Introduction to Research Efforts on

Robot AI for Elderly-Care

Talk @ CBNU

2020.10.15

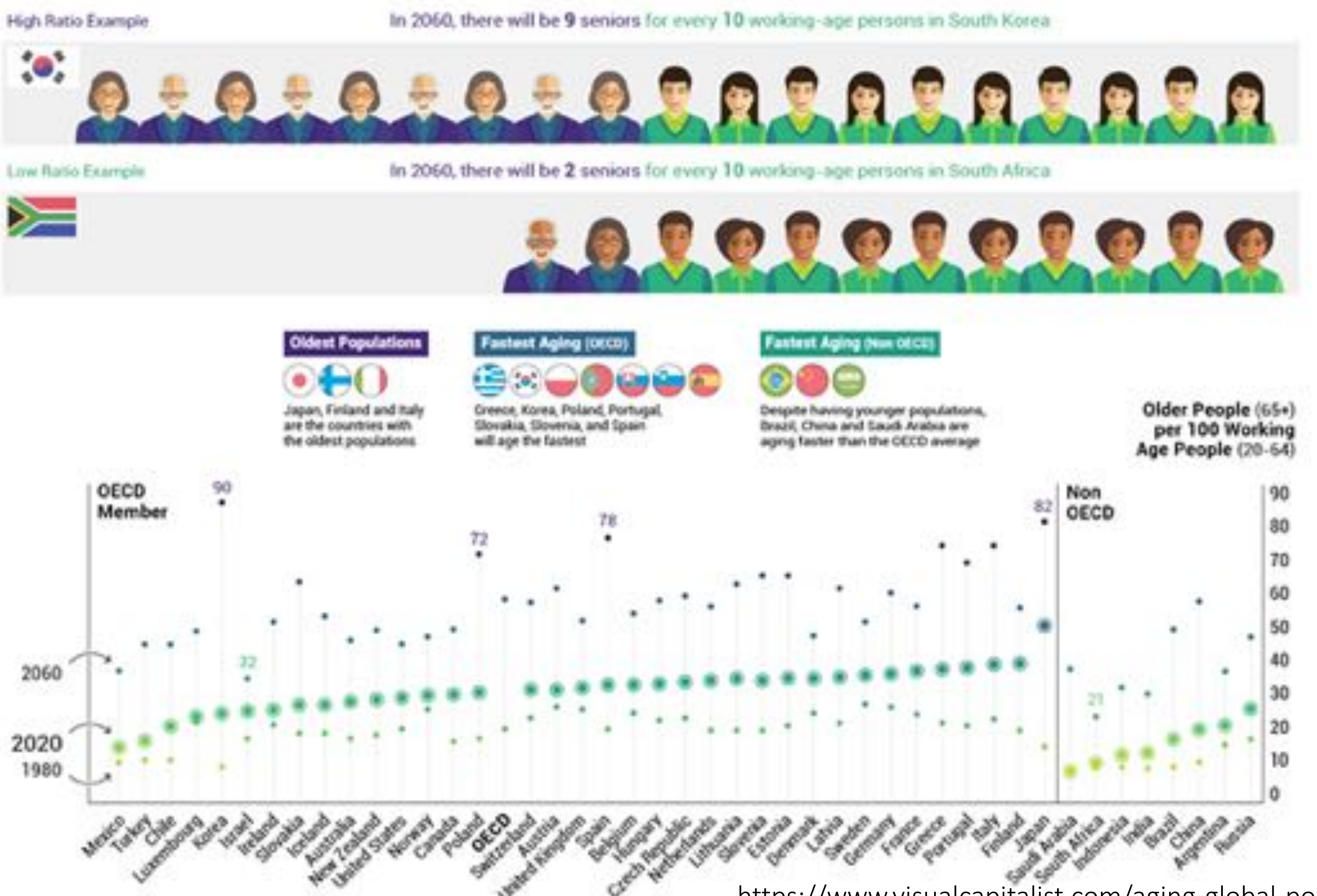
Minsu Jang (minsu@etri.re.kr)
Human-Robot Interaction Research Lab
Electronics and Telecommunications Research Institute

Outline

- **Motivation and Challenges**
- **Domain AI for Elderly-Care**
  - Daily Activity Detection
  - Human Detection and Tracking
  - Human Attributes Recognition
  - Object Instance Detection
  - Elderly Voice Recognition
- **Robot Social AI**
  - Co-Speech Gesture Generation
  - Non-Verbal Interaction Behavior Generation
- **Summary**

Motivation and Challenges

Aging society is a global problem

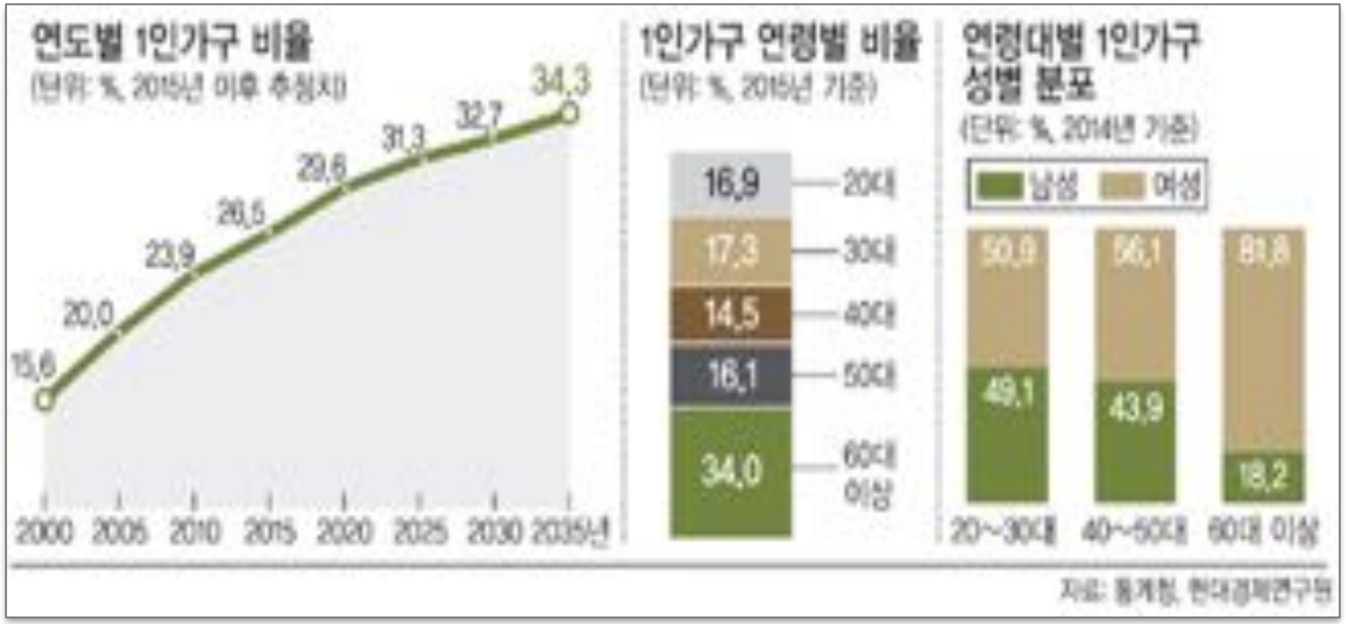


https://www.visualcapitalist.com/aging-global-population-problem/

ETRI

The problem of aging population in Korea

- Population of the elderly over 65 years of age: 13.8%('19) → 20%('25)
- More than half of the elderly will live alone in 2030



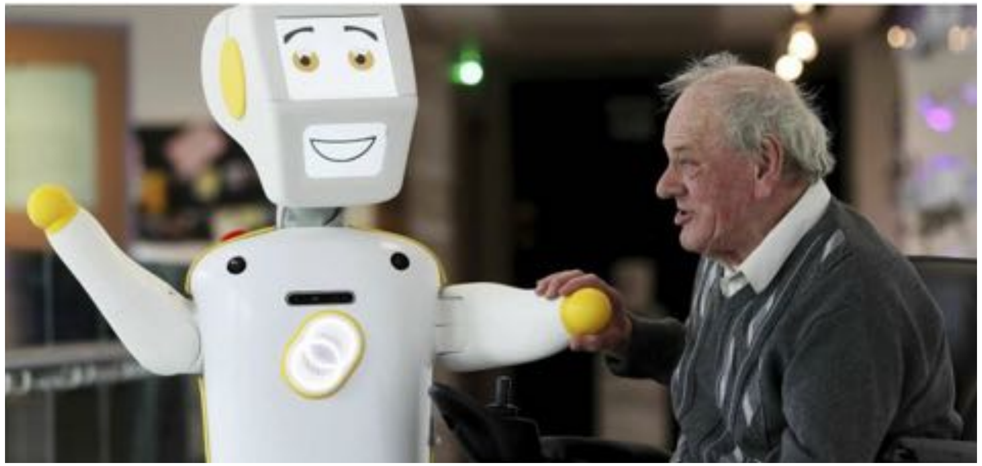
Elderly people are fragile

- Social isolation: more than 20% of the elderly
- **Mental health problems**: Loneliness, Psychological Distress, Depression
- Mental health and physical health have an impact on each other
  - Depression → Heart Disease

Assistive robots for elderly-care

https://www.mdpi.com/2079-9292/9/2/367/htm

http://www.seoulilbo.com/news/articleView.html?idxno=379516

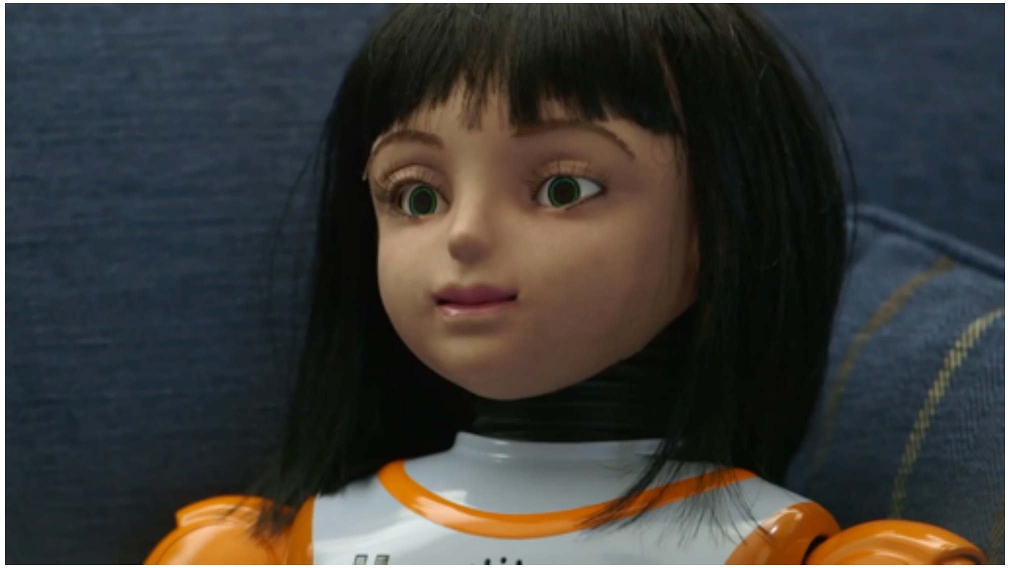


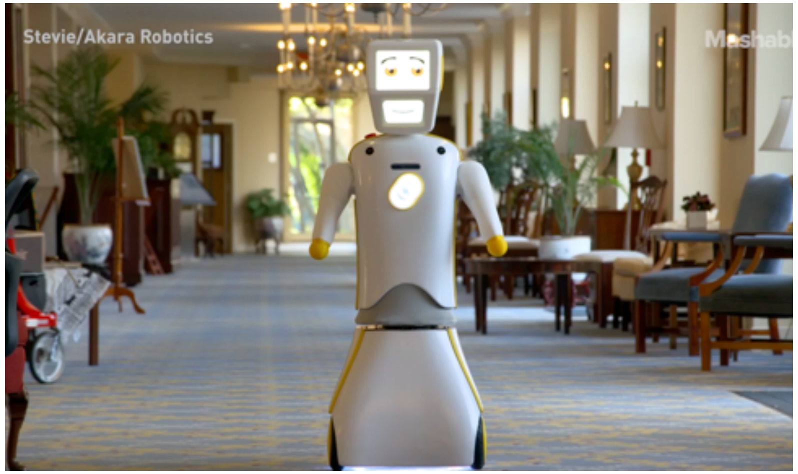
https://www.mdpi.com/2079-9292/9/2/367/htm

http://shorturl.at/qER39

PARs: Physically Assistive Robots

SARs: Socially Assistive Robots





We are trying to realize...

SARs (Socially Assistive Robots)

Human-aware Perception Understanding & Empathy	Human-like Behaviors Emotional & Sympathy

"You are dressed up today. Fedora hat looks great on you."

"I am very sorry to hear that..."

Challenges of Elderly Domain: STT

Subject	Non-elderly	Elderly	Difference
Women	10.4% (27 speakers)	40.3% (32 speakers)	+29.9%
Men	11.7% (25 speakers)	61.3% (11 speakers)	+49.6%
Average	11.0%	45.7%	34.7%
Standard deviation	6.4%	16.8%	10.4%

STT Performance on Non-Elderly vs Elderly Speech

- Imprecise in consonant pronunciation
- Tremors
- Slower Articulation

Vacher, M., Aman, F., Rossato, S. and Portet, F., 2015, August. Development of automatic speech recognition techniques for elderly home support: Applications and challenges. In International Conference on Human Aspects of IT for the Aged Population (pp. 341-353). Springer, Cham.

Challenges of Elderly Domain: Our Experiments

Speech Recognition

Table 3. Speech Recognition Result per Age Group

Age Group	Number of Subjects	WER Average ± SD (%)	*p* value when compared to 25-50 group
25-50	5	16.25 ± 6.42	-
50-64	6	17.89 ± 7.72	0.2607
65-69	6	17.45 ± 8.92	0.4513
70-74	6	18.12 ± 12.33	0.3537
78+	8	20.45 ± 10.23	0.0291*

- *Data: 12 hours of speech*
- *Speech Recognizer: Google Cloud Speech*

Challenges of Elderly Domain: Our Experiments

Facial Expression & Emotion Recognition (with Affdex[1])

Older Adults

attention 0.91	eye closure 0.83	smile 0.95	jaw drop 0.82	lip stretch 0.72
brow furrow 0.92	eye widen 0.81	cheek raise 0.83	lip corner depressor 0.73	mouth open 0.94
brow raise 0.89	lid tighten 0.82	chin raise 0.73	lip press 0.77	upper lip raise 0.75
inner brow raise 0.94	nose wrinkle 0.96	dimpler 0.84	lip pucker 0.78	smirk 0.82

anger 0.89	contempt 0.76	disgust 0.82	fear 0.62	joy 0.82
sadness 0.72	surprise 0.91	valence 0.88	arousal 0.75	

Younger Adults

attention 0.91	eye closure 0.85	smile 0.95	jaw drop 0.88	lip stretch 0.69
brow furrow 0.96	eye widen 0.86	cheek raise 0.85	lip corner depressor 0.82	mouth open 0.92
brow raise 0.94	lid tighten 0.79	chin raise 0.78	lip press 0.77	upper lip raise 0.82
inner brow raise 0.94	nose wrinkle 0.92	dimpler 1.00	lip pucker 0.82	smirk 0.92

anger 0.88	contempt 0.72	disgust 0.82	fear 0.69	joy 0.88
sadness 0.72	surprise 0.93	valence 0.90	arousal 0.83	

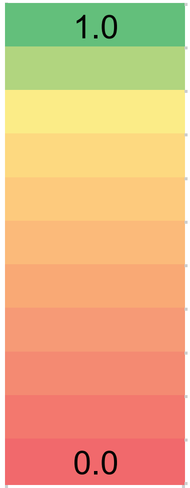
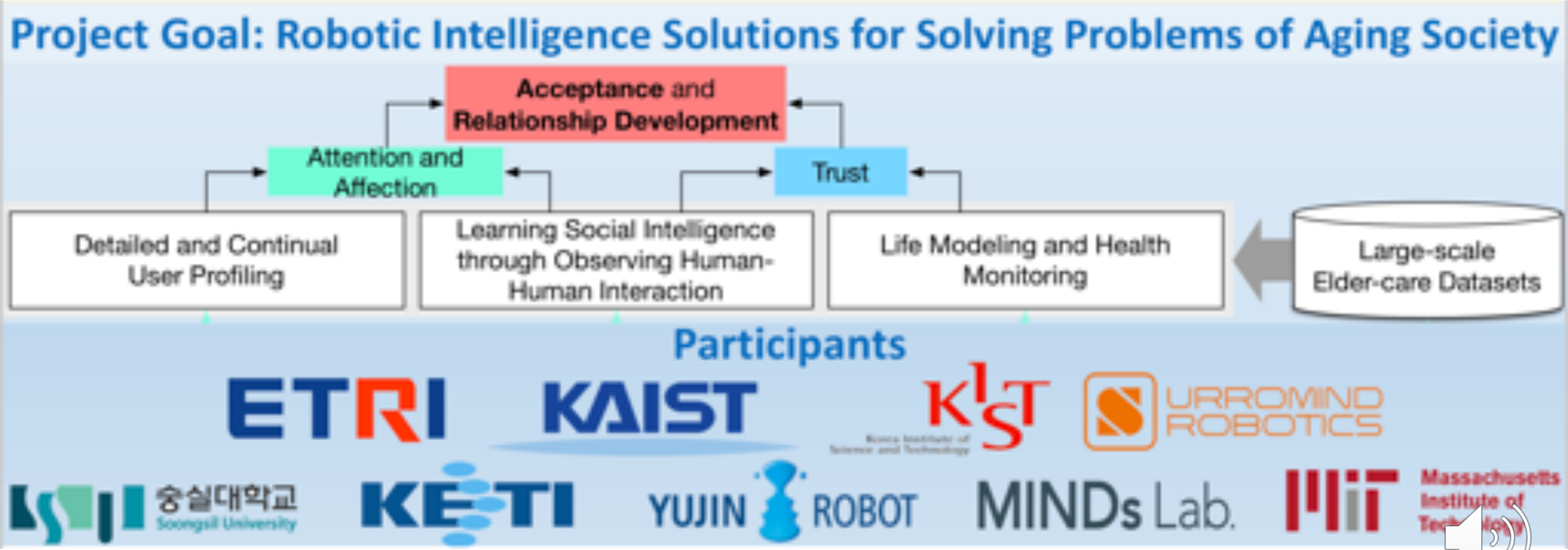


Figure 1. Facial expression and emotion recognition result (ROC)

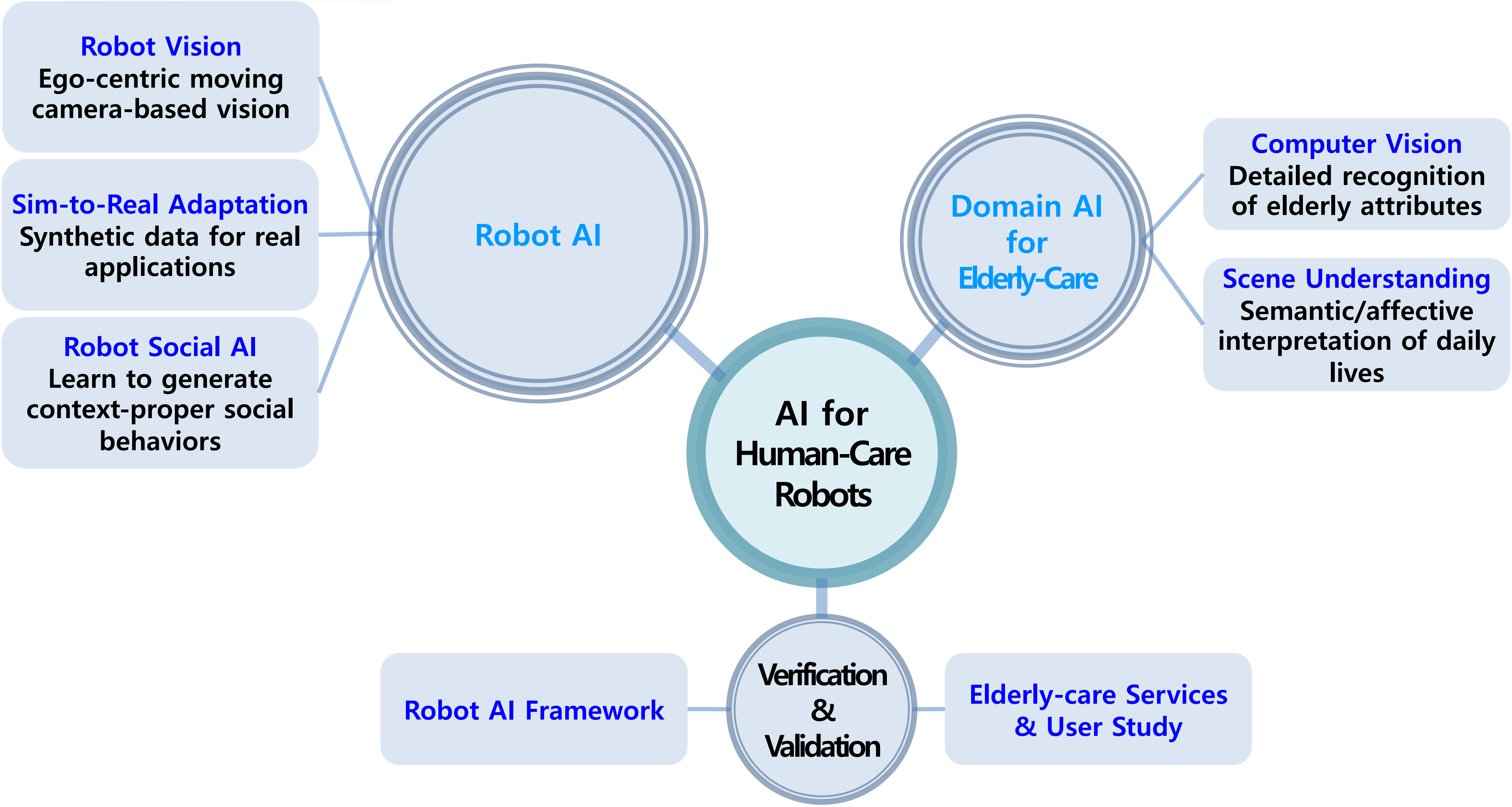
[1] McDuff, Daniel, Abdelrahman Mahmoud, Mohammad Mavadati, May Amr, Jay Turcot, and Rana el Kaliouby. "AFFDEX SDK: a cross-platform real-time multi-face expression recognition toolkit." In Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems, pp. 3723-3726. 2016.

AIR Project

"Development of Human-Care Robot Technology for Aging Society" (2017~2021, MSIT)



Research Issues



Domain AI for Elderly-Care

Domain AI for Elderly-Care

Elderly People

Home Environment

Human Detection/Tracking

Daily Activity Detection

Human Attribute Recognition

Personal Items Registration/Detection

Affective Scene Understanding

Elderly-Voice Recognition

Image-based Story Generation

Video QA

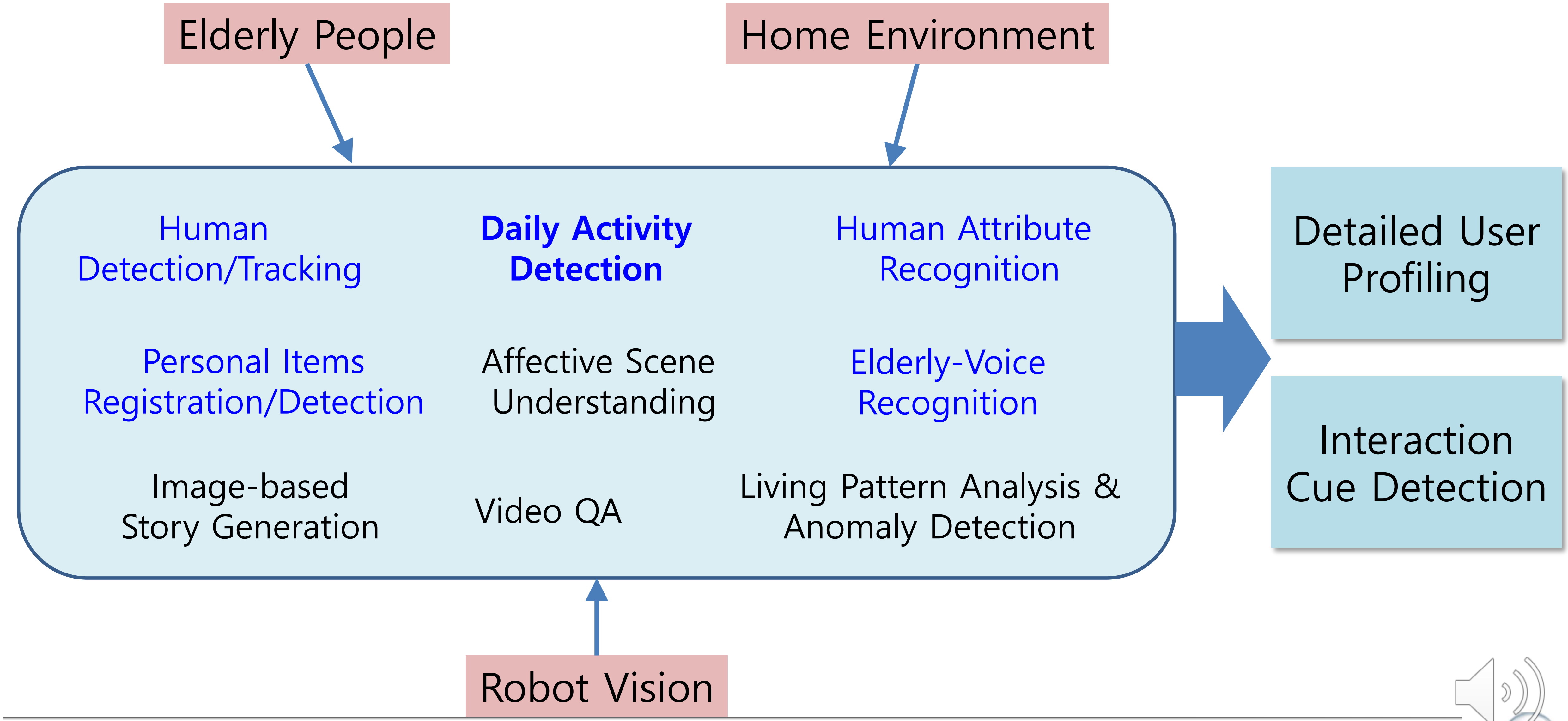
Living Pattern Analysis & Anomaly Detection

Robot Vision

Detailed User Profiling

Interaction Cue Detection

Domain AI for Elderly-Care



Daily Activity Detection for the Elderly

- Hypothesis: Motions of elderly people are very different from those of young adults.

We need data directly from elderly people.

Elderly Activity Dataset: What to collect?

53 Students
↓
Visit their own grand-parents (Age > 65)
↓
Observing and recording activities
↓
Selecting the most frequent 55 actions

Method	**Goal**	Select most frequent activities of older people
	How	Observing one day of older people
	Participants	53 Elderly People (age > 65)
	Dates	2017-06-15 ~ 2017-07-05
Result	**No. activities**	245
	Frequent activities	1. Watching TV 2. Meal-related activities (eating, preparing foods, washing dishes) 3. Defecation (using toilet) 4. Phone call 5. Taking medications 6. Washing face and brushing teeth 7. Wearing and taking off clothes
	Frequent objects	Mobile phone, Remote, Eyeglasses, Beds, Medicine, Cups

55 daily activities of the elderly

Category	ID	Activities
Foods	1	eating food with a fork
	2	pouring water into a cup
	3	taking medicine
	4	drinking water
	5	putting food in the fridge/taking food from the fridge
	6	trimming vegetables
	7	peeling fruit
	8	using a gas stove
	9	cutting vegetable on the cutting board
Clothing	10	brushing teeth
	11	washing hands
	12	washing face
	13	wiping face with a towel
	14	putting on cosmetics
	15	putting on lipstick
	16	brushing hair
	17	blow drying hair
	18	putting on a jacket
	19	taking off a jacket
	20	putting on/taking off shoes
	21	putting on/taking off glasses
Housework	22	washing the dishes
	23	vacuumming the floor
	24	scrubbing the floor with a rag
	25	wipping off the dinning table
	26	rubbing up furniture
	27	spreading bedding/folding bedding
	28	washing a towel by hands
	29	hanging out laundry

Category	ID	Activities
Leisure	30	looking around for something
	31	using a remote control
	32	reading a book
	33	reading a newspaper
	34	handwriting
	35	talking on the phone
	36	playing with a mobile phone
	37	using a computer
	38	smoking
Health	39	clapping
	40	rubbing face with hands
	41	doing freehand exercise
	42	doing neck roll exercise
	43	massaging a shoulder oneself
Interpersonal Communication	44	taking a bow
	45	talking to each other
	46	handshaking
	47	hugging each other
	48	fighting each other
Human-Robot Interaction	49	waving a hand
	50	flapping a hand up and down (beckoning)
	51	pointing with a finger
Etc	52	opening the door and walking in
	53	fallen on the floor
	54	sitting up/standing up
	55	lying down

Considerations on Data Acquisition

- Elderly Participants
- Real-world environments, Multi-modal, Robot vision

Systems for data acquisition: camera on moving cart (left), multiple Kinect v2 cameras (right)

Multi-Camera System in operation

Environments: Living Labs

- Real home where elderly participants are living
  - We could capture real life situations without intervention.
  - Slight interventions have been tried though.

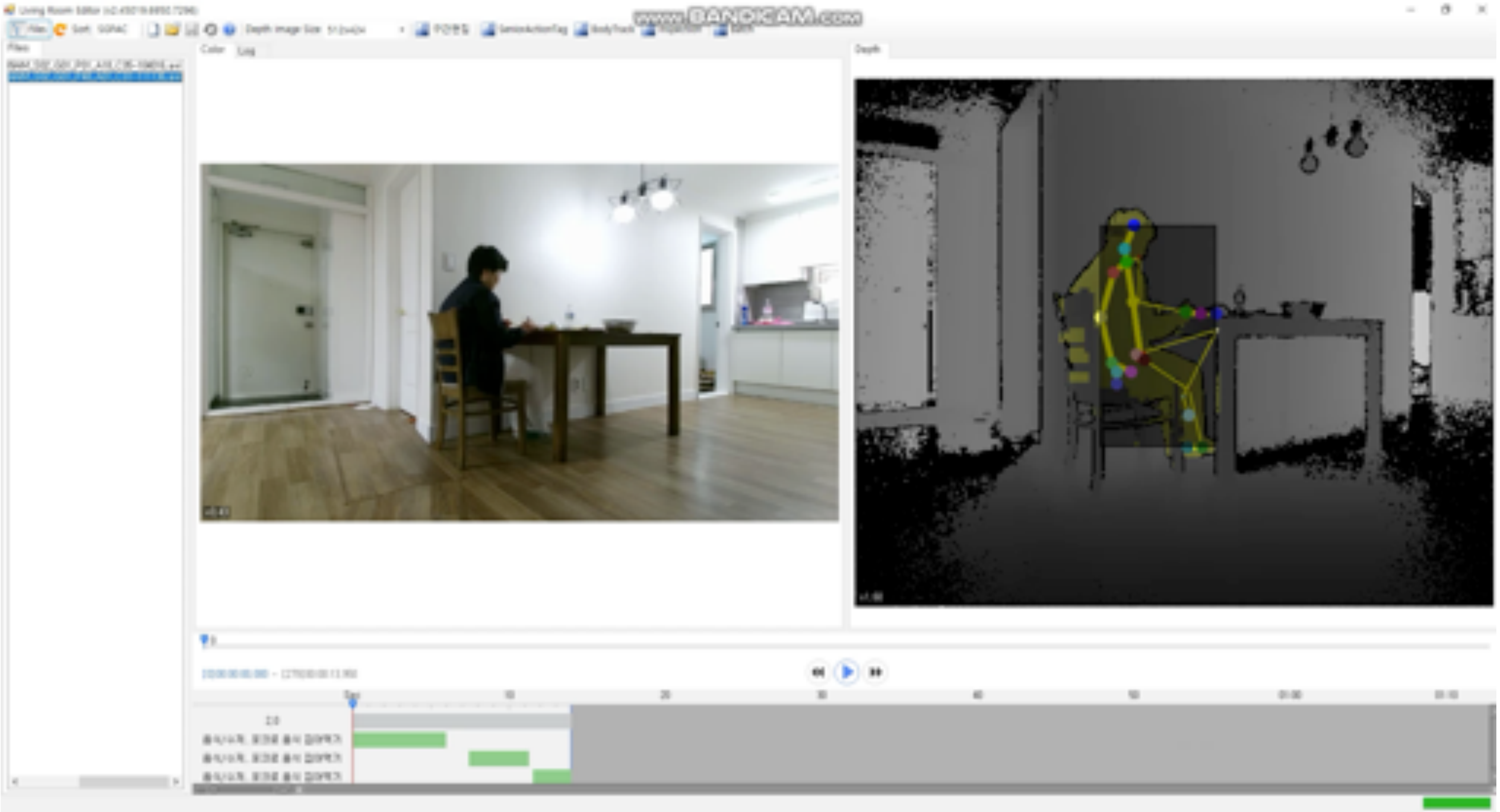
Environments: Apartment Testbed

- An apartment house for data collection and experiments
  - Daily activities intentionally performed by participants
  - Multiple RGB-D cameras for 8 different viewpoints

Data Acquisition at the Livings Labs

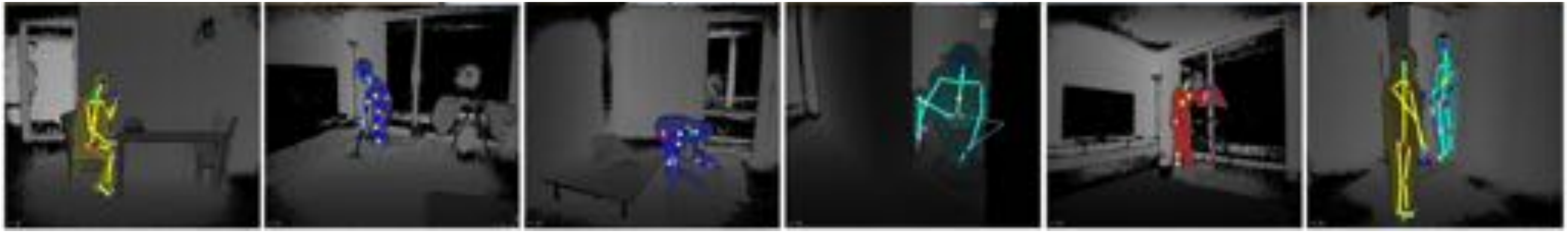
Data Acquisition at the Testbed

Annotations and Validation



ETRI-Activity3D Dataset

- Data acquisition environment: Test-bed
- Data format: RGB-DS video clips
- Participants: 50 older adults + 50 young adults
- Samples: 112,620 trimmed videos of 55 activities



ETRI-Activity3D is...

- The first **large-scale multi-modal elderly** activity dataset

Datasets	#Samples	#Sub	#Act	Modalities
RGBD-HuDaAct [3]	1,189	30	13	RGBD
MSRDailyActivity3D [4]	320	10	16	RGBDS
Act4^2 [5]	6,844	24	14	RGBD
CAD-120 [6]	120	4	10+10	RGBDS
Office Activity [7]	1,180	10	20	RGBD
UWA3D Multiview II [8]	1,075	10	30	RGBDS
NTU RGB+D [9]	56,880	40	60	RGBDSI
NTU RGB+D 120 [10]	114,480	106	120	RGBDSI
Toyota Smarthome [11]	16,129	18	31	RGBDS
ETRI-Activity3D	**112,620**	**100**	**55**	RGBDS

ETRI-Activity3D Availability

ETRI-Activity3D: A Large-Scale RGB-D Dataset for Robots to Recognize Daily Activities of the Elderly

Jinhyeok Jang, Dohyung Kim*, Cheonshu Park, Minsu Jang, Jaeyeon Lee, Jaehong Kim

Jang, J., Kim, D., Park, C., Jang, M., Lee, J., & Kim, ETRI-Activity3D: A Large-Scale RGB-D Dataset for Robots to Recognize Daily Activities of the Elderly. IROS 2020.

- Available at: https://ai4robot.github.io/etri-activity3d

ETRI-Activity3D extension coming this year…

- Data acquisition environment: Living Lab
- Data format: RGB-DS video clips
- Participants: 30 living labs
- Samples: 150 hours of untrimmed videos

Synthetic Dataset Generation Platform

Virtual Home Robot Environment

Parameter Variations →

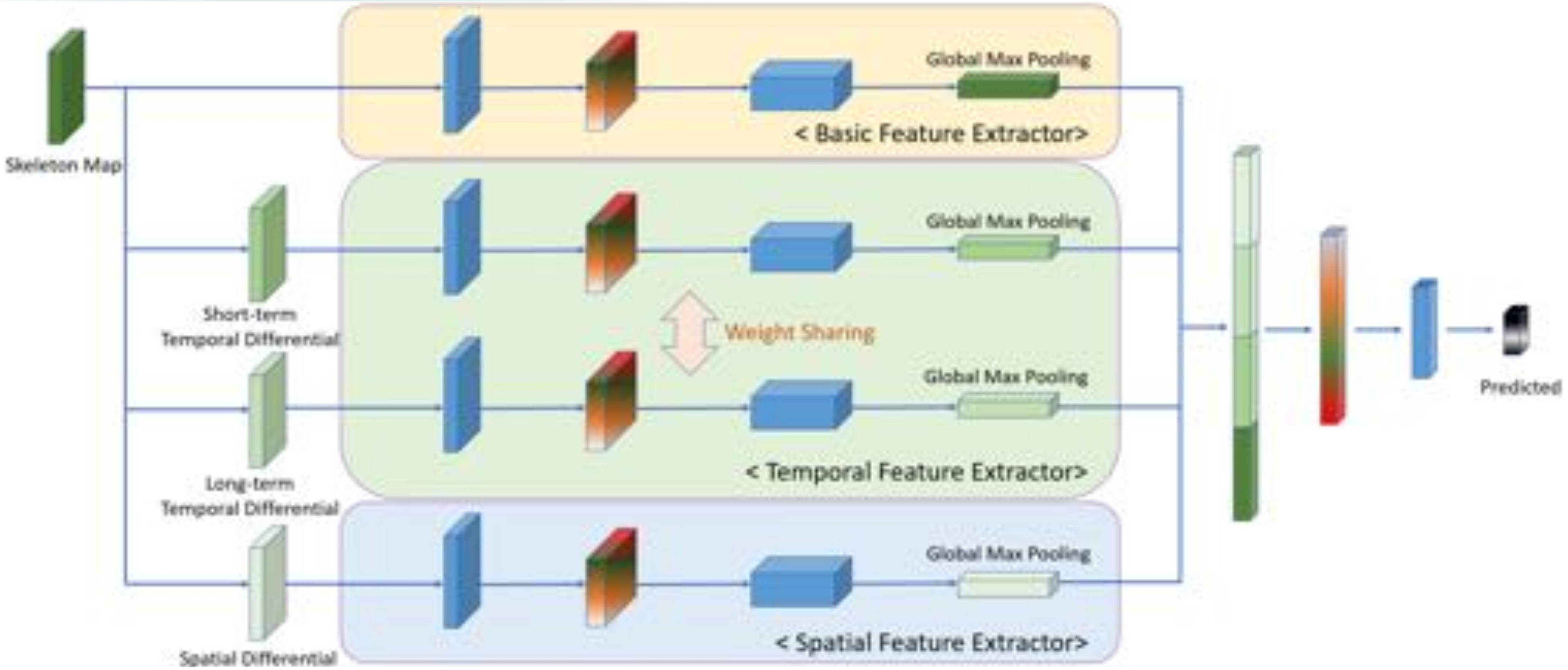
Large-scale Synthetic Human, Activity and Environment Data ←

← Robot Model Demonstrations Interaction

→ Robot AI Trained

You can generate infinite variations and scenarios

Elderly Daily Activity Recognition: FSA-CNN



- *Jang, Jinhyeok, Hyunjoong Cho, Jaehong Kim, Jaeyeon Lee, and Seungjoon Yang. "Deep neural networks with a set of node-wise varying activation functions." Neural Networks (2020)*
- *Jang, J., Kim, D., Park, C., Jang, M., Lee, J., & Kim. " ETRI-Activity3D: A Large-Scale RGB-D Dataset for Robots to Recognize Daily Activities of the Elderly. " IROS 2020. (2020) (accepted)*

Performance of FSA-CNN

Method	NTU RGB+D		ETRI-Activity3D
	CS (%)	CV (%)	CS (%)
IndRNN [18]	81.8	88.0	73.9
Beyond Joint [17]	79.5	87.6	79.1
SK-CNN [14]	83.2	89.3	83.6
ST-GCN [20]	81.5	88.3	86.8
Motif ST-GCN [21]	84.2	90.2	89.9
Ensem-NN [16]	85.1	91.3	83.0
MANs [19]	83.0	90.7	82.4
HCN [15]	86.5	91.1	88.0
FSA-CNN	**88.1**	**92.2**	**90.6**

Activities of the Elderly vs. Young

	Average activity length (sec)	Motion magnitude per time
Elderly	13.35	16.79
Young	9.45	20.28

Test data / Training data	$\text{TestData}_{\text{elderly}}$	$\text{TestData}_{\text{young}}$
$\text{TrainingData}_{\text{elderly}}$	87.69	68.99
$\text{TrainingRData}_{\text{young}}$	74.87	85.00
$\text{TrainingRData}_{\text{mixed}}$	84.78	82.05

"Is it plausible that activity patterns of elderly people are very different from those of young adults?" "Yes, maybe…"

Speech Recognition for the Elderly

- A large-scale – 400 hours of – Korean speech dataset
- Collected entirely from older adults
- Dialog Speech + Read Speech

Data Collection: Dialog Speech

- Conversations between a visiting social worker and an elderly living alone
- Recordings made with smartphones
  - Varying audio quality
  - Frequent environmental noises

Dialog Speech Data: Original Raw Data

- 873 hours, 3,381 participants, 12 regions

Region(R)	No. Participants	Len. (hrs)
Seoul-si(SE)	620	122
Busan-si(PS)	242	90
Daegu-si(DG)	202	33
Gwangju-si(GJ)	179	63
Daejeon-si(DJ)	275	66
Ulsan-si(WS)	80	28
Goyang-si(GG)	335	69
Gangwon-do(GW)	178	45
Chungcheongbuk-do(CB)	252	92
Chungcheongnam-do(CN)	317	46
Jeollanam-do(JN)	323	103
Gyeongsangbuk-do(GB)	378	116
Total	3,381	873

Dialog Speech Data: Post-Processing

- Quality Assurance
  - Speech segments inaudible or incomprehensible by human listeners were removed
- Screening
  - Every dialog including sensitive personal information were removed
- Transcription
  - An audio file was transcribed into a text file

Dialog Speech Data: Participants

- 1,170 participants, 79 years old in average

Region($\mathbb{R}$)	No. Participants	Age (μ/σ)
Seoul-si(SE)	251(F:210,M:41)	78.98/5.13
Daegu-si(DG)	108(F:95,M:13)	80.33/6.08
Gyoungki-do(GG)	110(F:83,M:27)	80.17/5.41
Chungcheongnam-do(CN)	6(F:6,M:0)	77.00/3.69
Jeollanam-do(JN)	70(F:56,M:14)	80.76/4.90
Busan-si(PS)	160(F:137,M:23)	78.70/5.51
Daejeon-si(DJ)	96(F:72,M:24)	78.81/5.24
Gangwon-do(GW)	109(F:94,M:15)	80.07/5.50
Gyeongsangbuk-do(GB)	98(F:95,M:3)	80.87/4.48
Gwangju-si(GJ)	87(F:70,M:17)	79.39/5.77
Chungcheongbuk-do(CB)	17(F:17,M:0)	80.47/5.51
Ulsan-si(WS)	58(F:49,M:9)	76.97/4.48
Total	1,170(F:984,M:186)	79.47/5.37

Dialog Speech Data: Statistics

- 300 hours, 15.4 minutes per a session in average

Region($\mathbb{R}$)	Len.(secs)	Len.(μ/σ)
Seoul-si(SE)	151,010	601.63/239.83
Daegu-si(DG)	60,740	562.42/228.14
Gyoungki-do(GG)	107,935	981.23/357.19
Chungcheongnam-do(CN)	5,193	865.62/293.98
Jeollanam-do(JN)	81,767	1,168.10/294.85
Busan-si(PS)	200,207	1,251.30/255.85
Gangwon-do(GW)	95,420	875.42/158.18
Daejeon-si(DJ)	123,138	1,282.70/293.83
Gyeongsangbuk-do(GB)	71,175	726.28/308.80
Gwangju-si(GJ)	92,699	1,065.52/276.53
Chungcheongbuk-do(CB)	20,135	1,184.41/309.54
Ulsan-si(WS)	70,754	1,219.90/254.43
Total	1,080,179	923.23/380.17

Dialog Speech Data: Data Formats

- Audio Data

Property	Value
Format.	PCM
Format Settings	Little/Signed
Codec ID	1
Bit Rate Mode	Constant
Bit Rate.	256
Channel(s)	1
Sampling Rate	16 kHz
Bit Depth	16 bits

Data Collection: Read Speech

- Pre-selected sentences were read by older adults
- Recordings made with a dedicated tablet app with on-line validation
  - Good quality overall
  - But, frequent mistakes by participants

Read Speech Data: Statistics

- 104 participants, 5 regions
- 111,814 sentences, 100 hours

Region(G)	No. Persons	No. Sent.	Len.(μ/σ)
Gyeongsangnam-do(GB)	20	22,575	3.18/1.38
Seoul-si(SE)	18	19,220	3.31/1.49
Jeollanam-do(JN)	21	21,393	3.36/1.52
Daegu-si(DG)	25	26,950	3.60/1.87
Gangwon-do(GW)	20	21,676	2.73/1.12
Total	104	111,814	3.25/1.54

Dialog Speech Data: Data Formats

- Audio Data

Property	Value
Format.	PCM
Format Settings	Little/Signed
Codec ID	1
Bit Rate Mode	Constant
Bit Rate.	705.6 kb/s
Channel(s)	1
Sampling Rate	44.1 kHz
Bit Depth	16 bits

STT Performance with VOTE400

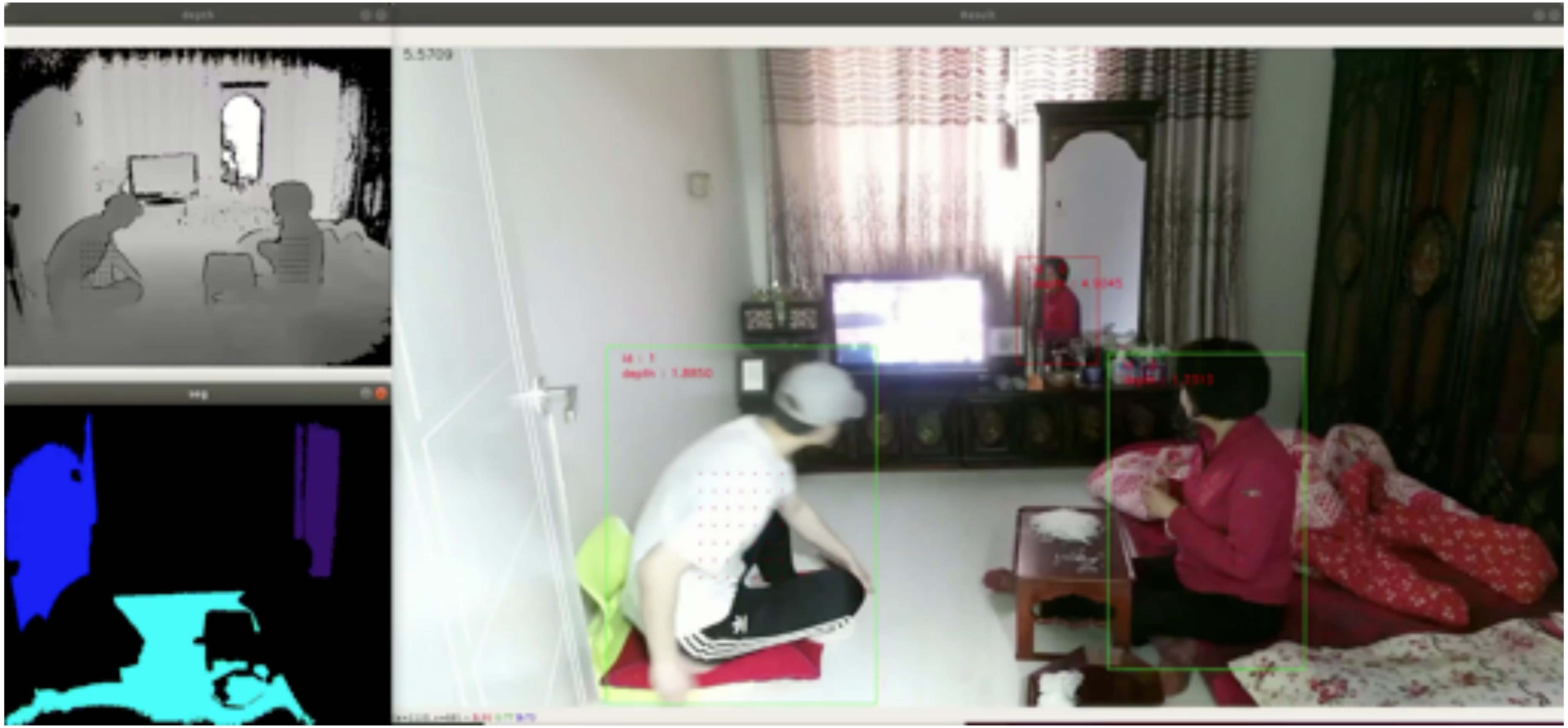
- Tested with MINDs Lab’s Baseline LSTM-based STT engine
- Fine-tuning with VOTE400 improves performance

Region	Gender	M(%)	G(%)
Seoul	Male	90	90
Seoul	Female	90	80
Gangwon	Male	80	90
Gangwon	Female	90	80
Daegu	Male	70	80
Daegu	Female	90	80
Milyang	Male	90	80
Milyang	Female	80	80
Jeonnam	Male	70	50
Jeonnam	Female	80	60
Total		83	77

✤ homepage: https://ai4robot.github.io/mindslab-etri-vote400/

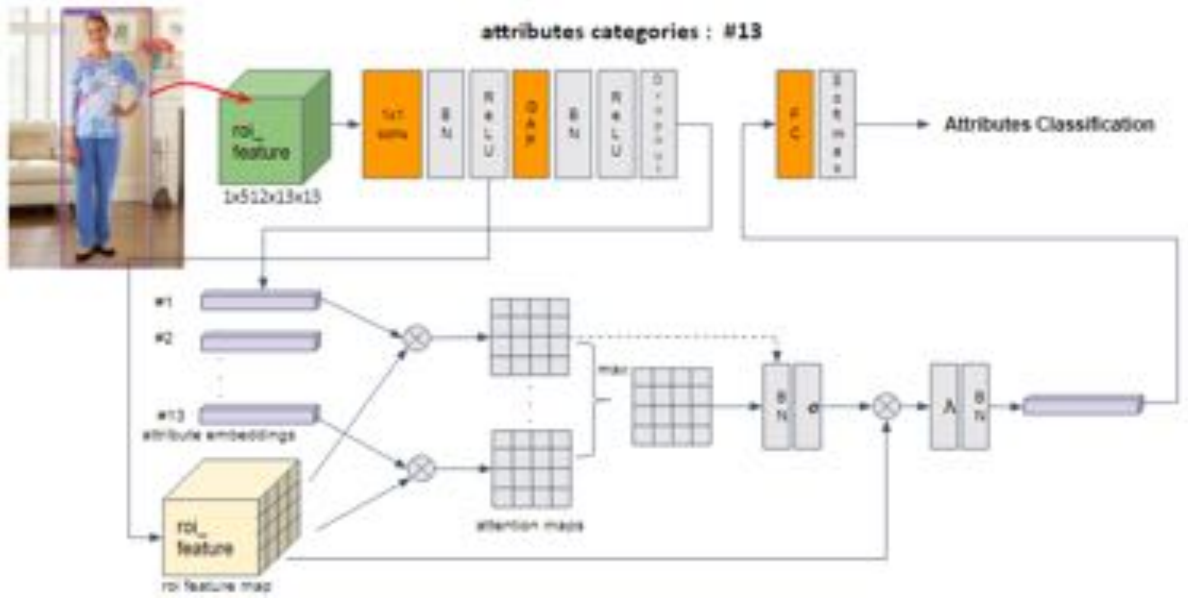
Human Detection and Tracking

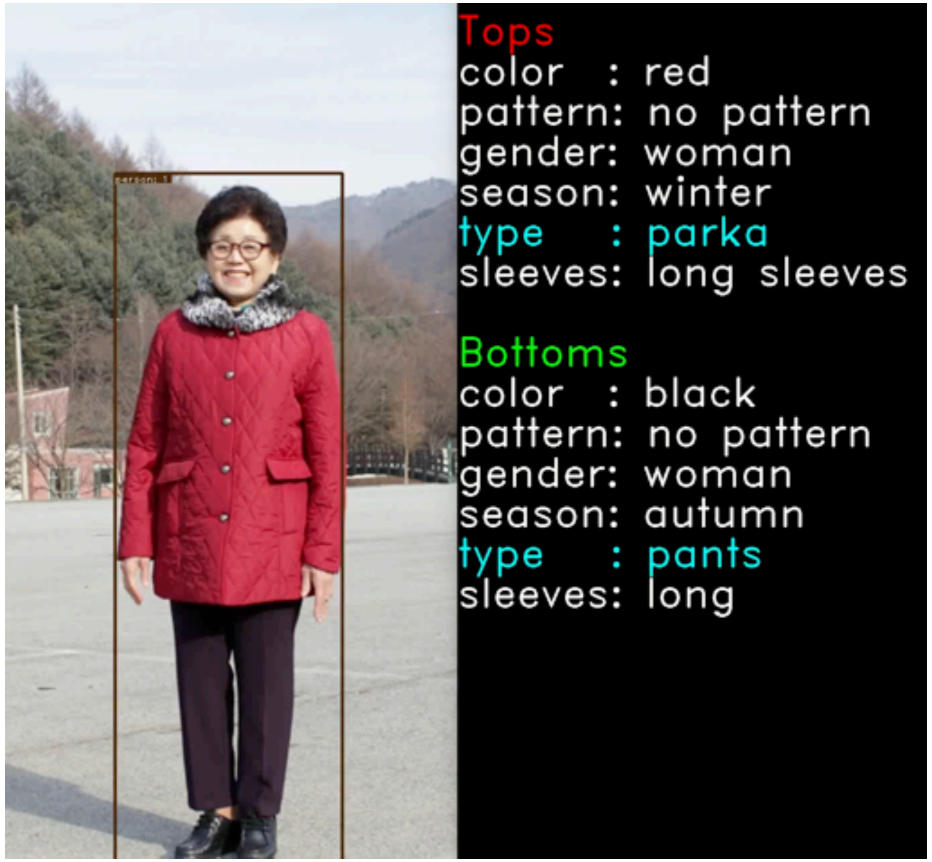
- Yolo + Online-learning for visual features in human ROIs
- Filtering out false human detections on reflective surfaces



Human Attribute Recognition

- Dataset: 35,000 elderly images with 80,000 ROIs
- 69 attributes





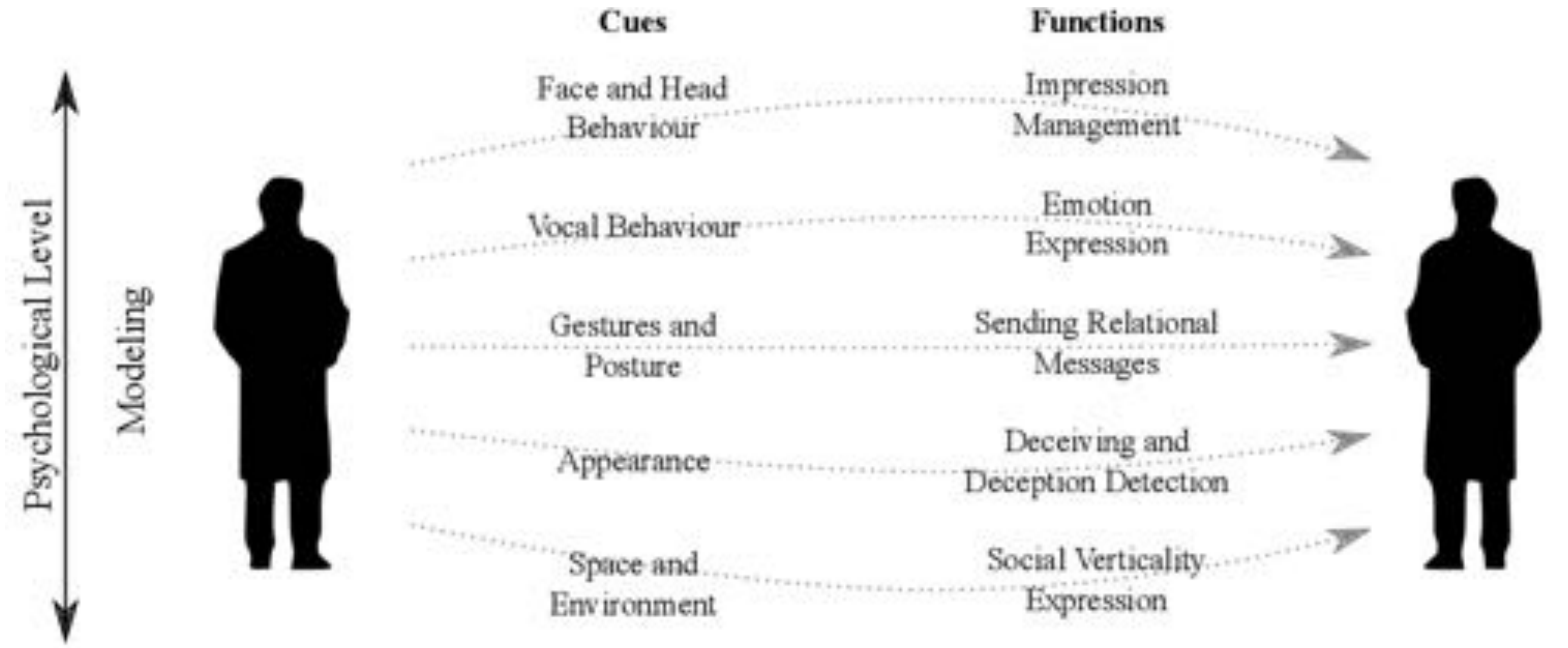
	Tops(상의)						Bottoms(하의)						
	color	pattern	gender	season	type	sleeves	color	pattern	gender	season	type	sleeves	leg_type
73.55	45.02	71.64	82.17	59.46	67.40	81.71	71.35	77.76	79.67	73.7	80.46	81.08	84.93

✣ homapage: https://github.com/ai4r/Air-Clothing-MA

Robot Social AI

Social Intelligence

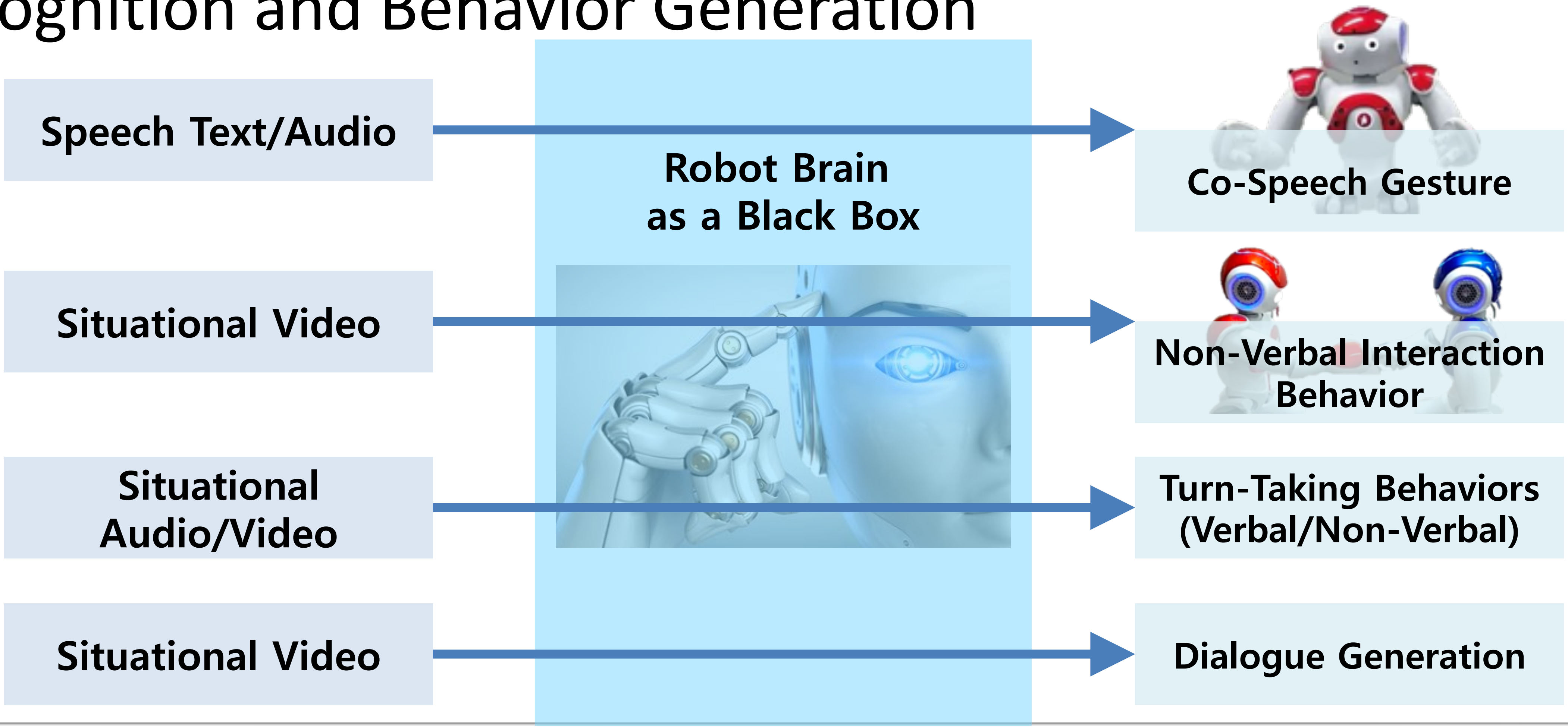
- Social Cognition and Social Behaviors



for Robots… HOW?

End-to-End Robot Social AI

- Learning from Human-Human Interaction for Social Cognition and Behavior Generation



End-to-End Robot Social AI

- Learning from Human-Human Interaction for Social Cognition and Behavior Generation

Speech Text/Audio

Situational Video

Situational Audio/Video

Situational Video

Robot Brain
as a Black Box

Co-Speech Gesture

Non-Verbal Interaction Behavior

Turn-Taking Behaviors (Verbal/Non-Verbal)

Dialogue Generation

What are Co-Speech Gestures?

- One of the key elements of social interaction

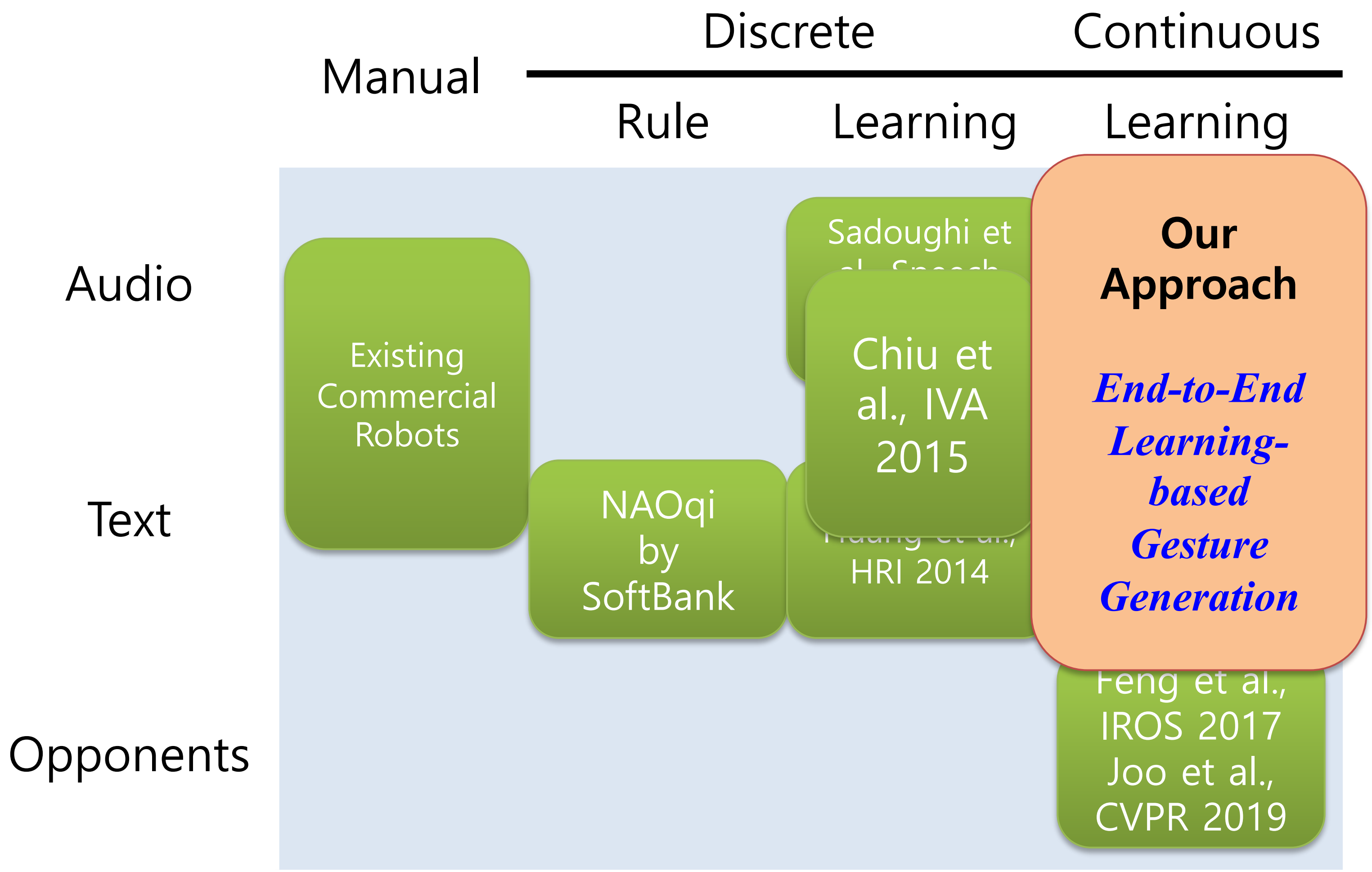
    Evaluation of Social Interaction (ESI) Assessment[1]

    – Approaches, Gaze, Conversation flow, **<u>Gesture</u>**, Facial expression, …
- More Attention[2], Help listeners comprehend[3], Human likeness

[1] Fisher, A.G. and Griswold, L.A., 2010. Evaluation of social interaction (ESI). Fort Collins, CO.
[2] Bremner, P., Pipe, A.G., Melhuish, C., Fraser, M. and Subramanian, S., 2011, October. The effects of robot-performed co-verbal gesture on listener behaviour. In *2011 11th IEEE-RAS International Conference on Humanoid Robots*.
[3] Cassell, J., McNeill, D. and McCullough, K.E., 1999. Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. Pragmatics & cognition.

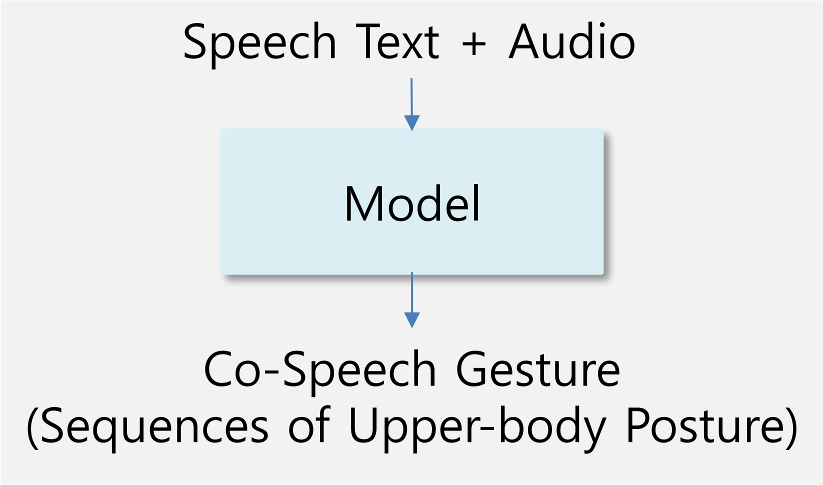
ETRI

Co-Speech Gesture Generation Methods



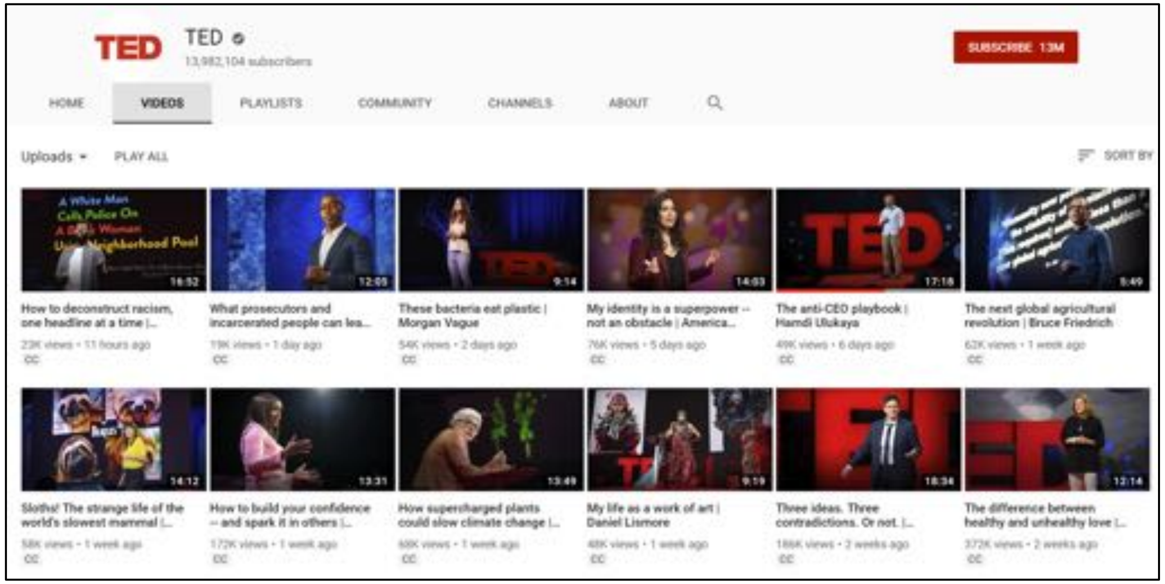
Goal

Generating natural and plausible co-speech gestures for multimodal speech context by end-to-end learning from in-the-wild videos



Data Acquisition

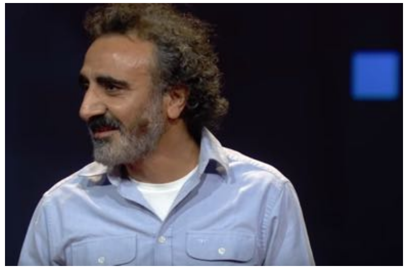
- TED Video Dataset
- First **large-scale** & **in-the-wild** dataset
- Why TED talks?
  - Large enough
  - Various speech content and speakers
  - Expect that the speakers use proper hand gestures
  - Favorable for automation of data collection and annotation



Automated Data Acquisition Pipeline

Automated Process

Download video and transcripts → Extract 2D poses → Shot filtering → Word-level transcript synchronization → Make training samples



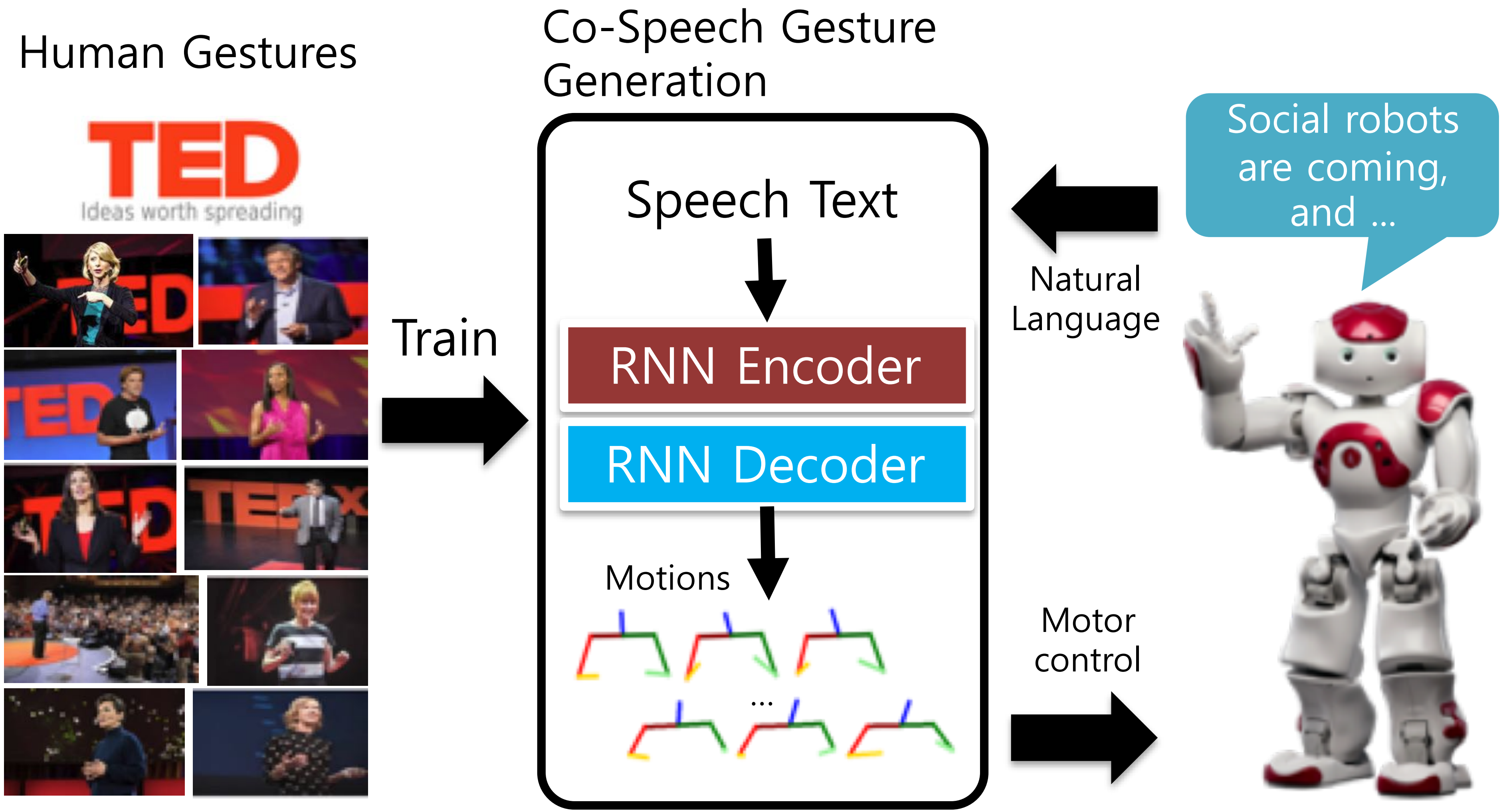
Excluded samples

Youtube TED Gesture Dataset

Number of videos	1,766
Average length of videos	12.7 min
Shots of interest	35,685 (20.2 per video on avg.)
Ratio of shots of interest	25% (35,685 / 144,302)
Total length of shots of interest	106.1 h

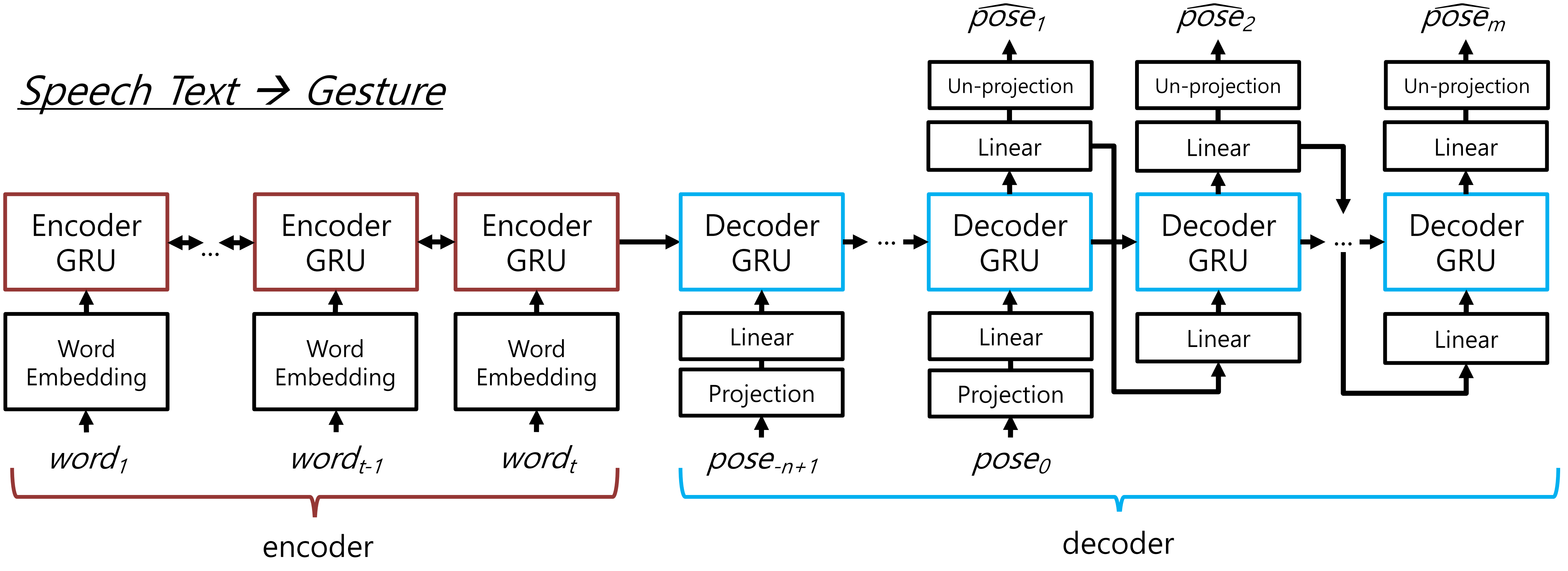
- homepage: http://ai4robot.github.io/datasets

System Architecture



Yoon, Y. et al., Robots Learn Social Skills: End-to-End Learning of Co-Speech Gesture Generation for Humanoid Robots, in the Proc. of The International Conference in Robotics and Automation (ICRA 2019).

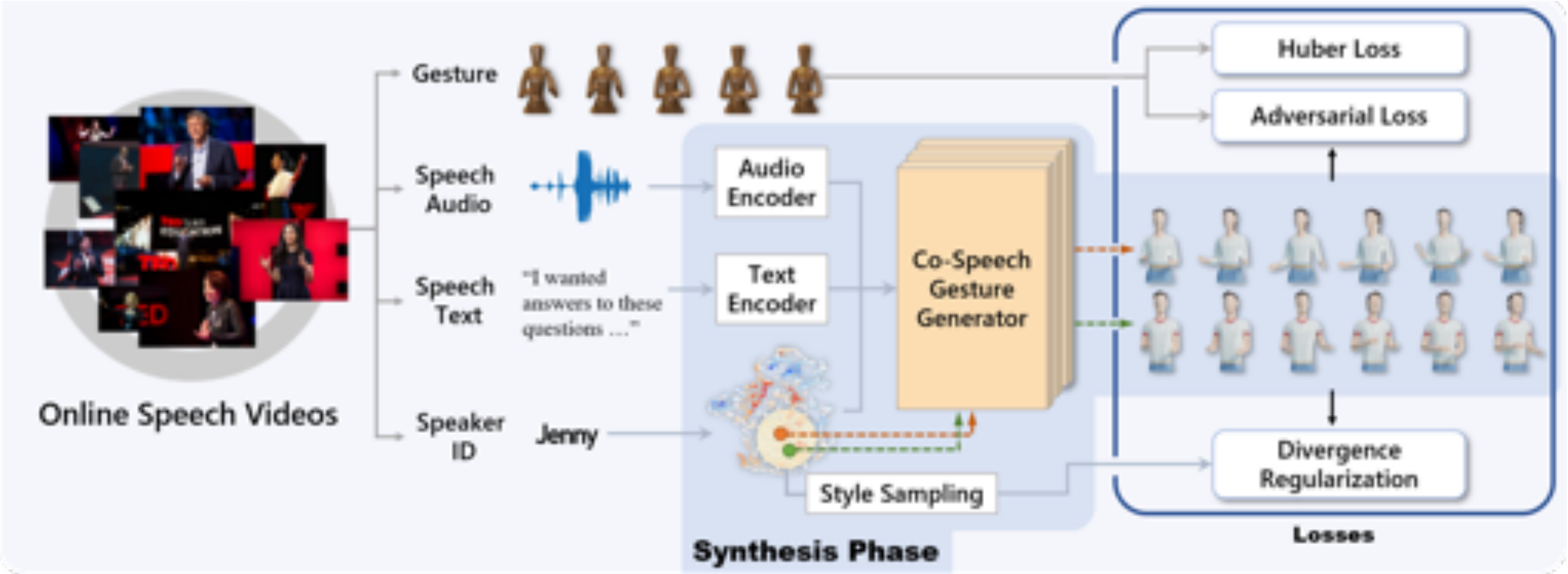
Text-to-Gesture Generation Model ('19)



Co-Speech Gesture Generation Demo ('19)

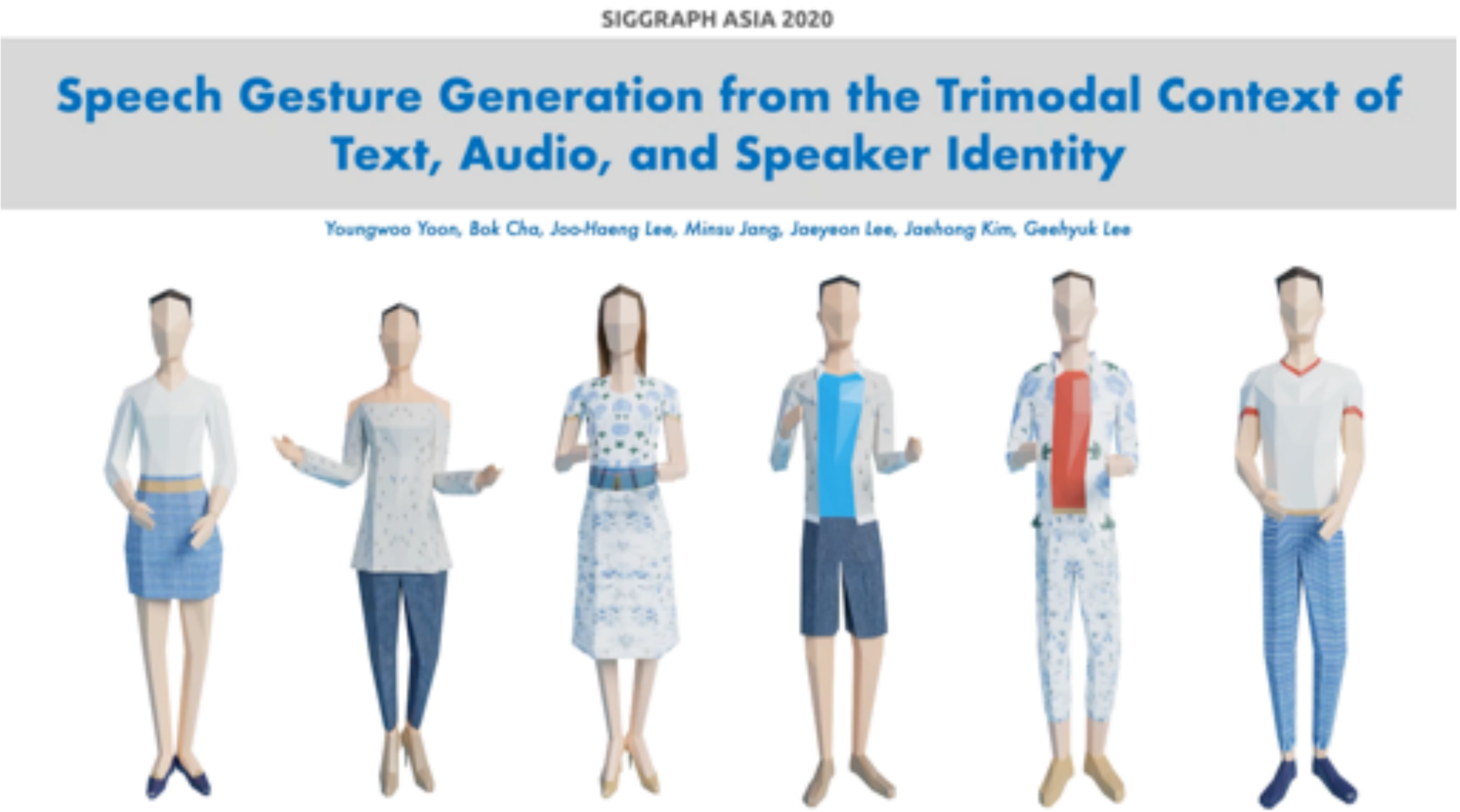
Trimodal-based Co-Speech Gesture Generation

Speech Text + Speech Audio + Speaker ID → Gesture



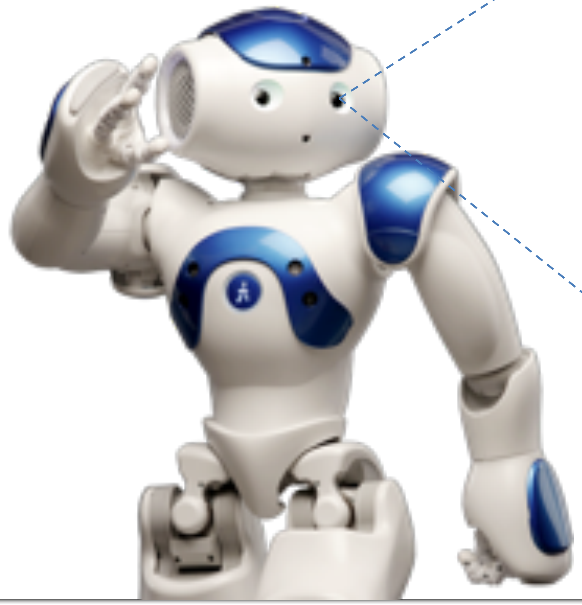
Yoon et al., "Speech Gesture Generation from the Trimodal Context of Text, Audio, and Speaker Identity." SIGRAPH ASIA 2020 (accepted)

Co-Speech Gesture Generation Demo ('20)



Act2Act: Non-Verbal Interaction Generation

**Learning to decide
when and how to perform which interaction behavior
by observing human-human interactions**



Act2Act Dataset

- Participants: 100 elderly people (age > 65)
- Data Format: RGBD-S/Robot Joint Angles Video Clips
- Samples: 7,500 sets (100 groups x 10 scenarios x 5 repetition x 3 views)

- homepage: https://ai4robot.github.io/air-act2act/

Act2Act Generation Model

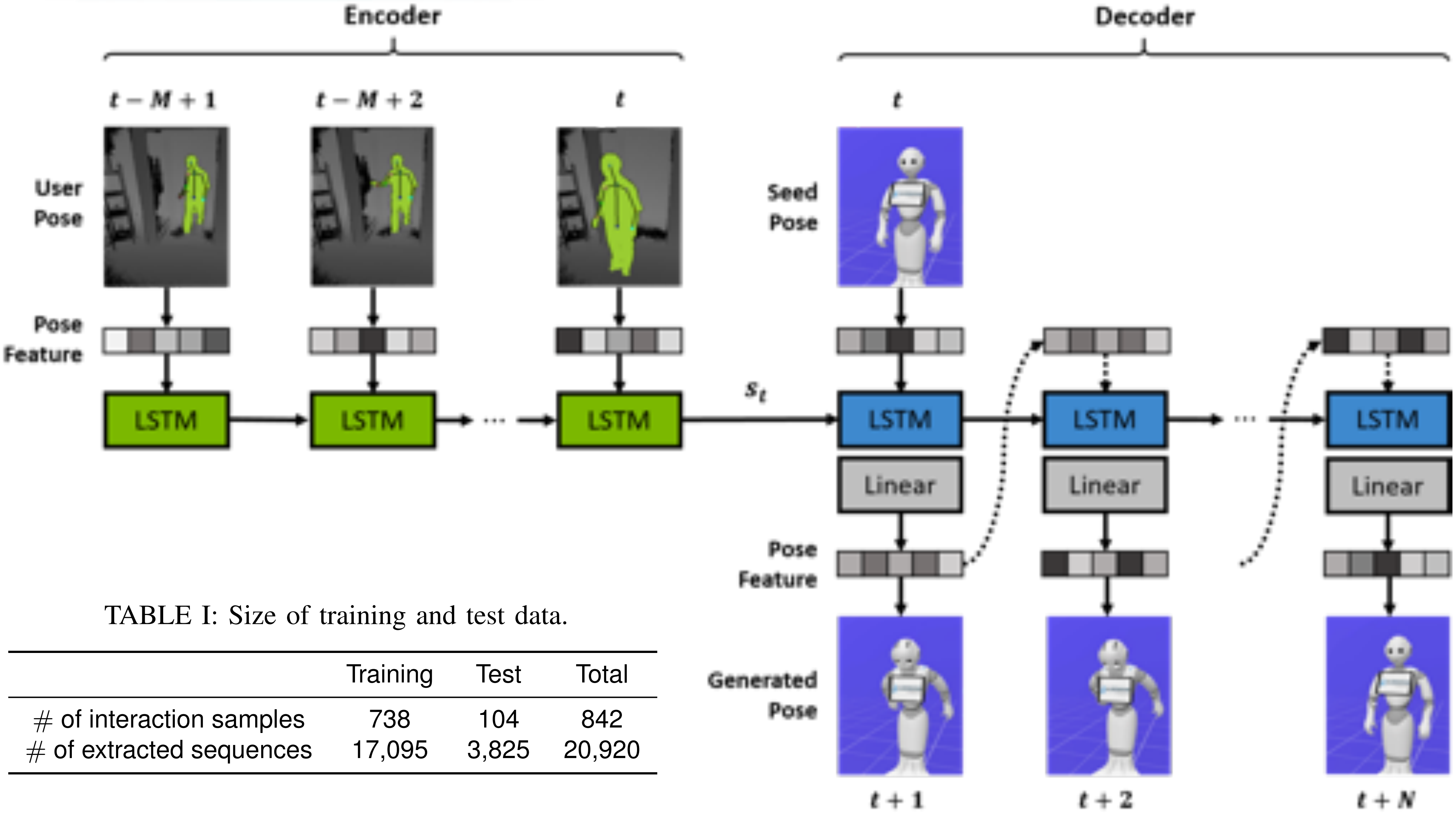


TABLE I: Size of training and test data.

	Training	Test	Total
# of interaction samples	738	104	842
# of extracted sequences	17,095	3,825	20,920

Act2Act Evaluation

TABLE II: Accuracy of behavior generation. (GT: ground truth, 1: bowing to the user, 2: staring at the user for a command, 3: shaking hands with the user, 4: stretching hands to hug the user, 5: no to all)

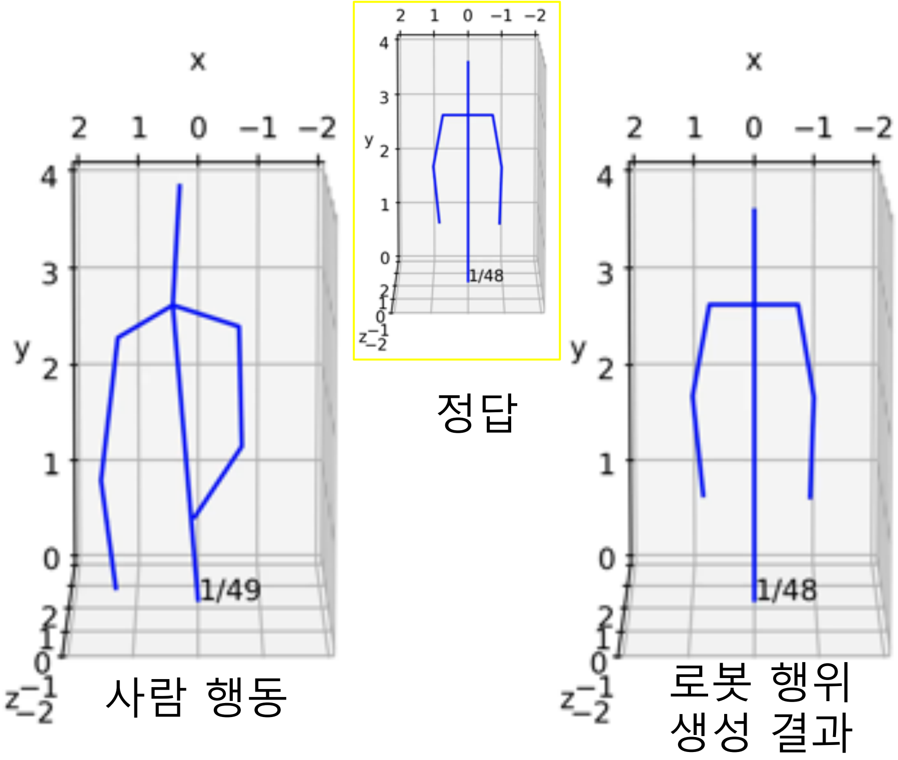
GT \ Answer	1	2	3	4	5	Total
1	**97.4**	0.0	0.0	0.0	2.6	100%
2	0.0	**85.1**	0.0	0.0	14.9	100%
3	1.8	10.5	**61.4**	0.0	26.3	100%
4	0.0	0.0	0.0	**71.9**	28.1	100%

TABLE III: Behavior satisfaction.

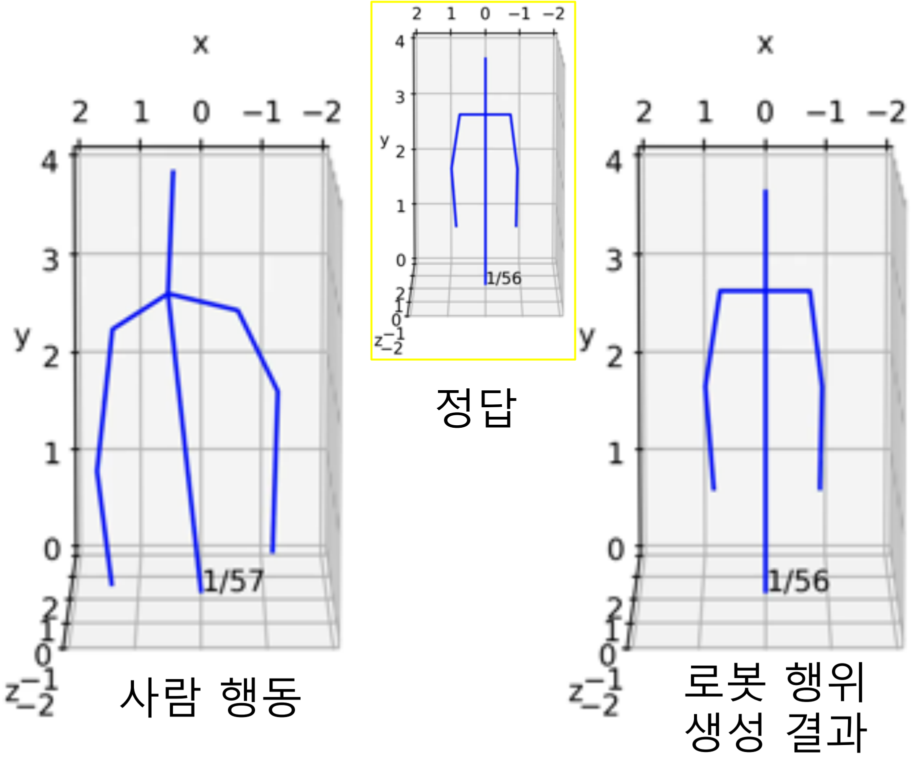
Behavior	Satisfaction
1	4.1
2	3.9
3	2.9
4	3.1

Woo-Ri Ko, Jaeyeon Lee, Minsu Jang, Jaehong Kim, "End-To-End Learning of Social Behaviors for Humanoid Robots" SMC 2020

Act2Act Demonstration

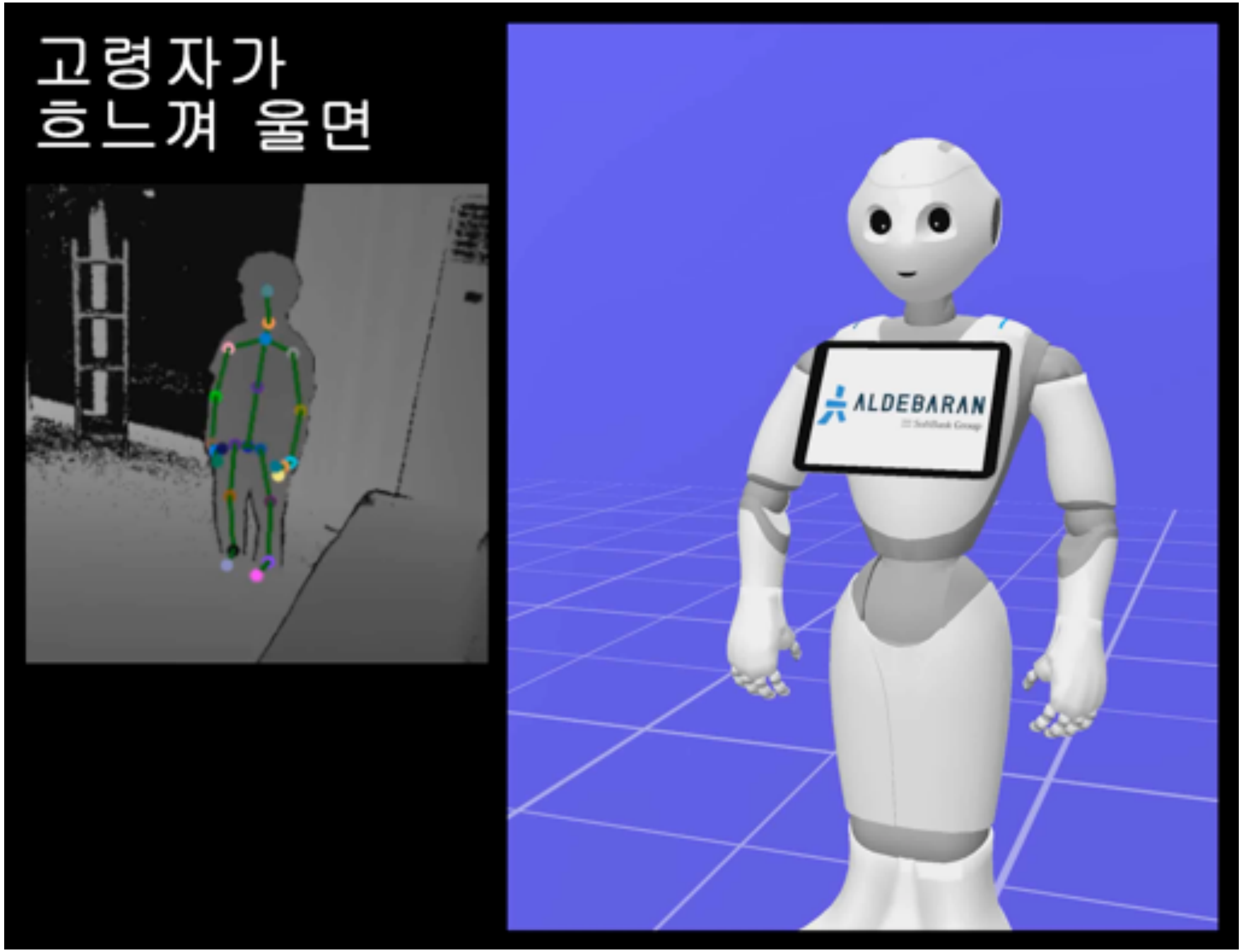


악수하기



울먹이면 안아주기

Act2Act Demonstration



Summary

Final Words…

- We are trying to build AI models and systems for elderly-care robots.
- Domain specific AI that really works in the real-world needs a lot of domain specific data collected from the real-world; we are doing it.
- You can find our results at:

https://ai4robot.github.com

https://github.com/ai4r

Thank you!

Contact: minsu jang (minsu@etri.re.kr)

AIR
Artificial Intelligence for Robots