SPCOM 2016

Indian Institute of Science, Bangalore, India

12^{th} June 2016

a tutorial on

multimodal gesture recognition

nassos katsamanis

http://cvsp.cs.ntua.gr/~nassos

… with the support of a fantastic group of collaborators!

ATHENA R.C. RPI Unit, CVSP

"The biggest enemy to learning is the talking teacher."
— **John Holt**

works

- multimodal speech recognition (2004-2008)
- multimodal speech inversion (2005-2009)

joint work with G. Papandreou, V. Pitsikalis, P. Maragos

works

- multimodal speech recognition (2004-2008)
- multimodal speech inversion (2005-2009)

- multimodal speech synthesis (2013-today)
- multimodal emotion recognition (2010-today)
- multimodal saliency modeling (2012-today)

"To be or not to be? That is the question."

joint work with P. Fildisis

works

- multimodal speech recognition (2004-2008)
- multimodal speech inversion (2005-2009)

- multimodal speech synthesis (2012-today)
- multimodal emotion recognition (2010-today)
- multimodal saliency modeling (2013-today)

- multimodal gesture recognition (2013-today)

ATHENA R.C.

Robotic Perception & Interaction Unit

Computer Vision, Speech Communication and Signal Processing Group

http://cvsp.cs.ntua.gr

You'll find us at ICASSP, ICIP, Interspeech, IROS, Eusipco or (more often) in Athens, Greece working on saving the world!... In our own (unique) way, of course. ☺

DIRHA

BABY ROBOT
NEXTGEN SOCIAL ROBOTICS

i-support

MOBOT

Intelligent Active MObility Assistance RoBOT integrating Multimodal Sensory Processing, Proactive Autonomy and Adaptive Interaction

SPCOM 2016

Indian Institute of Science, Bangalore, India

12th June 2016

a tutorial on

multimodal gesture recognition

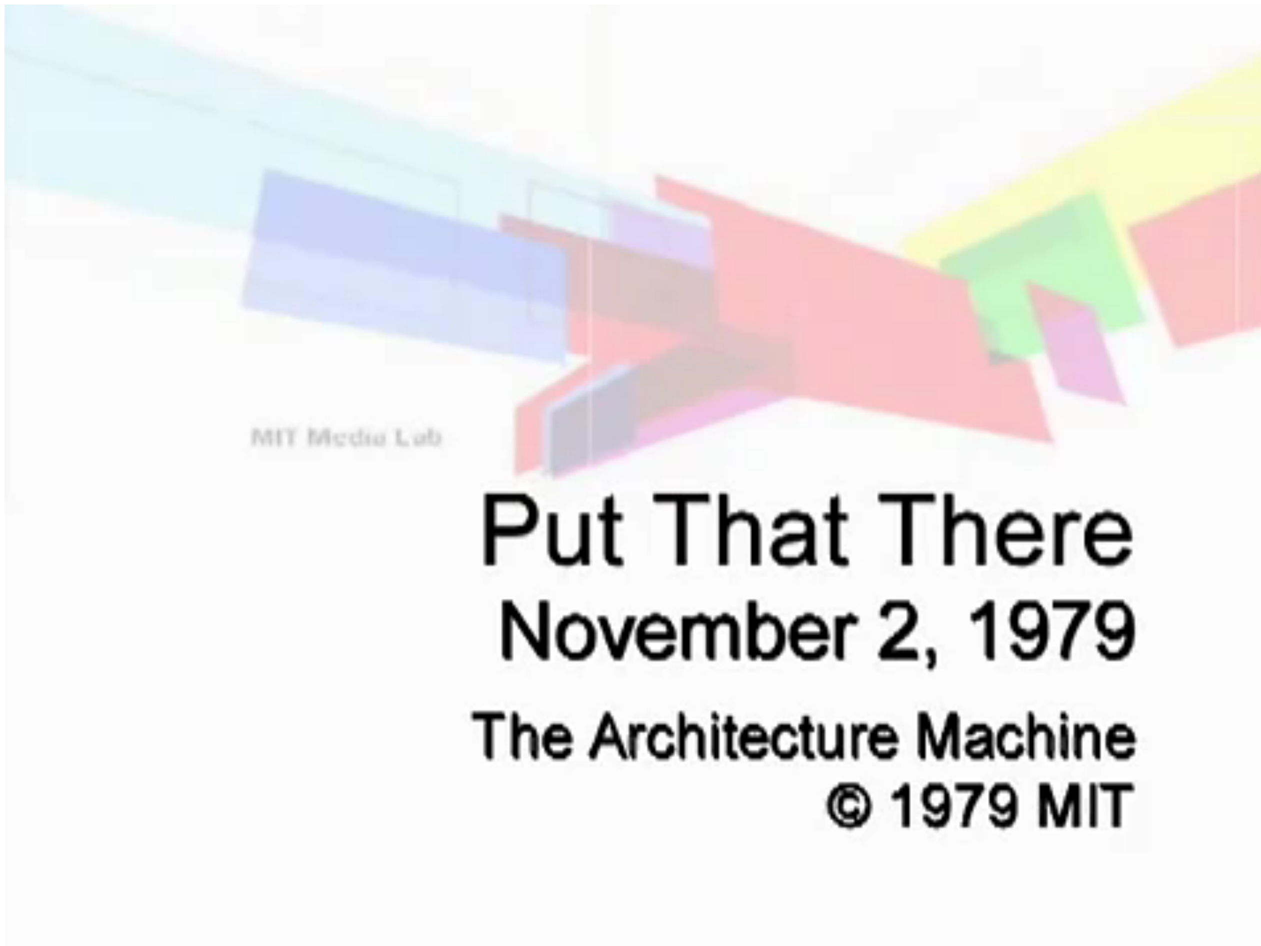
nassos katsamanis

http://cvsp.cs.ntua.gr/~nassos

… with the support of a fantastic group of collaborators!

ATHENA R.C. RPI Unit, CVSP

"put that there!"



Bolt, R. (1980). "Put that there": Voice and Gesture at the Graphics Interface speech as a whole includes lexical, emotional, semantic, phonological, syntactic, and motoric/gestural aspects

McNeill, D. (1985). So you think gestures are nonverbal?

a single unified classification scheme of gesture is merely impossible given the multitude of dimensions gesture can depend on

Kendon, A. (2004). Gesture. Visible action as utterance.

dimensions

- meaning independent of or only in conjuction with speech (Efron, 1941.)
- origin, usage, coding (Ekman & Friesen, 1969)
- form, meaning, communicative function (McNeill, 1992)
- topic related and interactive character (Bavelas, 1992)

iconic
metaphoric
beat
deictic
cohesive
emblem

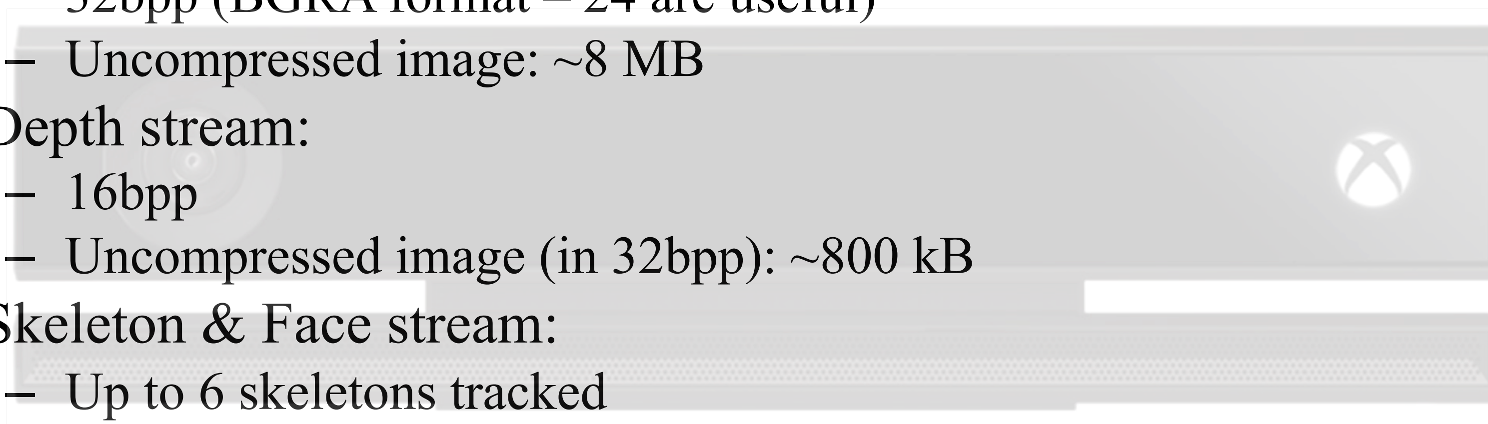
gestures help us communicate meaning and more easily retrieve words during speech gestures in computer interfaces have been viewed in the past as a language but it would be beneficial to consider them as part of a multimodal communicative event

… and the quest to create more natural and robust human-computer interfaces begins the majority of multimodal gesture recognition systems:

- first recognize events in each modality separately,
- and then fuse the decisions.

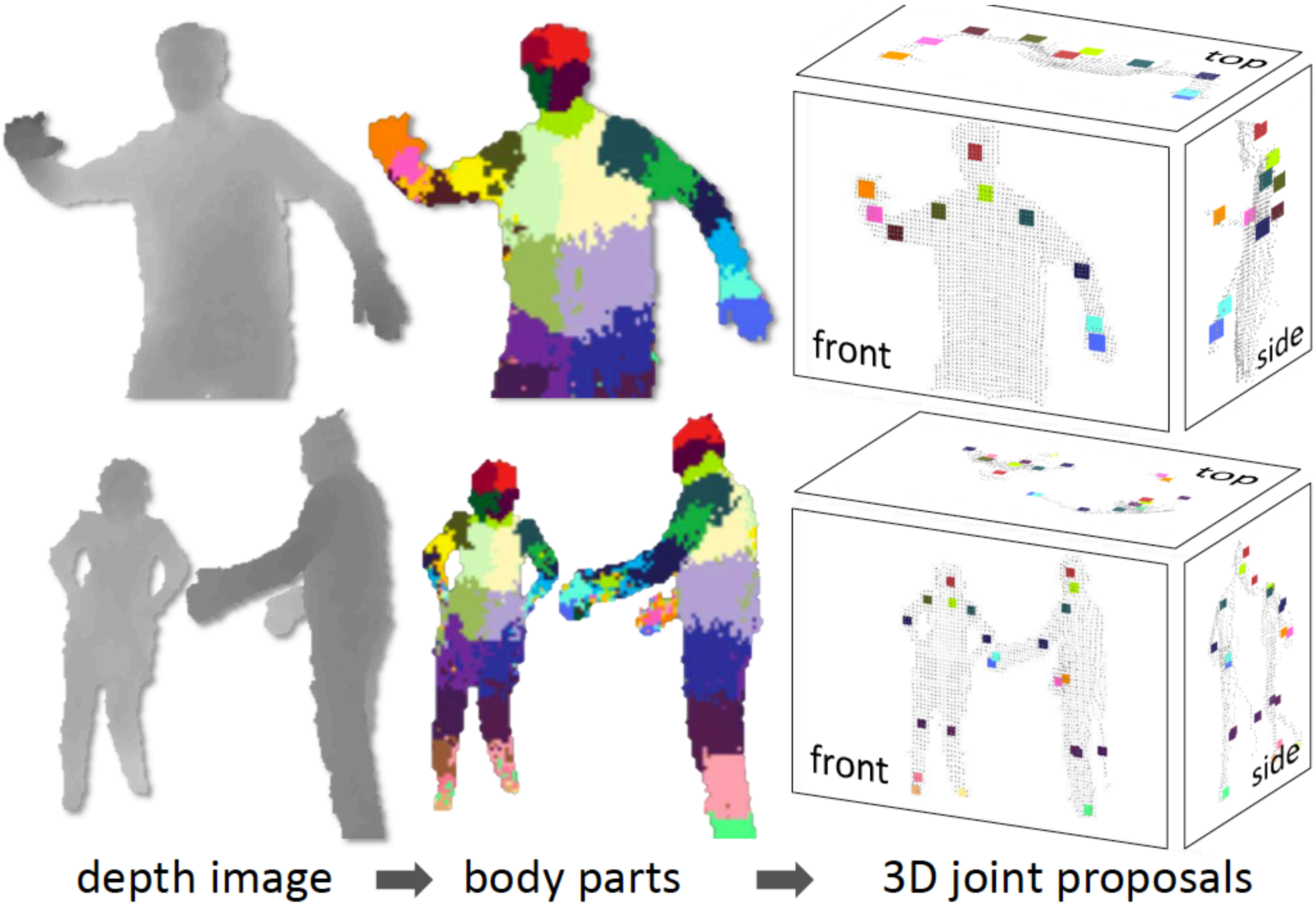
1, 2, 3… action!

Kinect details



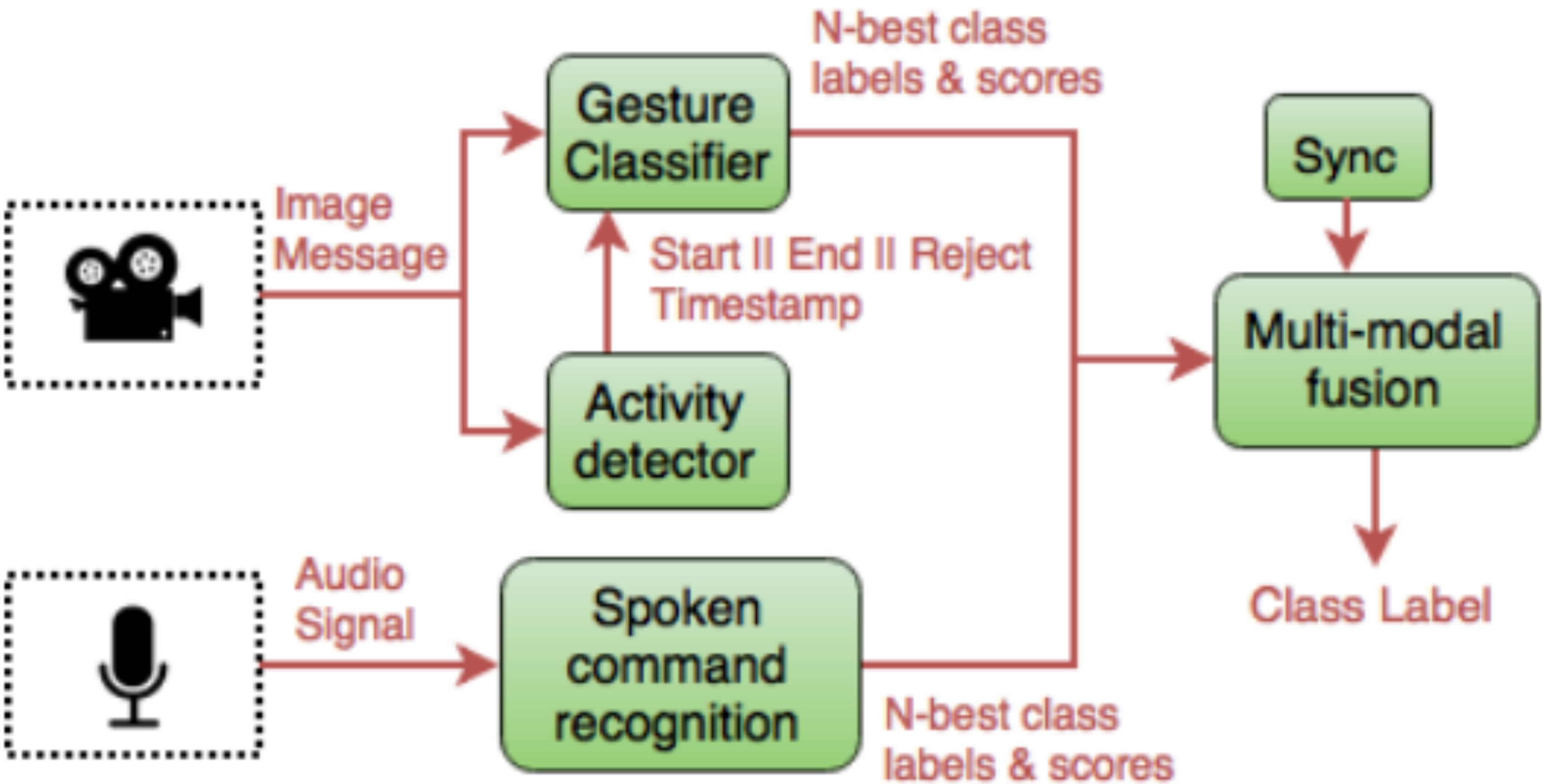
- Captures data at 30 fps
- Color stream:
  - 32bpp (BGRA format – 24 are useful)
  - Uncompressed image: ~8 MB
- Depth stream:
  - 16bpp
  - Uncompressed image (in 32bpp): ~800 kB
- Skeleton & Face stream:
  - Up to 6 skeletons tracked
  - Basic hand gestures included (closed, open, lasso)
  - Face position & properties
- Audio
  - 4 streams at 44100Hz (raw)
  - 1 clean audio stream (processed)

skeleton tracking



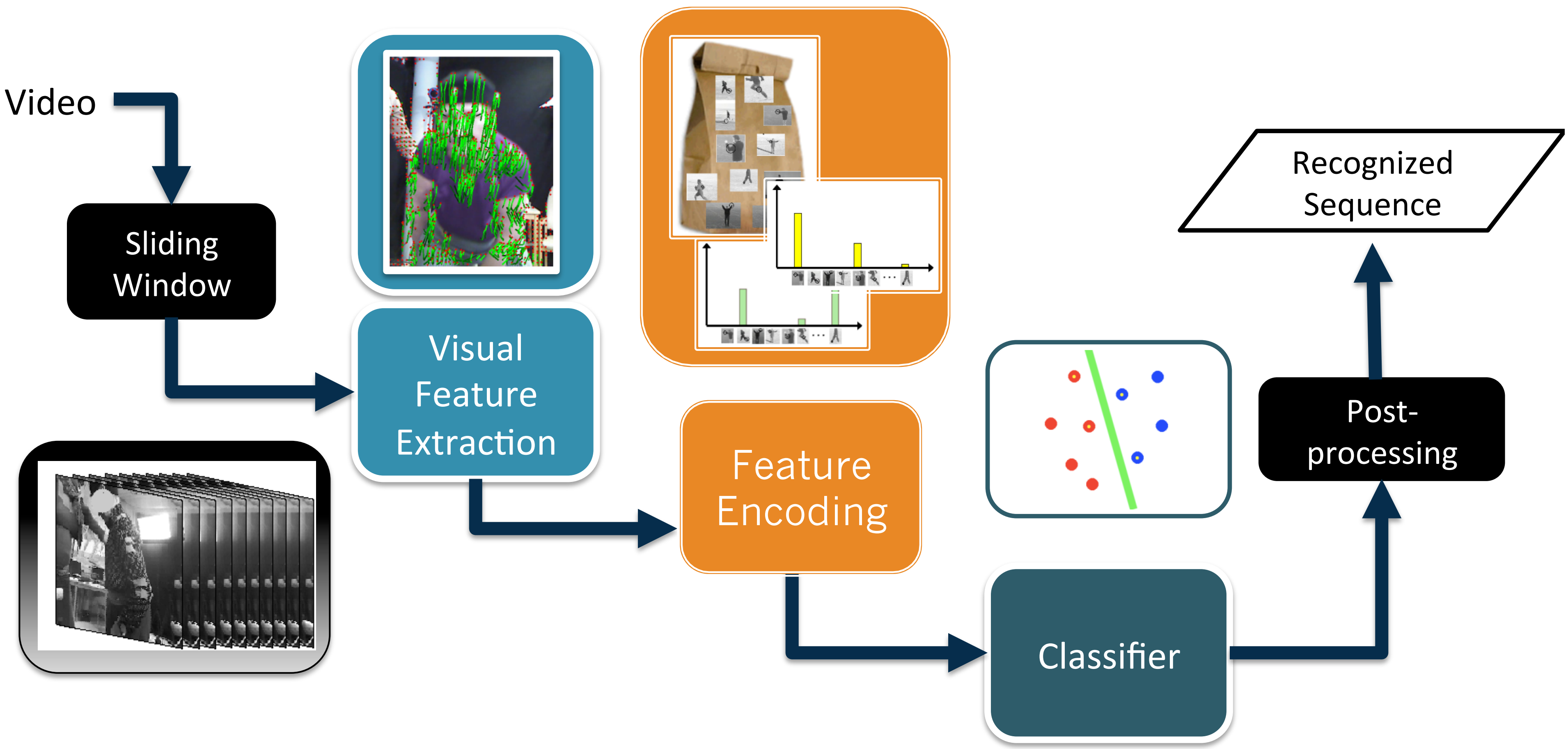
Shotton et al. (2011). Real-Time Human Pose Recognition in Parts from Single Depth Images

online gesture recognition system

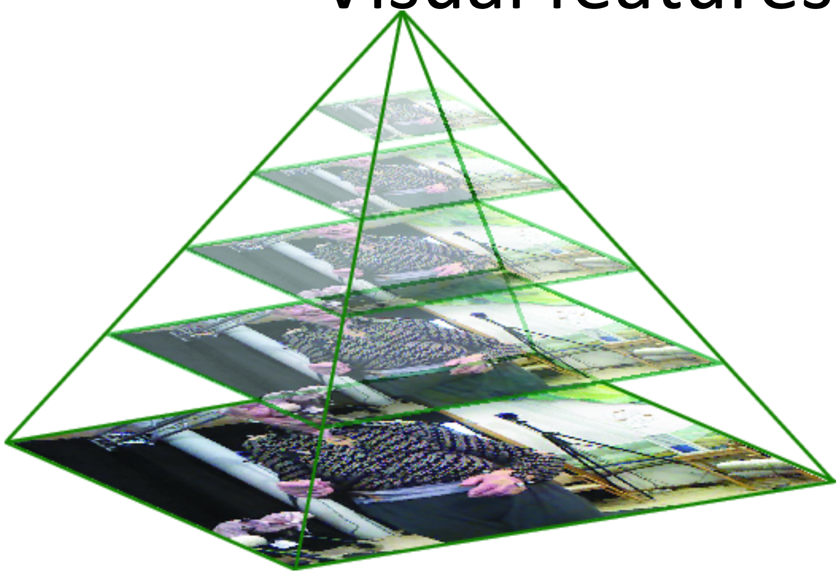


Rodomagoulakis et al. (2016). MM Human Action Recog. in Assistive Human-Robot Interaction

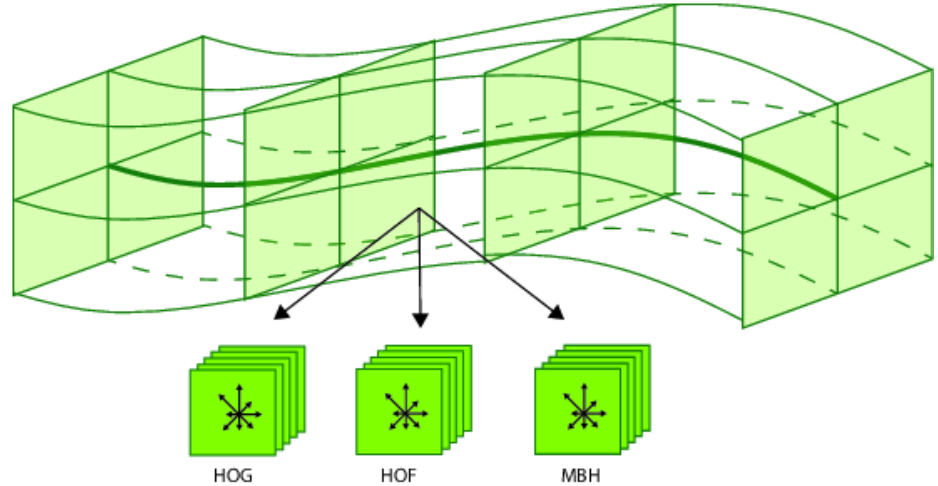
visual recognition pipeline



Visual features : Dense Trajectories



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

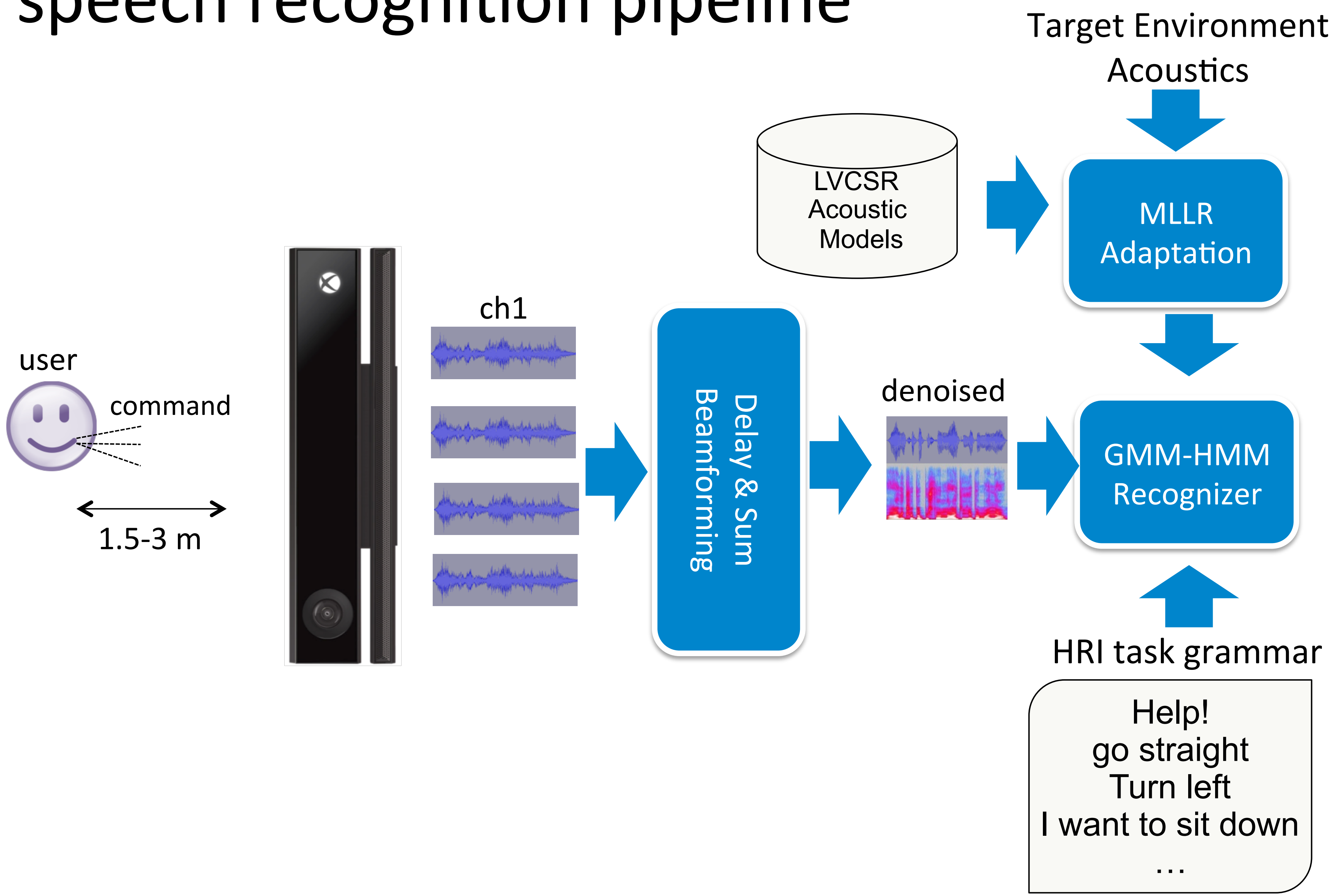


3. Descriptors are computed in space-time volumes along trajectories

2. Feature points are tracked through consecutive video frames

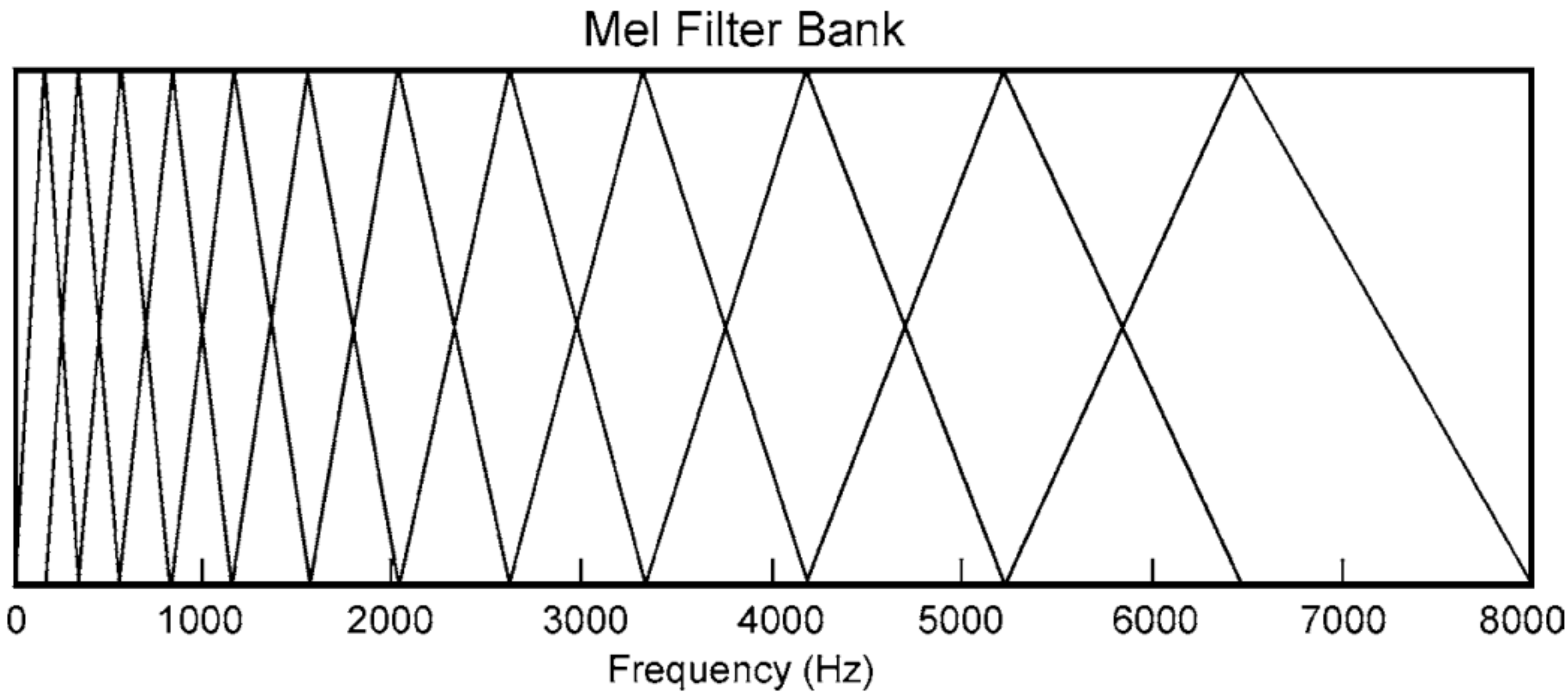
[Wang et al. IJCV 2013]

speech recognition pipeline

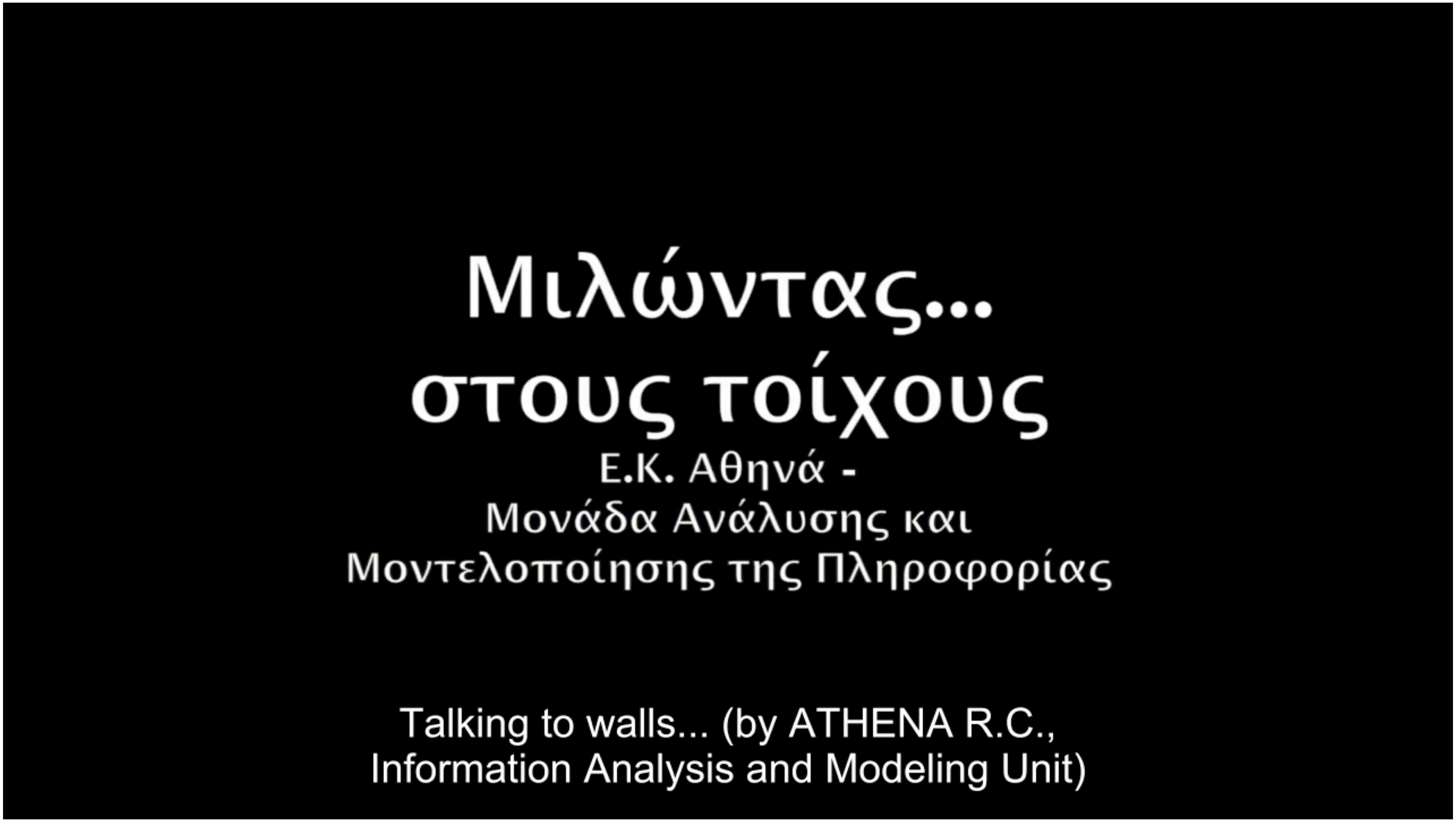


speech feature extraction

- Mel Frequency Cepstral Coefficients
- Mel Filterbank Energies

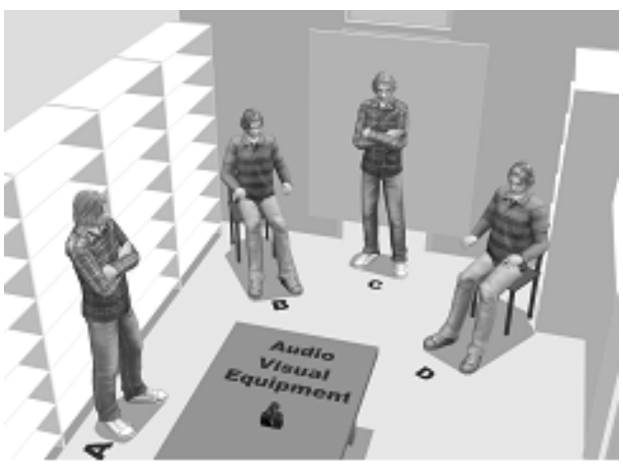


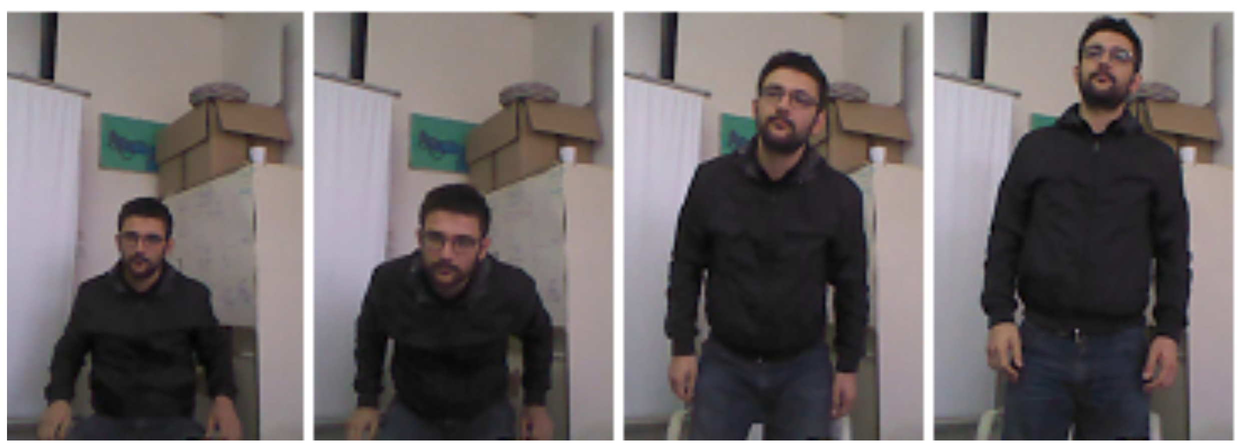
speech recognition pipeline @ work



multimodal gesture recognition data

- Multiple conditions- scenarios-setups: e.g., mixed sit/stand, near/far, angle of view,
- non-strict setups
- 13 subjects,
- 19 audio-gestural commands
- Greek spoken commands
- 5 iterations distributed in variable conditions



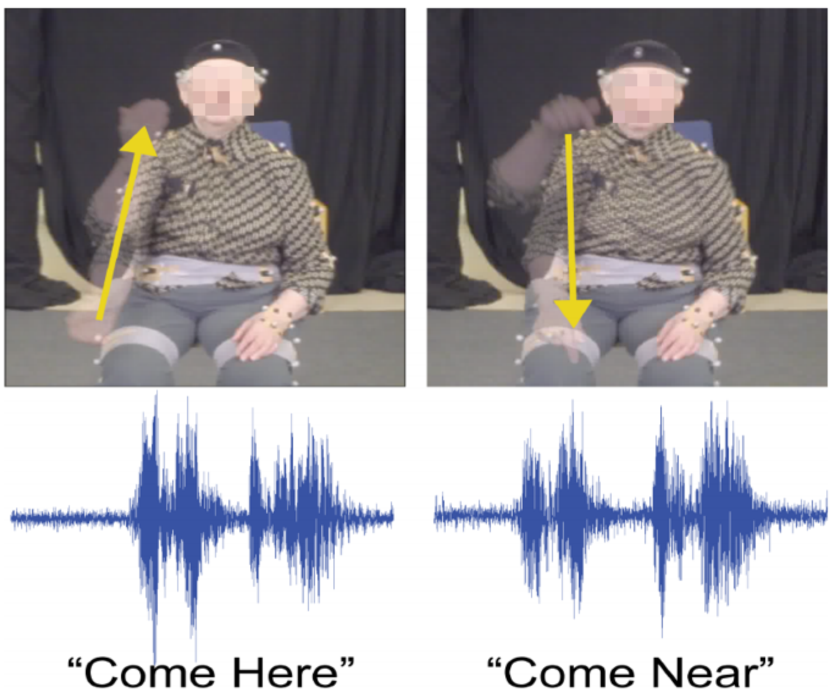


…and then

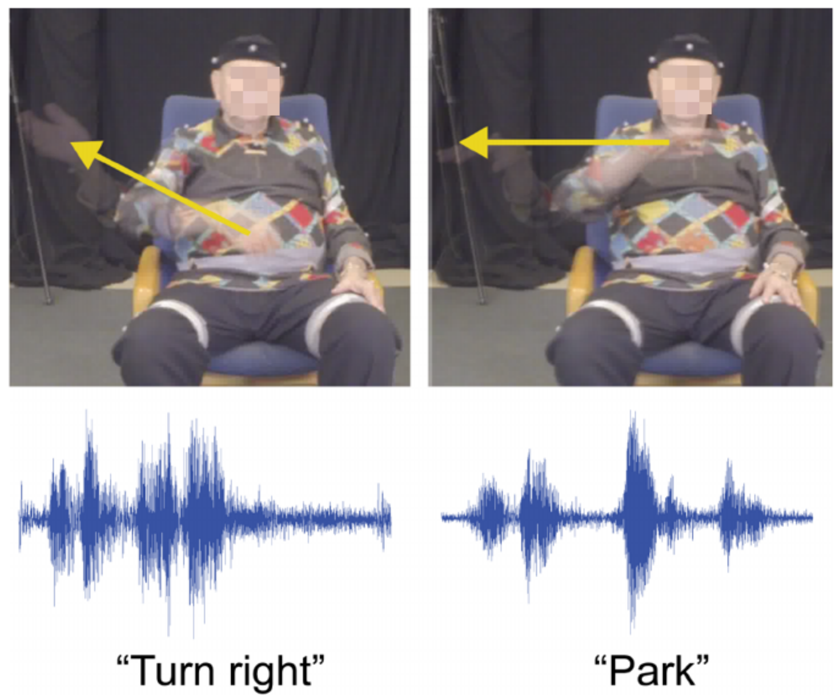
it’s fusion!

Multimodal fusion: Complementarity of visual and audio modalities

Similar audio,
distinguishable gesture



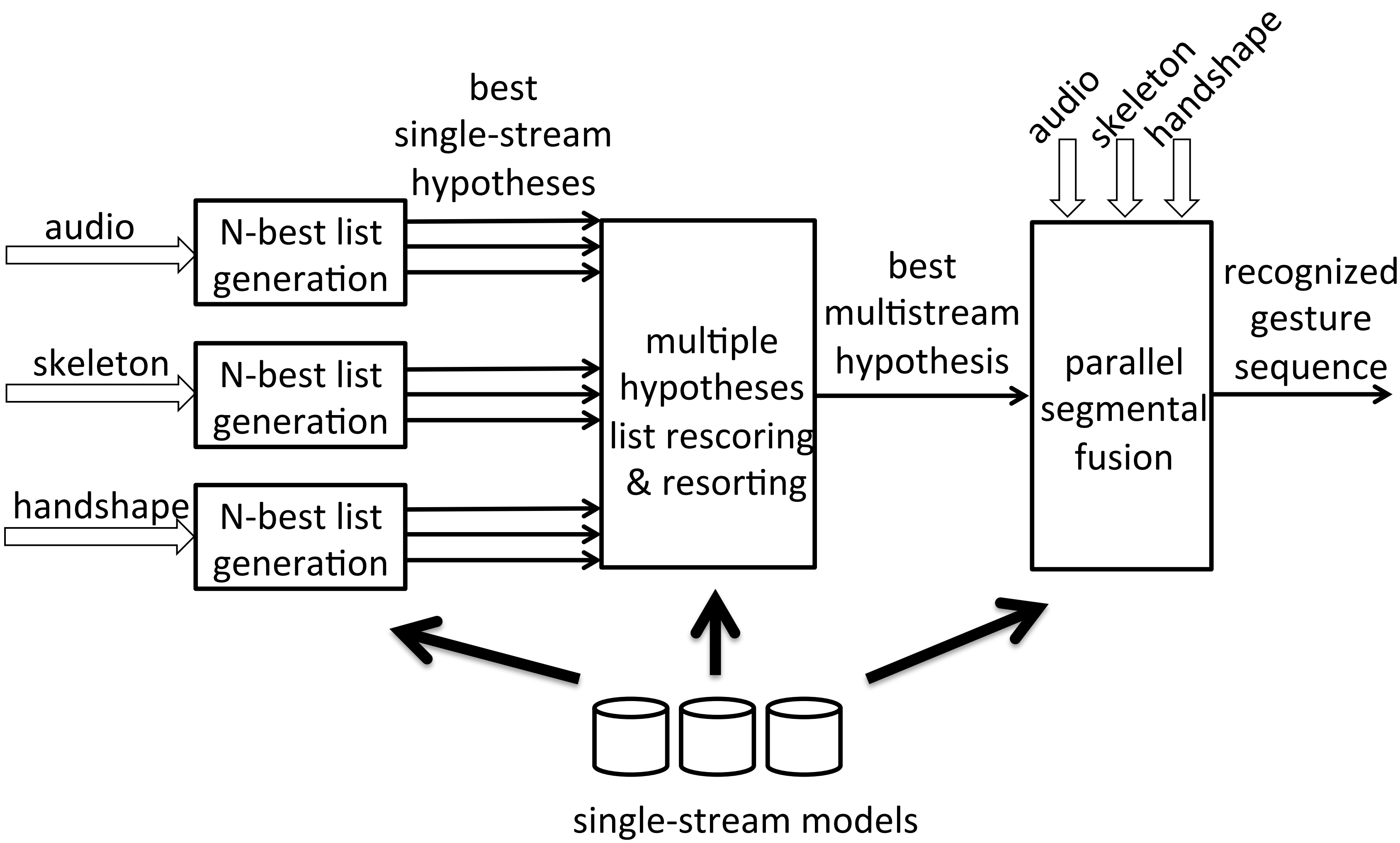
Distinguishable audio,
similar gesture



fusion approaches

- Early fusion
- Late fusion
  - Multiple hypotheses rescoring
  - Hypotheses rescoring with time constraints
  - Score normalization
  - …

overall fusion scheme



Pitsikalis et al. (2015). Multimodal Gesture Recognition via Multiple Hypotheses Rescoring

multiple hypotheses rescoring

Algorithm 1 Multimodal Scoring and Resorting of Hypotheses

% N-best list rescoring
for all hypotheses **do**
  % Create a constrained grammar
  keep the sequence of gestures fixed
  allow introduction/deletion of *sil* and *bm* occurences between gestures
  for all modalities **do**
    by applying the constrained grammar and Viterbi decoding:
    1) find the best state sequence given the observations
    2) save corresponding score and temporal boundaries
  % Late fusion to rescore hypotheses
  final hypothesis score is a weighted sum of modality-based scores
the best hypothesis of the 1st-pass is the one with the maximum score

Pitsikalis et al. (2015). Multimodal Gesture Recognition via Multiple Hypotheses Rescoring

segmental parallel fusion

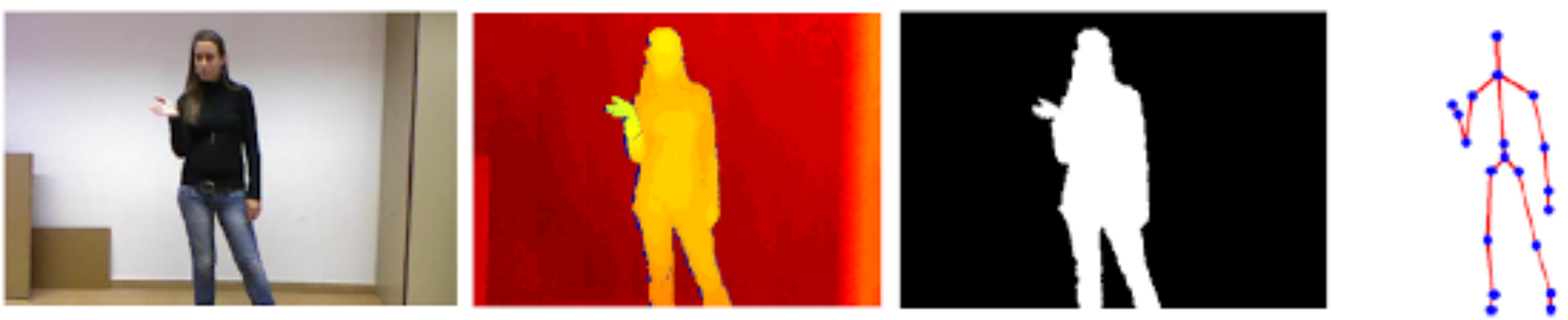
Algorithm 2 Segmental Parallel Fusion

% Parallel scoring
for all modalities **do** segment observations based on given temporal boundaries
    for all resulting segments **do**
        estimate a score for each gesture given the segment observations
        temporally align modality segments
        for all aligned segments **do**
            estimate weighted sum of modality-based scores for all gestures
            select the best-scoring gesture (*sil* and *bm* included)

Pitsikalis et al. (2015). Multimodal Gesture Recognition via Multiple Hypotheses Rescoring

a popular dataset

- ChaLearn 2013: using kinect for multimodal gesture recognition
  - RGB, depth, audio, skeleton



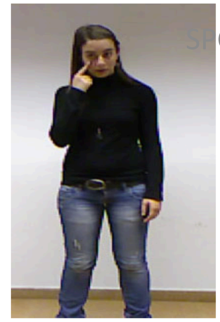
- 20 cultural/anthropological signs of Italian language
- 22 different users
- 20 repeats per user approximately (~1 minute for each gesture video)

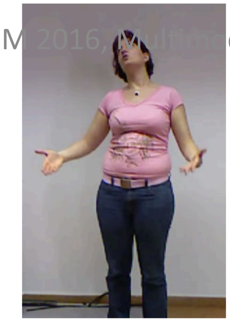
Escalera et al. (2013). Multimodal Gesture Recognition Challenge

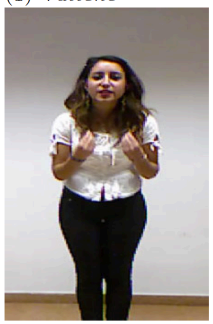

(1) *Vattene*


(2) *Viene qui*

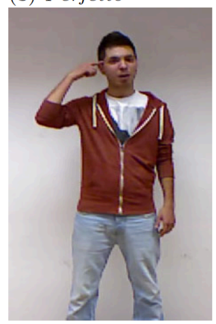

(3) *Perfetto*

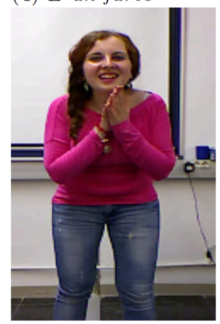

(4) *E un furbo*


(5) *Che due palle*

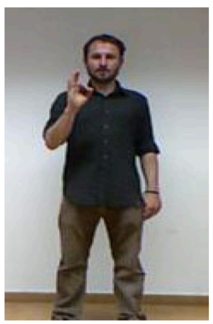

(6) *Che vuoi*

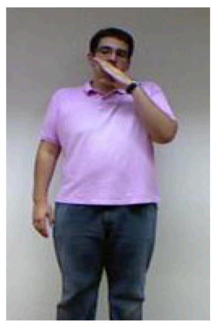

(7) *Vanno d'accordo*

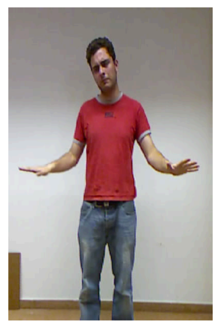

(8) *Sei pazzo*

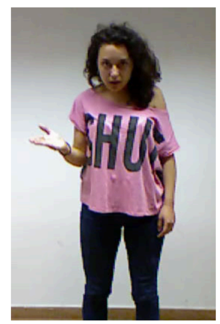

(9) *Cos hai combinato*

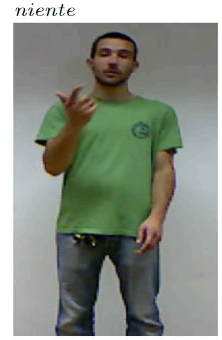

(10) *Nonme me frie niente*

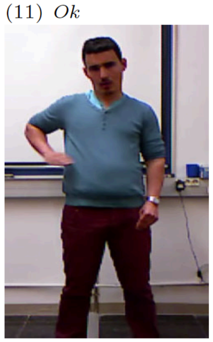

(11) *Ok*

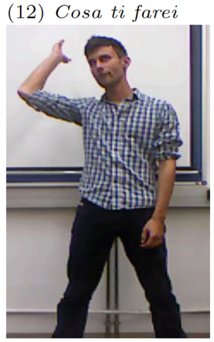

(12) *Cosa ti farei*

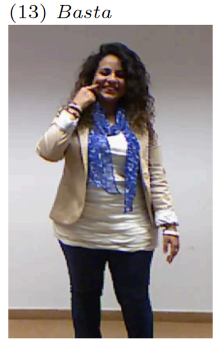

(13) *Basta*

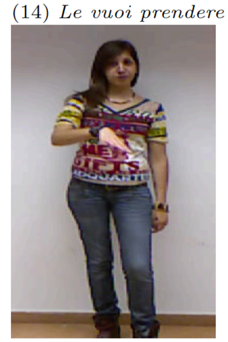

(14) *Le vuoi prendere*

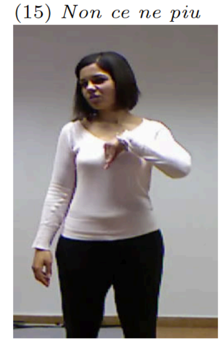

(15) *Non ce ne piu*


(16) *Ho fame*


(17) *Tanto tempo fa*

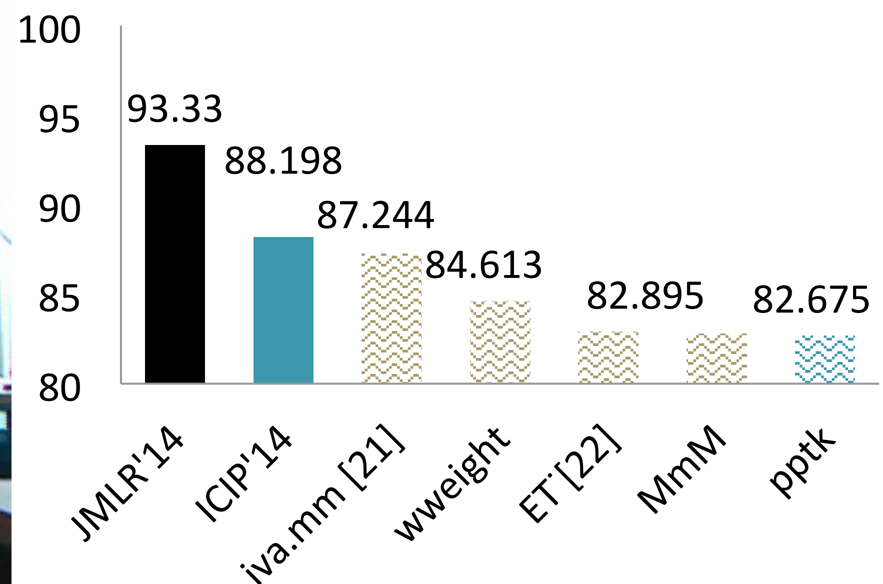

(18) *Buonissimo*


(19) *Si sono messi d'accordo*


(20) *Sono stufo*

results

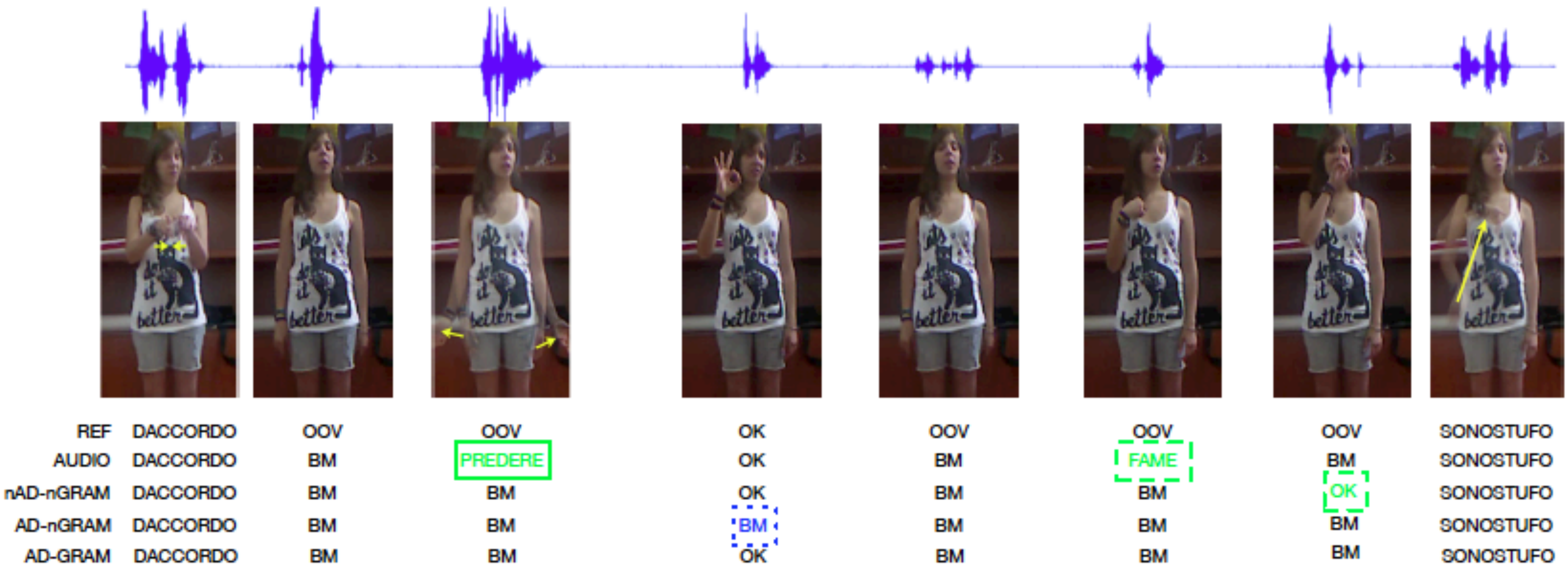
Decoding video example
ChaLearn challenge data



Best result in ChaLearn challenge: +7%

[21] Wu et al. (2013). Fusing multi-modal features for gesture recognition.
[22] Bayer and Silvermann (2013). A multi modal approach to gesture recognition

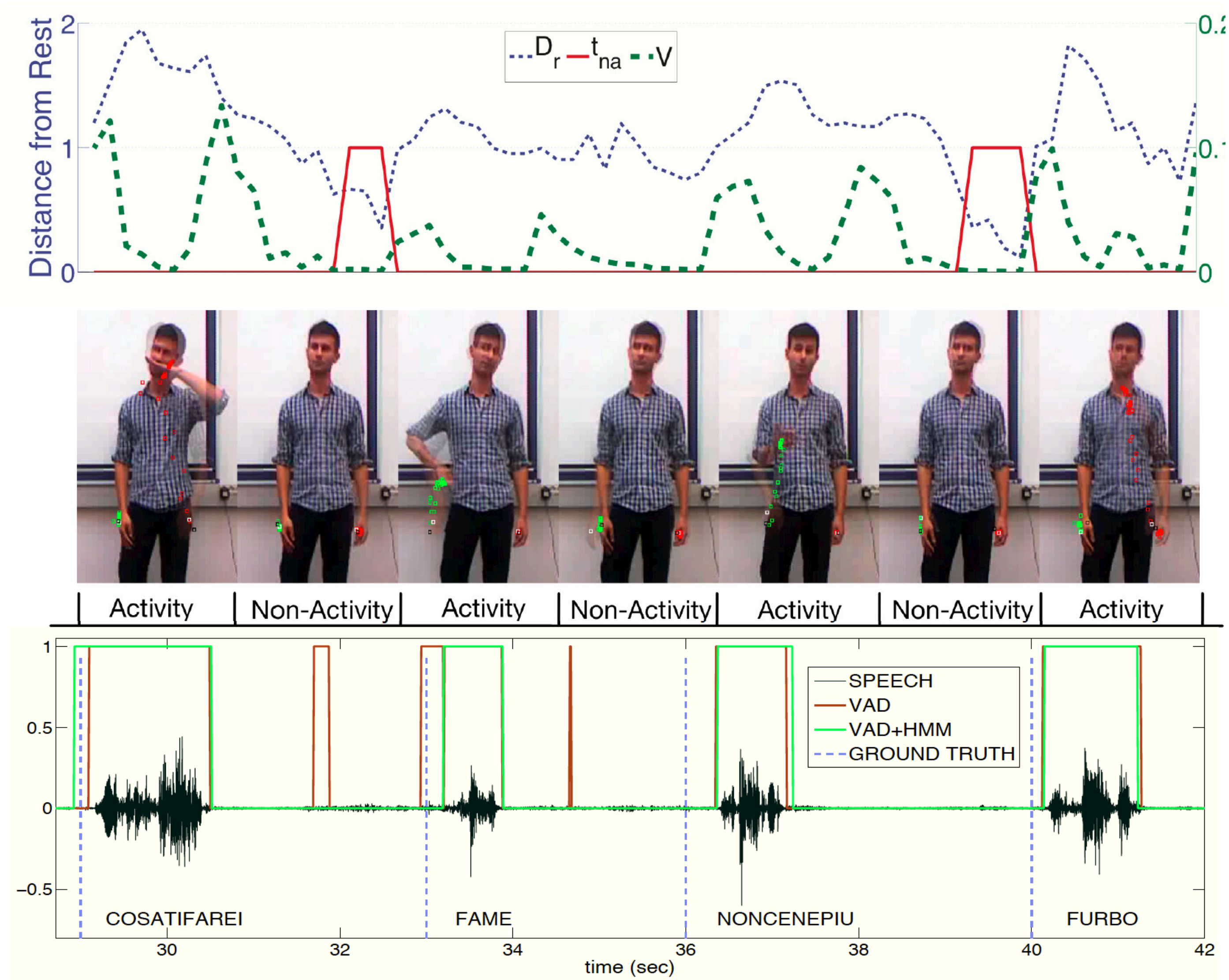
Audio-Visual Fusion & Recognition



- Audio and visual modalities for A-V gesture word sequence.
- Ground truth transcriptions ("REF") and decoding results for audio and 3 different fusion schemes.

Pitsikalis et al. (2015). Multimodal Gesture Recognition via Multiple Hypotheses Rescoring

activity detection



Pitsikalis et al. (2015). Multimodal Gesture Recognition via Multiple Hypotheses Rescoring

results (1)

AD	Single Modalities		
	Aud.	*Skel.*	*HS*
✗	78.4	47.6	13.3
✓	87.2	49.1	20.2

results (2)

	Method/ Exp. Code	Modality	Segm. Method	Classifier/ Modeling	Fusion	Acc. (%)	LD
Others	O1: 1st Rank*	SK, AU	AU:time-domain	HMM, DTW	Late:w-sum	87.24	0.1280
	O2: 2nd Rank†	SK, AU	AU:energy	RF, KNN	Late:posteriors	84.61	0.1540
	O3: 3rd Rank‡	SK, AU	AU:detection	RF, Boosting	Late:w-average	82.90	0.1710
2 Streams	s2-A1	SK,AU	HMM	AD, HMM	Late:SPF	87.9	0.1210
	s2-B1	SK,AU	-	AD,HMM,GRAM	Late:MHS	92.8	0.0720
	s2-A2	HS,AU	HMM	AD, HMM	Late:SPF	87.7	0.1230
	s2-B2	HS,AU	-	AD,HMM,GRAM	Late:MHS	87.5	0.1250
3 Streams	C1	SK,AU,HS	HMM	AD, HMM	Late:SPF	88.5	0.1150
	D1	SK,AU,HS	-	HMM	Late:MHS	85.80	0.1420
	D2	SK,AU,HS	-	AD,HMM	Late:MHS	91.92	0.0808
	D3	SK,AU,HS	-	AD,HMM,GRAM	Late:MHS	93.06	0.0694
	E1	SK,AU,HS	HMM	HMM	Late:MHS+SPF	87.10	0.1290
	E2	SK,AU,HS	HMM	AD,HMM	Late:MHS+SPF	92.28	0.0772
	E3	SK,AU,HS	HMM	AD,HMM,GRAM	Late:MHS+SPF	93.33	0.0670

*(Wu et al., 2013); † (Escalera et al., 2013b); ‡ (Bayer and Thierry, 2013)

results (2)

	Method/ Exp. Code	Modality	Segm. Method	Classifier/ Modeling	Fusion	Acc. (%)	LD
Others	O1: 1st Rank*	SK, AU	AU:time-domain	HMM, DTW	Late:w-sum	87.24	0.1280
	O2: 2nd Rank†	SK, AU	AU:energy	RF, KNN	Late:posteriors	84.61	0.1540
	O3: 3rd Rank‡	[illegible]	[illegible]	[illegible]	[illegible]	82.90	0.1710
2 Streams	s2-A1	SK,AU	HMM	AD, HMM	Late:SPF	87.9	0.1210
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	s2-A2	HS,AU	HMM	AD, HMM	Late:SPF	87.7	0.1230
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.1250
3 Streams	C1	SK,AU,HS	HMM	AD, HMM	Late:SPF	88.5	0.1150
	D1	SK,AU,HS	[illegible]	[illegible]	Late:MHS	85.80	0.1420
	D2	SK,AU,HS	-	AD,HMM	Late:MHS	91.92	0.0808
	D3	SK,AU,HS	-	AD,HMM,GRAM	Late:MHS	93.06	0.0694
	E1	SK,AU,HS	HMM	HMM	Late:MHS+SPF	87.10	0.1290
	E2	SK,AU,HS	HMM	AD,HMM	Late:MHS+SPF	92.28	0.0772
	E3	SK,AU,HS	HMM	AD,HMM,GRAM	Late:MHS+SPF	93.33	0.0670

*(Wu et al., 2013); † (Escalera et al., 2013b); ‡ (Bayer and Thierry, 2013)

From 87.2% using only audio performance improved to 93.33% which corresponds to a 50% relative error reduction!

approaches

Team	Score	Modalities	Fusion	Classifier
IVA MM	0123	AU, SK	Late	HMM, DP, KNN
WWEIGHT	0154	AU, SK	Late	RF, KNN
ET	0.169	AU, SK	Late	Tree, RF, ADA
MmM	0.172	AU,RGB+Depth	Late	SVM, GMM, KNN
PPTK	0.173	SK, RGB+Depth	Late	GMM, HMM
LRS	0.178	AU, SK, Depth	Early	NN
MMDL	0.244	AU, SK, RGB	Late	DBM+LR
TELEPOINTS	0.26	AU, SK, RGB	Late	HMM, SVM
CSI MM	0.29	AU, SK	Early	HMM

Escalera et al. (2013). Multimodal Gesture Recognition Challenge

yet another application

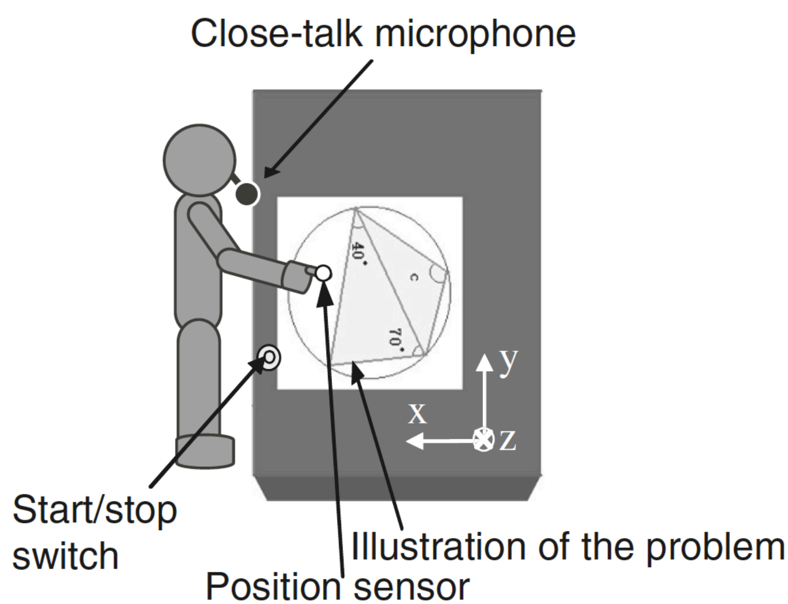


Figure 2 Recording setup.

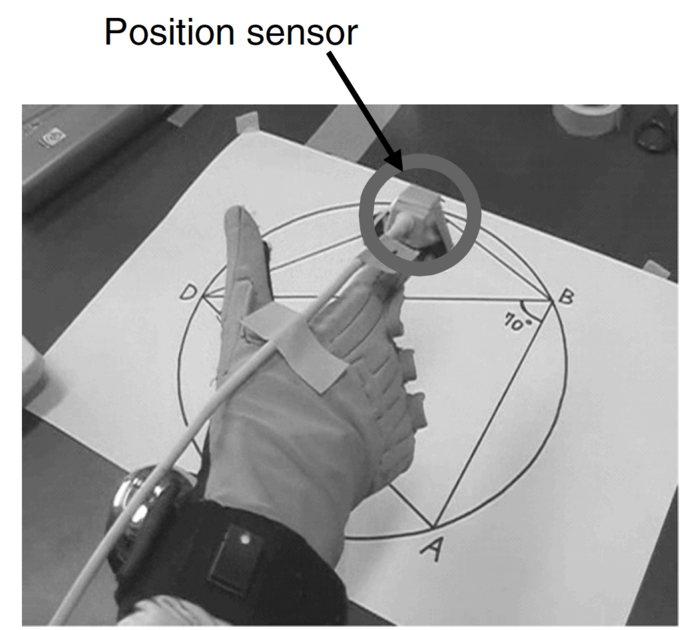


Figure 3 Position sensor apparatus.

Miki et al.(2014). Improvement of MM gesture and speech recog. performance

introducing time constraints

$$\tau = t_s - t_g$$

$$p_d(\tau) = \frac{1}{\sqrt{2\pi}\,\sigma_\tau} \exp\left\{ -\frac{(\tau - \mu_\tau)^2}{2\sigma_\tau^2} \right\}$$

$$L(u_i, g_j) = \begin{cases} \alpha L_s(u_i) + \beta L_g(g_j) + \gamma \log p_d(t_{s_i} - t_{g_j}), & \text{if } M(u_i, g_j) = 1, \\ -\infty, & \text{if } M(u_i, g_j) = 0 \end{cases}$$

Miki et al.(2014). Improvement of MM gesture and speech recog. performance

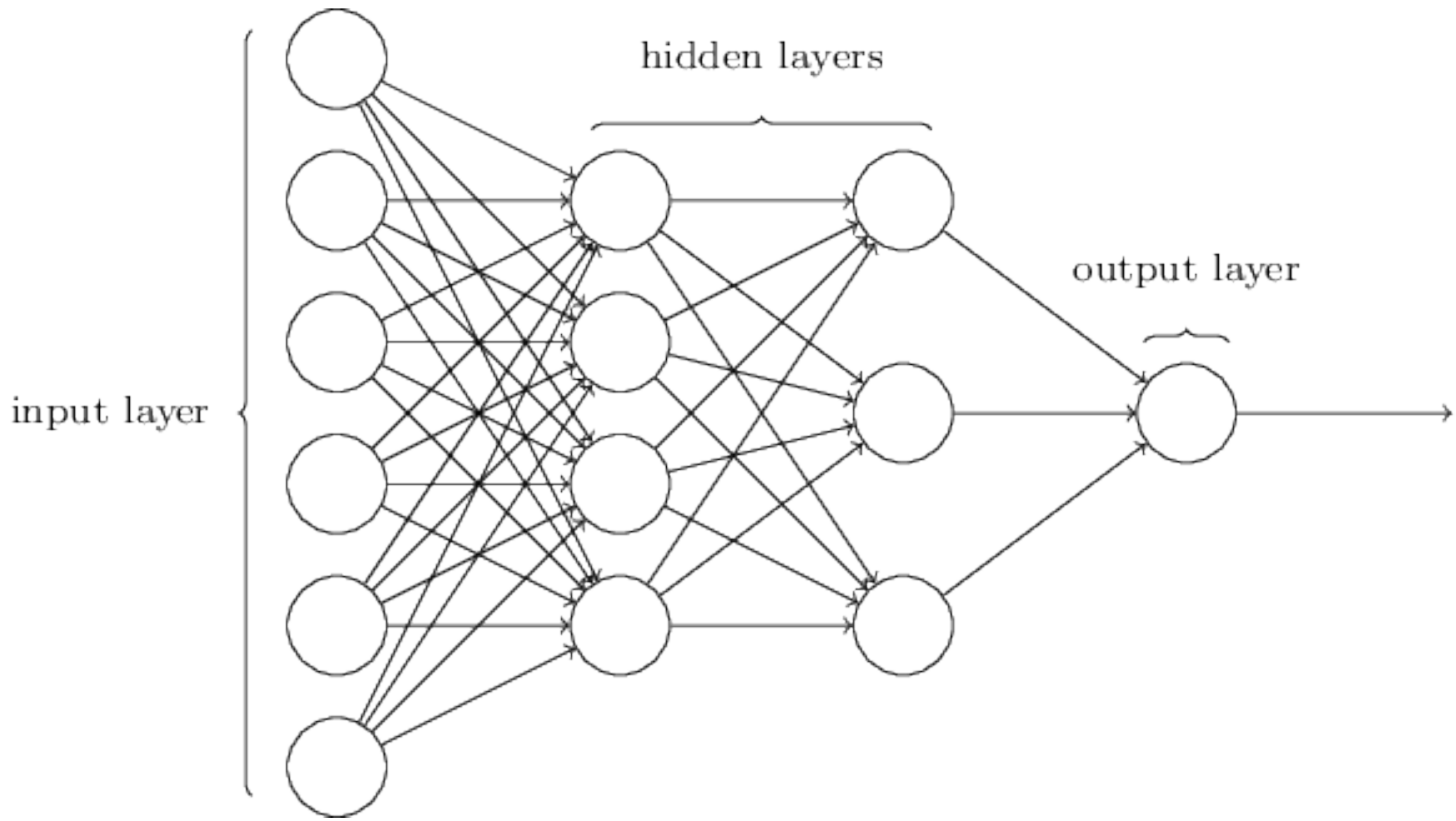
results

Modality		Recognition rate	
		Speech	Gesture
Speech	1-best	75.0	-
	20-best	80.0	-
Gesture	1-best	-	91.0
	20-best	-	94.7
Speech and gesture	-	*78.4*	*94.7*

Miki et al.(2014). Improvement of MM gesture and speech recog. performance

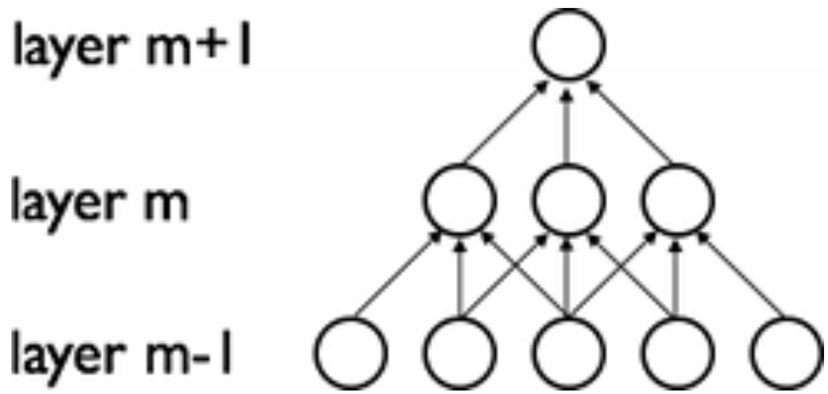
modeling

- Instead of GMMs, emission probabilities can be estimated by (deep) neural networks

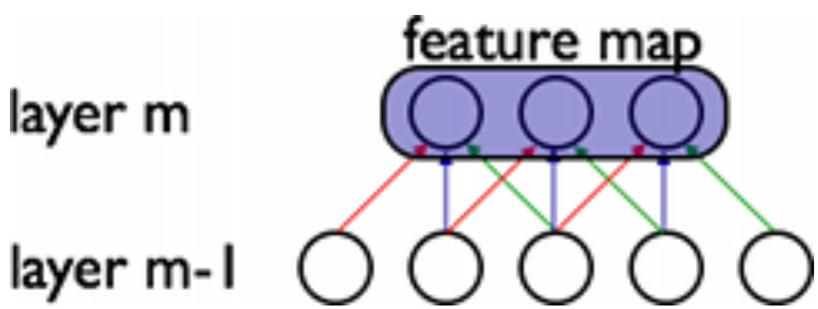


convolutional layers

- Local connectivity is enforced

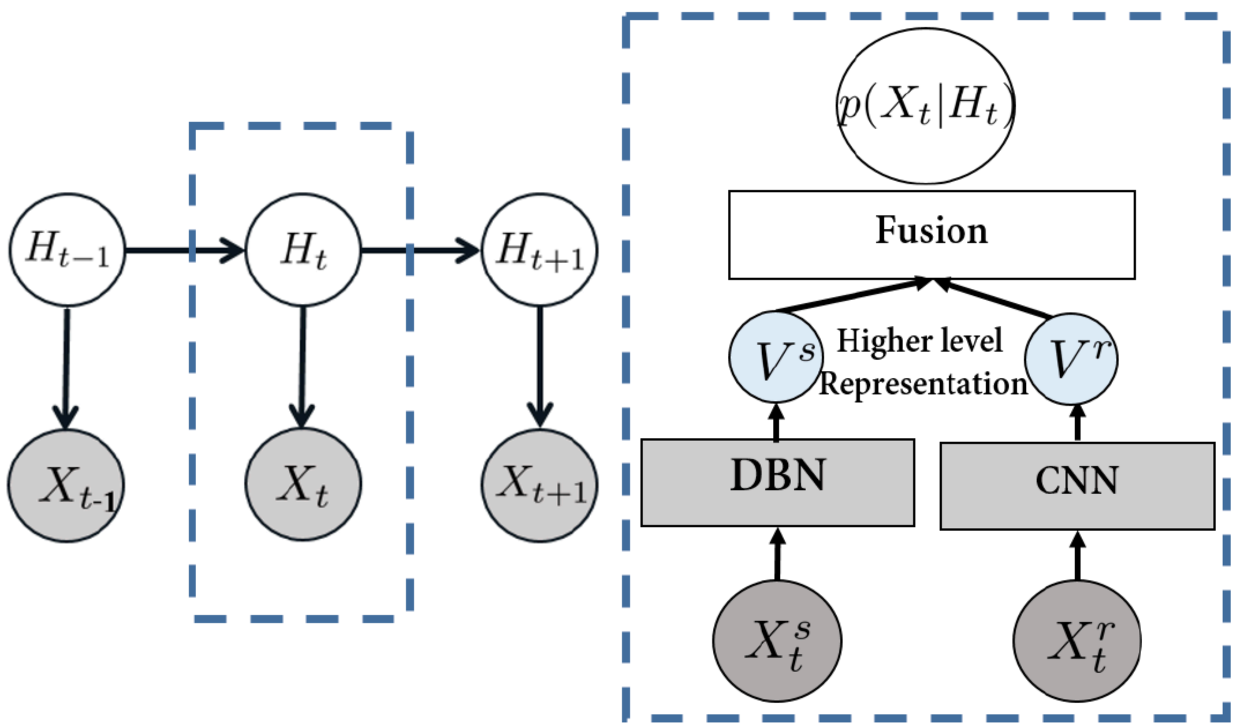


- Weights are shared



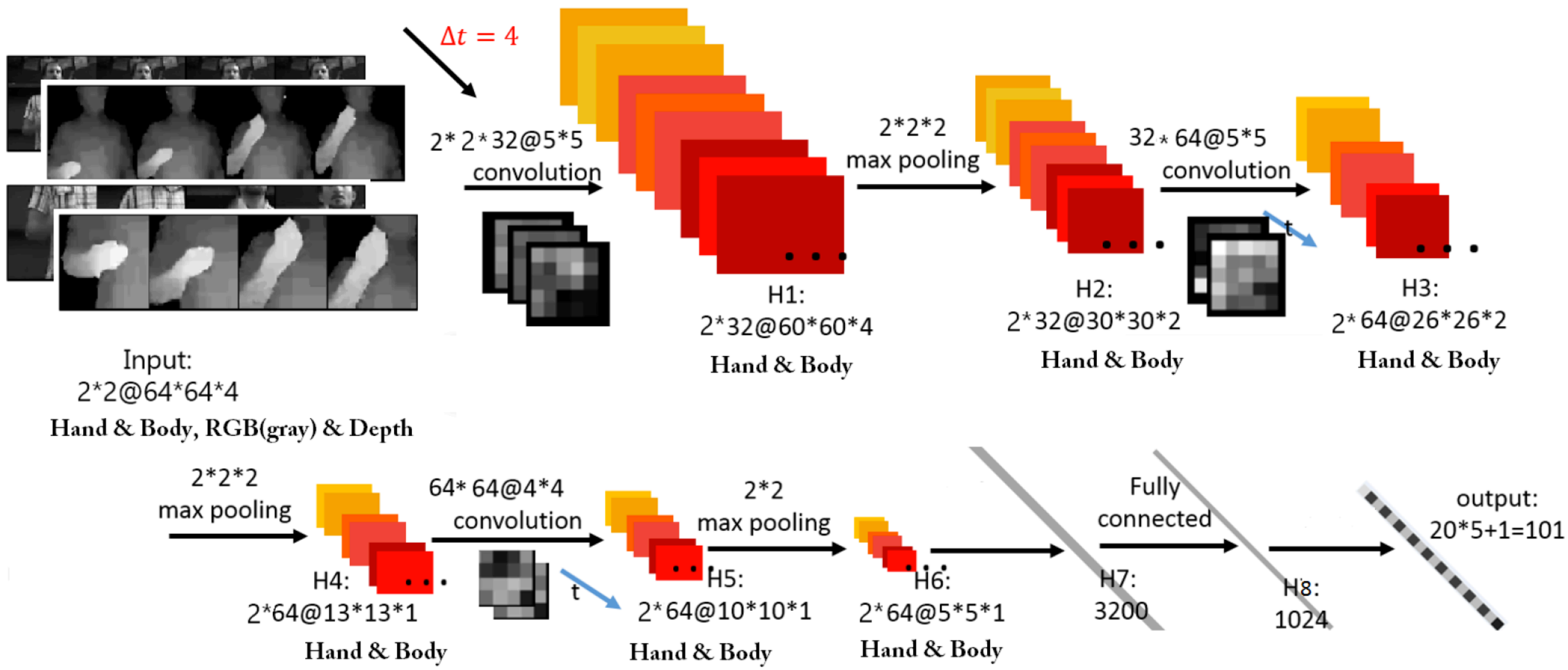
visual gesture recognition (1)

- Deep Dynamic Neural Networks for Gesture Recognition



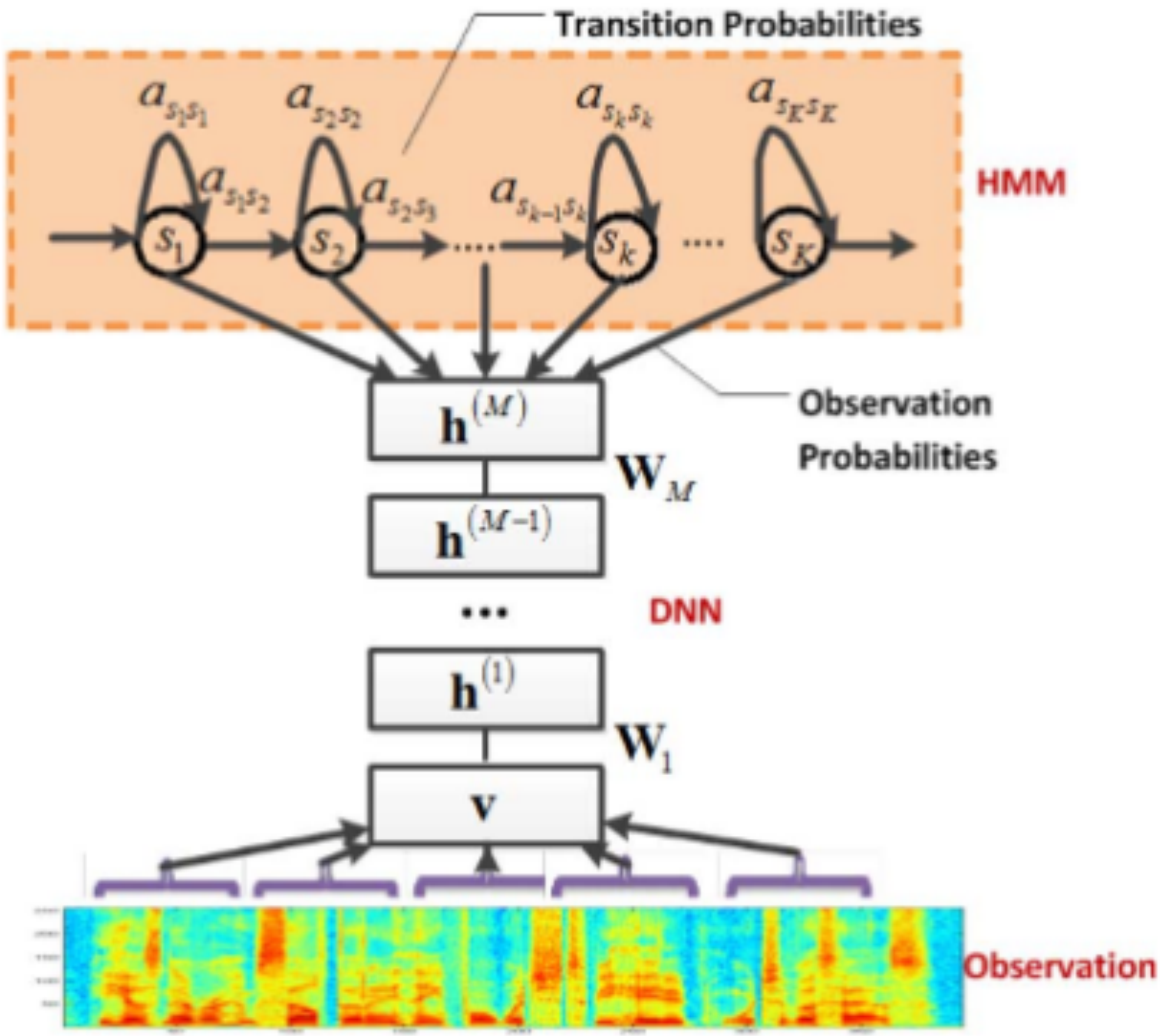
Wu et al.(2016). Deep Dynamic NNs for MM Gesture Segmentation and Recognition

visual gesture recognition (2)



Wu et al.(2016). Deep Dynamic NNs for MM Gesture Segmentation and Recognition

speech recognition



Slide from: Dong Yu, "Deep Learning for Automatic Speech Recognition"

challenges

- What if one of the available streams is noisy?
  - Or completely missing?
- Recognize gestures and enhance understanding during conversation
- Temporal modeling can possibly be significantly improved
  - Use HCRF or RNNs with LSTM nodes

thanks to collaborators!

- Niki Efthymiou
- Panagiotis Fildisis
- Nikos Kardaris
- Petros Koutras
- Petros Maragos
- Vassilis Pitsikalis
- Isidoros Rodomagoulakis
- Stavros Theodorakis
- Antigoni Tsiami

sponsors

Intelligent Active MObility Assistance RoBOT integrating Multimodal Sensory Processing, Proactive Autonomy and Adaptive Interaction