Computer Vision, Speech Communication & Signal Processing Group,
Intelligent Robotics and Automation Laboratory
Institute of Communication and Computer Systems (ICCS)
National Technical University of Athens, Greece (NTUA)

Part 5:
Audio-Visual HRI in Social Robotics for Child-Robot Interaction

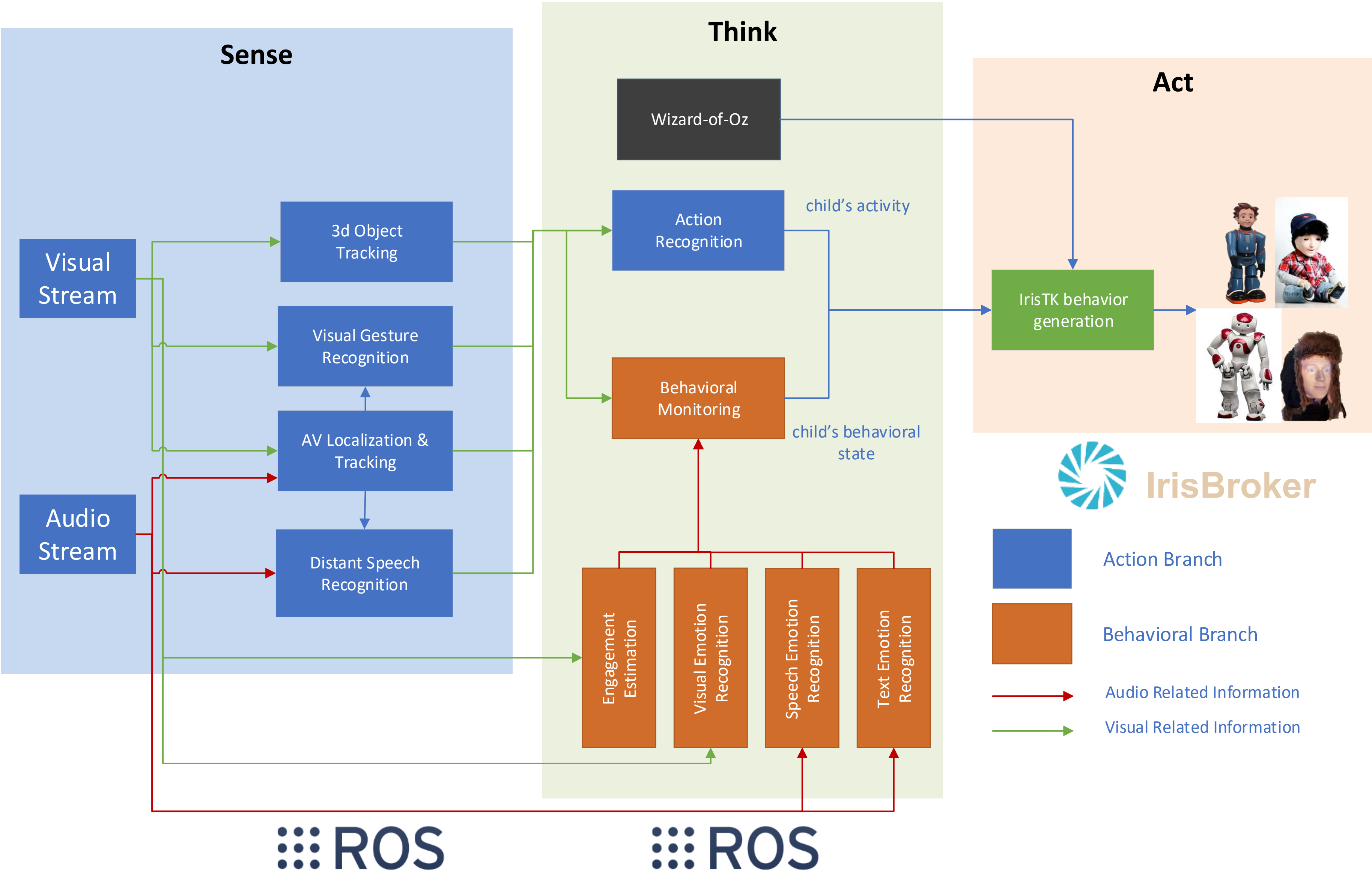
Petros Maragos and Petros Koutras

Tutorial at IEEE International Conference on Image Processing 2019,
Taipei, Taiwan, September 22, 2019

Child-Robot Interaction: Demo

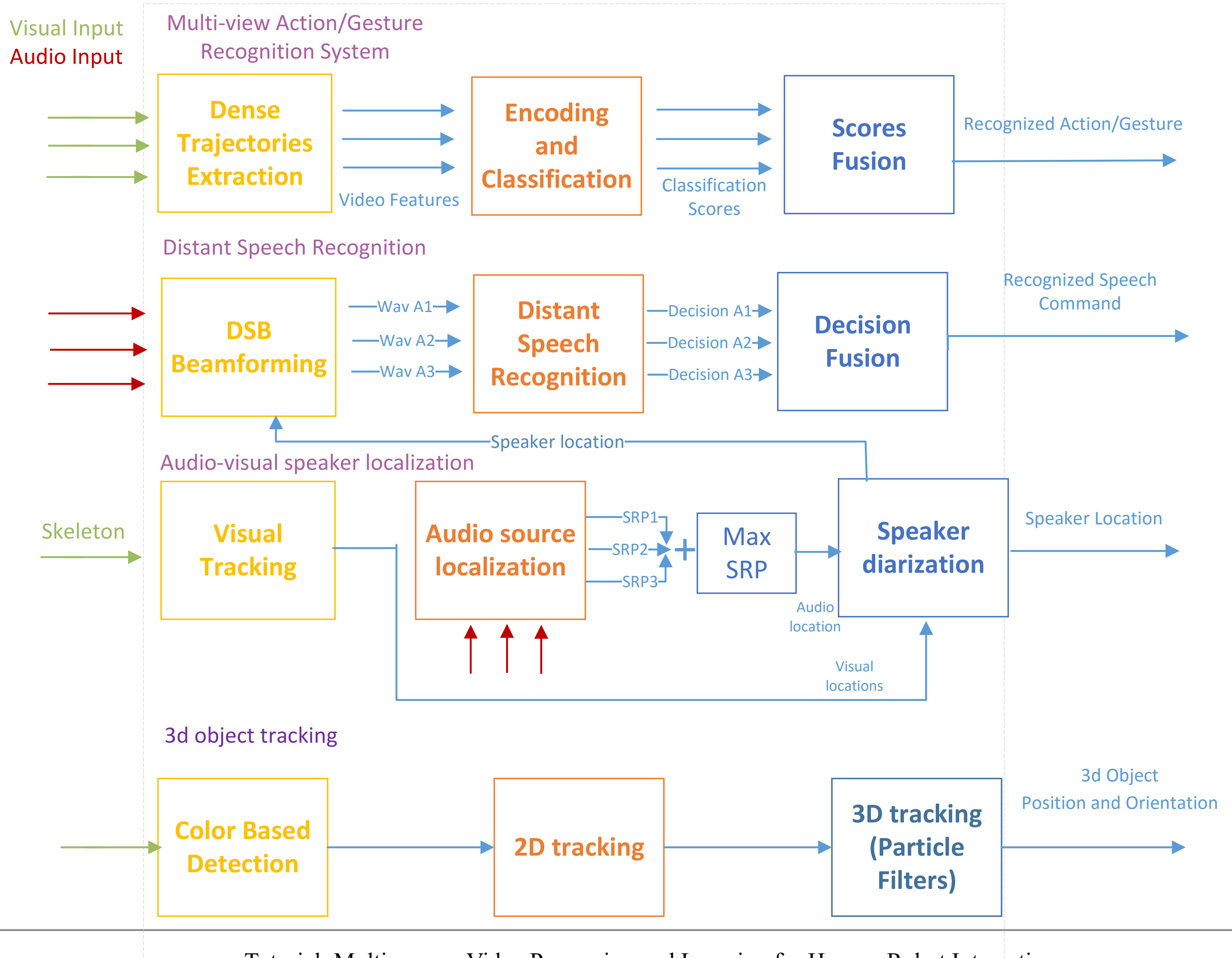
- Develop core audio-visual processing technology to extract low-, mid-, & high-level HRI information from AV sensors.

Perception System



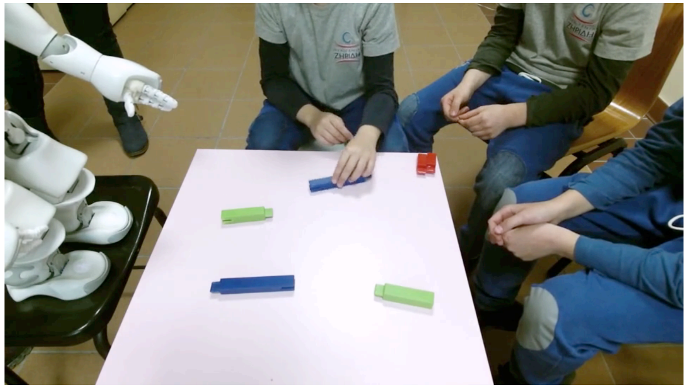
Action Branch: A closer look

- recognize the child's multimodal activities: e.g. speech, movements, gestures

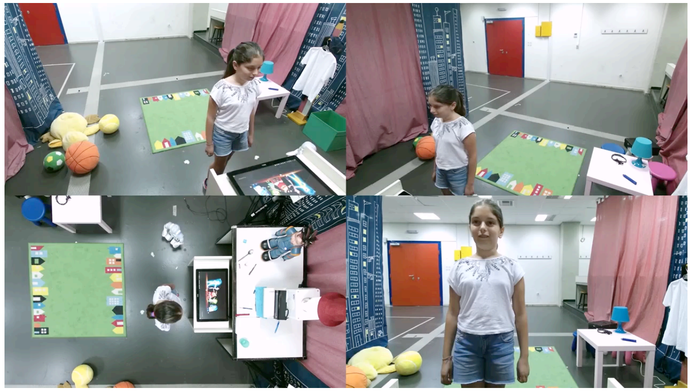


Action Branch: Developed Technologies

3D Object Tracking



Multi-view Gesture Recognition



Speaker Localization and Distant Speech Recognition

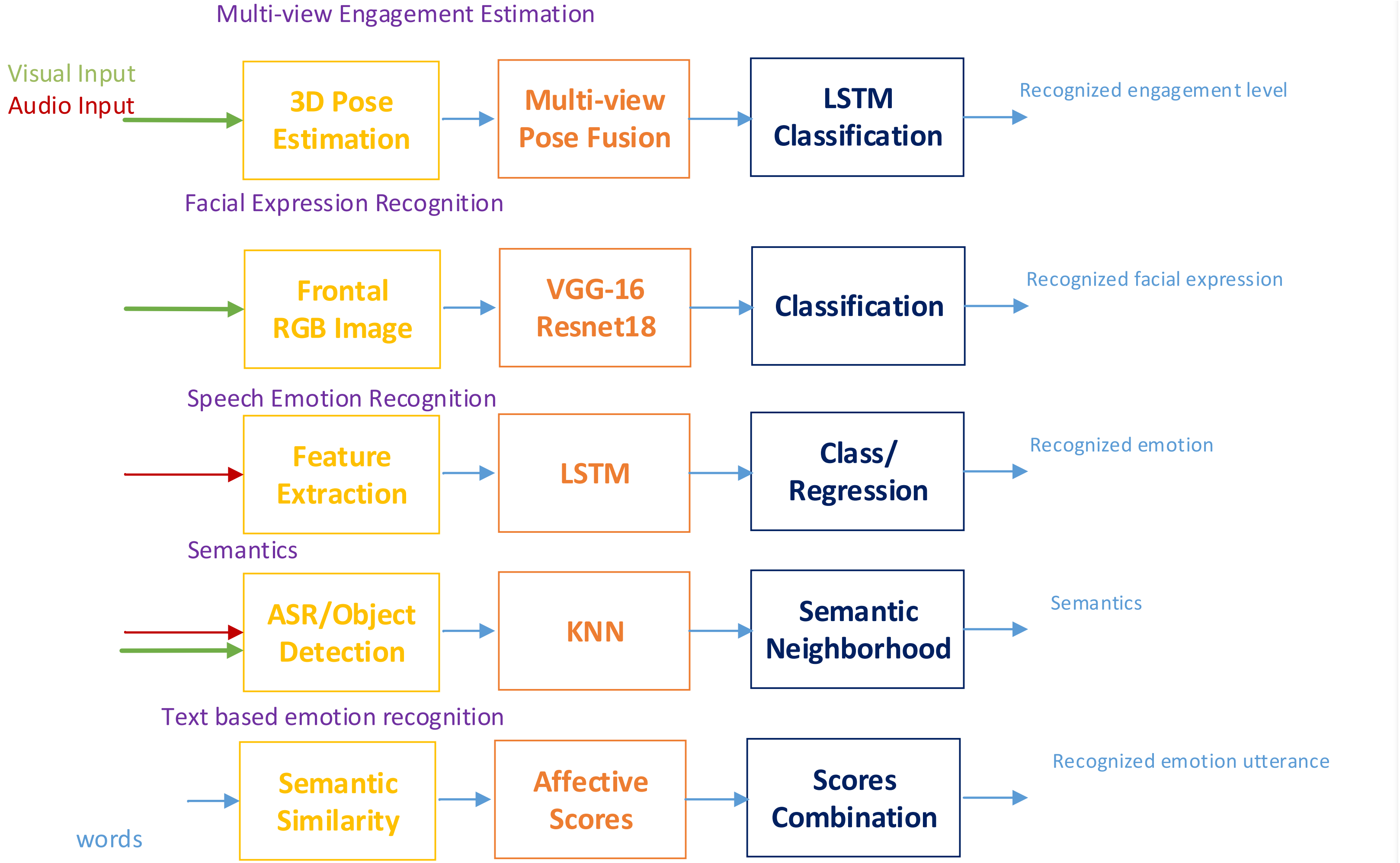
Multi-view Action Recognition

Hadfield et al. In Proc. IROS, 2018
Tsiami et al. In Proc. ICASSP, 2018.

Tsiami et al. In Proc. ICRA, 2018.
Efthymiou et al. In Proc. ICIP, 2018.

Behavioral Branch: A closer look

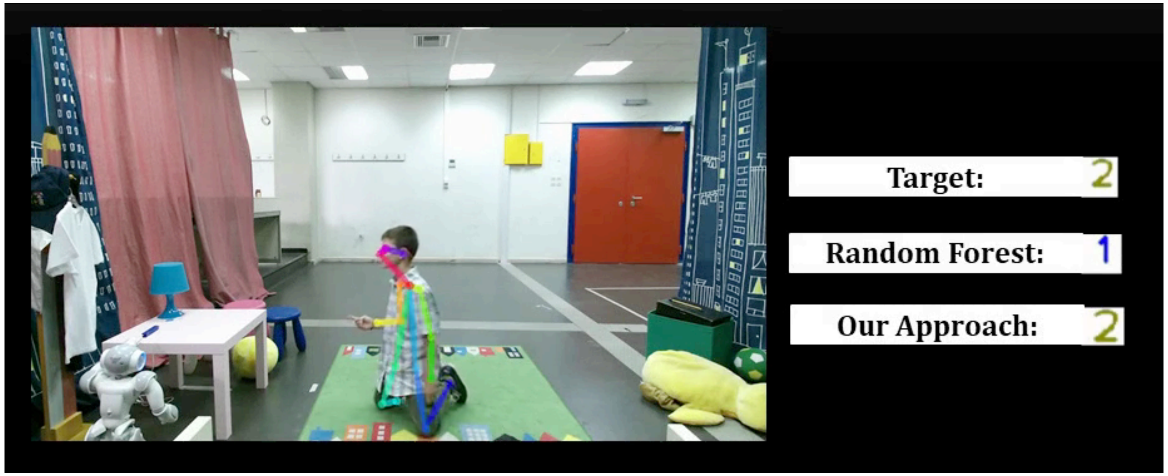
- recognize and identify the behavior of the child that is expressed as facial expressions and emotions, skeleton pose, engagement



Behavioral Branch: Developed Technologies

Engagement Estimation

Hadfield et al. In Proc. IROS, 2019.

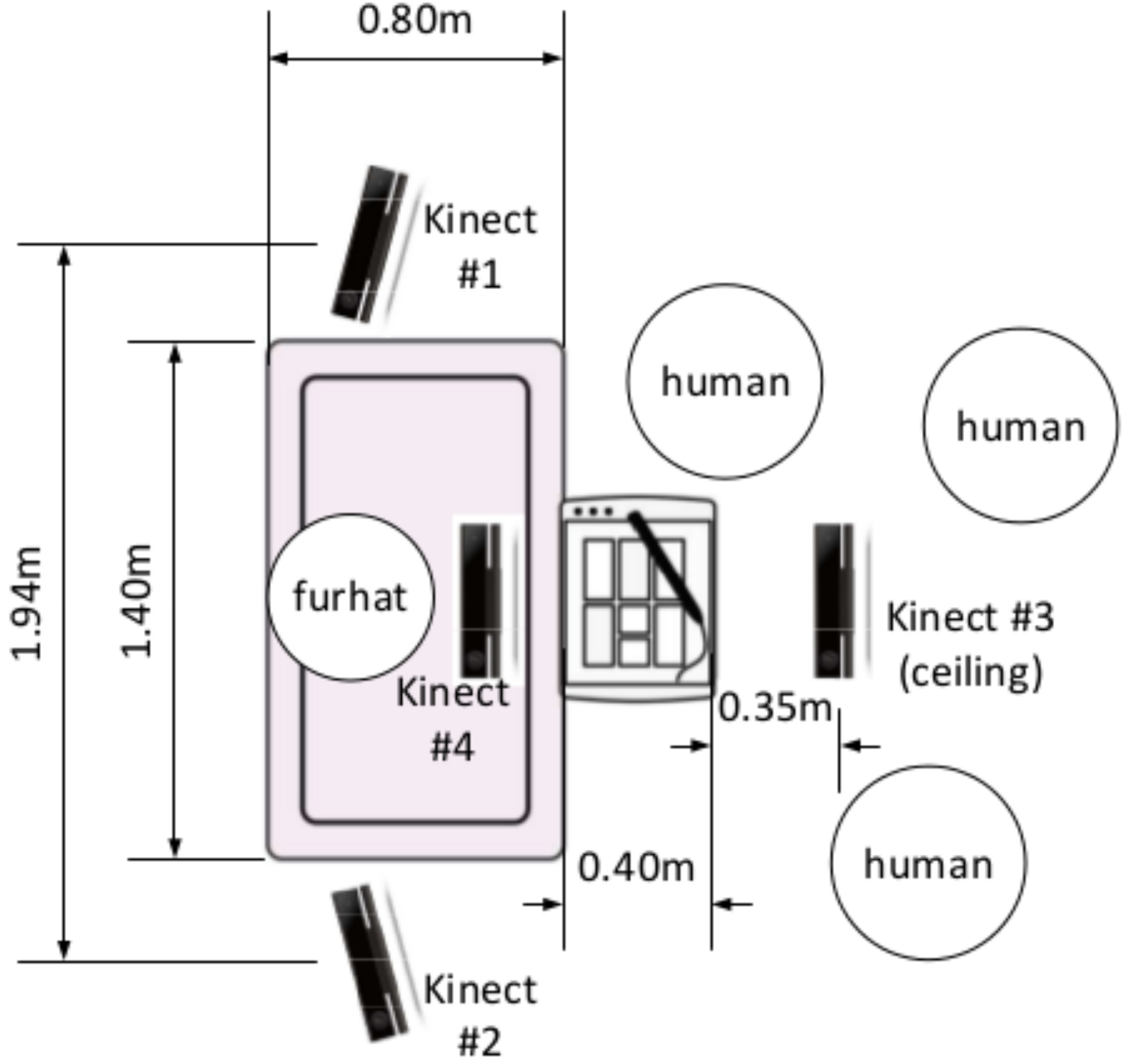


Multi-cue Visual Emotion Recognition

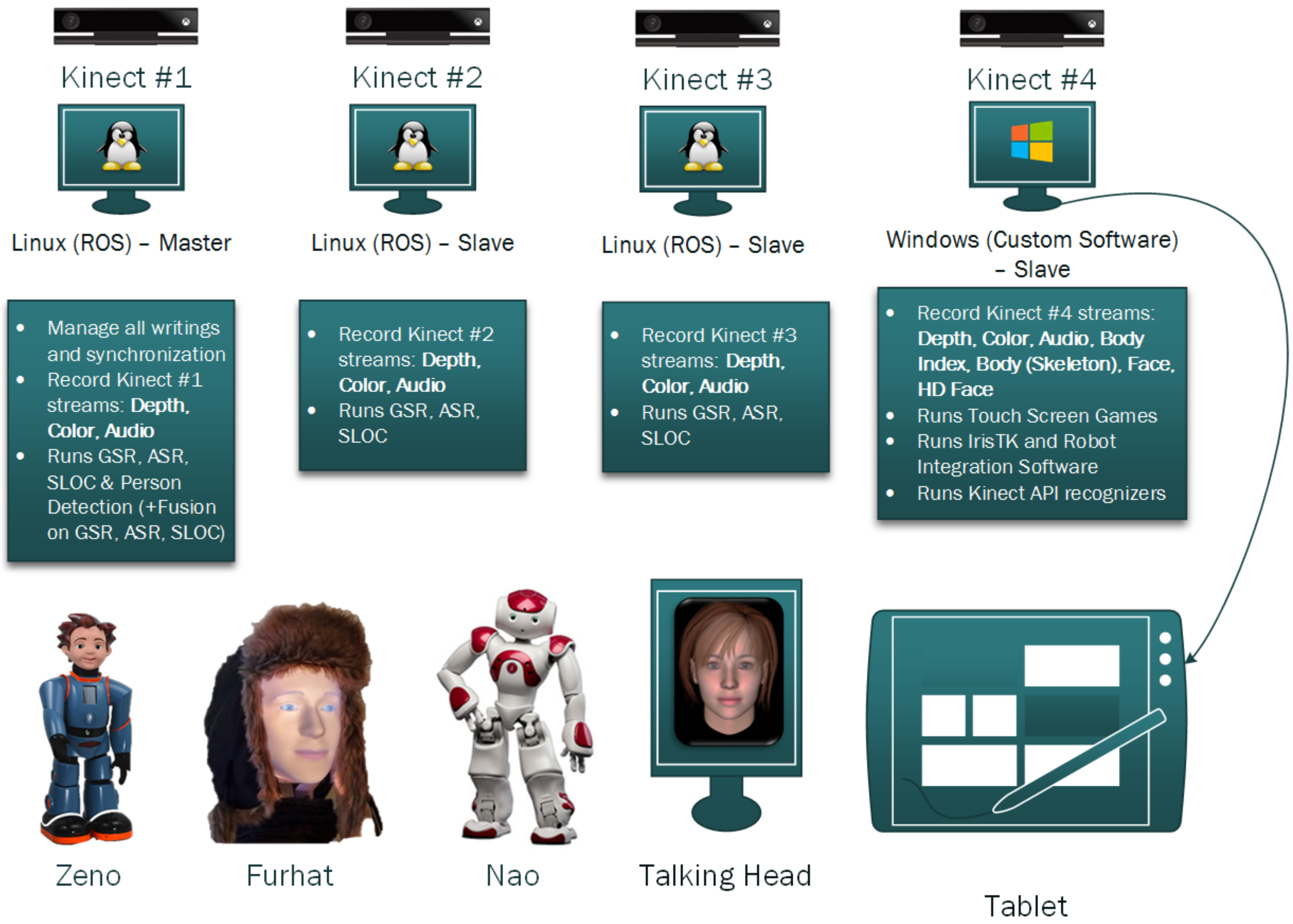
Filntisis et al., IEEE Robotics and Automation Letters, 2019.

Child Robot Interaction: Experimental Setup and Data

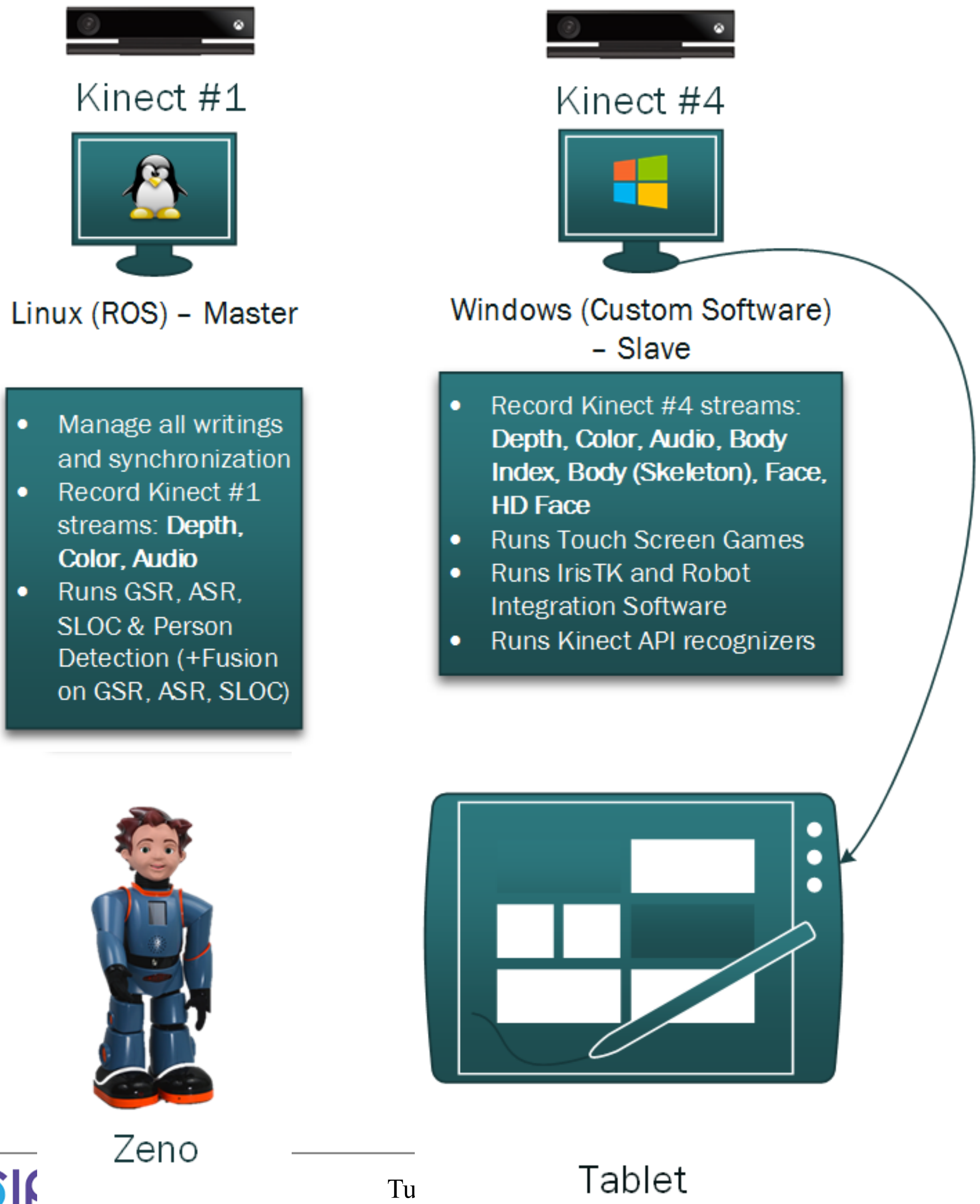
Experimental Setup: Floorplan



Experimental Setup: Hardware & Software



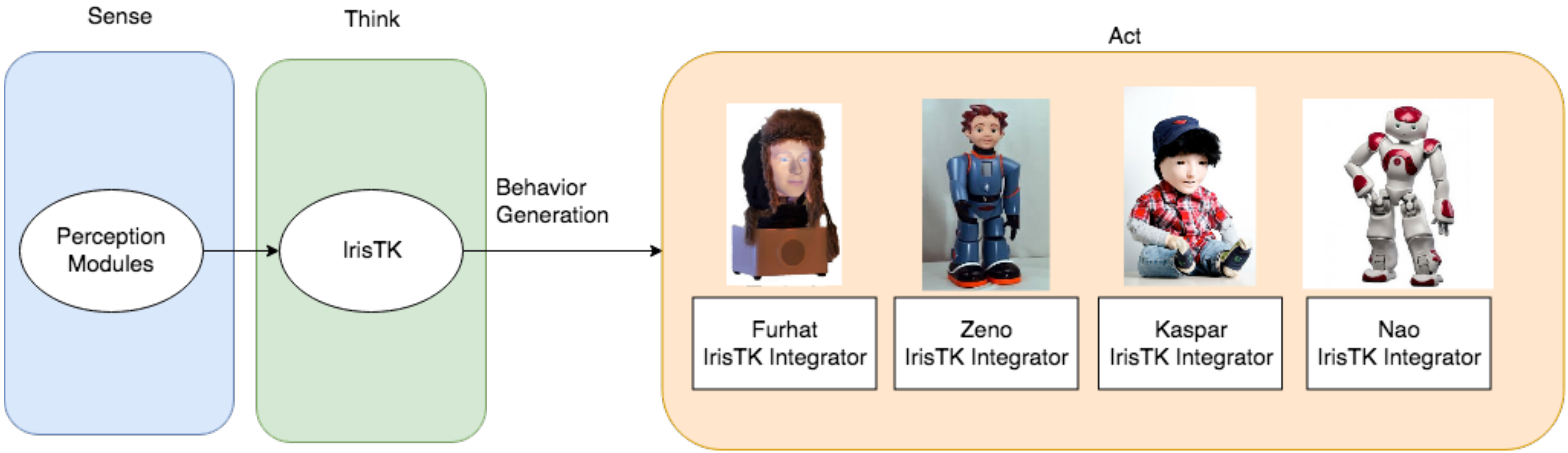
Experimental Setup: Lightweight



- Lightweight setup for school usage
- Designed for the needs of ASD experiments in schools
- Games implemented as tablet applications

Experimental Setup: Multi Robot Extension

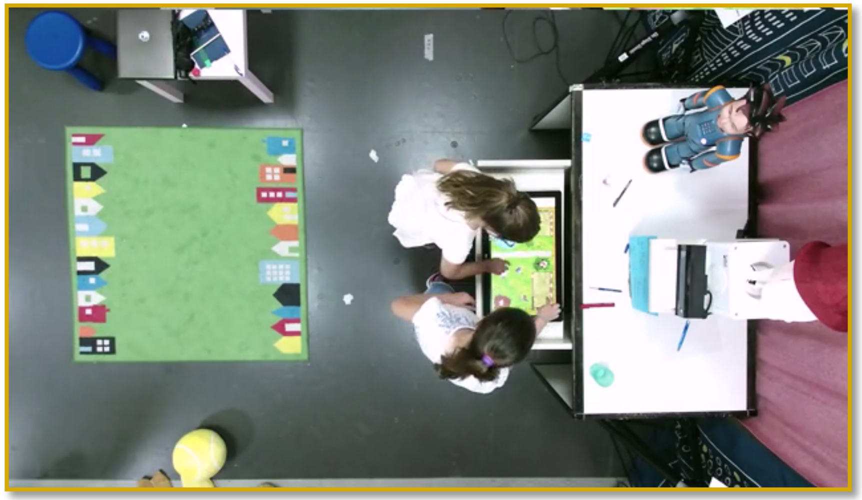
- "Act states":
  - Decouple robot details from core dialog flow
  - Intermediate layer between core dialog and robot
  - The robot is just another parameter
  - Easier to add new robots to the system



Data Collection – Overview

- 20 adults
- 31 TD children 6 – 10 years old (aver. 8 years old)
- 15 ASD children

- 2 types of data
  - Development corpus for Training/Testing algorithms (acted)
  - Experimental corpus related to usecases (spontaneous)



Data Collection – Development Corpus

- Children and adults were asked to do the following sequentially. The robots didn't interfere.
- TD Children data collection
  - 7 gestures
  - 6 emotions
  - 12 pantomimes
  - 40 phrases
- Adults data collection for comparisons
  - 7 gestures
  - 12 pantomimes
  - 100 phrases

Data Collection – Experimental Corpus

Kids interacted with robots by playing the following games. During the experiment, we used the WoZ and we also tested the integrated recognition modules

- Individual games (1 child each time)
  - Joint Attention
  - Introduction
  - "Show me the gesture"
  - Emotion Recognition
  - Pantomime
  - "Guess the object"
- Co-operative games (2 children)
  - "Form a Farm"
  - "Rock-Paper-Scissors"

	Distant Speech Recognition	Detect & Track	Speaker Localization	Visual Activity Recognition	
				Gesture	Action
Show me the Gesture	✓		✓	✓	
Pantomime	✓		✓		✓
Assembly Game		✓	✓		
Form a Farm	✓		✓	✓	

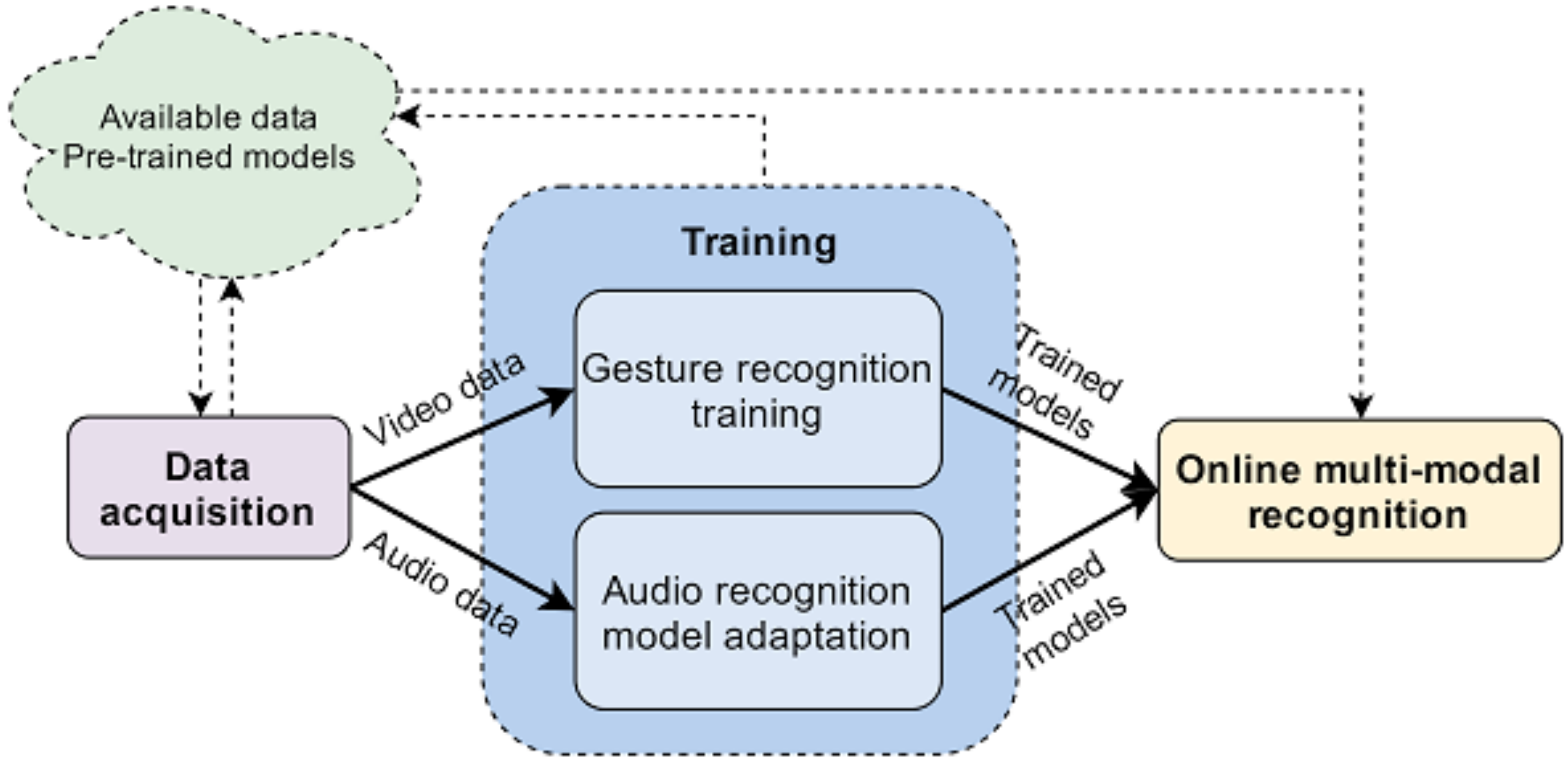
Experimental Corpus: Annotation

- Increasing the benefits of the collected data
- Provide ground truth for experimental corpus
  - Low level annotation for action branch technologies (Gesture, DSR, Action)
  - High level annotation for child behavior monitoring (emotion recognition, cognitive state, engagement)

Collected Data	Events' Type	# of Events
Development Data	Utterances	1120
	Gestures	196
	Pantomimes	336
Use-case Related Data	Utterances	630
	Gestures	143
	Pantomimes	109

CRI Modules: Building New Models

- Usages of development corpus
  - Training new models for gesture and action recogntion
  - Adapt pretrained models for speech recognition
  - Evaluate the developed modules



Gesture, DSR, Action: Children vs. Adult Models

different training schemes

- Adults models
- Children models
- Mixed model

need for children specific models

		Gesture Recognition Training scheme		
Test		Adults	Children	Mixed
Adults	Avg	86.49	56.32	87.36
	Fuse	92.19	62.08	**95.10**
Children	Avg	49.92	70.99	72.30
	Fuse	56.25	**83.80**	80.09

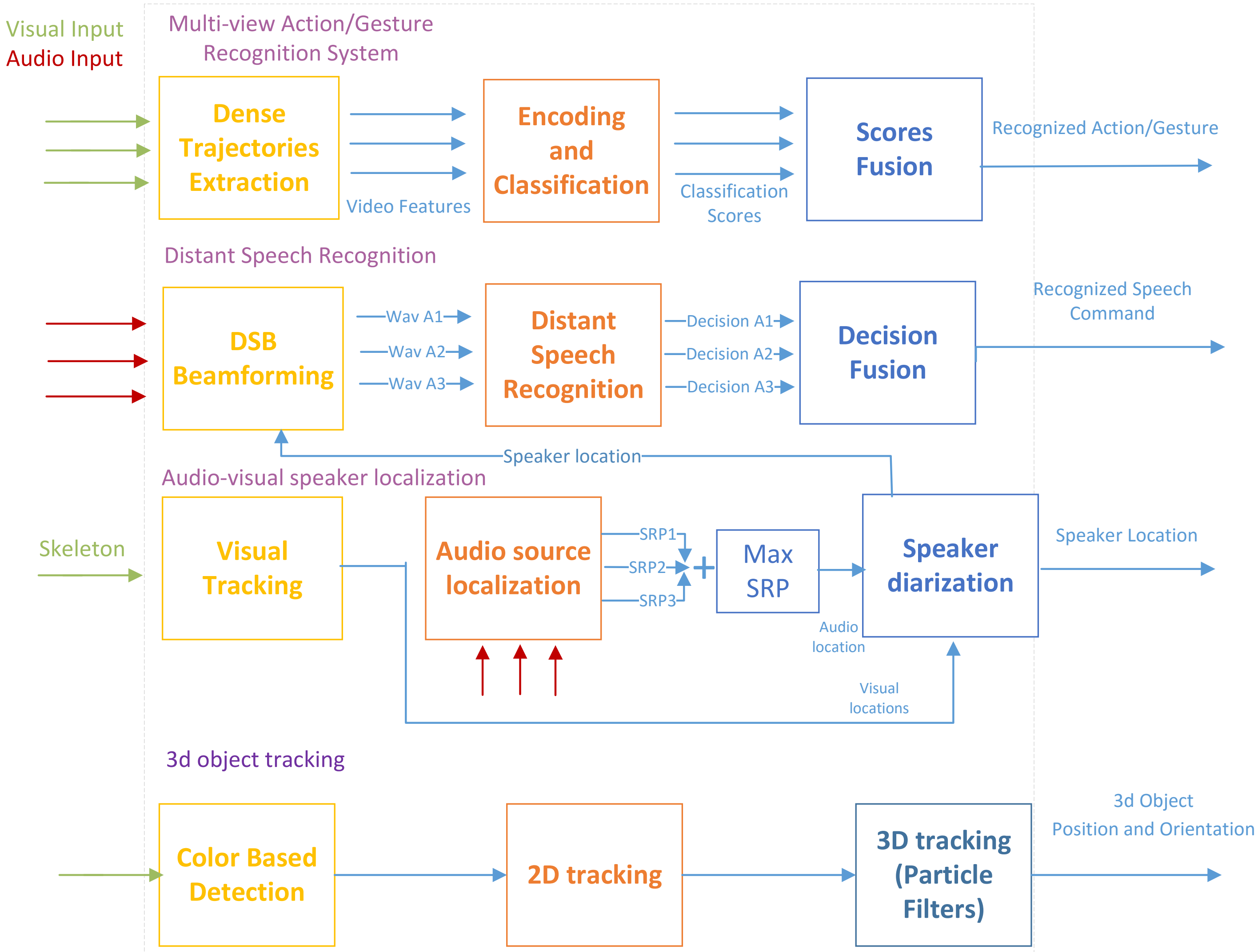
		Action Recognition		
		Training scheme		
Test		Adults	Children	Mixed
Adults	Avg	78.39	63.00	78.39
	Fuse	**87.36**	72.53	86.26
Children	Avg	46.55	65.74	65.88
	Fuse	56.51	**74.46**	74.26

	DSR-Adaptation scheme							
	No-adapt		Adults		Children		Mixed	
Test	WCOR	SCOR	WCOR	SCOR	WCOR	SCOR	WCOR	SCOR
Adults	97.54	91.25	99.58	**98.87**	96.73	93.20	99.50	98.43
Children	79.06	69.95	75.31	71.20	97.81	**95.50**	90.71	82.06

Child Robot Interaction: Perception Modules Details

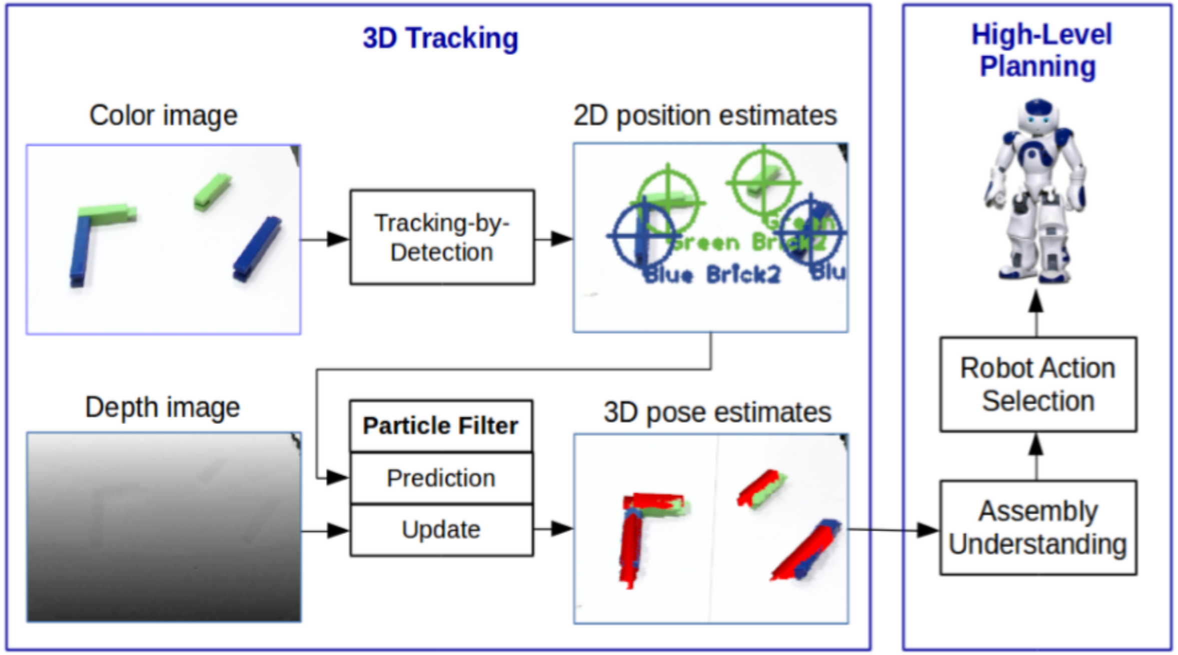
Action Branch: Modules

- recognize the child's multimodal activities: e.g. speech, movements, gestures



Object Detection & Tracking (I)

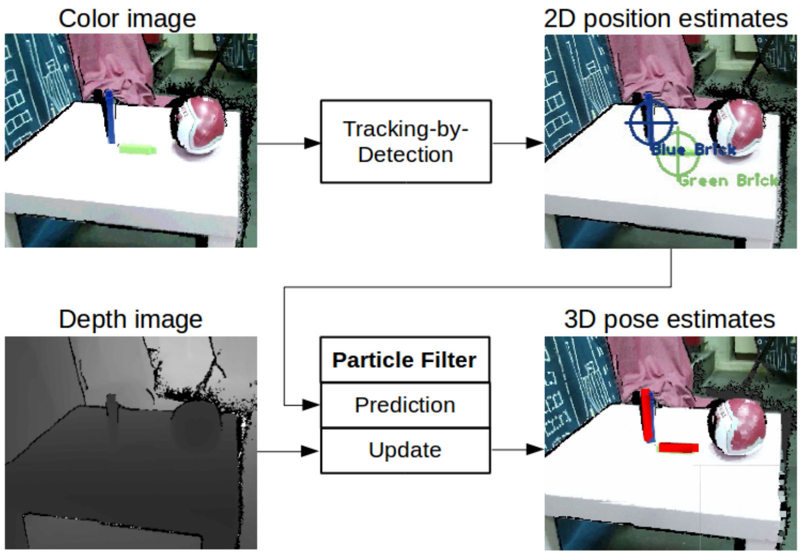
- 6-DoF tracking of multiple objects.
- Used to track bricks in assembly experiments.
- The resulting poses are used to understand the assembly state.



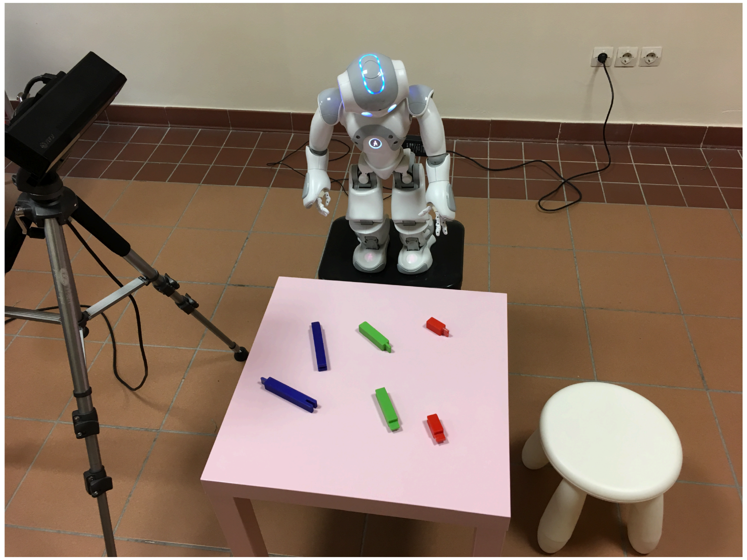
Hadfield et al. In Proc. IROS, 2018

Object Detection & Tracking (II)

- Detection using color histogram models.
- 2D tracking: Match detected regions to objects (Hungarian algorithm).
- 3D tracking: Particle Filter.
  - 2D estimates treated as process input, depth values as measurements.
  - Pixel-wise occlusions are also modeled.
  - Object intersections penalized in update phase.



Object Assembly: School Experiments



Rect. 1

Rect. 2

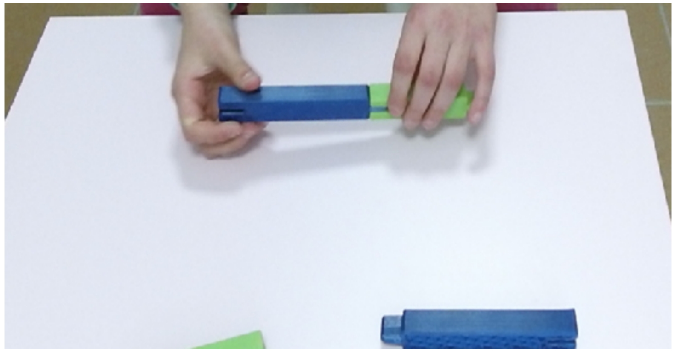
Square

- Children are responsible for the manipulation of the assembly's subcomponents, while the robot provides instructions and feedback
- Set up was placed in a Greek primary school
- 21 children (aged 9-10)
- 6 played on their own
  15 played in group of fives

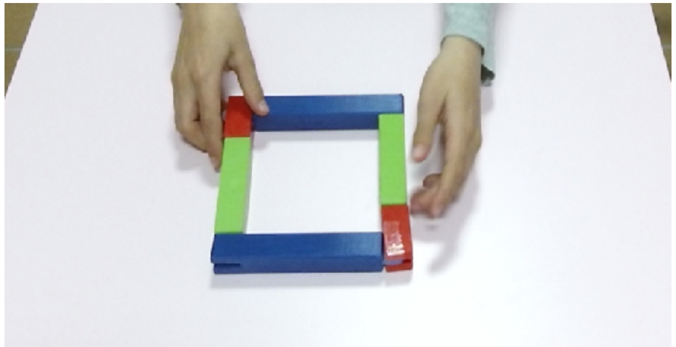
	Correct Connections		Total Connections			
	5s	20s	5s	20s	Avg mist/trial	Time
Rectangle	50.0	56.25	70.00	80.00	0.86	48.49
Square	43.24	59.46	39.39	57.58	0.67	110.09

Object Assembly: Snapshots

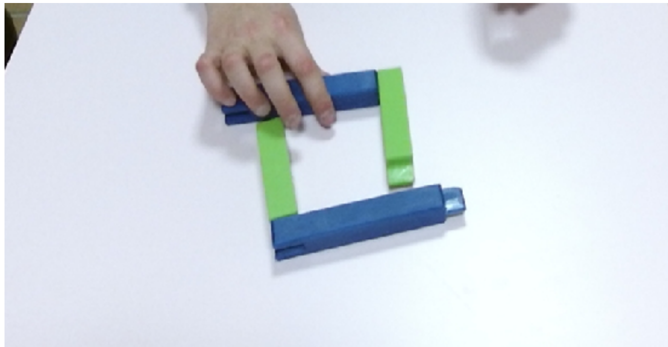
(a) correct connection successfully recognized



(b) incorrect action successfully recognized

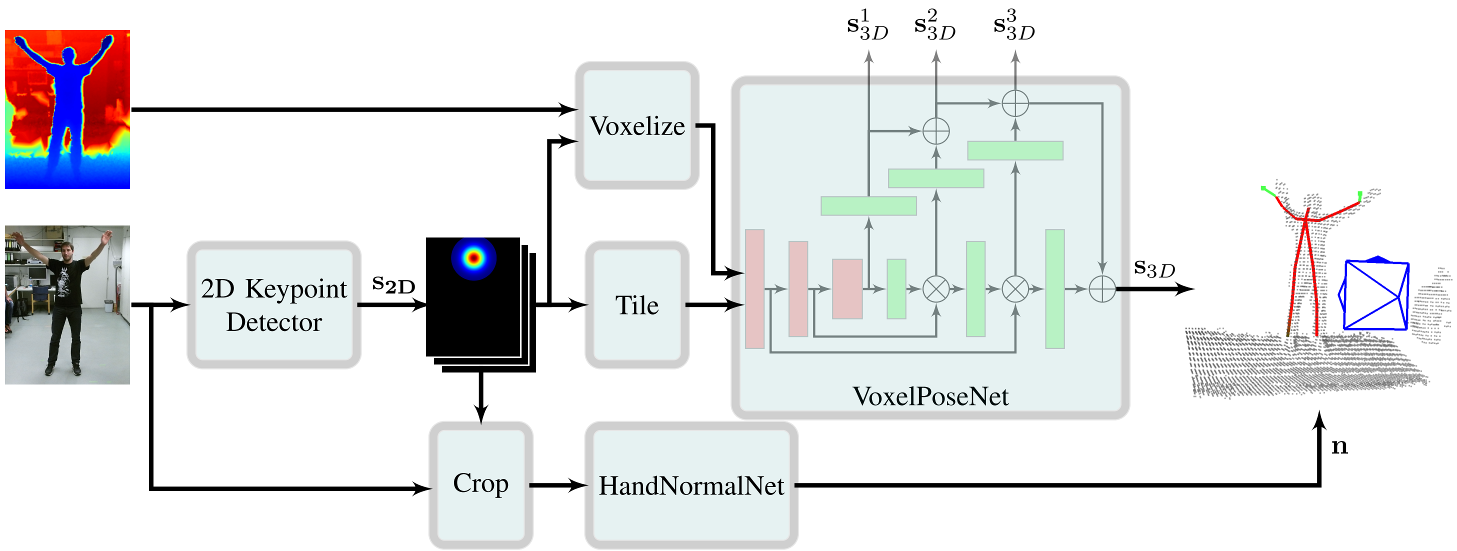


(c) correct action not recognized



(d) false alarm

3D Human Detection and Tracking

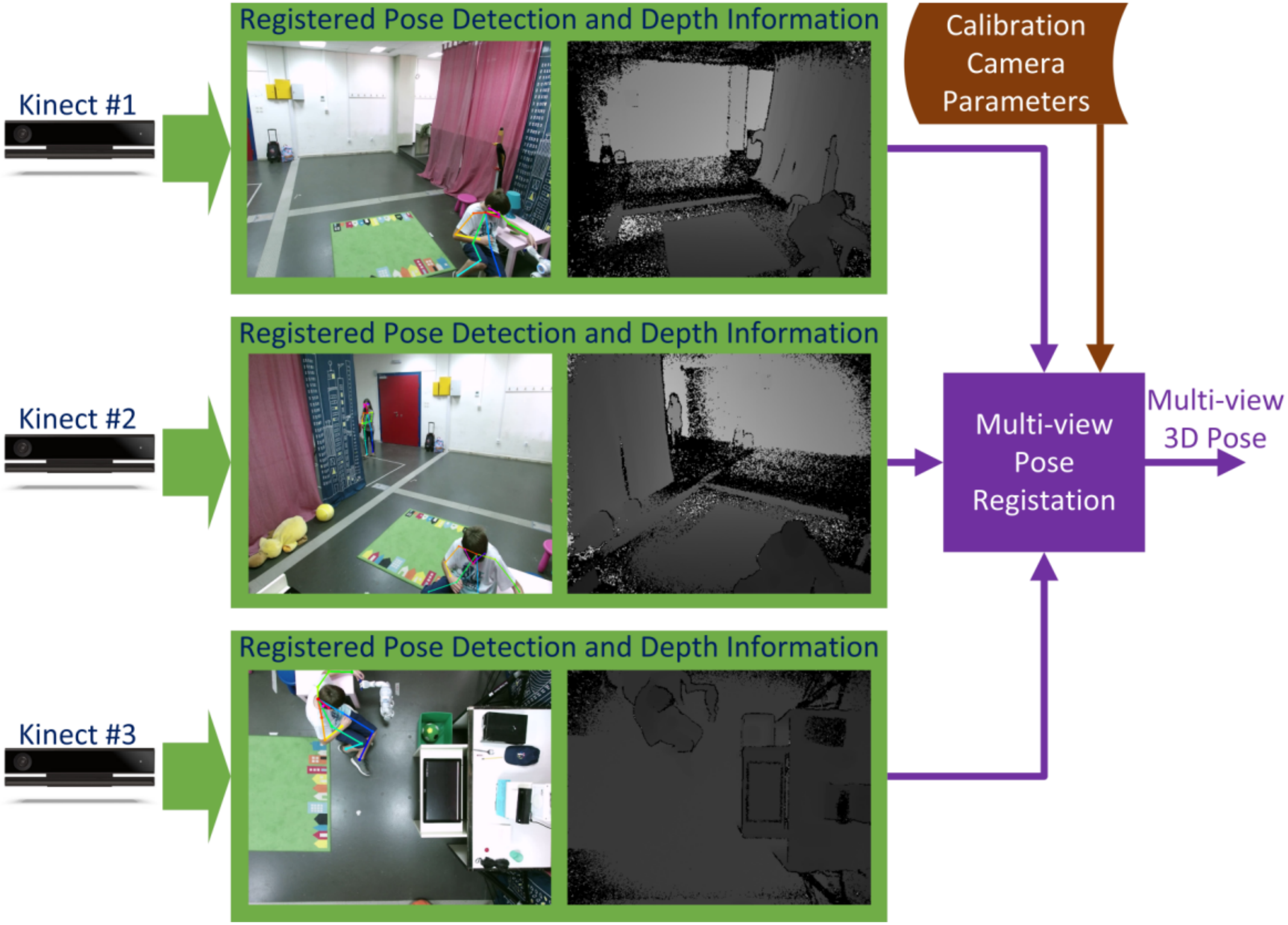


- Detect human joints in 3D space
  - Detect 2D body keypoints → OpenPose library
  - Employ depth stream → 3D pose estimation
- Transform points to global frame (detected using principal normals in point cloud)

Zimmermann et al. In Proc. ICRA, 2018.

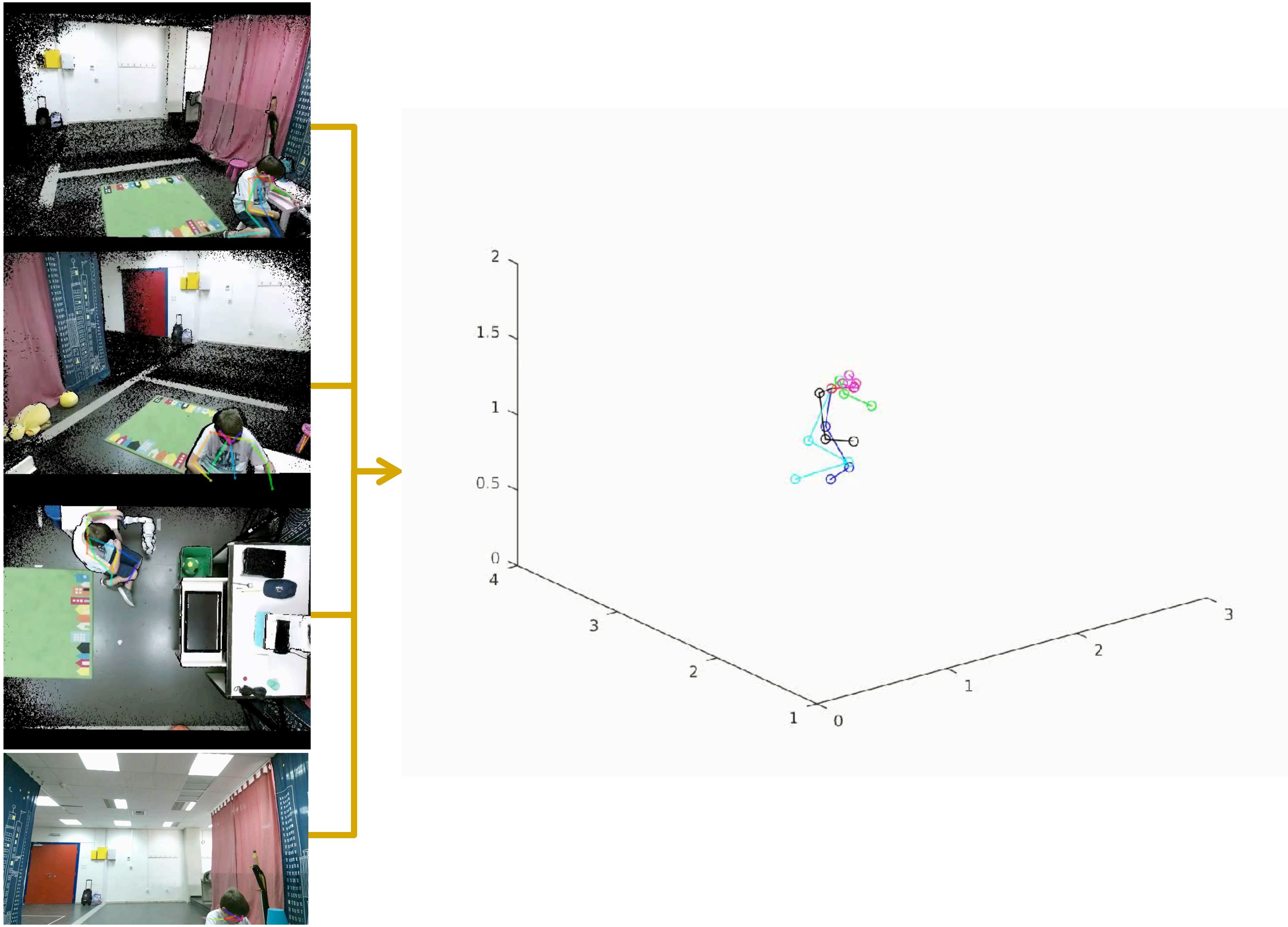
Child’s pose detection from each view

3D Child's Pose: fusion of detected poses

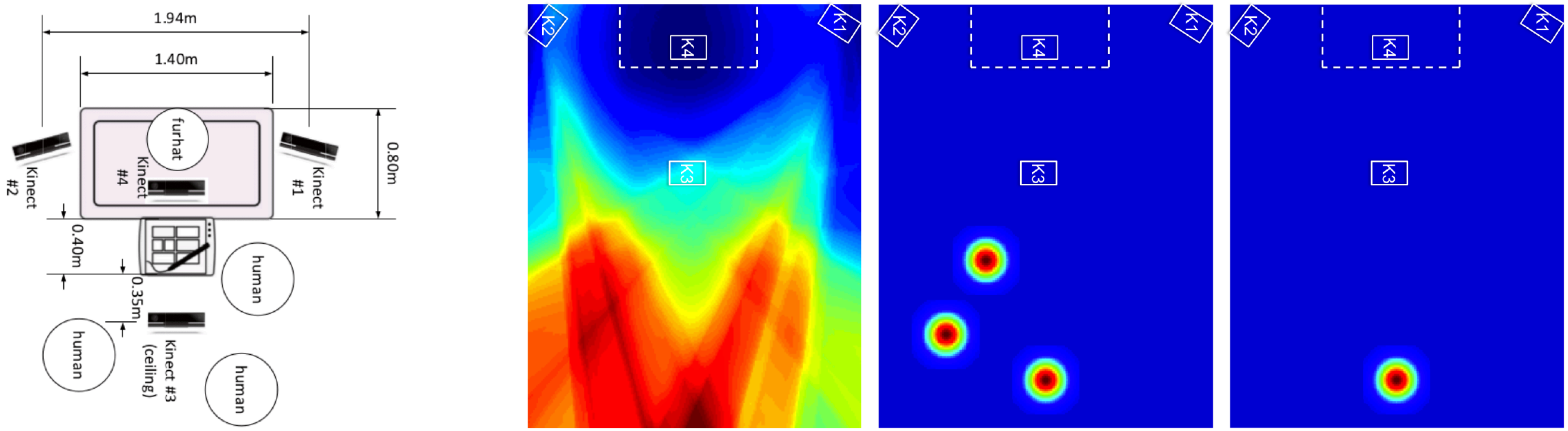


Hadfield et al. In Proc. IROS, 2019

3D Child's Pose: fusion results



Audio-Visual Active Speaker Localization

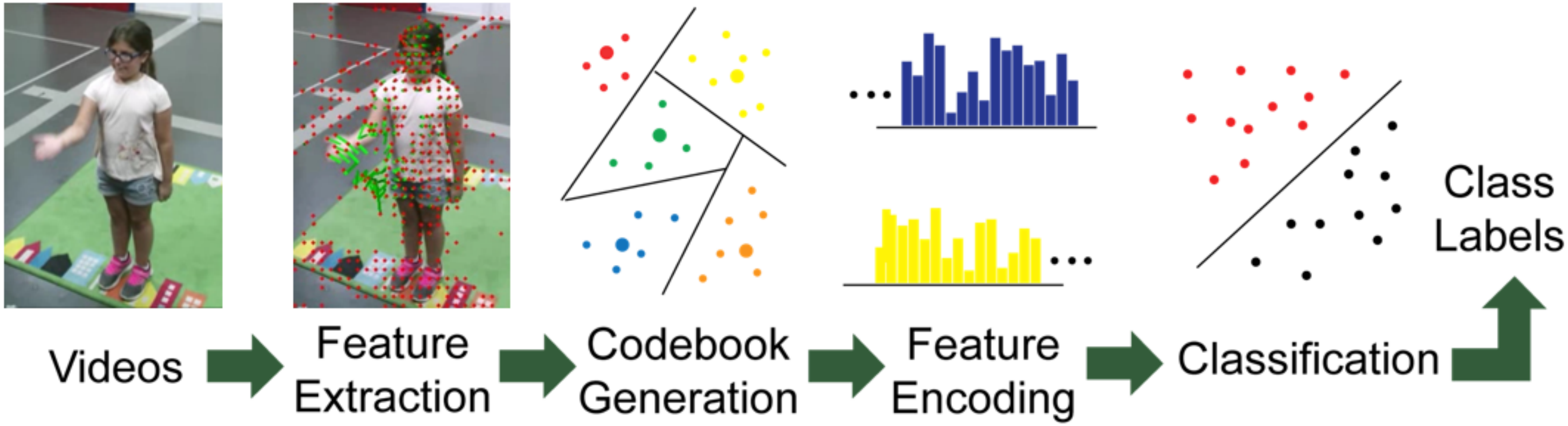


(a) Experimental setup (b) Audio-only (c) Visual-only (d) Audio-visual

- person tracking using 3D skeleton
- choosing the person closest to the auditory source position
- Rcor: percentage of correct estimations (deviation from ground truth less than 0.5m)
  - Audio Source Localization: 45.51%
  - Audio-Visual Localization: 85.58%

Tsiami et al. In Proc. ICASSP, 2018.

Visual Gesture Recognition System



- Gesture recognition system
  - Using only RGB (no depth) → can employ every camera
  - Trained using data from adults or children
- Based on state-of-the-art framework for action recognition
  - Dense trajectories from Optical Flow

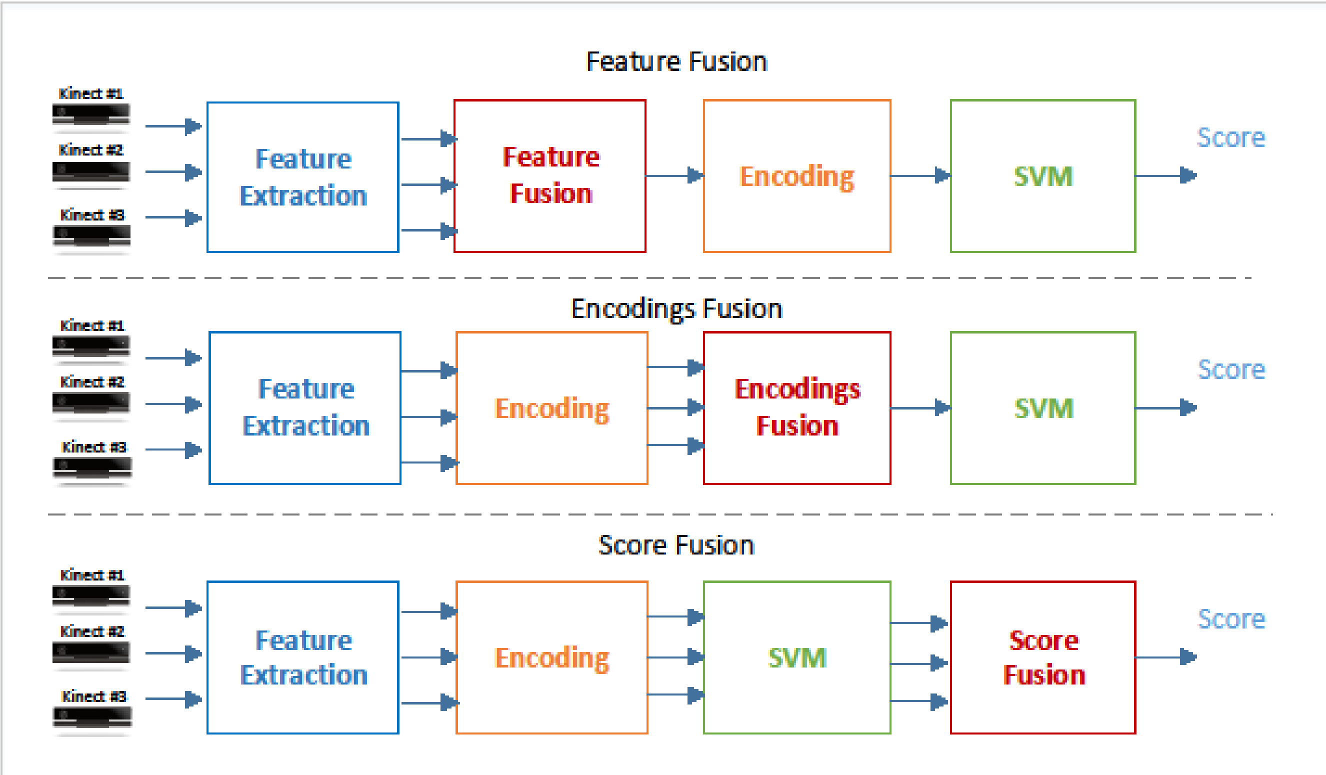
Tsiami et al. In Proc. ICRA, 2018.

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 fusion schemes

Multi-view Fusion for Gesture Recognition

- Feature Fusion: Early fusion of low-level descriptors
- Encodings Fusion: Middle fusion of encodings
- Score Fusion: Late fusion deploying the resulted probabilities for the recognition, from each sensor

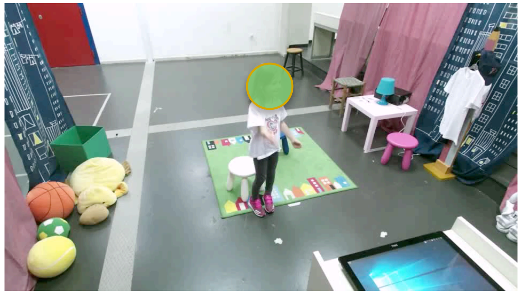


Efthymiou et al. In Proc. ICIP, 2018.

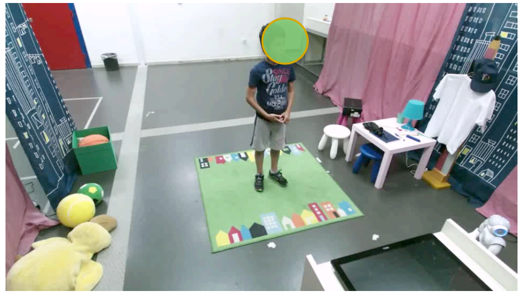
Gesture Recognition – Vocabulary

Nod

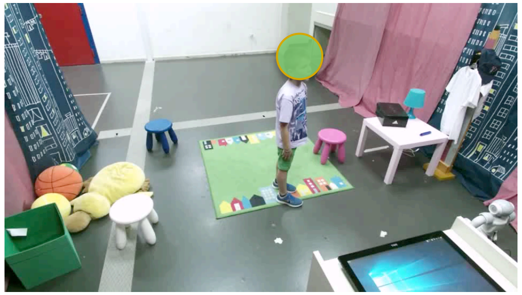
Greet



Come Closer

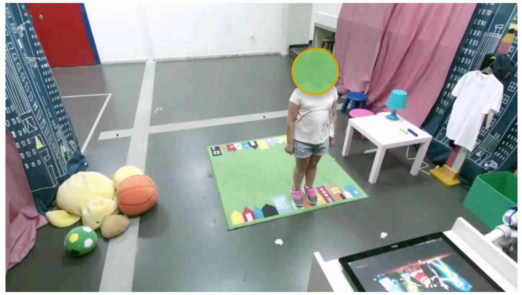


Sit

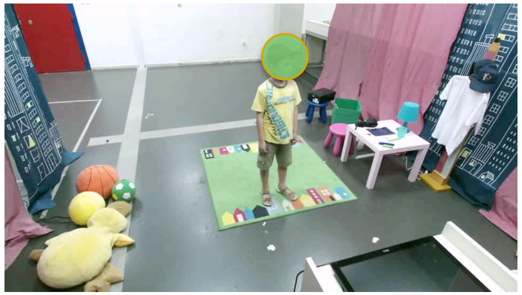


Stop

Point



Circle



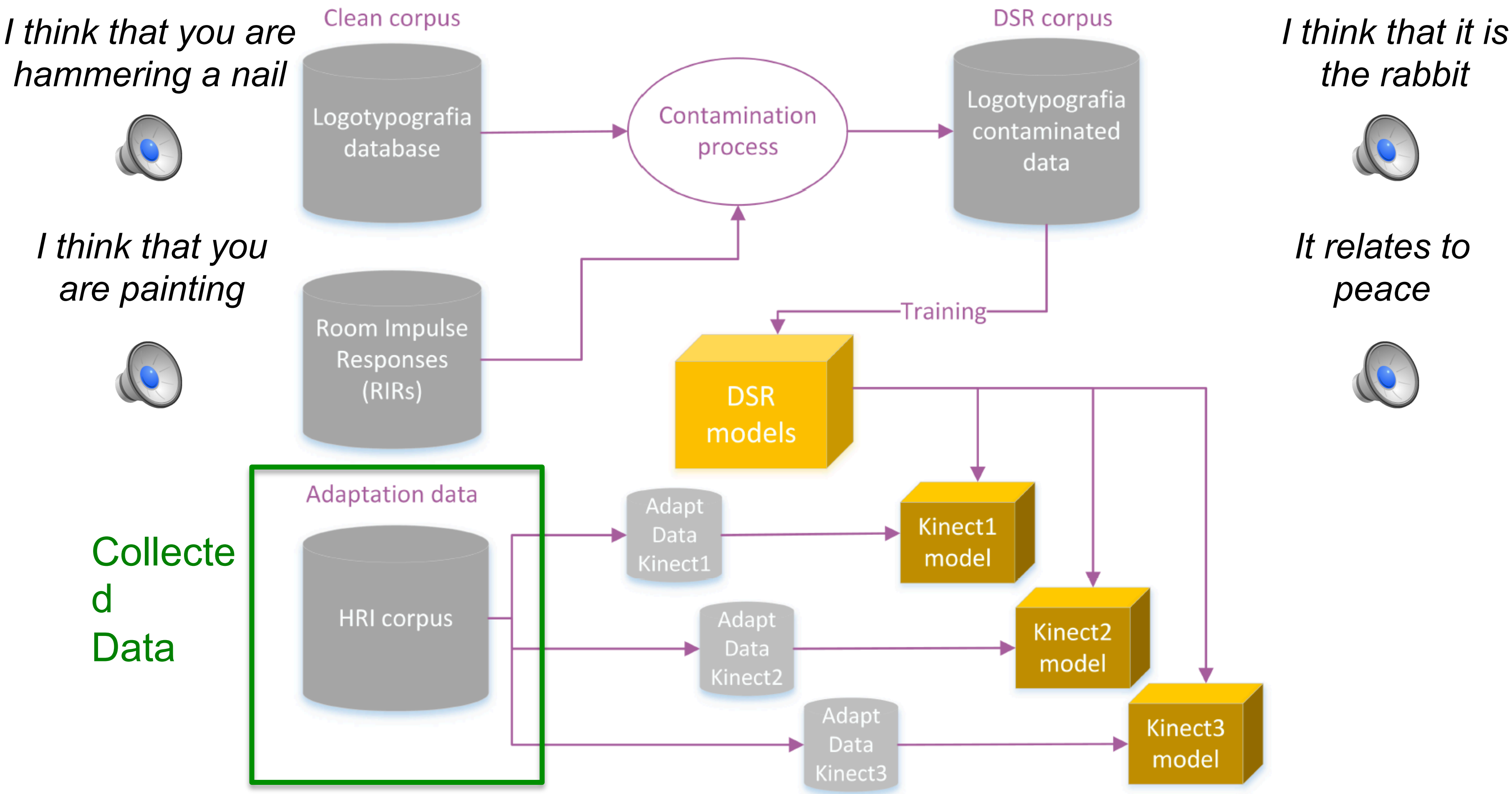
Multi-view Gesture Recognition - Evaluation

	Development Data													
	Single Camera								Fusion					
Feat.	Kinect #1		Kinect #2		Kinect #3		Kinect #4		Features		Encodings		Scores	
	BoW	VLAD	BoW	VLAD	BoW	VLAD	BoW	VLAD	BoW	VLAD	BoW	VLAD	BoW	VLAD
Traj.	68.75	70.83	66.90	64.58	65.74	65.74	68.52	62.96	75.00	76.39	76.85	79.63	77.31	78.24
HOG	40.74	33.33	33.33	31.25	29.40	30.79	41.20	31.94	39.81	40.28	41.67	39.35	41.20	37.04
HOF	70.83	72.69	70.37	71.76	69.21	66.67	63.43	53.70	71.76	74.07	77.78	81.48	75.93	81.94
MBH	76.85	75.93	67.82	73.38	68.29	68.75	65.28	57.41	76.39	76.85	81.02	81.48	82.87	83.80
Comb.	77.78	80.79	73.84	78.24	73.61	73.84	75.00	70.83	81.48	82.87	82.87	83.80	82.87	$\boxed{85.19}$

	Experimental Data													
	Single Camera								Fusion					
Feat.	Kinect #1		Kinect #2		Kinect #3		Kinect #4		Features		Encodings		Scores	
	BoW	VLAD	BoW	VLAD	BoW	VLAD	BoW	VLAD	BoW	VLAD	BoW	VLAD	BoW	VLAD
Traj.	45.76	50.55	42.12	45.19	45.41	58.85	45.56	43.84	54.16	58.90	51.88	58.39	49.00	65.37
HOG	24.13	28.32	29.70	22.26	17.09	19.13	41.25	27.79	37.84	35.59	31.79	27.76	34.48	31.78
HOF	56.92	66.93	54.49	65.93	57.10	63.01	51.97	37.14	54.56	71.61	58.01	74.73	63.26	74.83
MBH	62.70	63.00	56.47	65.95	60.15	68.04	54.25	56.33	65.32	72.70	67.72	72.52	66.73	72.72
Comb.	57.96	70.77	54.08	67.87	67.03	71.73	59.16	60.54	61.51	69.85	63.38	$\boxed{73.95}$	64.82	73.35

- Improved performance with VLAD encodings and multi-sensor fusion

Distant Speech Recognition System



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

Spoken Command Recognition Evaluation

Development Data: 40 phrases

	DSR-Adaptation scheme							
	No-adapt		Adults		Children		Mixed	
Test	WCOR	SCOR	WCOR	SCOR	WCOR	SCOR	WCOR	SCOR
Adults	97.54	91.25	99.58	**98.87**	96.73	93.20	99.50	98.43
Children	79.06	69.95	75.31	71.20	97.81	**95.50**	90.71	82.06

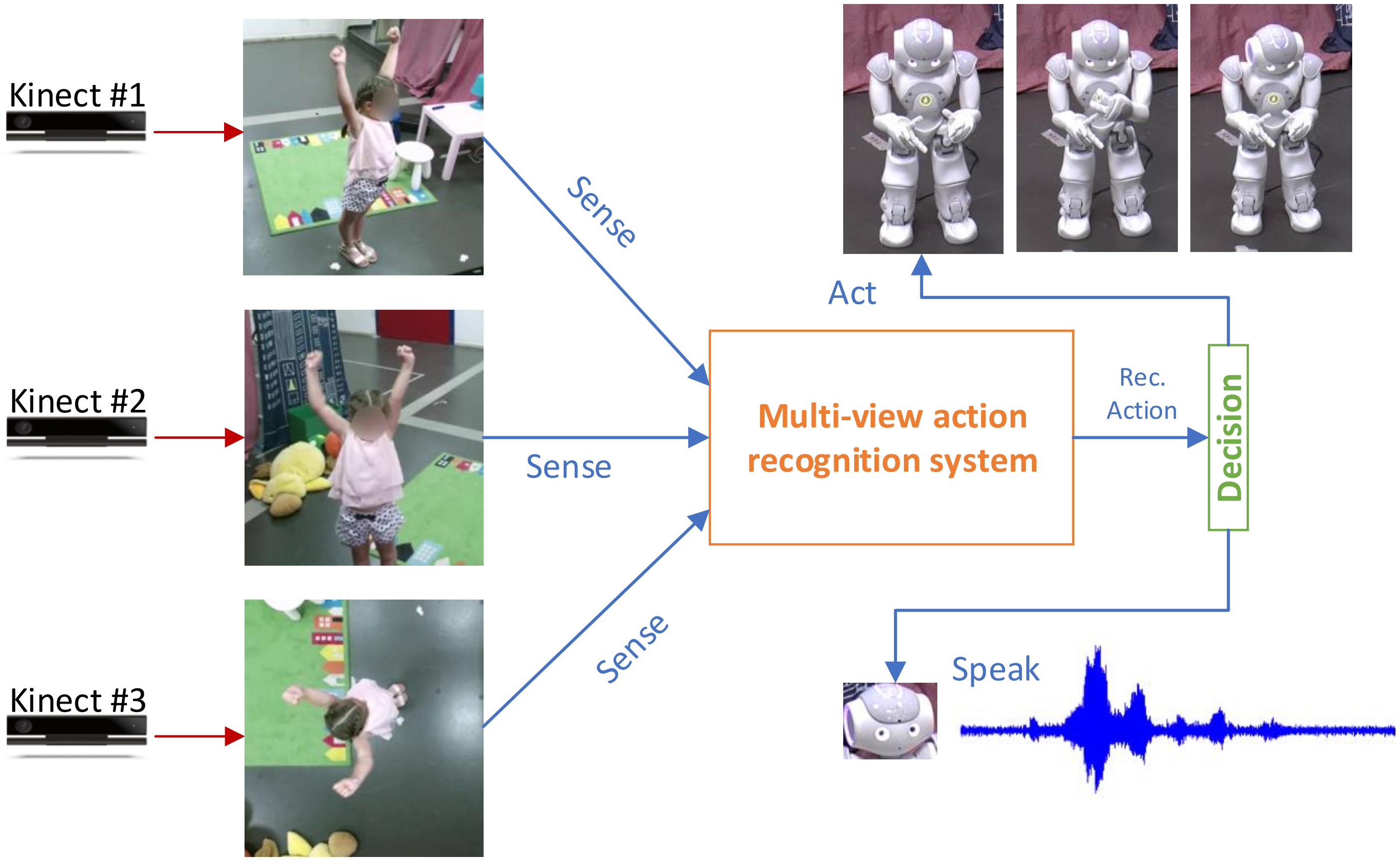
Experimental Data: All Annotated Utterenaces

	No-adapt			Adapt-all		
	WCOR	SCOR	LabelCOR	WCOR	SCOR	LabelCOR
Single Game	56.68	29.52	55.12	59.64	43.77	55.12
Cooperative Game	72.95	61.02	63.16	78.00	67.69	70.51

- Evaluation Metrics: average word accuracy (WCOR), sentence accuracy (SCOR), label accuracy (LabelCOR).
- Improved performance with children adapted models

Multi-view Child Action Recognition System

- Same frontend with Gesture Recognition
- Recognize challenging human movements like pantomime actions



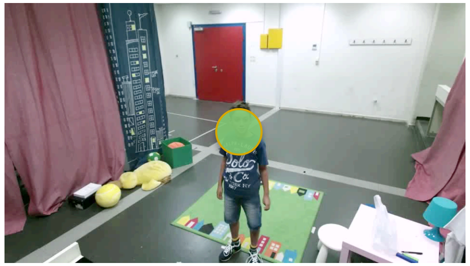
Efthymiou et al. In Proc. ICIP, 2018.

Multi-view fusion for action recognition

- Employ the same multi-view fusion approaches as in gesture recognition
- Compare with an additional implemented method for multi-view action recognition pre-trained CNN features

Action Recognition – Vocabulary

Cleaning a window

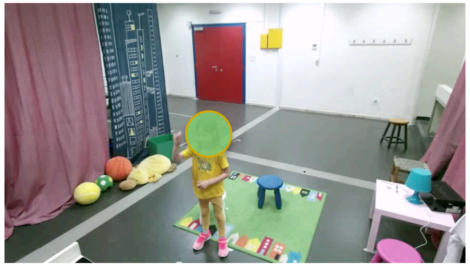


Ironing a shirt

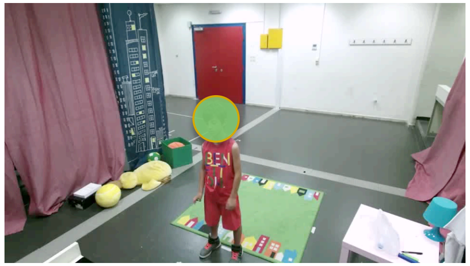
Digging a hole

Driving a bus

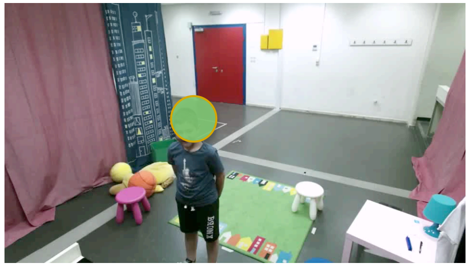
Painting a wall



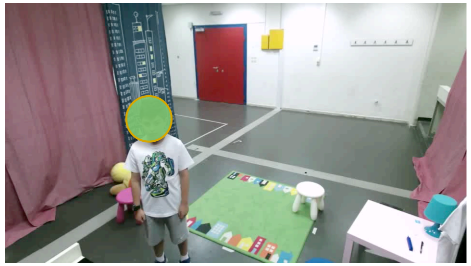
Hammering a nail



Wiping the floor



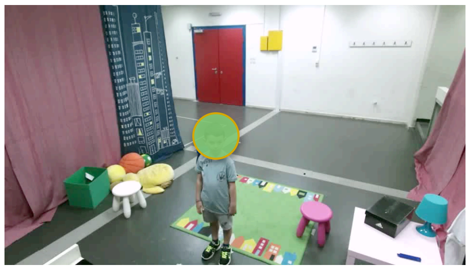
Reading



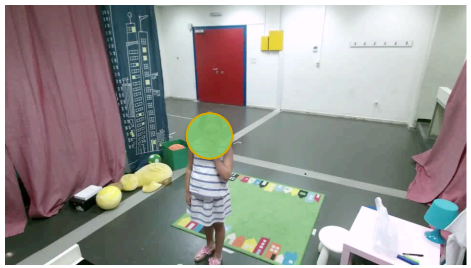
Swimming

Working Out

Playing the guitar



Dancing

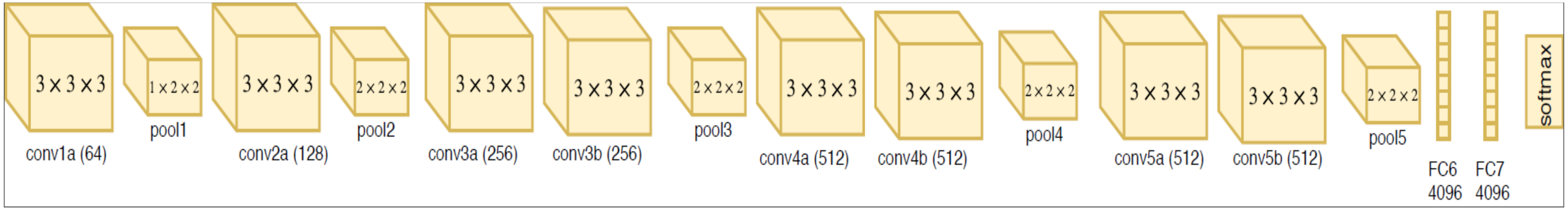


Multi-view Gesture Recognition - Evaluation

	Development Data													
	Single Camera								Fusion					
Feat.	Kinect #1		Kinect #2		Kinect #3		Kinect #4		Features		Encodings		Scores	
	BoW	VLAD	BoW	VLAD	BoW	VLAD	BoW	VLAD	BoW	VLAD	BoW	VLAD	BoW	VLAD
Traj.	63.08	60.31	48.62	48.62	45.45	46.46	49.73	55.08	62.15	60.00	66.15	66.77	64.31	69.50
HOG	36.69	36.69	32.00	38.15	27.69	34.46	28.30	50.15	48.62	50.15	49.85	54.15	44.31	58.00
HOF	68.31	69.85	56.31	63.08	48.62	50.46	53.85	63.69	66.77	67.08	68.00	69.23	68.62	75.50
MBH	70.77	72.92	60.92	68.62	61.85	60.00	55.22	72.92	76.00	76.69	76.92	76.92	74.46	76.50
Comb.	73.85	74.15	63.38	69.23	60.00	58.46	61.45	76.31	75.08	76.92	77.23	77.85	75.08	$\boxed{79.00}$
	Experimental Data													
	Single Camera								Fusion					
Feat.	Kinect #1		Kinect #2		Kinect #3		Kinect #4		Features		Encodings		Scores	
	BoW	VLAD	BoW	VLAD	BoW	VLAD	BoW	VLAD	BoW	VLAD	BoW	VLAD	BoW	VLAD
Traj.	47.89	50.84	38.49	43.75	25.52	22.61	44.14	44.64	58.26	52.30	50.79	54.87	47.45	56.99
HOG	30.98	23.10	23.95	27.51	16.53	19.73	21.76	34.14	37.41	31.49	26.95	36.26	29.14	31.44
HOF	46.34	51.00	46.19	51.78	25.50	26.13	47.70	44.67	63.08	61.02	49.87	56.17	52.59	57.99
MBH	61.42	56.86	46.28	45.82	31.59	29.44	45.57	56.11	70.25	67.97	57.70	59.04	62.18	62.49
Comb.	52.59	59.16	46.74	51.51	36.62	38.22	48.16	55.11	63.52	$\boxed{69.37}$	60.75	61.55	55.00	64.90

- Improved performance with VLAD encodings and multi-sensor fusion
  - Bigger improvement in the experimental data

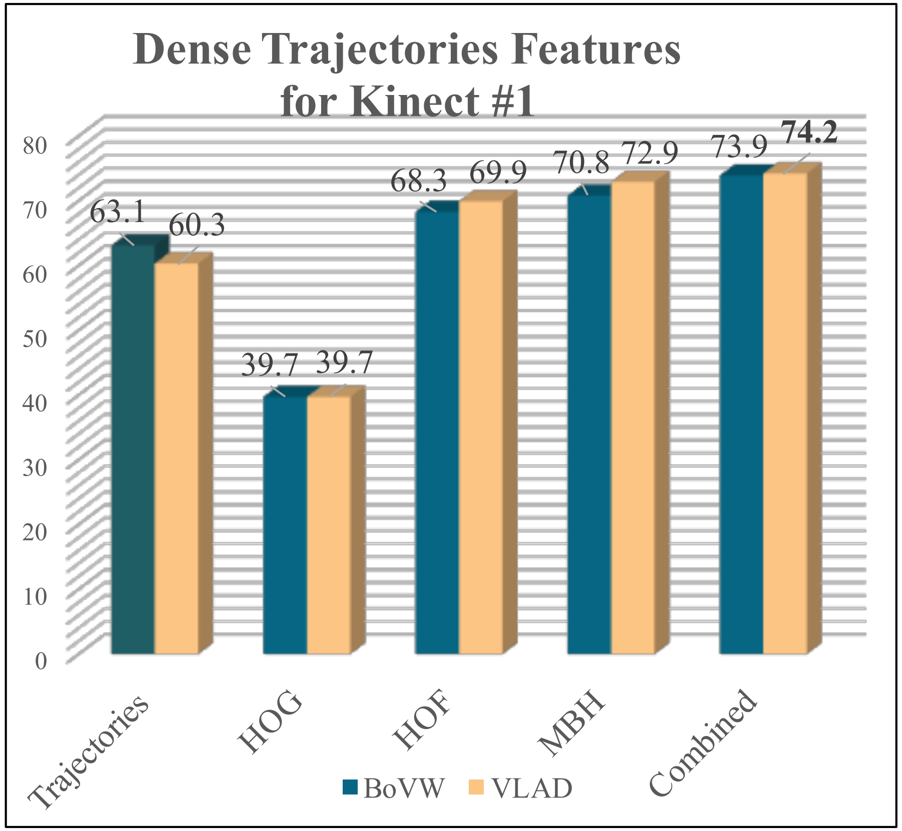
Convolutional Neural Network Approach

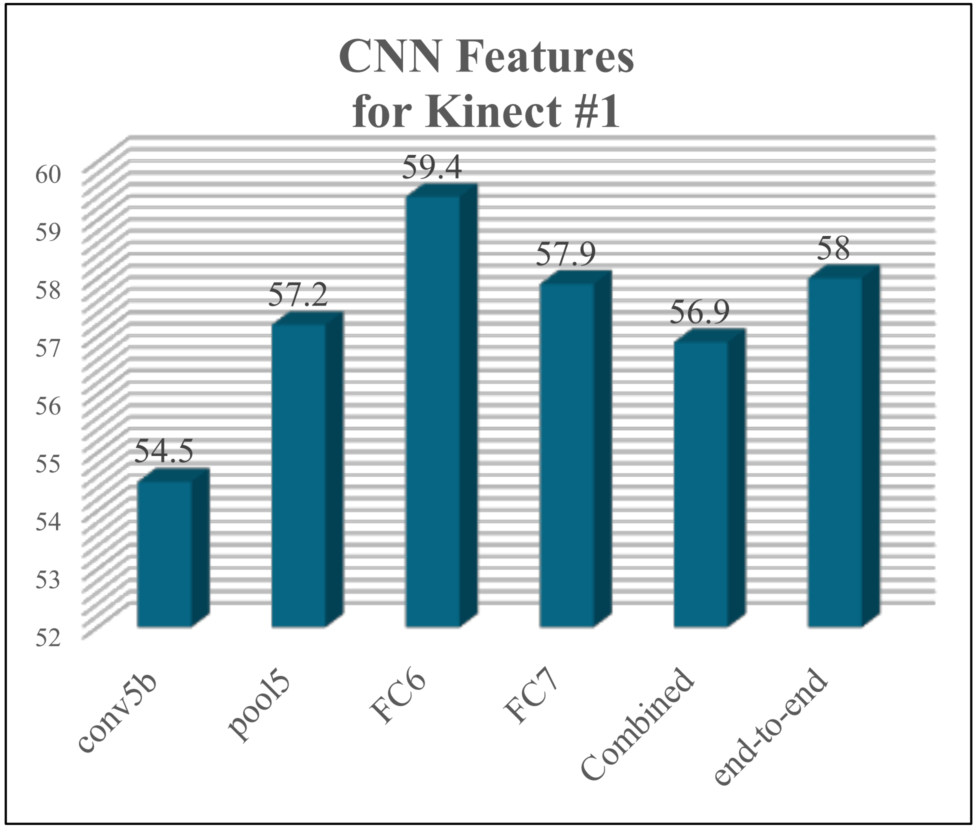


- Pretrained 3D CNNs on the Sports1M corpus* (487 classes)
- Finetuning:
  - Split each video in 16 frame clips with the 15 frames overlapping
- Testing:
  - Extract features from FC6, FC7, pool5, conv5b layers
  - Average over each clip to obtain a descriptor for each video
- Fusion: early, middle and late as in dense trajectories

Efthymiou et al. In Proc. ICIP, 2018.
D. Tran et al., In in Proc. ICCV, 2015.

CNN Approach: Single-view Evaluation





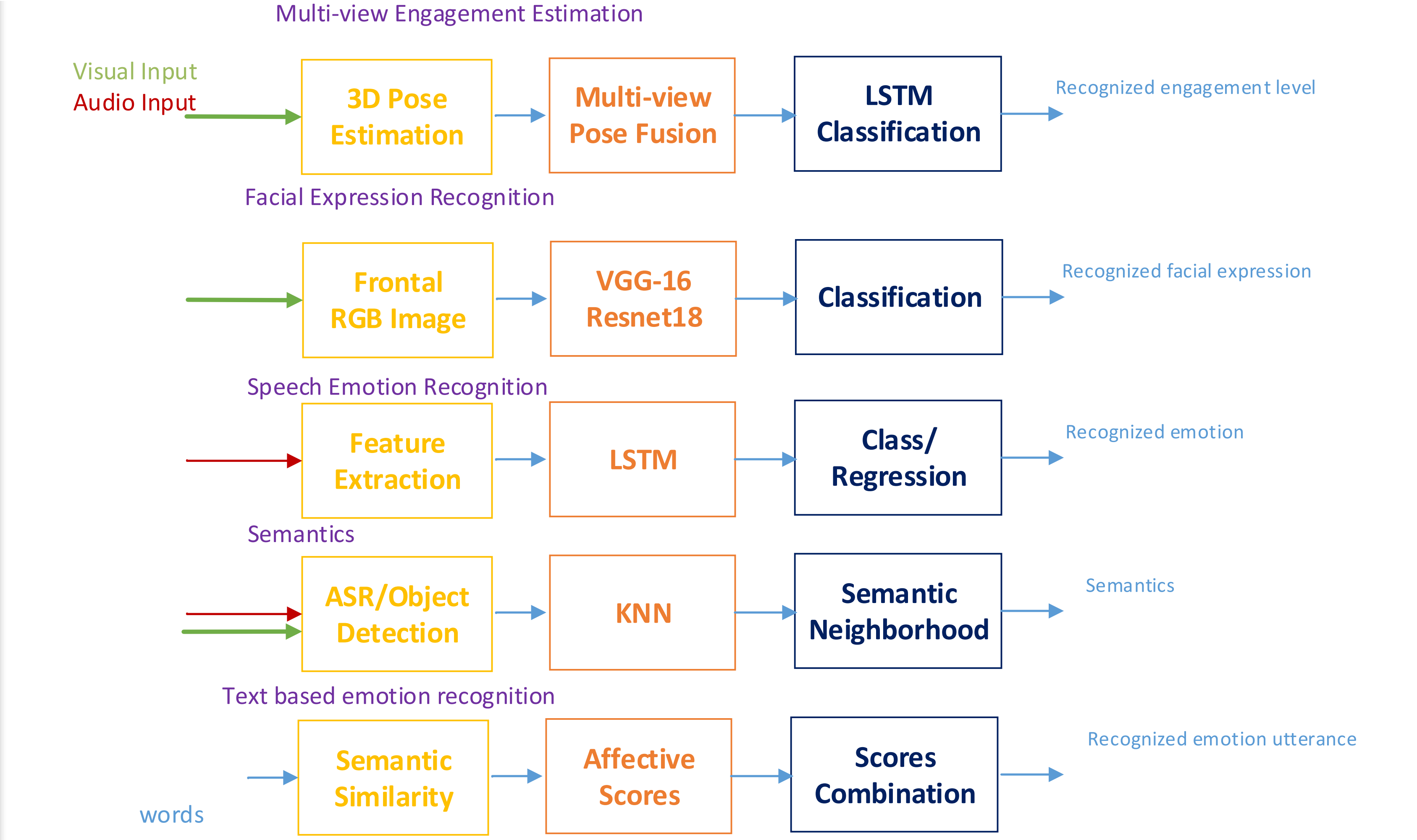
CNN Approach: Multi-view Evaluation

CNN Features			
Fusion / C3D Feats.	Feature Fusion	Encodings Fusion	Score Fusion
conv5b	58.77	61.23	62.46
pool5	60.31	61.23	63.08
FC6	60.31	63.08	62.46
FC7	63.08	63.08	62.15
Comb.	60.31	61.23	**63.69**
end-to-end	-	-	61.72

- Fusion schemes achieve to improve the performance of the single-view approach
- Recognition accuracy isn't sufficient for a CRI task
- Dense Trajectories perform better than Deep Learning features:
  - Pretrained models on very different datasets (sports, movies)
  - Not enough CRI data for training end-to-end deep models

Behavioral Branch: Modules

- recognize and identify the behavior of the child that is expressed as facial expressions and emotions, skeleton pose, engagement



Engagement in Child-Robot Interaction

- Engagement is important for lasting use of robots by children
  - ✓ Robot should be able to recognize the status of a child's engagement
  - ✓ Robot should be able to react to disengagement with re-engaging actions

Multi-view Engagement Estimation

(a) Class 1

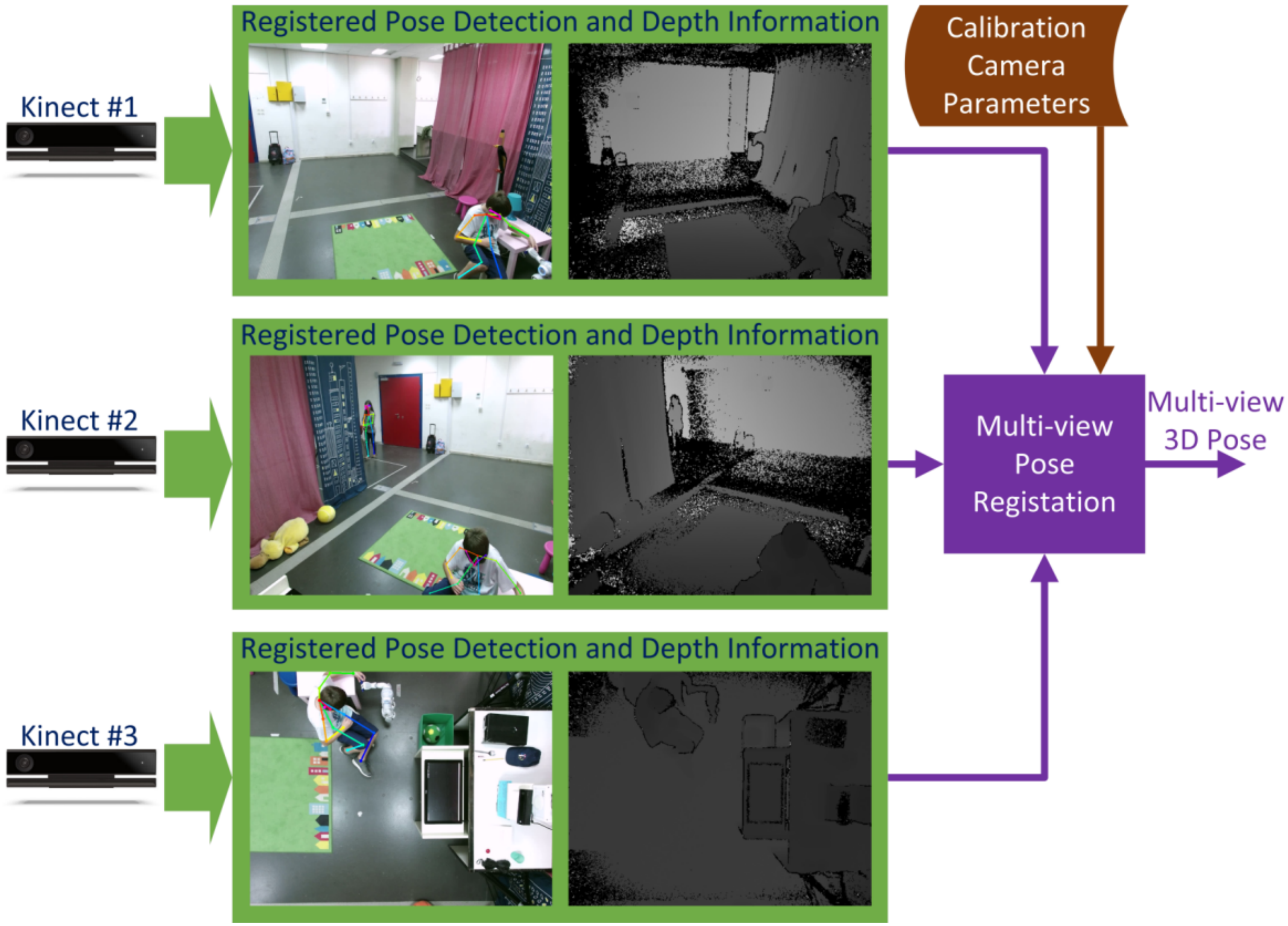
(b) Class 2

(c) Class 3

- Classify the engagement level of children in joint attention tasks

Hadfield et al. In Proc. IROS, 2019

Engagement: Fusion of 3D Child Pose



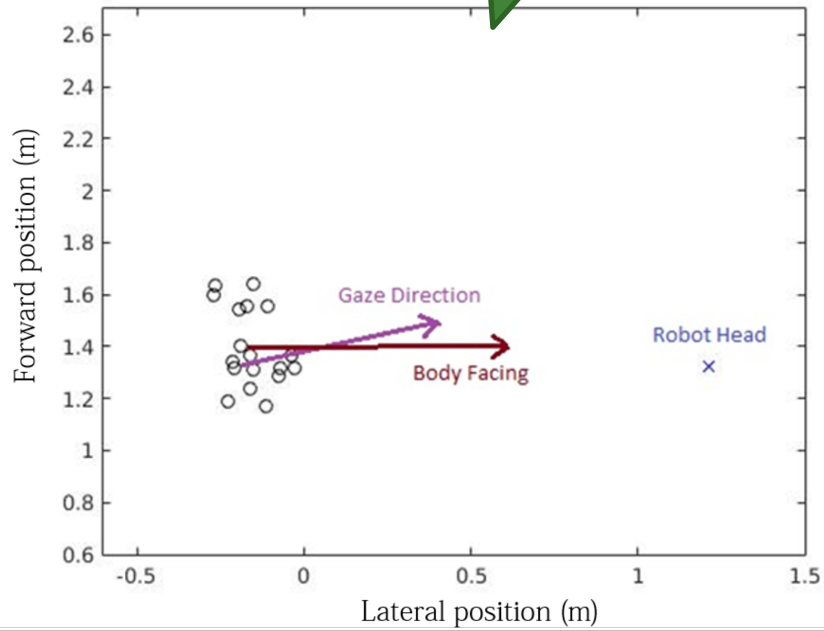
Engagement: Feature extraction

3D pose estimation

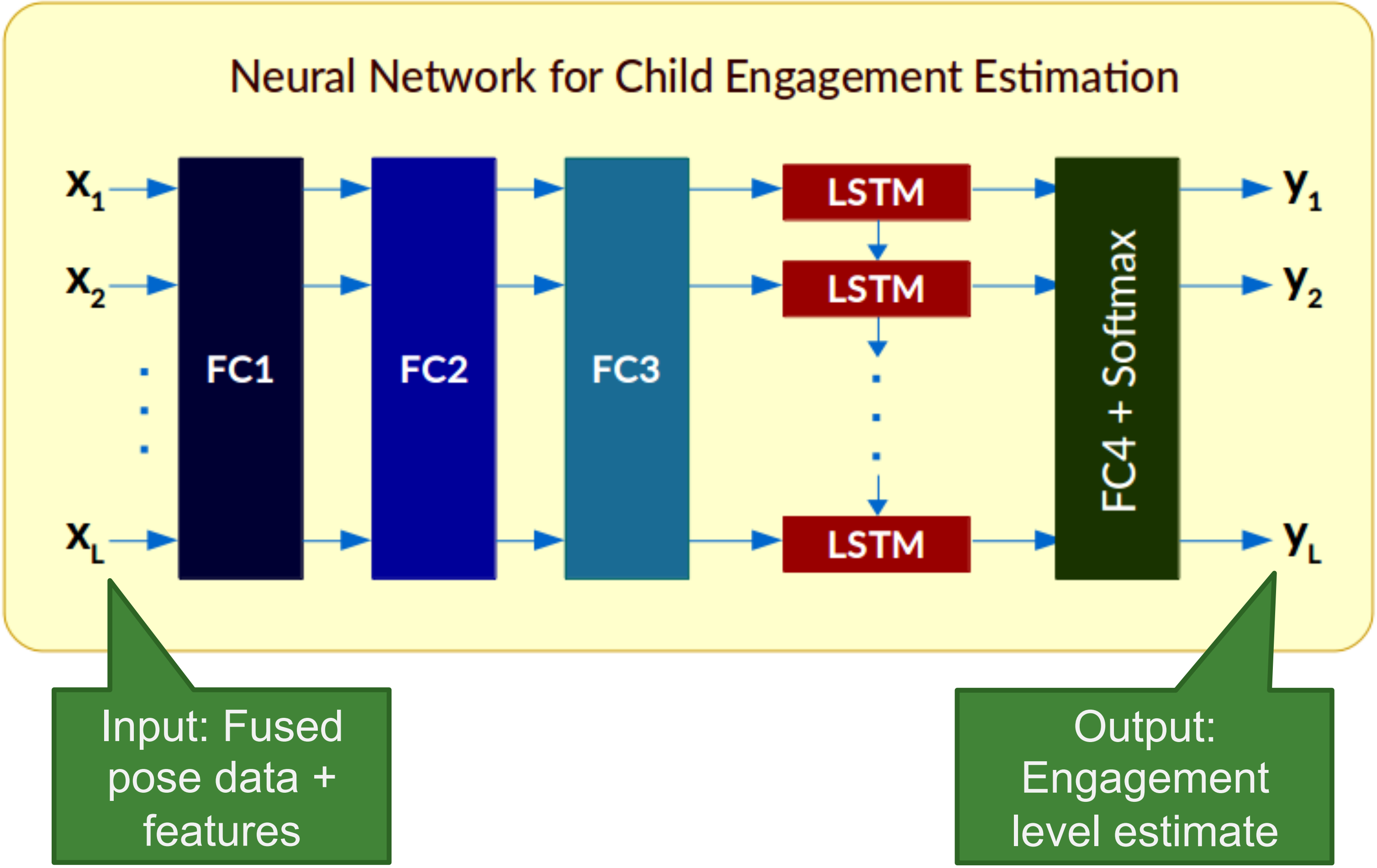
Nao detection: YOLO v3

Robot position is subtracted from keypoints

Features: Gaze, body facing, arm extension



Engagement: Classifier Architecture



Engagement: Evaluation Results

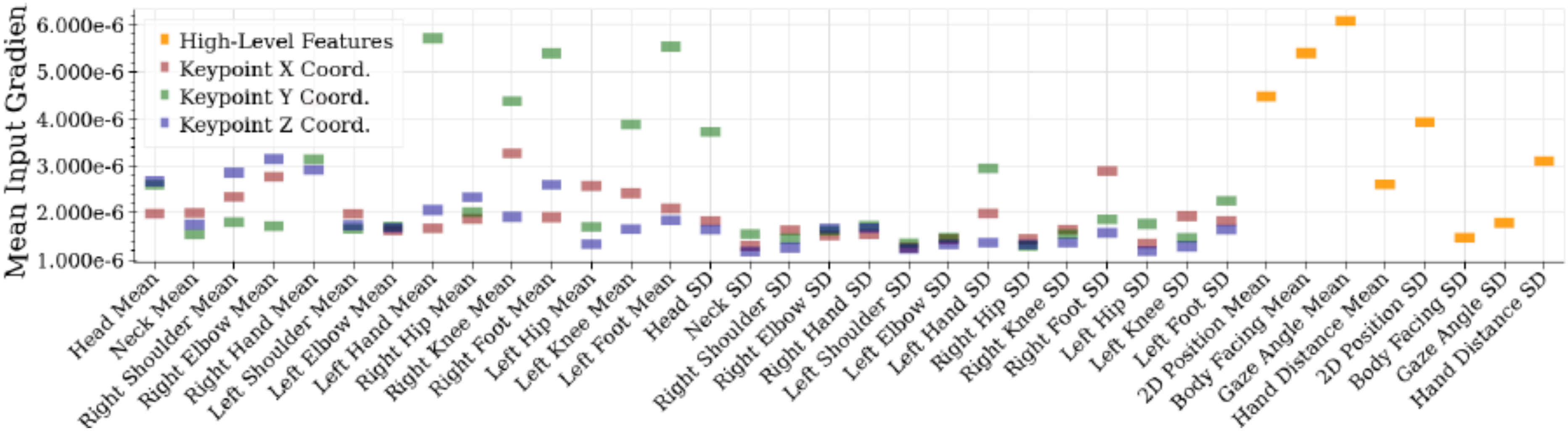
Net Architecture	Mean F-Score	Accuracy	Balanced Accuracy
$3 \cdot$ FC + LSTM	**62.18**	**77.11**	**61.88**
$2 \cdot$ FC + LSTM	56.23	71.86	58.46
$3 \cdot$ FC + $2 \cdot$ LSTM	54.78	70.60	56.30
$2 \cdot$ FC + $2 \cdot$ LSTM	54.45	69.71	56.91

Cross-validation results for different network architectures

Method	Mean F-Score	Accuracy	Balanced Accuracy
Majority class	27.90	71.97	33.33
Gaze LSTM	32.78	71.58	36.10
SVM	54.79	68.27	58.61
RF	56.41	68.60	61.78
3FC+LSTM	**62.18**	**77.11**	**61.88**

Performance results of different algorithms

Engagement: Features Importance



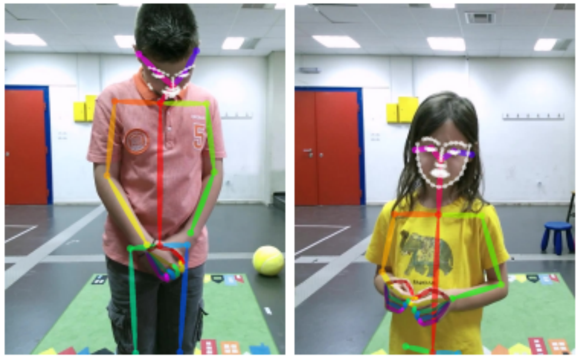
- Feature importance of input variables based on their average input gradient over 100 trained networks

Engagement: Demo Video

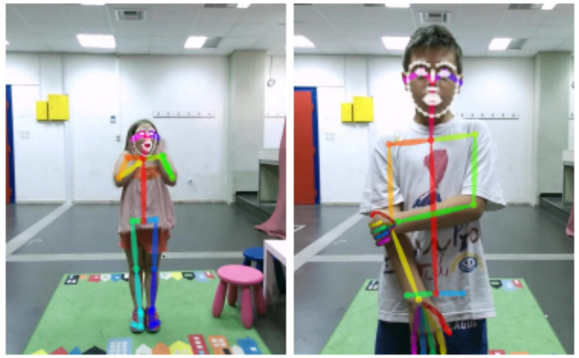
Visual Emotion Recognition - Patterns

happiness
mainly facial, rare jumping and/or open raised hands, body erect, upright head

sadness
crying (with hands in front on face), motionless, head looking down, contracted chest



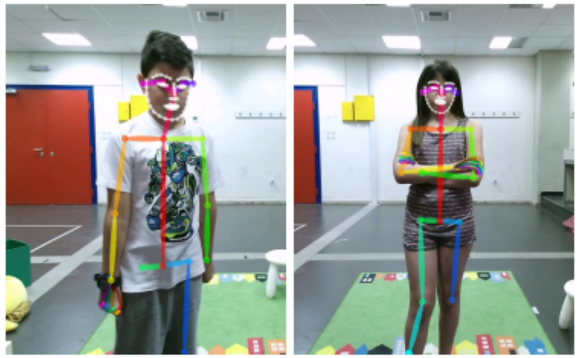
surprise
expanded chest, hand movement without specific patterns, either positive or negative surprise



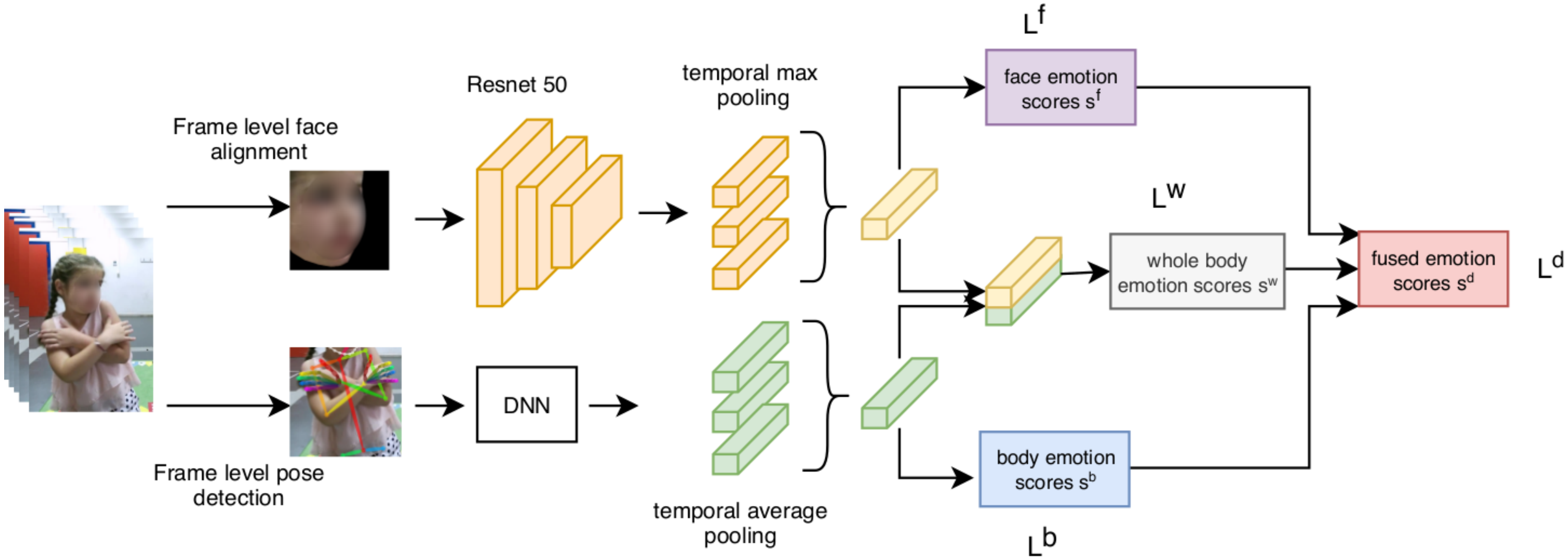
fear
quick eye gaze, weak facial expressions, arms crossed in front of body, head sink

disgust
mainly with facial expression (tongue out), movement away from/hands against robot

anger
clenched fists, arms crossed, squared shoulders



Visual Emotion Recognition – HMT network



Hierarchical Multi-label training (HMT) for recognition of affect from multiple visual cues.

$$\mathcal{L} = \mathcal{L}^f(y^f, \tilde{s}^f) + \mathcal{L}^b(y^b, \tilde{s}^b) + \mathcal{L}^w(y, \tilde{s}^w) + \mathcal{L}^d(y, \tilde{s}^d)$$

Filntisis et al., IEEE Robotics and Automation Letters, 2019.

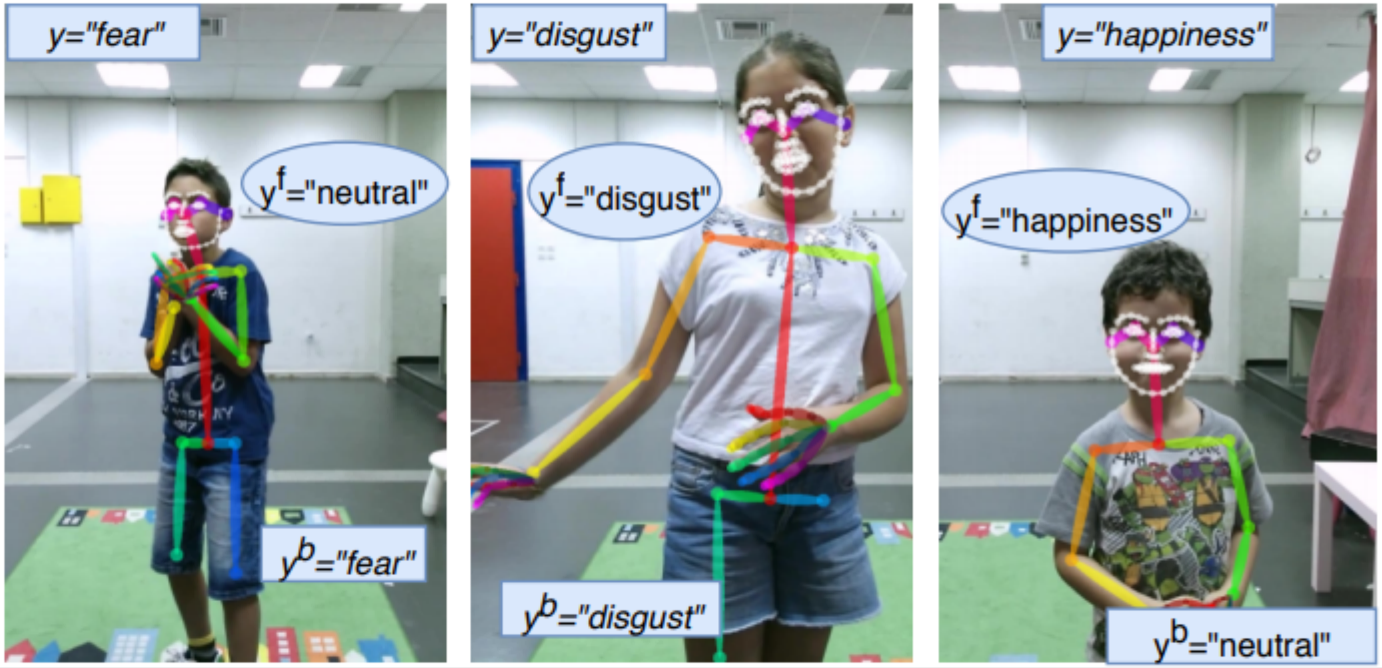
Visual Emotion Recognition - Database

30 children × 6 emotions for two sessions:
Acted and Spontaneous

	Total (#)	Facial (#)	Body (#)
Neutral	-	37	99
Happiness	44	44	9
Sadness	35	25	18
Surprise	30	30	13
Fear	32	14	31
Disgust	43	42	19
Anger	27	19	22

Statistics of multi-label annotations

Example multi-label annotations



Emotion Recognition with HMT - Results

	Label	y (6 classes)		y^f (7 classes)		y^b (7 classes)	
		F1	ACC	F1	ACC	F1	ACC
	Joint-1L	0.64 (0.66)	0.65 (0.67)	-	-	-	-
SEP	Body br.	0.29 (0.29)	0.35 (0.33)	-	-	0.36 (0.53)	0.38 (0.56)
	Face br.	0.57 (0.60)	0.61 (0.63)	0.50(0.59)	0.52 (0.61)	-	-
	Sum Fusion	0.63 (0.66)	0.65 (0.67)	-	-	-	-
HMT-3a	Body br.	0.32 (0.33)	0.36 (0.35)	-	-	0.35 (0.48)	0.39 (0.47)
	Face br.	0.54 (0.57)	0.59 (0.63)	0.48 (0.57)	0.52 (0.61)	-	-
	Fusion	0.64 (0.67)	0.66 (0.68)	-	-	-	-
HMT-3b	Body br.	0.32 (0.32)	0.36 (0.34)	-	-	0.36 (0.50)	0.39(0.49)
	Face br.	0.53 (0.56)	0.59(0.63)	0.51 (0.60)	0.54 (0.63)	-	-
	Whole body br.	0.64 (0.66)	0.66 (0.68)	-	-	-	-
HMT-4	Body br.	0.32 (0.32)	0.36 (0.34)	-	-	0.34 (0.47)	0.38(0.46)
	Face br.	0.53 (0.56)	0.58(0.62)	0.49 (0.58)	0.53 (0.62)	-	-
	Fusion	**0.69** (**0.71**)	**0.71** (**0.72**)	-	-	-	-

Visual Emotion Recognition: Demo Video

Examples of hierarchical recognitions of the HMT network in videos

Top Rectangle Box: final prediction

Middle Oval Box: face branch prediction

Bottom Rectangle Box: body branch prediction

Green Color: correct prediction

Red Color: incorrect prediction

Child-Robot Interaction: Multiple Children

Part 5: Conclusions

- Present and discuss many **perception modules** for CRI based on video processing and machine learning techniques
- Describe a **database for CRI applications** and highlight the importance for collecting children data to improve the performance of recognition algorithms
- Present **integrated systems** for CRI by employing multi-modal perception modules and multiple robots
- Future work:
  - Deploy the developed technology to more challenging CRI scenes
  - Explore ideas from zero-shot learning in order to design more generic interaction perception systems

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