Social Behavior Recongition for Nonverbal Human Robot Interaction

Topic, Related work, and Motivation

HoBeom Jeon

UST-ETRI
Social Robotics Research Section

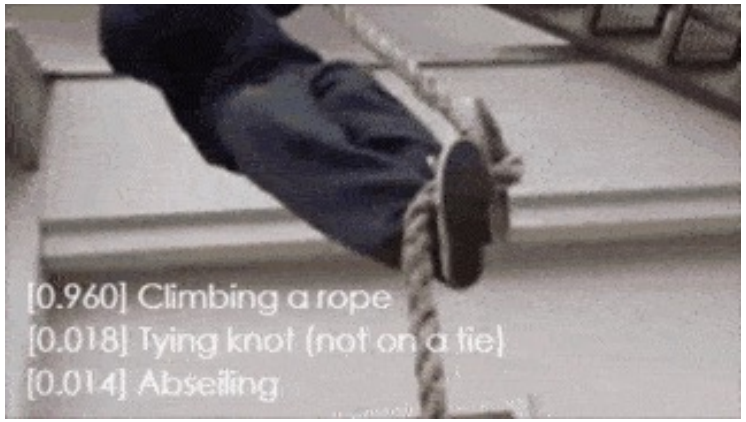
2023.09.14

Contents

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.



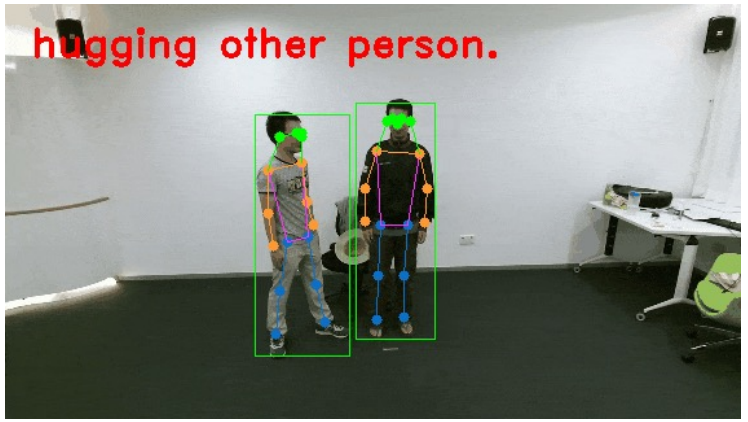


Image Source: [mmaction2]

Enhancing Human-Robot Interaction through HAR

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

Industrial collaborative robots

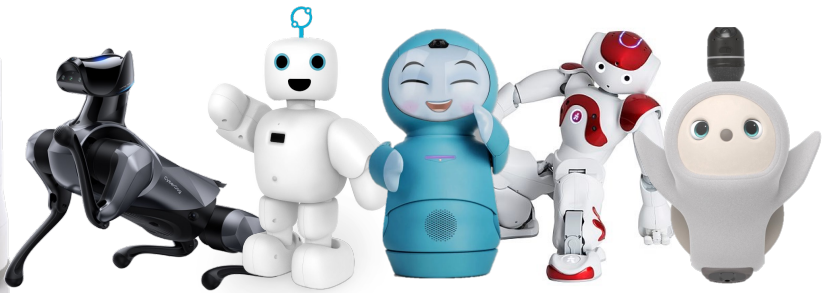
Image Source: MobileAutomation

Image Source: KUKA

Social robots

PAL ROBOTICS
TIAGo Pro

LG CLOi



Xiaomi
CyberDog2

PiBo

Moxie

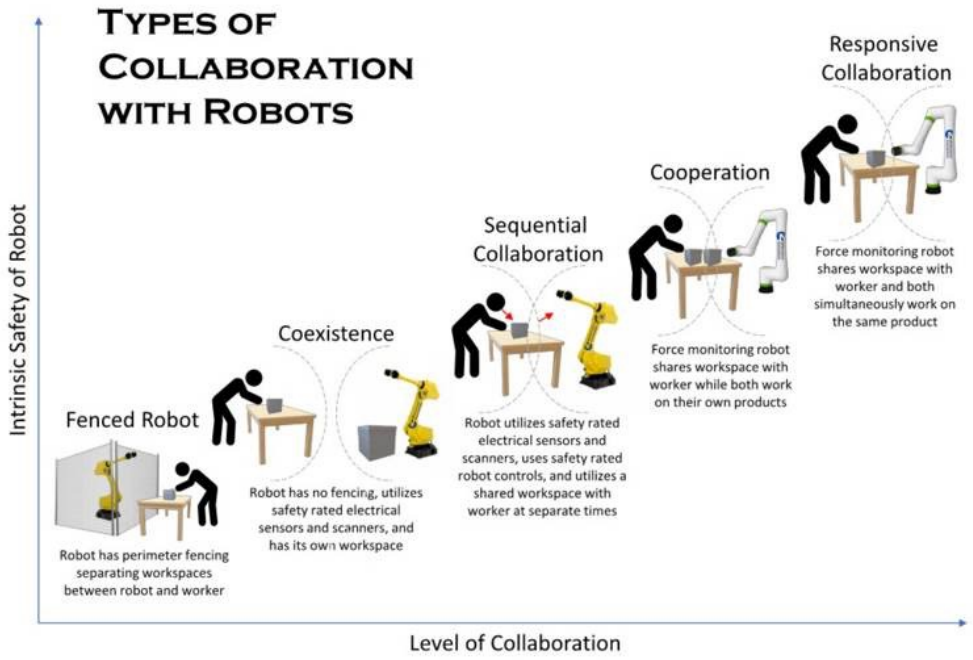
Nao

GrooveX
Lovot

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

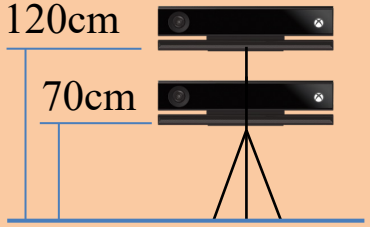
Social robots

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

Image Source: Moxie

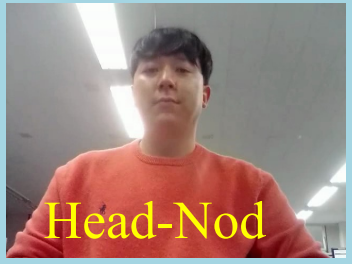
Different Viewpoint Scenario of Social Robots

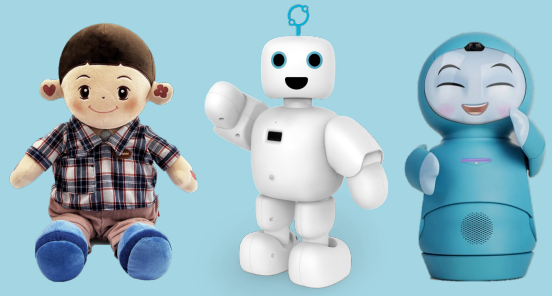
(a) Fixed camera: 3rd person robot view



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

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





Related Datasets – 1) 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

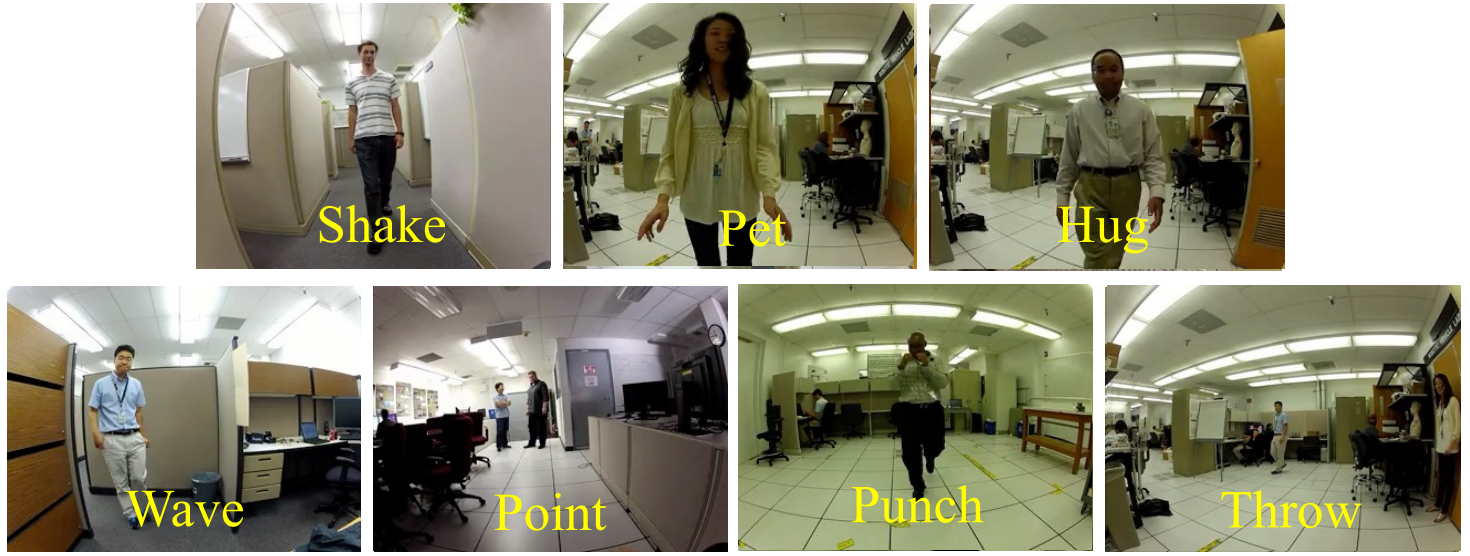
Michael S. Ryoo
Stony Brook University; Robotics at Google
cs.stonybrook.edu의 이메일 확인됨 - 홈페이지
Robotics Computer Vision Machine Learning

(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

Related Datasets – 1) 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

Related Datasets – 1) JPL Dataset

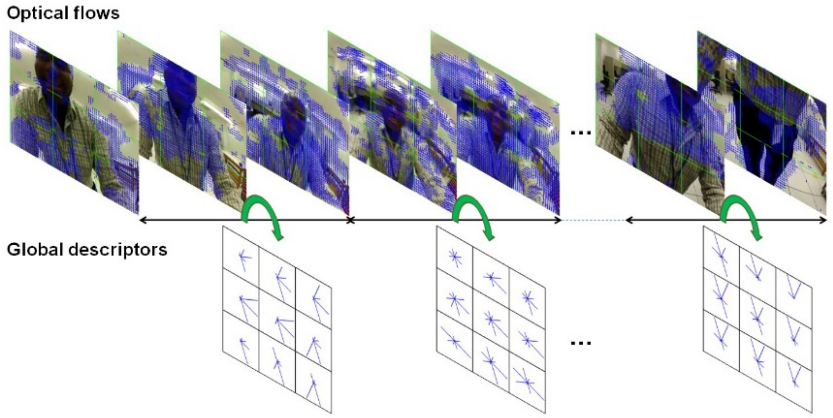


Figure 3. Example global motion descriptors obtained from a video of a human hugging the observer, which concatenates observed optical flows. These three descriptors (obtained during different types of ego-motion of the camera) are distinct, suggesting that our descriptors correctly captures observer ego-motion.

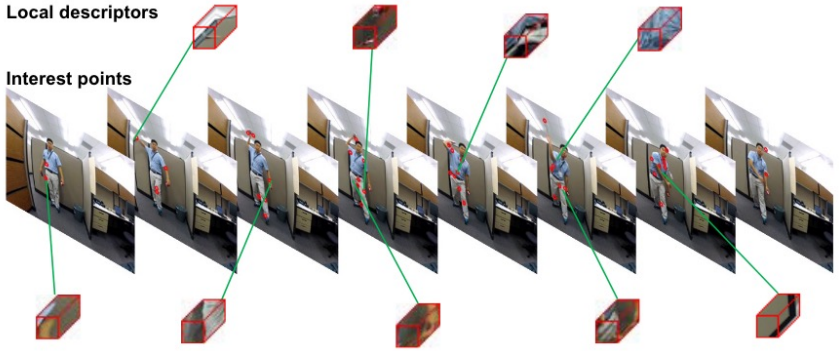


Figure 4. Example local motion descriptors obtained from our video of a person throwing an object to the observer. Locations with salient motion are detected, and their 3-D XYT volume patches are collected as our local descriptors.

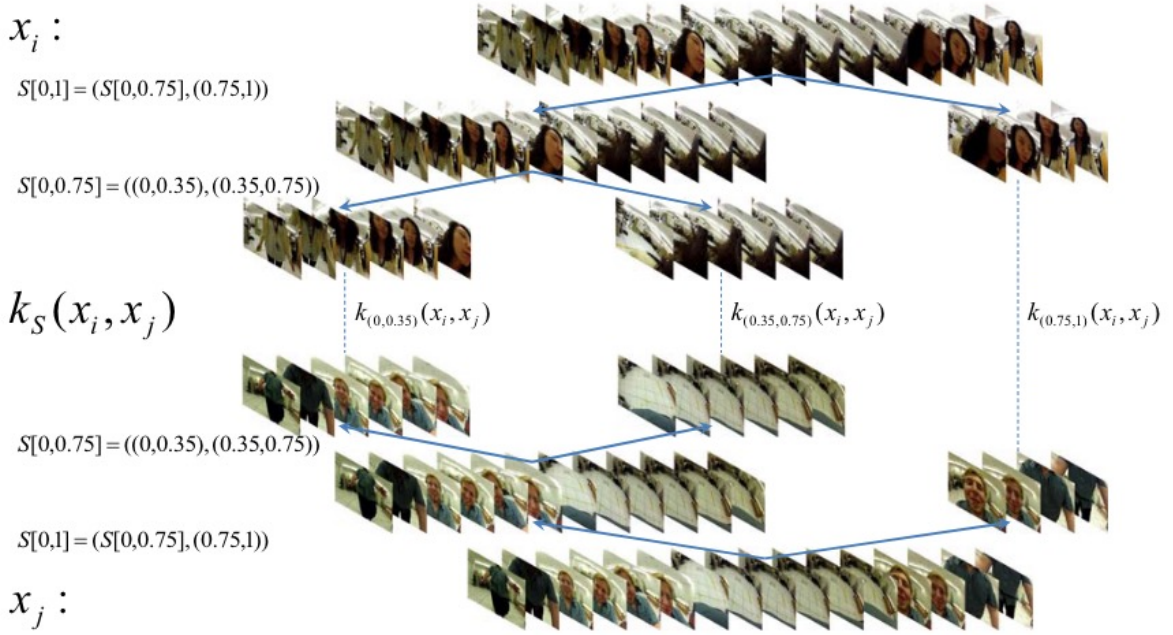


Figure 6. An example matching between two hugging videos, x_i and x_j, using the kernel K_S constructed from the hierarchical structure $S = (((0, 0.35), (0.35, 0.75)), (0.75, 1))$.

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

Related Datasets – 2) NUS Dataset

Action and Interaction Recognition in First-person videos

Sanath Narayan
Dept. of Electrical Engg.,
IISc, Bangalore
sanath@ee.iisc.ernet.in

Mohan S. Kankanhalli
School of Computing,
NUS, Singapore
mohan@comp.nus.edu.sg

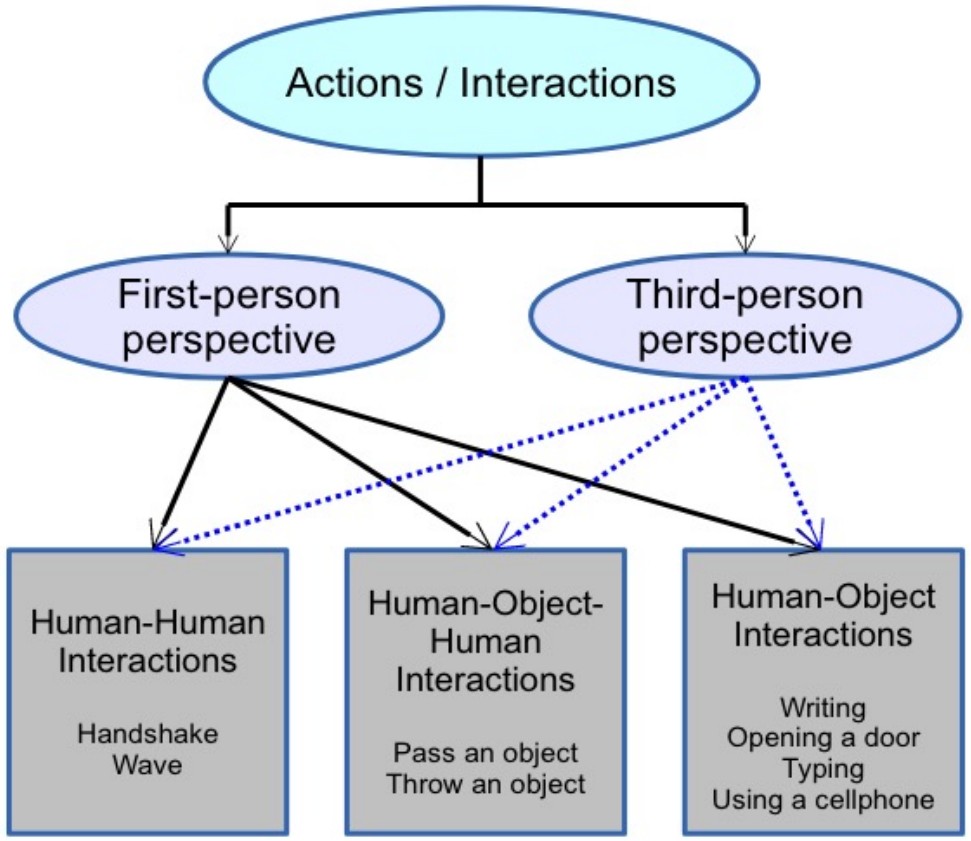
Kalpathi R. Ramakrishnan
Dept. of Electrical Engg.,
IISc, Bangalore
krr@ee.iisc.ernet.in

Class Labels

1. Cell
2. Door
3. Pass
4. Shake
5. Throw
6. Type
7. Wave
8. Write

[2] Action and Interaction Recognition in First-person videos (2014CVPRW)

Related Datasets – 2) NUS Dataset



[2] Action and Interaction Recognition in First-person videos (2014CVPRW)

Related Datasets – 2) NUS Dataset

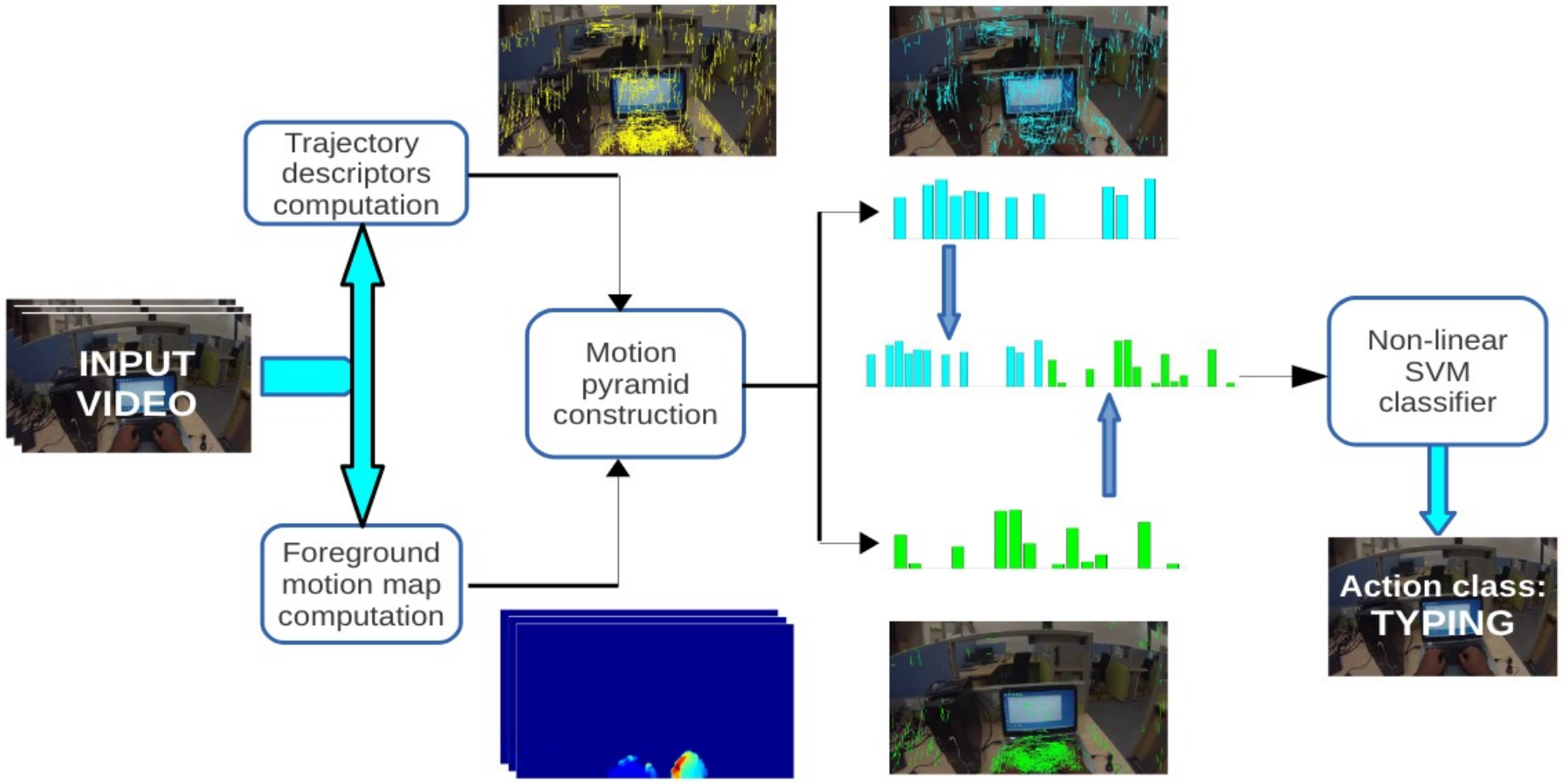


Figure 2. Illustration of the interaction recognition approach for first-person videos. The baseline improved trajectory features are computed. The foreground motion map for every frame is computed by multiplying the foreground mask and the motion magnitude map. The trajectory score is computed by adding the foreground motion map scores of the pixels the corresponding trajectory passes through. Based on the relative scores, different trajectories and corresponding features are grouped together and descriptors for each group (figure two groups) are computed. Concatenating the group descriptors results in a single descriptor for the video which is used for classification.

[2] Action and Interaction Recognition in First-person videos (2014CVPRW)

Related Datasets – 3) 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

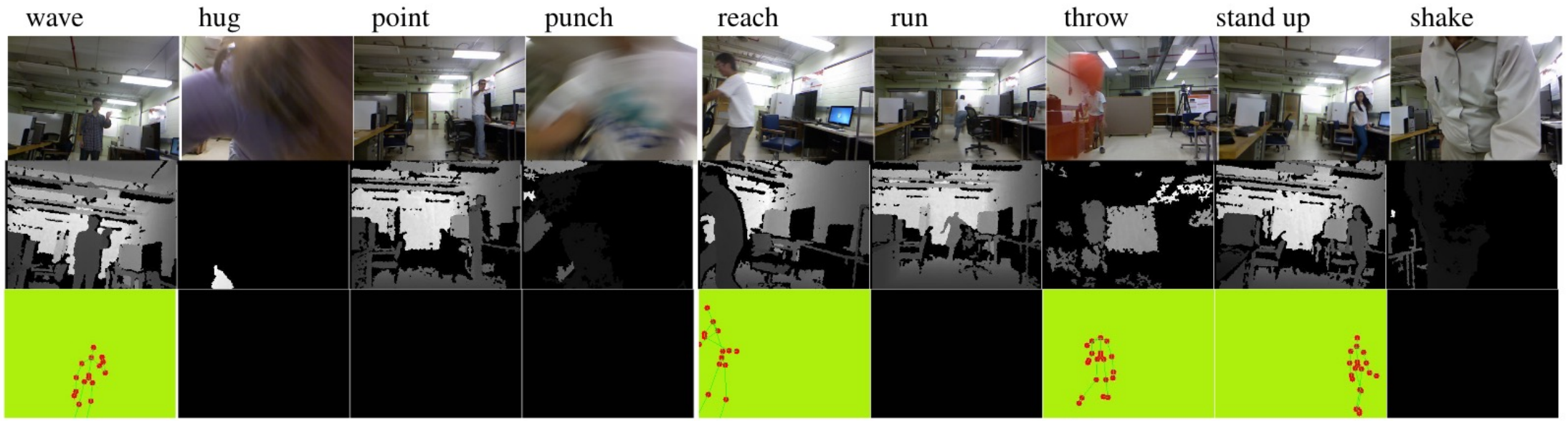
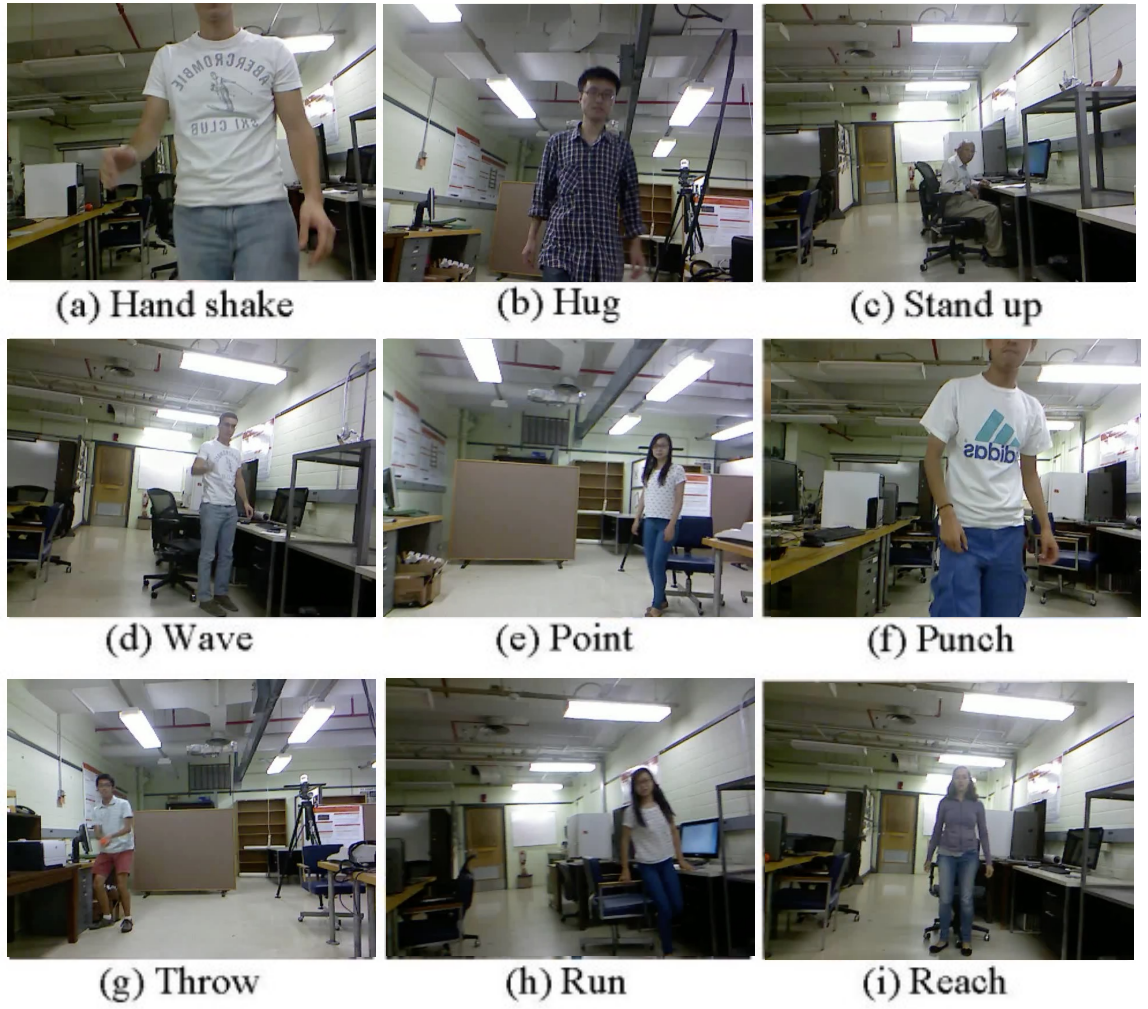


Table 1: Sample images of 9 activities in the humanoid robot first-person RGBD dataset. The first and second rows present the RGB and depth images, respectively. The last row represents skeleton images. If no skeleton is detected for a particular frame, a black image is shown.

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

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

Related Datasets – 3) 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

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

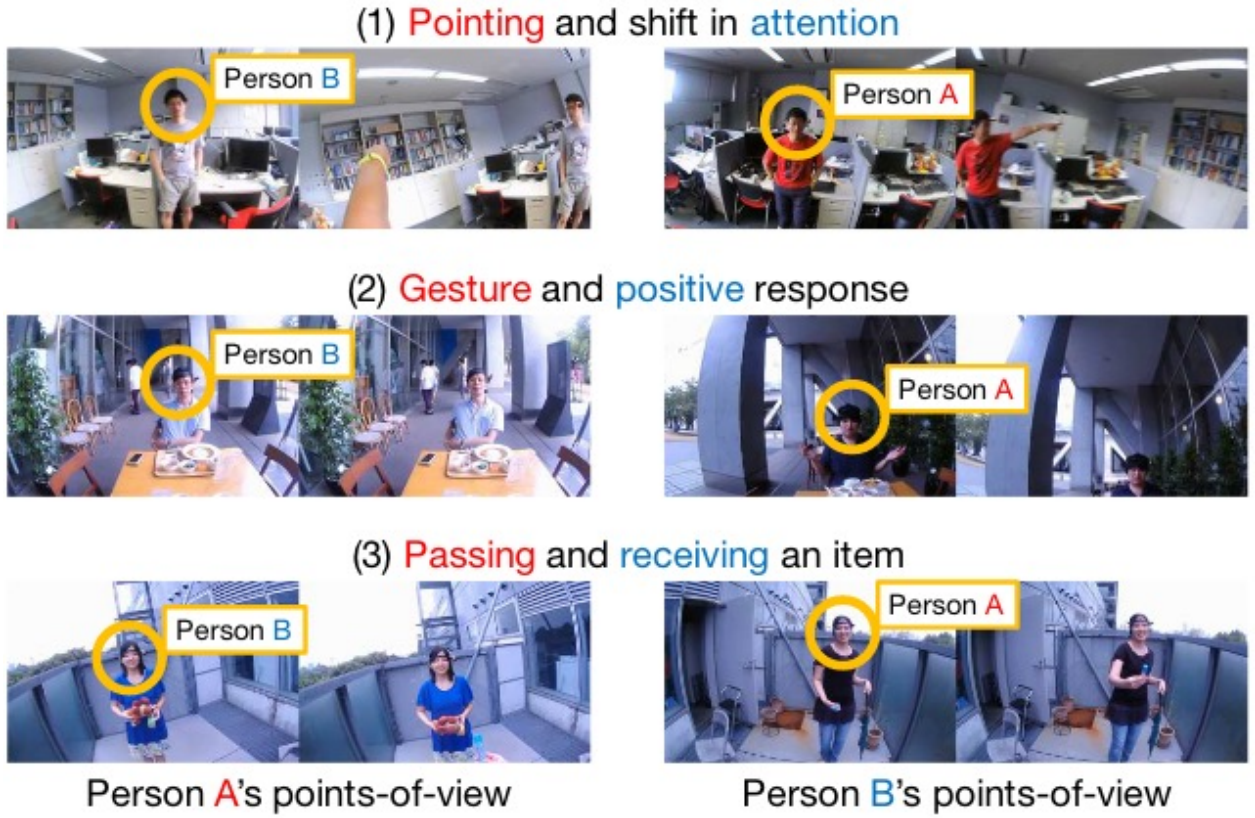
Related Datasets – 4) PEV

Recognizing Micro-Actions and Reactions from Paired Egocentric Videos

Ryo Yonetani
The University of Tokyo
Tokyo, Japan
yonetani@iis.u-tokyo.ac.jp

Kris M. Kitani
Carnegie Mellon University
Pittsburgh, PA, USA
kkitani@cs.cmu.edu

Yoichi Sato
The University of Tokyo
Tokyo, Japan
ysato@iis.u-tokyo.ac.jp



- Human-Human interaction on Ego-centric view
- Indirectly, robots can learn social behavior
- situations of communicating with others from afar

[4] Recognizing micro-actions and reactions from paired egocentric videos. CVPR2016 Yonetani

Related Datasets – 4) PEV

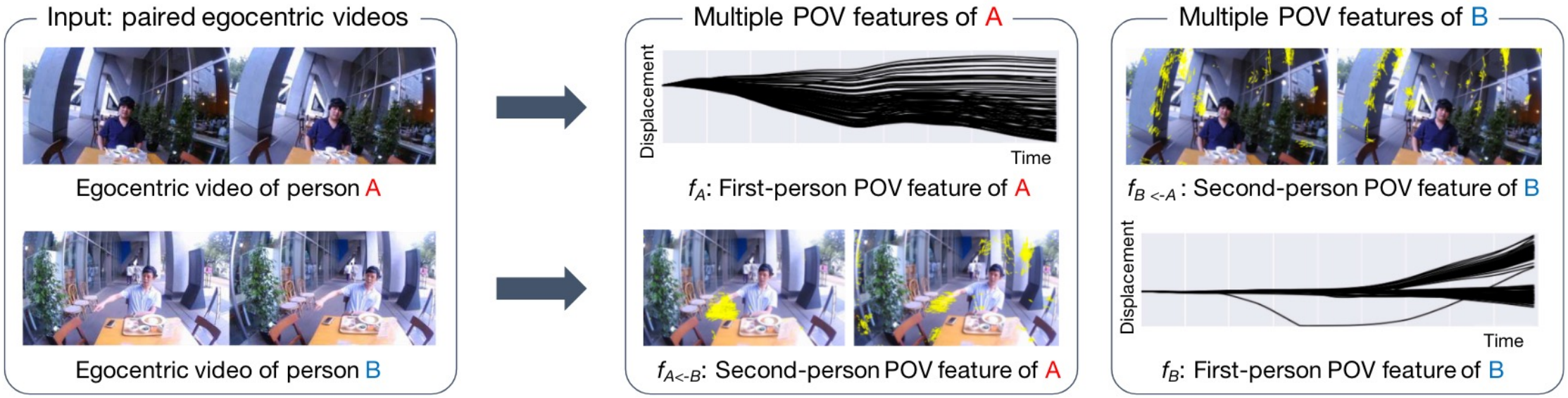


Figure 2. Our approach. Paired egocentric videos recorded by persons A and B are used to provide first-person and second-person POV features of both A and B, which are complementary and essential for recognizing micro-actions and reactions. Cumulative displacement patterns [27] and improved dense trajectories [39] are respectively visualized as examples of the first-person and second-person features.

- A new approach to using both person's videos.
- In contrast, difficult to use both videos in the real world.
- Also, it contains only human-human intentional interaction.

[4] Recognizing micro-actions and reactions from paired egocentric videos. CVPR2016 Yonetani

Related Datasets - Table

Datasets	action type	view	resolution	#videos	#clips	#actions	#subjects	year
JPL [1]	human-robot	robot-1st	320x240	57	84	7	8	2013
NUS [2]	human-human	human-1st human-3rd	1280×720	153 133	153 133	8		2014
UTKinect-FPD [3]	human-robot	robot-1st	640x480	127	268	9	8	2015
PEV [4]	human-human	human-1st	320x180	2452	1226	7	6	2016
Something v2 [5]	human-object	human-1st		220,847	220,847	174	-	2017
MHHRI [6]	human-human -robot	human-1st robot-1st	640x480	48	746	-	18	2017
Epic-kitchens [7]	human-object	human-1st	640x480	89,979	89,979	4,025		2020

[1] First-person activity recognition: What are they doing to me? Ryoo
[2] Action and interaction recognition in first-person videos. CVPRW2014, Narayan
[3] Robot-centric activity recognition from first-person rgb-d videos. WACV2015, Xia Lu
[4] Recognizing micro-actions and reactions from paired egocentric videos. CVPR2016 Yonetani
[5] The "something something" video database for learning and evaluating visual common sense. (2018) R. Goyal
[6] Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement. (2017) Celiktutan
[7] The epic-kitchens dataset: Collection, challenges and baselines. (2020) D. Damen

Related Works – 1) CNN + LSTM

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

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

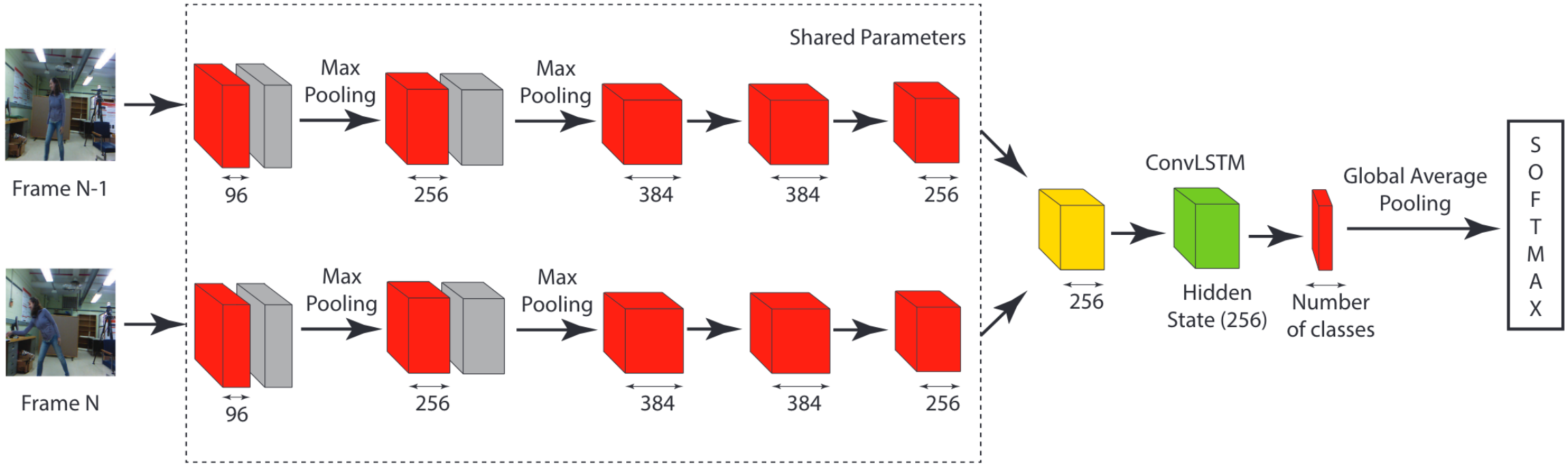


Figure 2. The architecture of the network. The convolutional layers are shown in red followed by normalization layer in gray. The 3D convolutional layer is shown in yellow and the convLSTM layer in green. We also experiment with a variant where, instead of raw frames we input difference-images obtained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

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

Related Works – 1) CNN + LSTM

		JPLFPID	NUSFPID	UTKinect First Person Dataset	
				Humanoid	Non-humanoid
[1]	Ryoo & Matthies [29]	89.6	-	57.1	58.48
Previous [1] iccv, 2011	Ryoo [28]	87.1	-	-	-
CVIU, 2016	Abebe *et al.* [1]	86	-	-	-
EUSIPCO, 2017	Ozkan *et al.* [24] + DogCentric	87.4	-	-	-
[2]	Narayan *et al.* [22]	**96.7**	61.8	61.9	57.6
ICCV, 2013	Wang and Schmid [38]	-	58.9	-	-
[1]	HOF [29]	-	-	45.92	-
CVPR, 2008	Laptev *et al.* [15]	-	-	48.46	50.83
	LRCN [6] (raw frames)	59.5	68.9	72.6	71.4
CVPR, 2015	LRCN [6] (difference of frames)	89.0	69.1	63.1	67.8
	Proposed Method (raw frames)	70.6	69.4	**79.6**	**78.4**
	Proposed Method (difference of frames)	91.0	**70.0**	66.7	69.1

Table 1. Comparison of the proposed method with existing techniques, on various datasets. Results are reported in terms of recognition accuracy in %. [22] on UTKinect are our reproduced results following the paper description. LRCN [6] results are produced from the authors' code following same training/testing protocol used with our method. All other results are those available from the authors' papers.

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

Related Works – 2) Body Part Detection

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

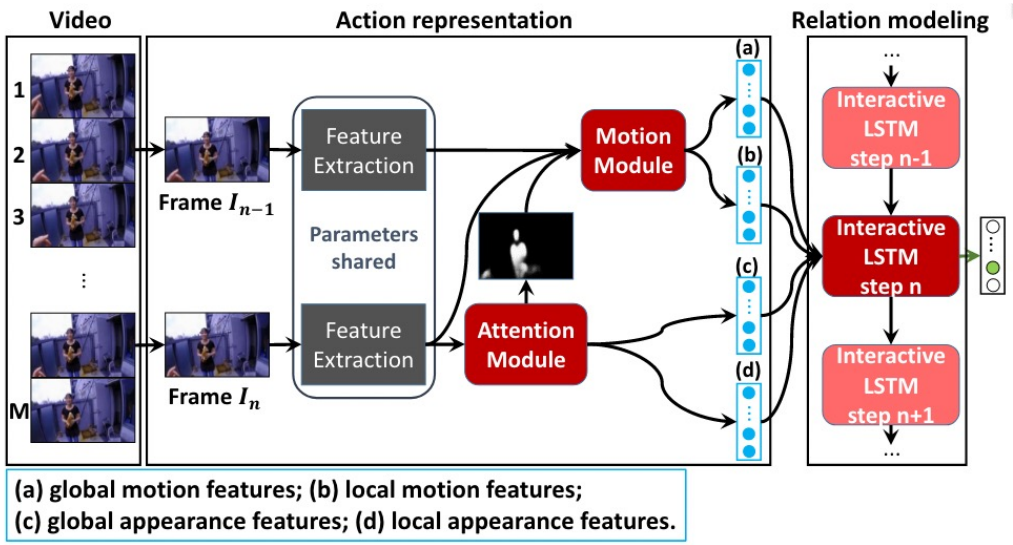


Figure 2. Proposed framework. Frames $\boldsymbol{I}_i (i = 1, ..., N)$ are sam-

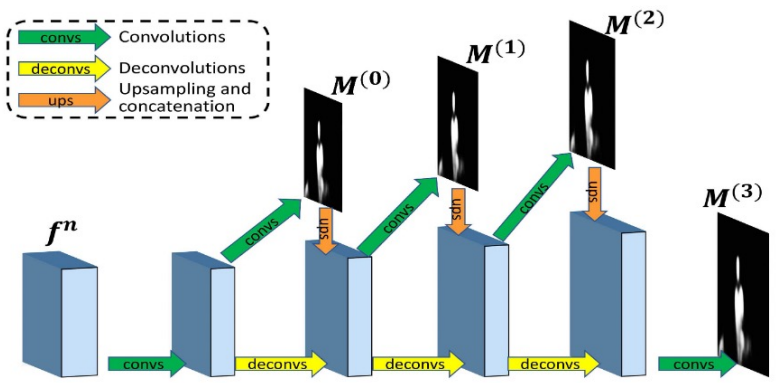


Figure 3. Structure of attention module. The module takes feature

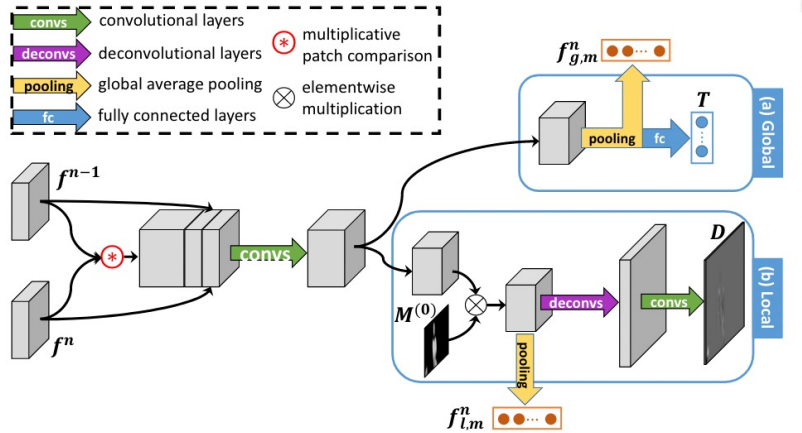


Figure 4. Structure of motion module. The module takes basic

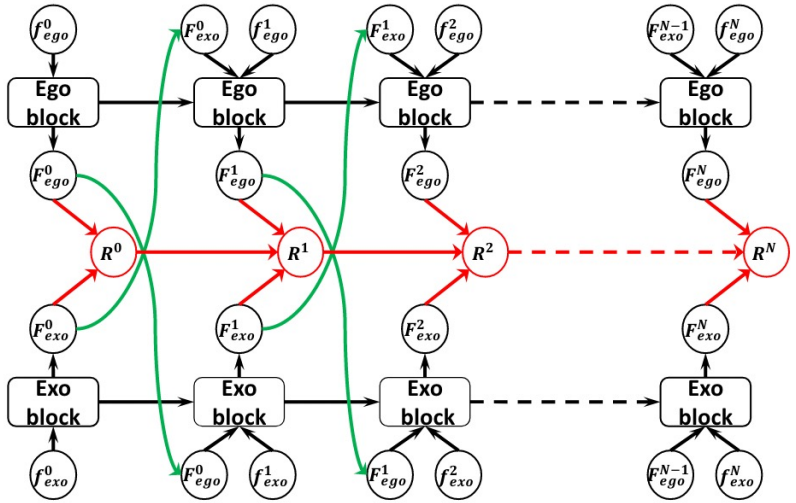


Figure 5. Diagram of Interactive LSTM. The unrolled symmetrical

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

Related Works – 2) Body Part Detection

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

	Methods	PEV	NUS(first h-h)	NUS(first)	JPL
[1]	RMF[1]	-	-	-	86.0
[1]	Ryoo and Matthies[31]	-	-	-	89.6
[2]	Narayan *et al.* [25]	-	74.8	77.9	96.7
	Yonetani *et al.* [39] (single POV)	60.4	-	-	75.0
[8]	convLSTM[35] (raw frames)	-	-	69.4	70.6
[8]	convLSTM[35] (difference of frames)	-	-	70.0	90.1
	LRCN[6]	45.3	65.4	70.6	78.5
	TRN[41]	49.3	66.7	74.7	84.2
	Two-stream[33]	58.5	78.6	80.6	93.4
	Our method	**64.2**	**80.2**	**81.8**	**98.4**
	Yonetani *et al.* [39] (multiple POV)	69.2	-	-	-
	Our method (multiple POV)	69.7	-	-	-

Table 1. State-of-the-art comparison (%) with existing methods. *NUS(first h-h)* denotes the first-person human-human interaction subset of NUS dataset and *NUS(first)* denotes the first-person subset. It is notable that only PEV dataset provides multiple POV videos so that no multiple POV result of other datasets is reported.

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

Related Works – 3) Human Detection

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

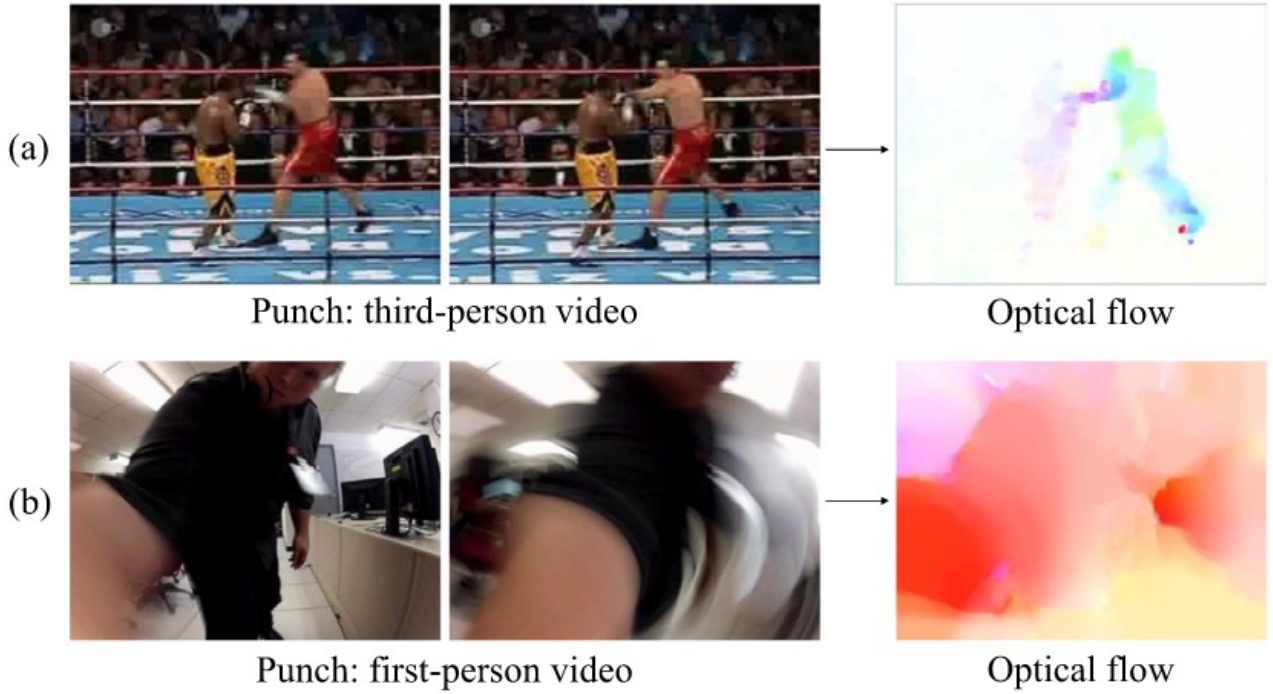


Fig. 1. Punch action in third-person and first-person videos. The two video clips have different characteristics in terms of appearance and motion. (a) The optical flow is extracted from a third-person video, where the camera is fixed. (b) The optical flow is extracted from a first-person video where the camera shakes considerably.

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

Related Works – 3) Human Detection

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

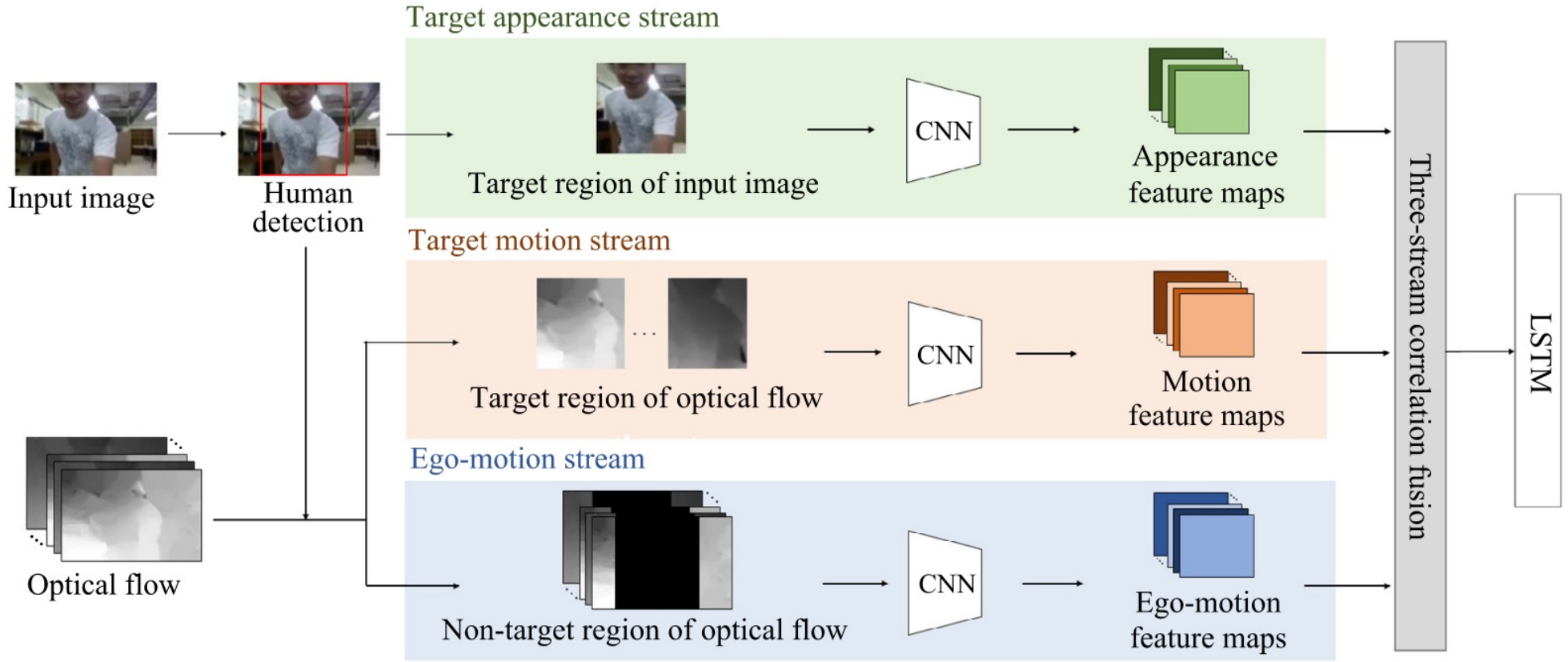


Fig. 2. Overview of the three-stream fusion network. Our proposed network is composed of three-stream architecture, the three-stream correlation fusion (TSCF), and a long short-term memory model. Each stream of the three-stream architecture respectively extracts appearance, motion, and ego-motion feature 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.

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

Related Works – 3) Human Detection

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

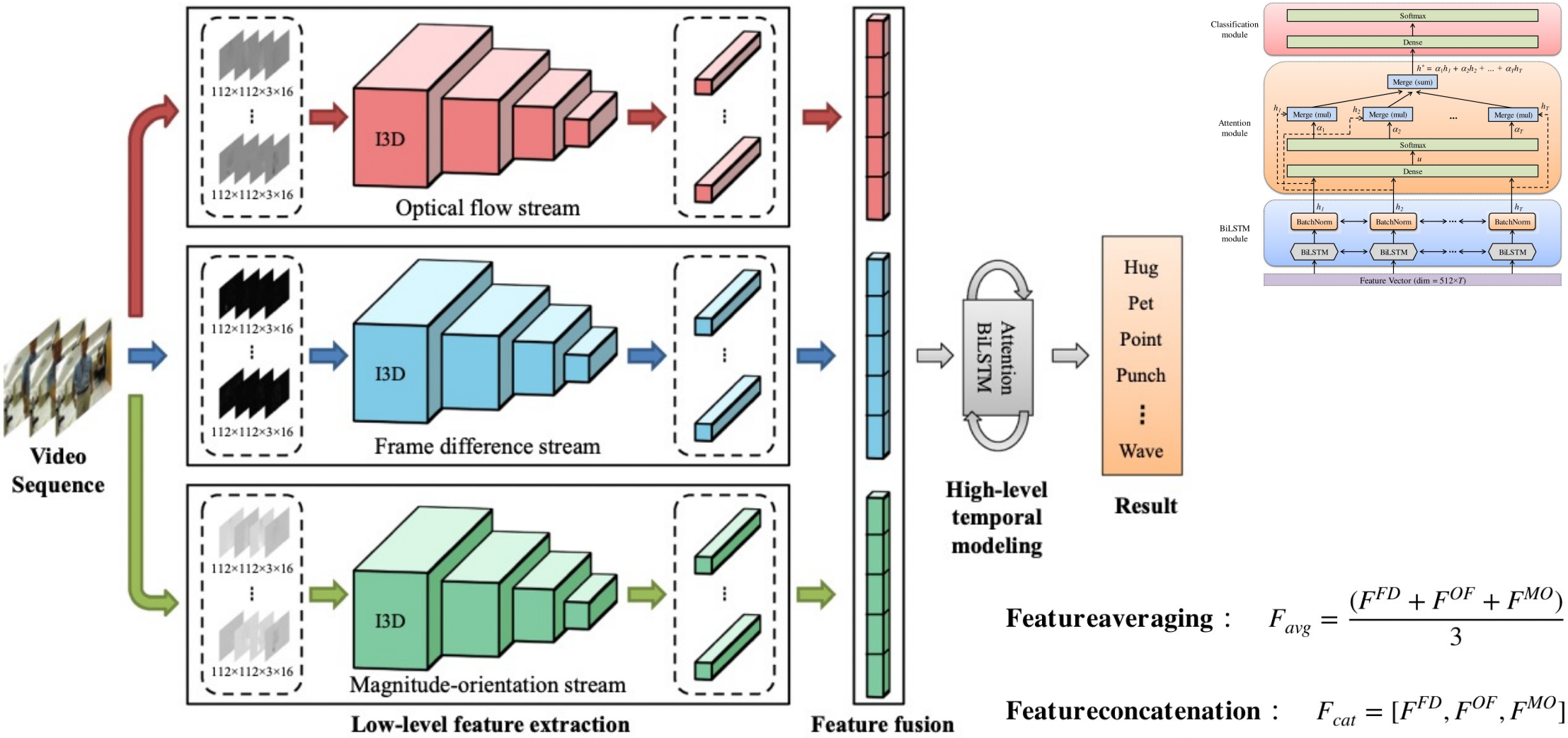
Methods	JPL (%)
Ryoo et al. [9]	89.6
Boosted MKL [38]	87.4
Two-Stream ConvNet [4]	54.2
LRCN (RGB frame) [25]	59.5
LRCN (difference of frames) [25]	89.0
KRP FS (RGB frame) [50]	73.8
KRP FS (difference of frames) [50]	85.7
SeDyn [10] + FTP [52]	92.9
Sudhakaran et al. (RGB frame) [41]	70.6
Sudhakaran et al. (difference of frames) [41]	91.0
Ours (RGB frame)	88.0
Ours (difference of frames)	**94.4**

Methods	UTK (%)
Ryoo et al. [9]	57.1
Laptev et al. [13]	48.4
Two-Stream ConvNet [4]	65.9
LRCN (RGB frame) [25]	72.6
LRCN (difference of frames) [25]	63.1
KRP FS (RGB frame) [50]	35.6
KRP FS (difference of frame) [50]	33.3
Sudhakaran et al. (RGB frame) [41]	79.6
Sudhakaran et al. (difference of frames) [41]	66.7
Ours (RGB frame)	83.1
Ours (difference of frames)	**84.4**

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

Related Works – 4) Optical Flow Magnitude

➢ CNN + LSTM (**Optical Flow** + **Magnitude Orientation**) (2022)



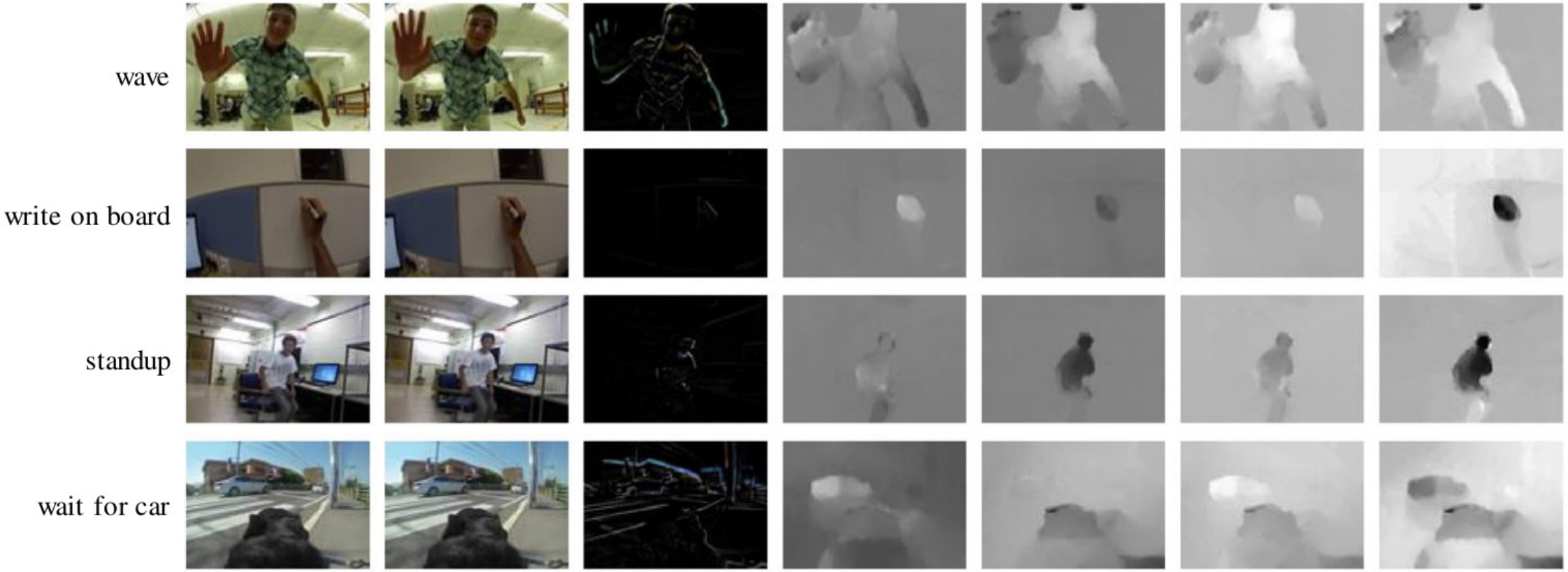
$$\textbf{Featureaveraging}: \quad F_{avg} = \frac{(F^{FD} + F^{OF} + F^{MO})}{3}$$

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

[11] 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

Related Works – 4) Optical Flow Magnitude

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



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

[11] 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

Related Works – 4) Optical Flow Magnitude

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

Table 7 Performance comparison on JPL dataset using different methods

Method	Accuracy (%)
Two-stream ConvNet (Simonyan and Zisserman 2014)	65.9
Global VRTD (Moreira et al. 2017)	84.0
RMF (Abebe et al. 2016)	86.0
Naïve VR + VRTD HOG (Moreira et al. 2020)	86.2
Boosted MKL (Özkan et al. 2017)	87.4
LRCN (frame difference) (Donahue et al. 2015)	89.0
Structure match (Ryoo and Matthies 2013)	89.6
PoT + ITF (Ryoo et al. 2015)	89.8
ConvLSTM (frame differnce) (Sudhakaran and Lanz 2017)	91.0
TDD + FV (Wang et al. 2015)	91.7
SeDyn+FTP (Zaki et al. 2017)	92.9
Global + local C3D (Fa et al. 2018)	92.9
TSCF (frame difference) (Kim et al. 2020)	94.4
ITF (Wang and Schmid 2013)	96.1
ITF + FV (Narayan et al. 2014)	96.7
DRM + Interactive LSTM (Li et al. 2019)	98.4
ConvNet + HHT (Purwanto et al. 2019)	**98.5**
Ours (Three-stream I3D + Attn-BiLSTM)	**98.5**

Table 9 Performance comparison on UTK dataset using different methods

Method	Accuracy (%)
LRCN (frame difference) (Donahue et al. 2015)	63.1
Two-stream ConvNet (Simonyan and Zisserman 2014)	65.9
ConvLSTM (frame difference) (Sudhakaran and Lanz 2017)	66.7
LRCN (frames) (Donahue et al. 2015)	72.6
ConvLSTM (frames) (Sudhakaran and Lanz 2017)	79.6
TSCF (frames) (Kim et al. 2020)	83.1
TSCF (frame difference) (Kim et al. 2020)	84.4
Ours (three-stream I3D + Attn-BiLSTM)	**91.5**

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

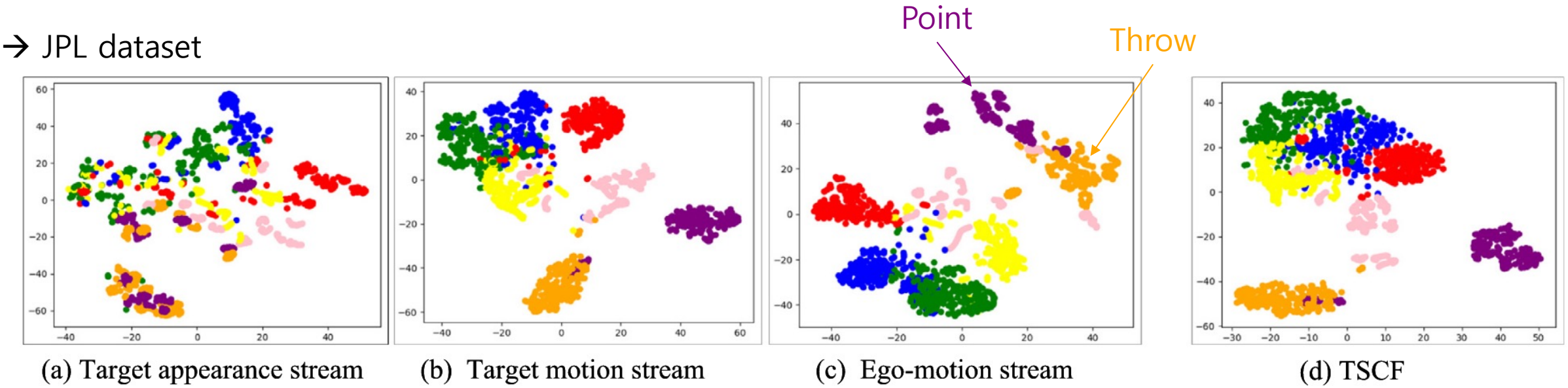


Fig. 9. t-SNE results for the JPL First-Person Interaction dataset. The results in (a), (b), and (c) represent the feature vectors of the target appearance stream, target motion stream, and ego-motion stream, respectively. (d) represents the feature vectors of the three-stream correlation fusion. The 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.)

[10] 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

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

Motivation - Dataset Analysis

→ UTKinect FPD

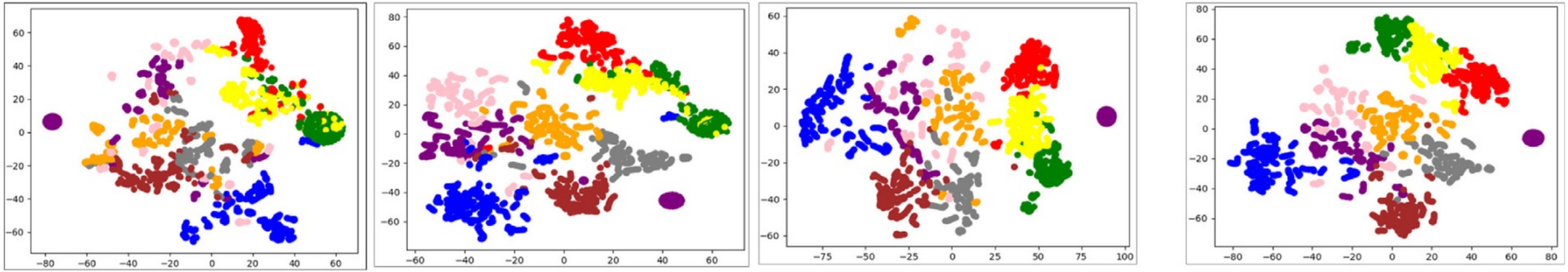


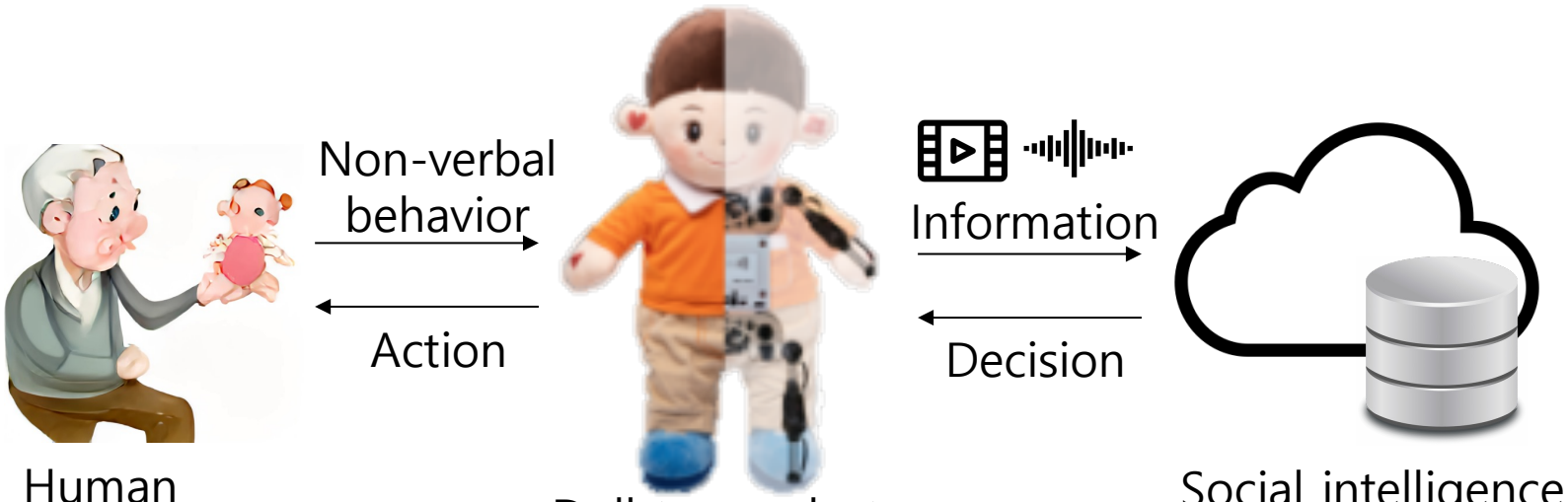
Fig. 8. t-SNE results for the UTKinect-FirstPerson (humanoid) dataset. (a), (b), and (c) represent the feature vectors of the target appearance stream, the target motion stream and the ego-motion stream, respectively. (d) represents the feature vectors of the three-stream correlation fusion. The color 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.)

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

Motivation

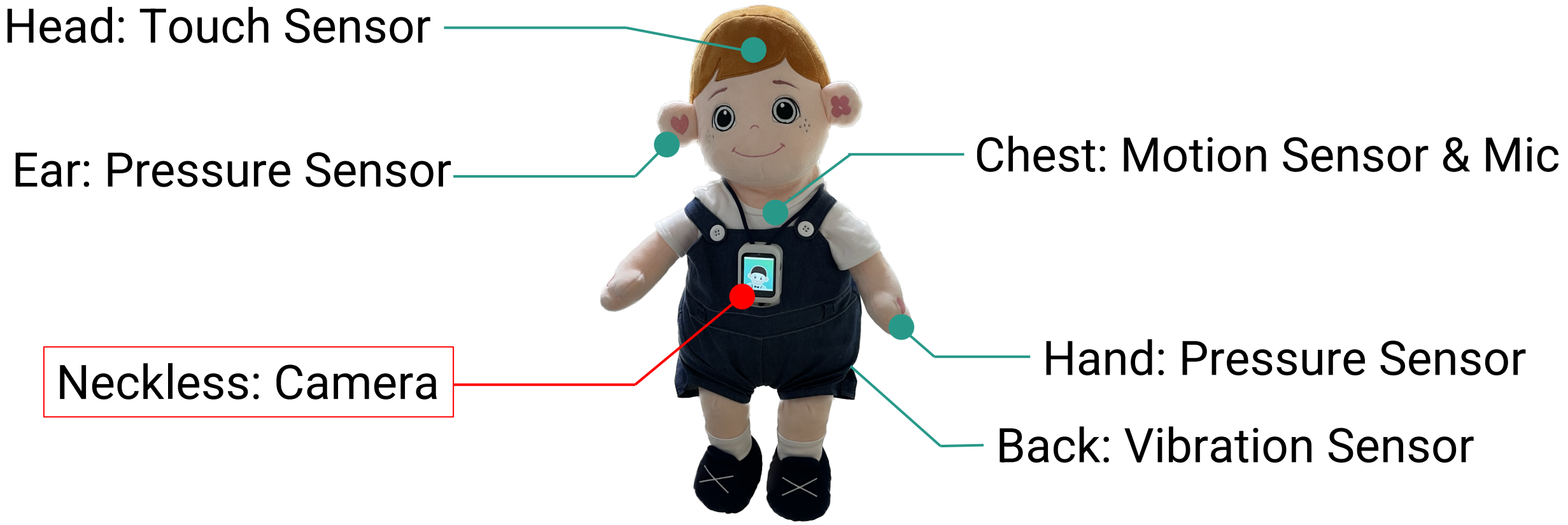
1. Enhancing Emotional Interaction Between Robots and Humans through non-verbal Behavior Recognition
   - Existing datasets have only included informational or expressive actions.
     → Excluding self-movement such as standing, reaching
     → Adding unconscious motions

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.



Preview of next presentation

- The function of Doll-type Social Robot



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

Preview of next presentation

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.

Thank you

Topic, Related work, and Motivation

HoBeom Jeon

UST-ETRI
Social Robotics Research Section

2023.09.14