Dense Parts-based Descriptors for Ingredient Detection

James Hahn Michael Spranger
Sony Corporation, Tokyo, Japan
jamhahn@yahoo.com Michael.Spranger@sony.com

1. Introduction

Food plays a vital role in everyday society, providing nutrients for healthy lives and convenience through pre-made meals. Food detection is a growing area of the computer vision community. This is further supported with the recent addition of the Recipe1M dataset, containing 1,029,720 recipes and 166,551 ingredients [5]. This comprehensive dataset lays the framework to develop new capabilities for advanced systems, such as in-home cooking assistants.

Two tasks cooking assistants must learn are food recognition and grasping point recognition. While researchers in these individual areas have achieved success, there is a surprising lack of integration in a cross-disciplinary setting (i.e. a cooking assistant). We hope to combine physical world sensor data (i.e. depth sensors), simulated environment data (i.e. 3D models), and 2D images of food into a system capable of learning dense descriptors for a diverse dataset of ingredients and food.

With dense descriptors, the goal is to learn parts-based models of individual ingredients in physical and simulated environments, then use those parts for two food recognition tasks: ingredient detection and segmentation. If we can learn descriptors for parts of ingredients, then we can hopefully achieve superior results in the fine-grained recognition tasks. This in turn will allow a robot to reconstruct a sandwich from an image. Thus, we plan to make advancements through a multi-discipline approach, with proprioception and images, gathered through the use of robotic manipulation. The hope is depth data will increase a robot's environmental awareness, uncommon in many vision applications, to boost results on benchmarks. We are not aware of any research similar to this, especially with the combination of depth and image data for food tasks.

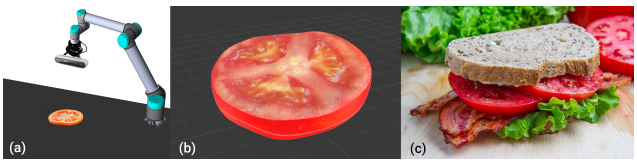

Figure 1. Tomato slice (a) in a physical scan with a depth sensor, (b) as a TurboSquid 3D model, (c) in an assembled recipe

2. Dataset

Due to the cross-disciplinary nature of this work, we will utilize one dataset in several variations, as seen in Figure 1. The first variation is a truncated version of the Recipe1M dataset, containing only RGB images. The second variation is a series of "common" ingredients bought at a local supermarket for physical scans, harvested from the truncated dataset. Finally, we gather a series of 3D ingredient models from an online database to mimic "common" ingredients in the truncated dataset's recipes.

For data collection, we define "common" by extracting the 200 most used ingredients in the Recipe1M dataset and selecting a subset of 50 ingredients that are rigid or semi-rigid (e.g. apples, eggs, peppers, chocolate chips, feta cheese, etc.), thus allowing us to complete 3D scans of the ingredients with depth sensors.

Next, we construct the new dataset. One portion of the dataset contains 2D RGB images of food. In [5]'s im2recipe task, they use RGB images and deep convolutional neural networks to predict ingredients. After extraction of the common ingredients, we find recipes contain anywhere from zero to sixteen of the ingredients. In order to compute a benchmark on the ingredient detection and segmentation tasks, we select four bins of 25 recipes, where each bin contains recipes consisting of 0-3 common ingredients, 4-7 ingredients, 8-11 ingredients, and 12-16 ingredients. Thus, the new dataset contains 100 RGB images of recipes containing anywhere from 0 to 16 of the common ingredients.

Next, in order to complete 3D ingredient scans with depth sensors, we require physical copies of the common ingredients. As such, the 50 ingredients are bought from a local supermarket.

Finally, due to the cross-disciplinary nature of this project, a cooking assistant may not have access to all ingredients in existence. As such, aggregating 3D models of the 50 common ingredients allows the system to learn textured, synthetic copies of the ingredients. Thus, future expansions of this work will be able to utilize any ingredient necessary, even if it is not common in a region or culture, as long as a 3D model exists. Ingredient models are manually retrieved from TurboSquid [1].

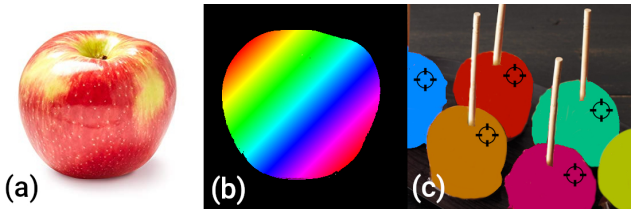

Figure 2. (a) A physical apple, (b) dense descriptor learned after a scan, (c) recipe image fully segmented with robot grasping points identified through use of dense descriptors

3. Methods

The goal is to advance two tasks: ingredient detection and segmentation. We aim to utilize additional depth sensor data, unseen in previous ingredient recognition literature, to generate dense descriptors. Then, we will learn parts-based models of recipes with these dense descriptors. The general pipeline is illustrated in Figure 2.

Recent advancements have led to learning pixel correspondences between two images [4]. These correspondences are achieved by learning dense descriptors for an RGB-D image. They can be utilized to learn grasping points in multi-class robotic manipulation tasks. With the use of RGB-D images collected with a Realsense D415 depth-sensing camera and [4], we transform a $W \times H \times 3$ image into a $W \times H \times D$ discriminative descriptor, where D is a hyperparameter dictating the length of a learned feature vector at each pixel in an image.

Once individual ingredients are recognized, their aggregation can be used to identify whole recipes. Generally, ingredients are almost never seen in their entirety. For example, when examining a burger, one might observe cheese and a tomato slice from a side angle. As such, with recent advancements in parts-based models [3], we hope to translate dense descriptors of ingredients into parts-based models. When observing a burger from a side angle, a dense descriptor will exist for ingredients. The aggregation of these ingredients (i.e. parts) into a model will allow more success for ingredient detection and segmentation.

4. Applications

Adding real-life interaction with food through robotic manipulation, scanning, and additional depth sensor data can open up new doors for improving accuracy on tasks tailored toward cooking assistants of the future.

First, there is limited literature in ingredient detection. For example, [2] utilizes a DCNN for ingredient recognition. However, their dataset is limited to Chinese dishes with labelled bounding box annotations for ingredients. Recipe1M converts an image to its ingredients and recipe, but fails to utilize data from its surrounding environment, such as depth data. With additional sensor data, we can learn dense descriptors to discriminate between several types of ingredients. Once dense ingredient parts are learned, a DCNN can hopefully be trained to boost prediction accuracy of ingredients in food.

The next task is ingredient segmentation. When ingredient detection is accomplished, one must know where ingredients in a recipe are located. Through the use of a dense descriptors in combination with semantic parts-based models [6], we aim to learn dense descriptors for partial views of ingredients in 2D images of recipes.

Finally, in order for a cooking assistant to construct recipes and food from example images, they must learn grasping points of food and ingredients from arbitrary angles. [4] learns these pixel correspondences through extensive data collection with depth and image sensors attached to a robotic arm. This research will allow a robot to grasp rigid and semi-rigid ingredients for construction of recipes (i.e. sandwiches), regardless of orientation.

5. Discussion and Future Work

With our proposed work, we aim to develop better dense food ingredient descriptors learned from various sources of food ingredient image data (2D, depth, 3D models). This will be used to advance benchmarks on ingredient detection (another form of fine-grained object detection), ingredient segmentation, and dense pixel correspondences for food. Additionally, we think the proposed framework will allow a robot to practically grasp ingredients and construct recipes (e.g. sandwiches).

References

[1] Turbosquid. `https://www.turbosquid.com/`. Accessed: 2019-07-19.

[2] J. Chen and C.-w. Ngo. Deep-based ingredient recognition for cooking recipe retrieval. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, pages 32–41, New York, NY, USA, 2016. ACM.

[3] S. K. Divvala, L. Zitnick, A. Kapoor, and S. Baker. Detecting objects using unsupervised parts-based attributes. Technical report.

[4] P. Florence, L. Manuelli, and R. Tedrake. Dense object nets: Learning dense visual object descriptors by and for robotic manipulation. *Conference on Robot Learning*, 2018.

[5] J. Marin, A. Biswas, F. Ofli, N. Hynes, A. Salvador, Y. Aytar, I. Weber, and A. Torralba. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.

[6] D. Modolo and V. Ferrari. Learning semantic part-based models from google images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1502–1509, June 2018.