Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

Bonn-Aachen
International Center for
Information Technology

Lifelong Action Learning for Socially Assistive Robots

November 28th, 2022

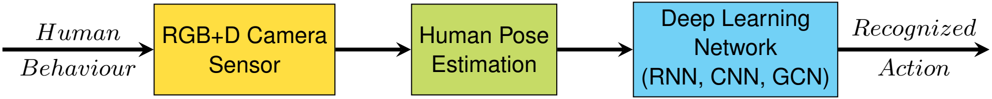
Hasnainali Walli

Advisors

Prof. Dr. Paul G. Plöger, Prof. Dr. Sebastian Houben, Alex Mitrevski

Motivation

- Action recognition is a key function for socially assistive robots
- **Challenge**: Conventional models' inability to learn new actions
- How can robotic systems learn new actions without forgetting?



Lifelong Action Learning

- Robotic systems fine-tune their knowledge with experience
- New actions are learnt while retaining the knowledge of the previous actions
- Concept of lifelong action learning was explored in the context of CRI
- **Objectives**:
  - Develop an action learning model using incremental learning
  - Integrate model on QTRobot for the MigrAVE project

Lifelong Action Learning

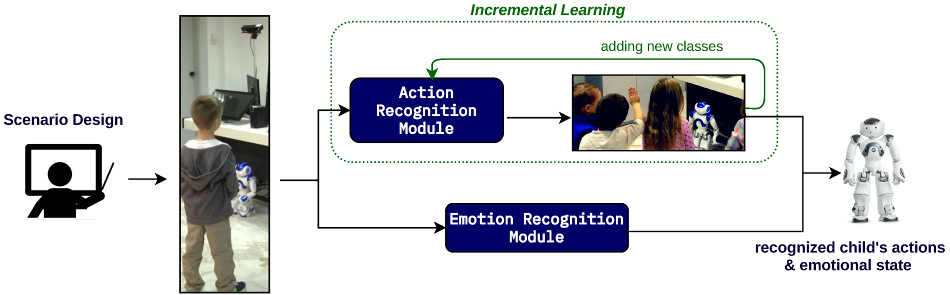


Figure 1: Incremental learning pipeline for action and emotion recognition[1]

[1] N. Efthymiou, P. P. Filntisis, G. Potamianos, and P. Maragos, "Visual Robotic Perception System with Incremental Learning for Child–Robot Interaction Scenarios," Technologies, vol. 9, no. 86, November 2021.

Lifelong Action Learning

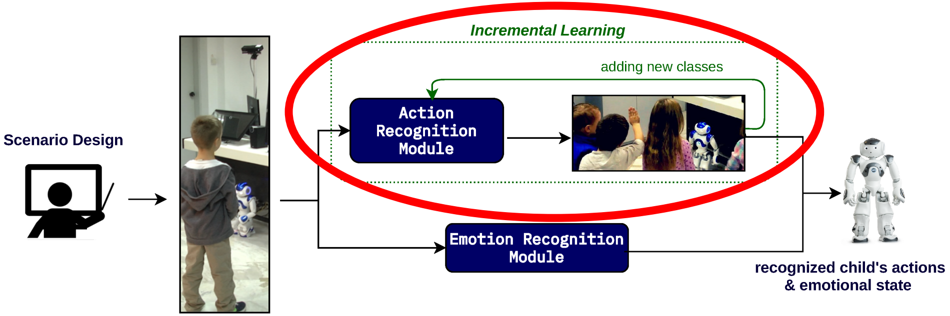


Figure 2: Incremental learning pipeline for action and emotion recognition[2]

[2] N. Efthymiou, P. P. Filntisis, G. Potamianos, and P. Maragos, "Visual Robotic Perception System with Incremental Learning for Child–Robot Interaction Scenarios," Technologies, vol. 9, no. 86, November 2021.

Lifelong Action Learning

Their Approach

- RGB+D and Optical Flow data
- TSN Network
- iCaRL Algorithm
- BabyRobot Dataset

Our Approach

- 3D Skeleton data
- CTR-GCN Network
- BiC Algorithm
- NTU RGB+D Dataset

Our Approach

Methodology

1. Perform comparative analysis on skeleton-based action recognition networks
2. Perform comparative analysis on class-incremental learning algorithms
3. Integrate final model on QTRobot

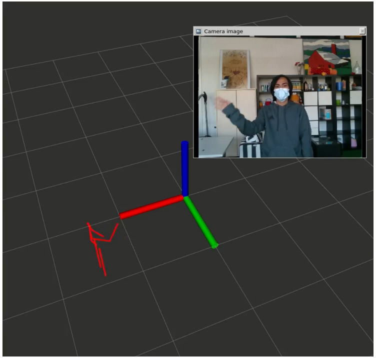


Figure 3: Hand waving action visualized in RVIZ

NTU Dataset

- Features 120 everyday actions
- 40 subjects; 3 cameras; 2 demos
- 25 skeletal joints tracked
- Evaluation:
  - Cross-Subject Accuracy: train on 20 subjects; test on 20 subjects
  - Cross-View Accuracy: train using 2 views; test on 1 view

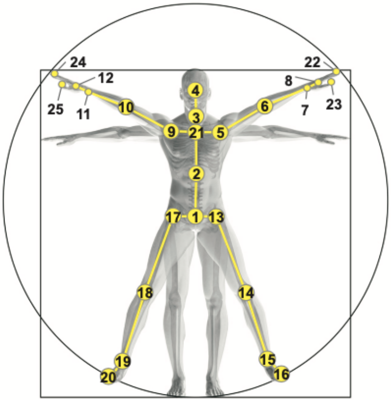


Figure 4: Joint configurations for NTU RGB-D dataset[3]

[3] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, “NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1010–1019.

Action Recognition Analysis

- Networks: CTR-GCN, MS-G3D, EfficientGCN, ViewAdaptive NN
- Joint, Bone and Joint Motions
- **Metrics**: Cross-Subject & Training Time

Drink Water	Eat Meal	Brush Teeth	Drop
Pick Up	Throw	Sit Down	Stand Up
Clapping	Hand Waving	Kick Something	Hopping
Jump Up	Play with Phone	Point to Something	Rub Hands
Nod Head/Bow	Shake Head	Wipe Face	Cross Hands

Table 1: Subset of action classes from the NTU RGB-D dataset

Action Recognition Analysis Results

Network	Cross Subject	Cross View
CTR-GCN (Joint)	92.63%	96.37%
CTR-GCN (Bone)	92.78%	96.02%
CTR-GCN (Motion)	92.51%	96.40%
MS-G3D (Joint)	91.27%	96.85%
MS-G3D (Bone)	90.90%	95.44%
EfficientGCN-B4 (SG Layer)	94.05%	97.47%
EfficientGCN-B4 (EpSep Layer)	94.43%	97.56%
VA-NN (CNN)	92.97%	92.20%

Table 2: Action recognition networks accuracy

Network	Training Time
CTR-GCN	4 hrs
MS-G3D	8 hrs
EfficientGCN-B4	5 hrs
VA-NN (CNN)	0.5 hrs

Table 3: Networks training time

Action Recognition Analysis Results

Action	CTR-GCN	MS-G3D	Action	CTR-GCN	MS-G3D
Drink Water	82.48%	83.94%	Kick Something	97.83%	94.93%
Eat Meal	78.91%	73.82%	Hopping	98.91%	95.27%
Brush Teeth	90.84%	91.21%	Jump Up	98.91%	98.55%
Drop	90.18%	91.64%	Play with Phone	86.91%	90.91%
Pick Up	98.91%	94.55%	Point to Something	92.39%	92.03%
Throw	96.36%	90.91%	Rub Hands	90.58%	89.49%
Sit Down	98.90%	97.80%	Nod Head/Bow	96.01%	95.65%
Stand Up	98.17%	98.90%	Shake Head	96.00%	95.64%
Clapping	82.42%	72.89%	Wipe Face	92.39%	94.20%
Hand Waving	94.16%	94.89%	Cross Hands	93.84%	94.57%

Table 4: Cross-Subject accuracy results per class for CTR-GCN and MS-G3D models

1. Introduction

2. Comparative Analysis: Action Recognition

3. Comparative Analysis: Class-Incremental Learning

4. Assistive Robot Integration

Incremental Learning

Class-Incremental Learning Problem

An algorithm that learns a given sequence of tasks, T:

$$T = [(C^1, D^1), (C^2, D^2), \ldots, (C^n, D^n)] \quad (1)$$

Tasks

- Set of actions to be learnt:

$$D^t = \{(x_1, y_1), \ldots, (x_{m^t}, y_{m^t})\} \quad (2)$$

- Action set is distinct per task

$$C^i \cap C^j = \emptyset, \; if \; i \neq j \quad (3)$$

Exemplars

- Memory of training data from previous tasks
- Augments to training data if $t > 0$
- Memory scenarios: fixed or growing
- Selection methods: random, herding, distance, entropy

Incremental Learning Metrics

Task-Aware Accuracy

Calculated with the knowledge of the action classes learnt within each task.

Task-Agnostic Accuracy

Calculated with the overall set of the action classes learnt.

Forgetting Percentage

Estimated percentage of data forgotten

$$F_{i,t} = max(A[i, 0 : t - 1]) - A[i, t], \; i \leq t \quad (4)$$

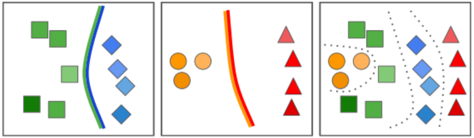


Figure 5: Incremental learning process of 4 classes split into 2 tasks[4]

[4] M. Masana et al., "Class-Incremental Learning: Survey and Performance Evaluation," CoRR, vol. abs/2010.15277, October 2020.

Incremental Learning Analysis

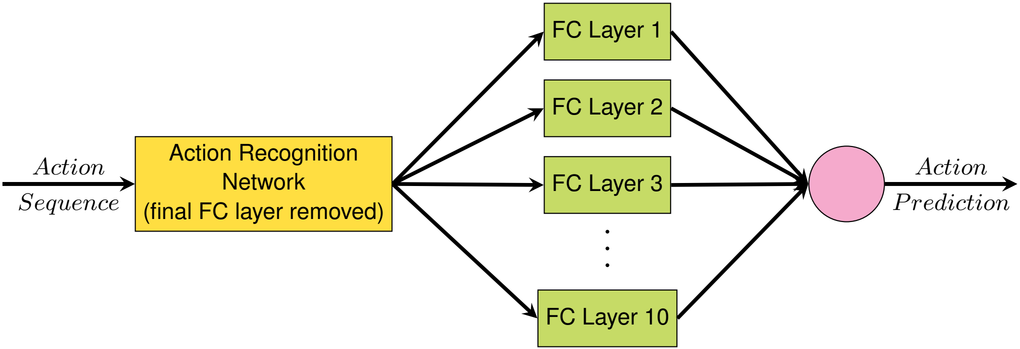
- IL Algorithms: LwF, iCaRL, LUCIR, BiC
- Memory Size: 400 (Fixed) & 20 per class (Growing)
- Selection Method: Random & Herding
- **Metrics**: Task-Aware & Task-Agnostic Accuracy

Task #	Action	Task #	Action
Task 1	Wipe Face Eat Meal	Task 6	Throw Point to Something
Task 2	Cross Hands Clapping	Task 7	Hand Waving Stand Up
Task 3	Kick Something Shake Head	Task 8	Nod Head/Bow Hopping
Task 4	Sit Down Play with Phone	Task 9	Drop Drink Water
Task 5	Pick Up Brush Teeth	Task 10	Rub Hands Jump Up

Table 5: Task sequence for class-IL comparative analysis

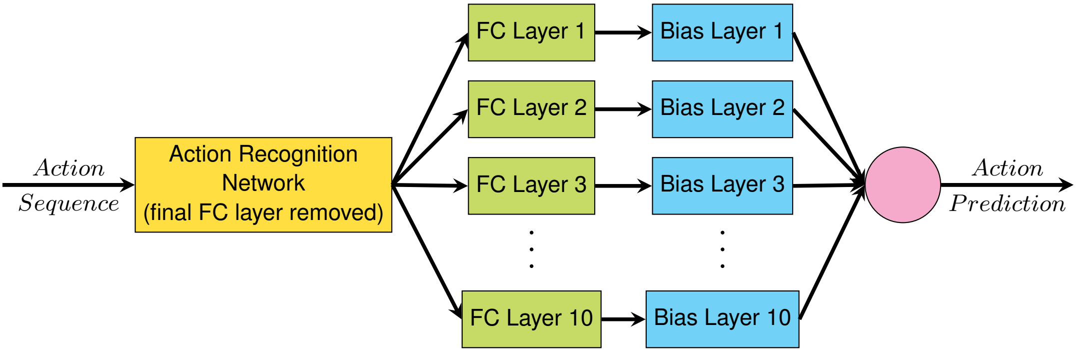
Incremental Learning Analysis

LwF, iCaRL, LUCIR Model Architecture



Incremental Learning Analysis

BiC Model Architecture



Incremental Learning Analysis Results

		CTR-GCN			MS-G3D		
Algorithm	Memory Config	Total Time (hrs)	Time per Task (min)	Time Incr per Task	Total Time (hrs)	Time per Task (min)	Time Incr per Task
iCaRL	Fixed	6.3	18.9 & 39.2	-	11.2	30.2 & 68.7	-
	Growing	5.6	-	2.8%	11.0	-	2.6%
LWF	Fixed	6.6	19.9 & 41.8	-	11.3	30.6 & 69.6	-
	Growing	5.8	-	2.8%	10.1	-	2.9%
BIC	Fixed	7.0	19.6 & 44.1	-	11.5	30.2 & 68.2	-
	Growing	6.2	-	3.3%	10.4	-	2.6%
LUCIR	Fixed	6.4	19.6 & 39.6	-	11.4	33.0 & 71.4	-
	Growing	5.8	-	2.9%	10.1	-	2.9%

Table 6: Incremental learning model training time

Incremental Learning Analysis Results

		Herding		Random	
Algorithm	Memory Config	CTR-GCN	MS-G3D	CTR-GCN	MS-G3D
iCaRL	Fixed	94.9%	97.2%	95.3%	98.1%
	Growing	91.1%	95.9%	91.2%	96.6%
LWF	Fixed	98.4%	97.8%	98.2%	97.5%
	Growing	97.6%	96.1%	97.1%	96.1%
LUCIR	Fixed	95.5%	97.5%	97.3%	97.5%
	Growing	96.3%	94.8%	95.5%	96.5%
BiC	Fixed	98.2%	98.1%	97.9%	96.9%
	Growing	97.3%	96.9%	97.5%	96.6%

Table 7: Average task-aware accuracy

Incremental Learning Analysis Results

		Herding		Random	
Algorithm	Memory Config	CTR-GCN	MS-G3D	CTR-GCN	MS-G3D
iCaRL	Fixed	52.1%	72.5%	50.9%	72.1%
	Growing	54.2%	67.8%	51.3%	66.9%
LWF	Fixed	73.8%	67.8%	75.4%	67.5%
	Growing	69.9%	61.5%	69.4%	63.5%
LUCIR	Fixed	66.2%	64.7%	68.4%	64.4%
	Growing	65.5%	43.6%	62.4%	59.9%
BiC	Fixed	79.1%	70.1%	74.8%	67.1%
	Growing	74.2%	61.1%	72.8%	61.6%

Table 8: Average task-agnostic accuracy

Incremental Learning Analysis

How would the model perform if we grouped similar actions?

- Task-Aware Accuracy:
  - Fixed: 98.2% $\Longrightarrow$ 94.6%
  - Growing: 97.3% $\Longrightarrow$ 93.4%
- Task-Agnostic Accuracy:
  - Fixed: 79.1% $\Longrightarrow$ 77.6%
  - Growing: 74.2% $\Longrightarrow$ 72.7%

<table>
<tr><th>Task #</th><th>Action</th><th>Task #</th><th>Action</th><th>Task #</th><th>Action</th></tr>
<tr><td rowspan="2">Task 1</td><td rowspan="2">Brush Teeth
Wipe Face</td><td rowspan="2">Task 4</td><td rowspan="2">Sit Down
Stand Up</td><td rowspan="3">Task 7</td><td rowspan="3">Kick Something
Hopping
Jump Up</td></tr>
<tr></tr>
<tr><td rowspan="2">Task 2</td><td rowspan="2">Drink Water
Eat Meal</td><td rowspan="3">Task 5</td><td rowspan="3">Clapping
Rub Hands
Cross Hands</td></tr>
<tr><td>Task 8</td><td>Play with Phone</td></tr>
<tr><td rowspan="3">Task 3</td><td rowspan="3">Drop
Pick Up
Throw</td><td rowspan="3">Task 9</td><td rowspan="3">Nod Head/Bow
Shake Head</td></tr>
<tr><td rowspan="2">Task 6</td><td rowspan="2">Hand Waving
Point to Something</td></tr>
<tr></tr>
</table>

Table 9: Task sequence with variable task size and sorting similar actions

Incremental Learning Analysis

Model Robustness

- Task-Aware Accuracy:
  - Fixed: 94.6% $\Longrightarrow$ 94.1% & 96.7%
  - Growing: 93.4% $\Longrightarrow$ 95.1% & 95.6%
- Task-Agnostic Accuracy:
  - Fixed: 77.6% $\Longrightarrow$ 76.3% & 78.5%
  - Growing: 72.7% $\Longrightarrow$ 74.5% & 70.6%

Test Sequence #1

Train Task 1 → Train Task 2 w/o Eat → Train Task 3 → Add Eat to Task 2 → Train Task 4-9

Test Sequence #2

Train Task 1-4 → Train Task 5 w/o Rub Hands → Train Task 5 w/o Point to Something → Add Rub Hands to Task 5 → Add Point to Something to Task 6 → Train Task 7-9

Incremental Learning Analysis

Memory Size

Growing						Fixed				
2	5	10	20	50	100	40	100	200	400	1000
78.3%	94.3%	97.2%	97.3%	98.5%	98.7%	94.5%	96.7%	97.6%	98.2%	98.5%

Table 10: Average task-aware accuracy for varying memory size

Growing						Fixed				
2	5	10	20	50	100	40	100	200	400	1000
11.3%	45.7%	63.8%	74.2%	79.8%	84.1%	35.8%	65.2%	70.3%	79.1%	80.0%

Table 11: Average task-agnostic accuracy for varying memory size

QTRobot Platform

- Teaching assistant for educators working with children
- Equipped with a RealSense 3D camera
- Skeletal tracking using Nuitrack SDK (19 joints tracked vs 25 joints in NTU)

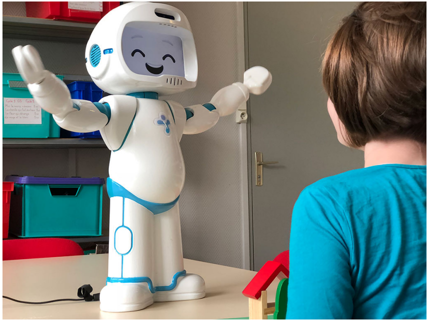


Figure 6: QTRobot interacting with a child[5]

[5] Image taken from: https://robots.ieee.org/robots/qtrobot/

Model Integration

- Task-Aware Accuracy: 97.3% $\Longrightarrow$ 97.4%
- Task-Agnostic Accuracy: 74.2% $\Longrightarrow$ 71.9%

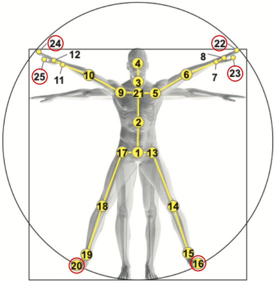


Figure 7: Joint configurations for NTU RGB-D dataset[6]

Figure 8: Joint configuration in the Nuitrack SDK[7]

[6] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1010–1019.

[7] Image taken from: https://github.com/3DiVi/nuitrack-sdk/tree/master/doc

CAL Server

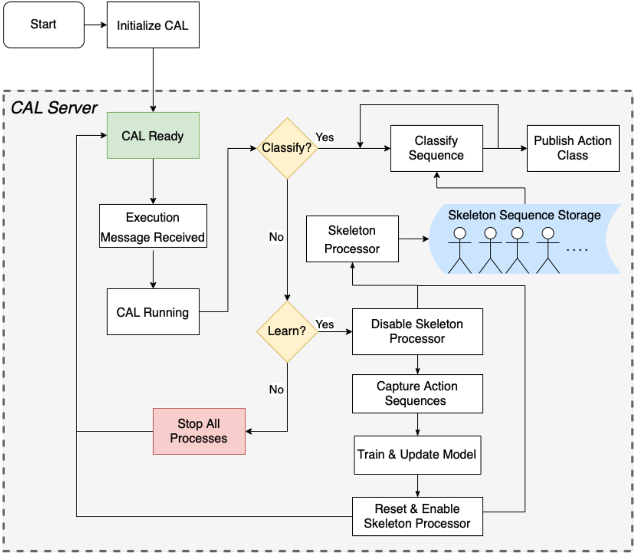


Figure 9: Workflow of the Continual Action Learning Server

Demo

Action Recognition

Demo

Action Learning

Thank You!

Questions?