A comprehensive dataset for monitoring germination of *Cannabis sativa* in greenhouse-controlled environments

José A. Brenes [a], Ana Codes [b], Javier Ferrández [c], Carmen Rocamora [b, *]

[a] CITIC-ECCI-PPCI, University of Costa Rica, San José, Costa Rica

[b] CIAGRO, Miguel Hernández University, Orihuela, Spain

[c] I2RC, University of Alicante, Alicante, Spain

* Corresponding author. Email: rocamora@umh.es

Abstract

Germination monitoring in crucial: it aids in planning and maximizing crop yields, contributing to sustainable agriculture practices. Furthermore, in the cultivation of *Cannabis sativa* L., precise control over the number of plants during the germination process is critical due to legal and regulatory restrictions. To assist growers in tracking germination progress, we planned to developed a classifier that informs them precisely when a plant has successfully germinated in the seedbed. A well-curated dataset is essential for teaching the algorithm to recognize different features. The dataset should encompass a diverse range of examples to ensure that the classifier can accurately identify and categorize instances it encounters during its operation. The quality and diversity of the dataset play a pivotal role in the performance and reliability of the developed classifier. Due to the limited information available on cannabis crops, we undertook the task of constructing an image dataset from the ground up. This dataset was meticulously crafted through a rigorous process. To build the dataset, we sow two varieties of cannabis, Finola and Kompolti, in separate seed trays, each containing 72 cells, in a greenhouse. A camera fixed above the seedbeds took one image every hour throughout the germination experiment. Four iterations were carried out: two without controlling climatic conditions, and two with controlled conditions and photoperiod. Then, we cropped the images and applied a homography process to correct perspective. The resulting images of each cell for each date and hour were labelled using six categories. first, the images were labelled considering the categories germination, non-germination, only cotyledon, and true leaves, then, they were labelled again using the categories invasion and non-invasion. The result was a comprehensive dataset comprising 80,640 images of seedbed cells showcasing plants at various growth stages. This paper outlines the step-by-step process employed in creating this image dataset.

Keywords: Cannabis sativa, germination, artificial vision.

1. Introduction

Seed is the most important input in agriculture, thus the need to ensure viability (Basu, 1995) and vigour (Dornbos, 1995). Seed vigour refers to both the ability and strength of a seed to germinate successfully and establish a normal seedling. Germination and seedling development monitoring is very time consuming and requires well trained technicians, resulting in very high personnel costs. Besides, classification process is not uniform and depends on skills and personal circumstances, which justifies the attempts to automate these tasks (Ureña et al., 2001).

Systems for monitoring seed germination based on artificial vision have been developed. Some of these systems are based on images of the seeds germinated in an incubator or in Petri dishes, recorded with a fixed frequency, in which seed evolution was analysed. Ducournau et al. (2004) designed a machine vision system to count the number of emergent radicle tips in seed lots under controlled lighting, temperature, and hygrometric conditions, performing continuous recording of sunflower seed germination at a frequency of one shot per hour. Dell'Aquila (2005) used a CCD camera to study broccoli and radish germination, monitoring the extent of imbibition phases through the assessment of seed area increase and timing of radicle emergence detected in individual seeds. Li et al. (2015) used image analysis to calculate the changes in seed length. Changes in seed area and length towards radicle emergence over time were used to identify the onset of germination.

Ureña et al. (2001) developed an automatic system for monitoring lettuce, cauliflower and tomato seeds germination based on the automatic measurement of leaf area, followed by classification of these measurements within a fuzzy logic-based framework. A mobile high resolution colour CCD camera captured

the image and the trays were placed manually. After image acquisition and pre-processing, automatic calculation of the leaf area of cotyledons and leaves was performed by using image processing algorithms: segmentation to obtain separate seedlings and description using leaf area. Seedling detection was based on the colour difference between the seedlings (green) and the trays' rooting media. A pixel was considered to be a green one if its proportion of green component was greater than that of its red and blue components. Fuzzy logic and expert knowledge was used to classify the degree of development of the seedlings. Since they had to work in realistic conditions, they captured images around solar midday trying to minimize the influence of light on the shadows and the colour of the images.

With the development of transplanting robots, machine vision systems to detect seedlings have been developed, based on leaf area. Artificial vision is used to identify and locate empty positions in the seedling tray. Ryu et al. (2001) developed a robotic transplanter that used a vision system to detect empty cells in high-density plug trays by counting and comparing the number of white pixels representing the seedling leaves with a predefined value. The vision system was also used to determine the leaf direction of the seedlings to minimise damage to them.

More recently, a vision system was used to measure the leaf area in each cell to distinguish "bad" and "good" plugs, achieving relative identification accuracies of seedling quality of 98.6%, 96.4%, 98.6% and 95.2% for tomato, cucumber, aubergine and pepper, respectively (Tong et al., 2013). An algorithm was developed that could segment overlapping leaves from neighbouring cells; additionally, an algorithm was developed to calculate the leaf area including intruding leaves.

Jin et al. (2022) proposed a low-damage transplanting method for leafy vegetable seedlings based on machine vision. Image processing based on Python-OpenCV was performed to obtain the seedlings' height and extreme edge points. The pixel coordinates in the RGB image were then obtained, and the depth image was aligned with the RGB image to acquire the depth information of the corresponding extreme points.

Cannabis cultivation is restricted around the world by governments. According to European Monitoring Centre for Drugs and Drug Addiction (2018), in the European Union, it is legal to cultivate and supply cannabis plants for hemp fibre if they have low levels of THC, the psychoactive substance present in cannabis plants. Additionally, medicinal cannabis can be produced if the farmers obtain a license from the Office of Medicinal Cannabis (OMC), the organisation in charge of controlling the production and supplying to pharmacies. Controlling the number of plants produced during the germination process could be beneficial for farmers to comply with license restrictions.

We propose implementing an integrated architecture to support farmers in the cultivation of cannabis for medicinal and industrial purposes. In this architecture, we take advantage of advances from emergent technologies like the Internet of Things (IoT), Wireless Sensor Networks, and Artificial Intelligence to provide a solution to support the germination process of cannabis plants. Our goal is that farmers can follow up on the germination of plants more easily, knowing when the germination occurs, the number of germinated plants, and other relevant information from the cultivation environment.

In order to assist farmers in monitoring germination, our aim is to develop a classifier that enables them to determine the number of plants that have germinated in the seedbed. Training the classifier requires a collection of labelled images.

Several datasets of cannabis images have been published for various purposes. Eranga et al. (2022) published a dataset containing different climate, soil, and yield data related to cannabis cultivation. Chumchu and Kailas (2023) released a dataset of cannabis seed images for machine learning applications. The same authors (Kailas and Chumchu, 2024) published a dataset for botanical exploration, including images of entire plants. However, none of these datasets include images of the plant's germination process.

Due to the limited information available on hemp cultivation, we have created a dataset. In this paper, we outline the rigorous process undertaken to capture, clean, label and organise the images to ultimately generate the dataset necessary for training the classification algorithm.

2. Materials and Methods

We work in a greenhouse where we deploy sensors to monitor ambient conditions. The greenhouse's size is approximately 56 square meters (9.75 m x 5.83 m), distributed in six crop tables, one of which was used for the germination experiment. The greenhouse can automatically control environmental conditions like temperature, humidity, and lightning. We have deployed a web platform to monitor and control the distinct

parameters in the greenhouse. We can access the sensors using the web platform and query the data gathered. We have also configured notification services to get alerts when a parameter is out of range or a sensor is disconnected. In Fig. 1, we show the distinct subsystems deployed in the greenhouse.

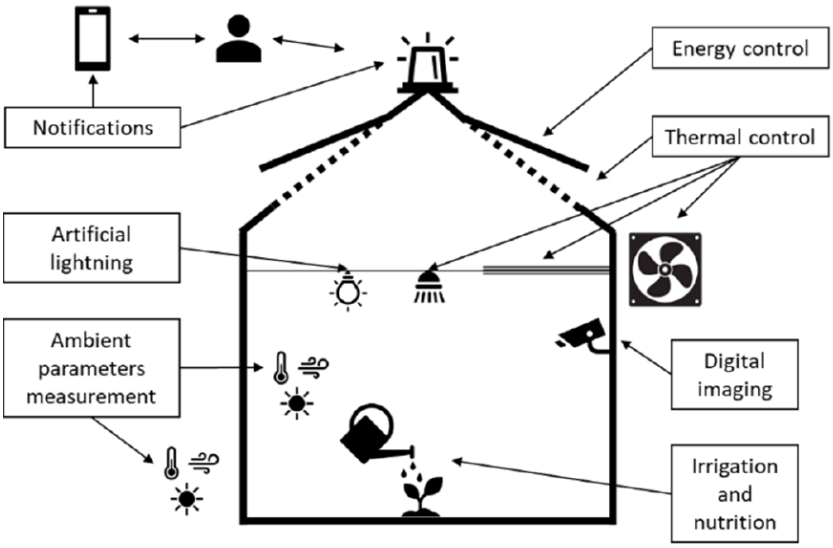

Figure 1. Greenhouse components

2.1. Germination experimental setup

Seeds of two varieties of *Cannabis sativa* L. were sown: Finola and Kompolti. The research group has a license from the Spanish Agency of Medicines and Medical Devices (AEMPS) to cultivate these hemp varieties. Finola is an oilseed hemp variety and was admitted to the European Union's list of subsidized crops in February of 2003. Kompolti is cultivated industrially and used for cosmetics, oils and even flours.

Each variety was sown in a 72-cell seedbed. The substrate was a mixture of blonde peat (65%), coconut fibre (20%), perlite (10%), and black peat (5%). During the germination tests, the substrate moisture was maintained by manually spraying the seedbeds with osmotised water with a pH of 6.2 and an electrical conductivity of 1.2 $mS \cdot cm^{-1}$.

We conducted four iterations each lasting 14 days since seeding. Two iterations were carried out without environmental control and two iterations were conducted under controlled conditions: the ambient temperature of the greenhouse was maintained between 18 and 26 ºC, the relative humidity between 55 and 80 %, and a photoperiod of 18 hours.

We put a camera above the seedbeds and connected it to a Raspberry Pi 4B. We created a python script to control the camera and take pictures. During each iteration, a photo was taken every hour. In Table 1, we detail the camera's specifications used in the experiment. Only the camera's focus was adjusted for the initial setup to avoid blur in the result images.

Table 1. Camera specification.

Camera model	Arducam 64 MP Auto-focus Camera
Optical size	Type 1/1.7” 9.25 mm Diagonal (7.4x5-55 mm)
Focus type	Manual/Auto
Sensor resolution	9152x6944
Color filter	Quad Bayer Coding (QBC)
Focus	8cm
Focal ratio (F-stop)	F1.8
Focal length	5.1 mm
View Angle	84 degrees (diagonal)

2.2. Images captured from overhead camera

Figure 2 shows an example of the image captured by the camera. We put drawing pins in the corners of the seedbeds for later processing of the photos. After completing each iteration, we collected the images and prepared the next iteration. Then, we created a python script to process the images. At the beginning of the process, we applied filters to clean the image and reduce the noise. Then, by using the drawing pins and Hough Circle Transform (Yuen et al., 1989), we detected the four corners of the seedbeds. After that, we used the coordinates of each corner and applied a Homography process (Hartley and Zisserman, 2004) to correct and align the image. At the end of the process, we cut the resultant image to get each cell of the seedbed. The whole process is detailed in Figure 3.

Figure 2. Example of an image captured by the camera. Left: Finola hemp variety samples. Right: Kompolti hemp variety samples.

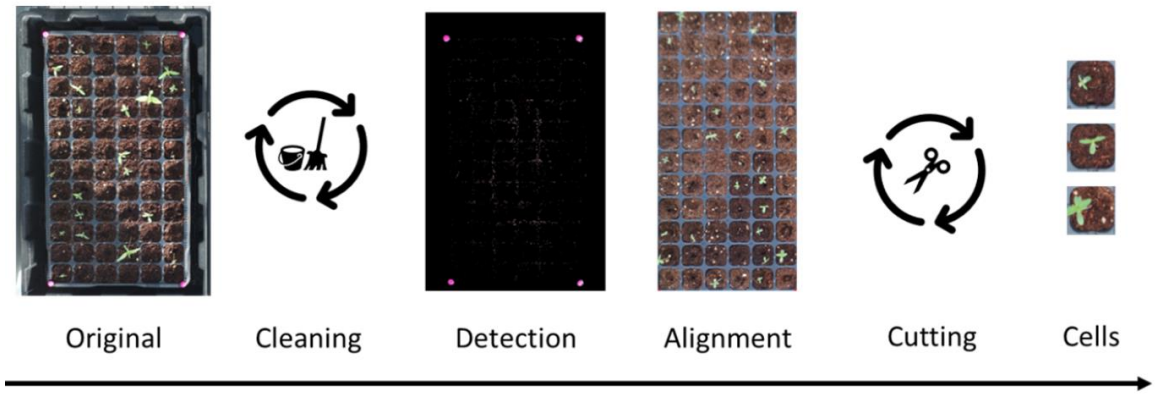

Figure 3. Images pre-processing steps.

To finish this process, we saved the images of cells. An automated labelling process was employed, wherein each image from the seedbed was cropped and assigned a name indicating the variety, iteration, day, time, and position within the seedbed. The cells were renamed as follows:

Variety_iteration_day_hour_cell-Row_cell-Column.jpg

For example: Finola_1_1_10_3_5.jpg.

Table 2 shows the possible values of the automatic labelling. Each iteration lasted 14 days. Every day 10 pictures were cropped, starting from 8 to 17 or from 9 to 18, according with the sun hours. Experiments were conducted during November 2022 to January 2023.

Table 2. Automatic labelling.

Label	Possible values
Variety	Finola / Kompolti
Iteration	1/2/3/4
Day	1/2/3/4/5/6/7/8/9/10/11/12/13/14
Hour	1/2/3/4/5/6/7/8/9/10
Row	1/2/3/4/5/6/7/8/9/10/11/12
Column	1/2/3/4/5/6

2.3. Pre-processed images

Figure 4 shows an example of the images pre-processed for a seedbed cell at days 6, 10, and 14. The pictures of Figure 4 correspond to a Finola hemp variety sample. Therefore, these images are easy to classify.

Figure 4. Example of images for the same cell at day 6 (left), 10 (center), and 14 (right).

Figure 5 shows examples of cells that are more difficult to classify. They correspond to different scenarios such as the presence of a plant sprouted in an adjacent cell, whether there is germination in the cell or not, or a plant whose leaves are not visible in the picture. This often occurs when the stems grow and bend.

Figure 5. Examples of scenarios where a plant is present, not always indicating germination in the cell. Left: plant leaves are missing, yet a germinated plant is present. Centre: neighbouring plant leaves alongside a germinated plant in the cell. Right: presence of a neighbouring plant, but no germinated plant within the cell.

Once we got images representing cells at different times, we labelled each one. To do that, we conducted two labelling phases and used six categories. The categories and stages are detailed in Table 2.

Table 2. Labelling phases and categories.

First labelling phase		Second labelling phase	
- Non-germination	NG		
- Germination	G	- Invasion	I
+ Cotyledon Only	CO	- Non-invasion	NI
+ Cotyledon and True Leaves	TL		

By combining two labels, we can support the correct classification of cells in germination.

For example: in general, a cell with the label Germination and Invasion means the presence of a plant. On the other hand, Non germination and Non Invasion correspond to not presence of a plant inside the cell. Meanwhile, there are another two cases in which a plant can be presented in a cell, when it is labelled as Germination and Non Invasion, and when it is labelled as Non germination and Invasion.

3. Results and discussion

A dataset has been created consisting in 80,640 images.

The images underwent thorough cleaning and organization processes, followed by the assignment of appropriate names and labelling. The dataset was meticulously curated to ensure consistency and accuracy in its contents.

The first automated labelling ensured efficient organisation and categorisation of the dataset, laying the groundwork for subsequent manual labelling and analysis. The systematic naming convention adopted facilitated easy retrieval and referencing of individual images during the dataset creation process, contributing to the overall coherence and accessibility of the dataset for research purposes.

Following the automated labelling process, manual labelling was conducted to further categorise the images based on several criteria. Each image was manually classified according to the presence of germinated plants, distinguishing between images depicting seedlings with only cotyledons and those exhibiting true leaves. Additionally, images were annotated to identify instances of plant invasion from adjacent cells. This meticulous manual labelling process enabled the creation of a richly annotated dataset, providing valuable insights into the progression of seedling development and interactions within the seedbed environment.

The combination of automated and manual labelling strategies ensured the accuracy and granularity of the dataset, enhancing its utility for a wide range of research applications in plant biology and agriculture.

Table 4. Example of a doubly labelled image from the dataset.

Image	Variety	Iteration	Day	Hour	Row	Column	Classes
Finola_it1_6_1_3_2.jpg	Finola	1	6	1	3	2	I/CO

The dataset provides researchers with a valuable resource for studying seedling development and related phenomena.

Having a dataset of seed germination images is of paramount importance for future research worldwide. These datasets serve as invaluable resources for scientists and researchers across various fields, providing a rich source of data to study the intricate processes involved in seed germination.

By collecting and curating such datasets, researchers can gain insights into the factors influencing germination rates, timing, and success, thereby advancing our understanding of plant biology and agricultural practices.

This dataset enables the development and validation of machine learning algorithms and computer vision techniques for automated seed germination detection and analysis, paving the way for innovative solutions in precision agriculture and crop management.

This dataset can be further enhanced by incorporating images of germinated cannabis plants exhibiting various stress conditions such as nutrient deficiencies, water stress, pest infestations, and diseases. It can also be supplemented by incorporating images of hemp plants in later stages of development.

By augmenting the dataset with these additional images, it will be possible to train more sophisticated algorithms capable of accurately detecting and diagnosing a range of pathophysiological conditions.

Having a dataset with images of cannabis plants at more advanced stages of development may allow to

determine the sex of the plants at an earlier stage, which is particularly important for cannabis cultivation, where the identification of male plants is crucial to prevent them from pollinating female plants. It would be also possible to perform a detailed analysis of the growth patterns and morphological changes over time and to predict the potential yield of the crop.

4. Conclusion

In this work, we have presented the creation of a dataset comprising images of cannabis seed germination for the Kompolti and Finola varieties. This dataset has been meticulously and rigorously developed, encompassing the entire process from image capture, pre-processing, and cropping of individual seedbed cells, to the systematic naming and labelling of the resulting images. Each cell was labelled to indicate the presence or absence of germination, whether the seedling exhibited only cotyledons or true leaves as well, and whether invasion from an adjacent cell was observed.

This dataset will enable the training of an algorithm for the automatic classification of seedbed images. It represents the first systematic dataset published with images of *Cannabis sativa* germination. The publication of this dataset will provide a valuable resource for researchers needing high-quality images to train their algorithms.

Moreover, the dataset can be further enhanced by incorporating images of plants affected by pests and diseases or images of plants at more advanced stages of development. This would facilitate the training of more sophisticated algorithms capable of detecting a broader range of conditions and traits, thereby expanding the utility and impact of this resource.

By making this dataset available, we aim to support the scientific community in advancing research in cannabis cultivation, ultimately contributing to improved agricultural practices and plant health monitoring.

Acknowledgements

This study is part of the AGROALNEXT program (AGROALNEXT/2022/048) and has been supported by MCIN with funding from the European Union NextGenerationEU (PRTR-C17.I1) and the Generalitat Valenciana. C. Rocamora has been funded by the Ministry of Universities and by the European Union-Next Generation EU within the framework of Grants for the Requalification of the Spanish University System, in the University teaching staff modality.

This study was partially supported by the Research Center for Communication and Information Technologies (CITIC), Research Project No. 834-B9-189. J.A. Brenes has been funded by the Office of International Affairs and External Cooperation OAICE (Short-Term Scholarship) and the Postgraduate Studies System SEP (Restricted Fund 082) of the University of Costa Rica.

References

Basu R. N. 1995. Seed viability. In *Seed Quality. Basic Mechanism and Agricultural Implications*, Food Products Press, The Haworth Press, Inc. New York, Ed. A.S. Basra, pp 1-44.

Chumchu P., K. Patil, 2023. Dataset of cannabis seeds for machine learning applications, Data in Brief, 47, 2023, 108954. https://doi.org/10.1016/j.dib.2023.108954.

Dell'Aquila, A, 2005. The use of image analysis to monitor the germination of seeds of broccoli (*Brassica oleracea*) and radish (*Raphanus sativus*). Annals of Applied Biology 146, 545–550. https://doi.org/10.1111/j.1744-7348.2005.040153.x.

Dornbos D. L. 1995. Seed Vigor. In *Seed Quality. Basic Mechanism and Agricultural Implications*, Food Products Press, The Haworth Press, Inc. New York, Ed. A.S. Basra, pp 45-80.

Ducournau, S., A. Feutry, P. Plainchault, P. Revollon, B. Vigouroux, M. Wagner, 2004. An image acquisition system for automated monitoring of the germination rate of sunflower seeds. Computers and Electronics in Agriculture 44, 189–202. https://doi.org/10.1016/j.compag.2004.04.005.

European Monitoring Centre for Drugs and Drug Addiction, Hughes, B., 2018. Cannabis legislation in Europe – An overview, Publications Office of the European Union, https://data.europa.eu/doi/10.2810/566650

Hartley, R.; A. Zisserman, 2004. Multiple View Geometry in Computer Vision; Cambridge University Press, 2004. https://doi.org/10.1017/cbo9780511811685.

Jin, X.; R. Li, Q. Tang, J. Wu, L. Jiang, C. Wu, 2022. Low-damage transplanting method for leafy vegetable seedlings based on machine vision. Biosystems Engineering 220, 159–171. https://doi.org/10.1016/j.biosystemseng.2022.05.017.

Li, C., A. Raheja, D. Still, 2015. Application of Computer Vision for Lettuce Seeds Germination Detection. In *WORLDCOMP'15-The 2015 World Congress in Computer Science, Computer Engineering, and Applied Computing*. WORLDCOMP, Las Vegas, NV, USA, pp.1-5.

Patil K., P. Chumchu, 2024. A comprehensive dataset of eight Thai cannabis classes for botanical exploration, Data in Brief, 54, 110292. https://doi.org/10.1016/j.dib.2024.110292.

Ryu, K.; G. Kim, J. Han, 2001. AE—Automation and Emerging Technologies: Development of a Robotic Transplanter for Bedding Plants. Journal of Agricultural Engineering Research 78, 141–146. https://doi.org/10.1006/jaer.2000.0656.

Tong, J.H., J.B. Li, H.Y. Jiang, 2013. Machine vision techniques for the evaluation of seedling quality based on leaf area. Biosystems Engineering 115, 369–379. https://doi.org/10.1016/j.biosystemseng.2013.02.006.

Ureña, R., F. Rodrıguez, M. Berenguel, 2001. A machine vision system for seeds quality evaluation using fuzzy logic. Computers and Electronics in Agriculture 2001, 32, 1–20. https://doi.org/https://doi.org/10.1016/S0168-1699(01)00150-8.

Wimalasiri E.M., E. Jahanshiri, T.A. Syaherah, N. Kuruppuarachchi, V.G.P. Chimonyo, S.N. Azam-Ali, P.J. Gregory, 2022. Datasets for the development of hemp (*Cannabis sativa* L.) as a crop for the future in tropical environments (Malaysia). Data in Brief, 40, 107807. https://doi.org/10.1016/j.dib.2022.107807.

Yuen, H.K., J. Princen, J. Dlingworth, J. A. Kittler, 1989. Comparative Study of Hough Transform Methods for Circle Finding. In *Proceedings of the Alvey Vision Conference*. Alvey Vision Club, Reading, U.K., September 1989. Ed K. D. Baker. https://doi.org/10.5244/c.3.29.