1st International Workshop on Evaluating Progress in Artificial Intelligence - EPAI 2020
In conjunction with the 24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

Design and validation of testing facilities for weeding robots as part of ROSE Challenge

Guillaume Avrin[1] and **Daniel Boffety** [2] and **Sophie Lardy-Fontan**[1] and **Rémi Régnier**[1] and **Rémi Rescoussié**[2] and **Virginie Barbosa**[1]

Abstract. The ROSE Challenge is the first robotics and artificial intelligence competition worldwide to implement a third-party performance evaluation of intra-row weeding robots in real conditions, under comparable conditions, to guarantee a credible and objective assessment of their effectiveness. This article reports on the design and validation of testing facilities for this competition, which presents a particular complexity : the experiments take place outdoors and act on living things (crops and weeds). Moreover, it implies to guarantee repeatable experimental conditions for a comparable and equitable evaluation. The article also discusses the opportunity that these competitions represent to define testing facilities in a consensual way. The method it proposes is very widely applicable to different fields of intelligent systems applications.

1 INTRODUCTION

In 2017, the French Ministries of Agriculture, Ecological Transition and Research, in partnership with the French National Research Agency (ANR), financed and defined the objectives of a robotics and artificial intelligence (AI) competition called *ROSE Challenge* (RObotics and sensors at the Service of Ecophyto) [2]. They commissioned the Laboratoire national de métrologie et d'essais (LNE) and Institut national de recherche pour l'agriculture, l'alimentation et l'environnement (INRAE) to define the evaluation plan and testing facilities, as well as to organize the competitions and analyze the results.

Challenges are specific funding instruments of the ANR that aim to compare, simultaneously and during evaluation campaigns, the performances of several technological and scientific solutions in relation to a specific theme and predefined objectives (see for example [7, 11, 13]. They are an essential tool for structuring and mobilizing industrial and academic players, making it possible to remove scientific obstacles and accelerate technological developments and transfers [4].

These competitions do not simply allow the development of intelligent systems to respond to a given problem, they are a unique opportunity to develop and test environments [10]. The latter are generally subject to rigorous specification, as for example in the ERL competitions [12], according to the recommendations of RoCKIn, and Robucup [3]. The NIST proposes to accompany the reproduction of these test environments on other geographical sites within the framework of intervention robotics [9]. Ultimately, these test environments are intended to be the subject of standards for testing and certification activities on intelligent systems [1].

However, the contribution of challenges in the field of agricultural robotics proved to be limited in comparison with those in the fields of industry, rescue or service. The agBOT Challenge [8] and the Field Robot Event are exceptions but are limited in scope and not addressing the problem of intra-row weed control which remains one of the major block in the field, especially for the small inter-row spacing crops. The ROSE Challenge represents a unique opportunity to promote the implementation of test facilities for agricultural robotics that are truly relevant to state-of-the-art, promising an acceleration of technological progress and a closer relationship with the market in the coming years.

The article will first present the need for rigorous and reproducible evaluation within the ROSE Challenge, whose context, objectives and evaluation tasks are recalled. It then presents the response that the ROSE organizers have provided to this need, detailing the testing facilities and tools deployed.

2 ROSE Challenge presentation

2.1 Context and objectives of the competition

The ROSE Challenge aims at encouraging the development of innovative technological solutions contributing to the objectives of the Ecophyto II plan carried out by the French Ministries in charge of Research, Agriculture and Ecology : to reduce the use of phytosanitary products by 50%.

The ROSE Challenge focuses on intra-row weed control (spacing between plants on the same crop line) in large-spacing crops and in row-cropped vegetables. Since the start of the challenge in January 2018 and for four years, the participating teams (see Section 2.4) have been competing on the experimental field of the AgroTechnoPôle at the Montoldre INRAE site in France during annual evaluation campaigns conducted by the LNE and INRAE.

2.2 Crops considered

The ROSE Challenge involves both large-spacing crops and vegetable crops (see Figure 1). The crops considered during the evaluation are:

- *Maize (Zea Mays)* (large-spacing crops): inter-row spacing 75 cm, 15 cm intra-row spacing. Two crop rows are planted on the 46.5 m long experimental plot (see Section 3.2.1).
- *Bean (Phaseolus vulgaris)* (vegetable crops): inter-row spacing 37.5 cm, intra-row spacing 7-8 cm. Three crop rows are planted on the 46.5 m long experimental plot.

[1] Evaluation of artificial intelligence Department, Laboratoire national de métrologie, France, email: guillaume.avrin@lne.fr
[2] Institut national de recherche pour l'agriculture, l'alimentation et l'environnement, France, email: daniel.boffety@inrae.fr

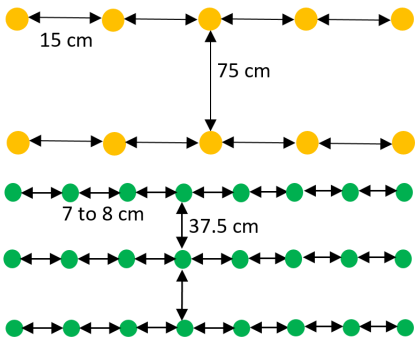

Figure 1: Implantation: two rows of maize (top) and three rows of bean (bottom) on the agricultural plot

The weeds planted are of two types:

- Spreading "type" (horizontal growth pattern): Wild mustard (Sinapis arvensis) and Matricaria (Matricaria chamomilla).
- Upright "type" (vertical growth pattern): Ryegrass (Lolium perenne) and Lamb's quarter (Chenopodium album).

For these weeds, the normal sowed density is 54 seeds per linear metre in the intra-row area.

2.3 Evaluation tasks

The ROSE Challenge takes into account the entire chain of intervention: observation and detection of crops and weeds, interpretation/decision-making, weeding action. Thus, three main evaluation tasks are considered and presented in more detail in the following paragraphs:

- the detection of weeds and crops;
- the weeding action;
- the global intervention of the robot (involving the whole detection-decision-action chain).

2.3.1 Detection task

Concerning detection, only terrestrial proxid detection is considered. Systems are evaluated and compared on their capacity to determine the position of weeds and/or crops during a trial on the same plot. This capacity is directly evaluated from the multimodal images (visible, multispectral, hyperspectral) generated by the participating solutions during their trial. The evaluation, for a given trial, is done in two steps:

- step 1: the consortium's robot passes over the crops row, collects the "raw" images, the algorithm automatically annotates the images (during or just after the trial). All the source images and automatic annotations (called "hypotheses") are transmitted to the organizers following the acquisitions;
- step 2: the panel of annotators qualified by the organizers annotates a sample of images selected at random from the images collected by the systems evaluated. These references are then compared to the systems' hypotheses to quantify the performance (see Section 3.3.3).

On each image, the automatic annotations generated by the detection systems evaluated must:

- recognize the classes of plants present (weeds or crops);
- locate weeds and/or crops.

2.3.2 Weeding action task

During the weeding action task (weeding while preserving surrounding crops), the state of the participant plot (see Section 3.2.4) before the robot trial is compared to the state after the trial. To make this task as independent as possible from the "detection task" (see 2.3.1), the weeds to be weeded and the crops are identified by easily detectable markers of different colors (see Section 3.2.4).

2.3.3 Global weeding task

In contrast to the task concerning the evaluation of the weeding action module (see Section 2.3.2), the plants is not be pre-located by easily identifiable markers during the global evaluation task. Thus, this task allows for the evaluation of the entire detection-decision-action chain: the detection system and the weeding system, but also all the decisions taken during the intervention. The overall evaluation criteria consider the weeds destroyed and crops damaged.

2.4 Participating robotics and AI systems

Four consortia participate to the ROSE Challenge (BIBBIP, PEAD, ROSEAU and WEEDELEC[3]). It is important to note that the ROSE challenge organizers, as an independent trusted third party, have not been involved in the selection of these participants. Their robots are very diverse and differentiated by their detection module (visible, infrared or hyperspectral cameras), their decision algorithms (dynamic mapping tools, use of multi-scale maps, partial or full autonomy, etc.) and their actuators (mechanical, electrical shock, etc.). It is undesirable to limit this heterogeneity, which makes it possible to explore different ways to overcome the scientific and technological obstacles identified by the ROSE initiative. On the other hand, this disparity makes the evaluation more complex and in particular the definition of testing facilities flexible enough and/or broad enough to cover all the robots participating in the challenge.

3 DEVELOPMENT AND VALIDATION OF TESTING FACILITIES

3.1 Design, validation and exploitation phases

The ROSE challenge is composed of several successive evaluation campaigns. These campaigns firstly allow the development of the testing facilities and secondly allow the participating robots to improve their performance thanks to reliable and regular performance measurements. Figure 2 presents the timeline of these different phases, which are also detailed in the follwing paragraphs.

3.1.1 Design of testing facilities during competitions

During the first six months of the challenge, the participating teams and the organizers work together to establish the best possible conditions for the evaluation campaigns. Meetings are organized for this purpose. These meetings enable a number of essential elements to be specified for the smooth running of the ROSE Challenge, although they are common to all the challenges:

- specifying the tasks on which the systems will be evaluated;
- designing the test environments (plots, crops and weeds to be set up, the color and size of the markers used during the weeding action task, etc.

[3] http://challenge-rose.fr/en

- formalizing the various technical, organizational and safety aspects;
- defining the metrics for measuring the performance of the systems that are quantitative, rigorous, comparable, repeatable and accepted by all;
- specifying the formats of the data input and output of the systems.

All of these elements constituting the evaluation protocol are transcribed in an evaluation plan, a reference document for the conduct of evaluation campaigns.

3.1.2 Validation of testing facilities during competitions

The two "dry-run" evaluation campaigns (in 2018 and 2019) are used to validate the evaluation protocol and facilities. They aim to test and correct the comparison tools and evaluation protocol set up by the challenge organizers (experimental plots, field data, etc.). It is also an opportunity for the participating teams get to know the testing facilities. Thus, participants can propose additions and modification, in particular concerning the layout of the plots (infrastructures and safety measures to be put in place, etc.).

3.1.3 Exploitation of testing facilities during competitions

The three official evaluation campaigns that follow the dry-run phase (one per year until the end of the project in 2021) will be used to evaluate the performance of the proposed solutions. Following each campaign, the results of the tests are announced during a workshop bringing together all the teams.

The evaluation plan, although largely established thanks to the dry-run campaigns, continue to be adapted throughout the challenge to accompany the evolution of the technological solutions proposed by the participants.

The next paragraphs describe the testing facilities used to allow a quantitative, rigorous, comparable, repeatable and accepted performance measurement of the systems.

3.2 Physical testing facility

3.2.1 Experimental field

Field evaluations are carried out on a four-hectare experimental field from INRAE site in Montoldre. On this field, protected, monitored, maintained, power supplied plots mixing the different types of crops and weeds are made available for the competition:

- a reference plot conducted in a conventional manner (using chemical products),
- a plot for each participant including different areas for the evaluation and another area for the robot setting (including adaptation to weather conditions) just before evaluation,
- a plot common to all participants in order to acquire images as part of the detection task (see Section 2.3.1).

At the end of the plots, impoundment areas about ten meters wide allow the technological solutions to move and make U-turns and/or change crop rows. Similarly, between each of the plots allocated to the participants, free areas six meters wide are available. In the longitudinal direction of the plot, free areas eight meters wide separate the areas allocated to each consortium to facilitate experimentation.

Before each evaluation campaign, a succession of technical interventions are carried out on the plot at the end of the winter period until the sowing of the crops/weeds planned two weeks before the evaluation period:

- destruction of the crops/weeds of the previous evaluation campaign by mowing and export,
- superficial tillage for mechanical destruction of weed growth (several passes depending on weeds cover density and weather conditions),
- soil preparation work for loosening and warming the soil,
- seedbed preparation with heat treatment (three different depths and speeds),
- sowing of crops and weeds at the desired densities (sowing carried out by an external service),
- the maintenance and delimitation of the plots allocated to the participants (delimitation of the intra-row areas by hoeing the inter-row areas, maintenance of the edges of the sown strips).

3.2.2 Reference plot

The conventionally treated reference plot (two meters wide and 46.5 meters long) is set up in order to constitute a gold-standard for the evaluation of the systems participating in the competition. Weeding of the reference plot is carried out using pre-emergence or post-emergence chemical intervention. The tillage, seedbed preparation without heat treatment and sowing of the crops on the reference plot are however similar to the plots assigned to the challenge participants.

3.2.3 Participant plot

For each type of crop selected (large-spacing and vegetable crops), a plot associated with a specific type of weed is made available to each participant. On these plots, a 10 cm wide intra-row area centered on the main crop seed line is provided with weeds for the participants' interventions. The inter-row crop area is weeded by the organizers.

The plot allocated to each team is divided into three areas corresponding to:

- a 10-meter area for robot settings prior to evaluation;
- a 10-meter area for evaluation of the weeding action task with the positioning of markers on crops and weeds (markers distributed over the two rows of maize or the three rows of bean);
- a 23-meter area for global evaluation;
- the remaining 3.5 meters allow to have buffer areas between the three previously defined areas.

In addition, for the detection task, two rows for large-spacing crops and three rows for vegetable crops with each specific type of weed chosen are shared by the participants.

3.2.4 Performance evaluation

Weeding action task: For the weeding action task, the metrics used are:

- the number of weeds before and after the weeding action to obtain the percentage of weeds destroyed,
- the number of crops with integrity before and after the weeding action to obtain the percentage of preserved crops.

Pictures are also taken to compare the evolution of the plants (before weeding action and just after weeding action) and keep a record of evaluations.

Table 1 presents the color-coded markers used for this task. These plastic disks are placed at the bottom of each plant (see Figure 3) with

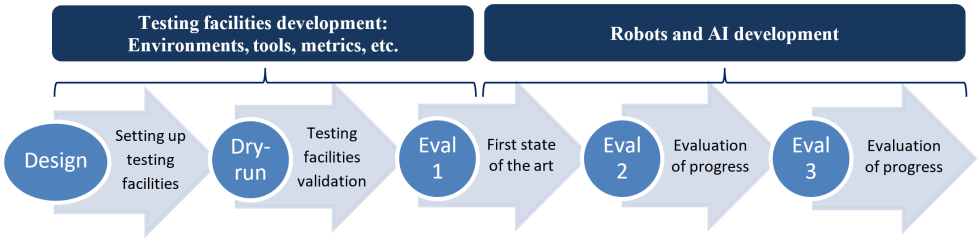

Figure 2: Testing facilities design, validation and exploitation phases during competition

the plant stem in the center of the marker (in the space provided). An opening (gutter) has been added to facilitate the installation of the marker without damaging the plant. The markers are made of opaque plastic with a thickness of about 1 mm.

Table 1: Characteristics of the markers according to the type of plant

Plant	Color(RGB)	Ext. Ø	Int. Ø	Opening width
Weed	Yellow(FFFF00)	2 cm	5 mm	3 mm
Crop	Blue(0000FF)	2.5 cm	8 mm	6 mm

In order to ensure maximum equity between participants, the organizers reproduce a similar difficulty on each plot by defining rules on the placement of markers (total number of markers, number of weed markers between two crops, distance between markers, etc.). Several configurations have been selected and are repeated a similar number of times for each of the participants plots:

- A single weed marker between two successive crops marked. To guarantee fairness in terms of difficulty, the markers are moved away from the crops (more than 3 cm) two times out of three (one time out of three the marker is placed less than 3 cm from the crop);
- Two weed markers between two successive crops marked, markers at a distance of more than 3 cm from the crops;
- Three weed markers between two successive marked crops, one marker close to the crops and two distant (more than 3 cm) from the crops.

Each set of markers is pictured and recorded so that, if necessary, it can be used for confirmation of the results after the weeding action.

Global evaluation task: Eight to ten weeks after the robot intervention on the plot during the evaluation campaign, an estimate of the dry biomass of the plot is made. The ratio between the biomass of the evaluated plot and the reference plot is used as a metric to assess the weeding performance of the robot. The measurements of the plot biomass is carried out using crop and weed still present in the intra-row.

As a complement, crops and weeds counts are carried out at different times (before the last trial of the solution on the plot, after the last trial of the solution on the plot, on D+3 with a tolerance of +/- 1 day depending on the constraints). A picture is taken every week to record the evolution of the plots for a minimum of two weeks (adapted according to the weather conditions).

In addition to the detection and action modules of the robot, this global evaluation assesses the technical itinerary:

- the choice of the intervention time, which depends on the constraints imposed by the crop (e.g. it is necessary to intervene when the weeds are at an early stage of development) and those imposed by the robotic system (e.g. the weeds must be sufficiently developed to be detectable);
- the choice of whether or not to proceed the weeding action on a weed when there is a risk to damage nearby crops.

Figure 3: Weeding action task: color-coded markers indicate weeds to be weeded (yellow) and crops to be preserved (blue)

3.3 Virtual testing facility

3.3.1 Collected images

The images collected by the detection systems of the different robots being evaluated are generated by visible, multispectral and hyperspectral cameras. The images show weeds and/or crops in varying proportions in each image, and may not even show any weed or crop at all. The images contain the different plants considered as part of the ROSE Challenge: crops (maize and beans) and weeds (matricaria, lamb's-quarters, mustard and ryegrass). Detection systems being evaluated must collect and submit to the organizers at least 5 images per meter traveled on the experimental plot. Image acquisitions are made by all participants on the same evaluation plot. The trials of the different robots are carried out in a very limited time frame in order to minimize variations in environmental conditions, the order of trials being drawn at random.

3.3.2 Automatic annotation

Hypothesis annotation: The systems have to produce their own hypotheses when acquiring images of the plot (see Section 2.3.1). After each trial on the plot, each consortium submit the files containing the raw images and the files containing the hypothesis annotations. For each collected image, the evaluated system has to return a file containing the hypothesis annotations as well as N bitmap images, N being the number of plants detected in the image. Each bitmap image defines the location of a specific plant and classically corresponds to a detection mask. A pixel of intensity 0 denotes a point not belonging to the plant. All pixels belonging to the plant have the same intensity and are of maximum intensity (intensity of 255).

Reference annotation: The images used for the evaluation are annotated a posteriori by human experts, under the supervision of the challenge organizers. A complete annotation guide is available to annotators. It contains details on the nature of the annotations. Each image annotated by the experts contains the following information:

- plant trimming (manual annotation): delimitation of each plant visible in the image by a polygonal box that will be as small as possible while encompassing the entire plant under consideration;
- plant type (manual annotation): a label is associated with each bounding box indicating the type of the plant (”weed”, ”crop” or, in rare cases, ”undetermined”), its common name (if the growth stage of the plant allows its identification) and the growth stage of the plant;
- growth stage (manual annotation): indication of the growth stage of each plant (”0”: early emergence; ”1”: seedling; ”2”: several leaves; ”3”: advanced);
- common name (manual annotation): indication of the common name of each plant (name from the list or ”indeterminable”).

The images with all these manual annotations constitute the so-called ”reference” data. During the evaluation, these references are compared with the automatic annotations produced by the detection systems being evaluated. To ensure the quality of the annotations, 10% of the images per team/plant type combination (random selection of images) are annotated by two different experts following the previous ”inter-annotation” comparison studies. An intra-anotator measurement on 10% of the images is performed.

LNE-DIANNE annotation software: The test images are annotated by human experts using the LNE-DIANNE annotation tool, which automatically pre-cuts the plants visible on the image to save annotation time. This pre-annotation is based on two classification methods: k-means method or thresholding. The user has the possibility to modify the configuration of these algorithms. This pre-annotation is corrected manually by the annotator.

3.3.3 Performance evaluation

Mapping phase: The masks defined by the systems as well as the bounding boxes defined by the human annotators to indicate the location of plants (weeds and crops) are used to assess the performance of the detection. The evaluation starts with a first mapping phase. The first mapping step aims at associating one by one the detection areas defined by the systems using the masks (hypotheses) with those manually annotated as bounding boxes (references). The mapping selected for the challenge is the one that minimizes the sum of the numbers of pixels located outside the intersection of the areas associated one by one. Note that two areas cannot be associated if they do not have any pixel in common. Once the best mapping has been identified, some hypothesis masks may not be associated with any reference bounding box, either because they have no pixels in common with a reference box, or because each reference bounding box is already associated with another hypothesis mask. These masks will be called false positives. Similarly, some reference bounding boxes may not be associated with any mask, either because they have no pixels in common, or because all the masks in the hypothesis are already associated with other reference bounding boxes. These boxes will be called false negatives. Figure 4 shows examples of hypotheses (masks submitted by a participant), which are compared to the annotated reference image.

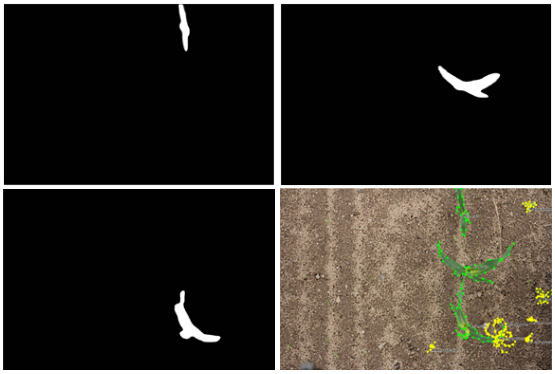

Figure 4: Detection task: Crops hypothesis annotations (top left and right, bottom left) are compared to reference (bottom right)

Performance metric: The evaluation metric is the Estimated Global Error Rate (EGER) [6]. For each annotated image of the reference, the lists of plants detected respectively by the system and by the annotators are generated automatically. These two lists are compared on the basis of the association of the detection areas defined by the mapping. An association between two plants classified in the same way by the system and by the annotators counts as correct. An association between two plants with different classes counts as confusion. Each plant of the non-associated assumption counts as a false positive, and each plant of the non-associated reference counts as an false negative. A penalty is associated with each confusion, each false negative and each false positive. The sum of the error counts per image gives the overall error count. The total number of expected entries is also counted by cumulating the number of plants present in the reference of each image. The error rate is then the global number of errors divided by the total number of expected entries. The penalties applied are as follows:

- penalty of 1 for forgetfulness/false positive,
- penalty of 2 for confusion.

The EGER metric used is:

$$EGER = \frac{\sum_{k=1}^{N} C_k + FA_k + O_k}{\sum_{k=1}^{N} NR_k} \quad (1)$$

Where C_k, FA_k and O_k represent respectively the sum of the penalties for confusion, false positive and false negative in the image k. NR_k represents the number of plants detected in the reference (weeds and crops). F-measurement, precision and recall scores are also be provided. The results of this evaluation is presented by type of plant (weeds or crops) and in a global manner taking into account both classes. The global evaluation process is presented in Figure 5.

LNE-MATICS software suite: LNE-MATICS is a free and open-source software suite designed for data mining and system evaluation. LNE-MATICS was originally designed for the evaluation of Automatic Language Processing systems [5]. It has been adapted to meet the evaluation needs of image processing systems and is used in the context of the ROSE Challenge.

4 INFLUENCE FACTORS

As the evaluation campaigns take place in an open environment and on living entities, many environmental factors may influence the performance of the solutions implemented. Some of these influencing

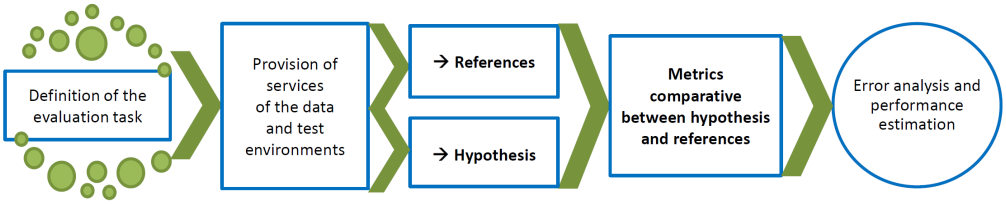

Figure 5: Automatic detection evaluation process

factors are controllable by campaign organizers. Others are difficult to control and at least need to be measured so that they can take them into account when analyzing the evaluation results. Still, the participants have to ensure that their robots are sufficiently robust with regard to non-controllable factors, so that they can be used in the real conditions by professionals in the agricultural sector. Two main types of influencing factors can be distinguished: testing modalities and agropedoclimatic factors.

4.1 Testing modalities

The testing modalities group together the parameters related to the experimental plot. First of all, the technical itinerary of the plot can be controlled because it is known before the intervention on the plot and is guaranteed to be the same for all participants. Crop and weed density and distribution are controllable factors. The actual state of the plantation is communicated to the consortia on a daily basis by means of image recording. The stage of development of crops and weeds is a measurable factor. The organizers communicate the level of plant growth on a daily basis, notably by taking images of the plot.

4.2 Agropedoclimatic factors

Agropedoclimatic conditions include weather and light conditions, as well as the soil's pedological and agronomic characteristics. The organizers monitor the weather conditions using weather stations together with other types of sensors such as soil moisture and temperature sensors. Light conditions, which cause shadow and glare problems disrupting the detection systems, will be measured using luxmeters. Soil characteristics such as humidity and temperature will be measured daily. Soil texture characteristics (clay content, etc.) of the plot were determined at the beginning of the challenge. These influence factors are presented Table 2. The participants are free to choose the date of intervention of their robot on the plot based on these measurements. In addition, the geo-referencing data acquired by GPS RTK when moving the seed drill on the plot during sowing are also shared with the participating robots to facilitate the location of the sowing lines and each intra-row area.

Table 2: Characteristics of the markers according to the type of plant

Type	Factor	Measurements taken
Test mod.	Technical itinerary	Described before eval.
Test mod.	Plants density and distribution	Pictures before eval.
Test mod.	Growth stage of plants	Daily image capture
Agropedoc.	Weather (temp., humidity, wind)	Daily
Agropedoc.	Brightness and solar radiation	Daily
Agropedoc.	Soil moisture and temp.	Daily
Agropedoc.	Leaf wetting, evapotranspiration	Daily
Agropedoc.	Clay content	Before first eval.

5 NEXT LEVEL EVALUATION CRITERIA

Consideration of some other evaluation criteria was premature in light of the current development of the technologies participating in the ROSE Challenge. These criteria are presented in the paragraphs below to open the door to future evaluations as part of the ROSE challenge (evaluation campaign of 2020 and 2021) and beyond. These new criteria will make it possible to bring even more singularity at this challenge compared to what is traditionally done in robotics testing.

Robustness Evaluation: The robustness of the detection systems could also be evaluated with respect to changes in the environment (weather, brightness, temperature, etc.). To evaluate this robustness, images acquired under different conditions could be considered for performance evaluation.

Flexibility: Although flexibility is related to the robustness of the solution, it may also involve evaluating the number of days available in the year for the use of robotic systems, taking into account the agro-environmental conditions.

Environmental impact: The environmental impact of the robot is an important criterion within ROSE Challenge which aims to reduce dependence on phytosanitary products. In case such products were used by a robot, measurements of the consumed products could be compared to those of conventional weeding of the reference plot in addition to a technical analysis of the solution (type of nozzle used, height/soil, pressure and flow rate used, forward speed, etc.). The quantity of product consumed could be estimated by the quantity of product in the tank before and after the evaluation trial. This quantity could be related to the area treated to estimate the efficiency of the use of these products and to allow for a comparison with the treatment frequency index (TFI). Other environmental criteria could also be considered:

- soil pollution (risks of leakage with the fuel, lubricant and hydraulic fluids, oils, etc.),
- carbon balance (cost of the production of the robot and the electric batteries, energy consumption while using the solution, etc.),
- soil settlement and compaction effects.

Theoretically, robot characteristics (tire size, inflation pressure, weight) would allow an evaluation of the risks of compaction incurred in relation to the type of soil encountered. Other specific measures in the field could also be considered (pressure, time and number of travels over an area).

Techno-economic criteria: An estimate could be made of the following aspects:

- the intervention time (working rate),

- the degree of automation (time spent by humans to plan/pilot the intervention),
- energy consumption (worst-case estimate),
- the energy autonomy of the solution; the duration of use (for each solution) between two refills could be used as a metric to evaluate the energy autonomy of the technological solutions.

A more in-depth study on the cost of using a specific solution (cost of techniques and materials used, human costs, operating costs, etc.) could also be considered.

Acceptability: Once the technologies have reached a sufficient level of maturity, an analysis of the acceptability of the technology by potential users (farmers, industry professionals, etc.) would be useful. This could be done through questionnaires. This acceptability analysis would be supplemented by an analysis of the risks incurred by users or local residents (exposure to products, maintenance of tools, proximity of the machine in operation, etc.) and an analysis of the arduousness of the work related to the use of the solutions (noise, need for maintenance, supervision, etc.).

6 CONCLUSION

The ROSE Challenge is the first initiative worldwide to put different robots in competition by including both image-based evaluation and field evaluation on agricultural plots. Indeed, this challenge makes it possible to carry out a modular evaluation of the different technological building blocks of the solutions participating in the challenge, as well as a global evaluation of weed control efficiency.

The testing facilities developed within this challenge will constitute useful consensual references for the characterization of future research and industrial projects in this field, which can be disseminate for standardization. In particular, the qualified and annotated test databases have, by the richness of their contents, a strong potential for dissemination. The creation of a reference corpus of images in the visible, multispectral and aligned hyperspectral is indeed a novelty that will allow comparative evaluations of different weed and crop detection technologies. These validated databases will be particularly useful to the community because, in the context of limiting the use of phytosanitary products, many innovative robotic machines wish to include automatic weed detection devices. These systems are based on algorithms learning from annotated images. Numerous image databases of weeds and crops in the visible spectrum exist, such as the free Pl@ntnet database, but for the moment no open hyperspectral image database is yet available, even though this technology is promising for digital agriculture.

The integration of technologies that are still not widely used in agricultural systems and tools, such as infra-red or hyperspectral cameras and their use in multimodal detection systems, dynamic mapping tools, automated platforms combined with precision treatment strategies, will make it possible to establish a major breakthrough in the process of providing farmers with multiple solutions to weed treatment problems on the crop rows. The research carried out will also be useful for applications other than those concerned by the ROSE Challenge. Future developments can indeed be imagined for other functionalities and tasks that can be carried out by these new tools at the service of all agricultural professionals.

Thus, the ROSE Challenge shows the real opportunity that competitions represent to develop innovative testing facilities, both for robotic systems and AI algorithms. It paves the way for other initiatives that will draw inspiration from it, starting with the H2020 METRICS (Metrological evaluation and testing of robots in international competitions) project started on 1 January 2020, which will in particular organize a competition in this field at European level. In order to help the construction of future challenges in the field, the created databases and the complete evaluation plan of the ROSE challenge are intended to become public at the end of the project.

7 ACKNOWLEDGMENTS

The authors wish to thank the G. Bernard, A. Delaborde, O. Galibert, B. Lalère, S. Lecadre and M. Veron for adapting the evaluation software LNE-MATICS and annotation software LNE-DIANNE to the needs of the ROSE Challenge and for their contribution to the definition of evaluation protocols. We would also like to thank all the people who contributed to the setting up of the ROSE Challenge testing facilities at the INRAE experimental site. This work is carried out with the financial support of the French Ministry of Agriculture and Food, the French Ministry for the Ecological and Solidary Transition, the French Ministry of Higher Education, Research and Innovation, and the French National Research Agency.

REFERENCES

[1] ASTM, 'E2566-17a, standard test method for evaluating response robot sensing: Visual acuity', Technical report, West Conshohocken, PA, (2017).
[2] G. Avrin, A. Delaborde, O. Galibert, and D. Boffety, 'Boosting agricultural scientific research and innovation', in *3rd RDV Techniques AXEMA February 23, 2019, SIMA, France*, number used., (2019).
[3] S Behnke, 'Robot competitions-ideal benchmarks for robotics research', in *Proc. of IROS-2006 Workshop on Benchmarks in Robotics Research. Institute of Electrical and Electronics Engineers*. IEEE, (2006. October).
[4] F Bonsignorio, A Del Pobil, and E Messina, 'Fostering progress in performance evaluation and benchmarking of robotic and automation systems', *IEEE Robotics and Automation Magazine*, **21**(1), 22–25, (2014).
[5] O Galibert, G Bernard, A Delaborde, S Lecadre, and J Kahn, 'Matics software suite: New tools for evaluation and data exploration', in *proc. 11th edition of the Language Resources and Evaluation Conference*, pp. 7–12, Miyazaki, (2018. May 2018). Japan.
[6] O Galibert and J Kahn, 'The first official repere evaluation', in *First Workshop on Speech, Language and Audio in Multimedia*, (2013).
[7] O Galibert, S Rosset, C Grouin, P Zweigenbaum, and L Quintard, 'Extended named entities annotation in ocred documents: From corpus constitution to evaluation campaign', in *LREC*, (2012).
[8] R Gerrish, 'Ready for the agbot challenge', *Resource Magazine*, **26**(3), 8–9, (2019).
[9] A Jacoff, E Messina, H Huang, A Virts, A Downs, and R Norcross, 'Standard test methods for response robots'. ASTM International Committee on Homeland Security Applications, (2010).
[10] A Jacoff, R Sheh, A Virts, T Kimura, J Pellenz, S Schwertfeger, and J Suthakorn, 'Using competitions to advance the development of standard test methods for response robots', in *Proceedings of the Workshop on Performance Metrics for Intelligent Systems*, pp. 182–189, (2012. March).
[11] J Kahn, O Galibert, L Quintard, M Carré, A Giraudel, and P Joly, 'A presentation of the repere challenge', in *2012 10th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pp. 1–6. IEEE, (2012. June).
[12] P Lima, D Nardi, G Kraetzschmar, R Bischoff, and M Matteucci, 'Rockin and the european robotics league: building on robocup best practices to promote robot competitions in europe', 181–192, (2016. June).
[13] I Oparin, J Kahn, and O Galibert, 'First maurdor 2013 evaluation campaign in scanned document image processing', in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5090–5094. IEEE, (2014. May).