Beet seedling and weed recognition based on convolutional neural network and multi-modality images

Jun Sun · Kaifeng Yang · Xiaofei He · Yuanqiu Luo · Xiaohong Wu · Jifeng Shen

Received: 29 June 2021 / Revised: 6 September 2021 / Accepted: 26 November 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract
Difficulties in the recognition of beet seedlings and weeds can arise from a complex background in the natural environment and a lack of light at night. In the current study, a novel depth fusion algorithm was proposed based on visible and near-infrared imagery. In particular, visible (RGB) and near-infrared images were superimposed at the pixel-level via a depth fusion algorithm and were subsequently fused into three-channel multi-modality images in order to characterize the edge details of beets and weeds. Moreover, an improved region-based fully convolutional network (R-FCN) model was applied in order to overcome the geometric modeling restriction of traditional convolutional kernels. More specifically, for the convolutional feature extraction layers, deformable convolution was adopted to replace the traditional convolutional kernel, allowing for the entire network to extract more precise features. In addition, online hard example mining was introduced to excavate the hard negative samples in the detection process for the retraining of misidentified samples. A total of four models were established via the aforementioned improved methods. Results demonstrate that the average precision of the improved optimal model for beets and weeds were 84.8% and 93.2%, respectively, while the mean average precision was improved to 89.0%. Compared with the classical R-FCN model, the performance of the optimal model was not only greatly improved, but the parameters were also not significantly expanded. Our study can provide a theoretical basis for the subsequent development of intelligent weed control robots under weak light conditions.

Keywords Object detection · Beets and weeds · Multi-modality images · Deformable convolution · Deep learning

1 Introduction
The presence of weeds in the field can cause great damage to crop seedlings. More specifically, weeds compete with crops for sunlight and nutrients, thus seriously affecting the photosynthesis of seedlings and increasing the spread of diseases and insect pests.
Therefore, the removal of weeds is of great significance in order to maintain crop yield [9]. Generally, the weeding methods based on machine vision are performed under good daylight conditions. However, in order to improve production efficiency and reduce the damage of weeds to crop seedlings, continuous operation at night is required to identify weeds. Traditional weed elimination methods are time-consuming and laborious, and mainly depends on artificial excavation or pesticide spraying. Moreover, it is difficult to identify weeds that are similar to crops [11]. Pesticide residues produced by spraying not only poses a great threat to human health, but can also damage the ecological environment [25]. With the increasing application of precision agriculture, it is particularly important to establish the real-time accurate identification of weeds and crops for the rational application of pesticides, crop yield increases, the reduction of environmental pollution, and the implementation of intelligent weeding [12, 28].

With the development of computer technology, the work efficiency is greatly improved by machine vision. Traditional machine vision methods classify objects in images by classifiers such as SVM [3, 5], random forest [5, 15, 34] and naive Bayes [17]. Traditional machine vision methods can only generate low-level image features, which limits the expression ability. However, the convolutional neural network (CNN) has a strong capability to represent image feature without manual feature selection. It can further extract high-level abstract features in images with a high accuracy and efficiency based on the gradient descent and back propagation algorithm [19]. Therefore, CNNs are widely used in image classification [22], object detection [20] and semantic segmentation [31]. The neurons in convolutional layers can extract the primary visual features of the image by local receptive fields, and reduce the number of parameters through weight sharing. The pooling layer can realize the invariance of displacement, scaling and distortion, while simultaneously performing feature dimensionality reduction.

At present, the computer vision-based detection of crops and weeds under weak light in the field is limited. This is a result of the noise commonly present in images of certain scenes (e.g. at night with insufficient light). Thus, the detection process is easily affected changes in light, thus misinterpreting the shape, color, texture and additional information of the detected object, resulting in a poor visual performance [29]. Moreover, traditional CNNs only use the original regular convolutional kernel to extract the features of crops and weeds, with a limited ability in geometric modeling. This may weaken the feature extraction ability of the model. Furthermore, during the training of the model, the uneven distribution of positive and negative samples will also lead to the poor generalization of the model. In view of the above limitations of the current methods, three improved methods were proposed in this study:

1. A deep fusion algorithm was adopted to fuse the RGB (Red, Green and Blue) and near-infrared (NIR) images of beets and weeds obtained under weak light conditions into three-channel multi-modality images, then the fusion images were sent into CNN for training.
2. In the feature extraction layer, the traditional convolutional kernel was replaced by the deformable convolution.
3. The hard negative samples were fully excavated by using online hard example mining (OHEM) in the detection process, and then they were sent into the network for re-identification.
The structure of the rest of this paper is as follows. Section 2 introduces the related work, and Section 3 introduces the materials and methods of this paper, including data source, improved weeds and beets detection model. Section 4 introduces the experimental environment. The fifth section shows the experimental results and discusses the results. Finally, the conclusion of this work is drawn in Section 6.

2 Related work

At present, most weed detection methods are developed based on machine vision. Zhao et al. [35] proposed a classification method for weed classification based on a back propagation (BP) neural network. Following the fuzzy classification of the features, a genetic algorithm was used to optimize the network for the identification of weeds. Yan et al. [32] designed a method to identify weeds based on machine vision during the maize seedling stage. After distortion correction, HSI (Hue, Saturation and Intensity) color space conversion and threshold segmentation, the collected images of maize plants and weeds were identified according to the shape and color features. Bakhshipour et al. [6] designed a weed segmentation network based on an artificial neural network. The single-stage wavelet transform was used to extract weed texture features with 14 texture features selected to optimize the algorithm by using the principal component analysis method. Finally, the features were sent into the neural network in order to identify the weeds. Akbarzadeh et al. [2] used Gaussian support vector machine algorithm to classify corn and weeds under laboratory conditions, and compared the accuracy with the traditional data aggregation method based on two different wavelength discrete normalized difference vegetation index. Abouzahir [1] improved weed detection performance by using directional gradient histogram (HOG). The accuracy of weed detection using back propagation neural network is 71.2% ~ 83.3%, which is 37.6% higher than the traditional HOG algorithm. The aforementioned studies combine shallow feature extraction and pattern recognition to identify weeds. However, the feature extraction of such methods is time consuming and the applicability is weak. In addition, due to the influence of the complex field background, the weed characteristics extracted by humans can be ambiguous and uncertain, which consequently results in weed identification based on traditional machine vision at low accuracies.

Recently, many networks based on pre-trained CNNs have achieved promising results in weed and crop seedling detection. Jiang et al. [13] proposed a weed identification method based on a deep CNN and hash code, which was able to effectively compress the high-dimensional features of the weed through a binary hash layer to detect weeds. Andrea et al. [4] used a classification method based on a CNN that identified maize seedlings and weeds by optimizing the number of convolutional kernels on the basis of the original classification network. Huang et al. [12] proposed a fully-connected CNN method by applying transfer learning to improve feature extraction, and a skip architecture structure for network optimization to detect weeds. Results from the aforementioned literature demonstrate that CNNs can not only automatically extract the shallow features (texture, color, etc.) of weeds and crops, but can also learn deeper abstract features. Moreover, CNNs are able to reduce the cost of feature extraction and are more robust for weed detection in a complex environment. Therefore, CNNs have the potential to be applied to detect beet seedlings and weeds. However, the models in the current literature are not sensitive to the feature information of weeds due to the weak light at night and the use of traditional convolutional kernels. This has resulted in feature extraction difficulties and low recognition accuracies.


3 Materials and methods

3.1 Data source

In order to investigate the detection and identification of beets and weeds in complex backgrounds, images of beets and weeds were collected at the University of Bonn, Germany, in 2016. For more details about the dataset, see the link and study below (http://www.ipb.unibonn.de/data/sugarbeets2016/). The images were collected via a multi-modality camera (JAI AD-130GE), equipped with two high-sensitivity CCD multispectral sensors of 1.3 million pixels. The camera can simultaneously collect visible (400nm~650nm) and near infrared (NIR) (760nm~1000nm), with an output image size of 1296×1296 pixels [18]. The dataset contains a total of 2,093 images of beets and weeds at different growth stages. In the process of data acquisition, beet seedlings and weeds with different levels of maturity under varying angle transformations were consider for the image acquisition. Moreover, the same plant (beet and weed) was imaged multiple times under different ranges of overlap and occlusion between beets and weeds. Some image examples are shown in Fig. 1. Since the original image was collected in a low-light environment and it was difficult to visualize, the image shown in this article had been enhanced the exposure and brightness.

3.2 Multi-modality image fusion

In general, object detection methods based on deep learning aim to understand the distribution of basic data via a large amount of training data and subsequently induce the optimizer to adjust the parameters of the network [26]. At present, RGB images on weeds and beets are commonly used to train deep learning models. However, RGB images are sensitive to variations in light, resulting in the loss of important information on the shape, color and texture of target objects [33]. Therefore, the performance of such models is poor under complex backgrounds at night.

In order to solve this problem, in the current study, the NIR and visible images of beets and weeds were fused into multi-channel images. In particular, multi-modality image fusion spatially registers the data of the same image from different sources, and subsequently combines the information in each image to generate an integrated data set of all the images [24].

3.2.1 Deep fusion algorithm frame diagram

The visible and near-infrared images (Input1 and Input2) are extracted by a denseblock composed of convolution layers, and then sent to fusion layer for pixel-level superposition. These fusion images are reconstructed through a decoding network also composed of convolutional layers to obtain the three-channel multi-modality images. As shown in Fig. 2, the encoding network consists of two sections: C1 and denseblock. In the C1 section, a 3×3 convolutional kernel is used to extract the rough features of the image. The denseblock section then uses three convolutional layers (the output of each layer is cascaded as the input of the next layer) to extract the high-level abstract semantic features of the image. The denseblock adopted in this research can retain image features as much as possible to ensure that all significant features can be used in the fusion strategy. In the fusion
Fig. 1 Examples of seedling dataset for weeds and beets
Fig. 2 Deep fusion method for multi-modality images
layer, $l_1 – norm$ fusion strategy is selected to fuse the visible and near-infrared images. Finally, four convolutional layers ($3 \times 3$ convolutional kernel) are used to reconstruct the final fusion image in the decoding network. What’s more, in this framework, the size of the input and output images is $1296 \times 1296$ pixels, and the number of feature mapping channels per convolutional layer is 16. More details of the deep fusion algorithm can be found in the study [14].

### 3.2.2 Multi-modality fusion image

Figure 3 presents the three-channel multi-modality image obtained via the depth fusion algorithm. Following the fusion of the data, the LabelImg software (https://tztutlin.github.io/LabelImg/) was used by experts in the agricultural field to label the beet seedlings and weeds based on the PASCAL Visual Object Classes Challenge [10]. A python script was then used to randomly divide the images and the corresponding tag files into training and testing sets at a ratio of 4:1.

### 3.3 Improved weeds and beets detection model

#### 3.3.1 Deformable convolution

Recently, the use of CNNs has made significant breakthroughs in many vision applications. However, due to the regular grid sampling and the fixed geometric structure in traditional convolution methods, it is difficult for networks to deal with geometric deformations. The adaptability of the existing models to process the geometric deformation of objects almost comes from the diversity of the data itself, and there is no mechanism to adapt to geometric deformation in the model. Thus, the ability of geometric transformation modeling is limited, and it cannot be adjusted adaptively according to the image content [8]. In order to overcome this limitation, a new module, deformable convolution, was adopted to improve the modeling ability of CNN for transformations in the current study. More specifically, an offset variable is added to the position of each sampling point in the convolutional kernels. The kernel with the offset variable can then be sampled randomly near the current position, and is thus not restricted to the previous regular grid points. Moreover, the offset variable can be obtained by learning within the target task without the need of an additional monitoring signal, improving the traditional convolution.

Figure 4 shows the sampling methods of traditional convolution and deformable convolution with convolution kernel size $3 \times 3$. Figure 4(a) demonstrates the regular sampling grid (green points) of traditional convolution, while (b) present the deformed sampling locations (black points) with augmented offsets (blue arrows) in deformable convolution. Then (c) and (d) are special cases of (b), showing that deformable convolution generalizes scale, aspect ratio and rotation transformations.

Figure 5 presents the internal structure of deformable convolution. First, the displacement required for deformable convolution is obtained through the output of a small convolutional layer, and the displacement is then applied to the convolutional kernel in order to achieve the effect of deformable convolution. This operation is able to add the offsets to the regular grid sampling locations in the standard convolution, thus enabling the free form deformation of the sampling grid. The offsets are learned from the preceding feature maps via additional convolutional layers.
Fig. 3 Examples of RGB and multi-modality images. The image triplet (a) shows the original visible dataset, and the image triplet (b) shows the corresponding fusion dataset.
Fig. 4 Schematic diagram of deformable convolution
Traditional convolution consists of two steps: (1) Sampling using a regular grid $R$ over the input feature map $x$; and (2) the summation of sampled values weighted by $w$. For each location $p_0$ on the output feature map $y$, traditional convolution is then performed as follows:

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n)$$  \hspace{1cm} (1)

where the grid $R$ defines the receptive field size and dilation of a $3 \times 3$ kernel with dilation 1, $R = \{(-1, -1), (-1, 0), \ldots, (0, 1), (1, 1)\}$, and $p_n$ enumerates the locations in $R$.

In deformable convolution, the grid $R$ has the offsets $\{\Delta p_n \mid n = 1, 2, \ldots, N\}$, and $N = |R|$, thus, formula (1) becomes:

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n + \Delta p_n)$$  \hspace{1cm} (2)

As the offset $\Delta p_n$ is typically fractional, formula (2) is implemented via bilinear interpolation [8].

### 3.3.2 The network structure of the improved model

In the target detection task, it is necessary to classify and locate targets. The classification task needs to increase the translation invariance of the object to classify the target at different positions, while the positioning needs to reduce the translation variation of the object to precisely locate the target position [7]. In order to balance the relationship between the two tasks, location information is fused by constructing position-sensitive score, and all information is fused by adding an RoI pooling layer to score maps. The network structure of the improved model is shown in Fig. 6. Following the input of the image, feature extraction is performed, resulting in a feature map of $k^2 \times (C+1)$ dimensions. The region proposal network (RPN) [23] is then used to extract the regions of interest (ROIs) in the feature map, with $C$ denoting the number of categories. The extracted ROIs are divided into $k \times k$ regions, with $k$ generally equal to 3, corresponding to 9 regions: top-left, top-center, … and bottom-right. Finally, the score of each region is determined by a pooling operation, and the output
Fig. 6. Key idea of the improved R-FCN for weeds and beets detection.
feature vector of the ROIs are then obtained by voting. This vector is subsequently used for classification and regression of the weeds and beets.

4 Experimental environment

4.1 Equipment and platform

The Ubuntu 16.04 system is used as the operating platform, and MXNet is adopted as the deep learning framework to train the network. The computer memory of the system is 32GB, with a 3.6 GHZ i7-9700k CPU processor. Additionally, the 11 GB GeForce GTX1080Ti GPU with Pascal architecture was used.

4.2 Model parameter setting

In order to reduce variation of parameter updates and to stabilize the convergence of the model, a mini-batch stochastic gradient descent (SGD) was used to train the network [30]. The parameters were set as follows: the number of each mini-batch of samples was 128, the momentum factor was fixed to 0.9, and the weight attenuation factor was 0.0005 to avoid over-fitting. Finally, the gradient descending learning rate was applied to all layers of the network, and was gradually reduced in stages to 0.1 times of the current learning rate. Additionally, the initial learning rate during the training process was set to 0.005, and the model was iterated for 100 epochs.

5 Results and discussion

5.1 Performance evaluation of the model

Precision and recall are widely used in the field of information retrieval. As with all machine learning problems, in order to calculate precision and recall, it is necessary to explain the following: True Positive (TP), the number of positive classes predicted as positive classes; False Positive (FP), the number of negative classes predicted as positive classes (error rate); True Negative (TN), the number of negative classes predicted as negative classes; and False Negative (FN), the number of positive classes predicted as negative classes (missing rate). Based on this, we define the following:

\[
Precision = \frac{TP}{TP + FP}\quad (3)
\]

\[
Recall = \frac{TP}{TP + FN}\quad (4)
\]

The average precision (AP) is then calculated as follows:

\[
AP = \int_{0}^{1} p(r)d(r)\quad (5)
\]

where \(p\) represents Precision, \(r\) represents Recall, and \(p\) is a function taking \(r\) as a parameter. The mean average precision (mAP) equals the average of all the AP categories, and is determined as follows:
where classes represents the detected objects, and N is the number of all categories of objects to be detected.

5.2 The impact of multi-modality fusion images on mAP

Since the algorithm proposed in the current study only supports three-channel images as the input data, while the NIR image was composed of a single channel, RGB and multi-modality images (three-channel fusion images) were used in all ablation experiments for verification. As reported in Table 1, the classical region-based fully convolutional network (R-FCN) model and our proposed model exhibited higher detection accuracies on the fusion data set compared with the RGB data set. This is attributed to the high sensitivity of the shallow features of beets and weeds in visible images under the low levels of light, as well as the complex environment. However, the near-infrared image is able to depict the thermal radiation of the target objects. The surface reflectivity of the object was completely different to that of the background, and was thus more robust to variations in light [21]. However, as the near-infrared images are of a low spatial resolution and contain limited texture information, the visible and NIR images were fused into a multi-modality image. These fusion images were sent to the CNN for training. The final detection accuracy of the model greatly improved with the use of the fusion data set, and the detection system was also more robust.

The detection results were visualized to further demonstrate the performance of multi-modality images. As can be seen from Fig. 7, image triplet (a) demonstrates the detection result of the improved R-FCN model on the RGB data set, while image triplet (b) depicts the result of the same model using the multi-modality data set. Due to the fusion of the near-infrared and visible images for the multi-modality image, the feature information of beet seedlings and weeds were able to be better characterized under the terrible light and complex field backgrounds. Thus, the detection performance of the multi-modality fusion images was better than that of the improved R-FCN model. Hence, the subsequent ablation experiments were conducted using on the fusion data set.

5.3 The impact of deformable convolution on mAP

Models 2 and 3 used deformable convolutions, while models 0 and 1 implemented traditional convolutions (Table 2). Compared with the traditional convolution model, the deformable convolution model was able to improve the mAP of beets and weeds by

\[
mAP = \frac{\sum AP}{N(\text{classes})}
\]  

(6)

| Methods          | Backbone      | Types of pictures used for training | APs (%) | mAP (%) |
|------------------|---------------|------------------------------------|---------|---------|
| R-FCN (best)     | Resnet-101    | RGB                                | 85.8    | 78.9    | 82.3    |
|                  |               | Fusion image                       | 88.6    | 80.7    | 84.6    |
| Ours (best)      |               | RGB                                | 87.4    | 79.5    | 83.4    |
|                  |               | Fusion image                       | 93.2    | 84.8    | 89.0    |
3-4% points. This can be attributed to the offset variables of the deformable convolution kernel, allowing for the feature expression of the CNN to automatically adapt to changes in the morphological of the target object.

Figure 8 depicts the convolution results for the multi-modality data set. In Fig. 8(a), the detection results of the traditional convolution method miss some targets. The deformable convolution results (image triplet (b)) exhibited a higher detection accuracy and a lower missing rate than the traditional convolutional model (image triplet (a)). This is a result of the ability of the deformable convolution model to adaptively detect irregular geometric edges of seedlings when detecting smaller target objects.

### Table 2  Model parameter settings and mAP

| Methods  | Number of the models | Types of the convolutional layer | OHEM\(^{[a]}\) | AP\(^{s}(\%\)) weed | AP\(^{s}(\%\)) beet | mAP\(^{(\%)}\) |
|---------|----------------------|----------------------------------|----------------|----------------------|----------------------|----------------|
| R-FCN   | 0                    | Traditional convolution          | O              | 87.9                 | 79.1                 | 83.5           |
|         | 1                    |                                  | Π              | 88.6                 | 80.7                 | 84.6           |
| Ours    | 2                    | Deformable convolution           | O              | 91.0                 | 82.1                 | 86.5           |
|         | 3                    |                                  | Π              | 93.2                 | 84.8                 | 89.0           |

\(^{[a]}\) OHEM = online hard example mining
The impact of OHEM on mAP

OHEM retrains the hard samples with large loss values, as such samples may lead to the misclassification of weeds and beets. Compared to other models without OHEM, the detection accuracy of models 1 and 3 with the OHEM algorithm were improved (Table 2). This indicates that the OHEM method can suppress the simple samples and the small number of samples, making the training process more efficient. In addition, the OHEM algorithm also eliminated several heuristics and hyper-parameters by automatically selecting hard examples, thus simplifying the training process [27].

| Methods   | Backbone    | APs (%) | mAP (%) |
|-----------|-------------|---------|---------|
|           |             | weed    | beet    |         |
| Faster R-CNN | Resnet-101  | 74.6    | 80.3    | 77.4    |
| RetinaNet   |             | 86.3    | 82.7    | 84.5    |
| Yolo V5     | CSPDarknet53| 86.8    | 81.7    | 84.2    |
| Ours        | Resnet-101  | 93.2    | 84.8    | 89.0    |

**Table 3** Performance comparison of different algorithms

Fig. 8 Testing results using deformable convolution and traditional convolution

5.4 The impact of OHEM on mAP

OHEM retrains the hard samples with large loss values, as such samples may lead to the misclassification of weeds and beets. Compared to other models without OHEM, the detection accuracy of models 1 and 3 with the OHEM algorithm were improved (Table 2). This indicates that the OHEM method can suppress the simple samples and the small number of samples, making the training process more efficient. In addition, the OHEM algorithm also eliminated several heuristics and hyper-parameters by automatically selecting hard examples, thus simplifying the training process [27].
5.5 Performance comparison of different algorithms

In order to verify the effectiveness of our model, we compared its performance with the classical algorithm (Faster R-CNN [23], RetinaNet [16]) and the latest algorithm (Yolo V5) of target detection. All comparative experiments used the fusion images. As can be seen from Table 3, the detection performance of Faster R-CNN on weeds was relatively poor, because Non-Maximum Suppression (NMS) was used for post-processing in order to avoid overlapping candidate boxes when RPN generated proposals. However, due to different weed scales and mutual occlusion, it was difficult to detect accurately. The performance of RetinaNet and Yolo V5 were similar, in which Yolo V5 had higher detection accuracy for weeds while RetinaNet had higher detection accuracy for beet.

The detection results were visualized to further demonstrate the performance of our model. As shown in Fig. 9, the Faster R-CNN had serious repeated detection boxes in weed detection, and the same weed was identified as multiple targets. There was not much difference between four algorithms in the detection results of beet. Due to the existence of very small beets, there was a certain amount of omission, so the overall accuracy was lower than that in weed detection accuracy. Our improved model showed optimal performance due to the use of OHEM and deformable convolution.

5.6 Detection results of the optimal model

In order to verify the prediction results of the optimal model under the actual field environment, six images of beets and weeds were selected from the test set. As demonstrated in Fig. 10, with the use of deformable convolution and multi-modality fusion images, our improved model was able to maintain a high detection accuracy. The detection results of beet and weeds with large targets was better while the classification confidence can reach 1. Small targets could also be detected successfully with a confidence level of more than 0.99. This indicated that the optimal proposed model exhibits strong generalization and robustness abilities for the detection of beet seedlings and weeds under the poor light and complex field backgrounds.

6 Conclusions

In the current study, an improved R-FCN model was proposed to detect and identify beet seedlings and weeds under poor light (at night) and complex field backgrounds. Based on the classic R-FCN network, the visible and near-infrared images of beets and weeds were fused into three-channel multi-modality images at pixel-level using a deep fusion algorithm. The fusion images were then sent to the convolutional neural network for training, so as to improve the mean average precision of beets and weeds detection. Furthermore, considering that the traditional convolutional kernel would restrict the geometric modeling ability of the model, deformable convolution was adopted in the feature extraction layer. Moreover, online hard example mining was introduced to excavate the hard negative samples in the detection process to retrain the misidentified samples. Through the aforementioned improvement measures, the average precision of beets and weeds were 84.8% and 93.2% respectively, and the mean average precision was increased from 82.3 to 89.0%. Compared with the original model, the detection accuracy was improved by approximately
Fig. 9. Detection results of fused images by different algorithms.
7% points. Our results demonstrate that the performance of the optimized model is not only greatly improved, but the quantity of the model parameters is also maintained at a reasonable amount. The model can be compressed and deployed to industrial equipment to lay the research foundation for automatic weeding or spraying robots. In the future, the robot can be built to work continuously all day long by collecting images under different lighting conditions for model learning.

Acknowledgements This work is partially supported by a project funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD-2018-87), Synergistic Innovation Center of Jiangsu Modern Agricultural Equipment and Technology (4091600002). Project of Faculty of Agricultural Equipment of Jiangsu University (4121680001).

References

1. Abouzahir S, Sadik M, Sabir E (2021) Bag-of-visual-words-augmented Histogram of Oriented Gradients for efficient weed detection. Biosyst Eng 202:179–194
2. Akbarzadeh P, Ahderom, Apopei A (2018) Plant discrimination by Support Vector Machine classifier based on spectral reflectance. Comput Electron Agric 148:250–258
3. Al-Smadi M, Qawasmeh O, Al-Ayyoub M, Jararweh Y, Gupta B (2017) Deep recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels’ reviews. J Comput Sci 27:386–393
4. Andrea CC, Daniel B, Misael J (2017) Precise weed and maize classification through convolutional neuronal networks. IEEE Second Ecuador Technical Chapters Meeting (ETCM)1–6
5. Baareh AK, Elsayad A, Al-Dhaifallback M (2021) Recognition of splice-junction genetic sequences using random forest and Bayesian optimization. Multimed Tools Appl 2021:1–18
6. Bakhshipour A, Jafari A, Nassiri SM, Zare D (2017) Weed segmentation using texture features extracted from wavelet sub-images. Biosyst Eng 157:1–12
7. Dai J, Li Y, He K, Sun J (2016) R-FCN: Object detection via region-based fully convolutional networks. In: Advances in Neural Information Processing Systems, pp 379–387
8. Dai J, Qi H, Xiong Y, Li Y, Zhang G et al (2017) Deformable convolutional networks. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp 764-773
9. Dinesh Kumar JR, Ganesh Babu C, Priyadharsini K (2021) An experimental investigation to spotting the weeds in rice field using deepnet. Mater Today: Proc 45:8041-53
10. Everingham M, Eslami S, Gool LV (2015) The pascal visual object classes challenge: a retrospective. Int J Comput Vis 111:98–136
11. Garcia B, Mylonas N, Athanasakos L, Fountas S (2020) Improving weeds identification with a repository of agricultural pre-trained deep neural networks. Comput Electron Agric 175:105593
12. Huang H, Deng J, Lan Y, Yang A, Deng X et al (2018) A fully convolutional network for weed mapping of unmanned aerial vehicle (UAV) imagery. PLoS ONE 13:e0196302
13. Jiang H, Wang P, Zhang Z, Mao W, Zhao B et al (2018) Fast identification of field weeds based on deep convolutional network and binary hash code. Trans Chin Soc Agric Mach 49:30–38
14. Li H, Wu X (2019) DenseFuse: A fusion approach to infrared and visible images. IEEE Trans Image Process 28(5):2614–2623
15. Li B, Bai B, Han C (2020) Upper body motion recognition based on key frame and random forest regression. Multimed Tools Appl 79:5197–5212
16. Lin T, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In: IEEE Trans Pattern Anal Mach Intell, pp 2999-3007
17. Maswadi K, Ghani NA, Hamid S, Rasheed MB (2021) Human activity classification using Decision Tree and Naïve Bayes classifiers. Multimed Tools Appl 80:21709–21726
18. Milio A, Lottes P, Stachniss C (2018) Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in CNNs. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp 2229-2235
19. Nogueira K, Penatti OAB, dos Santos JA (2017) Towards better exploiting convolutional neural networks for remote sensing scene classification. Pattern Recogn 61:539–556
20. Pearse GD, Tan AYS, Watt MS, Franz MO, Dash JP (2020) Detecting and mapping tree seedlings in UAV imagery using convolutional neural networks and field-verified data. ISPRS J Photogramm Remote Sens 168:156–169
21. Raghavendra R, Dorizzi B, Rao A, Kumar GH (2011) Particle swarm optimization based fusion of near infrared and visible images for improved face verification. Pattern Recogn 44:401–411
22. Raja R, Nguyen TT, Slaughter DC, Fennimore SA (2020) Real-time weed-crop classification and localisation technique for robotic weed control in lettuce. Biosyst Eng 192:257–274
23. Ren S, He K, Girshick R, Sun J (2017) Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39:1137–1149
24. Ren X, Meng F, Hu T, Liu Z, Wang C (2018) Infrared-visible image fusion based on Convolutional Neural Networks (CNN). Intelligence Science and Big Data Engineering; 301–307
25. Sandooval-Insauti H, Chiu YH, Dong HL, Wang S, Chavarro JE (2021) Intake of fruits and vegetables by pesticide residue status in relation to cancer risk. Environ Int 156:106744
26. Shin HC, Roth HR, Gao M, Lu L, Xu Z et al (2016) Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 35:1285–1298
27. Shrivastava A, Gupta A, Girshick R (2016) Training region-based object detectors with online hard example mining. In: IEEE Computer Society, pp 761-769
28. Sun J, He X, Tan W, Wu X, Shen J et al (2018) Recognition of crop seedling and weed recognition based on dilated convolution and global pooling in CNN. Trans Chin Soc Agric Eng 34:459–465
29. Wang H, Li Z, Yang L, Gupta BB, Chang C (2018) Visual saliency guided complex image retrieval. Pattern Recognit Lett 130:64–72
30. Wang T, Knap J (2020) Stochastic gradient descent for Semilinear elliptic equations with uncertainties. J Comput Phys 426:109945
31. Wu G, Li Y (2021) CyclicNet: an alternately updated network for semantic segmentation. Multimed Tools Appl 80:3213–3227
32. Yan B (2018) Identification of weeds in maize seedling stage by machine vision technology. J Agric Mechanization Res 40:212–216
33. Ying Z, Ge L, Ren Y, Wang R, Wang W (2017) A new image contrast enhancement algorithm using exposure fusion framework. In: Presented at International Conference on Computer Analysis of Images and Patterns, pp 36-46
34. Zhang J, Li M, Feng Y, Yang C (2020) Robotic grasp detection based on image processing and random forest. Multimed Tools Appl 79:2427–2446
35. Zhao P, Wei X (2014) Weed recognition in agricultural field using multiple feature fusions. Trans Chin Soc Agric Mach 45:275–281

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.