A Deep Learning-Based In-field Fruit Counting Method Using Video Sequences

Jiaqi Wang\textsuperscript{1}[0000–0002–1708–3573], Wenli Zhang\textsuperscript{1}[0000–0003–3151–5755], Kaizhen Chen\textsuperscript{1}[0000–0001–6871–4091], Huibin Li\textsuperscript{2}[0000–0002–4901–2104], Yun Shi\textsuperscript{2}[0000–0002–6294–0124], and Wei Guo\textsuperscript{3}[0000–0002–3017–5464]

\textsuperscript{1} Beijing University of Technology \texttt{zhangwenli@bjut.edu.cn}
\textsuperscript{2} Chinese Academy of Agricultural Sciences \texttt{shiyun@caas.cn}
\textsuperscript{3} The University of Tokyo \texttt{guowei@g.ecc.u-tokyo.ac.jp}

1 Introduction

In recent years, computer vision-based fruit counting in orchards has become a hot research topic in smart agriculture. Modern farms started to get benefits on fruit yield estimation and precision marketing strategy decisions from such technology. There are mainly two tasks for developing such techniques: precision fruit detection and counting from orchard images.

For fruit detection task, researchers have proposed deep learning-based image detection algorithms for fruit detection [1–4]. But they did not address the simultaneous presence of small-scale targets. For fruit localization and counting, researchers have proposed methods based on static images and video sequences[1, 3, 5–7]. The video-based counting method collects fruit images from multiple viewpoints and is considered as an efficient solution for fruit counting. However, the current video-based methods do not discuss the complex occlusion situations that may exist in global video sequences, which result in the loss of tracking targets.

Therefore, using orange as a study case, we propose the following solutions to the above two tasks: 1) We proposed an improved Yolov3 [8] detection model based on the principle of matching the feature map’s receptive field to the target scale [9]. 2) We first analyze the complex occlusion of orange fruits and define the counting region at each global video sequence frame. Then, using the multi-objective tracking algorithm Sort [10] to count the fruits that only appear in the pre-defined region.

2 Method

In this study, the video sequence was captured by the DJI Osmo Action camera (DJI Technology Co., Ltd., Shenzhen, China) in an orange orchard in Sichuan Province, China. The proposed fruit detection and counting method based on video include two steps: fruit detection and fruit tracking counting.
Table 1. Fruit Detection Performance

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>AP</th>
<th>FPPI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yolov3</td>
<td>0.926</td>
<td>0.90</td>
<td>0.911</td>
<td>0.960</td>
<td>2.294</td>
</tr>
<tr>
<td>improved-Yolov3</td>
<td>0.926</td>
<td>0.926</td>
<td>0.926</td>
<td>0.968</td>
<td>2.35</td>
</tr>
</tbody>
</table>

Table 2. Fruit Counting Performance

<table>
<thead>
<tr>
<th>Counting Method</th>
<th>Number of fruit counts</th>
<th>Inference time</th>
</tr>
</thead>
<tbody>
<tr>
<td>manual counting</td>
<td>90</td>
<td>30s</td>
</tr>
<tr>
<td>improved-Yolov3(No Track)</td>
<td>900</td>
<td>0.02s</td>
</tr>
<tr>
<td>improved-Yolov3+Sort(proposed)</td>
<td>102</td>
<td>0.08s</td>
</tr>
</tbody>
</table>

Fig. 2. Visualization of Fruit Detection

Step 1. Fruit detection method based on improved-Yolov3: Firstly, we calculate the size of the receptive field [11] of the Yolov3 network, and cluster the orange dataset to count the orange scale distribution. Secondly, we design the shallow prediction layer for detecting orange based on the principle of matching the feature map receptive field to the target scale. Then using a multi-level fusion strategy to fuse the shallow layer feature with the deep layer feature to enhance the semantic features of the shallow feature map. Finally, the fusion features are used to detect small-scale oranges in each image frame. The improved-Yolov3 network structure is shown in Figure 1, where the yellow region indicates the shallow prediction layer.

Step 2. Fruit tracking counting method based on specified area: Firstly, the orange detection results from step 1 are input to the tracking algorithm Sort, and determine whether these oranges are in the specified count area. If the fruit is in the count area, it will be assigned a unique number and tracked frame by frame until it leaves the count area. Finally, the number of orange ordinal numbers is counted as the final orange counting results.

3 Results and Discussion

In this study, we used 330 orange images and divided them into the train set and test set at the ratio of 8:2. Table 1 shows the comparison results between the improved-Yolov3 and the original Yolov3 for the five metrics of Precision, Recall, F1-score, FPPI, and AP. Figure 2 shows the detection results of the improved-Yolov3, where the red boxes correspond to ground truth and the blue boxes correspond to detection results. The orange counting results shown in Table 2, where the proposed improved-Yolov3 with tracking algorithms count 102 oranges at a speed of 0.08s per frame, is close to the manual count result.
References


