2019 IEEE 5th International Conference on Computer and Communications

Food Calorie and Nutrition Analysis System based on Mask R-CNN

Meng-Lin Chiang, Chia-An Wu, Jian-Kai Feng, Chiung-Yao Fang, Sei-Wang Chen
Department of Computer Science Information Engineering
National Taiwan Normal University
Taipei, Taiwan
e-mail: meng.chiang@gmail.com

***Abstract*—Over the past few decades, obesity has become a serious problem. Obesity is associated with many of the leading causes of death, such as chronic diseases including diabetes, heart disease, stroke, and cancer. The most effective way to prevent obesity is through food intake control, which involves understanding food ingestion, including the nutrients and calories of each meal. To assist with this issue, this study develops a food calorie and nutrition system that can analyze the composition of a food based on a provided image. Further, we introduce a newly collected dataset, Ville Cafe, for food recognition. This dataset contains 16 categories with 35,842 images, including salad, fruit, toast, egg, sausage, chicken cutlet, bacon, French toast, omelet, hash browns, pancake, ham, patty, sandwich, French fries, and hamburger. The system is based on a Mask Region-based Convolutional Neural Network (R-CNN) with a union postprocessing, which modifies the extracted bounding boxes and masks, without the non-maximum suppression (NMS), to provide a better result in both analytics and visualization. The recognition accuracy for the combination of Ville Cafe and the Food-256 Datasets was 99.86%, and the intersection over union (IoU) was 97.17%. The food weight estimation experiment included eight classes: salad, fruit, toast, sausage, bacon, ham, patty, and French fries. These classes respectively had 40, 40, 44, 40, 41, 49, 26, and 40 data points, adding up to 320 data points for the linear regression model. In the experimental results, the average absolute error was 8.22, and the average relative error was 0.13.**

Keywords-food image recognition; food nutrition analysis;food calorie analysis; Mask R-CNN; instance segmentation

I. INTRODUCTION

Lifestyle has a significant impact on physical health, with eating habits playing a major role. Harris et al. [2] grouped lifestyles into five categories, two of which relate to diet (health habits, avoiding harmful substances), indicating that eating habits have an impact on health. In clinical studies, Livingstone et al. [12] suggested that a typical approach to understanding eating habits is to record the type and amount of food in meals and to analyze calorie and nutrient intake. However, Lichtman et al. [15] pointed out that approximately 33% of subjects underestimate the amount of food intake, according to self-reports of their obese subjects' diets. The John Tung Foundation [8] investigated the understanding of nutritional labelling of available food in the Taipei area. Up to 73.1% of respondents did not understand the nutrition label, and 68.4% of those who thought they understood it actually did not. Therefore, an effective way to keep healthy is to monitor your own calorie and nutrition intake. This study proposes a system for analyzing and estimating calories and nutrients using food images, allowing users to conveniently and quickly understand the calorie and nutrient intake of each meal, with the aim of controlling diet and balancing nutrition.

The Taiwan Health Promotion Administration surveyed nutritional health changes in 2013–2016 [5]. The results showed that Taiwanese people aged 19–64 ate too little protein and starch, and did not reach a balanced diet. Regarding breakfast choice [6], Taiwanese people have a 25% fat intake for breakfast, which is 21% higher than other Asia-Pacific countries, making Taiwan one of the "most greasy" countries in terms of breakfast. According to KEYPO's Big Data Engine survey [3], Taiwan's most popular top 10 breakfasts are (1) fruit salad with yogurt, (2) toast, (3) sandwich, (4) quiche, (5) burger, (6) muffin, (7) taro, (8) rice ball, (9) fried noodles, and (10) radish cake. The results show that Taiwanese people prefer Western-style breakfasts. Therefore, this study limits the scope of food calorie and nutrient analysis to Western breakfasts.

Many studies have proposed food calorie and nutrient measurement systems. Villalobos et al. [4] presented an approach for calorie intake measurement to help understand dietary intake with respect to weight-loss and chronic diseases. Anthimopoulos et al. [10] proposed a food carbohydrate counting system. The system captures food images via a mobile phone and uses a bag-of-features model to obtain the color and texture features of various foods in the image.

Current food detection and recognition techniques can be divided into two types: the first type captures the dish in the image and then analyzes and recognizes the food class; the second type identifies the food directly in the image. Although the first approach can improve the accuracy rate of food recognition, the colors, sizes, and shapes of the dishes vary greatly, making it difficult to capture and detect the features of the dish object. Most researchers use the second approach.

According to the first approach, Dehais et al. [7] proposed a dish detection and segmentation system for dietary assessment. This system first detected the dish in the

978-1-7281-4743-7/19/$31.00 ©2019 IEEE

image and then segmented the food block in the dish. The system assumed that dish is circular, and the edge detection of the dish was mainly a multi-layered RANdom SAmple Consensus (RANSAC), which approximates the curve closest to the edge of the dish.

In line with the second approach, Anthimopoulos et al. [13] developed a food recognition system for diabetic patients that captured color, size, shape, and texture features based on histograms using a Support Vector Machine (SVM). The recognition accuracy was between 58% and 95%. The shortcoming of this study was that bag-of-features does not consider the positional relationship between features.

Pouladzadeh et al. [14] proposed a system of measuring calorie and nutrition from a food image, and it can be used on mobile phones to help dieters and patients understand each meal and its associated nutrient intake. After the user takes two differently angled images of the food on the mobile phone, the system can identify the type of food and calculate the area and volume of the food. The study identified 15 foods with an average accuracy of 86%. The limitation of this study was that the food dishes in the image could not overlap, and the food had to be placed separately in the dish.

Bolaños and Radeva [11] presented an approach of Simultaneous Food Localization and Recognition for the consumer's perspective. First, a deep learning neural network, GoogleNet-GAP, was used as a classifier to distinguish between food and non-food blocks in the image. The experimental results showed that the method had a recognition rate of 90.90% on the EgocentricFood dataset and 79.20% on the Food101 dataset. Thus, there is still room for improvement regarding food recognition.

Mask R-CNN [9] is an instance segmentation approach that improves on faster R-CNN. Its structure has four parts: the convolutional backbone, the RPN, the RoIAlign layer, and the head. Since Mask R-CNN uses RoIAlign to unify the RoI size, it improves the positional deviation of the frame selection object compared with faster R-CNN. Therefore, when an RoI enters the mask branch prediction mask, the accuracy rate is increased by at least 10%.

Regarding the above, in the image processing method, specific features of the food image are used as the basis for classification, and there is room for improvement regarding accuracy. In the deep learning based approach, color and texture are low-level specific features. A deep neural network can learn more abstract features, which can help detect and recognize food. Therefore, this study recognizes food images by using Mask R-CNN, estimates the food weight by a linear regression calculation, and uses the nutrient table to estimate the food calories and nutrients.

II. SYSTEM FLOWCHART

Fig. 1 shows the flow chart of the calorie and nutrition analysis system. It has four main steps: image resizing, food detection and classification, food weight prediction, and food calorie and nutrition analysis [1]. Once the food image is input into the system, the system scales the image to appropriate size. The resized image is then fed into Mask R-CNN to capture the food features and perform food detection and classification. At this step, Mask R-CNN detects and recognizes the food class and the box regression of the food based on the captured features. The system then estimates the weight of the object through the image of the recognized food. After obtaining the weight of the food, the food calorie and nutrition analysis system is estimated according to the Ministry of Health and Welfare [16] and the US Department of Agriculture's Food Nutrition Database [17].

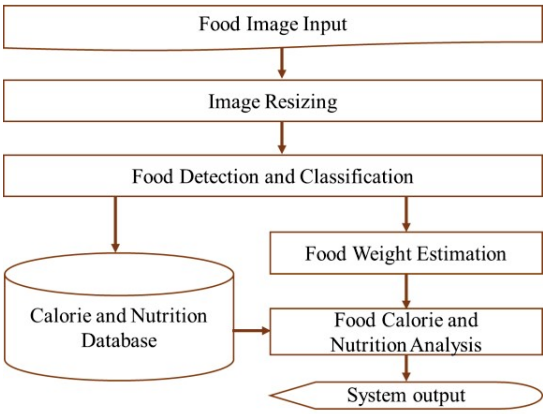


Figure 1. Flowchart of calorie and nutrition analysis system.

A. Image Resizing

Mask R-CNN requires an input image of 1024×1024 pixels, so the input image I_o is resized to a specific length and width to form I_r. To avoid reducing the proportion of food in the image when adjusting the image length and width, if the original input image scale is not 1:1, we adjust the long edge of the original image to 1024, and the short edge size is calculated according to the proportion of the original input image. For example, if the original input image size is 2788×2204 pixels, the long edge (2788) is adjusted to 1024 pixels, and the short edge (2204) is adjusted to 809 pixels according to the scale ratio.

B. Food Detection and Classification

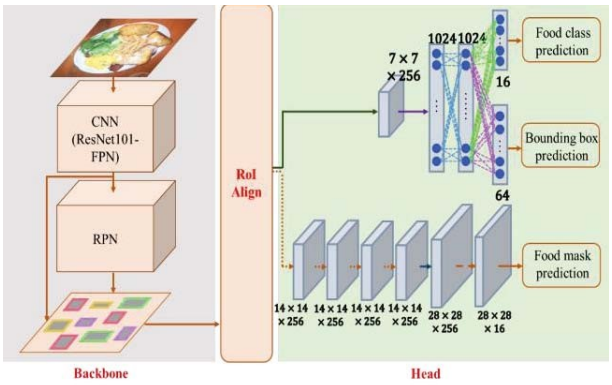


Figure 2. The architecture of mask R-CNN.

The food image I_r is input into Mask R-CNN to obtain the following prediction results: the food class, the food bounding box, and the food mask. Fig. 2 shows the architecture of Mask R-CNN. This architecture can be divided three main parts: Backbone, RoIAlign, and Head. The gray block is the Backbone part, the orange block is the RoIAlign part, and the green block is the Head part. The Backbone includes ResNet101-FPN and RPN components. Its main purpose is to extract RoIs. RoIAlign is used to adjust the size of the Backbone-output RoIs. The Head outputs three kinds of prediction results: class, box, and mask.

1) Mask R-CNN Backbone

The Backbone of Mask R-CNN contains ResNet101-FPN and RPN. ResNet101-FPN is a feature extraction method combining ResNet101 and FPN, including a bottom-up pathway (yellow block in Fig. 3), lateral connections (orange dotted line in Fig. 3), a top-down pathway (purple block in Fig. 3), and an aliasing effect reduction (blue block in Fig. 3).

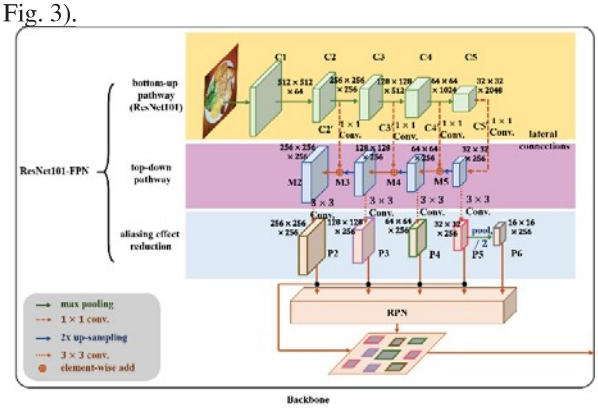


Figure 3. The architecture of the mask R-CNN backbone.

The bottom-up pathway extracts food features from low to high levels in the neural network layers (ResNet-101). The lateral connections output the feature maps of the bottom-up pathway to the number of channels. The top-down pathway involves the process of transferring high-order food features from the neural network to the lower order. The aliasing effect reduction adjusts the feature maps of the output of the top-down pathway by the convolution operation. This step is used to avoid the aliasing effect caused by the up-sampling operation in the top-down pathway. RPN uses the feature maps of the ResNet101-FPN output to determine the foreground block in the image.

2) Mask R-CNN RoI Align

RoIAlign is a technique for adjusting the size of RoIs, and its input is RoIs of different sizes. When the food image I_r is input to the Mask R-CNN Backbone, the feature map of the food block RoI in the image I_r can be obtained.

In this step, RoIs can be obtained by dividing the image into 14×14 grids, and then taking four sampling points (bin) for each grid at a fixed interval width and height. The values of the four sampling points are calculated by bilinear interpolation according to the four elements' values adjacent to the coordinates converted to the feature map. After the values of the four sampling points are obtained, maxpooling is performed on the four sampling points in the cell, and the outputs are RoIs of size 14×14.

3) Mask R-CNN Head

The Mask R-CNN Head is a neural network architecture used to predict food class, bounding box, and food masks, as shown in Fig. 4. There are two major branches in the Mask R-CNN Head: the food class and food bounding box prediction branch and the food mask branch. The food mask branch is used to predict the pixels of the food class in the input RoIs. In the first branch, the system performs a convolution operation using 1024 kernels for RoIs. The kernel size is 7×7, and the output is FC_1, which has a size of $1 \times 1 \times 1024$. FC_1 is the input of FC_2. Using FC_1 to perform a convolution operation using 1024 kernels, the kernel size is 1×1, its output size is $1 \times 1 \times 1024$, and it is the input of FC_2. To predict the food class, the output of FC_2 is passed through a fully connected layer (FC_3), and the food class of the RoIs ($class$) and the probability of belonging to the class ($class_{prob}$) can be obtained. Then, the output of FC_2 is passed through a different fully connected layer (FC_4), and the food bounding prediction is obtained.

The second food mask branch output is $RoIs_{mask}$, as shown in Fig. 4 (red-outlined block). Since a corresponding mask is predicted for each food class in this branch, the output of this branch is $28 \times 28 \times K$, where K is the number of food classes. In this study, $K = 16$. This branch is constructed by a Fully Convolutional Network (FCN). The FCN network architecture consists of a convolution operation and a deconvolution operation, which can be divided into two parts: food feature extraction (yellow block in Fig. 4) and food feature upsampling (green block in Fig. 4). The food feature extraction refers to the input of RoIs to obtain food feature maps after several convolution operations. The food feature upsampling is performed by the deconvolution calculation. The purpose of the upsampling is to capture the pixels in the input image that belong to the class of the food.

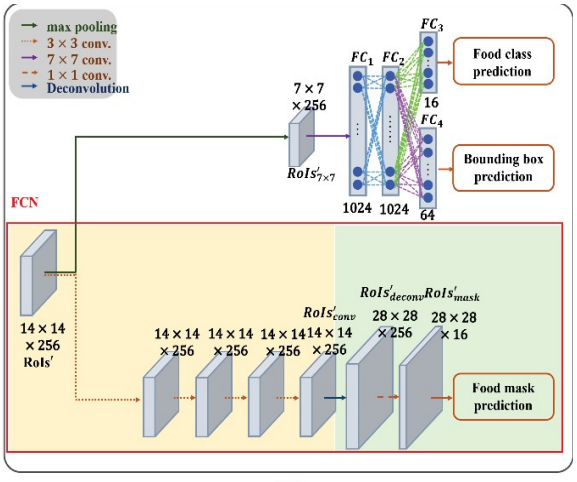
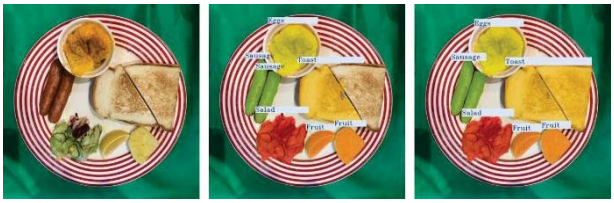


Figure 4. The architecture of the mask R-CNN head.

4) Union Postprocessing

When the food class, bounding box, and mask prediction are completed, the results are then integrated. In this step, the inputs are I_r, $class$, $class_{prob}$, box, and $RoIs_{mask}$. The output result, $class_{refine}$, box_{refine}, and $mask_{refine}$ are marked in the image I_r.

In Mask R-CNN, the output stage can be divided into three steps: filter out low confidence, non-maximum suppression, and apply mask. Filtering the low confidence level involves pointing to the $class_{prob}$. If the value is too low, the prediction box and the food mask are deleted. Non-maximum suppression points to a food bounding box that overlaps in the same class, keeping the result with the highest degree of confidence according to $class_{prob}$. The food mask is based on the $class$, and the food prediction box is selected from $RoIs_{mask}$.



(a) (b) (c)

Figure 5. Example food recognition images. (a) The input image. (b) The result of Mask R-CNN with NMS. (c) The result of Mark R-CNN with union postprocessing.

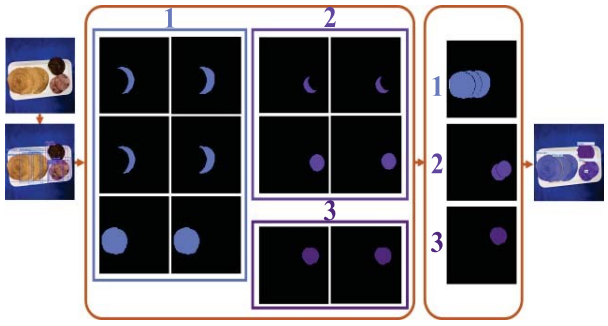


Figure 6. An example of union postprocessing.

After experimental observation, if the food is placed too close on the dish, once the image is input into the Mask R-CNN with non-maximum suppression, only the box with highest $class_{prob}$ will be retained, and the others boxes will be filtered out. Since the overlapping area of the prediction box is too large, the recall rate of Mark R-CNN is lower in this study. Fig. 5 shows an example of food recognition images. Here, the two pieces of toast were not individually detected, shown in Fig. 5(b).

Therefore, this system modifies the non-maximum suppression step. First, the boxes of the same class of food and the ratio of the overlap are obtained. If the ratio of overlap is higher than the threshold, the box is retained with its corresponding $class_{prob}$, box_{shift}, and $RoIs_{mask}$. In the apply mask step, $mask_{refine}$ is the result of the union of $RoIs_{mask}$ of the same class. Fig. 6 is an example result of the union postprocessing, where 1, 2, and 3 are pancake, ham, and hamburger food masks, respectively. The union of all food masks of the same class (six pancake images, four ham images, and two hamburger images) are marked on the original image. Fig. 5(c) shows the food recognition result of Mark R-CNN with union postprocessing.

C. Food Weight Estimation

This system uses Mask R-CNN to obtain food recognition and mask prediction, and then to capture the number of pixels in the image through the food mask. In this study, a fixed angle photographing dish was used, and the same food was placed on the dish. Photographs of different portions of the food were taken, and the weight and the amount of food in the image were recorded.

To prevent the estimation error being too large, the number of pixels was uniformly divided by 10,000, and the weight of the fruit was uniformly divided by 100. Here, the number of pixels in the image is x, the food weight is y, the linear regression to calculate all known actual values is $f(x)$, and the estimated error values are $(\overline{f(x)})$. Let $(\overline{f(x)})$ be the food weight estimate obtained from the linear regression equation, y be the actual food weight, n be the number of data points, and i be the index of n data, where $1 \le i \le n$. The least squares method (LSM) for calculating the error is

$$\mathrm{LSM} = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \overline{f(x_i)})^2. \quad (1)$$

LSM is the error between the estimated value and the predicted value, which is obtained by calculating the regression equation for each data point. The error value is squared, and the average value is calculated. To avoid the problem that positive and negative offsets occur for various error values, the error value is squared. In addition, in (1), the error value is divided by $2n$.

Let the linear equation be $(\overline{f(x)})$, where the slope is a and the intercept is b, that is, $\overline{f(x)} = ax + b$. After substituting $(\overline{f(x)})$ into (1), and obtaining the minimum value of (1), the slope a and the intercept b are taken. The result is shown in (2).

$$arg\min_{a,b}\left(\frac{1}{2n}\sum_{i=1}^{n}(y_i - ax_i - b)^2\right). \quad (2)$$

To find the values of a and b in (2), subdivide (2) with a and b for partial differential operations:

$$\frac{\partial \mathrm{LSM}}{\partial b} = \frac{1}{n}\sum_{i=1}^{n}(y_i - ax_i - b)(-1). \quad (3)$$

$$\frac{\partial \mathrm{LSM}}{\partial a} = \frac{1}{n}\sum_{i=1}^{n}(y_i - ax_i - b)(-x_i). \quad (4)$$

Then, let (4) equal zero to obtain b, and substitute the value of b into (3) to obtain the value of a. Substituting the values of a and b into $\overline{(f(x)}$ gives a linear equation for weight estimation.

According to $mask_{refine}$ of the Mask R-CNN output, the number of pixels in each image of the food, $pixel_{food}$, can be obtained occupied for each class of food in the image. Substituting $pixel_{food}$ into $\overline{(f(x)}$, the food weight estimate value is $weight_{food}$. Using the food nutrition and calorie tables, the calories and nutrients of the food in the image can be estimated.

D. Food Calorie and Nutrition Analysis

This study uses the Ministry of Health and Welfare Nutrition Database [16]. Since some typical Western breakfast foods are not listed in this database, this study uses the US Department of Agriculture's Food Nutrition Database [17] as a reference for nutrients. The food calorie and nutrition analysis step recognizes the food class and food weight, and estimates the calories and nutrients of the food in the image. The nutritional components include crude protein, crude fat, saturated fat, trans fat, carbohydrate, dietary fiber, sugar, and sodium. As the calories and nutrients are listed in the databases for every 100 grams of food, the totals for the food in the image are calculated based on the estimated weight.

III. Experimental Results

The experimental environment of the system is set to shoot a depression angle when the table is photographed. The phone is held directly above the dish, and the phone is parallel to the dish when the image is taken. A variety of foods are placed on the dish, some of which will cover each other, or the food will be placed beyond the dish. The resolutions-are 4000×3000, 4000×3,000, and 4608×3456. In addition, when estimating the weight of the food, this study takes a variety of different servings of the same food and records the weight of each serving. Linear regression is used to analyze the relationship between the image ratio in food and the food weight based on the proportion of food in the image and the actual weight of the previous record.

A. Ville Cafe Dataset

This study created 16 classes in the Ville Cafe dataset: salad, fruit, toast, egg, sausage, chicken cutlet, bacon, French toast, omelet, hash browns, pancake, ham, patty, sandwich, French fries, and hamburger. The Ville Cafe dataset contains five different food items from five restaurants, giving a total of 35,842 images and 9,776 food items. The length and width ratio of food images is 1:1 and non-1:1, and its length and width are 3024×302 and 962×1094 to 4385×2988 pixels, respectively. To increase the amount of data when collecting food images, in addition to the original background (table), the food images are taken using different sets of discs.

B. Breakfast Food Detection and Recognition

This section analyzes and explains the correct detection rate of the 16 classes of breakfast food in the Ville Cafe dataset. Table 1 shows the 16 classes of food, the number of images, and the numbers of food in each class. There are 35,842 images and 9,776 foods. In the 16 classes experiment, the Food-256 Dataset was used in combination with Ville Cafe for training and testing. This experiment used 850 training images and 428 validation images. The number of food items tested and validated were 4118 and 1974, respectively.

TABLE I. Numbers of food Image of 16 Classes for Training and Validation Data.

No.	Class	# of Training Images	# of Training Food Items	# of Validation Images	# of Validation Food Items
1	Salad	267	286	107	107
2	Fruit	236	516	104	266
3	Toast	274	561	142	265
4	Egg	196	196	64	64
5	Sausage	119	222	57	120
6	Chicken Cutlet	51	153	15	45
7	Bacon	240	618	51	124
8	French Toast	67	179	23	62
9	Omelet	64	67	21	21
10	Hash Browns	90	142	50	68
11	Pancake	130	253	43	86
12	Ham	126	219	134	363
13	Patty	261	405	146	255
14	Sandwich	81	107	36	47
15	French Fries	105	131	33	37
16	Hamburger	80	81	48	48
-	Total	850	4,118	428	1,978

Three parameters discussed in this experiment are step per epoch, RPN train anchors per image, and train RoI per image. Epoch refers to training all images in the training set once during training. Step per epoch refers to training batch size, or how many images are trained each epoch. RPN train anchors per image refers to the number of RoIs output by the RPN. Train RoI per image refers to the number of anchors used in RPN training. For the above three parameters, one parameter is adjusted at a time with the remaining two parameters fixed to calculate the optimal rate for each parameter.

First, we consider the parameter adjustment of step per epoch. The highest correct rate is selected as the subsequent experiment of RPN train anchor per image. Then, the result with the optimal rate after adjusting the second parameter is selected for experimental use to adjust the train RoI per image. Precision, Recall, F1 Measure, and IoU are used to evaluate the experimental results.

After experimenting, a pre-trained weights of the COCO dataset were used, the three parameters of step per epoch, RPN train anchors per image, and train RoI per image were set to 1000, 256, and 512, respectively. After 18 epochs training, the average precision rate was 98.48%, the average recall rate was 96.31%, and the average IoU was 97.17%. The average accuracy rate of food recognition for the 16 classes was 99%. Of the 3,680 testing foods, 3,568 foods were predicted, of which five were misrecognized and 112 foods were unsuccessfully detected.

C. *Food Detection and Recognition Model Improvement*

As the first experiment had a high precision rate and a relatively low recall rate, we aimed to improve on these results. In the first experiment, it was found that the food images were placed too close to each other, resulting in some food, such as pancake, ham, and toast, not being detected successfully, as shown in Fig. 7. Figs. 7(a) and 7(c) are the original input food images, and Figs. 7(b) and 7(d) are the corresponding results of detection and recognition. The toast detection failed in Fig. 7(b), and the pancake and ham detection failed in Fig. 7(d).

(a) (b) (c) (d)

Figure 7. Example of food placement too close to recognize failure. (a) The original input image. (b) Toast detection failed. (c) The original input image. (d) Pancake and ham detection failed.

In this experiment, the accuracy rate was calculated after the union operation, and compared with the method before the improvement. Table 2 shows the precision rate, recall rate, and F1 measure before and after improvements. The results indicate that the three classes of toast, pancake, and ham had improved accuracy; toast, sausage, chicken cutlet, French toast, pancake, and ham improved their recall and F1 measure. After improvement, the average precision rate, recall rate, and F1 measure were 99.09%, 97.91%, and 98.50%, respectively.

TABLE II. THE ACCURACY RATES OF 16 CLASSES WITH NMS AND UNION.

No.	Class	Mask Color	Precision (With NMS)	Precision (Union)	Recall (With NMS)	Recall (Union)	F1 Measure (With NMS)	F1 Measure (Union)
1	Salad		100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
2	Fruit		100.00%	100.00%	98.19%	98.36%	99.09%	99.17%
3	Toast		**95.92%**	**100.00%**	**92.17%**	**100.00%**	**94.01%**	**100.00%**
4	Egg		100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
5	Sausage		100.00%	100.00%	**96.93%**	**100.00%**	**98.44%**	**100.00%**
6	Chicken Cutlet		100.00%	100.00%	**94.65%**	**95.88%**	**97.25%**	**97.90%**
7	Bacon		98.98%	98.98%	99.49%	99.49%	99.23%	99.23%
8	French Toast		100.00%	100.00%	**97.03%**	**99.26%**	**98.49%**	**99.63%**
9	Omelet		100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
10	Hash Browns		99.60%	99.60%	100.00%	100.00%	99.80%	99.80%
11	Pancake		**97.63%**	**100.00%**	**95.38%**	**100.00%**	**96.49%**	**100.00%**
12	Ham		**93.38%**	**94.08%**	**87.58%**	**88.82%**	**90.38%**	**91.37%**
13	Patty		93.71%	93.71%	88.16%	88.16%	90.85%	90.85%
14	Sandwich		96.97%	96.97%	94.12%	94.12%	90.52%	95.52%
15	French Fries		97.44%	97.44%	100.00%	100.00%	98.70%	98.70%
16	Hamburger		95.16%	95.16%	90.77%	90.77%	92.91%	92.91%
-	Total		98.48%	99.09%	96.31%	97.91%	97.38%	98.50%

Regarding the precision rate, the toast class increased from 95.92% to 100.00%, the pancake class increased from 97.63% to 100.00%, the ham class increased from 93.38% to 94.08%, and the average accuracy increased from 98.48% to 99.09%.

Regarding the recall rate, the toast class increased from 92.17% to 100.00%, the sausage class increased from 96.93% to 100.00%, the chicken cutlet class increased from 94.65% to 95.88%, the French toast class increased from 97.03% to 99.26%, the pancake class increased from 95.38% to 100.00%, and the ham class increased from 87.58% to 88.82%. The average accuracy increased from 96.31% to 97.91%.

Regarding the F1 measure, the toast class increased from 94.01% to 100.00%, the sausage class increased from 98.44% to 100.00%, the chicken cutlet class increased from 97.25% to 97.90%, the French toast class increased from 98.49% to 99.63%, the pancake class increased from 96.49% to 100.00%, and the ham class increased from 90.38% to 91.37%. The average accuracy increased from 97.38% to 98.50%.

Figure 8. An example of the original input image and improvement results.

Fig. 8 shows an example of the improved experimental results. The first row (top) is the original input image, the second row (middle) is the pre-improved output, and the third row (bottom) is the improved output. No sausage is detected in the first (pre-improved output) image of the second row. After the improvement, the results are as shown in the first image of the third row, and all the sausages are successfully detected. Two French toasts are not detected in the second image of the second row. The improved results show that all French toasts are successfully detected as shown in the second image of the third row. No chicken cutlet is detected in the third image of the second row. The improved output is shown in the third image of the third row, with all chicken cutlets successfully detected.

D. Food Weight Estimation

In the food weight estimation experiment, linear regression was performed for eight foods: salad, fruit, French toast, sausage, bacon, ham, patty, and French fries. Different food images were taken for each of the eight foods, and the weight of the food and the number of pixels in the images were recorded. The results of the food weight estimation experiment were evaluated by absolute and relative error. In the classes of salad, fruit, toast, sausage, bacon, ham, patty, and French fries, the absolute errors were, respectively, 2.71, 8.45, 15.98, 8.00, 2.50, 1.79, 6.53, and 9.83. The relative errors were 0.34, 0.11, 0.19, 0.11, 0.08, 0.07, 0.06, and 0.04, respectively. In the food weight estimation experiment, the final average absolute error was 8.22, and the average relative error was 0.13. Fig. 9 shows the estimated result for the patty weight estimation experiments.

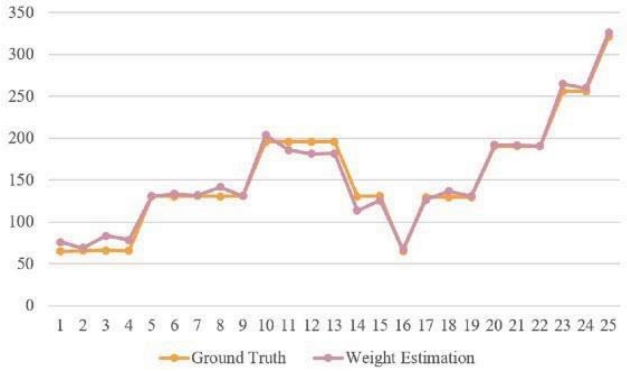


Figure 9. The weight estimation results of patty. The x-axis is the number of pixels in the food image and the y-axis is the estimated weight of the food.

IV. Conclusion and Future Work

The proposed system aims to help users manage their diet through food recognition and calorie nutrient analysis. This study uses food images as input to the system, based on Mask R-CNN to detect and recognize food class and food masks. The proportion of food in the image is obtained through the food mask, and the weight of the food is estimated by linear regression. The combination of food calories and estimated weights allows the system to ultimately label food calories and nutrients.

This study proposes the Ville Cafe dataset, which is divided into 16 food classes with 35,842 images and 9,776 food items. The Ville Cafe dataset collected five Western-style brunch restaurants with different food items, and most food images contain a variety of food. The accuracy of the combination of the Ville Cafe dataset and the Food-256 Dataset is 99.86% and IoU is 97.17%. In the weight estimation, the eight classes of salad, fruit, toast, sausage, bacon, ham, patty, and French fries had 40, 40, 44, 40, 41, 49, 26, and 40 data points, respectively, summing to 320 data points for the linear regression model. In the experimental results, the average absolute error was 8.22, and the average relative error was 0.13. In future, we hope to enhance the system to allow users to record each meal, not limited to breakfast food, and to provide dietary advice for patients with different conditions.

Acknowledgment

The authors would like to thank the Ministry of Science and Technology of the Republic of China, Taiwan for financially supporting this research under Contract No. MOST-107-2221-E-003-023 and MOST-108-2221-E-003-015. We would like to thank Uni-edit (www.uni-edit.net) for editing and proofreading this manuscript.

References

[1] C. A. Wu, "A Neural Network Model for Calorie and Nutrition Analysis based on Food Images." M.A. thesis, National Taiwan Normal University, Taiwan, 2019.

[2] D. M. Harris, S. Guten, "Health-protective behavior: An exploratory study," *Journal of Health and Social Behavior*, vol. 20, no. 1, pp. 17-29, 1979.

[3] Daily View, "Taiwan's most popular top ten breakfasts," Available at: https://dailyview.tw/Daily/2017/04/22?page=0pid=8388, accessed 2019.

[4] G. Villalobos, R. Almaghrabi, P. Pouladzadeh, and S. Shirmohammadi, "An Image Processing Approach for Calorie Intake Measurement," Proceedings of International Symposium Measurements and Applications, Budapest, pp. 1-5, May 2012.

[5] Health Promotion Administration - Ministry of Health and welfare, "Daily Diet Guide in 107," Available at: https://www.hpa.gov.tw/Pages/Detail.aspx?nodeid =1405&pid=8388, accessed 2019.

[6] Herbalife Nutrition Taiwan, "Asia Pacific Healthy Breakfast Survey," Available at: https://company.herbalife.com.tw/press-room/news-release/201610_wellnesstour, accessed 2019.

[7] J. Dehais, M. Anthimopoulos, and S. Mougiakakou, "Dish Detection and Segmentation for Dietary Assessment on Smartphones," *Proceedings of International Conference on Image Analysis and Processing*, Springer, pp. 433-440, 2015.

[8] John Tung Foundation, "Commercial Food Nutrition Labeling Understanding," Available at: https://nutri.jtf.org.tw/index.php?idd=10&aid=2&bid=34&cid=537, accessed 2019.

[9] K. He, G. Gkioxari, P. Doll´ar, and R. Girshick, "Mask R-CNN," *Proceedings of 2017 IEEE International Conference on Computer Vision*(*ICCV*), Italy, pp. 2980-2988, 2017.

[10] M. Anthimopoulos, J. Dehais, P. Diem, and S. Mougiakakou, "Segmentation and Recognition of Multi-food Meal Images for Carbohydrate Counting," *Proceedings of 2013 IEEE 13th International Conference on Bioinformatics and Bioengineering* (*BIBE*), Greece, pp. 1-4, 2013.

[11] M. Bolaños and P. Radeva, "Simultaneous Food Localization and Recognition," *Proceedings of 2016 23rd International Conference on Computer Vision and Pattern Recognition*(*CVPR*), Mexico, pp. 3140-3145, 2016.

[12] M. Livingstone, P. Robson, and J. Wallace, "Issues in Dietary Intake Assessment of Children and Adolescents," *Brit. J. Nutrition*, vol. 92, no. 2, pp. 213-222, 2004.

[13] M. M. Anthimopoulos, L. Gianola, L. Scarnato, P. Diem, and S. G. Mougiakakou, "A Food Recognition System for Diabetic Patients Based on an Optimized Bag-of-Features Model," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 4, pp. 1261-1271, 2014.

[14] P. Pouladzadeh, S. Shirmohammadi, and R. Al-Maghrabi, "Measuring Calorie and Nutrition from Food Image," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 8, pp. 1947-1956, Aug. 2014.

[15] S. W. Lichtman, K. Pisarska, E. R. Berman, M. Pestone, H. Dowling, E. Offenbacher, H. Weisel, S. Heshka, D. E Matthews, and S. B Heymsfield, "Discrepancy between self-reported and actual caloric intake and exercise in obese subjects," *New England Journal of Medicine*, vol. 327, no. 27, pp. 1893-1898, 1992.

[16] Taiwan Ministry of Health and Welfare - Food and Drug Administration, "Taiwan Food Nutrition Database," Available at: https://consumer.fda.gov.tw/Food/TFND.aspx?nodeID=178&rand=1530933874, accessed 2019.

[17] U.S. Department of Agriculture, "FoodData Central – Download Data," Available at: https://fdc.nal.usda.gov/index.html, accessed 2019.