Semear MAppleT FW: a dataset for apple detection and tracking in orchards under fruiting wall training system

Thiago Teixeira Santos[291], Lilian Nogueira de Faria[292], Luciano Gebler[293]

Abstract

Computer vision techniques for fruit detection and tracking are crucial for agricultural automation, yet most current datasets lack temporally consistent annotations needed for reliable tracking. Here we present Semear MAppleT FW, a dataset for apple detection and tracking in modern fruiting wall systems. The dataset comprises six video sequences of 100 frames each, captured by two RGB-D stereo cameras mounted on a tractor traversing orchard rows. Unlike previous datasets, Semear MAppleT FW features wide-angle images capturing entire tree lengths, ensuring complete canopy visibility within the field of view. To date, we provide over 53,000 bounding box annotations for 1,267 unique apple instances with temporal consistency across frames, stereo image pairs with known baseline calibration, and 3D reconstruction data. Our annotation method leverages structure-from-motion to estimate fruit positions in 3D space, enabling accurate tracking even when fruits are occluded by branches, leaves, or other fruits. The dataset includes visibility flags for each annotation, distinguishing between visible and occluded fruits. This approach maintains spatial consistency of annotations across frames while significantly reducing manual annotation workload. Semear MAppleT FW provides a valuable resource for developing artificial intelligence systems for automated yield estimation, fruit growth monitoring, and robotic harvesting in commercial orchards.

Keywords: Apples Production; Fruit Detection; Fruit Tracking; Digital Agriculture.

[291] Embrapa Digital Agriculture. ORCID: 0000-0002-9272-3403. Email: thiago.santos@embrapa.br.
[292] Scholarship holder at Embrapa Digital Agriculture. ORCID: 0009-0003-8802-1147.
Email: lilian.faria@colaborador.embrapa.br.
[293] Embrapa Uva e Vinho. ORCID: 0000-0001-9622-5578. Email: luciano.gebler@embrapa.br.

1. Introduction

The development of artificial intelligence-based systems for automation and assistance in fruit production tasks depends on datasets that accurately represent the conditions observed in commercial orchards. These datasets enable automated learning of patterns for tasks such as fruit detection, tracking, and counting in commercial orchards settings, which are essential for yield prediction and robotic harvesting. This data-driven approach replaces the need for explicit hard-coded rules for pattern identification.

Despite the scarcity of public image datasets in the agricultural domain (Lu; Young, 2020), a few datasets for image-based tasks in apple production have emerged recently. MinneApple (Häni et al., 2020) is an image dataset for apple detection and segmentation, containing more than 41,000 segmentation masks for individual fruits. As fruit tracking is not the goal of this dataset, association between apple detections in different images is not provided. Furthermore, parts of the tree canopy are often not visible because it was not possible to fit entire trees within the camera's field of view.

Datasets for inter-frame fruit detection and tracking have also been proposed. APPLE MOTS (Jong et al., 2022) is presented as the first dataset with temporal consistency in image sequences, containing almost 86,000 mask annotations for individual apples in images. However, most of the image sequences do not face the plants frontally, and some sequences do not capture the entire canopy within the field of view. Villacrés et al. (2023), for a detailed comparison of tracking algorithms, created a MOT dataset for apples using 9 video sequences captured with smartphones. Across these combined videos, tracks for more than 8,800 apples were annotated in 26,011 video frames. As with previous datasets, only portions of the apple trees are contained within the cameras' field of view.

The present work introduces *Semear MAppleT FW*, a dataset for multiple apple detection and tracking in orchards under modern *fruiting wall* training system (FW). Differently of other datasets, Semear MAppleT FW presents: 1) images grabbed by wide-angle cameras, able to fit the entire tree length in the field of view; 2) 3-D consistency, meaning the 2-D annotations are well-fitted to the fruit tridimensional position in space, *even under occlusion;* and 3) stereo images pairs are provided, with known baseline from camera factory calibration. In the present version, the dataset provides more than 50,000 annotations for more than 1,200 fruits.

2. Methods

The video footage was collected in the Embrapa's Temperate Climate Fruit Growing Experimental Station (EEFCT) located in Vacaria (one of the ten Agrotechnological Districts of the Semear Digital Project Embrapa, 2024), Rio Grande do Sul, Brazil (28°30'58.2"S, 50°52'52.2"W), on March 13, 2025. This footage was grabbed using the data collection system *SEEmear*[294]. The system was fixed in a tractor that performed an inter-rows path in an orchard under fruiting wall training, as seen in Figure 1A. Two ZED X RGB-D cameras (Stereolabs, Inc.), mounted in opposite directions, face the trees' row frontally (Figure 1)).

Video sequences 100 frames-long were extracted from the RGB-D cameras footage. Each sequence is composed of 1200 × 1920 pixels images in JPEG format, 100 images from the left camera and 100 images from the right camera on a ZED device. The COLMAP system for structure-from-motion was employed to retrieve the relative pose of the camera for each frame (the left camera was adopted as reference). The pose data is leveraged by the annotation tool, allowing the estimation of the fruits' positions in 3-D and reducing the workload on the manual annotation process.

The annotation process followed the methodology developed by Santos et al. (2024). Annotators begin by defining bounding boxes for each fruit in a few frames (typically three). Using these annotations and camera projection matrices previously generated with COLMAP, the system estimates the fruit's 3D center position and approximate size (represented as a radius in a spherical model). The projection matrices from remaining frames are then used to automatically generate bounding boxes by projecting the fruit center back, using the estimated size and depth information. Annotators can refine these annotations, re-estimate the fruit position and projections, and adjust fruit size in either the 2D bounding boxes or the 3D model. Additionally, annotators must mark each bounding box with a binary flag indicating whether the fruit is visible or occluded in the frame. A fruit is considered occluded when approximately one-third of it is not visible. Neither fallen fruits on the ground nor fruits from background trees (in rows behind the target foreground row) are annotated. Annotations are stored in text files following the MOT16 format (Milan et al., 2016). The time required to annotate a complete fruit track varies considerably, primarily depending on occlusions, but

[294] SANTOS, T. T.; KOENIGKAN, L. V.; GEBLER, L. SEEmear: a system for large scale geo-referenced stereo imaging of orchards. Apresentado no II Workshop Científico do Semear Digital, 2025, Campinas, SP. Embrapa Agricultura Digital.

generally takes between 40 seconds and 1 minute per fruit track. For more comprehensive details, readers are directed to Santos et al. (2024).

Figure 1. Image acquisition using the SEEmear on a tractor: (A) tractor path (from QGIS), collected by a GNSS system employing RTK correction; (B) the SEEmear facing two rows of trees using two ZED X cameras pointing in opposite directions. **Photos:** Thiago Santos (A); Lilian Faria (B)

3. Results and Discussion

The dataset is organized into directories called snippets. Each snippet contains: 1) 100 JPEG images from the left view and 100 JPEG images from the right view of the ZED RGB-D camera (1200 × 1920 pixels resolution); 2) camera pose data for each frame (using the left view as reference) from COLMAP; and 3) MOT annotation in MOT16 format, as described in Milan et al. (2016). Currently, six snippets (V01 to V06) have been completely annotated. Table 1 presents the number of visible apples and bounding box annotations for each of these snippets.

Figure 2A shows a sample frame with bounding boxes displayed in different colors for each fruit. The camera's wide angle captures the entire tree height within the field of view. While this means each apple occupies a relatively small area of the image, Figure 2B demonstrates that the fruits still appear with sufficient resolution for detection or even segmentation tasks. Figure 2B also illustrates the visibility annotation system: red markers indicate fully visible fruits, while white markers denote occluded fruits (where less than one-third of the fruit is visible).

Table 1. Semear MAppleT FW dataset (annotations until April 30, 2025).

Sequence	# Apples	# Bounding boxes
V01	217	8,437
V02	318	14,713
V03	85	3,277
V04	274	11,707
V05	121	6,119
V06	252	9,622
Total	1,267	53,875

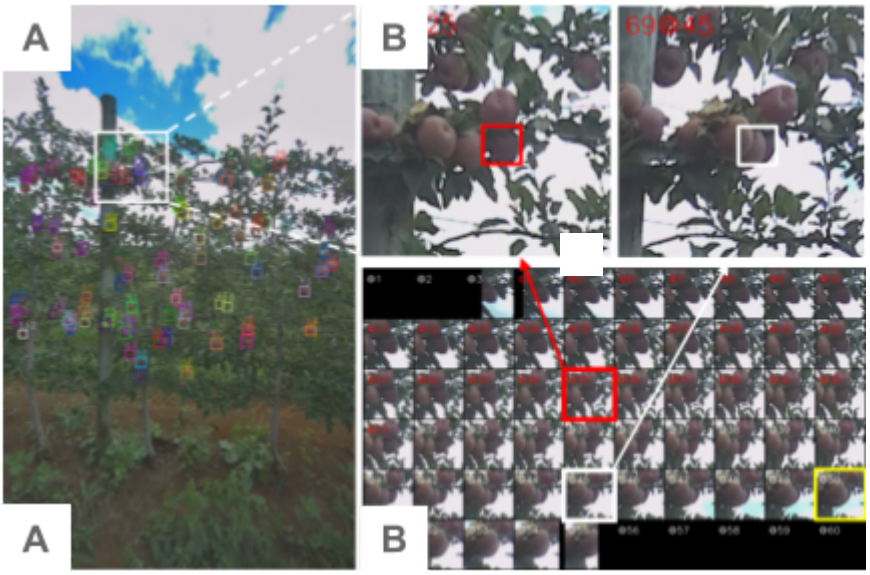

Figure 2. Example of MOT annotation in the dataset: (A) a video frame in sequence V04, where each individual apple annotation is shown in a different color; (B) details of the annotation procedure: apple 69 is visible at frame 25 (left), but it is significantly occluded at frame 45 (right).

4. Conclusion

The Semear MAppleT FW dataset provides annotations of apples across video frames in fruiting wall training systems. This dataset addresses critical gaps in existing resources by capturing entire tree lengths in the field of view and maintaining 3D spatial consistency of annotations even during occlusions. The novel annotation methodology leveraging structure from motion substantially reduces manual workload while ensuring tracking accuracy. Despite its contributions, limitations include the relatively small object size due to wide-angle imaging, the focus on a single training system, and the absence of

segmentation masks. Nevertheless, this dataset enables the development and evaluation of detection and tracking algorithms for yield estimation and robotic harvesting.

Acknowledgements

The authors would like to thank Fapesp (Proc. 2022/09319-9; 2024/19729-5) for the funding.

References

JONG, S. de; BAJA, H.; TAMMINGA, K.; VALENTE, J. Apple MOTS: detection, segmentation and tracking of homogeneous objects using MOTS. **IEEE Robotics and Automation Letters**, v. 7, n. 4, p. 11418-11425, Oct. 2022. DOI: https://doi.org/10.1109/LRA.2022.3199026.

HÄNI, N.; ROY, P.; ISLER, V. MinneApple: a benchmark dataset for apple detection and segmentation. **IEEE Robotics and Automation Letters**, v. 5, n. 2, p. 852-858, Apr. 2020. DOI: https://doi.org/10.1109/LRA.2020.2965061.

LU, Y.; YOUNG, S. A survey of public datasets for computer vision tasks in precision agriculture. **Computers and Electronics in Agriculture**, v. 178, 105760, Nov. 2020. DOI: https://doi.org/10.1016/j.compag.2020.105760.

MILAN, A.; LEAL-TAIXÉ, L.; REID, I.; ROTH, S.; SCHINDLER, K. **MOT16**: a benchmark for multi-object tracking. 2016. DOI: https://doi.org/10.48550/arXiv.1603.00831.

SANTOS, T. T.; SOUZA, K. X. S. de; CAMARGO NETO, J.; KOENIGKAN, L. V.; MOREIRA, A. S.; TERNES, S. Multiple orange detection and tracking with 3-D fruit relocalization and neural-net based yield regression in commercial sweet orange orchards. **Computers and Electronics in Agriculture**, v. 224, 109199, Sept. 2024. DOI: https://doi.org/10.1016/j.compag.2024.109199.

VILLACRÉS, J.; VISCAÍNO, M.; DELPIANO, J.; VOUGIOUKAS, S.; CHEEIN, F. A. Apple orchard production estimation using deep learning strategies: a comparison of tracking-by-detection algorithms. **Computers and Electronics in Agriculture**, v. 204, 107513, 2023. DOI: https://doi.org/10.1016/j.compag.2022.107513.