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)

Multimodal Processing and Learning with Applications in Audio-Visual Perception and Understanding

Petros Maragos

IEEE SPS DL Talk, Feb. 2018, Google Research, Los Angeles

Talk Outline

- Audio-Visual Perception and Fusion
- Applic 1: A-V-T Saliency & Video Summarization
- Applic 2: Multimodal Gesture Recognition and Human-Robot Interaction

Human versus Computer Multimodal Processing

- Nature is abundant with multimodal stimuli.
- Digital technology creates a rapid explosion of multimedia data.
- Humans perceive world multimodally in a seemingly effortless way, although the brain dedicates vast resources to these tasks.
- Computer techniques still lag humans in understanding complex multisensory scenes and performing high-level cognitive tasks. **Limitations**: inborn (e.g. data complexity, voluminous, multimodality, multiple temporal rates, asynchrony), inadequate approaches (e.g. monomodal-biased), non-optimal fusion.
- **Goal**: *develop truly multimodal approaches that integrate several modalities toward improving robustness and performance for anthropo-centric multimedia understanding and applications.*

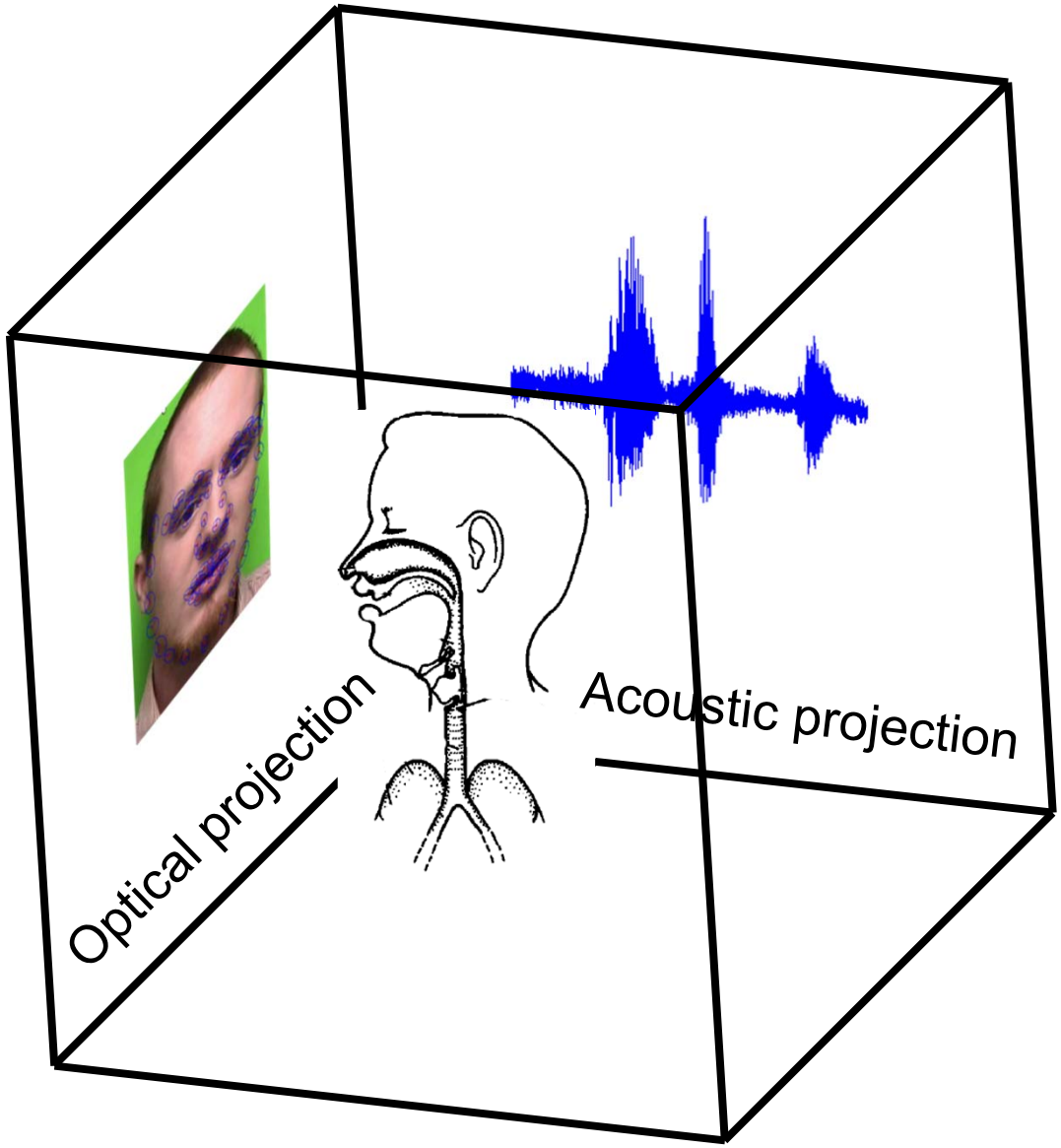
Multimodal Data Challenges

- **Data are Voluminous**:
  - 24 hrs of TV = 430 GB = ~2 millions still (frame) images
  - WWW: 300-hr videos are uploaded on YouTube per minute.
  - 300 millions images are uploaded on FaceBook per day.
  - Kinect sensor: 250 MB/sec uncompressed RGB
- **Data are Dynamic**
  - Temporal video, Website updating, News quickly get obsolete
- **Different Temporal Rates**
  - Video: 25-30 frames /second
  - Audio: 44000 sound samples/sec,
  - Speech: 100 feature-frames/sec, 4 syllables/sec
- **Cross-Media asynchrony**
  - image and audio scene boundaries are different

1. Audio-Visual Perception and Fusion

Perception: the sensory-based inference about the world state

Speech: Multi-faceted phenomenon



McGurk effect example

- [ba – audio] + [ga – visual] → [da] (fusion)
- [ga – audio] + [ba – visual] → [gabga, bagba, baga, gaba] (combination)
- Speech perception seems to also take into consideration the visual information. Audio-only theories of speech are inadequate to explain the above phenomena.
- Audiovisual presentations of speech create fusion or combination of modalities.
- One possible explanation: *a human attempts to find common or close information in both modalities and achieve a unifying percept.*

Multicue or Multimodal Perception Research

- ***McGurk effect**: Hearing Lips and Seeing Voices* [McGurk & MacDonald 1976]

- *Modeling Depth Cue Combination using Modified Weak Fusion* [Landy et al. 1995]
  - scene depth reconstruction from multiple cues: motion, stereo, texture and shading.
- *Intramodal Versus Intermodal Fusion of Sensory Information* [Hillis et al. 2002]
  - shape surface perception: intramodal (stereopsis & texture), intermodal (vision & haptics)

- ***Integration of Visual and Auditory Information for Spatial Localization***
  - Ventriloquism effect
  - Enhance selective listening by illusory mislocation of speech sounds due to lip-reading [Driver 1996]
  - Visual capture [Battaglia et al. 2003]
  - Unifying multisensory signals across time and space [Wallace et al. 2004]

- *AudioVisual Gestalts* [Monaci & Vandergheynst 2006]
  - temporal proximity between audiovisual events using Helmholtz principle

- ***Temporal Segmentation of Videos into Perceptual Events by Humans*** [Zacks et al. 2001]
  - humans watching short videos of daily activities while acquiring brain images with fMRI

- *Temporal Perception of Multimodal Stimuli* [Vatakis and Spence 2006]

Perceptual Aspects of Multisensory Processing

Multisensory Integration: unisensory auditory and visual signals are combined forming a new, unified audiovisual percept.

Goal: *Perceiving Synchronous and Unified Multisensory Events*

Principles: Multisensory integration is governed by the following rules:

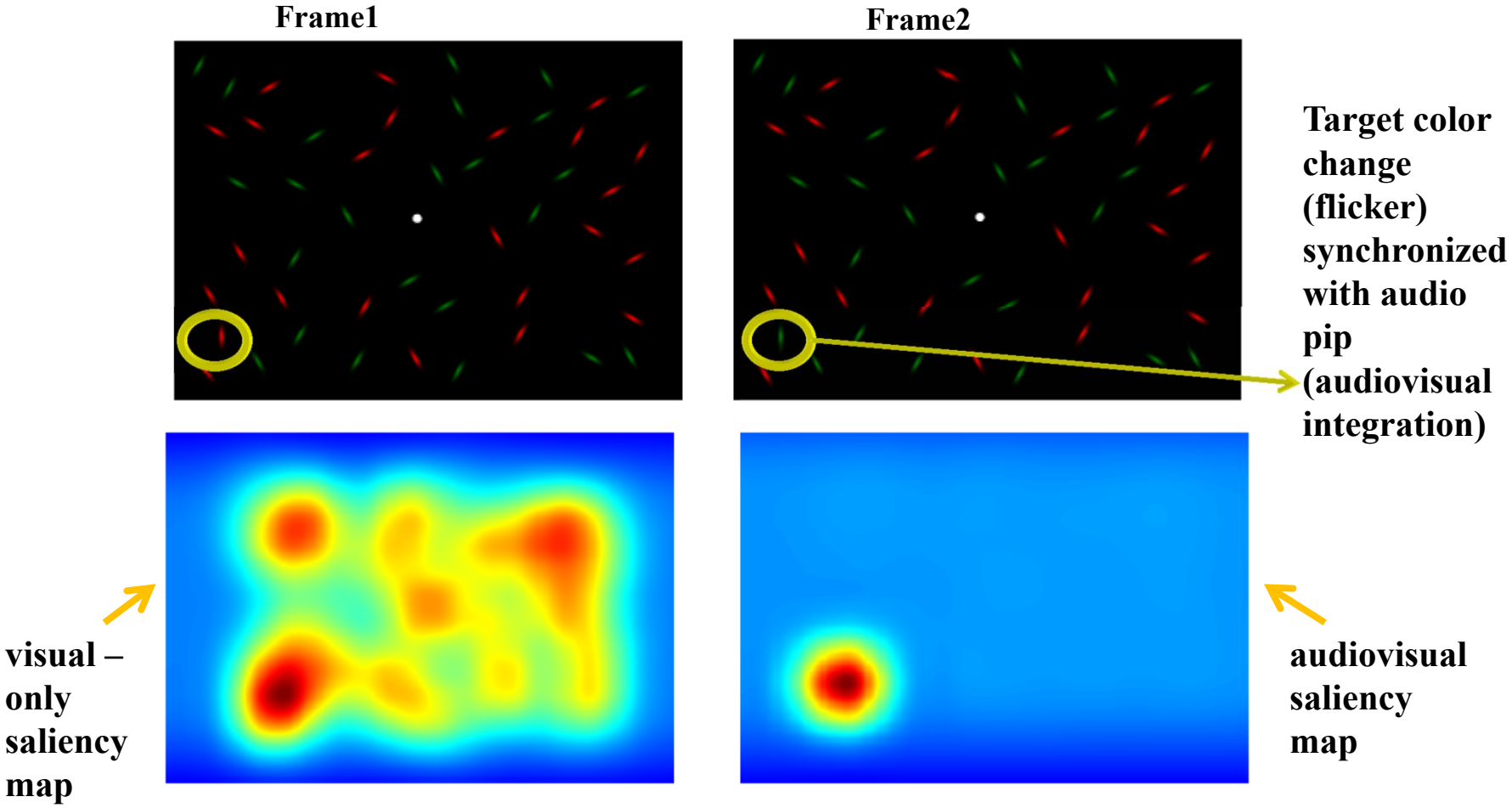
- **Spatial rule**,
- **Temporal rule**,
- **Modality Appropriateness**:
  - Visual dominance of spatial tasks.
  - Audition is dominant for temporal tasks.
- **Inverse effectiveness law**:
  - In multisensory neurons, multimodal stimuli occurring in close space-time proximity evoke supra-additive responses. The less effective monomodal stimuli are in generating a neuronal response, the greater relative percentage of multisensory enhancement.
  - Is this the case for behavior? Recent experiments indicate that inverse effectiveness accounts for some behavioral data.

Synchrony and **Semantics** are two factors (structural and cognitive) that appear to favor the binding of multisensory stimuli, yielding a coherent unified percept. Strong binding, in turn, leads to higher stream asynchrony tolerance.

[E. Tsilionis and A. Vatakis, "*Multisensory Binding: Is the contribution of synchrony and semantic congruency obligatory*?", COBS 2016.]

Computational audiovisual saliency model

- Combining audio and visual saliency models by proper fusion
- Validated via behavioral experiments, such as pip & pop:



[A. Tsiami, A. Katsamanis, P. Maragos & A. Vatakis, ICASSP 2016]

Bayesian Formulation of Perception

$$P(S|D) = \frac{P(D|S)P(S)}{P(D)}$$

S : configuration of auditory and/or visual scene of world
D : mono/multi-modal data or features.

P(S): Prior Distribution, *P(D/S)*: Likelihood, *P(D)*: Evidence

P(S/D): Posterior conditional distribution

S → D : World-to-Signal mapping

Perception is an ill-posed inverse problem

$$\hat{S}_{MAP} = \underset{S}{\operatorname{argmax}}\, P(D|S)P(S)$$

Models for Multimodal Data Integration

Levels of Integration:

- *Early* integration (as in strong fusion)
- *Intermediate* integration
- *Late* integration (as in weak fusion)

Time dimension:

- *Static*: CCA- Canonical Correlation Analysis: e.g. "cocktail-party effect"
  Max Mutual Information
  SVMs- Support Vector Machines: kernel combination

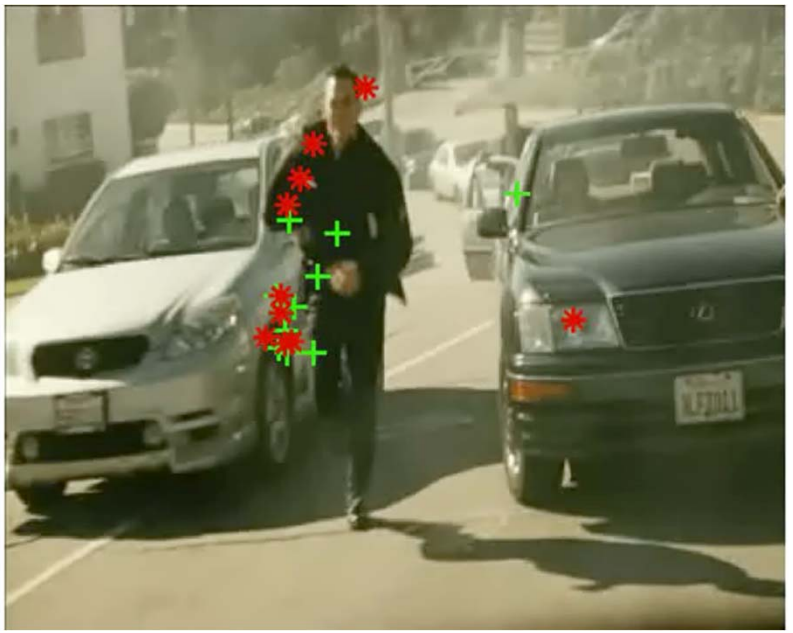
- *Dynamic*: HMMs (Hidden Markov Models)
  DBNs (Dynamic Bayesian Nets)
  DNNs (Deep Neural Nets)
  Multimodal Hypothesis Rescoring

2. Audio-Visual-Text Saliency and Video Summarization

Spatio-Temporal Visual Saliency Representations for Video Summarization

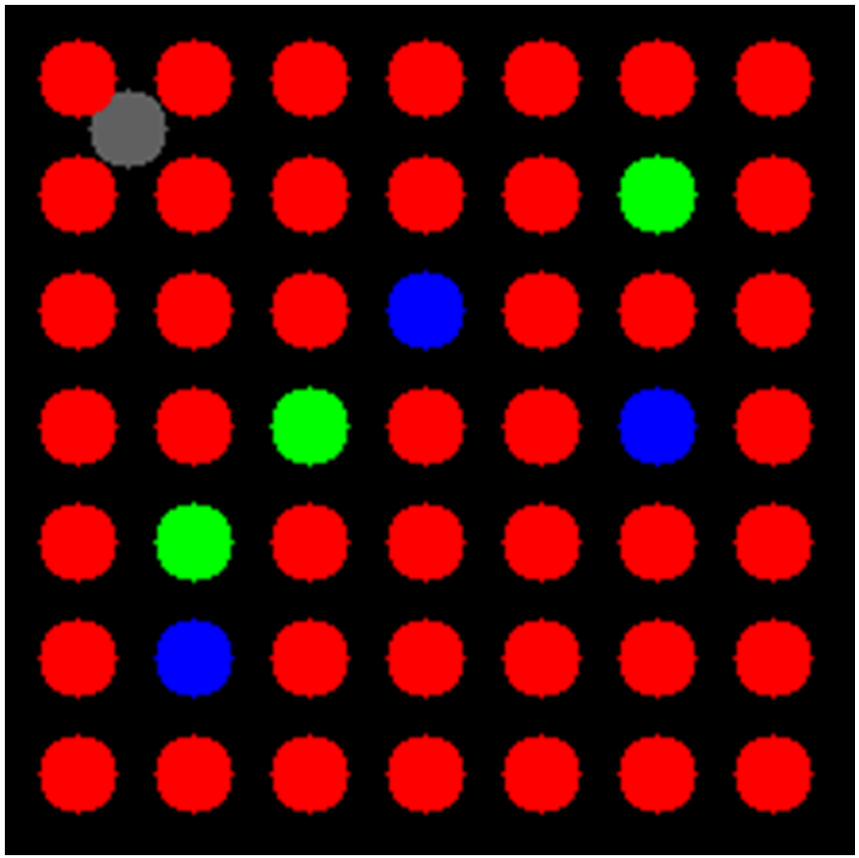
Introduction

- Visual Attention
  - Top-down, Task-driven
  - High level topics
- Visual Saliency
  - Bottom-up, Data-Driven
  - Low level sensory cues

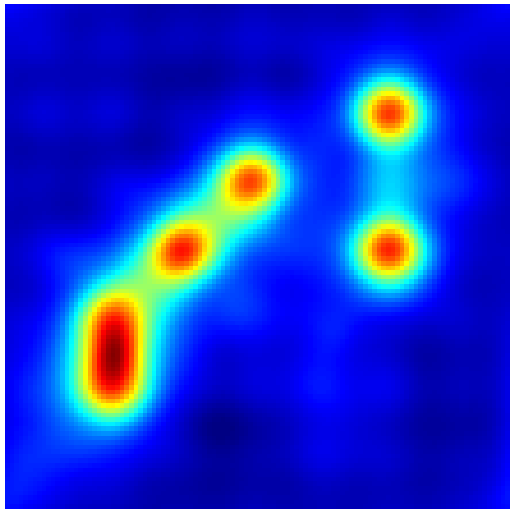


- Applications
  - Systems for selecting the most important regions of a large amount of visual data
  - Movie Summarization
  - Visual Frontend for other applications.

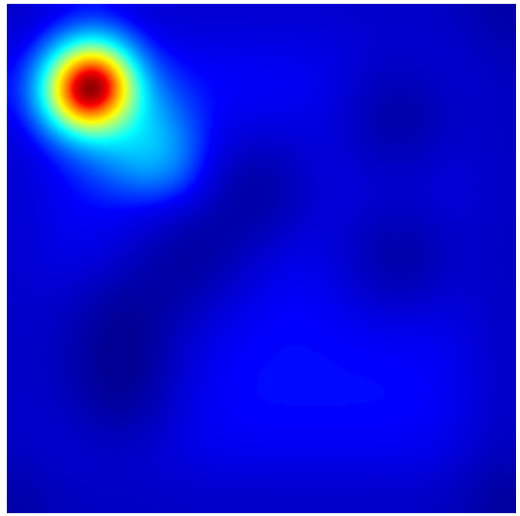
Why Spatio-Temporal Saliency?



AWS



Spatio-Temporal Energy



Visual Saliency Estimation

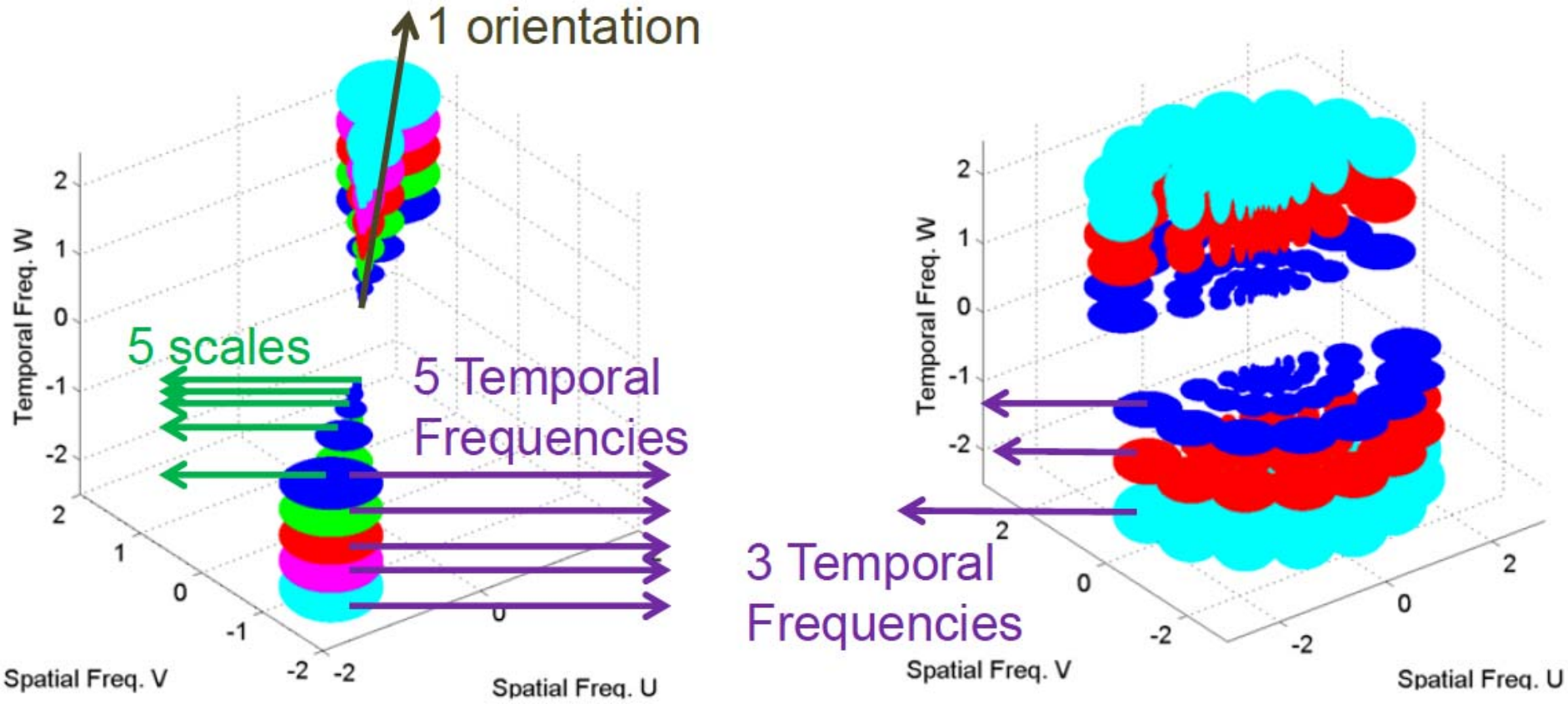
- **Visual Saliency: Approaches, Measurements, Applications**:
  - Predict Viewers' Fixations both in space and time
  - Detect Salient Objects
  - Framewise saliency: find the frames that are more salient than others

- **Spatio-Temporal Frontend for Visual Saliency**
  - Relevant to the cognition-inspired saliency methods, based on Koch & Ullman theory.
  - Uses biologically plausible spatio-temporal filters, like oriented 3D Gabor filters, in order to extract visual features (intensity, color, motion).
  - Detects both the fastest changes in the video stimuli (e.g. flicker) and the slowest motion changes related to action events.

[Koutras and Maragos, *"Spatio-Temporal Visual Saliency"*, *SP: Image Communication*, 2015.]

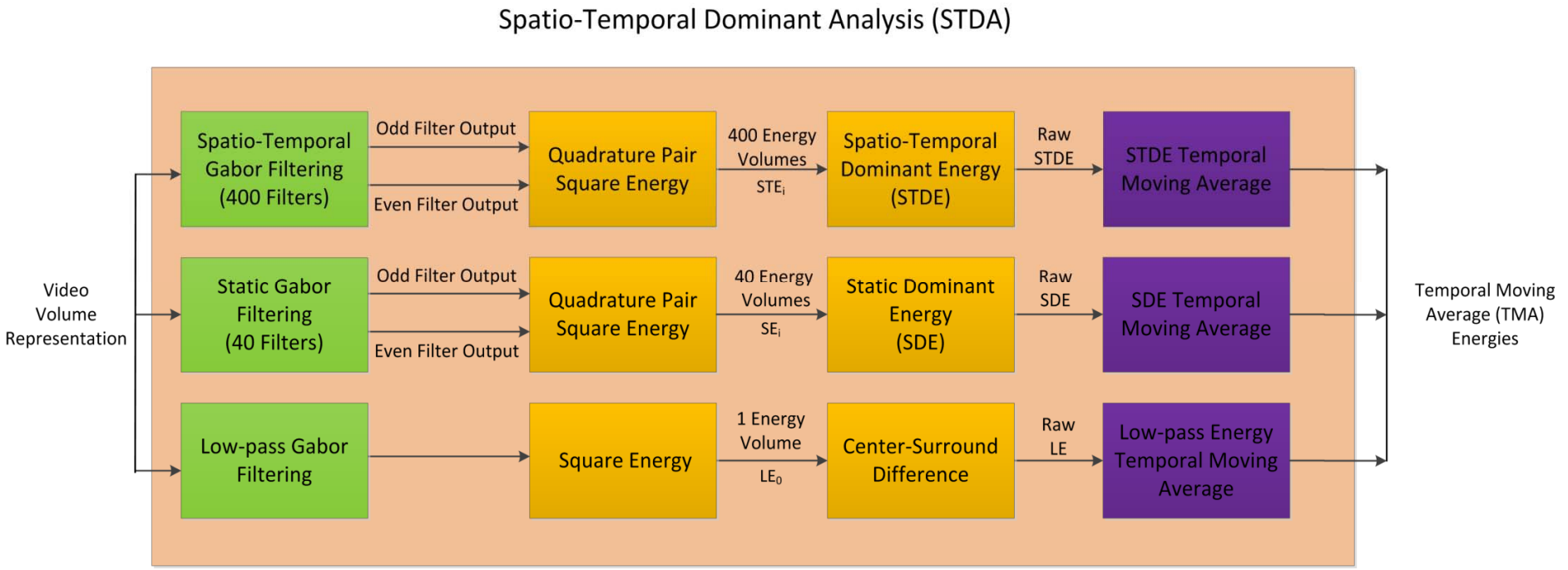
Spatio-Temporal Gabor Filterbank

Full Spatio-Temporal Filterbank in 3D Space



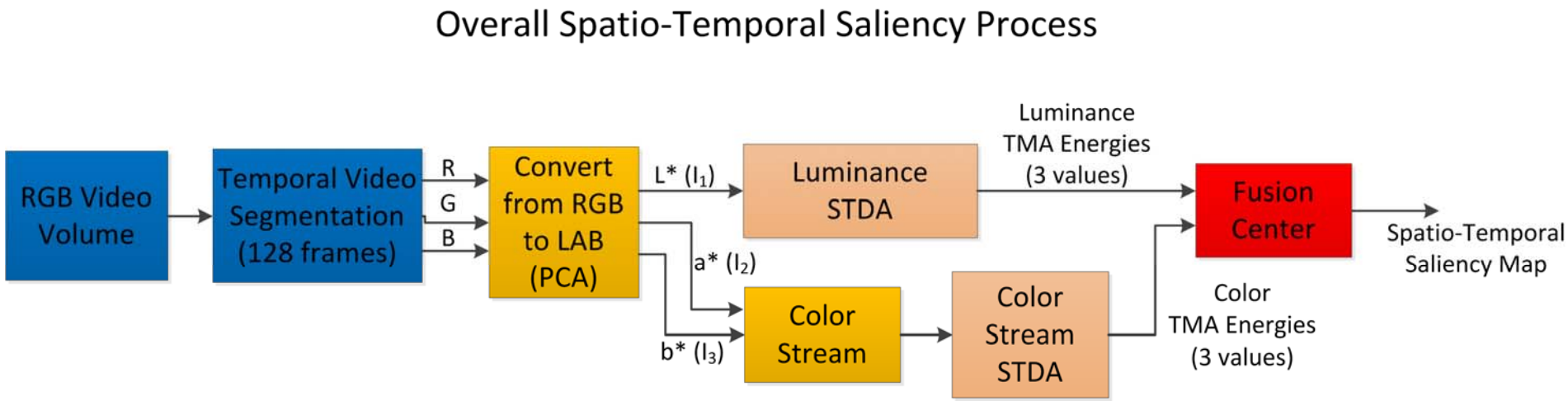
[Koutras and Maragos, *SP: Image Communication*, 2015]

Spatio-Temporal Dominant Analysis (STDA)



- Extract 3 dominant energy volumes for each stream (expressing basic perceptual concepts in visual saliency)
  - **Spatio-Temporal** related with motion
  - **Static** (or Spatial) related with frames' texture or edges
  - **LowPass** related with that other model called "intensity" (which can be either in luminance or color stream)

Spatio-Temporal Frontend for Visual Saliency Overview



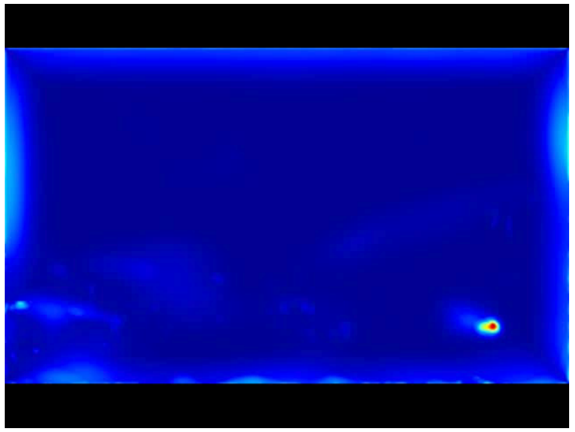
- Color Modeling
  - CIE-Lab or PCA projected color space
  - Luminance stream
  - Color steam

[Koutras and Maragos, SP: Image Communication, 2015]

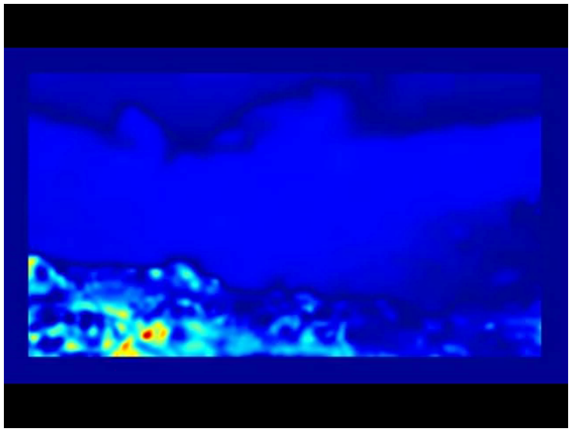
Visual Saliency in Movie Videos - Demo

Original RGB Frames

Luminance STDE



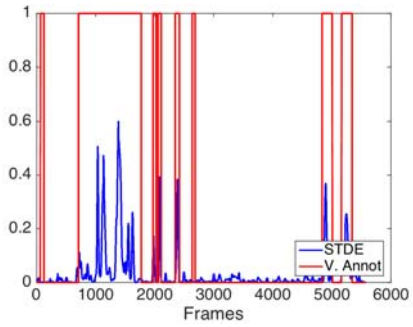
Color Contrast
Low-pass Energy



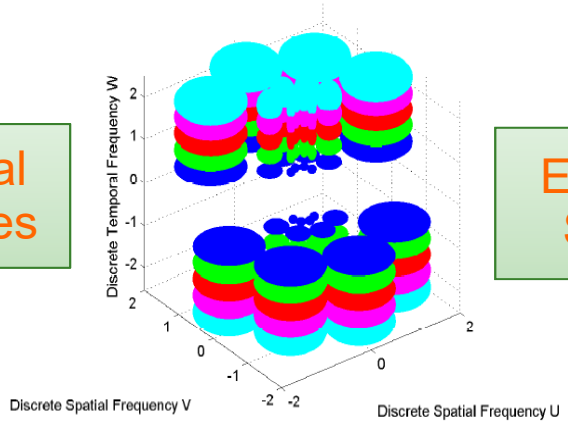
COGNIMUSE Database: Lord of the Rings

Visual Saliency Representations

Spatio-Temporal Frontend for Visual Saliency

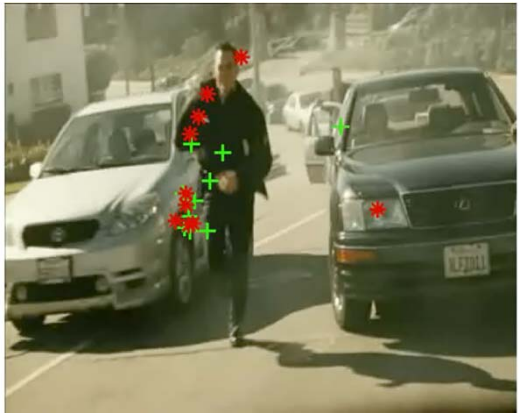


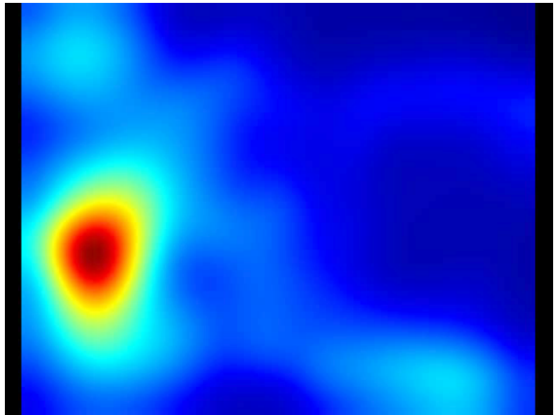
Visual Curves



Energy STIP

Eyes Fixation Prediction





Video Summarization

- Summarization task refers to producing a shorter version of a video:
  - containing all the necessary information required for context understanding
  - without sacrificing much of the original informativeness and enjoyability
- Automatic summaries can be created with:
  - key-frames, which correspond to the most important video frames and represent a static storyboard
  - video skims that include the most descriptive and informative video segments

Movie Summarization (Audio-Vision-Text)

Refs:

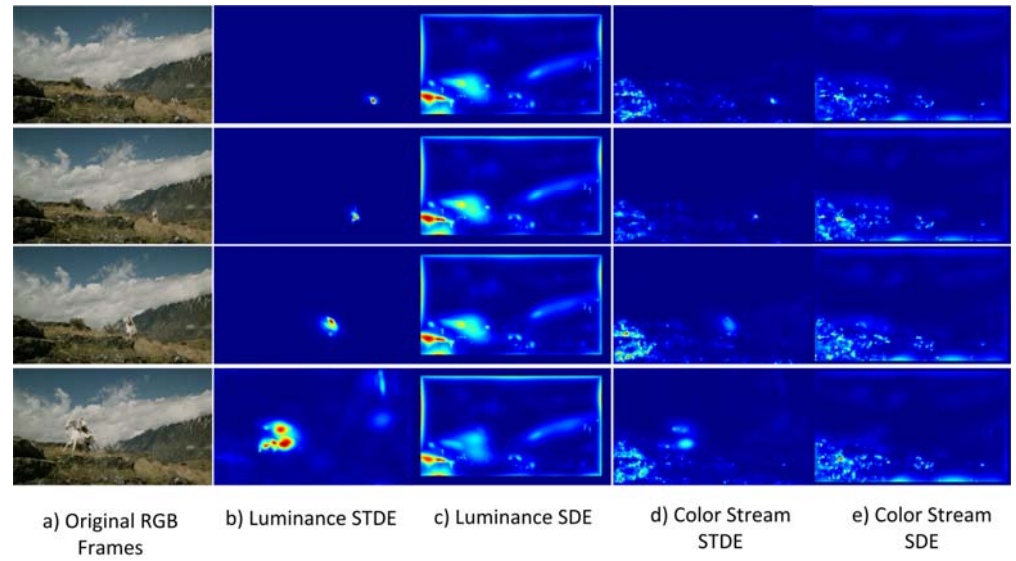
- G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, Y. Avrithis, "Multimodal Saliency and Fusion for Movie Summarization based on Aural, Visual, and Textual Attention", *IEEE Trans.-MM*, 2013.
- P. Koutras, A. Zlatintsi, E. Iosif, A. Katsamanis, P. Maragos, A. Potamianos, "Predicting Audio-visual Salient Events based on A-V-T Modalities For Movie Summarization", *ICIP* 2015.
- A. Zlatintsi, E. Iosif, P. Maragos, A. Potamianos, "Audio Salient Event Detection And Summarization Using Audio And Text Modalities", *EUSIPCO* 2015.
- A. Zlatintsi, P. Koutras, G. Evangelopoulos, N. Malandrakis, N. Efthymiou, K. Pastra, A. Potamianos and P. Maragos, "COGNIMUSE: a multimodal video database annotated with saliency, events, semantics and emotion with application to summarization", *EURASIP Journal on Image and Video Processing* , 2017.

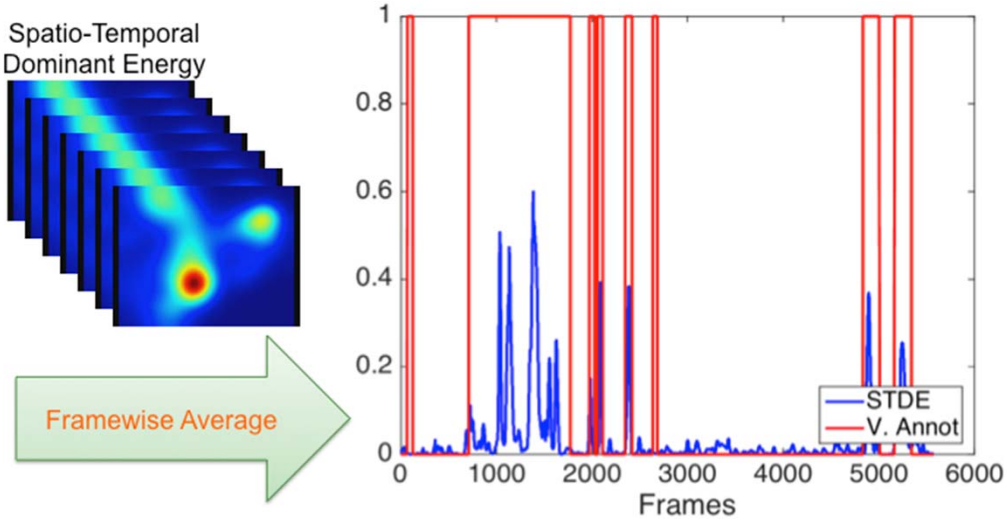
Visual Analysis for Saliency (synopsis)

3D Gabor Energy model

Visual Features

- Both luminance and color streams:
  - Spatio-Temporal Dominant Energies (Filterbank of 400 3D Gabor filters)
  - Spatial Dominant Energies (Filterbank of 40 Spatial Gabor filters)
- Energy Curves
  - Mean value for each 2D frame slice of each 3D energy volume
  - 4 temporal sequences of visual feature vectors.



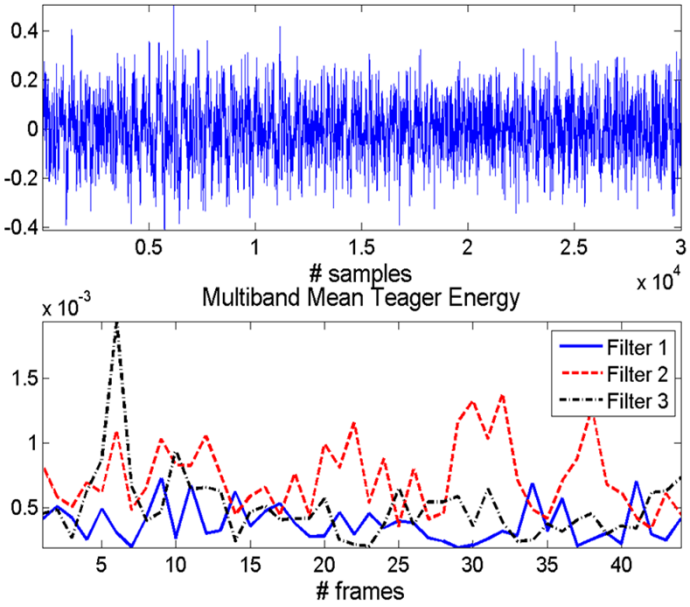


[P. Koutras and P. Maragos. *A Perceptually-based Spatio-Temporal Computational Framework for Visual Saliency Estimation*, Signal Proc.: Image Commun., 2015.]

Audio Analysis for Saliency (synopsis)

- **Multiband filtering** of the signal with **Gabor filters**
  - Teager energy: meaningful in narrowband signals only
  - Gabor filtering for isolation of narrowband components
  - Gabor filters: good joint time-frequency resolution
- **Teager-Kaiser Energy Operator (TEO):** $\Psi[x] = \dot{x}^2 - x\ddot{x}$, where $\dot{x} = dx/dt$
  - For energy estimation and AM-FM demodulation.
  - Robust to noise compared to the squared energy operator.
  - Multiband TECC (Teager Energy Cepstrum Coefficients) successful in speech recognition.
  - For AM-FM signals: $x(t) = \alpha(t)\cos(\varphi(t))$

    Ψ yields: $\Psi[x(t)] \approx a^2(t)\dot{\varphi}^2(t)$
    - Captures amplitude and frequency variation info.
    - Improves accuracy in speech & music recognition.
    - Detects robustly & discriminate various acoustic events due to its sharp time resolution and lowpass behavior.
    - Important for auditory scene analysis.



- **Roughness** (or sensory dissonance of sound)
- **Loudness** (perceived sound pressure level)

[P. Maragos, J.F. Kaiser and T.F. Quatieri, *Energy Separation in Signal Modulations with Application to Speech Analysis,* IEEE Trans. Signal Process., 1993]

[D. Dimitriadis, P. Maragos and A. Potamianos, *On the effects of filterbank design and energy computation on robust speech recognition*, IEEE Trans. ASLP, 2011]

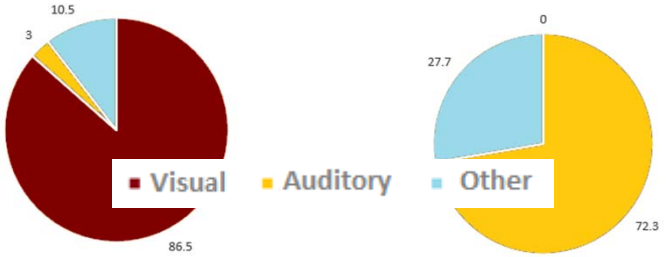
[A. Zlatintsi, E.Iosif, P. Maragos and A. Potamianos. *Audio Salient Event Detection and Summarization using Audio and Text Modalities*, EUSIPCO, 2015]

fMRI for Audio-Visual Saliency

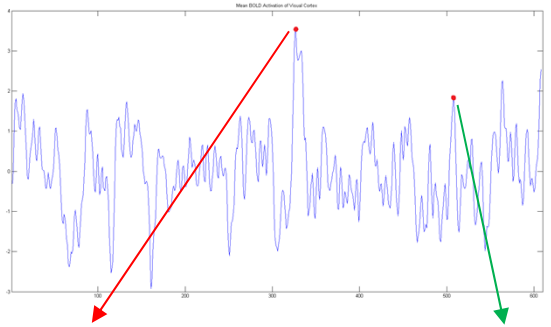
Use of fMRI data for the problem of audiovisual saliency extraction

- Popular computational models for visual & auditory saliency
- Brain activation data during stimulation
- Complex stimuli (movie video)

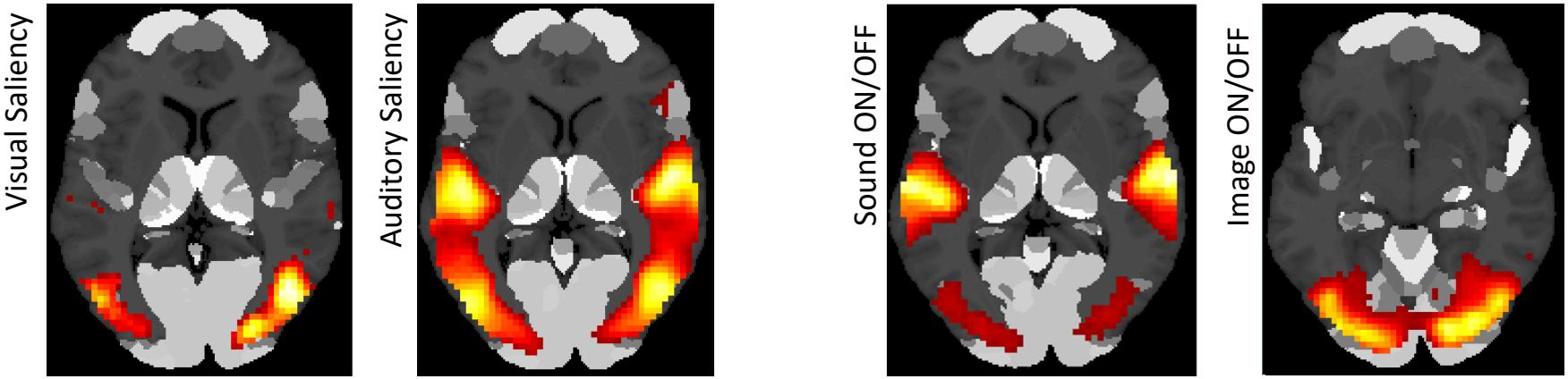
Movie free-viewing [DEP]



Visually salient action scenes [DEP]



Documentary ON/OFF design [BLB]



[Panagiotaropoulou et al., "fMRI-based Perceptual Validation of a Computational Model For Visual and Auditory Saliency In Videos", ICIP 2016]

Text Processing and Saliency

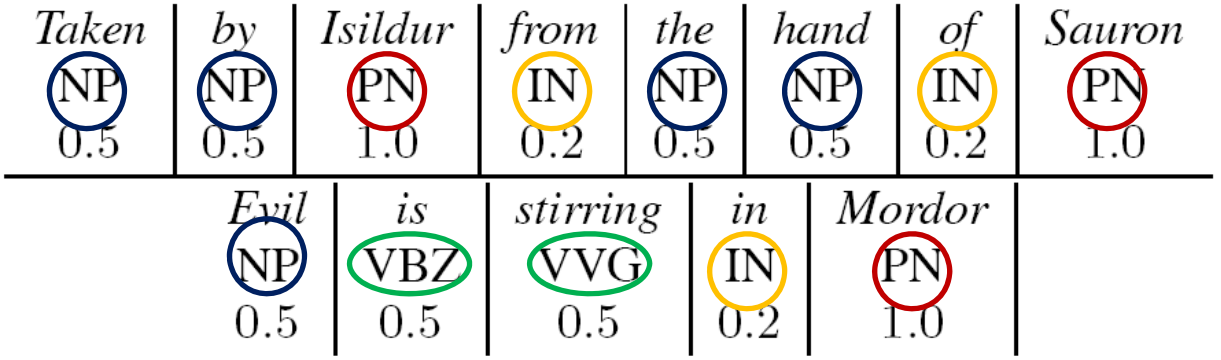
Movie Subtitles provide Text, Timestamps, Semantics.

Easy to process:

I. Extract movie transcript from subtitles and perform part-of-speech (POS) tagging.

II. Segment audio stream using automatic speech recognition & forced alignment, and find the beginning/ending frame for each word in the transcript.

III. Assign text saliency value to video frames based on parser tags.



Fusion (of normalized features/saliencies)

- **Fusion schemes**

$$S_A = \text{fusion}(S_1, S_2, S_3)$$

  - **Linear** (equal weights)
    (Low-level, memoryless)

$$S_{\text{LIN}} = w_1 S_1 + w_2 S_2 + w_3 S_3$$

  - **Variance-based** (adaptive weights)

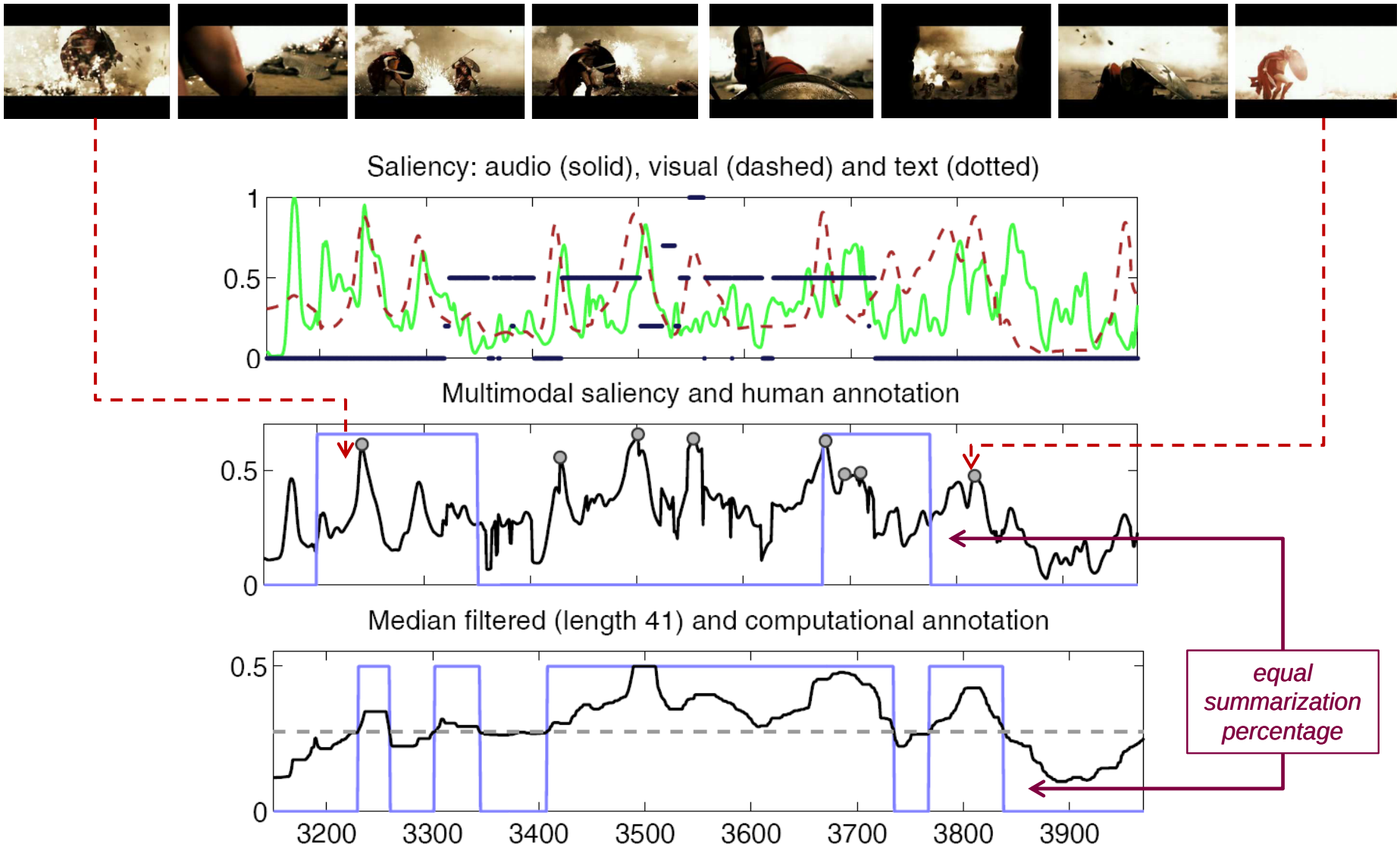
$$S_{VAR} = \sum_i \left(\frac{S_i}{\text{var}(S_i)} \right) / \sum_i \left(\frac{1}{\text{var}(S_i)} \right)$$

  - **Nonlinear**
    - MIN $S_{MIN} = \min\{S_1, S_2, S_3\}$
    - MAX $S_{MAX} = \max\{S_1, S_2, S_3\}$
    - Weighted MIN $S_{MIVA} = \min(S_1 - w_1, S_2 - w_2, S_3 - w_3\} + \max(w_1, w_2, w_3)$

$$\text{where } w_i = \log\left(\frac{1}{\text{var}(S_i)} \right)$$

- **Normalization**: Global (GL), Scene-based (SC), Shot-based (SH).
- **Dynamic Adaptation** (of Variance Weights): Global level (VA-GL), Scene (VA-SC), Shot (VA-SH).

Multimodal Fusion: Audio, Visual, Text



Demo: Movie Summaries

Baseline System: MovieSum 1 (Bottom-Up, Low-dim Features)

LOR VA-SH-F, rate: x5 (6:50 min from 37:33 min)
Inform: 78.7 %
Enjoy: 80.9 %

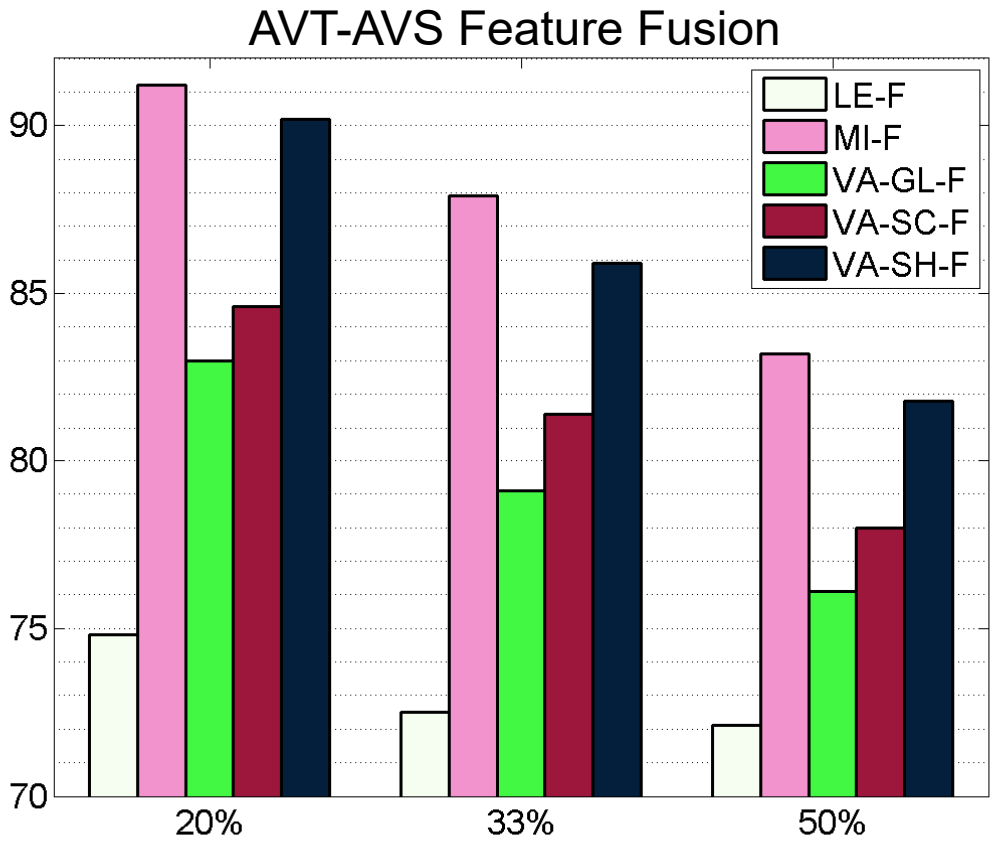
FNE MI-F, rate: x5 (5:07 min from 30:17 min)
Inform: 74.1 %
Enjoy: 78.3 %

[G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, Y. Avrithis, "*Multimodal Saliency and Fusion for Movie Summarization based on Aural, Visual, and Textual Attention*", IEEE Trans.-MM, 2013.]

Objective Evaluation (system 1)

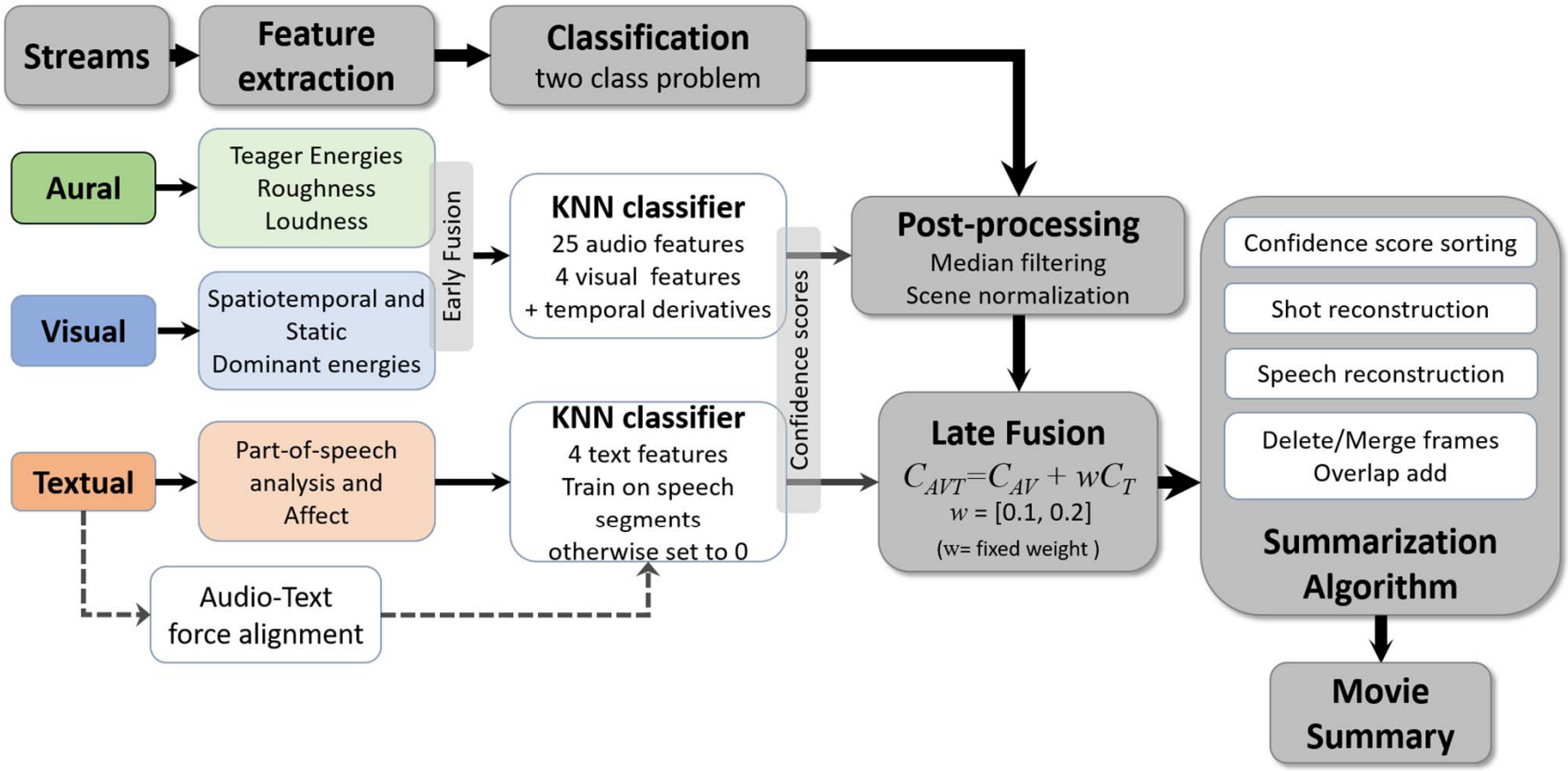
Automatic Movie Summarization (Database: 7 x ~30min movies)

Best Four Fusion schemes with GL-N + baseline (LE-F)
(in terms of frame-level precision)



[G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, Y. Avrithis, "*Multimodal Saliency and Fusion for Movie Summarization based on Aural, Visual, and Textual Attention*" IEEE Trans.-MM, 2013.]

Summarization System MovieSum 2 (w. Learning, improved frontend)



[P. Koutras, A. Zlatintsi, E. Iosif, A. Katsamanis, P. Maragos and A. Potamianos, *Predicting Audio-visual Salient Events based on A-V-T Modalities For Movie Summarization*, ICIP 2015]

COGNIMUSE Database http://cognimuse.cs.ntua.gr/database

An evolving multimodal video database annotated with:

Saliency

Semantic events

"Good to see you again old friend!"

Cross-media relations

Audio & Visual events

Experimental Results using the MovieSum System-2:
on 7 Hollywood movies clips (ca. 30 min./each), a full movie (ca. 100 min) and 5 travel documentaries (ca. 20 min./each).

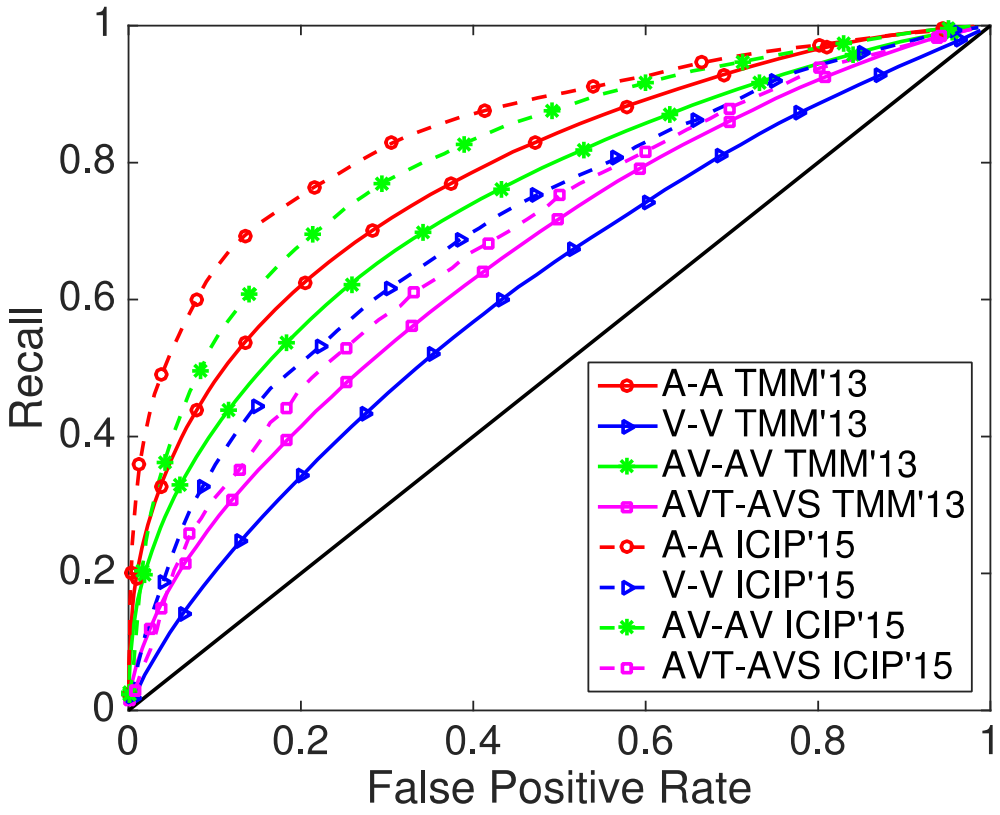
[A. Zlatintsi et al., EURASIP J. on Image and Video Processing, 2017]

Video summaries: Hollywood movies (system 2)

CRA (w0.1) ca 20%, ca 5'30"
informativeness up to 80%

CHI (w0.2) ca 20%, ca 7'
enjoyability up to 85%

System 2: Objective Evaluation on Movie videos

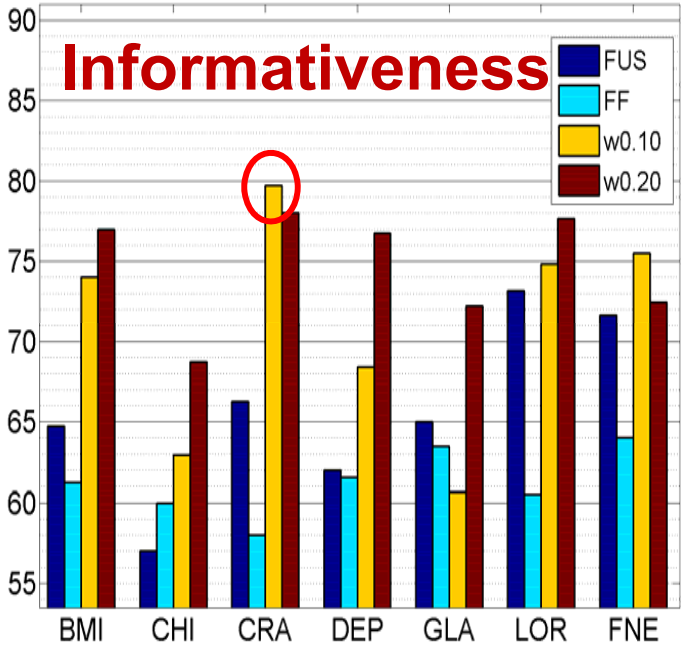


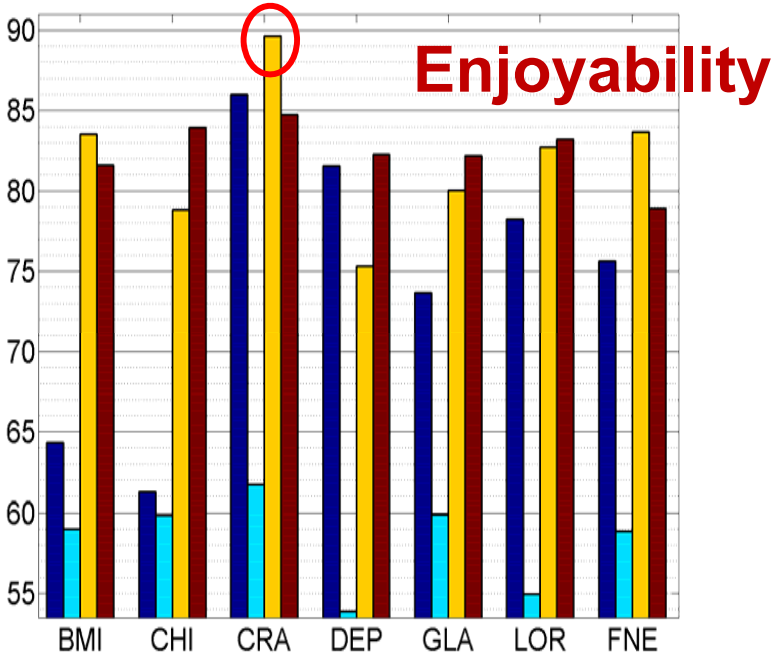
IEEE T-MM 2013 versus ICIP 2015

- The proposed system (ICIP 2015) outperforms the baseline MovieSum-1 System (T-MM 2013) for all evaluation setups
- Greater improvement for A-A
- **Improvements** due to:
  - Advanced monomodal frontends, in all modalities
  - Carefully designed movie summarization algorithm that:
    - corrects the boundaries and
    - results in smoother transitions

[G. Evangelopoulos et al., *Multimodal Saliency and Fusion for Movie Summarization based on Aural, Visual, and Textual Attention,* IEEE T-MM 2013]
[P. Koutras et al., *Predicting Audio-visual Salient Events based on A-V-T Modalities For Movie Summarization,* ICIP 2015]

Experimental results: (20) Human Evaluation on 7 movie clips





Setup: Summaries x5, ca. 6 min., 20 users

Evaluation on:

T_W0.1: text weight $T_W = 0.1$
T_W0.2: text weight $T_W = 0.2$
FUS: fusion method [TMM 2013]
FF: fast-forward (sub-sampling 2 sec. every 10 sec.)

Results:

- Different $\mathbf{T_W}$ is important and related to the movie genre
- Action movies need higher T_W
- Boundary correction contributed to enjoyability:

a) smoother transitions &

b) semantically coherent events

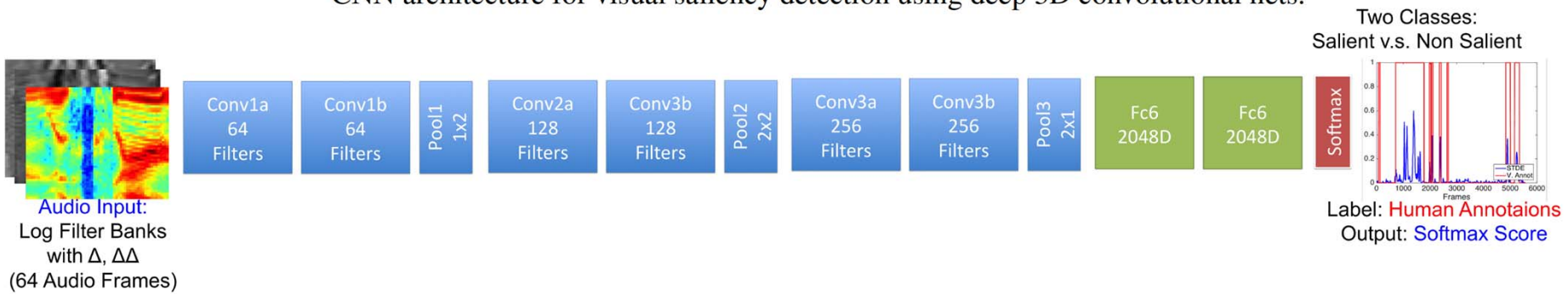
Video summaries: travel doc & gwtw (system 2)

AR London ca 16% ca 3'40"

GWTW ca 3% ca 3'
(3min from full duration movie**)**

CNN-based salient event detection

CNN architecture for visual saliency detection using deep 3D convolutional nets.



CNN architecture for audio saliency detection using 2D convolutional nets on audio time-frequency representations.

Visual Only

AUC Results	V-V	
videos	Hndcr.	CNN
BMI	0.739	0.765
GLA	0.718	0.772
CHI	0.644	0.706
FNE	0.608	0.502
LOR	0.688	0.738
CRA	0.719	0.726
DEP	0.777	0.741
Aver.	0.699	0.707

AUC Results	Visual	
videos	Hndcr.	CNN
Full length Movie		
GWW	0.580	0.618
Five Travel Documentaries		
LON	0.650	0.806
RIO	0.668	0.717
SYD	0.620	0.770
TOK	0.767	0.831
GLN	0.657	0.678
Aver.	0.673	0.767

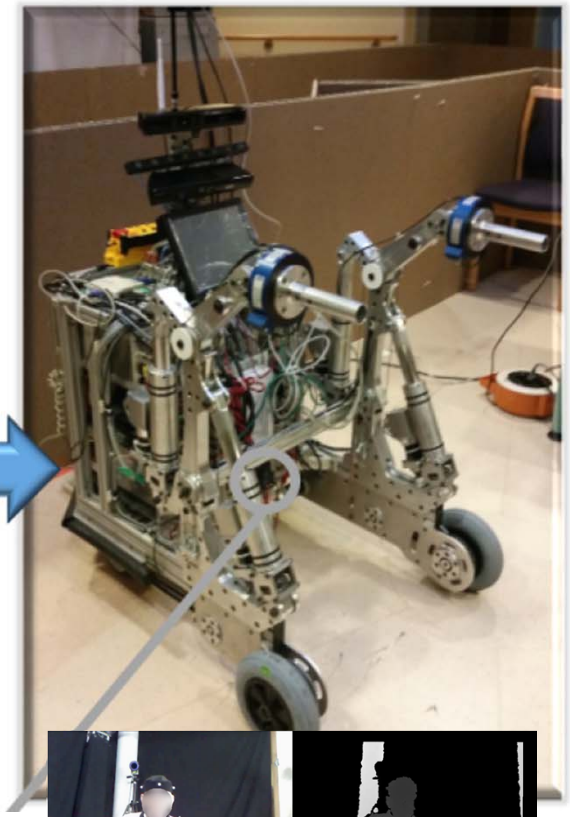
Audio-Visual

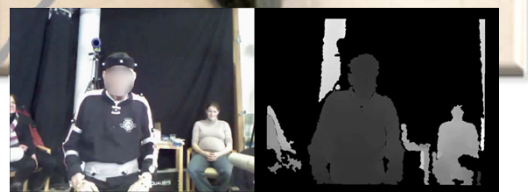
AUC Results	CNN				
movies	A-AV	V-AV	AV-AV mean	AV-AV min	AV-AV max
BMI	0.796	0.785	0.827	0.814	0.811
GLA	0.836	0.775	0.830	0.785	0.839
CHI	0.852	0.677	0.782	0.719	0.834
FNE	0.825	0.504	0.778	0.794	0.676
LOR	0.841	0.750	0.832	0.795	0.830
CRA	0.548	0.732	0.736	0.736	0.549
DEP	0.788	0.824	0.841	0.795	0.842
Aver.	0.781	0.720	**0.801**	0.775	0.767

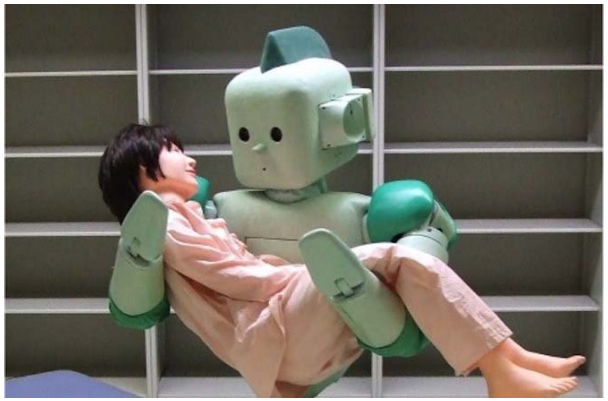
3.
Audio-Visual Gesture Recognition and Human-Robot Interaction

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

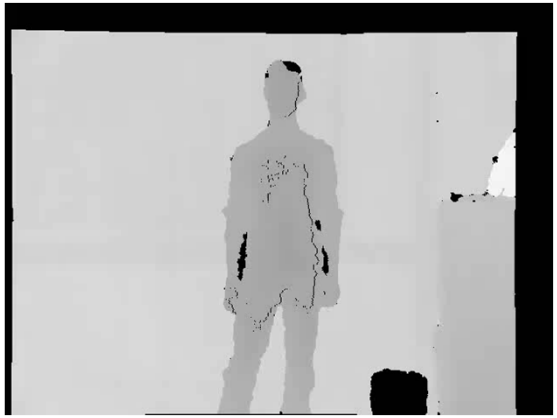
Multimodal Gesture Signals from Kinect-0 Sensor

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

RGB Video & Audio

Depth

(vieniqui - *come here*)



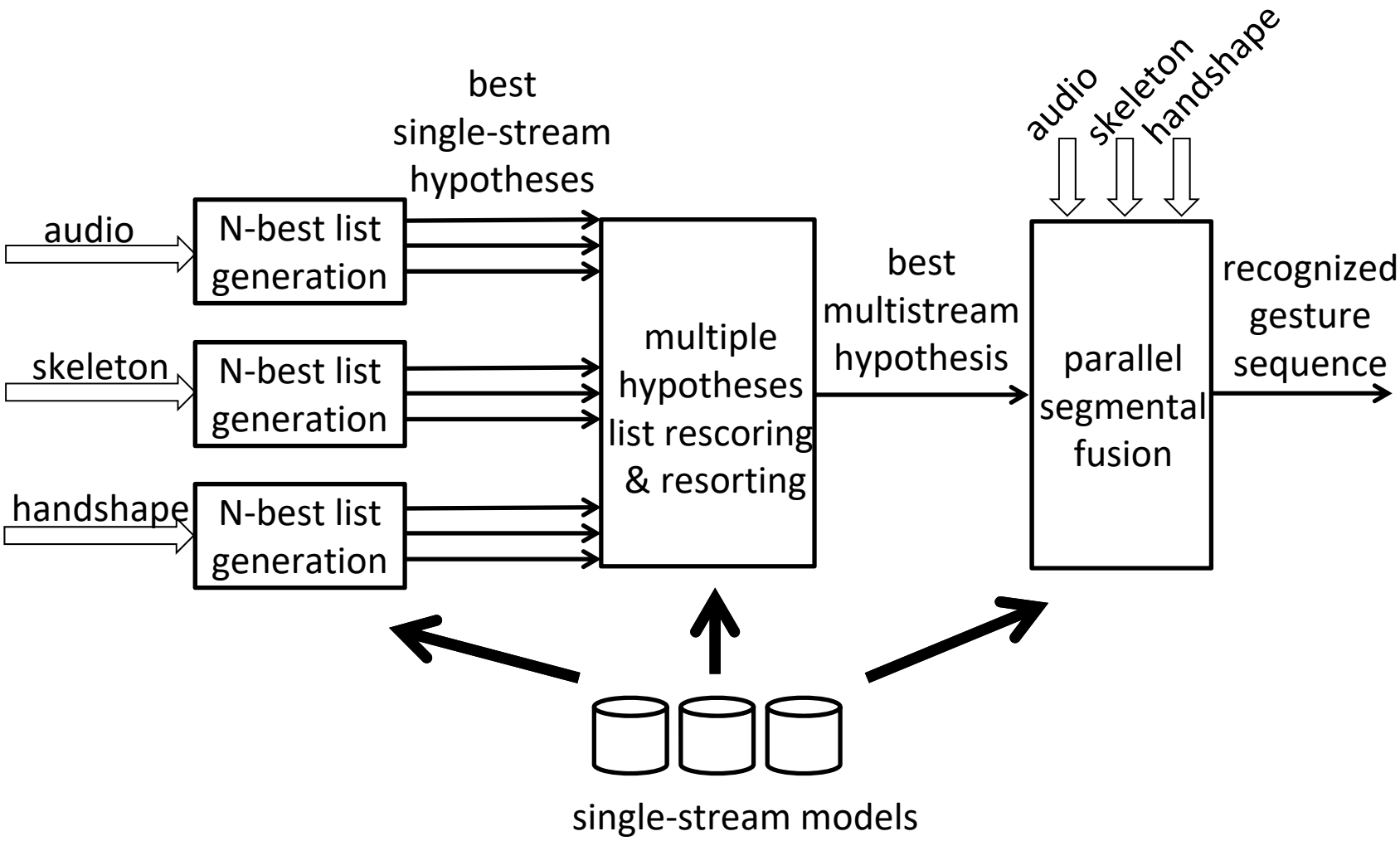
Skeleton

(vieniqui - *come here*)

User Mask

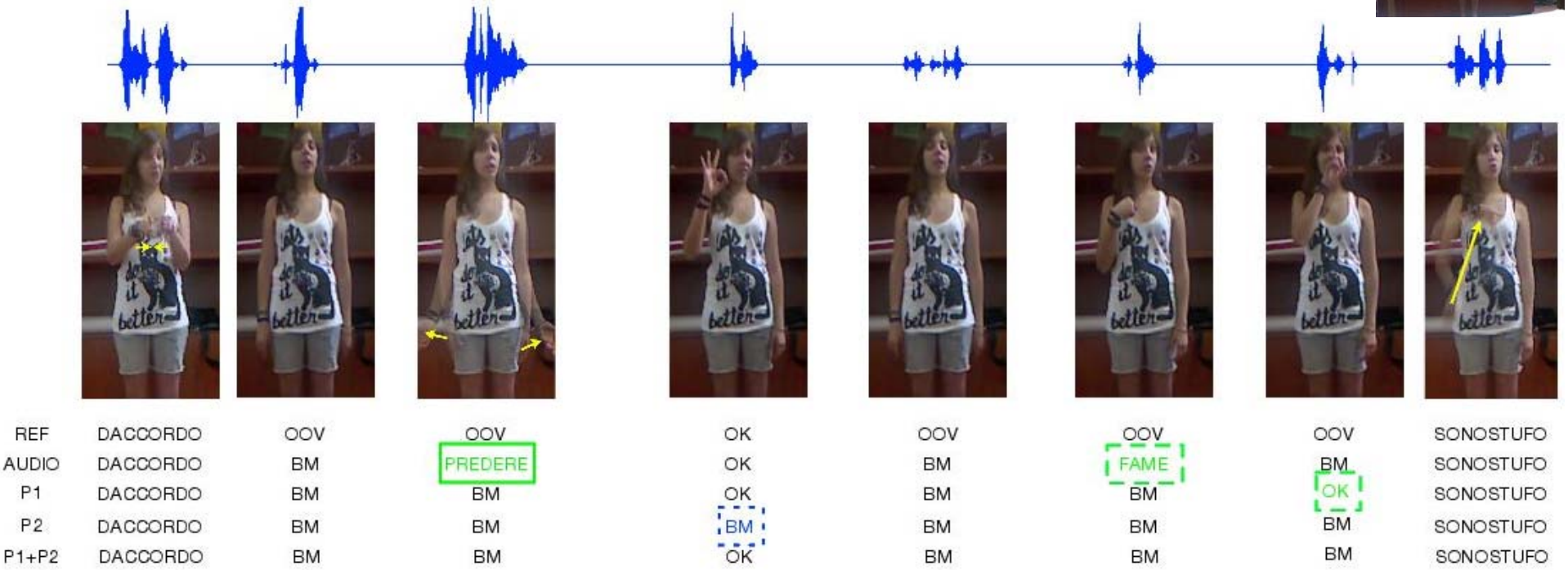
(vieniqui - *come here*)

Overview: Multimodal Hypothesis Rescoring + Segmental Parallel Fusion



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

Audio-Visual Fusion & Recognition



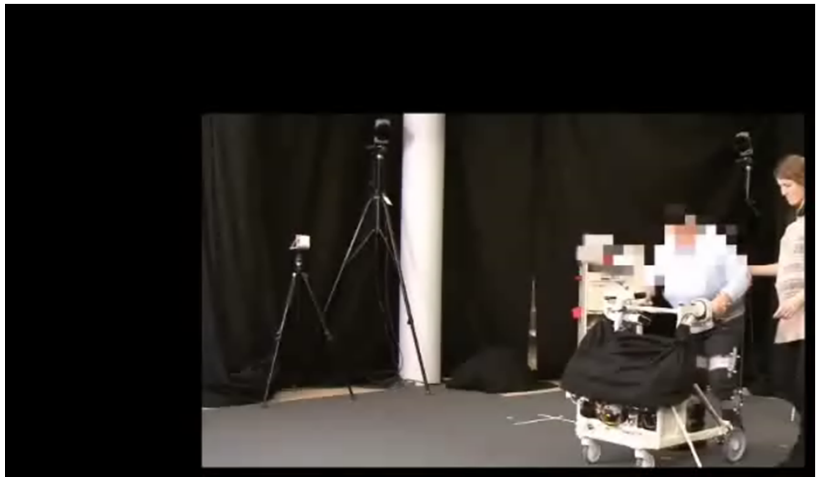
REF	DACCORDO	OOV	OOV	OK	OOV	OOV	OOV	SONOSTUFO
AUDIO	DACCORDO	BM	PREDERE	OK	BM	FAME	BM	SONOSTUFO
P1	DACCORDO	BM	BM	OK	BM	BM	OK	SONOSTUFO
P2	DACCORDO	BM	BM	BM	BM	BM	BM	SONOSTUFO
P1+P2	DACCORDO	BM	BM	OK	BM	BM	BM	SONOSTUFO

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

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

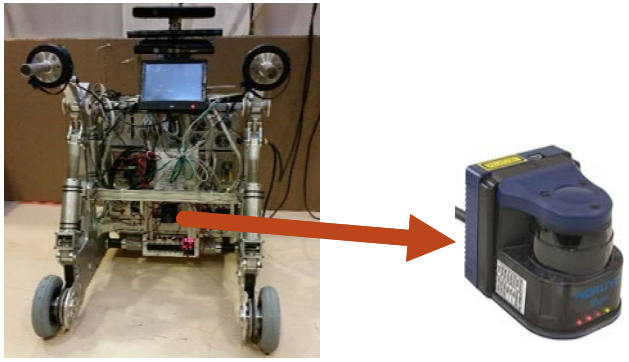
EU Project MOBOT: Motivation

Experiments conducted at
Bethanien Geriatric Center Heidelberg



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

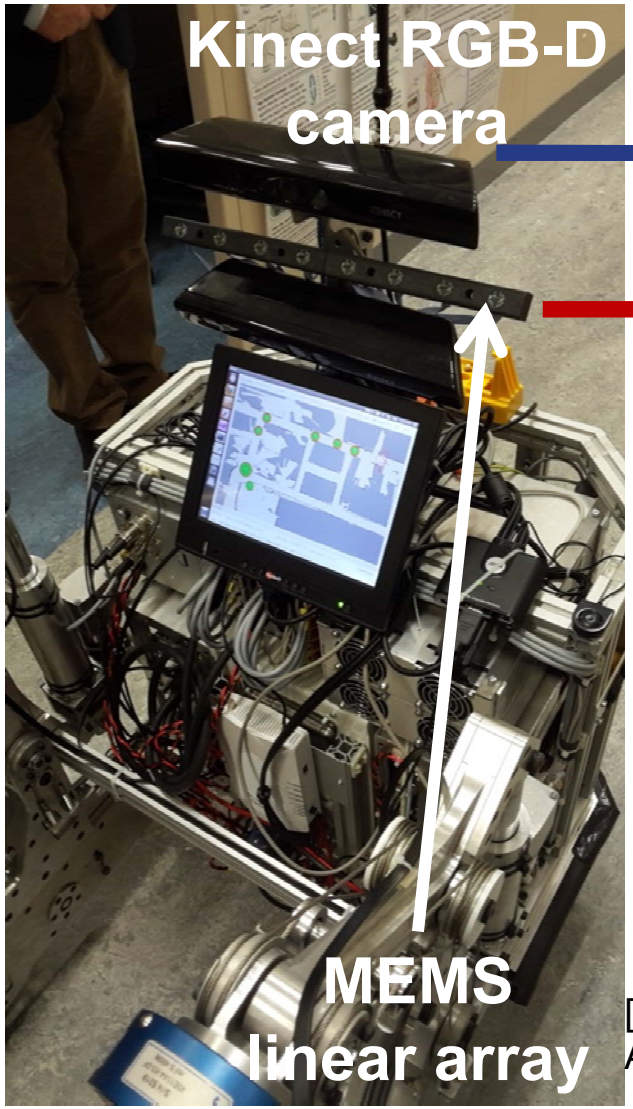
Intelligent assistive devices (robotic Rollator) aiming to provide ***context-aware*** and ***user-adaptive*** mobility (**walking**) assistance



MOBOT rollator

Audio-gestural command recognition: Overview of our multimodal interface

MOBOT robotic platform



Visual Action-gesture Recognition

Spoken Command Recognition

N-best hypotheses & scores

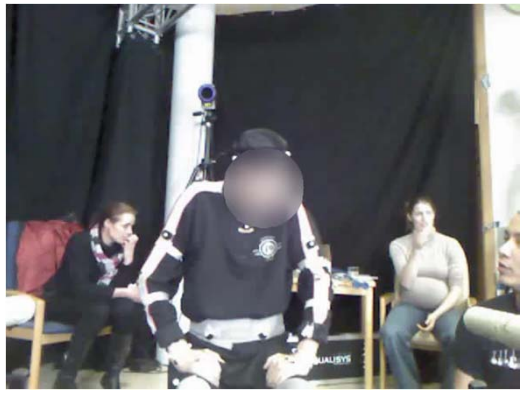
Multimodal Late Fusion

Best AV Hypothesis

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

Multi-Sensor Data for HRI

Kinect1 RGB Data

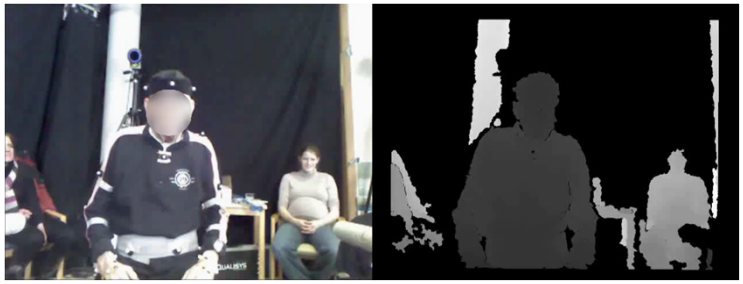


Kinect Depth Data

Kinect1 RGB

Kinect1 Depth

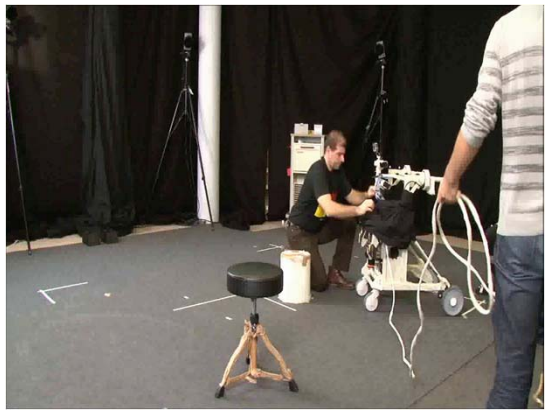
MEMS Audio Data



Go Pro RGB Data

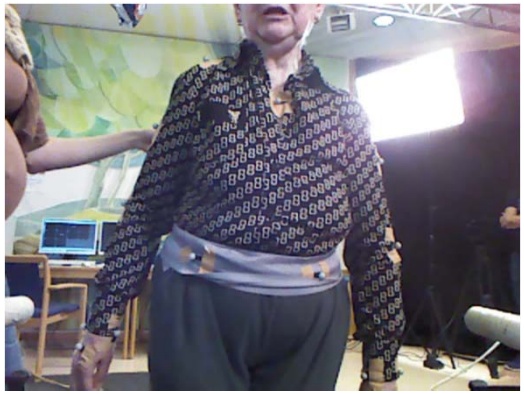
HD1 Camera Data

HD2 Camera Data

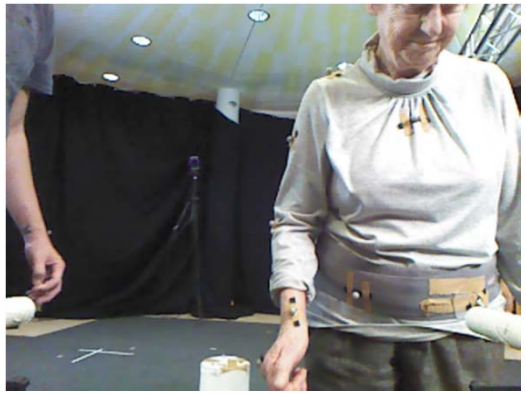


Action Sample Data and Challenges

- 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



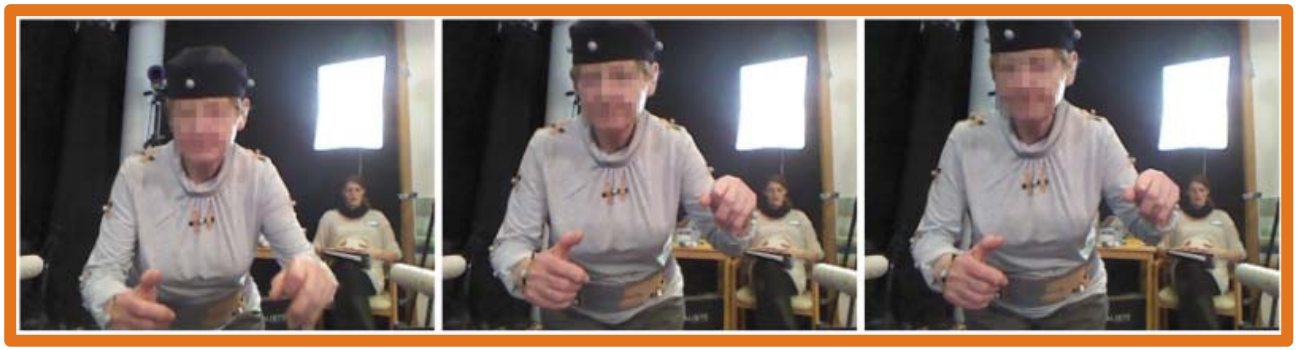
Stand-to-Sit – P3

Stand-to-Sit – P4

Visual Activity Recognition

action → gesture → sign

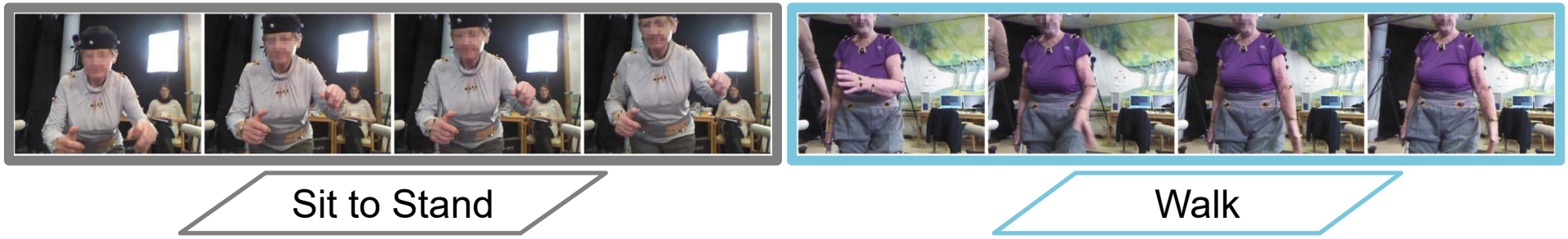
Action: sit to stand



Gestures: come here, come near

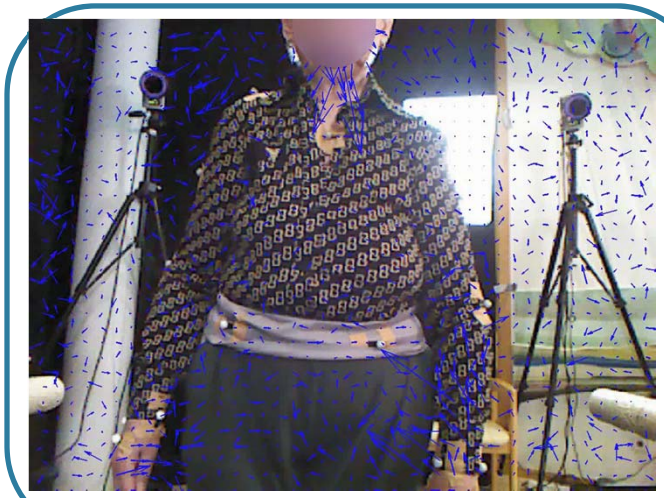
Sign:
(GSL) Europe

Visual action recognition pipeline



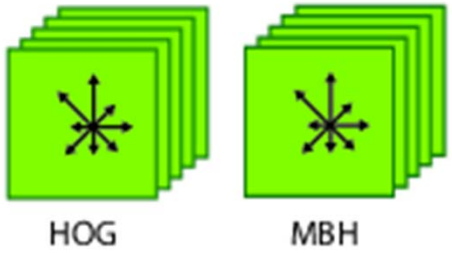
Visual Front-End

Video



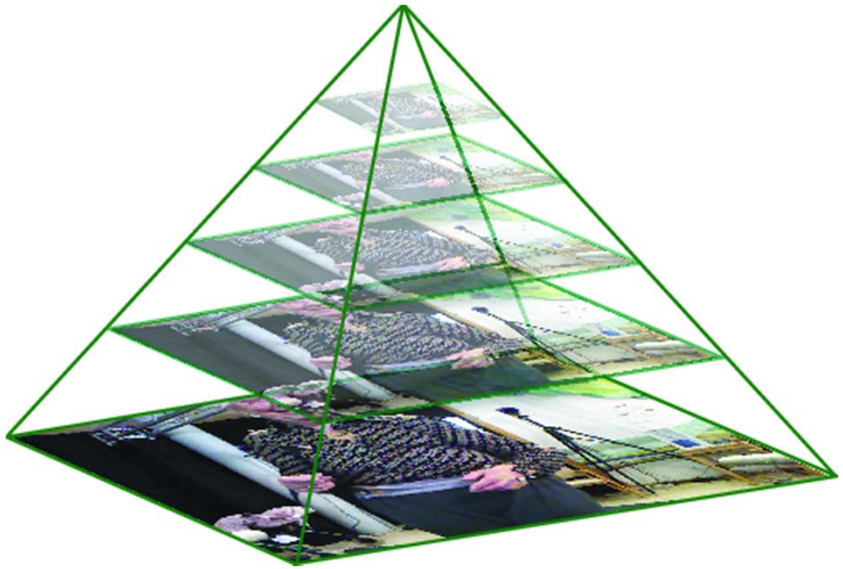
Optical Flow

Dense Trajectories

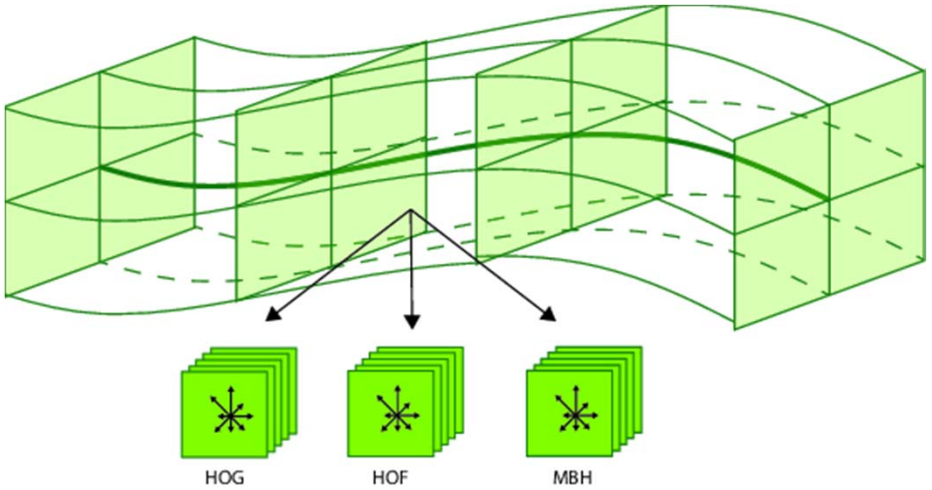


Feature Descriptors

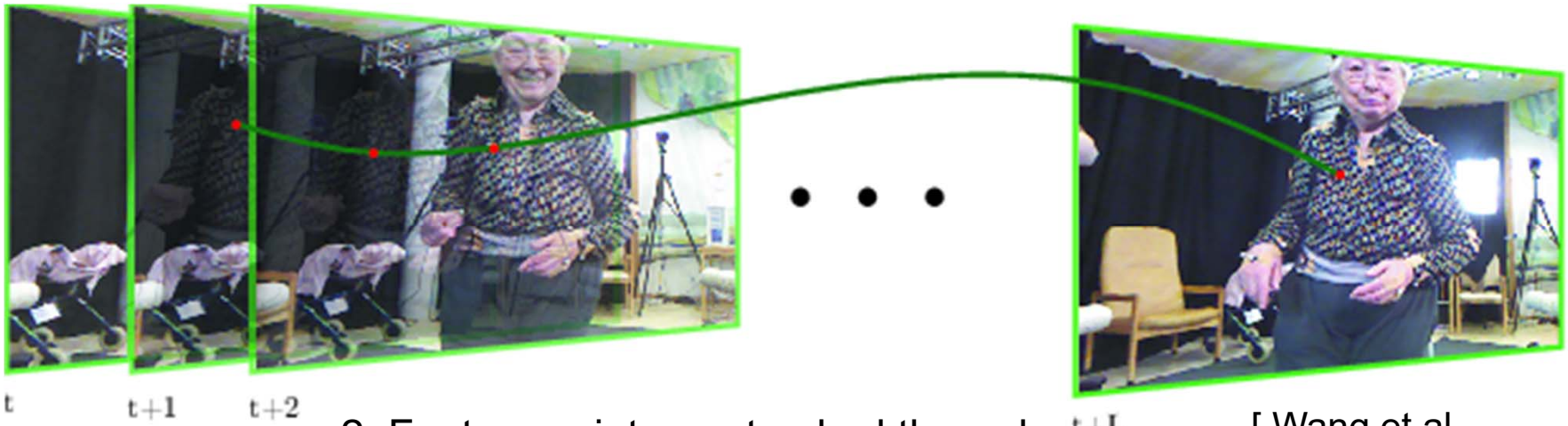
Features: Dense Trajectories



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



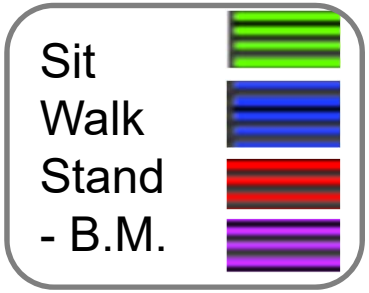
3. Descriptors are computed in space-time volumes along trajectories

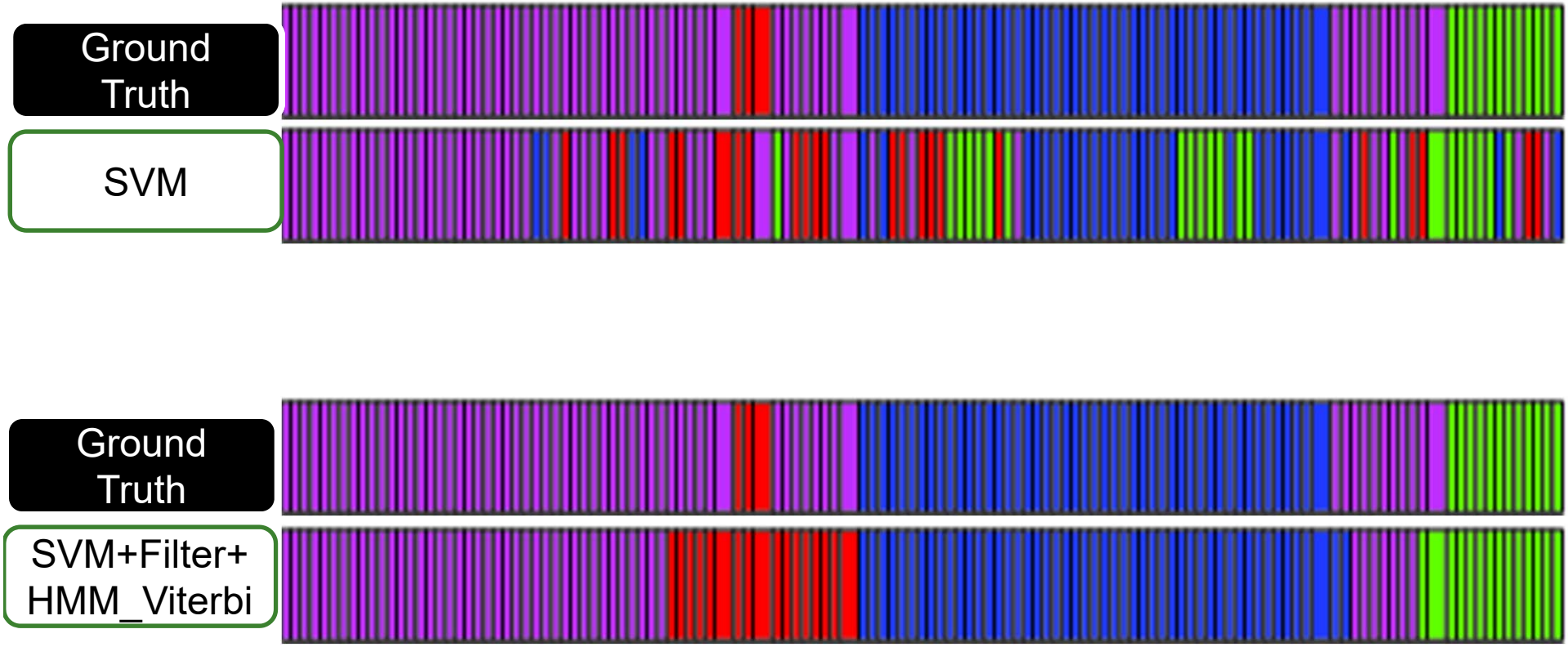


2. Feature points are tracked through consecutive video frames

[Wang et al. IJCV 2013]

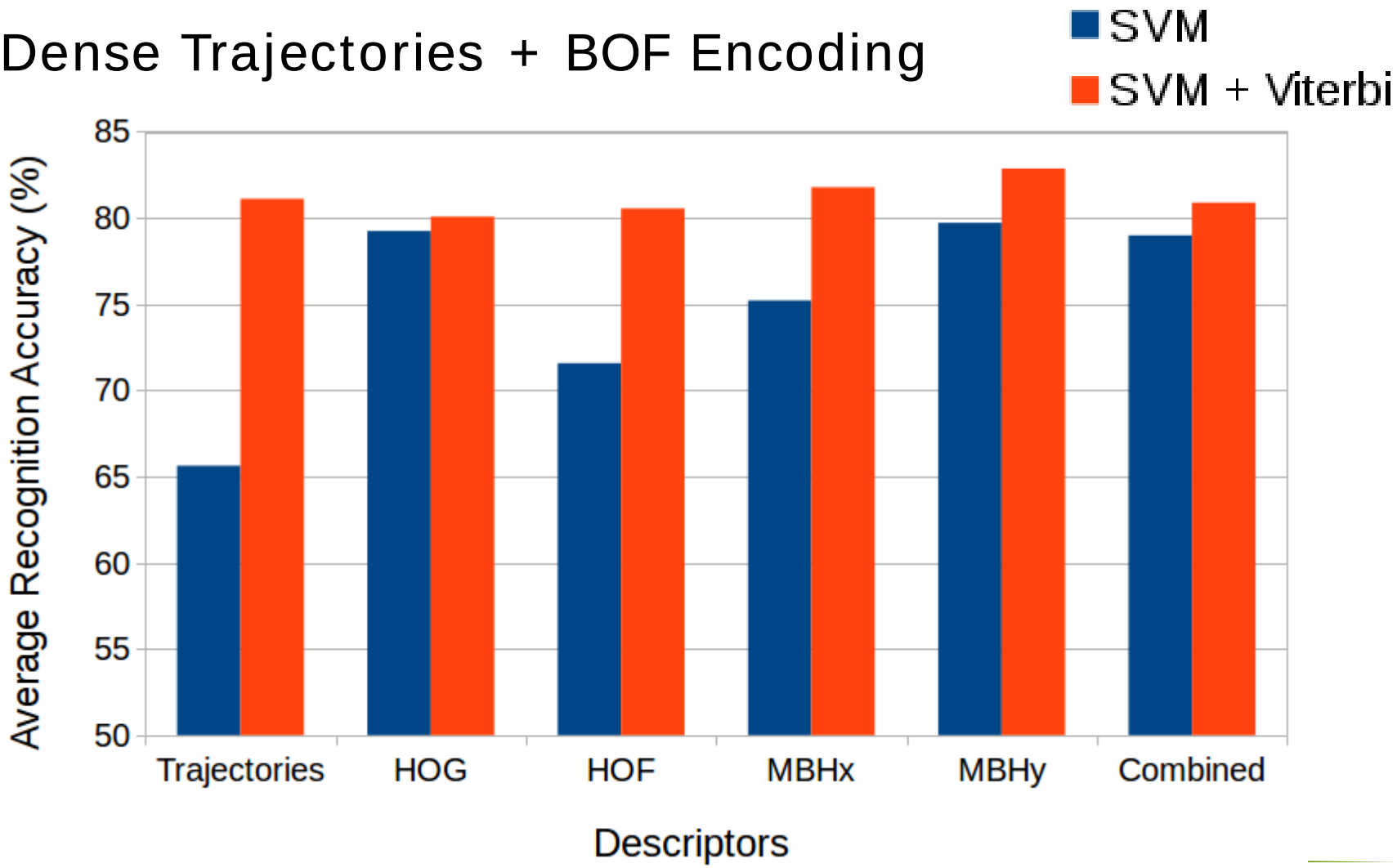
Temporal Segmentation Results





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.

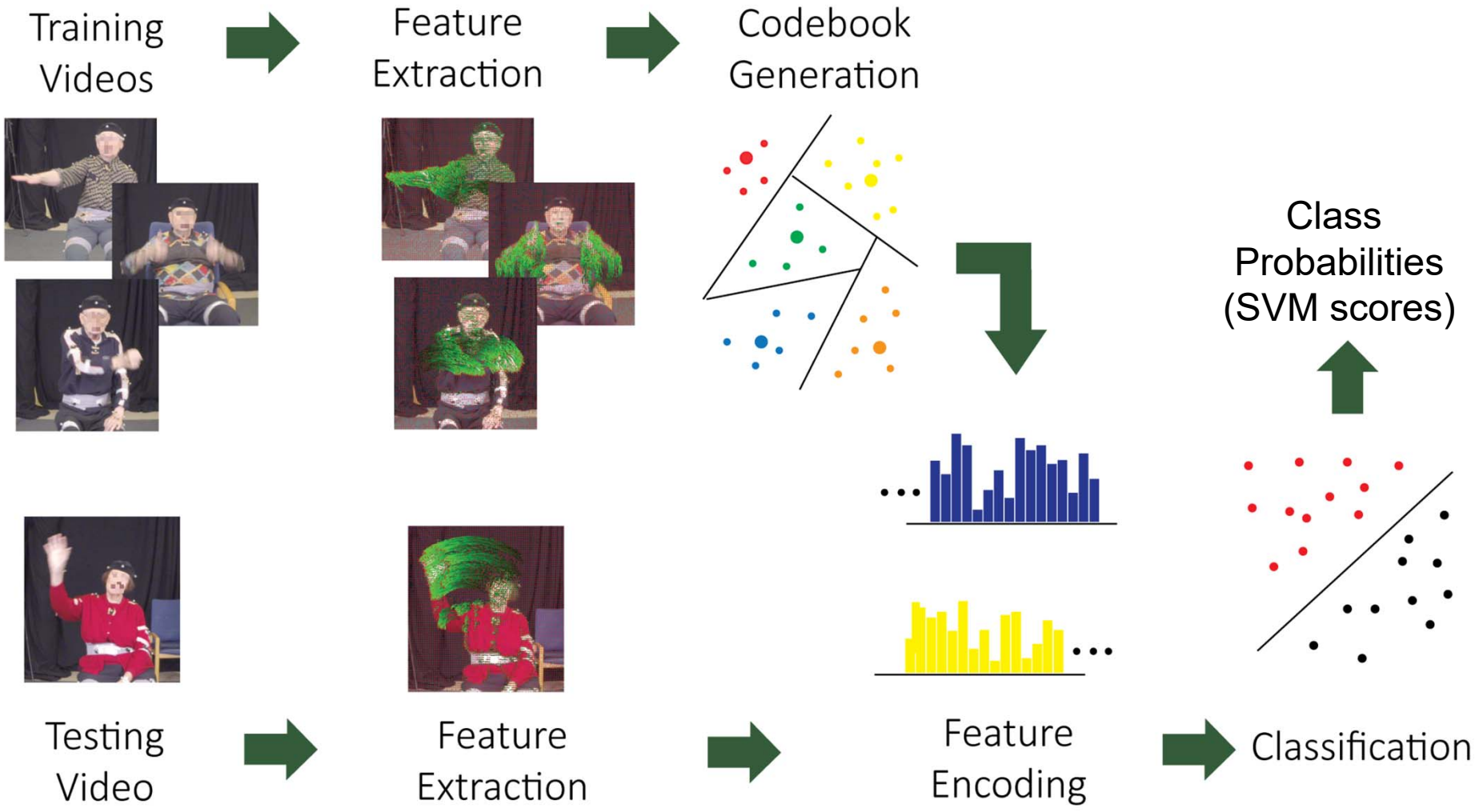
Come Closer

I want to Sit Down

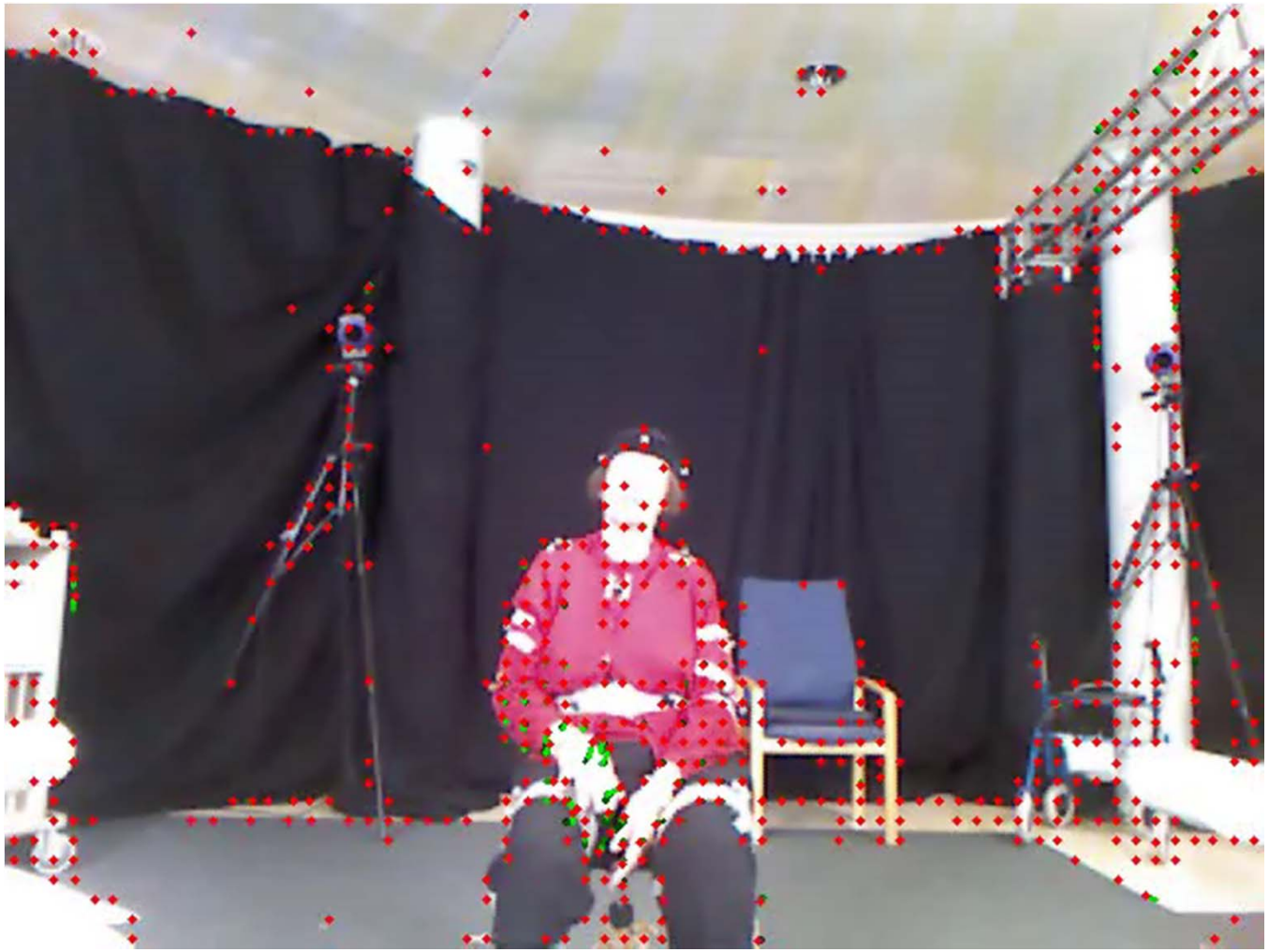
Park

I want to Perform a Task

Visual Gesture Classification Pipeline



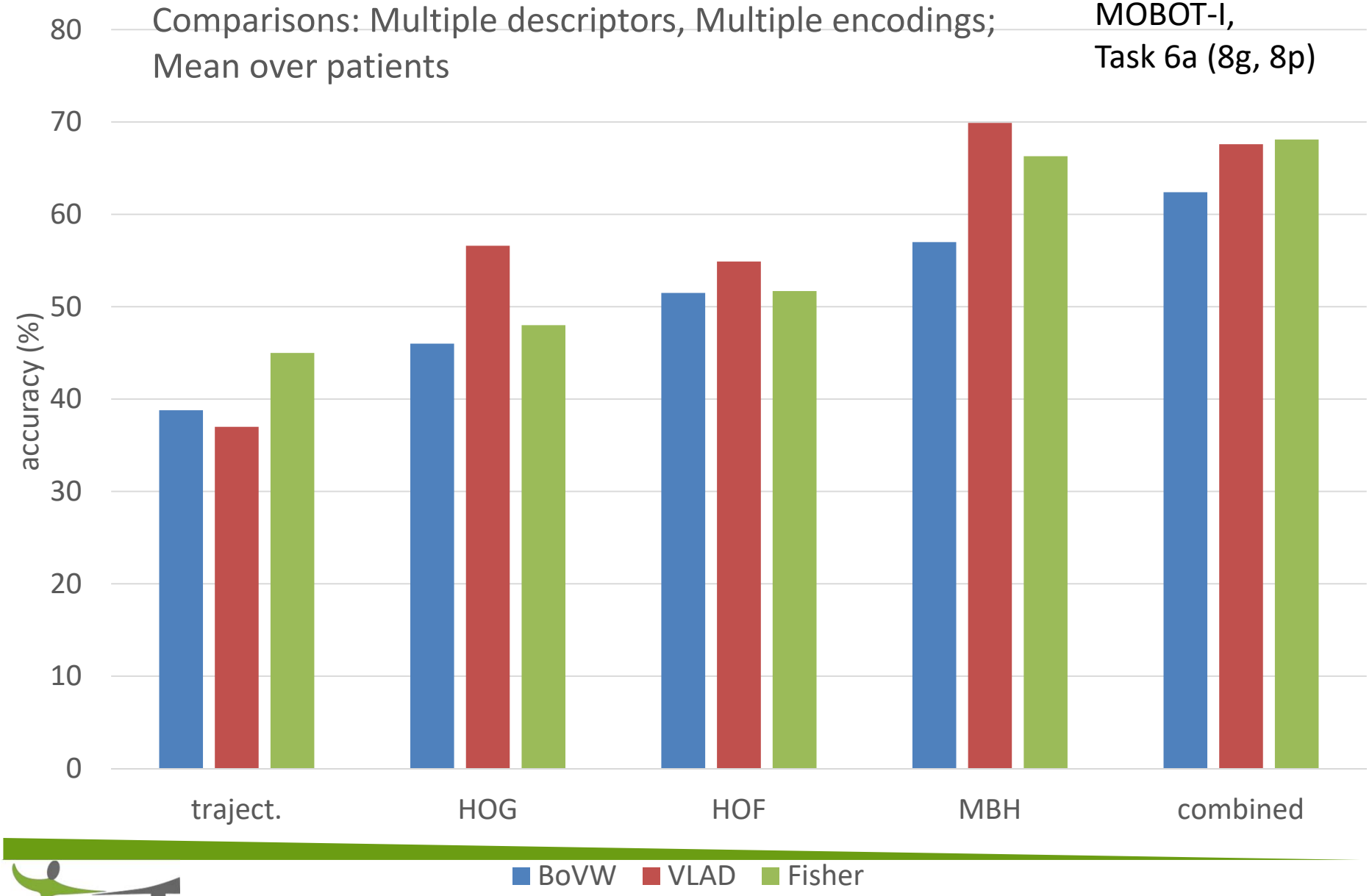
Applying Dense Trajectories on gesture data



Extended results on Gesture Recognition

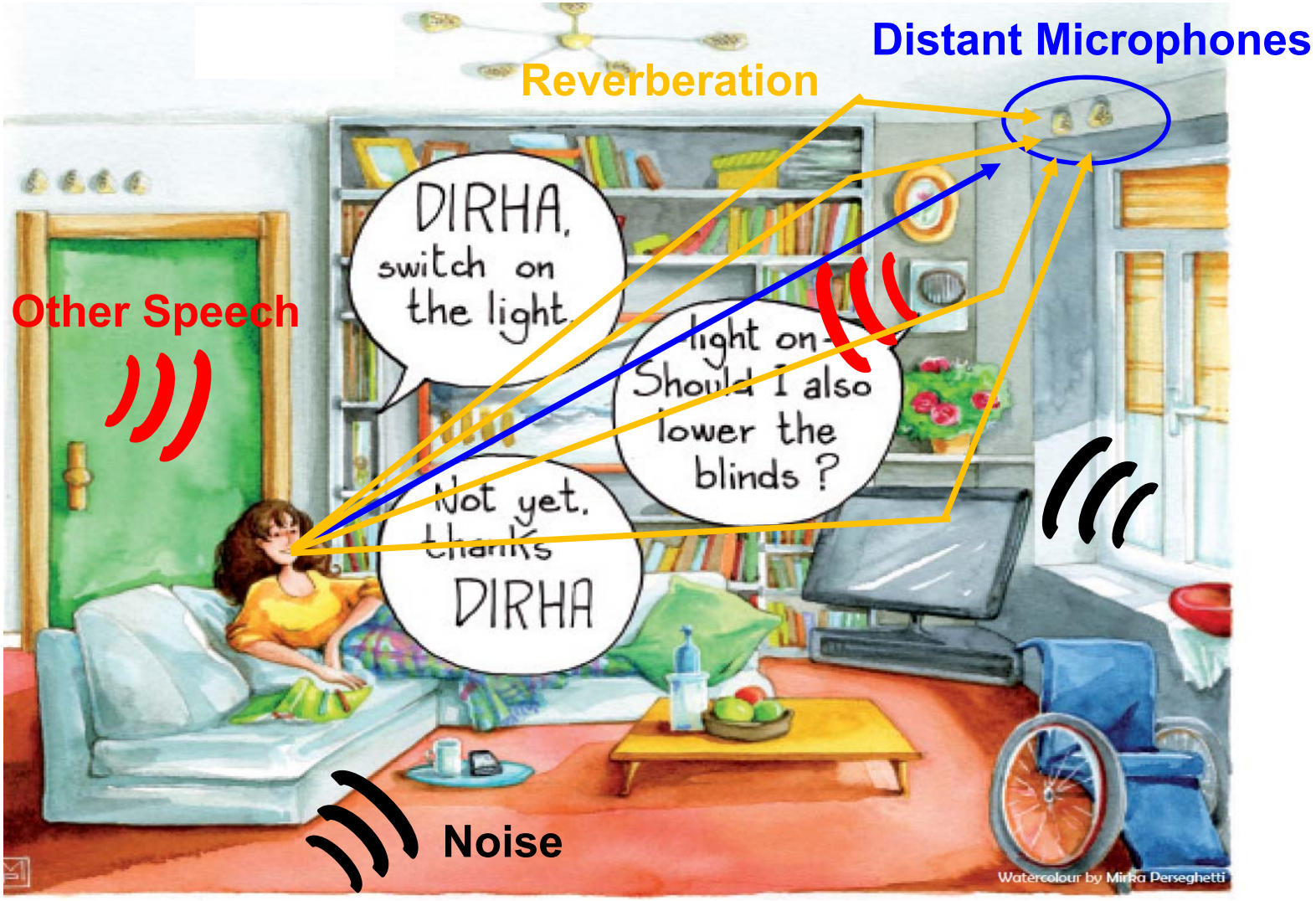
Comparisons: Multiple descriptors, Multiple encodings;
Mean over patients

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



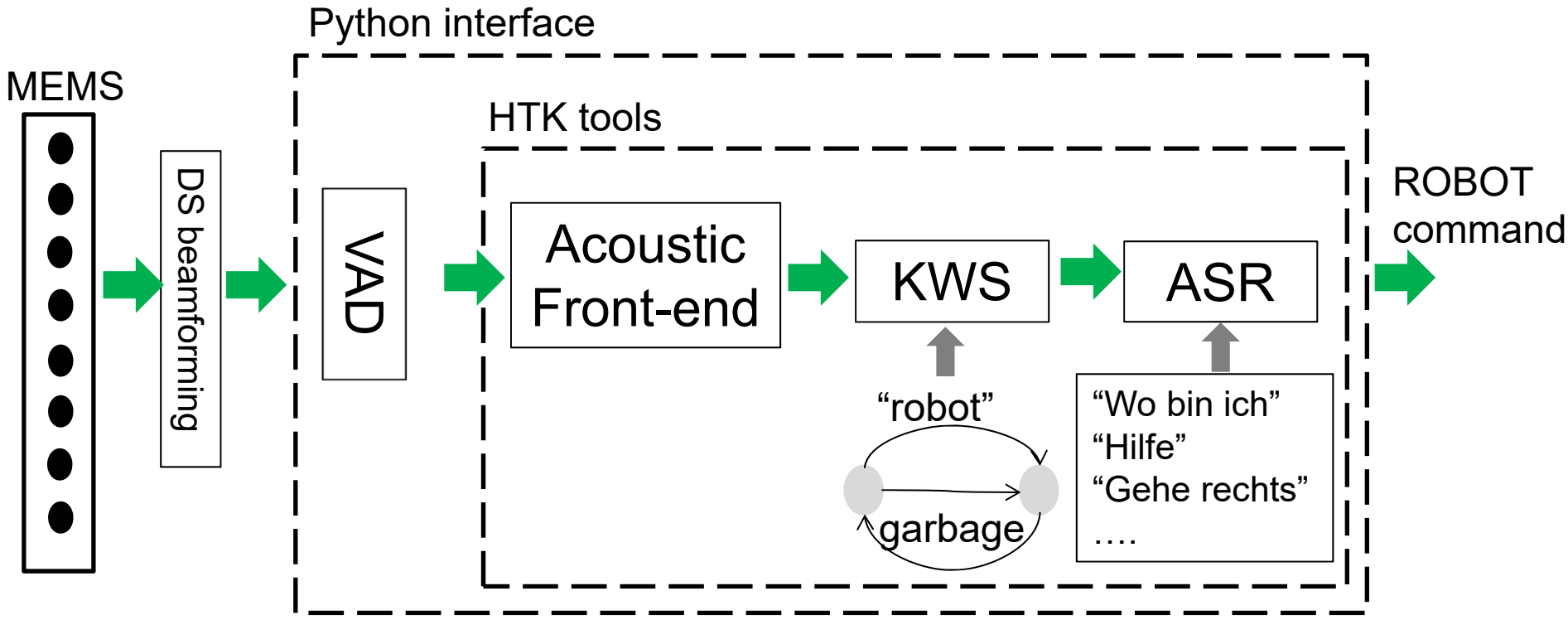
Spoken Command Recognition

Distant Speech Recognition in Voice-enabled Interfaces



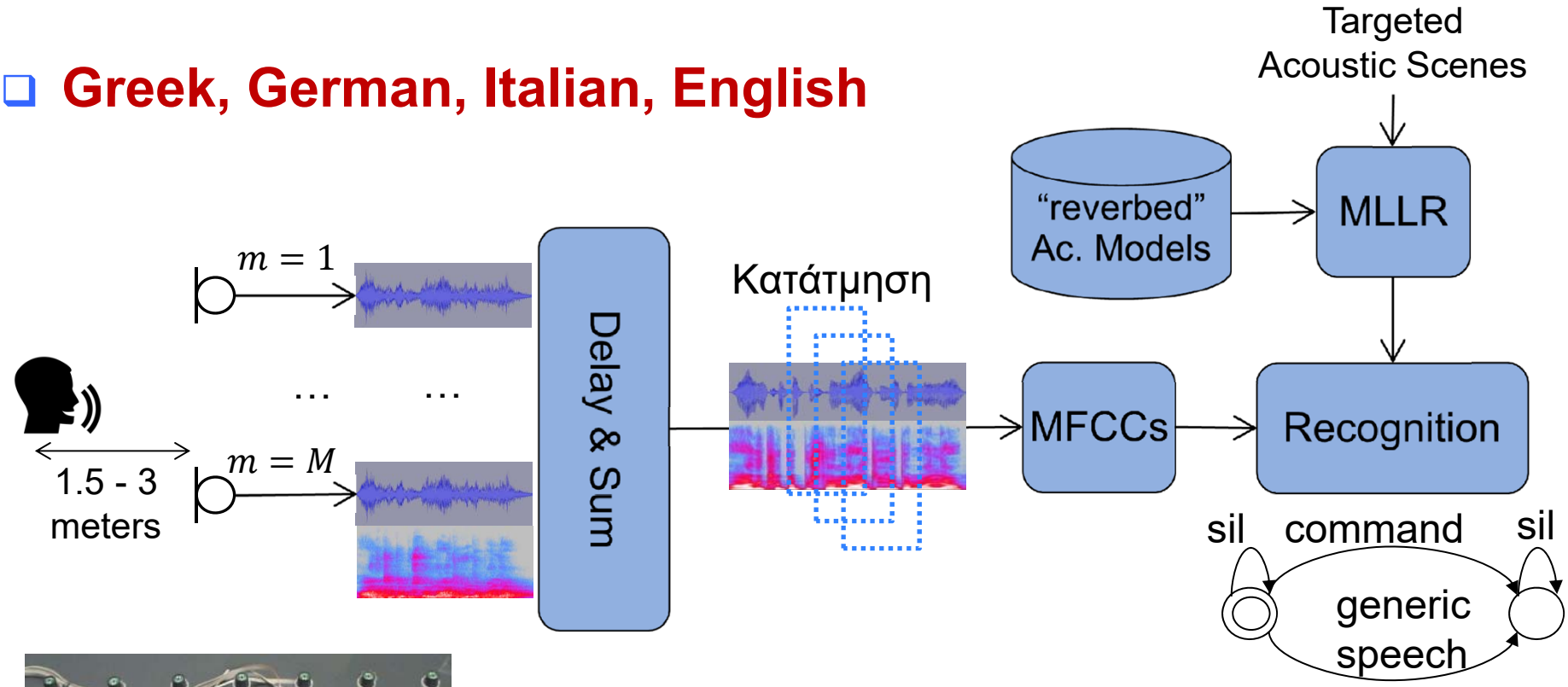
Spoken-Command Recognition Module for HRI

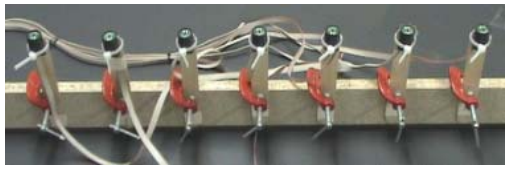
- integrated in ROS, always-listening mode, real time performance

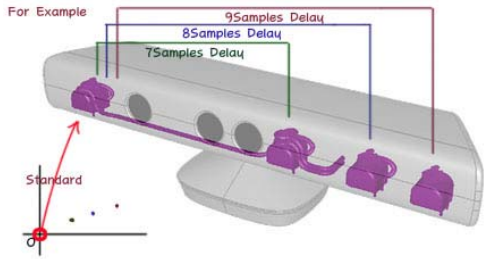


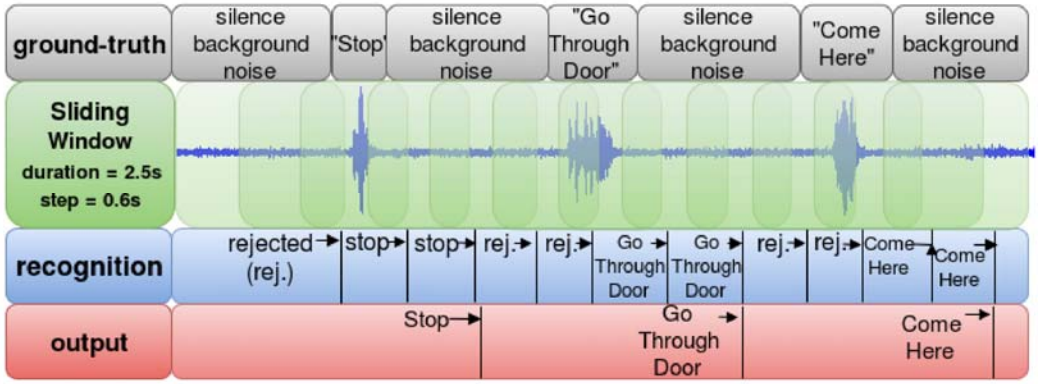
Online Spoken Command Recognition

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





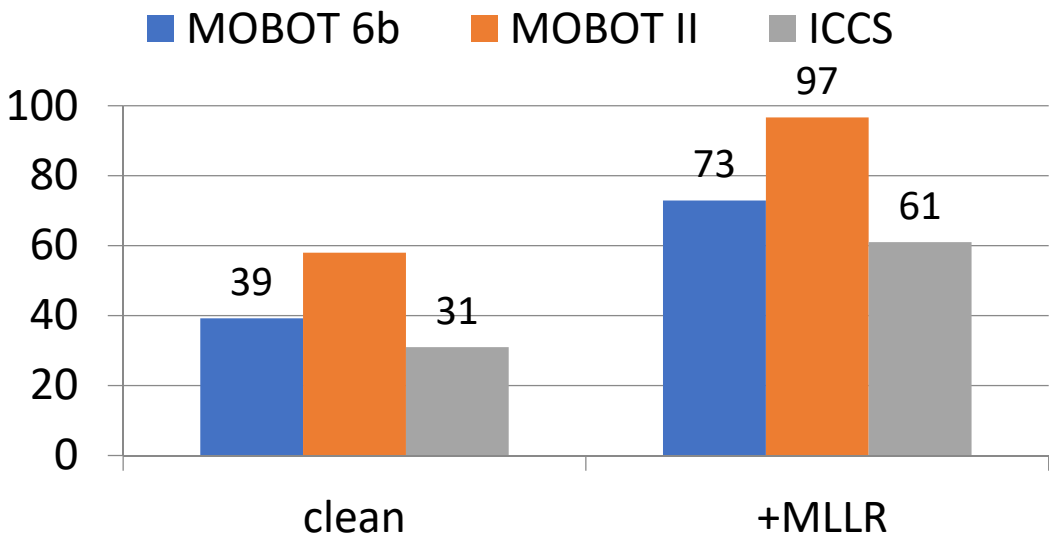


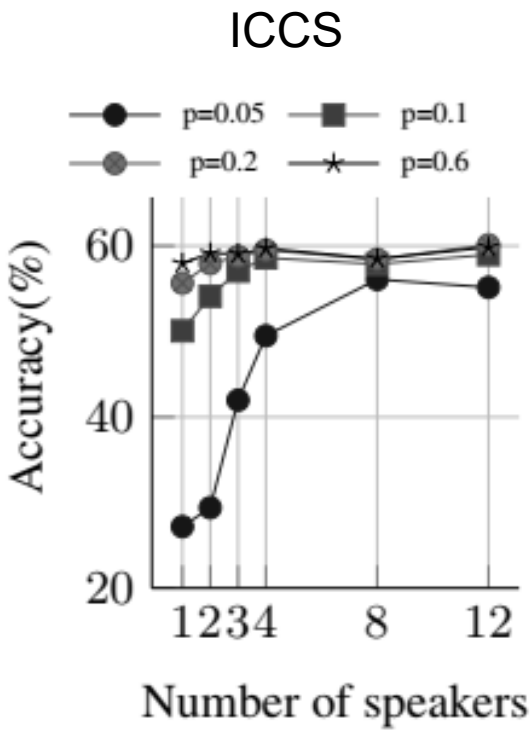


Environmental Adaptation of Acoustic Models

Σύνολο	Mic Setup	# Users	# Commands	# Repetitions	Language
MOBOT I.6b	8 MEMS	8 patients	19	3-4	German
MOBOT II	8 MEMS	8 healthy	21	5	German
ICCS	8 MEMS	13 healthy	19	~74	Greek

- round-robin leave-one-out adaptation/testing



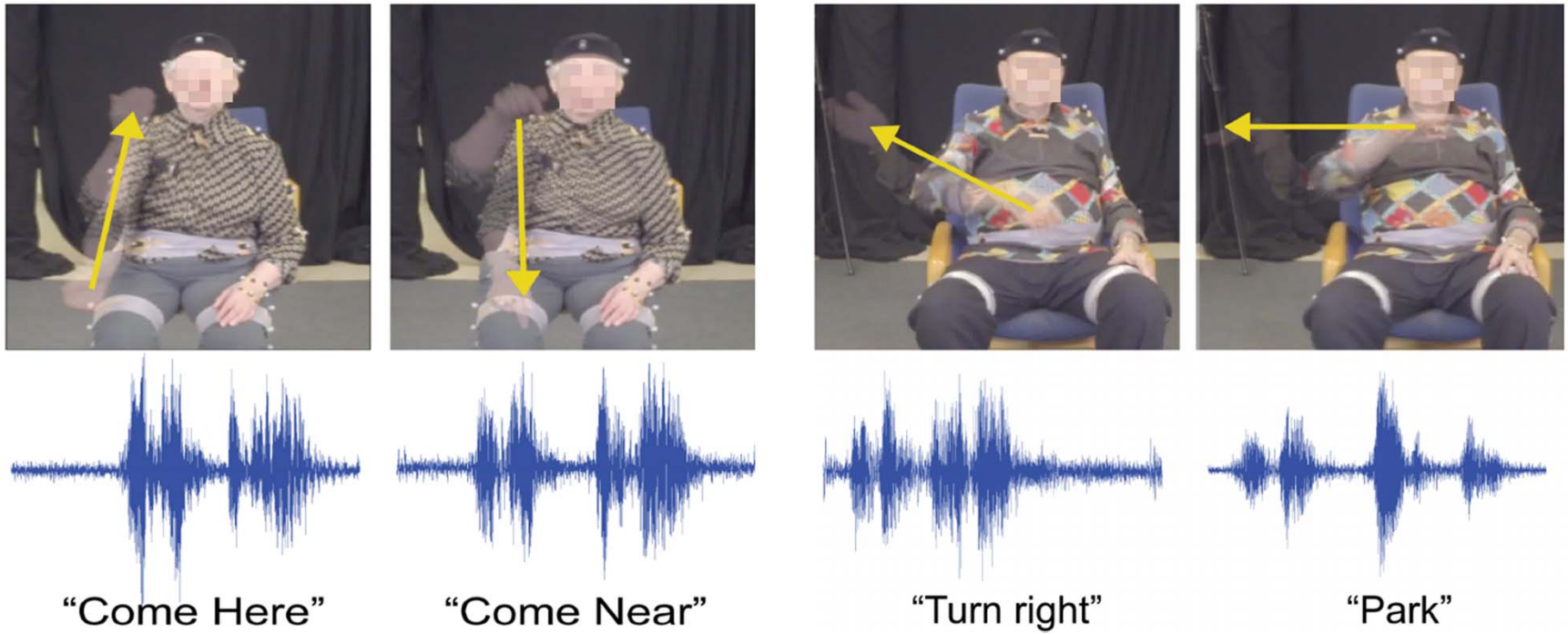


Audio-visual Fusion for Multimodal Gesture Recognition

Multimodal fusion: Complementarity of visual and audio modalities

Similar audio,
distinguishable gesture

Distinguishable audio,
similar gesture



Audio-Visual Fusion: Hypotheses Rescoring

speech & gesture recognition

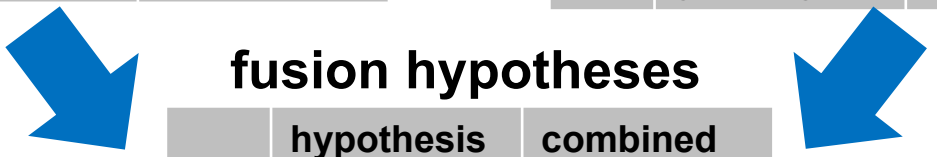
spoken commands hypotheses

N-best

	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

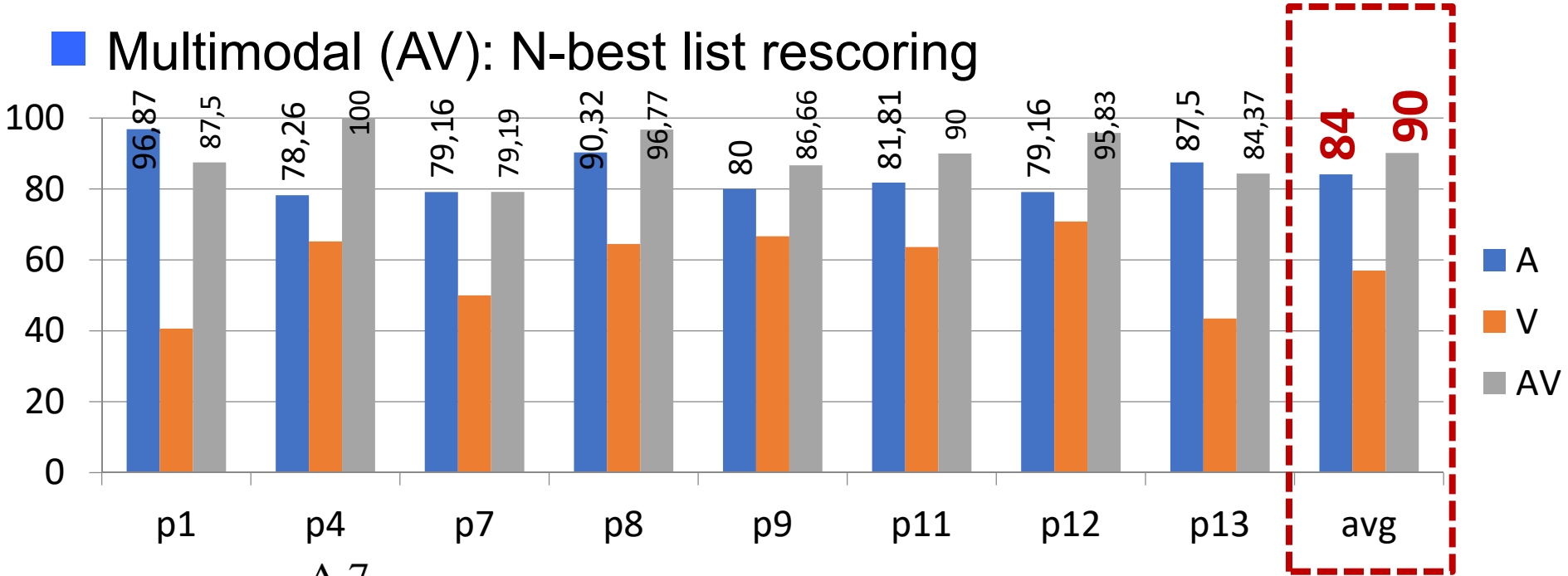
	hypothesis	combined score
F1	Stop	0.205
F2	help	0.196

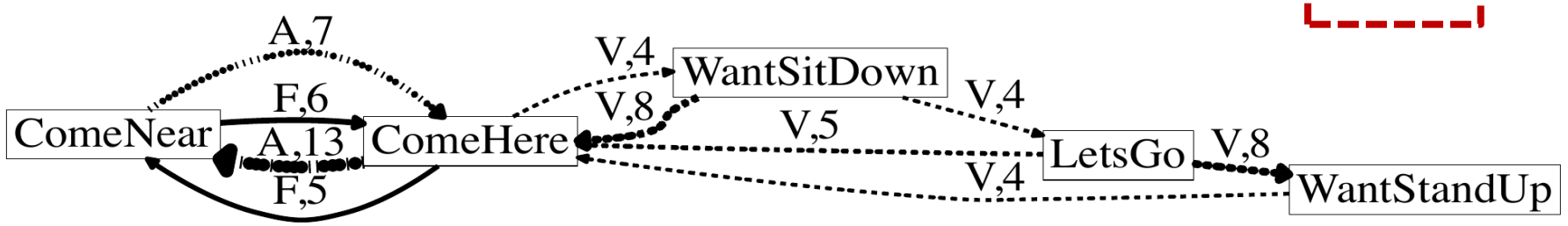
$$\mathrm{MAX}(w_a \times score(A_1) + w_v \times score(V_3), w_a \times score(A_2) + w_v \times score(V_1))$$

w_a, w_v : modality weights

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

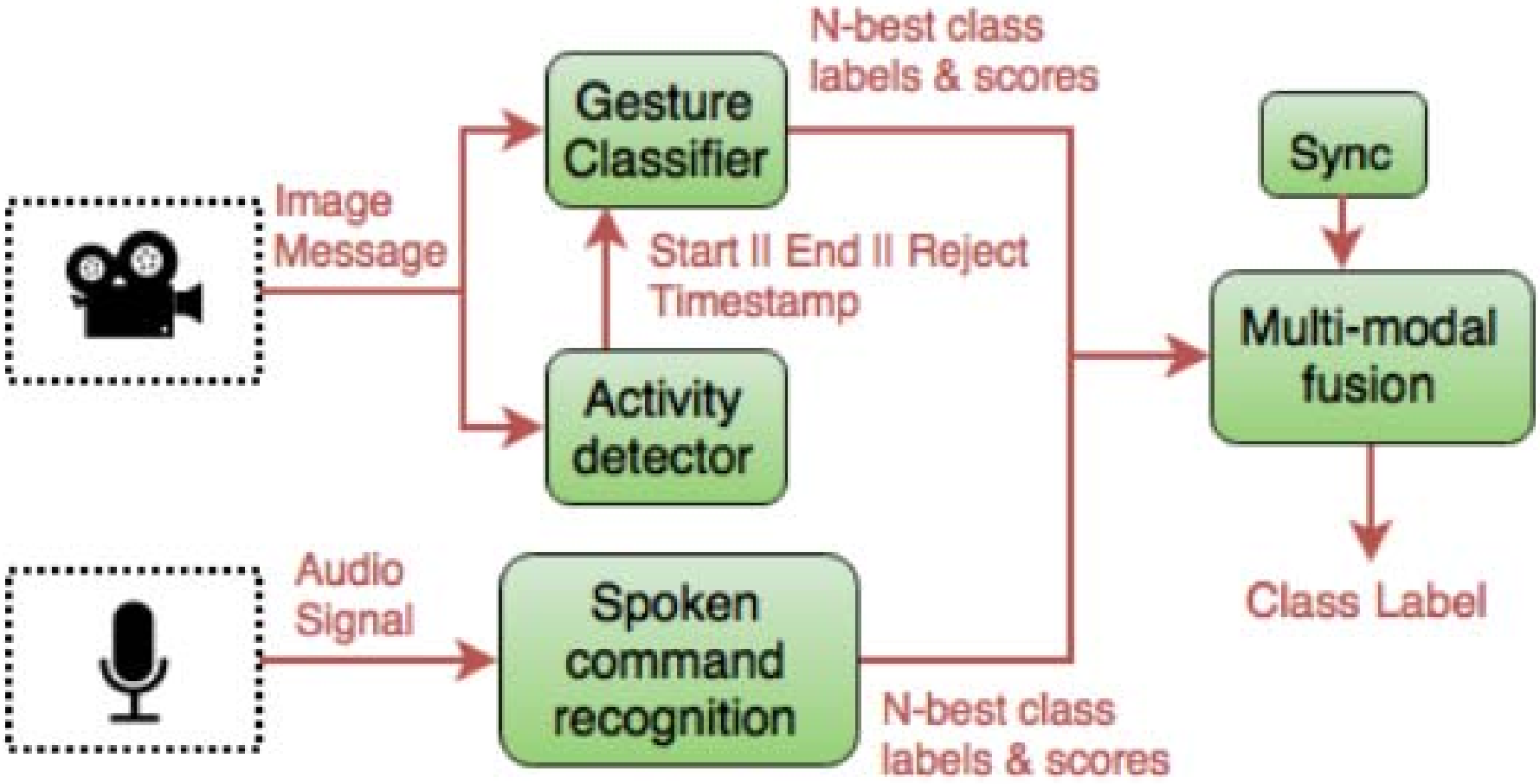




Multimodal confusability graph

HRI Online Multimodal System Architecture

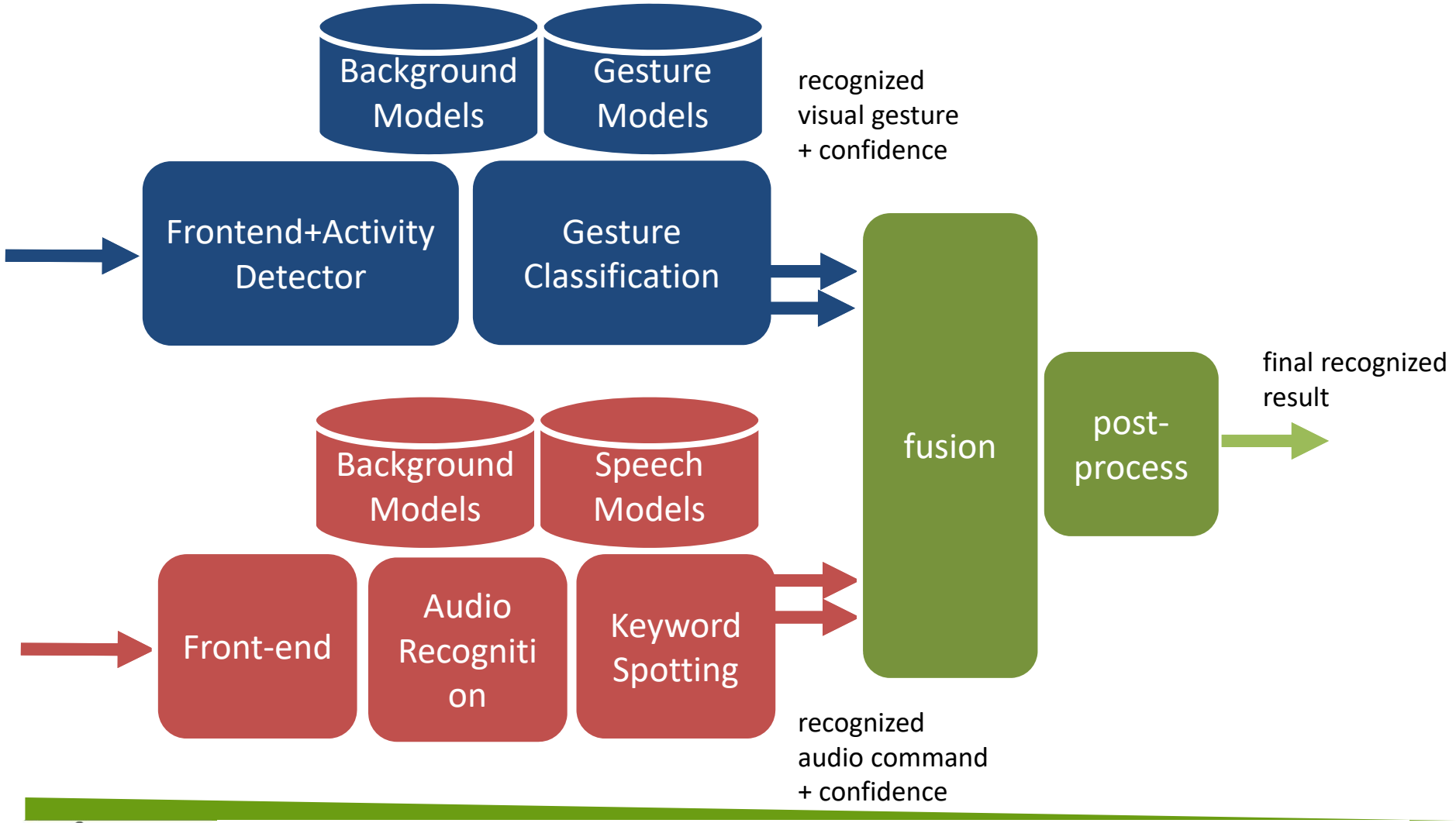
- ROS based integration
  - Spoken command recognition node
  - Activity detection node
  - Gesture classifier node
  - Multimodal fusion node
- Communication using ROS messages



Audio-Visual gesture recognition

Online processing system – Open Source Software

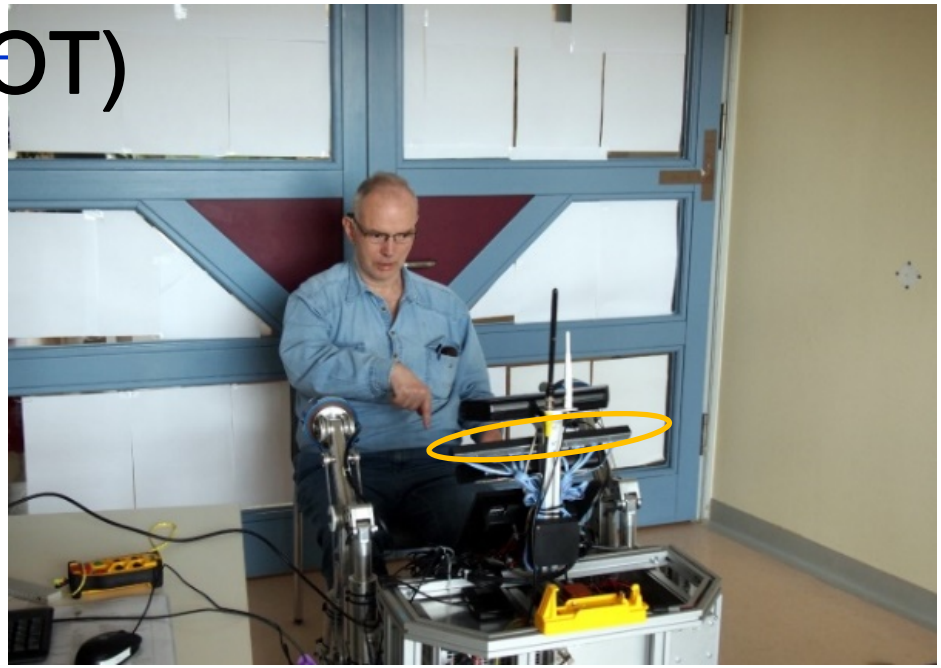
http://robotics.ntua.gr/projects/building-multimodal-interfaces



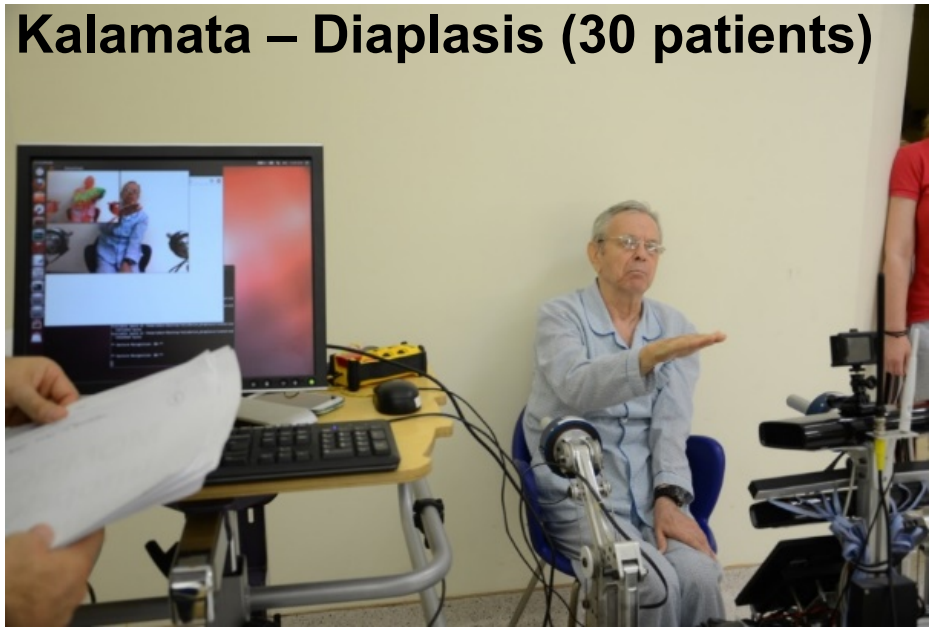
N. Kardaris, I. Rodomagoulakis, V. Pitsikalis, A. Arvanitakis and P. Maragos, *A platform for building new human-computer interface systems that support online automatic recognition of audio-gestural commands*, Proc. ACM Multimedia 2016, Amsterdam, The Netherlands, 15-19 Oct. 2016.

Clinical Studies (MOBOT)

Heidelberg – Bethanien (19 patients)



Kalamata – Diaplasis (30 patients)

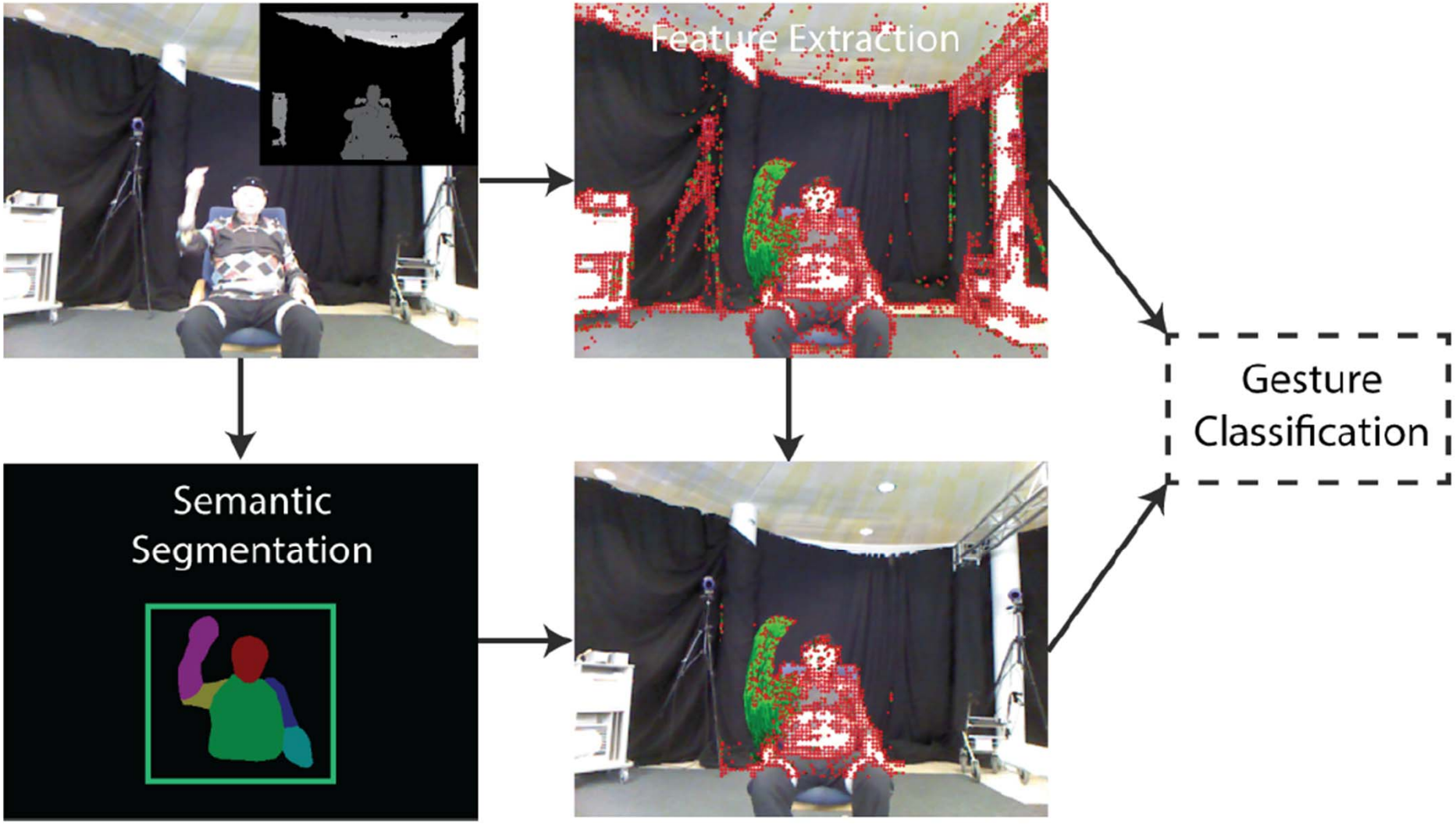


Speech, Gestures, Combination: 3 repetitions of 5 commands

Validation experiments (Bethanien, Heidelberg):

Audio-Gestural recognition in action (1/2)

Visual Synergy: : SS+GR

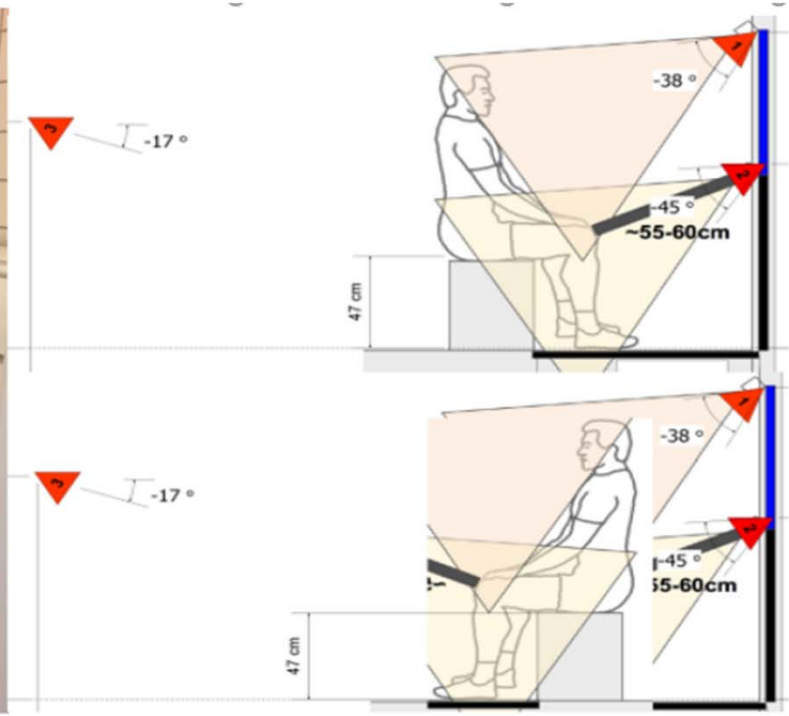


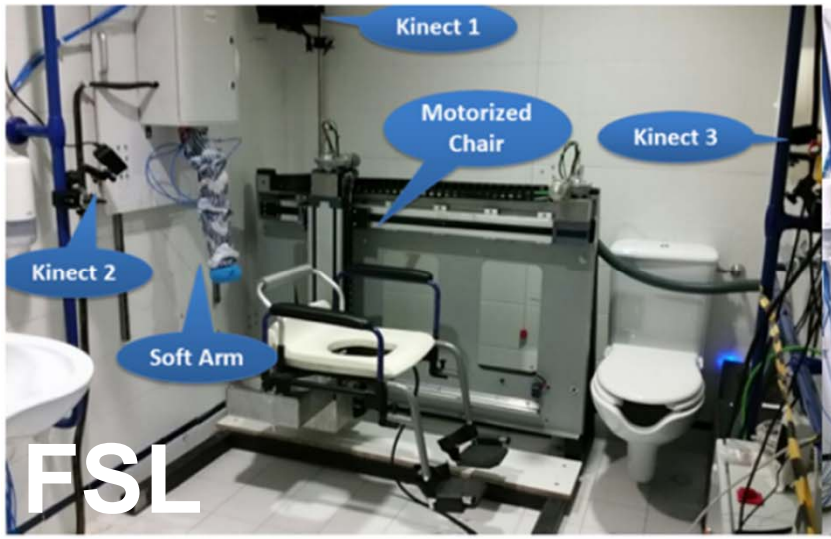
foreground/background+gesture recognition

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.

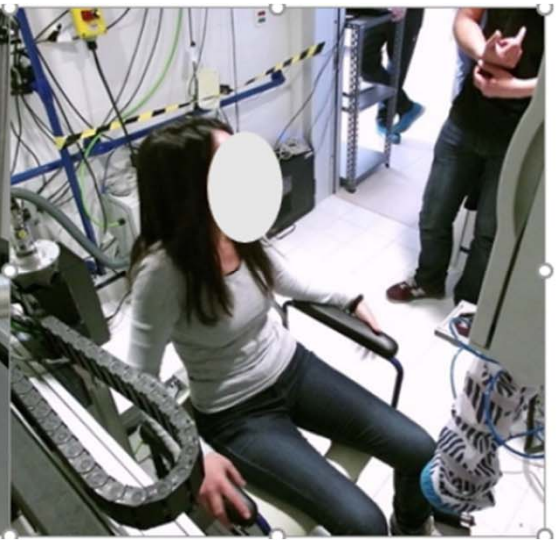
Clinical Studies (I-Support)

Bethanien





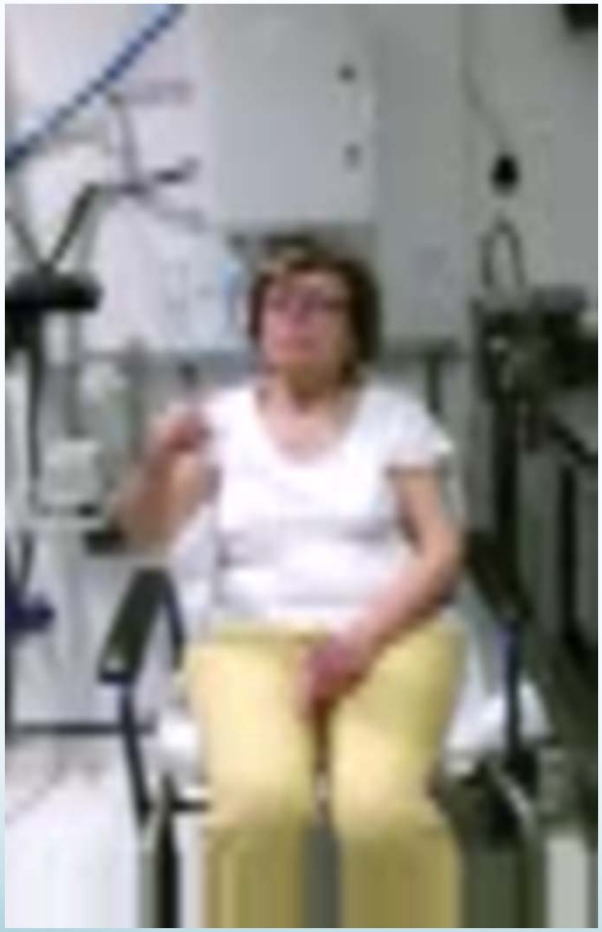
FSL



Washing back/legs: Sequences of 7 commands

FSL validation-1

Kinect 3 view for “Washing the Back”



Bethanien validation-2

I-SUPPORT - Entwicklung eines intelligenten robotischen Duschsystems

Projekttitel:	ICT-Supported Bath Robots (I-SUPPORT)
Projektnummer:	643666
Projektstart:	1. März 2015
Projektdauer:	36 Monate
Projektbudget:	3,5 Mio
Projektkoordination:	ROBOTNIK, Spanien
Projektpartner:	9 Europäische Institutionen aus 5 EU-Ländern
Projekt-URL:	www.i-support-project.eu
Förderung:	EU-Kommisssion
EU Projektmanager:	Jan Komarek

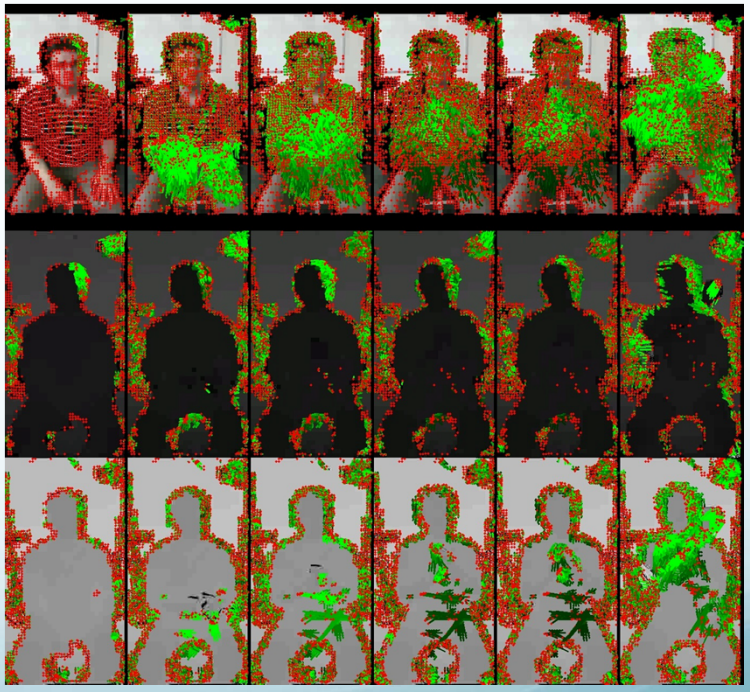
Gesture Recognition – Depth Modality

- 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** (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**

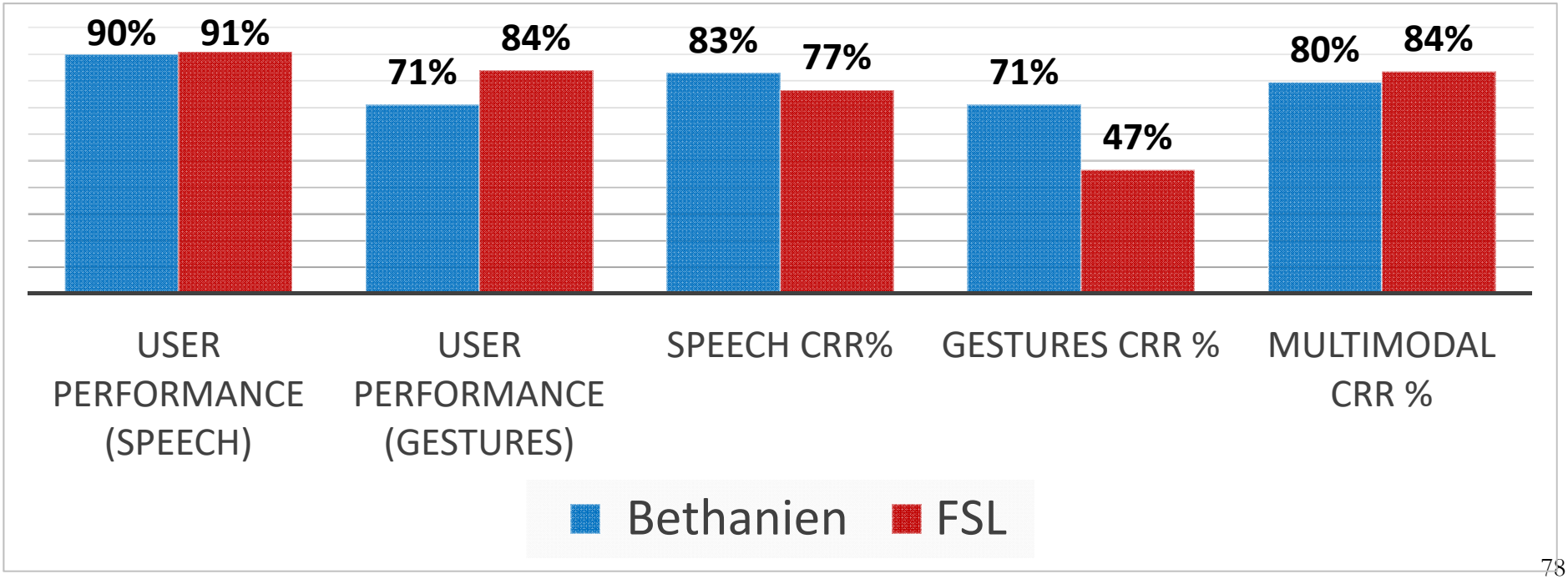
- **KIT Dataset (**8u, 8/10g)
  - Two different setups
  - Average gesture recognition accuracy:
    - Legs (8 gestures): 83%
    - Back (10 gestures): 75%

- **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%

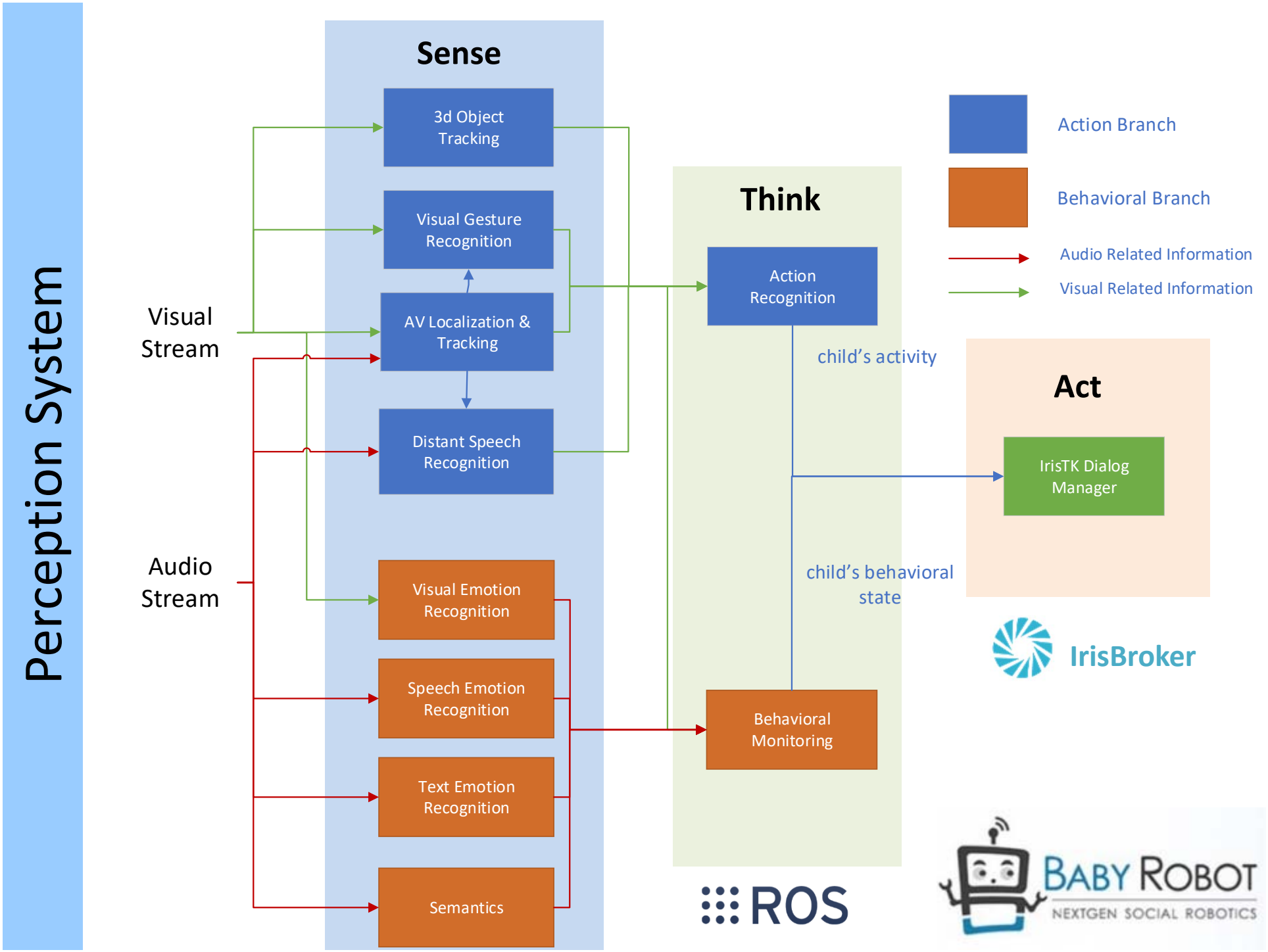
I-Support validation: Online Multimodal Recognition

Hospital	# Patients	# Commands	Language
Bethanien	29	5	German
FSL	25	5	Italian

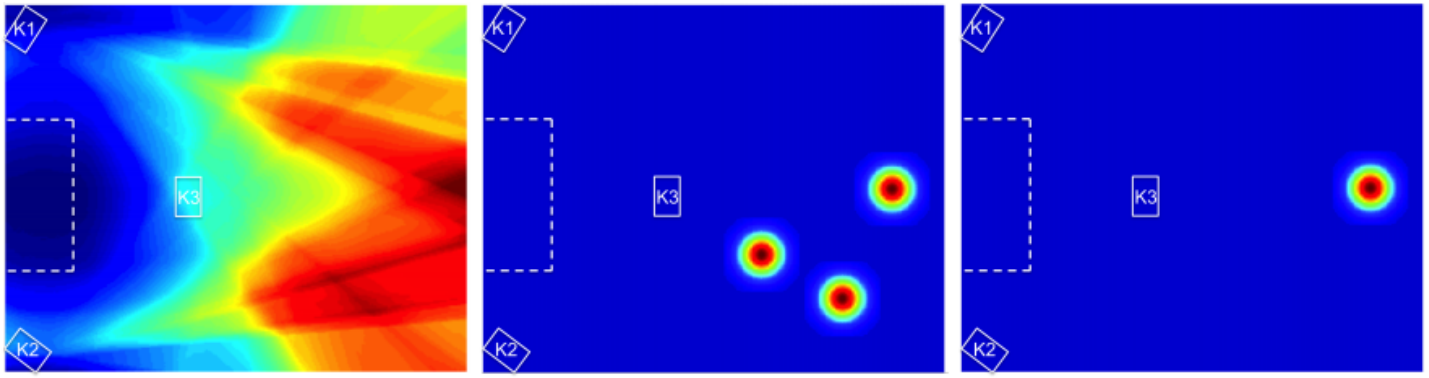
$$\text{CRR} = \frac{\#\ commands\ correctly\ recognized\ by\ system}{\#\ commands\ correctly\ performed\ by\ user}$$



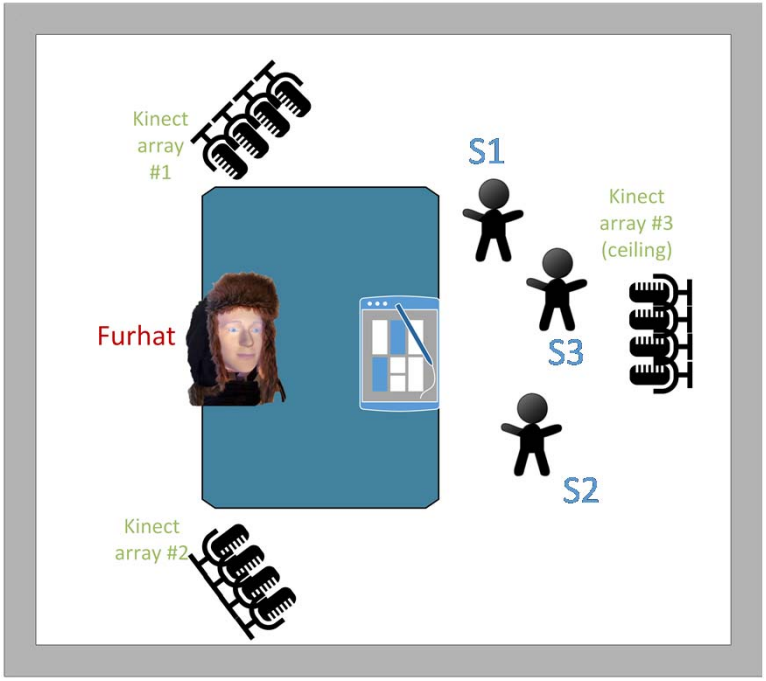
EU project BabyRobot: Experimental Setup Room


Perception System
Sense
3d Object Tracking
Visual Gesture Recognition
AV Localization & Tracking
Distant Speech Recognition
Visual Emotion Recognition
Speech Emotion Recognition
Text Emotion Recognition
Semantics
Visual Stream
Audio Stream
Think
Action Recognition
child's activity
Behavioral Monitoring
child's behavioral state
Act
IrisTK Dialog Manager
IrisBroker
Action Branch
Behavioral Branch
Audio Related Information
Visual Related Information
ROS
BABY ROBOT
NEXTGEN SOCIAL ROBOTICS

AudioVisual Localization Evaluation



- Track multiple persons using Kinect skeleton.
- Select the person closest to the auditory source position.
- Rcor: percentage of correct estimations (deviation from ground truth less than 0.5m)
  - Audio Source Localization: 45.5%
  - Audio-Visual Localization: 85.6%



Multiview Children-Robot Interaction: Demo

Multi-view Gesture Recognition-Children vs Adults

- different training schemes
  - Adults models
  - Children models
  - Mixed model

		Gesture Recognition -Training scheme		
		Adults	Children	Mixed
Test		Acc.	Acc.	Acc.
Adults	Kinect #1	84.79	60.21	87.81
	Kinect #2	89.27	53.13	92.19
	Kinect #3	85.42	55.63	82.08
	Avg	86.49	56.32	87.36
	Fuse	92.19	62.08	**95.10**
Children	Kinect #1	60.42	76.85	77.31
	Kinect #2	46.99	67.82	68.75
	Kinect #3	42.36	68.29	70.83
	Avg	49.92	70.99	72.30
	Fuse	56.25	**83.80**	80.09

A. Tsiami, P. Koutras, N. Efthymiou, P. Filntisis, G. Potamianos, P. Maragos, "*Multi3: Multi-sensory Perception System for Multi-modal Child Interaction with Multiple Robots*", ICRA 2018.

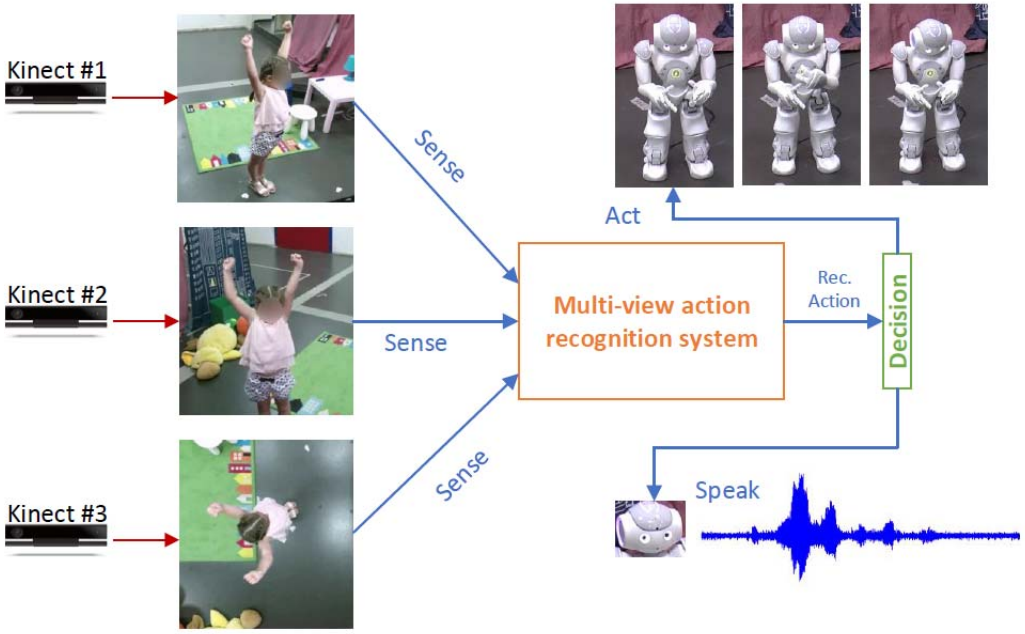
BABY ROBOT
NEXTGEN SOCIAL ROBOTICS

Spoken Command Recognition- Children vs Adults

- different training schemes
  - Adults models
  - Children models
  - Mixed model

		DSR-Adaptation scheme			
		No-adapt	Adults	Children	Mixed
Test		SCOR	SCOR	SCOR	SCOR
Adults	Kinect #1	91.76	98.95	94.52	98.69
	Kinect #2	90.60	98.70	90.99	97.85
	Kinect #3	91.39	98.95	94.11	98.75
	Avg	91.25	**98.87**	93.20	98.43
	Fuse	92.41	**99.82**	94.42	99.77
Children	Kinect #1	70.53	72.31	95.95	82.95
	Kinect #2	72.48	73.85	95.95	82.52
	Kinect #3	66.83	67.63	94.60	80.70
	Avg	69.95	71.20	**95.50**	82.06
	Fuse	64.17	66.02	**98.97**	95.51

BABY ROBOT
NEXTGEN SOCIAL ROBOTICS


Kinect #1
Kinect #2
Kinect #3
Sense
Sense
Sense
Act
Multi-view action recognition system
Rec. Action
Decision
Speak


Kinect #1
Kinect #2
Kinect #3

Conclusions

- **Synopsis**:
  - Audio-visual saliency and fusion for improved detection and recognition
  - More Big Data → Needs for Summarization
  - Multimodal Action Recognition and Human-Robot Interaction
    - Gesture Recognition
    - Spoken Command Recognition
    - Gait Analysis

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

For more information, demos, and current results:

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

Collaborators and References

Arvanitakis, Antonis
Dometios, Thanos
Efthymiou, Niki
Filntisis, Panagiotis
Kardaris, Nikos
Katsamanis, Nasos
Koutras, Petros
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