Food Nutrition Estimation

Shuqi Yang
Supervised by Professor Fadi Kurdahi and Doctor Minjun Seo
In UC Irvine

2018.07.02 – 2018.09.07

Contents

1 Food Category Classification

1.1 Introduction

The first part of this project is to classify the categories of the food. Our whole model is running on Google Vision Kit. However, Vision Kit has only limited resources. Therefore, it can only support very small networks, such as, squeezeNet and mobileNet.

Besides, the dataset we used contains many different categories. Some of them in different types looks very similar, making the classification difficult. Specifically, we use the "food-101" dataset, which contains 101 categories of food. Here is a part of this dataset:

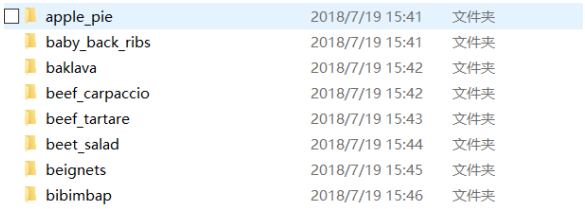

Figure 1: A part of the dataset

To illustrate the difficulty of food classification. Here cites an example from other paper's result [1]:

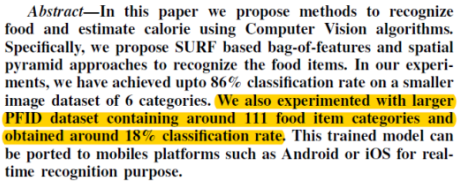
***Abstract*—In this paper we propose methods to recognize food and estimate calorie using Computer Vision algorithms. Specifically, we propose SURF based bag-of-features and spatial pyramid approaches to recognize the food items. In our experiments, we have achieved upto 86% classification rate on a smaller image dataset of 6 categories. We also experimented with larger PFID dataset containing around 111 food item categories and obtained around 18% classification rate. This trained model can be ported to mobiles platforms such as Android or iOS for real-time recognition purpose.**

Figure 2: Others' Results

From this result, we can see that when the number of classes is small, the model can perform relatively well. However, with the increasing of the number of classes, the performance becomes much worse.

1.2 Our Classification Model

According to the official document of Google Vision Kit, we first briefly tried SqueezeNet and MobileNet with the largest input size and width-multiplier that Vision Kit can support and compared their results.

Then, we decided to use MobileNet structure as our CNN model and trained it using the "food-101" dataset. [2] Mobilenet is an efficient and small-size convolutional neural networks specifically designed for mobile vision applications.

Also, due to the constraints of Vision Kit, we use the input size of $192 * 192$, width-multiplier of 1.0, and use no fully-connected layer.

The following picture shows the structure of this network:

[1] Pooja, Hattarki, and P. S. A. Madival. "Food recognition and calorie extraction using Bag-of-SURF and Spatial Pyramid Matching methods." Int. J. Comput. Sci. Mobile Comput 5 (2016): 387-393.

[2] Howard A G, Zhu M, Chen B, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications[J]. 2017.

Type / Stride	Filter Shape
Conv / s2	$3 \times 3 \times 3 \times 32$
Conv dw / s1	$3 \times 3 \times 32$ dw
Conv / s1	$1 \times 1 \times 32 \times 64$
Conv dw / s2	$3 \times 3 \times 64$ dw
Conv / s1	$1 \times 1 \times 64 \times 128$
Conv dw / s1	$3 \times 3 \times 128$ dw
Conv / s1	$1 \times 1 \times 128 \times 128$
Conv dw / s2	$3 \times 3 \times 128$ dw
Conv / s1	$1 \times 1 \times 128 \times 256$
Conv dw / s1	$3 \times 3 \times 256$ dw
Conv / s1	$1 \times 1 \times 256 \times 256$
Conv dw / s2	$3 \times 3 \times 256$ dw
Conv / s1	$1 \times 1 \times 256 \times 512$
$5\times$ Conv dw / s1 Conv / s1	$3 \times 3 \times 512$ dw $1 \times 1 \times 512 \times 512$
Conv dw / s2	$3 \times 3 \times 512$ dw
Conv / s1	$1 \times 1 \times 512 \times 1024$
Conv dw / s2	$3 \times 3 \times 1024$ dw
Conv / s1	$1 \times 1 \times 1024 \times 1024$
Avg Pool / s1	Pool 6×6
Softmax / s1	Classifier

Figure 3: The Structure of MobileNet

After training, this model can achieve an accuracy rate of 59%. Compared with the paper cited above, it is a very good result. However, compared with the results of other very deep neural network, it can not achieve such high accuracy rate.

To run this model on Vision Kit, it should be first converted to a frozen graph and the be complied by the compiler provided by Google Vision Kit. The details of this part can be found in my partner's report.

On Vision Kit, our model can be computed within several seconds.

2 Food Volume Estimation

2.1 Previous Research

Using deep learning to get 3D information [3] [4]

Getting the depth for a single image is a very popular research topic. If the depth of each pixel is known, it could be very easy to get the volume. Besides, Some people have also tried to use a deep learning method to get calories directly from a single image.

However, the network used in these projects are very large, consuming much memory and time. Due to the limited resources on our device, it's not feasible for our case. And the data (food images and the corresponding volume) is hard to get.

Using Multiview Images to Get 3D Information [5] [6] [7]

This kind of methods first takes a set of single 2D images at different positions above the plate. Then, they can use these images to conduct 3D reconstruction of the food to

[3] There are many studies on getting the depth of a single image, which can be searched by "Google Scholar".

[4] Ege, Takumi, and Keiji Yanai. "Simultaneous estimation of food categories and calories with multi-task CNN." Machine Vision Applications (MVA), 2017 Fifteenth IAPR International Conference on. IEEE, 2017.

[5] Puri, Manika, et al. "Recognition and volume estimation of food intake using a mobile device." Applications of Computer Vision (WACV), 2009 Workshop on. IEEE, 2009.

[6] Xu, Chang, et al. "Image-based food volume estimation." Proceedings of the 5th international workshop on Multimedia for cooking & eating activities. ACM, 2013.

[7] Dehais, Joachim, et al. "Two-view 3d reconstruction for food volume estimation." IEEE transactions on multimedia 19.5 (2017): 1090-1099.

calculate its volume.

However, taking several images from different positions costs lots of time of the users, which is not feasible in our project.

In the future, if we plan to attach more than one camera on the eyeglasses, this method can also be a good idea.

Shape-based [8] [9]

Another method is to use the shape template to calculate the volume. They first obtain the specific shape template for each food item based on its food label (such as, a sphere shape for an apple or an orange and a cylinder shape for a liquid.). Then they extract feature/corner points in order to size the food shape template (such as the corner points of a cube).

This kind of method is most similar to our method.

2.2 Summary of Our Method

- In the project, we use the plate size as reference.
- We first use the rgb color to separate the background (the desk), the plate and the food region.
- We divide the shape of the food into six types. Then design a formula for each of them.
- After we know the type of the food, the shape of it is know. Therefore, the next problem is how to get the volume, using the shape information.

2.3 Different Shapes of Food

Cube Shape This is kind of a very general shape. Not only the cubes, this shape contains all the shapes whose volume can be obtained by multiplying the top/bottom area by the height. The food, such as cake, bread are in this shape. The following picture shows some examples:

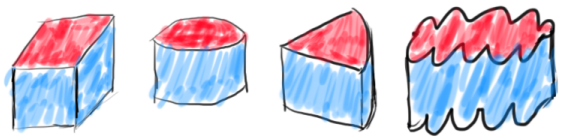

Figure 4: Examples for cube shape

In some previous research, they choose to extract the feature points from the image to calculate the volume of the cube. However, in our more general shape, some shape is not

[8] Chae, Junghoon, et al. "Volume estimation using food specific shape templates in mobile image-based dietary assessment." Computational Imaging IX. Vol. 7873. International Society for Optics and Photonics, 2011.

[9] He, Ye, et al. "Food image analysis: Segmentation, identification and weight estimation." Multimedia and Expo (ICME), 2013 IEEE International Conference on. IEEE, 2013.

so regular that it's hard to detect their feature points, such as the right most two shapes in the picture above.

In our project, we choose to use some image segmentation methods to separate its top part (the red parts in the above picture) and its side part (the blue parts). Then we can get the top area from the top part and height from the side part.

The original color and the lightness are different for the two parts. Therefore, it can be possible to separate them apart. In our project, we use K-means cluster for this segmentation task. The input feature for the algorithm is the r g b colors and horizontal and vertical positions of each pixel in the food region.

For more accuracy in the future research, we can also use some more advanced image segmentation methods, such as using CNNs to finish this tasks.

Ball and Half-ball shapes The apple can be an example of ball and the the bun can be an example of half-ball. For ball and halfball, it is very easy to get the diameter from the food region: The longest segment in the horizontal direction, as shown in the following picture:

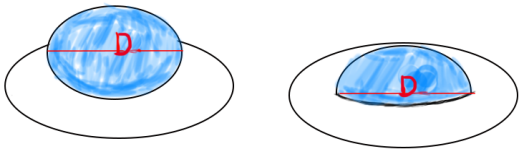

Figure 5: Examples for ball and half-ball shape

Cone Shape The fried rice can be an example of cone. For cone, we can get the longest segment in the horizontal and vertical direction, which can be used to get the height and diameter of the cone, as shown in the following picture:

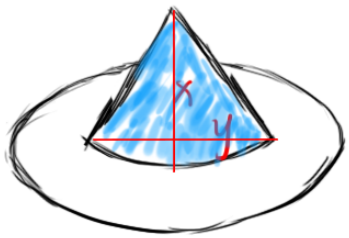

Figure 6: Examples for cone

Irregular Shape The shape of some food is irregular, such as the banana. In some previous research, people choose to multiply the average height by the area to calculate the volume. However, in our project, we find out that for much food with irregular shape, the shapes for all the food in the same category is the same. Such as the bananas, all the bananas are in the same but irregular shape. In fact, ball and half-ball can also be treated as two examples in this shape.

If a type of food has the same shape, then $volume \propto area^{3/2}$.

Therefore, if we can get the volume per unit area of this type of food, we can obtain its volume easily.

Figure 7: Examples for irregular shape (In two bananas, the length of x and y is proportional.)

Fixed Height Shape Some types of food is of fixed height, such as the pizza. These types are very simple as long as we have the pre-defined height.

If some food is not included in the previous five shapes and is of no fixed height, we can also use their average height to estimate their volume. (But the results would be less accurate.)

2.4 Detecting the Regions of Background, Plate and Food

Our previous shape-based method works when the region of food and the plate reference are given. Therefore, we should get the region of the food and plate before sending these data to the volume estimation algorithm.

Given the pre-defined color of the plate, we can first select the pixels with the similar color in the food image and then do some morphological processing to the result to make it more reasonable.

As for the background, we can first assume the pixels in the four corners of the images are background. Then, starting with these pixels, we can find their neighbor pixels. If the neighbor pixel is in the similar color of a background pixel, we can label it as background. This process is done recursively.

Other Possible Methods to Detect the Regions of Plate and Food

- One way is to detect the ellipse of the image. But if some parts of the plate is hidden (making the ellipse incomplete), this method can not work so well. An example result:

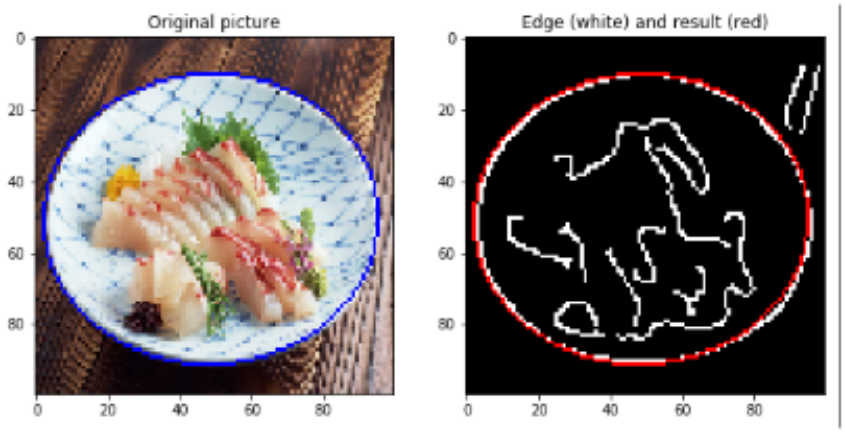

Figure 8: Ellipse Detection Example

- We can also use some saliency detection algorithm(the algorithm to detect the region of interest of an image) to segment the food parts. This can work better than the "rgb segmentation" if the background is more completed. But in some cases, it's not very stable. For example:

Figure 9: Example #1

Figure 10: Example #2

- Some people also use deep learning methods to detect the region of the food. This kind of method can usually reach good results. However, our limited resource is not feasible for deep CNNs.

Calibrate the Length In the Image We assume that the eyeglasses are only angled vertically and parallel to the ground in the horizontal direction. (That is, when people wear their eyeglasses, they look down at the food and create a vertical angle, but the eyeglasses (and both eyes) are parallel to the ground. Using the ratio of the long axis to the short axis of the plate in the image, we can estimate this angle. Then, we can get the actual length of a pixel in the image, in the x, y, z direction.

Getting the Food Nutrition Information We use an online database to get the nutrition information of each category of the food.

3 Server Program

3.1 GUI on server for showing results

Then, we made a GUI on server machine to show the results more clearly. We use socket to send data from Vision Kit to the laptop. And we use matlab to create the GUI to show the data sent to the server. Using this program, we can send original data and the results processed by Vision Kit to a remote server so that these data and results can be shown and stored in the server. This interface of this GUI program can be seen in the following "results" section.

3.2 More Possible Functions of Server

The server machine can be a much more powerful machine and it can receive data from different Vision Kit clients.

Therefore, after collecting the information sent by clients, it is also possible for us to conduct some more complex algorithms on the server to get more accurate results.

When creating the server program, we found that if we use UCInet Mobile Access, we could not connect our server and client. Then we used mobile phone hotspot to connect, which worked. If someone else wants to use the server in the future, he or she may need to avoid using UCInet Mobile Access.

4 Results

Some results for volume estimation:

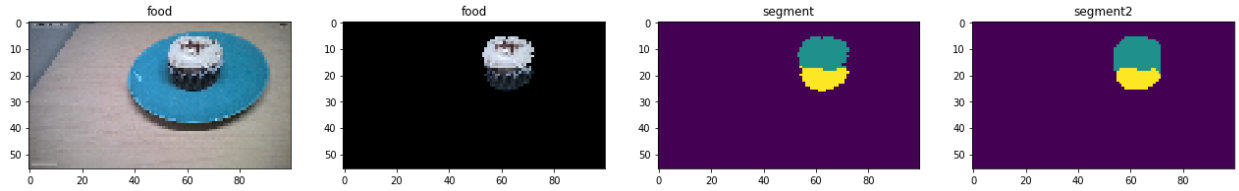

Figure 11: Example #1

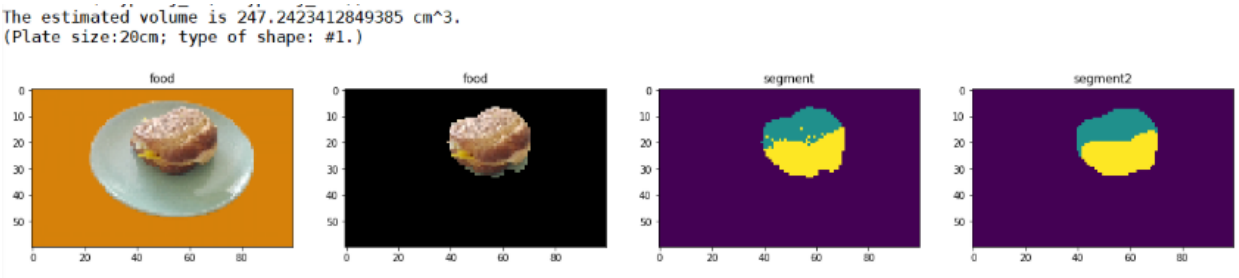

Figure 12: Example #2

The whole results:

(a) Original Image (b) Results processed by Vision Kit

Figure 13: Results

5 Some Possible Ideas for Future Work

Collect more needed information about the food Some information about food needs to be collected, as mentioned in the previous parts: the shape of each type of food; and some related information, such as the volume per unit area of foods with fixed shapes and the height of food with a fixed height.

Get some food image and volume information In order to better evaluate the performance of our algorithm and to provide the training data for training some machine learning models to estimate the volume end-to-end (if we are going to try this way in the future), some food images with volume information need to be collected. This can be done by using the volume or calorie information in the food receipt, or using some real food to measure their volume and then take some pictures of them.

Use more advanced methods for detecting the food region Detecting the region of the food is also an important part of this project. In our program, we mainly divide the food area based on the rgb color, which is a very simple method. However, in some cases, such as when the lighting conditions change a lot, this method is not very stable. Some other methods for identifying food areas have been mentioned in the previous section, such as some algorithms that detect the region of interest. If we can add the predefined information such as the color of the plates to these algorithms as a prior, we may then improve the performance of them in our food region detecting problem. Besides, using CNNs for food area identification is also a feasible method.

Use more advanced methods for segmentation For the segmentation for the cube shape food, we can also use some advanced segmentation algorithms, or use CNNs, or extract more features to improve the results.

Try to fit a large network into Raspberry Pi device For the classification problem, the performance of the deep CNNs are much better than the small ones. However, the resources of Vision Kit is limited. Therefore, how to fit the large model into this small device can be a very helpful and interesting problem.

Attach stereo camera to the eyeglasses It is also a helpful idea for this project if we can attach the stereo camera to the eyeglasses, which can provide the 3D information of the image.

Deal with the camera position problem When using this program, the position of the eyeglasses has a great influence on the quality of the picture. If the user who wears the eyeglasses sees the food in a improper angle, height or distance, it may cause some problems. For example, the food may be in the corner of the image, only a small part of the image is the food part, or the image is out of the image. For the problem that the food is not in the correct position in the picture, if we can guarantee that the algorithm for identifying the food region works well, then we can first crop the image to get a image with the food in the middle of the image for subsequent classification and volume estimation process. If the food is beyond the scope of the image, if we can identify if the image is unqualified after receiving the image, we can use the vision kit to ring to tell the user to press the button again to take another picture.