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

Part 2:
Audio-Visual HRI: Methodology and Applications in Assistive Robotics

Petros Maragos and Athanasia Zlatintsi

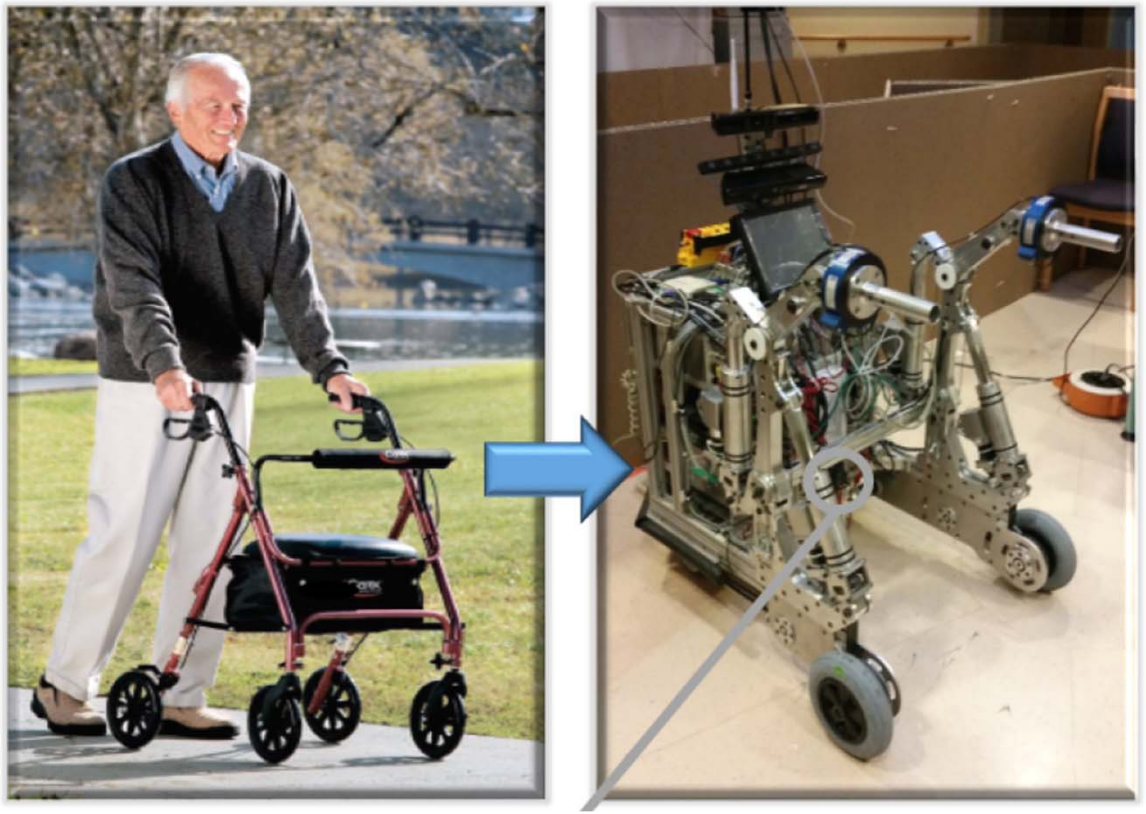
slides: http://cvsp.cs.ntua.gr/interspeech2018

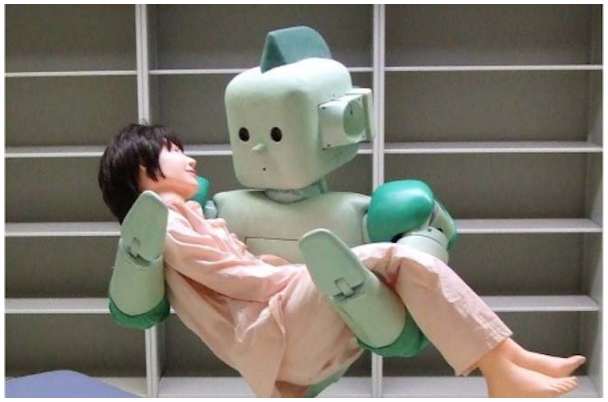
Tutorial at INTERSPEECH 2018, Hyderabad, India, 2 Sep. 2018

2A.
Audio-Visual HRI:
General Methodology

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

Database of Multimodal Gesture Challenge

(in conjunction with *ACM ICMI 2013*)

- 20 cultural/anthropological signs of Italian language


- ‘vattene’ (get out)

- ‘vieni qui’ (come here)

- ‘perfetto’ (perfect)

- ‘furbo’ (clever)

- ‘che due palle’ (what a nuisance!)

- ‘che vuoi’ (what do you want?)

- ‘d’accordo’ (together)

- ‘sei pazzo’ (you are crazy)

- ‘combinato’ (combined)

- ‘freganiente’ (damn)


- ‘ok’ (ok)

- ‘cosa ti farei’ (what would I make to you!)

- ‘basta’ (that’s enough)

- ‘prendere’ (to take)

- ‘non ce ne piu’ (there is none more)

- ‘fame’ (hunger)

- ‘tanto tempo’ (a long time ago)

- ‘buonissimo’ (very good)

- ‘messi d’accordo’ (agreed)

- ‘sono stufo’ (I am sick)

- 22 different users
- 20 repeats per user approximately

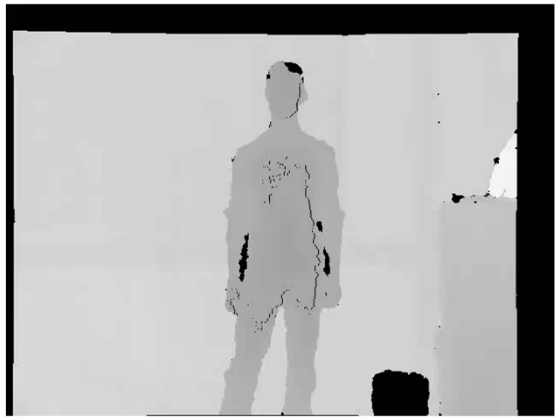
(~1 minute for each gesture video)

Multimodal Gesture Signals from Kinect-0 Sensor

RGB Video & Audio

Depth

(vieniqui - *come here*)



Skeleton

(vieniqui - *come here*)

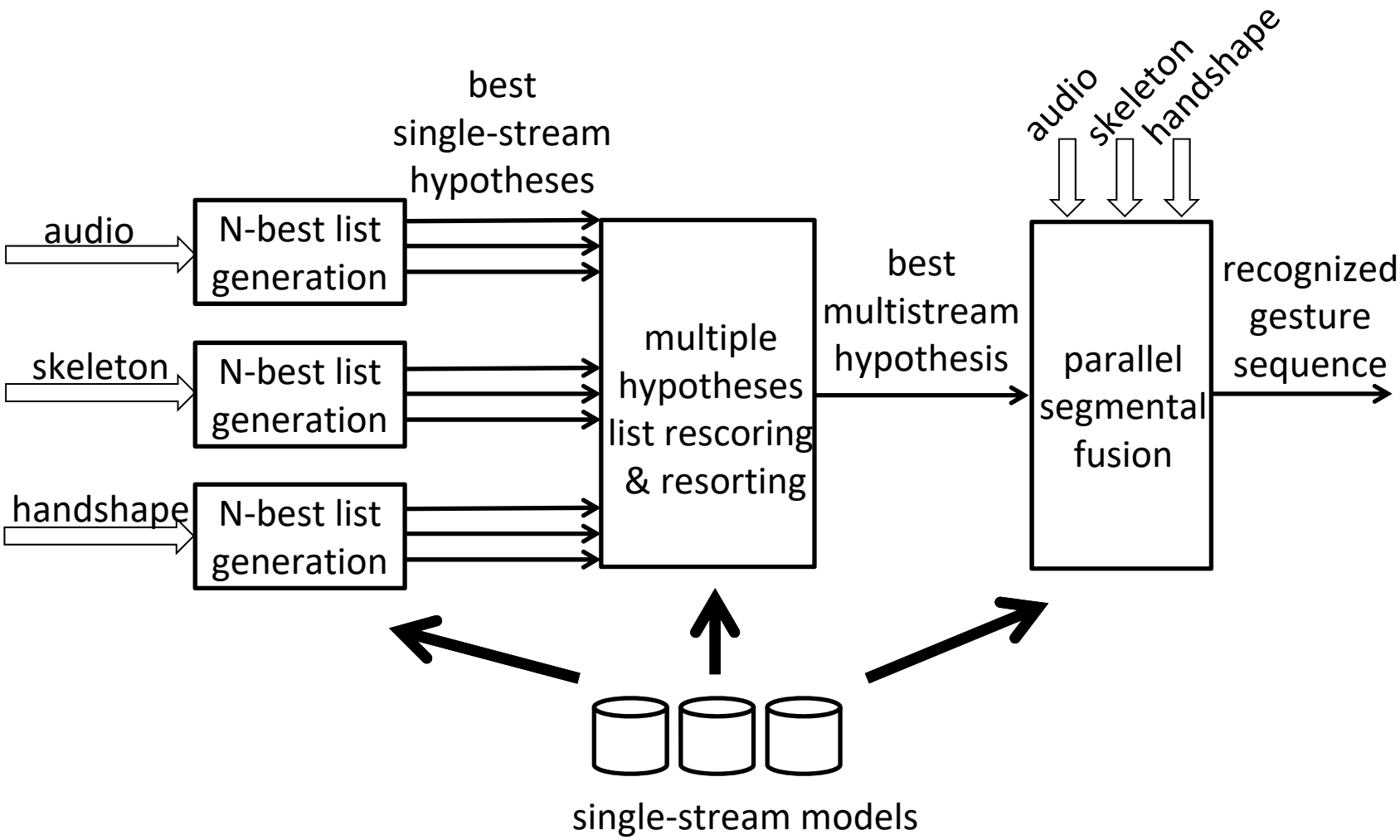
User Mask

(vieniqui - *come here*)

ChaLearn corpus

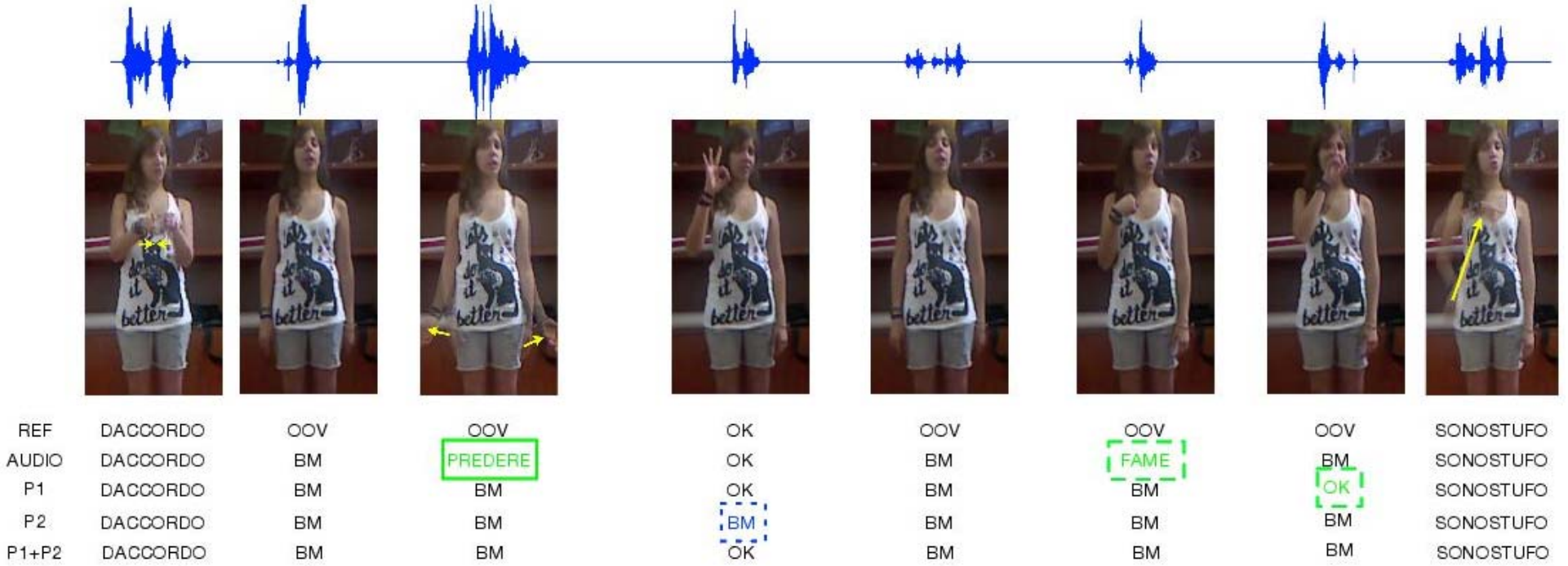
[S. Escalera, J. Gonzalez, X. Baro, M. Reyes, O. Lopes, I. Guyon, V. Athitsos, and H. Escalante, "*Multi-modal gesture recognition challenge 2013: Dataset and results*", Proc. 15th ACM Int'l Conf. Multimodal Interaction, 2013.]

Multimodal Hypothesis Rescoring + Segmental Parallel Fusion



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

Audio-Visual Fusion & Recognition



REF	DACCORDO	OOV	OOV	OK	OOV	OOV	OOV	SONOSTUFO
AUDIO	DACCORDO	BM	PREDERE	OK	BM	FAME	BM	SONOSTUFO
P1	DACCORDO	BM	BM	OK	BM	BM	OK	SONOSTUFO
P2	DACCORDO	BM	BM	BM	BM	BM	BM	SONOSTUFO
P1+P2	DACCORDO	BM	BM	OK	BM	BM	BM	SONOSTUFO

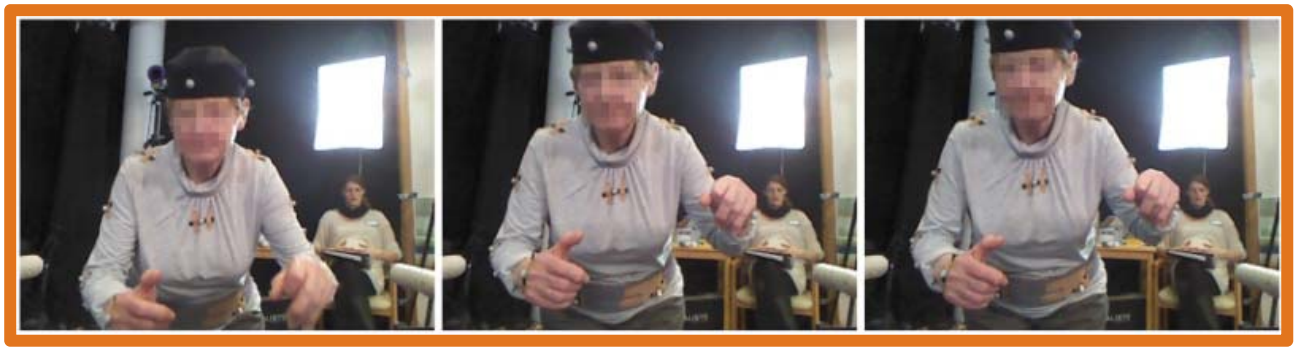
- Audio and visual modalities for A-V gesture word sequence.
- Ground truth transcriptions (“REF”) and decoding results for audio and 3 different A-V fusion schemes.
- Results in top rank of ChaLearn (ACM 2013 Gesture Challenge – 50 teams - 22 users x 20 gesture phrases x 20 repeats).

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

Visual Activity Recognition

dfwlrq jhvwxuh vljq

Action: sit to stand

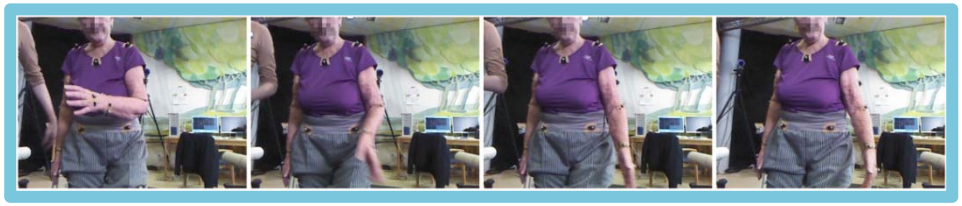


Gestures: come here, come near

Sign:
(GSL) Europe

Visual action recognition pipeline

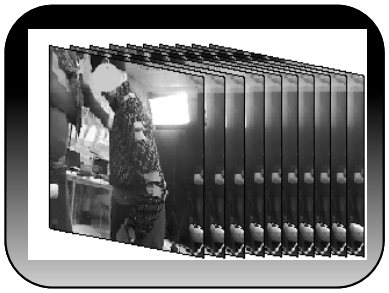
Sit to Stand

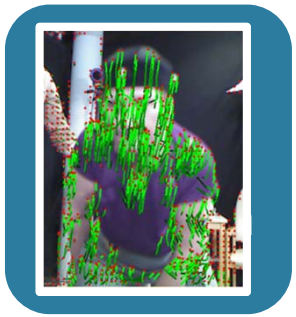


Walk

Video

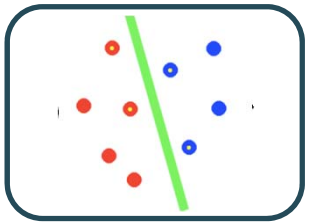
Sliding Window





Visual Feature Extraction

Ihdwxuh Hqfrglqj

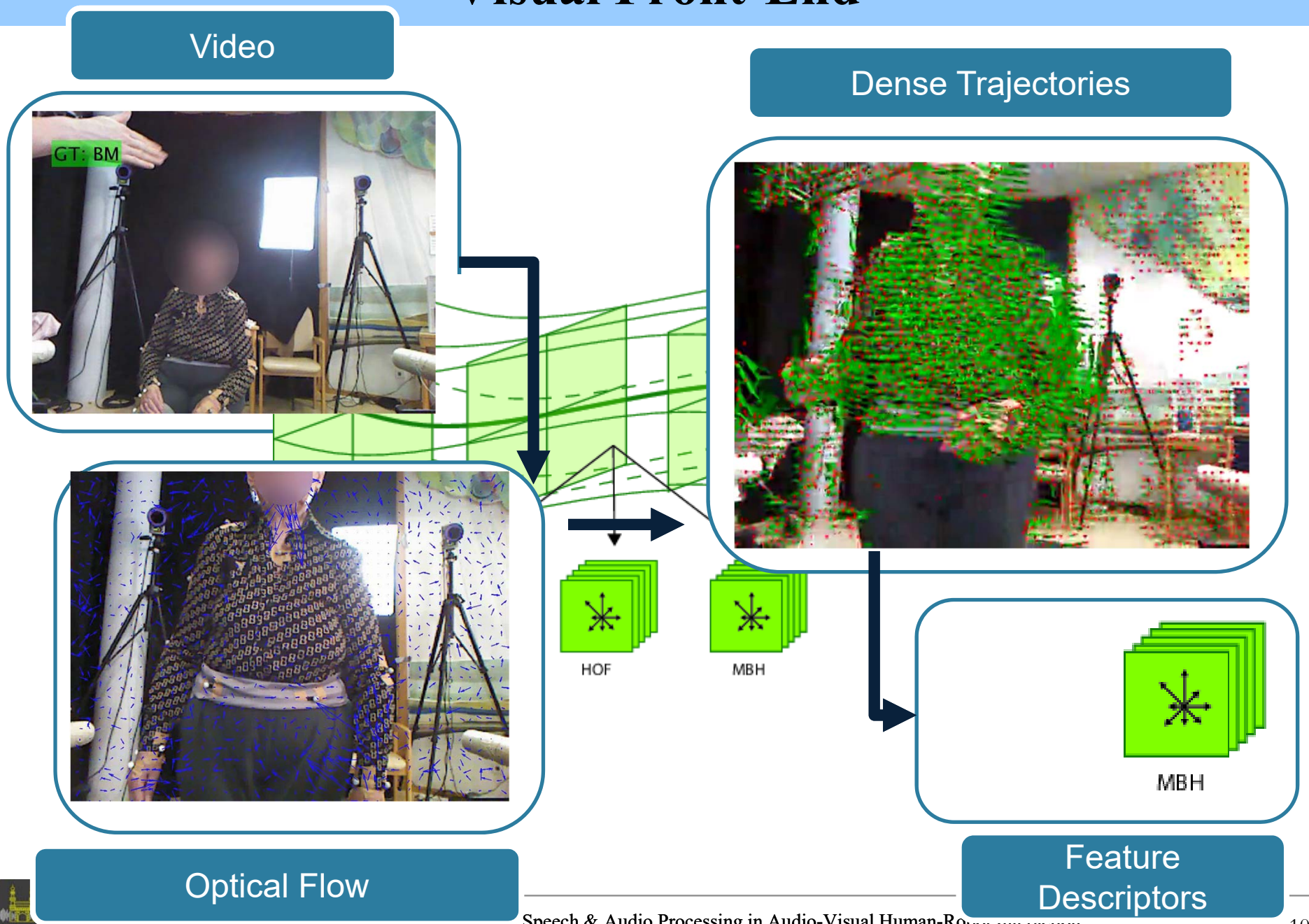


Classifier

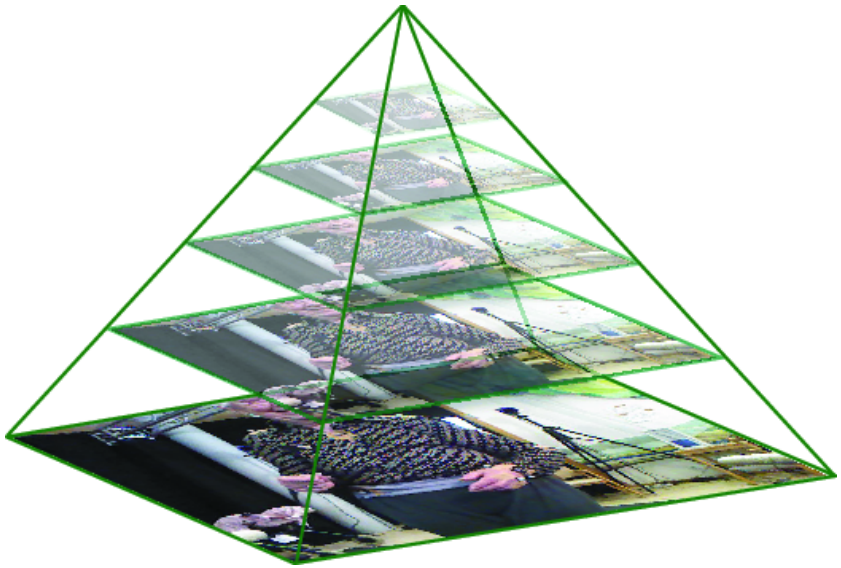
Post-processing

Recognized Sequence

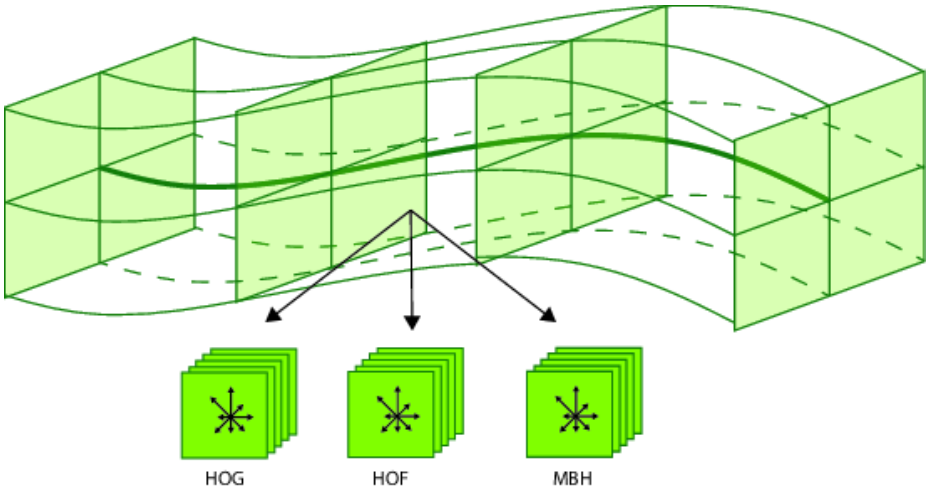
Visual Front-End



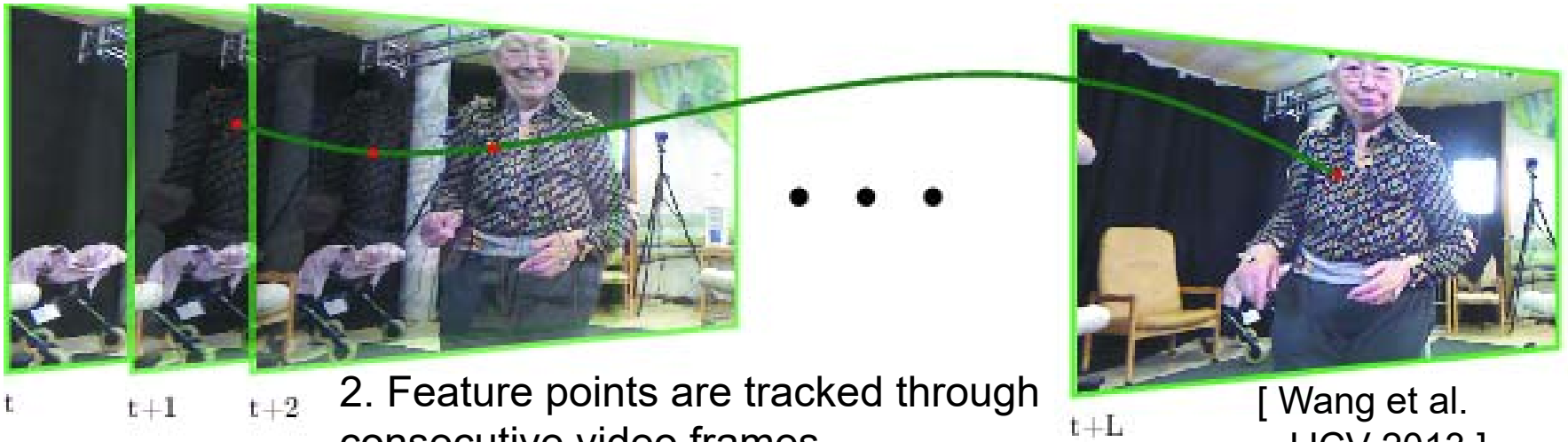
Features: Dense Trajectories



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



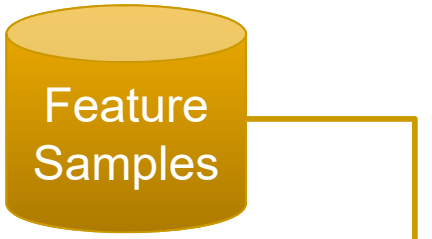
3. Descriptors are computed in space-time volumes along trajectories

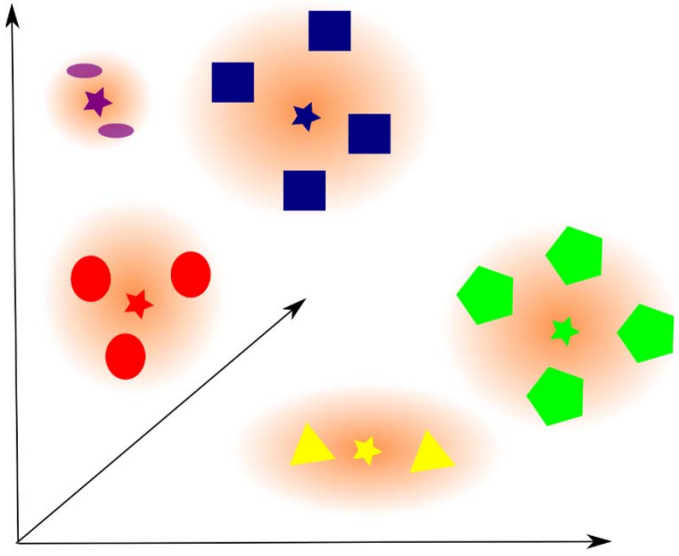


2. Feature points are tracked through consecutive video frames

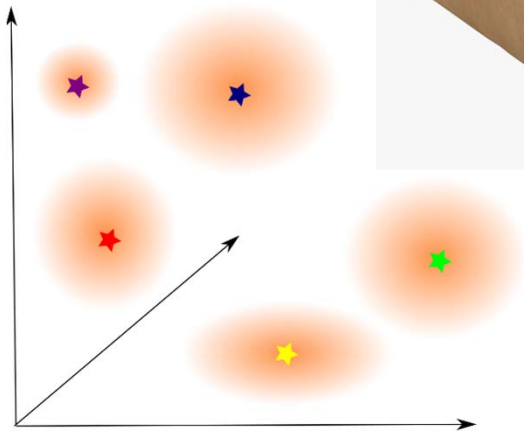
[Wang et al. IJCV 2013]

K-means Clustering and Dictionary



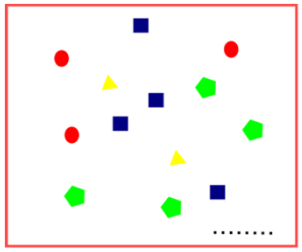


K-means

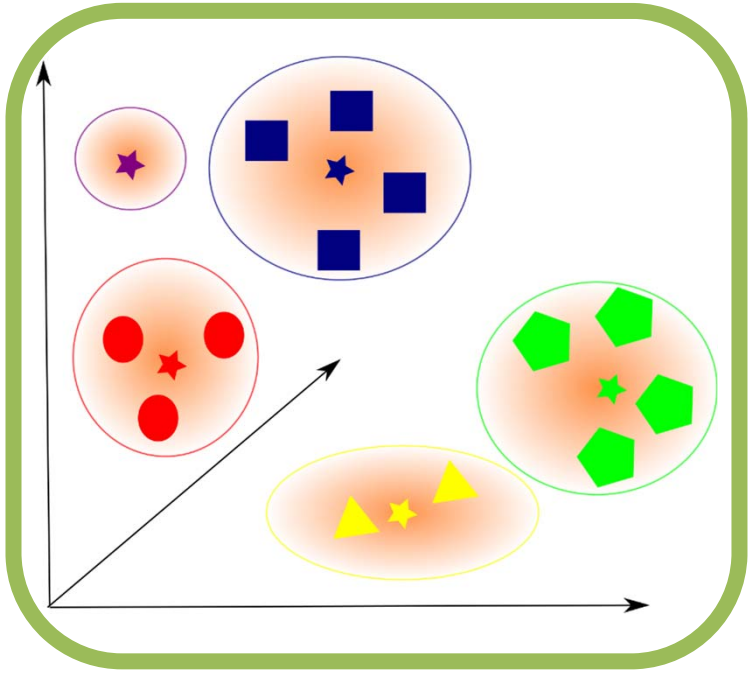


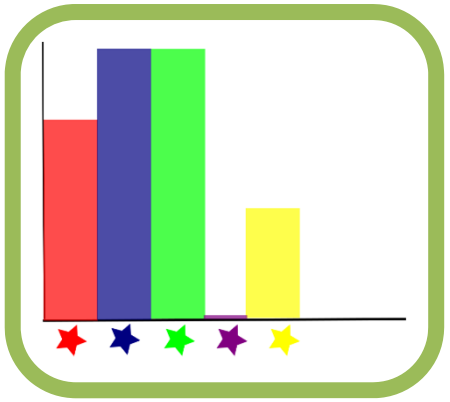
Dictionary

Feature Encoding



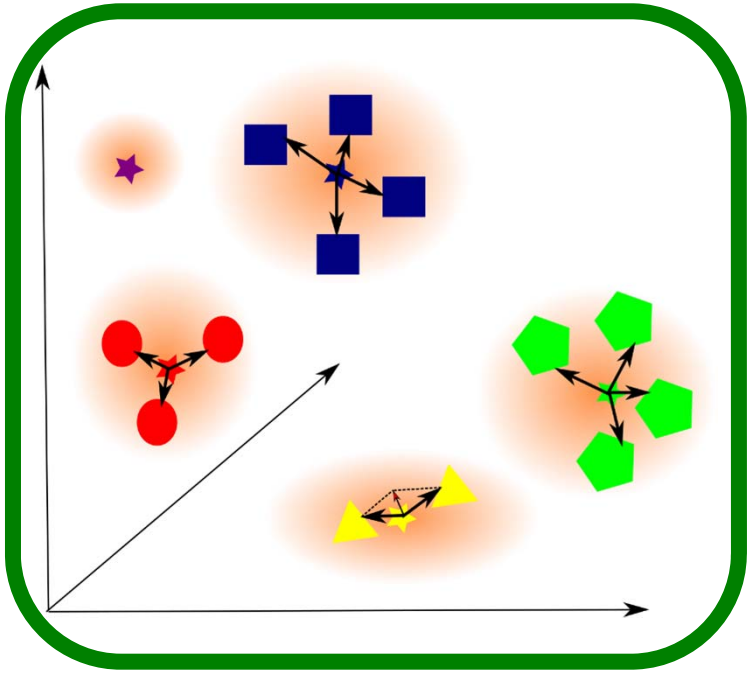
BOF

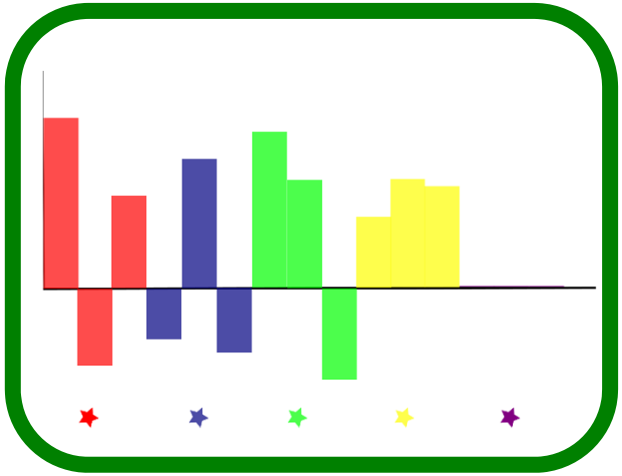




BOF - Size: K

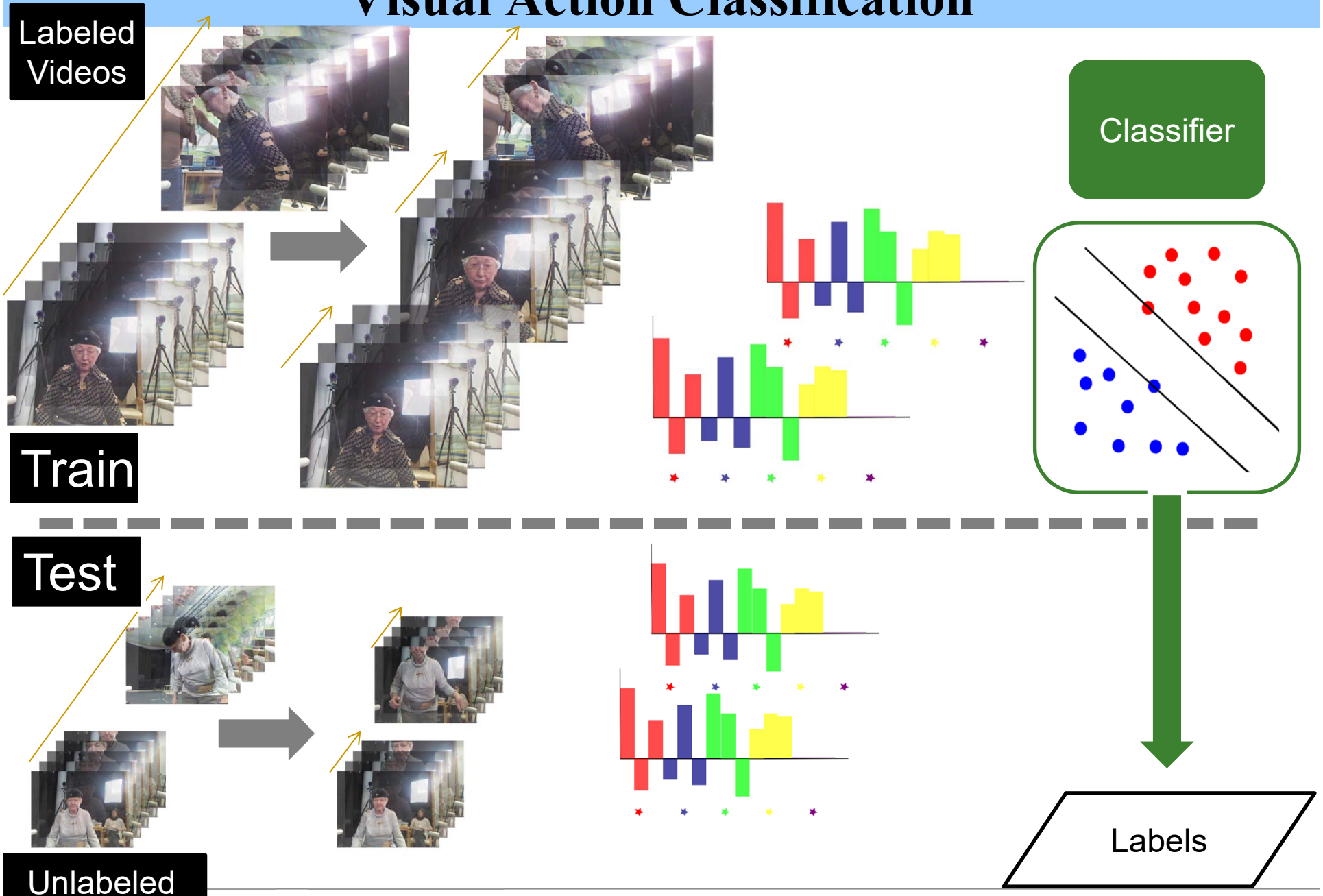
VLAD



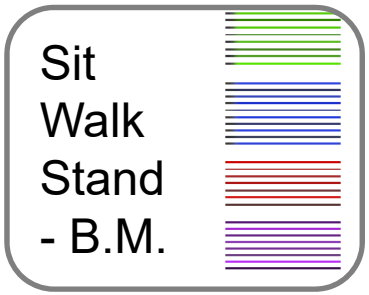


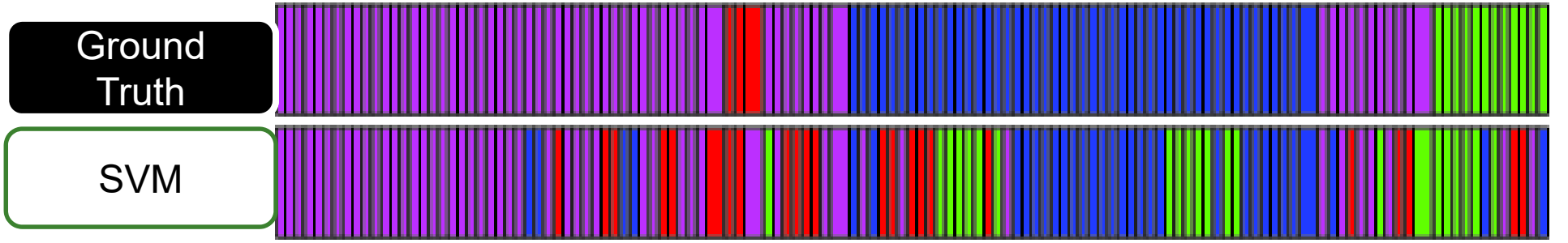
VLAD - Size: K*D

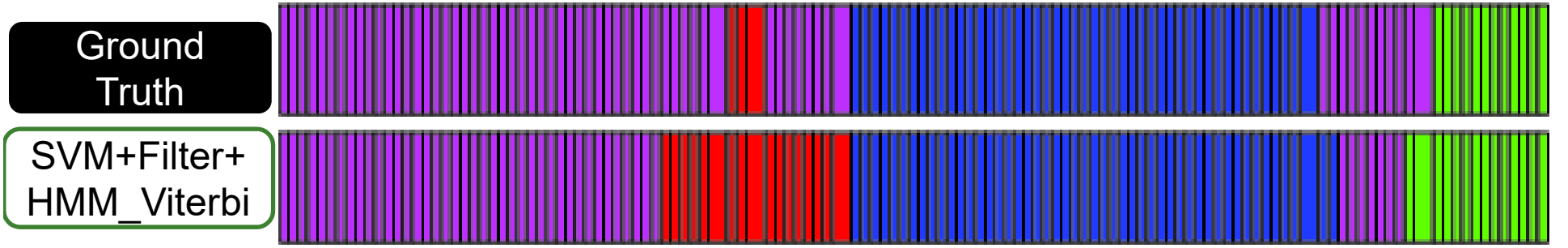
Visual Action Classification



Temporal Segmentation Results

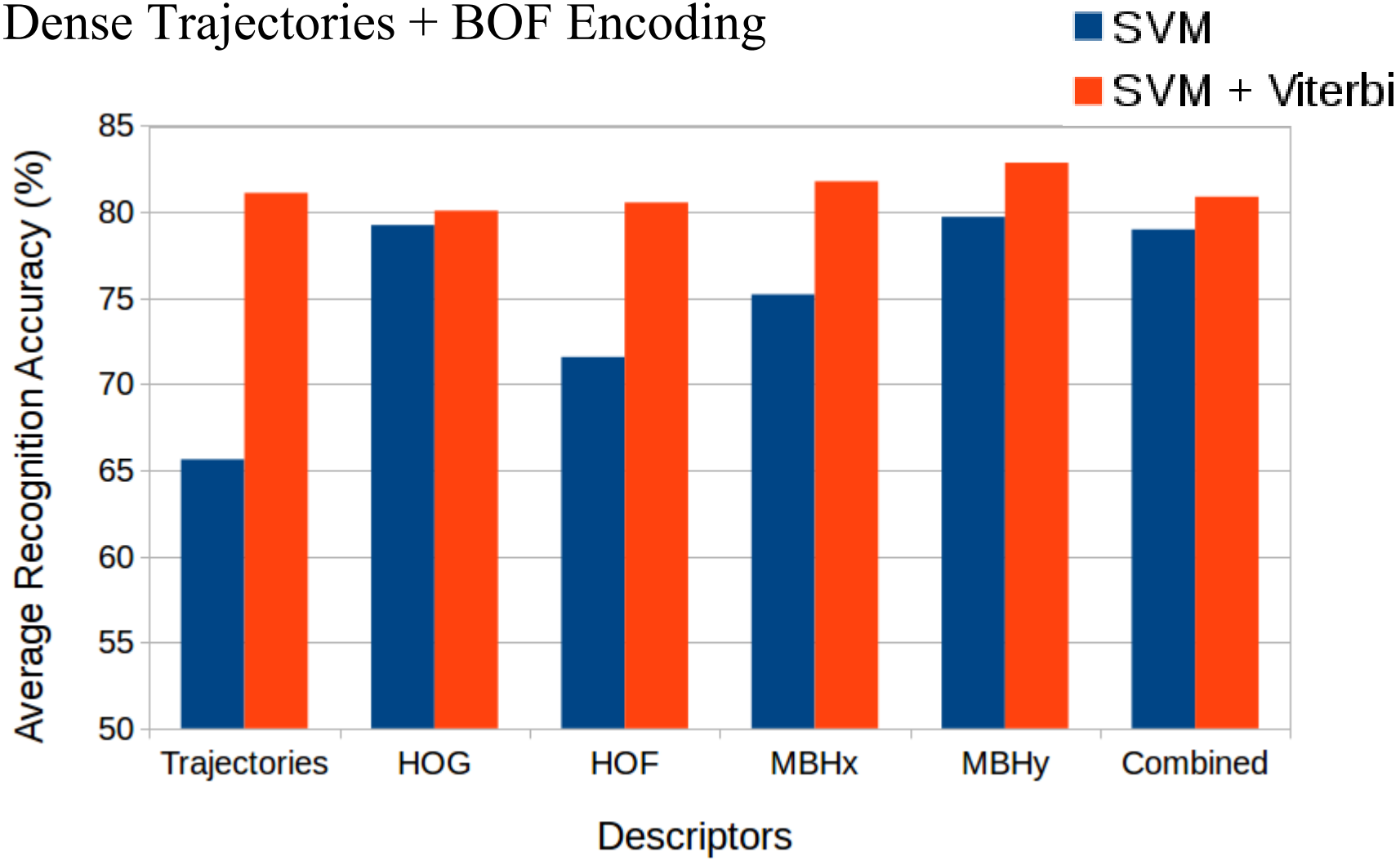






Action Recognition Results (4a, 6p): Descriptors + Post-processing Smoothing

- Dense Trajectories + BOF Encoding



Results improve by adding Depth and/or advanced Encoding

Gesture Recognition

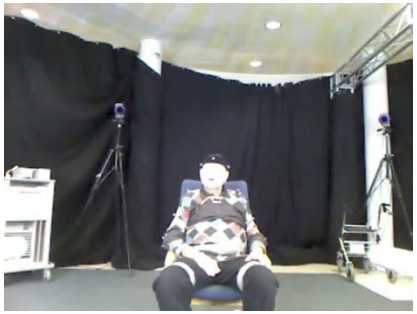
Gesture Recognition Challenges

Challenging task of recognizing human gestural movements:

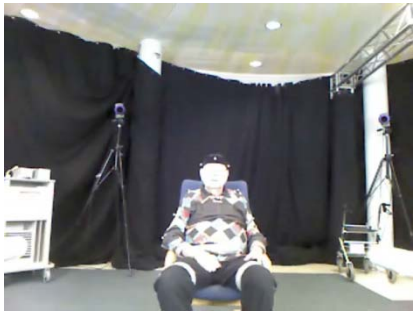
- Large variability in gesture performance.
- Some gestures can be performed with left or right hand.

Come Closer


I want to Sit Down

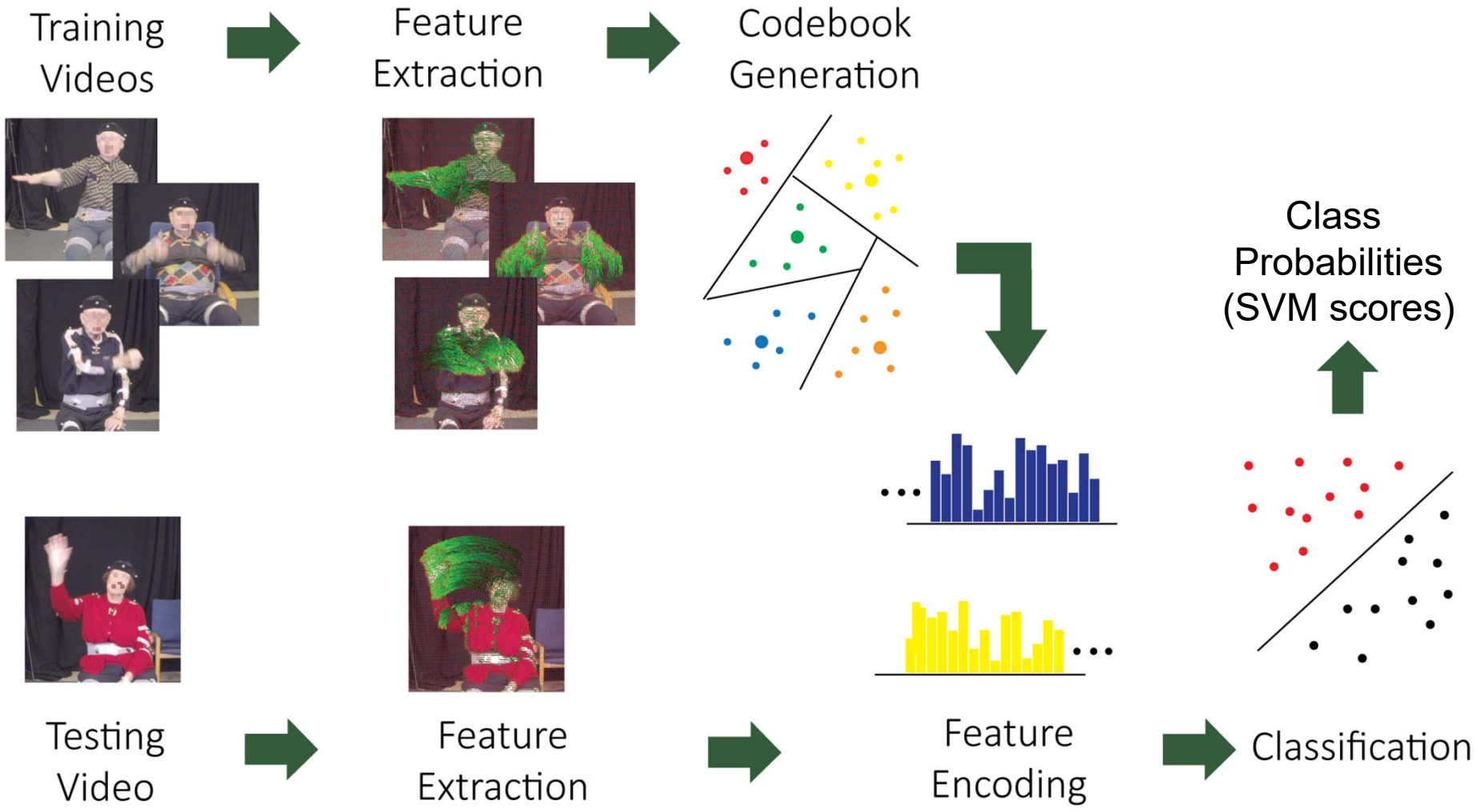




Park

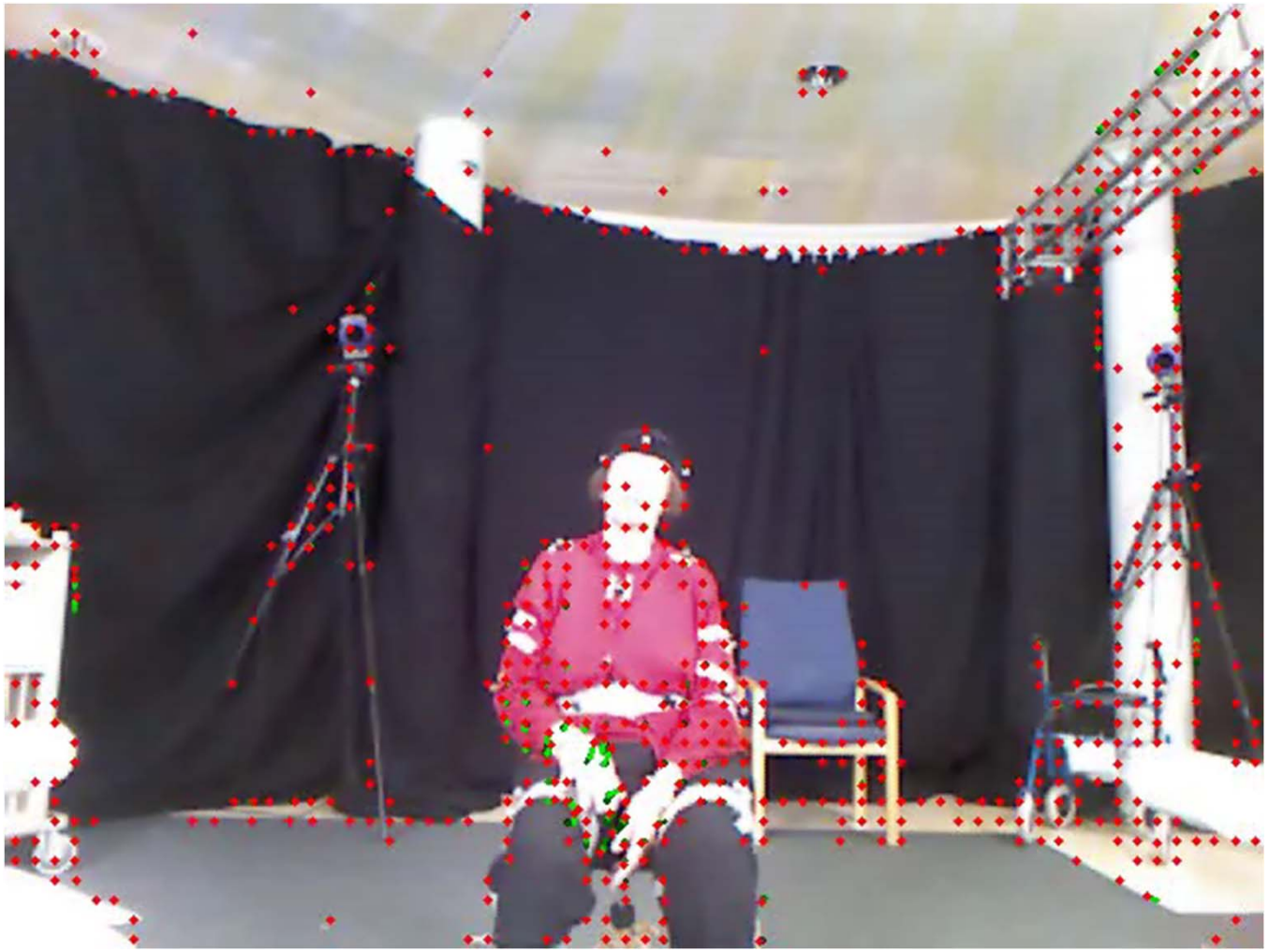



I want to Perform a Task

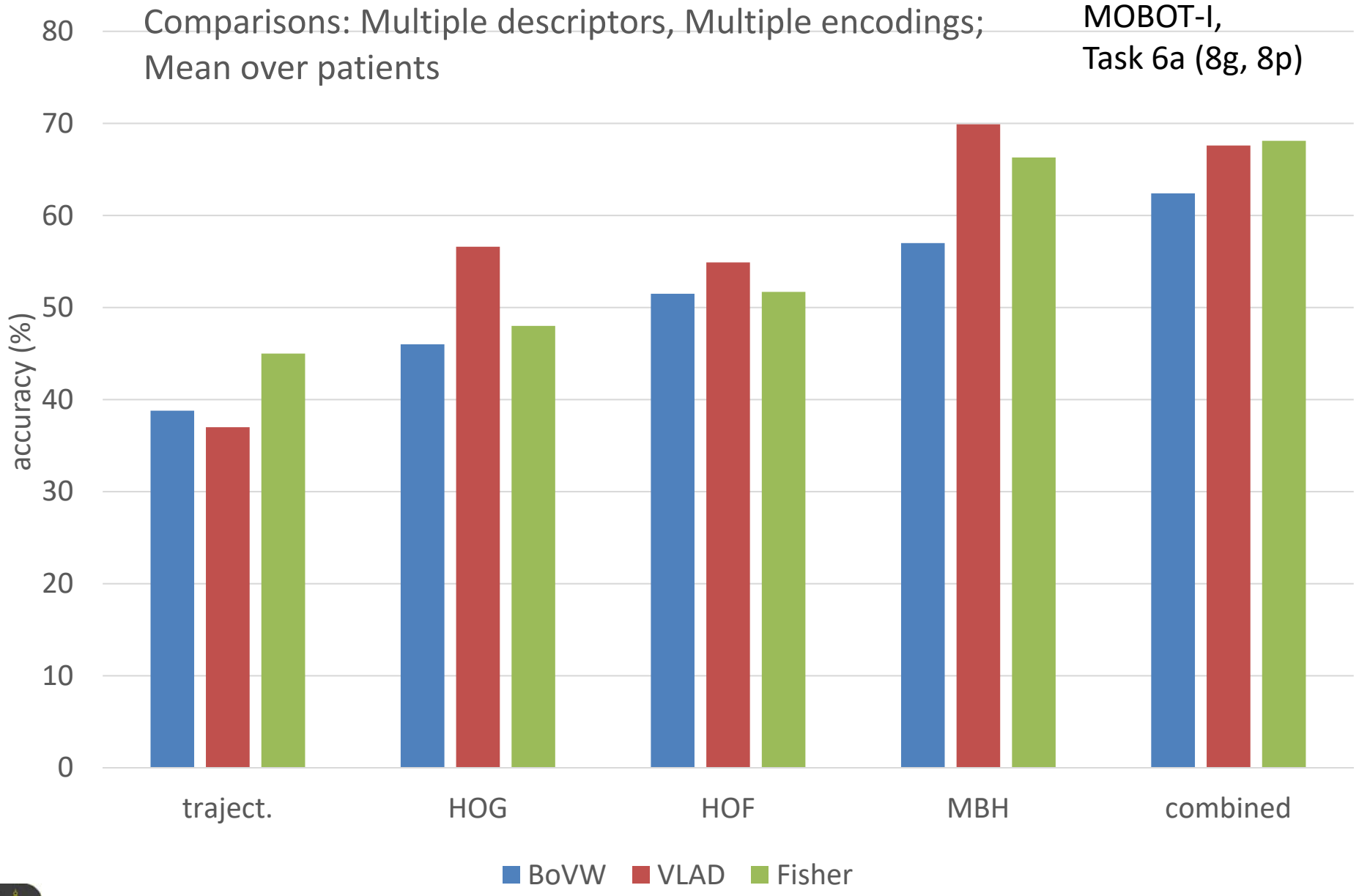
Visual Gesture Classification Pipeline



Applying Dense Trajectories on Gesture Data

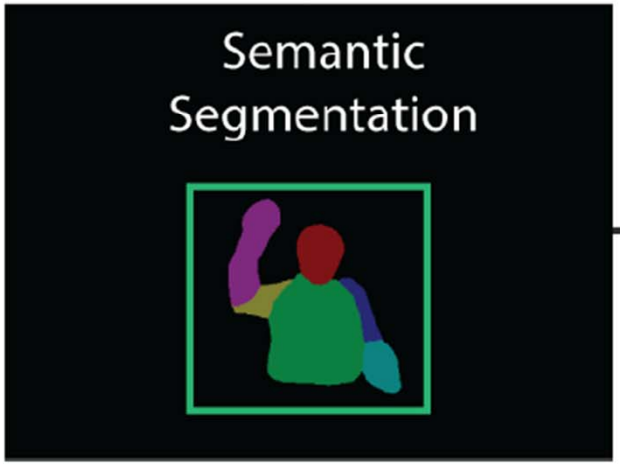


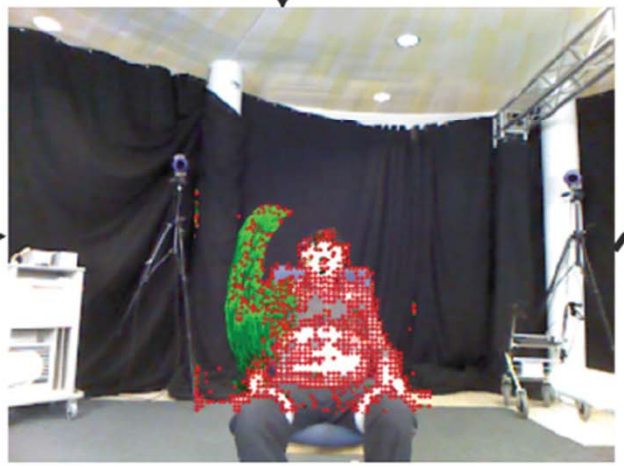
Extended Results on Gesture Recognition



Visual Synergy: Semantic Segmentation + Gesture Recognition

Gesture Classification





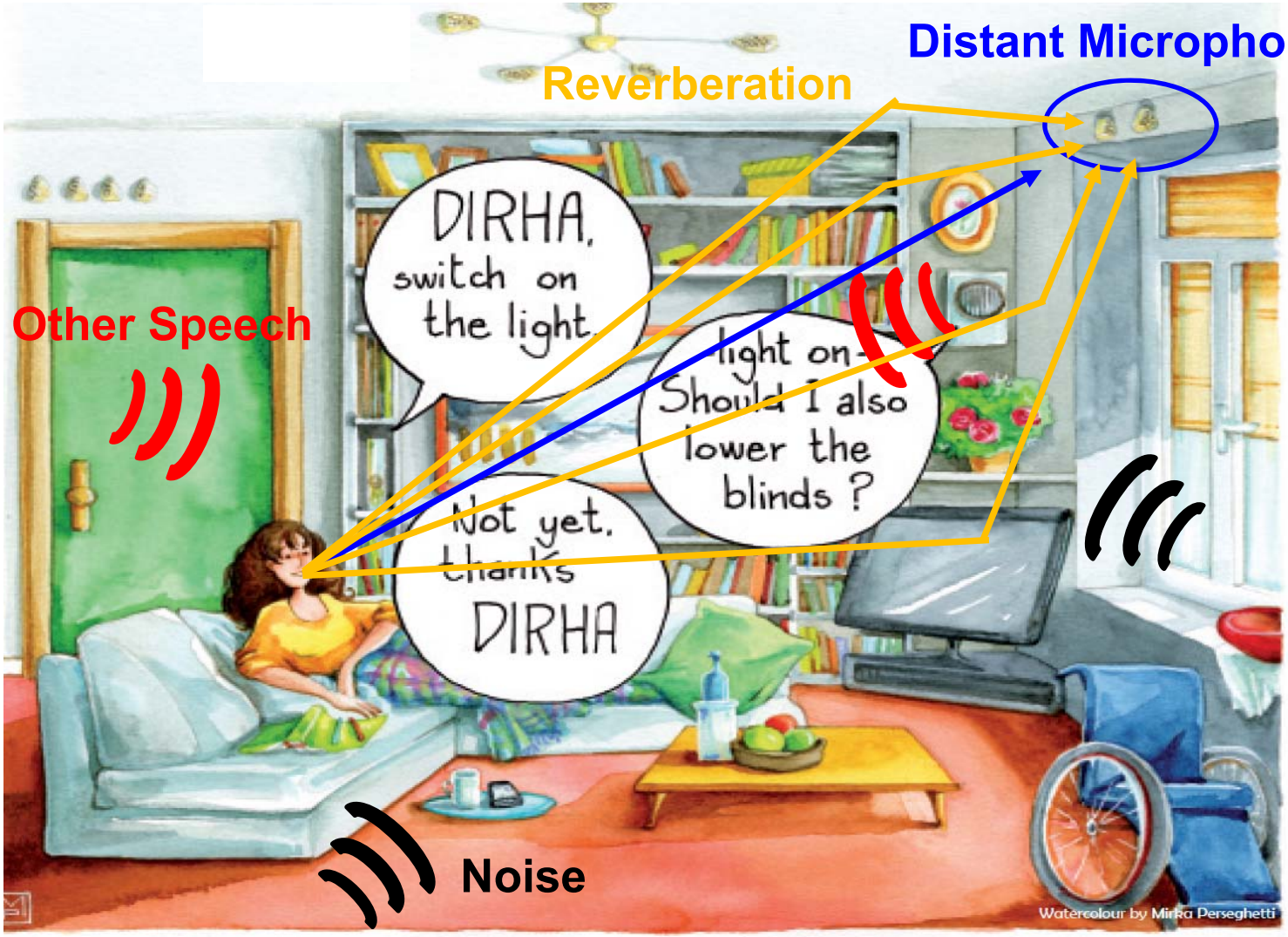
Median Relative Improve: 9 %

foreground/background+gesture recognition

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*" ECCV Workshop on Assistive Computer Vision and Robotics, 2016.

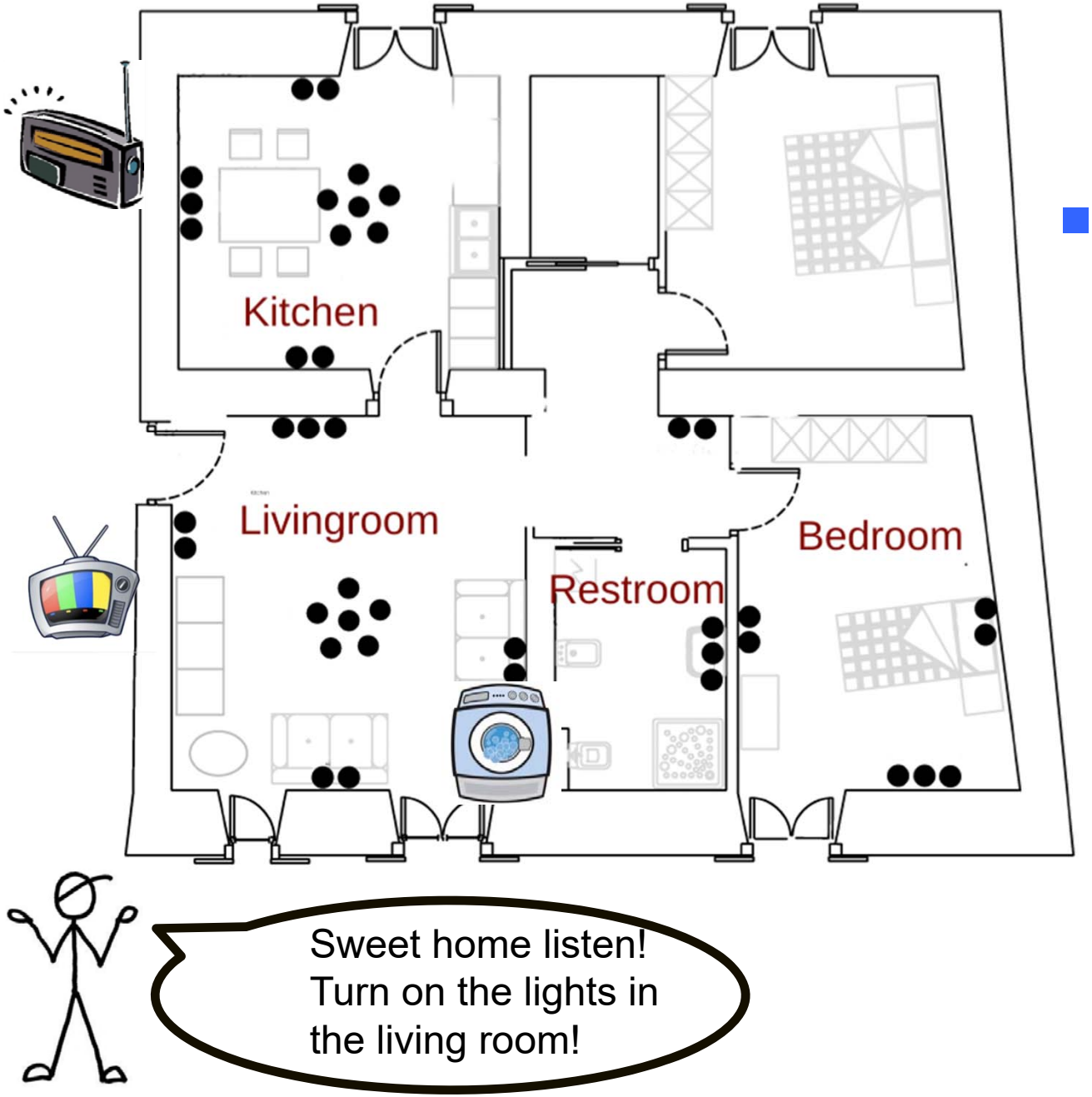
Spoken Command Recognition

Distant Speech Recognition inVoice-enabled Interfaces



https://dirha.fbk.eu/

Smart Home Voice Interface



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

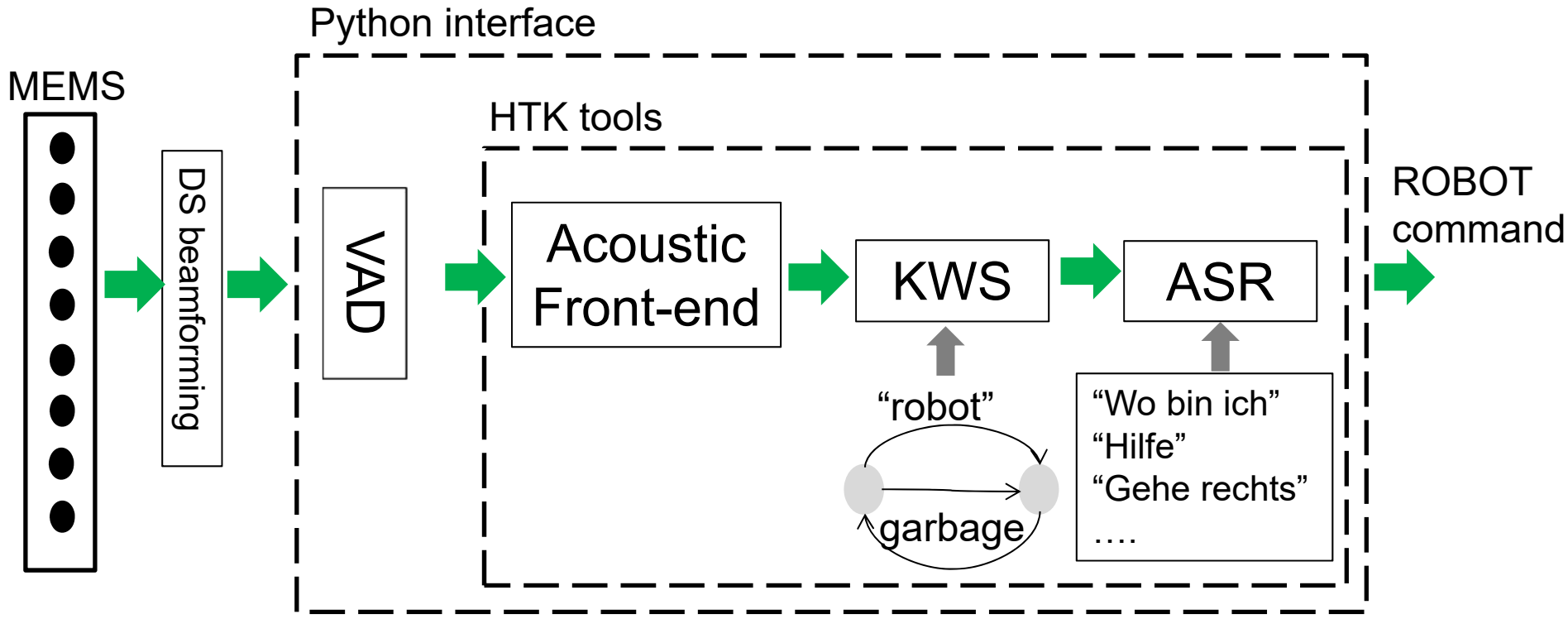
DIRHA demo (“spitaki mou”)

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

- 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.
- A. Tsiami, I. Rodomagoulakis, P. Giannoulis, A. Katsamanis, G. Potamianos and P. Maragos, “ATHENA: A Greek Multi-Sensory Database for Home Automation Control”, INTERSPEECH 2014.

Spoken-Command Recognition Module for HRI

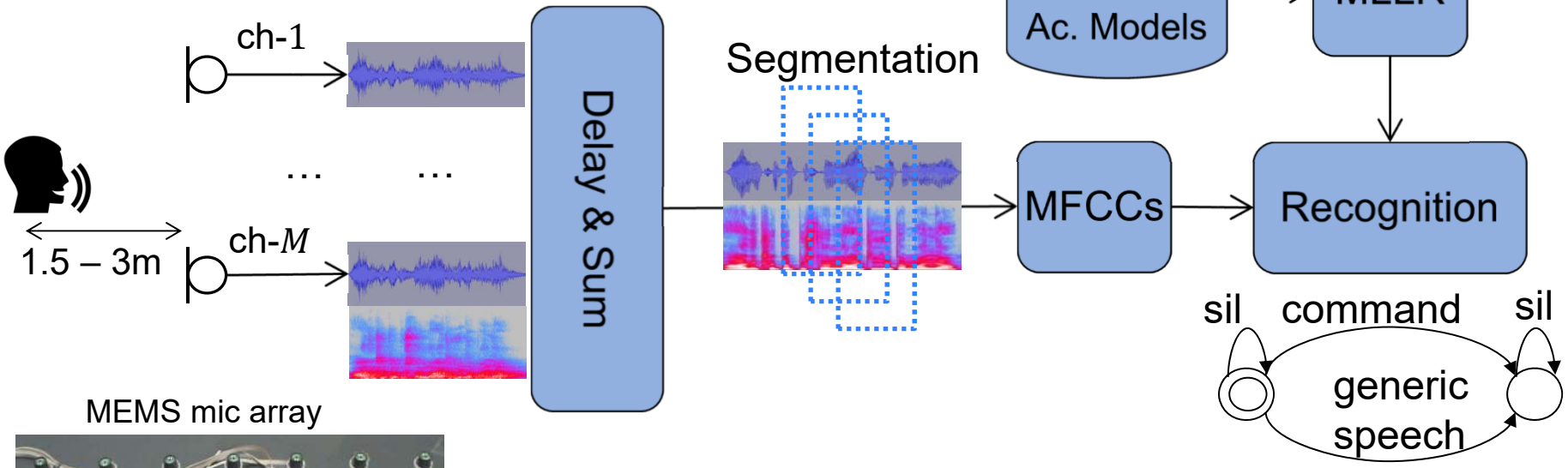
- integrated in ROS, always-listening mode, real time performance



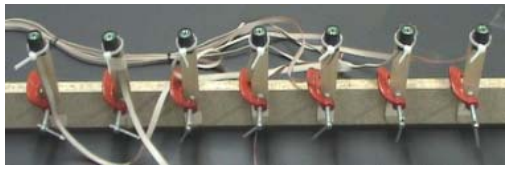
Online Spoken Command Recognition

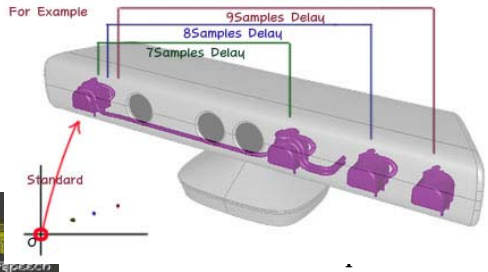
- **Greek, German, Italian, English**

Pentagon ceiling array (Shure)

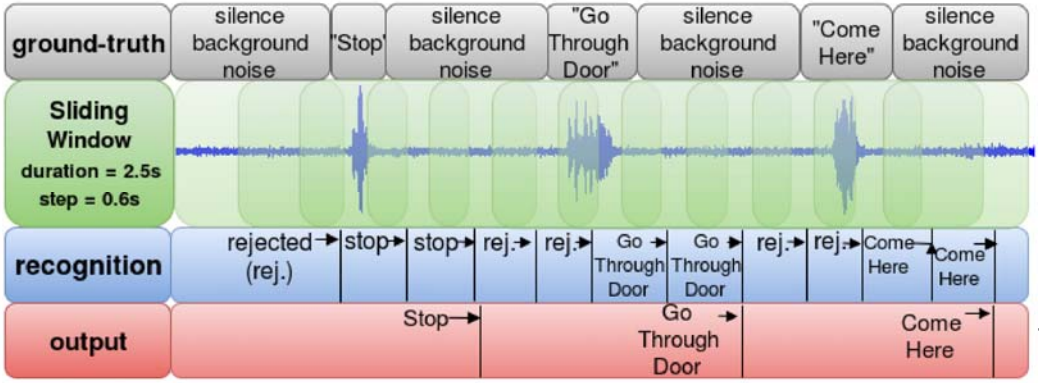


MEMS mic array





Kinect mic array

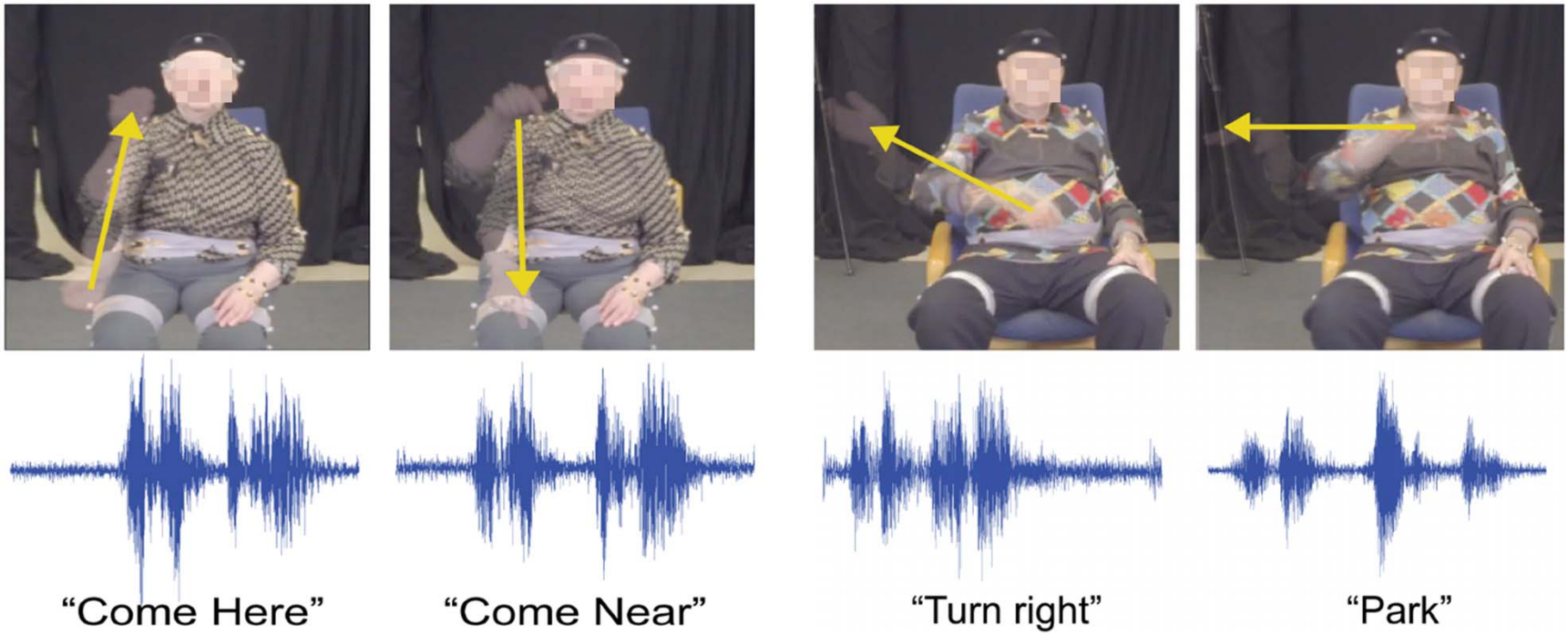


Audio-Visual Fusion for Multimodal Gesture Recognition

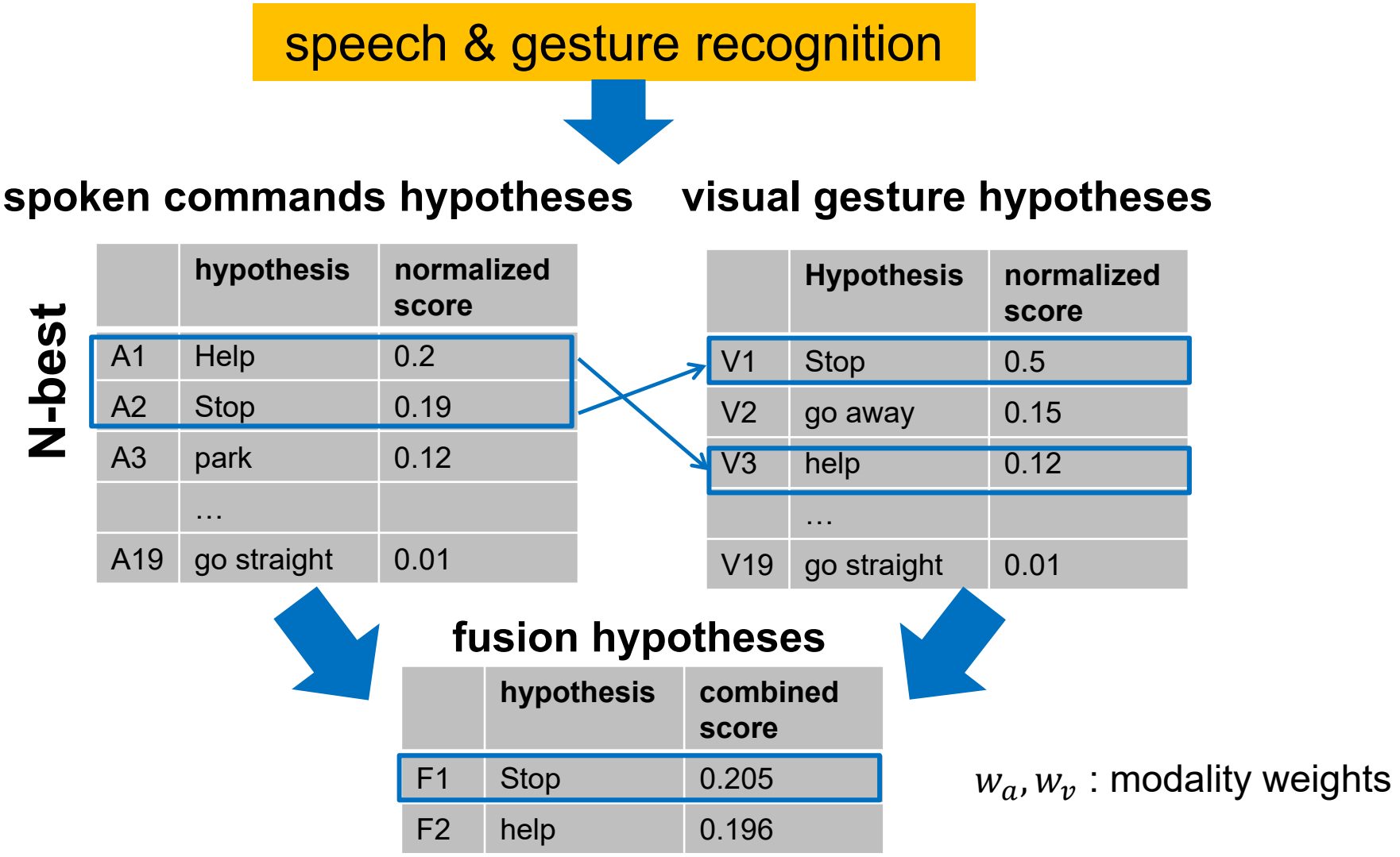
Multimodal Fusion: Complementarity of Visual and Audio Modalities

Similar audio, distinguishable gesture

Distinguishable audio, similar gesture

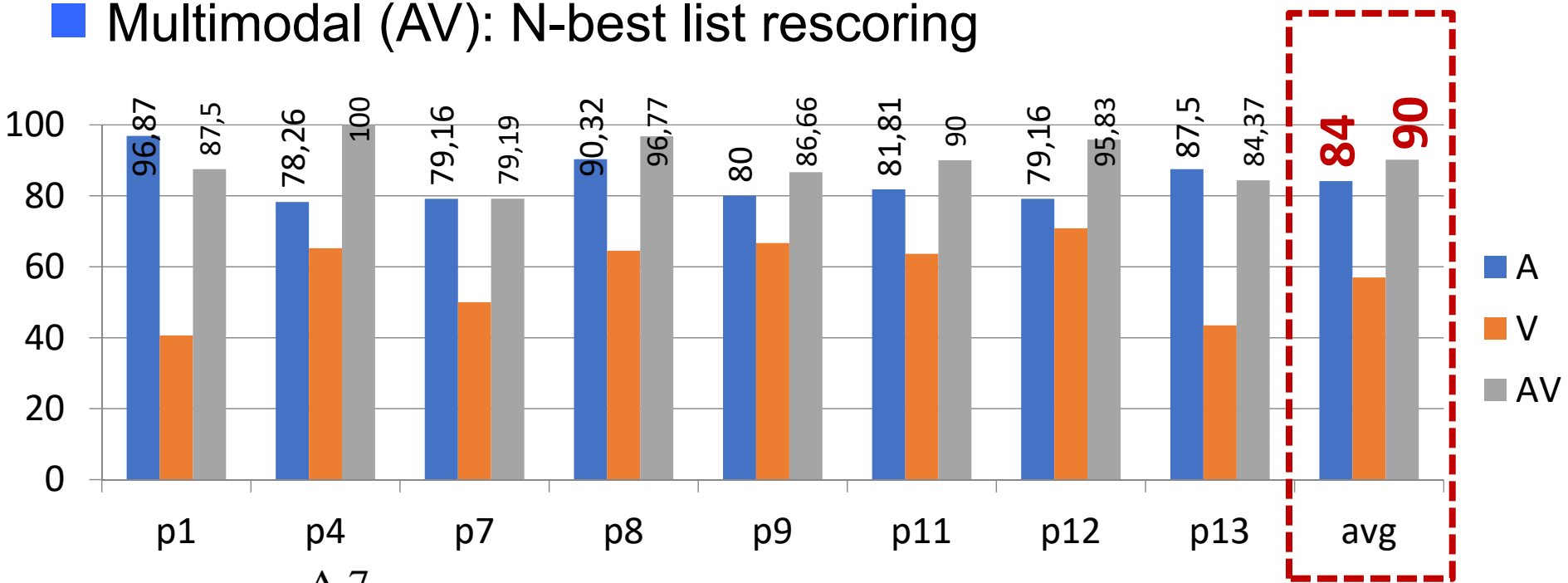


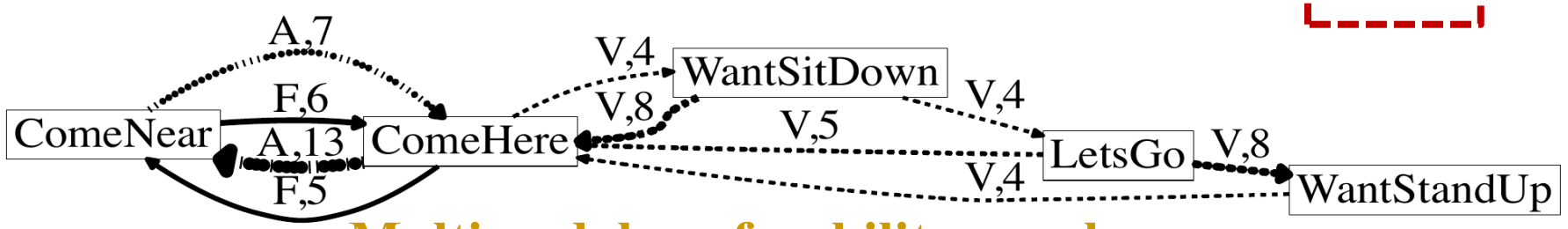
Audio-Visual Fusion: Hypotheses Rescoring



Offline Multimodal Command Classification

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

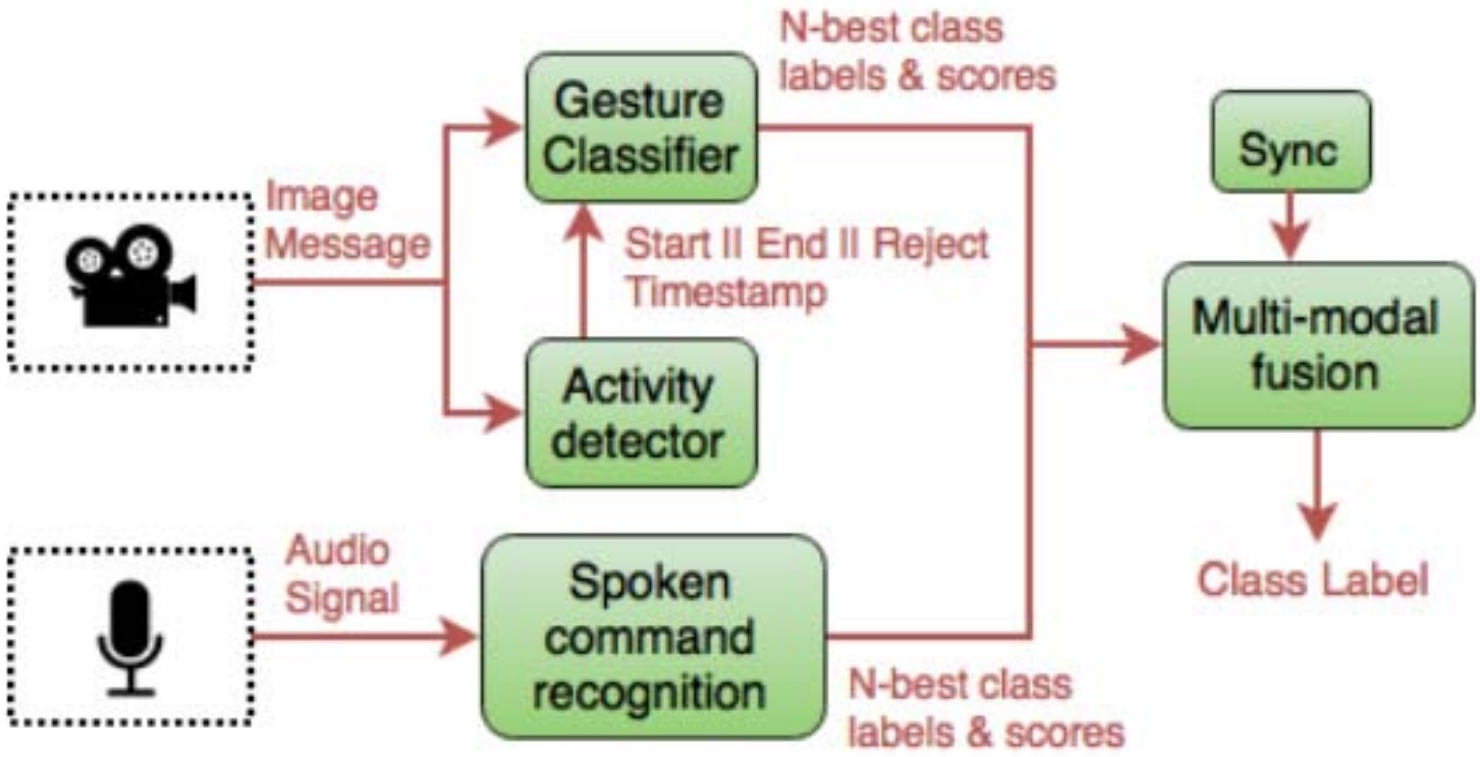




Multimodal confusability graph

HRI Online Multimodal System Architecture

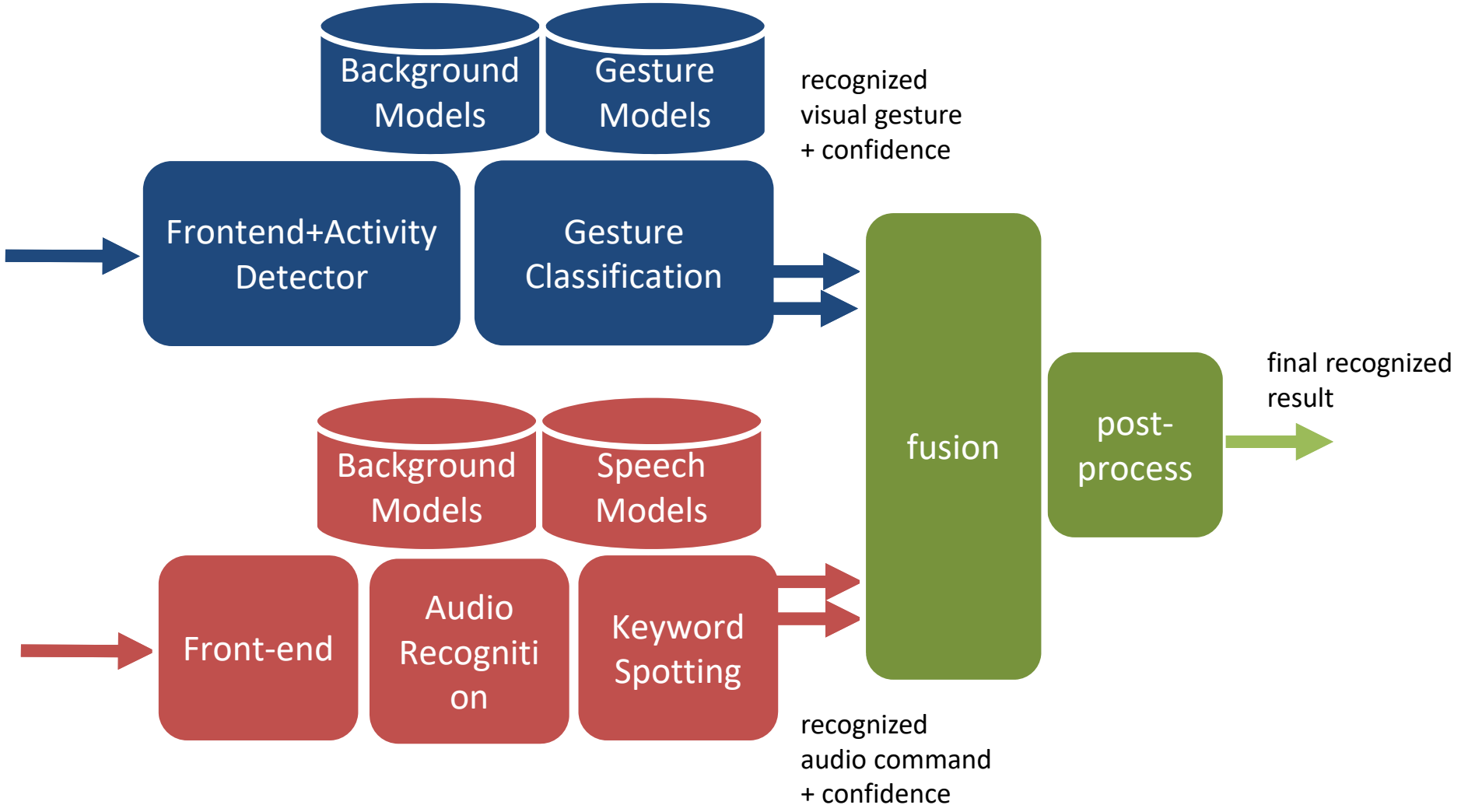
- ROS based integration
  - Spoken command recognition node
  - Activity detection node
  - Gesture classifier node
  - Multimodal fusion node
- Communication using ROS messages



Audio-Gestural Command Recognition

Online processing system – Open Source Software

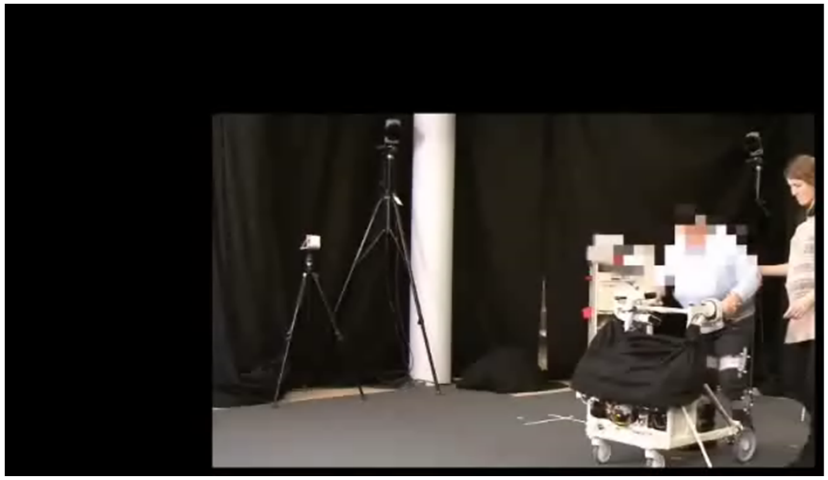
http://robotics.ntua.gr/projects/building-multimodal-interfaces



2B.
Audio-Visual HRI: Applications in Assistive Robotics

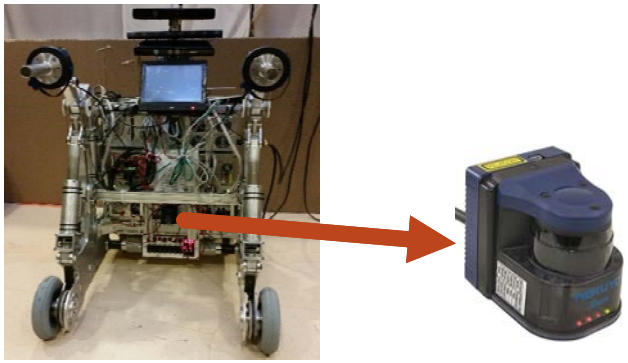
EU Project MOBOT: Motivation

Experiments conducted at
Bethanien Geriatric Center Heidelberg



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

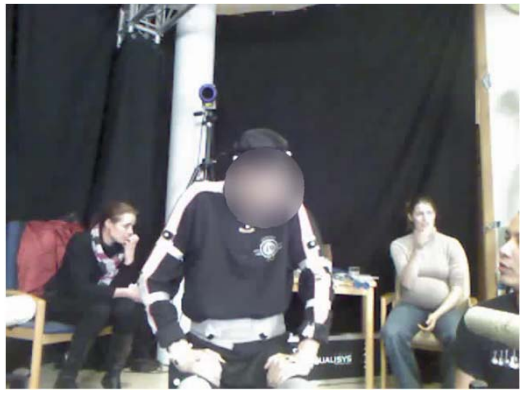
Intelligent assistive devices (robotic Rollator) aiming to provide ***context-aware*** and ***user-adaptive*** mobility (**walking**) assistance



MOBOT rollator

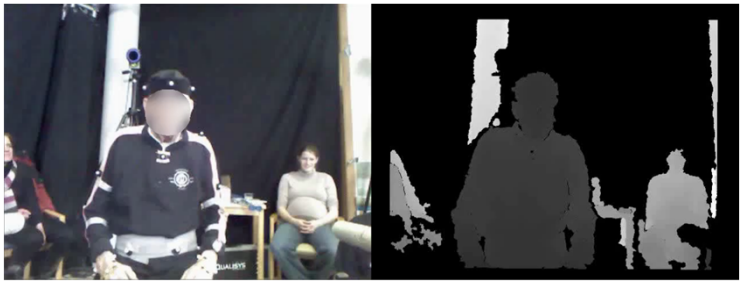
Multi-Sensor Data for HRI

Kinect1 RGB Data



Kinect Depth Data

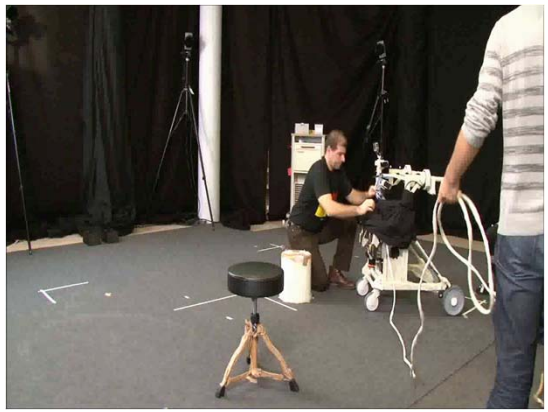
Kinect1 RGB
Kinect1 Depth
MEMS Audio Data



Go Pro RGB Data

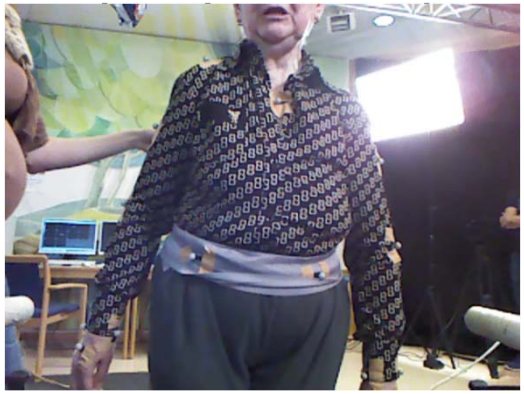
HD1 Camera Data

HD2 Camera Data

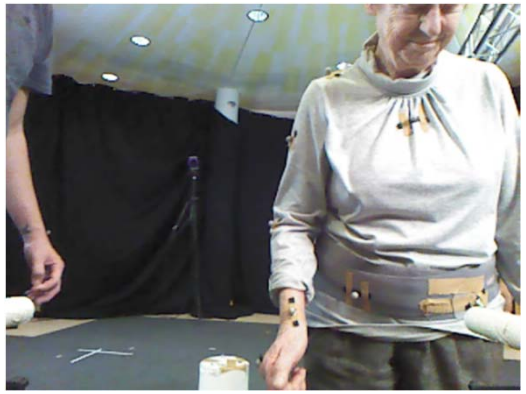


Action Sample Data and Challenges

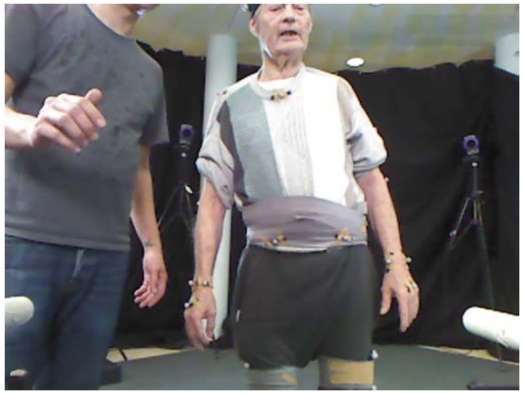
- Visual noise by intruders
- Multiple subjects in the scene, even in same depth level
- Frequent and extreme occlusions, missing body parts (e.g. face)
- Significant variation in subjects pose, actions, visibility,



Stand-to-Sit – P1



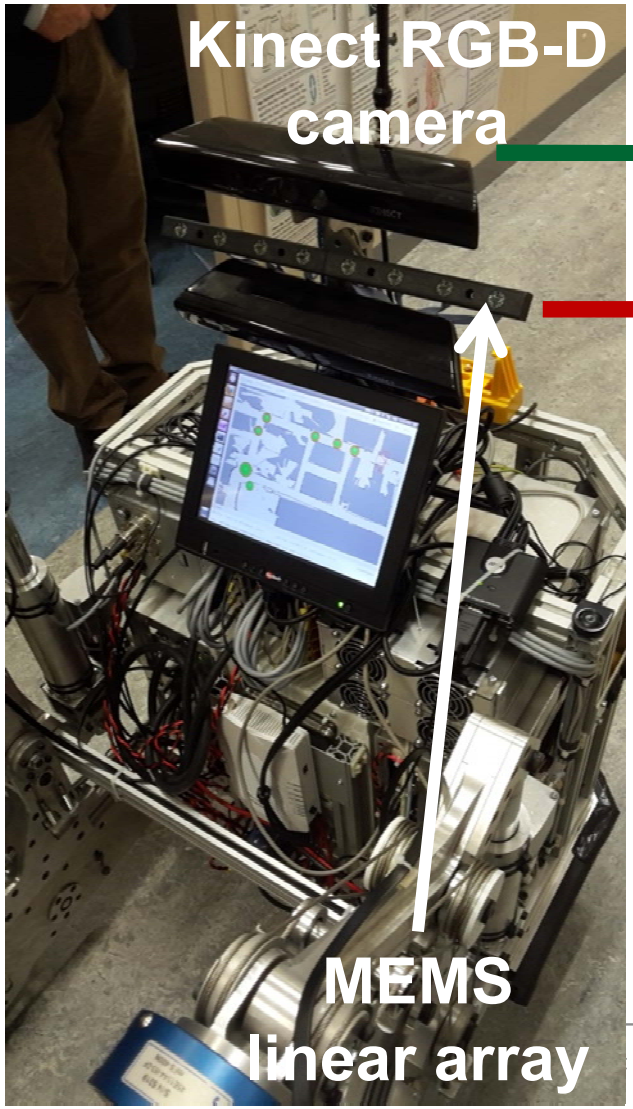
Stand-to-Sit – P3



Stand-to-Sit – P4

Audio-Gestural Command Recognition: Overview of our Multimodal Interface

MOBOT robotic platform



Visual Action-gesture Recognition

Spoken Command Recognition

N-best hypotheses & scores

Multimodal Late Fusion

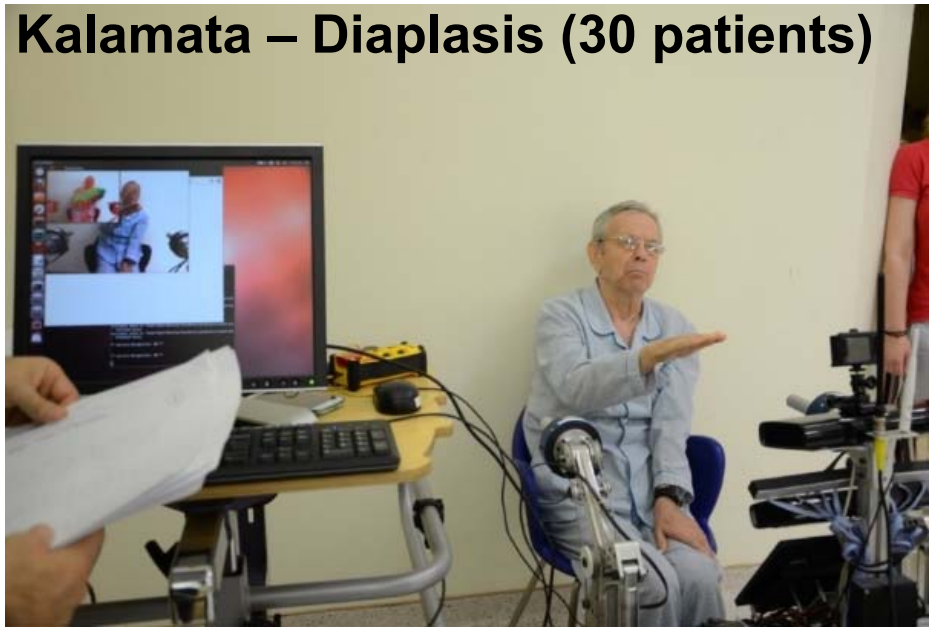
Best AV Hypothesis

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

Clinical Studies (MOBOT)

Heidelberg – Bethanien (19 patients)

Kalamata – Diaplasis (30 patients)

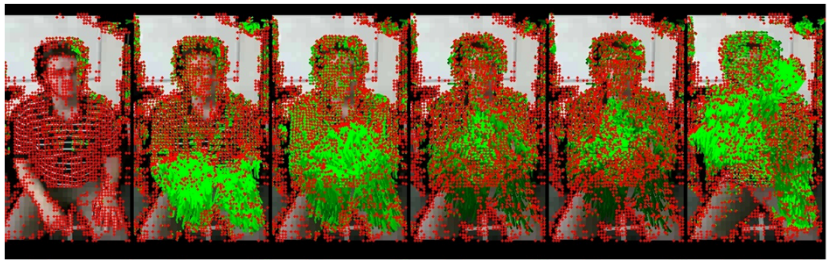


Speech, Gestures, Combination: 3 repetitions of 5 commands

Validation experiments (Bethanien, Heidelberg):

EU Project I-SUPPORT: Overview (Gesture & Spoken Command Recognition)

dense trajectories of visual motion



Audio-Gestural Recognition: Validation Experiments

(FSL, Rome)

Validation Setup

FSL, Rome

Bethanien, Heidelberg

Gesture Recognition

Challenges

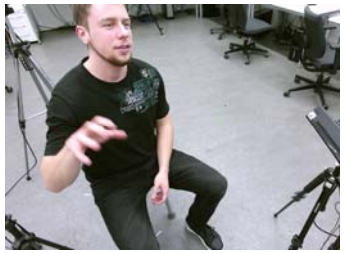
Different viewpoints

Poor gesture performance

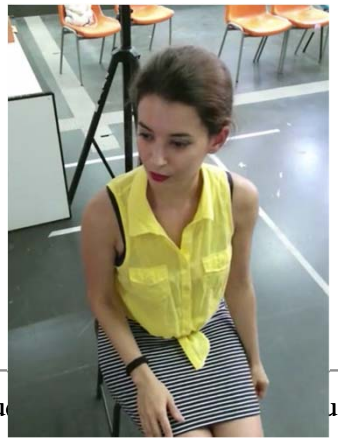
Random movements

Data collection

KIT



ICCS - NTUA

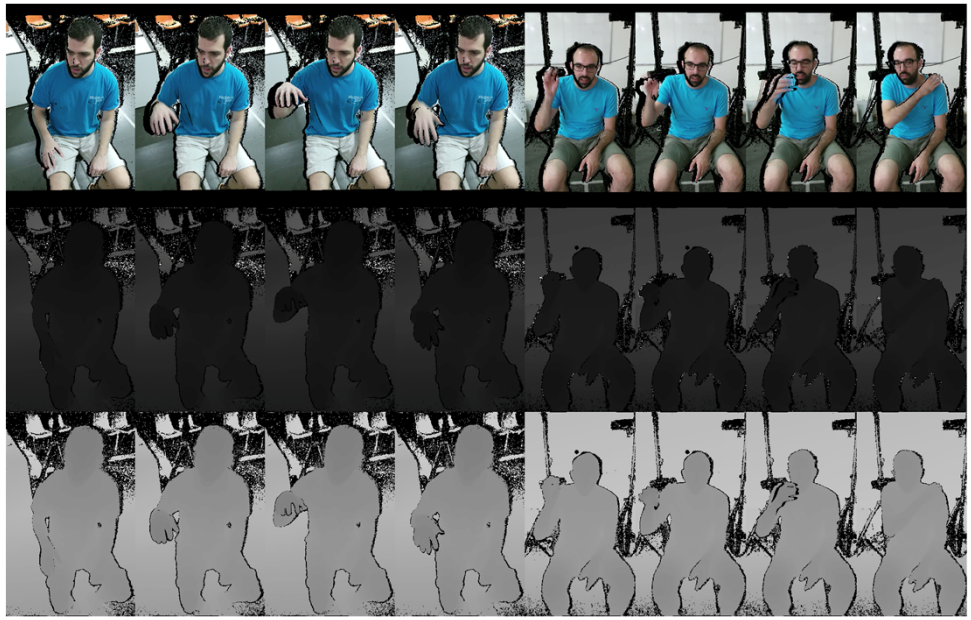


Pre-Validation
FSL - Bethanien

Gesture Recognition – Depth Modality

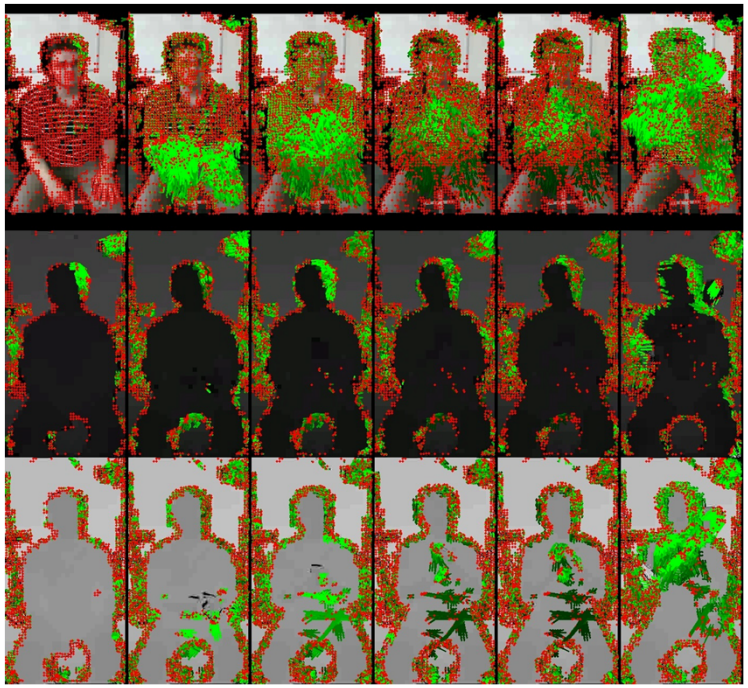
- Experiments with Depth and Log-Depth streams
- Extraction of Dense Trajectories performs better on the Log-Depth stream

RGB stream



Log-Depth stream

Dense Trajectories



Gesture Offline Classification – Results

- **ICCS Dataset** (24u, 28g)
  - Two different setups
  - Two different streams
  - Different encoding methods
  - Different features

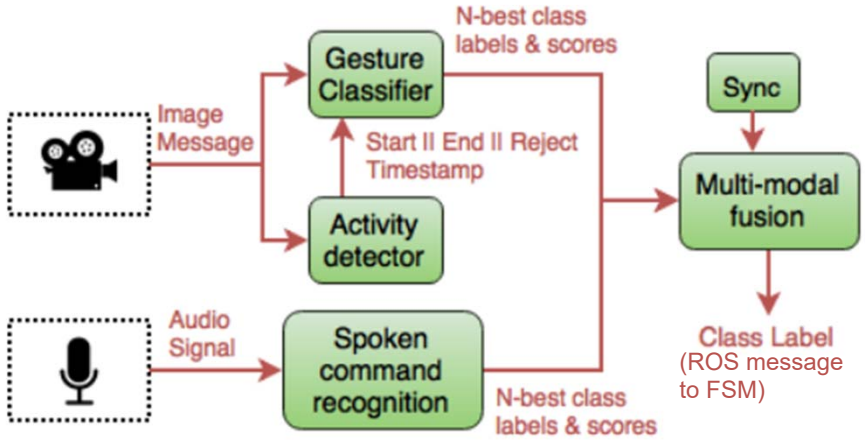
		Task: Legs		Task: Back	
Feat.	Encoding	RGB	D	RGB	D
Traj.	BoVW	69.64	60.52	77.84	60.87
HOG		41.01	53.34	58.51	57.14
HOF		74.15	66.26	82.92	71.58
MBH		77.36	65.31	80.81	65.73
Comb.		**80.88**	**74.41**	**83.92**	**75.70**
Traj.	VLAD	69.22	52.66	74.34	54.14
HOG		49.86	65.99	61.23	65.63
HOF		76.54	72.88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

- **KIT Dataset (**8u, 8/10g)
  - Two different setups
  - Average gesture recognition accuracy:
    - Legs (8 gestures): 83%
    - Back (10 gestures): 75%
- **FSL Pre-Validation Dataset** (5u, 10g)
  - 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%

Multimodal Fusion and On-line Integration

- Multimodal “late” fusion (Validation @ Bethanien, Heidelberg)

- ROS (Robot Operating System) based integration



Validation results

Command Recognition Rate (CRR)

(= accuracy only on **well** performed commands)

Bethanien, Heidelberg

Round 1 (no training, audio-gestural scenario)	
Back	73.8% (A)*
Legs	84.7%

Round 2 (“back” position)		
	Gesture-only scenario	**Audio-Gestural Scenario**
Without training	59.6%	86.2%
With training	68.7%	79.1%

FSL, Rome

Round 1 (no training, audio-gestural scenario)	
Back	87.2%
Legs	79.5%

Round 2 (no training, audio-gestural scenario, “legs” position)
83.5%

I-SUPPORT system video

Part 2: Conclusions

- **Synopsis**:
  - Multimodal Action Recognition and Human-Robot Interaction
    - Visual Action Recognition
    - Gesture Recognition
    - Spoken Command Recognition
    - Online Multimodal System and Applications in Assistive Robotics

- **Ongoing work**:
  - Fuse Human Localization & Pose with Activity Recognition
  - Activities: Actions – Gestures – SpokenCommands - Gait
  - Applications in Perception and Robotics

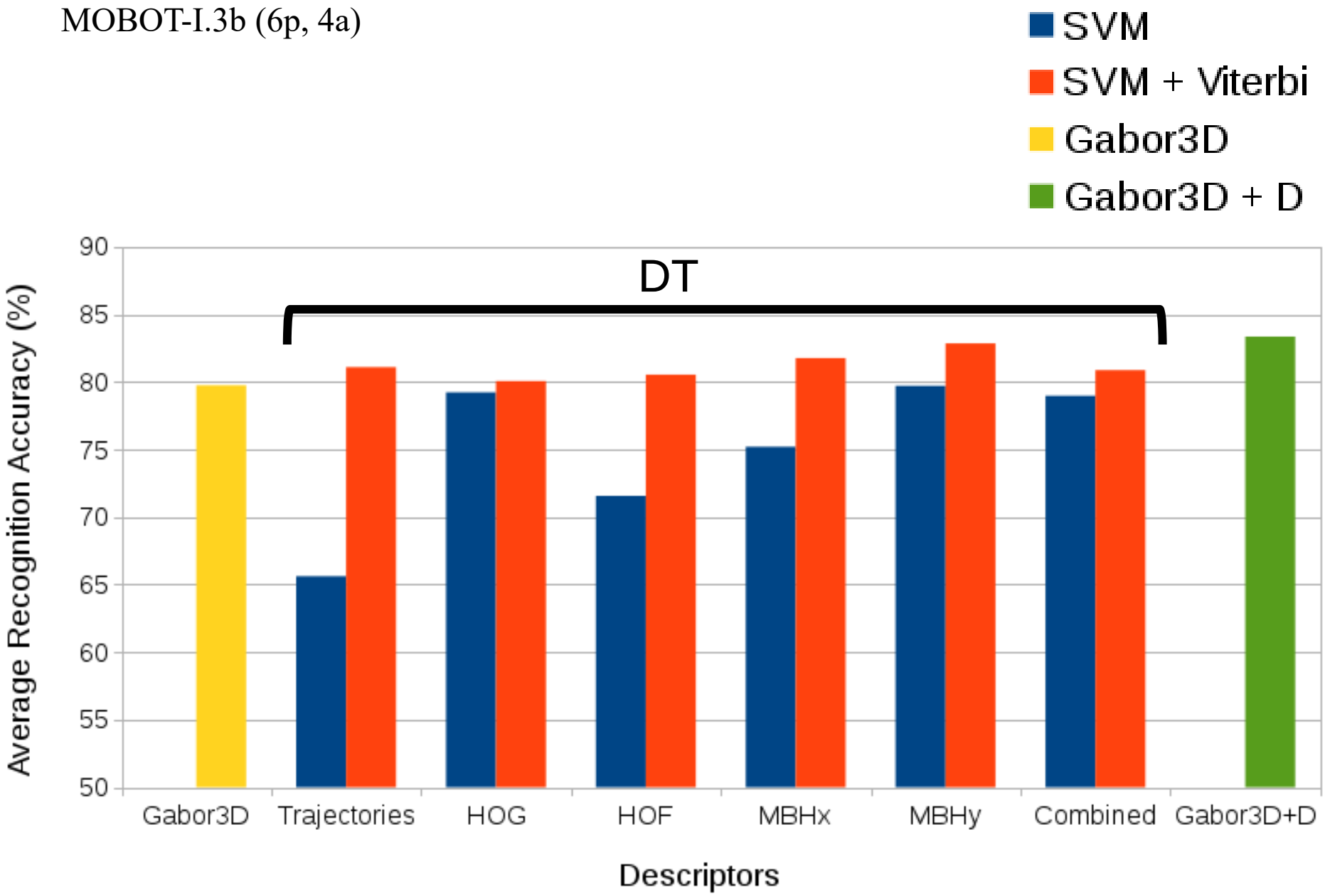
Tutorial slides: http://cvsp.cs.ntua.gr/interspeech2018

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

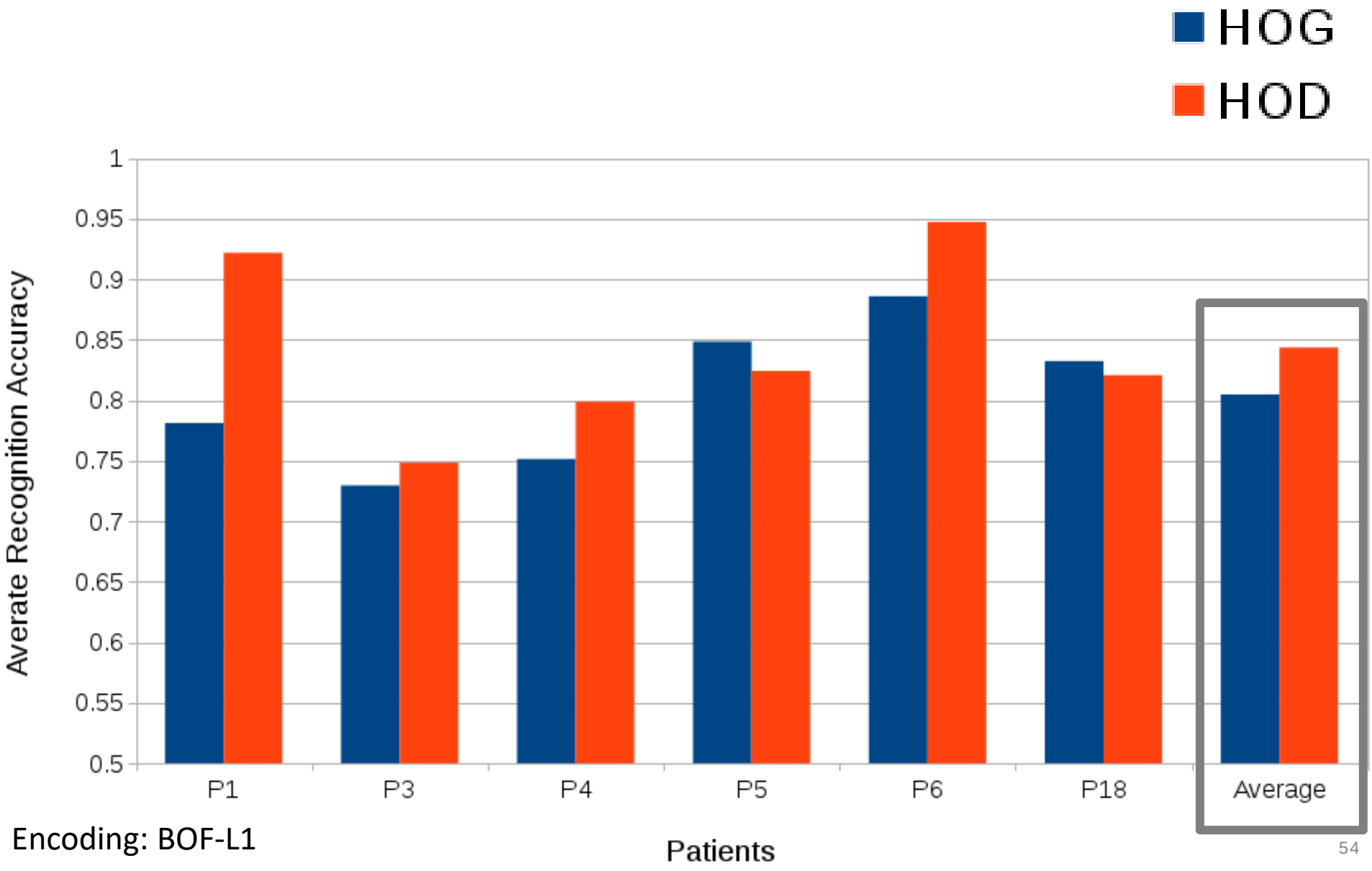
APPENDIX

More Action Recognition results: +Gabor3D+Depth

MOBOT-I.3b (6p, 4a)

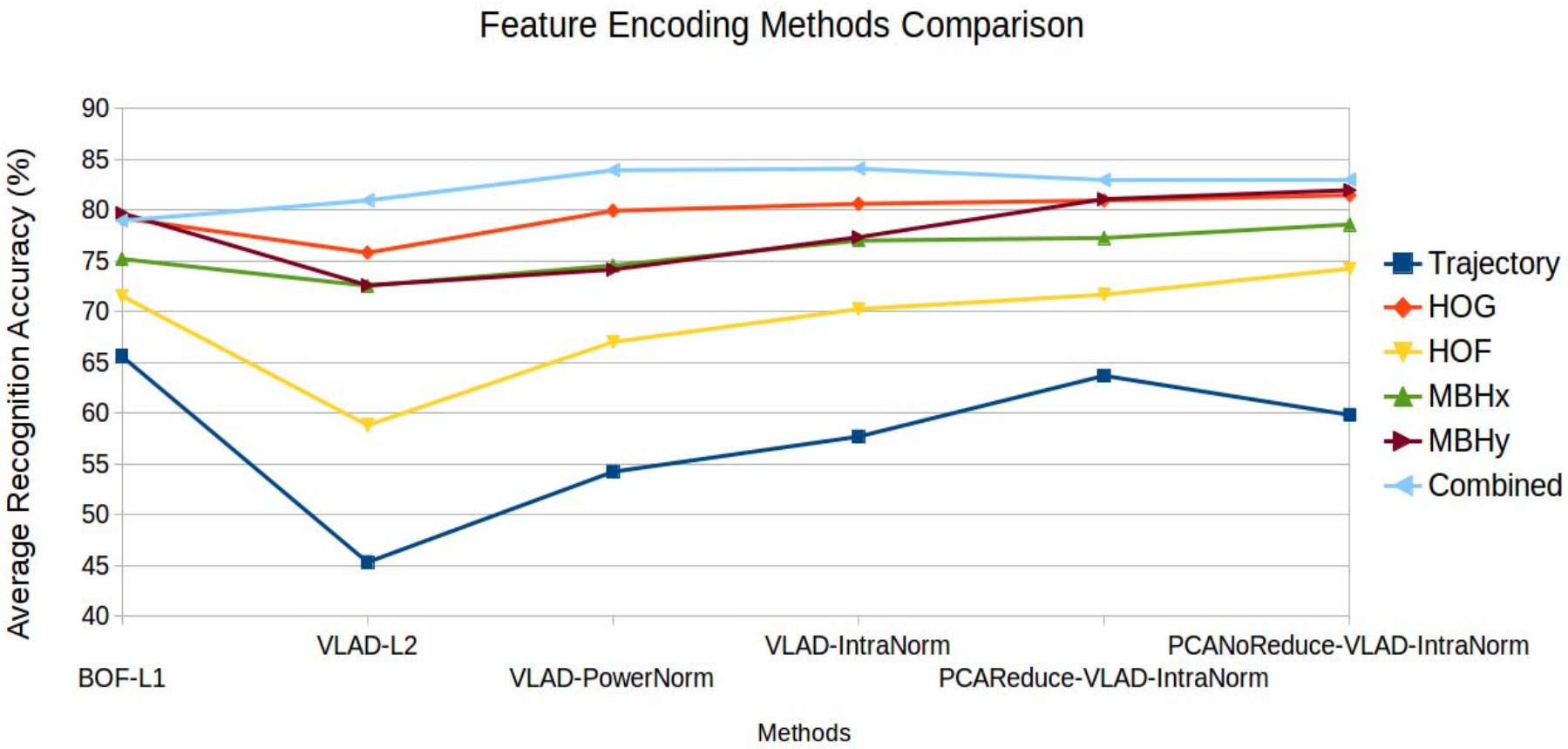


Action Recognition results: Comparison of HOG (RGB) vs HOD (Depth)



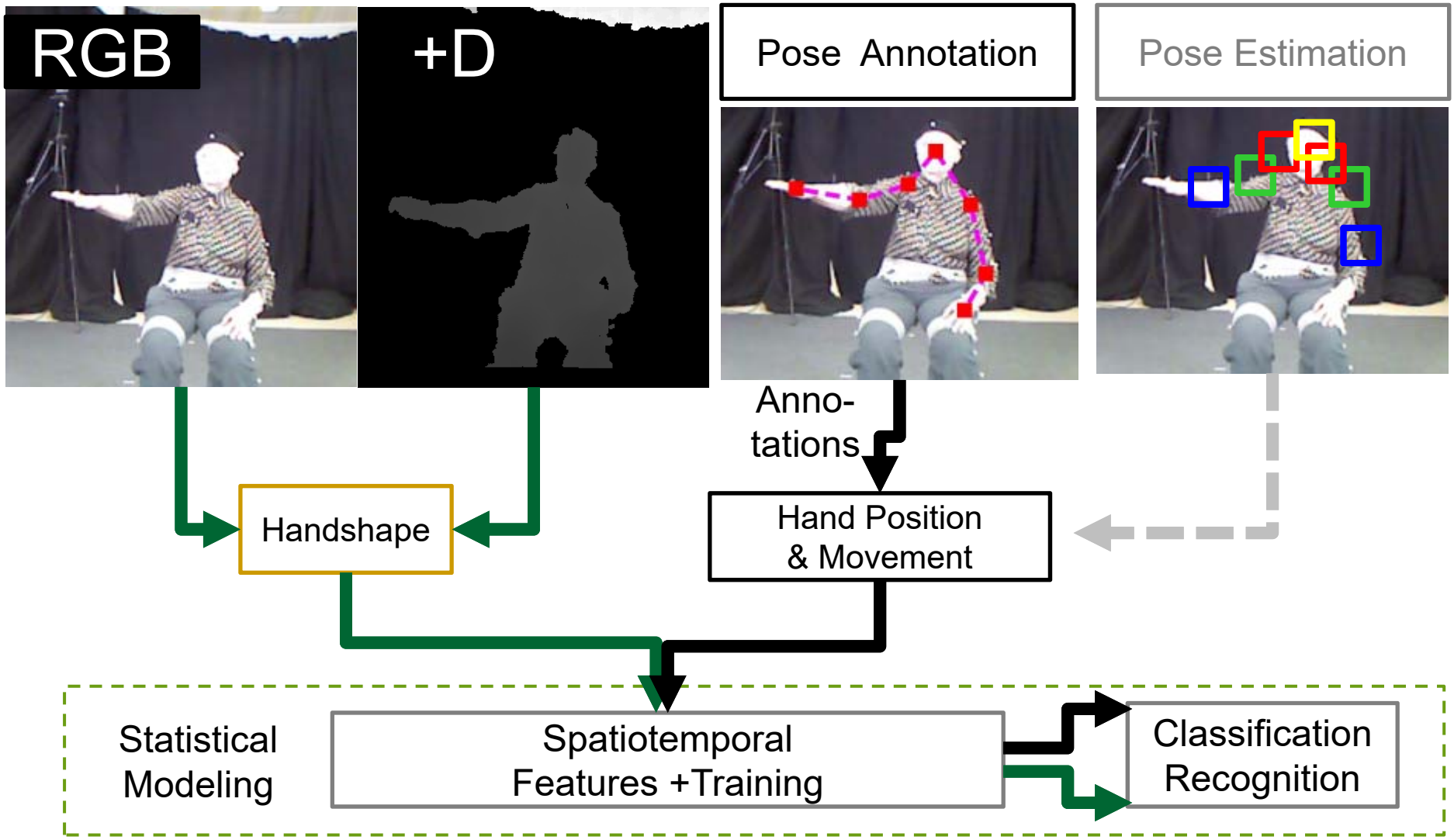
Encoding: BOF-L1

VLAD vs BOF comparison



MOBOT-I Scenario 3b (3 actions + BM, 6 patients).

Overview: Visual Gesture Recognition



GoPro Camera Data



"Help"





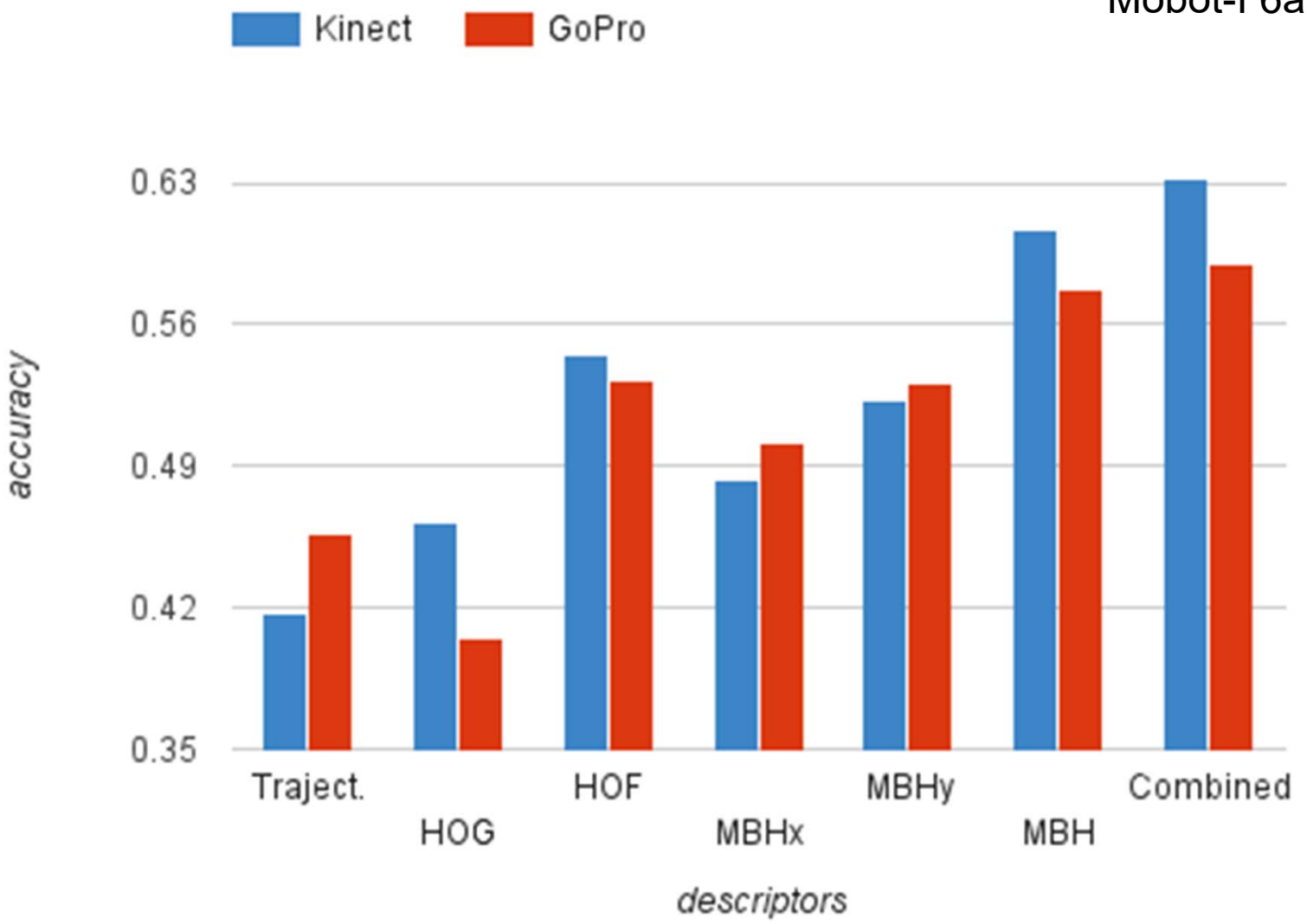
"I want to stand up"



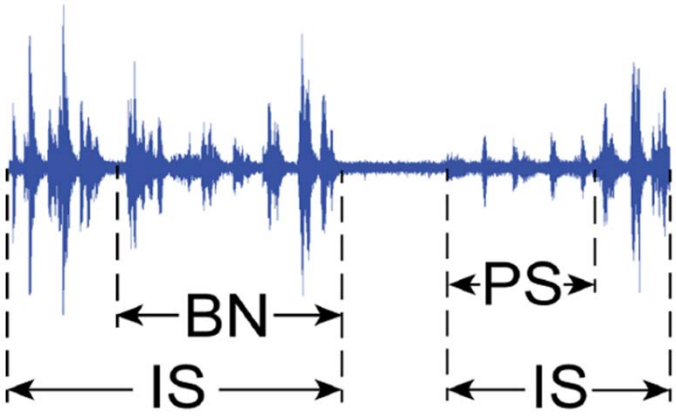


Gesture Classification results on GoPro data

Mobot-I 6a (8p, 8g)



Multimodal Gesture classification on Mobot 6.a dataset



- Task 6a
  - User is sitting & gesturing
  - 13 patients
- GoPro videos & MEMS audio
  - aligned with annotations
- Noisy audio-visual scenes
- Different speech & gesture pronunciations

IS: Instructor speaking
PS: Patient speaking
BN: Background Noise