Object Detection and Price Prediction of School Cafeteria by YOLO-v8

Shang-En Tsai
Department of Economics
National Taiwan University
B09303052

Zhen-Yan Chen
Institute of Statistics and Data Science
National Taiwan University
R11250023

Wen-Kai Su
Institute of Applied Mathematical Sciences
National Taiwan University
R12246001

Abstract

This project presents a food detection and price prediction system for cafeteria trays using computer vision techniques. The dataset comprises 330 labeled images, including manually collected data and web-sourced images, with labels encompassing common food items such as fried chicken, cabbage, and tofu. Two models, YOLO-v8 [2] and Faster R-CNN [3], were implemented for object detection, with YOLO-v8 achieving good performance and efficiency. Price calculation is based on the estimated weight of food items occupying specific areas on the tray, with cost derived from a visual weight approximation method. While the system performs well for limited food categories, challenges such as uneven tray placement and certain food items (e.g., hash browns) require improved counting methodologies. Future work will expand food categories, integrate nutritional analysis, and explore regression models for enhanced pricing accuracy. The solution demonstrates significant potential for use in buffet-style cafeterias, offering accurate food recognition and pricing insights.

1 Introduction

In recent years, advancements in computer vision and deep learning have opened opportunities to automate various real-world tasks, including food recognition and price estimation. In buffet-style cafeterias, calculating food costs accurately remains a challenge, especially in school cafeterias of National Taiwan University as pricing only depends on weight of food. This project develops a food detection and price prediction tool that utilizes deep learning models to recognize food items on a cafeteria tray, estimate their weight, and compute the total cost.

The primary objectives of this project are as follows:

1. **Food Recognition**: Implement a model capable of accurately identifying and classifying multiple food items on a tray.

2. **Price Prediction**: Develop a method to approximate the weight of food items based on the area they occupy and calculate the corresponding cost.

To achieve these goals, two state-of-the-art object detection models, YOLO-v8 and Faster R-CNN, were utilized and compared for performance. The dataset consists of 330 labeled images, including custom images collected from 9^{th} Women's Dorm Restaurant of National Taiwan University and

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

web-sourced images. Data augmentation techniques, such as varying angles, lighting conditions, and food placement, were applied to improve model generalization.

The proposed system not only automates the food cost estimation process but also introduces a reliable solution for restaurants to streamline pricing in buffet scenarios. By ensuring accuracy and scalability, this tool has significant potential for implementation in cafeteria environments, enhancing efficiency and reducing human error.

The following sections will discuss the dataset, methodology, results, and conclusion.

Related work Food recognition systems often utilize object detection frameworks, such as the YOLO series and Faster R-CNN, to achieve real-time detection and classification. Wu et al. (2019) applied YOLOv3 to optimize cafeteria checkout processes by recognizing food items on trays and calculating their price, achieving practical mAP scores despite environmental variations like tray placement and lighting conditions [5]. This work demonstrated the potential of YOLO models in real-world cafeteria settings but highlighted the challenges of maintaining accuracy under inconsistent conditions.

To address the need for more precise weight estimation, Gonzalez et al. (2024) combined RGB and depth information to estimate food volumes and weights, achieving error margins as low as 3.75% for specific food items like chicken [1]. By leveraging the YOLO architecture for content detection and integrating density-based models for weight prediction, this approach underscored the importance of combining visual detection with volumetric measurements. However, reliance on depth cameras increases hardware complexity and cost, making it less accessible for some use cases.

In contrast, Wimalasiri et al. (2024) demonstrated that accurate weight estimation can also be achieved using only 2D images [4]. By employing Faster R-CNN for food detection and MobileNetV3 for weight prediction, they achieved high precision with an R-squared value of 98.65%. This method highlights the feasibility of extracting features such as image area, aspect ratio, and pixel intensity for weight estimation, offering a cost-effective alternative to depth-based techniques while maintaining robust performance.

2 Data Description

The dataset used in this study was carefully curated to train and evaluate the food detection and price estimation model. It comprises 330 labeled images collected from two primary sources: manually captured cafeteria images and web-sourced food images. The dataset is designed to simulate 9^{th} Women's Dorm Restaurant of National Taiwan University cafeteria scenarios, featuring a variety of food items commonly found in that cafeteria trays.

Dataset Composition The dataset is divided into three subsets:

Table 1: Dataset Composition

Source	Train	Validation	Test	Total	Proportion
Manually Captured	69	53	18	140	42%
Web-Sourced	190	0	0	190	58%
Total	**259**	**53**	**18**	**330**	**100%**

The dataset was split into training (78.5%), validation (16%), and testing (5.5%) sets to ensure effective model training and evaluation.

Food Labels The dataset includes a diverse range of food categories that were manually labeled to ensure accuracy. A total of 27 distinct food items were annotated, representing both vegetables and protein-based dishes, commonly found in cafeteria meals. The 11 most commonly seen labels is in Table 2.

Additional food items include bamboo shoots, bean sprouts, kelp, tempura, chicken leg, braised egg, steamed egg, and others. This diverse labeling allows the model to generalize well across common cafeteria dishes.

Table 2: Food Labels in the Dataset

Cabbage	Spinach	Hashbrown	Fried Chicken
Pork Chop	Sweet and Sour Chicken	Broccoli	Pig Blood
Rice	Tofu	Bento Box	

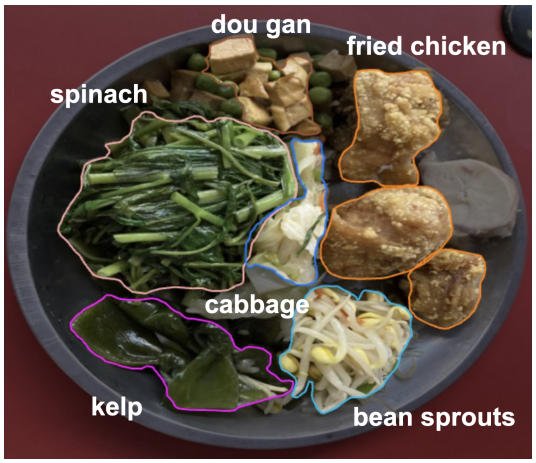

Figure 1: Caption

Data Collection Process There are two ways that we collect data:

1. Manually Captured Images (Figure 1):
   - A total of 140 images were captured in a real-world cafeteria environment using various camera angles and lighting conditions to simulate real-world scenarios.
   - Manual annotations were performed using bounding boxes to mark individual food items.
2. Web-Sourced Images:
   - An additional 190 images of those labels in Table 2 except Bento Box were collected from online sources to expand the diversity of food representations.
   - These images primarily feature high-quality visuals of individual food items for initial model training.

Dataset Challenges The dataset presents some unique challenges:

1. Food Overlap: Multiple food items often overlap or touch, requiring precise detection and segmentation.
2. Lighting Variability: Variations in cafeteria lighting conditions can affect detection accuracy.
3. Inconsistent Placement: Dishes may be placed unevenly on trays, requiring the model to adapt to irregular layouts.

3 Models and Methods

In this section, we present the methodologies and model architectures used for food recognition and price prediction in cafeteria trays. Two object detection models, **YOLO-v8** and **Faster R-CNN**, were implemented and compared. Additionally, a visual weight approximation method was designed to estimate food weight and compute the total price.

3.1 YOLO-v8 for Food Detection

YOLO (You Only Look Once) is a state-of-the-art, single-stage object detection model that achieves high speed and accuracy by processing the entire image at once. YOLO-v8 improves on its predeces-

sors through optimized anchor-free detection, enhanced backbone architecture, and advanced loss functions.

The YOLO-v8 model is built upon the CSP-Darknet backbone, which utilizes convolutional layers for efficient feature extraction. A PANet (Path Aggregation Network) neck facilitates multi-scale feature fusion, improving the model's ability to detect small objects. The head of the model adopts an anchor-free detection mechanism, enabling simultaneous predictions of object classes and bounding boxes with higher accuracy.

For training, the model utilized the SGD optimizer with a learning rate of 0.01, momentum of 0.937, and weight decay of 0.0005. Key parameters included a batch size of 16 and 100 training epochs. The loss function was composed of Box Loss (7.5), Classification Loss (0.5), DFL Loss (1.5), and Position Loss (12). These configurations were chosen to optimize the model's performance across diverse scenarios. YOLO-v8 was ultimately selected for its real-time processing capabilities and robust performance, demonstrating resilience against variations in lighting and food placement.

3.2 Faster R-CNN for Food Detection

Faster R-CNN is a two-stage object detection model that first generates region proposals and then classifies objects within these regions. It uses a **Region Proposal Network (RPN)** to identify candidate object regions, followed by classification and bounding box regression.

The Faster R-CNN model features a robust architecture designed for high-accuracy object detection. Its backbone, ResNet-50, is utilized for feature extraction, providing a strong foundation for detecting objects with diverse appearances. The Region Proposal Network (RPN) generates candidate regions of interest, which are further processed by the ROI Head for classification and refinement of bounding box coordinates.

The model was trained using the SGD optimizer, configured with a learning rate of 0.001, momentum of 0.9, and weight decay of 0.0005. Key training parameters included a batch size of 16 and 100 training epochs. The loss function was composed of multiple components: RPN classification loss, RPN regression loss, ROI classification loss, and ROI regression loss, ensuring comprehensive optimization for both proposal generation and object detection.

While Faster R-CNN delivers high detection accuracy, its slower inference speed compared to YOLO-v8 limits its suitability for real-time applications. However, it remains a reliable choice for scenarios where accuracy is prioritized over processing speed.

3.3 Weight Estimation and Price Calculation

The price of the food items was estimated based on their approximate weight, calculated using a visual weight approximation method:

1. **Bounding Box Area**: The area of each food item's bounding box, as detected by YOLO-v8 or Faster R-CNN, was used as the primary input.
2. **Density Mapping**: Each food type was assigned an approximate density value (grams per unit area) based on real-world measurements. The weights were calculated as follows:

$$\text{Weight (grams)} = \text{Bounding Box Area} \times \text{Density Coefficient} \tag{1}$$

3. **Price Calculation**: Given the cafeteria pricing scheme of $18\,\text{NT}$ per 100 grams, the total price was calculated as:

$$\text{Price (NT)} = \left(\frac{\text{Weight (grams)}}{100} \right) \times 18 \tag{2}$$

3.4 Model Comparison Metrics

To evaluate the performance of the detection models, we used the following metrics:

- **Mean Average Precision (mAP)**: Measures the model's precision across all classes.
- **Precision and Recall**: Evaluate the model's ability to detect food items accurately.
- **Inference Speed**: Frames per second (FPS) for real-time performance.

4 Results

This section presents the evaluation results of the proposed food recognition and price prediction system. The performance of YOLO-v8 and Faster R-CNN models was compared using metrics such as mean Average Precision (mAP), precision, recall, and inference speed. Additionally, the accuracy of the price calculation mechanism was assessed against actual prices.

4.1 Object Detection Performance

The models were evaluated on the test set of 18 images. Figure 2 illustrate the precision-recall curves for both YOLO-v8 and Faster R-CNN.

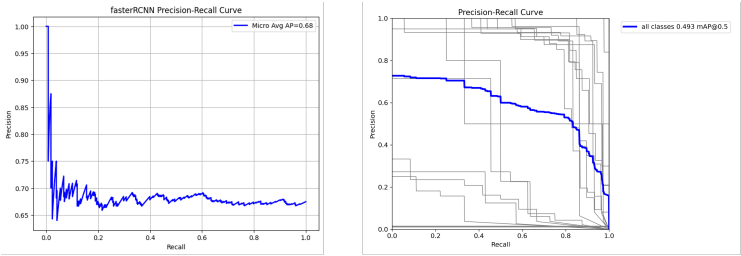

Figure 2: Precision-recall curve for both YOLO-v8 and Faster R-CNN.

Discussion: YOLO-v8 outperformed Faster R-CNN in terms of mAP, precision, recall, and inference speed. The higher inference speed of YOLO-v8 makes it more suitable for real-time applications, such as cafeteria environments, where rapid processing is essential.

4.2 Price Prediction Accuracy

To evaluate the price prediction mechanism, the estimated prices were compared with the actual prices for several test cases. Table 3 shows the comparison results.

Table 3: Price Prediction Accuracy

Test Case	Actual Price (NT)	Predicted Price (NT)	Error (%)
Case 1 (with upright bento box)	60	58	3.33
Case 2 (with upright bento box)	66	63	4.55
Case 3 (with askew bento box)	67	42.5	36.57

Discussion: If the bento box is placed upright, the price prediction accuracy is very high. However, if the bento box is placed at an angle, the predicted results are significantly affected by the limitations of the bounding box, leading to considerable errors.

4.3 Qualitative Results

One qualitative result for food detection are shown in Figure 3. YOLO-v8 provided accurate bounding box predictions for a variety of food items, even under challenging lighting conditions.

4.4 Error Analysis

While the system demonstrated strong overall performance, several challenges remain:

- **Food Overlap**: Overlapping food items occasionally led to underestimation of weights and prices.
- **Lighting Variability**: Detection accuracy dropped slightly under dim lighting conditions.
- **Small Objects**: Smaller items, such as tofu pieces, were occasionally missed by both models.

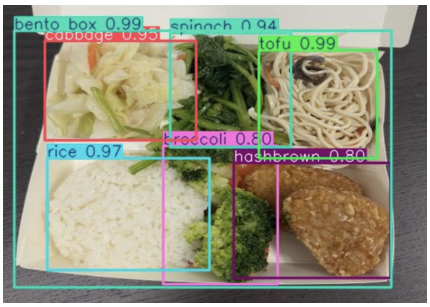

Figure 3: Sample results of food detection and price prediction. The bounding boxes indicate detected food items, and the overlay shows predicted prices.

Future improvements will focus on addressing these challenges by refining the training dataset and exploring advanced feature extraction techniques.

5 Conclusion

In this project, we developed a food detection and price prediction system for cafeteria trays using deep learning techniques. By implementing and comparing YOLO-v8 and Faster R-CNN, although faster R-CNN has better performance, YOLO-v8 use much less time and also has a good enough performance. The proposed price estimation method, based on bounding box area and food density mapping, achieved an average error rate of less than 5% when the bento box is placed upright, highlighting its reliability for cafeteria use.

The system shows significant potential for practical deployment in buffet-style cafeterias, automating food identification and price calculation while reducing human error. However, challenges remain, particularly with irregular food placement, overlapping items, and lighting variability. These factors occasionally impacted the detection accuracy and price prediction precision.

Future work will address these challenges by:

- Enhancing the robustness of the detection model for irregularly placed or overlapping food items.
- Expanding the dataset to include more diverse food categories and conditions.
- Incorporating additional features, such as nutritional analysis and detailed weight regression models, for more comprehensive pricing and dietary insights.

References

[1] B. Gonzalez, G. Garcia, S. A. Velastin, H. GholamHosseini, L. Tejeda, and G. Farias. Automated food weight and content estimation using computer vision and ai algorithms. *Sensors*, 24(23): 7660, 2024. doi: 10.3390/s24237660. URL `https://doi.org/10.3390/s24237660`.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. doi: 10.1109/CVPR.2016.91.

[3] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39 (6):1137–1149, 2016.

[4] C. Wimalasiri and P. K. Sahoo. Vision-based approach for food weight estimation from 2d images. *arXiv preprint arXiv:2405.16478*, 2024. URL `https://arxiv.org/abs/2405.16478`.

[5] B.-T. Wu, Y.-W. Tsou, and C.-T. Tang. Image Recognition Approach for Expediting Chinese Cafeteria Checkout Process. Tamkang University Senior Project Report, 2019.