Social Behavior Recognition for Nonverbal Human Robot Interaction

Our Approach

HoBeom Jeon

UST-ETRI
Social Robotics Research Section

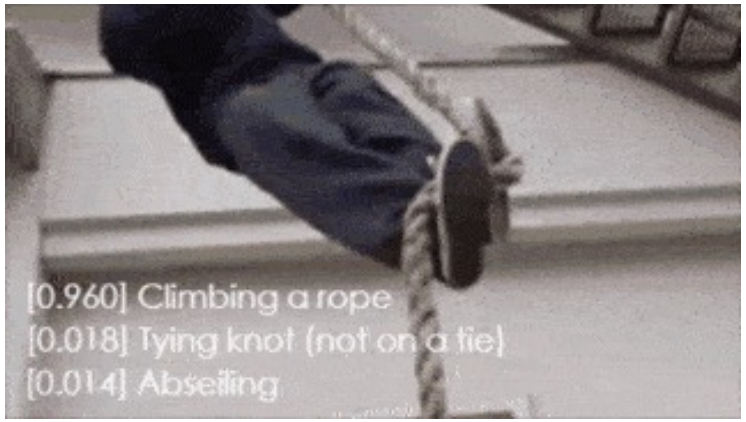
2023.10.26

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.



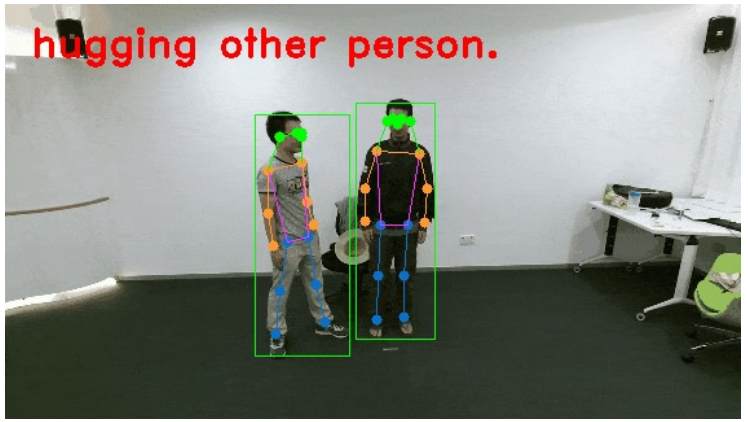


Image Source: [mmaction2]

Recap: Enhancing Human-Robot Interaction through HAR

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

Industrial collaborative robots

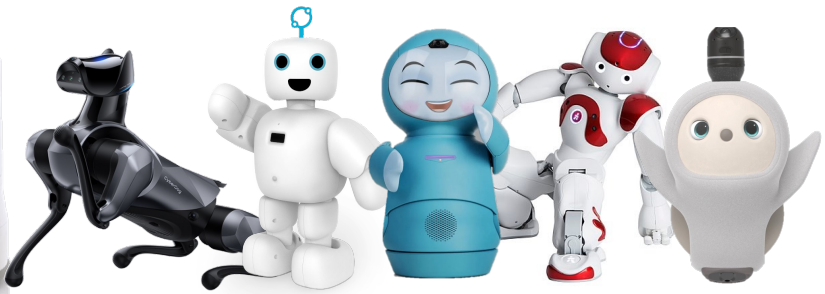
Image Source: MobileAutomation

Image Source: KUKA

Social robots

PAL ROBOTICS
TIAGo Pro

LG CLOi



Xiaomi
CyberDog2

PiBo

Moxie

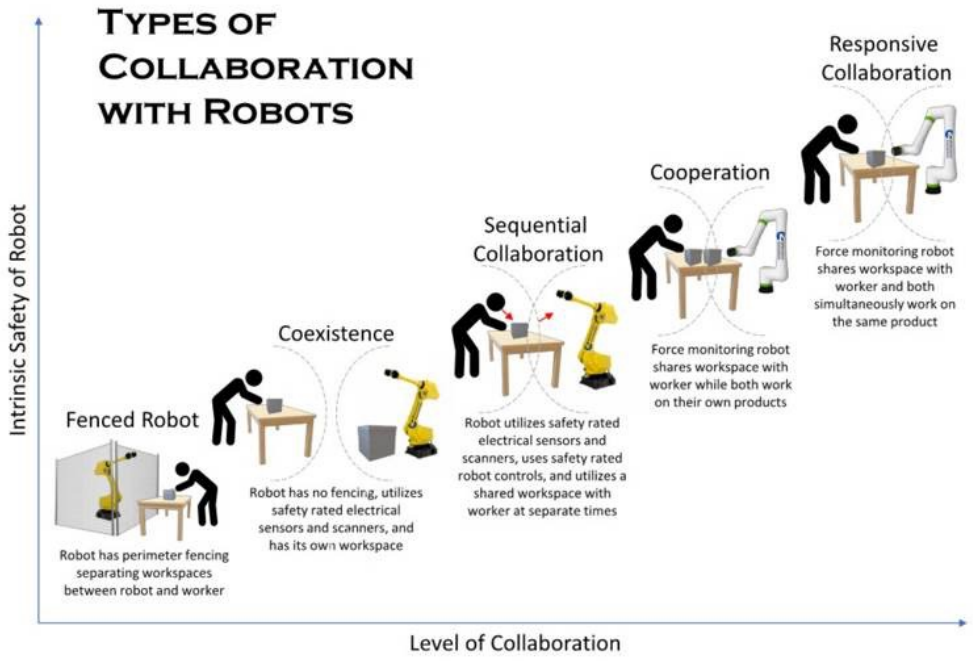
Nao

GrooveX
Lovot

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: ZetaGroupENG

Social robots

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

Image Source: Moxie

Recap: 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

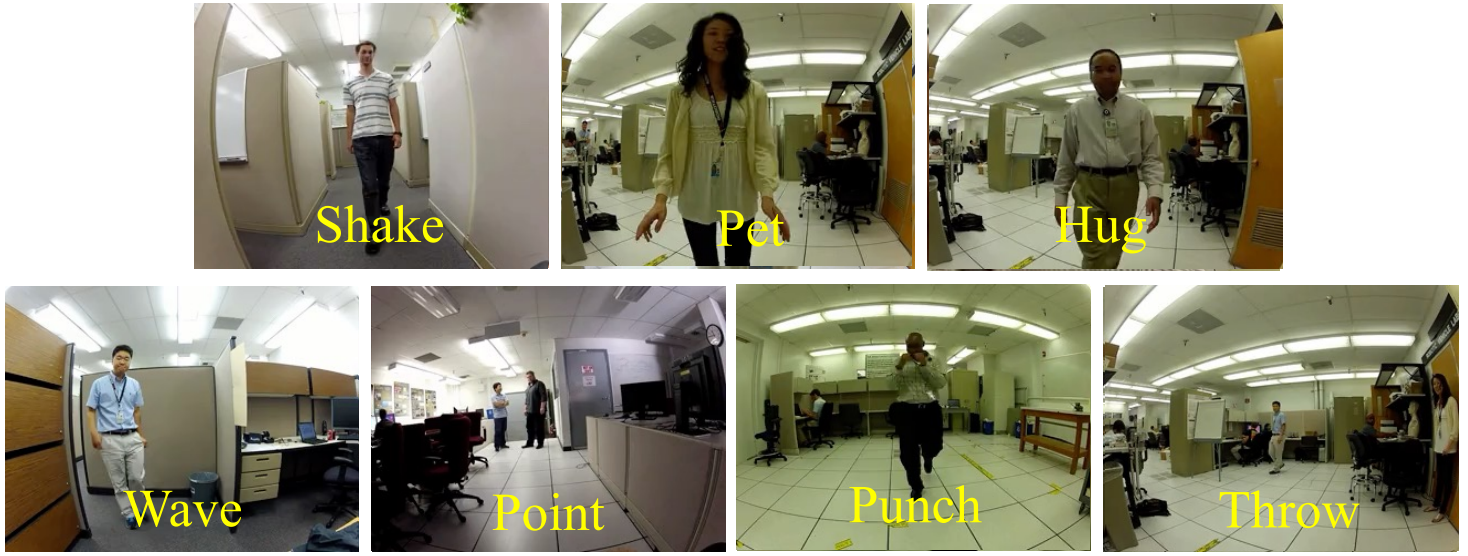
Michael S. Ryoo
Stony Brook University; Robotics at Google
cs.stonybrook.edu의 이메일 확인됨 - 홈페이지
Robotics Computer Vision Machine Learning

(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

Recap: JPL Dataset



- 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

Recap: 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

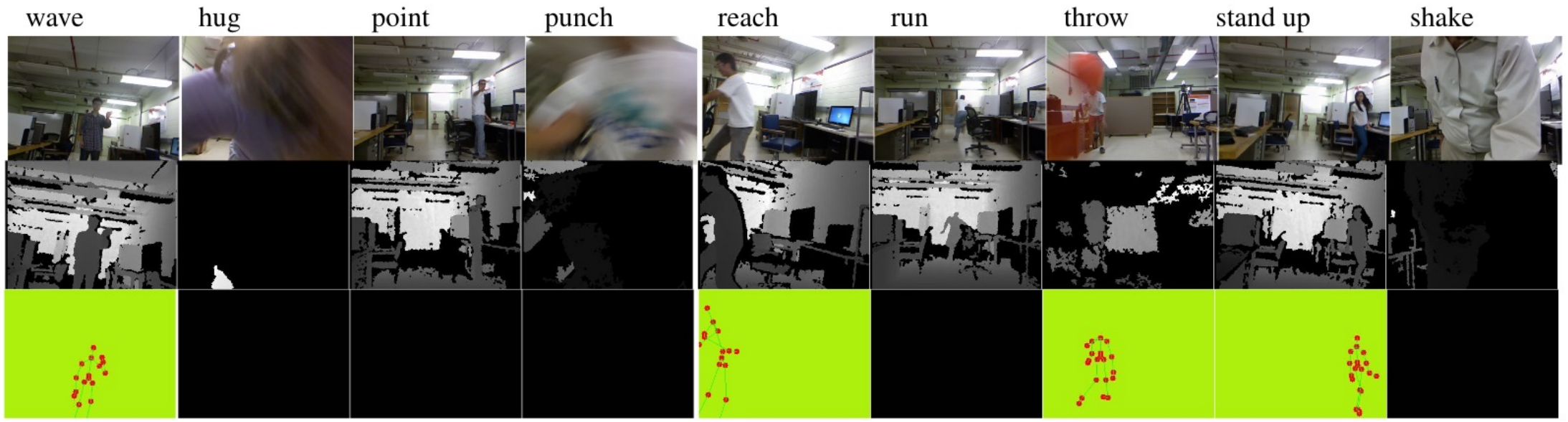
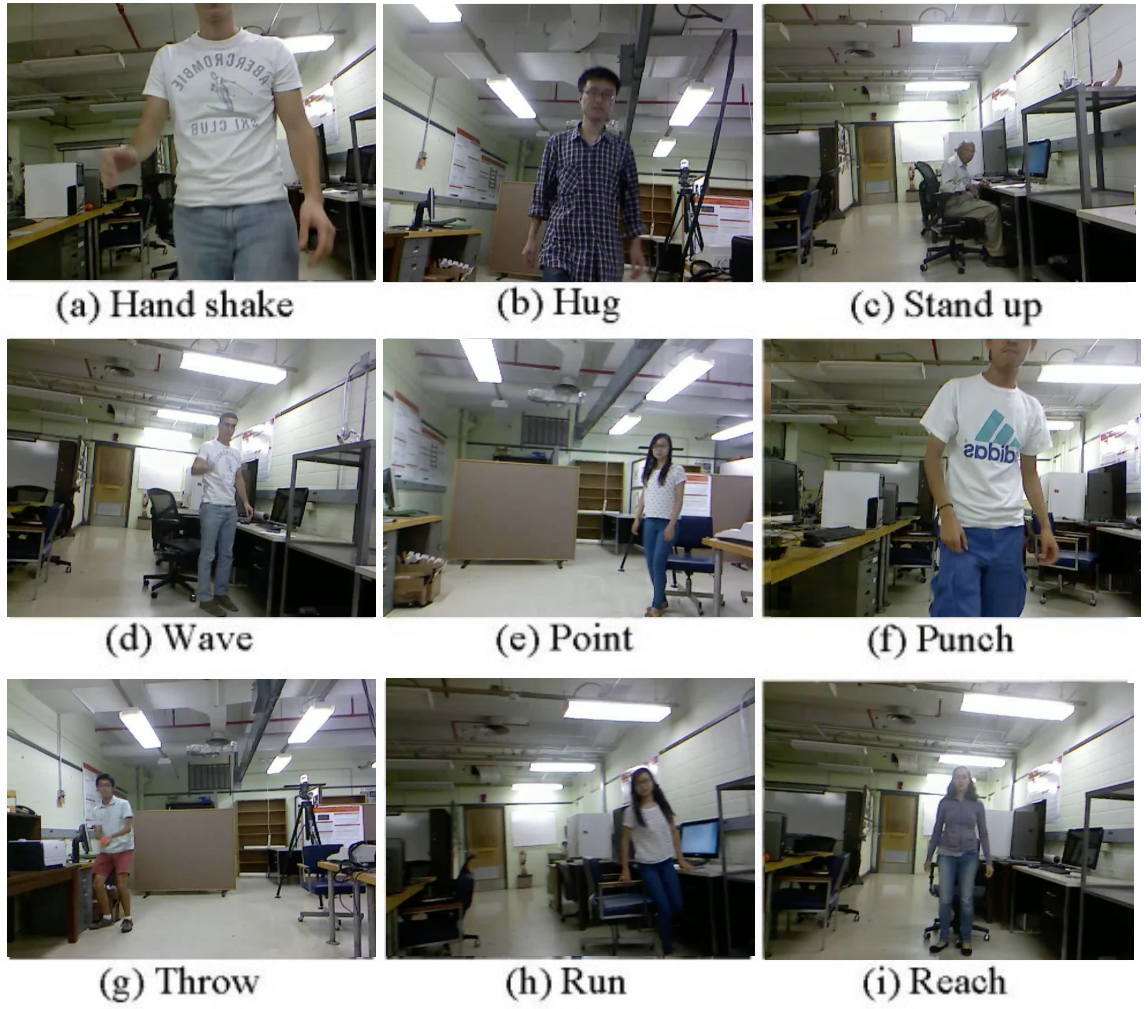


Table 1: Sample images of 9 activities in the humanoid robot first-person RGBD dataset. The first and second rows present the RGB and depth images, respectively. The last row represents skeleton images. If no skeleton is detected for a particular frame, a black image is shown.

https://cvrc.ece.utexas.edu/KinectDatasets/FirstPerson.html
[2] Robot-Centric Activity Recognition from First-Person RGB-D Videos (WACV2015)

Recap: 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)

Recap: convLSTM – first deep learning approach

- convLSTM: **CNN + LSTM** architecture (2017)

  Introduced deep learning model for the first time on first-person interaction datasets.

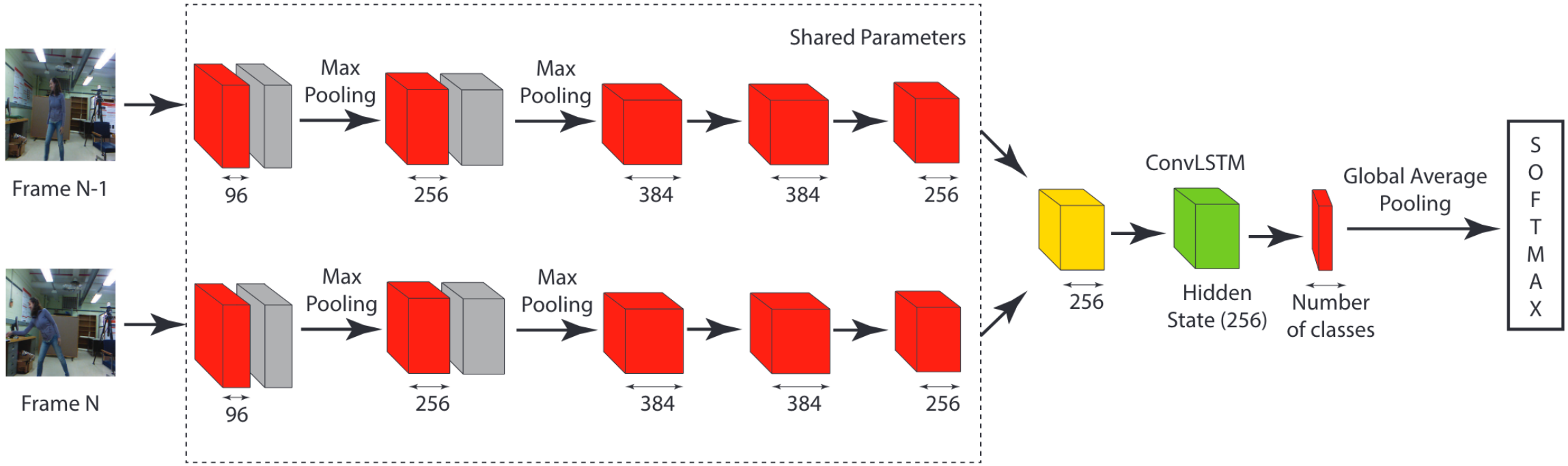


Figure 2. The architecture of the network. The convolutional layers are shown in red followed by normalization layer in gray. The 3D convolutional layer is shown in yellow and the convLSTM layer in green. We also experiment with a variant where, instead of raw frames we input difference-images obtained from pairs of successive frames.

[3] Convolutional Long Short-Term Memory Networks for Recognizing First Person Interactions (ICCV2017)

Recap: DRM - Body Part Detection

- DRM: CNN (**Human Body Part Attention**) + interactive LSTM

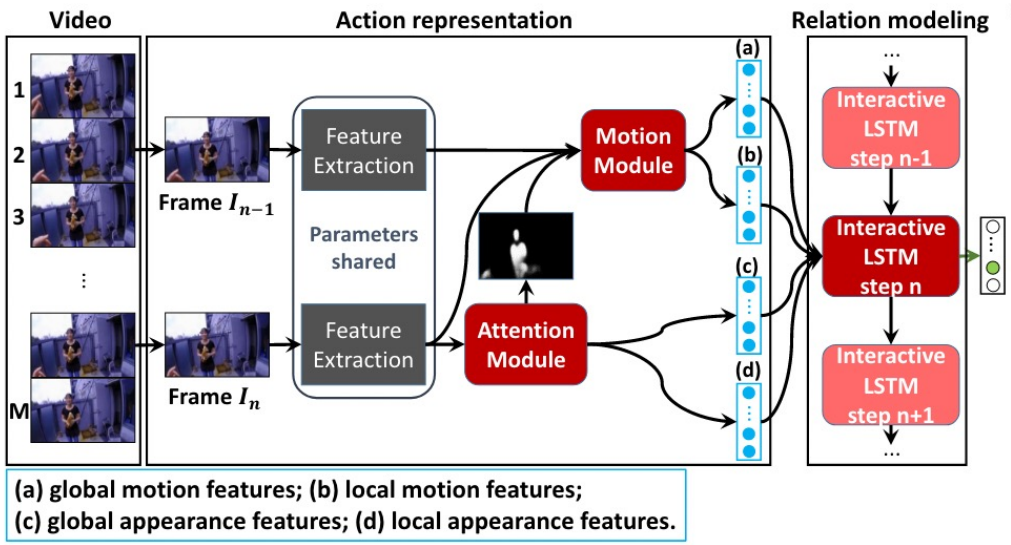


Figure 2. Proposed framework. Frames $\boldsymbol{I}_i (i = 1, ..., N)$ are sam-

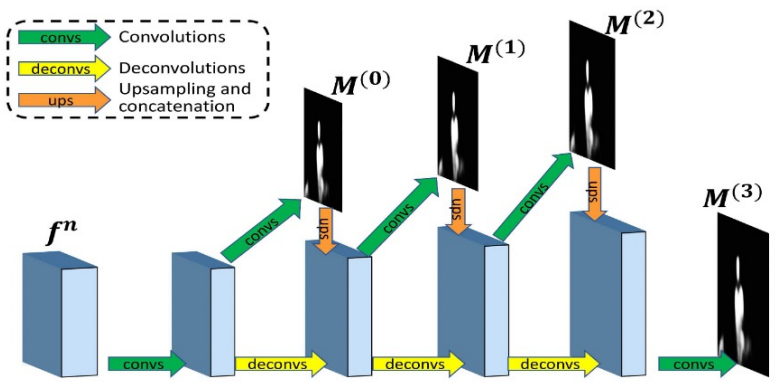


Figure 3. Structure of attention module. The module takes feature

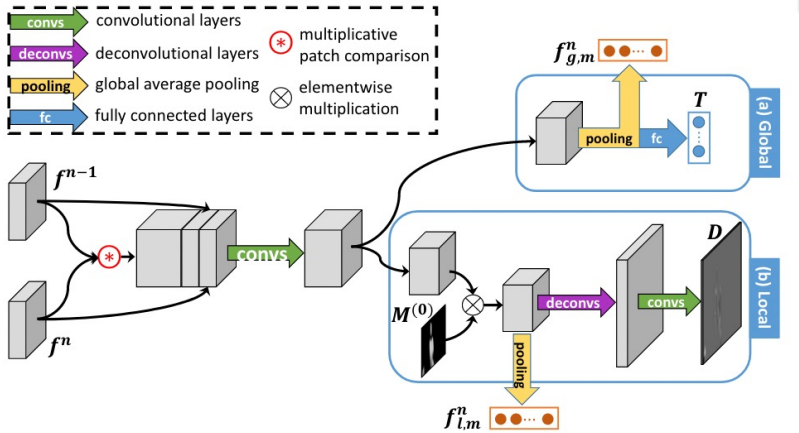


Figure 4. Structure of motion module. The module takes basic

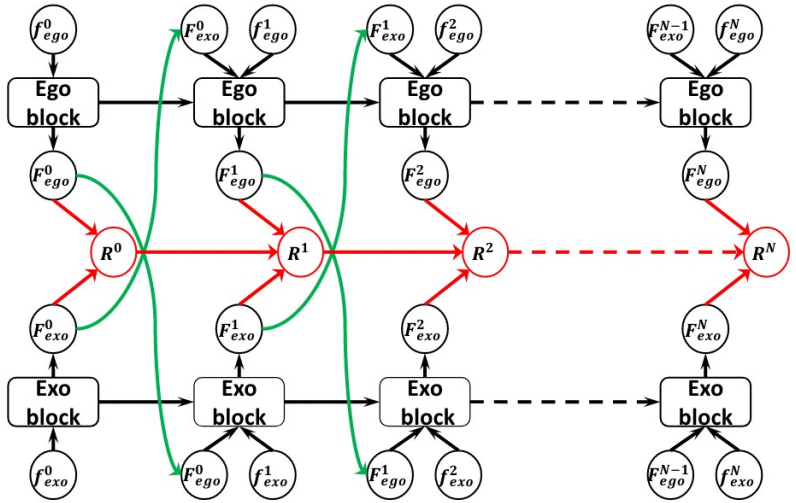


Figure 5. Diagram of Interactive LSTM. The unrolled symmetrical

[4] Deep Dual Relation Modeling for Egocentric Interaction Recognition (CVPR2019)

Recap: Three Stream Correlation Fusion

- TSCF: CNN + LSTM **(Human Detection + Optical Flow)** (2020)

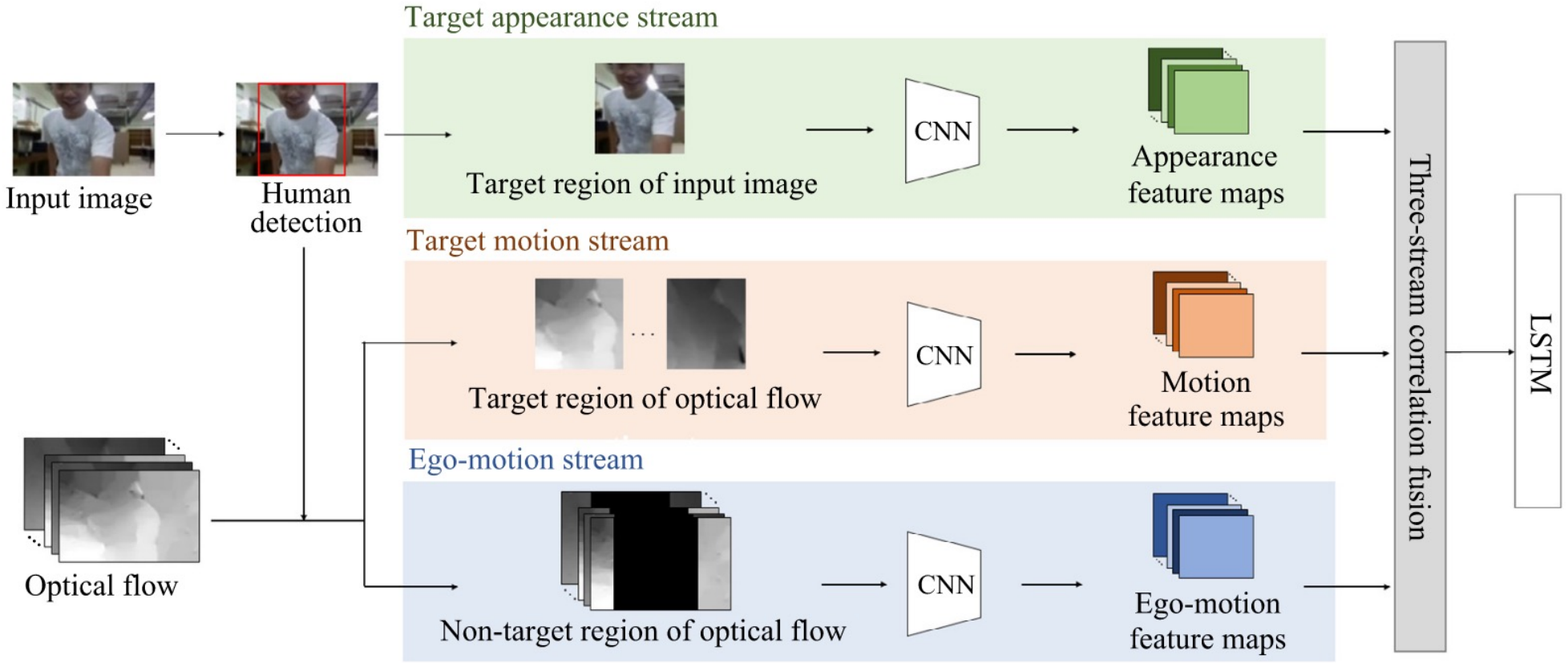
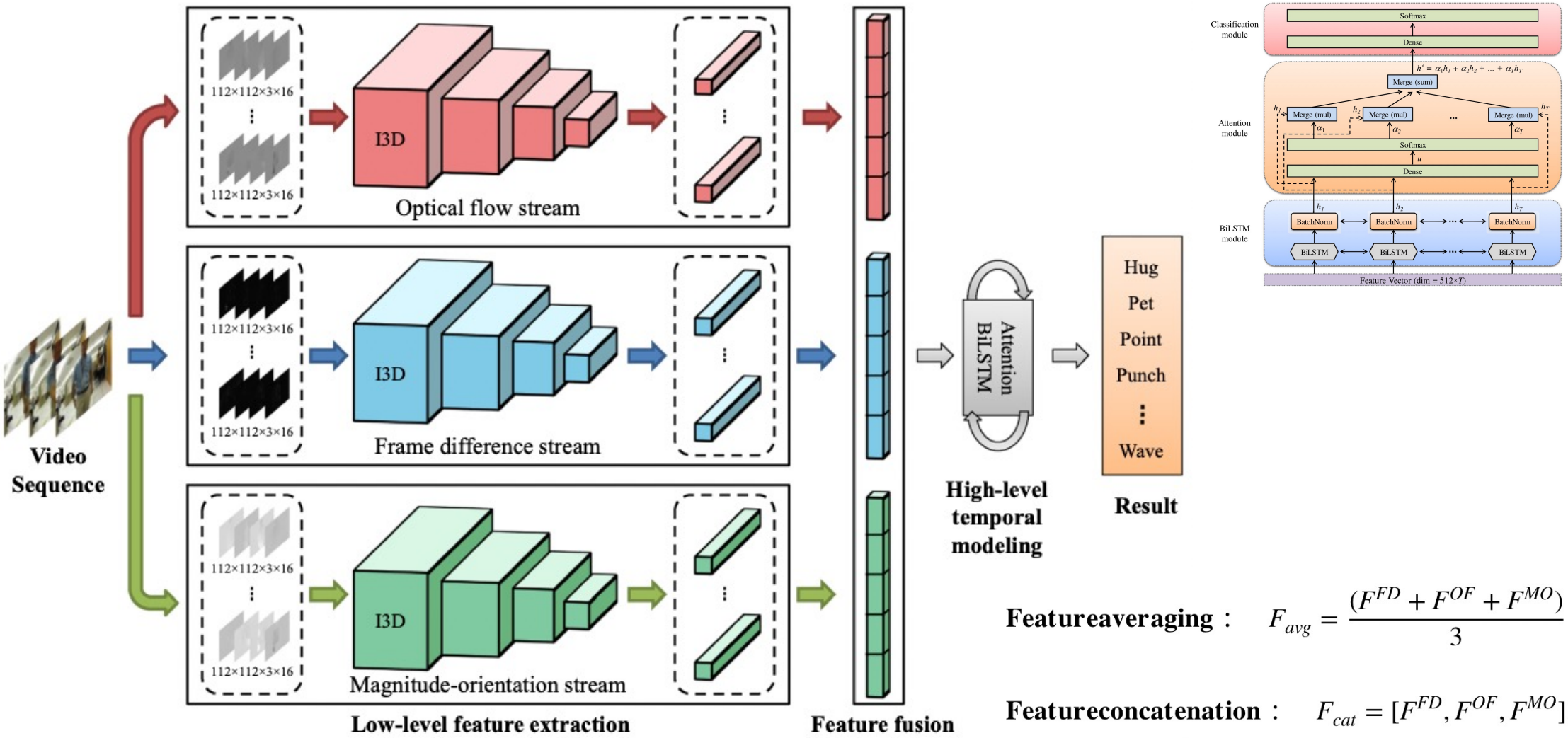


Fig. 2. Overview of the three-stream fusion network. Our proposed network is composed of three-stream architecture, the three-stream correlation fusion (TSCF), and a long short-term memory model. Each stream of the three-stream architecture respectively extracts appearance, motion, and ego-motion feature maps. Then, the proposed TSCF combines the output feature maps of the three streams. The LSTM model takes the fused features as an input value to classify the video class.

[5] Three-stream fusion network for first-person interaction recognition, Ye-Ji Kim (Pattern Recognition 2020)

Recap: Three Stream Spatio Temporal Attention

- TSSTA: CNN + LSTM (**Optical Flow** + **Magnitude Orientation**) (2022)



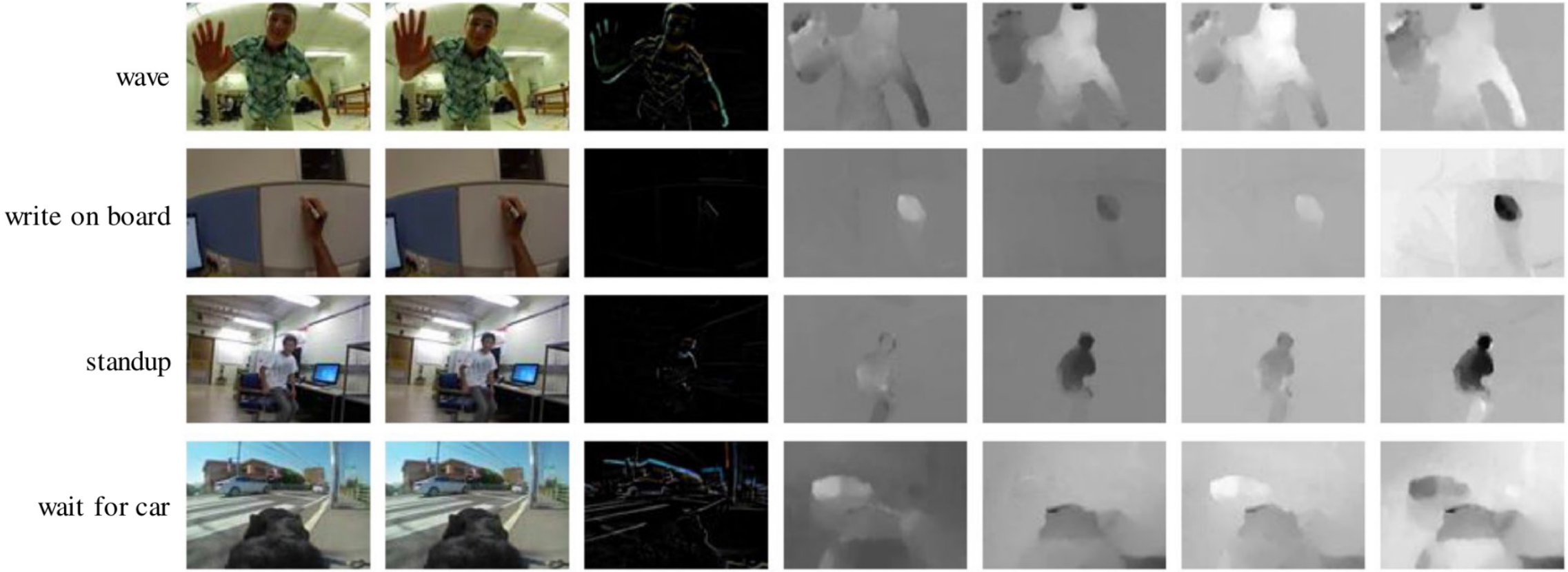
$$\text{Featureaveraging}: \quad F_{avg} = \frac{(F^{FD} + F^{OF} + F^{MO})}{3}$$

$$\text{Featureconcatenation}: \quad F_{cat} = [F^{FD}, F^{OF}, F^{MO}].$$

[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: Three Stream Spatio Temporal Attention

- TSSTA: CNN + LSTM (**Optical Flow** + **Magnitude Orientation**) (2022)



$$M_t = \sqrt{(O_t^x)^2 + (O_t^y)^2} \qquad \theta_t = tan^{-1}\left(\frac{O_t^y}{O_t^x}\right).$$

[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

Motivation - Dataset Analysis

→ 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

Motivation - Dataset Analysis

→ JPL dataset

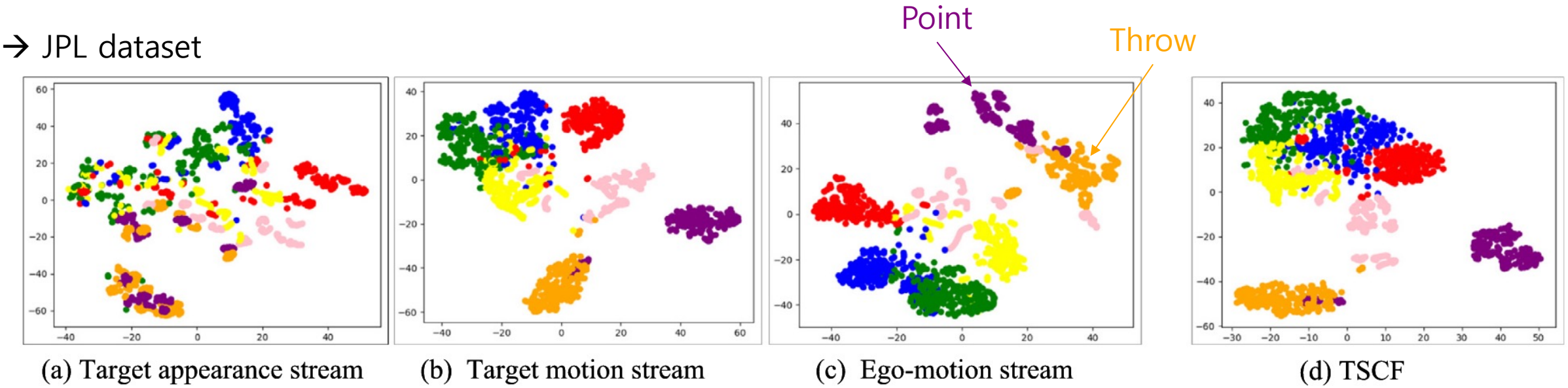


Fig. 9. t-SNE results for the JPL First-Person Interaction dataset. The results in (a), (b), and (c) represent the feature vectors of the target appearance stream, target motion stream, and ego-motion stream, respectively. (d) represents the feature vectors of the three-stream correlation fusion. The color list for seven classes is as follows: hand shake (red), hug (green), pet (blue), wave (pink), point-converse (purple), punch (yellow), and throw (orange). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

[5] Three-stream fusion network for first-person interaction recognition, Ye-Ji Kim (Pattern Recognition 2020)

Motivation - Dataset Analysis

→ 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)

Motivation - Dataset Analysis

→ UTKinect FPD

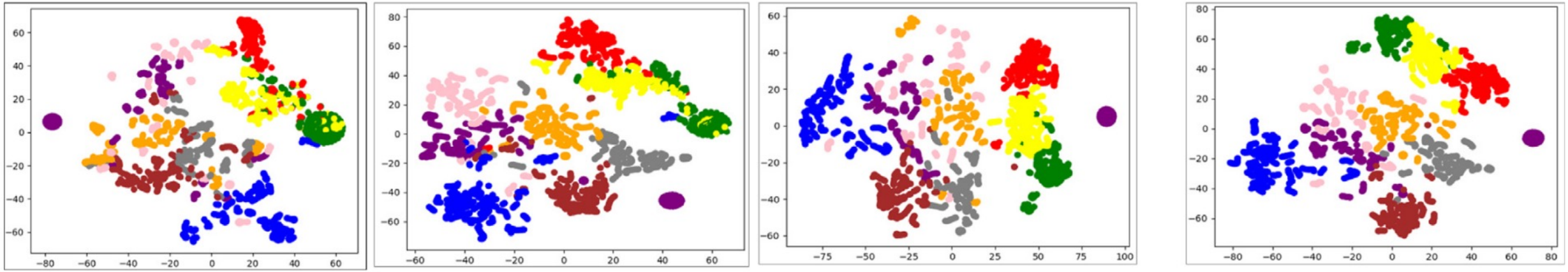


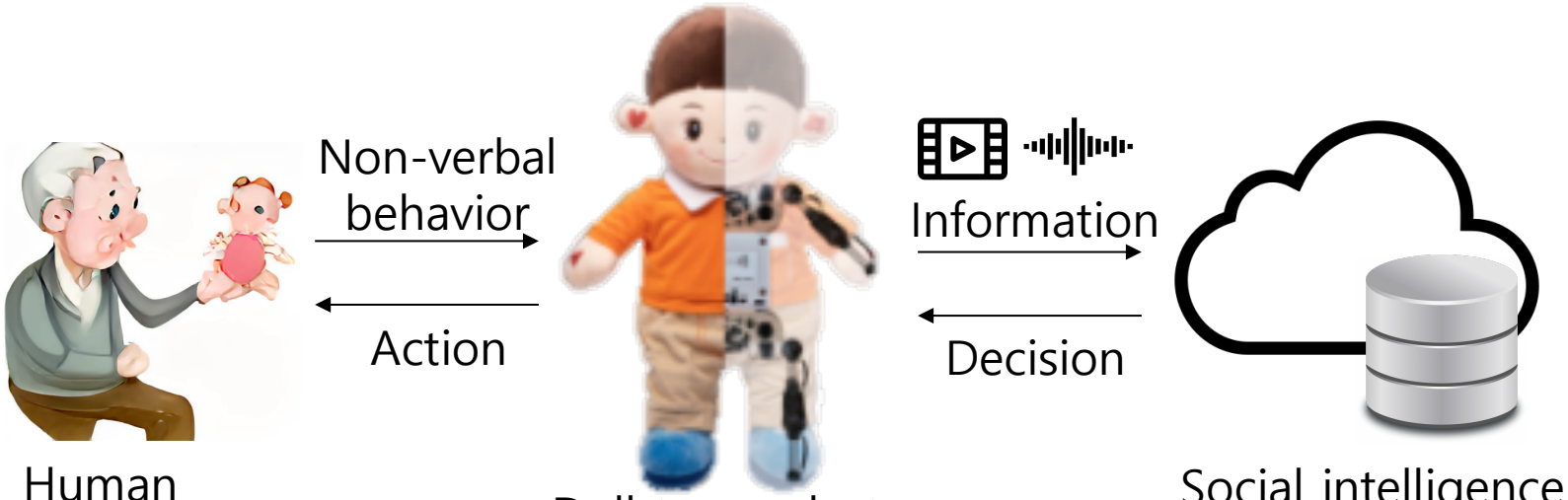
Fig. 8. t-SNE results for the UTKinect-FirstPerson (humanoid) dataset. (a), (b), and (c) represent the feature vectors of the target appearance stream, the target motion stream and the ego-motion stream, respectively. (d) represents the feature vectors of the three-stream correlation fusion. The color list for the nine classes is as follows: hand shake (red), hug (green), stand up (blue), wave (pink), point (purple), punch (yellow), throw (orange), run (brown), and reach (gray). (For interpretation of the references to colo in this figure legend, the reader is referred to the web version of this article.)

[5] Three-stream fusion network for first-person interaction recognition, Ye-Ji Kim (Pattern Recognition 2020)

Motivation

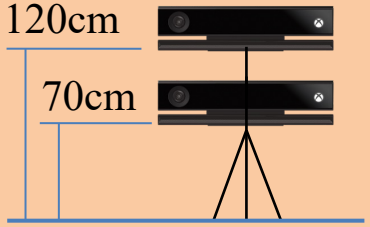
1. Enhancing Emotional Interaction Between Robots and Humans through non-verbal Behavior Recognition
   - Existing datasets have only included informational or expressive actions.
     → Excluding self-movement such as standing, reaching (not interaction)
     → Adding unconscious motions, such as head-noding, arm-crossing

2. Addressing the Shortage of Datasets for Recognizing Social Behavior in Robot Environments
   - Existing datasets have been filmed for conventional computer vision methods.
     → Constructing a new dataset within a first-person robot environment for deep-learning
     → The dataset is recorded under video streaming conditions.



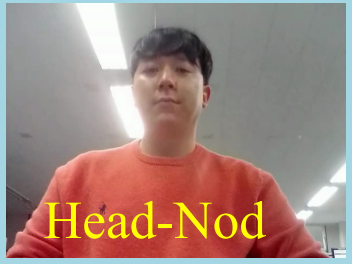
Our Approach: Different Viewpoint Scenario of Social Robots

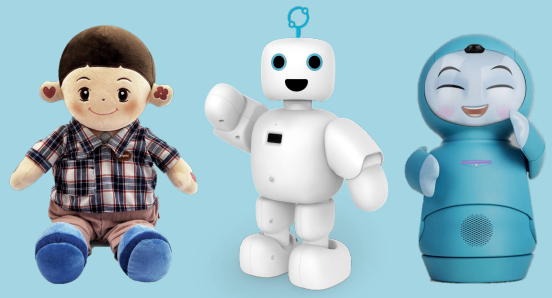
(a) Fixed camera: 3rd person robot view



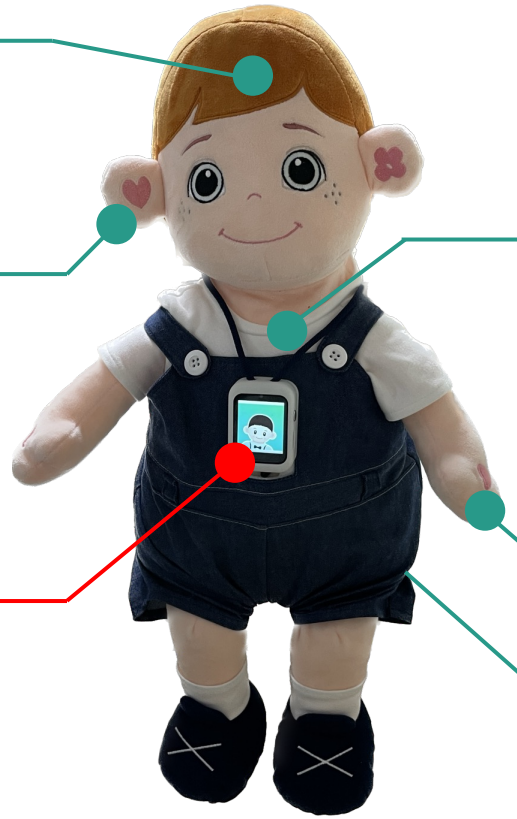
(b) Mobile humanoid robot: 1st person robot view

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





Our Approach: The function of Doll-type Social Robot



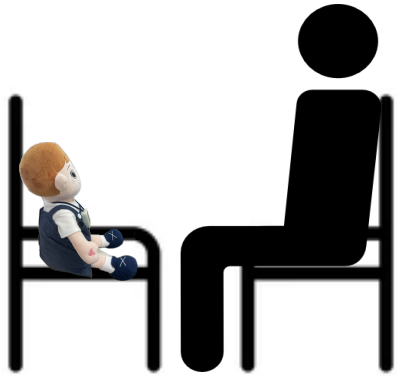
- The close proximity and narrow field of view of doll-type robots make it difficult to capture a comprehensive view of human actions.
- The intervention in camera motion caused by human touch introduces noise and further co mplicates the recognition of human behavior from a first-person robot perspective.

[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.

Our Approach: New Interaction Dataset

→ Two type of doll robot seating positions

Table Seating View
→ Equal Eye Position



Chair Seating View
→ Equal Seat Height

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.

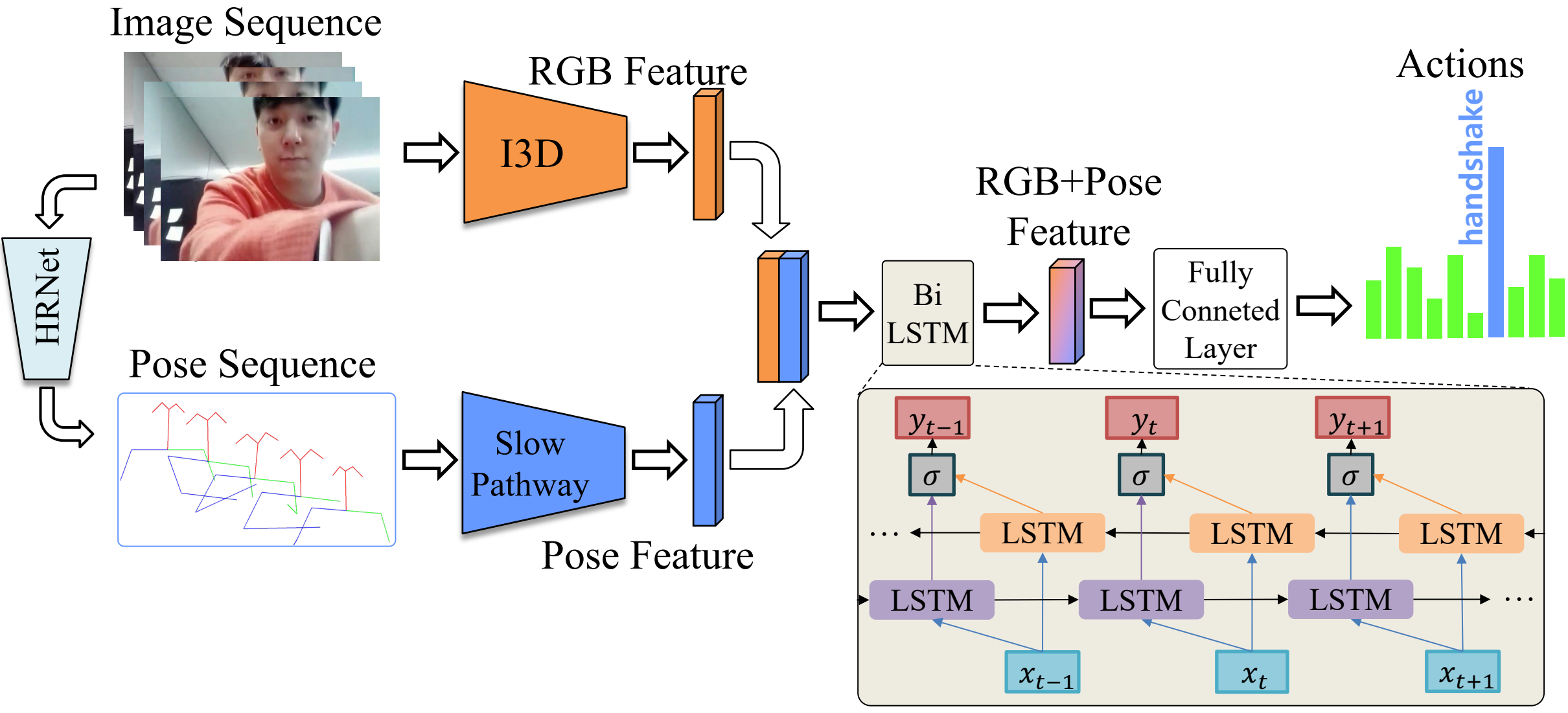
S. H. Kim, H. M. Kim, D. H. Lee, J. H. Hwang, et. al.

Our Approach: New Interaction 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**

Our Approach: RGB + Pose Modality



Overview of our action recognition model

Our Approach: RGB + Pose Modality

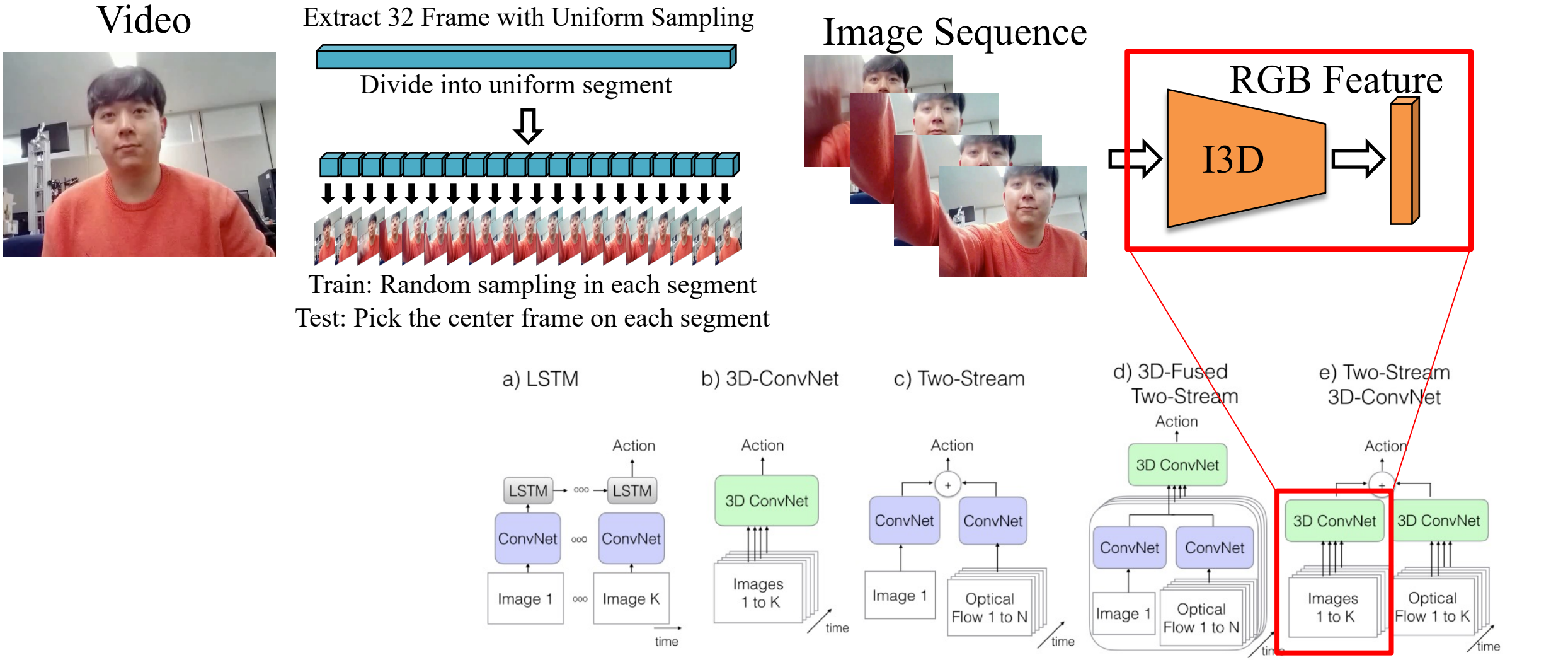


Figure 2. Video architectures considered in this paper. K stands for the total number of frames in a video, whereas N stands for a subset of neighboring frames of the video.

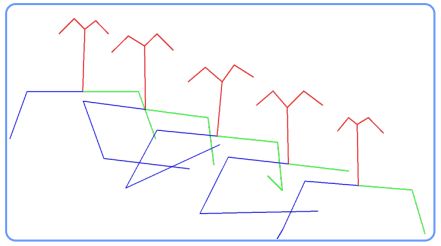
[8] Joao Carreira, and Andrew Zisserman, “Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset”, CVPR2017

Our Approach: RGB + Pose Modality

Image Sequence

HRNet

Pose Sequence

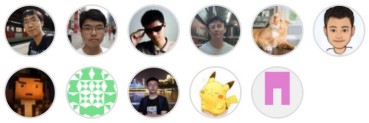


MMPose

Fork 1k Star 4.1k

HoBeom #21
3 commits 485 ++ 18 --

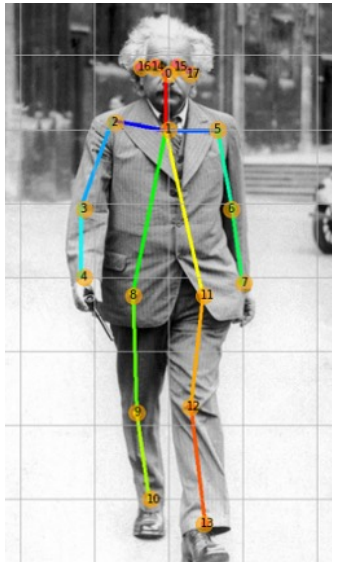
Contributors 108



+ 97 contributors

➢ Human Pose Estimation

HPE aim to estimate the **position of the human joint** in a image input



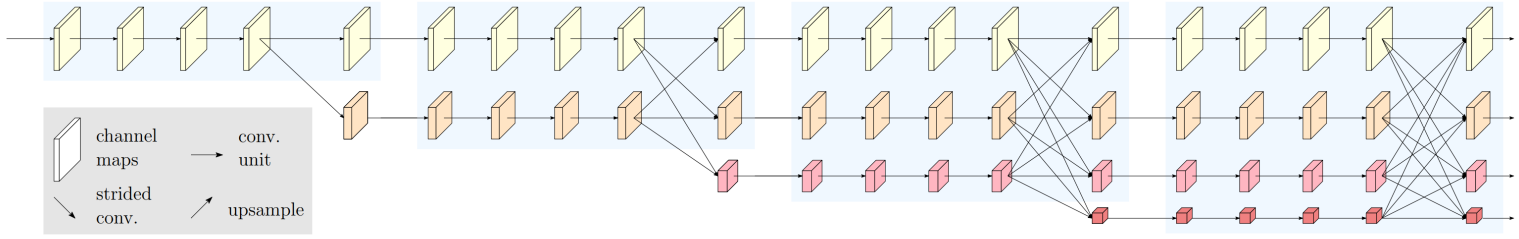
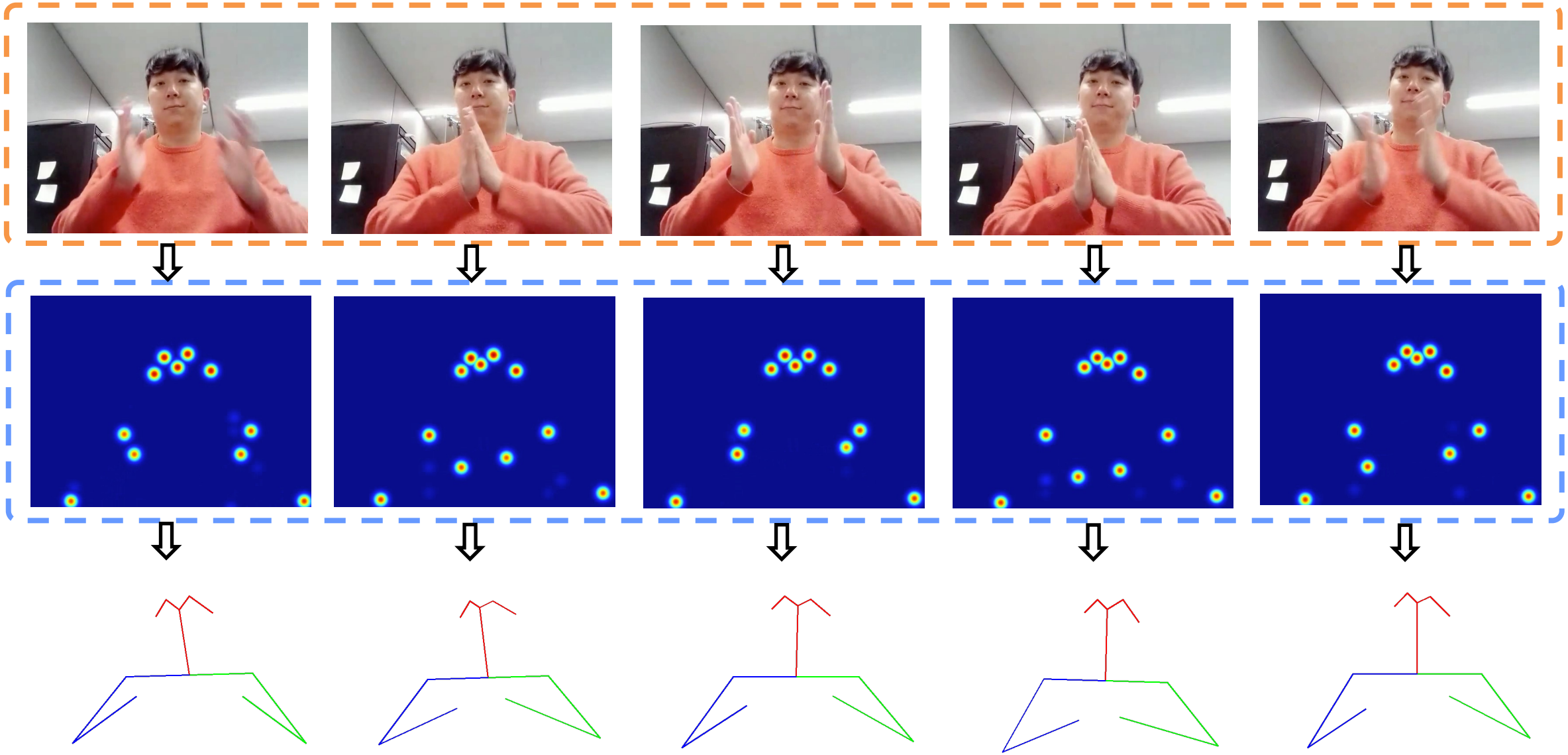


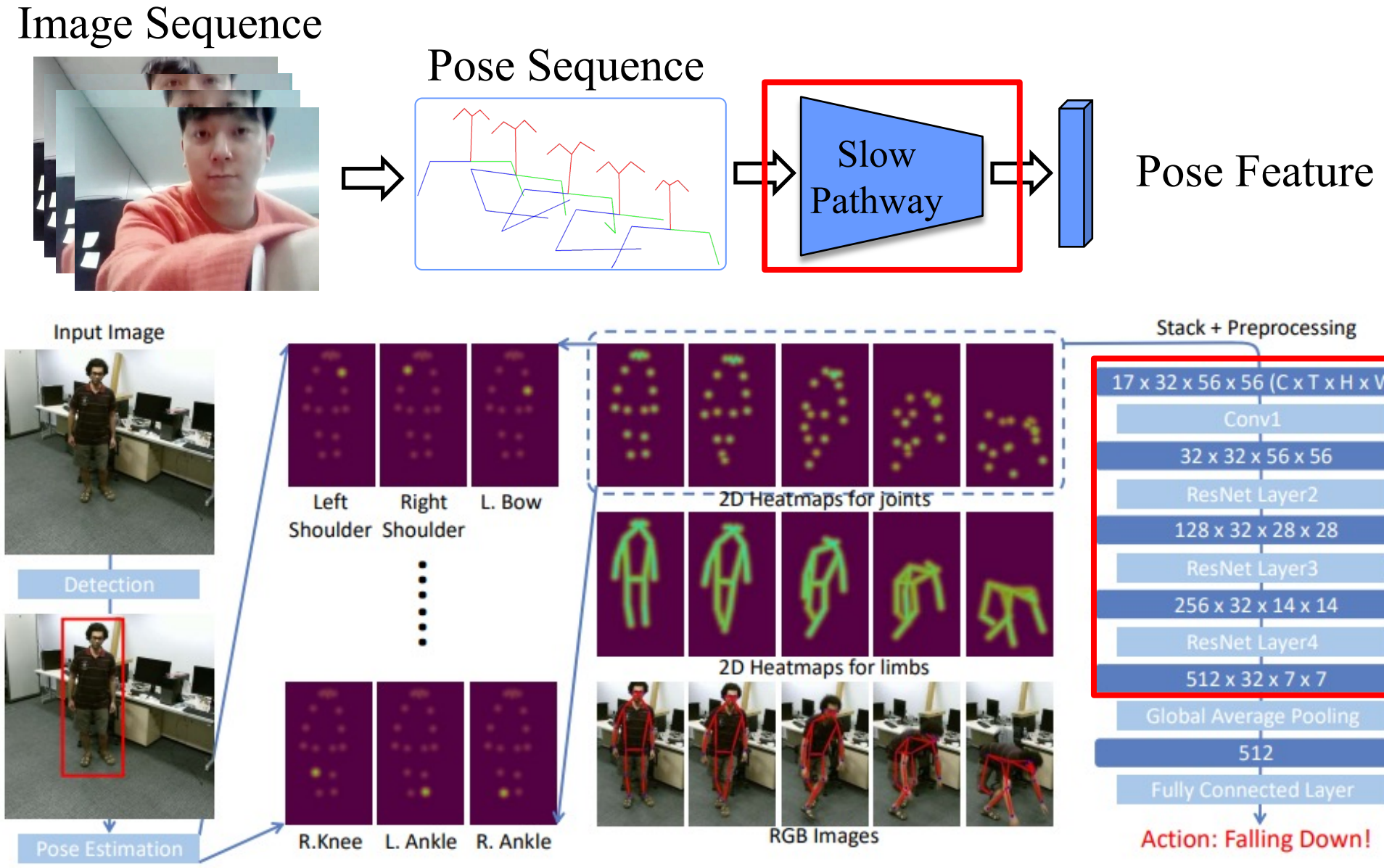
Fig. 2. An example of a high-resolution network. Only the main body is illustrated, and the stem (two stride-2 3×3 convolutions) is not included. There are four stages. The 1st stage consists of high-resolution convolutions. The 2nd (3rd, 4th) stage repeats two-resolution (three-resolution, four-resolution) blocks. The detail is given in Section 3.

[9] Deep High-Resolution Representation Learning for Human Pose Estimation, CVPR2019

Our Approach: RGB + Pose Modality



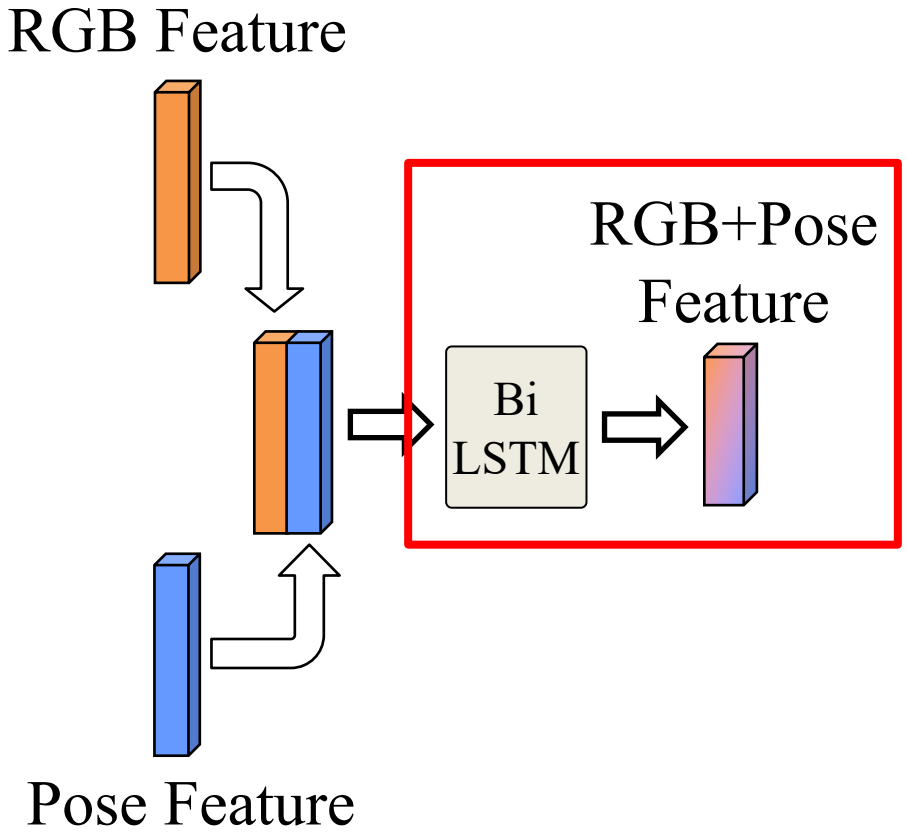
Our Approach: RGB + Pose Modality

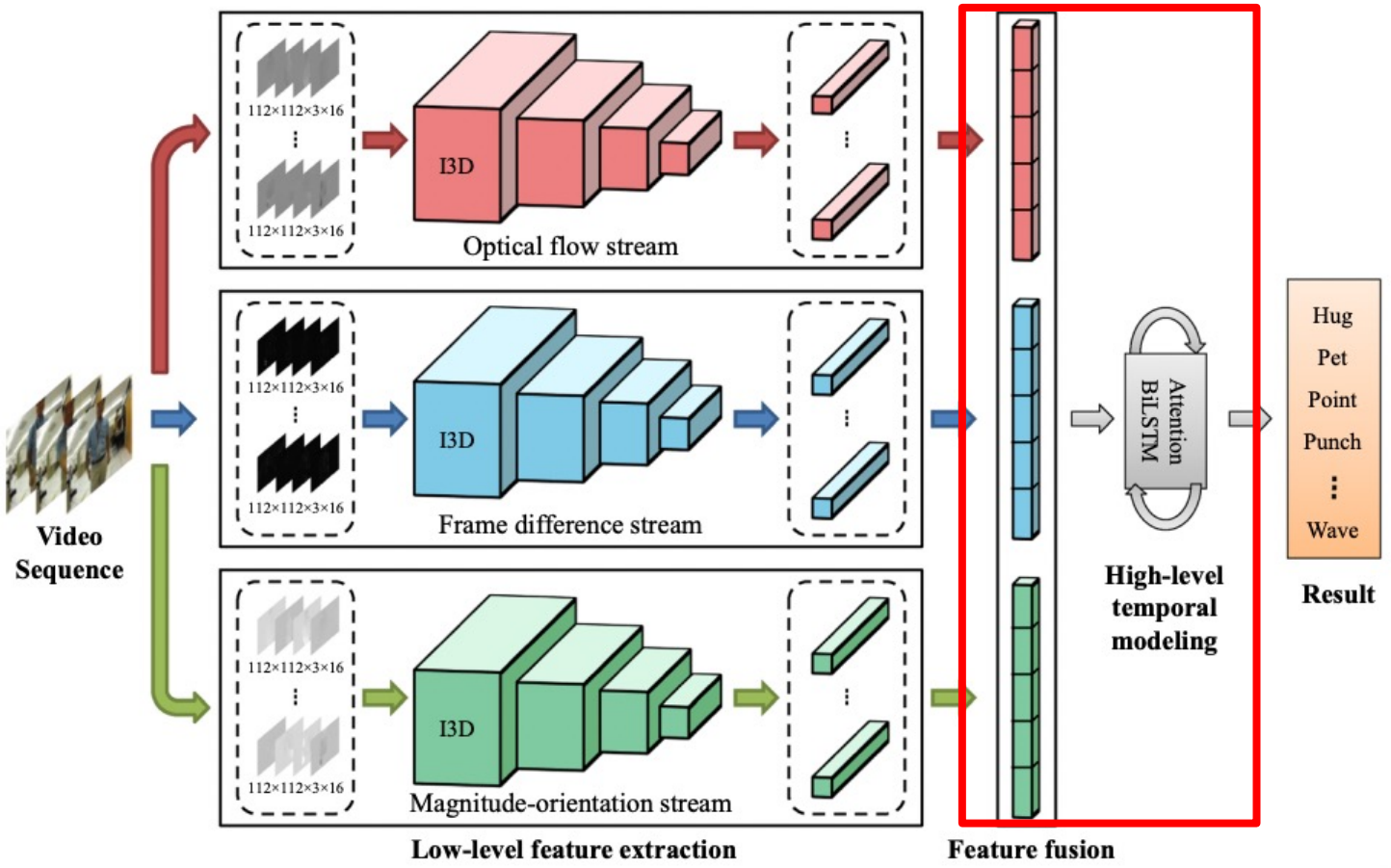


[10] SlowFast Networks for Video Recognition, ICCV2019

[11] Revisiting Skeleton-based Action Recognition, CVPR2022

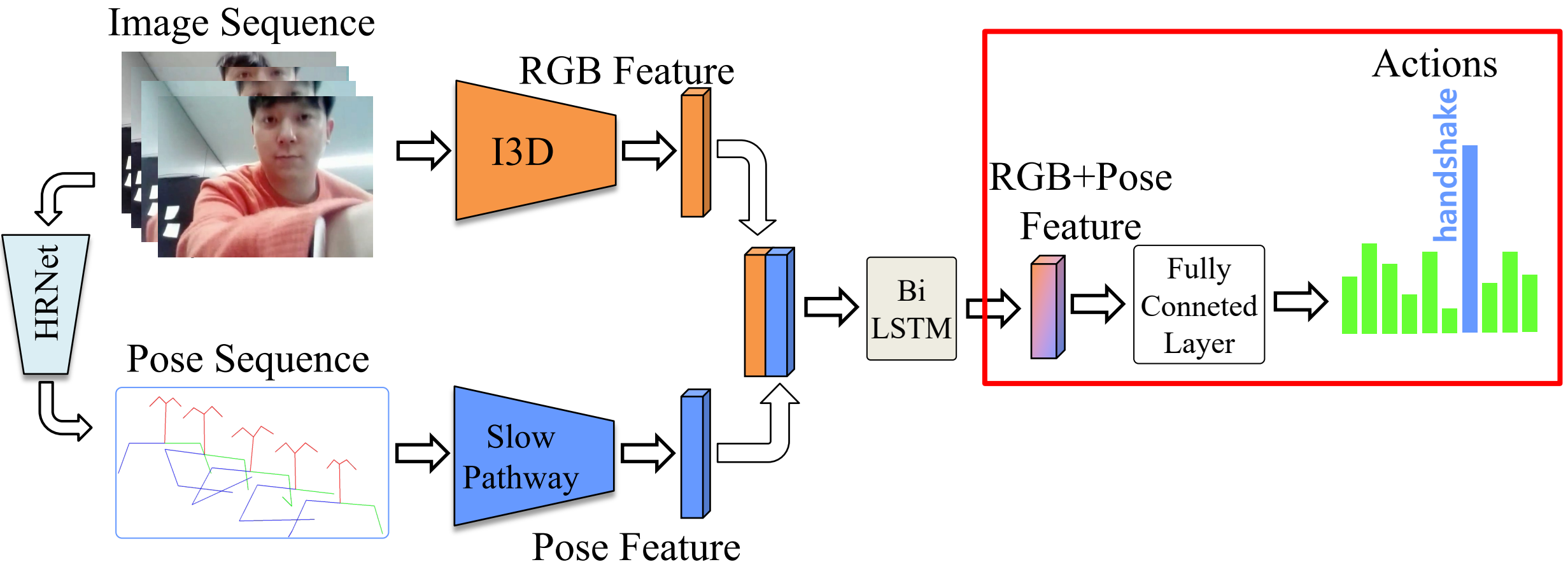
Our Approach: RGB + Pose Modality





[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

Our Approach: RGB + Pose Modality



Preview of next presentation

- Evaluation Results on the social interaction datasets

Method	JPL	UTKinect-FPD	ESI
Scratch-training			
TSN Flow	75.00	42.11	84.2
I3D RGB	73.81	62.11	85.0
MotionBert	79.17	70.53	60.4
I3D Flow	82.14	80.00	89.4
TSN RGB	77.38	83.16	80.0
PoseC3D Limb	84.52	86.32	86.8
PoseC3D Joint	80.36	88.42	91.8
Ours fc-head	87.49	93.68	92.0
Ours lstm-head	88.10	94.74	92.7
Pretrained on the Kinetics 400 dataset			
TSN Flow	80.36	73.68	89.4
TSCF	94.4	84.4	-
TSN RGB	89.29	89.47	74.4
I3D RGB	92.26	91.58	89.6
Three-stream I3D	98.5	91.58	-
PoseC3D Limb	95.24	92.63	90.0
PoseC3D Joint	95.24	93.68	92.2
Ours fc-head	95.24	96.84	92.4
Ours lstm-head	95.83	95.79	93.8

IROS 2023 – Late Breaking Results

Social Behavior Understanding for Nonverbal Human Robot Interaction: I See How You Feel

Hobeom Jeon[1], Hyungmin Kim[1], Dohyung Kim[1,2], and Jaehong Kim[2]

1. Department of Artificial Intelligence at the Korea University of Science and Technology
2. Social Robotics Research Section at the Electronics and Telecommunications Research Institute

Introduction

- we focus on recognizing human behavior from the camera perspective of a companion social robot for human-robot interaction.

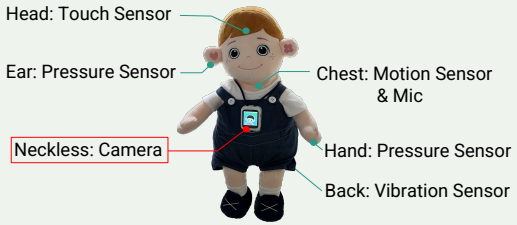


Fig 1. Functionality of Doll-type Social Robot

- 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 complicates the recognition of human behavior from a first-person robot perspective.

Related Work

Fig 2. Sample images on the JPL Dataset

Fig 3. Sample images on the UTkinect First-Person Dataset

- Previous studies have primarily concentrated on action recognition using optical flow features, utilizing datasets such as the JPL dataset [1] and the UTKinect-FPD [2], recorded by humanoid mobile robots.
- The limited number of videos in these datasets (only 82 and 177, respectively) has hindered the effectiveness of deep learning methods.

Visual Action Recognition on the Doll-type Social Robots

Enhancing Perception with Human Pose Modality

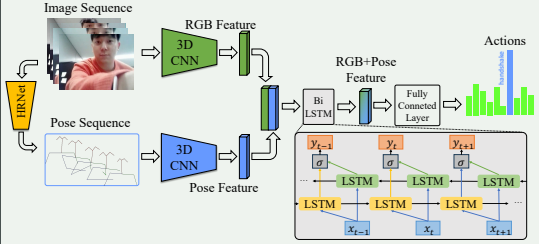


Fig 4. Architecture of Our Action Recognition Model

- We propose a multi-modality action recognition model that combines RGB-based and skeleton-based features by sampling 32 frames from the video, extracting RGB features using I3D and skeleton features using PoseC3D.
- Our approach, utilizing either LSTM or Fully Connected Layer for feature fusion across RGB and Pose, outperforms in recognizing human interactions and addressing challenges like partial truncation or camera movement effects.

Introducing Nonverbal Human Robot Interaction with the Doll-type Social Robot

Fig 5. ETRI Social Interaction (ESI) Dataset Overview

- We introduces a novel dataset that broadens the scope of human social behavior recognition in robot interaction by including implicit attitudes towards robots.
- Our ESI dataset covers a wide range of subtle yet significant behaviors, such as head nodding, head shaking, arm crossing, and Zone out, offering a comprehensive understanding of nonverbal communication for improved robot interaction.

Experimental Results

Method	JPL	UTKinect-FPD	ESI
Scratch-training			
TSN Flow	75.00	42.11	84.2
I3D RGB	73.81	62.11	85.0
MotionBert	79.17	70.53	60.4
I3D Flow	82.14	80.00	89.4
TSN RGB	77.38	83.16	80.0
PoseC3D Limb	84.52	86.32	86.8
PoseC3D Joint	80.36	88.42	91.8
Ours fc-head	87.49	93.68	92.0
Ours lstm-head	88.10	94.74	92.7
Pretrained on the Kinetics 400 dataset			
TSN Flow	80.36	73.68	89.4
TSCF	94.4	84.4	-
TSN RGB	89.29	89.47	74.4
I3D RGB	92.26	91.58	89.6
Three-stream I3D	98.5	91.58	-
PoseC3D Limb	95.24	92.63	90.0
PoseC3D Joint	95.24	93.68	92.2
Ours fc-head	95.24	96.84	92.4
Ours lstm-head	95.83	95.79	93.8

- Our model outperforms previous methods on UTKinect-FPD[2] with an outstanding accuracy of 96.84%, demonstrating the robustness of combining Pose and RGB modalities for social interaction recognition.
- Even without pre-training on the Kinetics400 dataset, our model achieves an impressive accuracy of 92.7% on the ESI dataset.

Conclusion

- We developed a multi-modal fusion by Bi-LSTM for the intricacies of human behavior recognition from the viewpoint of doll-type robots.
- The proposed ESI dataset provides comprehension of human nonverbal behavior, paving the way for improved interactions with social robots.

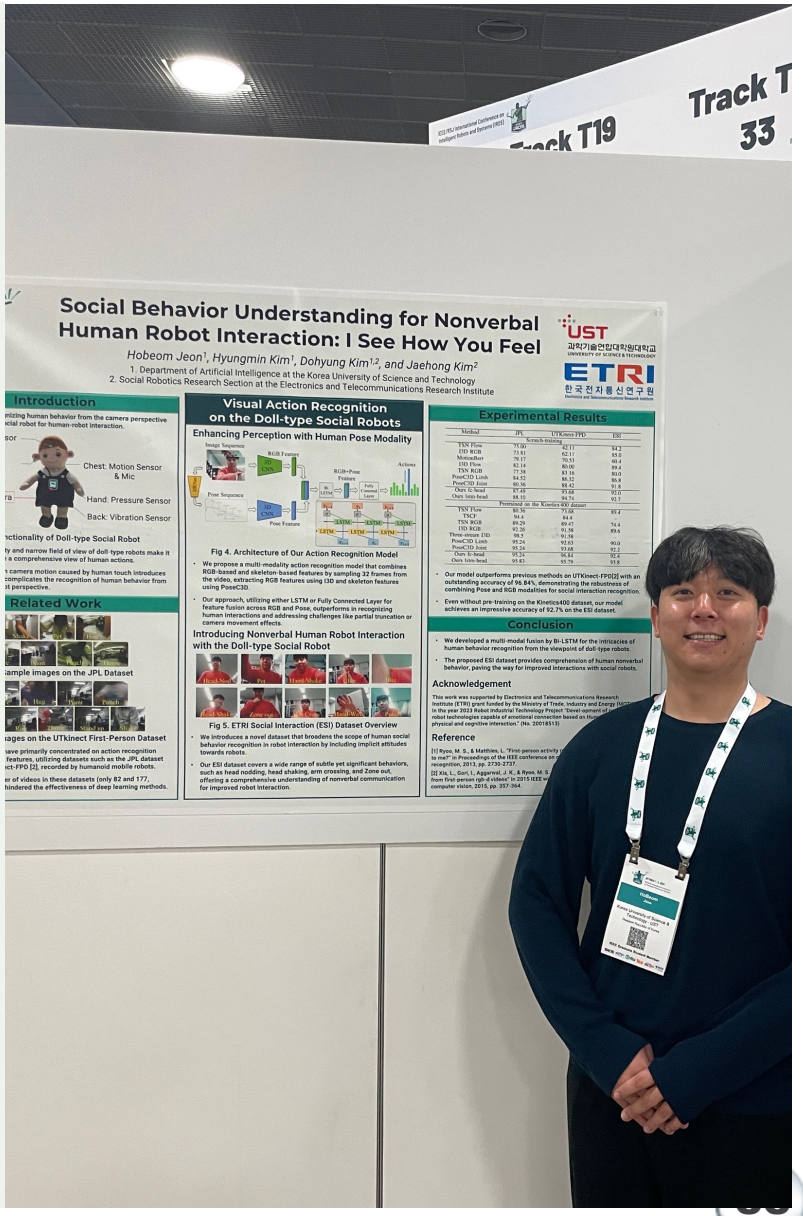
Acknowledgement

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Ministry of Trade, Industry and Energy (MOTIE) in the year 2023 Robot Industrial Technology Project "Devel-opment of companion robot technologies capable of emotional connection based on Human-Robot physical and cognitive interaction." (No. 20018513)

Reference

[1] Ryoo, M. S., & Matthies, L. "First-person activity recognition: What are they doing to me?" in Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 2730-2737.

[2] Xia, L., Gori, I., Aggarwal, J. K., & Ryoo, M. S., "Robot-centric activity recognition from first-person rgb-d videos" in 2015 IEEE winter conference on applications of computer vision, 2015, pp. 357-364.



Thank you

Q & A

HoBeom Jeon

UST-ETRI
Social Robotics Research Section

2023.10.26