Rémi Regnier, Guillaume Avrin, Virginie Barbosa, Daniel Boffety, Anne Kalouguine, Sophie Lardy-Fontan

Validation de méthodologies d'évaluation de solutions de désherbage autonomes, dans le cadre des projets Challenge ROSE et METRICS.

Volume 2, nº 1 (2021), p. 11-32.

<http://roia.centre-mersenne.org/item?id=ROIA_2021__2_1_11_0>

Revue Ouverte d'Intelligence Artificielle
Volume 2, n° 1, 2021, 11-32

Validation de méthodologies d'évaluation de solutions de désherbage autonomes, dans le cadre des projets Challenge ROSE et METRICS.

Rémi Regnier[a], Guillaume Avrin[b], Virginie Barbosa[b], Daniel Boffety[c], Anne Kalouguine[b], Sophie Lardy-Fontan[b]

[a] LNE, Département de l'évaluation de l'IA et de la robotique, 29 avenue Roger Hennequin, 78197 Trappes, France
E-mail : remi.regnier@lne.fr

[b] LNE, Trappes, France
E-mail : guillaume.avrin@lne.fr, virginie.barbosa@lne.fr, anne.kalouguine@lne.fr, sophie.lardy-fontan@lne.fr

[c] INRAE, Domaine des Palaquins, 40 route de Chazeuil, 03150 Montoldre, France
E-mail : daniel.boffety@inrae.fr.

Résumé. — Le Challenge ROSE est la première compétition mondiale de robotique et d'intelligence artificielle à mettre en place une évaluation par une tierce partie des performances des robots de désherbage intra-rang en conditions réelles et reproductibles, afin de garantir une évaluation crédible et objective de leur efficacité. Cet article rend compte de la conception et de la validation des installations d'essai pour cette compétition, qui présente une complexité particulière : les évaluations se déroulent en conditions réelles sur parcelles de cultures et visent des organismes (cultures et adventices). De plus, elles nécessitent de garantir la reproductibilité des conditions expérimentales pour assurer la comparabilité des résultats d'évaluation et l'équité de traitement des différents participants. L'article discute également de l'opportunité que représente ce challenge pour définir, de manière consensuelle, les moyens et méthodes de caractérisation de ces systèmes intelligents. Les outils développés dans le cadre de ce challenge établissent les références nécessaires à la conduite de recherches futures dans le domaine de la robotique agricole : les images annotées seront particulièrement utiles à la communauté et le protocole d'évaluation permettra de définir des méthodologies harmonisées au-delà du challenge ROSE.

Après avoir exposé les objectifs du challenge, l'article présentera la méthodologie et les outils développés et utilisés pour permettre une évaluation objective et comparable des performances des systèmes et solutions développées. Enfin, l'article illustrera ce potentiel d'harmonisation et de partage de références au travers de la compétition européenne ACRE du projet européen H2020 METRICS.

Mots-clés. — Intelligence Artificielle, évaluation, agriculture, robotique.

1. Introduction

L'organisation du Challenge ROSE et des compétitions METRICS poursuit deux objectifs complémentaires. Tout d'abord, dans une optique écologique, ROSE s'inscrit dans le cadre du plan Ecophyto II, qui vise la réduction de 50 % l'utilisation des produits phytopharmaceutiques pour réduire la dépendance de l'agriculture à ces produits et les risques associés. Ensuite, les compétitions ont été créées dans le cadre d'un projet européen H2020 pour l'évaluation et l'essai métrologique de robots dans les compétitions internationales. Un des domaines couverts par les compétitions METRICS étant la robotique pour l'agriculture, cette compétition « ACRE » est construite sur la base de ROSE afin de mettre à profit l'expérience acquise au cours du Challenge ROSE. En effet, les compétitions de robotique [5, 10, 21] et en IA [15, 14] sont un environnement idéal pour l'évaluation métrologique [1] de nouvelles solutions. Des méthodes de test standardisées, basées sur des compétitions, ont par exemple été proposées par le NIST [2, 12, 11] pour l'évaluation de robots d'intervention et de secours. L'organisation de ces compétitions doit suivre des règles de bonne pratique établies au fil des compétitions passées, au niveau Européen [17] ou mondial [16, 19, 23]. Un état de l'art plus complet de la conception d'environnements de test dans le cadre de compétitions de robotique est proposé dans [3].

Afin de susciter la mise au point de solutions technologiques innovantes contribuant à atteindre cet objectif, les ministères français en charge de l'Agriculture, de la Transition écologique, et de la Recherche ont lancé en 2017 en partenariat avec l'Agence Nationale de la Recherche (ANR), un appel à projets de recherche « Challenge ROSE ». Le challenge ROSE est le premier challenge au monde à mettre en œuvre une évaluation des performances des robots de désherbage de la zone d'intra-rang (zone d'une largeur de 10 cm centrée sur la ligne de semis de la culture principale) en grandes cultures à fort écartement et en cultures maraîchères de plein champ à faible écartement [4].

Depuis le début du challenge en janvier 2018 et pendant quatre ans, les quatre équipes participant au challenge se confrontent sur le terrain expérimental de l'Agro-Technopôle du site de l'institut national de recherche pour l'agriculture, l'alimentation et l'environnement (INRAE) de Montoldre dans l'Allier, à l'occasion de campagnes d'évaluation annuelles organisées sous la direction du consortium évaluateur constitué d'une tierce partie évaluatrice constituée du Laboratoire national de métrologie et d'essais (LNE) et d'INRAE.

L'ensemble de la chaîne d'intervention est évalué : détection des cultures et/ou adventices, interprétation/décision, action de désherbage. De manière complémentaire, des premières approches d'évaluation d'indicateurs économiques, sociologiques et écologiques sont également développés. Les outils développés dans le cadre du challenge tels que les banques d'images annotées et le protocole d'évaluation permettront notamment d'établir des outils et méthodes de référence pour l'évaluation des performances des systèmes robotiques agricoles.

2. Présentation du Challenge ROSE

Le projet ROSE a une durée totale de quatre ans. Il a débuté par une campagne de *dry-run* (campagne permettant de tester et valider le protocole d'évaluation, ainsi que de familiariser les consortiums à ce dernier) constituée de deux rencontres réalisées en 2018 et 2019. Cette dernière est suivie d'une rencontre par an jusqu'à la fin du challenge en 2021.

2.1. Campagnes d'évaluation

Le challenge ROSE se déroule sur 4 ans (2018-2021) et se compose de 5 campagnes de test (2 campagnes *dry-run* et 3 campagnes d'évaluation, la distinction entre les deux types de campagne étant présentée à la section 6.4). Une collaboration entre les équipes participantes et les organisateurs permet d'établir les conditions de déroulement de ces campagnes en accord avec les demandes spécifiques du financeur, à savoir les tests en plein champ sur plantes réelles. Cette définition incrémentale des méthodes d'évaluation en lien avec les participants rejoint les bonnes pratiques proposées par [12, 11]. Il s'agit notamment de :

- spécifier clairement les tâches sur lesquelles les systèmes seront évalués ;
- définir les environnements de test (parcelles, cultures et adventices à mettre en place) ;
- formaliser les différents aspects techniques, organisationnels et de sécurité ;
- fixer les métriques permettant une mesure des performances des systèmes qui soit quantitative, rigoureuse, comparable, répétable et acceptée par tous ;
- préciser les formats des données en entrée et sortie des systèmes qui pourront être traités par les outils d'évaluation.

L'ensemble des conditions définies au-dessus sont susceptibles d'évoluer entre les campagnes *dry-run* et les campagnes d'évaluation, afin de prendre en compte les retours d'expérience et l'évolution des systèmes évalués.

2.2. Choix de cultures et d'adventices

Dans le cadre du challenge ROSE sont considérées à la fois les grandes cultures à fort écartement entre les rangées de plantes et les cultures maraîchères de plein champ dont les écartements inter-rangs sont plus restreints. Les espacements entre les rangs ont évolué entre les campagnes, les valeurs présentées sont celles utilisées pour la campagne d'évaluation de 2020 (2.1).

Les plantes d'intérêt pour l'évaluation sont les suivantes :

- *Maïs (Zea maïs)* (cultures à fort écartement) : espace entre les rangées de 75 cm, espace entre les plants 15 cm. Une parcelle expérimentale de 46,5 m de long est constituée de deux rangées de plants de maïs.

- *Haricot (Phaseolus vulgaris)* (cultures maraîchères) : espace entre les rangées de 37,5 cm, espace entre les plants 7-8 cm. Une parcelle expérimentale de 46,5 m de long est constituée de 3 rangées de plants de haricot.

Les adventices implantées sont de deux types :

- Type « port étalé » (étalement horizontal) : Moutarde sauvage (Sinapis arvensis) et Matricaire (Matricaria chamomilla).
- Type « port érigé » (étalement vertical) : Raygrass (Lolium perenne) et Chenopode (Chenopodium album).

La densité de semis d'adventices est de 27 graines par mètre linéaire de rang de plantes d'intérêt. Cette valeur a également évolué relativement à la densité proposée lors de la première campagne *dry-run*.

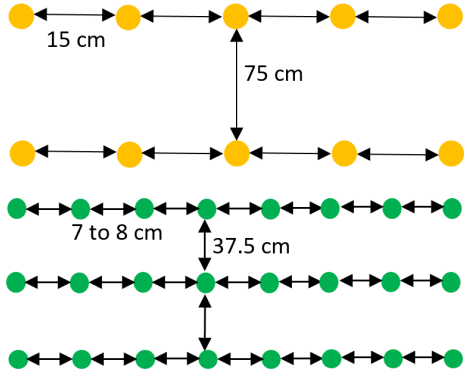

FIGURE 2.1. Implantation des deux cultures principales de maïs et haricot sur une parcelle expérimentale. (Deux rangées pour le maïs, trois pour le haricot.)

2.3. TÂCHES D'ÉVALUATION

Le Challenge ROSE se positionne sur l'ensemble de la chaîne d'intervention pour le désherbage robotisé : la détection des cultures et/ou des adventices, l'interprétation et la prise de décision en accord avec la situation, et l'action de désherbage. Ainsi, trois tâches d'évaluation principales sont considérées et présentées plus en détail dans les paragraphes suivants :

- la détection de plantes d'intérêt et/ou d'adventices ;
- l'action de désherbage (efficacité de l'effecteur) ;
- L'intervention du robot dans sa globalité sur la parcelle (incluant la chaîne détection–prise de décision–action).

2.3.1. *Tâche de détection*

Seule la proxi-détection est considérée dans ce challenge (ce qui exclut les solutions à base de drones pour cette tâche). Les systèmes sont évalués et comparés sur leur capacité à déterminer la position des adventices et/ou des cultures lors d'un essai sur une parcelle commune. Cette capacité est évaluée directement à partir des images multimodales (visibles, multi-spectrales, hyper-spectrales) générées par les systèmes. L'évaluation, pour chaque participant, se fait en deux étapes :

- étape 1 : la solution passe sur la ligne de culture commune à l'ensemble des participants, collecte les images « brutes » sans effectuer d'action de désherbage, l'algorithme annote automatiquement les images (pendant ou juste après le passage sur la ligne de cultures). L'ensemble ses images sources et des annotations automatiques (appelées « hypothèses ») est transmis aux organisateurs dans un délai restreint après cette phase d'acquisition ;
- étape 2 : un panel d'annotateurs qualifiés par les organisateurs annote un échantillon de 250 images tirées au sort parmi les images collectées par les systèmes évalués. Ces références sont ensuite comparées aux hypothèses des systèmes afin de quantifier les performances (voir section 3.3.3).

Sur chaque image, les annotations automatiques générées par les systèmes de détection évalués doivent :

- reconnaître les classes de plantes présentes (adventices et/ou cultures) ;
- localiser les adventices et/ou les cultures.

2.3.2. *Tâche de l'action de désherbage*

L'objectif de la tâche de désherbage est de détruire les adventices tout en préservant les plantes d'intérêt. L'évaluation de cette tâche est effectuée en comparant l'état de la parcelle (cf. Section 3.2.4) avant et après le passage de la solution. Afin de rendre cette tâche aussi indépendante que possible de la tâche de détection (voir section 2.3.1), les adventices et les plantes d'intérêt sont identifiées par des marqueurs physiques de différentes couleurs, faciles à détecter et positionnés au pied des plantes (voir section 3.2.4).

2.3.3. *Tâche de l'action globale*

Contrairement à la tâche concernant l'évaluation du système d'action de désherbage (voir section 2.3.2), les plantes ne sont pas pré-localisées par des marqueurs facilement identifiables lors de la tâche d'évaluation globale. Ainsi, cette tâche permet d'évaluer l'ensemble de la chaîne détection-décision-action : le système de détection et le système de désherbage, mais aussi toutes les décisions prises lors de l'intervention. Les critères de l'évaluation globale prennent en compte les ratios d'adventices détruites et de cultures endommagées mais aussi le débit de chantier ou le niveau d'automatisation de la solution.

2.4. Participants au Challenge ROSE

Quatre consortiums sélectionnés par l'ANR participent au Challenge ROSE :

- BIPBIP : `http://challenge-rose.fr/projet/bipbip/` ;
- PEAD : `http://challenge-rose.fr/projet/pead/` ;
- ROSEAU : `http://challenge-rose.fr/projet/roseau/` ;
- Weedelec : `http://challenge-rose.fr/projet/weedelec/`.

3. Développement et validation des environnements de test

Les compétitions robotiques sont une opportunité unique pour développer des outils et des environnements de test réellement pertinents vis-à-vis des technologies robotiques de pointe dans un domaine particulier. Les premiers résultats obtenus dans le cadre du Challenge ROSE sont présentés ci-dessous.

3.1. Phases de conception, validation et exploitation

Le Challenge ROSE est composé de plusieurs campagnes d'évaluation successives. Ces campagnes permettent tout d'abord le développement des environnements de test, et ensuite permettent aux robots participants d'améliorer leurs performances grâce à la fiabilité des mesures de performances qui sont reproductibles. La figure 3.1 présente le calendrier de ces différentes phases, qui sont également détaillées dans les paragraphes suivants.

3.1.1. *Conception d'environnements de test pour les compétitions*

Durant les six premiers mois du Challenge, les équipes participantes et les organisateurs ont travaillé ensemble pour établir les meilleures conditions possibles pour les campagnes d'évaluation. Les réunions entre les organisateurs et les participants ont permis de préciser un certain nombre d'éléments essentiels au bon déroulement du Challenge ROSE, mais qui sont communs à tous les challenges :

- concevoir les environnements de test (parcelles, cultures et adventices à mettre en place, couleur et taille des marqueurs utilisés lors de la tâche d'action de désherbage à définir, etc.) ;
- spécifier les tâches sur lesquelles les systèmes seront évalués ;
- définir les métriques de mesure des performances des systèmes qui soient quantitatives, rigoureuses, comparables, répétables et acceptées par tous ;
- spécifier les formats des données en entrée et en sortie des systèmes.

Même s'il est largement défini pendant les campagnes de *dry-run*, le plan d'évaluation continue d'être adapté et d'évoluer pendant toute la durée du challenge. Ces évolutions permettent d'accompagner l'évolution des solutions technologiques proposées par les participants et ainsi de suivre la montée en TRL des solutions.

3.1.2. *Validation des environnements de test pendant les compétitions*

Les deux campagnes d'évaluation *dry-run* en 2018 et 2019 ont permis de valider le protocole d'évaluation et l'environnement de test. Leur but est de tester et corriger les outils de comparaison et le protocole d'évaluation mis en place par les organisateurs du challenge.

3.1.3. *Exploitation des environnements de test pendant les compétitions*

Les trois campagnes d'évaluation officielles qui suivent la phase de *dry-run* seront utilisées pour évaluer les performances des solutions proposées. Les résultats des tests sont annoncés après chaque campagne d'évaluation, lors d'une réunion de travail rassemblant l'ensemble des participants.

Les paragraphes suivants décrivent les environnements de test utilisés pour réaliser une mesure de performance quantitative, rigoureuse, comparable, répétable et acceptable.

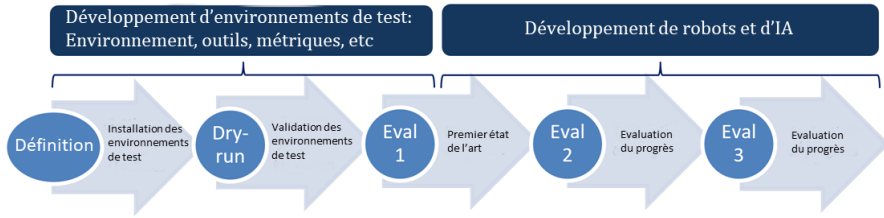

FIGURE 3.1. Définition, validation et exploitation des environnements de test durant la compétition.

3.2. ENVIRONNEMENT DE TEST PHYSIQUE

3.2.1. *Champ expérimental*

Les évaluations sont réalisées sur une parcelle expérimentale de quatre hectares du site de l'INRAE à Montoldre. Pour le déroulement du challenge, cette parcelle est organisée en zones d'expérimentations protégées, surveillées, entretenues et alimentées en électricité.

- une parcelle de référence conduite de manière conventionnelle (avec utilisation de produits chimiques) ;
- une parcelle pour chaque participant comprenant différentes zones pour l'évaluation et une zone de test et validation des paramétrages de la solution (y compris l'adaptation aux conditions climatiques) juste avant l'évaluation ;
- une parcelle commune à tous les participants afin d'acquérir des images dans le cadre de la tâche de détection 2.3.1.

À l'extrémité des parcelles, des zones de fourrière d'une dizaine de mètres de large permettent aux solutions technologiques de se déplacer et de faire des demi-tours et/ou de changer de rang de culture. De même, entre chacune des parcelles attribuées aux participants, des zones libres de six mètres de large sont disponibles. Dans le sens longitudinal de la parcelle, des zones libres de huit mètres de large séparent les zones allouées à chaque consortium pour faciliter l'expérimentation. En préparation de chaque campagne d'évaluation, une succession d'interventions techniques est réalisée sur la parcelle depuis la fin de la période hivernale jusqu'à l'ensemencement des cultures/adventices prévues deux à trois semaines avant la période d'évaluation :

- destruction des cultures/adventices de la campagne d'évaluation précédente par fauchage et exportation ;
- travail superficiel du sol pour la destruction mécanique de la croissance des adventices (plusieurs passages en fonction de la densité de couverture des adventices et des conditions météorologiques) ;
- travaux de préparation du sol pour l'ameublissement et le réchauffement du sol ;
- préparation du lit de semences avec traitement thermique (trois profondeurs et vitesses différentes) ;
- semis des cultures et des adventices aux densités souhaitées (semis effectué par un sous-traitant) ;
- l'entretien et la délimitation des parcelles attribuées aux participants (délimitation des zones intra-rangs par binage des zones inter-rangs et entretien des bords des bandes semées).

3.2.2. *Parcelle de référence*

Une parcelle de référence traitée de manière conventionnelle (deux mètres de large et 46,5 mètres de long) est mise en place pour l'évaluation des systèmes participant au challenge. Le désherbage de la parcelle de référence est effectué par une intervention chimique de pré-levée ou de post-levée. Le travail du sol, la préparation du lit de semence sans traitement thermique et le semis des cultures sur la parcelle de référence sont cependant similaires aux parcelles attribuées aux participants.

3.2.3. *Parcelles des participants*

Pour chaque type de culture retenu (grandes cultures et cultures maraîchères), une parcelle associée à un type d'adventice spécifique est mise à la disposition de chaque participant. Sur ces parcelles, une zone intra-rang d'une largeur de 10 cm centrée sur la ligne de semis de la culture principale est pourvue d'adventices. La zone de culture inter-rang est désherbée par les organisateurs.

La parcelle attribuée de façon aléatoire à chaque consortium participant est divisée en trois zones correspondant à :

- une zone de 10 mètres pour les réglages du robot avant l'évaluation ;

- une zone de 10 mètres pour l'évaluation de la tâche d'action de désherbage avec le positionnement de marqueurs sur les cultures et les adventices (marqueurs répartis sur les deux rangs de maïs ou les trois rangs de haricot) ;
- une zone de 23 mètres pour l'évaluation globale ;
- les 3,5 mètres restants permettent de disposer de zones tampons entre les trois zones précédemment définies.

En outre, pour la tâche de détection, deux rangées pour les cultures à grand espacement et trois rangées pour les cultures maraîchères avec chaque type spécifique d'adventice choisi sont partagées par les participants.

3.2.4. *Évaluation des performances*

Prise en compte des biais d'opérateur. — Afin que les évaluations soient reproductibles, les opérateurs chargés de la mesure de l'ensemble des métriques sont formés à la façon d'effectuer les mesures (par exemple, taille minimale des adventices à compter).

La reproductibilité des comptages de plantes est assurée par un double comptage indépendant par deux opérateurs différents. Si le nombre de cultures ou d'adventices obtenu par les deux opérateurs diverge de plus de 10 %, un troisième opérateur intervient pour déterminer le comptage correct.

Tâche d'action de désherbage. — Pour la tâche d'action de désherbage, les métriques utilisées sont :

- le nombre d'adventices avant et après l'action de désherbage pour obtenir le pourcentage d'adventices détruites ;
- le nombre de cultures intègres avant et après l'action de désherbage pour obtenir le pourcentage de cultures préservées,

Des photos sont également prises afin de comparer l'évolution des plantes (avant l'action de désherbage et après l'action de désherbage sur une période pouvant aller jusqu'à une semaine) et de garder une trace des évaluations.

Le tableau 3.1 présente les marqueurs à code couleur utilisés pour cette tâche. Ces disques en plastique sont placés au pied de chaque plante (voir Figure 3.2). Il est à noter qu'à la demande des participants, les marqueurs de couleur jaune peuvent être remplacés par des marqueurs de couleur rouge.

Table 3.1. Caractéristiques des marqueurs suivant le type de plantes

Plante	Couleur (RGB)	Ext. Ø	Int. Ø	Largeur de l'ouverture
Adventice	Jaune (FFFF00)	2 cm	5 mm	3 mm
Culture	Bleu (0000FF)	2,5 cm	8 mm	6 mm

Afin d'assurer une équité maximale entre les participants, les organisateurs reproduisent une difficulté similaire sur chaque parcelle en définissant des règles sur le placement des marqueurs (nombre total de marqueurs, nombre de marqueurs d'adventices entre deux cultures, distance entre les marqueurs, etc.). Plusieurs configurations ont été sélectionnées et sont répétées dans la zone d'intra-rang pour chacune des parcelles des participants :

- un seul marqueur d'adventice entre deux cultures successives marquées. Pour garantir l'équité en termes de difficulté, les marqueurs sont éloignés des cultures (plus de 3cm) deux fois sur trois (une fois sur trois, le marqueur est placé à moins de 3 cm de la culture) ;
- deux marqueurs d'adventices entre deux cultures successives marquées, les marqueurs étant éloignés de plus de 3 cm des cultures ;
- trois marqueurs d'adventices entre deux cultures successives marquées, un marqueur près des cultures et deux marqueurs à plus de 3 cm des cultures.

Chaque ensemble de marqueurs est photographié et les photos sont conservées pour une éventuelle confirmation ultérieure des résultats après l'action de désherbage.

FIGURE 3.2. Tâche d'action de désherbage : les marqueurs colorés indiquent les adventices à éliminer (jaune) et les cultures à préserver (bleu). Les combinaisons présentées ici sont, à gauche, trois marqueurs entre deux adventices successives marquées, et à droite, deux marqueurs entre deux adventices successives marquées.

TÂCHE D'ÉVALUATION GLOBALE. — Des comptages de cultures et d'adventices sont réalisés à différents moments (avant le dernier essai de la solution sur la parcelle, après le dernier essai de la solution sur la parcelle, à J+3 avec une tolérance de +/− 1 jour en fonction des contraintes). Une photo est prise chaque semaine pour enregistrer l'évolution des parcelles pendant un minimum de deux semaines (durée adaptée en fonction des conditions météorologiques). Dans le cadre de la campagne de 2020, les contraintes sanitaires ne permettant pas de réaliser l'évaluation durant la période de mai/juin, la campagne a été décalée à octobre. Les conditions climatiques différentes pour cette campagne d'évaluation rendent plus difficile l'interprétation des résultats. En plus des modules de détection et d'action du robot, cette évaluation globale tient compte de l'itinéraire technique :

- le choix du moment de l'intervention : il dépend de l'état de développement des adventices et des cultures. En effet, si les adventices ne sont pas suffisamment développées, certaines solutions ne sauront pas les détecter. Au contraire, si les adventices sont trop développées, certaines solutions de désherbage peuvent perdre en efficacité.
- le choix de poursuivre ou non l'action de désherbage sur une mauvaise herbe lorsqu'il y a un risque d'endommager les cultures voisines.
- le nombre d'interventions effectuées. Toutefois, la possibilité d'effectuer plusieurs interventions successives ou espacées de quelques jours n'était pas mise ne place lors des premières campagnes d'évaluation.

La campagne d'octobre 2020 a permis de surcroît de tester de nouvelles méthodologies pour l'évaluation de la performance à savoir le débit de chantier (chronométrage du temps d'intervention mis au regard de la surface traitée), la consommation énergétique (par installation de pince ampérométrique qui mesure en temps réel la consommation) ou le niveau d'automatisation (via une grille d'évaluation remplie par l'équipe organisatrice sur des observations visuelles lors des évaluations). Ces méthodologies seront perfectionnées lors de la dernière campagne d'évaluation.

3.3. Environnement de test virtuel

L'environnement de test virtuel permet de mener des campagnes d'évaluation des capacités du système robotisé présenté par chaque participant pour la détection, discrimination et localisation des plantes d'intérêt et des adventices, à partir des images acquises par des participants au Challenge sur le terrain.

3.3.1. *Images collectées*

Les robots participant au Challenge disposent de caméras pouvant être dans le spectre visible, multi-spectral ou hyper-spectral. Les images collectées par ces caméras peuvent contenir des plantes d'intérêt et/ou des adventices, ou même ne contenir aucune plante. Chaque équipe réalise une campagne d'acquisition d'images pour chaque combinaison de plantes d'intérêt (maïs, haricot) et d'adventices (matricaire, chenopode, moutarde et raygrass).

Un nombre minimal de 5 images par mètre linéaire parcouru sur la parcelle expérimentale est exigé. Les acquisitions de ces images sont réalisées sur la même parcelle expérimentale par tous les participants, et dans un laps de temps très court afin de minimiser les différences de conditions météorologique. L'ordre de passage des équipes pour l'acquisition des images est aléatoire.

3.3.2. *Annotation automatique*

Annotation d'hypothèses. — Les systèmes doivent produire leurs propres hypothèses lors de l'acquisition des images de la parcelle (voir section 2.3.1). Après chaque essai sur la parcelle, chaque consortium soumet les fichiers contenant les

images brutes et les fichiers contenant les annotations d'hypothèses. Pour chaque image collectée, le système évalué doit renvoyer un fichier contenant les annotations d'hypothèses ainsi que *N* images bitmap, *N* étant le nombre de plantes détectées dans l'image. Chaque image bitmap définit l'emplacement d'une plante spécifique et correspond classiquement à un masque de détection. Un pixel d'intensité 0 indique un point n'appartenant pas à la plante. Tous les pixels appartenant à la plante ont la même intensité et sont d'intensité maximale (intensité de 255).

Annotation de référence. — Les images utilisées pour l'évaluation sont annotées a posteriori par des experts humains, sous la supervision des organisateurs du Challenge. Un guide d'annotation complet est mis à la disposition des annotateurs. Il contient des détails sur la nature des annotations. Chaque image annotée par les experts contient les informations suivantes :

- le détourage des plantes (annotation manuelle) : délimitation de chaque plante visible dans l'image par une boîte polygonale (« bounding box ») qui sera aussi petite que possible tout en englobant la totalité de la plante considérée,
- le type de plante (annotation manuelle) : une étiquette est associée à chaque *bounding box* indiquant le type de la plante (adventice, culture ou, dans de rares cas, « indéterminée ») ;
- le stade de croissance de la plante (annotation manuelle) : indication du stade de croissance de chaque plante (« 0 » : émergence précoce ; « 1 » : plantule ; « 2 » : quelques feuilles ; « 3 » : avancé) ;
- le nom commun (annotation manuelle) : si le stade de croissance de la plante le permet, indication du nom commun de chaque plante (nom de la liste ou « indéterminable »).

Les images comportant toutes ces annotations manuelles constituent les données dites « de référence ». Lors de l'évaluation, ces références sont comparées avec les annotations automatiques produites par les systèmes de détection évalués. Pour s'assurer de la qualité des annotations, 10 % des images de chaque combinaison participant/type de plante (images sélectionnées au hasard) sont annotées par deux annotateurs différents. Une comparaison entre les annotateurs est effectuée sur cet échantillon.

Logiciel d'annotation LNE-DIANNE. — Les images de test sont annotées par des experts humains à l'aide de l'outil d'annotation LNE-DIANNE (Détourage, Identification et ANNotation pour l'Evaluation), qui pré-découpe automatiquement les plantes visibles sur l'image pour optimiser le temps d'annotation. L'interface principale du logiciel est présentée par la figure 3.3. Cette pré-annotation est basée sur deux méthodes de classification : la méthode des k-moyennes ou le seuillage par teinte. L'utilisateur a la possibilité de modifier la configuration de ces algorithmes. Cette pré-annotation est corrigée manuellement par l'annotateur.

Le logiciel LNE-DIANNE est mis à disposition des participants pour leurs travaux de recherche, et sera disponible publiquement à l'issue du Challenge ROSE.

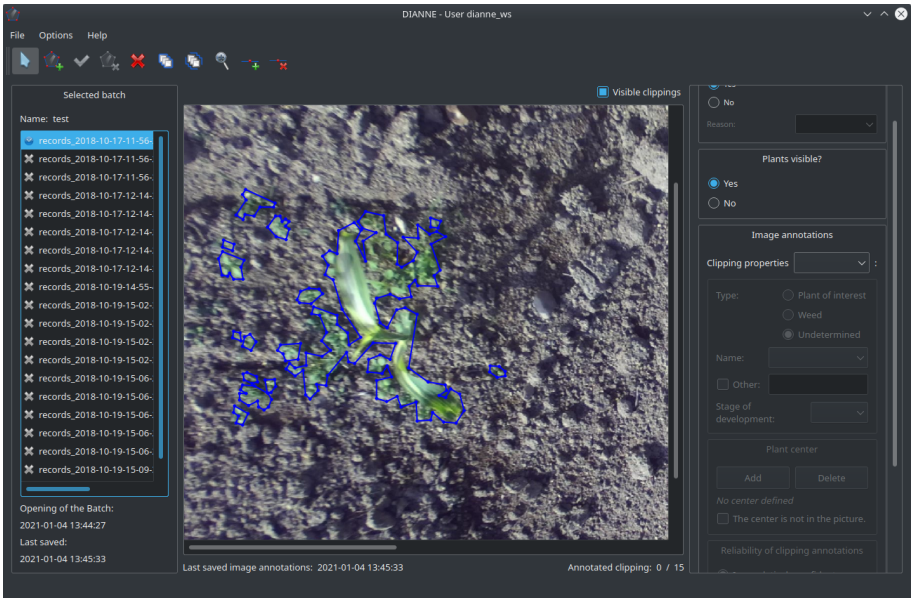

FIGURE 3.3. Capture d'écran de l'interface principale de LNE-DIANNE

3.3.3. *évaluation de performance*

PHASE DE MAPPING. — Les masques définis par les systèmes ainsi que les *bounding box* définies avec le logiciel LNE-DIANNE par les annotateurs humains pour indiquer l'emplacement des plantes (adventices et cultures) sont utilisés pour évaluer la performance de la détection. L'évaluation commence par une première phase de mapping, qui vise à associer une à une les zones de détection définies par les systèmes en utilisant les masques (hypothèses) avec celles annotées manuellement comme boîtes de délimitation (références). Le processus de mapping sélectionné pour le challenge est celui qui minimise la somme des nombres de pixels situés à l'extérieur de l'intersection des zones associées une par une. Notez que deux zones ne peuvent pas être associées si elles n'ont pas de pixel en commun. Une fois le meilleur mapping identifié, certains masques d'hypothèse peuvent ne pas être associés à une *bounding box* de référence, soit parce qu'ils n'ont aucun pixel en commun avec une *bounding box* de référence, soit parce que chaque *bounding box* de référence est déjà associée à un autre masque d'hypothèse. Ces masques seront appelés *faux positifs*. De même, certaines *bounding boxes* peuvent ne pas être associées à un masque, soit parce qu'elles n'ont pas de pixels en commun, soit parce que tous les masques de l'hypothèse sont déjà associés à d'autres boîtes de référence. Ces *bounding boxes* seront appelées *faux négatifs*. La figure 3.4 montre des exemples d'hypothèses (masques soumis par un participant), qui sont comparées à l'image de référence annotée.

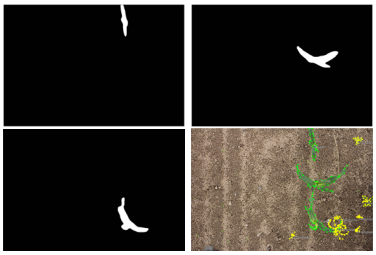

FIGURE 3.4. Tâche de détection : les hypothèses d'annotation des cultures (haut gauche et droite, bas gauche) sont comparées à la référence (bas droite).

MÉTRIQUE DE PERFORMANCE. — La métrique d'évaluation est l'*Estimated Global Error Rate* (EGER) [6]. Pour chaque image ayant reçu une annotation de référence, les listes des plantes détectées respectivement par le système et par les annotateurs sont générées automatiquement. Ces deux listes sont comparées sur la base de l'association des zones de détection définies par le mapping. Une association entre deux plantes classées de la même façon par le système et par les annotateurs est considérée comme correcte. Une association entre deux plantes de classes différentes est considérée comme une confusion. Chaque plante de l'hypothèse non-associée est considérée comme un faux positif, et chaque plante de la référence non associée compte comme un faux négatif. Une pénalité est associée à chaque confusion, à chaque faux négatif et à chaque faux positif. La somme des comptes d'erreurs par image donne le nombre total d'erreurs. Le nombre total d'entrées attendues est également compté en cumulant le nombre de plantes présentes dans la référence de chaque image. Le taux d'erreur est alors le nombre global d'erreurs divisé par le nombre total d'entrées attendues. Les pénalités appliquées sont les suivantes :

- pénalité de 1 pour les oublis/faux positifs ;
- pénalité de 2 pour la confusion.

La métrique EGER utilisée est la suivante :

$$\text{EGER} = \frac{\sum_{k=1}^{N} C_k + FA_k + O_k}{\sum_{k=1}^{N} NR_k}.$$

Où C_k, FA_k et O_k représentent respectivement la somme des pénalités pour confusion, faux positif et faux négatif dans l'image k. NR_k représente le nombre de plantes détectées dans la référence (adventices et cultures). Les scores de précision (proportion de vrais positifs parmi l'ensemble des positifs), de rappel (proportion de vrais positifs parmi l'ensemble des positifs de référence) et de F-mesure (moyenne harmonique de la précision et du rappel) sont également fournis. Les résultats de cette évaluation sont présentés par type de plantes (adventices ou cultures) et de manière globale en tenant compte des deux classes. Le processus d'évaluation de la détection automatique est présenté dans la figure 3.5.

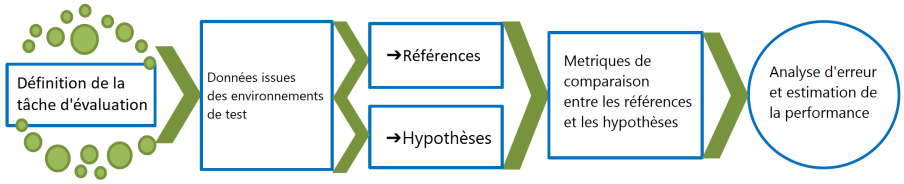

FIGURE 3.5. Processus d'évaluation de la détection automatique.

SUITE LOGICIELLE LNE-MATICS. — LNE-MATICS [7] est une suite logicielle gratuite et open source conçue pour l'exploration de données et l'évaluation de systèmes. LNE-MATICS a été conçu à l'origine pour l'évaluation de systèmes de traitement automatique du langage [8, 9, 18]. Il a été adapté pour répondre aux besoins d'évaluation des systèmes de détection et est utilisé dans le cadre du Challenge ROSE notamment en intégrant le système de mapping décrit précédemment ainsi que la métrique EGER.

4. FACTEURS D'INFLUENCE

Comme les campagnes d'évaluation se déroulent dans un environnement ouvert et sur des entités vivantes, de nombreux facteurs environnementaux peuvent influencer les performances des solutions mises en œuvre. Certains de ces facteurs d'influence sont contrôlables par les organisateurs des campagnes. D'autres sont difficilement contrôlables et doivent au moins être mesurés afin de pouvoir en tenir compte lors de l'analyse des résultats de l'évaluation. Les participants doivent néanmoins s'assurer que leurs robots sont suffisamment robustes par rapport aux facteurs non contrôlables, afin qu'ils puissent être utilisés dans les conditions réelles par les professionnels du secteur agricole. On peut distinguer deux grands types de facteurs d'influence : les modalités d'essai et les facteurs agropédoclimatiques.

4.1. MODALITÉS DE TEST

Les modalités de test regroupent les paramètres liés à la parcelle expérimentale. Tout d'abord, l'itinéraire technique de la parcelle peut être contrôlé car même s'il est déterminé par chaque consortium participant, il est connu avant l'intervention sur la parcelle. La densité et la répartition des cultures et des adventices sont des facteurs contrôlables. L'état réel de la parcelle expérimentale est communiqué quotidiennement aux consortiums par l'enregistrement d'images. Le stade de développement des cultures et des adventices est un facteur mesurable. Les organisateurs communiquent quotidiennement le niveau de développement des plantes, notamment par la prise d'images sur la parcelle.

4.2. Facteurs agropédoclimatiques

Les conditions agropédoclimatiques comprennent les conditions météorologiques et la luminosité, ainsi que les caractéristiques pédologiques et agronomiques du sol. Les organisateurs surveillent les conditions météorologiques à l'aide de stations dédiées, et les autres conditions à l'aide d'autres types de capteurs tels que des capteurs d'humidité et de température du sol. Les conditions de lumière, qui entraînent des problèmes d'ombre et d'éblouissement perturbant les systèmes de détection, seront mesurées à l'aide de luxmètres. Les caractéristiques du sol telles que l'humidité et la température sont mesurées quotidiennement. Les caractéristiques de la texture du sol (teneur en argile, etc.) de la parcelle ont été déterminées au début du Challenge. Les participants sont libres de choisir la date d'intervention de leur solution sur la parcelle en fonction de ces mesures. Par ailleurs, les données de géoréférencement acquises par GPS RTK lors du déplacement du semoir sur la parcelle pendant le semis sont également partagées avec les participants afin de faciliter la localisation des lignes de semis et de chaque zone intra-rang.

L'ensemble de ces facteurs d'influence sont présentés Tableau 4.1 (Test mod. fait référence aux facteurs d'influence liés aux modalités de test, et Acropedoc. aux facteurs agropédoclimatiques).

Table 4.1. Caractéristiques des marqueurs correspondant au type de plante

Type	Facteurs	Mesures effectuées
Test mod.	Itinéraire technique	Décrit avant l'évaluation
Test mod.	Densité et distribution des plantes	Images avant évaluation
Test mod.	État de croissance des plantes	Images quotidiennes
Agropedoc.	Conditions météo. (temp., humidité, vent)	Quotidiennes
Agropedoc.	Luminosité et radiation solaire	Quotidiennes
Agropedoc.	Humidité et temp. du sol	Quotidiennes
Agropedoc.	Humidité des feuilles, évapotranspiration	Quotidiennes
Agropedoc.	Taux d'argile	Avant la première évaluation

5. Critères d'évaluation du niveau supérieur

La prise en compte de certains autres critères d'évaluation était prématurée compte tenu du développement actuel des technologies participant au Challenge ROSE. Ces critères notamment les facteurs d'acceptabilité ou les facteurs sociaux-économiques seront abordée lors de l'évaluation de septembre 2021.

6. Mise en œuvre dans le cadre du projet METRICS

Le projet METRICS (`metricsproject.eu`) est un projet européen coordonné par le LNE et dont l'objectif est l'organisation de compétitions robotiques à visée métrologique, dans les quatre domaines prioritaires définis par la Commission Européenne :

- HEART-MET dans le domaine de la santé ;
- ACRE dans le domaine de l'agriculture et l'alimentation ;
- RAMI dans le domaine de l'inspection et la maintenance des infrastructures ;
- ADAPT dans le domaine de l'usine du futur.

La compétition ACRE (Agri-food Competition for Robot Evaluation) est organisée dans la continuité du Challenge ROSE, avec des partenaires supplémentaires et un plan d'évaluation comportant des éléments du plan d'évaluation de ROSE et complété par de nouvelles évaluations et une partie « évaluation cascade » c'est à dire à partir des images acquises lors des évaluations terrain. Le lien et la différence entre les deux organisations sont explicités dans la section 6.4.

L'ensemble des compétitions organisées dans le cadre de METRICS suivent une méthodologie unifiée avec un plan d'évaluation formalisé et validé par une relecture par les pairs. L'organisation suit plusieurs principes fondamentaux :

6.1. Compétitions sur le terrain et « en cascade »

Les compétitions METRICS reposent sur l'organisation de campagnes d'évaluation. Une campagne d'évaluation est une compétition de benchmarking : c'est-à-dire une compétition où les équipes participantes sont évaluées et classées en fonction des résultats de l'application de repères scientifiques à leur performance. Deux types de campagnes d'évaluation ont été mis en place dans l'ensemble des compétitions METRICS :

- les campagnes d'évaluation sur le terrain, qui se déroulent dans des environnements réels représentatifs du domaine étudié ;
- les campagnes d'évaluation en cascade, qui sont des compétitions basées sur des données auxquelles les équipes participent à distance.

Les jeux de données sur lesquels reposent les campagnes en cascade sont collectés lors des campagnes sur le terrain. Ainsi, lorsque le même *benchmark* est utilisé par les deux, il est possible de comparer directement les performances des équipes participant aux campagnes sur le terrain et aux campagnes en cascade.

La campagne d'évaluation *dry-run* basée sur les données collectées lors des évaluations sur le terrain implique des équipes participant à distance. En 2020, elle s'est appuyée sur les ensembles de données collectés lors de l'événement 2020 du Challenge ROSE. Les jeux de données ont été fournis par toutes les équipes participantes et validés par le LNE afin d'assurer leur représentativité.

6.2. Compétition « dry-run » lors de la première édition

L'objectif de cette campagne est de valider le plan d'évaluation et de produire des ensembles de données pour les prochaines campagnes.

Dans le cas de ACRE, la première campagne *dry-run* était co-localisée dans l'espace et le temps avec l'un des événements de terrain de ROSE, les deux évènements se

sont déroulés en Octobre 2020 à Montoldre. Cette cohabitation (ou association) était destinée à initier une implication des équipes du challenge ROSE pour amorcer un début à METRICS et à permettre de garantir la disponibilité des ensembles de données pour les campagnes en cascade à partir des images collectées par les robots des équipes du challenge ROSE.

6.3. Évaluation modulaire : TBM et FBM

METRICS intègre le cadre méthodologique « Benchmarking through Competitions » conçu par le projet européen RoCKIn [17] et développé dans les projets européens RockEU2 et SciRoc. Il s'agit du même cadre que les compétitions de robots de la Ligue européenne de robotique. En quelques mots, il repose sur la définition de deux types de *benchmarks* :

- les Functionality Benchmarks (FBM), axés sur les capacités spécifiques d'un robot et conçus pour rendre le *benchmark* aussi indépendant que possible des autres caractéristiques du robot qui ne sont pas directement impliquées dans la fonctionnalité examinée ;
- les Task Benchmarks (TBM), permettant l'évaluation de l'exécution de tâches complexes impliquant de multiples fonctionnalités, où le résultat final dépend à la fois de celles-ci individuellement et des caractéristiques du robot au niveau du système, telles que l'intégration entre les fonctionnalités.

6.4. Transition entre ROSE et ACRE

6.4.1. *Nouveaux partenaires et nouvel environnement expérimental*

Comme pour l'ensemble des compétitions METRICS, la compétition ACRE est organisée à l'international, par les deux partenaires français du Challenge ROSE et des partenaires italiens :

- Politecnico di Milano (POLIMI) ;
- Università degli Studi di Milano (UNIMI).

La présence de ces nouveaux partenaires permet de proposer un second environnement expérimental, à Cornaredo (Italie). Ceci permettra de valider les protocoles expérimentaux pour un terrain différent (structure et composition du sol différentes). Des adventices et des cultures supplémentaires seront également ajoutées.

6.4.2. *Fonctionnalité de navigation autonome*

Ce paragraphe présente un des *FBM* du Challenge ACRE : la fonctionnalité « Field navigation » est ajoutée dans le cadre du Challenge ACRE. La fonctionnalité vise à décrire la navigation autonome dans le champ. En effet le challenge ROSE ne s'intéressant pas à cette problématique de suivi de ligne. Pour tester la capacité du système à naviguer de façon autonome, une expérimentation a été envisagée sur une

parcelle spécifique. Le système doit effectuer une trajectoire complexe, présentée sur la figure 6.1.

(1) Parcourir une première parcelle droite.
(2) Effectuer un demi-tour au bout de la parcelle.
(3) Parcourir la parcelle droite voisine de la première.
(4) Effectuer un demi-tour au bout de la parcelle.
(5) Parcourir la parcelle suivante, de forme plus complexe (décrochage de 37,5 cm au milieu).

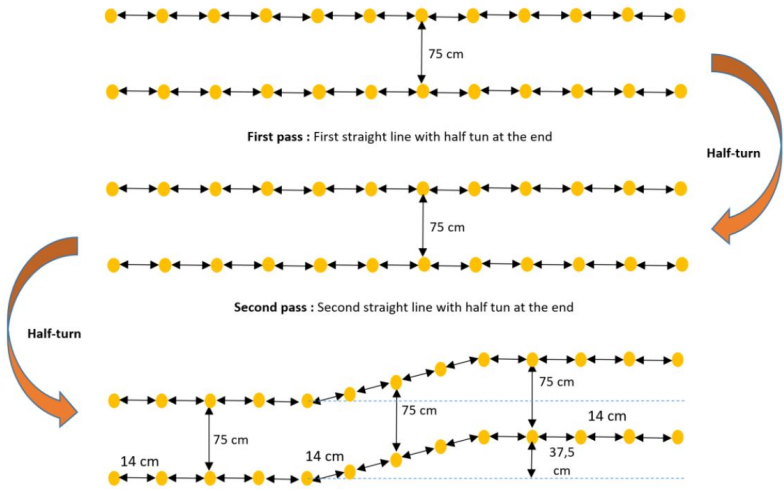

FIGURE 6.1. Expérimentation de navigation autonome.

7. CONCLUSION

Le Challenge ROSE est la première initiative mondiale à mettre en compétition différents robots en incluant à la fois une évaluation par l'image et une évaluation sur le terrain, sur des parcelles agricoles. En effet, ce Challenge permet de réaliser une évaluation modulaire des différentes briques technologiques des solutions participant au Challenge, ainsi qu'une évaluation globale de l'efficacité du désherbage. Les résultats de ce challenge ainsi que les outils développés seront rendus publics après la dernière campagne d'évaluation de 2021.

Les moyens d'essais développés dans le cadre de ce challenge constitueront des références consensuelles utiles pour la caractérisation des futurs projets de recherche et industriels dans ce domaine, qui pourront être diffusées en vue de la normalisation. En particulier, les bases de données de tests qualifiées et annotées ont, par la richesse de leur contenu, un fort potentiel de diffusion. La création d'un corpus de référence

d'images dans le visible, multi-spectrales et hyper-spectrales alignées est en effet une nouveauté qui permettra des évaluations comparatives de différentes technologies de détection des mauvaises herbes et des cultures. Ces bases de données validées seront particulièrement utiles à la communauté car, dans le cadre de la limitation de l'utilisation des produits phytosanitaires, de nombreuses machines robotisées innovantes auront besoin d'intégrer des dispositifs de détection automatique des adventices. Ces systèmes sont basés sur des algorithmes apprenant à partir d'images annotées. De nombreuses bases de données d'images d'adventices et de cultures dans le spectre visible existent, comme la base gratuite Plntnet, mais pour l'instant aucune base de données d'images hyper-spectrales ouverte n'est encore disponible, alors que cette technologie est prometteuse pour l'agriculture numérique. Les bases de données du challenge ROSE seront donc en libre accès sur le site de la compétition courant 2022 pour combler ce manque.

L'intégration de technologies encore peu utilisées dans les systèmes et outils agricoles, comme les caméras infrarouges ou hyper-spectrales et leur utilisation dans des systèmes de détection multimodale, des outils de mapping dynamique, des plateformes automatisées combinées à des stratégies de traitement de précision, permettra d'établir une avancée majeure dans le processus visant à fournir aux agriculteurs des solutions multiples aux problèmes de traitement des mauvaises herbes sur les rangs de culture. Les recherches menées seront également utiles pour d'autres applications que celles concernées par le ROSE Challenge. On peut en effet imaginer des développements futurs pour d'autres fonctionnalités et tâches qui pourront être réalisées par ces nouveaux outils au service de tous les professionnels de l'agriculture. Un de ces développements, la compétition ACRE de METRICS, est déjà lancé dans la continuité du Challenge ROSE. La première campagne d'évaluation ACRE a permis de confirmer la possibilité de transférer la méthodologie de ROSE à un cadre plus général.

Remerciements

Les auteurs tiennent à remercier G. Bernard, A. Delaborde, O. Galibert, B. Lalere, S. Lecadre, M. Veron et M. Kalouguine pour l'adaptation des logiciels d'évaluation LNE-MATICS et d'annotation LNE-DIANNE aux besoins du Challenge ROSE et pour leur contribution à la définition des protocoles d'évaluation. Nous tenons également à remercier toutes les personnes qui ont contribué à la mise en place des moyens de test du ROSE Challenge sur le site expérimental de l'INRAE. Le Challenge ROSE est réalisé avec le soutien financier du Ministère de l'Agriculture et de l'Alimentation, du Ministère de la Transition Écologique et Solidaire, du Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation, et de l'Agence Nationale de la Recherche. Nos remerciements vont aussi aux partenaires du projet METRICS et plus particulièrement ceux impliqués dans l'organisation de la compétition ACRE. Le projet METRICS est financé par le programme de recherche et d'innovation Horizon 2020 de l'Union européenne – subvention n° 871252. Le contenu de ce document relève de la seule responsabilité des auteurs et ne peut en aucun cas être considéré comme reflétant la position de l'Union européenne.

Bibliographie

[1] M. Anderson, O. Jenkins & S. Osentoski, « Recasting robotics challenges as experiments », *IEEE Robotics and Automation Magazine* **18** (2011), nº 2, p. 10-11.

[2] « Standard Test Method for Evaluating Response Robot Sensing : Visual Acuity », ASTM International, 2017, E2566-17a.

[3] G. Avrin, D. Boffety, S. Lardy-Fontan, R. Régnier, R. Rescoussié & V. Barbosa, « Design and validation of testing facilities for weeding robots as part of ROSE Challenge », in *Evaluating Progress in IA (EPAI)*, 2020.

[4] G. Avrin, A. Delaborde, O. Galibert & D. Boffety, « Boosting agricultural scientific research and innovation », in *3rd RDV Techniques AXEMA February 23, 2019, SIMA, France*, 2019.

[5] S. Behnke, « Robot competitions-ideal benchmarks for robotics research », in *Proc. of IROS-2006 Workshop on Benchmarks in Robotics Research*, IEEE, 2006. October.

[6] F. Bonsignorio, A. Del Pobil & E. Messina, « Fostering progress in performance evaluation and benchmarking of robotic and automation systems », *IEEE Robotics and Automation Magazine* **21** (2014), nº 1, p. 22-25.

[7] O. Galibert, G. Bernard, A. Delaborde, S. Lecadre & J. Kahn, « Matics Software Suite : New Tools for Evaluation and Data Exploration », in *proc. 11th edition of the Language Resources and Evaluation Conference* (Miyazaki), Japan, 2018, p. 7-12.

[8] O. Galibert & J. Kahn, « The first official repere evaluation », in *First Workshop on Speech, Language and Audio in Multimedia*, 2013.

[9] O. Galibert, S. Rosset, C. Grouin, P. Zweigenbaum & L. Quintard, « Extended named entities annotation in ocred documents : From corpus constitution to evaluation campaign », in *LREC*, 2012.

[10] R. Gerrish, « Ready for the agBOT Challenge », *Resource Magazine* **26** (2019), nº 3, p. 8-9.

[11] A. Jacoff, H. Huang, A. Virts, A. Downs & R. Sheh, « Emergency Response Robot Evaluation Exercise », in *Proceedings of the Workshop on Performance Metrics for Intelligent Systems*, 2012, p. 145-154.

[12] A. Jacoff, E. Messina, H. Huang, A. Virts, A. Downs & R. Norcross, *Standard test methods for response robots*, ASTM International Committee on Homeland Security Applications, 2010.

[13] A. Jacoff, R. Sheh, A. Virts, T. Kimura, J. Pellenz, S. Schwertfeger & 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*, 2012. March, p. 182-189.

[14] N. Jan, R. Cattoni, S. Sebastian, M. Negri, M. Turchi, S. Elizabeth, S. Ramon, B. Loic, S. Lucia & M. Federico, « The IWSLT 2019 evaluation campaign », in *16th International Workshop on Spoken Language Translation 2019*, 2019.

[15] J. Kahn, O. Galibert, L. Quintard, M. Carré, A. Giraudel & P. Joly, « A presentation of the REPERE challenge », in *10th International Workshop on Content-Based Multimedia Indexing (CBMI)*, IEEE, 2012, p. 1-6.

[16] H. Kitano, M. Asada, Y. Kuniyoshi, I. Noda & E. Osawa, « Robocup : The robot world cup initiative », in *Proceedings of the first international conference on Autonomous agents*, ACM, 1997, p. 340-347.

[17] P. Lima, D. Nardi, G. Kraetzschmar, R. Bischoff & M. Matteucci, « Rockin and the european robotics league : building on robocup best practices to promote robot competitions in europe », in *Robot World Cup*, Springer, Cham, 2016, p. 181-192.

[18] I. Oparin, J. Kahn & O. Galibert, « First maurdor 2013 evaluation campaign in scanned document image processing », in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2014, p. 5090-5094.

[19] G. Pratt & J. Manzo, « The DARPA robotics challenge », *IEEE Robotics and Automation Magazine* **20** (2013), nº 2, p. 10-12.

[20] L. Quintard, O. Galibert, G. Adda, B. Grau, D. Laurent, V. Moriceau, S. Rosset, X. Tannier & A. Vilnat, « Question Answering on web data : the QA evaluation in Quæro », in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010.

[21] C. Ross, N. Mccarthy, D. Beatty, A. Della, M. Valentine, R. Vyenielo & C. Fink, « agBOT 2017 Challenge Autonomous Corn Seeding Tractor », Cal Poly, 2017.

[22] B. Schuller, S. Steidl, A. Batliner, P. Marschik, H. Baumeister, F. Dong, C. Einspieler et al., « The Interspeech 2018 computational paralinguistics challenge : Atypical & self-assessed affect, crying & heart beats », in *Proceedings of the INTERSPEECH International Conference* (Hyderabad, India), 2018.

[23] K. Tee & H. Van Der Kooij, « The ICRA 2017 Robot Challenges Competitions », *IEEE Robotics and Automation Magazine* **24** (2017), n° 3, p. 15-21.

Abstract. — The ROSE Challenge is the first global robotics and artificial intelligence competition to implement a third-party evaluation of the performance of robotized intra-row weed control in real and reproducible conditions, to ensure a credible and objective assessment of their effectiveness. This paper reports on the design and validation of test facilities for this competition, which presents a particular complexity: the evaluations take place in real conditions on crop plots and target living organisms (crops and weeds). Moreover, the experimental conditions need to be reproductible to allow for comparison of evaluation results and for fair treatment of different participants. The article also discusses the opportunity this challenge offers to define, in a consensual manner, the means and methods for characterizing these intelligent systems. The tools developed in the framework of this challenge establish the necessary references for future research in the field of agricultural robotics: the annotated images will be particularly useful to the community and the evaluation protocol will allow to define harmonized methodologies beyond the ROSE challenge. After presenting the objectives of the challenge, the article will present the methodology and tools developed and used to allow an objective and comparable evaluation of the performances of the systems and solutions developed. Finally, the article will illustrate this potential for harmonization and sharing of references through the European competition ACRE of the European project H2020 METRICS.

Keywords. — Artificial Intelligence, evaluation, agriculture, robotics.

Manuscrit reçu le 30 mars 2021, révisé le 15 juillet 2021, accepté le 1er septembre 2021.