Robotica (2024), **42**, pp. 2108–2124
doi:10.1017/S0263574724000122

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Vision-based food handling system for high-resemblance random food items

Yadan Zeng, Yee Seng Teoh, Guoniu Zhu, Elvin Toh and I-Ming Chen

Robotics Research Centre of the School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore
Corresponding author: Yadan Zeng; Email: yadan001@e.ntu.edu.sg

Received: 29 April 2023; **Revised:** 7 December 2023; **Accepted:** 11 December 2023;
First published online: 21 February 2024

Keywords: food handling system; pose estimation; object detection; food dataset; high-resemblance item; closed-loop strategy

Abstract

The rise in the number of automated robotic kitchens accelerated the need for advanced food handling system, emphasizing food analysis including ingredient classification pose recognition and assembling strategy. Selecting the optimal piece from a pile of similarly shaped food items is a challenge to automated meal assembling system. To address this, we present a constructive assembling algorithm, introducing a unique approach for food pose detection–Fast Image to Pose Detection (FI2PD), and a closed-loop packing strategy. Powered by a convolutional neural network (CNN) and a pose retrieval model, FI2PD is adept at constructing a 6D pose from only RGB images. The method employs a coarse-to-fine approach, leveraging the CNN to pinpoint object orientation and position, alongside a pose retrieval process for target selection and 6D pose derivation. Our closed-loop packing strategy, aided by the Item Arrangement Verifier, ensures precise arrangement and system robustness. Additionally, we introduce our *FdIngred328* dataset of nine food categories ranging from fake foods to real foods, and the automatically generated data based on synthetic techniques. The performance of our method for object recognition and pose detection has been demonstrated to achieve a success rate of 97.9%. Impressively, the integration of a closed-loop strategy into our meal-assembly process resulted in a notable success rate of 90%, outperforming the results of systems lacking the closed-loop mechanism.

1. Introduction

Robotics has expanded significantly in response to the growing demand for higher efficiency and productivity in a range of production fields [1]. The field of kitchen automation has garnered substantial attention. In comparison to conventional labor-intensive approaches, automated robotic kitchens offer prospective advantages including extended operational hours, heightened safety measures, and superior quality assurance [2]. The mechanization of meal assembling procedures has emerged as a prominent area of focus within the subject of kitchen automation. This procedure involves the utilization of robots to identify and select the most suitable ingredient from piled high-resemblance random food items, such as the pattern shown in Fig. 1 [3–5].

Contemporary food service paradigms rely heavily on manual labor, with ingredient assembly often necessitating the collaborative efforts of 5–7 human operators, followed by a secondary team supervising assembly verification, transportation, and packaging. This procedure, which depends heavily on human labor, unintentionally demonstrates inefficiencies that are typified by extra waiting times and coordination difficulties.

Though current robotic kitchen systems exhibit commendable competencies like meal packaging, transporting ingredients, or dish component understanding, they often treat the serving container as the primary unit of operation. Limited attention has been given to the investigation of front-end robots interacting directly with individual ingredients. Instead, our study aims to address this gap, emphasizing

© The Author(s), 2024. Published by Cambridge University Press.

https://doi.org/10.1017/S0263574724000122 Published online by Cambridge University Press

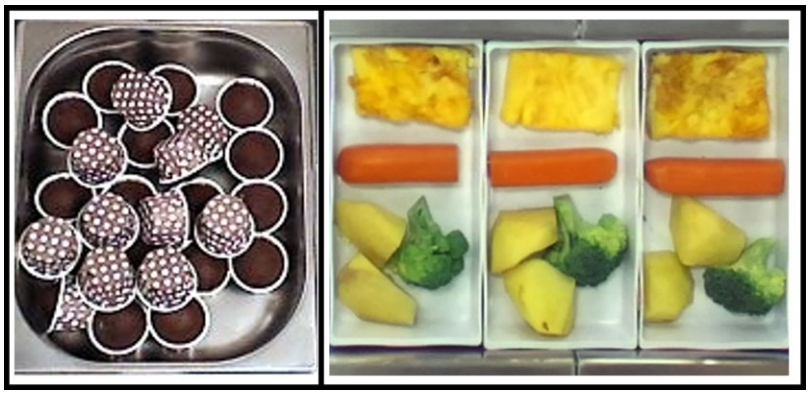


Figure 1. *Examples of piled high-resemblance food ingredients.*

the challenges intrinsic to assembling visually similar food items, each with its distinct shape and no fixed model, into specific containers. The task of accurately identifying and differentiating these visually comparable items is further complicated by their physical resemblances. Moreover, the intricate characteristics of certain food products and the increased difficulty in robotic manipulation resulting from the presence of moisture contribute to further complicating the process. To address these intricacies, we proposed a meal-assembly framework. Central to our methodology is the integration of vision detection at both the ingress and egress phases. This dual integration facilitates comprehensive pose comprehension of different input ingredients and allows iterative refinements in food arrangement via a closed-loop strategy. Such a framework highlights the critical challenge: how can one accurately determine the 6D pose of a target ingredient within a pile of randomly organized food products that share a high degree of similarity, thus ensuring its correct placement and concurrently maximizing the assembly success rate?

Recent studies in 6D object pose estimation have primarily relied on image-based techniques [6] and convolutional neural network (CNN) [7] methodologies. Traditional approaches often leaned on pattern matching with CAD models [8, 9]. Existing CNN-based methods, such as PoseCNN [10], have stringent requirements like detailed camera parameters. Meanwhile, methods like Normalized Object Coordinate Space (NOCS) [11] or Point-wise 3D Keypoints Voting Network (PVN3D) [12], primarily rely on 3D models. For items with inherent variations, such as food, these methods often fall short. Existing research on food grasping often relies on either multi-perspective imagery or CAD models, both of which might not be ideal for quick assembly tasks.

In the domain of food grasping, specific items, e.g., apples, guavas, and grapes, have been studied extensively. Fruit-harvesting methodologies [13–15] initiate the process by fruit segmentation and identification and then employ geometric relationships to infer the pose. Similarly, [16] also employs post-classification, utilizing the relationships of continuous camera captures to determine poses. On the other hand, [17] requires the generation of CAD models for apples, using projection correlations to achieve pose estimation. Notably, these methodologies distinguish between food recognition and pose estimation. For pose retrieval stage, they often necessitate images from multiple angles with detailed camera specifics or the use of CAD food models. Such prerequisites can be restrictive, particularly in rapid grasping and assembly applications, and might not adapt well in settings without predefined models.

In response to identified shortcomings in existing methodologies, this work introduces Fast Image to Pose Detection (FI2PD). This methodology, underpinned by a one-shot CNN paradigm, exclusively relies on image data. Effectively, FI2PD is engineered to derive a 6D pose from the immediate visual input. The design extracts planar rotation with object recognition, which is optimal for in-plate object retrieval using Delta robotic configurations. To further refine pose estimation, the methodology incorporates the Range of Orientation (ROO) data to extract 10 keypoints, enabling precise computation of the 6D pose. To enhance precision, we have constructed a comprehensive food dataset covering nine

https://doi.org/10.1017/S0263574724000122 Published online by Cambridge University Press

categories, incorporating fake, real-world, and synthetic variants, making it an excellent resource for 6D pose estimation in food selection tasks.

Given these advancements, we have also developed a vision-based food handling system to optimize the meal-assembly phase. However, challenges arise from inaccuracies in object positioning and orientation, and potential failures during manipulation, such as failed grasps or unintended movements. These challenges can significantly impact task execution, especially during packing. Complications may further ensue due to inconsistent item placements and the intrinsic constraints of the gripper, posing difficulties in the precise handling of various food items. To mitigate these challenges, we have incorporated a closed-loop approach leveraging our vision-driven Item Arrangement Verifier (IAV) equipped with a modifiable tolerance attribute. This integration empowers the robot with real-time corrective capabilities, bolstering system robustness. In addition, we have augmented our solution with a modular soft gripper, designed with versatile finger units. The gripper possesses the capability to adapt to diverse food products and has been specifically engineered to handle them with a gentle touch, minimizing potential harm. Our work has the following key contributions:

- **Advanced Framework for Meal Assembly:** Our study underscores the pivotal role of direct robot engagements with high-resemblance, randomly stacked food items. We introduce a comprehensive framework specifically for such intricate food assembly tasks.
- **FI2PD Technique and the *FdIngred328* Dataset:** Our FI2PD, a one-shot CNN-based pose estimation approach, concentrates on the intricate arrangement of visually similar food items using just image data. Complementing this, our *FdIngred328* dataset combines several scenarios involving synthetic, simulated, and authentic food items, thereby enriching the scope of object detection and 6D pose estimation in the food domain.
- **Vision-Guided Closed-Loop Handling:** The integration of our IAV into a closed-loop architecture serves to correct ingredient placement and robot interactivity, thereby solidifying the overall robustness of our system in real-world applications.

The rest of this paper is organized as follows. After reviewing related work in Section 2, we introduce our meal-assembly framework along with proposed method FI2PD for 6D object pose estimation, the vision-guided closed-loop strategy, and the proposed *FdIngred328* dataset in Section 3. Experimental results are reported in Section 4 to confirm the effectiveness of our method. Finally, conclusions are highlighted in Section 5.

2. Related work

2.1. Robot-assisted food assembly systems

In recent years, there has been a growing interest in robot-assisted food assembly systems [18]. The current rise in interest can be attributed to the increasing need for kitchen automation, which is further supported by notable developments in robotic manipulation and vision technology. The JLS Company provides a range of food assembly systems specifically designed for the processing of fresh & frozen meats, bakery products, chocolates, and similar items [19]. In the context of their production line, food products are arranged on a conveyor belt in a manner that ensures they are either aligned or distributed without any overlapping. Such equipment is not suitable for the assembly of ingredients or pre-cooked meals. The robotic systems described in ref. [20] and ref. [21] are designed specifically for granular food materials, and they don't consider the angle of food grasping. Moreover, a substantial amount of research has been primarily concentrated on designing grippers for food handling [22]. Ref. [23] developed a specialized pneumatically powered needle gripper designed specifically for handling finely shredded food. Meanwhile, a circular shell gripper, composed of a sturdy external shell and four soft internal air chambers, was devised to attain versatile handling postures and twisting maneuvers while gripping diverse food and beverage products [24].

https://doi.org/10.1017/S0263574724000122 Published online by Cambridge University Press

One notable differentiation between the existing solutions and our approach is that these food systems often focus on the categorization of non-overlapping food items and do not place significant emphasis on the rotational management of these products. Their grippers are designed to grasp single portions. In contrast, our system is focused on accurately identifying the 6D pose of individual items within stacks of similar-looking random food items. This enables the sequential arrangement of these items into one meal-tray.

2.2. 6D pose estimation

In terms of pose estimation, traditional methods used template matching with 3D models [25], but more recent studies have shifted towards deep learning-based methods [26–28], which provide accurate pose estimation of objects in cluttered scenes. The most common method, PoseCNN, calculates the 3D translation by localizing the center of the object to estimate its distance from the camera and regressing the convolutional features to a quaternion representation. Wang et al. [11] proposed a two-stage method that generates the NOCS map via CNN and then aligns it with the depth map to estimate the 6D pose and size of the objects. These recent developments demonstrate the potential of deep learning-based methods in achieving accurate and robust 6D pose estimation with image input.

Pose estimation for food has consistently been a challenging issue. The common methodology is initially classifying the food type, followed by employing post-processing techniques to determine the orientation of the meal. For instance, [15] identifies and segments grapes, and then calculates the central axis of a cylindrical model representing the grape. Ref. [16] also adopts a post-classification method, leveraging the relationships between consecutive camera captures to infer poses. Typically, these algorithms separate the task of food recognition from that of pose estimation. They often require images captured from multiple perspectives with detailed camera specifications or the incorporation of CAD models of food items.

2.3. Datasets of food

Datasets focusing on food ingredients are not uncommon. However, many existing datasets primarily serve the purpose of meal analysis. For instance, datasets such as ref. [29–32] offer images of dishes and their respective recipes, and are utilized for recognizing food ingredients. Other research endeavors have shifted toward food harvesting and picking, thus datasets that mainly encompass fruits like apples and mangoes have emerged [33, 34]. These datasets primarily focus on the identification of whole fruits.

The dataset presented in ref. [35] takes into account cluttered food ingredients. Recognizing that human annotations can be error-prone due to closely packed foods and indistinct boundaries; this dataset was crafted by populating trays with high-quality 3D models of actual food pieces. These are purely synthetic datasets, used specifically for training instance segmentation models.

In addition to the utilization of 3D models for data augmentation, neural network-based data augmentation techniques have gained traction as they can generate data automatically. The Deep Convolutional Generative Adversarial Network (DCGAN) [36] is an unsupervised learning method that uses CNNs to develop both the generator and discriminator, proving adept at producing unique images. However, the suitable original data for GAN-based method is single image on which there is only a single object or separate objects.

3. Methodology

This section describes the overall food assembly framework with the pose estimation algorithm and the item arrangement checking for the closed-loop packing strategy. We also introduce our proposed dataset which can be used for not only food item detection but also the pose retrieval of it. In this section, we present a detailed explanation of our all-encompassing food assembly framework. This framework

https://doi.org/10.1017/S0263574724000122 Published online by Cambridge University Press

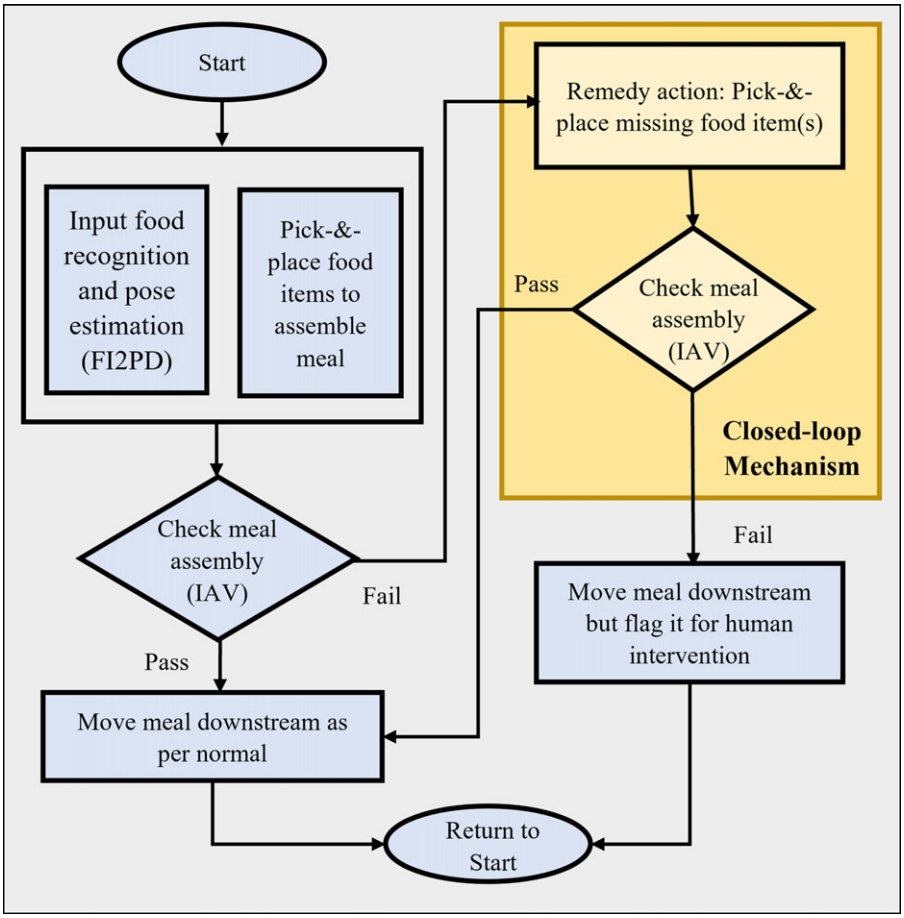


Figure 2. *Flowchart of the food assembly framework.*

effectively combines the posture estimate technique with our proprietary IAV to enable a closed-loop packing strategy. In addition, we present our proposed dataset, which serves as a flexible tool designed not only for accurate identification of food items but also for precise recovery of their spatial orientation.

3.1. Overall framework

In recent advancements in robotic kitchen systems, achieving precision and accuracy in meal assembly remains paramount. Our proposed system offers a structured, closed-loop approach, combining advanced robotics with intricate verification mechanisms to ensure the consistency of meal arrangements. The depicted structure can be observed in Fig. 2.

The process starts by recognizing the food and estimating its pose through our advanced FI2PD mechanism. This integral step provides the necessary information to accurately manipulate the food items. With the acquired information, the robot proceeds to pick and place the food items onto the designated meal-tray, aiming for a perfect assembly. The vision process and assembly action are controlled by the dual-thread architecture. Once the meal assembly is complete, our proprietary IAV checks the assembly for completeness and correctness. This verification stage is vital in the closed-loop manner to ensure that the meal meets the desired specifications. If the result meets the criteria for meal assembly, it signifies a successful arrangement. The meal then moves downstream for further processing or packaging, after which the system returns to its initial state, ready for the next meal assembly. However, if the IAV detects

https://doi.org/10.1017/S0263574724000122 Published online by Cambridge University Press

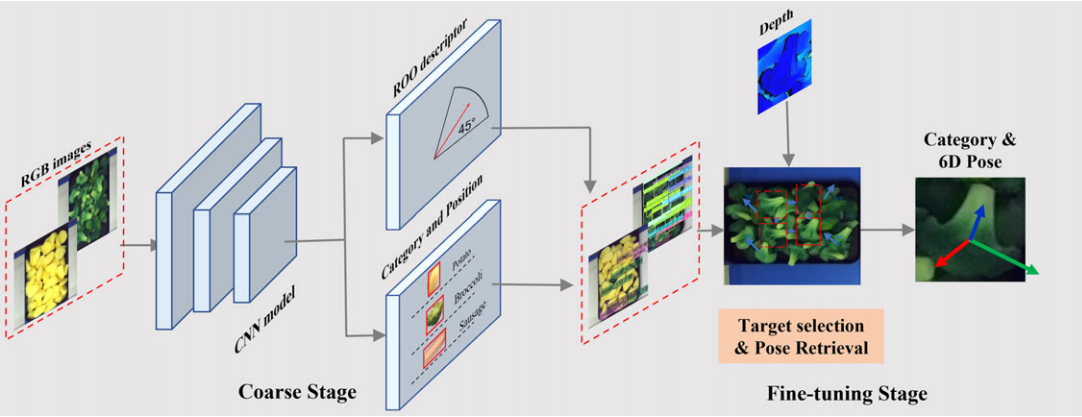


Figure 3. *Architecture of pose estimation process. The RGB images with announcement are put into the CNN model to predict the class, bounding box, and ROO of the items in the image. This information combined with the depth map is considered to calculate the 6D pose.*

an absence or misplaced item, the system takes corrective action. The system effectively identifies and transfers the absent food item from the input container again onto the tray. If the IAV still detects an error after the remedy action, instead of approving the meal for normal downstream processing, the system flags the meal for human intervention. This system is designed to effectively identify and correct any persistent irregularities, hence upholding the quality and accuracy of the assembly process.

This structured and closed-loop framework ensures both the efficiency and accuracy of our food assembly system, effectively integrating advanced robotics with intelligent verification for optimal results.

3.2. FI2PD pose determination framework

For robotic grasping in autonomous handling systems, accurate determination of position and orientation of the picking item is pivotal. In this instance, we proposed FI2PD method to search for the special information of the items in a coarse-to-fine process (Fig. 3). The method begins with the creation of a CNN-based object detector. The CNN model is adopted to the acquired data to obtain the category, position, and ROO descriptor for different food products. During the pose retrieval phase, categories, ROO, and 3D data are utilized to estimate the pose of the subject.

During the coarse phase, only the 2D RGB images are adopted as the input of CNN architecture. The CNN model outcomes the category, position, and ROO descriptor. The CNN model is built on YOLOv4 framework [37], is tailored to not only estimate the ROO descriptor but also categorize objects. Considering YOLOv4 is due to the fact that it has demonstrated performance on 2D object identification tasks, particularly when dealing with small-sized object. Our CNN model begins with down-sampling layers and generating feature maps with different scales. Within these maps, anchor boxes of distinct sizes are present. If the ground-truth center of the object lies within the grid, the grid is used to detect the object. Additionally, the yolo layer is used to predict the ROO, along with the bounding boxes, class, and confidence scores.

Fig. 4 demonstrates our main contribution to the network architecture. Compared to the YOLOv4 network, our framework introduces the ROO head architectures. This integration facilitates the capture of the angular features of objects, which are vital for accurate pose retrieval. Specifically, the head of our model employs three output scales, optimized to improve the detection of small objects, such as food ingredients, with corresponding original input sizes of 1/8, 1/16, and 1/32. The depth structure of the head encodes the bounding box offset, confidence, category, ROO, and anchor boxes. The output

https://doi.org/10.1017/S0263574724000122 Published online by Cambridge University Press

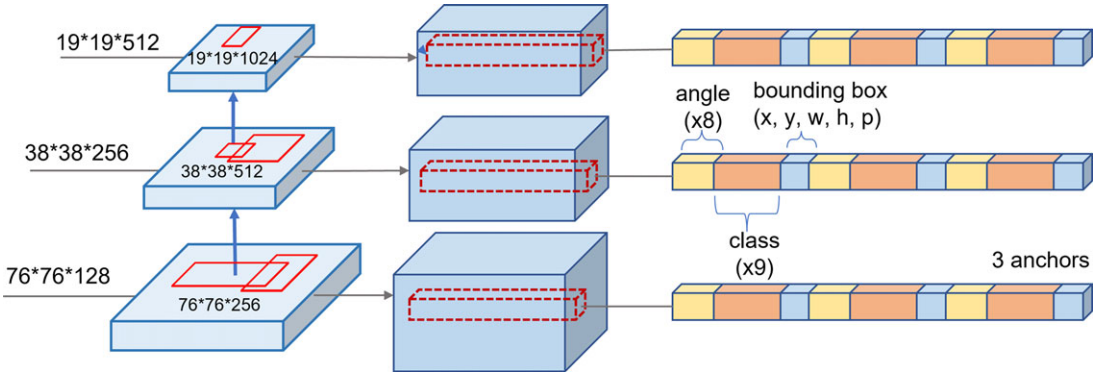


Figure 4. *Architecture of ROO head. The architecture incorporates three scales. The ROO component is predicted concurrently with the class and bounding box for each food item. Subsequently, non-maximum suppression is applied to determine the optimal anchor box among the available options, based on the acquired ROO, classes, and CIOU loss of bounding boxes.*

size of the angle head is $19 \times 19 \times N$, while the other two heads generate output sizes of $38 \times 38 \times N$ and $76 \times 76 \times N$, respectively, where N represents the depth of the head. Additionally, each scale output incorporates three distinct sizes of anchor boxes. The resolution of the ROO is set to 8, and for symmetrical objects, the resolution is increased to achieve this value. To improve the performance of model, we leveraged CSPDarknet53 as the backbone, along with Feature Pyramid Network and Path Aggregation Network, following the YOLOv4 architecture.

In the second stage, the information derived previously is considered to recover the 6D pose of the ingredient. The summary of the strategy of pose retrieval is listed in Algorithm 1. Given the typical piling arrangement of food ingredients, a grasping strategy is prioritized to optimize computational efficiency. The grasping strategy discerns the most appropriate target from a set of candidates first, factoring in the picking sequence and collision avoidance. By integrating the knowledge of position, course direction, and size of the candidates with depth information, the operating space of each candidate is determined. The picking target is chosen with the highest ranking of this operating space in terms of both height and collision risk. After that, the diagonal of the bounding box of the picking target is projected onto the object by using the 3D point cloud and sparsely sampled ten key points from it. Thus, the pose of object can be calculated using 10 key points that follow the direction indicated by the ROO. During this part of the procedure, the information from the 3D point cloud is used only for the 10 key points and the object's center, thereby significantly reducing the time spent searching and calculating.

3.3. Vision-guided closed-loop architecture

The arranging of objects in a sequential manner holds significant importance in product packaging applications, particularly when it comes to meal packaging. The difficulties encountered in robotic manipulations generally arise from the inherent properties of components, including their smooth and moist surfaces, as well as the lack of easily identifiable grab spots. During automated grasping procedures, there is an elevated likelihood of the misplacement or even omission of items, leading to an increased grasp failure rate. To address this concern, we propose the implementation of the IAV, which aims to provide real-time feedback regarding the efficacy of robotic grasping.

The IAV serves as an evaluative tool, assessing the results of food item grasping. The derived evaluation plays a crucial role in two key areas: the closed-loop packaging process and comprehensive quality assessment. In the framework of closed-loop packaging, the placement of each food item is rigorously monitored. The identification of irregularities, such as the absence of a designated ingredient within a container, is indicative of operational inconsistency. Upon the detection of such a failure, remedy action

https://doi.org/10.1017/S0263574724000122 Published online by Cambridge University Press

Algorithm 1: Pose Retrieval

Input: ROO, **B**, and **D**
Output: Q

Notations
ROO : *Range of Orientation*
B, **D** : Sets of *bounding boxes and Depth Information*
Bounding boxes contain class, position, and size
Q : Quaternion of the target
procedure
  ▷ *Determine the most desirable target from the candidates based on the grasping strategy*
  B_P ← *PriorityDetermination*(**B**, **D**)
  Target ← *CollisionAvoid*(**B_P**)
  ▷ *Compute the key points on the target*
  $\mathbf{P}_{key}$ ← *KeyPointSelect*(Target, **D**, ROO)
  Q ← *QuaternionCal*($\mathbf{P}_{key}$, ROO)
end procedure

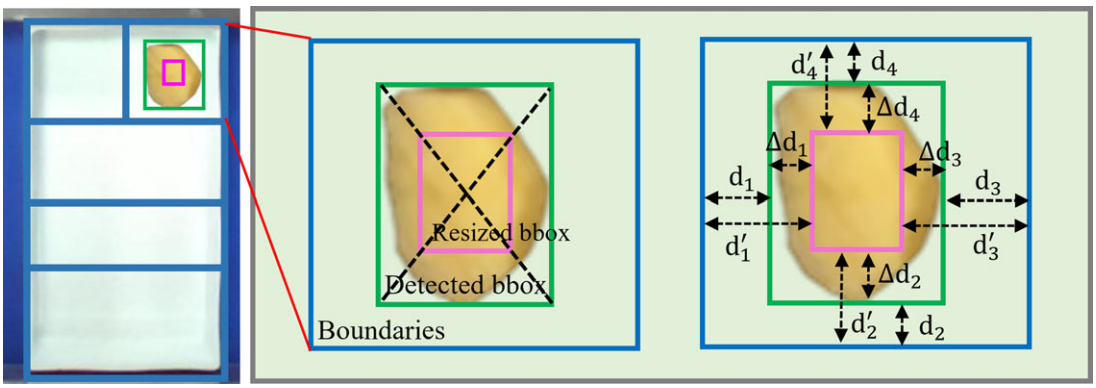


Figure 5. *Illustration of the item arrangement verifier and the tolerance mechanism.*

is initiated, wherein the specific category of the omitted food is re-grasped to address the deficiency. Simultaneously, there is an evaluation of the picking pose for the remaining items, with an emphasis on preserving alignment with the initial packaging plan. Further checks are conducted to verify whether any item partially extends beyond its designated boundary within the container. Despite these variations, the packing procedure for subsequent goods continues, with a persistent effort to ensure alignment with the preexisting plan. After the completion of the packing procedure, the IAV is moved to deliver a final evaluation score. Containers that do not fulfill the predetermined requirements are appropriately identified, eventually requiring manual intervention for rectification.

Fig. 5 illustrates our IAV approach with an emphasis on a distinct food item arrangement scenario. As depicted in left picture, a hypothetical boundary, denoted by the blue contour, is extrapolated to the geometric characteristics of the packaging box. This metric aids in ascertaining whether individual items are methodically positioned within their designated boundary confines. Simultaneously, the bounding region of individual food item is identified via our FI2FD technique. Importantly, in this context, the secondary phase of the FI2FD is deemed redundant, as the grasping pose of items is not pertinent.

To enhance the compactness of food item placement and bolster the robustness of grasping, a dynamically adjustable bounding box is introduced. This design modification extends greater latitude in positioning food items within their confines during the item arrangement verification process. As illustrated right schema in Fig. 5, both the bounding box and the boundary exhibit clear separations,

https://doi.org/10.1017/S0263574724000122 Published online by Cambridge University Press

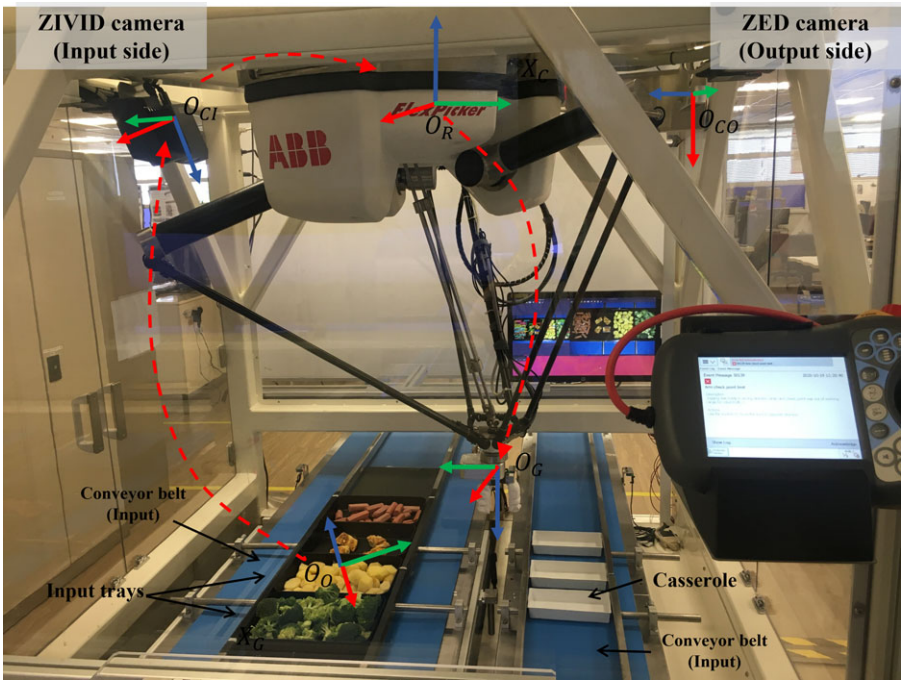


Figure 6. *System setup for automatic food handling process. This system contains image acquisition and detection (RGBD camera for both input side and output side), ABB Delta robotic, and soft gripper. The food ingredients are placed on the input conveyor belt and transferred into the casserole dishes on output conveyor belt for food assembling.*

denoted as d_i, where $i = 1, 2, 3, 4$, on all four cardinal directions. The cumulative extent of these separations represents the overarching gap, mathematically captured as $\sum_{i=1}^{4} d_i$. By shrinking the size of the bounding, it increases gaps in all four directions ($d_i^{'}, i = 1, 2, 3, 4$), allowing more area to place the resized bounding box within the boundary. The tolerance metric can be encapsulated as the summation of the gap differentials across the four directions, formulated as

$$Tolerance = \sum_{i=1}^{4} \Delta d_i, where \Delta d_i = d_i^{'} - d_i, \tag{1}$$

3.4. System setup and configuration

Fig. 6 presents a schematic illustration of the layout of our custom-built automatic food handling system, with both input and output conveyor belts for the assembly of food-trays and casserole meal-trays. Each food item that constitutes the meal is served in separate trays. The food items are gently picked and placed into each meal-tray with the assistance of Delta robot with a preset picking sequence (two pieces of potato cubes, followed by omelet, sausage, and broccoli).

It is crucial to identify the food items within the trays and determine their position for each operational cycle. The provided classification information is pivotal, facilitating the dynamic configuration of the gripper to adeptly handle different types of food. As shown in Fig. 7, the gripper adopts variable configurations depending on the shape of the food. For long-shaped food like sausage, the gripper will use a pinch mode, while for foods with an irregular geometry like broccoli, it will use a claw mode to ensure stable gripping. Furthermore, the use of pneumatic actuation facilitates precise control over the gripper via both positive and negative air pressures. This air pressure manipulation directly influences the bending of actuator, thereby producing varied grip strengths. Such adaptability is instrumental in

https://doi.org/10.1017/S0263574724000122 Published online by Cambridge University Press

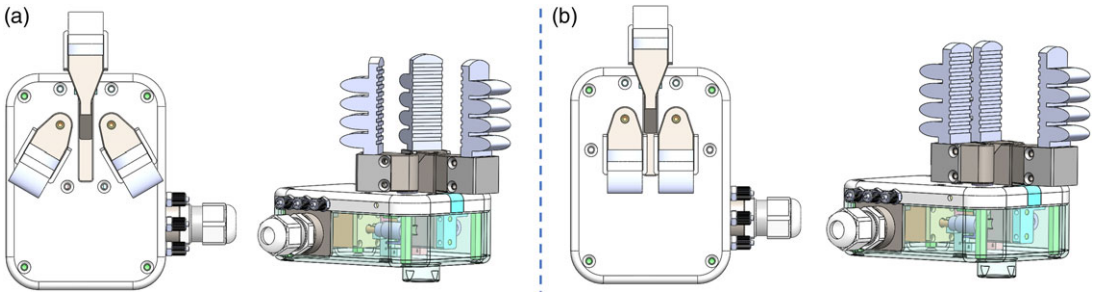


Figure 7. *Structure of the various grip poses. (a) top view and side view of claw position, thumb in and fingers rotate 30°; (b) top view and side view of pinch position, thumb out and fingers rotate 0°.*

ensuring a secure grasp while simultaneously safeguarding the integrity of food and minimizing any potential harm.

With regards to the item picking strategy, the robot was programed to handle the picking sequence and locate the target food item in the ideal gripping position using vision detection. During the placement phase, the food items were arranged in the proper sequence into a set of casserole meal-trays. The robotic conveyor tracking module relied on the analog input signal, which corresponds to the velocity of the conveyor belt, to precisely determine the spatial location of the meal-trays. The process of moving between several input trays and a single output meal-tray requires the implementation of a motion trajectory that is both efficient and effective [38]. By utilizing this dual-thread architecture, we have effectively separated the process of obtaining visual input from the processing of trajectory. The proposed architectural design guarantees real-time data exchange whenever the robot finishes placing a food item in its assigned compartment. Simultaneously, it ensures that the robot remains outside the camera's field of view while it is in operation. This design paradigm enables the simultaneous execution of the vision detection and the movement process, which significantly speeds up the entire pick-and-place procedure.

3.5. Dataset generation & augmentation

The absence of training data with adequate category, posture, combination, situation, and lighting diversity is a significant difficulty when training CNNs, particularly for datasets used in the application of meal assembling. For the application of food assembly, the location and pose of the food items are the most important information that is underrepresented in the current dataset [29, 30]. As such, we set up a dataset called *FdIngred328* regarding the high-resemblance random food items piled with different arrangements. The dataset consists of two types of data – real-world data and synthetic data. The real-world data comprises nine categories of food including potatoes, broccoli, sausage, and so on. To facilitate testing at any time, both fake and real food items are included in the dataset due to the perishability of real food. Some samples of different arrangements for each category are displayed in Fig. 8. Most of the food is stacked in multiple layers, although there are also some single-layer and individual piles. Compared to fake food with limited shapes, real food items are present in ever-changing shapes. For the annotation, ROO descriptor is added for use in the further pose estimation. ROO specifies the approximate orientation of the item based on the type of food, with a 45-degree resolution. Notably, such labels will be decreased by half if the object is symmetric, and only labels within 180 degrees will be retained to indicate orientation.

The real scenario in *FdIngred328* dataset was generated by manually collecting images using an RGBD sensor (ZED camera, Stereolabs Inc., San Francisco, USA) from a top view (as shown in Fig. 9). The images were then cropped to a suitable output size with a resolution of 328 × 328. To capture the natural lighting conditions, the items were placed on a black tray during image acquisition. However, setting up the dataset was challenging due to the time-consuming nature of data collection and annotation,

https://doi.org/10.1017/S0263574724000122 Published online by Cambridge University Press

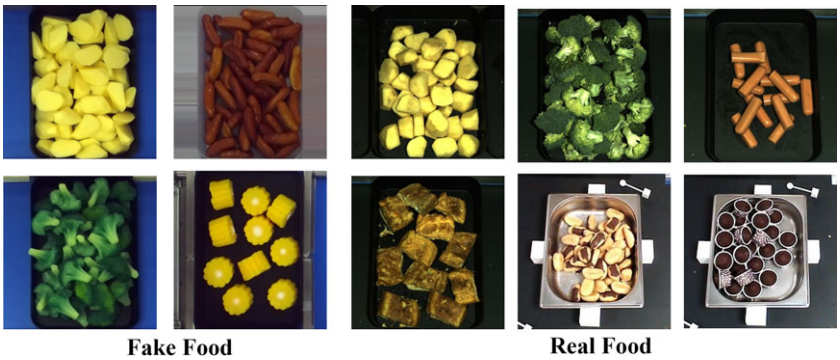


Figure 8. *Food samples from the FdIngred328 dataset. The left two columns show fake food items with multi-layer arrangement, while those of real food items are shown in the right three columns.*

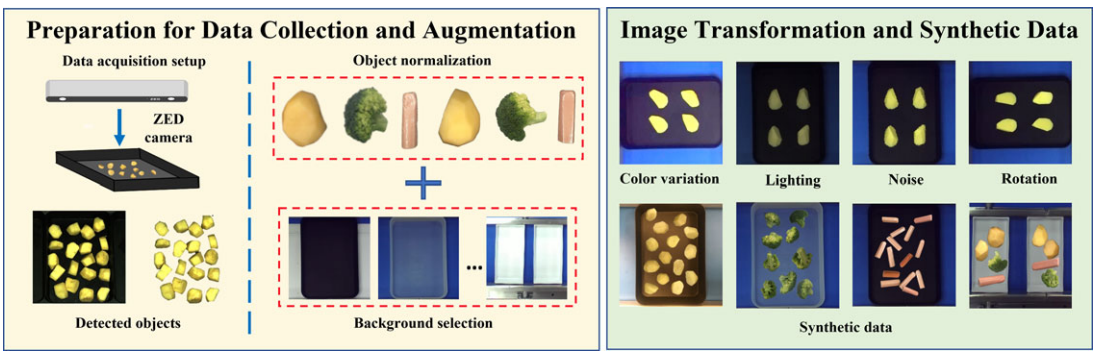


Figure 9. *Synthetic images generated by the combination of different objects and backgrounds. The original images of different types of items and background images are captured separately. The objects are randomly arranged on the background with known rotated angles after background removal and orientation initialization. These synthetic images create a great number of images with diverse scenarios.*

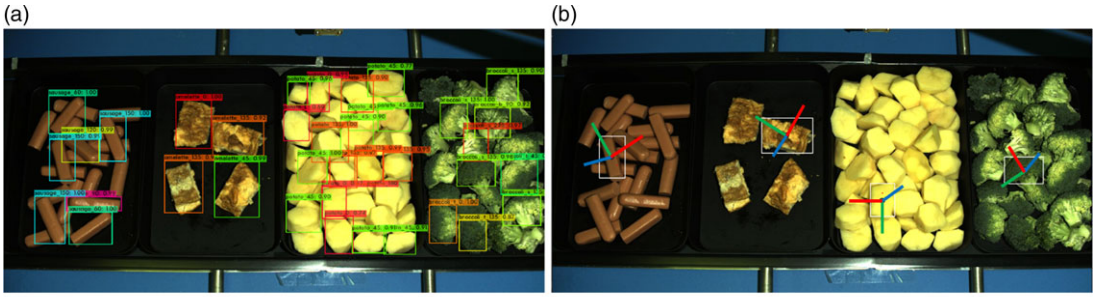
as well as the difficulty in preparing diverse food ingredients with varying shapes. To ensure the robustness of the dataset, synthetic data generated by the data augmentation techniques is also included. As depicted in Fig. 9, the generated data contains the one produced by image transformation techniques [39] such as rotation, image segmentation, color conversion, noise addition approaches, and even the combination of them. Furthermore, the synthetic data generation comprises four steps: (1) extracting the real food instances from real-world images, (2) collecting background data, (3) initializing the orientation, and 4) integrating the background data and the single items together in random pose. Simultaneously, annotation files are able to be generated as the position and orientation of the items that are known by us, through which the manual cost can be significantly reduced. Lastly, synthetic images can also be further increased by image manipulation techniques. These synthetic images create a great number of new images with diverse scenarios, making the dataset more robust.

4. Evaluation

4.1. Recognition and pose estimation

As described in Section 3.5 and Fig. 9, a synthetic dataset including both original and generated food images is adopted. Subsequent to the training phase, the CNN model predicts the object class, bounding

https://doi.org/10.1017/S0263574724000122 Published online by Cambridge University Press



Object detection Pose estimation

Figure 10. *Results of real-time object recognition and pose detection based on FI2PD algorithm. The detected items are labeled with their class names and bounding boxes. The left image shows the performance of object detection and the right one gives the results of pose estimation.*

boxes, and ROO, which are utilized to calculate the quaternion of the selected food items using the depth information from 3D point cloud according to Algorithm 1.

We evaluated the capability of FI2PD technique for simultaneous object detection and 6D posture estimation of several varieties of food. The algorithm was implemented on the standard workstation with the following configuration:

- Memory is 64 GB
- Processor is Intel® CoreTM i9-9900K CPU @ 3.6GHZ
- GPU: Quadro RTX 4000
- OS: Ubuntu 18.04

The results of real-time object detection and pose calculation are depicted in Fig. 10. Four types of food are stacked in separate trays and the food items in the top layer have all been detected and classified according to food type. Additionally, ROO is also predicted alongside the detection through the improved CNN model. The picking target is selected based on the grasping strategy and calculated as the 6D pose for the subsequent gripping operation. In this instance, the pose calculation costs 532ms for the whole process of food recognition and pose estimation while the pose retrieval process requires only 92ms. Moreover, the training process for the entire dataset, which includes all types of food, will take a total of 37 hours and consist of 72,000 iterations.

In our evaluation, we compared the performance of the YOLOv4-based method with the SSD-based approach using the mAP50 criterion. The mAP50 criterion is defined as the mean average precision, calculated by considering the intersection over union threshold to be more than 0.5. We test both methods on the fake data in our *FdIngred328* Dataset. As seen by the data presented in Table I, the methodology based on YOLOv4 demonstrated markedly superior results. This implies that YOLOv4-based method demonstrates improved accuracy in detecting partial positions and the ROO, hence enhancing the effectiveness of pose retrieval.

4.2. Item arrangement verification

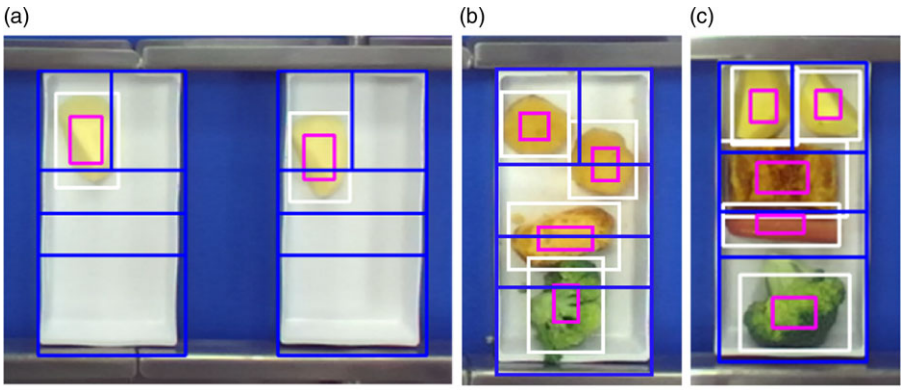
In our approach, we integrated the IAV at the output side. The FI2PD technique is first utilized to detect vacant casserole meal-trays. Following this, boundaries are created for each individual item on the meal-trays. The primary function of the IAV is to validate the positioning of each food item. Fig. 11 demonstrates the outcomes produced by the IAV under various scenarios.

Fig. 11a showcases the tolerance capacity of the IAV. As an illustration, although the original bounding box for the potato on the left meal-tray extends beyond its designated boundary, the adjusted

https://doi.org/10.1017/S0263574724000122 Published online by Cambridge University Press

Table I. *Comparison of mAP50 between YOLOv4-based and SSD-based algorithms on our FdIngred328 dataset.*

Methods \ Epoch	300	500	1500
YOLOv4-based	0.944	0.959	0.962
SSD-based	0.7794	0.7778	0.7837



tolerance capacity "Fail" sample "Success" sample

Figure 11. *Results of IAV under different scenarios.*

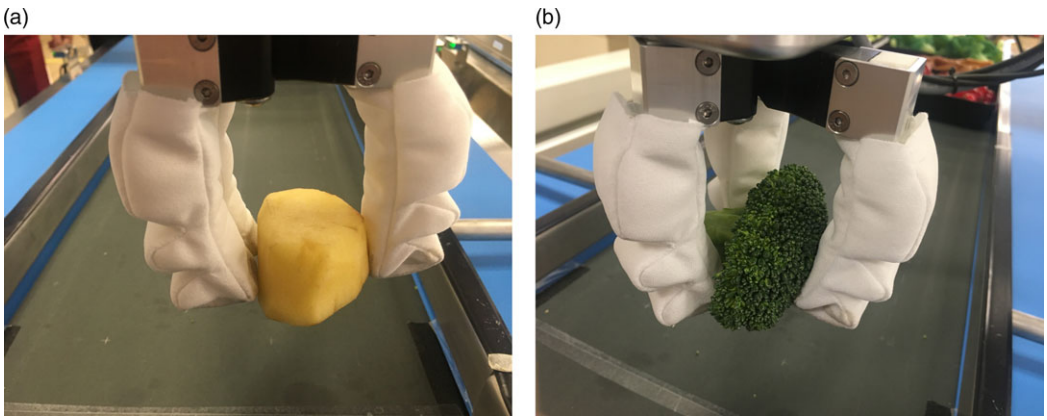


Figure 12. *Soft gripper configured with different modes of grasping food items with different shapes.*

bounding box remains within. Hence, the system considers this scenario to be acceptable. Conversely, in the meal-tray on the right, both bounding boxes intersect with their boundaries, resulting in a "Fail" assessment. The scenario depicted in Fig. 11b illustrates a situation that requires the implementation of corrective measures for the sausage. The robot is programed to reselect the sausage to compensate for the absence of the food item. Notably, even though items such as the potato, broccoli, and omelet are misaligned from their prescribed boundaries, as long as they remain within the casserole, the system will refrain from corrective action. However, a "Fail" flag is generated to request human intervention. Furthermore, fig. 11c presents an example "Success" instance, highlighting the accurate assembly of a meal-tray.

https://doi.org/10.1017/S0263574724000122 Published online by Cambridge University Press

Table II. *Comparison of success rate with and without closed-loop strategy.*

	With closed-loop strategy	Without closed-loop strategy
Success rate	0.90	0.367

Figure 13. *Comparison of generated data through DCGNN and our method. The revolution of output image is 328*328.*

4.3. Application on food handling system

We have effectively integrated our dataset and FI2PD strategy analysis into our custom-built automatic food handling system, as demonstrated in Fig. 6. The system utilizes a Delta robot to transfer food ingredients from different trays into a casserole meal-tray in real time. To achieve this, the system must recognize the food items in the trays and determine their position and orientation for each run. We employ an RGBD camera (ZIVID camera, Zivid Inc., Oslo, Norway) in the input end to acquire both 2D images and 3D point cloud data in real-time. Conversely, the ZED camera on the output end serves to verify both the presence and appropriate arrangement of the food items.

Upon identifying the pose of the target food items, the coordinates are converted from the coordinate system of the objects to that of the robot. Subsequently, the soft gripper then sequentially grips the food items using the corresponding configuration. Fig. 12 illustrates the gripper operating with different modes under different categories detected by vision. The IAV detects any instance of absence and subsequently triggers the appropriate remedial response. To test the performance of the system, we conducted 144 picking tests using potato and broccoli pieces. The system achieved a high success rate of 97.9% in grasping the food items. In order to evaluate the effectiveness of our closed-loop method, we conducted a total of 30 assembly runs for the meal-tray. In each iteration, four food items, specifically two potatoes, an omelet, broccoli, and sausage, were arranged in the casserole meal-tray in this order and arrangement delineated in Fig. 1. With the implementation of the closed-loop strategy, the system is designed to undertake remedial actions if it detects an absence of any food item in the meal-tray. We then cataloged the final status ("Success" or "Fail") of the meal-tray assembly to discern the success rate, both with and without the incorporation of the closed-loop strategy. The outcomes are presented in Table II. The results indicate that the implementation of the closed-loop technique has a significant beneficial effect on the success rate of meal assembly, thus enhancing the overall robustness of the system.

4.4. Data augmentation

We increase the number of datasets through fundamental image manipulation including image rotation, color conversion, and synthetic techniques. In the synthetic process, different food items are placed in

https://doi.org/10.1017/S0263574724000122 Published online by Cambridge University Press

multiple layers in known positions and rotations. The food models are set to have comparable dimensions depending on their respective backgrounds. Therefore, 4749 augmented images and the corresponding annotations are created through basic manipulation of the original data. Simultaneously, 1280 synthetic images were generated by the combination of different objects and backgrounds and the rotation of them. These augmented images are also utilized to balance the size of each class, a crucial consideration in deep learning techniques.

Fig. 13 presents the outcomes of DCGAN data enhancement after 672,400 iterations. The input images are the fake food with the resolution of 328*328. We can find that the DCGAN performs unsatisfactory on this type of dataset conducted with a pile of similar items and it cannot generate the corresponding annotation at the same time. However, through our proposed method, the synthetic data still provides sharp outlines of the food items and diverse backgrounds. Furthermore, the position and orientation of the items are randomly placed but known by us, which are reflected in the annotation via the bounding boxes and ROO information. Thus, the annotations of the generated data are able to be produced at the same time.

5. Conclusion

In this work, we have presented a vision-based robot-assisted framework for meal-assembly system. A novel 6D pose estimation method is also introduced for scenarios where high-resemblance objects are piled together. Our method employs a coarse-to-fine process, where a CNN network generates category, position, and ROO descriptors for different food products. A pose retrieval strategy then calculates the 6D pose of the objects using ten key points generated by ROO, category, and depth information. The addition of ROO is particularly helpful in predicting the rough direction, enhancing the real-time capability of the system. In an effort to enhance the reliability and adaptability of the packing system, we have developed an innovative closed-loop approach that utilizes an item arrangement verifier to improve the process of assembling food items. We have also established a dataset containing both original and synthetic data of different food ingredients. This dataset will enable future researchers to save much effort on pose dataset annotation. Our experimental results have demonstrated that our method can recognize all the items on the top layer of a pile and calculate poses to meet the real-time requirements of food automatic assembly systems. The implementation of our closed-loop technique has significantly enhanced the effectiveness of meal assembly.

However, our current food handling system still exhibits certain constraints, primarily concerning the remedial procedures and the stability of item grasping during high-speed operations. Specifically, the remedial procedures have focused solely on addressing instances of item absence, without considering possible inconsistencies in item placement within the meal-tray. Additionally, ensuring the stability of item grasping becomes challenging, particularly in high-speed grasping scenarios. In our future work, we intend to implement strategies to address these shortcomings. In the future work, for items slightly extending beyond the bounding box, a method of gently pushing will be utilized to return them to their intended placements. In cases where items have completely departed from their designated areas, regrasping will be undertaken to realign the arrangement of the meal-tray. Furthermore, we will conduct exploratory investigations into identifying optimal grasping key points for objects like potato chunks, enhancing the overall stability of high-speed grasping tasks.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0263574724000122.

Acknowledgments. The authors want to acknowledge the members of the Advanced Robotics Centre, for their contribution in the development to the soft gripper, especially Jin Huat Low and Han Qianqian.

Author Contributions. Yadan Zeng conceived and designed the study, and wrote the initial manuscript, Yadan Zeng, Yee Seng Teoh, Guoniu Zhu, and Elvin Toh carried out the experiment and data analysis. All authors contributed critically to the draft manuscript and gave final approval for publication.

https://doi.org/10.1017/S0263574724000122 Published online by Cambridge University Press

Financial Support. This work was supported by National Robotics Programme – Robotics Enabling Capabilities and Technologies Grant Ref. No: ASTAR SERC 1822500053.

Competing interests. The authors declare no conflicts of interest exist.

Ethical Approval. None.

References

[1] Q. M. Marwan, S. C. Chua and L. C. Kwek, "Comprehensive review on reaching and grasping of objects in robotics," *Robotica* **39**(10), 1849–1882 (2021).

[2] Z. Wang, S. Hirai and S. Kawamura, "Challenges and opportunities in robotic food handling: A review," *Front Robo AI* **8**, 789107 (2022).

[3] N. Lu, Y. Cai, T. Lu, X. Cao, W. Guo and S. Wang, "Picking out the impurities: Attention-based push-grasping in dense clutter," *Robotica* **41**(2), 470–485 (2023).

[4] H. Wang, D. Sahoo, C. Liu, E.-p. Lim and S. C. Hoi, "Learning Cross-Modal Embeddings with Adversarial Networks for Cooking Recipes and Food Images," **In:** *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA (IEEE, 2019) pp. 11564–11573.

[5] G.-N. Zhu, Y. Zeng, Y. S. Teoh, E. Toh, C. Y. Wong and I.-M. Chen, "A bin-picking benchmark for systematic evaluation of robotic-assisted food handling for line production," *IEEE/ASME Trans Mech* **28**(3), 1778–1788 (2022).

[6] Y. Hu, P. Fua, W. Wang and M. Salzmann, "Single-Stage 6D Object Pose Estimation," **In:** *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA (IEEE, 2020) pp. 2927–2936.

[7] A. S. Periyasamy, M. Schwarz and S. Behnke, "Robust 6D Object Pose Estimation in Cluttered Scenes Using Semantic Segmentation and Pose Regression Networks," **In:** *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Madrid, Spain (IEEE, 2018) pp. 6660–6666.

[8] A. Collet, M. Martinez and S. S. Srinivasa, "The moped framework: Object recognition and pose estimation for manipulation," *Int J Rob Res* **30**(10), 1284–1306 (2011).

[9] A. Zeng, K.-T. Yu, S. Song, D. Suo, E. Walker, A. Rodriguez and J. Xiao, "Multi-View Self-Supervised Deep Learning for 6D Pose Estimation in the Amazon Picking Challenge," **In:** *2017 IEEE International Conference on Robotics and Automation (ICRA)*, Singapore (IEEE, 2017) pp. 1386–1383.

[10] Y. Xiang, T. Schmidt, V. Narayanan and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes, (2017). arXiv preprint arXiv: 1711.00199.

[11] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song and L. J. Guibas. Normalized Object Coordinate Space for Category-Level 6D Object Pose and Size Estimation. **In:** *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA (IEEE, 2019) pp. 2637–2646.

[12] Y. He, W. Sun, H. Huang, J. Liu, H. Fan and J. Sun, "PVN3D: A Deep Point-Wise 3D Keypoints Voting Network for 6DoF Pose Estimation," **In:** *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA (IEEE, 2020) pp. 11632–11641.

[13] H. Kang, H. Zhou and C. Chen, "Visual perception and modeling for autonomous apple harvesting," *IEEE Access* **8**, 62151–62163 (2020).

[14] G. Lin, Y. Tang, X. Zou, J. Xiong and J. Li, "Guava detection and pose estimation using a low-cost rgb-d sensor in the field," *Sensors* **19**(2), 428 (2019).

[15] W. Yin, H. Wen, Z. Ning, J. Ye, Z. Dong and L. Luo, "Fruit detection and pose estimation for grape cluster–harvesting robot using binocular imagery based on deep neural networks," *Front Robot AI* **8**, 626989 (2021).

[16] J. Rong, P. Wang, T. Wang, L. Hu and T. Yuan, "Fruit pose recognition and directional orderly grasping strategies for tomato harvesting robots," *Comput Electron Agr* **202**, 107430 (2022).

[17] M. Costanzo, M. De Simone, S. Federico, C. Natale and S. Pirozzi, "Enhanced 6d pose estimation for robotic fruit picking, (2023). arXiv preprint arXiv: 2305.15856, 2023.

[18] Z. H. Khan, A. Khalid and J. Iqbal, "Towards realizing robotic potential in future intelligent food manufacturing systems," *Innov Food Sci Emerg* **48**, 11–24 (2018).

[19] JLS Automation, Pick-and-Place Robots Designed For Agility (2002), https://www.jlsautomation.com/talon-packaging-systems.

[20] H. Paul, Z. Qiu, Z. Wang, S. Hirai and S. Kawamura, "A ROS 2 Based Robotic System to Pick-and-Place Granular Food Materials," **In:** *2022 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Jinghong, China (IEEE, 2022) pp. 99–104.

[21] K. Takahashi, W. Ko, A. Ummadisingu and S.-i. Maeda. Uncertainty-Aware Self-Supervised Target-Mass Grasping of Granular Foods. **In:** *2021 IEEE International Conference on Robotics and Automation (ICRA)*, Xi'an, China (IEEE, 2021) pp. 2620–2626.

[22] J. H. Low, P. M. Khin, Q. Q. Han, H. Yao, Y. S. Teoh, Y. Zeng, S. Li, J. Liu, Z. Liu, P. V. y Alvarado, I-M Cheng, B. C. Keong Tee and R. C. Hua Yeow, "Sensorized reconfigurable soft robotic gripper system for automated food handling," *IEEE/ASME Trans Mech* **27**(5), 3232–3243 (2022).

https://doi.org/10.1017/S0263574724000122 Published online by Cambridge University Press

[23] Z. Wang, Y. Makiyama and S. Hirai, "A soft needle gripper capable of grasping and piercing for handling food materials," *J Robot Mech* **33**(4), 935–943 (2021).
[24] Z. Wang, R. Kanegae and S. Hirai, "Circular shell gripper for handling food products," *Soft Robot* **8**(5), 542–554 (2021).
[25] G. Pavlakos, X. Zhou, A. Chan, K. G. Derpanis and K. Daniilidis, "6-DoF Object Pose from Semantic Keypoints," **In:** *2017 IEEE International Conference on Robotics and Automation (ICRA)*, Singapore (IEEE, 2017) pp. 2011–2018.
[26] J. Wu, B. Zhou, R. Russell, V. Kee, S. Wagner, M. Hebert, A. Torralba and D. M. Johnson," Real-Time Object Pose Estimation with Pose Interpreter Networks," **In:** *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Madrid, Spain (IEEE, 2018) pp. 6798–6805.
[27] K. Park, T. Patten and M. Vincze, "Pix2Pose: Pixel-Wise Coordinate Regression of Objects for 6d Pose Estimation." **In:** *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, South Korea (IEEE, 2019) pp. 7668–7677.
[28] Y. Li, G. Wang, X. Ji, Y. Xiang and D. Fox, "DeepIm: Deep Iterative Matching for 6d Pose Estimation," **In:** *Proceedings of the European Conference on Computer Vision (ECCV)*, (Springer, 2018) pp. 683–698.
[29] G. G. Lee, C.-W. Huang, J.-H. Chen, S.-Y. Chen and H.-L. Chen, "AIFood: A Large Scale Food Images Dataset for Ingredient Recognition," **In:** *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*, Kochi, India (IEEE, 2019) pp. 802–805.
[30] S. Horiguchi, S. Amano, M. Ogawa and K. Aizawa, "Personalized classifier for food image recognition," *IEEE Trans Multi* **20**(10), 2836–2848 (2018).
[31] G. Ciocca, P. Napoletano and R. Schettini, "Food recognition: A new dataset, experiments, and results," *IEEE J Biomed Health Inform* **21**(3), 588–598 (2017).
[32] C. Güngör, F. Baltacı, A. Erdem and E. Erdem, "Turkish Cuisine: A Benchmark Dataset with Turkish Meals for Food Recognition," **In:** *2017 25th Signal Processing and Communications Applications Conference (SIU)*, Antalya, Turkey (IEEE, 2017) pp. 1–4.
[33] S. Bargoti and J. Underwood, "Deep fruit detection in orchards, (2016). arXiv preprint arXiv: 1610.03677.
[34] N. Häni, P. Roy and V. Isler, "Minneapple: A benchmark dataset for apple detection and segmentation," *IEEE Robot Auto Lett* **5**(2), 852–858 (2020).
[35] A. Ummadisingu, K. Takahashi and N. Fukaya, "Cluttered Food Grasping with Adaptive Fingers and Synthetic-Data Trained Object Detection," **In:** *2022 International Conference on Robotics and Automation (ICRA)*, Philadelphia, PA, USA (IEEE, 2022) pp. 8290–8297.
[36] A. Radford, L. Metz and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks, (2015), arXiv preprint arXiv: 1511.06434.
[37] A. Bochkovskiy, C.-Y. Wang and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection, (2020), arXiv preprint arXiv: 2004.10934.
[38] T. Su, X. Liang, X. Zeng and S. Liu, "Pythagorean-hodograph curves-based trajectory planning for pick-and-place operation of delta robot with prescribed pick and place heights," *Robotica* **41**(6), 1651–1672 (2023).
[39] S. Bang, F. Baek, S. Park, W. Kim and H. Kim, "Image augmentation to improve construction resource detection using generative adversarial networks, cut-and-paste, and image transformation techniques," *Automat Constr* **115**(3), 103198 (2020).

Cite this article: Y. Zeng, Y. S. Teoh, G. Zhu, E. Toh and I-M. Chen (2024). "Vision-based food handling system for high-resemblance random food items", Robotica **42**, 2108–2124. https://doi.org/10.1017/S0263574724000122

https://doi.org/10.1017/S0263574724000122 Published online by Cambridge University Press