Region-Based Food Calorie Estimation for Multiple-Dish Meals

Kaimu Okamoto Kento Adachi Keiji Yanai
{okamoto-ka,adachi-k,yanai}@mm.inf.uec.ac.jp
Department of Informatics, The University of Electro-Communications
Chofu-shi, Tokyo, Japan

ABSTRACT

One of the major tasks in food computing is vision-based food calorie estimation. However, unfortunately, food image datasets annotated with calorie amounts are very hard to obtain. In fact, no large-scale food datasets annotated with calorie amounts exits as long as we know. However, we can see some Web sites which provide photos of food set menus with only total calorie values. Then, in this work, we crawl such data from the Web and use them as training data of a vision-based food calorie estimation model. To estimate calorie amounts of food items, the calorie values of each food item in meal photos should be known in general. However, they are not available in this setting. Then, we propose a model employing food segmentation which can estimate calorie amounts of each food item from total calorie values of set meal photos. The experimental results showed that our region-segmentation-based calorie estimation model was able to estimate calorie amounts of individual food items roughly.

CCS CONCEPTS

• **Computing methodologies → Image segmentation**.

KEYWORDS

food image, food calorie estimation, food region segmentation, UEC-FoodPix Complete

ACM Reference Format:
Kaimu Okamoto Kento Adachi Keiji Yanai. 2021. Region-Based Food Calorie Estimation for Multiple-Dish Meals. In *Proceedings of the 13th International Workshop on Multimedia for Cooking and Eating Activities (CEA '21), August 21–24, 2021, Taipei, Taiwan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3463947.3469236

1 INTRODUCTION

Recently food computing [26] draws much attention in the field of multimedia as well as dietary assessment. The task of estimating the calorie values of meals from images is one of the most important and challenging problems in food computing. However, no large-scale food datasets annotated with calorie amounts exits as long as we know. In addition, unfortunately, food image datasets annotated with calorie amounts are very hard to create and obtain.

CEA '21, August 21–24, 2021, Taipei, Taiwan

ACM ISBN 978-1-4503-8532-9/21/08…$15.00
https://doi.org/10.1145/3463947.3469236

Recently, we can see some Web sites which provide photos of food set menus with only the total value of food calorie amounts. Some restaurant menu sites and some blogs on school lunch provide such information. Then, in this work, we crawl such data from the Web, and use them as training data of a vision-based food calorie estimation model. To estimate calorie amounts of food items, the calorie values of each food item in meal photos should be known in general. However, they are not available in this setting. Then, we propose a model which can estimate calorie amounts of each food item from the total calorie values of set meal photos. For the vision-based calorie estimation, food region segmentation is important because there exists a close relationship between the area size of food regions and the amount of food calories. Then, we integrate a food calorie estimation model and a food region segmentation model. In our proposed model, we estimate food regions by a semantic segmentation model and estimate the total calorie values of food items in a given photo simultaneously. The experimental results showed that our region-segmentation-based model was able to estimate calories amounts of individual food items more accurately than a direct regression model and a regression model with multi-label classification.

2 RELATED WORK

In this section, we mention related works on food image segmentation and food calorie estimation.

2.1 Food Region Segmentation

For food region segmentation, food segmentation datasets are required. Although some large-scale generic image segmentation datasets such as MS-COCO [21] contains food categories, the number of them are very limited. In fact, MS-COCO contains only 10 food categories such Pizza, banana and hot dogs out of all the 80 categories. Regarding food image datasets, there exists some large-scale datasets such as ISIA Food-500 [27] and Recipe1M+ [24]. However, most of them contains only image-level labels and no food calorie information, which means they cannot be directly applied to food calorie estimation tasks.

Currently, there are a few large-scale open food image datasets with segmentation masks. The UNIMIB2016 dataset [9] provides food region information as polygons which are equivalent to segmentation masks. However, its scale is not so large (1027 multiple-dish images with 73 food categories), and the food images in UNIMIB2016 are biased and not unconstrained since all the food images were taken at the same canteen.

Lu *et al.* [22] proposed a food volume estimation method by extending Mask R-CNN [17] which extracts food regions from a given RGB-D image. To train the proposed model, they used the MADiMa17 dataset [1] which consists of 21 food categories with

segmentation masks. However, all the images in the MADiMa17 dataset were taken in the laboratory environment which was different from uncontrolled real situations.

Okamoto *et al.* proposed a region-based food calorie estimation system, *CalorieCam*, running on a Android mobile phone [30]. They employed food image segmentation and estimated food calories based on the size of the reference card and food regions. However, at that time, no food image segmentation dataset on uncontrolled food images is available. Instead of using semantic region segmentation methods which requires training data, they used GrabCut [34], which was a hand-crafted segmentation method that divides the foreground and background by graph-based reasoning.

To change this situation, Ege *et al.* created a large-scale food image segmentation dataset, which was called "the UEC-FoodPix dataset" [14]. They added pixel-wise annotation to 10,000 food images included in the UECFood-100 dataset [25]. Regarding 1,000 food images for testing, they added pixel-wise labels by hand, while for the other 9,000 images they created pixel-wise labels automatically by applying GrabCut [34] on each of the bounding boxes originally annotated in the UECFood-100 dataset. Before applying GrabCut, they verified if the bounding box annotations were enough correct one by one, and revised them if needed. In addition, Ege *et al.* [14] proposed a method to estimate actual size of foods without a reference card for estimating food calorie amounts of uncontrolled food images. To do that, they proposed to estimate actual size of foods in a given image by using the size of rice grains as reference objects.

However, since the UECFoodPix created by Ege *et al.* [14] generated pixel-wise annotations semi-automatically, it may contain noisy annotations, which is expected to be harmful for training of CNN models. Then, Okamoto *et al.* improved it by revising pixel-wise annotation of training data in UEC-FoodPix by hand [31]. The improved dataset is called "UEC-FoodPix Complete", which consists of 10,000 fully-hand-annotated food image segmentation masks.

As the other dataset for unconstrained food images, Google Food-201 [28], SUEC Food Dataset [16], and Food segmentation benchmark [4] have been released so far. Food-201 [28] was created for the Im2Calories project by Google, and released to the public several years after the paper was published. They annotated 201 pixel-level labels to parts of the images in the ETH Food-101 dataset [5] with the help of the crowd-sourcing workers. SUEC Food Dataset [16] was created by GrabCut [34] based on the bounding box annotation of UEC-Food256 [20]. The dataset for segmentation benchmark created by the UMINIB group [4] contains 5,000 segmentation masks for all the images of 50-category Chinese food image dataset [8]. We listed the current public food segmentation datasets on unconstrained food images in Table 1.

2.2 Food Calorie Estimation

Some works for image-based food calorie estimation have been proposed so far. Ege *et al.* estimated the calorie amount directly from a single food image without depth information using a multi-task CNN simultaneously trained with recipe information (food categories, ingredients and cooking directions) and calorie amount [12, 13]. However, this method did not consider the size or volume of foods, and they assumed that a dish in a given image was a portion for one person. Being different from this work, in our work, we estimate the amount of food calories based on the estimated region size of the foods.

Because multiple view stereo was used to be one of the main topics of computer vision research 10 years ago, some old works tried to introduce multi-view 3D reconstruction into food calorie estimation. As one of such the existing works, Puri *et al.* proposed a volume-based food calorie estimation system employing multiple-view 3D reconstruction [32]. However, to reconstruct 3D shapes of dishes accurately, many images and heavy computation were needed. Dehais et al. [10] also proposed a food volume estimation by two-view 3D reconstruction.

On the other hand, Myers *et al.* used CNN-based depth estimation from a single image for food calorie estimation [28]. They estimated the volume of food with depth information estimated by a depth estimation CNN. They employed CNN-based segmentation [7] and CNN-based 3D volume estimation from 2D single images [15] in addition to CNN-based food category recognition. Although they achieved relatively high accuracy for volume estimation, their method needed a large amount of RGB-D food images and pixel-wise annotated segmentation masks of food images which were costly to obtain for training.

Allegra *et al.* [2] also tried depth and volume estimation of a food image from a single image employing a CNN. They created a new food RGB-D image dataet, Madima17, for training of a CNN. Lu *et al.* [23] proposed a multi-task CNN architecture to perform food segmentation and food volume estimation simultaneously.

Okamoto *et al.* developed *CalorieCam*, a system that estimated the calorie amount of food simply by shooting with the camera of the smartphone [29]. *CalorieCam* can estimate food regions automatically from a single image taken by a smartphone built-in camera, and estimate the amount of their calories based on the real size of food regions. To estimate real size of foods, a user needs to take a food photo with a reference object the real size of which is known in advance at the time of photographing.

While *CalorieCam* need to take a food photo with a size-known reference object, *DepthCalorieCam* proposed by Ando *et al.* [3] does not need to prepare any reference object since it employs stereo camera built in an iPhone. Real volume values of food items can be estimated by *DepthCalorieCam*, because all the camera parameters of the iPhone are known in advance. In addition, since *DepthCalorieCam* employs CNN-based image segmentation, the accuracy of food segmentation is much higher that *CalorieCam*.

Tanno *et al.* have developed *AR DeepCalorieCam V2* which does not need any reference object to estimate the calorie of the foods by using AR [36]. By using visual inertial odometry which is a fundamental technique of AR, the real size of things can be estimated by a smartphone having only a single camera.

3 METHOD

Currently, existing calorie amount estimation is based on the assumption that the calorie amounts for each of individual food categories is known. For example, when dealing with an image containing only a single food item, such as *pizza*, we only need to prepare one calorie amounts for one corresponding food image. However, when dealing with an image containing multiple food items, such as

Table 1: A list of the public food segmentation datasets on unconstrained food images.

Dataset name	release	#image	#class	annotation	Original dataset
Google Food-201 [28]	2017	12,093	208	crowdworker	ETH Food101 [5]
UEC-FoodPix [14]	2019	10,000	102	auto (GrabCut)	UEC-Food100 [25]
SUEC Food Dataset [16]	2019	28,897	256	auto (GrabCut)	UEC-Food256 [20]
Food segmentation benchmark [4]	2020	5,000	50	controlled	Chinese food 50 categories [8]
UEC-FoodPix Complete [31]	2021	10,000	102	controlled	UEC-Food100 [25]

rice, *miso soup*, and *grilled fish*, we need to prepare calorie amounts for each of the food item in a single meal image. However, in the real situation, only the total amounts of calories of food items is sometimes provided, and no calorie amounts of each of the items is given. Especially, at the Web sites on restaurant menus and school lunches, we can observe such tendency.

To take advantage of such data for food calorie estimation, in this work, we propose a model for estimating the calorie amounts of individual food items in a set meal image from the dataset containing meal set photos annotated with only total calorie values. This is a new problem setting in the food calorie estimation tasks. Since the calorie amount and the size of food regions are closely related, we use semantic food region segmentation to estimate the calorie amounts of individual food items.

3.1 Dataset Construction

In this study, we need a food image dataset with a segmentation mask in which the overall calorie content is known in order to perform both meal region estimation and calorie content estimation simultaneously. However, there is no public food segmentation dataset annotated with calorie amounts. Therefore, we have created a calorie-annotated food segmentation image dataset where only total calorie amounts are annotated to each of set meal photos. Originally, Ege *et al.* [11] used school lunch images collected from the Inzai City School Lunch Center website [1] in order to estimate food calories for multiple-dish food photos. Note that they trained a food calorie estimation model not from the school lunch data but from the other calorie-annotated single dish photos, and they used school lunch photos for training a food detection model and testing their model after adding 22-food-category bounding box annotations by hand. In this site, photos with the list of food items and total calorie amounts of school lunch are posted everyday.

However, no food segmentation masks are provided. Therefore, we added segmentation mask images to the gathered school lunch images using the same UI when we created the UECFood-Pix Complete dataset. Some samples in the dataset are shown in Figure 1. This dataset contains 593 school lunch images including 474 training images and 119 evaluation images annotated with 60 food category pixel-wise labels and the total calorie values. In addition, we associated the 60 food categories of the annotated segmentation masks with 22 food categories of the bounding boxed Ege *et al.* annotated in [11] in order to enable comparison on calorie estimation accuracy between ours and [11]. Note that we will release this dataset at the time of publication.

[1] http://inzai.ed.jp/kyusyoku/?page_id=32

Figure 1: School lunch images annotated with the total calorie amounts gathered from the school lunch blog of Inzai city, Chiba, Japan.

3.2 Overview of the Method

In this work, we estimate the amount of calories of individual food items from the total calorie amounts using region segmentation. We use a region segmentation model, Deeplab V3+ [6], as a base model. We propose to add a new calorie estimation branch at the location of the concatenated features marked with a red box in the decoder part of Figure 2 to estimate the amount of food calories. In this branch, we estimate the calorie amount vector v_c for each meal category by combining the calorie amount feature map created from the image features and the segmentation feature map calculated from the segmentation part of Deeplab V3+. An overview of the overall model used for calorie estimation is shown in Figure 3. The model is trained in two stages: first, the region segmentation part is trained using Deeplab V3+, and then the calorie amount estimation part is trained with the parameters of the region segmentation part fixed. The calculated calorie amounts consists of calorie values of each of the food items in the given photo, and training is performed using the summation of all the calorie values.

3.3 Segmentation Model

In the first step, we train the segmentation branch which is based on DeepLab V3+. We used ResNet-101 [18] as the backbone of the segmentation branch. We trained the model using the created school lunch dataset with both 60 categories and 22 categories. For evaluation, we used Accuracy and mIoU (mean intersection over union) . The evaluation results are shown in Table 2. For calorie estimation, the parameters of the trained Deeplab V3+ part are fixed in order to guarantee and learn the relationship and location information between meal categories.

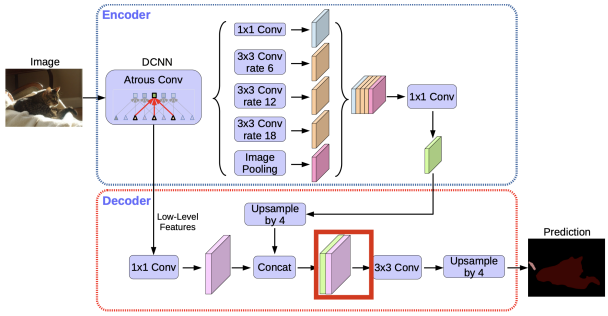

Figure 2: The architecture of Deeplab V3+ (cited from [6]).

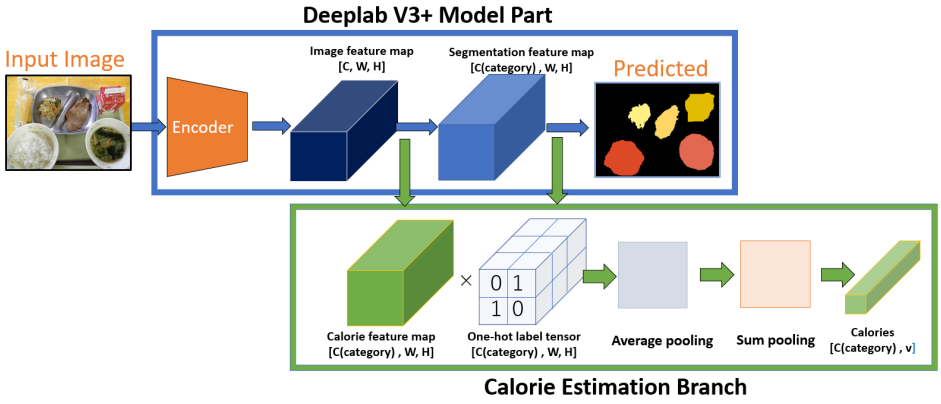

Figure 3: The proposed architecture consisting of the segmentation branch and the calorie estimation branch.

Table 2: Food segmentation performance.

traning data	Acc	mIoU
School Lunch (60 categories)	0.610	0.484
School Lunch (22 categories)	0.744	0.660

3.4 Calorie Estimation Branch

In the calorie amount estimation branch, the calorie amount vector v_c is calculated using the feature map of calorie amount $T_{calorie}$ and the tensor $T_{one-hot}$, which consists of one-hot vectors along the channel direction generated from the segmentation feature map. The $T_{calorie}$ is calculated in the same way as the segmentation feature map from the image features calculated from the encoder of Deeplab V3+. That is, the image features are convolved by passing them through two 3×3 Conv layers, and the number of channels is adjusted by a 1×1 Conv layer, resulting in a calorie feature map with size $C \times W \times H$. The C here corresponds to the number of categories in the data set, and this caloric feature map is the same size as the segmentation feature map.

In the segmentation feature map, the softmax function is applied in the channel direction and the channel with the maximum value is selected to determine the category at each pixel and generate the mask image. In the segmentation feature map, the softmax function is applied in the channel direction, and the channel with the maximum value is set to 1 and the other channels are set to 0. This creates a tensor $T_{one-hot}$ with one-hot values in the channel direction. This allows us to spatially preserve the relationship between categories that exist simultaneously in the image. In addition, it does not include the process of resizing the feature map, so it is the same size as the calorie feature map. By taking the Hadamard product of $T_{calorie}$ and $T_{one-hot}$ and performing Average pooling and Sum pooling, we calculate the calorie quantity vector v_c for each meal category. The $T_{one-hot}$ of this branch depends on the accuracy of the segmentation feature map. Therefore, we fix the parameters of the trained segmentation part, and calculate $T_{one-hot}$, which is used to guarantee the location of each meal category. The detailed network is shown in Table 3.

Table 3: The network of the calorie estimation branch.

input size	Operator	stride	padding
calorie map estimation part			
$129^2 \times 304$	conv2d 3x3	1	1
$129^2 \times 256$	BatchNorm	-	-
$129^2 \times 256$	ReLU	-	-
$129^2 \times 256$	conv2d 3x3	1	1
$129^2 \times 256$	BatchNorm	-	-
$129^2 \times 256$	ReLU	-	-
$129^2 \times 256$	conv2d	1	0
$129^2 \times 60$	BatchNorm	-	-
$129^2 \times 60$	Softplus	-	-
calorie vector estimation part			
$129^2 \times 60$	$T_{calorie} \otimes T_{one-hot}$	-	-
$129^2 \times 60$	AvgPool2d 7x7	5	0
$25^2 \times 60$	SumPooling	-	

3.5 Loss Function

In the calorie amount estimation branch, a vector v_c with the calorie amounts of individual meal categories is calculated. The total calorie amount v, which is the sum of the calorie amounts of each meal category, and the total calorie amount g, which is the correct answer, are used to learn the calorie amount estimation branch. The loss function L_{cal} used for learning is the weighted absolute error L_{ab} and the relative error L_{re}.

The weighted absolute error is the absolute error weighted against values close to zero to avoid calculating values where the caloric content is close to zero, and is defined as follows:

$$L_{ab} = \lambda |v - g| \begin{cases} \lambda = 1.0 & (v - g \geqq 0) \\ \lambda = 1.2 & (otherwise) \end{cases} \tag{1}$$

The relative error loss is defined as

$$L_{re} = \frac{|v - g|}{g} \tag{2}$$

The toral loss function is defined as

$$L_{cal} = \lambda_{ab} L_{ab} + \lambda_{re} L_{re} \tag{3}$$

In the experiment, we empirically set λ_{ab} and λ_{re} to $\lambda_{ab} = 0.1$ and $\lambda_{re} = 0.01$, respectively, for training.

4 EXPERIMENTS

4.1 Experimental settings

In the experiments, we estimate not only the total amount of calories in school lunch photos but also the calorie amounts of each of the food items in the lunch photos. For comparison, we prepare two baseline models in addition to our proposed model. Model A is a regression model that directly estimates the total amount of calories of all the food items in a given photo, and Model B is a multi-task model of estimating multi-label food categories and the total food calorie amounts. Model B is based on VGG16 [35]. First, we perform multi-label classification training using VGG16 as a backbone. Next, after the convolutional layer of the trained VGG16, we add a branch that estimates the amount of calories for each meal category using the FC layer. Finally, we estimate the amount of calories for each meal category by multiplying the results of the multi-label classification with the results of the calorie amount estimation branch. Let model C be the Deeplab V3+ model with the proposed calorie amount estimation branch added. The total calories are the sum of the calorie amounts for each meal category estimated by each model.

For training and evaluation, we used the created meal dataset, and conducted experiments with both 60 and 22 categories. Stochastic Gradient Descent (SGD) is used as the optimization method with a Momentum value of 0.9 and a learning rate of 10^{-4}. The batch size is set to 4, and 200 epoch iterative learning is performed.

4.2 Results

As a measure of results, we calculate the absolute error of the estimated total calorie amount and the mean of the calorie amount for each meal category. The mean absolute error of the total calorie amounts is shown in Table 4.

Table 4: Absolute error on total calorie amounts (kcal).

model	school lunch (22)	school lunch (60)
model A (direct regression)	-	45.0
model B (multi-label)	-	44.2
model C (Ours)	74.5	74.8
baseline (Ege *et al.* [11])	136.8	-

Regarding the absolute error of the total calorie amounts, Model A and Model B were close to each other with the accuracy of 45.0 kcal and 44.2 kcal. Since the average calorie amount of the school lunch is around 650 kcal, both are less than 10% in the relative error rate, which means good enough as results of the food calorie estimation. On the other hand, Model C (the proposed method) achieved the larger error with 74.8 kcal compared to the two models. This result may be due to the fact that the values for each meal category, which will be explained later, are somewhat constant for all images in Models A and B, while Model C calculates values that include variations. In fact, compared to [11], they achieved 136.8 kcal as the absolute error which is much larger than ours, although they estimated the calorie amounts of each of the food items in a school lunch photo by food item detection based on Faster RCNN [33] and summed them.

For the amount of calories in each food category, the average estimated calories of 13 representative food items among the 60 school lunch categories are shown in Table 5, and the average estimated calories of the 22 categories are shown in Table 6. Some results of the experiments with Model C for each image are shown in Figure 4 and Figure 5. Note that in these figures, the regions of the estimated food items are colored and bounded by red boxes with calorie amount values of food items for easy identification.

In our dataset, we have no annotation of calorie amounts on individual food items. Although the test data should have been annotated with the calorie amount of each food item, this time we could not prepare that. Instead, for reference, we prepared the reference calorie amounts of standard size of each of the individual food items by referring the Dietary Life Checkbook [2] and the Ministry of Education, Culture, Sports, Science and Technology's Food Composition Database[3]. Meal sizes were taken from the Web page [4], which lists standard calorie amounts and grams. We added them as "Reference Calorie" in the last columns of Table 5 and 6.

As shown in Table 5, Model A estimates a constant value of around 10 kcal for all of the food items, which are totally different from the reference values. In Model B, certain categories, especially milk, accounts for the majority of the total calories. On the other hand, Model C estimates the calorie values for each of the food items without being biased toward a certain meal category. Most of the values are close to the reference values. This tendency is also observed in the results for 22-categories as shown in Table 6.

We will discuss the amount of calories in each food category estimated by Model C. The calorie value of *milk* in all images was 123.3 kcal, which is close to the reference caloric value when using the school lunch data of 60 categories as shown in Table 5. In

[2] https://www.mhlw.go.jp/bunya/kenkou/pdf/eiyou-syokuji8.pdf
[3] https://fooddb.mext.go.jp
[4] https://www.eiyoukeisan.com

Table 5: Average estimated calorie amounts of 13 category food items among 60 categories (kcal).

Category	Model A	Model B	Model C	Ref. Calorie
Milk	11.5	457.6	123.3	130
Rice	12.1	50.9	209.4	250
Mixed rice	10.6	21.7	220.2	250
Bread	10.8	7.5	190.3	220
Japanese salad	11.9	33.1	60.2	50
Green salad	11.9	33.0	68.9	50
Grilled pork loin	12.2	2.5	74.4	240
Grilled chicken teriyaki	11.6	2.0	35.8	300
Miso soup	10.3	59.9	147.6	159
Minestrone	11.1	15.5	141.4	159
Fruit punch	11.1	5.2	68.1	50
Jelly	11.3	3.1	55.4	89
Oranges	9.6	0.5	45.5	50

Table 6: Average estimated calorie amounts of 22 category food items with Model C.

category	estimated calorie (kcal)	reference calorie (kcal)	category	estimated calorie (kcal)	reference calorie (kcal)
Background	-	-	Small tomatoes	-	6
Milk	112.0	130	Soup	157.8	159
Yogurt	108.0	60	Curry	196.9	118
Rice	159.8	250	Bean curd	182.5	230
Mixed rice	178.2	250	Bibimbap	74.1	50
Bread rolls	194.3	220	Yakisoba	214.4	130
Bread rolls	214.1	220	meat spaghetti	181.3	260
Udon noodles	198.6	139	Citrus	47.0	13
Fish	84.4	120	Apple	27.3	13
Meat	48.6	240	Cup dessert	47.8	-
Salads	77.1	50	Other	64.6	-

addition, the [11] study by Ege *et al.* also determined a value close to 134 kcal, which is a reasonable value. Using the school lunch data of 22 categories, we also estimated a value close to 112.0 kcal as shown in Table 6. For staple food categories such as *rice* and *mixed rice*, the 60-category school lunch data set is closer to the reference calorie value, while the 20-category school lunch data set is closer to the reference calorie value for *shaped bread* and *coppe bread*. Note that we estimated calorie amounts of *milk* in the same way as the other food items in the experiments, although we could have given a fixed calorie value to the paper pack of *milk*.

For the vegetable, soup, and dessert categories, similar values were estimated. For the meat categories, the number of the images was small for each category, and the amount of calories varied depending on the cooking method and the amount of food served.

The experimental results showed that the proposed method, Model C, was the most reasonable method for estimating the calorie amount for each meal category from the total calorie amount when comparing the models.

4.3 Discussions

In the experiments, we conducted comparative experiments using three models. First, in terms of the average absolute error of the total calorie amount, Model A and Model B have good accuracy, and the proposed method has the lowest value. This may be due to the fact that the total calories of the school lunch images are around 650 kcal, which is an easy data set to estimate. Therefore, it was possible to estimate the calorie amount with high accuracy even with a simple model, because the accuracy of the results was good by simply estimating the closest calorie amount without considering the detailed visual analysis. It is expected that the accuracy would be different when using a data set with variations in the total calorie amounts.

On the other hand, the proposed method, Model C, showed the best accuracy in estimating the caloric value for each meal category. This is probably because the proposed method, Model C, was able to estimate the calorie value with variations in each meal category because it considered the relationship between each category and the size of the regions. Model A is unable to learn explicit relationships between labels, resulting in only uniform values being calculated. Model B resulted in certain categories, especially *milk*, accounting for the majority of the total calorie content. This may be due to the fact that only the relationship between categories was taken into account, and also because multi-labeling alone does not allow the calorie amount to reflect the information of categories that contain only a small number of training samples. From these results, it was found that using region estimation to estimate the calorie amount for each meal category was more appropriate. In addition, it is highly likely that the accuracy of the estimated values will be improved by increasing the size of the data set.

The weak point of region-segmentation-based methods is that less calorie amounts than actual are expected to be estimated in case that a part of food items are out of the image like some *bread* and *rice* regions as shown in Figure 4 and Figure 5. This is one of the limitations of image-based calorie estimation.

In this work, we estimated the calorie amount for each meal category using a segmentation-masked school lunch dataset with overall calorie amounts. The proposed model consisting of Deeplab V3+-based region segmentation with a calorie amount estimation branch enabled us to estimate the calorie amount for each meal category.

Here, we summarize the operations that affect the estimation of calorie amount for each meal category. First of all, since the amount of calories is basically positive, it is necessary to make the value positive in the lower part of the branch. This time, we used the Softplus function instead of the ReLU function as the last activation function in the calorie feature map estimation section of Table 3. This is because we have observed that when the ReLU function is used, if a negative value is calculated in the early stage of branch training, training does not proceed and a specific meal category that should exist in the image becomes 0 kcal. To solve this problem, we used the Softplus function, which has a very small positive value even when a negative value is calculated. The next step is to adjust the number of parameters in the calorie amount vector estimator in Table 3. In the initial stage of the calorie vector estimation part, we tried to estimate the amount of calories in each meal category by directly performing Sum Pooling without including the Average Pooling layer. However, we observed a phenomenon where a huge amount of calories were instantaneously calculated during the training process, and the training process did not proceed. Also, when the size of the feature map after the Average Pooling layer

Figure 4: Estimated calorie amounts with the model trained with 60 categories. Each of the food items in the photos is annotated with the food category no. and the estimated calorie amounts.

Figure 5: Estimated calorie amounts with the model trained with 22 categories.

was set to a large value, the total amount of calories became zero. Therefore, it is necessary to adjust the size of the feature map after the Average Pooling layer, and it can be said that this method depends on the size of the input image and the number of categories in the data set.

5 CONCLUSIONS

In this paper, we proposed a model and dataset for dietary calorie amount estimation using multiple-dish food images annotated only total calorie amounts, which are relatively easy to obtain from the Web. The calorie amount estimation model was implemented by adding a calorie estimation branch to Deeplab V3+ which is a common semantic segmentation model. The dataset was constructed by collecting 593 images of multiple meals with total calorie amounts from school lunch blogs and adding pixel-wise food category annotations to them. In the experiment, we demonstrated the effectiveness of the proposed model which took account of food segmentation results for estimating a calorie amount of each of the food items in a given food image.

Currently we have only a small-scale dataset containing only school lunch images for estimating food calorie amounts of each of the food items from food images with total calorie amounts. To enable large-scale experiments, we plan to gather more diverse and larger number of multiple-dish food images annotated with total food calorie values from the Web, and apply a zero-shot food image segmentation method [19] into them for reducing the cost of pixel-wise annotation. With a large-scale dataset, we will improve our model by introducing recent techniques such as transformer networks.

REFERENCES

[1] D. Allegra, M. Anthimopoulos, J. Dehais, Y. Lu, F. Stanco, G. M. Farinella, and S. Mougiakakou. 2017. A Multimedia Database for Automatic Meal Assessment Systems. In *Proc. of the ICIAP Workshop on Multimedia Assisted Dietary Management.*

[2] D. Allegra, M. Anthimopoulos, J. Dehais, Y. Lu, F. Stanco, G. M. Farinella, and S. Mougiakakou. 2017. A Multimedia Database for Automatic Meal Assessment Systems. In *Proc. of International Workshop on Multimedia Assisted Dietary Management (MADiMa).*

[3] Y. Ando, Ege T., J. Cho, and K. Yanai. 2019. DepthCalorieCam: A Mobile Application for Volume-Based Food Calorie Estimation using Depth Cameras. In *Proc. of ACM MM Workshop on Multimedia Assisted Dietary Management.*

[4] Sinem Aslan, Gianluigi Ciocca, Davide Mazzini, and Raimondo Schettini. 2020. Benchmarking Algorithms for Food Localization and Semantic Segmentation. *International Journal of Machine Learning and Cybernetics* 11 (2020), 2827–2847.

[5] L. Bossard, M. Guillaumin, and L. Van Gool. 2014. Food-101 – Mining Discriminative Components with Random Forests. In *Proc. of European Conference on Computer Vision.*

[6] L. Chen, Y. Zhu, G. Papandreou, F Schroff, and H. Adam. 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Proc. of European Conference on Computer Vision.*

[7] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. 2014. Semantic image segmentation with deep convolutional nets and fully connected CRFs.

[8] Mei-Yun Chen, Yung-Hsiang Yang, Chia-Ju Ho, Shih-Han Wang, Shane-Ming Liu, Eugene Chang, Che-Hua Yeh, and Ming Ouhyoung. 2012. Automatic Chinese Food Identification and Quantity Estimation. In *Proc. of SIGGRAPH Asia.*

[9] G. Ciocca, P. Napoletano, and R. Schettini. 2017. Food recognition: a new dataset, experiments and results. *IEEE Journal of Biomedical and Health Informatics* 21, 3 (2017), 588–598.

[10] J. Dehais, M. Anthimopoulos, S. Shevchik, and S. Mougiakakou. 2017. Two-View 3D Reconstruction for Food Volume Estimation. *IEEE Transactions on Multimedia* 19, 5 (2017), 1090–1099.

[11] T. Ege and K. Yanai. 2017. Estimating Food Calories for Multiple-dish Food Photos. In *Proc. of Asian Conference on Pattern Recognition (ACPR).*

[12] T. Ege and K. Yanai. 2017. Image-Based Food Calorie Estimation Using Knowledge on Food Categories, Ingredients and Cooking Directions. In *Proc. of ACM Multimedia Thematic Workshops on Understanding.*

[13] T. Ege and K. Yanai. 2018. Image-Based Food Calorie Estimation Using Recipe Information. *IEICE Transactions on Information and Systems* E101-D, 5 (2018), 1333–1341.

[14] T. Ege and K. Yanai. 2019. A New Large-scale Food Image Segmentation Dataset and Its Application to Food Calorie Estimation Based on Grains of Rice. In *Proc. of ACM MM Workshop on Multimedia Assisted Dietary Management.*

[15] D. Eigen and R. Fergus. 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolution al architecture. In *Proc. of IEEE International Conference on Computer Vision (ICCV).* 2650–2658.

[16] Junyi Gao, Weihao Tan, Liantao Ma, Yasha Wang, and Wen Tang. 2019. MUSEFood: Multi-sensor-based Food Volume Estimation on Smartphones. *arXiv:1903.07437* (2019).

[17] K. He, G. Gkioxari, P. Dollar, and R. Girshick. 2017. Mask R-CNN. In *Proc. of IEEE International Conference on Computer Vision.*

[18] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *Proc. of IEEE Computer Vision and Pattern Recognition.* 770–778.

[19] Y. Honbu and K. Yanai. 2021. Few-shot and zero-shot semantic segmentation for food images. In *Proc. of ICMR Workshop on Multimedia for Cooking and Eating Activities (CEA).*

[20] Y. Kawano and K. Yanai. 2014. Automatic Expansion of a Food Image Dataset Leveraging Existing Categories with Domain Adaptation. *ECCV Workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV)* (2014).

[21] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P/ Dollár, and C/ Zitnick. 2014. Microsoft coco: Common objects in context. In *Proc. of European Conference on Computer Vision.*

[22] Y. Lu, D. Allegra, M. Anthimopoulos, F. Stanco, G. M. Farinella, and S. Mougiakakou. 2018. A Multi-Task Learning Approach for Meal Assessment. In *Proc. of the IJCAI Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management.* 46–52.

[23] Y. Lu, D. Allegra, M. Anthimopoulos, F. Stanco, G. M. Farinella, and S. G. Mougiakakou. 2018. A Multi-Task Learning Approach for Meal Assessment. In *Proc. of International Workshop on Multimedia Assisted Dietary Management (MADiMa).*

[24] Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. 2021. Recipe1M+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 1 (2021), 187–203.

[25] Y. Matsuda, H. Hajime, and K. Yanai. 2012. Recognition of Multiple-Food Images by Detecting Candidate Regions. In *Proc. of IEEE International Conference on Multimedia and Expo.* 25–30.

[26] Weiqing Min, Shuqiang Jiang, Linhu Liu, Yong Rui, and Ramesh Jain. 2019. A Survey on Food Computing. *ACM Computing Survey* 52, 5 (2019).

[27] Weiqing Min, Linhu Liu, Zhiling Wang, Zhengdong Luo, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. 2020. ISIA Food-500: A Dataset for Large-Scale Food Recognition via Stacked Global-Local Attention Network. In *Proc. of ACM International Conference Multimedia.* 393–401.

[28] A. Myers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and P. K. Murphy. 2015. Im2Calories: towards an automated mobile vision food diary. In *Proc. of IEEE International Conference on Computer Vision.* 1233–1241.

[29] K. Okamoto and K. Yanai. 2016. An Automatic Calorie Estimation System of Food Images on a Smartphone. In *Proc. of ACM MM Workshop on Multimedia Assisted Dietary Management (MADiMa).*

[30] K. Okamoto and K. Yanai. 2016. An Automatic Calorie Estimation System of Food Images on a Smartphone. In *Proc. of ACM MM Workshop on Multimedia Assisted Dietary Management.*

[31] K. Okamoto and K. Yanai. 2021. UEC-FoodPIX Complete: A Large-scale Food Image Segmentation Dataset. In *Proc. of ICPR Workshop on Multimedia Assisted Dietary Management(MADiMa).*

[32] M. Puri, Z. Zhu, Q. Yu, A. Divakaran, and H. Sawhney. 2009. Recognition and Volume Estimation of Food Intake Using a Mobile Device. In *Proc. of Workshop on Applications of Computer Vision (WACV).*

[33] S. Ren, K. He, R. Girshick, and J. Sun. 2017. Faster R-CNN:Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 6 (2017), 1137–1149.

[34] C. Rother, V. Kolmogorov, and A. Blake. 2004. GrabCut: Interactive Foreground Extraction Using Iterated Graph Cuts. *ACM Transaction on Graphics* 23, 3 (2004), 309–314.

[35] K. Simonyan and A Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations.*

[36] R. Tanno, T. Ege, and K. Yanai. 2018. AR DeepCalorieCam V2: Food Calorie Estimation with CNN and AR-based Actual Size Estimation. In *Proc. of ACM Symposium on Virtual Reality Software and Technology (VRST).*