SBIAGRO
2021

A methodology for detection and localization of fruits in apples orchards from aerial images

Thiago T. Santos[1], Luciano Gebler[2]

[1]Embrapa Informática Agropecuária
Caixa Postal 6041 – CEP 13083-886 – Campinas, SP – Brazil

[2]Embrapa Uva e Vinho
Caixa Postal 177 – CEP 95200-000 – Vacaria, RS – Brazil

{thiago.santos,luciano.gebler}@embrapa.br

Abstract. *Computer vision methods based on convolutional neural networks (CNNs) have presented promising results on image-based fruit detection at ground-level for different crops. However, the integration of the detections found in different images, allowing accurate fruit counting and yield prediction, have received less attention. This work presents a methodology for automated fruit counting employing aerial-images. It includes algorithms based on multiple view geometry to perform fruits tracking, not just avoiding double counting but also locating the fruits in the 3-D space. Preliminary assessments show correlations above 0.8 between fruit counting and true yield for apples. The annotated dataset employed on CNN training is publicly available.*

1. Introduction

Crop monitoring is essential for anomaly detection, yield prediction and risk assessment in agriculture, basing the farmer's interventions. A continuous data collection during the fruits' growth cycle would allow an accurate modeling of its development, identifying anomalies and bottlenecks. Recently, convolutional neural networks [LeCun et al. 2015] have been employed for ground-level, image-based detection for different fruits [Sa et al. 2016], as apples [Häni et al. 2020] and mangoes [Bargoti and Underwood 2017]. However, just a few works [Liu et al. 2019, Häni et al. 2020, Santos et al. 2020] have addressed the *data association* problem in fruit counting: how to properly integrate the detections found in multiple images for accurate, row-level fruit tracking.

The present work describes a methodology for detecting and locating apples in orchards from aerial images sequences. This methodology allows not only the detection of fruits in the images, but also their association between images, identifying apples already observed previously, an essential requirement for fruit counting. The identified apples are properly mapped in the three-dimensional space, enabling the analysis of the variability in the field. The present methodology was able to produce promising results from aerial images of about 1 cm per pixel, thus being an alternative for autonomous monitoring of entire plots in orchards by unmanned aerial vehicles (UAVs).

2. Materials and methods

The data employed in the development of the methodology came from a plot located at the Embrapa's Temperate Climate Fruit Growing Experimental Station at Vacaria-RS (28°30'58.2"S, 50°52'52.2"W). The plot, seen in Figure 1 (a), is composed of 10 rows of apple trees, of which the 8 inner rows contain the plants of interest (the first and last rows are border ones). The rows contain plants of the varieties *Fuji* (west facing) and *Gala* (east facing). The images were taken during December 13, 2018. For aerial shots, an UAV (DJI Phantom 4 Pro) performed a 12 m height flight over the orchard's rows, capturing imagery data in the form of a 4K resolution video (3840×2160 pixels). The camera tilt is not nadir, allowing a more extensive view of the canopy if compared to a top/nadir one. The terms *frame* and *image* will be employed interchangeably in this text.

2.1. Methodology

The methodology consists of three steps. The first one is apple detection performed on each image, using a deep convolutional neural network [LeCun et al. 2015]. The second step estimates the camera position and orientation at each frame, using the structure-from-motion framework from multiple view computer vision [Hartley and Zisserman 2003, Schönberger and Frahm 2016]. The last step, the main contribution in this work, uses projective geometry and directed graphs to represent multiple alternative associations between fruits observed in different frames. Each *path* in the graph represents an association hypothesis, determining the location of the same fruit in different images, and a greedy algorithm is used to choose the paths.

2.1.1. Apple detection

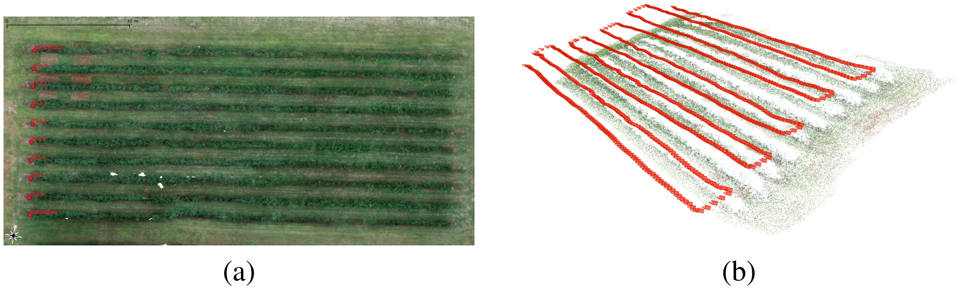
To this task, we have built an annotated dataset, formed by random selected 256×256 pixels samples from the frames extracted from the UAV video sequences. The dataset was split in training and test subsets for supervised machine learning, as shown in Table 1. This dataset is publicly available[1].

Table 1. Dataset for image-based apple detection training and evaluation.

	Number of images	Number annotated apples
Training	1025	2204
Test	114	267

For apple detection, we have employed a Faster R-CNN network [Ren et al. 2017], using a ResNet-50 backbone [He et al. 2016]. We employed the

[1] Available at `https://doi.org/10.5281/zenodo.5586329`.



(a) (b)

Figure 1. The orchard. (a) The plot presenting 10 rows. (b) Structure-from-motion performed by COLMAP - the UAV pose at capture time for each image is shown by the red frustums.

implementation available in PyTorch [Paszke et al. 2019] (see the `torchvision` library). The details of the training process, including data augmentations techniques, optimizer, batch sizes, number of epochs and hyperparameters can be seen in the publicly available code[2] and, due to text size restrictions, they will not be described here.

2.1.2. Relative camera pose estimation

To estimate the camera position at the time of capture for each video frame, we have employed the Structure-from-Motion (SfM) system COLMAP [Schönberger and Frahm 2016]. A SfM system estimates the *projection matrix* $\mathtt{P}_i$, a 3×4 matrix, for each image i: for each three-dimensional point $\mathbf{X} = (X, Y, Z, 1)^\intercal$ in the field, its 2-D projection $\mathbf{x}_i = (x_i, y_i, 1)^\intercal$ on the image plane of frame i can be computed[3] by the product

$$\mathbf{x}_i = \mathtt{P}_i\mathbf{X}. \tag{1}$$

The matrices $\mathtt{P}_i$ also allow the computation of the relative position $\mathbf{C}_i$ between cameras in the 3-D space by the property $\mathtt{P}_i\mathbf{C}_i = 0$. Figure 1 (b) illustrates the position of the UAV camera at the time of each frame capture in the flight over the plot.

2.1.3. Data association: tracking apples in the frames sequence

Projections matrices $\mathtt{P}_i$ and $\mathtt{P}_j$ allow the computation of the *fundamental matrix* $\mathtt{F}_{\mathtt{i,j}}$ [Hartley and Zisserman 2003]. Suppose that a point $\mathbf{X}$ in the 3-D space is mapped to the 2-D points $\mathbf{x}_i$ and $\mathbf{x}_j$ on the i-th and the j-th frames of the video sequence, respectively. The fundamental matrix maps $\mathbf{x}_i$ on frame i to a *epipolar line* $\mathbf{l}_{i,j}$ on frame j that contains $\mathbf{x}_j$. In our apple tracking problem, we have $\mathbf{x}_i^{(m)}$, the centroid of the m-th apple detected by the neural network on frame i. We can employ the fundamental matrix linking frames i and j to aid us in choosing the most suitable detections to correspond to $\mathbf{x}_i^{(m)}$, as seen in Figure 2.

[2] Available at `https://github.com/thsant/add256-fastercnn`.

[3] Points $\mathbf{X}$ and $\mathbf{x}_i$ are in *homogenous coordinates*, what explains the 1 in their last dimension.

Consider the centroids of the N apples detected by the neural network on the j-th frame, $\mathbf{x}_j^{(n)}, n = 1..N$. The detection corresponding to apple $\mathbf{x}_i^{(m)}$ should be close[4] to the line $\mathbf{l}_{i,j}^{(m)}$ in frame j, given by

$$\mathbf{l}_{i,j}^{(m)} = \mathtt{F_{i,j}} \cdot \mathbf{x}_i^{(m)}. \tag{2}$$

The fundamental matrix can be computed from the projection matrices by

$$\mathtt{F_{i,j}} = [\mathbf{e}_j]_\times \mathtt{P}_j \mathtt{P}_i^+, \tag{3}$$

where $\mathtt{P}_i^+$ is the pseudo-inverse of $\mathtt{P}_i$, and $\mathbf{e}_j = \mathtt{P}_j \mathbf{C}_i$ is the *epipole*, with $\mathtt{P}_i \mathbf{C}_i = 0$, i.e., $\mathbf{C}_i$ is projection center for the camera in frame i [Hartley and Zisserman 2003].

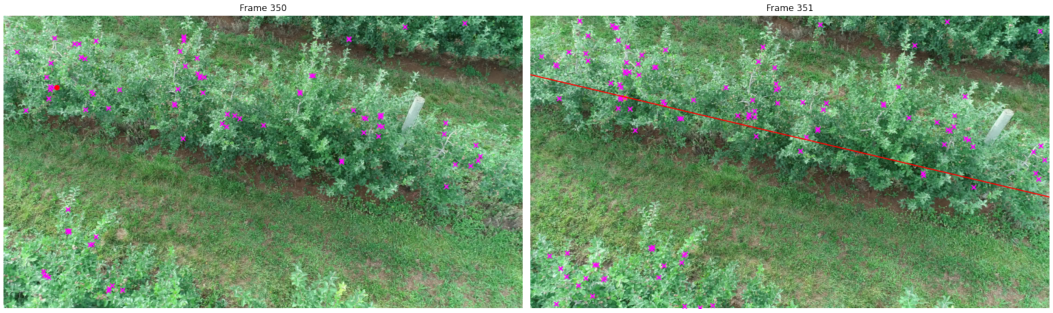


Figure 2. Epipolar restriction. Detected apples are shown as magenta 'x' markers. The point corresponding to the apple marked in red on Frame 350 defines the red epipolar line seen in Frame 351. The same apple should be observed near this line, limiting the number of options for apple tracking.

Our proposed apple tracking algorithm employs a *graph*, G, to represent multiple fruit associations hypothesis. Each *node* $v_i^{(m)} \in G$ corresponds to the centroid $\mathbf{x}_i^{(m)}$ of the m-th apple detected on a frame i. We add an *edge* $v_i^{(m)} \to v_j^{(n)}$ iif

$$\text{dist}(\mathbf{x}_j^{(n)}, \mathbf{l}_{i,j}^{(m)}) = \frac{\mathbf{x}_j^{(n)} \cdot \mathbf{l}_{i,j}^{(m)}}{\sqrt{a^2 + b^2}} \leq \tau_{\text{epipolar}}, \tag{4}$$

being $\mathbf{l}_{i,j}^{(m)} = (a, b, c)^\intercal = \mathtt{F_{i,j}} \cdot \mathbf{x}_i^{(m)}$. In other words, we are testing if the distance between the point and the epipolar line is below a threshold τ_{epipolar}. This procedure is performed by the lines 4–9 in Algorithm 1, FRUITASSOCIATION. So, *an edge in G represents a possible association between two detections in different frames*. As seen in line 5, for each frame i, the following k frames are evaluated for associations, what provides robustness to momentaneous misdetections of a fruit by the neural network.

A sequence of edges $v_i^{(m)} \to v_j^{(n)} \to \ldots \to v_k^{(o)}$ is a *path*. *Each path represents a possible association hypothesis for a fruit detected in frame i and the fruits detected in the following frames.* Lines 10-16 in Algorithm 1 implement a path selection process, employing a second algorithm, FRUITESTIMATION3D (Algorithm 2).

Algorithm 2 starts performing a depth-first search (DFS) from node $v_i^{(m)}$, getting all possible paths starting at $v_i^{(m)}$. An algorithm based on *random sample consensus*

[4] Ideally, in a noisy-free, perfect detection scenario, $\mathbf{x}_j^{(n)} \in \mathbf{l}_{i,j}^{(m)}$, i.e., $\mathbf{x}_j^{(n)} \cdot \mathbf{l}_{i,j}^{(m)} = 0$.

```
Data: The detected apples' centroids $\mathbf{x}_i^{(m)}$ for each frame $i$, $i = 1..F$
Result: A set of 3-D points (apples centers) $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \dots \mathbf{X}_L\}$ and their
        tracks
begin
    for $i \leftarrow 1$ to $F$ do
        foreach detected apple $\mathbf{x}_i^{(m)}$ in $i$ do Add node $v_i^{(m)}$ to $G$
    for $i \leftarrow 1$ to $F$ do
        for $j \leftarrow i + 1$ to $\min(i + k, F)$ do
            for all $\mathbf{x}_i^{(m)}$ and $\mathbf{x}_j^{(n)}$ do
                if $\mathrm{dist}(\mathbf{x}_j^{(n)}, \mathbf{l}_{i,j}^{(m)}) \leq \tau_{\text{epipolar}}$ then
                    Add the edge $v_i^{(m)} \rightarrow v_j^{(n)}$ to $G$
    $\mathcal{X} \leftarrow \emptyset$
    for $i \leftarrow 1$ to $F$ do
        for each $v_i^{(m)}$ do
            $\mathbf{X}, \mathcal{I}, r_{\mathcal{I}} \leftarrow$ FRUITESTIMATION3D($G$, $v_i^{(m)}$)
            if $\mathbf{X} \neq$ NIL then
                Add $\mathbf{X}$ to $\mathcal{X}$
                Remove from $G$ all edges $v_i^{(m)} \rightarrow v_j^{(n)}$ such that $v_i^{(m)}, v_j^{(n)} \in \mathcal{I}$
    return $\mathcal{X}$
```

Algorithm 1: FRUITASSOCIATION.

(RANSAC) [Fischler and Bolles 1981] is employed to estimate the tridimensional point $\mathbf{X}$ corresponding to a path (an apple's 3-D position in space). At each iteration, the TRIANGULATIONRANSAC algorithm pick three 2-D points, $\mathbf{x}_i^{(m)}$, $\mathbf{x}_j^{(n)}$ and $\mathbf{x}_k^{(o)}$ (corresponding to nodes $v_i^{(m)}$, $v_j^{(n)}$ and $v_k^{(o)}$ in a path T) and estimates the 3-D point $\mathbf{X}$. The estimation of X is performed by a least-squares minimization algorithm [Hartley and Sturm 1997]. Next, $\mathbf{X}$ is projected on each frame i in the path, defining the points $\hat{\mathbf{x}}_i^{(m)} = \mathtt{P}_i\mathbf{X}$ and their corresponding *geometrical errors*, i.e., the Euclidean distance between $\hat{\mathbf{x}}_i^{(m)}$ and $\mathbf{x}_i^{(m)}$. Nodes in the path T whose geometrical error is below the threshold τ_{geom} are considered *inliers*. At each iteration, the RANSAC procedure keeps the point $\mathbf{X}$ that delivered the largest number of inliers. Algorithm 2 looks for the path presenting the largest rate of inliers $\mathcal{I}$, keeping the longest path presenting the largest inlier rate. In other words, the inlier ratio acts as a quality measure for the inter-frame association hypothesis regarding fruit $\mathbf{x}_i^{(m)}$, represented by a path starting from $v_i^{(m)}$. Once a path is selected, the algorithm remove its edges from G (line 16 in Algorithm 1), avoiding those associations to be employed again. However, the nodes are preserved in the graph, allowing *fruits occlusions* to be considered: fruits that occlude each other can create crossing paths in G, i.e., paths sharing nodes.

```
Data: An association graph $G$ and a initial node $v_i^{(m)}$.
Result: The 3-D apple position $\mathbf{X}$, a set of inliers nodes $\mathcal{I}$, and the inlier ratio
        $r_\mathcal{I}$.
begin
    $\mathcal{T} \leftarrow$ DFS($G$, $v_i^{(m)}$)
    $\mathbf{X} \leftarrow$ NIL
    $\mathcal{I} \leftarrow \emptyset$
    $r_\mathcal{I} \leftarrow 0$
    for each track $T = \langle v_i^{(m)} \to v_j^{(n)} \to \ldots \rangle \in \mathcal{T}$, from the longest to the
     shortest do
        $\mathbf{X}_T$, $\mathcal{I}_T \leftarrow$ TRIANGULATIONRANSAC($T$)
        if $\mathcal{I}_T \neq \emptyset$ then
            $r_T \leftarrow \frac{\|\mathcal{I}_T\|}{\|T\|}$
            if $r_T > r_\mathcal{I}$ then
                $r_\mathcal{I} \leftarrow r_T$
                $\mathbf{X} \leftarrow \mathbf{X}_T$
                $\mathcal{I} \leftarrow \mathcal{I}_T$
    return $\mathbf{X}, \mathcal{I}, r_\mathcal{I}$
```

Algorithm 2: FRUITESTIMATION3D.

3. Results and discussion

Figure 3 displays the tracks determined by Algorithm 2 for two different apples. Each line in the figure corresponds to an apple's track (only the *inliers*). Note how the look of the fruit and its surroundings varies slightly as the pose (the UAV position) changes from frame to frame. Each track determines the three-dimensional position of an apple:

all *inliers* are used in the final estimation of the fruit's position $\mathbf{X}$ in the 3-D space, again by employing the least-squares algorithm [Hartley and Sturm 1997]. Figure 4 displays a total of 9,237 apples found in the plot. Fruits were automatically divided into the ten rows of the field by K-means clustering.

Disregarding two rows out of the UAV's field of view, caused by imprecision in the vehicle positioning system[5], the observed linear correlation between the counted apples in each row and the row's yield was 0.11 for *Fuji* and 0.80 for *Gala*, considering six rows. However, one of the rows (row 8) looks like a severe outlier: considering just the other five rows, linear correlation is 0.93 for *Fuji* and 0.88 for *Gala*. Although promising, the results should be viewed with caution, given that few rows were evaluated, at a single plot. More extensive experiments are yet necessary for a full characterization of uncertainty in yield prediction and the proportion of the fruits that is visible in imagery. It should also be noted that the images were captured in December and the harvest was carried out in February of the following year, which indicates that the methodology has the potential to provide yield estimated in early stages. Indeed, the presented methodology can be employed as a component of a more sophisticated yield prediction system.

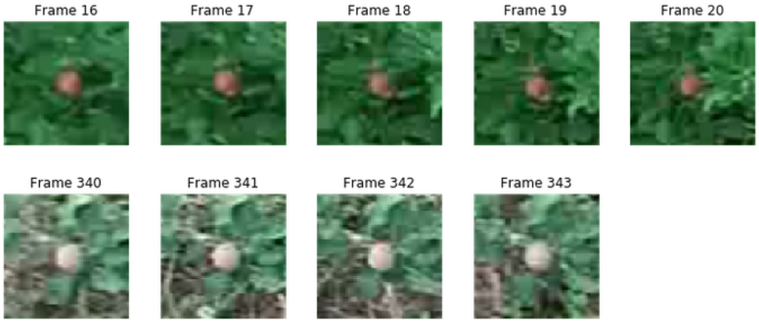


Figure 3. Two inter-frame fruit association examples found by Algorithm 1. Each row corresponds to an apple, observed in a few frames. Note the inter-frame variations, caused by UAV's pose changes during recording.

4. Conclusions

Fruit detection and tracking can be, in short term, applied to yield prediction and crop monitoring. In the long term, precise detection and 3-D localization can be employed on harvesting by autonomous agents. Detection and tracking allow autonomous agents to estimate their position relative to the fruit, so that accurate handling planning can be performed by the machine. The three-dimensional localization can also characterize the spatial variability of the fruits in the plots, helping on growing management according to precision agriculture practices.

The presented methodology is not restricted to aerial images: the same algorithms could be adapted to images obtained by ground vehicles with embedded cameras. Autonomous aerial vehicles with precise positioning control, such as devices equipped with

[5] Precise flights, able to keep the plants in the UAV's field of view, can be performed by vehicles presenting a precise position control, as a Real-Time Kinematic (RTK) Global Navigation Satellite System (GNSS). Unfortunately, the vehicle used in this work presented an ordinary GNSS system, without the positioning corrections provided by RTK.

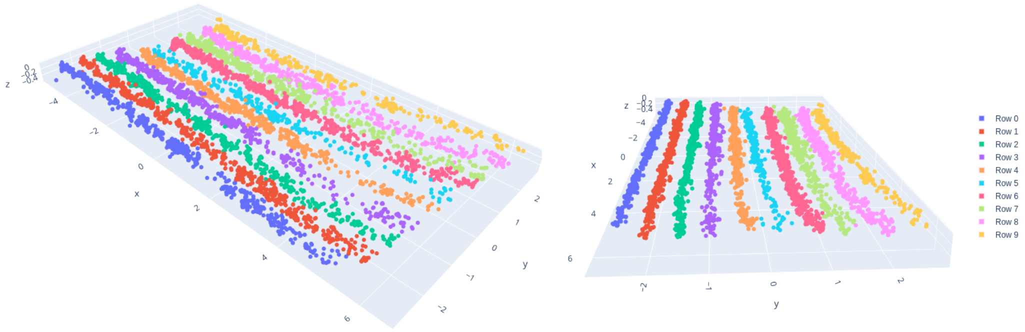


Figure 4. Fruit automatic localization. Each point represents the three-dimensional location determined for an apple. Colors represent different lines in the field, automatically identified using the K-means algorithm.

RTK GNSS, could be used as a row-scanning system able to perform automated field monitoring. New experiments, with a greater variability of plants, management regimes and plant architectures, should be carried out to validate and adapt the methodology for operation in different scenarios, and provide a better characterization of the estimation errors in yield prediction.

Acknowledgments

This work was supported by Brazilian Agricultural Research Corporation (Embrapa) under grant 01.14.09.001.05.04 and by FAPESP under grant (2017/19282-7).

References

Bargoti, S. and Underwood, J. (2017). Deep fruit detection in orchards. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3626–3633. IEEE.

Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395.

Hartley, R. and Zisserman, A. (2003). *Multiple View Geometry in Computer Vision*. Cambridge University Press, New York, NY, USA, 2 edition.

Hartley, R. I. and Sturm, P. (1997). Triangulation. *Computer vision and image understanding*, 68(2):146–157.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Häni, N., Roy, P., and Isler, V. (2020). A comparative study of fruit detection and counting methods for yield mapping in apple orchards. *Journal of Field Robotics*, 37(2):263–282.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Liu, X., Chen, S. W., Liu, C., Shivakumar, S. S., Das, J., Taylor, C. J., Underwood, J., and Kumar, V. (2019). Monocular Camera Based Fruit Counting and Mapping With Semantic Data Association. *IEEE Robotics and Automation Letters*, 4(3):2296–2303.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037.

Ren, S., He, K., Girshick, R., and Sun, J. (2017). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149.

Sa, I., Ge, Z., Dayoub, F., Upcroft, B., Perez, T., and McCool, C. (2016). Deepfruits: A fruit detection system using deep neural networks. *Sensors*, 16(8).

Santos, T. T., de Souza, L. L., dos Santos, A. A., and Avila, S. (2020). Grape detection, segmentation, and tracking using deep neural networks and three-dimensional association. *Computers and Electronics in Agriculture*, 170:105247.

Schönberger, J. L. and Frahm, J.-M. (2016). Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.