Plant Phenomics 7 (2025) 100013

Contents lists available at ScienceDirect

Plant Phenomics

journal homepage: www.sciencedirect.com/journal/plant-phenomics

Research Article

VRoot: A VR-Based application for manual root system architecture reconstruction

Dirk N. Baker [a,b,*], Tobias Selzner [c], Jens Henrik Göbbert [a], Hanno Scharr [d], Morris Riedel [b,a], Ebba Þóra Hvannberg [b], Andrea Schnepf [c], Daniel Zielasko [e]

[a] Jülich Supercomputing Centre, Forschungszentrum Jülich GmbH, Jülich, Germany
[b] School of Engineering and Natural Sciences, University of Iceland, Reykjavík, Iceland
[c] Institute of Bio- and Geosciences 3: Agrosphere, Forschungszentrum Jülich GmbH, Jülich, Germany
[d] Institute of Advanced Simulation 8: Data Analytics and Machine Learning, Forschungszentrum Jülich GmbH, Jülich, Germany
[e] Human-Computer Interaction, Trier University, Trier, Germany

ARTICLE INFO

Keywords:
Virtual reality
Root phenotyping
Root system architecture
3D image analysis
Immersive analytics

ABSTRACT

This article describes an immersive virtual reality reconstruction tool for root system architectures from 3D scans of soil columns. In practical scenarios, experimental conditions will be adapted to fit the need of the data analysis pipeline, including sieving and drying the soil before scanning. Based on previous reports of automatic systems that do not represent what experts would annotate, we developed a virtual reality system to assist with the extraction of root systems in cases in which automated approaches fall short of expert knowledge. The aim of the present study is to evaluate whether our immersive method is superior to classical annotation approaches when tested on synthetic data sets using untrained participants. Our laboratory user study consists of evaluating the root extractions of participants, along with their rating on central user experience and usability measures. We show significant improvement in *F*1 score across conditions (noisy or clear data) as well as an improved usability. Our study highlights that using virtual reality in root extraction improves accuracy, and we perform an in-depth evaluation of biases that occur when users trace roots in soil volumes.

1. Introduction

Understanding root systems is an underappreciated part of the process of sustainable agriculture [1,2]. Historically, it was not possible to access root systems except by using difficult excavation processes. However, more recent research highlights that an analysis of the spatial configuration of roots, i.e., the root system architecture (RSA), is a critical aspect of understanding plant response [3]. Notably, there is a large variance in root traits [4], which impacts functional aspects of plant behavior. The root traits of a single plant can be measured from its RSA. A digitized version of the RSA yields a full description of the functional and structural traits of the root system. Functional-Structural Plant Models (FSPMs) are coupled simulations of plant structure and functional traits. RSAs can be used as boundaries in these functional simulations to provide insight into plant performance that cannot be measured directly. FSPMs bridge gaps between measurable indicators that might not directly correlate unless the plant is viewed as a continuum model [5].

Ideally, non-destructive observations of RSAs are used in experiments to allow repeated measurements, such as is possible using rhizotrons for statistical descriptions of roots [6]. A full RSA reconstruction in the face of partial root obstruction or destructive measurements is challenging. Non-invasive 3D imaging methods, like Magnetic Resonance Imaging (MRI), can assist by giving a more complete insight into the root system architecture [7]. 3D imaging techniques do not require direct intervention into the plant growth, and they do not require the introduction of transparent obstructions. While the plant growth in a soil column is more restrictive than in a field, key insights can be gained from an in-depth analysis of 3D imaging data, such as the progression of diseases in the plant [8]. Another key aspect in the dissemination of information from plant image data is the potential to gain insight through root modeling, as both in-silico experiments [9] as well as quantification of the continuous processes between soil, plant and atmosphere can provide valuable insights [5].

The extraction of RSAa is more challenging and depends on the quality of the image data. Most approaches to extract RSAs from 3D image data result in tree-like or centerline structures that describe the

* Corresponding author. Jülich Supercomputing Centre, Forschungszentrum Jülich GmbH, Jülich, Germany.
E-mail address: d.baker@fz-juelich.de (D.N. Baker).

https://doi.org/10.1016/j.plaphe.2025.100013
Received 18 June 2024; Received in revised form 8 December 2024; Accepted 17 January 2025
Available online 26 March 2025

morphology of the root systems. Fully automatic approaches fall into the categories of topological analysis [10] or optimization-based approaches [7]. There are also semi-automatic approaches, that require user interaction for key aspects, or to correct the automatic propositions. For a fully automatic approach to function, the globally optimal solution to the extraction problem must reflect the correct RSA. This is not necessarily the case, as in some measurements, artifacts due to soil composition and soil water content can lead to a difference in measured and actual morphology [11].

Manual approaches are a way of dealing with challenging image data properties, as expert knowledge can be required to completely extract the RSA to a degree that it is useable in further analysis. Selzner et al. [12] show this in an analysis of MRI image segmentation and how it impacts automated tracings, an analysis which uses a previous version of VRoot as expert baseline. Past approaches typically aimed at solving this challenge through guided optimization [11] or semi-automatic correction [7]. To assist with manual RSA reconstruction, Virtual Reality (VR) software has been developed, which was used to model RSA functional properties [13]. However, usability of these systems was limited as they were less portable than current VR hardware, which have improved in accessibility and display properties. The use of modern Extended Reality (XR) or VR hardware and software has the potential to increase the space of experimental conditions that are useable in combination with 3D RSA extraction techniques, thus closing the gap between data analysis requirements and realistic experimental conditions. Within VR, RSAs can be visualized directly in a more intuitive embedding, as the 3D displays will be able to accommodate and visualize depth as well as spatial configuration. VR is a promising tool, as it has been shown to improve extraction quality for similar tasks in other disciplines, such as neural imaging [14]. To improve our workflows of manual RSA reconstruction, and to investigate the applicability of VR in these workflows, we have developed VRoot, a VR application that assists in the reconstruction of RSAs by visualizing the 3D volume and providing intuitive toolsets for the reconstruction and adaption of RSAs.

We have built VRoot with the toolsets needed for exact and expedient RSA reconstruction. There are many tasks for which VR improves the quality of task completion in comparison to desktop applications. In this work, we investigate two research questions: Does VRoot improve the data extraction workflow for users annotating 3D MRIs? Furthermore, is the RSA reconstruction using VRoot more exact than using classical methods?

To answer these questions, we have conducted a laboratory user study with participants on-site. With an in-silico 3D root image, we have tasked participants with extracting the root system using VRoot and NMRooting, a state-of-the-art desktop RSA extraction and analysis application [7]. With the resulting RSA reconstructions, we have quantified key traits of the root system, and more importantly, the accuracy of the extraction in comparison to the original RSA.

This work makes several key contributions to 3D plant phenomics. First, we present a new VR-based method to interactively reconstruct root systems from 3D imaging techniques. We quantify the user-based errors and reconstruction artifacts in a laboratory user study. Lastly, we evaluate the use of VRoot based on user feedback obtained through controlled questionnaires.

In the remainder of this section, we briefly describe basic terminology for VR and provide a background for our reference application, NMRooting. In Sec. 2, we describe VRoot in the context of our day-to-day extraction workflow as well as the setup of our laboratory user study. Results of the study and our data analysis results are shown in Sec. 3, while we discuss the findings and implications as well as future directions in Sec. 4.

1.1. Immersive Analytics

Visual Analytics is a discipline of supporting data analysis and reasoning through visualization and graphical interfaces [15]. A subset of this field, and the collection of techniques it encompasses, is Immersive Analytics (IA), which involves the use of immersive interfaces, such as VR, which itself is a subcategory of XR. There have been a large variety of use cases for IA in science [16], and the VR application described in this work is another example. We focus on use-cases that are comparable to the use of VRoot, to provide context and an overview of instances in which IA has provided increased insight for data analysis pipelines. The neuron tracing application developed by Usher et al. [14] use Head-Mounted Displays (HMDs) for the sparse annotation of 3D image data. Usher et al. [14] developed an application to trace neuron connections in 3D space with handheld controllers and consumer-grade VR hardware. Usher et al. show that there is a significant speedup for experts to annotate neuron traces in VR in comparison to a classical desktop application. This result has been further improved by the introduction of topological features assisting with the extraction of neuron traces as shown by McDonald et al. [17].

Immersive displays are varied, ranging from room-scale installations to small portable devices. In comparison to the ImFlip150 system used in Stingaciu et al. [13] for root annotation, HMDs require less space, are mobile/movable, and can be comfortably used at any office workplace. A disadvantage of HMDs in this comparison can be the loss of reference in the real world [18]. The HTC Vive Pro used in this work is a wearable low-persistence display [19], similar to other HMDs. It is tracked using base stations that help the HMD infer user position and orientation. VR controllers are similarly tracked, in addition to containing buttons that users can press for interactions. Interaction in VR is commonly done using interaction *metaphors*, which are user actions done using controllers or gestures that impact the virtual world, as the user cannot directly interact with it. These include grabbing to virtually pick up items, as is common in VR applications, but also pointing for movement outside of the restrictions of the installation or walkable space [16]. Our central consideration of choosing VR techniques without involving pass-through or see-through augmented reality devices involves reduced depth perception of 3D renders in see-through devices [20] and reduced text readability in pass-through devices [21]. As such, any use-case of multitasking between the virtual and real world that directly relates to the RSA data would suffer from these drawbacks. Additionally, performance on difficult tasks tends to be reduced when multitasking [22].

Evaluating the use of human-centered techniques, particularly in the case of immersive systems, is challenging [23]. However, there is a long history of formal analysis in human-computer interaction that we can make use of. For example, a common method of evaluation by users is the System Usability Scale (SUS) [24], which introduces the notions of usability regarding task completion. Questionnaires such as SUS have been predominantly designed for software system evaluation but can be used for immersive software, as the inherent effects are similar [25]. Evaluation metrics used in this work are described in Sec. 2.4.

1.2. NMRooting

NMRooting [7,26] is a framework and application for the extraction of root system architecture by extracting the minimum-weight shortest paths with additional functionalities, such as gap closing and semi-manual extraction. We chose this application as a baseline as it is a well-established application that is also a proxy for similar applications and approaches, such as those developed by Horn et al. [11] or Zeng et al. [10]. Furthermore, NMRooting has seen regular use, including recently by Le Gall et al. [27], who used the non-destructive investigation of the RSA through NMRooting to analyze whether the root water uptake profile is an indicator for plant development. NMRooting uses desktop user interaction metaphors such as clicking and dragging to fulfill 3D annotation. Clicking in NMRooting traces a selection ray from the viewing surface to the isosurface data, marking the first surface that it hits. Dragging is a metaphor that allows users to turn the camera around the isosurface data, as well as zoom and pan. Within NMRooting, users can alter the automatic tracing of the data set, which is more in-line with

the use of applications such as TopoRoot [10] or the application developed by Horn et al. [11]. However, more fine-granular alteration to the reconstructions are possible by directly interacting with the data, yielding higher extraction accuracy in cases such as larger gaps as reported by Horn et al. [11].

2. Methods

Fig. 1 shows the total setup of our workflow, from data extraction to eventual use. Plant containers are typically plastic containers with a single plant growing in sieved soil. We are aiming to measure and extract RSAs from a large variety of soils and water contents. The resulting soil volume is generally a slice based 3D volume. This volume might require stitching, depending on the explicit setup of the MRI scanner. The soil water content together with the soil type (such as either loam or sandy loam) impacts the overall signal-to-noise ratio [26,28]. Furthermore, ferromagnetic particles within the soil might impact the measurement quality and might even disrupt root signal continuity, resulting in gaps in the root system [11,12].

Segmentation is generally a voxel-based mapping that labels a part of the image as either foreground or background, thus reducing the complexity of the data. For our pipeline, we generally use 3D segmentation by deep-neural networks, as implemented by Zhao et al. [29] and Uzman et al. [30]. Previously, we have shown that this step improves both automatic as well as manual RSA reconstructions [12]. Fig. 1B contains an illustration of this step. Generally, network architectures vary, but the key challenge for segmentation networks is providing a full view of the root system in the soil, while removing noise from it.

This paper uses a semi-automated method as reference, and most automatization methods rely on sparse extraction from volumetric data. NMRooting developed in Van Dusschooten et al. [7] uses a voxel signal cutoff to determine what areas could include root voxels, followed by a signal-strength-weighted shortest-path algorithm to extract the root system. A key aspect that drove our implementation of VRoot is a challenge presented in the automatic tracing implementation by Horn et al. [11], who included a comparison to manually annotated RSAs. Their algorithm optimizes signal-based features. They highlight cases in which extractions that differ from manual annotation are computed, because the algorithm chooses smaller (more optimal) gaps based on signal strength, while the expert annotation bridges a comparably much larger gap in the segmented image data. This kind of expert knowledge is hard to incorporate into automated algorithms in cases where the information cannot be gained otherwise, for example through repeated measurements or measurements at a later point in time when the roots have developed further.

Our output structure relies on the Root System Markup Language (RSML) [31]. It is based on the XML standard and can be extended to other use-cases. We use RSML both as data output as well as the primary source of information when rendering RSA structures in 3D.

Ultimately, the VR application presented in this work provides an immersive visualization of root systems and can easily be coupled to automatic tracing algorithms through their standardized output. Our application furthermore benefits from segmentation approaches that enhance the spatial visibility of the root morphology [12]. Lastly, using our approach, one can use previously unusable image data, either because automated tracing algorithms still fail in certain cases, or because a level of precision needs to be reached that would be otherwise unobtainable with desktop software.

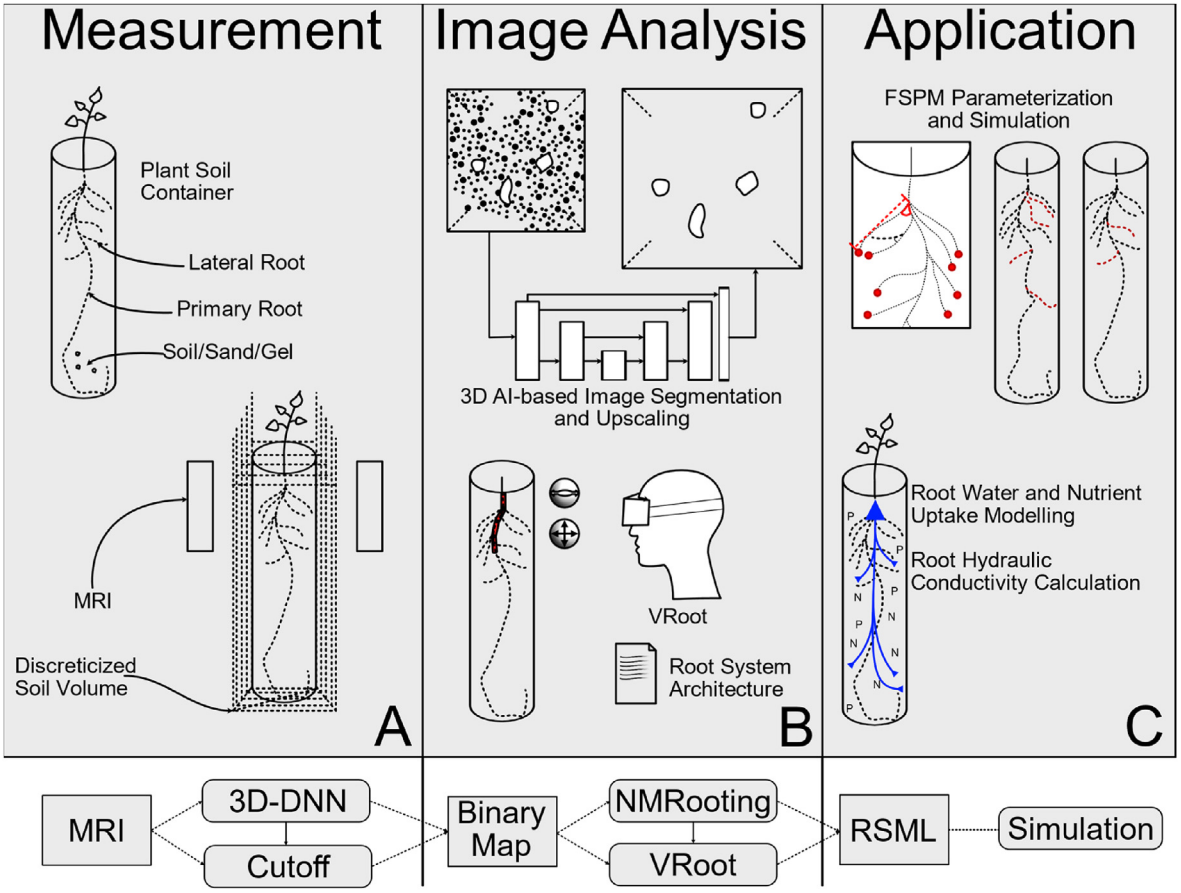

Fig. 1. Process overview of our MRI analysis pipeline. A: Plant containers are non-magnetic tubes with sieved soil. These are scanned using an MRI scanner, resulting in a voxelized soil volume. B: If the data is too noisy, 3D-U-Net segmentation can be employed in addition to the explorative functions of VRoot. C: The resulting RSAs can be used for FSPM calibration or functional simulation boundary. Below, we present a flowchart of the overall steps in our pipeline.

2.1. Virtual reality root tracing

We present *VRoot*, a novel method for manual RSA reconstruction and correction. To obtain the RSAs with the accuracy that we require, we developed a tool that allows expedient fine-tuning. Manual annotation in VRoot is loosely based on methods developed and used by Stingaciu et al. [13] and Usher et al. [14]. To support a wider usability in terms of target hardware, and customization options, we developed[2] VRoot on top of Unreal Engine[1], primarily to allow for reproduction even if hardware constraints are altered. A key aspect of the rendering of 3D imaging data in VR is the ability to look at the data from different perspectives while retaining the dimensionality of the data regarding their perception. We implemented VRoot with the key idea that different users might want to interact on different scales and in different postures, as previously highlighted by Zielasko and Riecke [32].

VRoot is an application that assists with very accurately tracing RSAs in 3D volumes, by providing immersive visualization and intuitive interaction. VRoot uses the Unreal Engine for cross-platform rendering and porting, while the interface to VR hardware uses the OpenXR abstraction standard. The data analysis system is a python server, communicating with the application via a remote connection implemented in ZeroMQ. VRoot almost exlusively uses geometric representations of data, augmented by custom interaction that allows editing of graph structures and visualization of soil scans. In the application, the soil volume is thresholded and visualized as isosurface computed around a cutoff value that can be chosen dynamically within the application. We use the Visualization ToolKit (VTK) [33] implementation for isosurfacing. VRoot is used to manually extract and edit RSAs, which are visualized on top of the 3D volume using the geometrization scheme described in Baker et al. [34]. Our implementation of VRoot consists of a full analysis pipeline using interaction metaphors, implemented[2] using the Unreal Engine. We implemted an RSML authoring system in VR, along with the possibility to extract, change, as well as fine-tune RSAs.

Fig. 2 shows sample views from the user's perspective in the application. We chose a darker environment to reduce eye strain. Users interact with two controllers, one for selecting as well as tracing, while the other controller is used for grabbing metaphors.

Fig. 2A shows the basic user interaction components. Interaction with the RSA is done via the nodes. All nodes in the RSA are selectable and while tracing depends on manual user interaction, changing of properties is selection set-based, meaning that users can mark as many nodes as desired to change their properties. The widgets (grey) change the selection set's properties, such as diameter and position. Fig. 2B shows a snapshot of a user drawing a root, indicated by the pink interaction. This is automatically available once the number of selected nodes is exactly one, or without any tracing present. The widgets to edit node properties always work on the selection set, changing the diameter or position of the selected nodes. The volume itself is displayed as an isosurface, as seen in Fig. 2C, whose signal cutoff value can be changed from within VR through the use of a slider widget. The full visualization of the RSA always uses the RSML topology and assigns colors to root order. We are using dithered translucency for the isosurface to avoid depth-perception issues with the surface. More general root functionality that is outside of node editing is accessible via a point-and-click menu, such as assistance tools for time series annotation or editing tools for RSA topology. Users can generally place nodes freely within the 3D environment, though it needs to be stressed that there is a degree of freedom in VR, the rotation, that is not captured by common data types, such as RSML or VTK format. To enable a more expedient workflow, many more complex tasks, including the visualization algorithms, are offloaded onto a server to allow the interactive application to run smoothly while still allowing the completion of complex tasks. For a selected data set, the system searches for the most recent tracing and displays it on top of the 3D image.

We designed the application with expert workflows in mind and have been continuously improving the workflow to assist with observations such as by Stingaciu et al. [13] and Selzner et al. [12]. However, the assessment of a system that depends on human interaction is not as straight forward as assessing the quality of an algorithm, even with a ground truth data set present. We are evaluating this in a more controlled fashion by performing a user study that is being guided by synthetic data and user questionnaires, under the restriction of using a pool of potential users that all have a similar knowledge level about the applications. We chose to evaluate untrained user performance for this reason, allowing us to focus on the relative performance of the applications without needing to balance the data for previous experience.

2.2. Laboratory user study

To answer our question on whether VR annotation can outperform state-of-the-art desktop annotation for RSA reconstruction, we performed a mixed design laboratory study, assessing the applications within-subjects with the between-subject condition of water noise. We compare our software against NMRooting described above, since NMRooting is not only state-of-the-art for 3D annotation, but also is similar to other applications. The evaluation of the study is aimed to answer the question on whether the VR software yields a higher reconstruction accuracy as well as a higher usability. We map performance to reconstruction accuracy in a virtual MRI scan: Our comparison between applications and conditions relies on the assumption that how closely a participant (after a short training phase) follows the ground truth with their annotation is a direct indicator of the usefulness of the application. In this instance, the term laboratory study refers to a controlled setting in which human participants with similar starting conditions could perform tasks and evaluate the applications. Through the use of a sufficient number of individual participants, effects that are individual to certain people should be eliminated, and the overall usability of the software can be evaluated. To enable this process, the set of possible options within a single application has to be restricted, so we exclusively use "tip-to-tree" and node annotation in NMRooting and basic drawing without correction in VRoot.

The user study was designed to answer our questions and assumptions on the improvement of software and measurement quality from MRI scans. We postulate the following hypotheses on the application performance on the software level as well as on the data level.

Improved Workflows: We expect that the major indicators for software quality will be im-proved when using VR software. These indicators are an improvement (**H1**) of System Usability as well as an improvement in the subjective pragmatic performance (**H2**) of the software. These hypotheses will be tested using the participants' evaluation using the questionnaire after task completion.

Extraction Accuracy: For the extracted root systems, we postulate that key relevant measures will be impacted by the use of VR. The total root length is expected to be different (**H3**), which also applies to the branching density (**H4**). We believe that the VR software will result in a higher overall accuracy (**H5**) which is further impacted by the presence of water noise (**H6**). Water noise and impact of signal-to-noise ratio have well-reported impacts on reconstruction accuracy [7,10–12], which is why we assume that it is a significant factor in the reconstruction accuracy of participants. We expect that the total root length is closer to ground truth when using VR (**H7**) and that the VR extraction of the branching density does not differ from the branching density within the virtual MRI (**H8**).

2.2.1. Tasks & measures

The main task that participants were asked to perform is extraction of the RSA from an MRI soil column scan. This includes the extraction of the pathway of individual roots as exhibited within the MRI scan, loosely

[2] VRoot implementation: https://github.com/dhelmrich/VRoot.

[1] Unreal Engine, developed by Epic Games Inc. https://www.unrealengine.com/, accessed 2025-03-27.

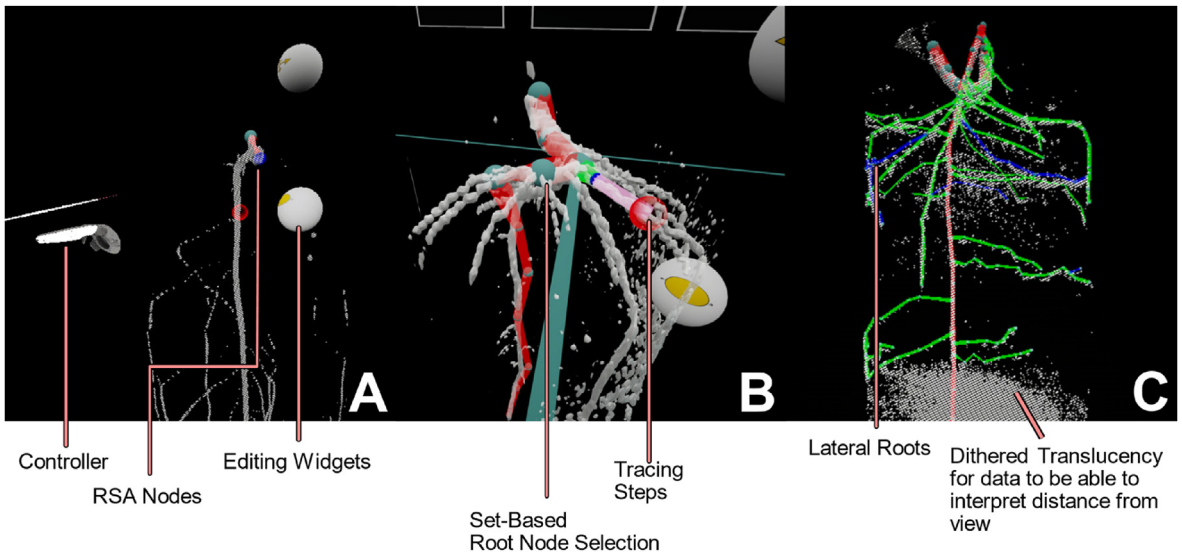

Fig. 2. Annotated screenshots of VRoot. A: Two-handed use of the drawing function. New position and connection is indicated in pink. B: Selection is set-based and changes are done on all selected nodes, drawing is only possible with one. C: Subsequent root orders have high color contrast and MRI is dithered for depth-preserving translucency.

based on signal strength. Participants were asked to mitigate noise effects if present and extract a fairly simple explanation for the signal that they were shown. Furthermore, participants created a *labeled* RSA, which includes the order of the root explicitly. Participants were asked to label both primary and lateral roots as such. The tracing of the RSA resulted in each case in a full RSA, including positioning but excluding diameter. Participants were asked to provide their demographic information as well as a subjective evaluation of the software they were tasked with using. In total, participants performed the extraction task two times, with the evaluation of a questionnaire in between and at the end.

Participants were tasked with extracting a root system from an MRI scan, once using the VR application and once using NMRooting. The water noise condition spanned both data sets, meaning that independent of the order, a participant either completed the task for each application with water noise, or without. Participants were tasked with tracing a virtual MRI scan, as described in Sec. 2.3. Extraction of the RSA was done with both applications, and the resulting structures were compared against ground truth.

Participants evaluated each application with the System Usability Scale (SUS) [24], the User Experience Questionnaire (UEQ) [35] as well as the NASA Task Load Index Short (TLXs) [36]. These are described in Sec. 2.4.

2.2.2. Procedure

Participants gave their informed consent. Participants were divided into four groups by ID. The first distinction was made on whether a participant received data with water-like noise. Furthermore, it is varied which application a user tested first, resulting in four conditions. The conditions were *order of application* and *water noise*. The study procedure consisted of five steps. In the first step, participants would quantify their own previous experience and calibration measurements were made to setup the HMD. Afterwards, participants would be introduced to the first application (Desktop or VR) and after this initial training phase, the study data set would be loaded, and the participant performed the task without help. Participants then evaluated the application using questionnaires. Lastly, these two steps would be repeated with the other application. For the full description of all steps involved in the individual phases, see App. B.

2.2.3. Apparatus

In our experiment, we ran VRoot on an HTC Vive Pro HMD. In our tests, the application framerate was typically within 80–90 frames per second. The entire study was conducted in the "Virtual Reality Laboratory" in the Institute of Bio- and Geosciences 3 of the Forschungszentrum Jülich GmbH. The study was conducted sitting at the laboratory desk, facing the monitor. The VR software was used sitting by all participants. For considerations on whether to support or design a system for standing or seated setups, we refer to current literature [37].

The questionnaires as well as the NMRooting application were used on a desktop PC with a desktop resolution of 1920×1080 with mouse and keyboard. Our tests were performed on a PC with an Intel i7-8700K CPU, 32 GB of RAM and an NVIDIA 2060 RTX SUPER GPU.

2.2.4. Participants

The user study data set was acquired from over 20 participants working on-site at Forschungszentrum Jülich. We have contacted potential participants, pre-emptively excluding anyone with either previous VRoot experience or knowledge of the goal of the study. Furthermore, we required normal or corrected-to-normal vision.

Total participation in the user study was $n = 20$. These include in total 15 male, 4 female, 1 non-binary and 0 other. Age distribution was almost uniform from 20 to 43 years, with a median age of 33. Self-reported experience using 3D applications was 4 participants with no experience, 8 users who reported using 3D applications at least once, 6 sporadic users and 2 experts. Self-reported experience using VR applications was 5: None, 5: Once, 7: Sporadic and 3: Expert. None of the participants had previous experience in the specific application NMRooting or the specific application VRoot.

2.3. Evaluation using FSPM simulated root data

We are evaluating user-based extractions of root systems in the context of a virtual MRI scan. These virtual MRI scans were designed specifically for this study and the RSAs were simulated using the FSPM CPlantBox [5]. We calibrated the simulation for the task and slightly increased the inter-lateral distance for the first-order lateral roots. This RSA serves as a ground truth measurement. With noise that we typically see on a larger scale, such as Fig. 3A, we modeled a smaller bean root system seen in Fig. 3B and imposed a noise model on it.

We computed the signal strength of the resulting MRI scan by using a simple heuristic based on the total volume of a root segment in a certain voxel. To avoid non-uniform task performance in the extraction task, the

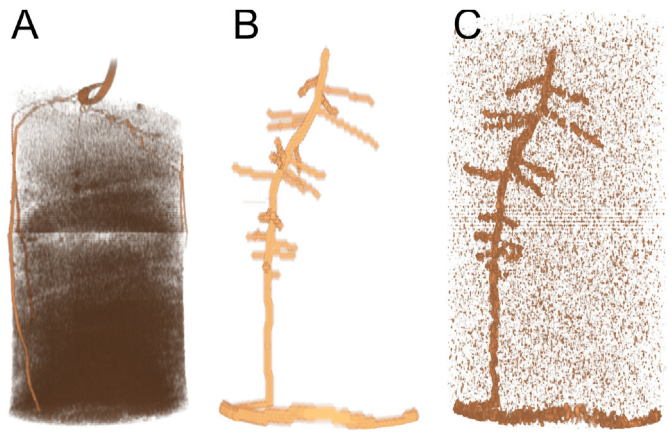

Fig. 3. Side-by-side comparison of the study image data sets, scaled to image height as opposed to real height. A: We typically observe, depending on soil type as well as soil water content, highly ununiform noise. Locally, this might be expressed as smudges around the roots. B: Our simulation of a faba bean is being rendered with respect to the relative root length/diameter through a voxel. This image data set is the use-case for the noise-free participants. C: We added and subtracted noise features using a Weierstrass transformation. This results in slightly more complex, but uniformly complex, image data.

between-subject condition of water noise was chosen such that either the whole data set had noise effects or no part of the data. Within a soil cylinder of 1.5 cm diameter, a soil volume with water noise was seeded. The noise was locally scaled with the signal-to-noise ratio of 4.3. Noise addition/substraction was followed by applying a weierstrass transformation.

Since the original root system has been created by an FSPM, we can directly compare it with the manual extractions. This allows for an exact quantification of errors, which helps in assessing the factors in the decision-making process on what application to use for the pipeline. CPlantBox has been well-researched in terms of its stochastic properties [38], and is fit to be used as baseline for a user-based evaluation of extraction software. Additionally, any effects we would see in terms of the impact of the synthetic nature of the plant we would see in both applications equally.

As described in Horn et al. [11], to be able to confidently match between user-annotated RSAs and those generated by software or simulations, the correspondence between the architectures needs to be calculated. In this evaluation, we have used the distance-matching threshold of $d = 15$ in voxel units that was used by Horn et al. [11]. Since any length of simulated root could correspond to one or more manually annotated segments, and likewise one manually annotated segment might correspond to more than one simulated root, we compute the full segment distance matrix as

$$(D)_{i,j} = \min_{x,y} \left\| P_i + x \cdot (P_i - P_{i-1}) - P_j + y \cdot (P_j - P_{j-1}) \right\|_2 \text{ where } x, y \in (0,1) \subset \mathbb{R} \tag{1}$$

where P_i is the point coordinate of the ith segment, x and y are optimization parameters and the resulting distance matrix D only captures the minimal euclidean distance between two lines. This will result in a base distance metric that we use to match segments of different lengths. The matching process first assigns 1-to-n correspondences to simulated segments before it tries to match any unmatched manual segments to those simulated segments that were previously matched to exactly one segment. Organ-level label continuity is provided through matching roots. Two roots are considered matched if their cumulative distance $D \cdot (e_n \times e_m) \mapsto d^{(m,n)} \in \mathbb{R}$ is the lowest among all other possible assignments with respect to the sets e of the segment indices.

We computed the accuracy based on root matching to ensure that the correct identification of roots is rewarded and to measure extraction differences that contribute to differences in root length. This topologically-aware accuracy is computed using a ground truth root set of I_{GT} and a set of traced roots I_T, with a set of $I_c := \{n \in I_{GT} \,|\, \text{argmin}_{m \in |R|} d^{(m,n)} \cap d^{(m,n)} \leq d\}$, signifying the correctly matched roots.

To compute the measures for the correctness of the extracted RSAs, we first match the root systems to the ground truth. This ensures that there is topological information in the resulting scores. Root Lengths L_n generally refer to the total length of all segments corresponding to the root, namely $L_n := \sum l_i^n$ of the $i \in I_n$ segments of organ index n. We use the ground truth L_{GT} as the reference for the scores, where L_{GC} is the total length of correctly matched roots computed from $I_{GC} \subseteq I_{GT}$. The root(s) that are matched to a root in the ground truth are a subset of the root set of the tracing, $I_{TC} \subseteq I_T$. It follows that the total length of false negative roots that were not traced is $L_{FN} = \sum l_i^n I_{GT} \setminus I_{GC}$. We especially highlight that the total length of correctly, which means matched, ground truth roots is not necessarily the same length as the sum of matched roots in the tracing, meaning that $L_{GC} \neq L_{TC}$ because $I_{GC} \subseteq I_{GT}$ whereas $I_{TC} \subseteq I_T$. L_{LP} is the total length of false positive roots, i.e., computed from segments that were not present in the ground truth but present in the tracing. The *recall* value R is a measure that encapsulates how much (in length) of the root system was traced, defined as

$$R = \frac{L_{GC}}{L_{GC} + L_{FN} + \min(0, L_{GC} - L_{TC})} \in [0,1] \tag{2}$$

where we further penalize the tracing in cases where the extracted root length is smaller than the length of the ground truth. On the other hand, the *precision* value P encapsulates whether the manual extraction contains only as much length as the ground truth root system:

$$P = \frac{L_{GC}}{L_{GC} + L_{FP} + \min(0, L_{TC} - L_{GC})} \in [0,1] \tag{3}$$

For the precision, we further penalize roots that were extracted correctly but are too long in comparison to the ground truth. The precision and recall values are asymmetrical, decreasing with different metrics, but can be summarized by the symmetrical F_1 score, which decreases with both false positives and false negatives:

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} \in [0,1] \tag{4}$$

The F_1 score is a comparison score that yields relatively similar values for different deviations from the ground truth. This score allows the comparison of applications that exhibit different characteristics, but due to its symmetrical nature, allows the comparison between them.

2.4. Measures for application comparison

The measures on usability of the software as well as user experience are difficult to measure objectively, and thus, questionnaires are typically utilized. These questionnaires include the System Usability Scale Questionnaire (SUS) [24], the User Experience Scale (UEQ) [35], and the NASA Task Load Index Short (TLXs) [36]. Due to the human interaction component of the system, we chose to use standard methods of evaluation with the added component of knowing the ground truth of a virtually generated MRI scan. Thus, we obtain a combination of subjective and objective measurements for assessing the software. We attached the full participant survey in the supplemental material.

The SUS is primarily aimed at quantifying the subjective user assessment of whether the given system is fit to help the user solve the problem. The resulting score is scaled within $[0,100] \subset \mathbb{N}$. Commonly, the software is considered to rate well on this scale if it is above 85 [39].

UEQ scores are a way of evaluating the quality of the subjective user experience regarding basic descriptions of the software. The UEQ is a mix of adjectives that are presented in a contrasting manner. The adjective

sometimes has overlapping meanings, and the general assessment of the software regarding these properties is very subjective. The NASA Task Load Index Short is a questionnaire to assess the subjective difficulty and strain on the user when completing the tasks. Users evaluated the applications in an online questionnaire, which we have attached to the supplemental material. We measured the extracted RSA, as well as camera data from participants, in addition to participants completing the questionnaire.

2.5. Data analysis

We aggregated the data into two groups, based on the water noise condition. We computed the subjective scores per participant according to the respective guidelines. This applied to SUS [24], UEQ [35], and the TLXs [36]. We performed the data analysis entirely in Python. In tables or figures, we will refer to VRoot simply as VR, and to NMRooting as Water/No Water Conditions are referred to as + W or -W respectively. As such, VR + W refers to all data points of the VR software that have the water noise condition. In the following, total (T) refers to all data points. For hypotheses testing, our confidence cutoff is $p = .05$.

We tested all measures for normal distribution using the Shapiro-Wilk test, to ensure subsequent tests are informative and valid. For the sake of uniformity and comparability, we are using non-parametric tests in cases where not every condition is normally distributed. We chose the Mann-Whitney (Summed Rank) test for differences in median in cases where we do not find a normal distribution.

Statistical reporting includes the test statistic $t(\mathrm{DoF})$, the critical value p, the effect size value (Cohen's d [40], defined as for differences in statistics T_i), as well as degrees of freedom (DoF). In cases where ground truth is available, we test for a specific means using t-tests. The test values for normal distribution can be found in App. A. We use t-tests for SUS, UEQ, and TLXs. Mann-Whitney tests will be used for total and average root length, F_1 score, and inter-lateral distance.

3. Results

We omitted one subject (female, +W) from the study after the subject succeeded the task *trace the taproot* within the training phase but did not succeed with that task in the study phase. This means that we have 10 data points for the condition -W and 9 for the condition + W. The tracing and details on reasons of omission can be found in App. C.

3.1. Descriptive data

We include box-plot descriptions of the relevant measures. Fig. 4 shows the SUS scores over all conditions as well as the TLXs scoring and UEQ pragmatic quality. The SUS scores are scaled within [0, 100], though must not be understood as percentages. We have computed the median in instances of data points that have no normal distribution, which is indicated by the orange line in the boxplots. Data sets that contain a ground truth (simulation) value indicate this value with a red line. Data points outside of the inter-quartile range ($1.5 \cdot (Q_3 - Q_1)$) have been included and marked as x-symbol.

The within-subject condition of the order of the application was combined it served as balancing of the applications against learning effects. The subjective scores that are relevant to the applications can be seen in Fig. 4. Herein, we present the SUS, the Task Load as well as the pragmatic quality. Fig. 5 shows the accuracy scores of the data sets. For a more in-depth understanding of the individual effects, we further present relevant RSA measures in Fig. 6. Herein, we have a ground truth measurement for all conditions. Ground truth measures were extracted from the virtual MRI algorithmically, meaning that we extracted measures with the MRI mapping and voxelization in mind. For the comparison, we show as a red line the ground truth value from the actual simulated data set as opposed to the parameterization. For the number of lateral roots, we filtered roots of a length $l_i < 3$ [cm] to allow for the evaluation of the extraction without unnecessarily including tracing artifacts in NMRooting (seen in Fig. 7). The inter-lateral distance was calculated on the taproot only.

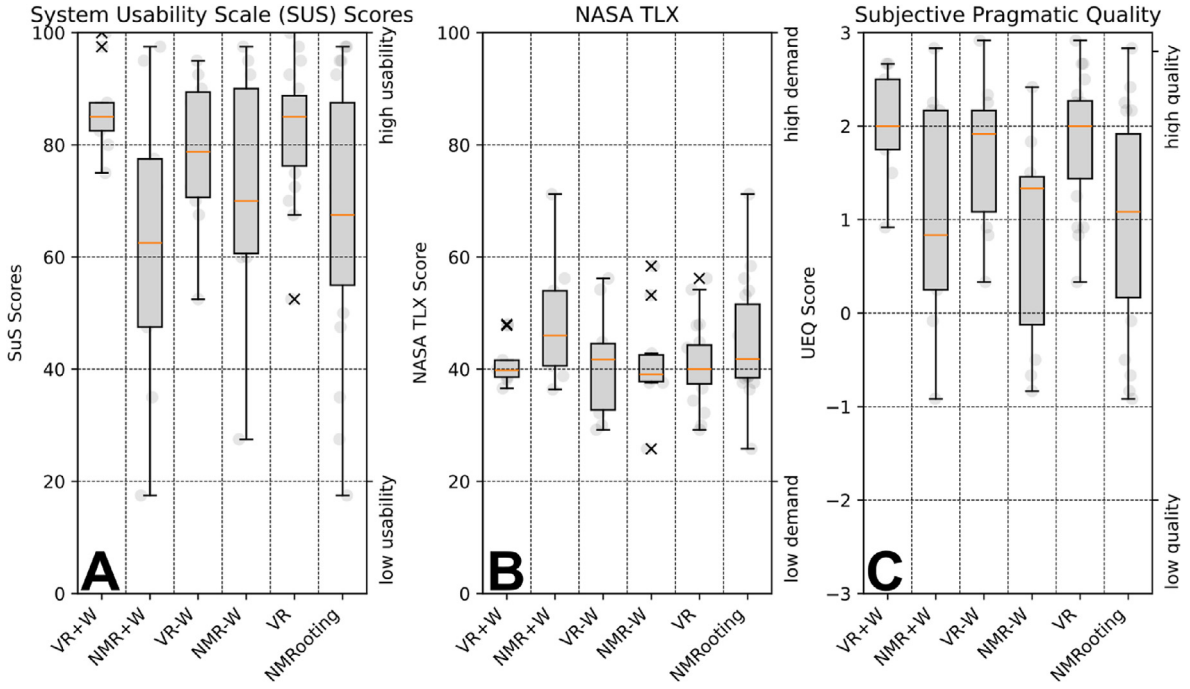

Fig. 4. Boxplot with conditions on x axis and score on y axis, orange line is the median. A: Overview of System Usability Scoring across conditions. Scoring across order conditions (within-subject) was summed. B: Overview of Task Load Index Short Questionnaire score, sum of all questions except perceived success, which was inverted. C: Overview of UEQ participant scoring for the pragmatic quality of the data.

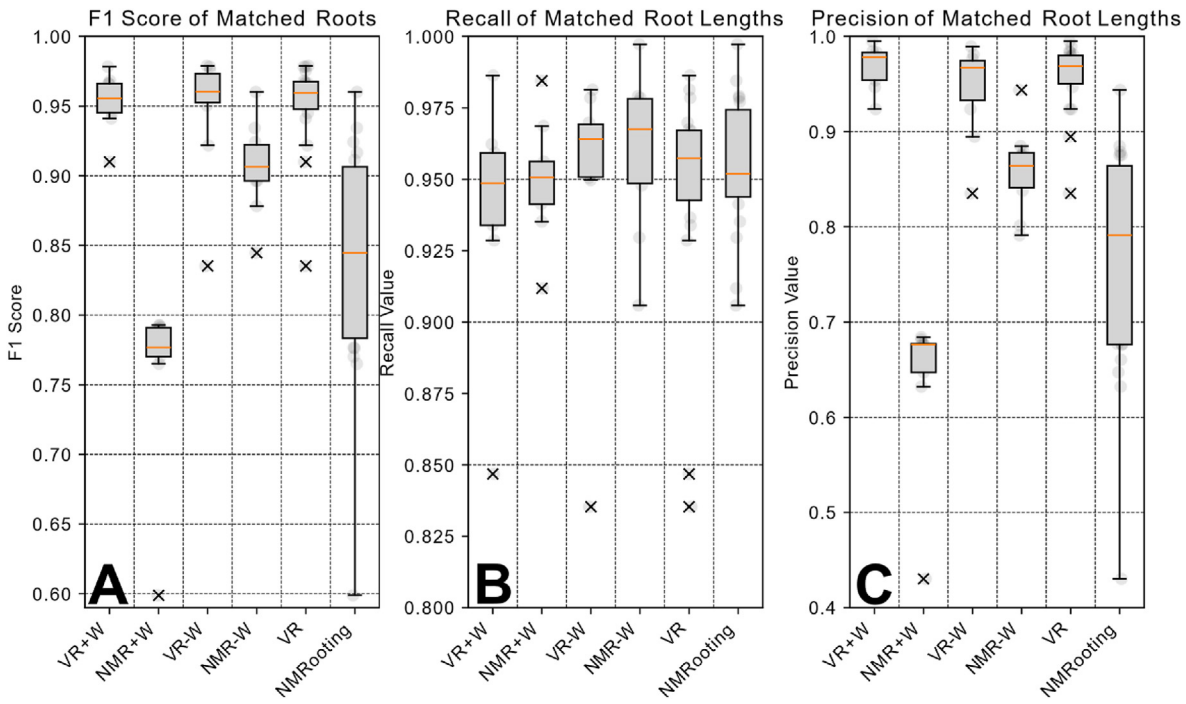

Fig. 5. Boxplots showing the median (orange) and distribution of the data with outliers marked as x. A: F_1 scores of the extracted root systems. B: Recall value of the matched root systems. C: Precision score of the matched root system.

3.2. Results regarding user study hypotheses

We summarized the statistics and hypothesis data in Table 1, which includes the most important data points. This section will give a brief overview of what results we measured on the hypotheses we had described in Sec. 2.2, and further indicators regarding ground truth comparison.

SUS: We observe an average usability of 82 for VR and 67 for NMRooting. VRoot has scores significantly better than NMRooting as measured using a one-sided *t*-test (H1). Notably, VR + W has a average usability of 86 and NMR + W of 62, with tests showing significant improvement by using VR, which is not the case for the comparison VR-W > NMR-W.

UEQ: Pragmatic experience that can be extracted from the UEQ is the average of the three statistics *Perspicuity*, *Efficiency*, *Dependability*. VR scored 1.841 on average and NMRooting scored 0.912, resulting in a significant difference (H2).

Root Length: The length measures of the root system are average and total root length, seen in Fig. 6A and B respectively. We computed the differences in extraction quality regarding these two measures, allowing for an assessment of overall extracted biomass and the correct identification of individual roots. The average and total root length contain useful information about what challenges participants encountered during the annotation. We observe a difference between NMRooting and VR in terms of extracted root lengths (H3). Deviation from ground truth was tested as well as the difference between the methods. The extraction of the average root length was, on average, correctly done in NMR-W, by a single-sample *t*-test ($t(9) = -0.979, p = .353$). The total root length extraction in VR-W yielded no significant difference to ground truth, yielding $t(9) = -0.426, p = .680$. Other conditions, including aggregates, yield significant differences. Differences between the applications exist in the case of the individual conditions (+W/-W) as well as the aggregate. We furthermore observe that the extraction of the total root length yields larger differences in the +W condition.

Root Topology: Computing the extracted number of laterals that have a minimum length of 3 cm, we find that only NMR-W found the correct number of lateral roots. For the branching density (H4), we observe a difference between the extracted number of lateral roots in the +W case, but not in the -W case nor in the aggregate case.

***F*$_1$:** We find that the F_1 score of the extracted root systems is significantly higher for VR in both + W and -W conditions (H5). Additionally, we find an increase in the difference between the applications once water noise is present (H6).

Influence of Water Noise: We observe a significant increase in the difference between the applications when water noise is present. This is the case for the average root length, the total root length, and the F_1 score. The computation for the F_1 score test required random pairing between individual scores.

Inter-Lateral Distance: This measure is defined as the average distance between two consecutive lateral organs O_i, O_j. VRoot users correctly extracted the inter-lateral distance, tested using a two-sided Mann-Whitney-U Test. NMRooting yielded a significant difference to ground truth in the +W condition, by single-sample *t*-test with $t(8) = 1.619, p = .002$ and an effect size of $d = 1.575$. There were no significant differences in both VR cases as well as the NMR-W case.

4. Discussion

We postulated that the VR software would yield a different usability, which has been confirmed in H1 for W+ and overall use of the software. The effect was much smaller for W-, resulting in a very small effect size and no significant difference. However, while no significant difference was measured, there is furthermore no indication that classical applications perform better in any of the study conditions. The overall measured variance for the measurement of pragmatic quality (H2) was fairly high, see Fig. 4B, resulting in no significant difference between the applications in the W+ condition.

Objective measurements were more uniformly successful, with the notable exception being the detection of the correct number of lateral roots. In that metric, all median extractions were below ground truth, with the exception of NMR-W, which was slightly higher. Interestingly, NMR + W extractions consist of less roots overall, which might be a result of the noisy data inhibiting the participant's ability to extract roots. We confirm other findings, such as by Selzner et al. [12] and Horn et al. [11]

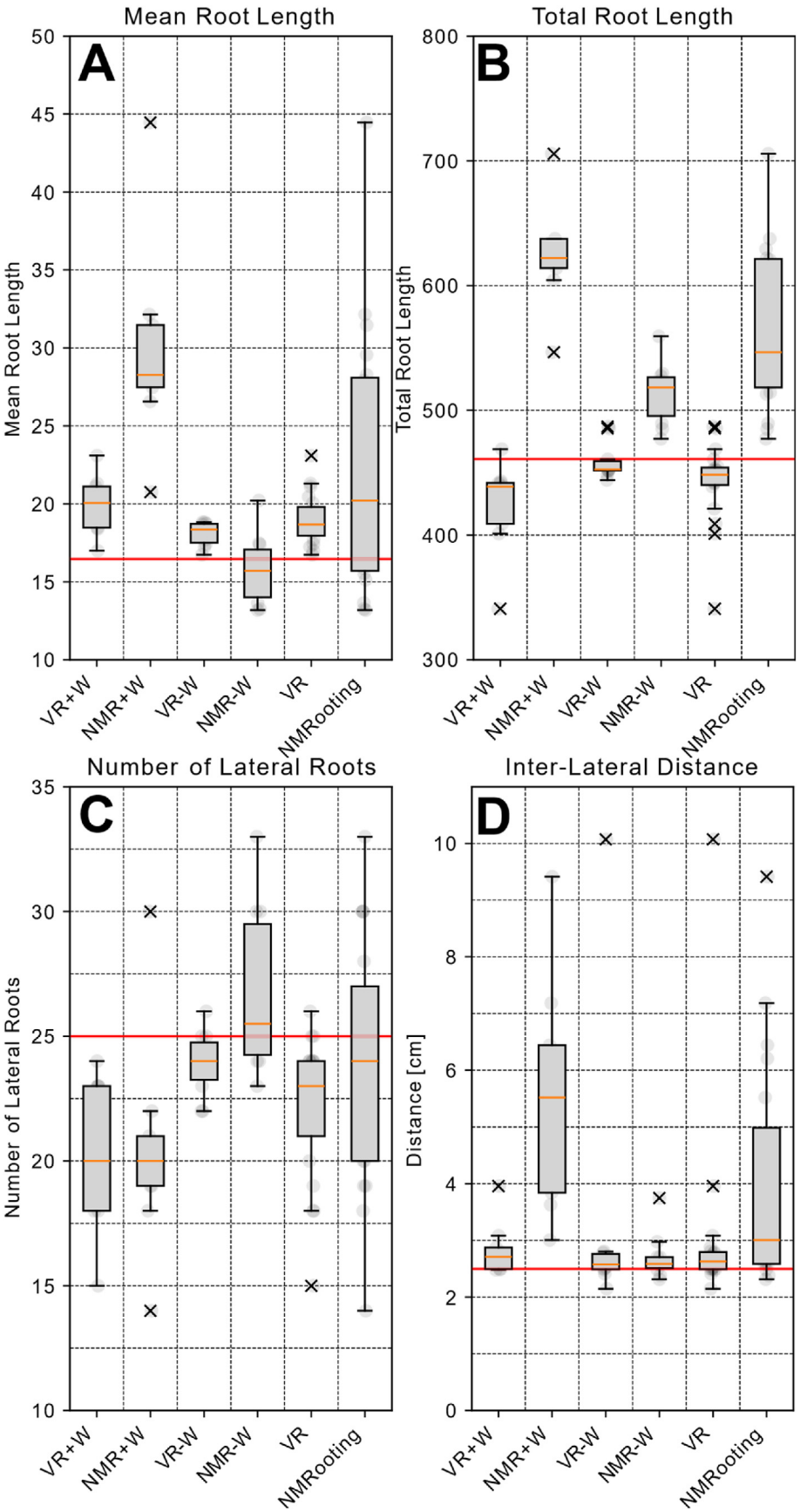

Fig. 6. Box-plot graphs show ground truth in red, median in orange, and outliers as x. A: Boxplots of average root length B: Boxplots of total root length $\sum l_i$ C: Number of lateral roots ($l_i \leq 4$ cm) D: Inter-lateral distance d_i.

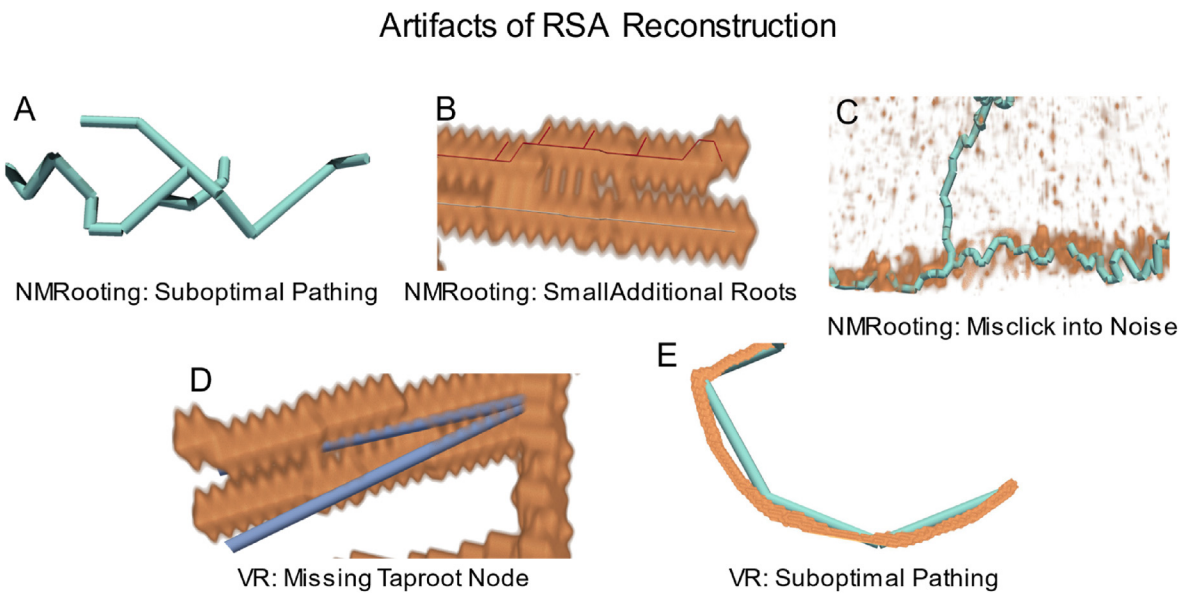

Fig. 7. Artifacts of RSA reconstruction in each application that occurred with participants.

about the impact of noise on the RSA reconstruction.

4.1. Task execution

Generally, task execution posed no problems for users. Users needed a few minutes to get accustomed to using the HMDs. Though pre-emptive measurement and calibration were done for the interpupillary distance, some users reported issues with depth perception, or depth-perception issues became apparent during the training phase. Though explicit introduction and prompting to repeat a certain interaction were done during the training phase, some users did not make use of all available options, particularly navigation, during task completion. This occurred in equal parts with NMRooting and VRoot.

4.2. Interpretation of results

The simulated root contained certain artifacts that would have made it fairly hard to trace for new users. Particularly, we note that the matching-based F_1 score is remarkably good for the VR software, which is in part due to users matching the correct root length for the individual organs successfully. The spatial distribution of the root system is more obvious in VR, which reflects in the root systems that were drawn, even under the condition that users do not edit root nodes that were already placed. The restricted set of functionalities was mitigated by the introduction of a training session, during which the procedure was explained, leading to most participants already pre-planning the taproot in a way that made it possible to achieve a high accuracy.

One aspect of the results we would like to highlight is the fact that with no water noise, the VR application still yielded better results in terms of matched length-based F_1 scoring, see Fig. 5A. Furthermore, the differences between the individuals were not very high, resulting in a standard deviation for the VR conditions of 0.02 (W+), 0.04 (W-), and 0.03 (T) respectively. However, this is partially due to the fairly 'destructive' nature of the F_1 score, leading to different user-based tracing errors to result in a similar score. The recall value R, as seen in Fig. 5, is fairly uniform across applications, with slightly more "under-tracing" done in the VR application. The F_1 score and particularly the precision P was low (but not extremely so) for NMRooting in the +W condition. The primary reason for this was the fact that each time a user clicked into the data and did not continue a root from the tip but rather from the closest segment regarding the signal strength, this induced a small lateral root at that point that users might have missed. For the actual comparison of the number of lateral roots, we discarded shorter roots to avoid comparing against this technicality. We note that the F_1 scores closely follow the power law distribution, tested by means of Cressie-Read test for goodness-of-fit at $\chi = 0.021$ and $p = 1.0$.

There is a significantly higher spread in the perceived system usability measured from NMRooting in the +W condition as well as overall. Tested variances were significantly higher according to the F-test for the overall condition ($F(18) = 4.330, p = .002$) as well as + W ($F(8) = 11.493, p = .001$) but not for the conditions -W ($F(9) = 2.571, p = .087$), even though this condition does fail the test for equal variance ($F(9) = 2.571, p = .175$). Between the VR + W and VR-W conditions, there is an increase in variance that is, while not significant, at least notable. Moreover, the inter-lateral distance has a large variance in the NMR + W condition due to several artifacts. We will note that, for the estimation of the inter-lateral distance, while there were no significant differences in VR + W, VR-W, and NMR-W, that most users (68% for VR and 84% for NMRooting) over-estimated the inter-lateral distance compared to the ground truth. We cannot make assumptions on the applicability of this regarding a real MRI scan, but our findings provide some insight into user bias when extracting parameters, particularly for FSPM simulation, from 3D imaging data. The average relative error for inter-lateral distance for VR users was 0.234 and for NMRooting it was 0.617.

The average and total root length is, in part, influenced by the presence of water noise, in both applications. NMR + W exhibited a larger total root length while still yielding a lower than ground truth number of laterals. However, the increased inter-lateral distance partially relates to misidentification later in the data. In VR, the presence of water noise caused under-identification of roots. We will note, that in the simulation data, there was one very thin root that would have been very hard to identify. There was a systematic issue with users not being able to identify that root and thus, the VR-W case was lower in median regarding the number of lateral roots users were able to identify.

4.3. Artifacts of the RSA reconstruction

Our general results show an improvement in the extraction quality using VR. More specific phenomena can often be explained taking into account observations from the study or by closer inspection of the data. There are a few instances of false positives within the NMRooting annotation that can be attributed to users clicking on surfaces they did not intend to. It is important to note that during the eventual study task, no further assistance was provided unless prompted, as opposed to the

Table 1
Hypotheses tests, reported with statistic T/U, critical value p, effect size d and degree of freedom (dof). We indicated what statistical tests were used by their abbreviation, namely U for Mann-Whitney U-Statistical Test and T for T-Test.

	W-				**W+**			
H1 SUS (T)	t	p	d	dof	t	p	d	dof
	0.716	.491	0.245	9	2.526	.017	1.249	8
	T							
	t		p		d		dof	
	5.112		< .001		0.812		18	
	W-				**W+**			
H2 UEQ (T)	t	p	d	dof	t	p	d	dof
	2.442	.017	1.019	10	1.808	.054	0.987	8
	T							
	t		p		d		dof	
	0		< .001		-2.859		9	
	W-				**W+**			
H3 ∑ L (U)	t	p	d	dof	t	p	d	dof
	0	< .001	-2.859	9	4	< .001	-2.537	8
	T							
	t		p		d		dof	
	4		< .001		-1.654		18	
	W-				**W+**			
H4 |O^1| (U)	t	p	d	dof	t	p	d	dof
	44	.789	0	8	22	.034	-1.223	9
	T							
	t		p		d		dof	
	148.5		.355		-0.386		18	
	W-				**W+**			
H5 F1 (U)	t	p	d	dof	t	p	d	dof
	84	.011	1.179	9	81	< .001	4.426	8
	T							
	t		p		d		dof	
	336		< .001		1.749		18	
	W+ and W-							
H6 F1 (U)	U		P		d		dof	
	89.821		< .001		2.494		9	
	W-				**W+**			
H7 ∑L (U)	t	p	d	dof	t	p	d	dof
	-7.898	< .001	-2.859	18	-5.239	< .001	-2.537	16
	T							
	t		p		d		dof	
	-5.021		< .001		1.749		36	
	W+ and W-							
H8 d_ij (U)	U		P		d		dof	
	13		.7		0.346		18	

training task, which included guidance and repetition until no mistake was made.

This effect was exasperated by the presence of water noise. Some users corrected their annotation to a certain degree, resulting in fewer false positives but still suboptimal pathing, as seen in Fig. 7A. In contrast, a case of supoptimal pathing in VR, which directly results from a coarser user interaction, is seen in Fig. 7E. A total of 5 users attempted to separate the proximal roots, which yielded a few perfect annotations, but also artifacts such as Fig. 7B, which includes a few small roots that are due to participants progressing the lateral by clicking in smaller steps, such that the algorithm does not use the connecting signal between the roots to path to the point indicated by the user. If water noise was present, a few users misclicked into the water volume during the study task but failed to remove such a tracing at a later point, as seen in Fig. 7C.

While the VR application had better F_1 scores on average, ultimately more training is required for users to produce high-quality RSA reconstructions. Some participants struggled with depth perception in VR, which occurred in approximately equal parts in people with corrected vision and normal vision. This is likely an experience effect that would only be resolved by further use of the VR system - participants with previous VR experience did not encounter this. Targeting objects in the virtual scene as well as the correct placement of segments was challenging for new users.

The generally low variance of individual scores might be an indicator of a systematic effect that is uniform among individuals. In the case of VR, the inability to edit nodes was likely a large contributor to the overall score of participants. This caused issues in cases where there was a node missing in the taproot, as exemplified in Fig. 7D. In the case of

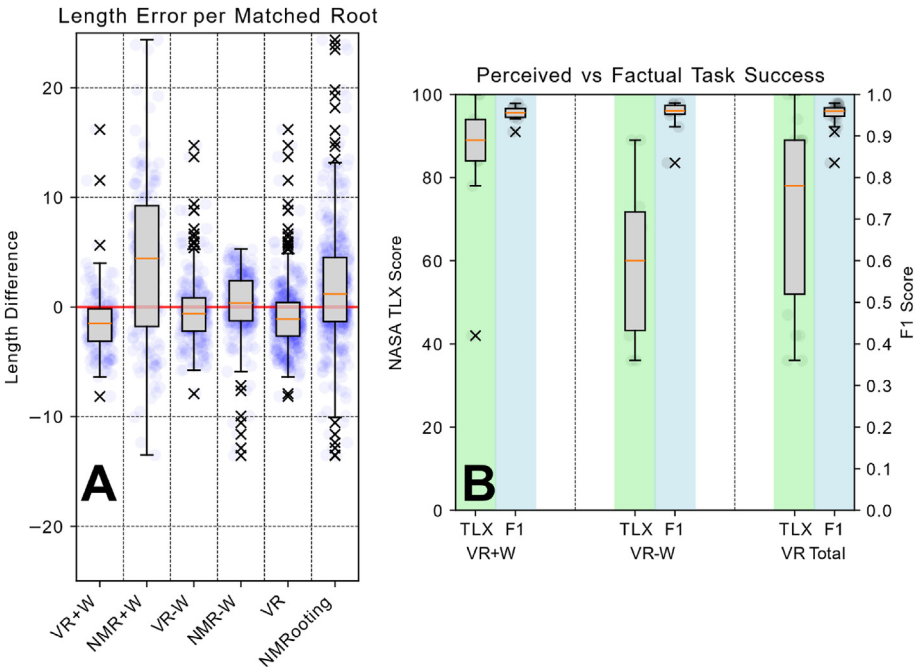

Fig. 8. A: Box-plots of length differences between ground truth and correctly annotated roots. Ground truth is red, median is orange, raw data is indicated blue, outliers are marked as x. B: Comparison of self-assessment of extraction quality in VR to actual extraction quality.

NMRooting, participants were told to manually trace the nodes and were able to delete nodes that were mistakenly inserted. The low score in NMRooting is mostly due to smaller effects accumulating to a lower score in total, including suboptimal pathing, misidentification of roots, as well as the presence of proximal roots and water noise in certain conditions.

4.4. Explanations for length differences

VR users tended to estimate the ground truth total root length correctly when dealing with noise-free data. However, there is a slight indication of overestimation of the individual root length within noisy data, as found in Fig. 6. VR users might not have drawn the taproot correctly, resulting in a slight overestimation of average root lengths, but still an underestimation of the total. It has to be noted that the task was performed sitting, which meant that users were forced to use the navigation in VRoot as an alternative to bending down. Some users chose to trace the root system less effectively as parts of it were out of reach, resulting in a higher variance of the total length measurements. A few users requested to be able to stand, but for the sake of uniformity, we did not allow this.

On the other hand, the large overestimation in average root length in NMRooting was in part due to supoptimal pathing, while the overestimation in total root length was caused by false-positive lateral roots, which is indicated by the shift in distribution in the inter-lateral distance, as seen in Fig. 6. We further investigate this by filtering the roots for only true positive identifications and subsequently compute their length difference, as shown in Fig. 8. While we do observe a higher-than- zero length extraction, by means of double-sided t-test with $t(198) = 2.379$, $p = .018$, $d = 0.169$, we observe a significantly higher variance in the +W condition of the desktop software ($F(192) = 0.651$ and $p = .001$). We believe that a combination of issues, most notably inability to effectively navigate in a desktop setting, caused the differences in the +W condition.

4.5. Subjective measures

The quantification of responses to the questionnaires yielded mixed results. Particularly, we want to highlight the increased usability through the use of VRoot. While there was no significant increase in the no-water condition, as seen in Fig. 4, there was a larger spread of responses for the NMRooting software, resulting in a higher average usability score of using VR in comparison. NMRooting generally had a larger spread in responses in contrast to the more unified responses for VR both in total ($F(18) = 4.330$ and $p = .002$) and in the +W condition ($F(8) = 11.493$ and $p = .002$). The TLXs yielded no significant difference between NMRooting and VRoot, but it did showcase a higher variance in the -W cases.

The user experience is challenging to compare, as certain measures (such as novelty) are not appropriate in the assessment of the desktop software. This is especially true since participants will already have the expectation of using VR software even if using the desktop software first, leading to influences measurements of those metrics.

The task, while the root was generally simple with only a taproot and 25 laterals, there were particularities about the data that caused issues for certain participants, especially when using NM- Rooting. However, artifacts like proximal roots, or smaller roots further down the taproot, have gone unnoticed to some participants. We will note that there was no difference in the individual ratings depending on the order of applications tested. Generally, it is quite natural that the tasks would appear equally demanding between the individual conditions. Interestingly, the VR + W condition has the smallest variance, indicating a more universal agreement even between the within-subject conditions.

One aspect of the TLXs questionnaire is the question on the self-assessment on whether the task was completed. This assessment is shown in Fig. 8B. The F_1 score shows that users performed similarly whether there was water noise present, or not. However, self-assessment of the accuracy was much lower than the actual accuracy with no water noise present. This is likely less a self-assessment and more a comparative assessment depending on how easy the problem appears to be solvable with automatic means. We highlight this to underline the issue that the subjective assessment of data extraction done by users is seldomly representative of the actual data quality. While our users estimated their own performance as worse than it actually was, this is a more general issue, as manually annotated public data sets also contain label errors, which generally are assumed to be perfectly labeled [41]. However, manual annotation work is incredibly valuable, especially in plant science. The human self-assessment of data quality measured through

manual means, especially if used as training basis, is seldomly accurate.

4.6. Usability of VR for annotation

Even if a virtual reality workflow improves the quality of extraction, manual tasks remain more tedious and time-consuming than automatic extraction. Our workflow is ultimately aimed at correcting rather than tracing. VRoot functions best with a mixture of high throughput pre-tracing and features that assist with the analysis of root architectures as the basis. In the future, we would like to combine automated, semi-automated, and manual tracing methods. In the case of NMRooting, this would require the introduction of additional interaction metaphors suitable for tasks that are voxel-based. There are other methods, such as TopoRoot [10], that could improve the manual extraction pipeline. This would be in line with published literature on similar topics, namely by Zeng et al. [17], who improved the VR application developed in Usher et al. [14] through the use of topological features.

Furthermore, there are technical aspects to consider. Geometry visualization is ultimately constrained by the physical device memory, which primarily includes GPU memory. While geometry reduction in terms of triangle merging can help, the complexity of the data through number of roots or noise will impact rendering performance. Resolution is also a factor, which is a device attribute of the MRI scanner. While we did not observe resolution-based effects that had clear origin, Selzner et al. [12] had already studied the effects of higher resolution (through upscaling) for manual annotation. Furthermore, using commonly available software such as Unreal Engine, it is possible to target a wider range of consumer-grade hardware, and the application is portable to different HMDs in principle.

5. Conclusion

In this work, we presented a pipeline to extract RSAs from MRI images using VR. We established the need for a more immersive manual analysis tool for complex data sets and showed the advantages of using VR to enable new users to achieve high-quality reconstructions faster. We evaluated the use of our VR software in comparison to contemporary desktop applications, and semi-automated analysis. Furthermore, we quantified how well participants with a uniform knowledge base performed in these tasks, both on the desktop as well as in VR. Our results show an increased usability and accuracy through the use of VR for manual root workflows, especially in instances where automatic tools need more assistance. This enables the analysis of root systems in more diverse soil conditions. Here, immersive annotation is a very valuable method in 3D root image analysis and helps to increase the variety of analyzed data to more soil types and soil water contents. In the future the goal is to combine the tools offered by our VR application with the benefits of an automatic extraction to provide a user-friendly and fast workflow to correct automatic tracing results. A repetition experiment specifically designed to estimate the actual errors made when manually extracting data would be needed to provide a robust quantification beyond relative comparison.

6. Ethical Statement

The study conception and planning got approval from the ethics committee of Trier University. Participants provided their explicit consent and were instructed on how to contact us to have their data deleted should they wish to do so. Data is stored and published anonymously.

Author contributions

This work was primarily authored by Baker. All authors discussed the results, provided feedback, and contributed to the text of this work. Study conception and execution was done by Baker and Zielasko. The implementation of the VR application and development of its features was done by Baker, Selzner, Göbbert, Zielasko, and Schnepf. Scientific counsel and directions for data analysis, study procedure, as well as results were provided by Selzner, Scharr, Riedel, Hvannberg, Schnepf, and Zielasko. Funding for hardware was provided by Schnepf and Göbbert. Zielasko has primarily supervised this project and guided implementation of its tasks.

Data availability

The data needed to reproduce the results of this work has been uploaded to 10.26165/JUELICH-DATA/B9SBOS. We have included a meta-description of the approach but encourage researchers aiming to reproduce our results to reach out. The Software VRoot will be found on Github: dhelmrich/VRoot. The questionnaires have been attached to the supplemental material. We have published a video description of the software, seen in 10.6084/m9.figshare.26003494.

Funding

The authors would like to acknowledge funding provided by the German government to the Gauss Centre for Supercomputing via the InHPC-DE project (01—H17001).

This work has partly been funded by the EUROCC2 project funded by the European High- Performance Computing Joint Undertaking (JU) and EU/EEA states under grant agreement No 101101903.

This work has partly been funded by the German Research Foundation under Germany's Excel-lence Strategy, EXC-2070 - 390732324 - PhenoRob and by the German Federal Ministry of Education and Research (BMBF) in the framework of the funding initiative Soil as a Sustainable Resource for the Bioeconomy BonaRes, the project BonaRes (Module A): Sustainable Subsoil Management - Soil3; subproject 3 (grant 031B1066C).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.plaphe.2025.100013.

A Normal Distribution Tests

The shapiro-wilk test for normal distribution tests whether the data is drawn from a normal distri- bution. This is a test where the alternative hypothesis is exhibiting a normal distribution, meaning that the H_0 hypothesis is that $X \sim N$ for some normal distribution N, referring to a significant difference to data produced by a normal distribution. For comparative measures, the residual of the two statistics, $R: = T_1 - T_2$, needs to be normally distributed, meaning that the difference between applications is tested as opposed to the statistics T of each application itself.

As seen in Tab. 4, there are some instances of data where a normal distribution is not present, which are the +W conditions for the root lengths,

resulting in a difference from a normal distribution for the total data. The inter-lateral distance d_i does not exhibit normal distribution for the -W condition.

These effects were likely caused by a non-uniform error that is present within the data, as opposed to the subjective measurements that are mostly centered around an average value, the objective distributions are non-symmetric regarding the differences between the applications.

Tab. 2
Goodness-of-fit tests for normal distribution on different measures, including combined measures. The pairing of residuals is always within subject.

Measure	Water		No Water		Total	
	t	p	t	p	t	p
SUS	0.902	.264	0.959	.783	0.944	.311
UEQ	0.939	.577	0.938	.497	0.897	.051
TLXs	0.933	.515	0.953	.713	0.976	.897
TRL	0.798	.019	0.870	.101	0.744	< .001
ARL	0.798	.019	0.870	.101	0.744	< .001
F_1	0.722	.002	0.931	.462	0.909	.072
d_i	0.915	.357	0.490	< .001	0.901	.051

Tab. 3
Goodness-of-fit tests for individual statistics for the testing against ground truth measure-ments, for each VR condition.

Measure	VR+W		VR-W		VR	
	t	p	t	p	t	p
SUS	0.902	.264	0.959	.783	0.944	.311
UEQ	0.932	.504	0.961	.790	0.953	.074
TLXs	0.877	.157	0.907	.259	0.913	.074
TLR	0.679	.001	0.777	.011	0.719	< .001
ARL	0.986	.988	0.879	.129	0.928	.159
F_1	0.908	.308	0.686	.001	0.720	< .001
d_i	0.756	.006	0.443	< .001	0.404	< .001

Tab. 4
Goodness-of-fit tests for individual statistics for the testing against ground truth measure-ments, for each NMRooting condition.

Measure	NMR+W		NMR-W		NMR	
	t	p	t	p	t	p
SUS	0.963	.824	0.884	.146	0.931	.160
UEQ	0.920	.394	0.918	.302	0.944	.289
TLXs	0.839	.043	0.924	.390	0.882	.019
TLR	0.744	.007	0.814	.014	0.714	< .001
ARL	0.930	.453	0.854	.083	0.885	.026
F_1	0.539	.007	0.979	< .001	0.890	.032
d_i	0.949	.687	0.771	.006	0.807	.001

B Study Procedure

In the questionnaire for the study, participants were informed about what data would be stored and how. Participants gave informed consent for the participation in the study. Participants were assigned an ID at the start of the study, and we measured their inter-pupillar distance for the purposes of calibrating the HMD to their respective measurements. Depending on the ID, participants then used either the application VRoot or NMRooting to extract a root system. For both applications, participants were verbally explained the functionality of the applications and what features they were to use to extract the root system. Each participant was tasked with performing all actions on a training data set that were required later on, which included navigation, tracing, and deletion in the case of NMRooting. After task completion with the study task, during which no intervention took place, participants filled out the questionnaires attached in the supplemental material. Participants will then repeat the process with the other application.

The study aimed to assess the direct interaction with the system in cases in which manual annotation was necessary. As such, participants were told to only use these tools to ensure comparability, as otherwise the task would be more complex, and user interaction would require more functionality and a deeper understanding of the application. In the VR application, participants were taught the draw root functionality as well as data set navigation techniques to ensure that all of the data set is reachable. Successful completion of the VR training task required the following steps.

1. Move head position in VR
2. Rotate the root system

3. Move the root system up/down
4. Select a node
5. Deselect a node
6. Draw a node
7. Deselect a node and draw a lateral

In NMRooting, we restricted the functionality to the tip-to-tree option, with the correct starting point already being set pre-emptively. Users did label the root topology manually. Successful completion of the NMRooting training phase required the following steps.

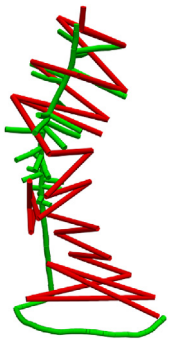

Fig. 9. Removed outlier data set after subject failed to trace the taproot. While the cause of this issue is unknown, the participant did trace a fully functional root system in the training stage of the task, which included being prompted to place nodes on the taproot of the root system.

1. Turn the view
2. Zoom in/out
3. Pan the view (lateral movement respective to the screen view)
4. Click into the data to add a root
5. Click into the data to delete a root
6. Open the root order menu
7. Label roots by their root order (here just taproot and first order lateral)

Our previous tests yielded an average completion time for the tasks of 45 min. From a previous test with the application with mixed, i.e., expert and non-expert, participants ($n = 16$), we concluded that 1 h would be sufficient for untrained participants. While we did not provide a timing nor a time for task completion in the individual conditions, we did allocate time slots for participants which were an hour long.

C Outlier

The outlier data set, exemplified by the VR performance of the participant. While the participant was able to annotate both the taproot as well as annotate lateral roots in the training task, the participant has not retained the knowledge of the annotation steps and produced an annotation (red) as shown in Fig. 9. It is unclear whether this was caused by the presence of water noise.

References

[1] J.A. Atkinson, M.P. Pound, M.J. Bennett, D.M. Wells, Uncovering the hidden half of plants using new advances in root phenotyping, Curr. Opin. Biotechnol. 55 (2019) 1–8, https://doi.org/10.1016/j.copbio.2018.06.002. Analytical Biotechnology.

[2] E.D. Rogers, P.N. Benfey, Regulation of plant root system architecture: implications for crop advancement, Curr. Opin. Biotechnol. 32 (2015), https://doi.org/10.1016/j.copbio.2014.11.015. Food Biotechnology • Plant Biotechnology:93–8.

[3] E.L. Fry, A.L. Evans, C.J. Sturrock, J.M. Bullock, R.D. Bardgett, Root architecture governs plasticity in response to drought, Plant Soil 433 (2018) 189–200, https://doi.org/10.1007/s11104-018-3824-1.

[4] T.H. Nguyen, T. Gaiser, J. Vanderborght, et al., Responses of field-grown maize to different soil types, water regimes, and contrasting vapor pressure deficit, EGUsphere 2024 (2024) 1–53, https://doi.org/10.5194/egusphere-2023-2967.

[5] M. Giraud, S.L. Gall, M. Harings, et al., CPlantBox: a fully coupled modelling platform for the water and carbon fluxes in the soil–plant–atmosphere continuum, silico Plants 5 (2023) diad009, https://doi.org/10.1093/insilicoplants/diad009.

[6] F.M. Bauer, L. Lärm, S. Morandage, et al., Development and Validation of a deep learning based automated minirhizotron image analysis pipeline, Plant Phenomics 2022 (2022), https://doi.org/10.34133/2022/9758532.

[7] D van Dusschoten, R. Metzner, J. Kochs, et al., Quantitative 3D analysis of plant roots growing in soil using magnetic resonance imaging, Plant physiology 170 (2016) 1176–1188, https://doi.org/10.1104/pp.15.01388.

[8] T. Tuomainen, A. Toljamo, H. Kokko, M. Nissi, Non-invasive assessment and visualization of Phytophthora cactorum infection in strawberry crowns using quantitative magnetic resonance imaging, Sci. Rep. 14 (2024), https://doi.org/10.1038/s41598-024-52520-7.

[9] X.R. Zhou, A. Schnepf, J. Vanderborght, D. Leitner, H. Vereecken, G. Lobet, Phloem anatomy restricts root system architecture development: theoretical clues from in silico experiments, silico Plants 5 (2023) diad012, https://doi.org/10.1093/insilicoplants/diad012.

[10] D. Zeng, M. Li, N. Jiang, et al., TopoRoot: a method for computing hierarchy and fine-grained traits of maize roots from 3D imaging, Plant Methods 17 (2021) Zeng2021, https://doi.org/10.1186/s13007-021-00829-z.

[11] J. Horn, Y. Zhao, N. Wandel, M. Landl, A. Schnepf, S. Behnke, Robust skeletonization for plant root structure reconstruction from MRI, in: 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 10689–10696, https://doi.org/10.1109/ICPR48806.2021.9413045.

[12] T. Selzner, J. Horn, M. Landl, et al., 3D U-net segmentation improves root system recon- struction from 3D MRI images in automated and manual virtual reality work flows, Plant Phenomics 5 (2023) 76, https://doi.org/10.34133/plantphenomics.0076.

[13] L. Stingaciu, H. Schulz, A. Pohlmeier, et al., In situ root system architecture extraction from magnetic resonance imaging for water uptake modeling, Vadose zone journal 12 (2013) 1–9, https://doi.org/10.2136/vzj2012.0019.

[14] W. Usher, P. Klacansky, F. Federer, et al., A virtual reality visualization tool for neuron tracing, IEEE Trans. Visual. Comput. Graph. 24 (2018) 994–1003, https://doi.org/10.1109/TVCG.2017.2744079.

[15] T. Chandler, M. Cordeil, T. Czauderna, et al., Immersive Analytics, in: 2015 Big Data Visual Analytics (BDVA), 2015, pp. 1–8, https://doi.org/10.1109/BDVA.2015.7314296.

[16] A. Fonnet, Y. Prié, Survey of immersive Analytics, IEEE Trans. Visual. Comput. Graph. 27 (2021) 2101–2122, https://doi.org/10.1109/TVCG.2019.2929033.

[17] T. McDonald, W. Usher, N. Morrical, et al., Improving the usability of virtual reality neu- ron tracing with topological elements, IEEE Trans. Visual. Comput. Graph. (2020), https://doi.org/10.1109/TVCG.2020.3030363.

[18] D. Zielasko, B. Weyers, T.W. Kuhlen, Travel your desk? An office desk substitution and its effects on cybersickness, presence and performance in an HMD-based exploratory anal- ysis task, in: 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), IEEE, 2019, pp. 1285–1286, https://doi.org/10.1109/VR.2019.8798068.

[19] Alex Vlachos, Advanced VR rendering. Game Developers Conference, 2015.

[20] H. Adams, J. Stefanucci, S. Creem-Regehr, B. Bodenheimer, Depth perception in augmented reality: the effects of display, shadow, and position, in: 2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), 2022, pp. 792–801, https://doi.org/10.1109/VR51125.2022.00101.
[21] S. Baceviciute, T. Terkildsen, G. Makransky, Remediating learning from non-immersive to immersive media: using EEG to investigate the effects of environmental embeddedness on reading in Virtual Reality, Comput. Educ. 164 (2021) 104122, https://doi.org/10.1016/j.compedu.2020.104122.
[22] R.F. Adler, R. Benbunan-Fich, The effects of task difficulty and multitasking on perfor- mance, Interact. Comput. 27 (2014) 430–439, https://doi.org/10.1093/iwc/iwu005.
[23] S. Mystakidis, J. Besharat, G. Papantzikos, et al., Design, development, and evaluation of a vir- tual reality serious game for school fire preparedness training, Educ. Sci. 12 (2022), https://doi.org/10.3390/educsci12040281.
[24] J.R. Lewis, J. Sauro, The factor structure of the system usability scale, in: International Conference on Human Centered Design, Springer, 2009, pp. 94–103, https://doi.org/10.1007/978-3-642-02806-9_12.
[25] P.T.K. Aaron Bangor, J.T. Miller, An empirical evaluation of the system usability scale, ternational Journal of Human–Computer Interaction 24 (2008) 574–594, https://doi.org/10.1080/10447310802205776.
[26] D. Pflugfelder, R. Metzner, D van Dusschoten, R. Reichel, S. Jahnke, R. Koller, Non-invasive imaging of plant roots in different soils using magnetic resonance imaging (MRI), Plant Methods 13 (2017) 102, https://doi.org/10.1186/s13007-017-0252-9.
[27] S le Gall, D van Duschoten, A. Lattacher, et al., Investigating the impact on water fluxes and physiological development of the combination of contrasted root wheat cultivars from isotopic analysis, in: EGU General Assembly 2024, 2024, https://doi.org/10.5194/egusphere-egu24-15966.
[28] R. Metzner, A. Eggert, D. Dusschoten, et al., Direct comparison of MRI and X-ray CT technologies for 3D imaging of root systems in soil: potential and challenges for root trait quantification, Plant Methods 11 (2015) 17, https://doi.org/10.1186/s13007-015-0060-z.
[29] Y. Zhao, N. Wandel, M. Landl, A. Schnepf, S. Behnke, 3D U-Net for Segmentation of Plant Root MRI Images in Super-resolution, 2020, https://doi.org/10.48550/arXiv.2002.09317.
[30] A.O. Uzman, J. Horn, S. Behnke, Learning Super-resolution 3D Segmentation of Plant Root Mri Images from Few Examples, 2019, https://doi.org/10.48550/arXiv.1903.06855 arXiv preprint arXiv:1903.06855.
[31] G. Lobet, M.P. Pound, J. Diener, et al., Root system Markup Language: toward a unified root architecture description language, Plant Physiology 167 (2015) 617–627, https://doi.org/10.1104/pp.114.253625.
[32] D. Zielasko, B.E. Riecke, Sitting or standing in VR: about comfort, conflicts, and hazards, IEEE Computer Graphics and Applications 44 (2024) 81–88, https://doi.org/10.1109/MCG.2024.3352349.
[33] W. Schroeder, K. Martin, W. Lorensen, The design and implementation of an object-oriented toolkit for 3D graphics and visualization, in: Proceedings of Seventh Annual IEEE Visualization '96, 1996, pp. 93–100, https://doi.org/10.1109/VISUAL.1996.567752.
[34] D.N. Baker, F.M. Bauer, M. Giraud, et al., A scalable pipeline to create synthetic datasets from functional–structural plant models for deep learning, silico Plants 6 (2023) diad022, https://doi.org/10.1093/insilicoplants/diad022.
[35] B. Laugwitz, T. Held, M. Schrepp, Construction and evaluation of a user experience ques- tionnaire, in: Holzinger A. Berlin (Ed.), HCI and Usability for Education and Work, Springer Berlin Heidelberg, Heidelberg, 2008, pp. 63–76, https://doi.org/10.1007/978-3-540-89350-9_6.
[36] S.G. Hart, L.E. Staveland, Development of NASA-TLX (task Load index): results of empiri- cal and theoretical research, in: P.A. Hancock, N. Vol 52 Meshkati (Eds.), Human Mental Workload, Advances in Psychology, North-Holland, 1988, pp. 139–183, https://doi.org/10.1016/S0166-4115(08)62386-9.
[37] D. Zielasko, B.E. Riecke, To sit or not to sit in VR: analyzing influences and (Dis) Advantages of posture and embodied interaction, Computers 10 (2021), https://doi.org/10.3390/computers10060073.
[38] A. Schnepf, K. Huber, M. Landl, F. Meunier, L. Petrich, V. Schmidt, Statistical characteriza- tion of the root system architecture model CRootBox, Vadose Zone J. 17 (2018) 170212, https://doi.org/10.2136/vzj2017.12.0212.
[39] J.R. Lewis, The system usability scale: past, present, and future, Int. J. Hum. Comput. Interact. 34 (2018) 577–590, https://doi.org/10.1080/10447318.2018.1455307.
[40] J. Cohen, A power primer, Psychol. Bull. 112 (1992) 155–159, https://doi.org/10.1037/0033-2909.112.1.155.
[41] C.G. Northcutt, M. ChipBrain, C.A. Athalye, J. Mueller, Pervasive label errors in test sets destabilize machine learning benchmarks, in: 35th Conference on Neural Information Processing Systems (NeurIPS 2021), 2021, https://doi.org/10.48550/arXiv.2103.14749.