Data in Brief 57 (2024) 111055

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

BEEHIVE: A dataset of Apis mellifera images to empower honeybee monitoring research

Massimiliano Micheli[a], Giulia Papa[b], Ilaria Negri[b], Matteo Lancini[c], Cristina Nuzzi[a,*], Simone Pasinetti[a]

[a] *Department of Mechanical and Industrial Engineering (DIMI), University of Brescia, Via Branze 38, 25123 Brescia, Italy*
[b] *Department of Sustainable Crop Production (DI.PRO.VE.S.), Catholic University of the Sacred Heart, Via E. Parmense 84, 29122 Piacenza, Italy*
[c] *Department of Medical and Surgical Specialties, Radiological Sciences, and Public Health (DSMC), University of Brescia, Viale Europa 11, 25123 Brescia, Italy*

A R T I C L E I N F O

Article history:
Received 27 September 2024
Revised 14 October 2024
Accepted 14 October 2024
Available online 22 October 2024

Dataset link: BEEHIVE: a public dataset of Apis mellifera images to empower honeybee monitoring research (Original data)

Keywords:
Entomology
Precision agriculture
Computer vision
Object detection
YOLO

A B S T R A C T

This data article describes the collection process of two sub-datasets comprehending images of Apis mellifera captured inside a commercial beehive ("Frame" sub-dataset, 2057 images) and at the bottom of it ("Bottom" sub-dataset, 1494 images). The data was collected in spring of 2023 (April–May) for the "Frame" sub-dataset, in September 2023 for the "Bottom" sub-dataset. Acquisitions were carried out using an instrumented beehive developed for the purpose of monitoring the colony's health status during long periods of time. The color cameras used were equipped with different lenses accordingly (liquid lenses for the internal one, standard lens of 8 mm focal length) and actuated by an embedded board, alongside red LED strips to illuminate the inside of the beehive. Images captured by the internal camera were mostly out-of-focus, thus a filtering procedure based on the adoption of focus measure operators was developed to keep only the in-focus ones. All images were manually labelled by experts using 2-class bounding boxes annotations representing full visible bees (class "bee") and blurred or occluded bees according to the sub-dataset ("blurred_bee" or "occluded_bee" class). Annotations are provided in YOLO v8 format. The dataset can be useful for entomology research

* Corresponding author.
E-mail address: cristina.nuzzi@unibs.it (C. Nuzzi).

https://doi.org/10.1016/j.dib.2024.111055

empowered by computer vision, especially for counting tasks, behavior monitoring, and pest management, since a few occurrences of Varroa destructor mites could be present in the "Frame" sub-dataset.

Specifications Table

Subject	Entomology and Insect Science, Computer Science.
Specific subject area	Entomology, Apis mellifera, pest management, insect monitoring, agricultural automation, data science
Type of data	Raw Images Annotations
Data collection	*Images were collected using two cameras of the same model (DFK ECU010-M12, Imaging Source), one positioned inside the beehive ("Frame" sub-dataset) and one at the bottom of it ("Bottom" sub-dataset). The internal camera mounts a liquid lens (Caspian M12-316-26 with built-in Arctic 316 lenses, Corning Varioptic) since the operating distance from the acquisition plane is of just 1.5* cm*, while the camera at the bottom of the beehive was equipped with standard lens with 8* mm *focal length. The instrumented beehive also mounts a red LED ring around the internal camera and red LED strips inside the beehive to illuminate the acquisition area. Lights were turned on only during acquisition, which were timed according to a duty cycle of 50 % for the LED ring, 70 % for the LED strips, lasting each time only 60* s*. Bounding boxes annotations of the bees in the two sub-datasets were manually labelled and saved in YOLO v8 format, using 2-classes for each sub-dataset.*
Data source location	Experiments were carried out in 2023 using the custom-made sensorized beehive described in [1]. Collected data are stored in hard-drives of the Laboratory of Mechanical and Thermal Measurements, University of Brescia, Department of Mechanical and Industrial Engineering, Via Branze 38, 25,123, Italy, Latitude: 45.562959, Longitude: 10.23156
Data accessibility	Repository name: BEEHIVE: a public dataset of Apis mellifera images to empower honeybee monitoring research Data identification number: 10.17632/5yz78xxpmy.1 Direct URL to data: https://data.mendeley.com/datasets/5yz78xxpmy/1 Instructions for accessing these data: Download the .zip files from the Mendeley Data page.
Related research article	M. Micheli, G. Papa, I. Negri, M. Lancini, C. Nuzzi, S. Pasinetti, Sensorizing a Beehive: A Study on Potential Embedded Solutions for Internal Contactless Monitoring of Bees Activity, Sensors, 24, (2024), 5270. 10.3390/s24165270

1. Value of the Data

The dataset contributes to the field of entomology (beekeeping) empowered by computer vision models and machine learning applications to facilitate tasks such as bees counting, bees' behavior monitoring, pest management, and general beehive's health monitoring.

The dataset can aid the research community providing new sources to train computer vision and machine learning models, especially regarding close-up images taken inside the beehive which are a novelty in the field. Additionally, the provided images can be useful for studies related to pest infections such as Varroa destructor and other diseases affecting bees.

The provided annotations are suitable to train, test and validate object detection models such as YOLO.

The dataset adds research value to the automation and instrumentation of the beekeeping field, that can result in novel technologies for automated monitoring of beehives during the whole season.

2. Background

The European Apis mellifera honeybee is commonly found in artificial for honey and wax commercial production as well as research purposes to study the hive behaviour during the whole season or to develop pest management systems (either chemical, biological or hardware-based) [2,3]. Open-source datasets of bees' images are not suited for this specific application; for example, the dataset in [4] contains images of bees in the wild at different angles, while the one in [5] only contains images of the bees' wings for entomology studies related to the bees' body structure. The scientific literature is lacking data taken directly on the beehive, preventing both research and commercial-oriented advances in the development of computer vision systems tailored for beehive monitoring. This is why the BEEHIVE dataset was assembled, thus becoming a useful help for research purposes since it contains images taken inside the beehive ("Frame" sub-dataset), and at the bottom of the beehive ("Bottom" sub-dataset).

3. Data Description

The dataset published in [6] is structured as depicted in Fig. 1. Inside the root folder there are two sub-folders named "bottom_dataset" and "frame_dataset" with the same structure, containing the images of the "Bottom" (1494 images) and "Frame" (2057 images) sub-datasets respectively. For each of them, the data is already provided subdivided into three sub-folders named "train", "valid" and "test" following a 70–20 % to 10 % splitting protocol. In addition, we provide a "README" .txt file containing Roboflow annotations notes (e.g., the size of the images, augmentation techniques adopted, pre-processing techniques adopted, total number of images in the sub-dataset, annotation format) and a "data" .yaml file containing the information needed by YOLO v8 models to start training (e.g., where the train, validation and testing sub-folders are located, the number of classes and their labels, information about Roboflow labeling environment).

Inside the "train", "valid" and "test" sub-folders, there are two sub-sub-folders named "images" and "labels" containing the images and the text annotations respectively. Annotations were provided for object detection tasks in YOLO v8 format [7]. For the "Bottom" sub-dataset, we provide two classes named "occluded_bee" and "bee". For the "Frame" sub-dataset, we provide two classes named "blurred_bee" and "bee".

4. Experimental Design, Materials and Methods

The acquisition campaigns started in the last days of April 2023 until the first week of May 2023 for the "Frame" sub-dataset, and during whole September 2023 for the "Bottom" sub-dataset. The internal camera captured a total of 17,210 images during the campaign; however, only 2057 of it are in-focus (filtered after collection). On the other hand, the "Bottom" sub-dataset contains a total of 1494 images captured by the corresponding camera, in which the metallic grid of the beehive was always present.

4.1. Instrumented set-up

The two sub-datasets were acquired during the experimental campaign conducted in 2023 using a custom-made instrumented beehive, as described in [1]. The beehive design includes the following hardware (Fig. 2):

1. Two color cameras manufactured by Imaging Source (model DFK ECU010-M12), with a pixel resolution of 1280 $\times$ 720 px, pixel size 3 $\times$ 3 μm, a frame rate of 30 frames per second (fps) and overall dimensions of 22 $\times$ 13 $\times$ 60 mm. The internal camera was mounted inside

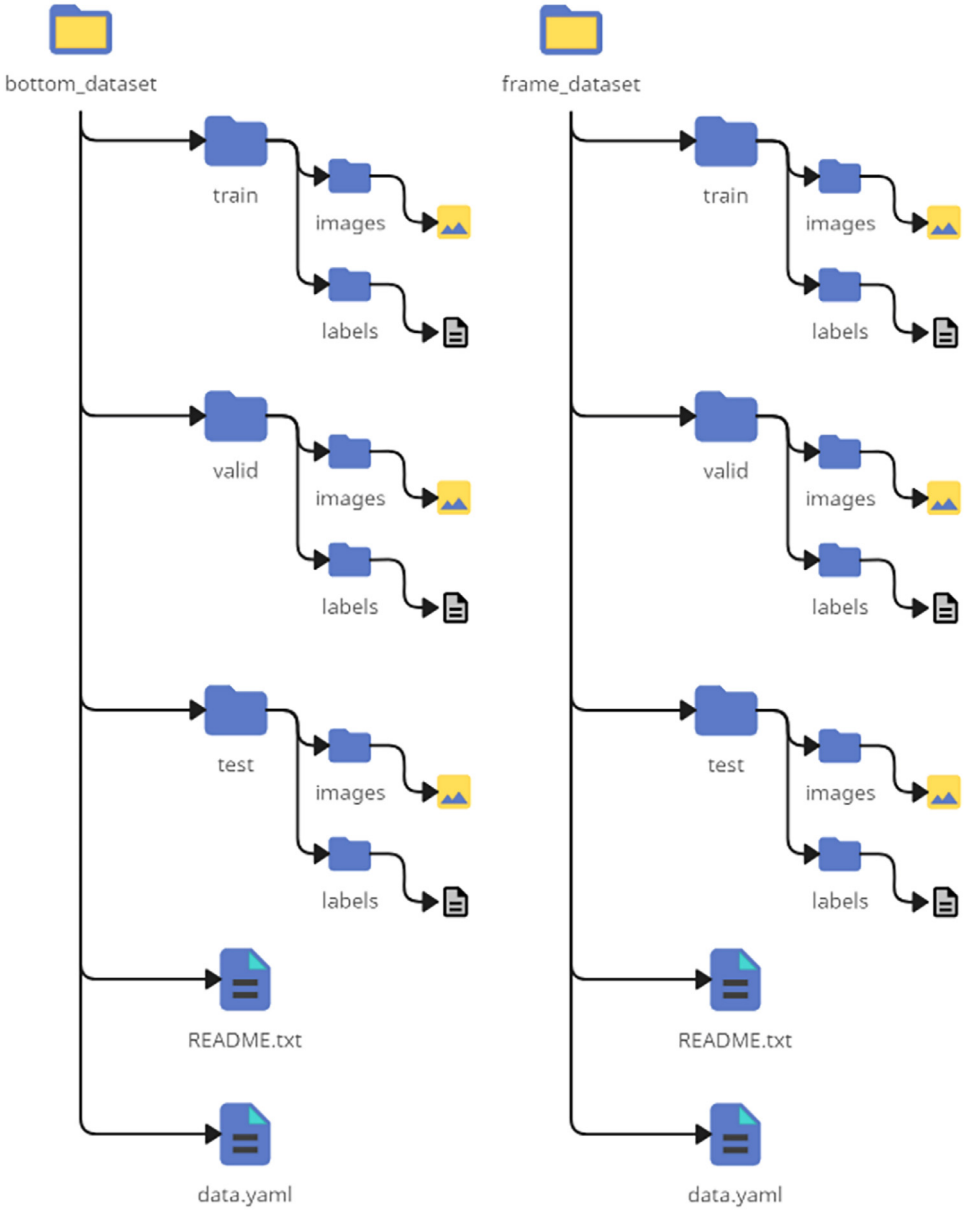

Fig. 1. Image depicting the folders structure for easy comprehension.

a comb in a seamless design to monitor the inside of the beehive, while the other one was placed at the bottom of the beehive to monitor the overall behavior of the bees (fixed with aluminum profiles).

2. Fixed optics with 8 mm focal length equipped on the camera at the bottom of the beehive.
3. Liquid lens optics (Varioptic Caspian M12–316–26, mounting a built-in Arctic 316 liquid lens) equipped on the internal camera. The minimum focal length of this model is 2.6 mm up to infinity, with an angular field of view of 160°, an iris aperture ranging from 2.5 to 10 mm,

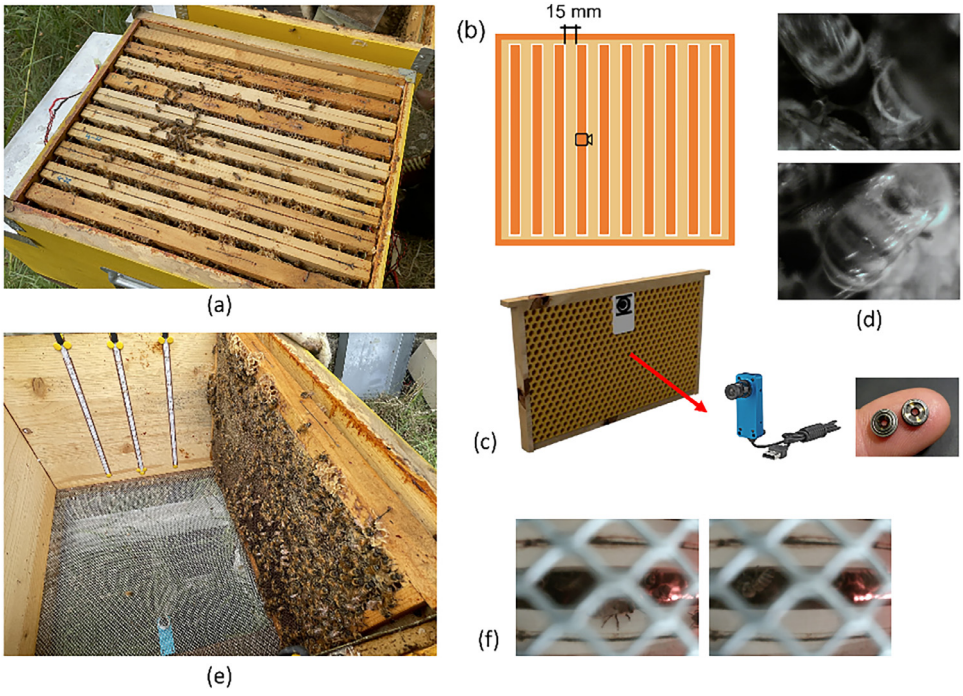

Fig. 2. (a) Top-view of the instrumented beehive. (b) Scheme of the position of the camera inside the beehive. Intra-comb distance is around 15 mm. (c) Image of the instrumented comb and of the camera and liquid lenses used. (d) Examples of the in-focus images taken from the "Frame" camera. (e) View of the inside and bottom set-up of the instrumented beehive, highlighting the position of the LED strips and of the "Bottom" camera. (f) Examples of the images taken from the "Bottom" camera. Please note the presence of the metallic grid.

4 mm to infinity focus range, and optical power that spans from −15 diopters to +38 diopters, resulting in a 53 diopters dynamic range, depending on the supply voltage. The device's response time is less than 15 ms after a new tension is provided to change the focal length, and the overall power consumption is 0.1 mW. The device is compatible with up to 1/2.5″ image sensors and M12 × 0.5 mounts (S-Mount), as well as off-the-shelf FPC connectors.

4. Control board Supertex HV892 to actuate the liquid lenses, operating voltage 25 – 60 VDC. Connected to the liquid lenses via I2C protocol using the dedicated software driver.
5. Red LED ring mounted around the internal camera to illuminate the inside of the beehive during acquisition. Red LED strips fixed inside the beehive (on the walls) to illuminate the inside of the beehive during acquisition.
6. Mosfet driver to control illuminators, operating at 12 VDC. The Mosfet actuates the LEDs according to PWM with duty cycle of 50 % for the LED ring and 70 % for the LED strips, thus avoiding overexposed effects and reduce the stress on the bees.
7. A single Raspberry Pi 4 which manages the Supertex control board, starts the cameras' acquisition via USB interface and operates the accension of the illuminators.

Since the bees move around the combs, their relative position with respect to the camera is always changing and the response time of the liquid lenses is not sufficient to keep up. Autofocus algorithms were explored to address this issue [8], but produced sub-optimal results. As a consequence, we decided to actuate the liquid lenses using a sinusoidal law with 1 Hz frequency in their voltage range (sweep). In this way, we were sure to acquire both in-focus and

out-of-focus images which we will filter out afterwards. A graphical example of the sweep with real images is shown in Fig. 3.

4.2. Acquisition software

The custom acquisition software run by the Raspberry Pi 4 was developed to manage the acquisition subprocesses of each connected camera independently as follows:

- The main program detects the number of cameras connected to the Raspberry Pi and creates the saving folders.
- The main program manages the illuminators (LED strings and LED ring) at the predefined duty cycle frequency set during configuration. It also manages the activation signal of the liquid lenses (sweep).
- For each camera, an "acquisition" subprocess is launched on a separated core. The task of these subprocesses is to trigger the camera acquisition at a rate of 30 fps. To limit the board's memory usage, the frames are resized to 640×480 px and stored in a queue. The acquisition lasts for 60 s only.
- At the same time, for each camera a "saving" subprocess is started with the task of saving on disk the frames temporarily stored in the queues.

4.3. Data processing

The processing procedure developed and described in [1] was used to discard the out-of-focus images (not present in the published dataset), obtaining from the original 17,210 images only the in-focus ones (2057, 12 % of the total images). The procedure was developed in MATLAB 2023b. We employed 5 focus measure operators from the literature, which were the Gaussian derivative (GDER), the Gray-level local variance (GLLV), the Steerable filters (SFIL), the Tenengrad (TENG), and the Tenengrad variance (TENV). We manually labelled 200 random images of the sub-dataset as "in-focus" or "out-of-focus", then plotted the distribution of each focus measure operator highlighting two distinct clusters. As a result, the threshold values used to separate the two types of images were GDER $\geq$ 26, GLLV $\geq$ 0.5, SFIL $\geq$ 1.1 $\cdot$ 1018, TENG $\geq$ 400, and TENV $\geq$ 480 0·103.

It is worth noting that among the in-focus images in the resulting "Frame" sub-dataset, bees were not always present or fully visible.

4.4. Data annotation

The data for the object detection task was manually annotated by an expert, identifying several rectangles (bounding boxes) that correspond to the bees present in the image. These bounding boxes are specified within the image plane (measured in pixels) using four parameters: the coordinates of the box center, C_X and C_Y, along with the box dimensions, W (width) and H (height). To incorporate these parameters into the model, their original values are normalized based on the image width I_W (for C_X and W) and image height I_H (for C_Y and H), resulting in values ranging between 0 and 1. Fig. 4 illustrates an example, showing the image coordinate system with the origin located in the top-left corner.

The annotations were provided for each image in the dataset following the YOLO v8 format [7]. For each image, a .txt file is provided in the corresponding folder of the dataset, containing for each line of text the following information in this order: (i) the class ID, (ii) normalized C_X, (iii) normalized C_Y, (iv) normalized W, and (v) normalized H. The class ID is a numerical indicator, in our case 0 means "blurred_bee" or "occluded_bee" (according to the sub-dataset) and 1 means "bee".

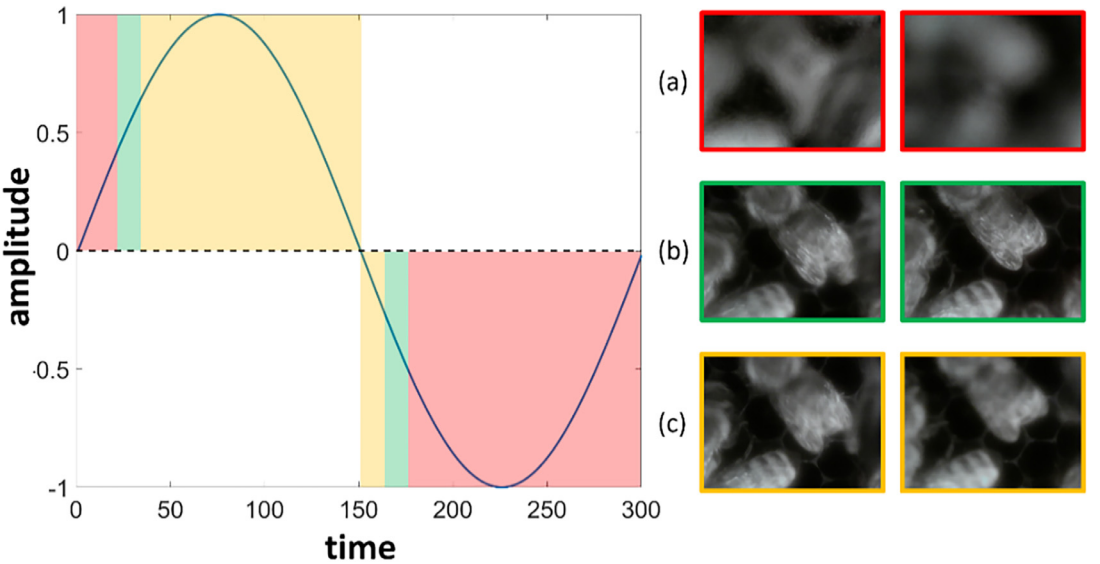

Fig. 3. Example of the sweep actuation of the liquid lenses. (a) Out-of-focus images acquired in the red areas. (b) In-focus images acquired in the green areas. (c) Slightly defocused images acquired in the yellow areas.(For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

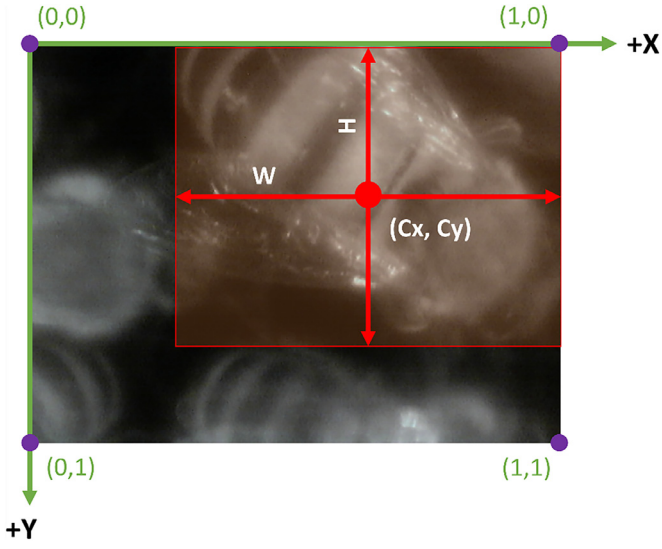

Fig. 4. Graphical example of the bounding boxes annotation with respect to the image coordinates. Image taken from [1].

Limitations

Due to the structure of the sensorized beehive, images belonging to the "Bottom" sub-dataset are partially occluded by a metallic grid.

Due to the activity of the colony, the bees inside the beehive were moving during acquisition and thus it is not guaranteed that the images in the "Frame" sub-dataset depict full-visible bees; in fact, a lot of the images contain portions of bees and even the honeycombs in the background when no bees were moving in front of the camera.

Ethics Statement

Authors have read and follow the ethical requirements for publication in Data in Brief. The current work does not involve human subjects, animal experiments, or any data collected from social media platforms since experiments with insects (in our case, bees of species Apis mellifera) are currently not regulated. We ensure that the experiment conducted did not cause any harm to the bees involved and complied with local regulations for ethical research.

CRediT Author Statement

Conceptualization, Micheli M., Pasinetti S. and Lancini M.; methodology, Micheli M., Negri I., Papa G., Pasinetti S. and Lancini M.; software, Micheli M.; validation, Micheli M., Nuzzi C. and Lancini M.; formal analysis, Micheli M. and Nuzzi C.; investigation, Micheli M., Papa G., Negri I., Pasinetti S. and Lancini M.; resources, Micheli M., Papa G. and Negri I.; data curation, Nuzzi C.; writing—original draft preparation, Micheli M. and Nuzzi C.; writing—review and editing, Pasinetti S., Lancini M. and Negri I.; visualization, Micheli M. and Nuzzi C.; supervision, Pasinetti S.; project administration, Lancini M.; funding acquisition, Lancini M. All authors have read and agreed to the published version of the manuscript.

Data Availability

BEEHIVE: a public dataset of Apis mellifera images to empower honeybee monitoring research (Original data) (Mendeley Data).

Acknowledgments

This work was supported by the European Union, FSE-REACT-EU, PON "Research and Innovation 2014–2020", D.M. 1062/202, contract number 46-G-13219-3.

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] M. Micheli, G. Papa, I. Negri, M. Lancini, C. Nuzzi, S. Pasinetti, Sensorizing a beehive: a study on potential embedded solutions for internal contactless monitoring of bees activity, Sensors 24 (2024) 5270, doi:10.3390/s24165270.
[2] C. Uthoff, M. Homsi, M. von Bergen, Acoustic and vibration monitoring of honeybee colonies for beekeeping-relevant aspects of presence of queen bee and swarming, Comput. Electron. Agricult. 205 (2023) 107589, doi:10.1016/j.compag.2022.107589.
[3] B.P. Oldroyd, Domestication of honey bees was associated with expansion of genetic diversity, Mol. Ecol. 21 (2012) 4409–4411, doi:10.1111/j.1365-294X.2012.05641.x.
[4] Yang, J. (2018). The BeeImage Dataset: annotated Honey Bee Images. Accessed in 2024. https://www.kaggle.com/datasets/jenny18/honey-bee-annotated-images.
[5] A. Oleksa, et al., Honey bee (Apis mellifera) wing images: a tool for identification and conservation, GigaScience 12 (2023) giad019, doi:10.1093/gigascience/giad019.
[6] M. Micheli, C. Nuzzi, G. Papa, I. Negri, M. Lancini, S. Pasinetti, BEEHIVE: a public dataset of Apis mellifera images to empower honeybee monitoring research, Mendeley Data V1 (2024), doi:10.17632/5yz78xxpmy.1.
[7] Torres, J. (2024). YOLOv8 Label Format: a Step-by-Step Guide. Accessed in 2024. https://yolov8.org/yolov8-label-format/.
[8] M. Micheli, S. Pasinetti, M. Lancini, G. Coffetti, Development of a monitoring system to assess honeybee colony health, in: 2022 IEEE Workshop on Metrology for Agriculture and Forestry (MetroAgriFor), Perugia, Italy, 2022, pp. 234–238, doi:10.1109/MetroAgriFor55389.2022.9964541.