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)

# Synergy between Computer Vision and Language Processing for Multimodal Perception and Understanding

**Petros Maragos**

**SLT-2018, Keynote**: IEEE Workshop on Spoken Language Technology, Dec. 2018

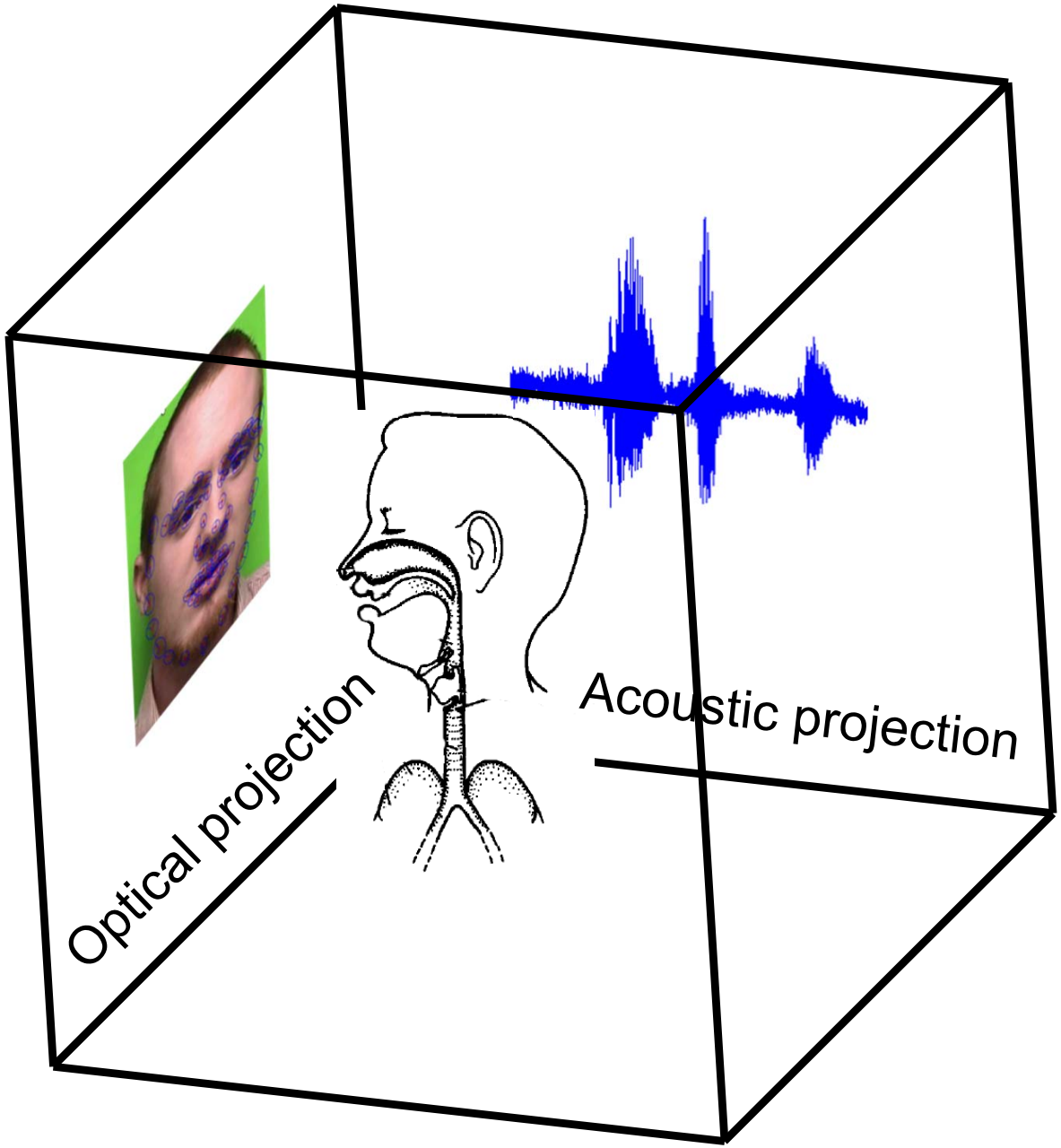
# Talk Outline

- Audio-Visual Perception and Fusion
- Applic 1: A-V-T Saliency & Video Summarization
- Applic 2: Visual-Textual Concept Learning in Videos with Weakly Supervised Techniques
- Applic 3: Audio-Gestural Recognition for Human-Robot Interaction

# 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:

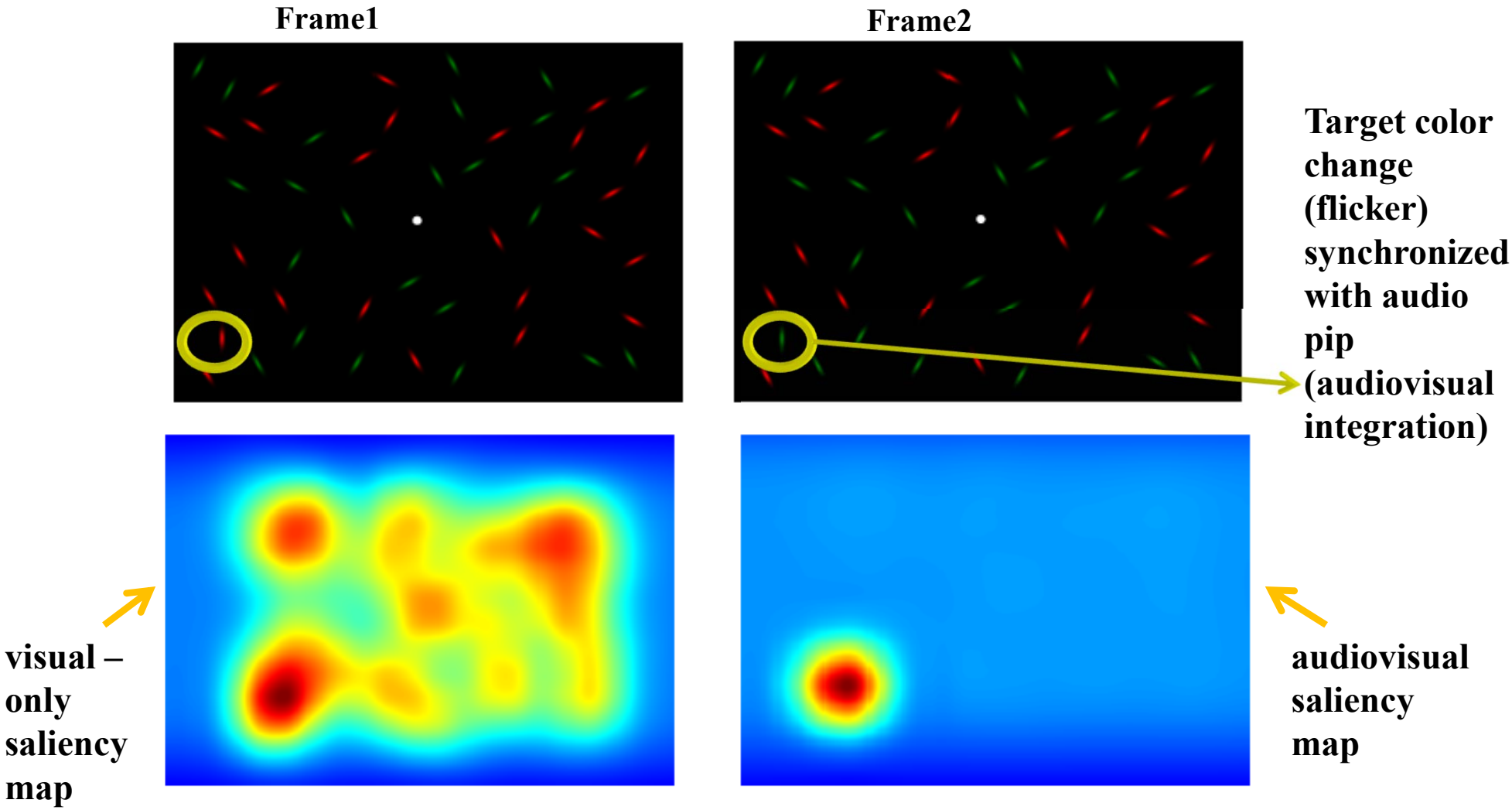
- **Spatial rule**,
- **Temporal rule**,
- **Modality Appropriateness:**
  - Visual dominance of spatial tasks.
  - Audition is dominant for temporal tasks.
- **Inverse effectiveness law:**
  - In multisensory neurons, multimodal stimuli occurring in close space-time proximity evoke supra-additive responses. 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:

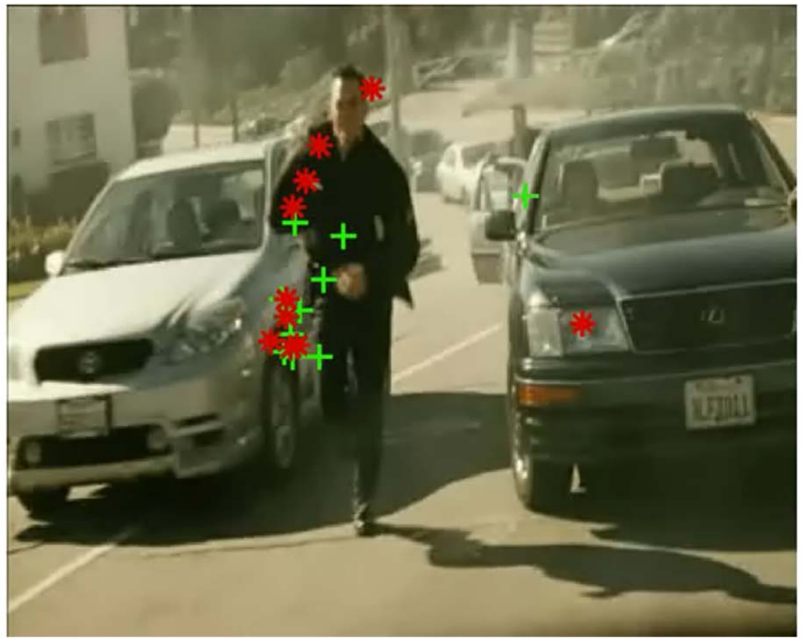
- *Static*: CCA- Canonical Correlation Analysis: e.g. "cocktail-party effect"
  Max Mutual Information
  SVMs- Support Vector Machines: kernel combination

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

# 1. Audio-Visual-Text Saliency and Video Summarization

# Human Attention and Summarization

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



- Applications
  - Systems for selecting the most important regions/segments of a large amount of visual data
  - Video/Movie Summarization
  - Frontend for other applications like action recognition.

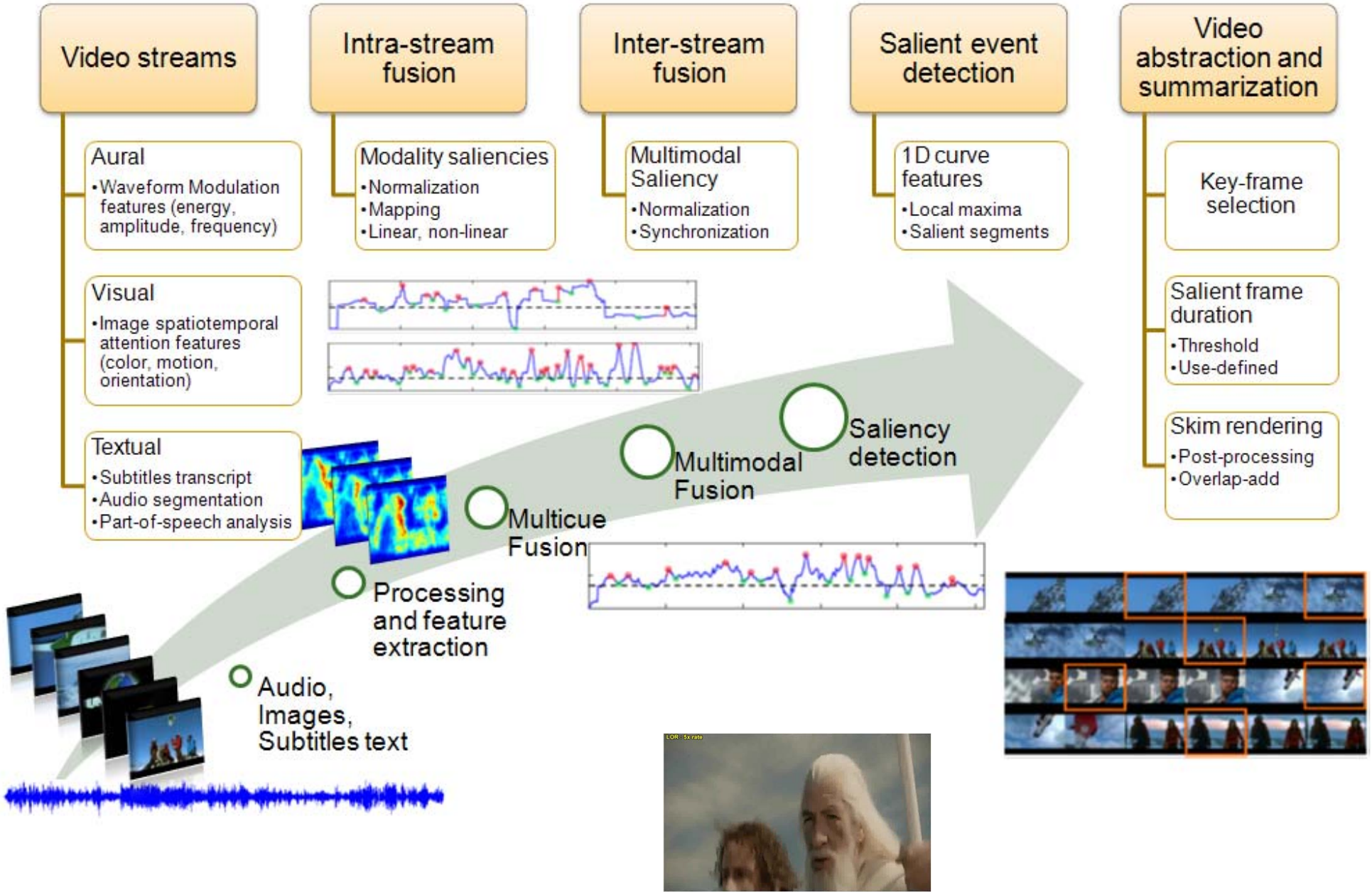
# Video Summarization

- Need for summarization:
  - 400 hours of video are uploaded to YouTube every minute
  - Need to search for relevant content quickly in large amounts of video
- Summarization goal: produce a shorter version of a video:
  - containing only the necessary and non-redundant information required for context understanding
  - covering the interesting and informative frames or segments
  - without sacrificing much of the original enjoyability

# Multimodal Saliency & Movie Summarization

**COGNIMUSE: Multimodal Signal and Event Processing In Perception and Cognition**

website: http://cognimuse.cs.ntua.gr/



# Demo: Movie Summaries

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

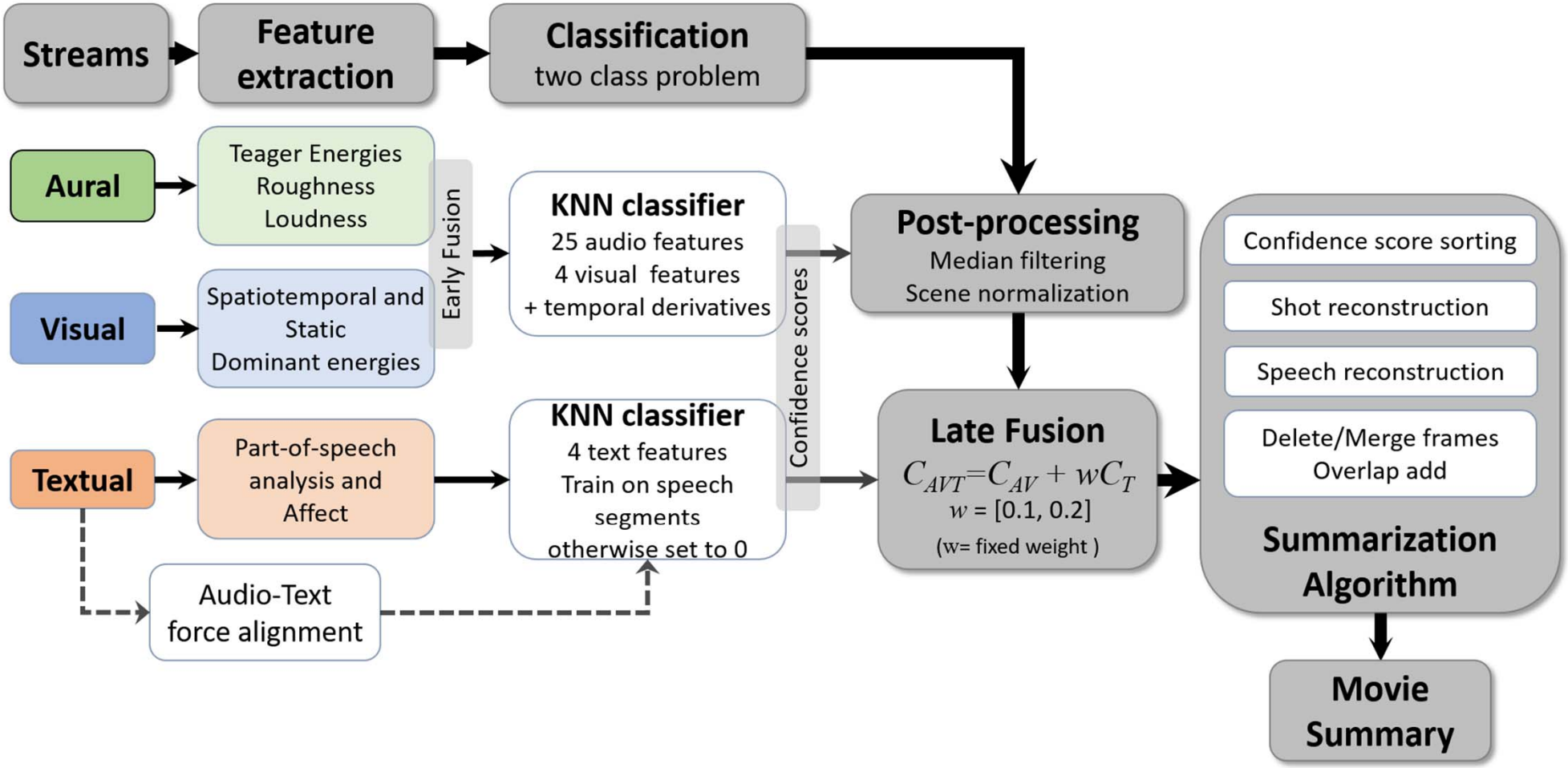
**LOR VA-SH-F,** rate: x5 (6:50 min from 37:33 min)
Inform: 78.7 %
Enjoy: 80.9 %


**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. ]

# Summarization System MovieSum 2

(with 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!"*

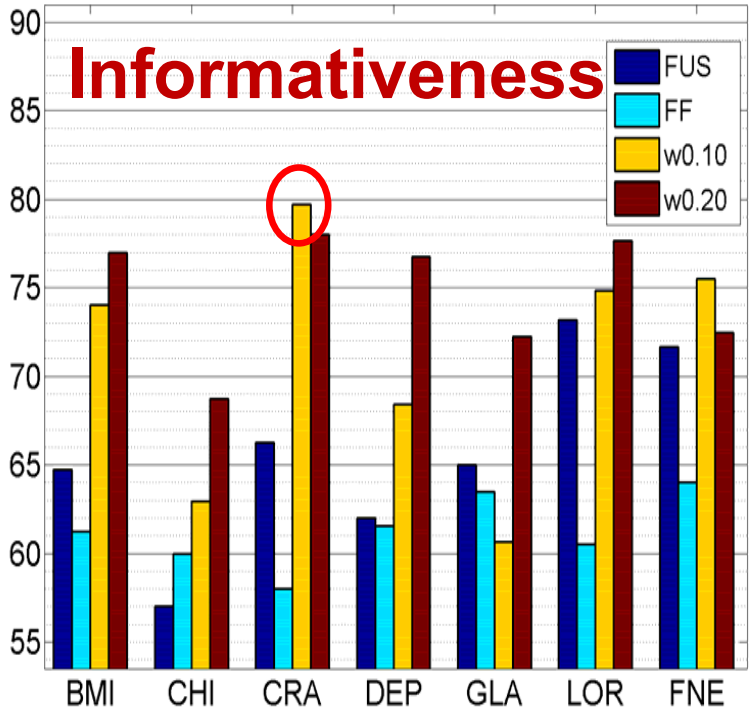
Cross-media relations

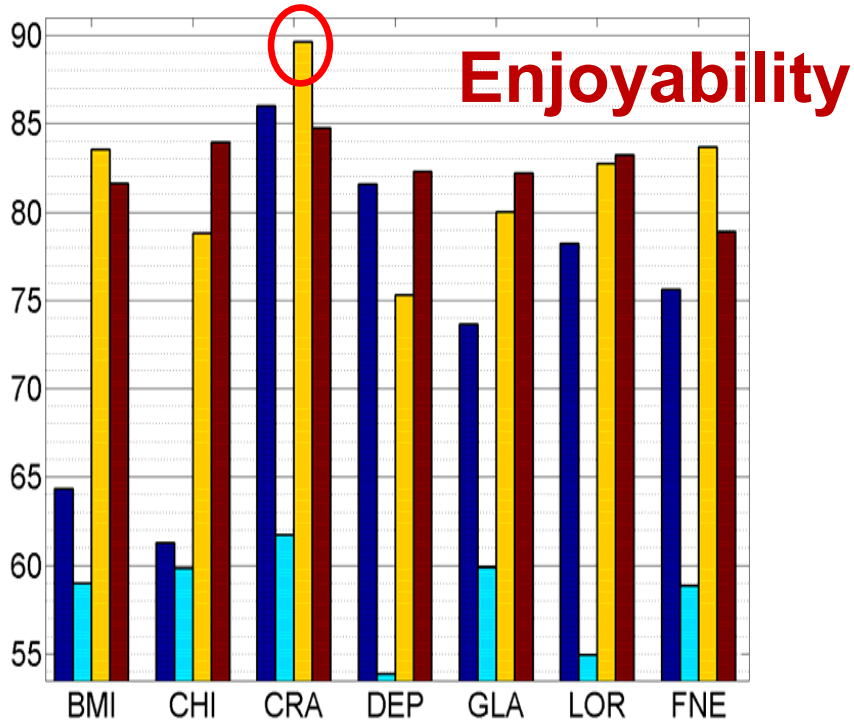

Audio & Visual events


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 ]

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




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

Evaluation on:

**$T_W$0.1**: text weight $T_W = 0.1$
**$T_W$0.2**: 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 Summarization Approaches

- Automatic summaries can be created with:
  - key-frames, which correspond to the most important video frames and represent a static storyboard

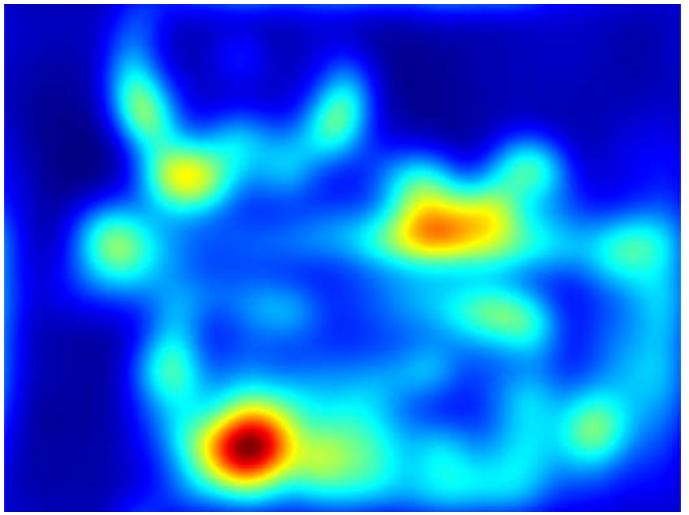
  - video skims that include the most descriptive and informative video segments

# Visual Saliency: Approaches, Evaluation

- Spatial Saliency
  - predict viewers fixations in image plane
  - static eye-tracking datasets: Toronto data set, MIT CAT200, SALICON, …
- Spatio-Temporal Saliency
  - predict viewers fixations both in space and time
  - dynamic eye-tracking datasets: CRCNS, DIEM, DHF1K, …
- Temporal Saliency
  - find the frames or segments that contain the most salient events
  - visual, audio and text streams → multimodal salient events
  - human annotated databases: COGNIMUSE multimodal video database

Original Image

Spatial Saliency Map



# Visual Saliency: Approaches, Evaluation

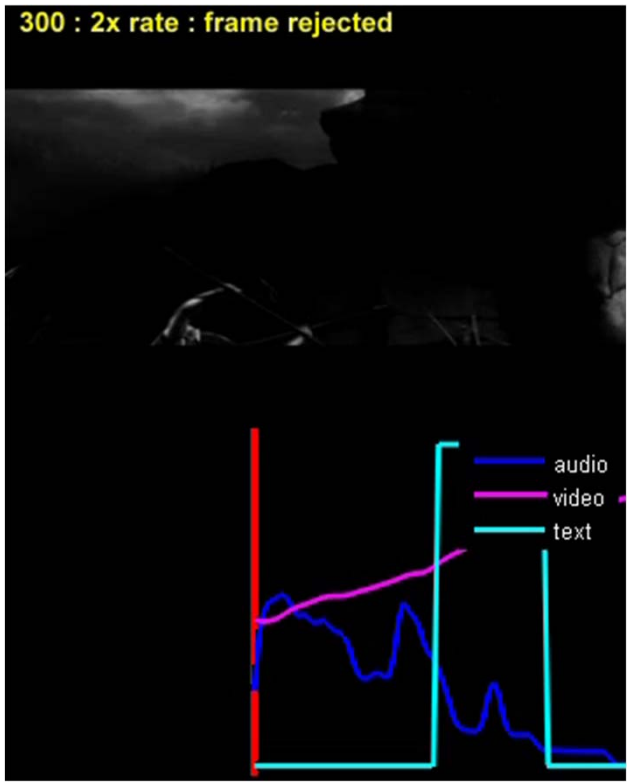
- Spatial Saliency
  - predict viewers fixations in image plane
  - static eye-tracking datasets: Toronto data set, MIT CAT200, SALICON, …
- Spatio-Temporal Saliency
  - predict viewers fixations both in space and time
  - dynamic eye-tracking datasets: CRCNS, DIEM, DHF1K, …
- Temporal Saliency
  - find the frames or segments that contain the most salient events
  - visual, audio and text streams → multimodal salient events
  - human annotated databases: COGNIMUSE multimodal video database

Spatio-Temporal Saliency Map

# Visual Saliency: Approaches, Evaluation

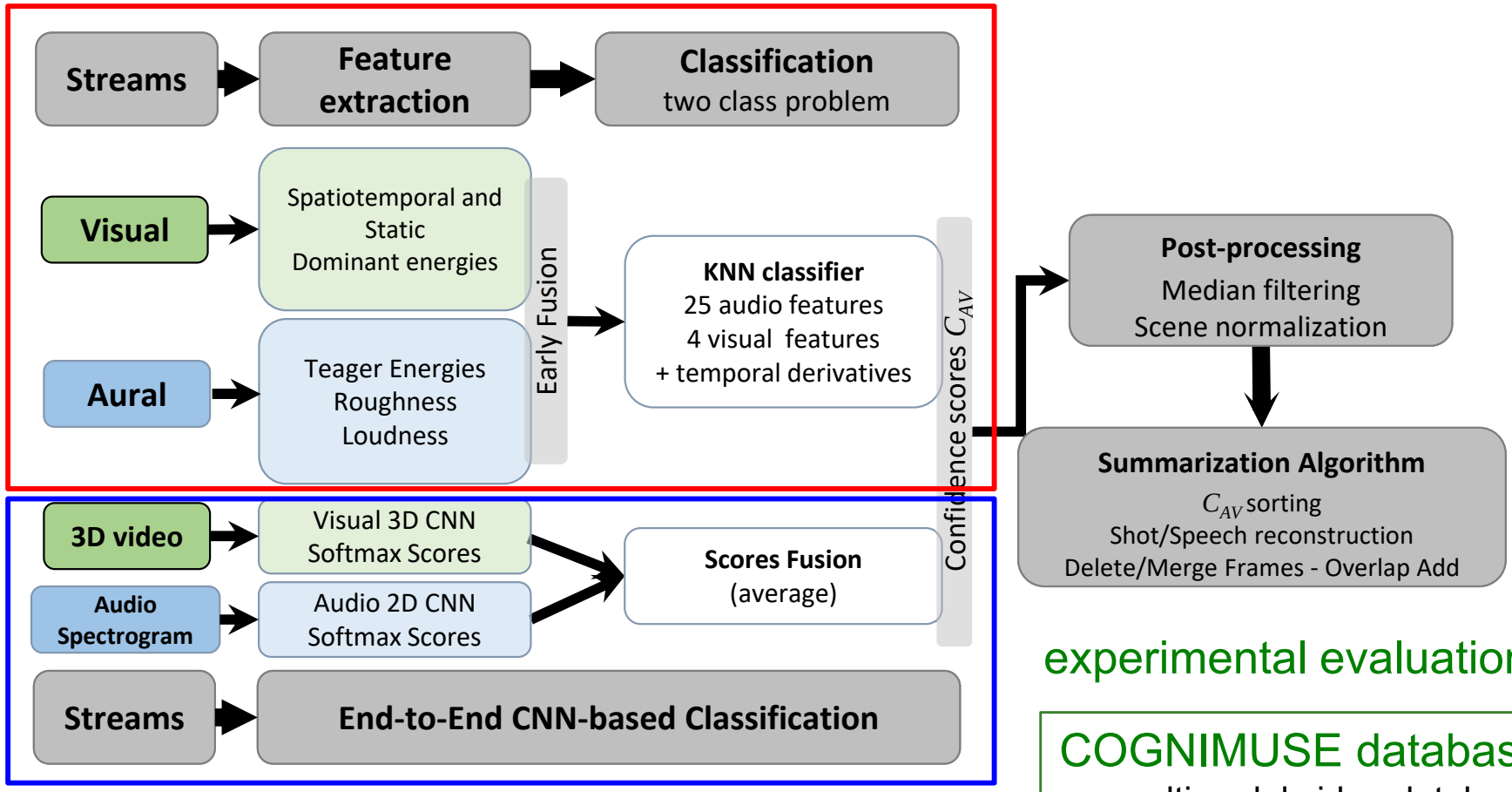
- Spatial Saliency
  - predict viewers fixations in image plane
  - static eye-tracking datasets: Toronto data set, MIT CAT200, SALICON, …
- Spatio-Temporal Saliency
  - predict viewers fixations both in space and time
  - dynamic eye-tracking datasets: CRCNS, DIEM, DHF1K, …
- Temporal Saliency
  - find the frames or segments that contain the most salient events
  - visual, audio and text streams → multimodal salient events
  - human annotated databases: COGNIMUSE multimodal video database

Multimodal Saliency Curves



# Multimodal Salient Event Detection - Contributions

handcrafted features + classification algorithms




multimodal CNN-based architectures for saliency detection

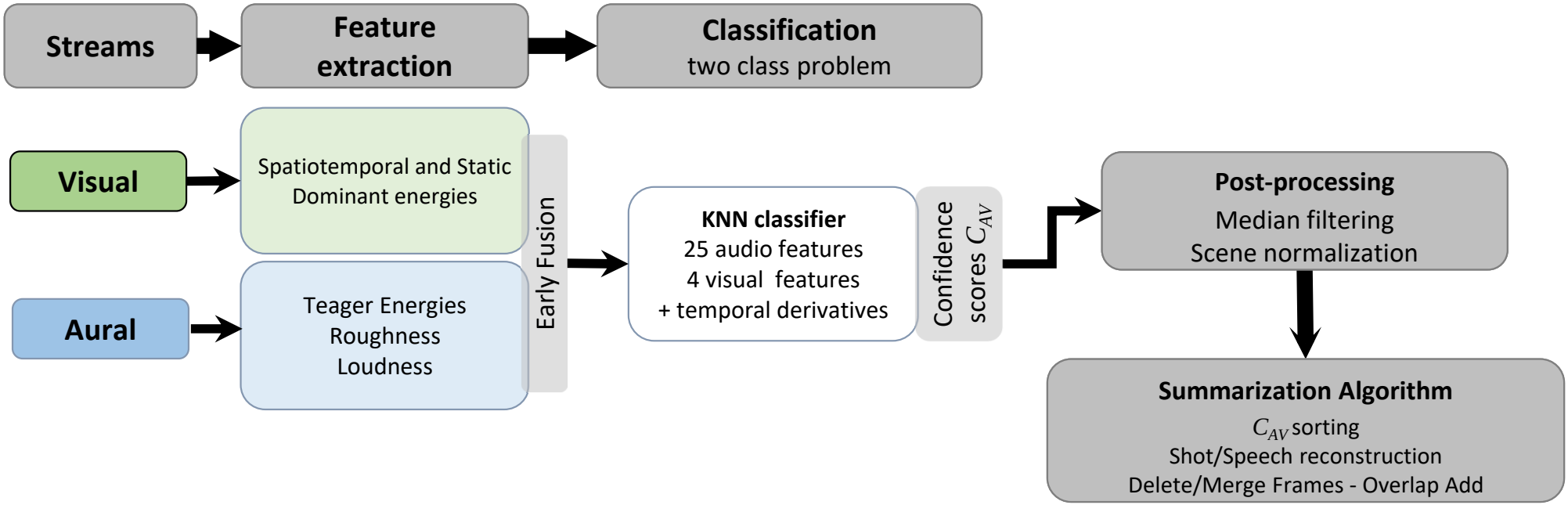
experimental evaluation

COGNIMUSE database:

- multimodal video database with multilevel saliency annotation
- *http://cognimuse.cs.ntua.gr/database*

[ P. Koutras, A. Zlatintsi and P. Maragos, IEEE Workshop on IVMSP, 2018. ]

# Hand-crafted Frontend



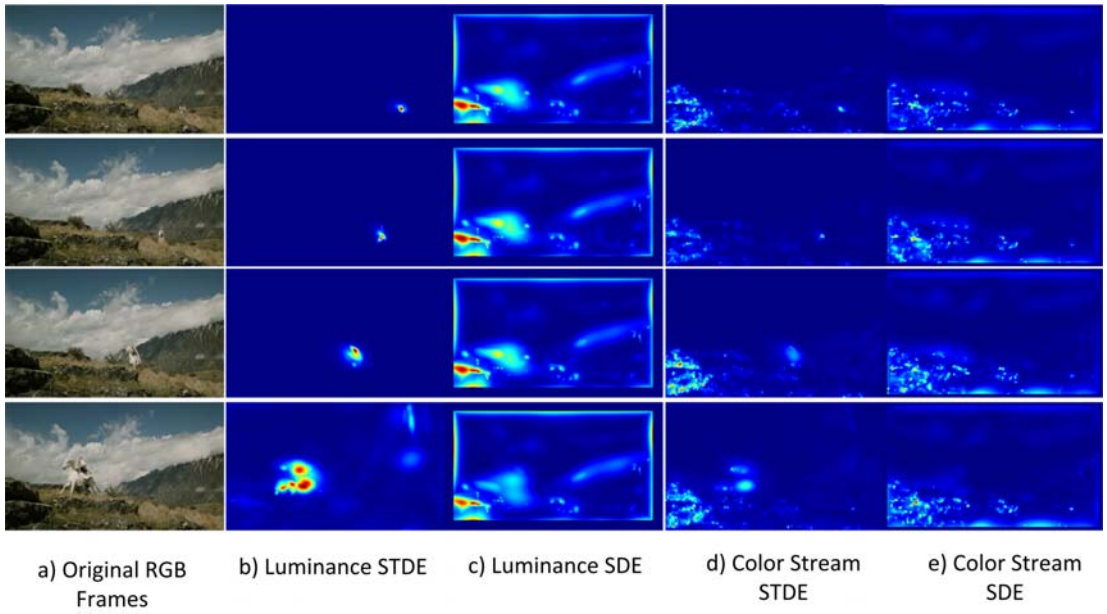

- Signal processing methods for feature extraction
  - unified energy-based framework for audio-visual saliency
  - perceptually inspired and carefully designed filterbanks
- Machine learning algorithms for salient events classification
- Postprocessing of the final scores

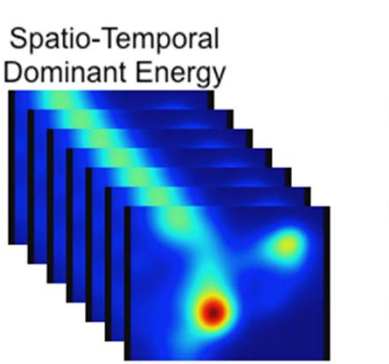
# Handcrafted Visual Saliency Model

## 3D Gabor Energy model

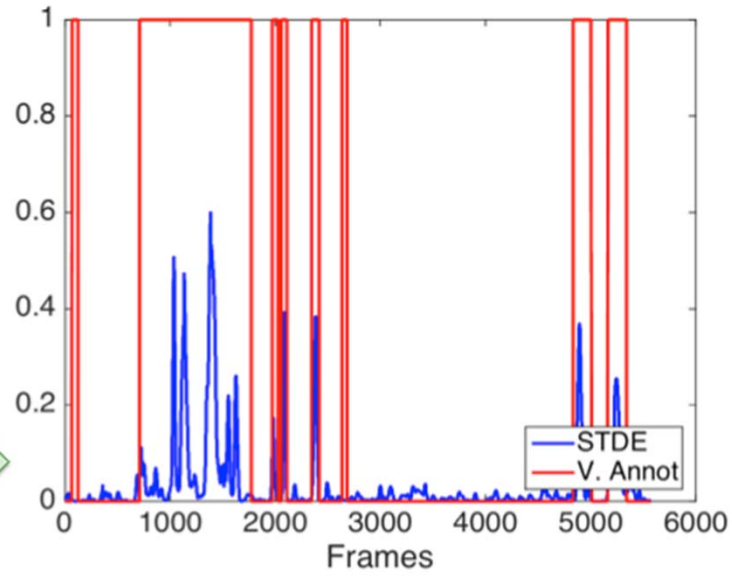
### Visual Features

- 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
  - Simple 3D to 1D Mapping
  - Mean value for each 2D frame slice of each 3D energy volume
  - 4 temporal sequences of visual feature vectors.





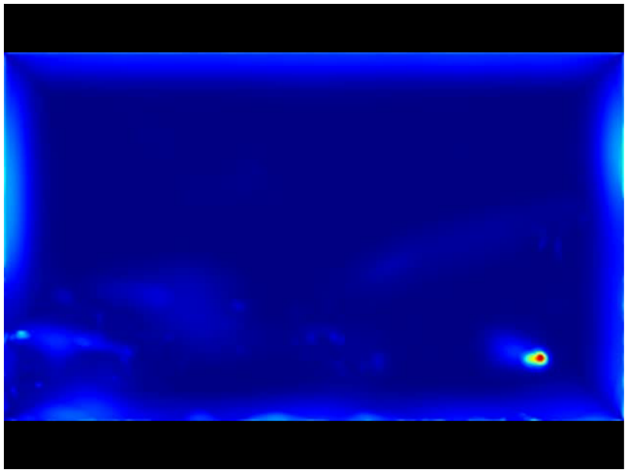



[P. Koutras and P. Maragos. *A Perceptually-based Spatio-Temporal Computational Framework for Visual Saliency Estimation*, Signal Proc.: Image Comm, 2015]

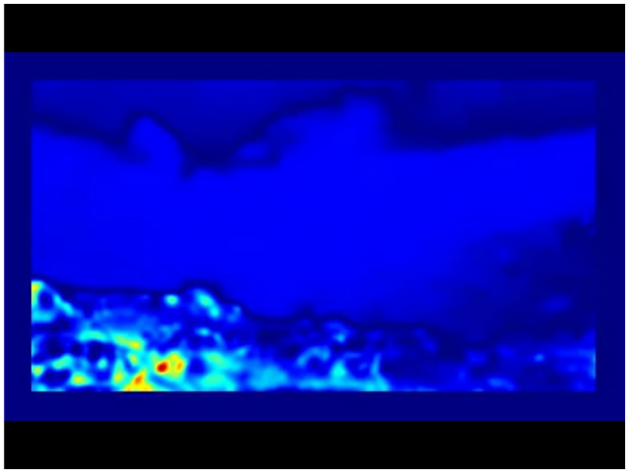
# Visual Saliency in Movie Videos - Demo

Original RGB Frames

Luminance STDE



Color SDE



COGNIMUSE Database:
Lord of the Rings: The Return of the King

# Handcrafted Audio Analysis

Teager-Kaiser Operator: $\Psi[t] = \dot{x}^2 - x\ddot{x}$

AM-FM Modulated Audio Signal *(narrow-band)*:

$$x(t) = \alpha(t)\cos\left(\int_0^t \omega(\tau)d\tau\right)$$

$$\frac{\Psi[x(t)]}{\sqrt{\Psi[\dot{x}(t)]}} \simeq |\alpha(t)| \qquad \frac{\sqrt{\Psi[\dot{x}(t)]}}{\sqrt{\Psi[x(t)]}} \simeq \omega(t)$$

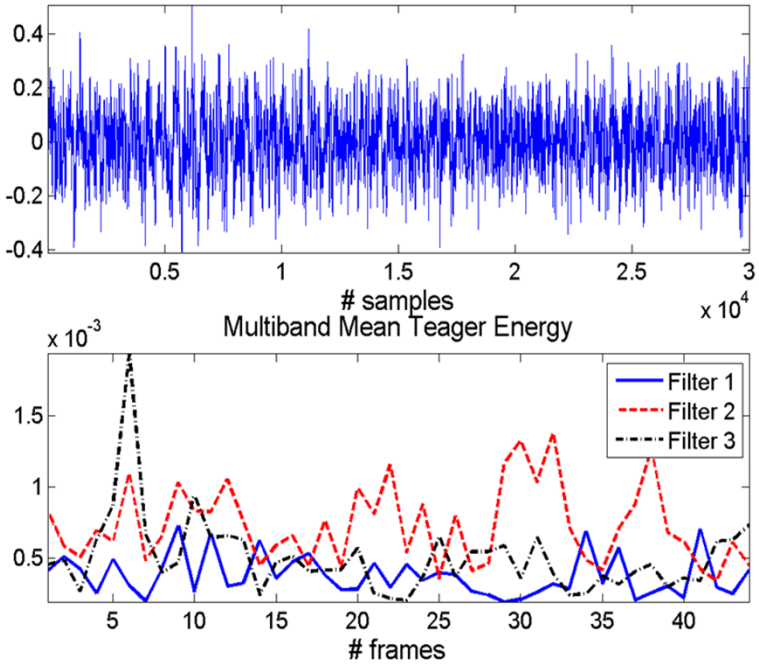
narrow-band→ Filterbank of 25 Mel arranged Gabor filters

- **Roughness** (or sensory dissonance)
  - expresses the "stridency" of a sound due to rapid fluctuations in the amplitude
- **Loudness** (perceived sound pressure level)
  - Loudness model for time-varying sounds by Zwicker & Fastl (1999)

## AM-FM model

## Auditory Features

- 25 Gabor Energies
- Roughness
- Loudness

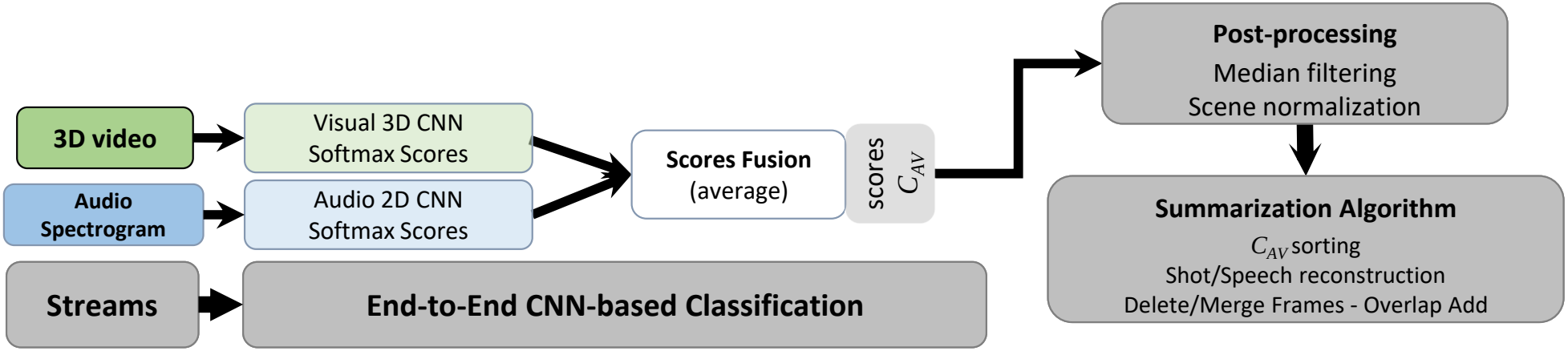



[A. Zlatintsi, E.Iosif, P. Maragos and A. Potamianos. *Audio Salient Event Detection and Summarization using Audio and Text Modalities*, EUSIPCO, 2015]

# Salient Events Classification

- Multimodal features vectors:
  - standardize features (zero mean, unit covariance)
  - compute 1st and 2nd order derivatives (deltas)
  - concatenate audio and visual features
- Classification based conventional machine learning:
  - binary classification problem (salient / non salient video segments)
  - K-Nearest Neighbor Classifier (KNN)
    - confidence scores for every classification result
    - continuous indicator function curve → represents the most salient events
- Scores post-processing
  - median filtering and scene normalization of saliency measurement
  - sorting the frames based on saliency measurement
  - summarization algorithm

# CNN-based Architectures for Saliency Detection




- Two-stream Convolutional Networks: video and audio
  - inspired from two-stream CNNs for action recognition (RGB + Flow, RGB + Depth)
  - replace the stages of feature extraction and classification with one single network
  - softmax scores of the CNN output as visual and audio saliencies
- use monomodal or multimodal annotations as ground-truth labels
  - multinomial logistic loss for binary classification:

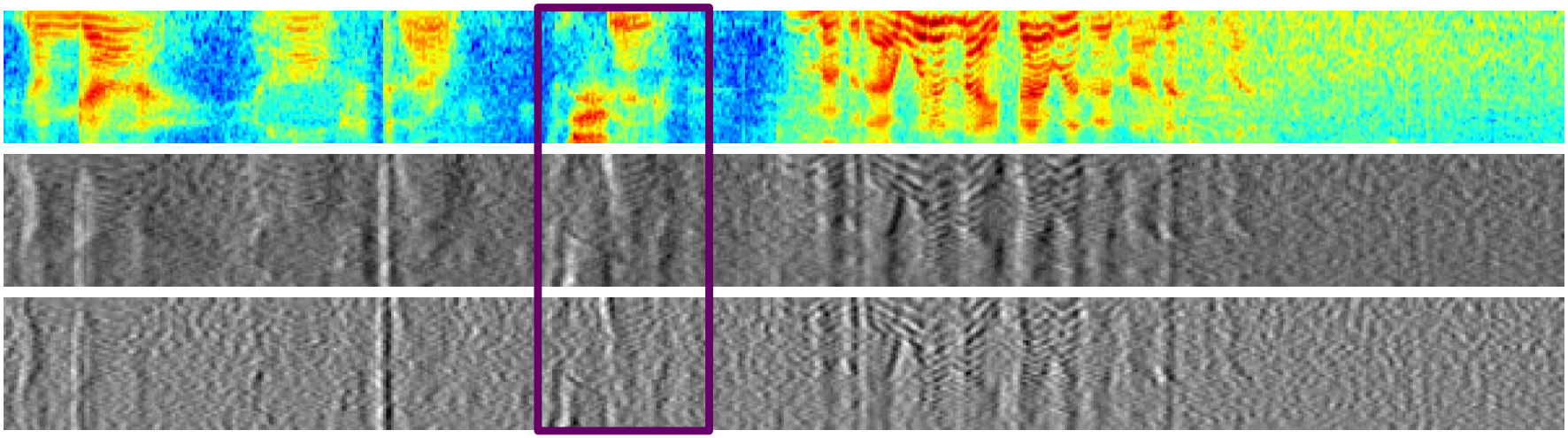
$$\mathcal{L}(\mathbf{W}) = -\sum_{j \in Y_+} \log P(y_j = 1 | X; \mathbf{W}) - \sum_{j \in Y_-} \log P(y_j = 0 | X; \mathbf{W})$$

- employ trained models for computing saliency curves in a new video
  - same postprocessing of the final scores as in handcrafted frontend

# CNN Architecture for Visual Saliency

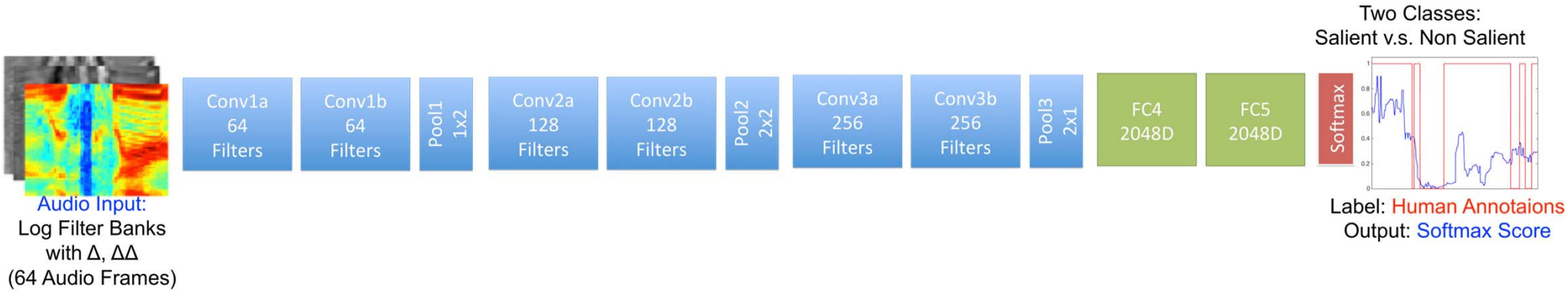

- deep end-to-end CNN architecture
  - input: split video into 16-frame RGB clips
    - total ~18000 clips for training
  - output: softmax score as visual saliency curve
- learn filterbank parameters as a sequence of 3D convolutional networks (C3D)
  - convolutions and pooling operations are applied inside spatio-temporal cuboids
- learn spatio-temporal patterns related to visual saliency
- train end-to-end using the visual-only or the audio-visual human annotation

# Audio 2D Time-Frequency Representation



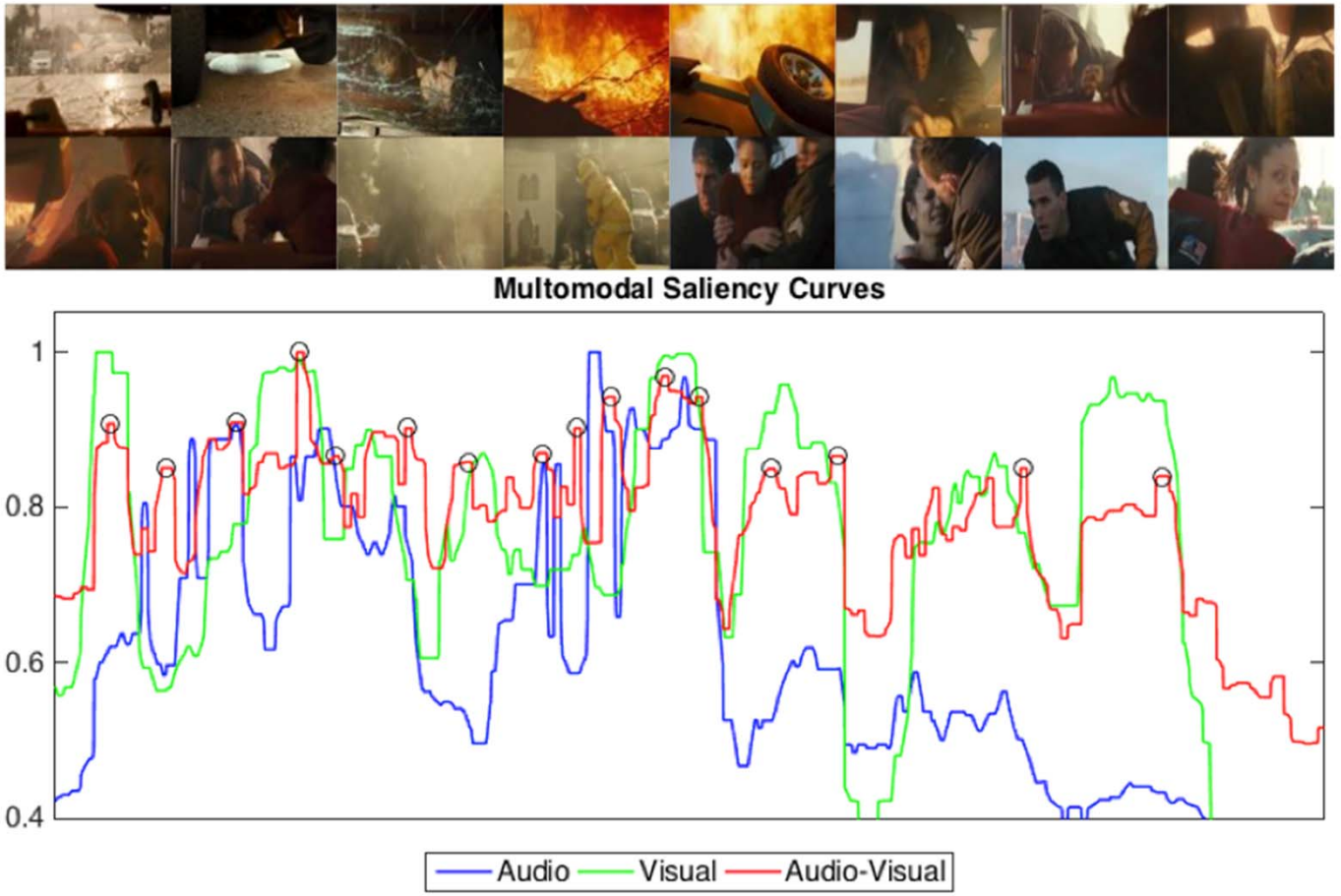

- represent the raw audio signal in the 2D time-frequency domain
  - preserve locality in both time and frequency axis
  - conventional MFCCs representation cannot maintain locality to frequency axis due to the DCT projection
- employ log-energies using 25 ms frames with 10 ms shift
  - compute first and second temporal derivatives
- temporal segments of 64 audio frames
  - synchronized with the 16-frames video clips

# CNN Architecture for Auditory Saliency




- deep 2D CNN architecture
  - input: 3 channel 2D input, similarly to the RGB image
    - synchronized with visual clips
  - output: softmax score as auditory saliency curve
- 2D convolutional and 2D max-pooling operations over time and frequency
  - based on the VGG idea of small kernels
- train end-to-end using the audio-only or the audio-visual human annotation

# CNN Estimated Audio-Visual Saliency Curves




- Audio-Visual Saliency Curves
  - two-stream CNNs trained with the audio-visual annotation labels
  - average the softmax scores
- Keyframes extracted as local extrema of the audio-visual curve

# COGNIMUSE Database

## Saliency, Semantic & Cross-Media Events Database

*http://cognimuse.cs.ntua.gr/database*

Including:

- Saliency annotation on multiple layers
- Audio & Visual events annotation
- COSMOROE cross-media relations annotation
- Emotion annotation

Database Content:

- **7** 30-min **movie clips** from: Beautiful Mind (BMI), Chicago (CHI), Crash (CRA), The Departed (DEP), Gladiator (GLA), Lord of the Rings III: The return of the king(LOR), Finding Nemo (FNE)
- **5** 20-min **travel documentaries**
- **1** 100-min **movie**: Gone with the Wind (GWTW)

# Database Annotation: Saliency & Structure

**Movie Structure:**

- Shots 370-699 (~540/movie)
- Scenes 7-23 (~14/movie)

**Generic Sensory Saliency:**

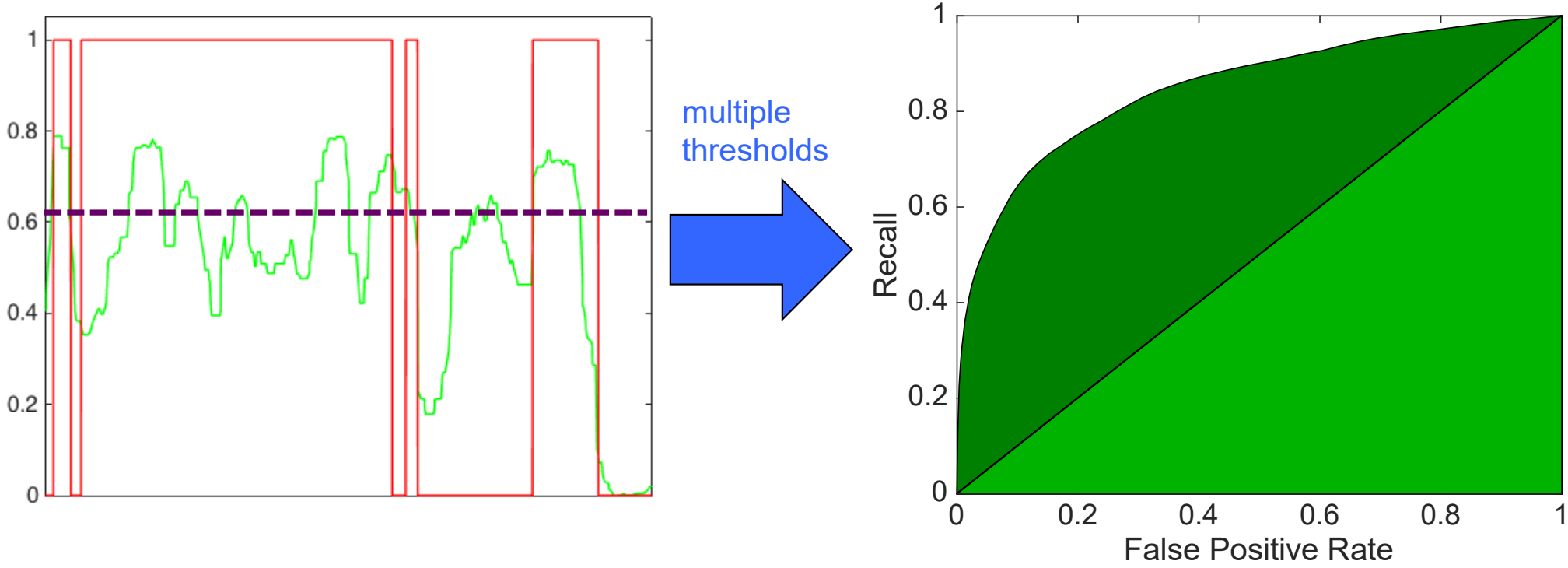
1) Audio-only
2) Visual-only
3) Audio-Visual (AV)

- Based on movie elements that capture the viewers' attention instantaneously or in segments
- Done quickly/effortlessly & without any focused attention or though
- Little or no searching required

**Attentive Saliency (Cognitive Attention):**

1) Semantics: segments that are conceptually important, e.g., phrases, actions, symbolic information, sounds….

# Salient Event Detection: Evaluation Metric




- Compare continuous saliency curves with binary annotations
- Area Under Curve (AUC)
  - apply threshold to saliency curves
  - area under the Receiver Operating Characteristic (ROC) curve (False Positive Rate – Recall)
  - binary classification problem: (salient / non salient segments)

# Evaluation Results – Hollywood Movies

A Beautiful Mind

Gladiator

Chicago

Lord of the Rings

Crash

The Departed

| AUC Results | V-V | | A-A | | AV-AV (mean) | |
|---|---|---|---|---|---|---|
| videos | Hndcr. | CNN | Hndcr. | CNN | Hndcr. | CNN |
| Six Hollywood Movies | | | | | | |
| BMI | 0.718 | 0.765 | 0.823 | 0.844 | 0.842 | 0.839 |
| GLA | 0.739 | 0.772 | 0.840 | 0.849 | 0.850 | 0.830 |
| CHI | 0.645 | 0.706 | 0.847 | 0.815 | 0.819 | 0.820 |
| LOR | 0.688 | 0.738 | 0.873 | 0.872 | 0.811 | 0.832 |
| CRA | 0.720 | 0.726 | 0.848 | 0.874 | 0.804 | 0.799 |
| DEP | 0.778 | 0.741 | 0.822 | 0.861 | 0.824 | 0.856 |
| Aver. | 0.715 | **0.742** | 0.842 | **0.853** | 0.825 | **0.830** |

- six fold cross-validation
  - five movies were used for training and tested on the sixth
- CNN-based architecture outperforms the hand-crafted frontend
  - for audio modality only in CHI we did not achieve improvement:
    - musical containing mostly music segments
    - CNN training on the other movies that do not contain this information

# Evaluation Results – Travel Documentaries

AR - London

AR - Rio

AR - Sydney

AR - Tokyo


GoT - London

| AUC Results | V-V | | A-A | | AV-AV (mean) | |
|---|---|---|---|---|---|---|
| videos | Hndcr. | CNN | Hndcr. | CNN | Hndcr. | CNN |
| Five Travel Documentaries | | | | | | |
| LON | 0.650 | 0.806 | 0.794 | 0.830 | 0.777 | 0.814 |
| RIO | 0.668 | 0.718 | 0.690 | 0.737 | 0.821 | 0.805 |
| SYD | 0.621 | 0.771 | 0.726 | 0.787 | 0.734 | 0.863 |
| TOK | 0.767 | 0.831 | 0.796 | 0.849 | 0.819 | 0.856 |
| GLN | 0.657 | 0.679 | 0.809 | 0.894 | 0.693 | 0.810 |
| Aver. | 0.673 | **0.761** | 0.763 | **0.819** | 0.769 | **0.830** |

- five fold cross-validation
  - four movies were used for training and the fifth for testing
- CNN-based architecture outperforms the hand-crafted frontend
  - Greater improvements for visual modality

# Evaluation Results – Full Length Movie

Gone with the Wind – Part 1

| AUC Results | V-V | | A-A | | AV-AV (mean) | |
|---|---|---|---|---|---|---|
| videos | Hndcr. | CNN | Hndcr. | CNN | Hndcr. | CNN |
| Full Length Movie | | | | | | |
| GWW* | 0.589 | 0.644 | 0.714 | 0.706 | 0.664 | **0.735** |
| GWW** | 0.626 | **0.660** | 0.706 | **0.740** | 0.648 | 0.710 |

- For the "Gone with the Wind" movie two different setups were adopted:
  i. only the six Hollywood movies were used for training (GWW*)
  ii. all data was used for training (GWW**), thus six movies and five travel documentaries
- CNN-based architecture outperforms the hand-crafted frontend for all modalities
  - Better improvements when CNN models are trained in all data

[ P. Koutras, A. Zlatintsi and P. Maragos, "Exploring CNN-based architectures for Multimodal Salient Event Detection in Videos", IEEE Workshop on IVMSP, 2018. ]

# Video summaries: travel doc & gwtw (system 2)

**AR London ca 16% ca 3'40"**

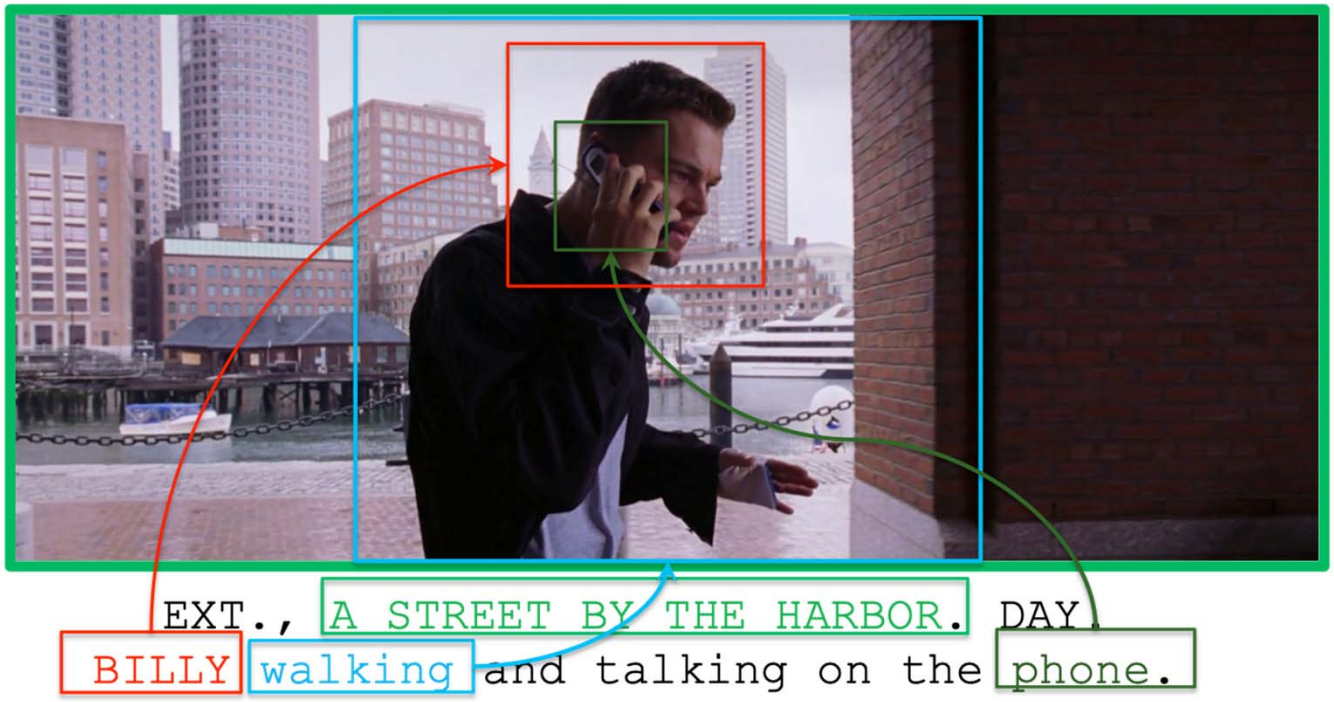

**GWTW ca 3% ca 3'**
**(**3min from full duration movie**)**

# 3.
# Multimodal (Visual + Textual) Concept Learning in Videos with Weakly Supervised Techniques

# Visual Concepts

- Detect and recognize visual concepts in videos in a weakly supervised manner, mining their labels from an accompanying descriptive text.
- **Visual Concepts:** *Spatio-temporally localized video segments that carry a specific structure in the visual domain.*

1. Faces
2. Actions
3. Scenes
4. Objects




G. Bouritsas, P. Koutras, A. Zlatintsi and P. Maragos, Multimodal Visual Concept Learning with Weakly Supervised Techniques, CVPR 2018

# Weak Supervision with Natural Language

- **Motivation:**
  - **Why Natural Language?**
    - Rich semantics – interpretable – easy to extract.
  - **Why Weak Supervision?**
    - Reduce the time-consuming and costly procedure of manual annotation.

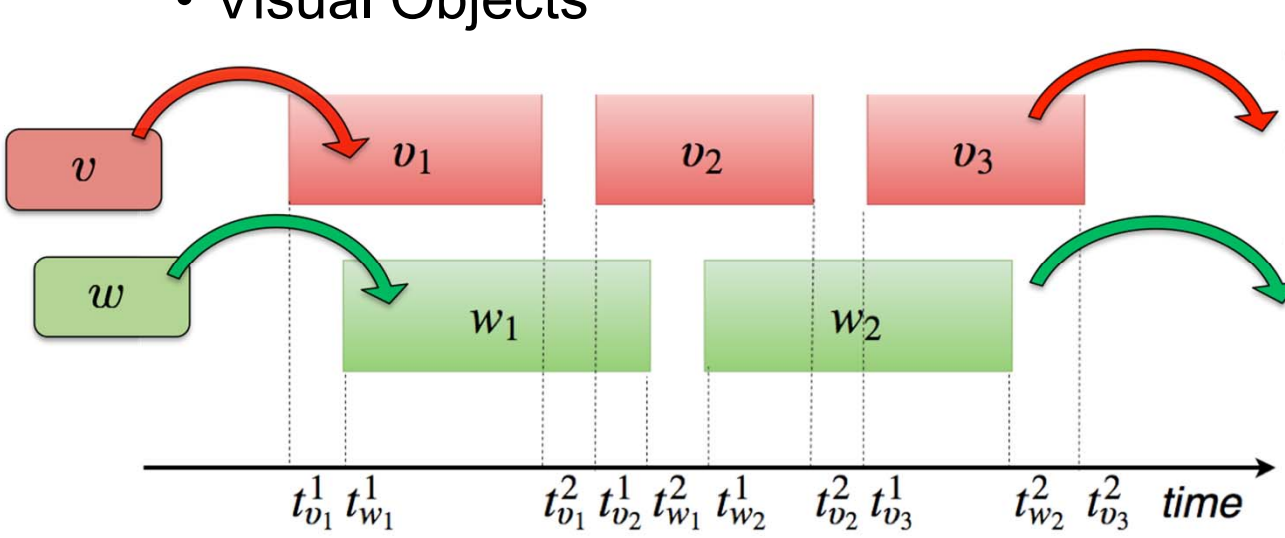
a) Achieve recognition in data annotated sparsely/imprecisely.

b) Collect new data to train fully supervised models.

- **Challenges:**
  - Spatio-Temporal ambiguity: absence of specific spatio-temporal correspondence between visual and textual objects.
  - Semantic ambiguity: Words/sentences may have several different meanings.

# Multimodal Visual Concept Learning

➢ **Dual Modality scheme:** Two data streams flowing in parallel.




• Representation: features $\boldsymbol{x_v}$

• Temporal interval: $T_v = [t_v^1, t_v^2]$

• Class: latent variable $z_v = ??$

• Implied label: $y_w = ??$

• Temporal interval: $T_w = [t_w^1, t_w^2]$

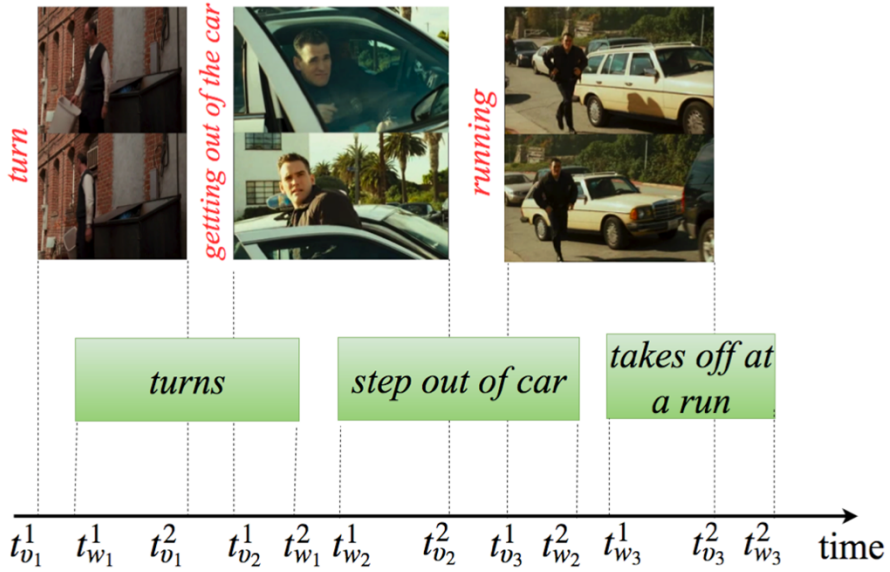
# Weakly Supervised frameworks

- Fuzzy Sets MIL (**FSMIL**): Fuzzy bags of Multiple Instances.

$$\mathcal{V}_w = \{(v, \mu_w(v)) \mid v \in \mathcal{V}, \mu_w(v) = g(\frac{|T_w \cap T_v|}{|T_v|})\}$$

*Visual objects overlapping with the textual one*

*Membership grade*



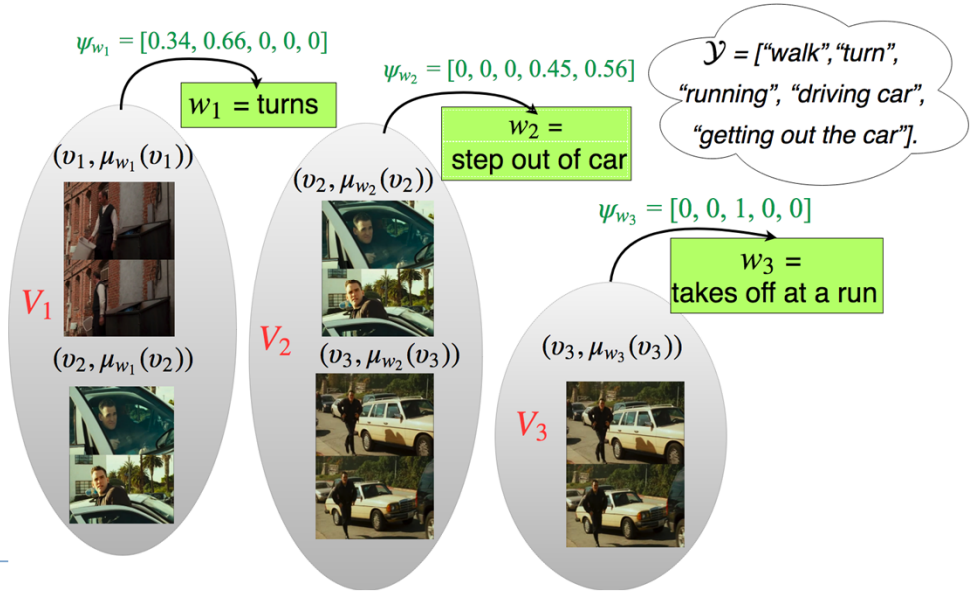

- Probabilistic Label MIL (**PLMIL**):

*Each bag is assigned a probabilistic label.*

*Unsupervised estimation via semantic similarity.*

$$\psi_w(y) = \mathbb{P}[y_w = y|w]$$

$$\psi_w(y) = s_{wy} / \sum_{\ell \in \mathcal{Y}} s_{w\ell}$$



# Multimodal Visual Concept Learning with Weakly Supervised Techniques

- **Discriminative clustering model (DIFFRAC):**
  - Ridge regression with linear classifier $f(x) = x^T \omega + b$

$$\min_{Z,\omega,b} \frac{1}{2V} \|Z - X^T\omega - 1_V b\|_F^2 + \frac{\lambda}{2} Tr(\omega^T \omega)$$

*Closed form w.r.t classifier*

Loss + Regularizer

$\mathcal{Y}_w = \{y \mid \psi_w(y) \neq 0\}$

Weighted bag members – FSMIL

$$\min_{Z,\xi} Tr(ZZ^T A(X,\lambda)) + \kappa \sum_{w \in \mathcal{W}} \sum_{y \in \mathcal{Y}_w} f(\psi_w(y)) \xi_{wy}^2,$$

**s.t**

$$\forall w \in \mathcal{W}, \forall y \in \mathcal{Y}_w, \sum_{v \in \mathcal{V}_w} z_{vy} \mu_w(v) \geq 1 - \xi_{wy}$$

Relaxing the constraints

Relaxation factor – PLMIL

# Results: Face Recognition

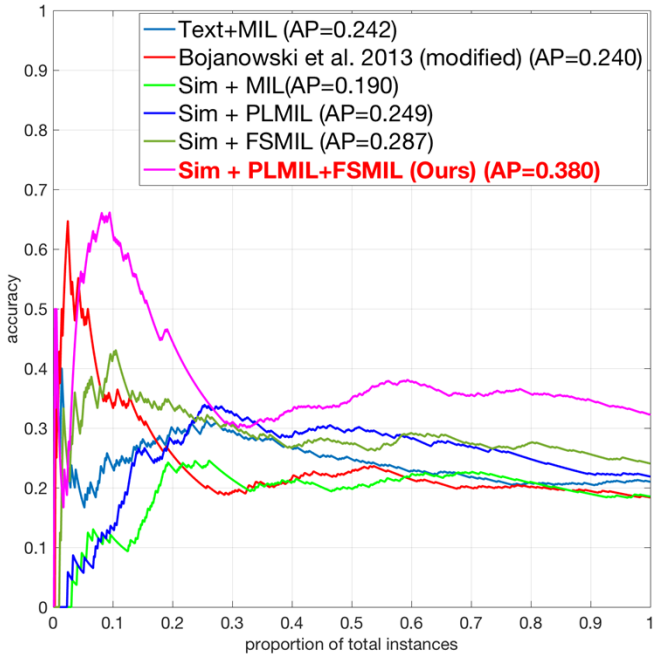
- COGNIMUSE Dataset: 5 movies + scripts

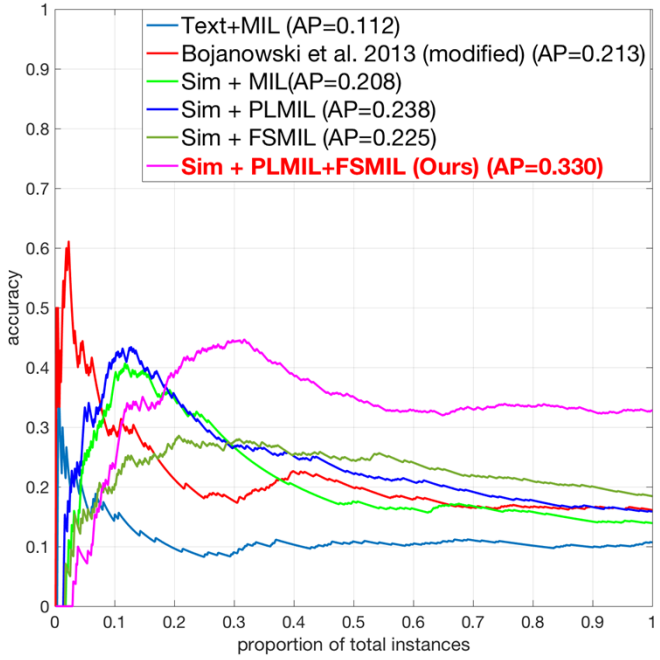
| Set | Development | | | Test | | | | All |
|---|---|---|---|---|---|---|---|---|
| | DEP | LOR | MAP | BMI | CRA | GLA | MAP | MAP |
| Text+MIL | 0.433 | 0.656 | 0.544 | 0.551 | 0.434 | 0.437 | 0.474 | 0.502 |
| SIFT+MIL [*Bojanowski et al. 2013*] | 0.630 | 0.879 | 0.755 | 0.724 | 0.644 | 0.681 | 0.683 | 0.711 |
| SIFT+FSMIL | 0.693 | 0.881 | 0.787 | 0.770 | 0.691 | 0.746 | 0.736 | 0.756 |
| VGG+MIL | 0.834 | **0.954** | 0.894 | 0.825 | 0.696 | 0.830 | 0.784 | 0.828 |
| VGG+FSMIL (Ours) | **0.864** | 0.952 | **0.908** | **0.857** | **0.731** | **0.901** | **0.830** | **0.861** |
| [*Miech et al. 2017*]+VGG: fg | 0.788 | 0.898 | 0.843 | 0.666 | 0.479 | 0.577 | 0.574 | 0.682 |
| [*Miech et al. 2017*]+VGG+FSMIL: fg | **0.810** | **0.913** | **0.862** | **0.696** | **0.505** | **0.651** | **0.617** | **0.715** |
| [*Miech et al. 2017*]+VGG: bg | **0.185** | **0.189** | **0.187** | **0.304** | 0.047 | **0.052** | 0.134 | 0.155 |
| [*Miech et al. 2017*]+VGG+FSMIL: bg | 0.184 | **0.189** | **0.187** | 0.269 | **0.278** | 0.038 | **0.195** | **0.192** |

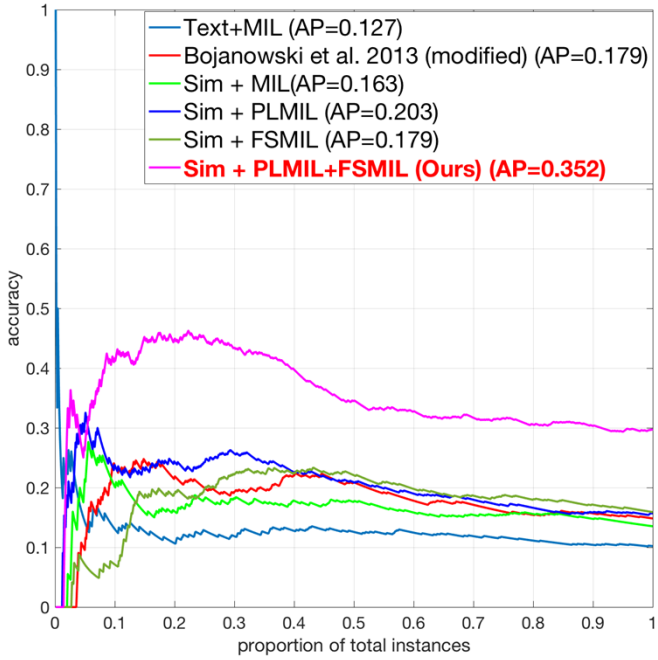
- Bojanowski et al. 2013: treats both ambiguities with hard constraints (MIL).
- Miech et al. 2017: extra constraint for background concepts.
- Bouritsas et al. 2018: FSMIL extension

# Results: Action Recognition

- COGNIMUSE Dataset: 5 movies + scripts
  - mean per sample accuracy curves for 6, 8 & 10 action classes.



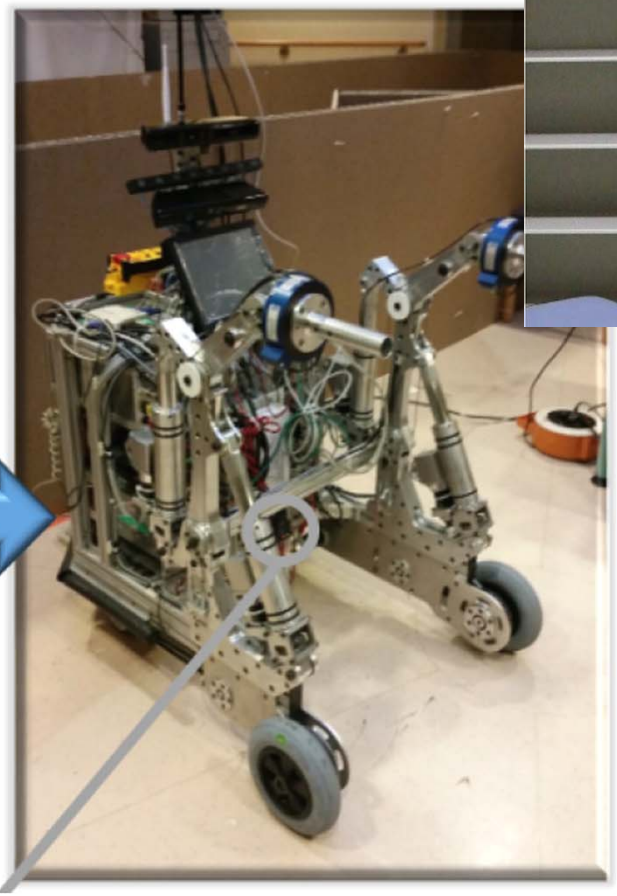



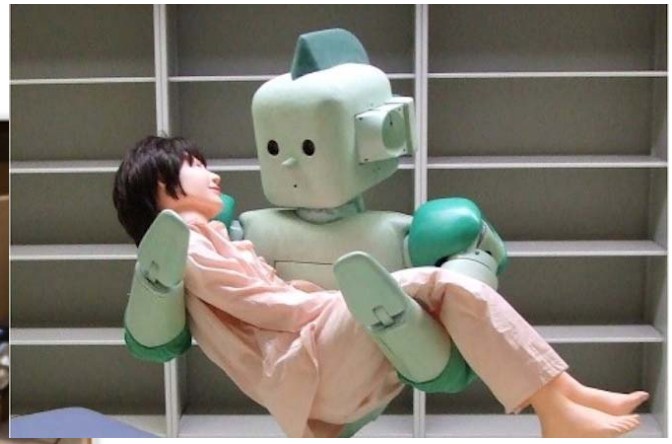



# 3.
# Audio-Visual Gesture Recognition and Human-Robot Interaction

# Multimodal HRI: Applications and Challenges

**assistive robotics**





**cation,**
**rtainment**

- 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

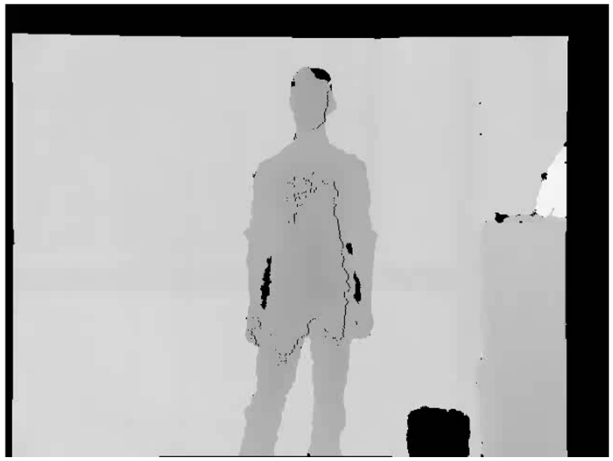
# Multimodal Gesture Signals from Kinect-0 Sensor

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

## RGB Video & Audio

## Depth

(vieniqui - *come here*)



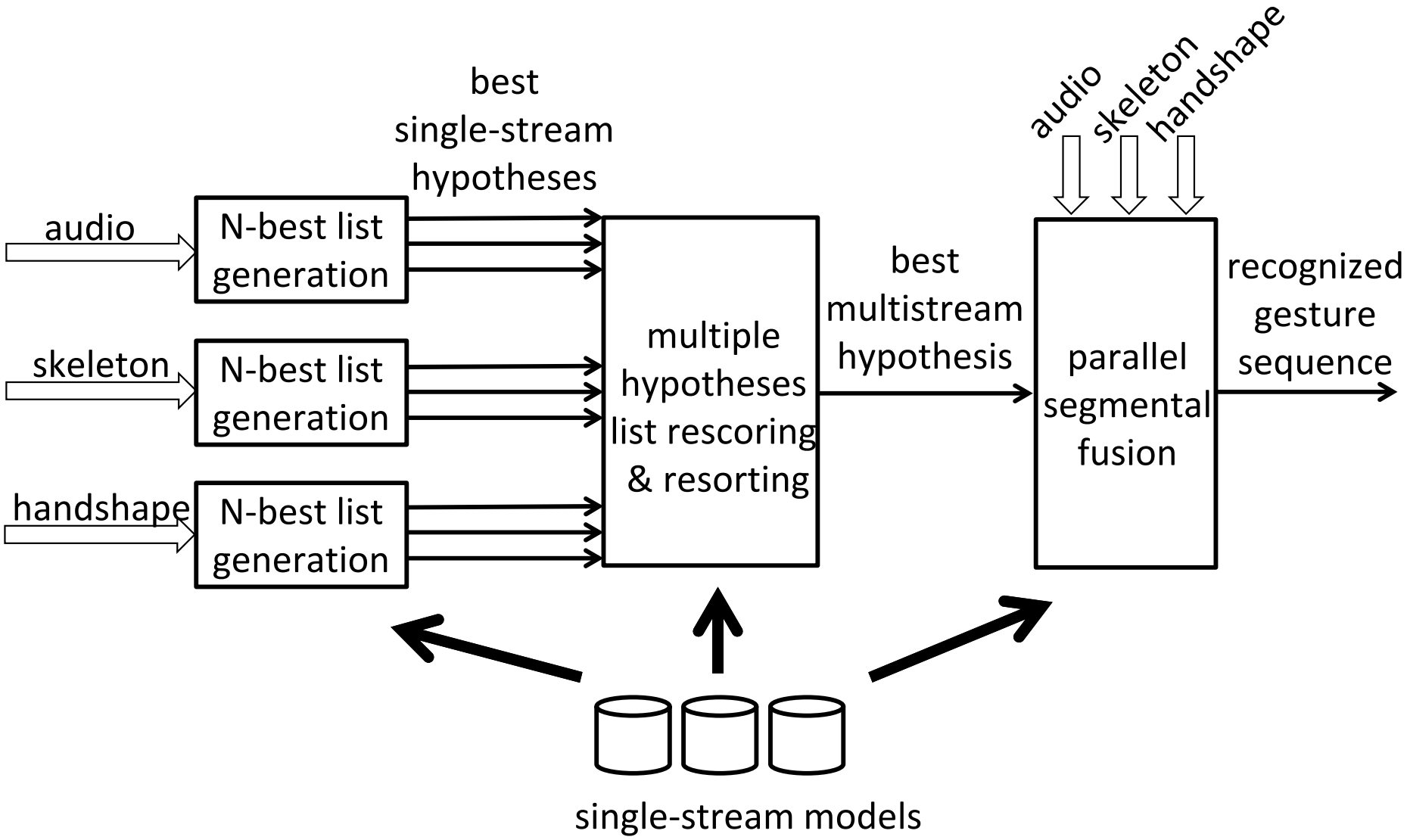
## Skeleton

(vieniqui - *come here*)

## User Mask

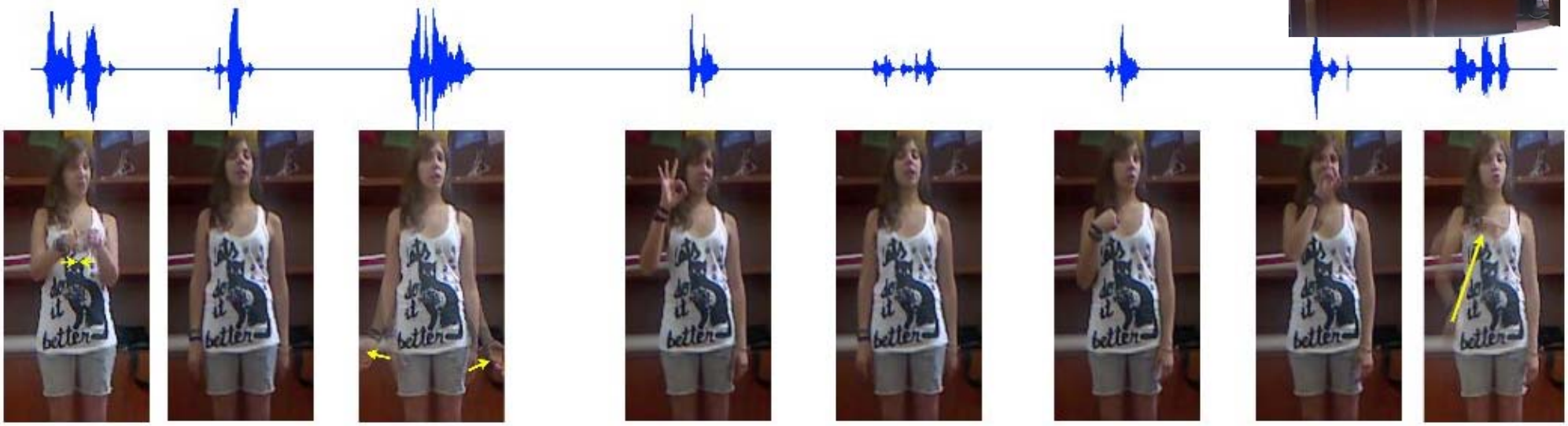
(vieniqui - *come here*)

# Overview: Multimodal Hypothesis Rescoring + Segmental Parallel Fusion




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

# Audio-Visual Fusion & Recognition



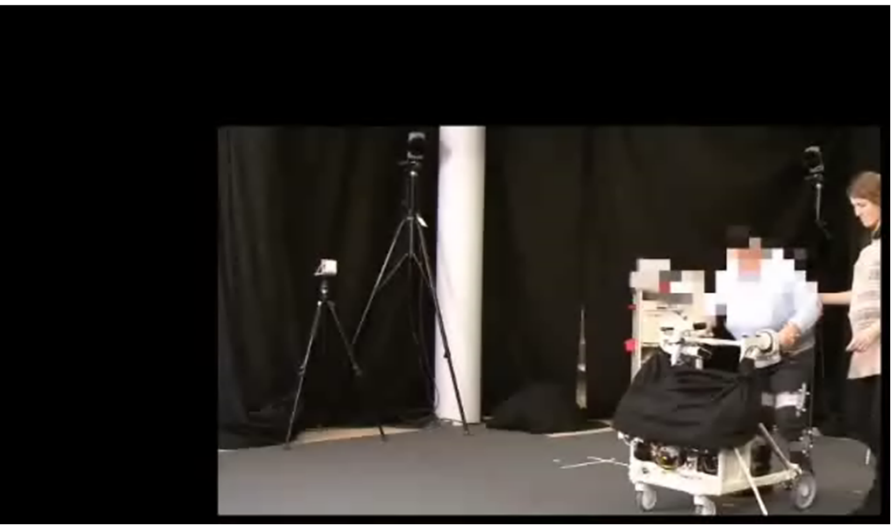
| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| REF | DACCORDO | OOV | OOV | OK | OOV | OOV | OOV | SONOSTUFO |
| AUDIO | DACCORDO | BM | PREDERE | OK | BM | FAME | BM | SONOSTUFO |
| P1 | DACCORDO | BM | BM | OK | BM | BM | OK | SONOSTUFO |
| P2 | DACCORDO | BM | BM | BM | BM | BM | BM | 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 ]

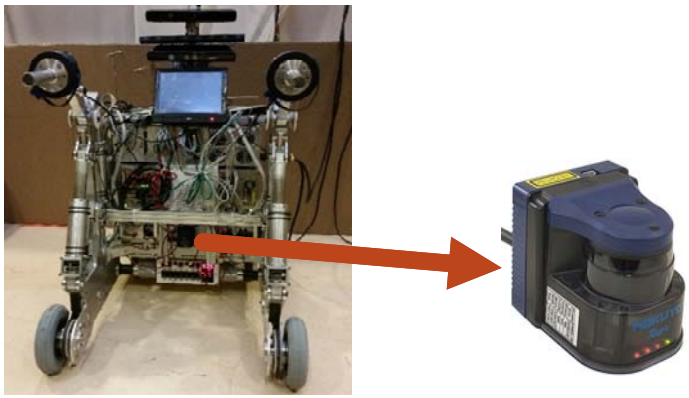
# EU Project MOBOT: Motivation

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 ]

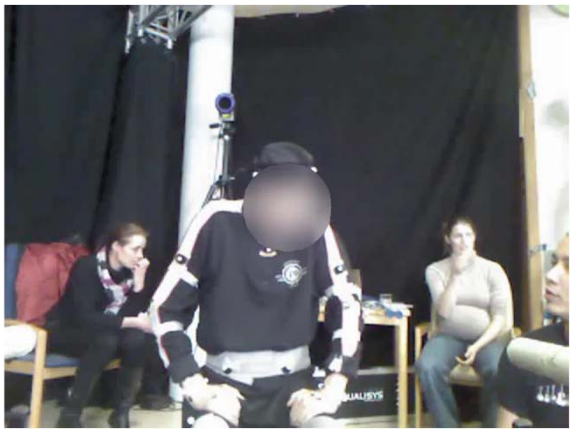
# Multi-Sensor Data for HRI

Kinect1 RGB

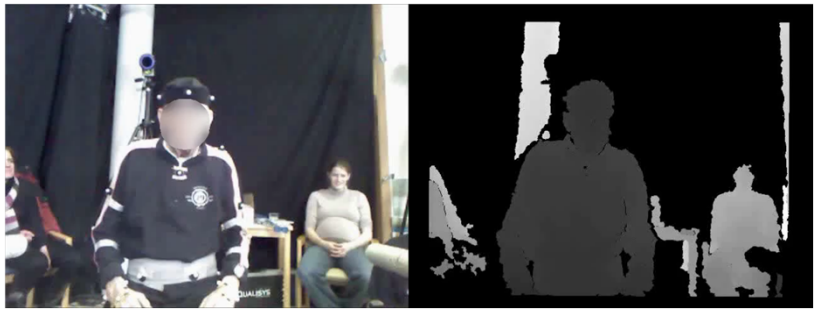
Kinect1 Depth

MEMS Audio Data

Kinect1 RGB Data



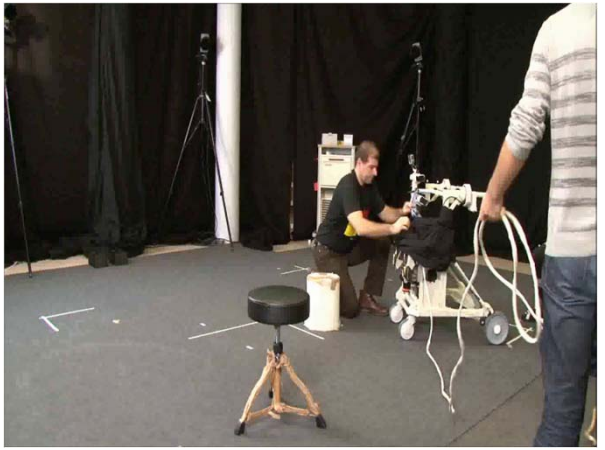
Kinect Depth Data



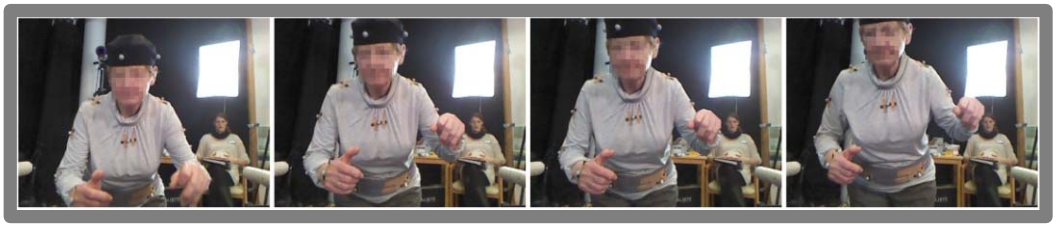
Go Pro RGB Data

HD1 Camera Data

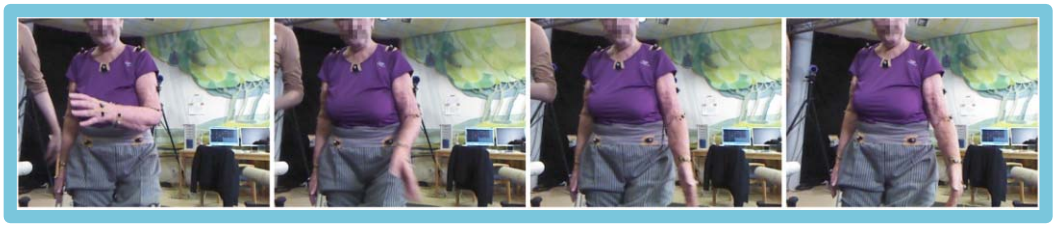
HD2 Camera Data



# Visual action recognition pipeline



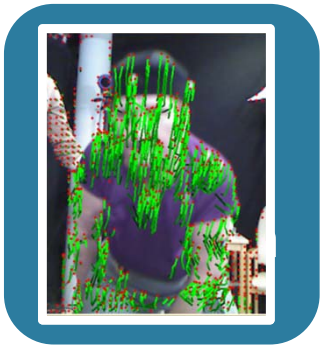
Sit to Stand



Walk

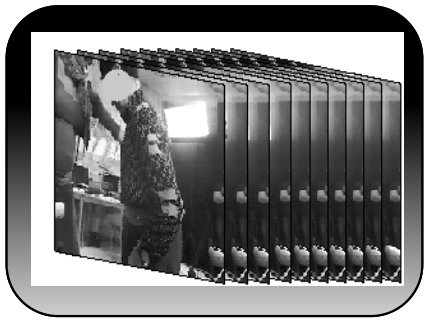
Video

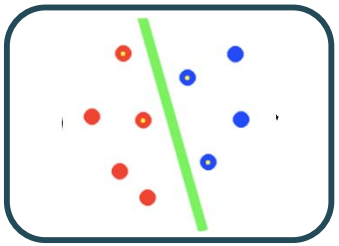
Sliding Window



Visual Feature Extraction

Feature Encoding





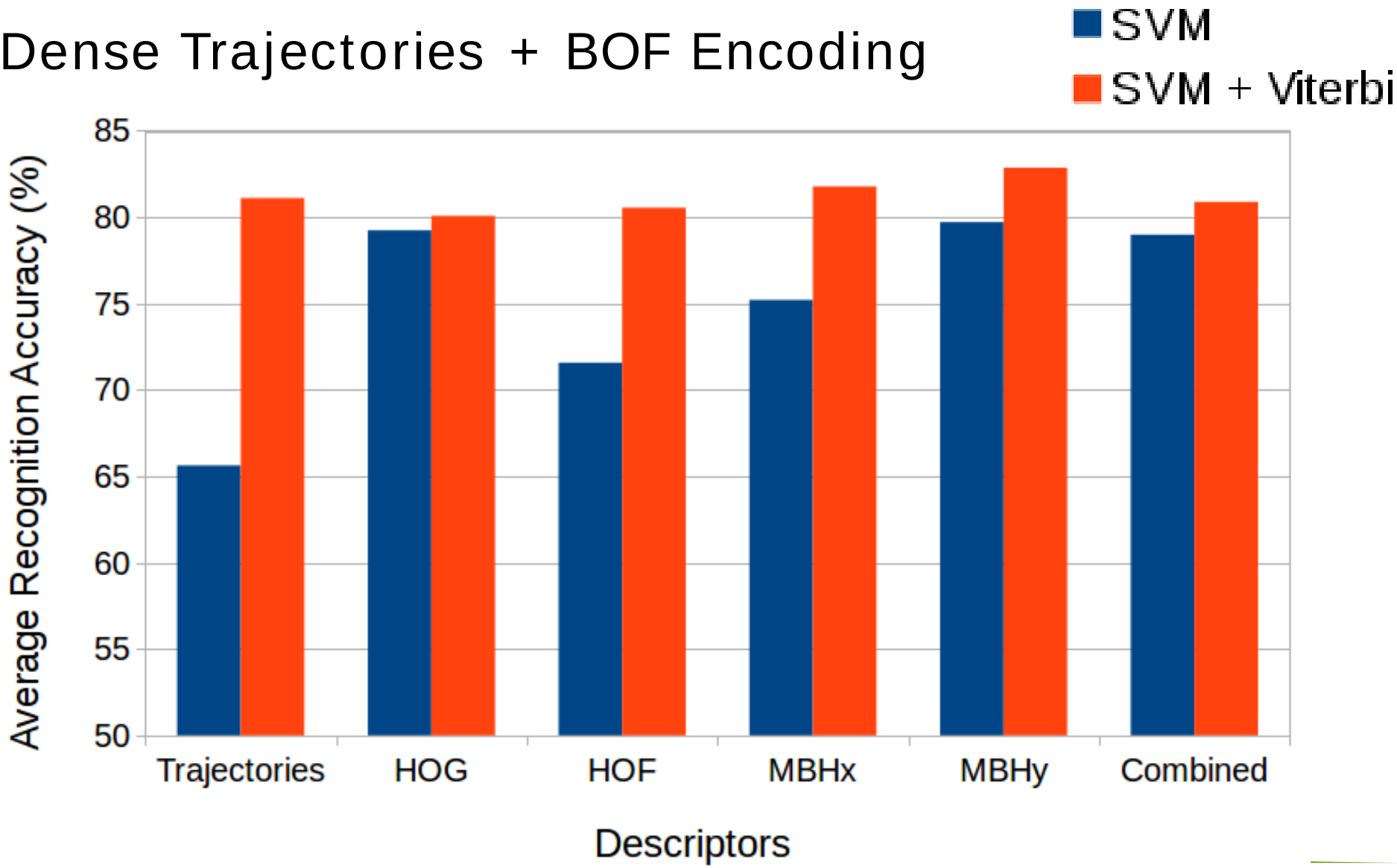
Classifier

Post-processing

Recognized Sequence

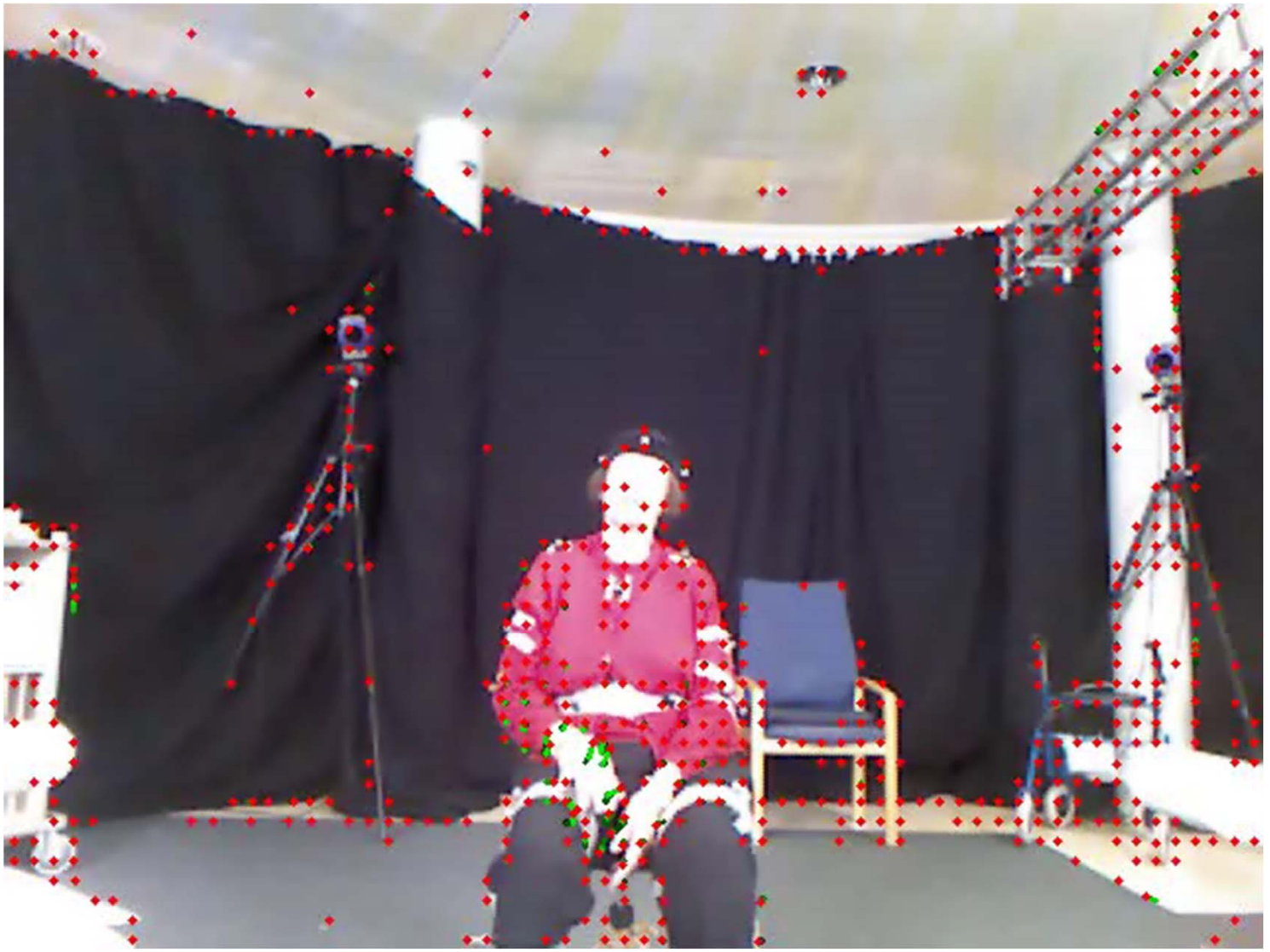
# Action Recognition Results (4a, 6p): Descriptors + Post-processing Smoothing

- Dense Trajectories + BOF Encoding




Results improve by adding Depth and/or advanced Encoding

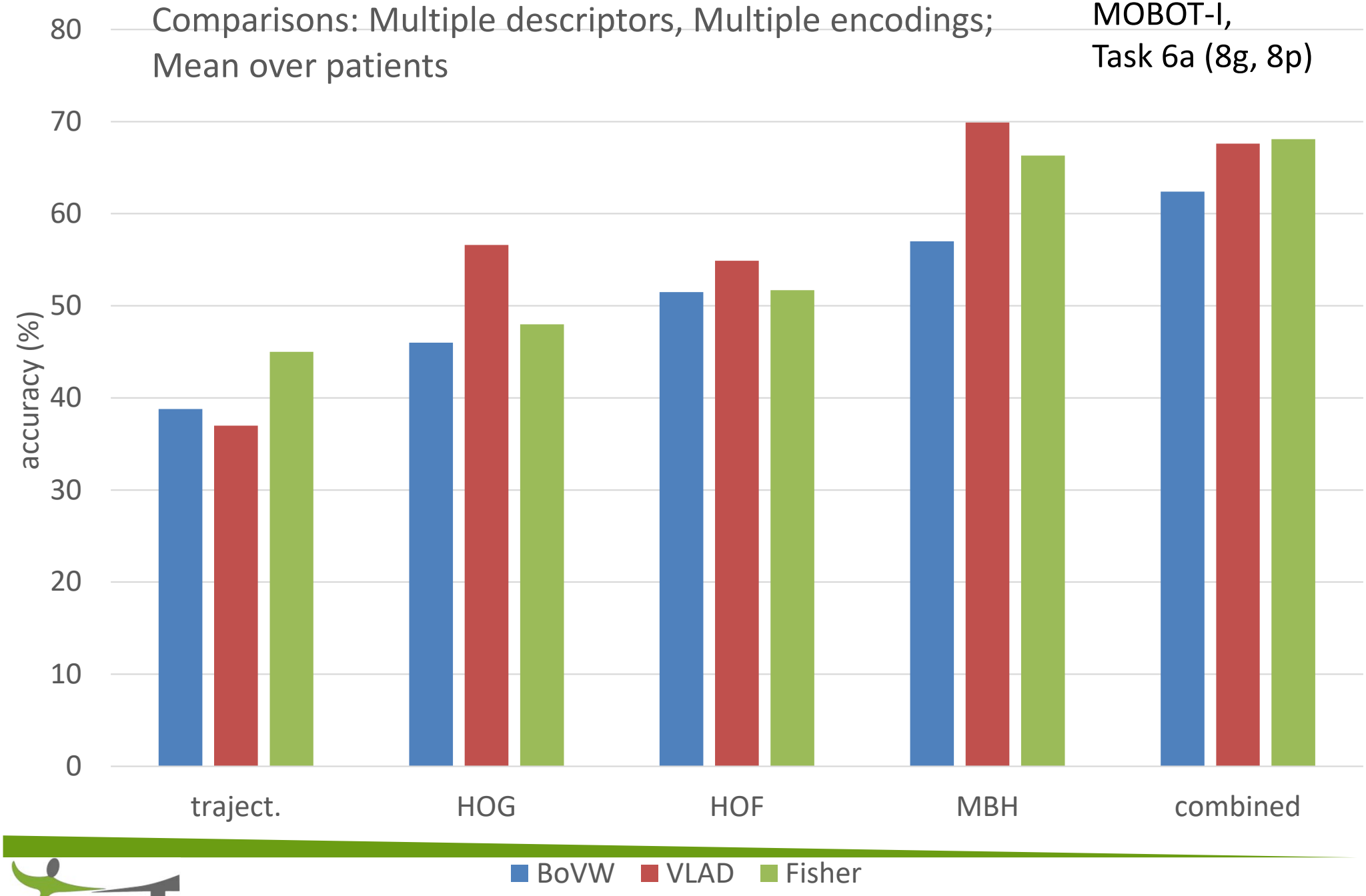
# Applying Dense Trajectories on Gesture data



# Extended results on Gesture Recognition

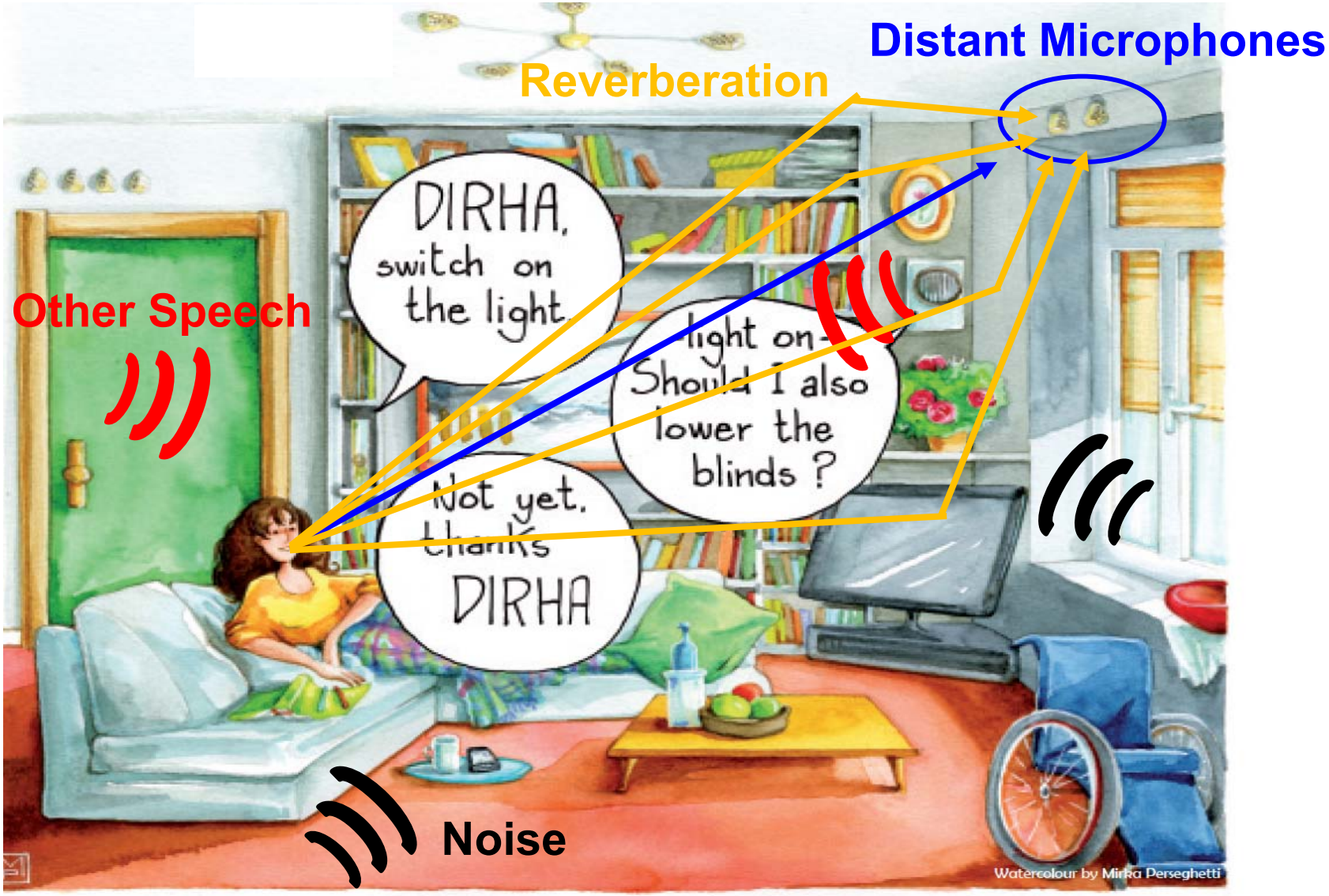
Comparisons: Multiple descriptors, Multiple encodings;
Mean over patients

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



# Spoken Command Recognition

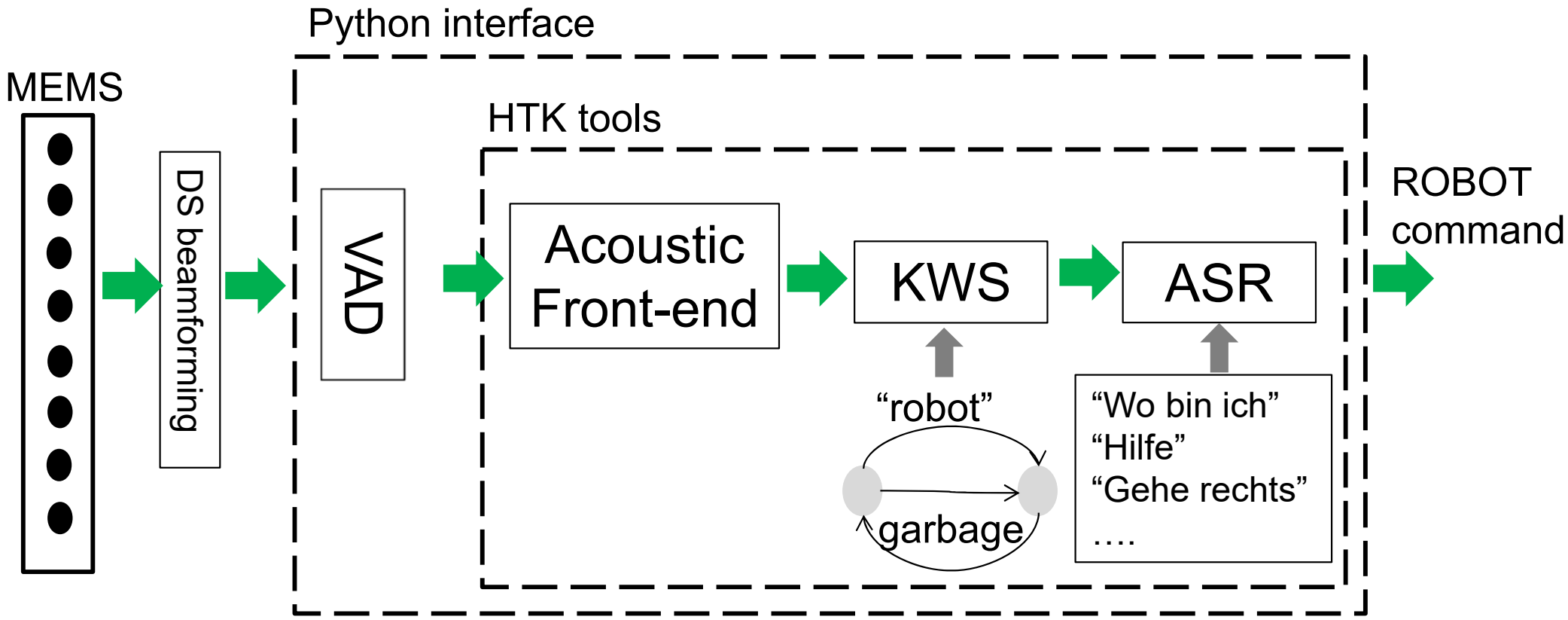
# Distant Speech Recognition in Voice-enabled Interfaces




https://dirha.fbk.eu/

# Spoken-Command Recognition Module for HRI

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



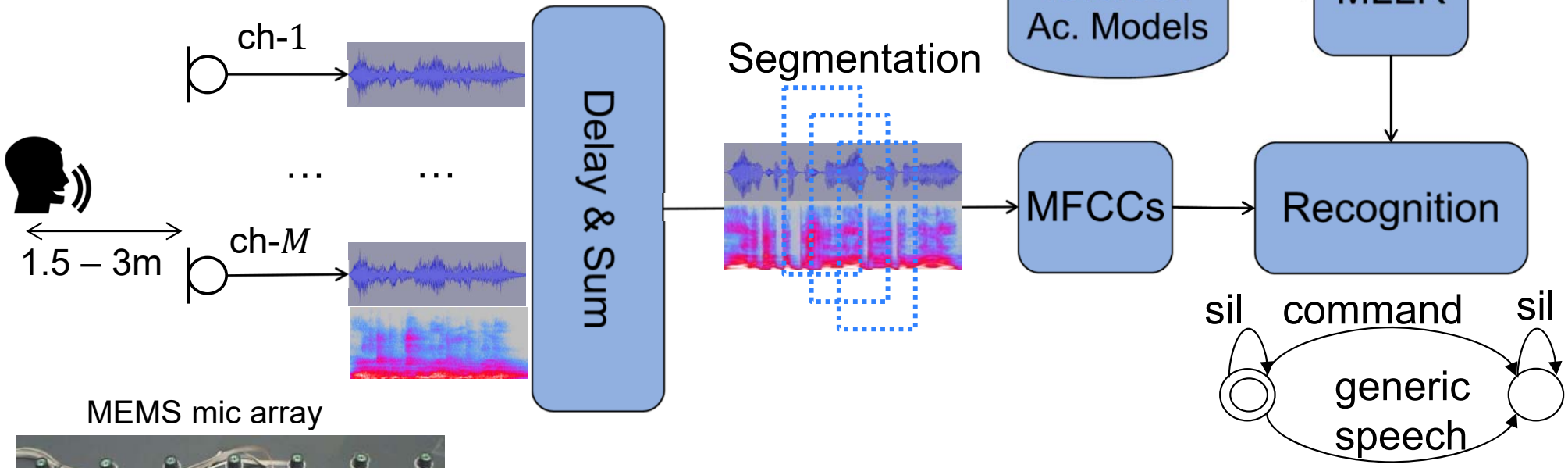

[ I. Rodomagoulakis, A. Katsamanis, G. Potamianos, P. Giannoulis, A. Tsiami, P. Maragos, "Room-localized spoken command recognition in multi-room, multi-microphone environments", *Computer Speech & Language*, 2017. ]

# Online Spoken Command Recognition

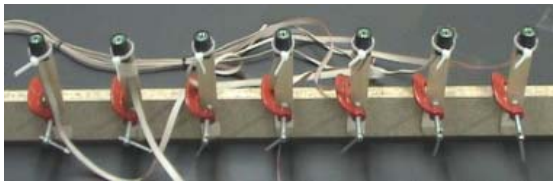
- **Greek, German, Italian, English**

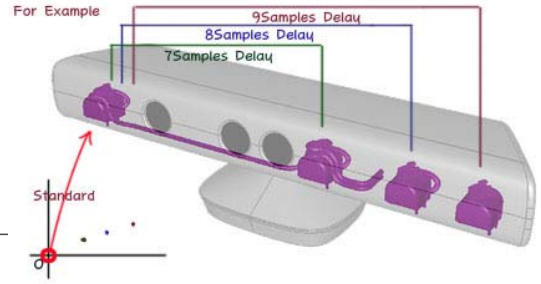
Pentagon ceiling array (Shure)

Targeted Acoustic Scenes

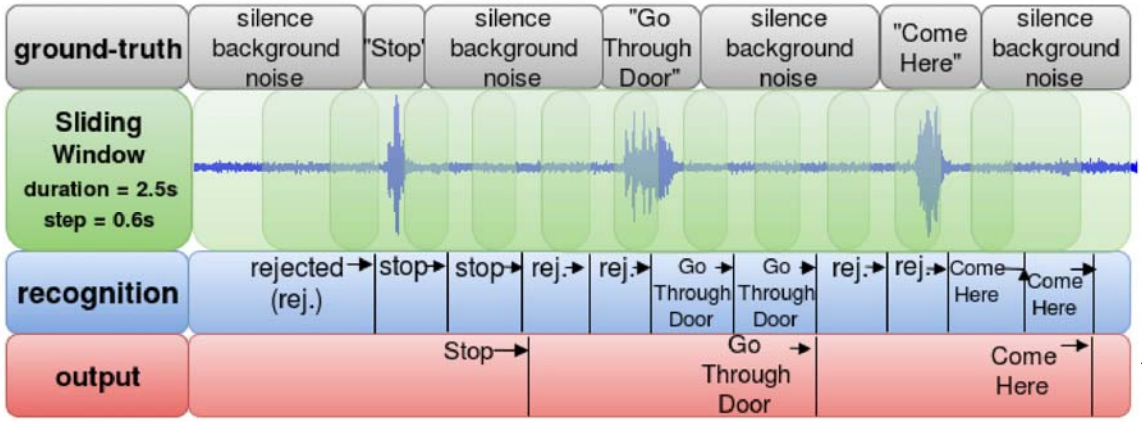



MEMS mic array






Kinect mic array

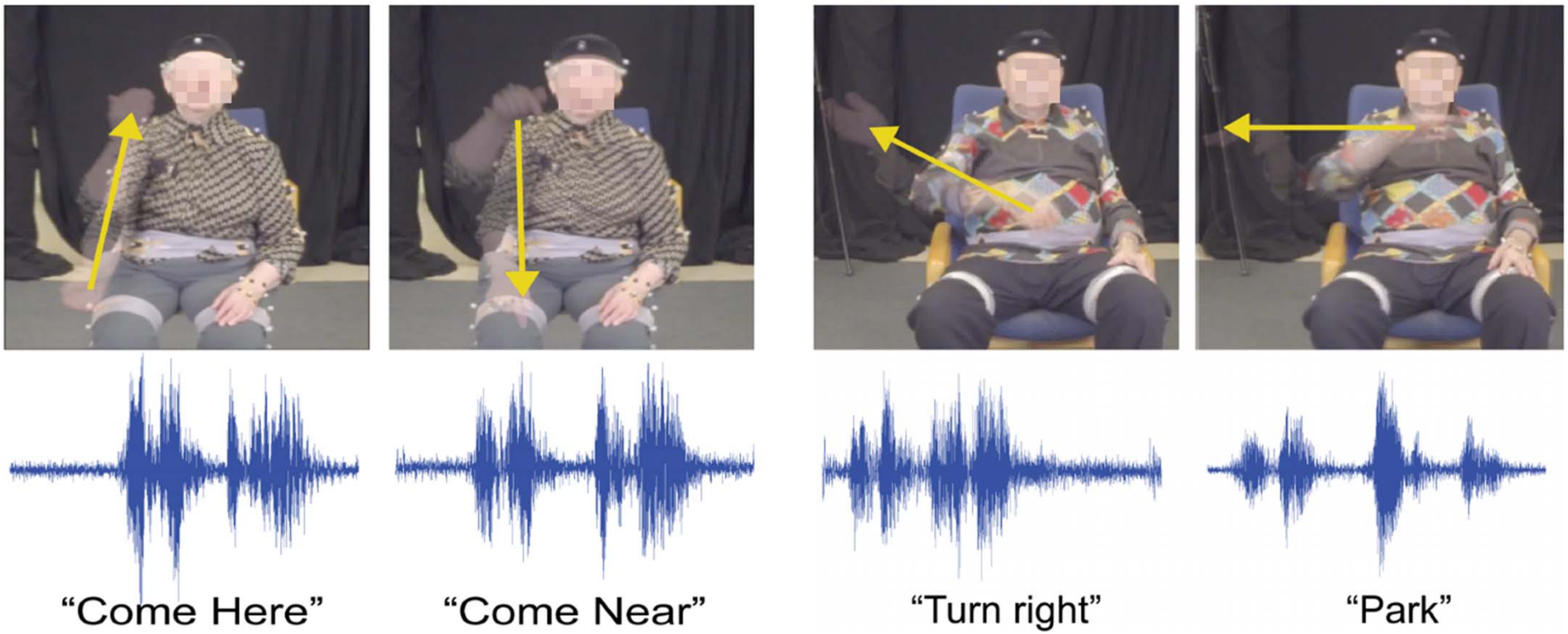


# Audio-visual Fusion for Multimodal Gesture Recognition

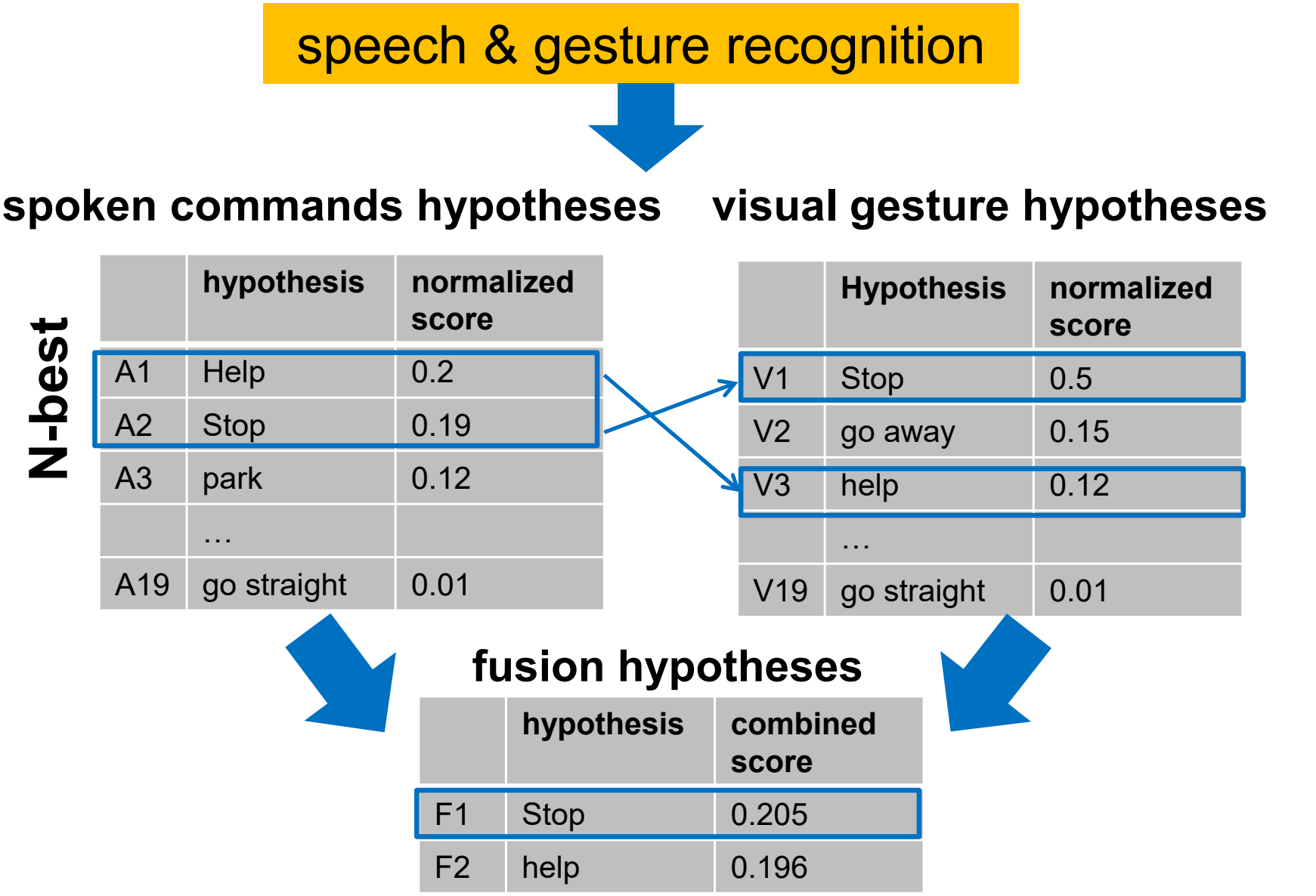
# Multimodal fusion: Complementarity of visual and audio modalities

Similar audio, distinguishable gesture

Distinguishable audio, similar gesture



# Audio-Visual Fusion: Hypotheses Rescoring

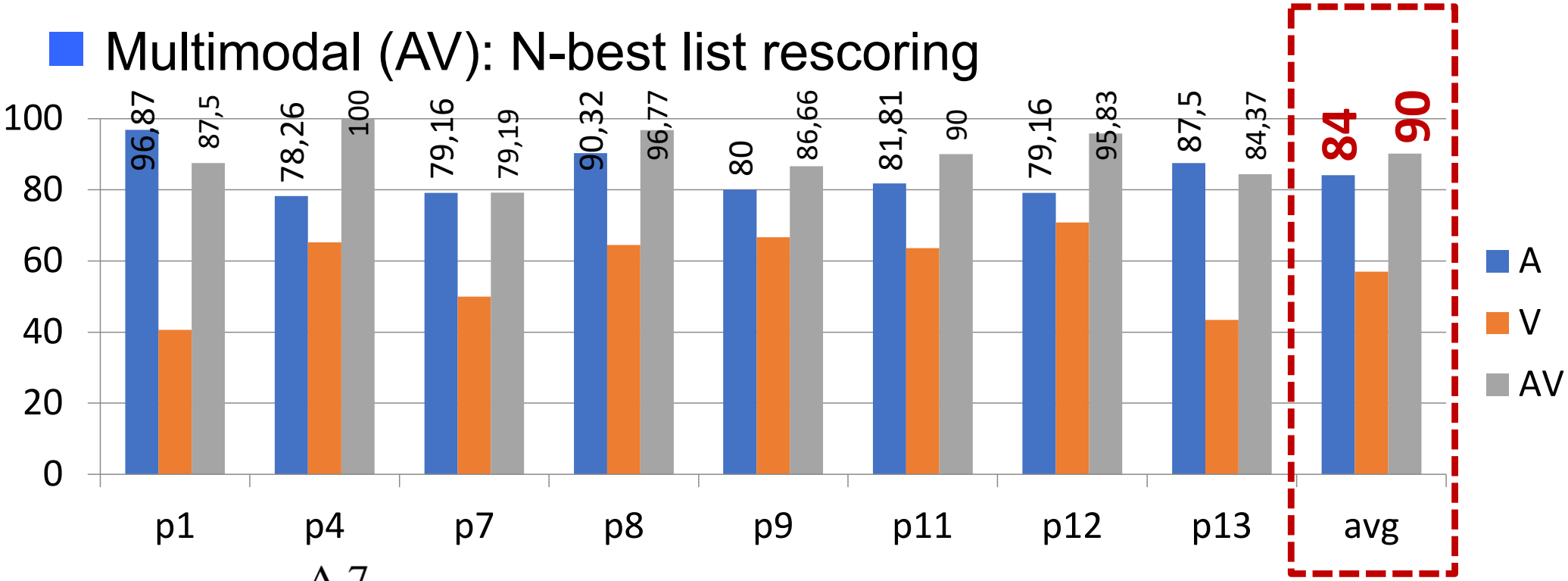


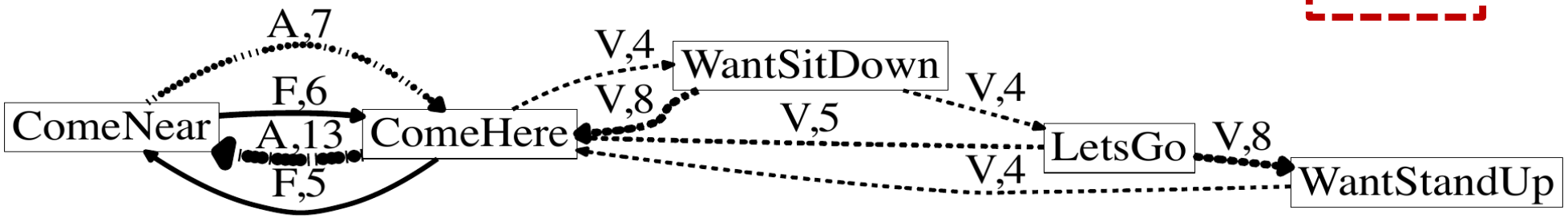

$$\mathrm{MAX}(w_a \times score(A_1) + w_v \times score(V_3), w_a \times score(A_2) + w_v \times score(V_1))$$

$w_a, w_v$ : 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



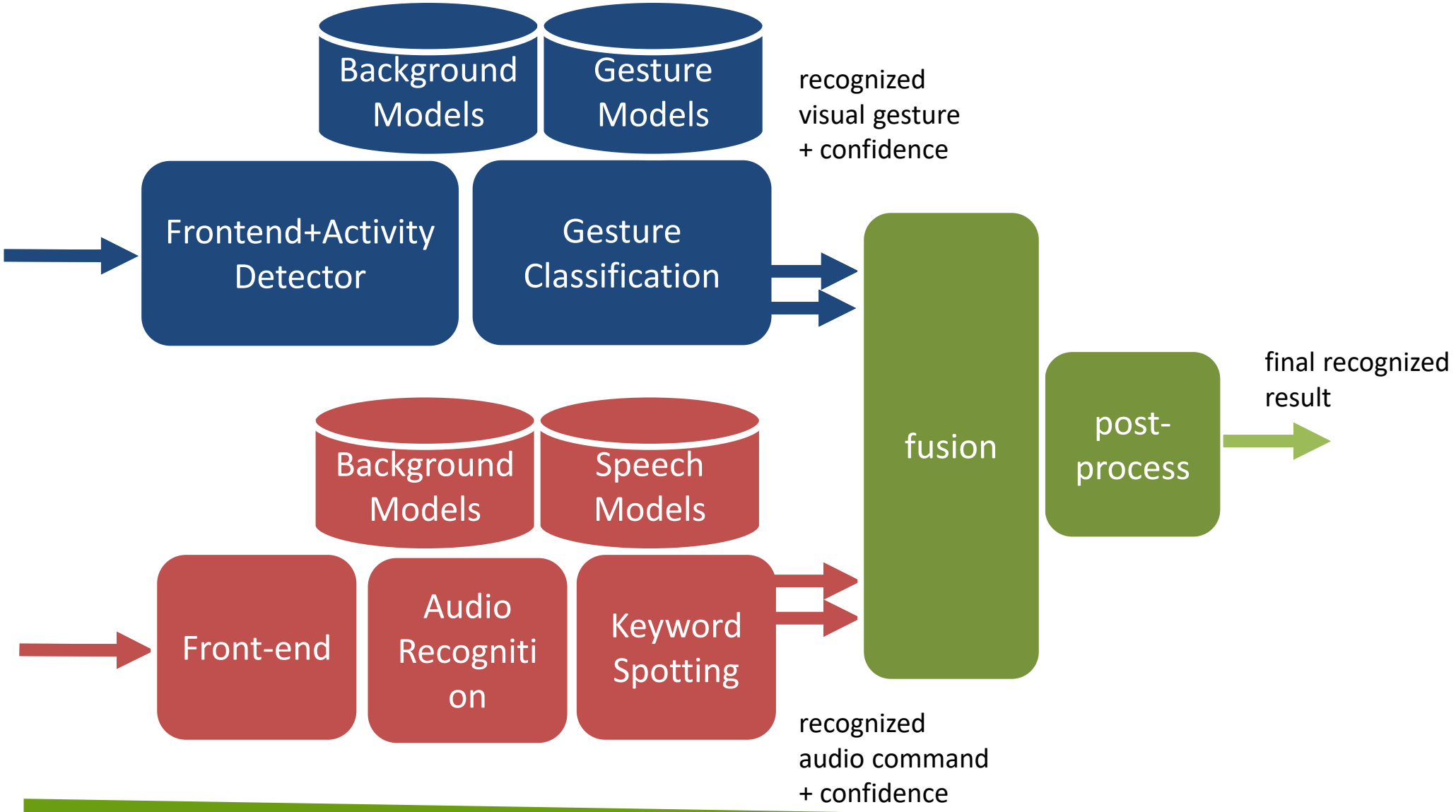




**Multimodal confusability graph**

# Audio-Visual gesture recognition

## Online processing system – Open Source Software

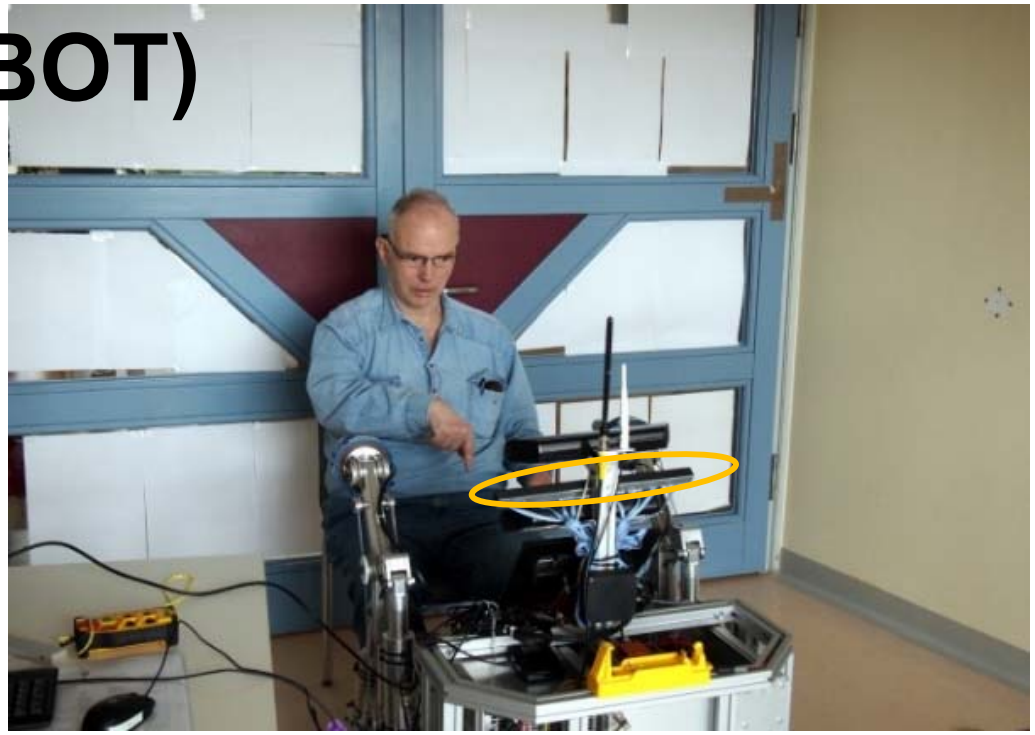
http://robotics.ntua.gr/projects/building-multimodal-interfaces



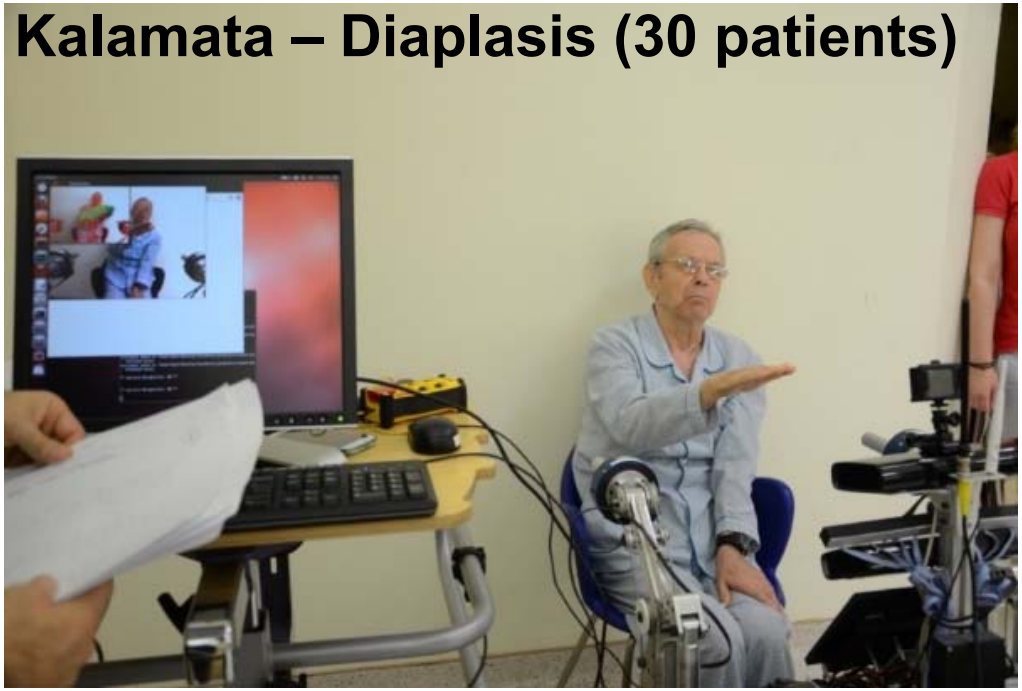

[ N. Kardaris, I. Rodomagoulakis, V. Pitsikalis, A. Arvanitakis and P. Maragos, *A platform for building new human-computer interface systems that support online automatic recognition of audio-gestural commands*, Proc. ACM Multimedia 2016. ]

# Clinical Studies (MOBOT)

Heidelberg – Bethanien (19 patients)



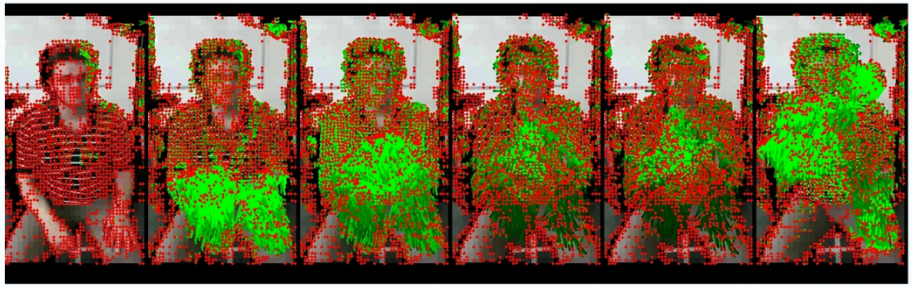
Kalamata – Diaplasis (30 patients)

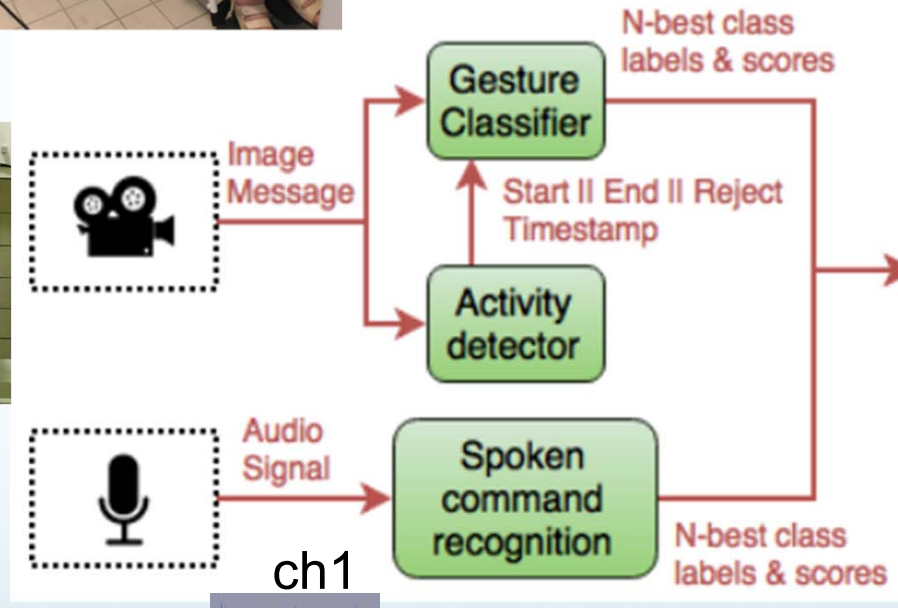


Speech, Gestures, Combination: 3 repetitions of 5 commands

# Gesture & Spoken Command Recognition

**dense trajectories of visual motion**



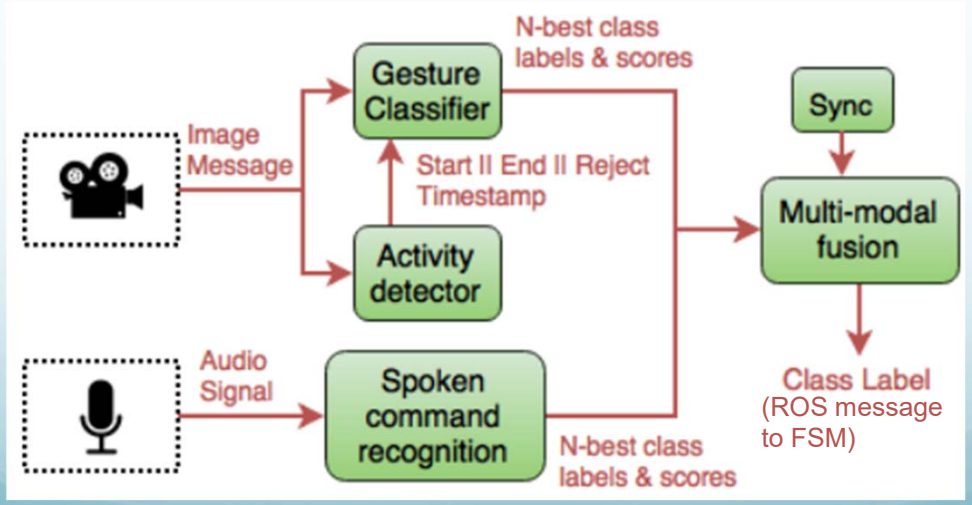





# Multimodal Fusion and On-line Integration

- Multimodal “late” fusion


- ROS (Robot Operating System) based integration



# Validation results

## 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% |

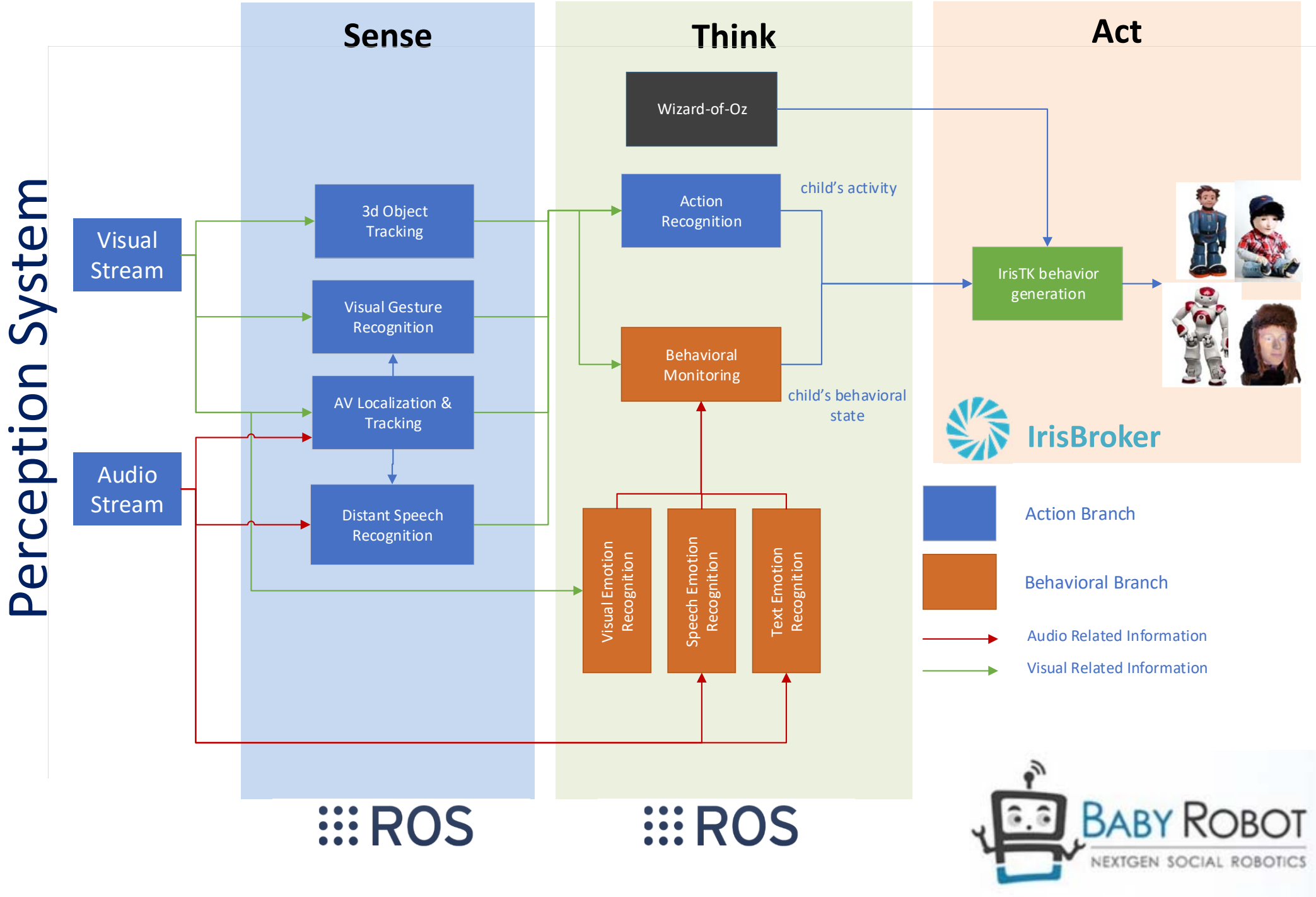
### Fondazione Santa Lucia @ Rome

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

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

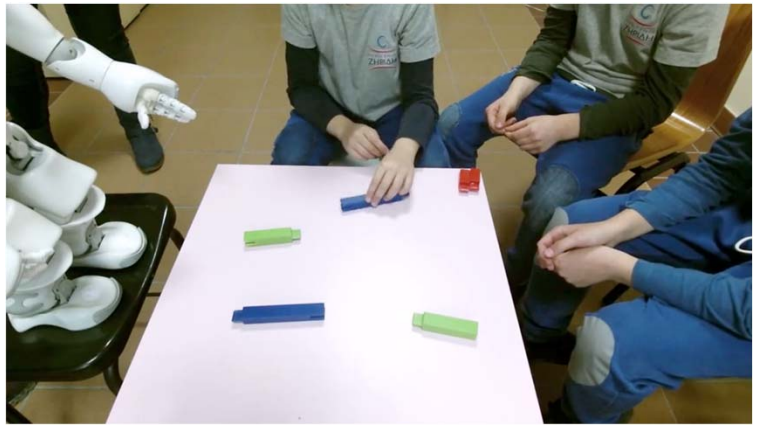
# I-SUPPORT system video

# EU project BabyRobot: Experimental Setup Room

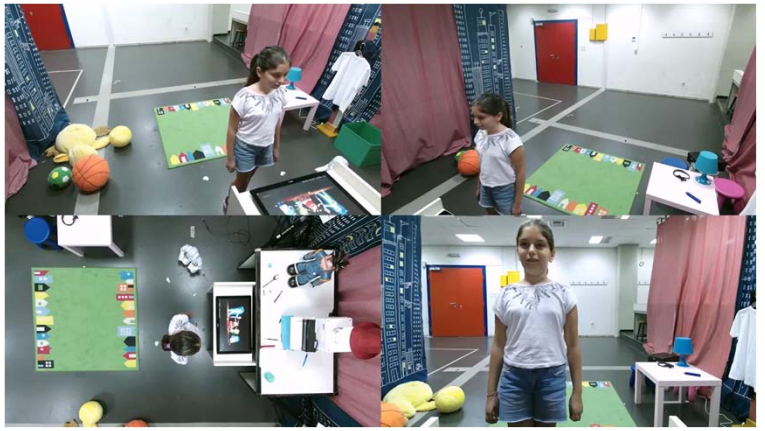

Perception System
Sense
Think
Act
Visual Stream
Audio Stream
3d Object Tracking
Visual Gesture Recognition
AV Localization & Tracking
Distant Speech Recognition
Wizard-of-Oz
Action Recognition
child's activity
Behavioral Monitoring
child's behavioral state
Visual Emotion Recognition
Speech Emotion Recognition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

# Action Branch: Developed Technologies

## 3D Object Tracking



## Multiview Gesture Recognition

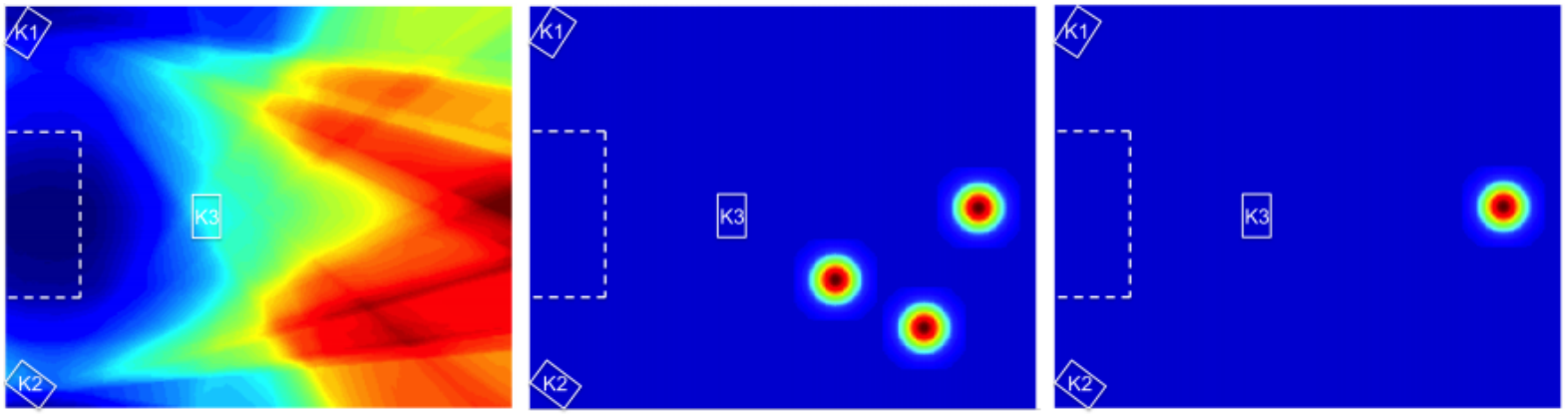


## Speaker Localization and Distant Speech Recognition

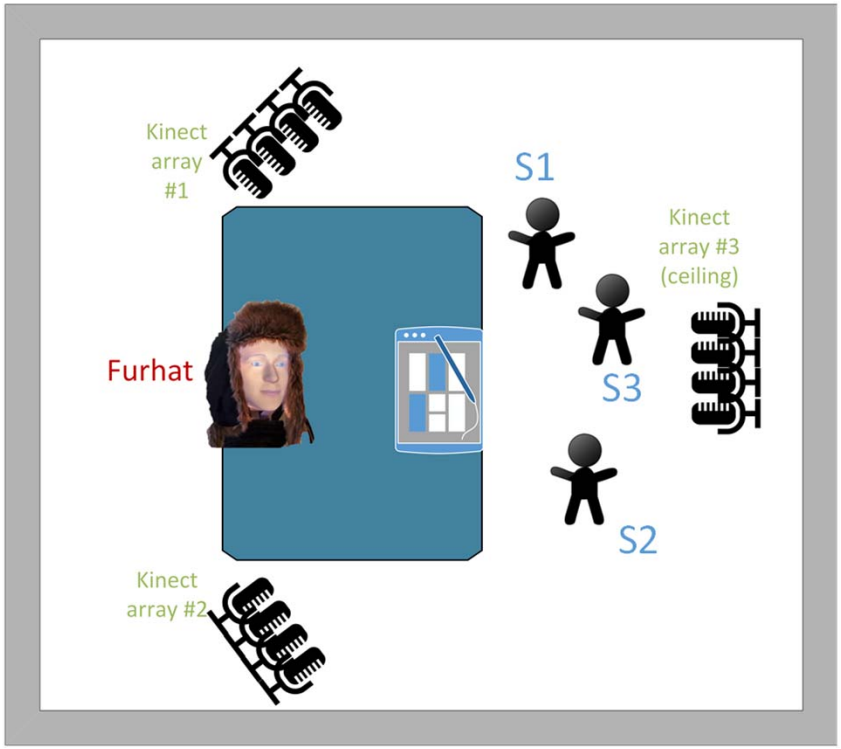
## Multiview Action Recognition

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%



# Multi-view Gesture Recognition

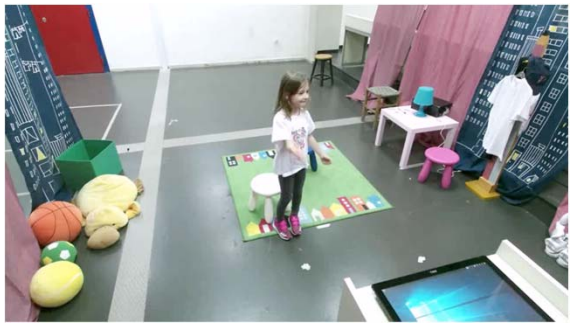
- Multiple views of the child's gesture from different sensors
- Extract Dense Trajectory features from each view
- Encoding Frameworks:
  - Bag of Visual Words (BoW)
  - Vector of Locally Aggregated Descriptors (VLAD)
- Employ different FusionSchemes

BABY ROBOT
NEXTGEN SOCIAL ROBOTICS

# Gesture Recognition – Vocabulary

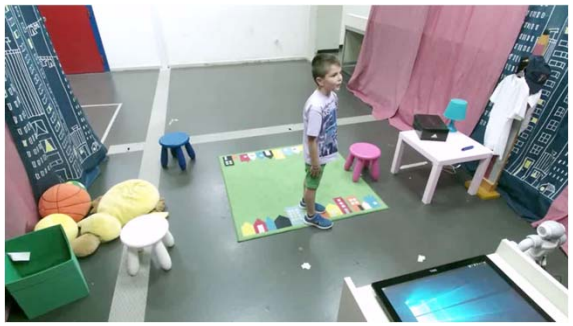
Nod

Greet



Come Closer

Sit



Stop

Point

Circle

BABY ROBOT
NEXTGEN SOCIAL ROBOTICS

# Multi-view Gesture Recognition - Evaluation

| | Single Camera | | | Fusion | | |
|---|---|---|---|---|---|---|
| Feat. | Kinect #1 | Kinect #2 | Kinect #3 | MEAN | MIN | MAX |
| Traj. | 68.75 | 66.90 | 65.74 | 76.62 | 75.00 | 71.53 |
| HOG | 40.74 | 33.33 | 29.40 | 39.58 | 36.57 | 39.58 |
| HOF | 70.83 | 70.37 | 69.21 | 78.01 | 77.55 | 76.39 |
| MBH | 76.85 | 67.82 | 68.29 | 83.80 | 80.09 | 78.24 |
| Comb. | 77.78 | 73.84 | 73.61 | 81.94 | **83.56** | 77.55 |

- 7 classes (+1 Bg): nod, greet, come closer, sit, stop, point, circle
- Average classification accuracy (%) for the employed gestures performed by 28 children (development corpus).
- Results for the five different features for both single and multi-stream cases.
- Results on spontaneous corpus (31 children): ~74%

BABY ROBOT
NEXTGEN SOCIAL ROBOTICS

# Multi-view Gesture Recognition - Children vs. Adults

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

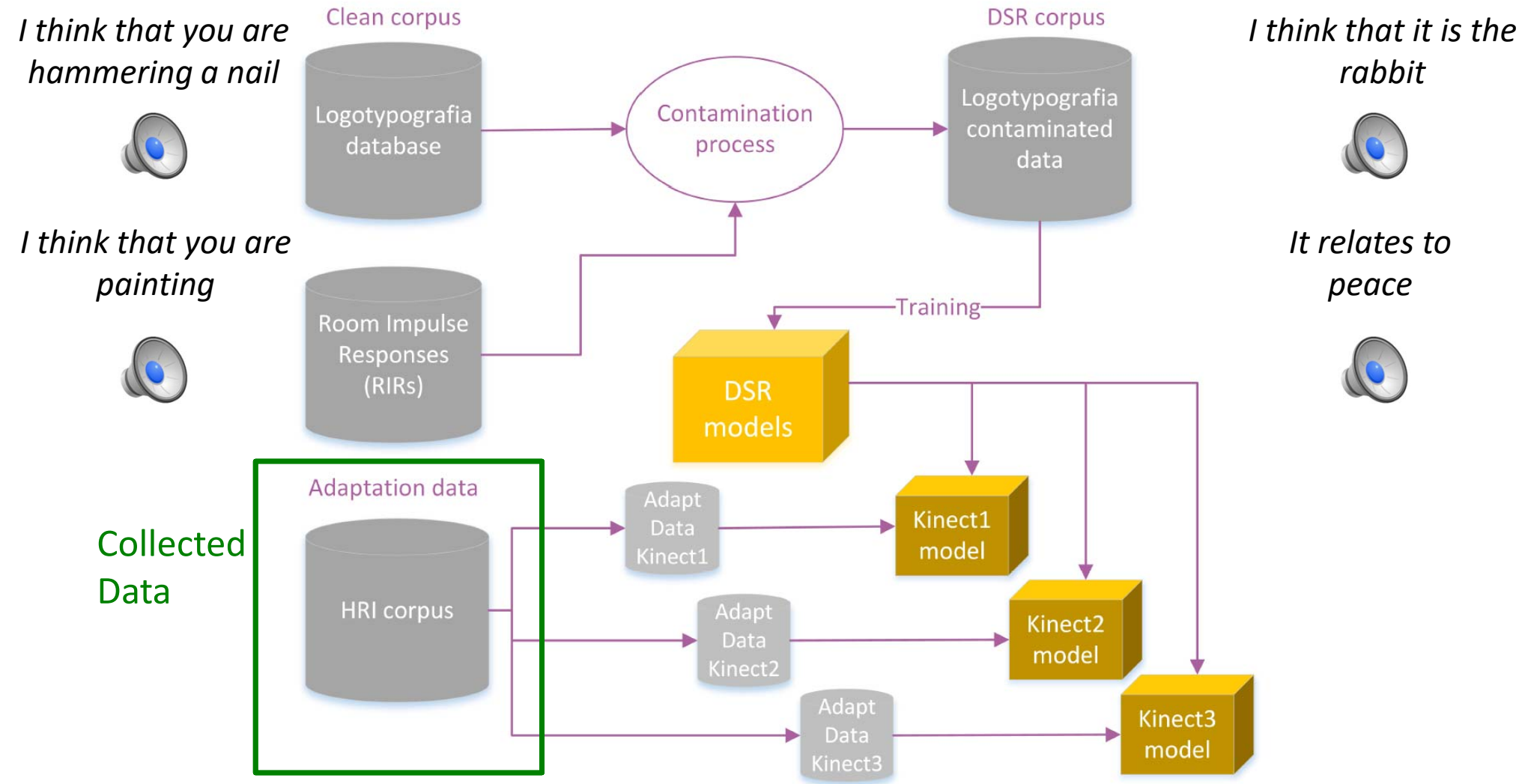
Employed Features: **MBH**

| | | 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

# Distant Speech Recognition System




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



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

# Action Recognition – Vocabulary

Cleaning a window

Ironing a shirt

Digging a hole

Driving a bus

Painting a wall

Hammering a nail

Wiping the floor

Reading

Swimming

Working Out

Playing the guitar

Dancing

BABY ROBOT
NEXTGEN SOCIAL ROBOTICS

# From Single-view to Multi-view Action recognition

Kinect #1
Side view 45º left

Kinect #2
Side view 45º right

Kinect #3
Top view

Example of the extracted Dense Trajectories for the “Swimming” pantomime

# Multi-view Action Recognition - Evaluation

| | Single Camera | | | Fusion | | |
|---|---|---|---|---|---|---|
| Feat. | Kinect #1 | Kinect #2 | Kinect #3 | MEAN | MIN | MAX |
| Traj. | 63.08 | 48.62 | 45.54 | 64.00 | 61.23 | 62.15 |
| HOG | 39.69 | 32.00 | 27.69 | 43.38 | 35.38 | 41.85 |
| HOF | 68.31 | 56.31 | 48.62 | 68.31 | 65.54 | 68.92 |
| MBH | 70.77 | 60.92 | 61.85 | 74.46 | 73.54 | 72.31 |
| Comb. | 73.85 | 63.38 | 60.00 | **74.46** | **74.46** | 73.85 |

- 13 classes of pantomime actions
- Average classification accuracy (%) for the employed gestures performed by 28 children (development corpus).
- Results for the five different features for both single and multi-steam cases.
- Results on spontaneous corpus (31 children): ~69%

BABY ROBOT
NEXTGEN SOCIAL ROBOTICS

# Multi-view Evaluation for Action: Dense Trajectories

| Dense Trajectories (DT) Features | | | | | | |
|---|---|---|---|---|---|---|
| Fusion / Desc. | Feature Fusion | | Encodings Fusion | | Score Fusion | |
| | BoVW | VLAD | BoVW | VLAD | BoVW | VLAD |
| Traj. | 59.38 | 54.15 | 65.85 | 64.62 | 63.02 | 65.23 |
| HOG | 43.08 | 45.54 | 44.00 | 50.15 | 42.60 | 45.54 |
| HOF | 62.46 | 65.84 | 68.31 | 70.77 | 67.16 | 70.15 |
| MBH | 73.85 | 75.38 | 74.77 | 76.31 | 73.08 | 74.46 |
| Comb. | 74.46 | **77.54** | 74.46 | 75.69 | 73.08 | 75.08 |

- Fusion schemes improve the performance of the best single-view cases (74.2%).
- Early fusion has the best performance.

N. Efthymiou, P. Koutras, P. Filntisis, G. Potamianos, P. Maragos, "Multi-view Fusion for Action Recognition in Child-Robot Interaction", *Proc. ICIP,* 2018.

# Children-Robot Interaction: TD video-Rock Paper Scissors

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



BABY ROBOT
NEXTGEN SOCIAL ROBOTICS

# Conclusions

- **Synopsis**:
  - Audio-visual saliency and fusion for improved detection and recognition
  - More Big Data → Needs for Automatic Summarization & Video Understanding
  - 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, References, Research Projects / Sponsors

Efthymiou, Niki
Filntisis, Panagiotis
Kardaris, Nikos
Koutras, Petros
Rodomagoulakis, Isidoros
Tsiami, Antigoni
Zlatintsi, Nancy
Pitsikalis, Vassilis
Katsamanis, Nasos
Potamianos, Gerasimos
Potamianos, Alexandros
Tzafestas, Costas

For more information, demos, and current results:

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

**COGNIMUSE:** http://cognimuse.cs.ntua.gr/

**MOBOT:** http://mobot-project.eu/

**I-SUPPORT**: http://www.i-support-project.eu/


**BabyRobot**: http://www.babyrobot.eu/