Enhancing Plant Variety Discovery Process with Visual Trait Assessment in VR

Muhammad Moiz Sakha*
DFKI
Saarland Informatics Campus

Florian Daiber†
DFKI
Saarland Informatics Campus

Matthias Enders‡
NPZ Innovation GmbH

Christoph Tieben§
DFKI

Benjamin Kisliuk¶
DFKI

Antonio Krüger‖
DFKI
Saarland Informatics Campus

Abstract

Plant breeders use field trials across locations and years to identify superior plant varieties with traits such as disease resistance and higher yield. However, comparing breeding candidates across locations and years is challenging and resource demanding. To address this, we developed an integrated system that combines data acquisition through a field robot with an immersive virtual reality (VR) interface for remote assessments. The robot autonomously collects images, spectral data and 3D scans of canola breeding trials. Our VR application, developed through a user-centered approach, offers photo-realistic 3D visualizations, enabling breeders to compare candidates across locations and growth stages—capabilities unavailable in field assessments. In a user study, five breeders conducted visual trait scoring in VR to evaluate how well the system supported typical field trial tasks. The results demonstrated consistent scoring patterns among raters. Feedback from breeders indicated that the ability to compare candidates across locations and growth stages enhanced their decision making in trait assessment. This work highlights the potential of combining robotics and VR to transform data-intensive processes in agriculture.

Index Terms: VR in Agriculture, Immersive Visualization, Agricultural Robotics, Plant Breeding.

1 Introduction

Plant breeding field trials are typically conducted across multiple locations, each with varying environmental conditions, and must account for temporal (seasonal or yearly) and spatial (site specific) variations [15]. To identify potential new high-performing varieties with sufficient statistical support, it is recommended to conduct these trials at a minimum of 12 distinct locations over a two-year period [35]. The collection of data from these large-scale field trials is highly time-consuming, as trials are distributed across extensive geographic areas. Although some private breeding organizations have adopted electronic devices, such as tablets with dedicated software[1] that can record comments, photos, and GPS coordinates of breeding plots [29], many public breeding organizations still rely on manual data collection with paper and pen, followed by subsequent transcription [28, 7]. Data collection in field trials mainly involves the regular visual evaluation of variety candidates by breeders. A large number of parameters and characteristics are assessed at different stages of plant development through a process known as visual traits assessment or visual traits scoring. The current methods of data collection and visual traits assessment have the following limitations:

- collecting data is time-consuming, as field trials are replicated across a large geographic region. It increases the likelihood of errors, particularly because of the large number of plots.
- comparing the same breeding candidates planted at different locations to negate environmental effects is challenging with current methods.
- visual traits assessment in the field by plant breeders is dependent on favorable weather conditions.

To address the limitations of existing data acquisition methods, we developed an automated robotic system equipped with multiple modality sensors capable of capturing data throughout all plant growth stages. This system is supported by a mobile base station that provides connectivity and renewable energy sources, enabling the robot to remain in close proximity to the field. This setup facilitates more frequent data acquisition missions compared to manual methods [16], significantly enhancing the temporal resolution of collected data. Autonomous navigation in breeding fields is achieved through a two-step mapping process: first, a UAV drone captures aerial imagery to generate a digital layout map, which is then processed into a detailed navigation map. The robot navigates using this map in conjunction with onboard sensors for collision avoidance.

To address the challenge of comparing breeding plots of the same genotype across different locations, we developed a VR application in collaboration with plant breeders. This application allows plant breeders to perform visual trait assessments of breeding plots remotely, regardless of their geographic location. It offers photorealistic visualizations of 3D reconstructed data of breeding plots within an immersive VR environment. The application allows breeders to perform side-by-side comparisons of the breeding plots of the same genotype planted at different locations, facilitating a direct assessment of environmental impacts on plant features of a genotype. Additionally, it enables breeders to observe various developmental stages of the plants throughout the growing season, providing a more comprehensive and detailed basis for decision making in plant breeding. The decision to use VR for data visualization and interaction of breeding plots is driven by the need to offer an immersive experience that enhances the ability of breeders to accurately assess complex traits. For example, tasks such as estimating the volume of breeding plots are challenging to perform accurately on traditional 2D screens. Research has shown that the immersive nature of VR interfaces significantly improves the

*e-mail: moiz.sakha@dfki.de
†e-mail: florian.daiber@dfki.de
‡e-mail: m.enders@npz-innovation.de
§e-mail: christoph.tieben@dfki.de
¶e-mail: benjamin.kisliuk@dfki.de
‖e-mail: antonio.krueger@dfki.de

[1] https://genovix.io/

user's ability to interpret 3D data compared to similar 2D interfaces [34, 6, 1, 13, 8].

In particular, our contributions include the following.

- preliminary interviews with domain experts in plant science and breeding to identify key opportunities and requirements for multimodal data acquisition of breeding plots and data visualization in VR (see Section 3, 5);
- the design and use of an autonomous robot that continuously monitors breeding plots by regularly acquiring GPS-localized, multi-modal data from field trials (see Section 4);
- the iterative design of a VR application for plant breeders to view 3D visualization of the breeding plots and do visual trait assessment (see Section 5);
- findings from a task-oriented user evaluation focusing on real scenario of visual scoring of traits of breeding plots using VR app (see Section 6, 8);
- a report on limitations of visual trait assessment in VR (see Section 9).

2 Background

Plant breeding is a data-driven field that heavily relies on phenotypic data for decision-making [10]. These decisions ultimately lead to the selection of new plant varieties with superior traits, such as higher yield or better tolerance to diseases and abiotic stresses such as drought. To ensure sufficient statistical support for these selections, large-scale field trials spanning multiple years and locations are conducted, paired with extensive data collection by human experts. The traditional process of visual plant characterization is increasingly complemented—and in some cases replaced—by sensor-based measurements. Drones are frequently used to capture RGB, multispectral, and 3D data, as demonstrated in studies such as [20, 38, 36]. The use of raw sensor data to support breeder decision-making varies, from simple documentation to advanced artificial intelligence methods that infer plant characteristics [21]. Compared to related work in plant breeding that uses drones for data acquisition, our approach is the first to integrate an autonomous field robot for data acquisition with a VR application, which plant breeders can use for remote visual trait assessment.

2.1 Robotics in Agriculture

In recent years, many new robotic systems have been developed to perform different field operations in the area of agriculture [12]. The ability to carry heavy payloads and generate high-resolution images and sensor data is the key feature that makes field robots advantageous [24]. These field robots vary in scale: from small-scale robots that navigate between plant rows, to mid-scale robots operating across multiple rows or plots, and even to large robots capable of replacing tractors [9]. All of these systems aim to operate autonomously for a specific task in the field [25]. BoniRob, a well-discussed mid-scale autonomous field robot platform [31], uses a top-down approach for phenotyping small or young plants. In contrast, small-scale robots such as Robotanist [23] employ a side-based viewpoint for observing taller plants or crops. We will present a hybrid solution that combines top-down and side-based viewing approaches to accommodate plants at different growth stages.

2.2 VR in Agriculture

The use of VR or augmented reality solutions in agriculture is a relatively new field, with limited published articles. Franzluebbers et al. [13] proposed a VR-based application for visualization and annotation of laser scans of plants in VR. Hurst et al. [17] conducted a review of the literature related to applied augmented reality solutions within agriculture. In the context of plant breeding, the Avatars project [2] demonstrated VR's potential for exploring multidimensional genome datasets. Most directly relevant to our work, Sakha et al. [32] proposed a preliminary VR prototype, Virtual Breeding Nursery, for remote assessment of plant traits by breeders.

3 System Requirements

Our initial review of the literature and interviews with experts in plant breeding provided key insights that shaped the development of the robot for data acquisition and VR application, as follows.

3.1 Plant breeding activity

In plant breeding, field trials are conducted using predefined rectangular plots, each containing multiple plants of the same genotype. This method allows breeders to evaluate different genotypes, with the goal of identifying superior candidates. In the following, we outline the workflow and summarize the challenges breeders face, which our robot and VR prototype are designed to address.

3.1.1 Workflow

Plant breeding field trials are conducted according to the following workflow:

- **Trial setup.** Before the growing season begins, the trial plots are prepared, and seeds of different genotypes are sown in predefined rows [11]. In our study, the field was divided into 22 x 11 plots, each plot measuring 3.8 meters in length and containing five rows of winter canola plants.
- **Visual trait assessments.** Throughout the growing season, breeders regularly visit the field to assess traits such as yield potential, disease resistance, and growth rate. These assessments, known as visual trait assessments, are performed visually and scores are assigned based on established criteria (e.g., a scale of 1 (low production) - 9 (high production) for yield) [2]. In our study, plant breeders evaluated traits such as the number of plants, early vigor, development before winter, leaf diseases, and yield parameters.
- **Data collection.** The breeders record the scores for each plot and compile the data for further analysis. This data helps to determine which genotypes perform better under the given environmental conditions.
- **Final assessment.** At the end of the growing season, a final evaluation is performed to assess the overall performance of each genotype, and decisions are made regarding which genotypes will be advanced for further breeding.

3.1.2 Challenges

- **Time-intensive data collection.** Frequent field visits are required to manually gather data throughout the growing season, making the process labor-intensive and time-consuming.
- **Inconsistent visual assessments across locations.** In breeding trials conducted at multiple locations, different individuals typically evaluate genotypes at each site, leading to variability in scoring due to subjective assessments.
- **Tracking developmental changes.** Monitoring trait changes over time is challenging with manual methods, making it difficult to capture growth progress accurately.
- **Scalability.** As trials expand, relying on human data collection becomes increasingly difficult to scale.

[2]https://www.avatars-project.de/immersive-analytics

Figure 1: The Valdemar robot equipped with multi-modal sensors for data capture from multiple viewpoints.

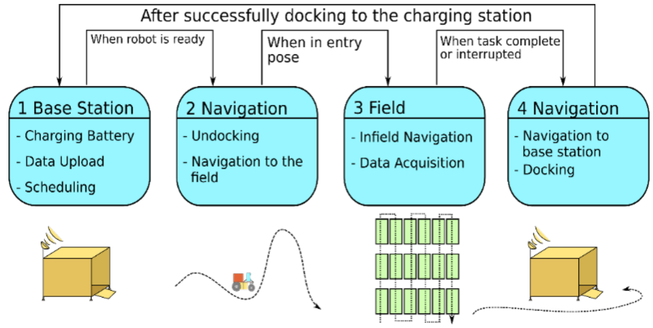


Figure 2: General operation concept of the Valdemar robot.

4 Robotic System for Data Acquisition

To be able to monitor canola plants throughout the entire season, we needed a specific robot setup with multiple views. The Valdemar robot, based on the Thorvald robotic platform [14], features a four-wheel configuration with full steering capability, enabling holonomic movement and point turning (see Figure 1). It was set up to fit the width of the canola breeding plots and equipped with an array of sensors. For autonomous navigation, we use a dual RTK-GNSS solution (Advanced Navigation Certus Evo) for localization, and a set of 360° Lidars (Velodyne Puck) and depth cameras (Luxonis OAK-D PoE) are intended to be used for navigation and collision avoidance. By combining the information of two GNSS receivers and a magnetometer, the absolute heading of the robot is known without using the wheel encoders.

Top-Down View. On the underside of the chassis, an array of four stereo plus 4k cameras (Luxonis OAK-D S2 PoE) is mounted to capture the plants while in motion. Additionally, on each of the long sides there are stereo depth cameras (StereoLabs ZED 2i), thermal (FLIR Boson 640) and multispectral cameras (JAI FS-3200T-10GE-NNC), and lidars (Ouster OS1-32G) to capture multimodal data of the plants.

Side View. To record data from the stems of the plants and to get a good perspective of the pods at the end of the growing season there are high-resolution cameras (Baumer VLXT-650C.I.EF) on each long side of the robot and a high-resolution terrestrial lidar (Z+F IMAGER 5016) on top of it. The robot has 2 Terabytes worth of data storage capacity and a battery runtime of 3-4h of deployment depending mainly on soil conditions and surrounding temperature.

4.1 Mobile base station

The base station described in the previous work [16] provides the robot with power, connectivity, and shelter. Solar panels, a wind turbine, and a power cable connection as a backup are used to charge a battery. Using an inductive charging system, power is transferred to the robot. Typically, it is deployed in the vicinity of a field so that the robot can autonomously drive from there to the field and return when it needs to recharge, or when the mission is completed, as can be seen in Figure 2.

4.2 Autonomous Approach

To enable fully autonomous missions, a navigation map derived from drone data is used. The navigation itself is executed with a set of specific proportional controllers to cover the four default movement directions. The controllers are implemented in ROS1 Noetic within the MoveBaseFlex [27] framework. For each plot lane, a rosbag[3], which contains all sensor data in a streamline, is recorded and manually transferred to a central storage to provide the data for the VR application.

[3]http://wiki.ros.org/Bags/Format

5 VR Application for Visual Trait Assessment

We developed a VR application to enable plant breeders to conduct visual trait assessments of breeding plots in field trials. We chose a VR interface for its immersive capabilities, which are essential for visualizing and interacting with 3D data, as research has shown its advantages for tasks that require 3D analysis [13, 8]. Our application has 2D and 3D visualizations of breeding plots, allowing breeders to evaluate plots from different locations and across multiple growth stages. For 3D data, we reconstructed point clouds and 3D textured meshes using high-resolution images of each plot. Since the captured data is GPS-localized, we can precisely visualize each plot in its spatial context. Key features of the application include side-by-side plot comparison, visualization of growth stages over time, and multiple view modes. These enhanced capabilities offer deeper insights into breeding field trials, which would not be feasible through traditional field assessments. Throughout the development process, we prioritized simplicity in the user interface (UI) of the VR app to ensure that it did not distract users from the assessment task, recognizing that many plant breeders in the target group had limited experience with VR.

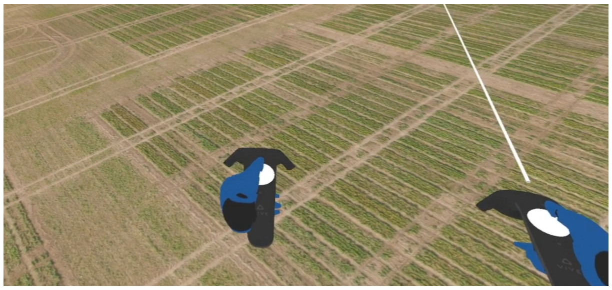
5.1 User-Centered Design

Given the limited experience that most plant breeders have with VR, we adopted a user-centered design approach. We consulted plant breeding experts throughout development, gathering feedback from interviews and usability studies to ensure that the VR application supported existing work practices, such as plot inspection and visual trait assessment, while adapting these tasks to a virtual environment.

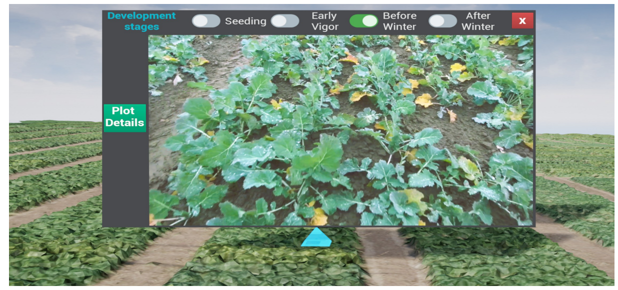
5.2 Prototype I

The first prototype was a proof of concept to demonstrate the potential of VR in visual trait assessment. It allowed users to explore two different breeding trial locations using 3D textured meshes reconstructed from images captured by a drone. The goal was to introduce plant breeders to VR as a new tool for their assessments and gather initial impressions.

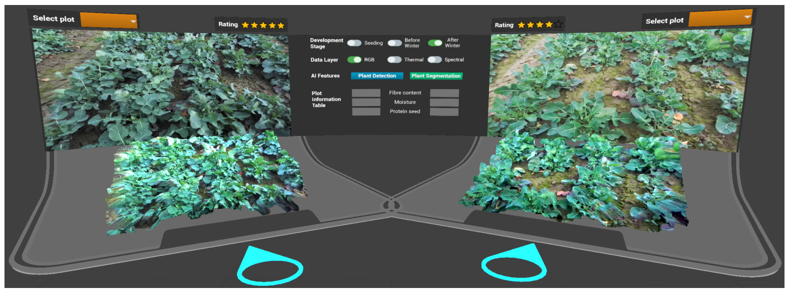
Feedback from semi-structured interviews with three breeders highlighted both the advantages and limitations of this early version. Although the immersive environment was well received, users were concerned about the lack of photorealistic detail in the visualizations. In addition, users suggested including an elevated view to get a better overall perspective of environmental conditions on plots at different locations. Another major suggestion was the integration of plot information waypoints to display detailed information about each plot, such as breeding database entries or 2D images.



(a) Aerial view mode.



(b) Walking view mode.



(c) Plot detail view mode.

Figure 3: Three view modes in VR: (a) Aerial view of the canola breeding field, (b) Walking mode with navigation via teleportation and interactive plot information waypoints for accessing plot data, and (c) Plot detail mode showing 3D data on a raised platform alongside 2D data display.

5.3 Prototype II

The second prototype aimed to address some of the limitations identified in Prototype I, while introducing new features to enhance the user experience. At this stage, we introduced two main view modes: an aerial view mode that offered a broad overview of the trial fields and a field-level view mode for closer inspection of individual plots. This provided users with the ability to switch between a high-level perspective and a more detailed view, offering flexibility depending on their assessment needs.

In this iteration, plot information waypoints were integrated into the system. When clicked, these waypoints displayed high-resolution images and relevant plot data, giving users immediate access to key information. However, due to the ongoing development of the data acquisition robot, photo-realistic data was still limited. We primarily relied on UAV drone data and side-view images captured by a small robot equipped with 3D LiDAR, which resulted in less detailed 3D meshes than initially desired. The breeders appreciated the added aerial view and information waypoints, but requested an improvement in data quality.

5.4 Prototype III

The third prototype addressed the issues identified in earlier feedback from plant breeders and was refined to support real visual assessment tasks (see Figure 3). With data now available from the new robot (see Section 4), we integrated photorealistic GPS-localized data and introduced the new "plot detail" mode for detailed plot inspection. In addition, we improved the UI of the aerial and walking modes to make them more intuitive.

5.4.1 Multiple View Modes

We implemented three display modes in our VR application—aerial mode, walking mode, and plot-detail mode—based on feedback from plant breeders. These modes provide different perspectives and levels of detail to enhance data interpretation. Users can seamlessly switch between these modes via the main menu in the UI.

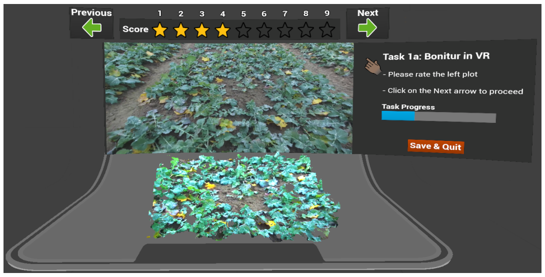
The aerial mode gives a bird's-eye view of the entire breeding field. The field is visualized as a 3D textured mesh from a top-down perspective. This mode allows users to get an overview of the trial, and observe how environmental factors at different locations affect breeding candidates. An example of this view is shown in Figure 3a.

The walking mode simulates the in-field experience, allowing users to virtually walk through the breeding field using teleportation. In this mode, users can inspect the plots up close, interact with them by clicking on plot information waypoints, conduct visual evaluations, and assign ratings to the plots. This interactive functionality mirrors how breeders assess plots during actual field trials, but in a virtual, immersive environment. An illustration of this mode is shown in Figure 3b.

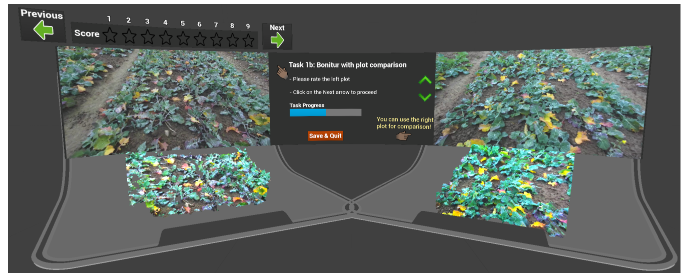
The plot detail mode offers the most in-depth analysis, focusing on one plot at a time. This mode allows users to seamlessly browse through plots by pressing a button on the VR controller. Users can view visualizations of a plot across different growth stages. In addition, users can compare the current plot with other plots. This feature is especially valuable for comparing replicated trials, where the same breeding candidates are evaluated in diverse environments. The mode also displays important metadata, such as moisture levels and other plot-specific information, on a 2D screen. An example of this mode, featuring both 3D and 2D visualizations, is shown in Figure 3c.

5.4.2 Plot Comparison

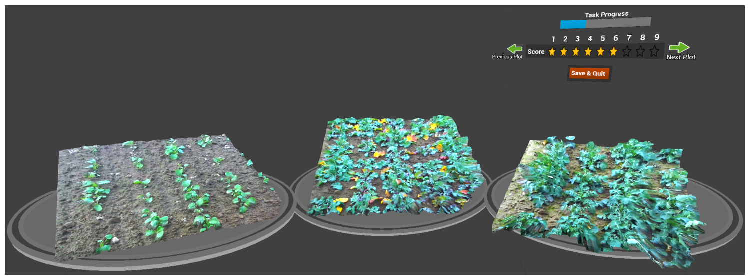
Plot comparison is one of the key features of our VR app, enabling side-by-side comparisons of breeding plots from different or the same locations. During initial interviews, breeders expressed a need to compare plots across locations during visual assessments, a task that is either difficult or impossible with current field-based methods. The data is captured by the robot at various trial locations, and the VR app brings these data together, allowing direct comparisons. It can help breeders assess how the same breeding candidates per-



(a) Task 1: Visual trait assessment in VR.



(b) Task 2: Visual trait assessment with plot comparison.



(c) Task 3: Visual trait assessment with plot history.

Figure 4: (a) In Task 1, users view a single plot and assign a score to a given trait. (b) In Task 2, users view two plots and rate the left plot using the right plot for comparison. (c) In Task 3, users view three developmental stages of the same plot and rate the latest stage, using the earlier stages for temporal context.

form under diverse environmental conditions throughout the growing season.

5.4.3 Plot History

Plot history enables users to view the development of individual plots over time, providing visualizations for multiple stages of crop growth. This feature helps breeders track the progression of traits from early growth to later stages, aiding in the analysis of how breeding candidates perform at each stage. By displaying historical data within the VR environment, users can compare the plot performance across different time points, allowing a more comprehensive assessment of plant development.

5.4.4 Plot Information Waypoints

Interactive plot information waypoints provide quick access to detailed plot information, including images and data from the breeding database. In walking mode, these interactive elements are displayed above each plot, becoming visible when the user approaches. Upon interaction, they trigger a 2D user interface panel that contains metadata from the database and plot images, as can be seen in Figure 3b.

5.4.5 Plot Rating

This feature allows plant breeders to perform visual trait assessments (see Section 3) and assign ratings to plots within the VR app on a scale of 1 to 9. The assigned ratings are automatically saved to the database.

6 User Study

The study was reviewed and approved by the Ethical Review Board of the Department of Computer Science of Saarland University, which determined there were no ethical concerns regarding the implementation of this research project. In our user study, we performed a task-based evaluation to determine how effectively our VR prototype supports the visual trait assessment process within plant breeder workflows (see Section 3.1.1). Rather than focusing only on individual UI components, we evaluated how well the prototype facilitates tasks that breeders typically perform during field trials, but in a virtual environment. We gathered quantitative data on task performance along with user feedback, providing insights into the system's performance and identifying areas for improvement.

6.1 Participants

We conducted our study with five female participants, all professionals in plant breeding, with varying levels of experience in visual trait assessment of canola. Their ages ranged from 23 to 56 years (M = 33.6, SD = 11.59), and their professional experience ranged from less than a year to more than seven years. Their skills in visual trait assessment ranged from novice to advanced, and only one participant had previous experience with VR. The participants were affiliated with NPZ Innovation GmbH and voluntarily participated during their work hours to contribute to improving a VR application for future integration into plant breeding workflows, without additional compensation. To avoid confusion with the plots, which we denote as P1, P2, and so on, we refer to the participants as "users": U1, U2, and so on.

6.2 Procedure

The users were first informed about the overall study and asked to complete a consent and demographic form. After agreeing to participate, they were introduced to the three visual trait assessment tasks, described in Section 6.4. They were also familiarized with the VR environment and controller options and given time to explore the VR application before beginning the tasks.

After each task, users filled out a questionnaire to provide feedback on their experience and preferences. A semi-structured interview followed to gather additional subjective insights. To simulate field conditions, users stood while performing the tasks, and breaks were provided between tasks to minimize fatigue.

6.3 Apparatus

We used the HTC Vive Pro headset with controllers to evaluate our VR prototype, which was developed in Unreal Engine. The application ran on a desktop PC equipped with an NVIDIA GeForce RTX 3080 graphics card. To capture user interactions and verbal feedback, we recorded screen activity and audio throughout the sessions.

6.4 Tasks

Users evaluated breeding plots in VR through three tasks, each testing different methods of visual trait assessment. For each task, they assigned a score from 1 to 9, with 9 representing the highest rating.

6.4.1 Task 1: VR assessment of individual plot

Users viewed one plot at a time, displayed as a 2D image and a 3D point cloud, as seen in Figure 4a. They assigned a score based solely on the visual information presented. A total of 10 randomly selected plots were evaluated. This task aimed to assess how effectively users could perform trait assessments in the VR environment.

In this task, users were asked to evaluate the trait *development before winter*. This trait refers to the biomass development of plants before the onset of winter. The trait was rated on a scale from 1 to 9, with 1 representing little biomass and few leaves, and 9 indicating substantial biomass and many leaves.

6.4.2 Task 2: VR assessment with plot comparison

Users compared two plots side-by-side in VR, using 2D and 3D visualizations, as seen in Figure 4b. The left plot remained fixed, while users could change the right plot from a selection of available plots to provide context for comparison. They then scored the plot on the left, considering the visual differences between the two. This task aimed to explore whether breeders found it helpful to have comparison plots. In this task, the users assessed the same trait as in Task 1, *development before winter*.

6.4.3 Task 3: VR assessment with plot history

Users viewed a plot's development over three different growth stages based on 3D point cloud visualizations, as seen in Figure 4c. They scored the latest stage while considering the earlier developmental stages. This task aimed to explore whether breeders found it helpful to have the temporal context of a plot's previous growth history when conducting assessments. In this task, users assessed the trait *development after winter*. This trait reflects the ability of the plant to survive winter conditions and quickly resume growth in spring.

6.5 Exploratory Phase: Evaluation of System Usability

In the final exploratory phase, users evaluated the complete VR prototype with all features, including multiple view modes (see Section 5.4). There were no specific tasks or time limits during this phase; users were free to explore the system at their own pace. Afterward, they completed a positive version of the System Usability Scale (SUS) questionnaire, designed to reduce cognitive load and collect reliable feedback on ease of use [33].

7 RESULTS

In this section, we present the results of our study, organized by the three tasks outlined in the user study. We present both quantitative results, focusing on the performance data, and subjective results, capturing users' feedback and insights gathered from the post-task questionnaires and semi-structured interviews.

7.1 Task 1

In Task 1, five users conducted visual trait assessment of 10 breeding plots in VR and assigned scores using a 1-9 scale.

Table 1: Visual trait assessment scores for Task 1

Plot ID	User Scores					Mean	SD
	U1	U2	U3	U4	U5		
P1	8	7	7	7	5	6.8	0.98
P2	6	5	6	5	4	5.2	0.75
P3	5	6	8	8	6	6.6	1.20
P4	6	5	7	6	5	5.8	0.75
P5	5	6	5	6	6	5.6	0.49
P6	4	4	4	4	3	3.8	0.40
P7	8	9	9	9	8	8.6	0.49
P8	5	3	3	5	3	3.8	0.98
P9	5	6	6	6	5	5.6	0.49
P10	6	7	7	6	6	6.4	0.49

Low score (1-4) Medium score (5-6) High score (7-9)

(a) Plot 7, Mean score: 8.6.

(b) Plot 6. Mean score: 3.8.

Figure 5: Images of the highest and the lowest rated plots in Task 1. (a) Highest-rated plot 7, (b) Lowest-rated plot 6.

7.1.1 Quantitative Results

VR Assessment Scores. The scores assigned by the users to each plot are summarized in Table 1. Plot 7 consistently received the highest score (M = 8.6), followed by Plot 1 (M = 6.8), suggesting that users found that these plots exhibit a superior trait. In contrast, Plots 6 and 8 received the lowest mean scores (M = 3.8), indicating that these plots were perceived to have the poorest performance in traits. These assessments align with the visual representations provided in Figure 5, which shows the images of the plots that received the highest and lowest average scores. Figure 6 illustrates the trend of individual user scores, with the graph showing alignment in user ratings across different plots, pointing to consistency in trait assessment.

VR vs. Field Assessment. To assess the reliability of visual assessments in VR, we compared the mean VR scores of the users with the available field assessment data from a single breeder (see Figure 7). For five of nine plots, VR scores closely matched field scores, showing less than a 1-point difference. Due to the unavailability of field data for Plot 8, it could not be compared. Since field data came from only one breeder and were compared to the average VR scores of five users, we cannot statistically confirm the alignment between the two methods.

Inter-Annotator Agreement: We evaluated the reliability of VR assessment scores using the Within-Group Agreement (rWG) [18]. It offers a plot-specific measure of agreement, allowing us to assess consistency among the raters for each individual breeding plot. This enables the identification of any specific plots with lower agree-

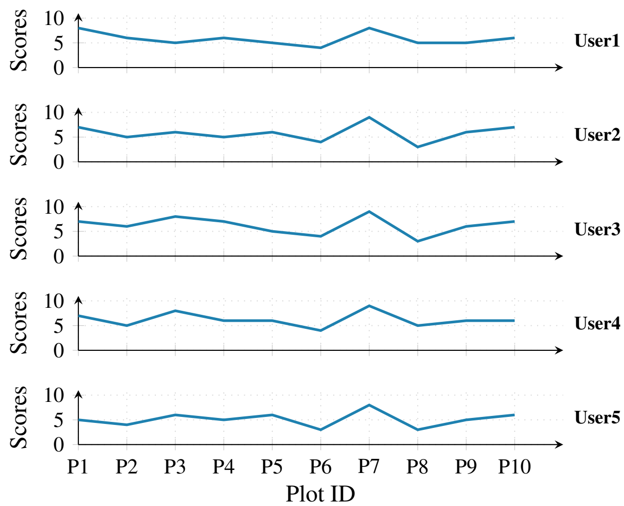
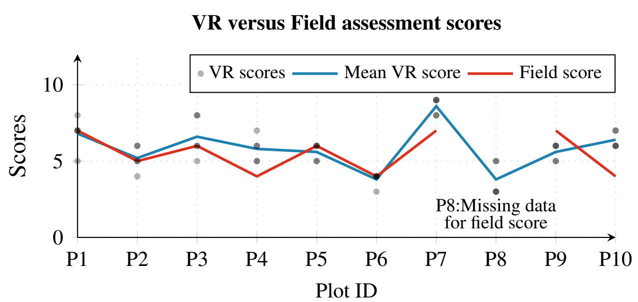


Figure 6: Trend analysis of visual trait assessment scores in VR, showing individual scoring patterns for each user.



Plot ID	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
VR	6.8	5.2	6.6	5.8	5.6	3.8	8.6	3.8	5.6	6.4
Field	7	5	6	4	6	4	7	-	7	4

Figure 7: Comparison of VR and field assessment scores. Plots with score differences less than 1 point are highlighted in blue in the table below.

ment.

Table 2 shows the rWG scores alongside the mean ratings for each plot. The mean rWG score (0.895) and median rWG score (0.925, range: 0.730 to 0.970) were well above the 0.70 benchmark for strong agreement [19], reinforcing consistency in the scoring between users.

VR Assessment Efficiency. We measured the time it took users to assess each plot. To ensure more accurate measurements, we excluded the time spent on the first plot, as users were still familiarizing themselves with the VR environment, which typically made the first assessment longer. The results revealed a significant range in assessment speed. The fastest user (U3) took an average of 5 seconds per plot, while the slowest (U1) averaged 37 seconds (see Figure 8). Across all users, the mean assessment time was 18 seconds per plot.

7.1.2 Subjective Results

To complement our quantitative findings, we collected subjective feedback through post-study questionnaires focusing on users' perceptions of the trait assessment in VR. The questionnaire addressed three key aspects: perceived accuracy, perceived speed of assessment compared to field-based methods, and the usefulness of 3D data visualization.

Table 2: Within-Group Agreement (rWG) Scores and Mean Scores for plots in Task 1.

Plot ID	P1	P2	P3	P4	P5
rWG	0.820	0.895	0.730	0.895	0.955
Mean Score	6.80	5.20	6.60	5.80	5.60

Plot ID	P6	P7	P8	P9	P10	Mean
rWG	0.970	0.955	0.820	0.955	0.955	0.895
Mean Score	3.80	8.60	3.80	5.60	6.40	5.82

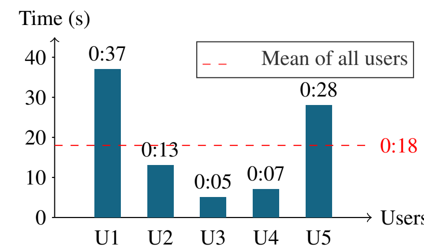


Figure 8: Average time per plot taken by users for Task 1.

Perceived Accuracy: When asked about the perceived accuracy of the VR-based assessment compared to the field-based method, the responses were mixed but generally positive. Three users rated the VR-based method as equally accurate as the field-based approach, while two perceived it as slightly less accurate (Q1 in Figure 9).

Perceived Speed: Users were also asked to compare the efficiency of VR-based assessment with field assessment. Two users indicated that the VR-based assessment was slightly faster than the field-based method (Q2 in Figure 9). Two others found the speed to be comparable to the field-based method, while one user perceived the VR-based method to be slightly slower.

3D Data Visualization. Given that we provided 2D (images) and 3D (point clouds) visualizations, we sought to understand the value of 3D data for breeders during assessments. Users generally found the 3D data visualization useful. One user rated it as *very helpful*, three as *helpful*, and one as *neither helpful nor unhelpful* (Q3 in Figure 9).

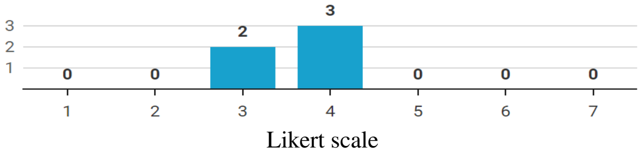
7.2 Task 2

In Task 2, users performed visual trait assessments in VR with the option to compare the plot they were evaluating against any available plot. This task allowed us to examine user preferences and assess how helpful the comparison feature was in the assessments. In the post-task questionnaire, we asked users which method they preferred. Three out of five users favored the side-by-side plot comparison offered in Task 2, while the other two preferred evaluating a single plot, as in Task 1. When asked whether the plot comparison feature aided their decision-making in trait assessment, one user rated it as *significantly improved*, three rated it as *improved*, and one was *neutral* (Q4 in Figure 10).

7.3 Task 3

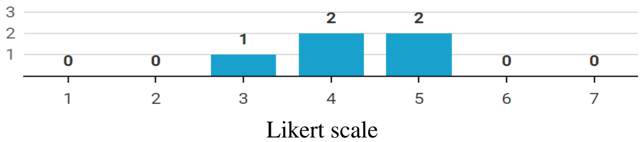
In Task 3, users evaluated breeding plots using the *"plot history"* feature, which allowed them to view and assess the development of a plot across three different growth stages in VR. After completion of the task, we asked users if they agreed with the statement that the plot history aided their decision-making during visual trait assessment. The responses were generally positive, but varied: two users *agreed* with the statement, two *somewhat agreed*, and one user remained neutral, *neither agreeing nor disagreeing* (Q5 in Figure 11).

Q1. Were VR assessments as accurate as field?



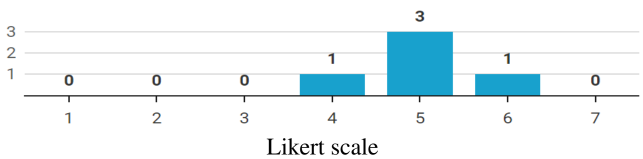
1: Much less accurate, 2: Less accurate, **3: Slightly less accurate, 4: About the same**, 5: Slightly more accurate, 6: More accurate, 7: Much more accurate.

Q2. Were VR assessments as quick as field?



1: Much slower, 2: Slower, **3: Slightly slower, 4: About the same, 5: Slightly quicker**, 6: Quicker, 7: Much quicker.

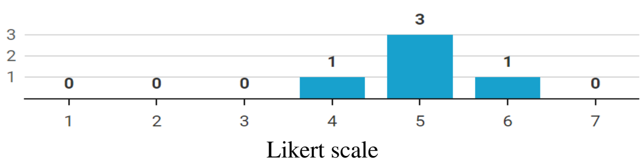
Q3. How helpful was 3D visualization?



1: Extremely unhelpful, 2: Very unhelpful, 3: Unhelpful, **4: Neither helpful nor unhelpful, 5: Helpful, 6: Very helpful**, 7: Extremely helpful.

Figure 9: Post-task questionnaire responses following Task 1.

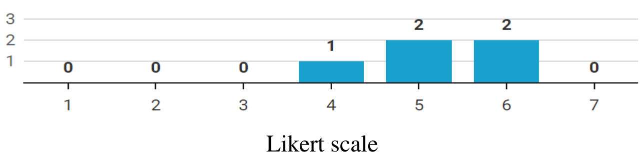
Q4. Did plot comparison aid your decision-making in the VR assessment?



1: Significantly worsened, 2: Worsened, 3: Slightly worsened, **4: Neither improved nor worsened, 5: Improved, 6: Significantly improved**, 7: Greatly improved.

Figure 10: Post-task questionnaire responses following Task 2.

Q5. The plot history aided my decision making in the VR assessment.



1: Strongly disagree, 2: Disagree, 3: Somewhat disagree, **4: Neither agree nor disagree, 5: Somewhat agree, 6: Agree**, 7: Strongly agree.

Figure 11: Post-task questionnaire responses following Task 3.

Table 3: System Usability Scores for the VR application

Statement	U1	U2	U3	U4	U5
1. I would like to use the VR app frequently.	4	4	4	3	4
2. I found VR app to be simple.	4	4	4	4	4
3. I thought VR app was easy to use.	5	4	4	4	4
4. I could use VR app without technical support.	4	2	4	2	2
5. VR app functions were well integrated.	4	3	4	4	4
6. There was consistency in the VR app.	5	4	4	3	3
7. Most people would learn to use VR app quickly.	4	3	3	5	4
8. I found VR app very intuitive.	5	4	3	4	4
9. I felt confident using the VR app.	5	3	4	4	4
10. I could use VR app without learning anything new.	5	2	3	4	3
SUS Score	**87.5**	**57.5**	**67.5**	**67.5**	**65.0**
Average SUS Score	**69.0 (Above Average)**				

Scale: 1: Strongly disagree, 2: Disagree, 3: Neutral, 4: Agree, 5: Strongly agree

Low score: 1-2 Medium score: 3 High score: 4-5

7.4 System Usability Assessment

The SUS evaluation revealed scores ranging from 57.5 to 87.5, with a mean score of 69 (SD = 11.12), as seen in Table 3. According to established benchmarks, scores above 68 indicate high usability [3], placing our VR prototype in the *'Above Average'* category. Three users gave low ratings (2 out of 5) when asked if they "could use the VR app without the support of a technical person." In particular, all participants, except one, were first-time VR users, which likely influenced their confidence in using the system independently.

8 Discussion

The results of our studies indicate both opportunities and challenges of VR for performing visual traits assessment, as follows:

VR offers potential for reliable and efficient trait assessment. Our evaluation demonstrated that VR can be a promising tool for consistent and reliable visual trait assessment. The ratings across different plots in Task 1 were mostly consistent, indicating that trait assessment in the VR environment is reliable. This is further supported by high inter-annotator agreement scores, suggesting that the traits presented in VR were observable, enabling multiple breeders to make comparable assessments.

In terms of efficiency, users took an average of 18 seconds per plot, with times ranging from 5 to 37 seconds. This variability reflects how quickly different users adapted to the VR environment and how much time they spent evaluating the plots. Since users were not required to prioritize speed, some took extra time to familiarize themselves with the interface. In particular, U3, who completed the assessments in 5 seconds per plot, had prior VR experience, while U1, who took 37 seconds per plot, noted: *"As I have never used VR, I think I am slower, but if I get used to it, it would be as fast as the field assessment."* This suggests that VR familiarity may significantly impact the speed of the assessment and could be improved through training.

Subjective insights highlighted both the strengths and limitations of the VR experience. Users generally appreciated the realistic visual representations, but certain areas for improvement were also noted. For example, a breeder commented, *"It would be better to have a complete plot (visualization)."* This feedback points to limitations in autonomous data acquisition, particularly in obtaining accurate and complete 3D reconstructions of entire 3.8 m plots. To ensure consistency in the VR environment, we presented users with visualizations of partial plot areas rather than full-length plots. Despite this, users were able to make consistent assessments.

VR versus field assessments. We found a close alignment between VR and field assessment scores, although this comparison was limited to data from a single breeder (see Figure 7). Although this consistency between VR and field scores is encouraging, it is not sufficient to establish statistical significance. In subjective feedback, three out of five users rated the VR-based method as equally accurate to the field-based approach, while two perceived it as slightly less accurate. One potential reason for the varied perceptions of accuracy could be the lack of tactile feedback (the ability to physically inspect plant leaves and stems for texture or subtle damage) in VR, which some users noted as crucial when assessing later growth stages. In terms of efficiency, users generally perceived VR assessments to be as fast or slightly faster than field assessments. One user noted that VR was faster because it eliminated the need to go into the field and physically walk between plots.

VR-enhanced capabilities. We integrated two advanced features into our VR system, which are unavailable in traditional field assessments. Task 2 enabled users to compare one plot with another in the same or a different field, while task 3 allowed users to view past developmental stages of a plot during its evaluation. Our objective was to determine whether these features provided practical benefits for plant breeders during assessments.

Users gave mixed feedback on the plot comparison feature. One user noted that while comparisons were interesting, they preferred to focus on the plot they were actively evaluating, indicating that additional context could distract. Another breeder proposed that showing both a *good* and *bad* plot for reference could enhance decision making, suggesting that benchmark plots may help provide clearer comparisons.

Feedback on the plot history feature in Task 3 was generally positive. User 3 highlighted the benefit of tracking developmental changes over time, noting that *"Being able to visualize earlier growth stages is really helpful, especially if there are significant changes over time."* This comment and the feedback about this feature (see Figure 11) suggest that the ability to visualize a plot's progression across time is useful, offering insights that are often hard to capture in the field, where such temporal context is not available.

VR can be used as a training aid for new plant breeders. Users recognized the potential of VR to train new plant breeders. User 1 pointed out that VR can be used to demonstrate how to perform visual trait assessments, stating: *"you can show them how to rate the plots. On the field, every year is different, sometimes you have (plant) diseases, sometimes not."* This highlights how VR can offer a standardized training environment that field conditions, with their yearly variability, cannot. U2 and U3 also saw its value, and U3 emphasized that it could supplement traditional field training by providing consistent and repeatable scenarios for learners. VR is especially useful in agricultural domains, where training opportunities are often constrained by seasonal factors.

9 Limitations

Due to the specialized nature of plant breeding and the practical challenges of working with domain experts, our study included a limited number of plant breeders. Although this sample size is typical for HCI research involving expert users (e.g., [26, 37, 30, 22, 4, 5]), the small pool of participants may limit the broader applicability of our findings. As part of the evaluation, we compared the mean assessment scores of five breeders in VR with the corresponding field scores of one breeder across nine plots. However, statistical analysis was constrained by the methodological limitation of comparing mean scores with individual ratings and the small sample size. Furthermore, comparative analyses between 2D image-only and 3D point cloud-only visualizations, which could more effectively quantify the efficacy of VR-based visualization for visual trait assessment, were beyond the scope of this study.

One technical limitation was inherent to the VR hardware, including a limited field of view and the weight of the device. Additionally, our prototype focused on visual trait assessment at specific growth stages, particularly lacking coverage of later developmental phases, such as pre-harvest. These later stages are crucial for evaluating traits such as yield, which significantly influence breeding decisions. This limitation arose from constraints in data collection with our robotic system. The increased plant height during later growth stages necessitated adjustments to the robot's navigation algorithm and sensor placements—modifications that could only be implemented in subsequent growing seasons due to the seasonal nature of field trials.

10 Conclusion

In this study, we explore the use of VR as a tool for visual trait assessment in plant breeding, investigating how immersive technology can improve traditional agricultural workflows. Our VR prototype offers capabilities not possible with current methods, such as visualizing plant development over time and comparing breeding plots across various locations. Task-based evaluation suggests that breeders can perform reliable visual assessments of plant traits in VR. However, some limitations remain, including technical limitations of VR hardware and reduced data quality at later developmental stages. Moving forward, we plan to continue to collaborate with plant breeders to improve both the data acquisition process and the VR prototype. Our goal is to integrate these systems with existing breeding tools and use data from VR assessments to train machine learning models. Overall, our findings show that combining autonomous data collection with immersive VR environment can effectively complement traditional field assessments, offering new opportunities for remote, data-driven decision making in plant breeding. This work highlights the potential of robotics and VR in agriculture, offering valuable insights for designing immersive systems in other data-intensive domains. It represents the first practical application of combining these technologies in plant breeding, and more broadly in agriculture.

Acknowledgments

We would like to thank the plant breeders who generously took the time to participate in our user study and provided valuable feedback; Alice Chase for early feedback that greatly improved this article; and Nina Knieriemen for helpful discussions. This work is funded by the Bundesministerium für Ernährung und Landwirtschaft (BMEL) as part of the PORTAL research project (FKZ: 28DK111B20).

References

[1] I. J. Akpan and M. Shanker. A comparative evaluation of the effectiveness of virtual reality, 3d visualization and 2d visual interactive simulation: an exploratory meta-analysis. *Simulation*, 95(2):145–170, 2019. 2

[2] C. F. Azevedo, L. F. V. Ferrão, J. Benevenuto, M. D. V. de Resende, M. Nascimento, A. C. C. Nascimento, and P. R. Munoz. Using visual scores for genomic prediction of complex traits in breeding programs. *Theoretical and Applied Genetics*, 137(1):9, 2024. 2

[3] A. Bangor, P. T. Kortum, and J. T. Miller. An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction*, 24(6):574–594, 2008. 8

[4] L. Besançon, P. Issartel, M. Ammi, and T. Isenberg. Hybrid tactile/tangible interaction for 3d data exploration. *IEEE transactions on visualization and computer graphics*, 23(1):881–890, 2016. 9

[5] L. Besançon, A. Semmo, D. Biau, B. Frachet, V. Pineau, E. H. Sariali, M. Soubeyrand, R. Taouachi, T. Isenberg, and P. Dragicevic. Reducing affective responses to surgical images and videos through stylization. In *Computer Graphics Forum*, vol. 39, pp. 462–483. Wiley Online Library, 2020. 9

[6] A. Bueckle, K. Buehling, P. C. Shih, and K. Börner. 3d virtual reality vs. 2d desktop registration user interface comparison. *PloS one*, 16(10):e0258103, 2021. 2

[7] J. N. Cobb, R. U. Juma, P. S. Biswas, J. D. Arbelaez, J. Rutkoski, G. Atlin, T. Hagen, M. Quinn, and E. H. Ng. Enhancing the rate of genetic gain in public-sector plant breeding programs: lessons from

the breeder's equation. *Theoretical and applied genetics*, 132:627–645, 2019. 1

[8] A. Doula, T. Güdelhöfer, A. Matviienko, M. Mühlhäuser, and A. S. Guinea. Pointcloudlab: An environment for 3d point cloud annotation with adapted visual aids and levels of immersion. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11640–11646. IEEE, 2023. 2, 3

[9] T. Duckett, S. Pearson, S. Blackmore, B. Grieve, W.-H. Chen, G. Cielniak, J. Cleaversmith, J. Dai, S. Davis, C. Fox, et al. Agricultural robotics: the future of robotic agriculture. *arXiv preprint arXiv:1806.06762*, 2018. 2

[10] D. A. Fasoula, I. M. Ioannides, and M. Omirou. Phenotyping and plant breeding: overcoming the barriers. *Frontiers in plant science*, 10:1713, 2020. 2

[11] W. Fehr. *Principles of cultivar development: theory and technique*. Macmillan Publishing Company, New York, 1991. 2

[12] S. Fountas, N. Mylonas, I. Malounas, E. Rodias, C. Hellmann Santos, and E. Pekkeriet. Agricultural robotics for field operations. *Sensors*, 20(9):2672, 2020. 2

[13] A. Franzluebbers, C. Li, A. Paterson, and K. Johnsen. Virtual reality point cloud annotation. In *Proceedings of the 2022 ACM Symposium on Spatial User Interaction*, pp. 1–11, 2022. 2, 3

[14] L. Grimstad and P. J. From. The thorvald ii agricultural robotic system. *Robotics*, 6(4):24, 2017. 3

[15] J. L. Hatfield and C. L. Walthall. Meeting global food needs: realizing the potential via genetics× environment× management interactions. *Agronomy Journal*, 107(4):1215–1226, 2015. 1

[16] M. Höllmann, B. Kisliuk, J. C. Krause, C. Tieben, A. Mock, S. Pütz, F. Igelbrink, T. Wiemann, S. F. Martinez, S. Stiene, et al. Towards context-aware navigation for long-term autonomy in agricultural environments. In *Proceedings of the International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA*, pp. 69–74, 2020. 1, 3

[17] W. Hurst, F. R. Mendoza, and B. Tekinerdogan. Augmented reality in precision farming: Concepts and applications. *Smart Cities*, 4(4):1454–1468, 2021. doi: 10.3390/smartcities4040077 2

[18] L. R. James, R. G. Demaree, and G. Wolf. Estimating within-group interrater reliability with and without response bias. *Journal of applied psychology*, 69(1):85, 1984. 6

[19] J. M. LeBreton and J. L. Senter. Answers to 20 questions about interrater reliability and interrater agreement. *Organizational research methods*, 11(4):815–852, 2008. 7

[20] B. Li, X. Xu, J. Han, L. Zhang, C. Bian, L. Jin, and J. Liu. The estimation of crop emergence in potatoes by uav rgb imagery. *Plant Methods*, 15:1–13, 2019. 2

[21] J. Ma, Y. Li, K. Du, F. Zheng, L. Zhang, Z. Gong, and W. Jiao. Segmenting ears of winter wheat at flowering stage using digital images and deep learning. *Computers and Electronics in Agriculture*, 168:105159, 2020. 2

[22] N. McCurdy, J. Lein, K. Coles, and M. Meyer. Poemage: Visualizing the sonic topology of a poem. *IEEE transactions on visualization and computer graphics*, 22(1):439–448, 2015. 9

[23] T. Mueller-Sim, M. Jenkins, J. Abel, and G. Kantor. The robotanist: A ground-based agricultural robot for high-throughput crop phenotyping. In *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 3634–3639. IEEE, 2017. 2

[24] L. F. Oliveira, A. P. Moreira, and M. F. Silva. Advances in agriculture robotics: A state-of-the-art review and challenges ahead. *Robotics*, 10(2):52, 2021. 2

[25] L. F. P. Oliveira, A. P. Moreira, and M. F. Silva. Advances in agriculture robotics: A state-of-the-art review and challenges ahead. *Robotics*, 10(2), 2021. doi: 10.3390/robotics10020052 2

[26] V. Pooryousef, M. Cordeil, L. Besançon, C. Hurter, T. Dwyer, and R. Bassed. Working with forensic practitioners to understand the opportunities and challenges for mixed-reality digital autopsy. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2023. 9

[27] S. Pütz, J. S. Simón, and J. Hertzberg. Move base flex a highly flexible navigation framework for mobile robots. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3416–3421. IEEE, 2018. 3

[28] T. W. Rife and J. A. Poland. Field book: an open-source application for field data collection on android. *Crop Science*, 54(4):1624–1627, 2014. 1

[29] F. Röckel, T. Schreiber, D. Schüler, U. Braun, I. Krukenberg, F. Schwander, A. Peil, C. Brandt, E. Willner, D. Gransow, et al. Phenoapp: A mobile tool for plant phenotyping to record field and greenhouse observations. *F1000Research*, 11, 2022. 1

[30] J. Rogers, A. H. Patton, L. Harmon, A. Lex, and M. Meyer. Insights from experiments with rigor in an evobio design study. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1106–1116, 2020. 9

[31] A. Ruckelshausen, P. Biber, M. Dorna, H. Gremmes, R. Klose, A. Linz, R. Rahe, R. Resch, M. Thiel, D. Trautz, et al. Bonirob: an autonomous field robot platform for individual plant phenotyping. In *Precision agriculture'09*, pp. 841–847. Wageningen Academic, 2009. 2

[32] M. M. Sakha, F. Daiber, C. Tieben, and M. Enders. Virtual breeding nursery: Towards a vr digital twin for plant breeding. In *Proceedings of the 2024 ACM Symposium on Spatial User Interaction*, SUI '24. Association for Computing Machinery, New York, NY, USA, 2024. doi: 10.1145/3677386.3688876 2

[33] J. Sauro and J. R. Lewis. When designing usability questionnaires, does it hurt to be positive? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, p. 2215–2224. Association for Computing Machinery, New York, NY, USA, 2011. doi: 10.1145/1978942.1979266 6

[34] Y. Shu, Y.-Z. Huang, S.-H. Chang, and M.-Y. Chen. Do virtual reality head-mounted displays make a difference? a comparison of presence and self-efficacy between head-mounted displays and desktop computer-facilitated virtual environments. *Virtual Reality*, 23:437–446, 2019. 2

[35] M. Talbot. Yield variability of crop varieties in the uk. *The Journal of Agricultural Science*, 102(2):315–321, 1984. 1

[36] L. Volpato, F. Pinto, L. González-Pérez, I. G. Thompson, A. Borém, M. Reynolds, B. Gérard, G. Molero, and F. A. Rodrigues Jr. High throughput field phenotyping for plant height using uav-based rgb imagery in wheat breeding lines: Feasibility and validation. *Frontiers in Plant Science*, 12:591587, 2021. 2

[37] X. Wang, L. Besançon, D. Rousseau, M. Sereno, M. Ammi, and T. Isenberg. Towards an understanding of augmented reality extensions for existing 3d data analysis tools. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2020. 9

[38] J. Zhou, D. Yungbluth, C. N. Vong, A. Scaboo, and J. Zhou. Estimation of the maturity date of soybean breeding lines using uav-based multispectral imagery. *Remote Sensing*, 11(18):2075, 2019. 2