Metrological evaluation and testing of robots in international competitions

HEART-MET
Dry-run Cascade Campaign Results

September 2021

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

Contents

1 Introduction **2**

2 Gesture Recognition Challenge **3**

3 Activity Recognition Challenge **6**

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

1 Introduction

HEART-MET is one of the competitions in the METRICS project, which has received funding from European Union's Horizon 2020 research and innovation program under grant agreement No 871252. The competition aims to benchmark assistive robots performing healthcare-related tasks in unstructured domestic environments. Two types of competitions are held, namely, the field evaluation campaign and the cascade evaluation campaign. This report describes the results of the dry-run cascade evaluation campaign, which consists of dataset-based challenges conducted entirely online. This edition of the cascade campaign for HEART-MET comprised of four challenges, namely, Gesture Recognition, Activity Recognition, Object Detection, and Handover Failure Detection, and ran from April to June 2021. The challenges were hosted on the Codalab[1] platform, and were publicly available.

Gesture recognition is one of the functional benchmarks by which assistive robots are evaluated. It is necessary for non-verbal communication with a user. The datasets for this challenge were collected from real robots performing gesture recognition in domestic environments, with several different volunteers performing the gestures. The gestures are meant to communicate intentions or commands to the robot, and include the stop sign, nodding, pointing etc.

Activity recognition is also an important skill for a robot which is operating in an assistive capacity for persons who may have care needs. In addition to recognizing daily living activities, it is important for the robot to detect activities or events in which the robot may need to offer help or call for assistance. The datasets for this challenge are also collected from real robots performing activity recognition in domestic environments.

AcRec@UniKoblenz from the University of Koblenz, comprising of 8 members, had the winning entries for both challenges with a true positive rate of 0.522 for the Gesture Recognition Challenge and 0.460 for the Activity Recognition challenge. RAAR3D from Macau University of Science and Technology and University of Chinese Academy of Sciences, comprising of 2 members, finished second in the Gesture Recognition challenge with a true positive rate of 0.436. In the following chapters, the detailed results from both challenges are presented, along with baseline results.

[1] `https://codalab.org/`

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

2 Gesture Recognition Challenge

The objective of the Gesture Recognition challenge[2] is to classify short video clips of persons performing gestures into one of nine classes. The dataset consists of video clips, up to 8 seconds long, with each clip containing one person performing a single gesture. The nine gestures listed in Table 1 were performed by 24 different people at several locations and in various poses.

The challenge was divided into two phases: validation and competition. In the validation phase, a subset of the dataset was provided for testing, with the expectation that teams would use external datasets for training, and test with the provided dataset. In the competition phase, the competition dataset was used for testing, and the previously provided validation dataset could be used for fine-tuning and validation.

Scoring Submissions were evaluated on the true positive rate of recognized gestures.

Table 1: Count of videos from each class in the gesture recognition dataset

Gesture	Validation videos	Test videos
Pulling hand in	16	38
Pushing hand out	12	40
Nodding	16	37
Stop sign	4	30
Thumb up	12	38
Waving	18	50
Thumb down	13	39
Pointing	7	27
Shaking head	2	15
Total	100	314

2.1 Results

A total of 19 submissions were made for the Gesture Recognition challenge. The results from the two winning teams and a baseline result are shown in Table 2. The baseline results are obtained using an I3D model[3] pretrained on the Charades[4] dataset, and fine-tuned using the validation set. Table 3 shows the class-wise true positive rates (TPR).

Table 2: Leaderboard results and baseline for Gesture Recognition

Team	TPR
AcRec@UniKoblenz	0.522
RAAR3D	0.436
Baseline	0.229

Figures 1, 2 and 3 show the confusion matrices for the three results.

[2] `https://competitions.codalab.org/competitions/30454`
[3] Paper: `https://arxiv.org/abs/1705.07750`, Code: `https://github.com/piergiaj/pytorch-i3d`
[4] `https://prior.allenai.org/projects/charades`

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

Table 3: Class-wise true positive rates for the Gesture Recognition challenge

Gesture	AcRec@UniKoblenz	RAAR3D	Baseline
Nodding	0.919	0.811	0.703
Stop sign	0.667	0.100	0.000
Thumb down	0.538	0.667	0.077
Waving	0.680	0.560	0.560
Pointing	0.259	0.111	0.037
Pulling hand in	0.184	0.184	0.158
Thumb up	0.395	0.421	0.158
Pushing hand out	0.375	0.425	0.050
Shaking head	0.733	0.467	0.000

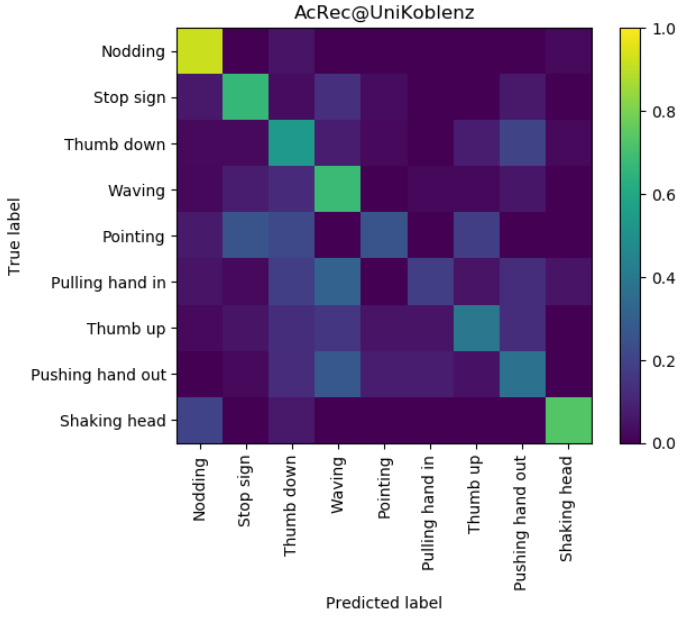

Figure 1: Confusion matrix for the best performing Gesture Recognition result (Team AcRec@UniKoblenz)

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

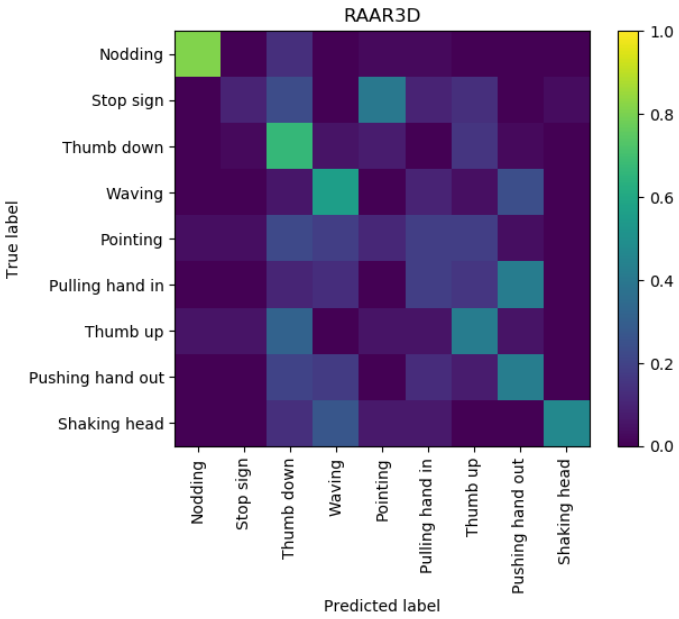

Figure 2: Confusion matrix for the second best performing Gesture Recognition result (Team RAAR3D)

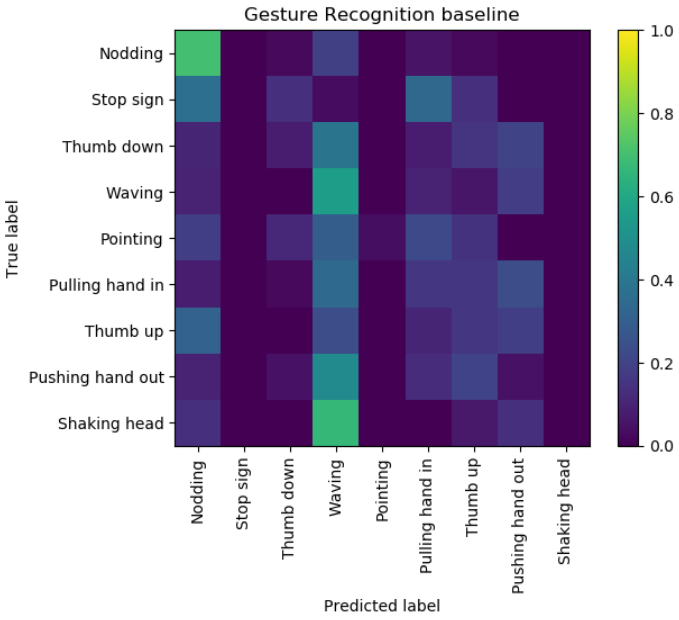

Figure 3: Confusion matrix for the baseline I3D model trained on the validation set

Both AcRec@UniKoblenz and RAAR3D show comparable performance for *Nodding*, *Thumb down*, *Waving* and *Shaking head*, all of which have a comparatively higher detection rate. AcRec@UniKoblenz is able to recognize *Stop sign* at a higher rate compared to RAAR3D, and has generally better true positive rates for the other classes. *Pulling hand in* was detected poorly by all three; the confusion matrices show that it was confused with *Waving* and *Pushing hand out* quite often. The baseline model performs poorly for most classes except *Nodding* and *Waving*; both teams perform significantly better than the baseline.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

3 Activity Recognition Challenge

The objective of the Activity Recognition challenge[5] is to classify short video clips of persons performing daily living activities into one of 22 activity classes. The dataset consists of video clips, up to 8 seconds long, with each clip containing one person performing a single activity. The 22 activities listed in Table 4 were performed by 29 different people at several locations and in various poses. Just as for Gesture Recognition, the Activity Recognition challenge was divided into two phases: validation and competition. In the validation phase, a subset of the dataset was provided for testing, with the expectation that teams would use external datasets for training, and test with the provided dataset. In the competition phase, the competition dataset was used for testing, and the previously provided validation dataset could be used for fine-tuning and validation.

Table 4: Count of videos from each class in the activity recognition dataset

Activity	**Validation videos**	**Test videos**
Writing	26	54
Putting on a jacket	8	10
Brushing teeth	30	48
Lying down	12	34
Using a computer	40	58
Doing freehand exercise	10	8
Reading a book	26	70
Doing neck roll exercise	8	10
Fallen on the floor	2	14
Colliding against furniture	8	10
Eating food with a fork	26	72
Putting/taking food in/from the fridge	2	16
Cutting vegetable on the cutting board	4	12
Drinking water	34	63
Limping	6	14
Opening the door and walking in	8	12
Rubbing face with hands	4	14
Talking on the phone	24	74
Someone is coughing/sneezing	2	16
Putting on/taking off shoes	10	8
Wiping off the dining table	8	10
Touching a hot surface	4	18
Total	302	645

Scoring Submissions were evaluated on the true positive rate of recognized activities. The top-3 and top-5 true positive rates were also calculated since participants had the option of submitting up to 5 activity classes for a given video clip. In case of ties, the top-3 and top-5 true positive rates were to be used for breaking the ties.

The final result of team AcRec@UniKoblenz for the activity recognition challenge is shown in Table 5. Also included are baseline results obtained using an I3D model which was pretrained on the Charades dataset, and fine-tuned using the validation set. Table 6 shows the class-wise true positive rates (TPR) for both results. Figure 4 and 5 show the confusion matrices for AcRec@UniKoblenz and the baseline.

Table 5: Leaderboard results and baseline for Activity Recognition

Team	**TPR**
Baseline	0.603
AcRec@UniKoblenz	0.460

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

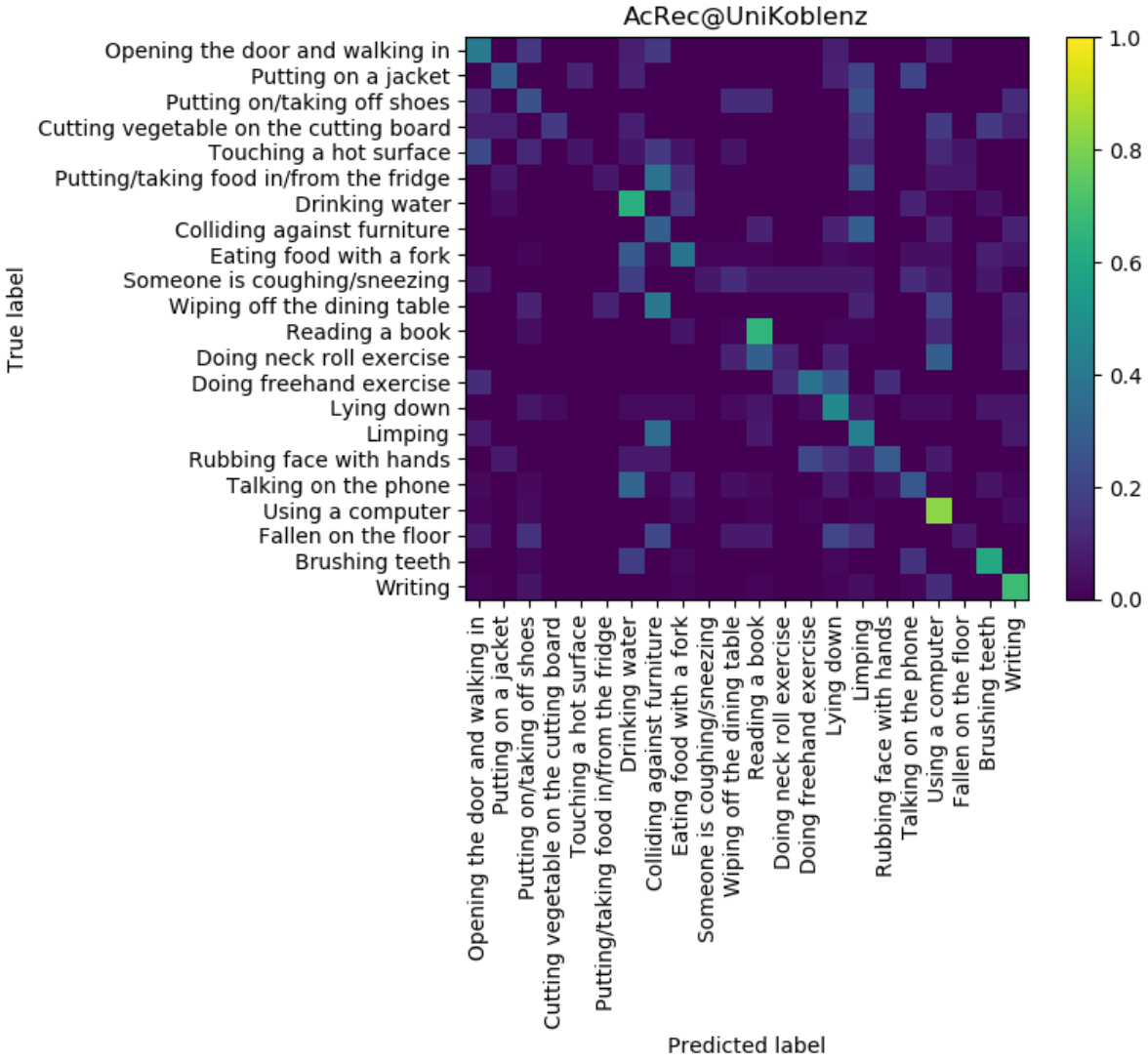

Figure 4: Confusion matrix for the best performing Activity Recognition result (Team AcRec@UniKoblenz)

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

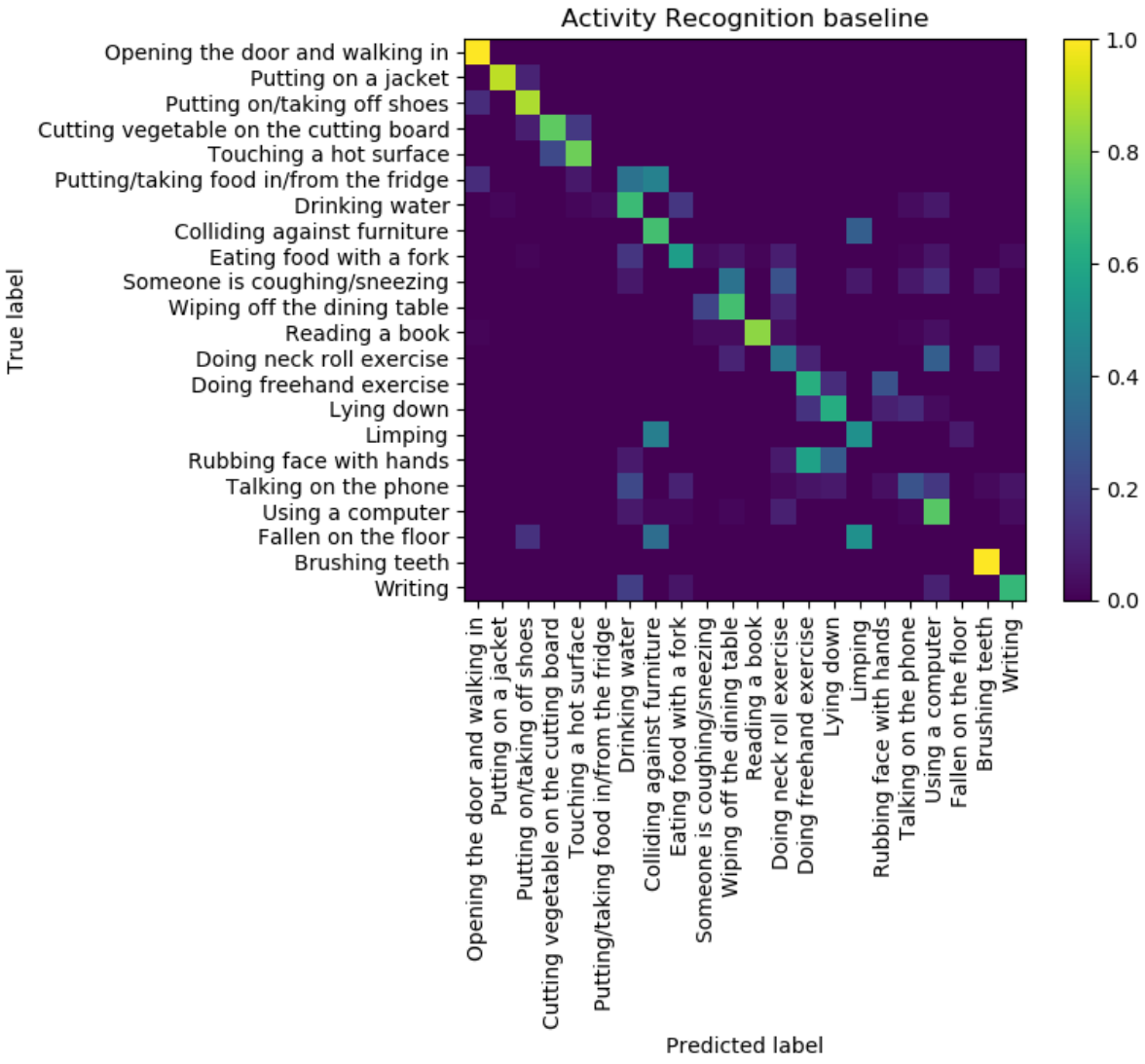

Figure 5: Confusion matrix for the baseline I3D model fine-tuned on the validation set

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

Table 6: Class-wise true positive rates for the Activity Recognition challenge

Activity	**AcRec@UniKoblenz**	**Baseline**
Opening the door and walking in	0.417	1.000
Putting on a jacket	0.300	0.900
Putting on/taking off shoes	0.250	0.875
Cutting vegetable on the cutting board	0.167	0.750
Touching a hot surface	0.056	0.778
Putting/taking food in/from the fridge	0.062	0.000
Drinking water	0.635	0.683
Colliding against furniture	0.300	0.700
Eating food with a fork	0.389	0.556
Someone is coughing/sneezing	0.062	0.000
Wiping off the dining table	0.000	0.700
Reading a book	0.657	0.829
Doing neck roll exercise	0.100	0.400
Doing freehand exercise	0.375	0.625
Lying down	0.471	0.618
Limping	0.429	0.500
Rubbing face with hands	0.286	0.000
Talking on the phone	0.270	0.257
Using a computer	0.828	0.741
Fallen on the floor	0.071	0.000
Brushing teeth	0.604	1.000
Writing	0.685	0.667

Some similarities can be observed between the two results. Both results show good performance on certain activities such as *Reading a book*, *Drinking water*, *Writing*, *Using a computer* and *Brushing Teeth*. Similarly, poor performance is observed in both results for activities such as *Taking food from the fridge*, *Someone is coughing/sneezing* and *Fallen on the floor*. The improved performance of the baseline can be explained from activities such as *Opening the door and walking in*, *Touching a hot surface*, *Wiping off the dining table* and *Putting on a jacket*. The activities that have a high true positive rate typically contained more validation videos, which were allowed to be used for fine-tuning. On the other hand, the poorly detected activities had very few samples in the validation set. If no external datasets are used, the low sample size converts the task into a few-shot learning problem, which is significantly more challenging.

[5] `https://competitions.codalab.org/competitions/30423`

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252