Metrological evaluation and testing of robots in international competitions

---

METRICS HEART-MET Field Evaluation Campaign Rulebook

Latest update: May 24, 2023

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

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/Acronym	**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

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

Contents

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

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 2nd Field Campaign, taking place in Florence, Italy from September 25 - 29, 2023.

2 Competition Structure

The second field campaign for HEART-MET takes place in a test-bed which resembles a living environment. 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 and person detection. The TBMs evaluate the capability of the robot to execute tasks such as sorting objects, which requires the robot to integrate functionalities.

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.

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 several functionality benchmarks with their associated metrics and procedure for execution.

[1] http://rockinrobotchallenge.eu/

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

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.

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 20 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 20 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.

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

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.

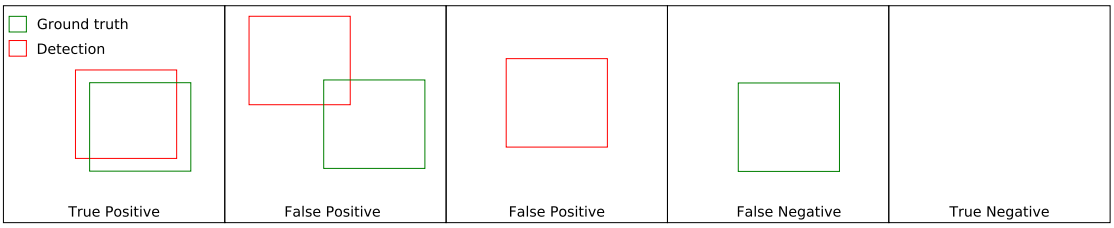

Figure 1: Metrics for object detection

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.

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.

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

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

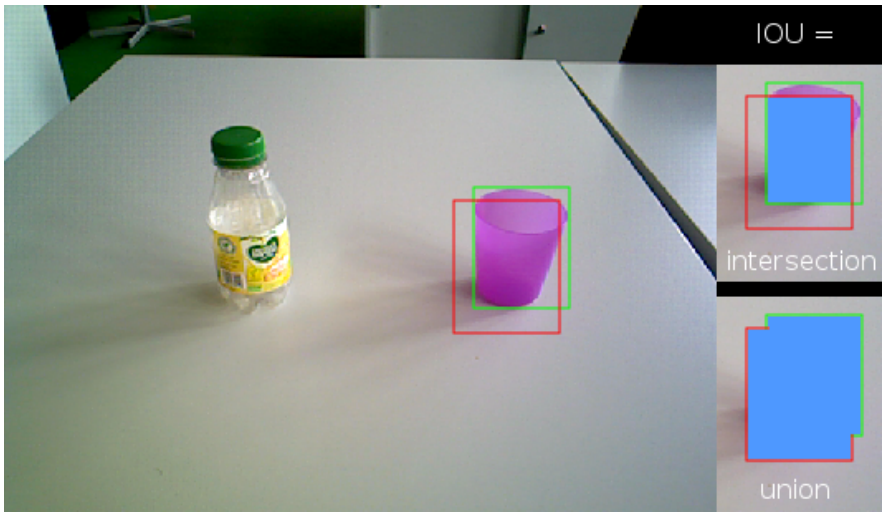

Figure 2: Example object detection; the green box shows the ground truth, and the red box shows the detection

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: 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 [straight, 30° left, 30° right]

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

- 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.

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 metrics used are identical to those used for the Object Detection benchmark. Therefore, teams are ranked based on

- the sum of 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.

Penalties

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

Disqualifying Behaviours

- the robot damages the environment

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

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 Object Sorting Task

4.1.1 Task Description

This task benchmark evaluates a robot's ability to perceive, grasp and sort objects based on a given criteria. The robot is placed in front of a table with a set of objects and containers for each category of the objects to be sorted. The robot needs to recognize the objects, grasp them, and place them in the container corresponding to the category of the object. The criteria for sorting the objects will differ based on the objects, but will be defined beforehand; for example, colour of the object, material of the object (plastic, metal, etc.), class of the object (forks, spoons, knives), etc.

4.1.2 Healthcare Relevance

An assistive robot operating in a home or healthcare environment may need to assist persons with physical impairments with household tasks which involve sorting objects. These may include sorting clothes, separating trash for recycling, storing groceries[3], storing cutlery and dinnerware from a dishwasher, etc.

4.1.3 Feature Variation

The independent variables for this functionality are:

- the set of objects and their poses
- the criteria by which the objects need to be sorted
- the pose and size of the target containers

The dependent variable is the number of correctly sorted objects.

4.1.4 Communication with the Referee Box

- The robot waits for a start message from the referee box. This message contains the criteria by which the objects need to be sorted.
- 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:
  - grasped an object
  - placed an object in the correct container
- The robot sends a message indicating the completion of the benchmark

[3]See example in Sec. 5.9 in the RoboCup@Home rulebook: `https://athome.robocup.org/wp-content/uploads/2022_rulebook.pdf`

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

4.1.5 Procedures and Rules

Before starting a trial, the referee places a set of objects and target containers on a table. As far as possible, the arrangement of objects is similar for all teams, by taking a picture of the arrangement, and replacing them for the next team by visual inspection. Objects may be cluttered, with objects overlapping each other. The configuration of objects, containers and the sorting criteria for a particular trial is fixed for all teams.

The robot is placed in front of the table by the teams, and waits for the start message. Once the start message is received, the robot can grasp objects and place them in their corresponding target containers. The target container for a category is specified beforehand by the referee.

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

4.1.6 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

4.1.7 Scoring and Ranking

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

1. Successful placement of an object in the right container [Achievements: 1]

Since the robot should sort multiple objects in a single trial, the total achievements for a trial is the total number of correctly sorted objects. In case multiple trials are executed during a competition, the sum of achievements for all trials is calculated. For a given trial configuration, all participating teams will execute the trial sequentially at the same location and time of day.

Penalties

- the robot collides with environment or human in an uncontrolled manner
- the robot drops an object off the table
- the robot stops responding

Disqualifying Behaviours

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

The final score is calculated as Achievements - 0.1 * Penalties. In case of a disqualifying behaviour, the final score for that trial is zero. In case of a tie, the number of incorrectly placed objects is counted, and the team with the lower number of incorrectly placed objects wins the tie-break.

5 Testbed

5.1 Testbed Environment

The testbeds resemble typical living environments with areas such as a living room, dining room and kitchen. For example, the testbed at Heriot Watt 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

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

system, CCTV cameras, RFID floor and a device-free-sensing system for indoor localization and monitoring are also available. At UWE, the Assisted Living Studio is a modular apartment instrumented with a network of Z-wave sensors, Wi-Fi cameras PIR sensors etc. At the University of Nottingham, the Cobot Maker Space includes a simulated apartment with smart home sensors. At BRSU, the test bed consists of a living room, dining room, a lounge area and a fully-functional kitchen. The test bed at Florence will contain similar elements as required for the benchmarks.

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. The exact list of objects which need to be recognized or interacted with will be released closer to the date of the field evaluation campaign.

5.3 Benchmarking Equipment in the Environment

5.3.1 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 features of such a referee box are as follows:

- 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
- records start and end time of each run

5.3.2 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

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[4].

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.

The description of the robot should be included in the team description paper.

[4] http://rockinrobotchallenge.eu/rockin_d2.1.6.pdf

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

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.*
- *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

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

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. An external device is allowed for additional computational power. 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

- The robot needs to have the software packages to run *rosbag record*
- A ros package will be provided which makes it easier to trigger starting and stopping the recording of bagfiles autonomously

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.

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