Data in Brief 55 (2024) 110659

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

# A Brazilian native bee (*Tetragonisca angustula*) dataset for computer vision

Rodolfo Rocha Vieira Leocádio[a,*], Alan Kardek Rêgo Segundo[a,b], Gustavo Pessin[a,b]

[a] *PPGCC - Programa de Pós-Graduação em Ciência da Computação, Campus morro do Cruzeiro - ICEB, Rua Quatro, S/N - CEP 35402-136, Universidade Federal de Ouro Preto (UFOP), Ouro Preto MG, Brazil*
[b] *Instituto Tecnológico Vale – Mineração, Campus morro do Cruzeiro – Escola de Minas, Rua Nove, S/N - CEP 35402-163, Ouro Preto MG, Brazil*

ARTICLE INFO

*Article history:*
Received 28 December 2023
Revised 8 June 2024
Accepted 13 June 2024
Available online 26 June 2024

Dataset link: Tetragonisca angustula (Original data)

*Keywords:*
Jataí
Multiple bee tracking
Bee detectors
Precision agriculture
Smart monitoring

ABSTRACT

Jataí is a pollinator of some crops; therefore, its sustainable management guarantees quality in the ecosystem services provided and implementation in precision agriculture. We acquired videos of natural and artificial hives in urban and rural environments with a camera positioned at the hive entrance. In this way, we obtained videos of the entrance of several colonies for multiple bee tracking and removed images from the videos for bee detectors. This data, their respective labels, and metadata make up the dataset. The dataset displays potential for utilization in computer vision tasks such as comparative studies of deep learning models. They can also integrate intelligent monitoring systems for natural and artificial hives.



* Corresponding author.
*E-mail address:* rodolfo.leocadio@ufop.edu.br (R.R.V. Leocádio).

https://doi.org/10.1016/j.dib.2024.110659

Specifications Table

| | |
|---|---|
| Subject | Computer Science |
| Specific subject area | Computer Vision and Pattern Recognition |
| Data format | Raw data from natural and artificial hives |
| Type of data | Video (.mp4)<br>Video label (.txt)<br>Image (.png)<br>Image label (.txt)<br>Image metadata (.txt) |
| Data collection | We positioned the video camera, trying to optimize the capture of movements around each hive. We used the SONY HDR-AS20 camera, with a resolution of 1920×1080p and a recording speed of 60 fps. In an attempt to represent the natural environment of the hives, the recordings took place at different times of movement at the hive's entrance. The motion in the hive's entrance guarantees the variability of the poses of the bees. We withdraw videos of one beehive from the BeeKeep CS database. We acquired videos of natural and artificial beehives in urban and rural environments relatively close, but climatic conditions are different. Each hive is located in a different natural environment, making it necessary to change the acquisition methods with each collection. Natural hive 001 is positioned at a height of 3 m in a stone wall. The entrance to natural hive 002 is at ground level, with indications suggesting that a portion of the hive is buried. Natural hive 003 is located 1.5 m above ground on a movable tree trunk. Conversely, hive 004 is artificial and positioned atop a table on the porch. Natural hive 005 is situated on the exposed foundation of a house, 0.3 m above ground. Finally, artificial hive 006 was procured from the internet. |
| Data source location | Beehive 001 - Horto Parque dos Contos: Ouro Preto, MG, Brazil. Geographical coordinates obtained from Google (−20.381099, −43.506333). Natural beehive in a urban environment.<br>Beehive 002 - Sítio São Lázaro: Santa Rita de Ouro Preto, MG, Brazil. Geographical coordinates obtained from Google: (−20.502591, −43.520515). Natural beehive in a rural environment.<br>Beehive 003 - Sítio São Lázaro: Santa Rita de Ouro Preto, MG, Brazil. Geographical coordinates obtained from Google: (−20.502507, −43.520330). Natural beehive in a rural environment.<br>Beehive 004 - Sítio Bela Vista: Cachoeira do Campo, MG, Brazil. Geographical coordinates obtained from Google: (−20.345266, −43.608529). Artificial beehive in a rural environment.<br>Beehive 005 - Sítio Bela Vista: Cachoeira do Campo, MG, Brazil. Geographical coordinates obtained from Google: (−20.345640, −43.608437). Natural beehive in a rural environment.<br>Beehive 006 - BeeKeep CS: https://beekeep.pcs.usp.br. BeeKeep CS artificial beehive. |
| Data accessibility | Repository name: Tetragonisca angustula<br>Data identification number: 10.5281/zenodo.10439007<br>Direct URL to data: https://zenodo.org/records/10439007<br>Instructions for accessing these data: Access the dataset by visiting the Zenodo repository. We assure anonymity for all users. |
| Related research article | [1] R.R.V. Leocádio, A.K.R. Segundo, G. Pessin, Multiple Object Tracking in Native Bee Hives: A Case Study with Jataí in the Field. In: Naldi, M.C., Bianchi, R.A.C. (eds) Intelligent Systems. BRACIS 2023. Lecture Notes in Computer Science, vol 14197 (2023). Springer, Cham. https://doi.org/10.1007/978-3-031-45392-2_12 |

## 1. Value of the Data

- Using data on native Brazilian bees, ecologists can preserve endangered species of native fauna and flora.
- To our knowledge, this is the first dataset containing images of natural hives of native Brazilian bees for computer vision.
- The data allows comparative studies of computer vision models and intelligent monitoring of natural and artificial hives.
- Jataí is a significant pollinator of crops such as acapu, avocado, coffee, carrot, cupuaçu, guava, orange, mango, watermelon, strawberry, moressuma, pepper, cucumber, tangerine, umbu, and

annatto. Therefore, the data serves as a basis for the monitoring hives used in ecosystem services automation (such as recovery of degraded areas) and precision agriculture.

## 2. Background

Hive monitoring using computer vision holds the potential to minimize human errors and repetitive tasks. In their study, Bilik and his team [2] outlined the steps of robust monitoring across various application areas: pollen detection, infestations monitoring, traffic monitoring, and general inspection. Bee detection serves as the foundation for many of these monitoring steps and faces challenges that hinder the implementation of large-scale applications [2]. A significant obstacle to the deep learning approach - increasingly used - is the lack of a robust approach and high-resolution annotated datasets.

Some studies focused on pest insect detection [5-9] include *Apis mellifera* bees as one of their classes. The initial three obtain images of insects on plants from the internet, while the remaining two capture images from plants and traps, respectively. There are also specialized datasets for bee detection, such as [10-15], all featuring *Apis mellifera*. In a previous study [16], we developed a dataset comprising 459 images of *Melipona seminigra*. All these previous studies obtained images from artificial hives.

There exist datasets featuring videos for bee Multiple Object Tracking (MOT), encompassing species such as Apis mellifera [17,18] from artificial hives, Bombus terrestris [19], and Melipona quadrifasciata [20] filmed within laboratory settings, and Osmia bicornis [21] within artificial cavities. Predominantly, studies have centered on Apis mellifera, with a limited number [19,20] addressing Brazilian native bees like ours. Only our dataset includes labeled images and metadata to assist in management, alongside video labels conforming to the benchmark outlined in [1].

We believe this is the first dataset dedicated to the Brazilian native bee species known by the indigenous name Jataí (*Tetragonisca angustula*), primarily focused on natural hives within their habitats. This approach is advantageous for ecological studies, as it is impossible to accurately consider habitats using artificial hives. Additionally, invasive techniques cannot be employed on natural hives without causing destruction. Consequently, studies like ours provide a valuable alternative for the ecological research community facilitating the study of species and their habitats while also helping to reduce repetitive work.

Convolutional Neural Networks (CNNs) necessitate large datasets during the training phase to achieve performance levels comparable to humans. However, currently available datasets exhibit multi-scale properties [5-7], are characterized by classes imbalance [5-8,10], include few bee species [5-7,10-21], contain a small number of bee samples [5-9,13,16,18,20,21], lack sufficient standardized and labeled data [5-7,13,15,16,18,20,21], and have restricted accessibility [5-7,12-14,16-20], agreeing with [3,4]. The dataset presented in this work overcomes most of these challenges except for including a few bee species, which we aim to address in future work with other case studies and supports our previous study about multiple objects tracking in native beehives [1].

## 3. Data Description

The dataset consists of videos of the entry of several colonies for multiple bee tracking identified as XXXY-Y, where XXX identifies the hive location and Y-Y the recording interval. Each video has labels, as explained in Table 1.

Human annotations were conducted by a biologist as follows: we separated each video into frames, and from these frames, the biologist manually annotated the parameters (Table 1) using an image viewer. On the other hand, automated annotations do not require a biologist, as the

**Table 1**
Description of video labels.

| Process | Parameter | Data | Meaning |
|---|---|---|---|
| Human annotations | Total bees | INT | Total number of bees |
| | Out bees | INT | Total number of bees that left the hive |
| | In bees | INT | Total number of bees that entered the hive |
| | Pollen | INT | Total number of bees with load |
| | Invaders | 0/X | Does it have invaders? 0 - No, X - Yes |
| | Phorids | INT | Total number of phorids |
| | Ants | INT | Total number of ants |
| | Other bees | INT | Total number of other bee species |
| Automated annotations | Total bees | INT | Total number of bees |
| | Out bees | INT | Total number of bees that left the hive |
| | In bees | INT | Total number of bees that entered the hive |
| | Pollen | INT | Total number of bees with load |

**Table 2**
Description of image metadata.

| Parameter | Data | Meaning |
|---|---|---|
| Species | String | Specie of bee contained in the image |
| Total bees | INT | Total number of bees |
| Bee shadows | 0/X | Does it have bee's shadows? 0 - No, X - Yes |
| Invaders | 0/X | Does it have invaders? 0 - No, X – Yes |
| Phorids | INT | Total number of phorids |
| Ants | INT | Total number of ants |
| Other bees | INT | Total number of other bee species |
| Shadows | 0/X | Does it have invaders shadows? 0 - No, X - Yes |
| Occlusions | 0/X | Does it have bees occluded? 0 - No, X - Yes |
| Lighting | 0/X | Does it have lighting differences? 0 - No, X - Yes |
| Background | 0/X | Is it a background image? 0 - No, X - Yes |
| Data augmentation | 0/X | Did the image undergo data augmentation? 0 - No, X - Yes |
| Season | String | Season of the year of the acquire |
| Period | String | Period of the day during the acquire |
| Weather-cloudy | String | Climatic conditions at the time of acquisition |

objects of interest are pre-marked. These annotations comprise the following: first, we applied the EuTrack model [1] to the videos for automated annotations, and then we followed the same procedure as conducted by the biologist.

We identified images taken from the videos for bee detectors as the complement of the previous XXXY-YZ..., where Z... is an id associated with each image. The image directory is standardized according to the You Only Look Once (YOLO) family of detectors [22], plus a folder called “info” containing image metadata explained in Table 2.

Fig. 1 illustrates the pipeline of the image annotation protocol. We selected 120 images from the videos, randomly dividing them into 100 for training and 20 for validation. The training process involved 100 epochs using the YOLOv5 model [22], with images manually labeled by an operator using LabelImg [23]. The trained model was then utilized to detect Jataís in new images, and the labels were saved. These saved labels underwent human inspection to correct detector errors, also with LabelImg [23]. Following these corrections, the annotations were finalized and ready for subsequent use.

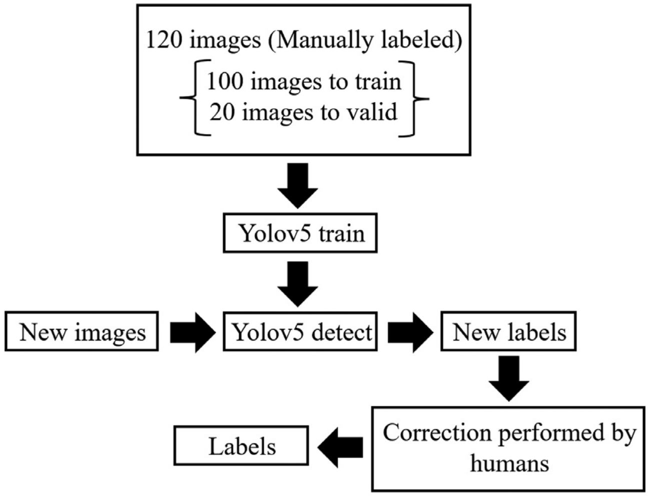


**Fig. 1.** Pipeline of the image annotation protocol.


a) Jataí b) Entrance of Jataí hives

**Fig. 2.** Tetragonisca angustula.
Source: Adapted from A.B.E.L.H.A.

## 4. Experimental Design, Materials and Methods

### *4.1.* Tetragonisca angustula *(Jataí)*

The Jataí (Fig. 2. a) is a bee native to Brazil, and its geographic distribution occurs throughout the Brazilian territory. Considering its natural environment, this species has the habit of nesting in tree hollows and rock cavities. Its adaptability promotes nesting in hollow walls, light boxes, letter boxes, stone walls, and other unusual places, considering an urbanized environment.

Its hive is formed by superimposed horizontal brood combs, with the presence of real cells surrounded by an envelope. There are deposits of viscous resin and pure white wax; the food pots are oval. It is a honey-producing species. The inlet tube is constructed with cerumen, has a Variable size, and can be inclined or in the shape of a hook with the inlet facing upwards (Fig. 2.b).

The colony has a queen mother and an average of 5000 (from 2000 to 8000) workers. Workers live about 24 days (maximum 76 days) and do not have a stinger. Guards are a caste with larger sizes than other workers. During the day, the workers remain positioned over the inlet tube or hover close to it. Virgin queens can be imprisoned in royal chambers. From 2 to 7 days after mating, the queen starts laying eggs. It has an average size of 4 mm and can forage 600 m from the hive.

It is a pollinator of some crops; therefore, sustainable management of hives guarantees the productivity of these crops and protects the colony from natural enemies. Jataís have three natural enemies: i) Phorids - are small and fast flies that lay eggs on the pollen stored in the hives. After hatching, its larvae consume all protein items, including the brood; ii) Ants – are attracted by the smell of food (honey and pollen), ants can attack the colonies, even exterminating them and; iii) Looting (mandatory and occasional) - weak hives can be looted by colonies of the same or another species. The breeder must be especially careful with the plundering done by thief bees (Lestrimelitta limao - lemon bee).

When the outside temperature reaches limits lower than 14 °C, the population gathers around the queen ("winter cluster"), keeping her at a temperature around 33 °C. The colony feeds on the pollen and honey stored in the hive during these periods. The workers do not leave the nest in periods of rain. During the night, they close the entrance to the nest with a thin wax film with holes, which, along with the constant flapping of the wings, cause air circulation and prevent the entry of enemies.

The dataset comprises a total of 26,487 instances in 2100 images. We propose partitioning the dataset into 19,000 instances within 1500 images for training, 3719 instances within 300 images for validation, and 3768 instances within 300 images for testing.

Observations regarding Jataí bees indicate that monitoring is unnecessary during cold, rainy days, and nighttime. Also, we avoid windy days. Additionally, it is crucial to monitor for the entry of invaders. The dataset reflects climate variations that provided differences in lighting (957 images), with a preference for days featuring intermittent sunlight (1750 images), and includes specific periods such as mornings (350 images) and afternoons (1400 images), as well as seasonal variations with 350 images from winter and 1400 images from spring. The dataset is free from duplicates and leaks, with a Structural Similarity Index Measure (SSIM) of 0.97 [24]. Each hive has a corresponding background image. For images without bees, data augmentation techniques were applied to remove them (5 images), filling the resulting space with textures from nearest neighbors. Instances of occlusions were noted due to overlapping bees, background items, entrance pipes, light beams, bee-to-bee contact, and partial bee images (1737 images). Anomalies such as shadows (361 images) and distortions during bee flight were also recorded. Incomplete data files were attributed to missing information from BeeKeep CS.

## Limitations

The context of the dataset is monitoring natural and artificial native beehives; therefore, it is only possible to use it for these purposes. The acquisition of videos and images containing invaders has not been possible due to difficulties in obtaining these media for prolonged periods. We don't know what caused the anomaly due to increased flight speed, but we believe it is due to obstacles near the hive entrance.

## Ethics Statement

The authors have read and follow the ethical requirements for publication in Data in Brief and confirming that the current work does not involve human subjects, animal experiments, or any data collected from social media platforms.

## Data Availability

Tetragonisca angustula (Original data) (Zenodo).

## CRediT Author Statement

**Rodolfo Rocha Vieira Leocádio:** Methodology, Investigation, Resources, Data curation, Writing – original draft, Visualization, Project administration; **Alan Kardek Rêgo Segundo:** Resources, Writing – review & editing, Supervision; **Gustavo Pessin:** Conceptualization.

## Acknowledgements

The authors would like to thank the Universidade de São Paulo (USP - https://beekeep.pcs.usp.br, the Empresa Brasileira de Pesquisa Agropecuária (Embrapa), and the Associação Brasileira de Estudo das Abelhas (A.B.E.L.H.A. - https://abelha.org.br) for the information and manuals on native bees that contributed to learning about Jataí and for the BeeKeep CS dataset that contributed with videos to compose the dataset this work. People who allowed filming to take place in beehives contained on their properties. To the Laboratório Multiusuário de Práticas Simuladas (LaMPS - https://lamps.medicina.ufop.br and the Laboratório de Controle e Automação Multiusuário (LABCAM - https://www.labcam.ufop.br/) for the infrastructure and equipment provided. This research received financial support from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Financing code 001.

## 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] R.R.V. Leocádio, A.K.R. Segundo, G. Pessin, Multiple object tracking in native bee hives: a case study with Jataí in the field, in: Intelligent Systems, 2023, pp. 176–191. Cham.
[2] S. Bilik, et al., Machine Learning and Computer Vision Techniques in Continuous Beehive Monitoring Applications: A Survey, 2023.
[3] R. Yasmin, A. Das, L.J. Rozario, Md.E. Islam, Butterfly detection and classification techniques: a review, Intell. Syst. Appl. 18 (2023) 200214 [Online] https://www.sciencedirect.com/science/article/pii/S266730532300039X.
[4] T.T. Høye, et al., Deep learning and computer vision will transform entomology, Proc. Natl. Acad. Sci. 118 (2021) e2002545117 [Online] doi:10.1073/pnas.2002545117.
[5] K. Li, J. Zhu, N. Li, Insect detection and counting based on YOLOv3 model, in: 2021 IEEE 4th International Conference on Electronics Technology (ICET), 2021, pp. 1229–1233.
[6] H. Liu, J.S. Chahl, Proximal detecting invertebrate pests on crops using a deep residual convolutional neural network trained by virtual images, Artif. Intell. Agric. 5 (2021) 13–23 [Online] https://www.sciencedirect.com/science/article/pii/S2589721721000106.
[7] I. Ahmad, et al., Deep learning based detector YOLOv5 for identifying insect pests, Appl. Sci. 12 (2022) [Online] https://www.mdpi.com/2076-3417/12/19/10167.
[8] A. Sava, L. Ichim, D. Popescu, Detection of halyomorpha halys using neural networks, in: 2022 8th International Conference on Control, Decision and Information Technologies (CoDIT), 2022, pp. 437–442. vol. 1.
[9] N. Zhao, et al., Development of an automatic pest monitoring system using a deep learning model of DPeNet, Measurement 203 (2022) 111970 [Online] https://www.sciencedirect.com/science/article/pii/S0263224122011666.

[10] S. Stefan, M. Kampel, VarroaDataset, 2020 [Online], doi:10.5281/zenodo.4085044.
[11] F. Hickert, Dataset for a Camera Based Bee-Hive Monitoring, 2021 \url.
[12] V. Kulyukin, Audio, image, video, and weather datasets for continuous electronic beehive monitoring, Appl. Sci. 11 (2021) [Online] https://www.mdpi.com/2076-3417/11/10/4632.
[13] R. Tashakkori, A.S. Hamza, M.B. Crawford, Beemon: an IoT-based beehive monitoring system, Comput. Electron. Agric. 190 (2021) 106427 [Online] https://www.sciencedirect.com/science/article/pii/S0168169921004440.
[14] G. Voudiotis, A. Moraiti, S. Kontogiannis, Deep learning beehive monitoring system for early detection of the Varroa Mite, Signals 3 (2022) 506–523 [Online] https://www.mdpi.com/2624-6120/3/3/30.
[15] S. Bilik, A. Ligocki, K. Horak, L. Kratochvila, T. Zemcik, Bee Dataset BUT-2, 2023 [Online] https://www.kaggle.com/dsv/5655997.
[16] R. Leocádio, et al., Detecção de abelhas nativas em colmeias em campo utilizando Visão computacional, in: Anais Do XII Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais, 2021, pp. 59–68. https://sol.sbc.org.br/index.php/wcama/article/view/15737. Porto Alegre, RS, Brasil[Online].
[17] T.N. Ngo, D.J.A. Rustia, E-C. Yang, T-T. Lin, Automated monitoring and analyses of honey bee pollen foraging behavior using a deep learning-based imaging system, Comput. Electron. Agric. 187 (2021) 106239 [Online] https://www.sciencedirect.com/science/article/pii/S0168169921002568.
[18] C.I. Narcia-Macias, et al., IntelliBeeHive: An Automated Honey Bee, Pollen, and Varroa Destructor Monitoring System, 2023.
[19] L.N. Stiemer, A. Thoma, C. Braun, MBT3D: deep learning based multi-object tracker for bumblebee 3D flight path estimation, PLoS ONE 18 (2023) 1–27 [Online], doi:10.1371/journal.pone.0291415.
[20] R.C. Bernardes, M.A.P. Lima, R.N.C. Guedes, C.B. da Silva, G.F. Martins, Ethoflow: computer vision and artificial intelligence-based software for automatic behavior analysis, Sensors 21 (2021) [Online] https://www.mdpi.com/1424-8220/21/9/3237.
[21] A.C. Knauer, J. Gallmann, M. Albrecht, Bee Tracker—an open-source machine learning-based video analysis software for the assessment of nesting and foraging performance of cavity-nesting solitary bees, Ecol. Evol. 12 (2022) e8575 [Online] https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.8575.
[22] G. Jocher, A. Chaurasia, J. Qiu, YOLO by Ultralytics, 2023.
[23] TzutalinLabelImg Free Software: MIT License, 2015.
[24] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (2004) 600–612.