Metrological evaluation and testing of robots in international competitions

---

METRICS HEART-MET Field Evaluation Campaign Rulebook

Latest update: March 22, 2022

Santosh Thoduka
Deebul Nair
Nico Hochgeschwender
Praminda Caleb-Solly
Mauro Dragone
Filippo Cavallo

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

List of Abbreviations and Acronyms

ABBREVIATION	MEANING
FEC	Field Evaluation Campaign
CEC	Cascade Evaluation Campaign
FBM	Functionality Benchmark
TBM	Task Benchmark
IoU	Intersection over Union
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative

Contents

1 Introduction

The METRICS project organises robotics competitions in four priority areas, namely: Healthcare, Inspection and Maintenance, Agri-Food and Agile Production. The main goal of the competitions is to evaluate robots using objective metrological principles in each of the priority areas. This document describes the rules for the field campaign of the healthcare competition - HEART-MET (Healthcare Robotics Technologies - Metrified).

Past and current projects such as RoCKIn, RockEU2 and SciRoc have already defined methods for conducting competitions for the purpose of benchmarking robots. In RoCKIn [1], two classes of evaluation were introduced: Task Benchmarks (TBM) and Functionality Benchmarks (FBM), which are adopted for the HEART-MET competition. The methodology aims to ensure both reproducibility and repeatability of the benchmarks conducted in the context of a competition. It was also recognized that the popularity and already established infrastructure for robotics competitions provides an opportunity to introduce more rigorous benchmarking of both robot systems and their individual functionalities.

HEART-MET conducts both field campaigns at physical test-beds with real robots, and cascade campaigns, which are dataset-based competitions conducted entirely online. This rulebook defines the functionality and task benchmarks for the field campaigns only. Table 1 provides a short summary of the benchmarks for the field campaign, which are explained in more detail in Sections 3 and 4. In Section 2, we briefly describe the structure of the field evaluation campaign. Sections 5 and 6 describe the characteristics of the test bed and specifications and constraints of robots that can participate in the competition.

Benchmark	Type	Description
Object Detection	FBM	Detect a target object
Person Detection	FBM	Detect a person at a known location
Activity Recognition	FBM	Recognize an activity performed by a person
Gesture Recognition	FBM	Recognize a gesture performed by a person
Cluttered Pick	FBM	Pick an object from a cluttered surface and place it in a container
Handover Object	FBM	Hand over an object to a person
Receive Object	FBM	Receive an object from a person
Assess Activity State	TBM	Assess the activity state of a person visually and through dialogue with the person
Item Delivery	TBM	Pick a target object from a defined location and deliver it to a person at another defined location

Table 1: Summary of benchmarks

2 Competition Structure

The field campaign for HEART-MET takes place in a test-bed which resembles a living environment, such as the Living Space at the Cobot Maker Space[1]. The two complementary types of benchmarks, FBMs and TBMs, evaluate both the individual functionalities of the robot and the ability of the robot to integrate several functionalities into a fully functional system to complete a task. The FBMs evaluate individual functionalities of the robot such as object detection, speech understanding, grasping etc. The TBMs evaluate the capability of the robot to execute tasks such as delivery of medicine to a person, which requires the robot to integrate functionalities such as object detection, grasping, navigation and person detection.

2.1 Trials

Each robot will be evaluated on a particular benchmark by multiple trials. The trials may be split up into several runs over multiple days, thus allowing participants to modify their robot system (software and hardware) between runs. Each trial will introduce variations to independent variables such as the type of object used, pose of the person performing an activity, lighting conditions, etc. The final evaluation results are calculated by averaging the results across trials and runs.

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

2.2 Autonomy Levels

Robots can operate in a healthcare setting with various levels of autonomy, ranging from fully autonomous to teleoperated. For task benchmarks, teams will be allowed to participate with varying levels of autonomy. However, fully-autonomous and non-autonomous executions of tasks will be categorised separately.

For example, in the Assess Activity State TBM, one robot could perform the task fully autonomously by recognizing the human, visually determining their activity state and engaging in a natural language dialogue to confirm their activity state. Another robot could be remotely controlled, with team members assessing the human's activity via the robot's camera and conversing by using a speaker and microphone on the robot. Both methods of executing the task are feasible and likely in a healthcare setting, though they address slightly different use cases. In the second case, a human is directly involved in assessing a person's activity state and can make immediate decisions based on their observations.

Since different autonomy levels require different skills and target different use-cases, the evaluations of the task benchmarks will categorized based on the autonomy level chosen by the teams.

2.3 Data Collection

The testbeds used in the HEART-MET competitions are designed to facilitate the creation of datasets capturing complex, interleaved and hierarchical naturalistic activities, collected in environments instrumented with a rich variety of sensors. These infrastructures will be used to collect benchmarking data (from external RGB cameras in addition to logs from robots' sensors, including proprioceptive sensor data from the robots' manipulators and base). The data will also include recordings of images and videos suitable for supporting participatory design and the evaluation of human-robot interaction, in addition to dissemination and outreach activities. A further opportunity will include exploiting the existing infrastructures and the replicability offered by our standardised set of test benchmarks, to provide more data on the humans interacting with the robots.

3 Functionality Benchmarks

This chapter describes the functionality benchmarks which are meant to evaluate specific, standalone capabilities of a robot. We provide a general description of the functionality, the variations in terms of the independent variables for that functionality, communication with the referee box, procedure for conducting the benchmark and finally the evaluation criteria for the benchmark.

The variations mentioned in the benchmark are specific controllable aspects of the task that will be selected by the referee before the benchmark begins and remains fixed for all teams. Besides these variations, it is expected that factors such as lighting conditions, locations and volunteers used during the benchmark are also varied without prior specification.

While the evaluation criteria mentioned in each benchmark is the primary method of comparing performance, the penalties and eventually the duration for executing the benchmark is taken into account in case there is a tie between teams; i.e. if the scores are tied, the team with lower penalties is ranked higher and if the penalties are also tied, the team with lower execution time is ranked higher.

The following sections describe the functionality benchmarks with their associated metrics and procedure for execution.

3.1 Object Detection Functionality

3.1.1 Functionality Description

This functionality benchmark assesses the robot's capability of locating a target object in a given location. A common task for a healthcare robot is to locate a particular object, possibly in a particular location. In typical object detection benchmarks, the task is to detect all objects in a given image. Here, we instead require the robot to find a particular object among a set of objects. Therefore, one of the possible variations in this benchmark is one in which the target object is not present at the target location.

Several secondary objects are placed on a flat surface with no minimum distance between objects. The target object is either included in this set of objects or not, depending on the variation selected for a particular

trial. The robot is placed in front of the location, and must locate the target object if it exists, or indicate that the target object is not present.

3.1.2 Healthcare Relevance

Several tasks for a healthcare robot require the robot to locate and fetch an item. Such items could include medicines, reading glasses, a walking cane etc. The task would typically involve moving to a known location where the item usually is, detecting it, grasping it and delivering it to a person. This functionality hence deals with only detecting a target item, once the robot is already at the location where it expects the item to be.

3.1.3 Feature Variation

The independent variables for this functionality are:

- the set of objects and their poses
- is the target object present [yes, no]

The dependent variable is the detected pose of the object, or the detection of no object. Some sample objects which may be used for this benchmark can be found in Appendix A. The exact list of objects will be released atleast one month before the competition.

3.1.4 Communication with the Referee Box

For each trial in the run:

- The robot waits for a start message from the referee box
- The robot sends a confirmation that it has received the start message
- The robot sends a message to the referee box with the location of the object (or indicate the object is not found)
- In case of a timeout, the referee box sends a stop message to the robot

3.1.5 Procedures and Rules

A set of target objects, secondary objects and their poses is specified and fixed for all teams.

The maximum time allowed for one trial is 10 seconds. The time is calculated from the moment the robot confirms the start message has been received, until the robot sends the result message. If 10 seconds is exceeded, a timeout is recorded for that trial, and the robot must prepare for the next trial.

The robot has the option of specifying the location of the object in one of two ways: a) the 3D bounding box of the object with respect to the robot's base b) the 2D bounding box of the object in an image. In the first case, the robot must provide, in the benchmarking data, a point cloud of the scene (transformed to the robot's base frame) corresponding to the provided 3D bounding box of the object. The 3D bounding box must also be in the robot's base frame. In the second case, the robot must provide the raw RGB image of the scene corresponding to the provided 2D bounding box. Only a single bounding box per point cloud / image must be specified. If multiple bounding boxes are specified, only the first one is considered.

3.1.6 Benchmarking Data

The internally recorded data must include (at minimum):

- RGB camera stream of the robot
- Point cloud of scene (if 3D bounding box is provided)
- RGB image of the scene (if 2D bounding box is provided)

In addition, the teams should provide the data used for training machine learning models (if any).

3.1.7 Metrics for Evaluation

We calculate the following four metrics for evaluating the performance of the robot:

- True Positive (TP): detects target object when it is present
- False Positive (FP): detects wrong object whether target object is present or not
- False Negative (FN): does not detect any object when target object is present
- True Negative (TN): does not detect any object when target object is not present

These metrics are illustrated in Figure 1.

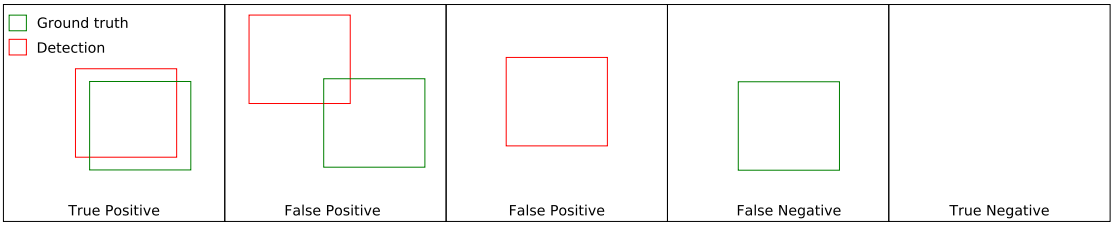


Figure 1: Metrics for object detection

For TP and FP, a measurement of the degree of similarity between the ground truth and detected bounding boxes is required. Here we use the Jaccard Index, also known as the Intersection over Union (IoU),

$$IoU = \frac{B_{GT} \cap B_{DET}}{B_{GT} \cup B_{DET}} \tag{1}$$

where B_{GT} and B_{DET} are the ground truth and detected bounding boxes respectively, and the intersections are calculated as areas and volumes for 2D and 3D bounding boxes respectively. In the example in Figure 2, the green box shows the ground truth bounding box, and the red box shows the predicted bounding box. If the IoU is greater than a threshold, the detection is considered to be a true positive, and a false positive if the IoU is lower than the threshold. The same applies to 3D bounding boxes, except that the areas will be replaced by volumes.

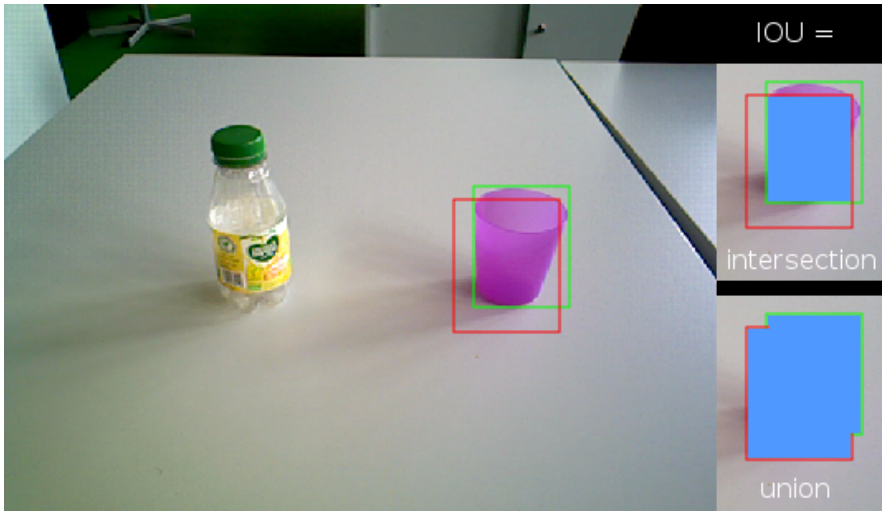


Figure 2: Example object detection; the green box shows the ground truth, and the red box shows the detection

Since the IoU threshold is a configurable parameter, we consider a range of thresholds from 0.5 to 0.95 with a step size of 0.05 (this is the approach taken by the COCO challenge[2]). The metrics TP, FP are then averaged over all IoU thresholds.

Teams will be ranked based on:

- the sum of the TP and TN;
- in case of a tie, the team with a lower FP count is ranked higher;
- in case teams are still tied, the team with the lower FN count is ranked higher.

It must be noted that the reason for considering FN a lower priority compared to FP is that it is preferable for a robot not to detect an object than to incorrectly detect an object. In the former case, the robot can simply retry detecting the object, whereas in the latter case, it might result in a task failure (for example, the robot might deliver an incorrect medicine to the person).

Penalties

- the robot collides with environment in an uncontrolled manner
- the robot stops responding

Disqualifying Behaviours

- the robot damages the environment

3.2 Person Detection Functionality

3.2.1 Functionality Description

This functionality benchmark assesses the robot's capability of detecting people in a living environment. For each trial, the robot is expected to detect a single person within a predefined distance range from the robot.

3.2.2 Healthcare Relevance

Robots operating in a home or healthcare environment must be able to detect people in their environment that they may need to assist or interact with.

3.2.3 Feature Variation

The robot is expected to detect a single person in its vicinity within a maximum distance. The person may not be located in the line of sight of the robot, therefore the robot may need to rotate on the spot, or move for a short distance to detect the person. The independent variables that will be varied for each execution are:

- Distance of the person to the robot
- Pose of the person [standing, sitting, laying]
- Pose of the person's face with respect to the robot [for example, straight, 30° left, 30° right]
- Presence of eye wear
- Presence of face mask
- Presence of head covering (such as a cap)

The dependent variable is the location of the person.

[2] http://cocodataset.org/#detection-eval

3.2.4 Communication with the Referee Box

For each trial in the run:

- The robot waits for a start message from the referee box
- The robot sends a confirmation that it has received the start message
- The robot sends a message to the referee box with the location of the person as a 2D bounding box and an accompanying image which the 2D bounding box refers to
- In case of a timeout, the referee box sends a stop message to the robot

3.2.5 Procedures and Rules

The configurations of all trials in a run are selected and fixed for all teams. If multiple runs are executed during a competition, new configurations are selected.

The maximum time allowed for one trial is 10 seconds. The time is calculated from the moment the robot confirms the start message has been received, until the robot sends a message with the location of the person. If 10 seconds is exceeded, a timeout is recorded for that trial, and the robot must prepare for the next trial.

3.2.6 Benchmarking Data

The internally recorded data must include (at minimum):

- RGB camera stream of the robot

In addition, the teams should provide the data used for training machine learning models (if any).

3.2.7 Metrics for Evaluation

The performance of the robot is measured by the F1-score. The following three metrics are computed as follows:

- True Positive (TP): detects the person
- False Positive (FP): detects incorrect object as a person
- False Negative (FN): does not detect any person

Similar to the Object Detection benchmark, for TP and FP, we use the IoU as a measurement of the degree of similarity between the ground truth and detected bounding boxes. We again consider a range of thresholds for IoU, from 0.5 to 0.95 with a step size of 0.05 and average TP and FP over all IoU thresholds. Based on these definitions, we calculate the F1-score as:

$$F1 = 2 * \frac{precision * recall}{precision + recall} \tag{2}$$

where

$$precision = \frac{TP}{TP + FP}, recall = \frac{TP}{TP + FN} \tag{3}$$

If multiple runs are performed, the F1-score is calculated using the sum of the TP, FP and FN over all runs.

Penalties

- the robot collides with environment in an uncontrolled manner
- the robot stops responding

Disqualifying Behaviours

- the robot damages the environment

3.3 Activity Recognition Functionality

3.3.1 Functionality Description

This functionality benchmark assesses the robot's capability of recognizing the activities of a human. The robot is placed in front of a human who performs an activity. The robot needs to recognize the activity being performed by the human.

3.3.2 Healthcare Relevance

When a robot is operating in an assistive capacity, it is often required to monitor the state of the person who has caring needs. Such monitoring could include detecting when the person is sleeping, detecting a fall, or simply other daily living activities. The robot could then make decisions based on the detected activities; for example by calling for help if a person has fallen down.

3.3.3 Feature Variation

The human activity (dependent variable) will be chosen from a list consisting of:

- All classes in the Charades dataset[3] [2], which consist of daily living activities
- A set of activities which are relevant in the healthcare context, listed below:
  - Coughing
  - Walking on crutches
  - Sitting in a wheelchair
  - Moving from a couch to a wheelchair
  - Falling down
  - Limping
  - Hopping
  - Reacting to getting hurt
  - Colliding against furniture

For a benchmark run, a fixed number of activities will be selected and performed by a set of actors, which the robot has to recognize. To maintain uniformity, the same actors will perform a given activity for all teams. The independent variables that could be varied for each execution are:

- Distance of the person to the robot
- Pose of the person [standing, sitting, laying]

3.3.4 Communication with the Referee Box

For each activity in the run:

- The robot waits for a start message from the referee box
- The robot sends a confirmation that it has received the start message
- The robot sends a message to the referee box with the classified activity label
- In case of a timeout, the referee box sends a stop message to the robot

[3] `https://prior.allenai.org/projects/charades`

3.3.5 Procedures and Rules

The referee selects the list of activities, their locations, distances to the robot, and poses of the person. The human actor performs the activity when the robot confirms it has received the start message.

The maximum time allowed for classifying one activity is 20 seconds. The time is calculated from the moment the robot confirms the start message has been received, until the robot replies with the activity class. If 20 seconds is exceeded, a timeout is recorded for that activity, and the robot must prepare for the next activity.

3.3.6 Benchmarking Data

The internally recorded data must include (at minimum):

- RGB camera stream of the robot
- Classified activity label

In addition to the recorded internal robot data, an external RGB camera will record each run.

3.3.7 Metrics for Evaluation

The performance of the robot is based on the following metrics:

1. True positive (TP): correctly identified activities

In addition, the overall true positive rate are calculated as:

1. True positive rate $TPR = \frac{TP}{N}$

where N is the number of trials.

Penalties

- the robot collides with environment in an uncontrolled manner
- the robot stops responding

Disqualifying Behaviours

- the robot damages the environment

3.4 Gesture Recognition Functionality

3.4.1 Functionality Description

This functionality benchmark assesses the robot's capability of recognizing gestures performed by a human. The robot is placed in front of a human who performs a gesture. The robot needs to recognize the gesture being performed by the human.

3.4.2 Healthcare Relevance

Gestures are a form of interaction that a human might use to communicate with the robot. Especially when the human has an impairment, it might be necessary to communicate via gestures as opposed to speech. In such a case, a robot needs to be able to recognize common gestures such as waving, pointing etc.

3.4.3 Feature Variation

The gestures (dependent variable) will be chosen from a list consisting of:

- Nodding
- Stop sign
- Thumb down
- Waving
- Pointing
- Pulling hand in
- Thumb up
- Pushing hand out
- Shaking head

For a benchmark run, a set of gestures will be selected and performed by actors, which the robot has to recognize. To maintain uniformity, the same actors will perform a given gesture for all teams. The independent variables that could be varied for each execution are:

- Distance of the person to the robot
- Pose of the person [standing, sitting, laying]

3.4.4 Communication with the Referee Box

For each gesture in the run:

- The robot waits for a start message from the referee box
- The robot sends a confirmation that it has received the start message
- The robot sends a message to the referee box with the classified gesture label
- In case of a timeout, the referee box sends a stop message to the robot

3.4.5 Procedures and Rules

The referee selects the list of activities, their locations, distances to the robot, and poses of the person.

The maximum time allowed for classifying one gesture is 20 seconds. The time is calculated from the moment the robot confirms the start message has been received, until the robot replies with the gesture class. If 20 seconds is exceeded, a timeout is recorded for that gesture, and the robot must prepare for the next gesture.

The human actor performs the gesture when the robot confirms it has received the start message.

3.4.6 Benchmarking Data

The internally recorded data must include (at minimum):

- RGB camera stream of the robot
- Classified gesture label

In addition to the recorded internal robot data, an external RGB camera will record each run.

3.4.7 Metrics for Evaluation

The performance of the robot is based on the following metrics:

1. True positive (TP): correctly identified gestures

In addition, the overall true positive rate are calculated as:

1. True positive rate $TPR = \frac{TP}{N}$

where N is the number of trials.

Penalties

- the robot collides with environment in an uncontrolled manner
- the robot stops responding

Disqualifying Behaviours

- the robot damages the environment

3.5 Cluttered Pick Functionality (ERL Professional)

This FBM is defined for participation in the ERL Professional[4] local tournament. Teams may participate in this and other benchmarks regardless of whether they are participating in the ERL tournament or not.

3.5.1 Functionality Description

This functionality benchmark assesses the robot's capability of picking objects from a set of cluttered objects on a surface, with the intention of placing it in a container, such as a basket. Several objects and a container are placed on a flat surface such as a table. For a single run, the robot must pick all objects on the surface and place it in the container in pre-defined orientations.

3.5.2 Healthcare Relevance

An assistive robot may be tasked with tidying up a table, or asked to collect several items for a person. Since objects in a home or healthcare environment are often cluttered with other objects, the focus of this benchmark is on assessing the robot's ability to pick objects from a cluttered space. The placement of the objects in the container is included in the benchmark to emphasize the *task-oriented* nature of the grasp - namely, the robot must grasp the objects in such a way that they can be placed in the container in a desired orientation.

3.5.3 Feature Variation

The configuration for a single run consists of the following independent variables:

- The number of objects placed on the table
- The location and orientation of the objects
- The level of clutter, estimated by the distance between objects on the table

[4] https://www.eu-robotics.net/robotics_league/erl-professional/about/index.html

3.5.4 Input Provided

The team will be provided with the following information:

- The list of possible objects used in the task;
- The difficulty level of the objects;
- The desired orientation of the objects when placed in the container;
- The location of the table and container

Some sample objects, including the container, which may be used for this benchmark can be found in Appendix A.

3.5.5 Communication with the Referee Box

For each run:

- The robot waits for a start message from the referee box
- The robot sends a confirmation that it has received the start message
- The robot sends a message to the referee box when an object has been picked
- The robot sends a message to the referee box when an object has been placed
- In case of a timeout, the referee box sends a stop message to the robot

ERL teams have the option of using the METRICS HEART-MET referee box, or the previously used logging tool[5].

3.5.6 Procedures and Rules

Step 1 The robot starts 1 meter away from the table.

Step 2 The robot moves towards the table and has to pick one object.

Step 3 The robot moves towards the container, and places the object in the container

Step 4 The robot repeats step 3 and 4 till no objects remain on the table.

Step 5 The competition time starts when the robot starts moving.

Step 6 The competition ends when the last object is placed.

3.5.7 Metrics for Evaluation

Achievements The set A of achievements for this task consists of:

- The robot moves to the table.
- The robot grasps an object (points as per difficult level mentioned in Table 2). (Repeatable)

Difficulty	Achievements
Easy Object	1
Medium Object	2
Hard Object	3

Table 2: Achievements as per difficulty level of objects.

[5] https://github.com/EuropeanRoboticsLeague/SciRocEpisode05PickPack

- The robot logs the name of the picked object. (Repeatable)
- The robot correctly places the object in the container (Repeatable)
- The orientation of the placed object matches with the desired orientation. (Repeatable)
- The robot logs the name of the placed object after placing.(Repeatable)
- The robot delivers all the K objects.
- The robot logs the end of the run.

Runs

- Each team has to run a minimum of 5 runs.
- The teams can run more than 5 runs if they want to provided there is sufficient time.
- From the 5 (or more) runs the best 3 are selected and the median is used as the score.

Penalties

- the robot collides with environment or human in an uncontrolled manner
- the robot drops the object causing it to drop to the floor through no fault of the human (note: this is a subjective evaluation, and it will be up to the referees to decide if the robot was at fault)
- the robot does not release the object
- the robot stops responding

Disqualifying Behaviours

- the robot damages the environment
- the robot damages the object

3.6 Handover Functionality

3.6.1 Functionality Description

This functionality benchmark assesses the robot's capability of handing over objects to a person. An object is placed in the robot's gripper and the robot is placed in front of a person. The robot needs to hand over the object to the person and verify that the object has been received.

3.6.2 Healthcare Relevance

An assistive robot is often tasked with bringing an item (such as medicine, reading glasses etc.) to a person. When delivering items to a human, the handover is a complicated interaction that requires awareness from both the human and the robot about each other's intentions in order to be successful. For this benchmark, an emphasis is placed on the ability of the robot to monitor the interaction and verify its success.

3.6.3 Feature Variation

The configuration for a single trial consists of the assignment of the following independent variables:

- Object to be handed over
- Human actor
- Human pose [standing, sitting, laying down]
- Human behaviour before grasp [reaches out, does not reach out]
- Human behaviour during grasp [grasps object, does not grasp object]
- Human behaviour after grasp (up to 5 seconds) [keeps object in hand, lets object fall]

3.6.4 Communication with the Referee Box

For each trial:

- The robot waits for a start message from the referee box
- The robot sends a confirmation that it has received the start message
- The robot sends a message to the referee box with the outcome of the execution (specified below)
- In case of a timeout, the referee box sends a stop message to the robot

The message defining the outcome of the execution should include the following information:

- human pose [standing, sitting, laying]
- human reached out for object [yes, no]
- object was grasped successfully [yes, no, undefined]
- object fell down after grasp [yes, no, undefined]

3.6.5 Procedures and Rules

The maximum time allowed for one trial is 30 seconds. The time is calculated from the moment the robot confirms the start message has been received, until the robot replies with the outcome of the action. If 30 seconds is exceeded, a timeout is recorded for that trial, and the robot must prepare for the next trial.

The human will place themselves at most 1 m in front of the robot.

3.6.6 Benchmarking Data

The internally recorded data must include (at minimum):

- RGB camera stream of the robot
- Proprioceptive sensor data from the robot's manipulator and base
- Position of end-effector with respect to the robot base
- Events defining the start and end of each phase of the handover (approach, handover, retract)

In addition to the recorded internal robot data, an external RGB camera will record each run.

3.6.7 Metrics for Evaluation

The performance of the robot is based on the following achievements for each execution:

1. Detection of human pose [**Achievements: 1**]
2. Detection of human (not) reaching out for object [**Achievements: 1**]
3. Detection of (un) successful grasp [**Achievements: 1**]
4. Detection of object (not) falling after grasp [**Achievements: 1**]
5. Successful handover [**Achievements: 1**]
6. In addition to the achievements described above, the referee, the volunteer who interacts with the robot or third-parties shall score the human-robot interaction aspect of the task by evaluating usability, social acceptance and user experience, by using the USUS evaluation framework [3]. Teams should consider the motions of the manipulator, intuitiveness of the handover, timing of the release, etc.

Penalties

- the robot collides with environment or human in an uncontrolled manner
- the robot drops the object before the person makes contact with the object. In case of ambiguous outcomes, the referee makes the final decision to decide if the robot was at fault.
- the robot does not release the object
- the robot stops responding

Disqualifying Behaviours

- the robot damages the environment
- the robot damages the object

3.7 Receive Object Functionality

3.7.1 Functionality Description

This functionality benchmark assesses the robot's capability of receiving objects from a person. The robot is placed in front of a person who is holding an object. The robot must receive the object from the person when it is offered. The robot is allowed to initiate a dialogue to facilitate the exchange. In case no dialogue is initiated, the person will hand over the object without prompting.

3.7.2 Healthcare Relevance

Similar to the handover task, a robot in a healthcare setting is often required to receive objects from a human. The interaction must be as smooth and intuitive as possible, as would be expected from a robot interacting with persons with impairments.

3.7.3 Feature Variation

The configuration for a single execution consists of the assignment of the following independent variables:

- Object to be handed over
- Human actor
- Human pose [standing, sitting, laying down]
- Human behaviour before grasp [reaches out with object, does not reach out, drops the object]
- Human behaviour during and after grasp [releases the object, does not release the object]

3.7.4 Communication with the Referee Box

For each trial in the run:

- The robot waits for a start message from the referee box
- The robot sends a confirmation that it has received the start message
- The robot sends a message to the referee box with the outcome of the trial (specified below)
- In case of a timeout, the referee box sends a stop message to the robot

The message defining the outcome of the trial should include the following information:

- human pose [standing, sitting, laying]
- human reached out with object [yes, no]
- object fell down [yes, no]
- object was grasped successfully [yes, no, undefined]
- object was released [yes, no, undefined]

3.7.5 Procedures and Rules

The maximum time allowed for one trial is 30 seconds. The time is calculated from the moment the robot confirms the start message has been received, until the robot replies with the outcome of the action. If 30 seconds is exceeded, a timeout is recorded for that trial, and the robot must prepare for the next trial.

The human will place themselves at most 1 m in front of the robot.

3.7.6 Benchmarking Data

The internally recorded data must include (at minimum):

- RGB camera stream of the robot
- Proprioceptive sensor data from the robot's manipulator and base
- Position of end-effector with respect to the robot base
- Events defining the start and end of each phase of the handover (approach, handover, retract)

In addition to the recorded internal robot data, an external RGB camera will record each run.

3.7.7 Metrics for Evaluation

The performance of the robot is based on the following achievements for each trial:

1. Detection of human pose [**Achievements: 1**]
2. Detection of human (not) reaching out with object [**Achievements: 1**]
3. Detection of fallen object [**Achievements: 1**]
4. Grasping of object [**Achievements: 1**]
5. Detection of human (not) releasing object [**Achievements: 1**]
6. Successful receipt of object[**Achievements: 1**]
7. In addition to the achievements described above, the referee, the volunteer who interacts with the robot or third-parties shall score the human-robot interaction aspect of the task by evaluating usability, social acceptance and user experience, by using the USUS evaluation framework [3]. Teams should consider the motions of the manipulator, intuitiveness of the handover, timing of the grasp, etc.

Penalties

- the robot collides with environment or human in an uncontrolled manner
- the robot drops the object after the person releases the object. In case of ambiguous outcomes, the referee makes the final decision to decide if the robot was at fault.
- the robot stops responding

Disqualifying Behaviours

- the robot damages the environment
- the robot damages the object

4 Task Benchmarks

The task benchmarks aim to evaluate the performance of a robot in the execution of a full task. The full task includes several subsystems of the robot, which have been individually evaluated in the functional benchmarks. Hence, the focus is on the integration of the functionalities and the capability of the robot to account for failures in individual functionalities to successfully complete the task.

The following sections describe some task benchmarks and the evaluation procedure. The evaluation is typically in the form of achievements for having reached certain checkpoints in the task.

4.1 Assess Activity State Task

4.1.1 Task Description

This task benchmark assesses the robot's capability of integrating several FBMs to assess a person's activity state through both visual cues and a natural language dialogue. The robot must locate a particular person in a given location and initially visually assess their activity state. The robot must then approach the person and initiate a natural language dialogue to verify their activity state. The functionalities required to complete this task include the FBMs Person Detection 3.2 and Activity Recognition 3.3, in addition to speech generation and understanding. All variations of the individual FBMs will considered for the task benchmark as well.

4.1.2 Healthcare Relevance

In addition to visually monitoring the activity state of a person, this task requires the robot to confirm the assessed state by engaging in a dialogue. In addition to increasing the level of engagement between the robot and the human, this task is a more comprehensive way for the robot to evaluate the activity state of a human, instead of simply observing visually.

4.1.3 Communication with the Referee Box

- The robot waits for a start message from the referee box. This message contains the semantic location of the person (for example, living room, kitchen, etc.)
- The robot sends a confirmation that it has received the start message
- The robot sends a feedback message to the referee box to indicate its progress, when it has:
  - located the person
  - visually assessed their activity state
  - completed the assessment via natural language dialogue with the person
- The robot sends a message indicating the completion of the benchmark

4.1.4 Procedures and Rules

For each trial, the referee selects a person, the location, and activity of the person. The configuration for a particular trial is fixed for all teams.

The maximum time allowed for one task execution is 5 minutes. The time is calculated from the moment the robot confirms the start message has been received, until the robot indicates the end of the benchmark. If 5 minutes is exceeded, a timeout is recorded for that execution, and the robot must prepare for the next execution.

4.1.5 Autonomy Level

Teams are allowed to choose any level of autonomy for this task, from fully autonomous to fully remote-controlled. The teams must specify the autonomy level beforehand for each functionality. For this task benchmark, autonomy levels could include:

- Navigation: [fully remote controlled, remote waypoint specification, fully autonomous]
- Visual Activity Recognition: [remotely assessed by team member, autonomous]
- Natural Language Dialogue: [remote conversation via microphone and speaker, autonomous]

4.1.6 Benchmarking Data

The internally recorded data must include (at minimum):

- RGB camera stream of the robot
- Odometry and global pose of the robot
- Output of person detection
- Output of activity recognition
- Audio used for recognizing speech
- Recognized speech
- Transcript of spoken text

In addition to the recorded internal robot data, an external RGB camera will record each run.

4.1.7 Scoring and Ranking

The performance of the robot is based on the following achievements for each execution:

1. *Navigation to the person*: The robot is considered to have successfully navigated to the person if it is less than 2.5 m away from the person, and is visible in the robot's camera. The robot must return an image from its camera to complete this achievement. [**Achievements: 1**]
2. *Detection of the person*: The robot must return a 2D bounding box of the detected person and a corresponding RGB image to complete this achievement. [**Achievements: 1**]
3. *Visual recognition of the activity*: The robot must visually recognize the activity performed by the person. The robot must return the activity recognition result to complete this achievement. [**Achievements: 1**]
4. *Initiation of dialogue with the person*: The robot must initiate a dialogue with the person by saying or asking something and recording a spoken response from the person. The robot must return the text of the spoken response to complete this achievement. In case the person does not respond, the referee will decide whether a sufficient effort was made by the robot to initiate a dialogue. [**Achievements: 1**]

5. *Verification of activity state through dialogue*: Through dialogue, the robot must determine the activity performed by the person. The robot should return the activity being performed to complete this achievement. In case the person does not respond, no achievements are awarded. [**Achievements: 1**]
6. In addition to the achievements awarded for the intermediate checkpoints, the referee, the volunteer who interacts with the robot or third-parties shall score the human-robot interaction aspect of the task by evaluating usability, social acceptance and user experience, by using the USUS evaluation framework [3]. Teams should consider the manner of approaching the person, manner of speech and usability of interfaces used to communicate with the person, etc.

In case multiple runs are executed during a competition, the sum of achievements for all runs is calculated.

Penalties

- the robot collides with environment or human in an uncontrolled manner
- the robot stops responding

Disqualifying Behaviours

- the robot damages the environment
- the robot damages the object

4.2 Item Delivery Task

4.2.1 Task Description

This task benchmark assesses the robot's capability of integrating several FBMs to safely deliver a healthcare related item to a human. The robot must locate and grasp a specified item, transport the item to a human, hand over the item and verify that the item has been received. The functionalities required to complete this task include the FBMs Object Detection 3.1 and Handover 3.6 in addition to navigation and grasping. All variations of the individual functionalities will considered for the task benchmark as well.

4.2.2 Healthcare Relevance

An assistive robot can aid a person by fetching items from around the living area, which would be particularly helpful for persons with physical impairments.

4.2.3 Communication with the Referee Box

- The robot waits for a start message from the referee box. This message contains the required item and its location, and the location of the human
- The robot sends a confirmation that it has received the start message
- The robot sends a feedback message to the referee to indicate its progress, when it has:
  - located the item
  - grasped the item
  - reached the human
  - handed over the item

  The feedback messages will be identical to the ones specified in the individual functionality benchmarks, if applicable. For example, the feedback for handing over the item should include the human pose, whether the human reached out for the item, whether the item was successfully grasped, and whether the item fell down after the grasp.
- The robot sends a message indicating the completion of the benchmark

4.2.4 Procedures and Rules

For each trial, the referee selects an item, the locations, pose of the human and intended human behaviour. The configuration for a particular trial is fixed for all teams.

The maximum time allowed for one task execution is 5 minutes. The time is calculated from the moment the robot confirms the start message has been received, until the robot indicates the end of the benchmark. If 5 minutes is exceeded, a timeout is recorded for that execution, and the robot must prepare for the next execution.

4.2.5 Autonomy Level

Teams are allowed to choose any level of autonomy for this task, from fully autonomous to fully remote-controlled. The teams must specify the autonomy level beforehand for each functionality. For this task benchmark, autonomy levels could include:

- Navigation: [fully remote controlled, remote waypoint specification, fully autonomous]
- Object Detection: [remote detection, fully autonomous]
- Grasping: [fully remote controlled, semi-autonomous with fine-tuned remote control, fully autonomous]
- Handover: [fully remote controlled, semi-autonomous with remote-controlled release, fully autonomous]

4.2.6 Benchmarking Data

The internally recorded data must include (at minimum):

- RGB camera stream of the robot
- Proprioceptive sensor data from the robot's manipulator and base
- Output of object detection (if any)
- Output of human or face detection (if any)
- Position of end-effector with respect to the robot base
- Events defining the start and end of different phases of the task

In addition to the recorded internal robot data, an external RGB camera will record each run.

4.2.7 Scoring and Ranking

The performance of the robot is based on the following achievements for each execution:

1. *Navigation to item location*: The robot is considered to have successfully navigated to the location of the item if it is less than 1 m away from the item and it is visible in the robot's camera. The robot must return an image from its camera to complete this achievement. [**Achievements: 1**]
2. *Detection of item*: The robot must return a 2D bounding box of the detected item and a corresponding RGB image to complete this achievement. [**Achievements: 1**]
3. *Pick item*: The robot is considered to have picked the item if the item makes contact with only parts of the robot (fingers, tray, etc.) and no other surface for a minimum of 5 seconds. [**Achievements: 1**]
4. *Navigation to the person with item*: The robot is considered to have successfully navigated to the person if it is less than 2.5 m away from the person, and is visible in the robot's camera. The item must be transported with the robot and the robot must return an image from its camera to complete this achievement. [**Achievements: 1**]
5. *Detection of the person*: The robot must return a 2D bounding box of the detected person and a corresponding RGB image to complete this achievement. [**Achievements: 1**]

6. *Detection of person's pose*: The robot must determine whether the person is standing, sitting or laying down and return the result. [**Achievements: 1**]

7. *Handing over of item*: The robot must successfully hand over the item to the person to complete this achievement. A successful hand-over is one in which the person has the item in their hand at the end of the interaction. [**Achievements: 1**]

8. *Detection of failure cases*: In case the person does not reach out for the object, or drops the item during or after the handover, the robot must detect this and report it in the result (Note: see penalties below for dropped objects) [**Achievements: 1**]

9. In addition to the achievements awarded for the intermediate checkpoints, the referee, the volunteer who interacts with the robot or third-parties shall score the human-robot interaction aspect of the task by evaluating usability, social acceptance and user experience, by using the USUS evaluation framework [3]. Teams should consider the manner of approaching the person, motions of the manipulator, intuitiveness of the handover, timing of the release, etc.

In case multiple runs are executed during a competition, the sum of achievements for all runs is calculated.

Penalties

- the robot collides with environment or human in an uncontrolled manner
- the robot drops the object before the person makes contact with the object. In case of ambiguous outcomes, the referee makes the final decision to decide if the robot was at fault.
- the robot does not release the object to the human
- the robot stops responding

Disqualifying Behaviours

- the robot damages the environment
- the robot damages the object

5 Testbed

5.1 Testbed Environment

The testbeds for METRICS HEART-MET resemble typical living environments with areas such as a living room, dining room and kitchen. The field campaign at the University of Nottingham takes places in the Cobot Maker Space[6], which is a 99 m^2 facility with robots and equipment for human-robot interaction research. The Living Space, which is one of the sections of the Cobot Maker Space, resembles a standard living environment, in which all of the benchmarks for METRICS HEART-MET will be conducted.

The Robotic Assisted Living Testbed at Heriot-Watt University, which is made available to participating teams for testing, consists of a 60m^2 simulated apartment with an open-plan living, dining and kitchen area, along with a bedroom and bathroom. Several sensors such as a motion capturing system, CCTV cameras, RFID floor and a device-free-sensing system for indoor localization and monitoring are also available. Robots in the lab are remotely accessible via Anydesk[7]. External users can run their code on the robots by providing a Dockerfile with the necessary software packages. While participating teams are not required to have access to such facilities for preparation, such test beds are open for teams to test their robots.

[6] https://cobotmakerspace.org/
[7] https://anydesk.com/

5.2 Objects in the Environment

The objects in the environment which the robot has to interact with or recognize include both general domestic objects and healthcare-related objects. Some examples of domestic objects include towels, cups, plates and cutlery, general food items, pillows, etc. Healthcare related objects include medicine boxes or bottles, insulin pen, first-aid kit, inhaler, crutches etc. Some sample objects can be seen in Appendix A.

5.3 Referee Box

Communication with the testbed will be done via a referee box. In particular, this will be used to indicate the start and end of a benchmark to the robot, and for the robot to send feedback if required by a benchmark.

The referee box:

- is able to communicate wirelessly to the robot, both to send and receive messages,
- can be controlled by the referee to initiate or end a benchmark,
- stores feedback sent by the robot, and
- records start and end time of each run

The referee box, and the counterpart client which runs on the robot can be found on Github[8].

5.3.1 Benchmark Data Collection

METRICS benchmarking is based on the processing of data collected in two ways:

- **internal benchmarking data**, collected by the robot system under test, such as video and proprioceptive sensors ;
- **external benchmarking data**, collected by the equipment embedded into the testbed

Instructions for recording data on the robot are included referee box software package.

The external benchmarking data collection equipment will include video cameras recording the runs, ambient sensors (if available) in the testbed and messages and timestamps recorded by the referee box.

6 Robots and Teams

The content of this section has been adapted from the RoCKIn@Work rulebook[9].

The purpose of this section is twofold:

1. It specifies information about various robot features that can be derived from the environment and the targeted tasks. These features are to be considered at least as desirable, if not required for a proper solution of the task. Nevertheless, we will try to leave the design space for solutions as large as possible and to avoid premature and unjustified constraints.
2. The robot features specified here should be supplied in detail for any robot participating in the competition. This is necessary in order to allow better assessment of competition and benchmark results later on.

6.1 General Specifications and Constraints on Robots and Teams

Robot Specification 6.1 (*System*)

A competing team may use a single robot or multiple robots acting as a team. It is not required that the robots are certified for industrial use. At least one of the robots entered by a team is capable of:

- *mobility and autonomous navigation.*

[8] https://github.com/HEART-MET/metrics_refbox
[9] http://rockinrobotchallenge.eu/rockin_d2.1.6.pdf

- *manipulate and grasp at least several different task-relevant objects. The specific kind of manipulation and grasping activity required is to be derived from the task specifications.*

The robot subsystems (mobility, manipulation and grasping) should work with the environment and objects specified in this rule book.

Robot Specification 6.2 (*Sensor Subsystems*)

Any robot used by a team may use any kind of ***onboard*** *sensor subsystem, provided that the sensor system is admitted for use in the general public, its operation is safe at all times, and it does not interfere with other teams or the environment infrastructure. A team may use the sensor system in the environment provided by the organizer by using a wireless communication protocol specified for such purpose. Sensor systems used for benchmarking and any other systems intended for exclusive use of the organizers are not accessible by the robot system.*

Robot Specification 6.3 (*Communication Subsystems*)

Any robot used by a team may ***internally*** *use any kind of communication subsystem, provided that the communication system is admitted for use in the general public, its operation is safe at all times, and it does not interfere with other teams or the environment infrastructure. A robot team must be able to use the communication system provided* ***as part of the environment*** *by correctly using a protocol specified for such purpose and provided as part of the scenario.*

Robot Specification 6.4 (*Power Supply*)

Any mobile device (esp. robots) must be designed to be usable with an onboard power supply (e.g. a battery). The power supply should be sufficient to guarantee electrical autonomy for a duration exceeding the periods foreseen in the various benchmarks, before recharging of batteries is necessary. Charging of robot batteries must be done outside of the competition environment. The team members are responsible for safe recharging of batteries. If a team plans to use inductive power transmission devices for charging the robots, they need to request permission from the event organizers in advance and at least three months before the competition. Detailed specifications about the inductive device need to be supplied with the request for permission.

Robot Constraint 6.1 (*Computational Subsystems*)

Any robot or device used by a team as part of their solution approach must be suitably equipped with computational devices (such as onboard PCs, microcontrollers, or similar) with sufficient computational power to ensure safe autonomous operation. Robots and other devices may use external computational facilities, including Internet services and cloud computing to provide richer functionalities, but the safe operation of robots and devices may not depend on the availability of communication bandwidth and the status of external services.

Robot Constraint 6.2 (*Safety and Security Aspects*)

For any device a team brings into the environment and/or the team area, and which features at least one actuator of any kind (mobility subsystems, robot manipulators, grasping devices, actuated sensors, signal-emitting devices, etc.), a mechanisms must be provided to immediately stop its operation in case of an emergency (emergency stop). For any device a team brings into the environment and/or the team area, it must guarantee safe and secure operation at all times. Event officials must be instructed about the means to stop such devices operating and how to switch them off in case of emergency situations.

Robot Constraint 6.3 (*Operation*)

In the competition, the robot should perform the tasks autonomously, unless the team chooses an autonomy level which allows for human control An external device is allowed for additional computational power. In case of fully-autonomous operation, it must be clear at all times that no manual or remote control is exerted to influence the behavior of the robots during the execution of tasks.

Robot Constraint 6.4 (*Environmental Aspects*)

Robots, devices, and apparatus causing pollution of air, such as combustion engines, or other mechanisms using chemical processes impacting the air, are not allowed. Robots, devices, and any apparatus used should minimize noise pollution. In particular, very loud noise as well as well-audible constant noises (humming, etc.) should be avoided. The regulations of the country in which a competition or benchmark is taking place must be obeyed at all times. The event organizers will provide specific information in advance, if applicable. Robots, devices, and any apparatus used should not be the cause of effects that are perceived as a nuisance to humans in the environment. Examples of such effects include causing wind and drafts, strong heat sources or sinks, stenches, or sources for allergic reactions.

6.2 Benchmarking Equipment in the Robots

Hardware

- Teams might have to install a USB-stick during the runs for storing the data.
- The robots need to have WiFi-connectivity for communication with the Referee Box

Software A referee box client and a ROS bagfile recorder are expected to run on the robot. The client relays messages from the referee box to internal software components on the robot, and the ROS bagfile recorder starts and stops recording of ROS bagfiles at the start and end of a benchmark trial execution.

Recorded Data The data required to be recorded internally by the robot is dependent on the FBM or TBM. Some common data streams that must be recorded (if available) include the following:

- base velocity commands
- odometry
- TF tree
- joint states
- camera (RGB, depth, camera calibration)
- sound
- laser (if applicable for the benchmark)
- other sensors such as force-torque, tactile, IR

While a particular sensor might not be used by the teams for a particular task, it is preferable to record it since it might be used by competing teams during the cascade evaluation campaigns. It is expected that participating robots will be heterogeneous, therefore the exact set of variables and sensors that will be recorded will be determined on site per robot.

References

[1] F. Amigoni, E. Bastianelli, J. Berghofer, A. Bonarini, G. Fontana, N. Hochgeschwender, L. Iocchi, G. Kraetzschmar, P. Lima, M. Matteucci, *et al.*, "Competitions for benchmarking: Task and functionality scoring complete performance assessment," *IEEE Robotics & Automation Magazine*, vol. 22, no. 3, pp. 53–61, 2015.

[2] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *European Conference on Computer Vision*. Springer, 2016, pp. 510–526.

[3] A. Weiss, R. Bernhaupt, M. Lankes, and M. Tscheligi, "The usus evaluation framework for human-robot interaction," in *AISB2009: proceedings of the symposium on new frontiers in human-robot interaction*, vol. 4, no. 1, 2009, pp. 11–26.

A Sample Objects

Table 3 lists some sample object which may be used for the various benchmarks. Additional household and healthcare-related objects may be introduced before and during the competition.

Table 3: Sample objects

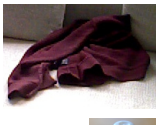
Object	Sample image
Cup	
Plate	
Bowl	
Towel	
Shoes	
Sponge	
Bottle	
Toothbrush	
Toothpaste	
Tray	
Sweater	
Medicine bottle	

Table 3: Sample objects

Object	Sample image
Reading glasses	
Flashlight	
Pill box	
Book	
Basket