Social Behavior Recognition for Nonverbal Human Robot Interaction

Experiments & Conclusion

HoBeom Jeon

UST-ETRI
Social Robotics Research Section

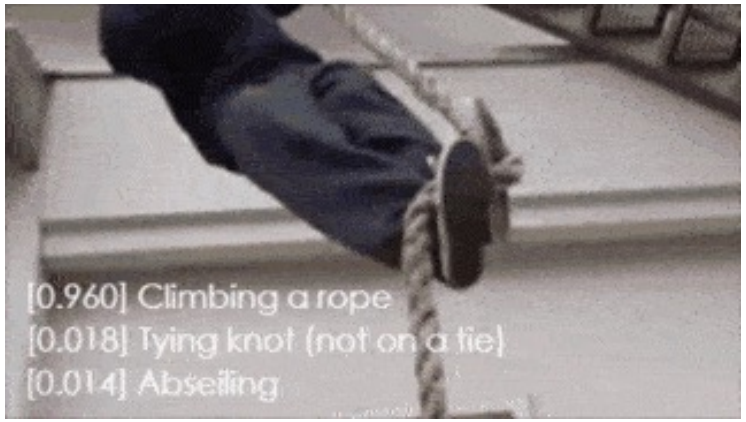
2023.11.23

Contents

Recap) What is Human Action Recognition?

HAR aims to predict the behavior of a human in a given sequence of image

1. Objective: HAR recognizes human actions in videos by analyzing sequences of images, encompassing everything from simple activities to complex tasks.
2. Applications: HAR finds utility in enhancing security measures, monitoring healthcare, optimizing sports analytics, and enhancing human-computer interaction.
3. Challenges: HAR faces challenges such as handling diverse activities and the requirement for large, annotated datasets to train effective models.



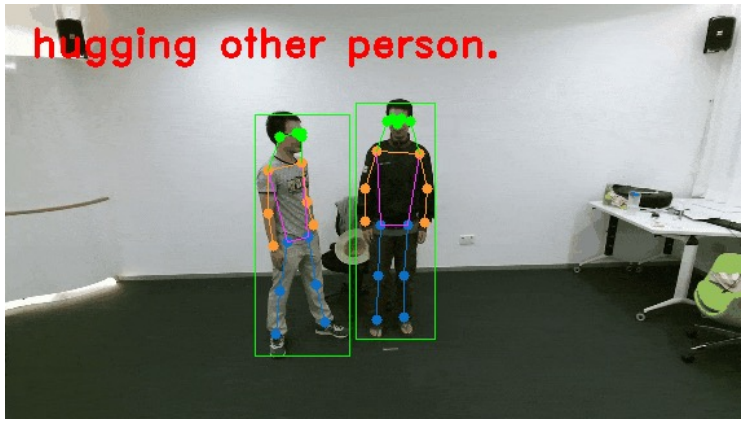


Image Source: [mmaction2]

Recap) Enhancing Human-Robot Interaction through HAR

Human–robot interaction (HRI) is the study of interactions between humans and robots

Industrial collaborative robots

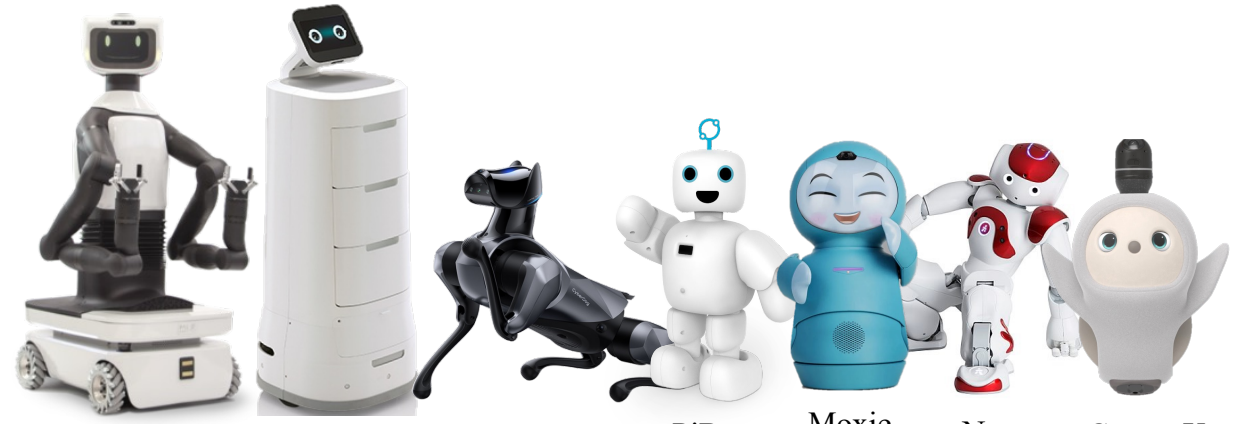
→ Mainly focuses Human's intentional movement
Ex) Grabbing heavy objects, Driving screw in

Image Source: MobileAutomation

Image Source: KUKA

Social robots

→ Mainly focuses Human's emotional expression
Ex) Hand-shaking, Hand-waving, Happy face



PAL ROBOTICS TIAGo Pro

LG CLOi

Xiaomi CyberDog2

PiBo

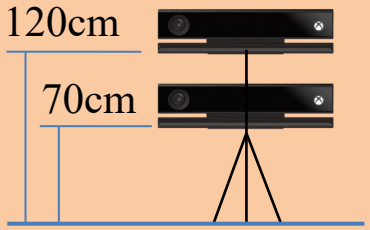
Moxie

Nao

GrooveX Lovot

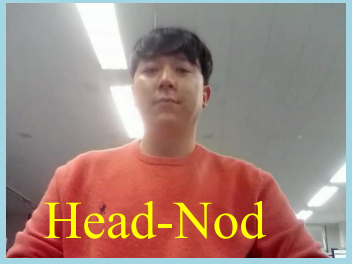
Recap) Different Viewpoint Scenario of Social Robots

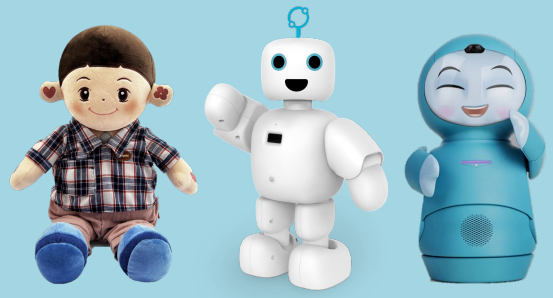
(a) Fixed camera: 3rd person robot view



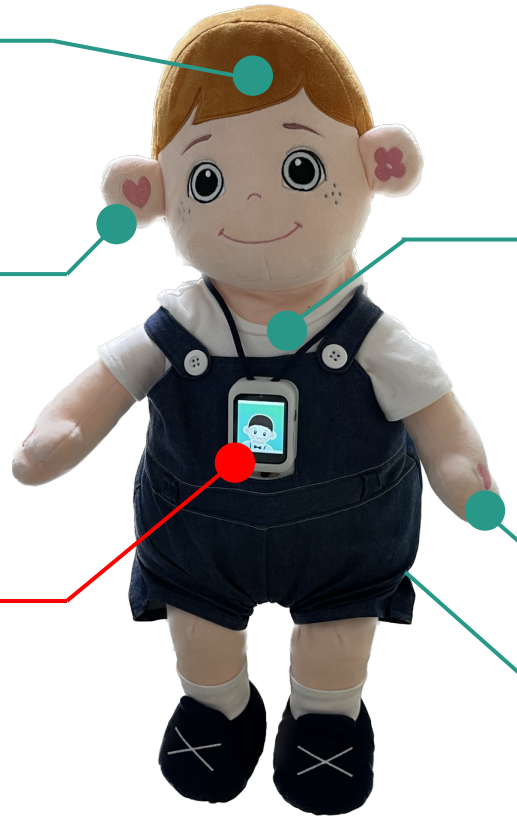
(b) Mobile humanoid robot: 1st person robot view

(c) Companion doll robot: 1st person robot view





Recap) Our Approach: The function of Doll-type Social Robot



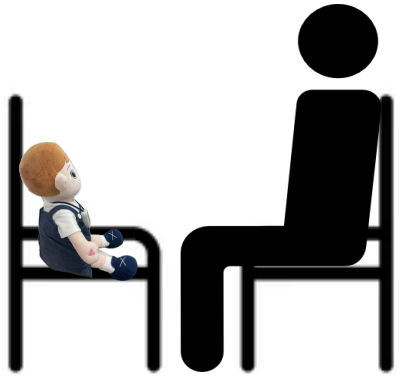
- The close proximity and narrow field of view of doll-type robots make it difficult to capture a comprehensive view of human actions.
- The intervention in camera motion caused by human touch introduces noise and further co mplicates the recognition of human behavior from a first-person robot perspective.

[7] Kim S.Y., et al. (2022) “Effectiveness of Social Robot by Family Caregivers Nursing for the Older Adult with Cognitive Impairment: A Randomized Controlled Trial”, Journal of Korean Gerontological Nursing 24.2.

Recap) Our Approach: New Interaction Dataset

→ Two type of doll robot seating positions

Table Seating View
→ Equal Eye Position



Chair Seating View
→ Equal Seat Height

Recap) Our Approach: New Interaction Dataset

Etri Social Interaction Dataset

→ A total of 1000 clips are recorded by repeating 10 individual actions 10 times. (10 subjects)

Head-nod

Pet

Hand-shake

Clap

Hug

Head-shake

Zone out

Arm cross

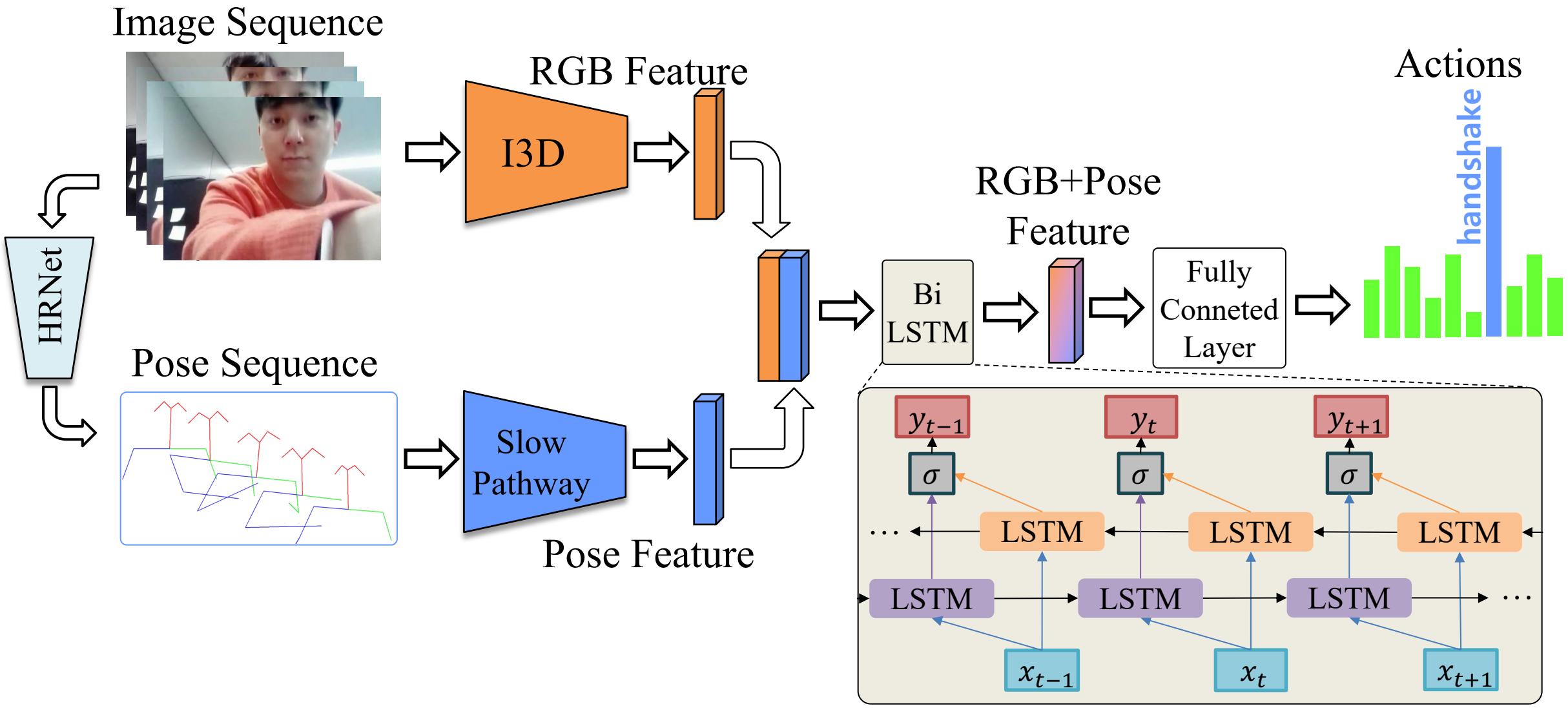
Hand wave

Punch

Thank you for your participating.

S. H. Kim, H. M. Kim, D. H. Lee, J. H. Hwang, et. al.

Recap) Our Approach: RGB + Pose Modality



Overview of our action recognition model

Recap) Our Approach: RGB + Pose Modality

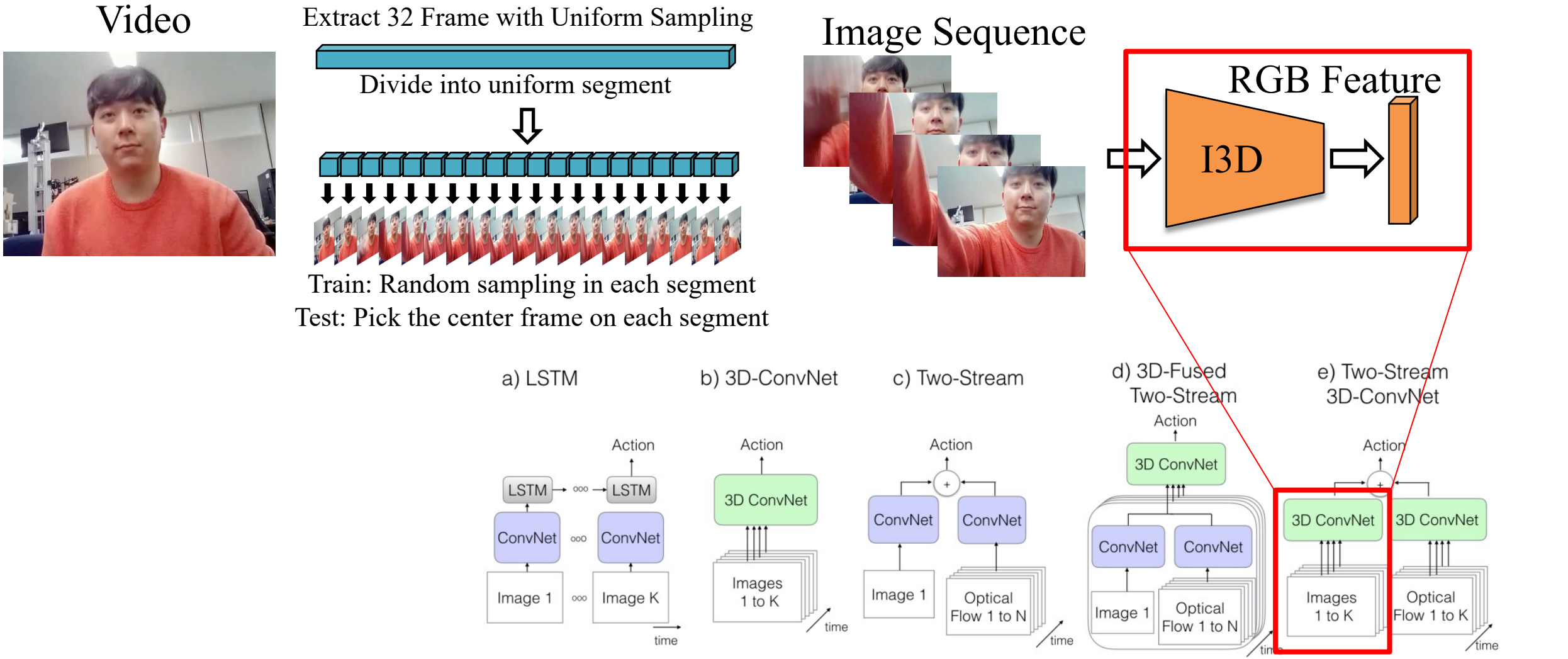
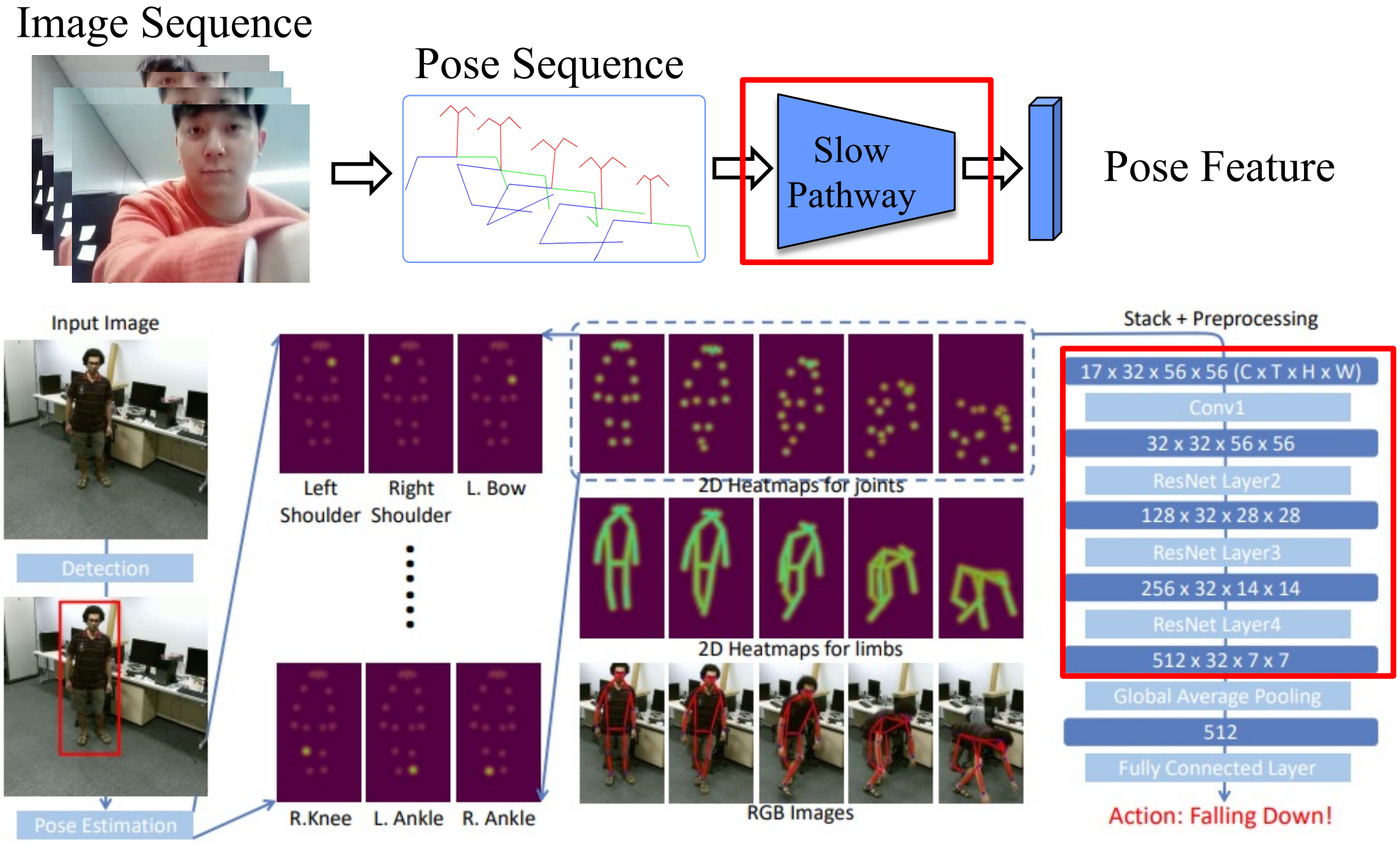


Figure 2. Video architectures considered in this paper. K stands for the total number of frames in a video, whereas N stands for a subset of neighboring frames of the video.

[8] Joao Carreira, and Andrew Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset", CVPR2017

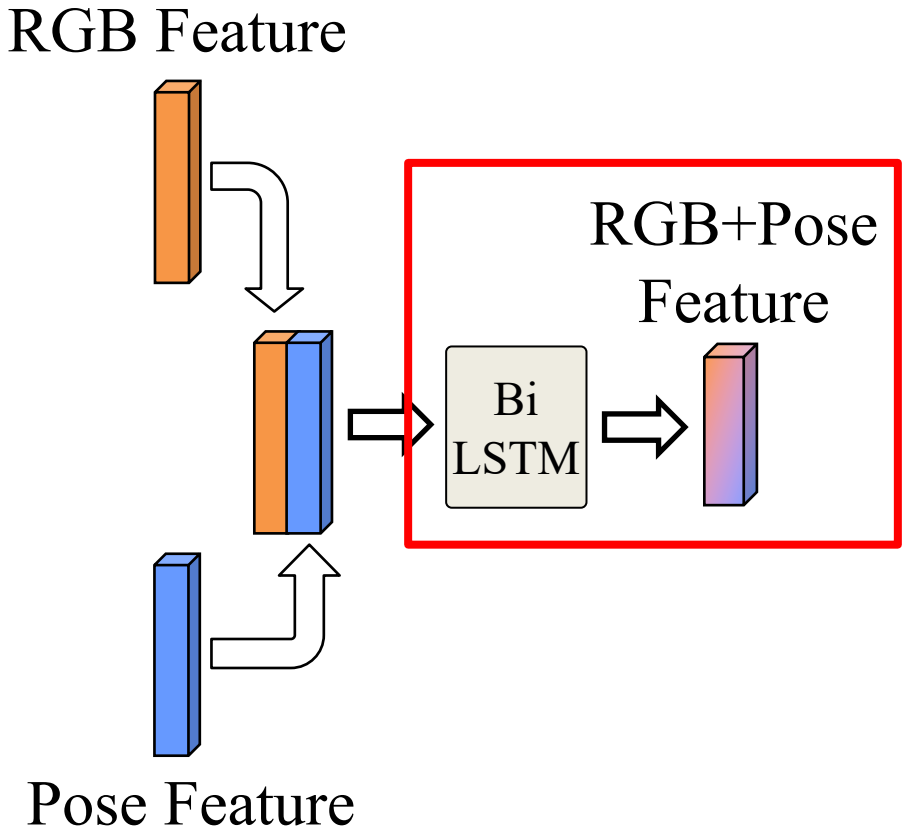
Recap) Our Approach: RGB + Pose Modality

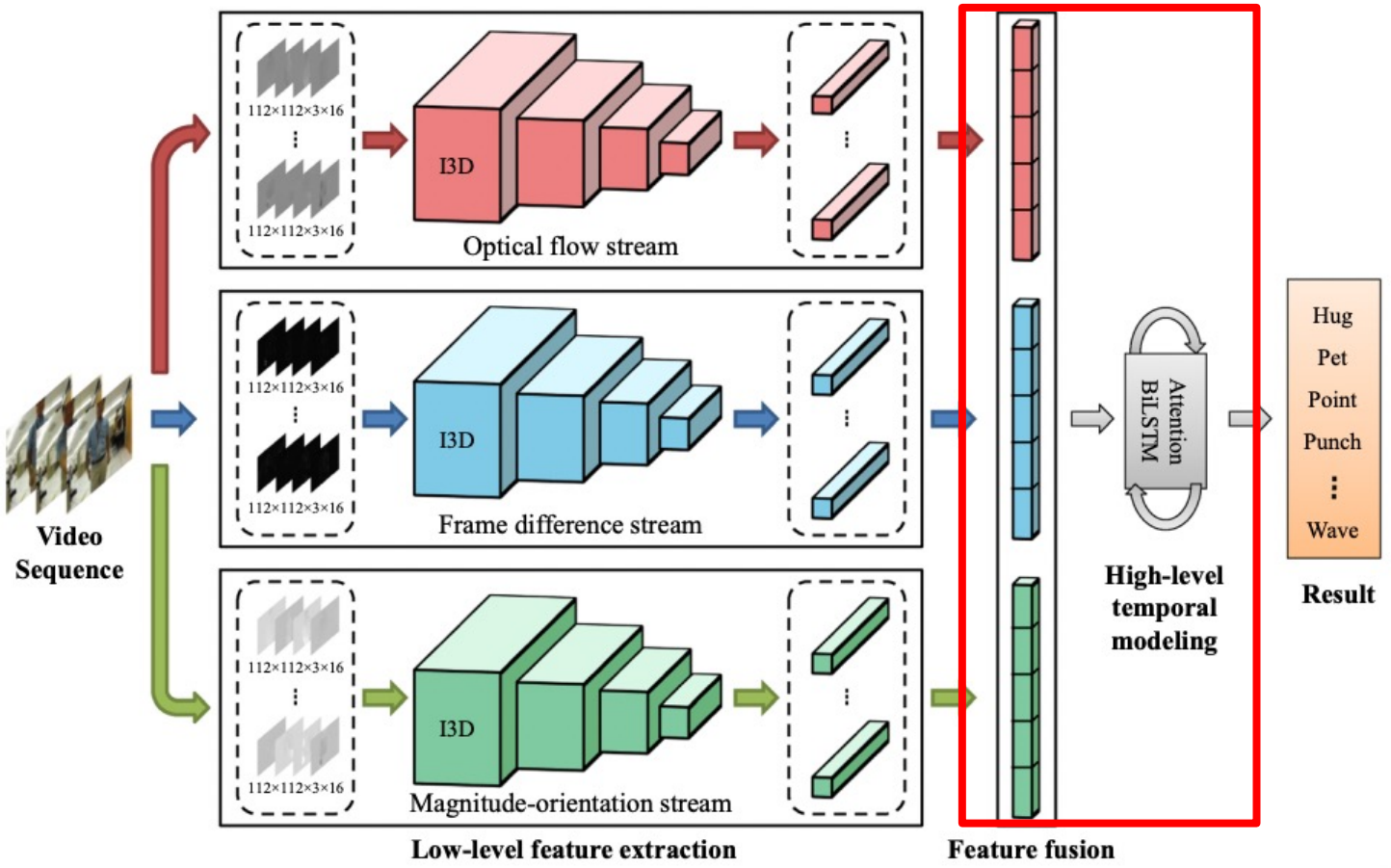


[10] SlowFast Networks for Video Recognition, ICCV2019

[11] Revisiting Skeleton-based Action Recognition, CVPR2022

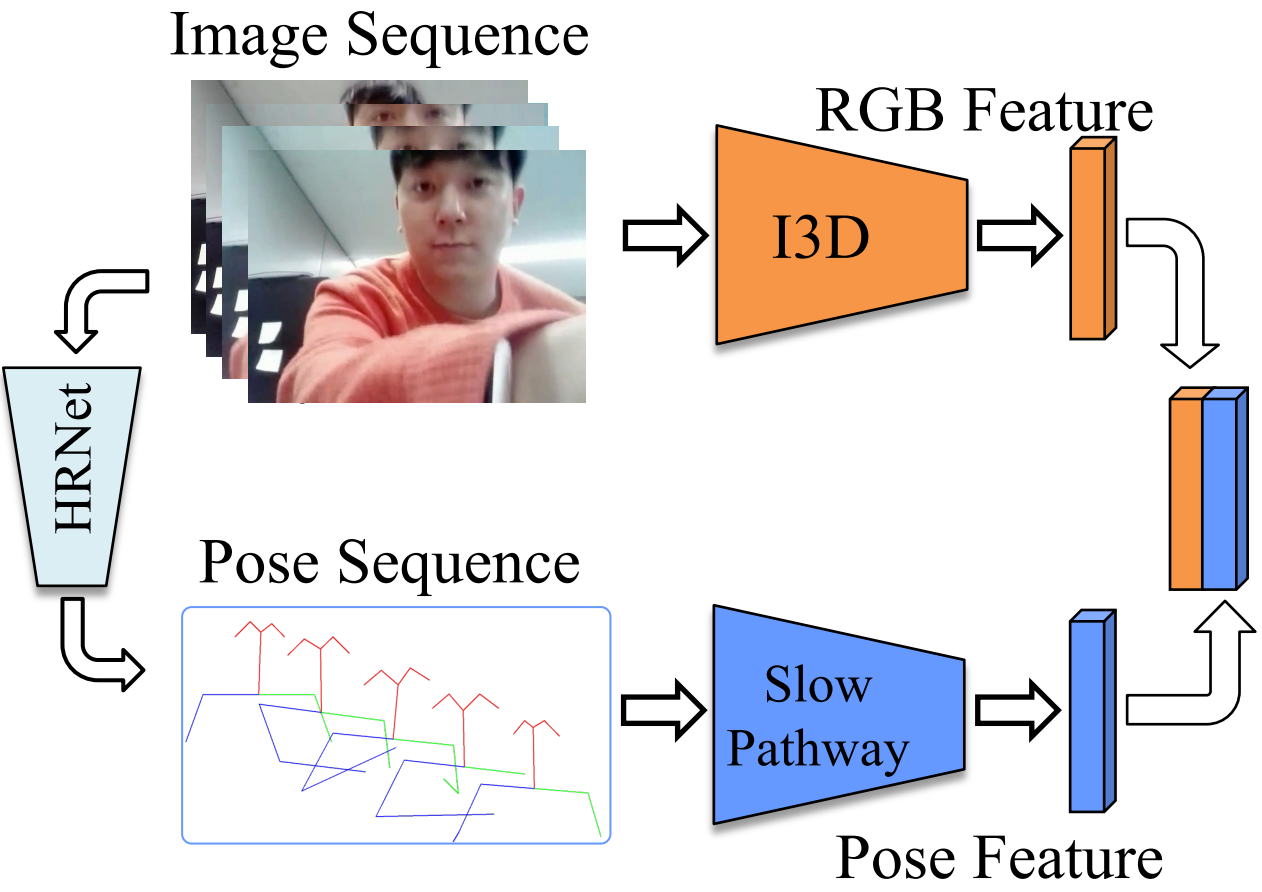
Recap) Our Approach: RGB + Pose Modality



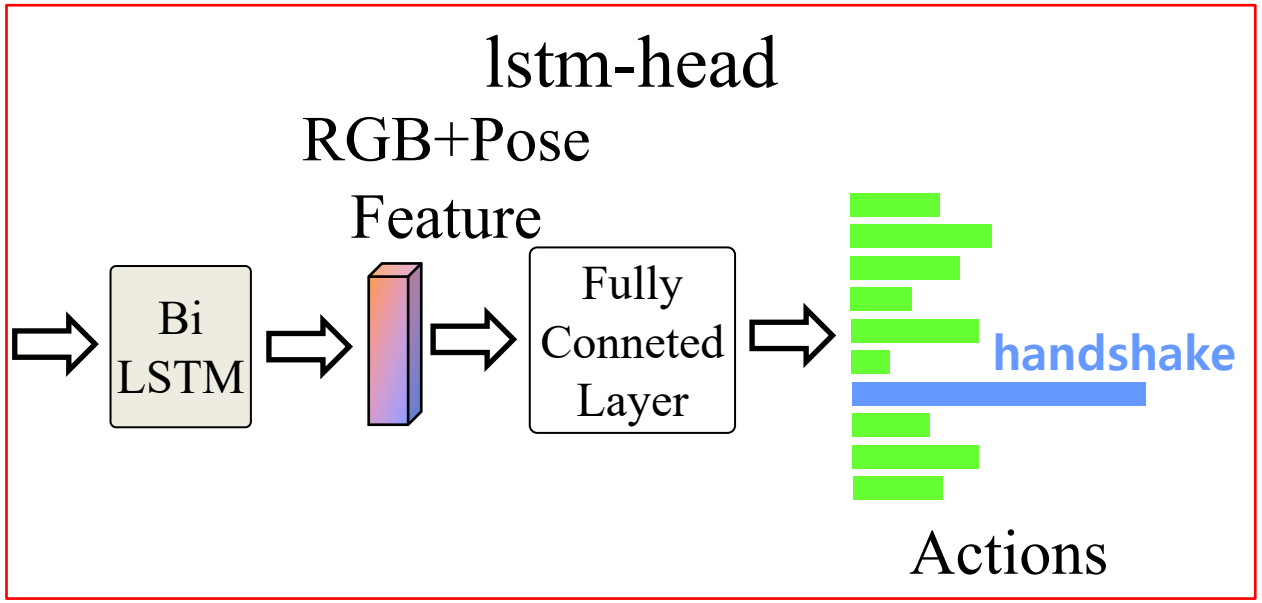


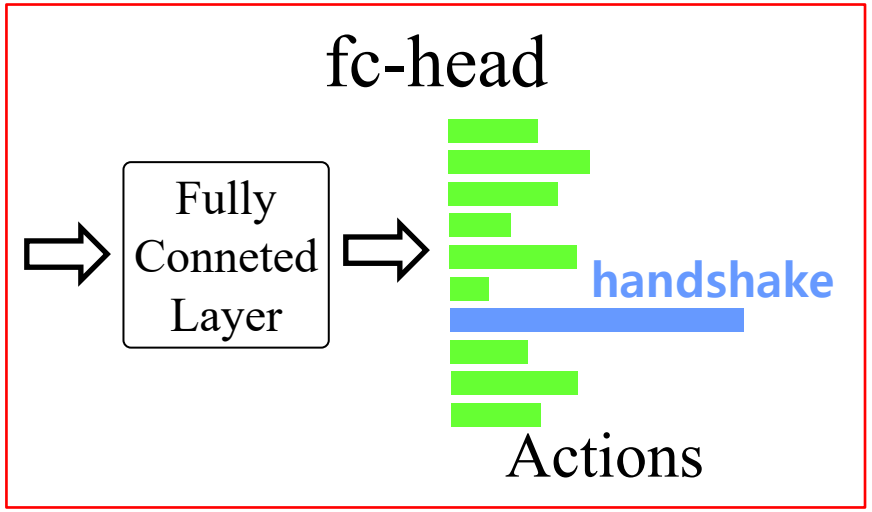
[6] Three-stream spatio-temporal attention network for first-person action and interaction recognition, Javed Imran
Journal of Ambient Intelligence and Humanized Computing (2022) 13:1137–1152

Recap) Our Approach: RGB + Pose Modality



❖ Tested with two types of heads





Experiments – JPL Dataset

First-Person Activity Recognition: What Are They Doing to Me?

M. S. Ryoo and Larry Matthies
Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA
{mryoo,lhm}@jpl.nasa.gov

Michael S. Ryoo
Stony Brook University; Robotics at Google
cs.stonybrook.edu의 이메일 확인됨 - 홈페이지
Robotics Computer Vision Machine Learning

(a) Our observer setup (b) Video snapshots from the observer

Figure 1. Picture of our setting, and its example observations obtained during a person punching it. The humanoid was placed on a rolling chair to enable its operator emulate translation movements.

[1] First-Person Activity Recognition: What Are They Doing to Me? (CVPR2013)
http://michaelryoo.com/jpl-interaction.html

Experiments – JPL Dataset

Positive interactions

Negative interactions

- First dataset recognizing interaction-level human activities from a first-person viewpoint.
- Actions are divided into friendly interactions (shake, pet, hug, wave) and hostile interactions (point, punch, throw).
- The limited dataset size, consisting of only 82 samples.

[1] First-Person Activity Recognition: What Are They Doing to Me? (CVPR2013)
http://michaelryoo.com/jpl-interaction.html

Experiments – JPL Dataset

Action	id	#Videos	#Frames			Avg sec
			Min	Max	Avg	
Hand-shaking	0	12	111	243	166.91	5.56
Hugging	1	12	**198**	512	340.75	11.35
Pet	2	12	157	485	256.33	8.54
Hand-wave	3	12	31	84	59.00	1.96
Pointing	4	12	180	**1058**	**607.50**	**20.25**
Punch	5	12	53	96	70.00	2.33
Throw Object	6	12	75	164	128.16	4.27

action	0	1	2	3	4	5	6
sub1	1	1	1	1	1	1	1
sub2	1	1	1	1	1	1	1
sub3	1	1	1	0	1	1	1
sub4=1	1	1	1	2	1	1	1
sub5	1	1	1	1	1	1	1
sub6	1	1	1	1	1	1	1
sub7	1	1	1	1	1	1	1
sub8	1	1	1	1	1	1	1
sub9=1	1	1	1	1	1	1	1
sub10=5	1	1	1	1	1	1	1
sub11=3	1	1	1	1	1	1	1
sub12	1	1	1	1	1	1	1

[1] First-Person Activity Recognition: What Are They Doing to Me? (CVPR2013)
http://michaelryoo.com/jpl-interaction.html

Experiments – JPL Dataset

- Evaluation Results on the JPL Dataset

1. pretrain on k400
2. pretrain on sports1m
3. pretrain on ucf101
4. Human parsing mask supervision: pretrain with LIP (Look into People)
5. unsupervised pertrain with AMASS, Human3.6M, PoseTrack, InstaVariety

Train Sub	1	2	3	4	5	6	7	8	9	10	11	12
Split 1	✓	✓	✓	✓					✓			✓
Split 2					✓	✓	✓	✓		✓	✓	
Split 3	✓	✓						✓		✓	✓	✓
Split 4			✓	✓	✓	✓	✓		✓			

Method	Split 1	Split 2	Split 3	Split 4	JPL Avg
Scratch-training					
I3D RGB	69.05	73.81	76.19	76.19	73.81
TSN Flow	66.67	73.81	88.10	71.43	75.00
TSN RGB	69.05	73.81	80.95	85.71	77.38
MotionBert	80.95	80.95	80.95	73.81	79.17
PoseC3D Joint	85.71	83.33	73.81	78.57	80.36
I3D Flow	80.95	78.57	85.71	83.33	82.14
PoseC3D Limb	83.33	85.71	85.71	83.33	84.52
Ours fc-head	95.24	92.86	80.95	80.95	87.49
Ours lstm-head	92.85	88.10	88.10	83.33	88.10
Fine-tuning					
TSN Flow[1]	73.81	80.95	88.10	78.57	80.36
TSN RGB[1]	85.71	88.10	92.86	90.48	89.29
I3D RGB[1]	92.86	85.71	95.24	95.24	92.26
Global + local C3D[2]	-	-	-	-	94.4
PoseC3D Limb[1]	97.62	92.86	95.24	95.24	95.24
PoseC3D Joint[1]	95.24	95.24	95.24	95.24	95.24
TSCF[3]	-	-	-	-	94.4
DRM+Interactive LSTM[4]	-	-	-	-	98.4
Three-stream I3D[1]	-	-	-	-	98.5
MotionBert[5]	100.00	97.62	97.62	100.00	98.81
Ours fc-head[1]	100.0	90.48	97.62	92.86	95.24
Ours lstm-head[1]	100.0	95.24	100.0	88.10	95.83

Experiments – UTKinect First Person Dataset

Robot-Centric Activity Recognition from First-Person RGB-D Videos

Lu Xia[1], Ilaria Gori[1,2], J. K. Aggarwal[1], and M. S. Ryoo[3]
[1]Department of ECE, The University of Texas at Austin, USA
[2]iCub Facility, Istituto Italiano di Tecnologia
[3]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, USA
xialu@utexas.edu, ilaria.gori@iit.it, aggarwaljk@mail.utexas.edu, mryoo@jpl.nasa.gov

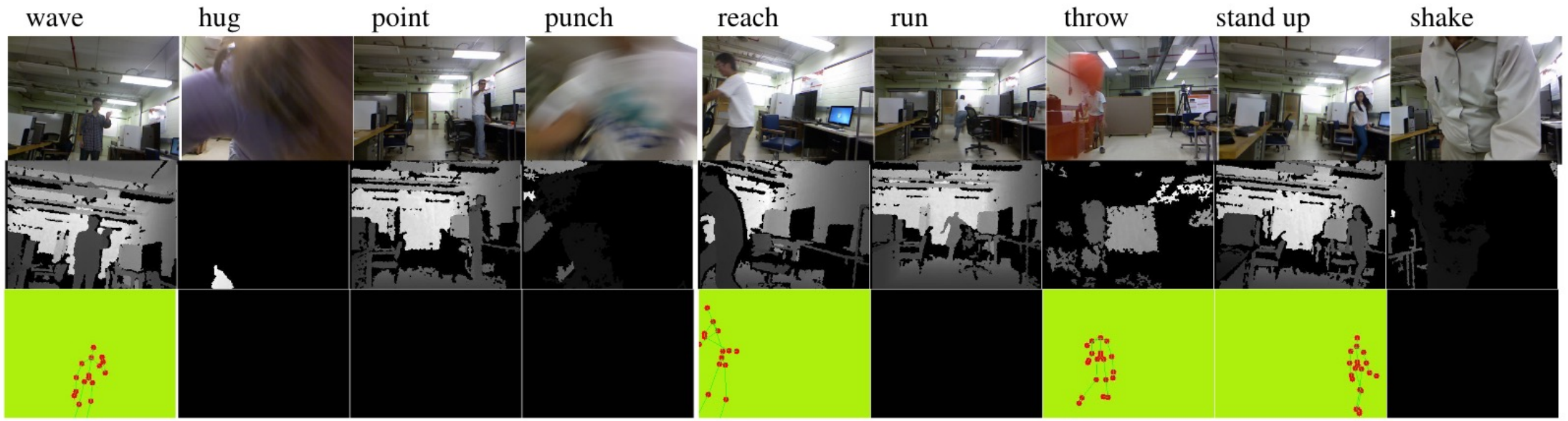
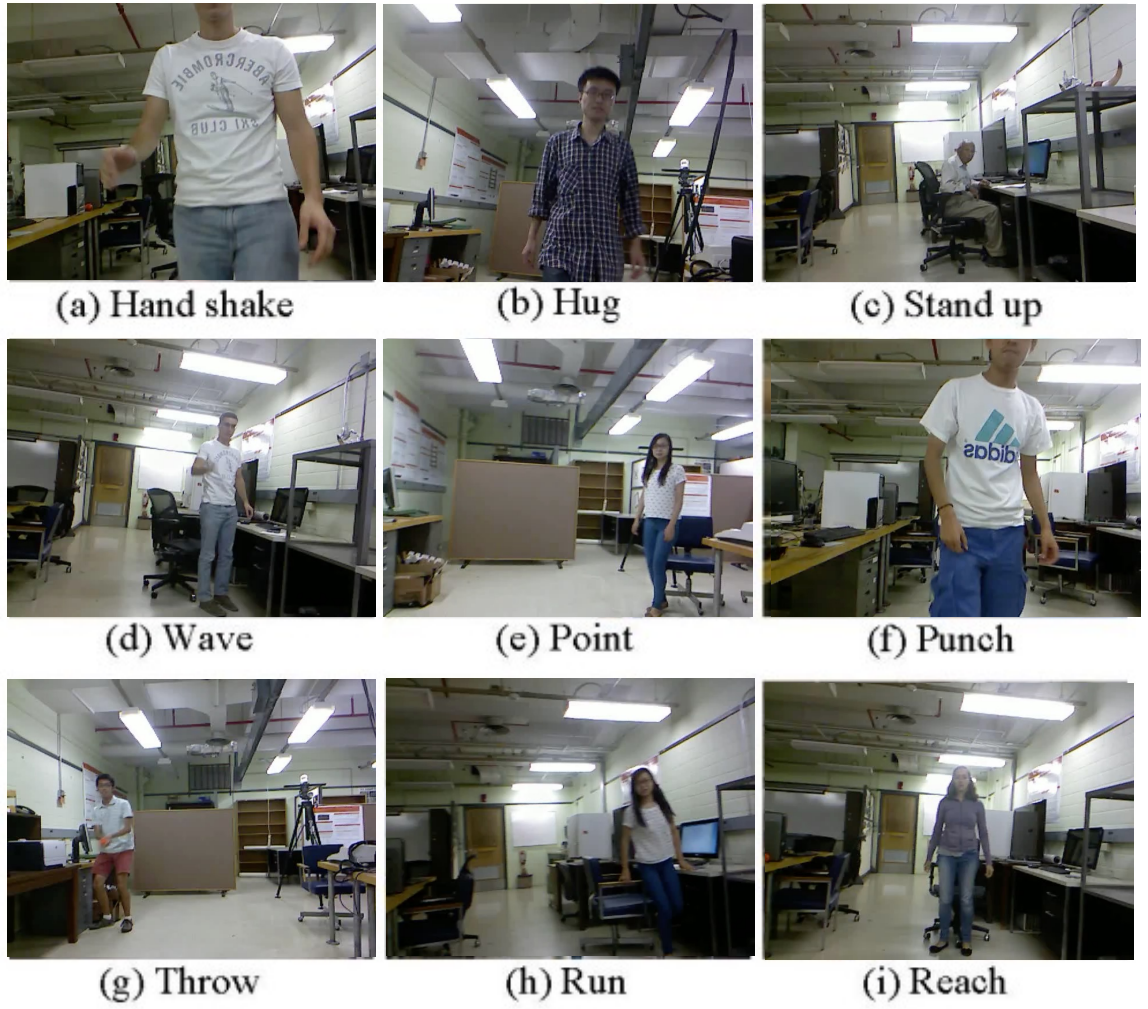


Table 1: Sample images of 9 activities in the humanoid robot first-person RGBD dataset. The first and second rows present the RGB and depth images, respectively. The last row represents skeleton images. If no skeleton is detected for a particular frame, a black image is shown.

https://cvrc.ece.utexas.edu/KinectDatasets/FirstPerson.html

[2] Robot-Centric Activity Recognition from First-Person RGB-D Videos (WACV2015)

Experiments – UTKinect First Person Dataset



- Extended version of JPL Dataset (on my op)

Two different robots → Kinect device

1. a humanoid robot : 177 video clips
2. an autonomous non-humanoid robot: 189 video clips

- 8 subjects, between the ages of 20 to 80
- Skeleton data was sparsely detected.

Kinect device provide depth image and human skeletons

https://cvrc.ece.utexas.edu/KinectDatasets/FirstPerson.html

[2] Robot-Centric Activity Recognition from First-Person RGB-D Videos (WACV2015)

Experiments – UTKinect First Person Dataset

→ UTKinect FPD

Action	id	#Videos	#Frames			Avg sec
			Min	Max	Avg	
Hand-shaking	0	18	41	125	74.44	2.48
Hugging	1	15	53	130	79.73	2.65
Stand Up	2	**39**	**20**	77	40.82	1.36
Hand-wave	3	19	24	85	45.10	1.50
Pointing	4	22	14	**132**	38.54	1.28
Punch	5	18	26	83	53.27	1.77
Reach object	6	19	29	75	47.52	1.58
Throw object	7	19	21	56	37.05	1.23
Run away	8	17	31	116	**83.58**	**2.78**

action	0	1	2	3	4	5	6	7	8
sub1	2	2	**6**	1	2	3	2	2	2
sub2	3	2	**6**	3	4	2	3	3	2
sub3	2	2	4	2	2	2	2	2	2
sub4	2	2	**6**	2	1	2	2	2	2
sub5	2	2	4	3	3	2	2	2	2
sub6	3	1	5	3	4	2	2	2	2
sub7	2	2	2	3	3	2	3	3	3
sub8	2	2	**6**	2	3	3	3	3	2

https://cvrc.ece.utexas.edu/KinectDatasets/FirstPerson.html

[2] Robot-Centric Activity Recognition from First-Person RGB-D Videos (WACV2015)

Experiments – UTKinect First Person Dataset

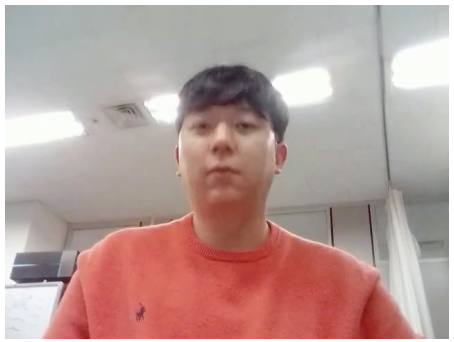
- Evaluation Results on the UTKinect First Person Dataset

1. pretrain on k400
2. pretrain on ucf101
3. unsupervised pertrain with AMASS, Human3.6M, PoseTrack, InstaVariety

Method	UTKinect-FPD
Scratch-training	
TSN Flow	42.11
I3D RGB	62.11
Two-stream ConvNet	65.9
MotionBert	70.53
LRCN RGB	72.6
ConvLSTM RGB	79.6
I3D Flow	80.00
TSN RGB	83.16
PoseC3D Limb	86.32
PoseC3D Joint	88.42
Ours lstm-head	93.68
Ours fc-head	94.74
Fine-tuning	
TSN Flow[1]	73.68
TSCF Diff[2]	84.4
TSN Flow[1]	89.47
TSN RGB[1]	89.47
Three-stream I3D[1]	91.5
I3D RGB[1]	91.58
PoseC3D Limb[1]	92.63
PoseC3D Joint[1]	93.68
MotionBert[3]	96.84
Ours lstm-head[1]	95.79
Ours fc-head[1]	96.84

Experiments - Our ESI dataset

Head-nod



Pet

Hand-shake

Clap

Hug

Head-shake

Zone out

Arm cross

Hand wave

Punch

Experiments - Our ESI dataset

→ Video Clip Length Information

Action	id	#Videos	#Frames			Avg sec
			Min	Max	Avg	
Head nodding	0	100	**35**	146	70.25	2.341
Pet	1	100	54	171	87.54	2.918
Hand shaking	2	100	54	196	89.55	2.985
Clapping	3	100	37	173	72.8	2.427
Hugging	4	100	66	**332**	**127.32**	**4.244**
Head shaking	5	100	39	145	69.22	2.307
Zone Out	6	100	45	222	84.36	2.812
Arm Cross	7	100	53	212	98.39	3.280
Hand wave	8	100	35	104	61.81	2.060
Punch	9	100	41	146	**60.44**	**2.015**

Experiments - Our ESI dataset

- Evaluation Results on the social interaction datasets

1. pretrain on k400
2. unsupervised pertrain with AMASS, Human3.6M, PoseTrack, InstaVariety

Method	ESI-dataset
Scratch-training	
MotionBert	60.4
TSN RGB	80.0
TSN Flow	84.2
I3D RGB	85.0
PoseC3D Limb	86.8
I3D Flow	89.4
PoseC3D Joint	91.8
Ours fc-head	92.0
Ours lstm-head	92.6
Fine-tuning	
TSN RGB[1]	74.4
MotionBert[2]	87.2
TSN Flow[1]	89.4
I3D RGB[1]	89.6
PoseC3D Limb[1]	90.0
PoseC3D Joint[1]	92.2
Ours fc-head[1]	92.4
Ours lstm-head[1]	93.8

Future work

Future work

Thank you

Q & A

HoBeom Jeon

UST-ETRI
Social Robotics Research Section

2023.11.23