Vision Guided Food Assembly by Robot Teaching from Target Composition

Jain, Siddarth; Corcodel, Radu; Jha, Devesh K.; Romeres, Diego

TR2023-067 June 07, 2023

Abstract

Automation in the food-serving industry has challenging requirements. The article deals with the application of vision-guided food assembly of lunch box containers by robot teaching from a demonstration of a given target composition for the container. Vision guidance is used for parsing the target composition, bin picking with instance segmentation, and computing pose information for similarity-based assembly for food serving with tracking of containers on a moving conveying system. A control system based on learning from demonstration technique is employed in the object frame for similar pose object dropoff of the grasped item for assembly. Integration of vision-guided robot control into high-speed automated food assembly can be highly productive. A pilot experiment scenario for automated food assembly demonstrates the functionality of the vision-guided food-serving system.

ICRA 2023 Workshop on Task-Informed Grasping IV (TIG-IV): From Farm to Fork

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Vision Guided Food Assembly by Robot Teaching from Target Composition

Siddarth Jain*, Radu Corcodel*, Devesh K. Jha, and Diego Romeres†

Abstract— **Automation in the food-serving industry has challenging requirements. The article deals with the application of vision-guided food assembly of lunch box containers by robot teaching from a demonstration of a given target composition for the container. Vision guidance is used for parsing the target composition, bin picking with instance segmentation, and computing pose information for similarity-based assembly for food serving with tracking of containers on a moving conveying system. A control system based on learning from demonstration technique is employed in the object frame for similar pose object dropoff of the grasped item for assembly. Integration of vision-guided robot control into high-speed automated food assembly can be highly productive. A pilot experiment scenario for automated food assembly demonstrates the functionality of the vision-guided food-serving system.**

I. Introduction

Lunch boxes play a vital role in the food culture of Asian markets and have gained immense popularity as a take-out meal [1]. Several million lunch boxes are produced and consumed daily in Japan. For example, Bento, a Japanese lunch box, while portable and convenient, is also intended to be visually appealing. These lunch boxes are typically assembled by human labor. The workers identify individual food items in various bin containers and assemble a predetermined pattern by picking and placing (Figure 1). The food industry demands automation for lunch box packaging to reduce labor costs and cater to growing demands. There are many challenges for automation in this sector. In this article, we focus on robotic systems and present a pipeline (Figure 2) for visually guided automated assembly of lunch boxes given a single target composition of an assembled lunch box. While a typical operation for human laborers, it presents many difficulties for a robotic system [2], requiring planning, dexterous manipulation, and robust perception capabilities. Generalization is a primary challenge, and we present a solution that can work with various food items using feature embeddings without requiring large annotated datasets. Another critical aspect of Bento packaging is the visual appeal, and we match the target composition with learned primitives for dropoff with relative pose estimation.

II. Method

A. Perception of Target Composition

The target composition comprises a manually assembled, visually appealing demonstration of a lunch box containing food items—for example, a typical customer order. There are two main components for performing automated parsing of the target composition: segmentation and feature extraction. A Fully Convolution Network (FCN) can be used for image segmentation tasks. Mask RCNN [3] is an extension of Faster RCNN, which can be used to segment food items. These methods require labeled datasets. For generalization, we use a zero-shot learning method [4]. Figure 3(a) shows some example segmentation masks for a target composition, where the model learned a general notion of segmenting objects. Segmented objects are then encoded with a backbone network to n-dimensional normalized embeddings as feature vectors. Residual networks [5] are popular backbones to extract instance color features. PointNet [6] can generate latent representations using raw point clouds from the segmented masks. These embeddings are utilized to associate objects for bin picking and embeddings can be combined in the feature space to achieve robustness.

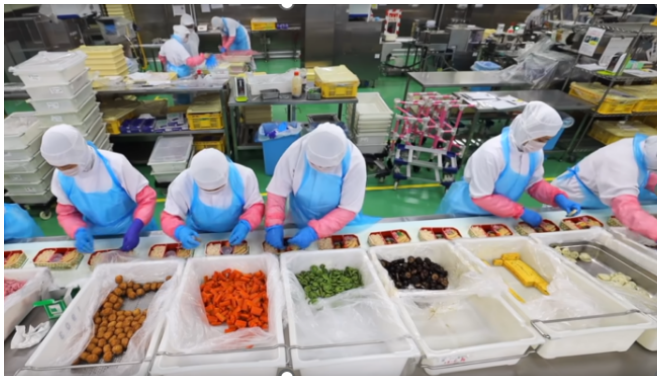

Fig. 1. Visualization of a typical factory scenario of food handling operations to assemble lunch boxes. Source: YouTube -“How a Train Bento Box is Made in Japan”.

B. Bin Picking

Once the target composition is parsed, the next step involves picking food items with a robot from the bin containers. Few shot instance segmentation [7] method can be used to identify a best pickable instance from a bin full of items Figure 3(c). A grasp detection approach [8], [9] can be utilized for grasp selection to pick the desired instance with a robot using motion planning (Figure 3(c)).

C. Assembly

Our automatic assembly framework aims to replicate the target composition using identical items and place them in the same pattern as demonstrated. Maximizing Cosine-similarity in the embeddings space can associate the instance to grasp with items parsed from the target composition, and

*Equal contribution. †All authors are with Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA 02139 {sjain,corcodel,jha,romeres}@merl.com

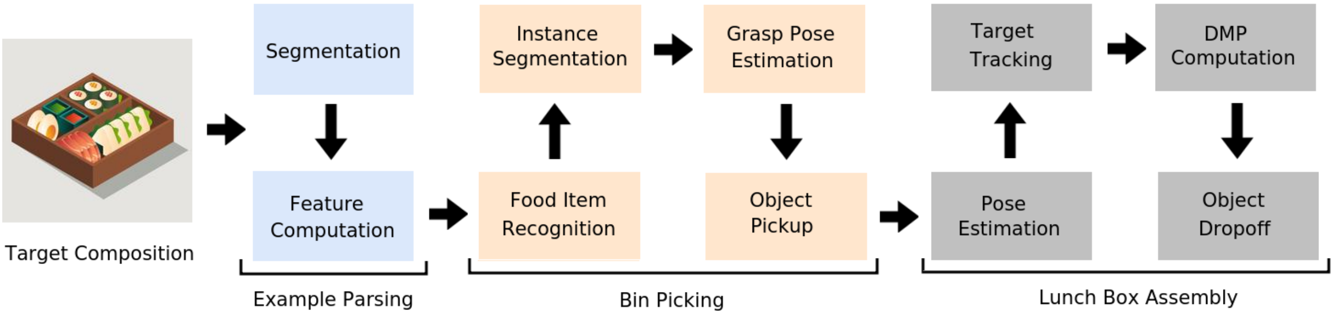

Fig. 2. Visually guided food assembly pipeline.

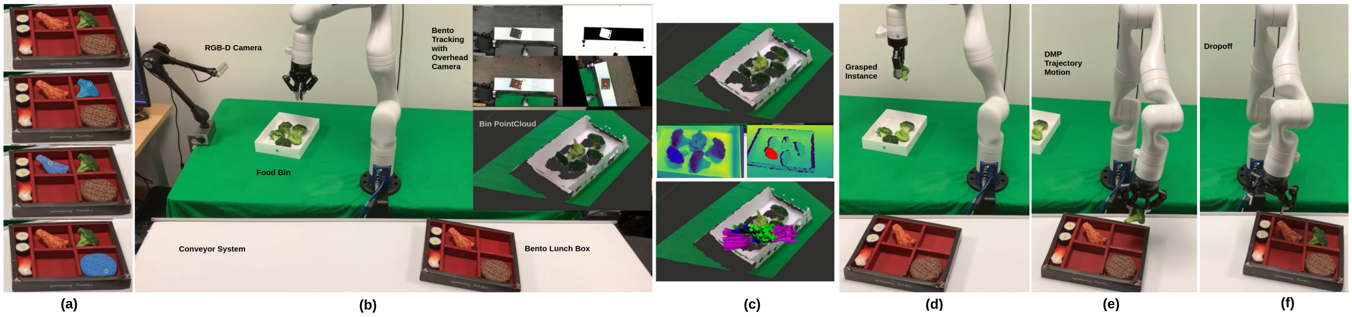

Fig. 3. Pilot. (a) Target composition parsing. (b) Assembly System. (c) Perception. (d) Bin Picking. (e) Robot learned dropoff. (f) Completed assembly.

we calculate a pose difference error using sample consensus initial alignment (SAC-IA) [10]. The registration provides an orientation difference for the object's placement. We compute the destination position as a composition of rigid body transformations between a mobile frame unambiguously attached to the lunch box and the relative pose of each respective item with respect to the box, as observed in the demonstration. The mobile frame is detected by an overhead camera in the color space using binary thresholding and geometry fitting. The grasp pose of each item is computed based on the initial pose to which a suitable motor skill is added (section II-D). We favor grasp poses with vertical picks to minimize bin disturbance Figure 3(d). Once an item is picked, our system computes the destination position of each item by minimizing the signed Hausdorff distance of the corresponding registered 3D points of the bin. Dynamical movement primitives (DMP)[11] are then used to compute a trajectory for placing the food items for lunch box assembly. Our system generates robot trajectories with fixed-time execution and, thus, simplifies the lunch box tracking problem, as the motion of the arm may interrupt the camera's line of sight, and it also allows us to schedule the moving speed of the conveyor belt. Based on the current pose of the box and a controlled conveyor speed, the system schedules a robot trajectory that will rendezvous with the lunch box for object dropoff at the destined pose.

D. Learning Motor skills for food placement

Specific food items must be handled using particular manipulation techniques. Consequently, we use Dynamic Movement Primitives (DMPs) (see [11] for more details) to generate suitable trajectories specific to each food class (*i.e.* scooping, pinching, sliding etc.). For learning the motor skills for the placement of different food items, kinesthetic demonstrations are provided using the robotic arm for individual food items given a target composition. These demonstrations are used to learn DMPs which are parameterized by the goal pose for the placement of the food item. During test time, once the goal pose is obtained from the vision system, a new DMP (corresponding to the style of a particular food item) is computed online that is used to move the robot arm for placement in the corresponding bin (Figure 3(e and f)).

III. Discussion and Conclusions

We deployed our system using a Kinova Gen3 robot arm and bins of assorted food items. Each bin is observed using an RGB-D camera, while an overhead camera tracks lunch boxes moving on a conveyor belt. We limit the scope of food items graspable with a parallel jaw gripper. The most significant source of errors was object slip during picking and, to a lesser degree, during transport. The source of these errors was the grasping controller enforcing gentle clamping forces on the food items, a concept we call *critical grasping*. We theorize that possible instrumentation of the gripper with tactile sensors may help detect and correct slipping with closed-loop manipulation. We report on the challenges and solutions encountered while designing and deploying a self-contained system for the robotic assembly of food platters. We discuss various machine vision approaches and suitable trajectory generators for automated lunch box assembly from a target composition. Finally, we present a pilot experiment scenario for automated food assembly, which demonstrates the functionality of the vision-guided food-serving system.

References

[1] T. Sakata and J. Working Group 2 of the Healthy Diet Research Committee of International Life Sciences Institute, "Current situation and perspectives of ready-to-eat food/meal suppliers," *Nutrition Reviews*, vol. 78, no. Supplement_3, pp. 27–30, 2020.

[2] Z. Wang, S. Hirai, and S. Kawamura, "Challenges and opportunities in robotic food handling: A review," *Frontiers in Robotics and AI*, vol. 8, p. 433, 2022.

[3] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017.

[4] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[6] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[7] A. Cherian, S. Jain, T. K. Marks, and A. Sullivan, "Discriminative 3d shape modeling for few-shot instance segmentation," 2023.

[8] A. Ten Pas, M. Gualtieri, K. Saenko, and R. Platt, "Grasp pose detection in point clouds," *The International Journal of Robotics Research*, vol. 36, no. 13-14, pp. 1455–1473, 2017.

[9] S. Jain and B. Argall, "Grasp detection for assistive robotic manipulation," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 2015–2021.

[10] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (fpfh) for 3d registration," in *2009 IEEE international conference on robotics and automation*. IEEE, 2009, pp. 3212–3217.

[11] S. Shaw, D. K. Jha, A. Raghunathan, R. Corcodel, D. Romeres, G. Konidaris, and D. Nikovski, "Constrained dynamic movement primitives for safe learning of motor skills," *arXiv preprint arXiv:2209.14461*, 2022.