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

Part 3:
Audio-Visual Child-Robot Interaction

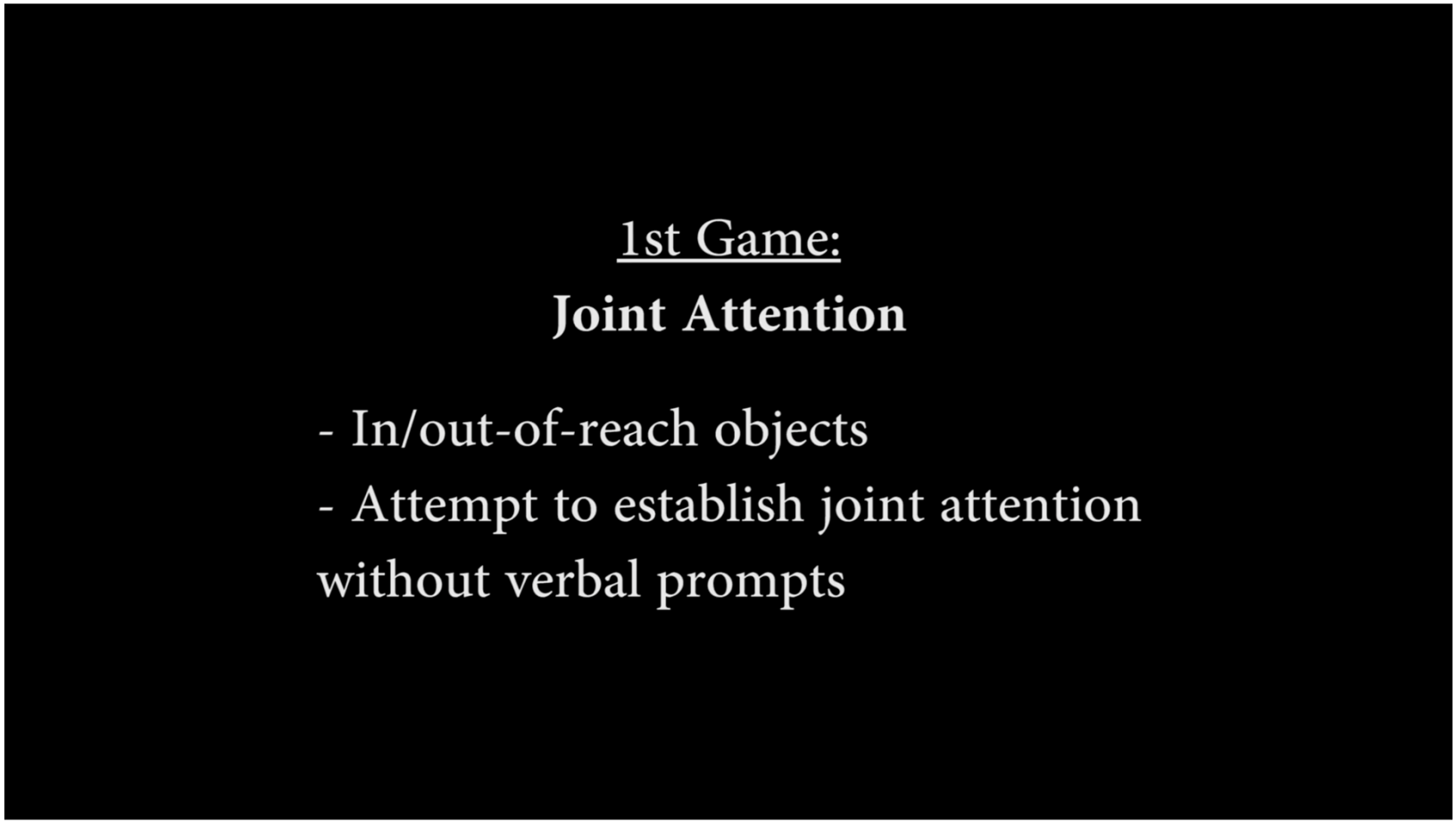
Petros Maragos

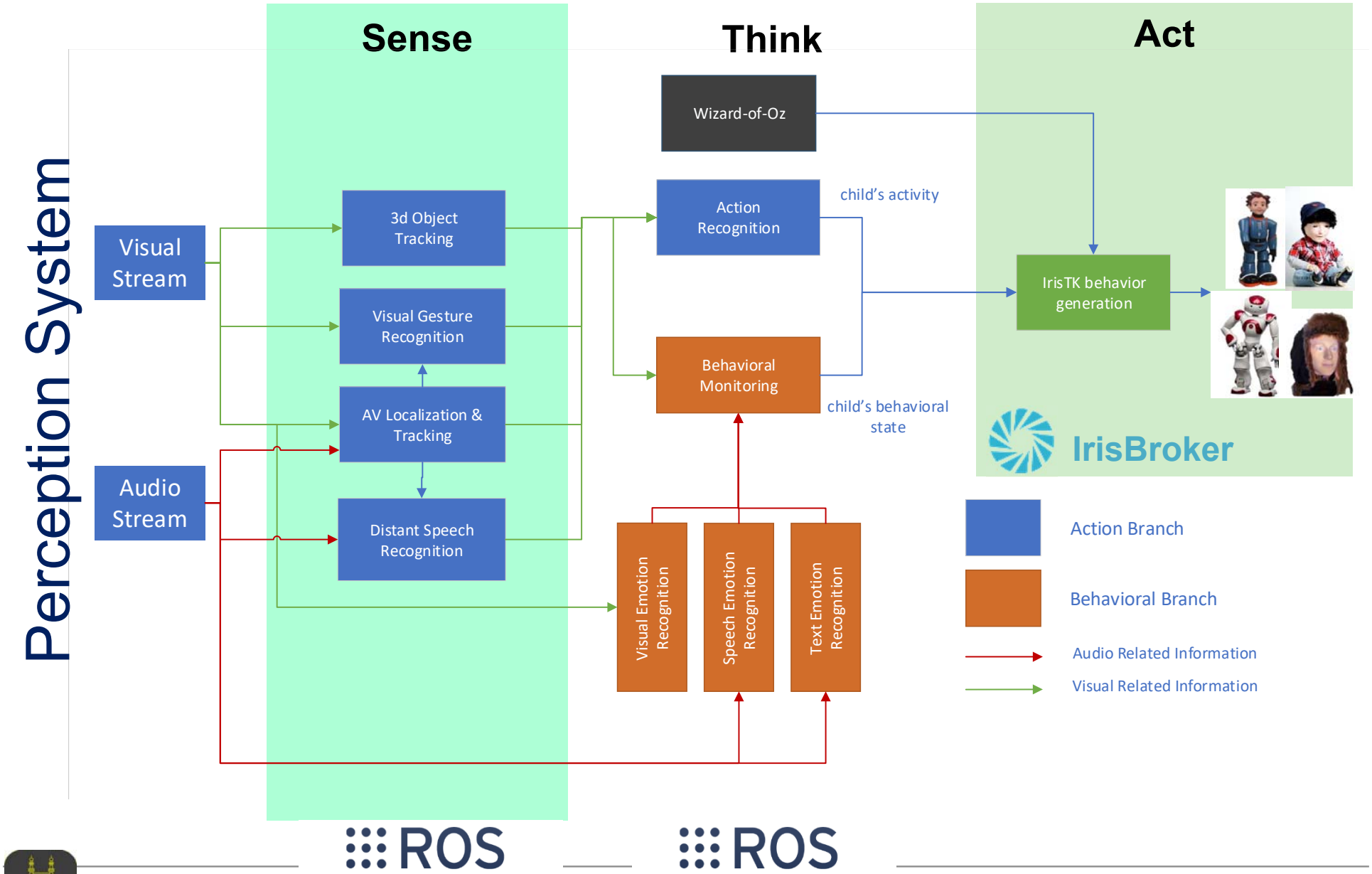
slides: http://cvsp.cs.ntua.gr/interspeech2018

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

EU project BabyRobot: Experimental Setup Room

TD experiments video




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

Experimental Setup: Hardware & Software

Kinect #1

Linux (ROS) – Master

- Manage all writings and synchronization
- Record Kinect #1 streams: **Depth, Color, Audio**
- Runs GSR, ASR, SLOC & Person Detection (+Fusion on GSR, ASR, SLOC)

Kinect #2

Linux (ROS) – Slave

- Record Kinect #2 streams: **Depth, Color, Audio**
- Runs GSR, ASR, SLOC

Kinect #3

Linux (ROS) – Slave

- Record Kinect #3 streams: **Depth, Color, Audio**
- Runs GSR, ASR, SLOC

Kinect #4

Windows (Custom Software) – Slave

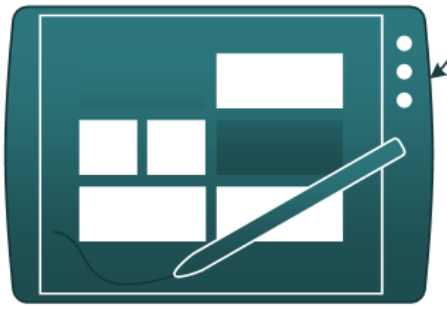
- Record Kinect #4 streams: **Depth, Color, Audio, Body Index, Body (Skeleton), Face, HD Face**
- Runs Touch Screen Games
- Runs IrisTK and Robot Integration Software
- Runs Kinect API recognizers

Zeno

Furhat

Nao

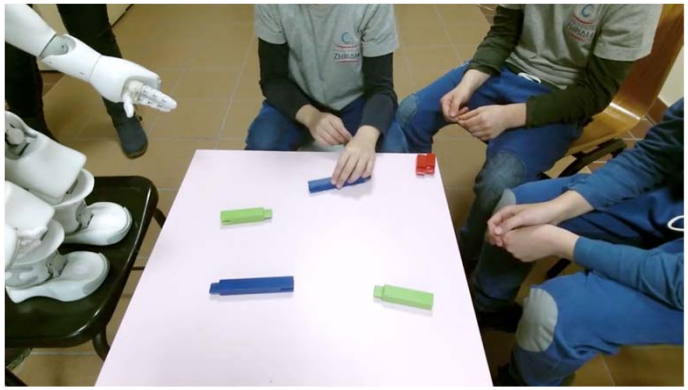
Talking Head



Tablet

Action Branch: Developed Technologies

3D Object Tracking

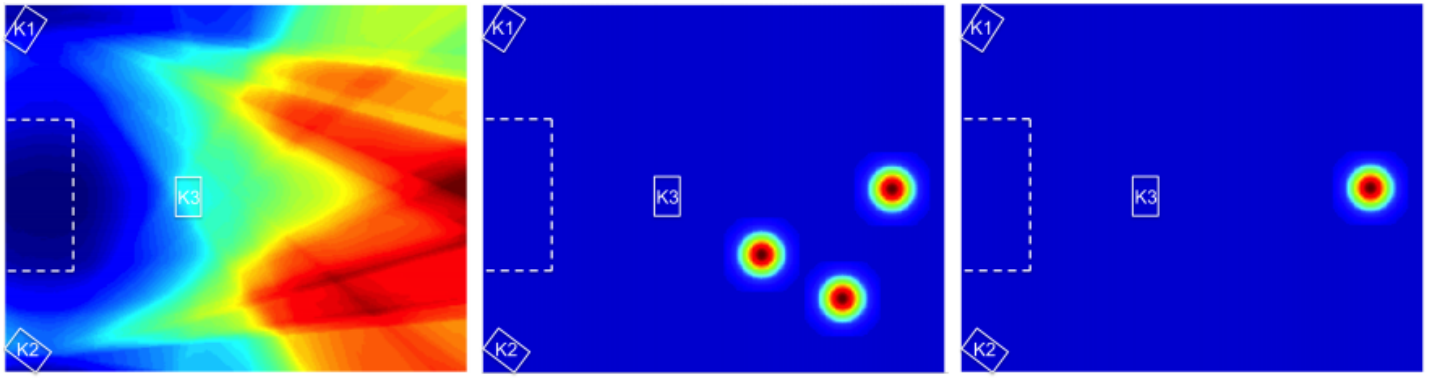


Multiview Gesture Recognition

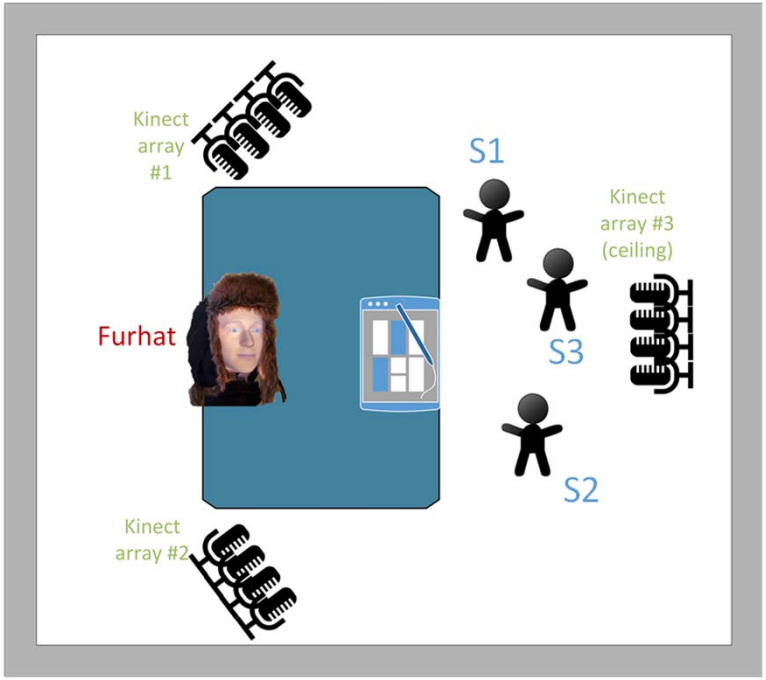
Speaker Localization and Distant Speech Recognition

Multiview Action Recognition

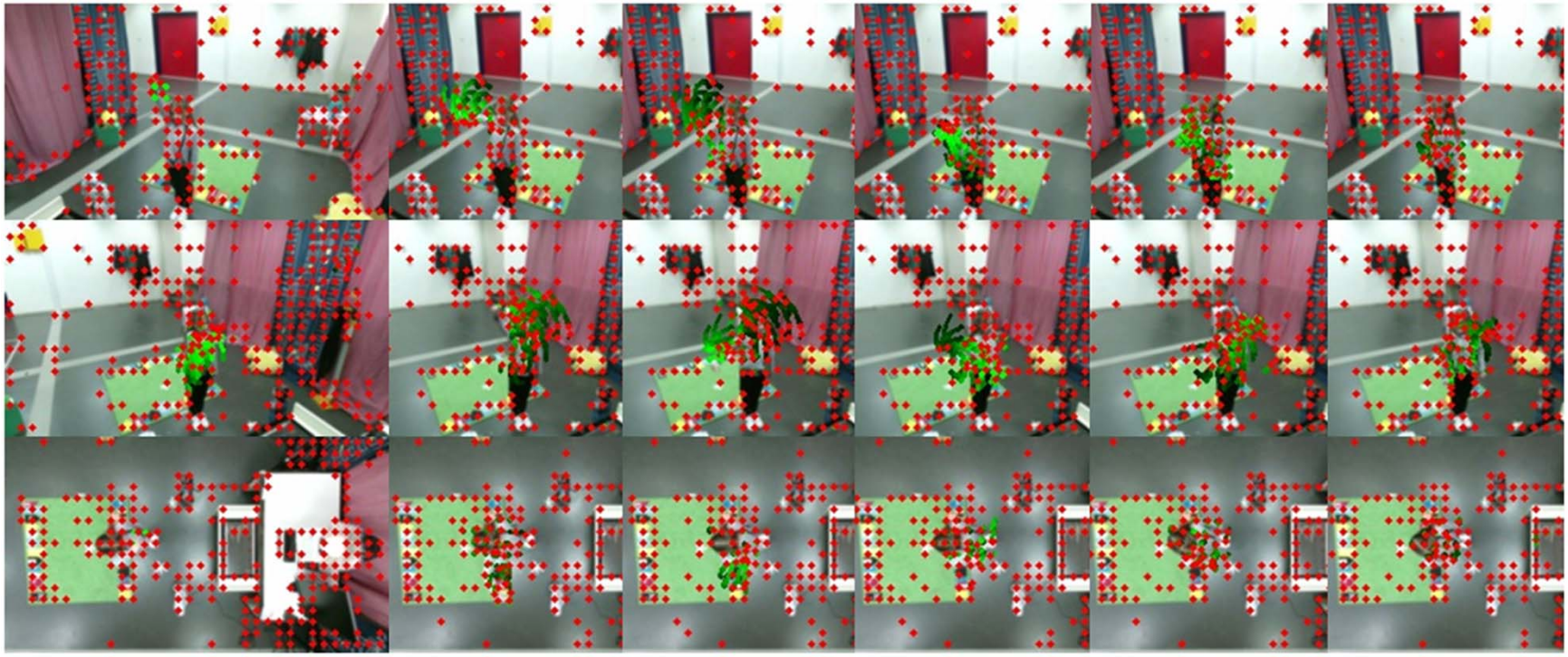
Audio-Visual Localization Evaluation



- 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

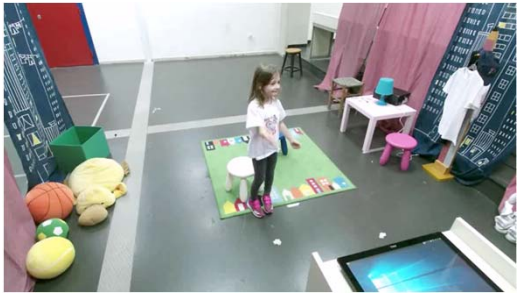


- Multiple views of the child's gesture from different sensors
- Fusion of the three sensors' decisions

Gesture Recognition – Vocabulary

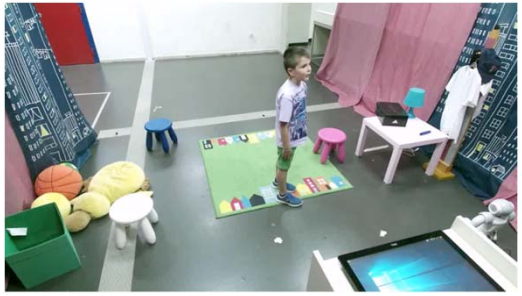
Nod

Greet



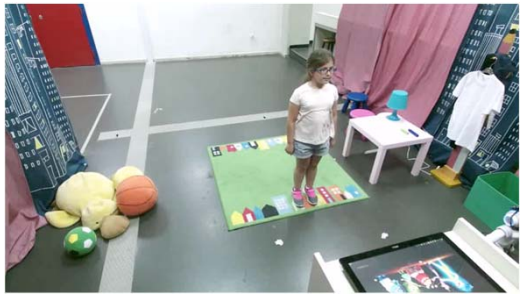
Come Closer

Sit



Stop

Point



Circle

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: 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-steam cases.

Multi-view Gesture Recognition - Children vs. Adults

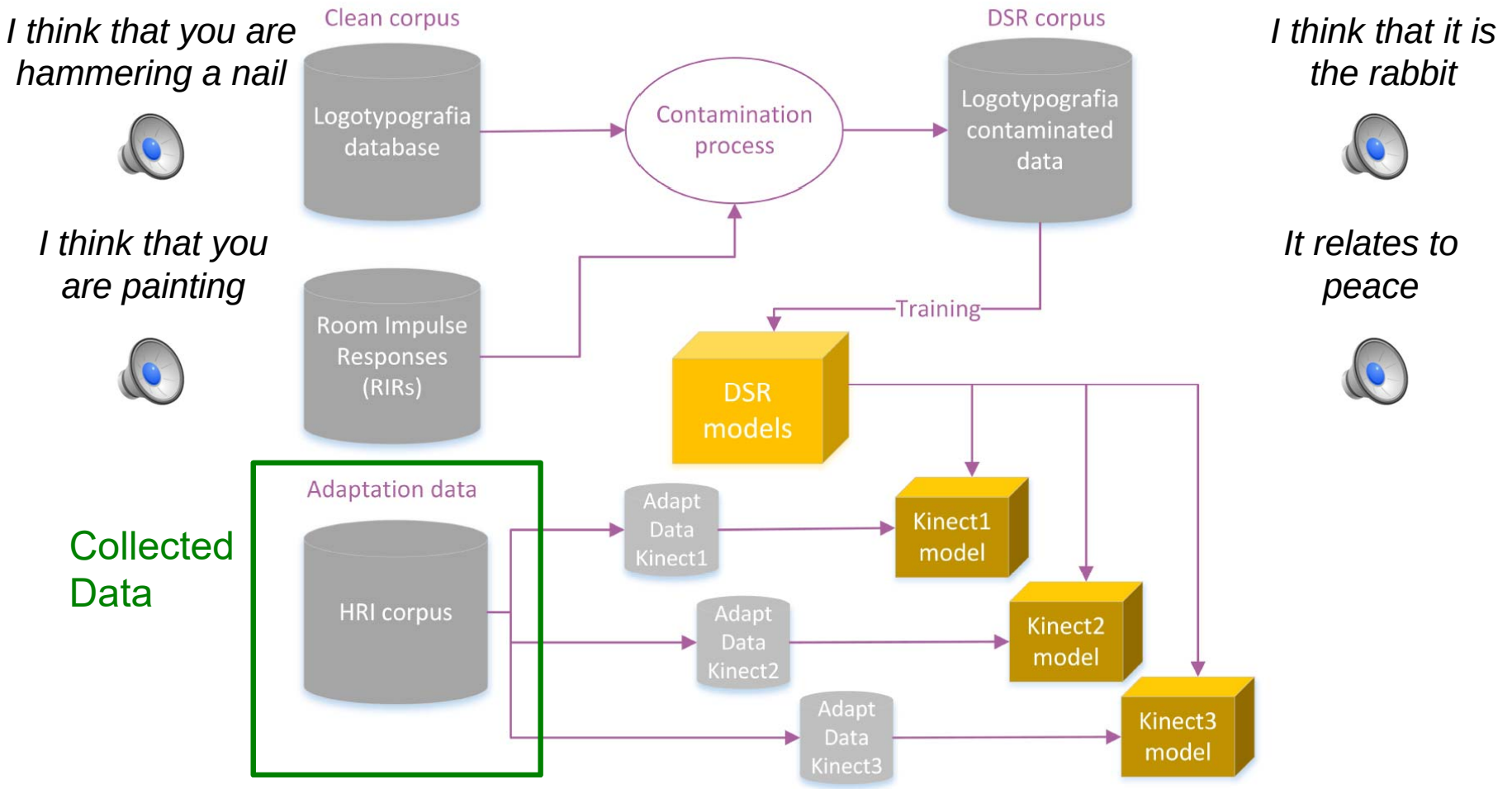
- different training schemes
  - Adults models
  - Children models
  - Mixed model

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", *Proc. ICRA,* 2018.

Distant Speech Recognition System



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

Spoken Command Recognition Evaluation

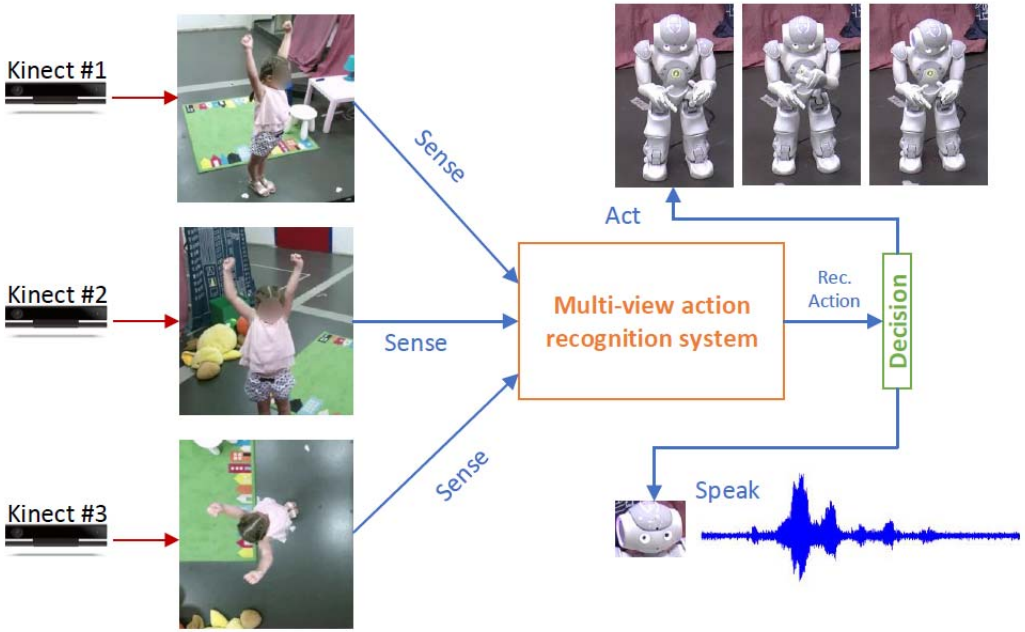
	No-adapt		Adapt-all		Adapt-per-array	
	WCOR	SCOR	WCOR	SCOR	WCOR	SCOR
Kinect #1	79.30	70.53	98.41	95.95	98.30	95.95
Kinect #2	81.04	72.48	97.56	95.95	97.35	95.95
Kinect #3	76.85	66.83	97.45	94.60	97.56	94.60
Fusion	-	64.17	-	**98.97**	-	96.30

- TD (Typically-Developing) children data: 40 phrases
- average word (WCOR) and sentence accuracy (SCOR) for the DSR task, per utterance set for all adaptation choices.
- 4-fold cross-validation

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


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

Action Recognition- Vocabulary

Cleaning a window

Ironing a shirt

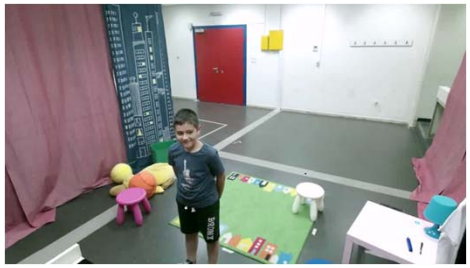
Digging a hole

Driving a bus

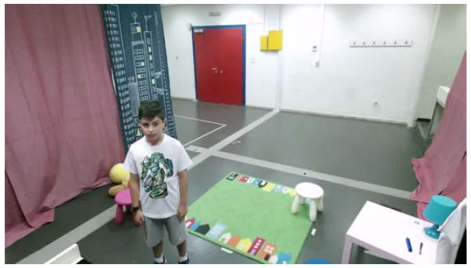
Painting a wall

Hammering a nail

Wiping the floor



Reading



Swimming

Working Out

Playing the guitar

Dancing

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.

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

Multi-view Action Recognition – Children vs Adults

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

		Action Recognition -Training scheme		
		Adults	Children	Mixed
Test		Acc.	Acc.	Acc.
Adults	Kinect #1	79.67	67.58	78.02
	Kinect #2	83.52	62.09	79.12
	Kinect #3	71.98	59.34	78.02
	Avg	78.39	63.00	78.39
	Fuse	**87.36**	72.53	86.26
Children	Kinect #1	53.85	73.85	73.67
	Kinect #2	47.63	63.38	64.20
	Kinect #3	38.18	60.00	59.76
	Avg	46.55	65.74	65.88
	Fuse	56.51	**74.46**	74.26

Employed Features: **MBH**

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. ICRA*, 2018.

Part 3: Conclusions

- **Synopsis**:
  - Data collection and annotation: 28 TD and 15 ASD children (+ 20 adults)
  - Audio-Visual localization and tracking
  - 3D Object tracking
  - Multi-view Gesture and Action recognition
  - Distant Speech recognition
  - Multimodal Emotion recognition

- **Ongoing work**:
  - Evaluate the whole perception system with TD and ASD children
  - Extend and develop methods for engagement and behavioral understanding

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

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