




IndianPotatoWeeds: An Image Dataset of Potato Crop to Address Weed Issues in Precision Agriculture

Rajni Goyal^(✉) , Amar Nath , and Utkarsh 

Sant Longowal Institute of Engineering and Technology, Longowal, India
{rajni_pcs2103, amarnath, utkarsh}@sliet.ac.in

Abstract. Over the past ten years, precision agriculture has raised much awareness in the agricultural sector. It automates and optimizes almost all agriculture practices. But the success of this technology depends upon the data. The more accurate and extensive the data is, the more accurate the system will be. Despite large datasets available online, there is still a lack of datasets from the Indian perspective. One of the main roadblocks to advancement is the shortage of publicly available statistics. This paper proposes a benchmark dataset named IndianPotatoWeeds for Potato crops and weeds from Indian farms. The dataset comprises 270 images with annotations and is available online <https://www.kaggle.com/datasets/rajni88/indianpotatoweed-dataset>. All images were acquired with the Sony CyberShot W830 20.1 M camera and mobile phone. There were intra and inter-row weeds present at the time of data collection. We have provided mask and manual annotation of the plant type (crop vs. weed) for every dataset image using VIA annotation tool. Images can be split into background and foreground via masking, enabling us to concentrate on the areas of the image that interest us. By making this information available to the public, we hope to encourage study in this field.

Keywords: Precision Agriculture · Weed Dataset · Annotation · Image Masking

1 Introduction

Food and fiber are mainly produced via agriculture. Agriculture provides all the world's inhabitants with the nutrition it needs. Also vital to the economy is agriculture. A significant share of the global workforce is employed in the agriculture sector. 50% of the Indian population is employed in the agriculture sector. The economic survey estimates that agriculture's share of the GDP for the fiscal year 2020–2021 was 19.9%¹. The population has increased exponentially over the past three decades (almost 10 billion by 2050, according to a U.N. study),

¹ <https://statisticstimes.com/economy/country/india-gdp-sectorwise.php>.

which has led to a sharp rise in food demand [1]. To feed such a vast population, conventional food production methods are insufficient.

Hence, there is a need for innovative and intelligent ways of farming. Smart and Precision farming is one of the best solutions to meet the current food demands of the large population. Precision agriculture provides the automation of various agriculture practices, which reduces time and labor costs. Artificial intelligence(AI) and IoT (Internet of things) are advanced technological tools that help to digitize agricultural activities. To cut waste, boost revenues, and protect the environment, precision agriculture manages each crop production input (water, fertilizer, herbicide, seed, pesticide, etc.) on a site-specific basis. (Ess & Morgan 2013). Precision agriculture has a wide application area. It helps in crop management, weather forecasting, weeds, and pest control, intelligent spraying, livestock farming, remote sensing, storage management, innovative harvesting, etc. It automates almost all agriculture practices. It helps to detect weeds and pests in crop fields and provides site-specific spraying of pesticides and herbicides to remove weeds and pests.

Weeds are a fundamental problem that affects crop yields to a large extent. Weeds are unwanted plants that compete for nutrients, water, and other resources with valuable plants. These unwanted plants are always needed to be removed from fields. The type of weeds depends upon the location, season, and crop. Weeds vary from country to country. So, the kind of weeds present in Germany's land does not need to also be present in the land of India. There are many datasets available from various countries. But from the Indian perspective, there is a lack of weed/crop datasets.

So in this paper, we provide a thoroughly annotated and masked crop/weed dataset from potato fields. The dataset contains 270 images manually annotated using VIA (VGG Image Annotator) [30]. The annotations made available with this dataset enable the development of weed detection and classification solutions and many types of image processing, including edge detection, motion detection, and noise reduction. The information presented is crucial from a computer vision standpoint. On the one hand, the process of picture collecting in the agricultural industry is challenging since it necessitates sophisticated hardware systems, access to fields, and lightning conditions, and the timing of the acquisition must be accurate and linked with the crop growth cycle (only once a year for many cultures). On the other hand, defining appropriate ground truth requires the assistance of agricultural professionals [2].

The dataset comprises field images in a top-down view that were acquired with a Sony CyberShot W830 20.1 M and mobile phone camera. The images are collected from the potato fields of Punjab Agriculture University (Precision farming) fields in Punjab, India. The crop was photographed at a stage of development where many genuine leaves were visible. The manual weeding was done in this field after a few hours of data collection. Here, we focus on potatoes, but wheat, peas, onions, and other cultivars also need manual weed control procedures. Every image has annotations, and the dataset contains crop/weed annotation JSON file, CSV file, Coco format file, and annotation mask for each image.

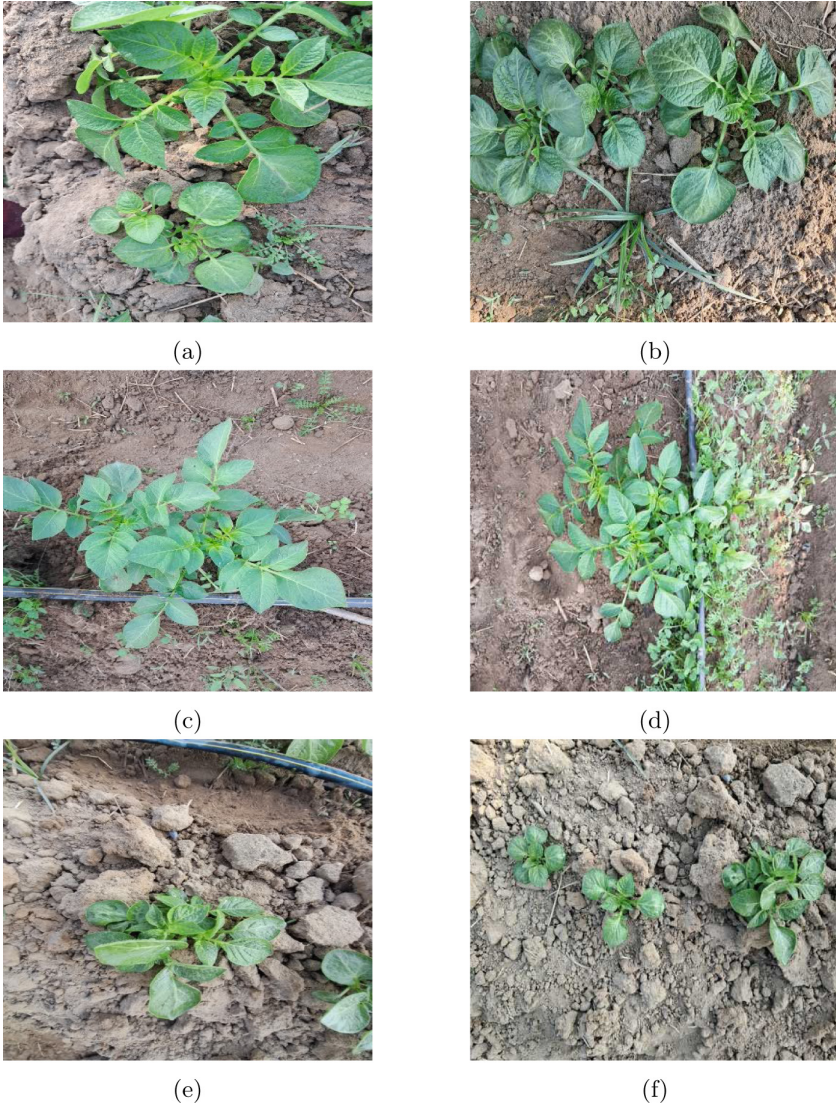


Fig. 1. Sample images from dataset

The dataset is available online at <https://www.kaggle.com/datasets/rajni88/indianpotatoweed-dataset> Fig. 1 provides sample images from the dataset.

2 Literature Survey

In general, there is a lack of open datasets accessible by researchers and academicians. Data sets are like the food for classification and detection problems

in machine learning models [3]. These technologies are used in a variety of agricultural fields, including crop disease detection, weed classification and identification, plant seedling classification, fruit identification and accounting, management of water resources and soil, weather forecasting (climate) [3–6]. Accurately classifying and detecting weed species in their natural environment may be the most significant barrier to the general adoption of robotic weed management.

The more data included in these databases, the more effective artificial intelligence systems can govern robotic weed growth, provide more accurate plant growth, and allocate scarce resources. Potato/weed dataset [8] is an open-access dataset having 411 images taken from potato fields. But this dataset contains separate images for crop and weed and cannot be used for segmentation problems. It is valid for classification problems only. Another dataset for weed detection has 202 images [9] that can be used for classification problems in deep learning. Another

Table 1. Public Crop/Weed datasets

| Dataset | Type of crop | Number of Images | Data Location | Reference |
|---------------------------------------|-----------------------------|---------------------------------------|-----------------------|------------|
| Potato Plant Dataset | Potato | 411 | not specified | [8] |
| Weed detection Dataset | not specified | 202 | not specified | [9] |
| Crop/Weed Field Image Dataset | Carrot | 60 | Germany | [10] |
| Dataset of food crop and weed | Six crop | 1118 | Latvia | [11] |
| DeepWeeds | Not Specified | 17,509 | Australia | [12] |
| Perennial ryegrass and weed | Perennial ryegrass | 33086 | USA | [15] |
| Early crop weed dataset | Tomato, cotton | 508 | Greece | [14] |
| Soybean and weed dataset | Soybean | 400 | Brazil | [16] |
| Open Plant Phenotype Dataset | Not Specified | 7,590 | Denmark | [17] |
| Sugar beet and hedge bindweed dataset | Sugar beet | 652 | Belgium | [18] |
| Sugar beet fields dataset | Sugar beet | 12340 | Germany | [19] |
| UAV Sugar beet 2015-16 Datasets | Sugarbeets | 675 | Switzerland | [20] |
| Corn, Lettuce and weed dataset | Corn and lettuce | 6800 | China | [22] |
| Carrot weed dataset | Carrot | 39 | Republic of Macedonia | [23] |
| Bccr-segset dataset | Canola, corn, radish | 30,000 | Australia | [24] |
| Carrots 2017 dataset | carrots | 20 | UK | [25] |
| Onions 2017 dataset | onions | 20 | UK | [25] |
| GrassClover image dataset | Red Clover and white clover | 31,600 real and 8000 synthetic images | Denmark | [26] |
| Leaf counting dataset | not specified | 9372 | Denmark | [27] |
| CNU weed dataset | not specified | 208,477 | Republic of Korea | [28] |
| Plant Seedling dataset | three crops | 5539 | Denmark | [29] |
| CNU weed dataset | not specified | 208,477 | Republic of Korea | [28] |
| IndianPotatoWeeds | Potato | 270 | India | This paper |



(a)



(b)

Fig. 2. The polygon annotation of images. The yellow color specify crop, and the blue color specify weed.

dataset named `cwfid` [10], having 60 shots, is available on GitHub for crop /weed classification and segmentation for computer vision in precision agriculture.

Sudras et al. [11] annotated 1118 images having six food crops and eight weed species from different locations in Latvia. DeepWeeds [12] is an extensive dataset having 17,509 images taken from different crop fields in Australia. Table 1 represents the various datasets available online from fields of other countries for different crops.

This paper aims to provide a real-world image dataset for image segmentation and classification model like Faster Region-based Convolutional Neural Network (RCNN) and Mask RCNN. This enables researchers to acquire research on the perception of data acquisition and treatment for weeds in potato fields.

3 Problem Description

Data presented in this paper shows how the dataset is distributed among food crops and weeds. The crop selected for this work is potato. Two hundred seventy images presented in this paper are manually annotated using the VGG image annotator (VIA) tool. The dataset is split into **train** and **val** folders containing 80:20 images. Each folder contains the JSON file having annotations. Raw images and mask for each image is also included in the dataset. Figure 3 displays images from a dataset with polygon annotation with yellow color specifying crop and blue color specifying weed (Better visible in color image).

4 Material and Methods

4.1 VIA (VGG Image Annotator)

VGG Image Annotator (VIA) is an easy-to-use standalone program for manually annotating images, audio files, and videos. There is no setup or installation needed with VIA; it simply runs in a web browser. The complete VIA program is included in a single self-contained HTML page that is less than 400 kilobytes and works as an offline application in most modern web browsers [30]. Using the VIA tool, we have annotated the images. We have also classified images into weed and crop categories, shown in Fig. 3. The region shape used for annotation is the polygon. The total number of annotations is 776, of which 393 are crop annotations and 383 for weed. The extent of the dataset is represented in Table 2.

Table 2. Extent of dataset

| Parameter | Value | Format |
|------------------|-------|-------------|
| Image count | 270 | .jpg |
| Annotations | 776 | json, csv |
| Crop annotations | 393 | json, csv |
| Weed annotations | 383 | .json, .csv |
| Mask of images | 270 | .png |

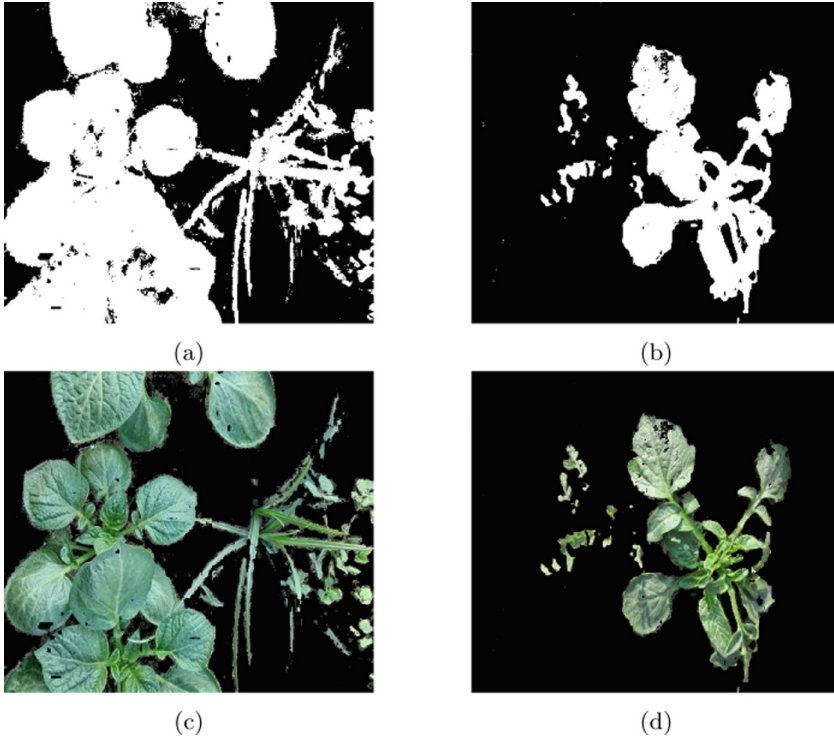


Fig. 3. Annotated and masked images into crop/weed using VIA tool and python (a) & (b) mask of images (c) & (d) masked images

4.2 Masking

A mask allows us to focus only on the portions of the image that interests us. It can be defined as setting specific pixels of an image to some null value such as 0 (black color). So, only that portion of the image is highlighted where the pixel value is not 0. In this program, we begin with reading the image using the `cv2.imread()` function in python. Then we convert the image to HSV format as all the operations can only be performed in HSV format.

During masking, the images can be segmented into background and foreground. Figure 3 shows the mask and the masked image from the dataset.

4.3 Field Setup and Acquisition Method

The 270-image dataset was captured at a precision agriculture potato farm in Northern India in December 2022 before manual weeding was applied. The potato plants were grown in a single row on small soil beds. Small close-to-close intra-row weeds were present at data acquisition time. Sony CyberShot W830

Table 3. Specifications of Dataset

| | |
|-----------------------|---|
| Subject | Precision Agriculture, Computer vision, Agronomy, and Science |
| Specific Subject Area | Image classification, segmentation, Object detection, crop growth and development |
| Type of data | Images Annotations Image Mask and masked images |
| Camera specifications | Sony CyberShot W830 20.1 MP and mobile phone camera |
| Data Format | Raw images: .jpg format Manually annotated images: JSON files Mask and masked image: .png format |
| Description of data | Dataset consists of Directory Raw images -270 images and mask for each image, train folder - 214 images, JSON file, CSV file test folder - 56 images, JSON file, CSV file |
| Data source location | Precision farms Punjab Agriculture University, Ludhiana, Punjab, India |
| Data accessibility | https://www.kaggle.com/datasets/rajni88/indianpotatoweed-dataset |

20.1 MP and mobile cameras captured the images in an unregulated environment. During data collection, the weather was clear, with no clouds. Specifications of the dataset are provided in Table 3.

5 Work Flow

Sony Cyber-shot cameras and mobile devices were initially used to capture the raw photos. There were 600 pictures altogether. The data were cleaned to eliminate duplicate photos, blurry images, and noise. After cleaning, 270 images in total were collected. The data were divided 80:20 between `train` and `val` folders. Using VIA Annotator, each image was manually annotated. The annotation tool exported JSON and CSV files. We manually constructed a mask for each image and used Python to mask each image. The files were all uploaded to <https://www.kaggle.com/datasets/rajni88/indianpotatoweed-dataset>.

6 Value of the Data

- The dataset presents images of potato crops and weeds in their early growth stages, which can be used by agronomists and researchers in different fields for computer vision and smart farming.
- The open-access dataset can be used for weed recognition and segmentation algorithms.

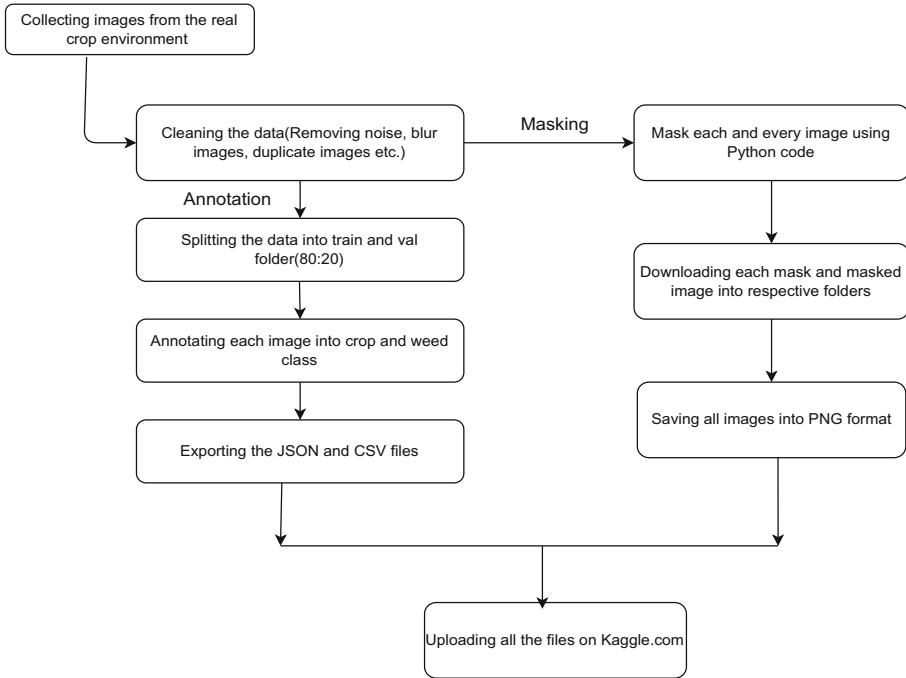


Fig. 4. Work Flow Diagram of the Proposed Approach

- The dataset can train, test and validate convolutional neural networks(CNN) models.

7 Conclusion

A potato crop and weeds dataset for addressing the weed issues in precision agriculture is collected, masked and posted on Kaggle. The images of crops and weeds are acquired using a Sony digital camera and mobile camera in Punjab, India. During the collection of data, there were inter and intra-row weeds were present in the field. The images are manually annotated using VIA (VGG Image Annotator) Tool. There are a total of 270 images in the dataset divided into train and val folders.

This dataset can be used for weed detection, segmentation, and classification problem. We hope this will help increase the progress in the required data acquisition domain and generate ground truth. It will help researchers and agriculture experts to develop ground truth of weed management. In the future, this dataset can be extended with more images from different regions in different seasons and growth days.

Acknowledgement. The authors thank the following colleagues for their comments and help with the acquisition of the dataset: Dr. Rakesh Sharda (Principal Scientist,

Punjab Agriculture University, Ludhiana Punjab (India), Dr. Pankaj Punjab Agriculture University, Ludhiana Punjab (India)).

References

1. Lal, R.: Soil structure and sustainability. *J. Sustain. Agric.* **1**(4), 67–92 (1991)
2. Haug, S., Ostermann, J.: A crop/weed field image dataset for the evaluation of computer vision based precision agriculture tasks. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) *ECCV 2014*. LNCS, vol. 8928, pp. 105–116. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16220-1_8
3. Yan, J., et al.: Robust multi-resolution pedestrian detection in traffic scenes. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2013)
4. Boulent, J., et al.: Convolutional neural networks for the automatic identification of plant diseases. *Front. Plant Sci.* **10**, 941 (2019)
5. Jeon, W.-S., Rhee, S.-Y.: Plant leaf recognition using a convolution neural network. *Int. J. Fuzzy Logic Intell. Syst.* **17**(1), 26–34 (2017)
6. Koirala, A., et al.: Deep learning for real-time fruit detection and orchard fruit load estimation: benchmarking of ‘MangoYOLO’. *Precision Agric.* **20**, 1107–1135 (2019)
7. Nkemelu, D.K., Omeiza, D., Lubalo, N.: Deep convolutional neural network for plant seedlings classification. arXiv preprint [arXiv:1811.08404](https://arxiv.org/abs/1811.08404) (2018)
8. ALI Hassan Kaggle datasets. <https://www.kaggle.com/datasets/ali7432/potato-weed-plants-classification>. Accessed 23 Nov 2022
9. AjinJayan. https://github.com/AjinJayan/weed_detection/blob/master/dataset_updated.zip. Accessed 6 Mar 2023
10. Sebastian Haug, Jörn Ostermann github.com. <https://github.com/cwfid/dataset>. Accessed 23 Nov 2022
11. Sudars, K., Jasko, J., Namatevs, I., Ozola, L., Badaukis, N.: Dataset of annotated food crops and weed images for robotic computer vision control. *Data Brief* **31**, 105833 (2020)
12. Olsen, A., et al.: DeepWeeds: a multiclass weed species image dataset for deep learning. *Sci. Rep.* **9**(1), 1–12 (2019)
13. Hasan, A.M., Soheli, F., Diepeveen, D., Laga, H., Jones, M.G.: A survey of deep learning techniques for weed detection from images. *Comput. Electron. Agric.* **184**, 106067 (2021)
14. Espejo-Garcia, B., Mylonas, N., Athanasakos, L., Fountas, S., Vasilakoglou, I.: Towards weeds identification assistance through transfer learning. *Comput. Electron. Agric.* **171**, 105306 (2020)
15. Yu, J., Schumann, A.W., Cao, Z., Sharpe, S.M., Boyd, N.S.: Weed detection in perennial ryegrass with deep learning convolutional neural network. *Front. Plant Sci.* **10**, 1422 (2019)
16. dos Santos Ferreira, A., Freitas, D.M., da Silva, G.G., Pistori, H., Folhes, M.T.: Unsupervised deep learning and semi-automatic data labeling in weed discrimination. *Comput. Electron. Agric.* **165**, 104963 (2019)
17. Leminen Madsen, S., Mathiassen, S.K., Dyrmann, M., Laursen, M.S., Paz, L.C., Jørgensen, R.N.: Open plant phenotype database of common weeds in Denmark. *Remote Sensing* **12**(8), 1246 (2020)

18. Gao, J., French, A.P., Pound, M.P., He, Y., Pridmore, T.P., Pieters, J.G.: Deep convolutional neural networks for image-based *Convolvulus sepium* detection in sugar beet fields. *Plant Methods* **16**(1), 1–12 (2020)
19. Chebrolu, N., Lottes, P., Schaefer, A., Winterhalter, W., Burgard, W., Stachniss, C.: Agricultural robot dataset for plant classification, localization and mapping on sugar beet fields. *Int. J. Robot. Res.* **36**(10), 1045–1052 (2017)
20. Chebrolu, N., Läbe, T., Stachniss, C.: Robust long-term registration of UAV images of crop fields for precision agriculture. *IEEE Robot. Autom. Lett.* **3**(4), 3097–3104 (2018)
21. Madakam, S., Lake, V., Lake, V., Lake, V.: Internet of Things (IoT): a literature review. *J. Comput. Commun.* **3**(05), 164 (2015)
22. Jiang, H., Zhang, C., Qiao, Y., Zhang, Z., Zhang, W., Song, C.: CNN feature based graph convolutional network for weed and crop recognition in smart farming. *Comput. Electron. Agric.* **174**, 105450 (2020)
23. Lameski, P., Zdravevski, E., Trajkovik, V., Kulakov, A.: Weed detection dataset with RGB images taken under variable light conditions. In: Trajanov, D., Bakeva, V. (eds.) *ICT Innovations 2017. CCIS*, vol. 778, pp. 112–119. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67597-8_11
24. Le, V.N.T., Ahderom, S., Apopei, B., Alameh, K.: A novel method for detecting morphologically similar crops and weeds based on the combination of contour masks and filtered Local Binary Pattern operators. *GigaScience* **9**(3), giaa017 (2020)
25. Bosilj, P., Aptoula, E., Duckett, T., Cielniak, G.: Transfer learning between crop types for semantic segmentation of crops versus weeds in precision agriculture. *J. Field Robot.* **37**(1), 7–19 (2020)
26. Skovsen, S., et al.: The GrassClover image dataset for semantic and hierarchical species understanding in agriculture. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2019)
27. Teimouri, N., Dyrmann, M., Nielsen, P.R., Mathiassen, S.K., Somerville, G.J., Jørgensen, R.N.: Weed growth stage estimator using deep convolutional neural networks. *Sensors* **18**(5), 1580 (2018)
28. Trong, V.H., Gwang-hyun, Y., Vu, D.T., Jin-young, K.: Late fusion of multi-modal deep neural networks for weeds classification. *Comput. Electron. Agric.* **175**, 105506 (2020)
29. Giselsson, T.M., Jørgensen, R.N., Jensen, P.K., Dyrmann, M., Midtiby, H.S.: A public image database for benchmark of plant seedling classification algorithms. *arXiv preprint arXiv:1711.05458* (2017)
30. Dutta, A., Zisserman, A.: The VIA annotation software for images, audio and video. In: *Proceedings of the 27th ACM International Conference on Multimedia* (2019)