Data in Brief 55 (2024) 110585

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Comprehensive wheat coccinellid detection dataset: Essential resource for digital entomology

Ivan Grijalva[a,*], Nicholas Clark[a], Emma Hamilton[a], Carson Orpin[b], Carmen Perez[c], James Schaefer[d], Kaylynn Vogts[e], Brian McCornack[a]

[a] Department of Entomology, Kansas State University, Manhattan, KS 66506, USA
[b] Department of Agronomy, Kansas State University, Manhattan, KS 66506, USA
[c] Department of Horticulture Natural Resources, Kansas State University, Manhattan, KS 66506, USA
[d] Department of Biology, Kansas State University, Manhattan, KS 66506, USA
[e] Department of Animal Science and Industry, Kansas State University, Manhattan, KS 66506, USA

ARTICLE INFO

Article history:
Received 16 February 2024
Revised 7 May 2024
Accepted 29 May 2024
Available online 4 June 2024

Dataset link: Wheat_Coccinellid (Original data)

Keywords:
Machine learning
Automation
Cereal
Lady beetles

ABSTRACT

Wheat (*Triticum aestivum*) is a major cereal crop planted in the Southern Great Plains. This crop faces diverse pests that can affect their development and reduce yield productivity. For example, aphids are a significant pest in wheat, and their management relies on pesticides, which affect the sustainability and biodiversity of natural predators that prey on aphids. Coccinellids, commonly named lady beetles, are the most abundant natural predators of wheat. These natural enemies contribute to the natural predation of aphids, which can reduce the use of excessive pesticides for aphid management. Usually, visual observations of these natural enemies are performed during pest sampling; however, it is time-consuming and requires manual labor, which can be expensive. An automation system or detection models based on machine learning approaches that can detect these insects is needed to reduce unnecessary pesticide applications and manual labor costs. However, developing an automation system or computer vision models that automatically

* Corresponding author.
E-mail address: grijalva@ksu.edu (I. Grijalva).
Social media: @iagrite (I. Grijalva)

https://doi.org/10.1016/j.dib.2024.110585

detect these natural enemies requires imagery to train and validate this cutting-edge technology. To solve this research problem, we collected this dataset, which includes images and label annotations to help researchers and students develop this technology that can benefit wheat growers and science to understand the capabilities of automation in Entomology. We collected a dataset using mobile devices, which included a diverse range of coccinellids on wheat images. The dataset consists of 2,133 images with a standard size of 640 × 640 pixels, which can be used to train and develop detection models for machine learning purposes. In addition, the dataset includes annotated labels that can be used for training models within the YOLO family or others, which have been proven to detect small insects in crops. Our dataset will increase the understanding of machine learning capabilities in entomology, precision agriculture, education, and crop pest management decisions.

Specifications Table

Subject	Agricultural sciences (agronomy and crop science). Biological sciences (entomology and insect science). Computer science (artificial intelligence, computer science applications, computer vision, and pattern recognition). Data science (applied machine learning). Environmental science (ecology).
Specific subject area	Our dataset contributes mainly to the subject of entomology with the integration of computer and data science, specifically in integrated pest management, insect biodiversity, agricultural automation, and computer vision models based on machine learning approaches.
Data format	The dataset includes images in a .JPG format at 640 × 640 pixels with annotated labels, specifically bounding boxes in .txt format. The dataset includes 2133 images with 2133 annotated labels. The dataset is ready to train detection models or develop automated monitoring systems using images.
Type of data	The dataset includes images and annotated labels of each coccinellid found in each image. The background of the images is mainly wheat plants.
Data collection	The dataset was collected across subplots planted with different wheat varieties in the North Agronomy Farm at Kansas State University during normal weather conditions. The subplots were located at GPS coordinates of 39°12′16.7″N 96°35′50.3″W ( ). Each sampler performed transects of 1 m within row alleys of subplots to capture individual images with coccinellid presence on wheat plants using mobile devices, specifically iPhone models of 6, 10, 12 pro, 13, and a Samsung model S21 FE 5G. All images collected include a background of wheat plants at different camera angles (Fig. 2), which helps diversify the dataset for training processes and obtain better accuracy results of detection using computer vision models. Each image captured includes a different number of coccinellids, ranging from 1 to 3 coccinellids per image. All collected datasets were processed and cleaned to avoid blurry images or images where coccinellid are absent. In addition, images were processed at a standard size of 640 × 640 pixels using the Adobe Bridge software, and manually annotated using bounding boxes found in the label toolbox of the cloud labeling environment Roboflow [1]. The dataset includes images and annotated labels ready for training computer vision models that can be deployed into web applications, automated monitoring systems or robotics for field scouting.
Data source location	Institution: Kansas State University, North Agronomy Farm. City, Region: Manhattan, Flint Hills. Country: United States. GPS coordinates for collected dataset: 39°12′16.7″N 96°35′50.3″W.
Data accessibility	Repository name: Mendeley data repository. Data identification number: 10.17632/j9735xjspw.1 Direct URL to data: https://data.mendeley.com/datasets/j9735xjspw/1

1. Value of the Data

- The dataset contributes to build and develop computer vision models and applications using machine learning approaches, including web-mobile applications and robotics for field scouting. In addition, the dataset helps integrate computer science with entomology, which can increase the automation of different activities, such as detecting insects, counting, and pest management decisions, which are common activities in entomology research and pest management in crops.
- The dataset is needed to develop computer vision models that can reduce the sampling time of common coccinellids in wheat fields, a time-consuming activity that can be solved with automation. Furthermore, the dataset contributes to an alternative sampling method, in which images can be used to acquire information about the coccinellid population in wheat fields to help management decisions instead of using manual observations that can be subjective during field sampling.
- Using machine learning approaches and this dataset, researchers and students can use the dataset to train, validate, and test computer vision models. This collected dataset and label annotations will reduce the development time of this cutting-edge technology.
- The dataset contributes to developing automated monitoring systems, such as robotic vehicles, that can detect coccinellids using sensors and reduce unnecessary pesticide applications for aphid management in wheat.
- The dataset can be used in outreach events and educational courses. Specifically, increase the knowledge of coccinellids and their importance for agriculture in controlling pests.
- The dataset adds research value to the implications of automation that can result in the precise management of crops in current agriculture.

2. Background

Coccinellids are natural predators of many crop pests, especially aphids [2,3]. Gathering images of these insects under field conditions can be used to train computer vision models to detect coccinellids, which can decrease the time of sampling in wheat using automation. For example, computer vision models have been developed to detect pests [4–7] and natural enemies [2,8] in other crop systems. The critical problem in developing this cutting-edge technology is the need for imagery to apply computer vision models that use machine learning approaches. Additionally, the generation of a training dataset with label annotations can be a time-consuming and expensive task to perform (Fig. 1).

Our dataset solves these issues and provides a comprehensive resource to train and validate computer vision models that, with further development, can be deployed into web-mobile applications and robotics for field scouting. Additionally, training of computer vision models with this dataset, wheat growers can make informed decisions on aphid management when coccinellids are present to avoid unnecessary pesticide application in their crops. Usually, a high population of coccinellids can decrease the presence of aphids [3] when aphids do not reach an economic threshold level for spraying.

3. Data Description

The dataset can be used to train computer vision models, which include image files with a .JPG format and were carefully reviewed to ensure quality (e.g., no blurriness) and coccinellid presence. The dataset includes 1–3 coccinellids per image, and most of the backgrounds are wheat plants (Fig. 2). The dataset was standardized to 640 $\times$ 640 pixels using Adobe Bridge software. Each image file was carefully annotated using a bounding box found in the labeling toolbox of the cloud labeling environment Roboflow [1] (Fig. 3). The label annotations of the 2133 images are in .txt format and are included with the imagery collected. The structure of

Fig. 1. Aerial image of the wheat subplots located at GPS coordinates of 39°12′16.7″N 96°35′50.3″W.

Fig. 2. Image examples with coccinellid presence on wheat plants.

the dataset includes a main zipped folder titled "Wheat_Coccinellid," with a subsequent folder titled "train" and subfolders titled "images," which include 2133 images in conjunction with a folder titled "labels," which include 2133 label annotations (Fig. 4). The entire dataset is ready to be used for training computer vision models and further deployment into automatic detection systems or robotics for field sampling. Further below, we show specific steps that provide an overview of the imagery collected, the processing and annotations process:

1. We downloaded all images (raw data) taken in the replicated wheat subplots from the mobile devices to a central server.
2. We carefully evaluated the dataset by deleting any undesired image (e.g., blurriness, mistaken insect, or no coccinellid presence).
3. We resized the images to 640 × 640 pixels using Adobe Bridge software.

Fig. 3. Images annotated with a bounding box using the cloud labeling environment Roboflow [1].

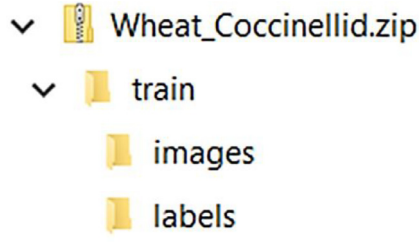

Fig. 4. Directory structure of the dataset.

4. We imported the dataset to the cloud labeling environment Roboflow [1]. This software helps annotate images and is suitable for training computer vision models. Each image was annotated with a bounding box to detect coccinellids within an image.
5. We exported the dataset with all the 2133 images and 2133 label annotations. Then, the dataset was published into the Mendeley data repository for public use.

4. Experimental Design, Materials and Methods

4.1. Experimental design

Sampler activities consisted of transects of 1 m within row alleys of replicated subplots of winter wheat. The subplots were on a randomized complete block design planted with 12 varieties. Each sampler used a mobile device to gather images of coccinellids on wheat plants. A total of 2133 images were collected and downloaded from each mobile device to a central server for further processing, analysis, and labelling.

5. Materials and Methods

5.1. Data acquisition process and timeline

The steps for gathering the dataset and their respective timeline are described in Fig. 5. All the imagery was collected during the summer of 2022 and 2023. The raw data was processed

Tasks	Year: 2022 and 2023							2024
	June	July	August	September	October	November	December	January
1. Dataset collection								
2. Processing raw dataset								
3. Processing and creating dataset for labeling								
4. Dataset labeling using Roboflow [1]								
5. Dataset published in Mendeley data repository [9]								

Fig. 5. Steps for dataset acquisition and timeline.

yearly to generate a dataset for further processing. The processing included eliminating images with blurriness or images with no coccinellid presence or other insects, such as beetles. Specifically, the final dataset with all 2133 images and label annotations was published and be available in 2024.

5.2. Collection methods

Images were collected under normal weather conditions with different angles and wheat plant backgrounds using mobile devices weekly during the summer of 2022 and 2023. The dataset includes images taken approximately 0.5 m from the wheat plants within row alleys using various zoom magnifications, ranging from 1x – to 5x across subplots.

The images have a dimension of 640 × 640 pixels, which can be reduced in size if needed for further training computer vision models. In addition, each image was annotated using the bounding box found in the label toolbox from the cloud labeling environment Roboflow [1] for each coccinellid found in each individual image. Finally, the dataset was exported from Roboflow [1] with all the imagery and label annotations generated.

5.3. Model specifications of mobile devices and image parameters

Different mobile devices were used in the collection of the dataset. The goal was to include variability with a diverse range of camera sensors and accessibility for potential users or growers with the further development of web mobile applications that can automatically detect coccinellids with computer vision models in images for sampling purposes. The specifications of one of the mobile device cameras used in the dataset collection are described in Table 1, with specifications of the image parameters in Table 2.

Table 1
Specifications of mobile device used to collect the dataset.

Camera specifications	Details
Camera manufacturer	Apple
Camera model	iPhone 12 Pro
F-stop	f/1.6
Exposure time	1/812 s
ISO Speed	ISO-32
Exposure bias	0 step
Focal length	4 mm
Metering mode	Spot
Flash mode	No flash, compulsory
35 mm focal length	39

Table 2
Specifications of image parameters after processing.

Image parameters	Details
Dimensions	640 × 640
Width	640
Height	640
Horizontal resolution	72 dpi
Vertical resolution	72 dpi
Bit depth	24
Resolution unit	2
Color representation	Uncalibrated

5.4. Future applications using the dataset collection

The dataset collected with the label annotations will be available publicly to train computer vision models to detect coccinellids on wheat plants. After models are trained with our collected dataset and label annotations, the models can be deployed into web applications, mobile applications or be deployed into robotic or sensor systems to automatically detect coccinellids. For example, a web application can be helpful to upload images and see the detection of a coccinellid in an image and provide the count number of the object within the image. This application will not only help understand the capabilities of automation but also will contribute to automatic counting, which currently is performed by humans. The future applications using the dataset consist of integration with robotics, which will enhance the use of automation and robotics in agriculture. We envision a time where imagery and other types of data will be captured automatically and management decisions for crops will be performed using machine learning approaches and robotics.

Limitations

Not applicable.

Ethics Statement

Authors affirm that they adhere to ethical guidelines for publishing. The present article does not include human subjects, animal experiments, or data obtained from social media platforms.

Credit Author Statement

Ivan Grijalva: Investigation, Data curation, Resources, Conceptualization, Methodology, Software, Formal analysis, Writing – original draft. **Nicholas Clark:** Conceptualization, Data Curation, Methodology, Software, Formal analysis, Writing – original draft. **Emma Hamilton:** Investigation, Data Curation, Methodology, Software. **Carson Orpin:** Investigation, Data Curation, Methodology, Software. **Carmen Perez:** Investigation, Data Curation, Methodology, Software. **James Schaefer:** Investigation, Data Curation, Methodology, Software. Investigation, Data Curation, Methodology, Software. **Kaylynn Vogts:** Investigation, Data Curation, Methodology, Software. **Brian McCornack:** Investigation, Conceptualization, Methodology, Resources, Writing – original draft.

Data Availability

Wheat_Coccinellid (Original data) (Mendeley Data).

Acknowledgments

We want to thank the research members of the Field Crops IPM Lab, including Kent Hampton, Anna Keenan, Luke Carney, Amanda Drouhard, and Trevor Witt, for their support, and data collection. The National Robotic Initiative partially supported this project grant no. 2019-67021-28995. This is contribution no. 24-193-J from the Kansas Agricultural Experiment Station.

Declaration of Competing Interest

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

References

[1] Dwyer, B., Nelson, J., Hansen, T., et. al. (2024). Roboflow (Version 1.0) [Software]. Available from https://roboflow.com.
[2] C. Wang, I. Grijalva, D. Caragea, B. McCornack, Detecting common coccinellids found in sorghum using deep learning models, Sci. Rep. 13 (2023) 9748, doi:10.1038/s41598-023-36738-5.
[3] M.J. Brewer, N.C. Elliott, Biological control of cereal aphids in North America and mediating effects of host plant and habitat manipulations, Annu. Rev. Entomol. 49 (2003) 219–242, doi:10.1146/annurev.ento.49.061802.123149.
[4] I. Grijalva, B.J. Spiesman, B. McCornack, Computer vision model for sorghum aphid detection using deep learning, J.f Agric. Food Res. 13 (2023) 100652, doi:10.1016/j.jafr.2023.100652.
[5] I. Grijalva, B.J. Spiesman, B. McCornack, Image classification of sugarcane aphid density using deep convolutional neural networks, Smart Agric. Technol. 3 (2023) 100089, doi:10.1016/j.atech.2022.100089.
[6] T. Zhang, K. Li, X. Chen, C. Zhong, B. Luo, I. Grijalva, B. McCornack, D. Flippo, A. Sharda, G. Wang, Aphid cluster recognition and detection in the wild using deep learning models, Sci. Rep. 13 (2023) 13410, doi:10.1038/s41598-023-38633-5.
[7] I. Ahmad, Y. Yang, Y. Yue, C. Ye, M. Hassan, X. Cheng, Y. Wu, Y. Zhang, Deep learning based Detector YOLOv5 for identifying insect pests, Appl. Sci. 12 (2022), doi:10.3390/app121910167.
[8] M. Vega, D.S. Benítez, N. Pérez, D. Riofrío, G. Ramón, D. Cisneros-Heredia, Coccinellidae beetle specimen detection using convolutional neural networks, in: 2021 IEEE Colombian Conference on Applications of Computational Intelligence (ColCACI), 2021, pp. 1–5, doi:10.1109/ColCACI52978.2021.9469588.
[9] I. Grijalva, N. Clark, E. Hamilton, C. Orpin, P. Carmen, J. Schaefer, K. Vogts, B. McCornack, Wheat_Coccinellid, (2024). https://data.mendeley.com/datasets/j9735xjspw/1.