Fruit Size Estimation by Integrating Depth Data calculated using Monocular Images

Seema Shrawne[a], Dishie Vinchhi[a], Smit Shah[a], Ishaan Chandak[a], Yatharth Dedhia[a], Vijay Sambhe[a]

[a] *Computer and IT Department, Veermata Jijabai Technological Institute, Mumbai, India*

Abstract

Precise estimation of fruit size is important for yield estimation and autonomous harvesting. We introduce a technique to estimate fruit sizes from monocular RGB images, bypassing the limitation of object detection models such as Faster R-CNN, which fail to classify size. Our technique classifies fruit as small, medium, and large from RGB input and boosts accuracy with depth information from MiDaS, a monocular depth estimation model. We have tried ensemble learning, clustering, and artificial neural networks (ANN) to determine the effect of depth. Experiments using a self-assembled mango dataset show that a Random Forest classifier that includes depth features has an accuracy of 67.47%, outperforming traditional methods. The results show that depth information consistently enhances classification performance in all models studied. The research contributes to the development of precision agriculture by allowing automatic fruit size classification without the need for specialized depth-sensing hardware.

Keywords: Fruit size estimation, Monocular depth, Object Detection, Precision Agriculture

1. Introduction

Accurate estimation of fruit size is very important in most agricultural uses such as yield estimation, autonomous farming, and quality assessment. Conventional methods of measuring fruit size are usually human-based and prone to human errors, labor-intensive, and inefficient to apply in large-scale farming practices. The latest developments in computer vision and deep learning technologies have improved the automation of fruit detection, classification, and segmentation processes in a very impressive way. However, to estimate the fruit size from a single monocular image and without the use of specialised hardware and sensors with reasonable accuracy is still a big challenge due to the lack of depth information.

Object detection models, such as Faster R-CNN [13], YOLO [12], and SSD [4], have reported high accuracy rates for object detection over cluttered backgrounds. However, classification heads of Faster R-CNN [13]and similar models tend to fail at estimating the correct size of the object from direct 2D RGB images. This is because there is a lack of depth vision in 2D images, making it difficult to tell fruits that seem the same with different sizes when placed at varying distances from the camera.

To solve this issue, we present a novel approach by combining monocular depth estimation and image classification to estimate the fruit size. Unlike utilizing more expensive stereo or LiDAR sensors [6], we use MiDaS [2], which is a state-of-the-art advanced deep learning techniques capable of inferring depth maps from a single RGB image. This strategy eliminates the cost of depth specialty cameras and offers depth-enhanced features that enhance classification accuracy. This work involves the classification of fruits into small, medium, and large classes using both RGB and depth (RGB-D) information. We investigate different methodologies, including ensemble learning, clustering, and artificial neural networks (ANN), in order to measure the impact of depth information on classification performance. Experimental evaluation using a specially designed mango dataset is shown to affirm that the integration of depth-derived features significantly increases classification accuracy. Specifically, a Random Forest classifier trained over spatial depth features reaches a top accuracy of 67.47%, which compares favorably against conventional classification methodologies that utilize merely RGB data. Our findings ascertain that the use of depth information improves model performance in all the learning methodologies examined.

This work adds to precision agriculture by allowing automated, accurate, and inexpensive fruit size classification with the help of common depth-perceiving hardware. As autonomous fruit sorting systems' reliability improves, our method can render agriculture more efficient and thus save farmers and agricultural businesses labor costs.

2. Related Work

Several approaches have been studied over the years to address the problem of inaccurate Object Size Estimation using 2D images, ranging from the traditional image processing techniques to deep learning approaches.

2.1. Traditional Image Processing Methods

Traditional computer vision techniques like contour detection [14], edge detection [16], and shape analysis [10] were used earlier to measure the size of fruits. Methods like the Hough Transform [3] and morphological processing [15] were commonly employed to detect the edges of fruits and determine their size. These techniques typically encountered problems with occlusions, lighting changes, and background noise, resulting in incorrect results.[7]

2.2. Deep Learning-Based Approaches

Deep learning has led to widespread application of Convolutional Neural Networks (CNNs) to detect and classify objects in agriculture applications. Models such as Faster R-CNN [13], YOLO (You Only Look Once) [12] , and SSD (Single Shot MultiBox Detector) [4] have turned out to be attractive in segmenting and detecting fruits. The models can identify single fruits from an image and thus automatically derive bounding boxes enabling the estimation of fruit size [5].. The approaches are, however, not endowed with the intrinsic feature of capturing depth information and therefore the estimation of size is erroneous, especially if fruits are placed at different distances from the camera.

2.3. Limitations in Existing Work

Despite tremendous advances, existing models still find it challenging to address problems such as varying fruit shapes, occlusions, and inconsistency in object distance. Further, most approaches require extensive use of Sensors and Specialized hardware along with annotated datasets for supervised learning, and therefore impede deployment in practical applications.

2.4. Our Contribution

To solve these problems, our paper introduces a method that combines Traditional Image Classification and monocular depth estimation to accurately estimate the size of fruits. Unlike other studies [9] [8], this manuscript presents a new feature engineering technique that incorporates both the depth and the area of the fruit to develop a proportionality measure for enhancing size classification. In addition, we investigate the application of clustering algorithms coupled with machine learning models to achieve greater accuracy in size estimation than traditional methods.

3. Materials and Method

3.1. Dataset Description

The data used in the research contains around 250 mango images taken under various real-world scenarios. Mango trees, fruit in their natural environments, and different occlusion and lighting scenarios are present in the images. Mangoes were also isolated and kept in front of different backgrounds for systematic assessment. There are various backgrounds in the data, where fruits are captured in various settings like leaves, plates, hands, and natural surfaces.

Figure 1: Sample Image from the Dataset

We used data augmentation techniques to improve the model's ability to generalize across various conditions. We applied scaling operations to marginally reduce or enlarge the size of the fruits. Changes in brightness were performed to change the light conditions, mimicking different environmental conditions. These transformations proved to be useful in expanding the dataset, allowing better learning of size and shape variations of mangoes under different conditions by the model.

Each image from the dataset was hand-annotated to provide accurate bounding box coordinates for all the visible mangoes. Annotation was performed via the CVAT tool to provide a more accurate localization of the fruits. Each annotated bounding box was defined in terms of four fundamental coordinates: top-left x (xtl) and top-left y (ytl) describing the top-left corner of the fruit bounding box, and bottom-right x (xbr) and bottom-right y (ybr) describing the bottom-right corner.

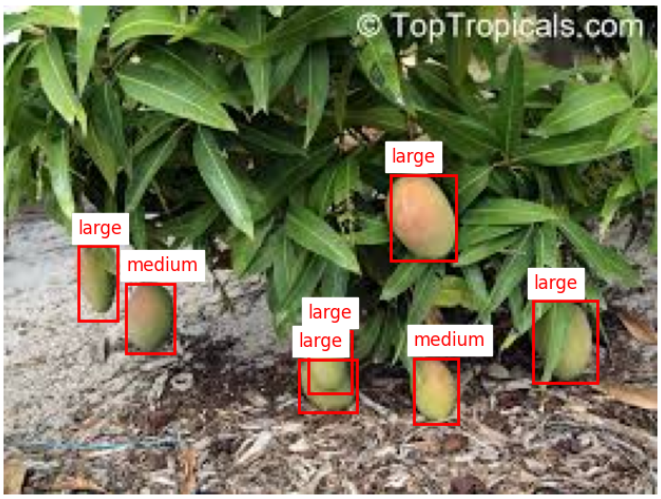

Figure 2: Sample Annotated Image from the Dataset

With these bounding box coordinates, the area, length and width of the bounding box was extracted, which was instrumental in estimating the size. Once bounding box annotation was performed, each fruit was labeled with a size label as ground truth. The size labels were small, medium, or large. These labels were based on physical measurements from actual samples to ensure that the ground truth classification was valid.

As monocular images do not contain intrinsic depth information, the MiDaS depth prediction model was utilized to estimate the relative depth of every mango. MiDaS was run on every image to generate a corresponding depth map and for every mango detected, the average depth inside the provided bounding box was calculated.

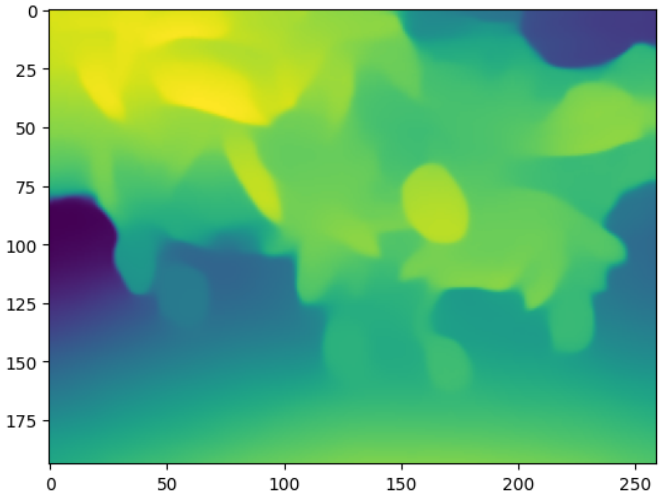

Figure 3: Depth Map Generated by MiDaS for a Sample Image

By integrating all these attributes, a CSV was created to act as the training basis for the model. The CSV contains major attributes like image detection, bounding box coordinates (xtl,ytl,xbr,ybr), bounding box area (calculated from coordinates), size class (small, medium, large), and average depth metric (calculated from MiDaS output). Moreover, length, width columns were introduced that were calculated by taking the number of pixels that the mango occupies divided by the total pixels of the image. Analogous scaling was done so that features would be presented uniformly.

All of these aspects were used to investigate a variety of classification methods, utilizing both spatial characteristics (area of the bounding box) and depth features to improve the accuracy of the estimation of fruit size.

3.2. Methodology

To improve object size classification accuracy in Faster R-CNN, we suggest incorporating depth information with the MiDaS monocular depth estimation model into the classification head. Faster R-CNN is mostly based on 2D image features extracted using its backbone network, which can cause misclassifications when objects of varying sizes look similar due to perspective distortion. With MiDaS, which gives pixel-wise depth estimates of an input RGB image, we can include depth-aware feature representations into spatial understanding. The suggested modification is to enrich the feature maps used in the classification head by adding or fusing depth maps with the extracted CNN feature maps prior to the region proposal network (RPN) and classification layers. This allows the model to distinguish objects more precisely based on true size instead of apparent size in 2D space.

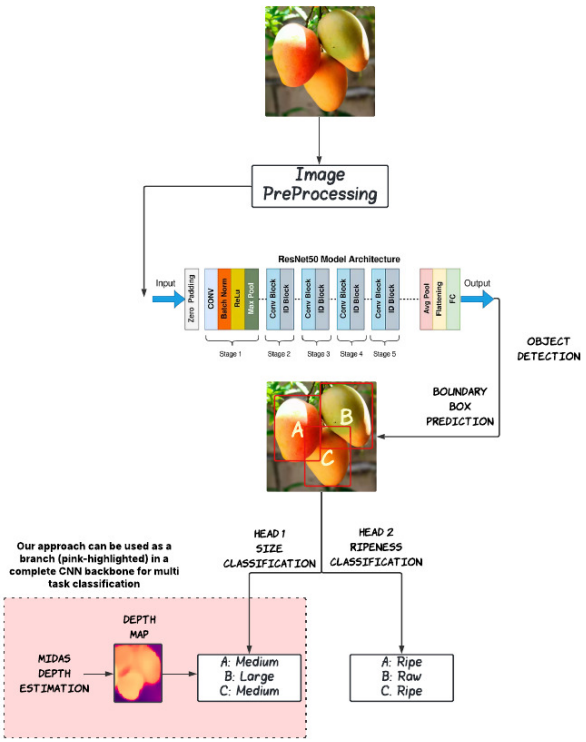

Figure 4: Highlighted portion indicates proposed changes to FasterRCNN Object Detection Classification Head to incorporate Depth Data obtained from MiDAS Monocular depth Estimation Model for increased accuracy in Object Size Classification

3.3. Method I - Breakpoint Method Using Quantiles

The simplest method to categorize the fruit size was to calculate breakpoints from quantile-derived feature values such as bounding box area and depth and divide images into small, medium, and large based on these values. The most important steps of the approach were to calculate the bounding box area using the formula:

$$\text{Area} = (x_{br} - x_{tl}) \times (y_{br} - y_{tl}) \tag{1}$$

In addition, an area-to-depth ratio measure was computed as follows:

$$Metric = Area/Depth^2 \tag{2}$$

In accordance with finding breakpoints, the data set was grouped into three classes according to values of area. Quantiles rather than medians were utilized for finding the threshold values. For the first breakpoint, the lower quantile (e.g., the 33rd percentile) was used and for the second breakpoint, the upper quantile (e.g., the 66th percentile) was employed.

For classification, one fruit was small if the computed area was below the first breakpoint. When the computed area was between the first and the second breakpoint, the fruit was medium. When the computed area was above the second breakpoint, the fruit was large.

This approach enhanced stability in response to data distribution shifts, reducing sensitivity and enhancing generalizability.

3.4. Clustering-Based Classification

In order to further enhance the breakpoint method, we used K-means clustering with K=3 so that the model could automatically create clusters from features obtained. Instead of using a single measure, we tried a number of variations of the feature for the purpose of classification like: the area-to-depth product ($F = \text{Area} \times \text{Depth}$), and the square root of their product ($F = \sqrt{\text{Area} \times \text{Depth}}$).

After finding the best feature transformation through clustering statistics, K-means clustering then segmented the dataset into three classes. The closest ground truth class small, medium, or large was then mapped against each cluster via the application of majority voting against the labeled dataset. The above application revealed stronger classification accuracy as compared to breakpoint approach, main reason being due to its capability to handle

more flexible decision borders. It also had the disadvantage of requiring feature transformation, which interfered with the flexibility in managing varied dataset distribution.

3.5. Ensemble Methods for Classification

Random Forest Classifier (RFC)[17] and XGBoost[11] are a few examples of ensemble learning that creates many decision trees while training and outputs the mode of the prediction of each individual tree. We chose these methods because of their stability to noise and ability to manage nonlinear relationships within the dataset.

The primary characteristics used were the **bounding box area** and the **depth** of the identified fruits. The data was divided into an 80-20 training-testing ratio to enable a fair assessment. A Random Forest Classifier with 100 trees was employed. The model was optimized using **Gini impurity** as the node split criterion. The trained model was then tested on the held-out test set. Accuracy was computed using:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Test Samples}} \times 100 \tag{3}$$

The RFC approach significantly improved the classification accuracy compared to the clustering approach; nonetheless, there was still potential to further improve it. As a result, this generated interest in feature engineering techniques like incorporating depth data for improving the predictive power of the model.

3.6. Feature Engineering on RFC for Improved Accuracy

To enhance the efficiency of classification, more features were developed that not only considered the size and depth of an object but also its other attributes. Additional features were developed to capture the shape, and size of the detected fruits so that it would possess a better description to be employed in classification. A significant feature that was integrated is the aspect ratio of the bounding box since it assists in describing the elongated or curved shape of the fruit. It was defined as:

$$\text{Aspect Ratio} = \frac{\text{xbr} - \text{xtl}}{\text{ybr} - \text{ytl}} \tag{4}$$

This feature was relatively less useful in identifying fruits of varying sizes and positions.

In addition, an area-to-depth ratio measure was computed as follows:

$$Metric = Area/Depth^2 \tag{5}$$

The extra features were added to the training set, and the **Random Forest Classifier (RFC)** was trained on the new feature set. In contrast to the first approach, this step yielded a notable improvement in classification accuracy, reaching about 67%. The increased number of shape, size and depth based features enabled the model to make more accurate decisions, ultimately resulting in better performance.

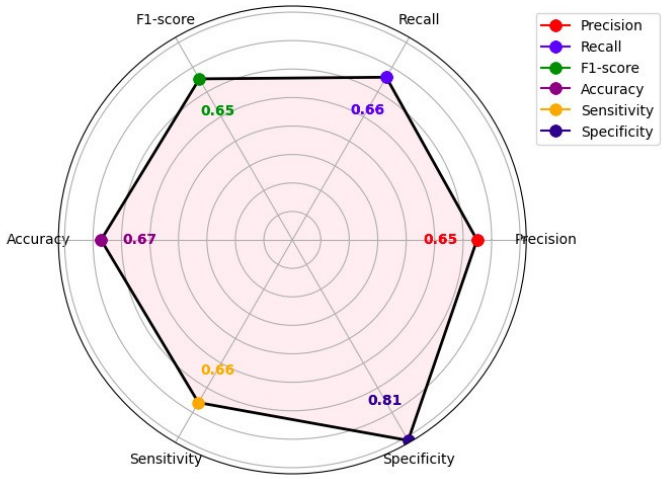

Figure 5: Random Forest Performance Radar Diagram

3.7. Artificial Neural Network

The Artificial Neural Network (ANN)[1] designed here is a feedforward neural network used for fruit size classification based on five input features. It consists of an input layer with 5 neurons, two hidden layers with 16 and 8 neurons respectively, both using the ReLU activation function to introduce non-linearity, and an output layer with 3 neurons utilizing softmax activation, indicating a three-class classification problem (e.g., small, medium, large fruit). The network learns patterns from input features to categorize fruit sizes effectively by adjusting its weights through backpropagation and optimization techniques.

4. Results and Discussion

This section presents the evaluation results of different fruit size estimation methods discussed earlier. The performance of each method was analyzed based on classification accuracy. The experiments were conducted on the custom RGB-D dataset containing mango images with their respective size labels.

4.1. Comparison of Methods

Below Table provides a comparative study of the increase in accuracy obtained using the different approaches, with and without the addition of Depth Data.

Method	**Accuracy (%) Without Depth**	**Accuracy (%) With Depth**	**Increase (%)**
$BreakpointMethod$	–	48.10	–
$Clustering\ Approach$	49.74	49.95	+0.42
$XGBoost\ Classifier$	57.56	64.84	+12.64
$RandomForest\ Classifier$	59.51	**67.47**	+13.37
$ArtificialNeural\ Network$	62.05	64.10	+3.30

This table clearly shows the improvement in accuracy for each method when incorporating estimated object depth. Let me know if you need any modifications!

Comparison of fruit size estimation methods shows that machine learning methods are better than unsupervised clustering. Without Depth based features, ANN performed best (62.05%), followed by Random Forest (59.51%) and XG Boost (57.56%), whereas Clustering-Based Approach performed the worst (49.74%). With Depth based features, all the machine learning models improved greatly, with Random Forest performing the best (67.47%), followed by XG Boost (64.84%) and ANN (64.10%). The Clustering-Based Approach improved very minimally, making its weakness more pronounced.

4.2. Performance Analysis

Machine learning models performed much better than clustering, especially with depth based features. Random Forest performed better than ANN with depth based features, showing its advantage in dealing with structured data. XG Boost also performed much better, showing its dependence on spatial context for improved classification. ANN performed reasonably well but experienced a lesser accuracy gain with depth features, potentially due to feature sensitivity. Clustering continued to be ineffective, showing the need for supervised learning for successful fruit size classification.

4.3. Error Analysis

Mistakes in the Clustering-Based Approach are because it cannot create strong decision boundaries, hence cannot classify properly. Machine learning algorithms performed poorly without the depth data due to overlapping distributions, but with the addition of these features, the number of misclassifications was greatly minimized. Random Forest was the most reliable method, while ANN, with high accuracy, can further be optimized to use Depth Data to the best. Optimization in the future can be achieved through hyperparameter tuning and ensemble methods to attain maximum classification accuracy and minimum errors.

4.4. Visualizing Results

Figure 6 presents a confusion matrix for the top-performing model, showing the distribution of the correctly and incorrectly classified instances.

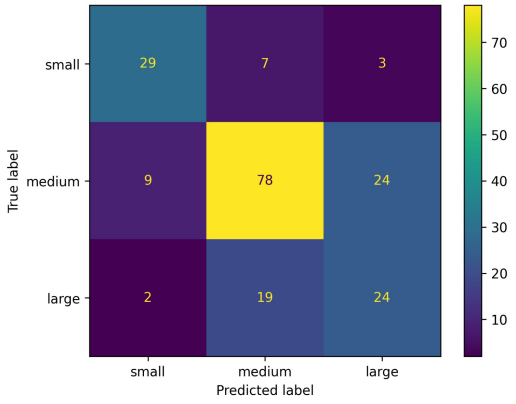

Figure 6: Confusion Matrix for Feature-Engineered Random Forest

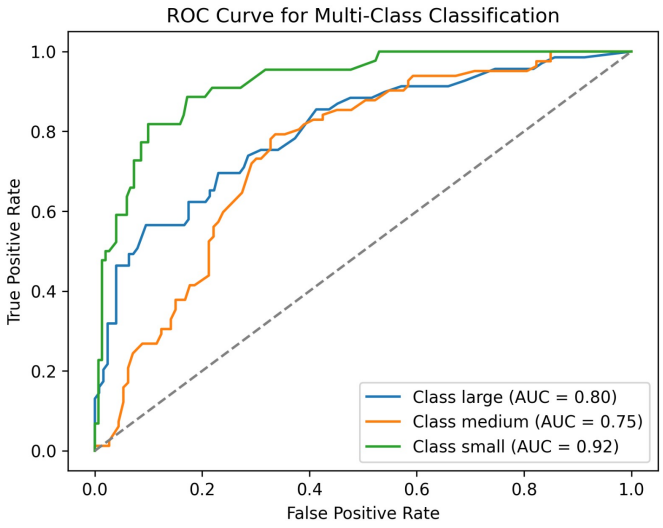

Figure 7: Receiver-Operator Charactertistics for Random Forest Classifier

Figure 7 depicts the receiver operating characteristic (ROC) curve, employed to graph the trade-off between the true positive rate and the false positive rate for all classes.

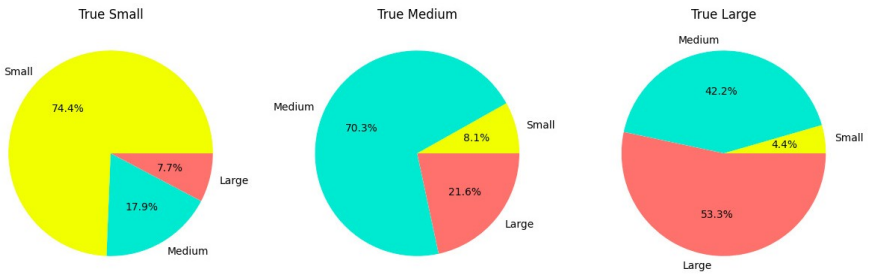

Figure 8: Class Wise Pie Charts of Predictions

Figure 9 shows the significance of each feature in the classification using RFC. It is clear that the feature depth is the most significant one, and thus, its inclusion in the classification makes the proposed method novel.

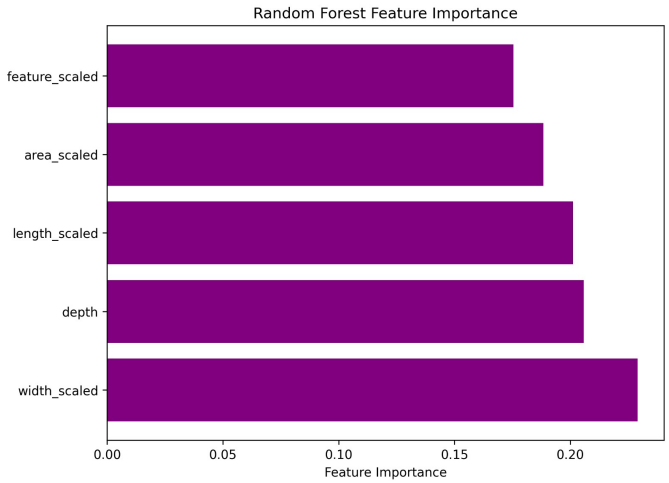

Figure 9: Feature Importance for Feature-Engineered Random Forest

4.5. Discussion and Future Improvements

The outcome suggests that Integration of Depth features and additional feature engineering greatly improves the classification performance. Further improvements can be investigated, such as the use of deep learning with a CNN-based method to further enhance accuracy by learning more intelligent feature representations automatically. Depth estimation algorithms can be made more robust to eliminate noise and enhance consistency to provide improved classification performance. The dataset can be further extended to support a variety of fruit types and environmental conditions to generalize better, leading to a more robust model for a wide variety of real-world applications.

5. Conclusion

This work introduced a system to infer fruit size from monocular depth images. Techniques employed varied from a basic breakpoint analysis to elaborate machine learning in proportion to improvement in classification performance in steps. The highest accuracy of about 67% was obtained employing the Random Forest Classifier, proving the role of Spatial Features like Depth. The key contributions of this paper are the significant impact of feature engineering, i.e., aspect ratio, center coordinates and use of **Depth Data**, on classification performance. Machine learning, particularly Random Forest, outperformed heuristic, ANN and clustering-based methods, corresponding to the power of data-driven models.

The paper contributes to precision agriculture through automatic fruit size classification, which can benefit yield estimation and quality assessment in horticulture. Furthermore transfer learning and pre-trained deep learning models may further enhance classification accuracy, without requiring extensive feature engineering. Addressing these challenges and exploring more advanced methodologies, future studies can enhance automatic fruit classification as stronger, scalable, and efficient for agricultural needs.

References

[1] Mar Ferrer-Ferrer et al. "Simultaneous fruit detection and size estimation using multitask deep neural networks". In: *Biosystems Engineering* 233 (2023), pp. 63–75. ISSN: 1537-5110. DOI: 10.1016/j.biosystemseng.2023.07.010. URL: https://www.sciencedirect.com/science/article/pii/S1537511023001526.

[2] Katrin Lasinger et al. "Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer". In: *CoRR* abs/1907.01341 (2019). arXiv: 1907.01341. URL: http://arxiv.org/abs/1907.01341.

[3] Guichao Lin et al. "Fruit detection in natural environment using partial shape matching and probabilistic Hough transform". In: *Precision Agriculture* 21 (Feb. 2020). DOI: 10.1007/s11119-019-09662-w.

[4] Wei Liu et al. "SSD: Single Shot MultiBox Detector". In: *CoRR* abs/1512.02325 (2015). arXiv: 1512.02325. URL: http://arxiv.org/abs/1512.02325.

[5] Rushabh Maru et al. "Improved Faster RCNN for Ripeness and Size Estimation of Mangoes with multi-label output". In: *2024 International Conference on Computational Intelligence and Network Systems (CINS)*. 2024, pp. 1–8. DOI: 10.1109/CINS63881.2024.10864434.

[6] Valeriano Méndez et al. "In-Field Estimation of Orange Number and Size by 3D Laser Scanning". In: *Agronomy* 9.12 (2019). ISSN: 2073-4395. DOI: 10.3390/agronomy9120885. URL: https://www.mdpi.com/2073-4395/9/12/885.

[7] Juan Miranda et al. "Fruit sizing using AI: A review of methods and challenges". In: *Postharvest Biology and Technology* 206 (Sept. 2023), p. 112587. DOI: 10.1016/j.postharvbio.2023.112587.

[8] John Olusegun. "Image-Based Size Estimation Using Computer Vision Technique". In: (Nov. 2024).

[9] John Olusegun. "Measuring Object Size Using Depth Estimation from Monocular Images". In: (Nov. 2024).

[10] Hetal Patel, R.K. Jain, and M.V. Joshi. "Automatic segmentation and yield measurement of fruit using shape analysis". In: *International Journal of Computer Applications in Technology* 45 (May 2012), pp. 19–24.

[11] Hari Pichhika, Priyambada Subudhi, and Raja Yerra. "On-tree mango detection and size estimation using attention-enhanced mangoYOLO5 and XGBoost regression". In: *Journal of Food Measurement and Characterization* (Jan. 2025), pp. 1–17. DOI: 10.1007/s11694-025-03114-y.

[12] Joseph Redmon et al. "You Only Look Once: Unified, Real-Time Object Detection". In: *CoRR* abs/1506.02640 (2015). arXiv: 1506.02640. URL: http://arxiv.org/abs/1506.02640.

[13] Shaoqing Ren et al. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: *CoRR* abs/1506.01497 (2015). arXiv: 1506.01497. URL: http://arxiv.org/abs/1506.01497.

[14] Seppe Rogge et al. “A 3D contour based geometrical model generator for complex-shaped horticultural products”. In: *Journal of Food Engineering* 157 (2015), pp. 24–32. ISSN: 0260-8774. DOI: https://doi.org/10.1016/j.jfoodeng.2015.02.006. URL: https://www.sciencedirect.com/science/article/pii/S0260877415000539.

[15] Bingkai Wang et al. “A Smart Fruit Size Measuring Method and System in Natural Environment”. In: *Journal of Food Engineering* 373 (2024), p. 112020. ISSN: 0260-8774. DOI: 10.1016/j.jfoodeng.2024.112020. URL: https://www.sciencedirect.com/science/article/pii/S0260877424000864.

[16] Dandan Wang et al. “Deep Learning Approach for Apple Edge Detection to Remotely Monitor Apple Growth in Orchards”. In: *IEEE Access* 8 (2020), pp. 26911–26925. DOI: 10.1109/ACCESS.2020.2971524.

[17] Hossam M. Zawbaa et al. “Automatic fruit classification using random forest algorithm”. In: *2014 14th International Conference on Hybrid Intelligent Systems*. 2014, pp. 164–168. DOI: 10.1109/HIS.2014.7086191.