural Modelling and Cognitive Development

H2020-ICT-2015-1, GA # 687831, Duration: 2016-2018, website: http://www.babyrobot.eu/



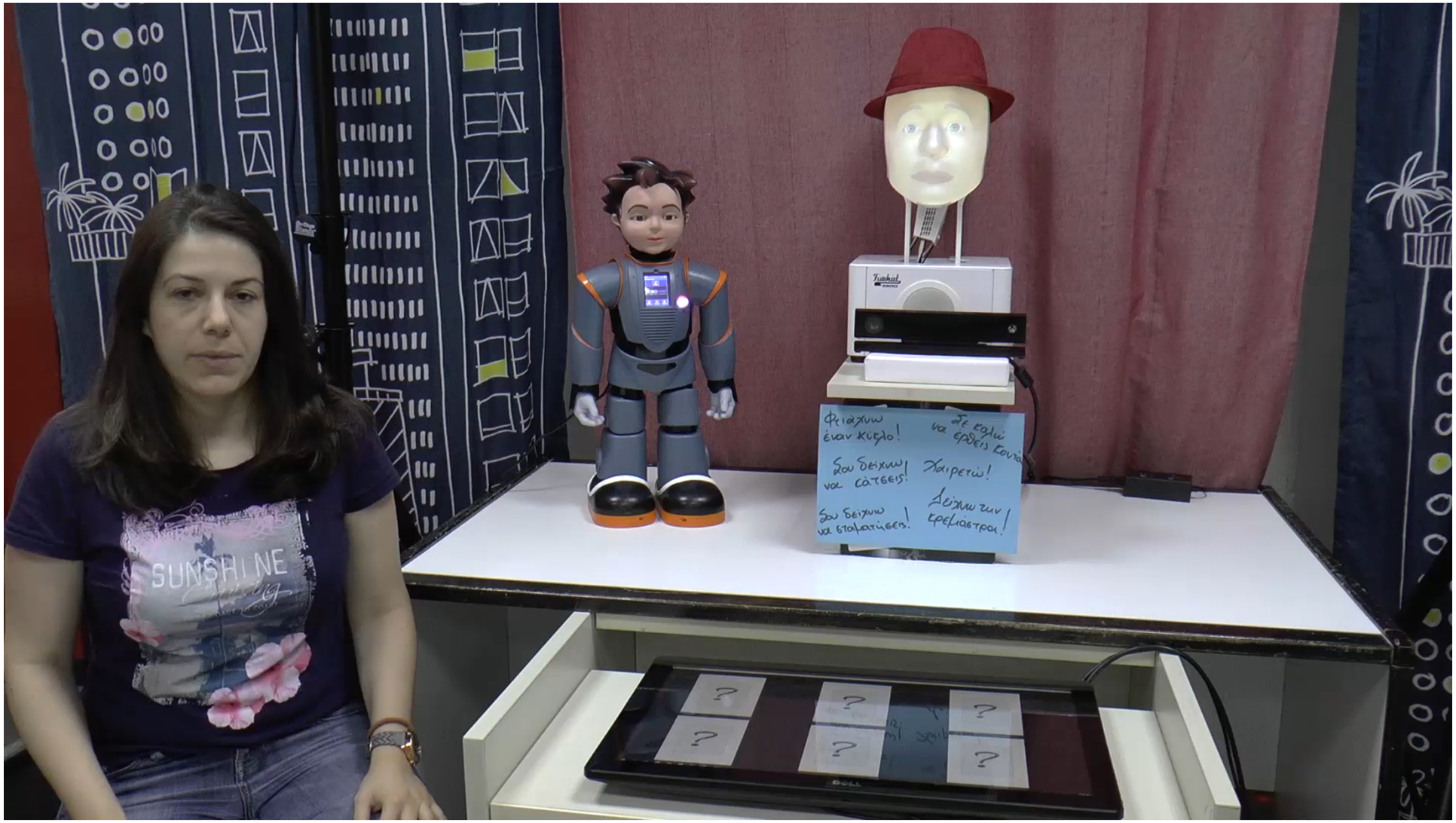
Our Team's Contribution

- Human Localization/Tracking
- Gesture Recognition
- Action Recognition
- Behavioral modeling and intention

Partners: ICCS-NTUA (Greece), AthenaRC-RPI (Greece), KTH (Sweden), UH (U.K.), UNIBI (Germany), BlueOcean (Denmark), USTL (France), Furhat (Sweden)

BabyRobot project-Setup Room

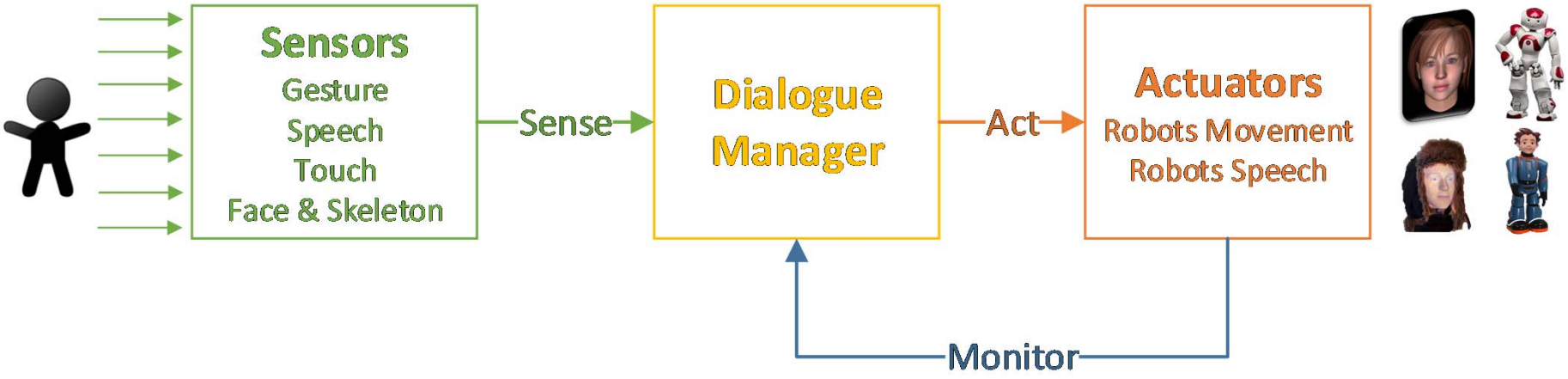
Emotion recognition



Pantomime

Experimental Setup: Dialog Management for HRI

- IrisTK manages dialog through **events**:
  - *Sense:* what the system perceives
  - *Act:* things the system should do
  - *Monitor:* feedback from actuators



System Evaluation (submitted to ICRA-2018)

- Experiments with 28 TD children 6 – 10 years old

- Objective Evaluation
  - 83.5% average classification accuracy for 8 gestures
  - 74.5% average classification accuracy for 13 pantomimes
  - 97.8% average word accuracy
  - 97.0% average sentence accuracy

- Subjective Evaluation

Children:

  - want to play with robots frequently
  - feel good when playing with robots
  - believe that it wasn't difficult to play with robots

Conclusions

- **Synopsis**:
  - Visual Action Recognition
  - Gesture Recognition
  - Spoken Command Recognition
  - Fusion for improved recognition
  - Gait Analysis

- **Ongoing work**:
  - Couple Human Localization & Pose with Activity recognition
  - Activities: Actions – Gestures – SpokenCommands - Gait
  - Applications in Robotics (EU projects: MOBOT, BabyRobot, I-Support)

For more information, demos, and current results:

http://cvsp.cs.ntua.gr and http://robotics.ntua.gr

Collaborators and References

Arvanitakis, Antonis
Chalvatzaki, Georgia
Dometios, Thanos
Efthymiou, Niki
Filntisis, Panagiotis
Kardaris, Nikos
Katsamanis, Nasos
Koutras, Petros
Papageorgiou, Xanthi
Pitsikalis, Vassilis
Potamianos, Gerasimos
Rodomagoulakis, Isidoros
Theodorakis, Stavros
Tsiami, Antigoni
Tzafestas, Costas
Zlatintsi, Nancy

For more information, demos, and current results:

http://cvsp.cs.ntua.gr and http://robotics.ntua.gr