An image analysis technique for wheat head count detection using machine learning

Ramadevi Kalluri, Prabha Selvaraj
School of Computer Science Engineering, Vellore Institute of Technology VIT-AP, Amaravati, India

ABSTRACT

Deep learning (DL) techniques have significantly enhanced the potential for wheat head detection in recent times. The different development phases of canopy, genotype, wheat heads, and wheat head orientation provide considerable obstacles. The overlapping density of wheat heads and wind-induced picture blurring complicate wheat head recognition. This study describes an effective wheat head detection and counting method. Due to its high throughput in agricultural field analysis, remote sensing phenotyping has grown in popularity. Applying DL methods for image processing and other technological advancements has increased the scope for the quantitative evaluation of various crop traits. The ability to detect and characterize wheat heads in the industry is an important part of the wheat breeding process for selecting high-yielding cultivars. The proposed method uses the Mask region-based convolutional neural network (RCNN) framework to detect and classify the wheat ears. The complete detection task is done in three steps: region proposal generation, region of interest alignment, and mask generation. The global wheat head detection (GWHD) dataset is used for the experimental analysis of the dataset. The proposed method achieved an accuracy of 95.11% on the GWHD dataset, demonstrating its effectiveness in wheat head detection and classification tasks.

Keywords: Mask region-based convolutional neural network Phenotyping Region of interest alignment Wheat head Wheat spikelet

This is an open access article under the CC BY-SA license.

Corresponding Author:
Prabha Selvaraj
School of Computer Science Engineering, Vellore Institute of Technology, VIT-AP University
G-30, Inavolu, Beside AP Secretariat Amaravati, Andhra Pradesh-522237, India
Email: Prabha.s@vitap.ac.in

1. INTRODUCTION

Wheat, rice, and maize are the most common grain crop in the world. After Norman Borlaug produced the semi-dwarf wheat species in the 1950s and added a method of agricultural science double-green revolution, it saved 300 million people in World War II [1]. However, after having grown to a staggering few years, wheat yields have declined since the mid-1990s. A few wheat products will most likely be in your pantry. This cereal can be used in your morning toast and cereal. Wheat is well-studied based on its needs, such as food and agricultural crop residues. The crop uses scientists notice the images of the “wheat-heads,” an outlier at the top of the plant, save the seed, to obtain large and detailed information about the wheat in the fields worldwide. These images are estimates of the density, the size of the wheat, and the seeds of the different varieties [2]. The farmers can use this information to determine the condition and situation in the country and make management decisions. A guide to note is still widely used in traditional breeding.

Genomic control, a new high-performance phenotyping technique, or a combination can help you improve the genetic advantage. This method is the choice of the importance of the wheat characteristics, such as yield, disease resistance, tolerance, and adaptation to abiotic stresses. Developing an effective and...
sustainable model to reproduce the current raw data is challenging. It has become a reality even for the high-throughput accumulation of phenotypic data. The density of the wheat head (the number of wheat ears per unit area) is the most important element, which is still manual to be evaluated and selected smaqlard, which is time-consuming and will lead to a measurement error of 10%. To help the breeders to manipulate the balance between the ingredients of the capital, many of the plants, the density, spike, grains, and per capita grain mass) in the breeding choices, you'll have to imagine the methods to improve productivity and accuracy in calculating the wheat ears in the field.

Deep learning is emerging as a contemporary solution for numerous computer vision tasks, encompassing object recognition, segmentation (such as semantic segmentation), image regression models, leveraging the efficiency of graphics processing units (GPUs) and benefiting from the abundance of large datasets. A number of authors have recently suggested a deep learning approach utilizing image data for plant phenotyping [3]. The quantification of wheat production with high-resolution red, green, and blue (RGB) images has been proposed using various methods. With a faster-than-RCNN object detection network, the authors demonstrate the ability to identify wheat in the early stage. In the resolution, the surveillance, the value of the relative calculation error of about 10% to allow for such a practice. The authors have developed a convolutional neural network (CNN)-coder-decoder model [4], [5], which has surpassed the traditional, manual computer methods regarding the semantic segmentation of a wheat field. To describe the movement of the wheat plants in the field, the researchers used a model for developing the wheat crop in the discovery and probabilistic tracking.

Even though in the past, the research, testing, and methods of detection were used for the individual units, the deep learning model is information that is difficult to scale in a real-time phenotyping platform [6], in practice, trained as they are in the small details of the fixtures, which causes the expected difficulties to extrapolate to new situations. Most training data are limited regarding the genotype, geographic area, and observation of the conditions. The morphology of the prime of wheat can vary widely between genotypes to provide the differences in the size, grade, paint, and awns. Then, due to the species, the heads of the background and the trees change dramatically, depending on the pay and population aging. In addition, the density and the trends vary across the globe, with various farming systems and expert start-ups, in the prime of wheat, and areas with higher crop density, often cover and overlap.

For modest data sets, training a CNN-based model on a portion of field-testing phenotyping and testing the rest of the area is common. To yet, there is no theoretical assurance that the CNN model [7] is a basic flaw of counter cause-effect model empirical approaches. In addition, comparing the approaches of various authors requires large amounts of data. Unfortunately, complex and large-scale phenotypic head-count data do not exist today, as they usually gather in front of non-governmental organizations, which will limit the number of genotypic and environmental conditions and observations used to train and evaluate the model. In addition, since the labeling process takes a long time, and has been repeated, only a small fraction of the images is processed. We have established the global wheat head of the detection system (GWHD) datasets [8]–[10] to meet the need for a broad and diverse community to live in, a Wheat Head, datasets with a clear record, which you can use to compare the methods provided by the computer vision community. The GWHD data set is created by combining the data from nine organizations spread across seven countries and three continents worldwide.

Ensuring long-term food security is a top priority, and one way to achieve this is by prioritizing the development of global wheat production. To improve wheat production, innovative plant breeding methods must be used to create new wheat varieties with disease resistance, climatic resilience, and higher yields [11]. However, conventional wheat breeding methods persist, predominantly manual and susceptible to errors. The process is extremely time-consuming and tedious, which can hinder progress. Plant phenotyping is one method that can help overcome these limitations. Various structural and functional plant features are assessed to discover wheat traits related to yield potential, disease resistance, or abiotic stress adaptation [12]. Assessing wheat head count per unit ground area is crucial to breeding trail yield evaluation and is presently done manually [13]. To choose which wheat types to cross for a better progeny, plant breeders depend significantly on the wheat head count.

To overcome human wheat head identification constraints in wheat breeding, automated detection is necessary. Several studies have used computer vision and image processing to identify wheat heads [14], [15]. Environmental and inherent wheat issues remain. Environmental difficulties include wind or motion blurring, observing circumstances, picture scale mismatches, and undesired shadow and brightness [16]. However, wheat-specific problems such genotype-induced variances in wheat head forms and colors, development phases, overlaps, and orientation make this endeavor more difficult. Accurately and automatically recognizing and counting wheat heads in the field may help yield estimate under field settings [17]. Due to the variety of cultivation methods and appearances, generating an accurate dataset is difficult. Also difficult is creating lightweight deep learning models for edge device applications to recognize wheat.
An image analysis technique for wheat head count detection using machine learning (Ramadevi Kalluri)
increase the capacity of the backbone network, it can be used in the past for the convenience of the inter-
stage network (CSPNet). Bhagat et al. [31] proposed a method to improve CNN training. This network will
improve CNN learning and minimize competition by 20%. We will deploy many CSPNet networks to boost
the backbone network's learning capacity. Figures 1 and 2 demonstrate the flow diagram and network
architecture of CSPDARKNET 53. It mostly consists of skip connections and keeps the DarkNet53 topology.
Data transfer was the most critical aspect of the concat layer network and one of the few remaining blocks to
build up. This is SPP; i) the core network, namely SPP-2. SPP-2 also contradicts SPP-1’s goal. This section
highlights the integration of global knowledge about the function and its functions to simplify the network
neck. From the top of the layer, SPP-2 convolutionally processes things. And then it will operate on
maximum connection with varying sizes of activities. Size of pool, k-1’s, swimming pools, and ii) swimming
pools, 3, 5, 7, and 13. And the step-size is 4, 6, and 13. SPP-2, to integrate the three common shares' features
and represent them to the next convolutional module to provide instruction and training to gain a lot of local.

![Figure 1. Architectural flow of CSPDARKNET 53](image1)

![Figure 2. The structure of CSPDARKNET 53](image2)

3. METHOD

Mask RCNN is a variant of CNN and state-of-the-art in terms of image segmentation. The key
characteristic of deep NN algorithms is object detection and segmentation [33]. Mask RCNN algorithm is
generally applied for object detection task that precisely detects the object location. The working principle of
Mask RCNN is theoretically simple to understand as shown in Figure 3.

The working is related to a faster RCNN algorithm, i.e., it has 2 output data, one for classification
label and the other for bounding box; and in Mask RCNN, one more output is added, i.e., object mask. The
difference is that masked outcome varies with labels and box outcome, and it requires three-dimensional
feature extraction than the essential components are included, i.e., a pel-to-pel arrangement which is the
drawback of faster RCNN.

As seen in the Figure 4, it is a three steps process, i.e., region of interest (ROI) align, boundary box,
and mask labeling with positive labels. The first phase creates region suggestions using the RPN framework,
while the second generates ROI classification labels, bounding boxes, and masks. Mask R-CNN predicts the
class and box offset while outputting a binary mask for each ROI, unlike most previous algorithms. This method
parallelizes bounding-box classification and regression like Fast R-CNN. Original R-CNN's multi-stage pipeline is simplified by this method. Multi-task loss is determined on each sampled ROI during training.

\[ L = L_C + L_B + L_M \]

![Architecture of Mask RCNN](image)

The classification and bounding-box losses \( L_C \) and \( L_B \) are from [9]. \( K \times m^2 \) dimensional outputs from the mask branch encode \( K \) binary masks of resolution \( m \times m \) for each ROI, one for each of the \( K \) classes. A per-pixel sigmoid and average binary cross-entropy loss (\( L_{Mis} \)) are used. For a ROI with ground-truth class \( k \), \( L_M \) is only specified on the \( k \)-th mask since other mask outputs do not contribute to the loss. \( L_M \) lets the network create masks for all classes without competition. The output mask-selected classification branch predicts the class label. Friendships, commerce, and navigations (FCNs) are often used for semantic segmentation; however, this method decouples mask and class prediction. If per-pixel softmax and multinomial cross-entropy loss are used, masks across classes compete. No competition exists with a per-pixel sigmoid and binary loss. For excellent instance segmentation, per-pixel sigmoid and binary loss formulation are necessary.

![Representation of classification and labeling of wheat head using Mask-RCNN](image)

When processing an input object, a mask can be used to represent its spatial layout. Compared to class labels or box offsets, which are typically transformed into compact output vectors by fully-connected layers, masks can be more naturally analyzed using convolutional layers that enable pixel-to-pixel correspondence. Convolutional layers scan the input image in small regions, allowing them to capture the spatial structure of masks at a high level of detail. This makes masks a powerful tool for object detection and segmentation tasks, as they can accurately indicate the exact location and shape of an object in an image, facilitating its identification and tracking. By leveraging the spatial information provided by masks, CNN can achieve superior performance in various computer vision applications.

*An image analysis technique for wheat head count detection using machine learning (Ramadevi Kalluri)*
4. WORKING OF MASK REGION-BASED CONVOLUTIONAL NEURAL NETWORK

A deep learning model called Mask R-CNN detects and segments objects in several areas. The model is very good for wheat head detection in images. Initial bounding boxes for picture objects are generated using RPN. The RPN is a tiny CNN that predicts item probability and bounding box coordinates. The RPN creates candidate bounding boxes by moving a window across the picture and using the convolutional layers' visual characteristics to forecast if an item resides there. The RPN then generates ideas with scores reflecting object probability. Mask R-CNN uses a segmentation head to predict binary masks for each candidate bounding box after the RPN creates them. The segmentation mask detects object pixels in the bounding box. The segmentation head is a CNN that predicts item masks from image and bounding box coordinates. The segmentation head convolutionally processes visual characteristics from the convolutional layers to anticipate an object's binary mask. A binary picture the same size as the bounding box, the mask shows which pixels are object pixels and which are not. Mask R-CNN creates candidate bounding boxes for all picture objects to identify wheat heads.

It then predicts a binary mask for each bounding box using the segmentation head. If the mask indicates that the bounding box contains a wheat head, Mask R-CNN considers the bounding box as a detection. First step: feature maps of feature pyramid network (FPN) are taken as input and generate bounding boxes as rectangular objects with labels. Region proposals are produced through sliding window over the image, predicting several regions and representing it as k, where these generated regions concerning boundary boxes are known as anchors. Each feature map of a convolutional layer with size cap Wxcap H, cap Wxcap H, x k anchors is generated and labeled through the label to demonstrate if it contains an object. An anchor is labeled with a positive number for the following two conditions: i) if the anchor overlaps through the ground truth boundary box with maximum IoU (i.e., Intersection of Union) and ii) if the anchor's IoU value is more than 0.7.

In the two conditions, the second one is commonly utilized to identify the anchor labels, as assigning positive labels for several anchors as a single ground-truth value is possible. The anchors are labeled as negative when the anchor IoU value >3. Set of unlabeled anchors is excluded and not considered while training.

Second step: in this step, ROI alignment is done by feature map extraction from RPN-generated region proposals. Faster and fast RCNN utilized the pool of ROI because of the false alignments among ROI and features extracted. This step, i.e., ROI alignment, is designed to extend faster and fast RCNN. Bi-linear intersection method is employed to calculate the input features such as texture, shape, color, and size at frequently sampled positions, i.e., 't' value for all ROI, and then all the outputs are coupled through average or max pooling.

Third step: is mask generation; this step extends faster and fast RCNN, providing additional details of a given image. Generation of mask works well and assists in detecting objects as it produces the specific object position in an image contrary to faster and fast RCNN algorithm. Pel-to-Pel representation of the ROI is used, for instance, image segmentation. FCN generates a mask for all ROI of m × m size by considering the spatial features instead of changing features to vectors as manifestly in spatial features. FCN is applied for the segmentation task, and it contains an input layer of size h × w × d, where d represents the feature, and w and h are spatial dimensions. Bi-linear interpolation method is applied to link the outcome of hidden network pixels. The stride convolution is the method to convert minimized feature maps to higher-dimension feature maps. Then to produce a binary mask, the feature maps should be equal to the input image size, and it is made through convolutional up-sampling over the 'f' factor and $1/f$ fractional stride and this convolutional up-sampling procedure is called deconvolution or backward convolution.

Generating accurate bounding boxes for small, occluded, or cluttered objects using RPN can be challenging. However, to overcome this issue, one technique that has been employed is the use of a CNN extracting image features. This approach has proven to be effective in enhancing the accuracy of RPN. ROI alignment is also challenging when dealing with rotated or skewed ROIs, geometric transformations are used to rotate and deskew the ROIs to address this issue.

5. RESULTS AND DISCUSSION

Dataset description: The GWHD dataset is used for the experimental analysis which contains labelled images that were collected between 2016 and 2019 by nine institutions across ten different locations. The images cover a diverse range of genotypes that are representative of Europe, North America, Australia, and Asia. The dataset includes images from a broad range of environments, including irrigated and rainfed conditions, different soil types, and diverse climatic conditions. It is one of the largest wheat head detection datasets consisting of varieties of wheat field visual images cultivated in various territories. Labeling wide
An image analysis technique for wheat head count detection using machine learning (Ramadevi Kalluri)

Bounding boxes is a complicated task. Hence, it's necessary to figure out the box, including total wheat head image pixels, and the labeled region holds a minimum of one wheat head. The GWHD dataset is a comprehensive collection of RGB images that have been captured using diverse ground-based phenotyping platforms and cameras. The images have been taken from a range of heights, varying from 1.8 m to 3 m above the ground. The cameras used in the image acquisition process have different focal lengths ranging from 10 mm to 50 mm and sensor sizes that vary across the range of cameras. As a result, the ground sampling distance (GSD) of the images varies from 0.10 to 0.62 mm, covering a broad range of resolutions. The half field of view along the image diagonal ranges between 10° to 46°, providing a wide-angle coverage of the scene. The GSD of the images is high enough to detect wheat heads and even awns visually, assuming they are 1.5 cm in diameter. This implies that the dataset is useful for tasks like head detection, counting, and size estimation. However, the varying camera setups and acquisition heights may introduce geometric distortion in a few sub-datasets, which should be taken into account while processing the images.

Model specification and implementation details: to develop and train the system for wheat head detection, we utilized Google Colab, an online platform that offered GPU specifications to facilitate the processing of large amounts of data. For programming, we used Python 3 and a variety of libraries specifically designed for pre-processing, segmentation, and detection of wheat heads. These libraries enabled us to manipulate and analyze the data in an efficient and streamlined manner, making it easier for us to train the system and achieve accurate results. Overall, the combination of Google Colab, Python 3, and specialized libraries provided us with the necessary tools and resources to develop a high-quality wheat head detection system.

Proposed model is trained with 50 epochs. The learning rate was tuned to check whether the model was overfitting. In pre-processing, the mean pixel is calculated on the training set, and these mean pixels are subtracted from the training, testing, and validation dataset to ensure that the images are zero-centering. The mean subtraction helps the model to converge speed on the domain data and improves the accuracy. The heads of the model were trained first with a 0.001 learning rate for the first 30 epochs for the entire dataset per epoch. To develop the model, various hyperparameters were adjusted, including the number of filters, convolutional blocks, filter size, pooling size, learning rate, dropout rate, number of epochs, optimizer, and batch size. For more detailed information on the hyperparameter settings utilized during model development, please refer to Table 1.

<table>
<thead>
<tr>
<th>Hyperparameters</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimizer</td>
<td>Adam</td>
</tr>
<tr>
<td>Activation function</td>
<td>ReLU</td>
</tr>
<tr>
<td>Learning rate</td>
<td>0.001</td>
</tr>
<tr>
<td>Dropout rate</td>
<td>0.1</td>
</tr>
<tr>
<td>Number of epochs</td>
<td>50</td>
</tr>
<tr>
<td>Number of neurons</td>
<td>20</td>
</tr>
<tr>
<td>Batch size</td>
<td>32</td>
</tr>
<tr>
<td>Filter size</td>
<td>3x3</td>
</tr>
<tr>
<td>Pooling size</td>
<td>3x3</td>
</tr>
<tr>
<td>Strides</td>
<td>1x1</td>
</tr>
<tr>
<td>Batch size</td>
<td>16</td>
</tr>
</tbody>
</table>

Evaluation metrics: for experimental analysis of a proposed model for wheat head detection, various metrics like recall, precision, accuracy, and mAP.

- Precision: true positives (TP)/all positives ratio. TP samples are projected to be positive. False positive (FP) samples are anticipated positive but negative. Predicted negative samples are often called false negatives (FN).

\[
\text{precision} = \frac{TP}{TP+FN}
\]

- Recall: is the measure of correctly identifying TP. Thus, for the proposed model, it represents the wheat heads in an image. Thus recall tells us how the proposed model correctly identified the wheat heads. Mathematically it is represented as follows:

\[
\text{recall} = \frac{TP}{TP+FN}
\]
Accuracy: is a metric for evaluating classification models. Informally, accuracy is the fraction of predictions, mathematically it is represented as follows:

\[
\text{accuracy} = \frac{\text{number of correct predictions}}{\text{total number of predictions}}
\]

For binary classification, accuracy can also be calculated in terms of positives and negatives as follows:

\[
\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]

- mAP: means adding the average accuracy of all categories and dividing by the number of categories:

\[
mAP = \frac{\sum_{i=1}^{N} AP_i}{N}
\]

where N is the number of object classification.

Performance of the proposed model, i.e., mask RCNN on GWHD dataset, obtained 0.97% precision, 0.95% recall, 95% accuracy and 96.47% mAP with 0.001 learning rate as demonstrated in Table 2 and Figures 5 to 7 respectively. Table 3 demonstrates the comparison of the proposed model with YOLO3, YOLO4, fast RCNN, and faster RCNN, and it represents that the proposed model recorded better results for detecting wheat heads.

Table 2. Experimental results of wheat heads detection using various metrics like recall, precision, accuracy, and mAP

<table>
<thead>
<tr>
<th>Wheat head detection</th>
<th>Mask RCNN</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.97%</td>
<td>0.95%</td>
<td>95.11%</td>
<td>96.47%</td>
</tr>
</tbody>
</table>

Figure 5. precision and recall obtained for a proposed model for detecting wheat head using Mask RCNN

Figure 6. Accuracy obtained for a proposed model for detecting wheat heads using Mask RCNN
An image analysis technique for wheat head count detection using machine learning (Ramadevi Kalluri)

6. CONCLUSION

Mask RCNN is used to train and assess a deep learning-based wheat head detection technique using the GWHD dataset. The strategy based on fast RCNN and faster RCNN models outperformed other models in wheat head identification and classification. RPN, ROI alignment, and mask creation comprise the suggested technique. The region proposal network creates wheat head-containing potential areas initially. The ROI alignment stage aligns these candidate areas to a predetermined size and feeds them to a CNN for feature extraction. Finally, the mask creation stage generates a binary mask for each candidate area indicating wheat head presence or absence. The suggested strategy improved wheat head recognition and classification on the GWHD dataset with a mean average precision (mAP) of 96.47% and accuracy of 95.11%. In future work, the model may be made a self-attention machine to improve detector feature learning.

ACKNOWLEDGEMENTS

The authors also thank to family and friends for their unwavering support, VIT-AP for providing resources, and the librarians and support staff for their assistance. This research is a collective effort, and be grateful for the contributions of each individual and entity mentioned.
REFERENCES


An image analysis technique for wheat head count detection using machine learning (Ramadevi Kalluri)

BIOGRAPHIES OF AUTHORS

Ramadevi Kalluri is a research scholar pursuing Ph.D. in the Department of Computer Science and Engineering in VIT-AP Campus, Amaravathi, Andhra Pradesh. She has completed her B.Tech. (Computer Science and Engineering) from VR Siddhartha Engineering College, Vijayawada in 2015. M.Tech. (Computer Science and Engineering) from Sri Padmavathi Mahila Viswavidyalayam, Tirupati in 2017. She has a teaching experience of about 4 years. Her research interests include IoT and image processing. She can be contacted at email: Ramadevi.20phd7007@vitap.ac.in.

Prabha Selvaraj is working as a professor in the Department of Computer Science and Engineering in VIT-AP Campus, Amaravathi, Andhra Pradesh. She has completed her B.E. (Computer Science and Engineering) from Kongu Engineering College, Perundurai in 1998. M.E. (Computer Science and Engineering) from Anna University, Chennai in 2005. She has completed her Ph.D. programme under the area Data Mining in Anna University, Chennai in the year 2016. She has a teaching experience of about 20 years. Her research interests include IoT, wireless sensor network, database systems, system modelling, compiler design, network security, data mining and information retrieval system. She has published papers in various national and international journals and conferences. She is a life member of ISTE and member of CSI. She can be contacted at email: Prabha.s@vitap.ac.in.