DepthGrillCam: A Mobile Application for Real-time Eating Action Recording Using RGB-D Images

Kento Adachi Keiji Yanai
The University of Electro-Communications, Tokyo, Japan
{adachi-k,yanai}@mm.inf.uec.ac.jp

ABSTRACT

An automatic meal recording is one of typical applications of image recognition technology. In fact, some mobile apps on meal recording have been released so far. Most of the apps assume that a user takes a meal photo before start eating. However, this approach is not appropriate for the meals in which foods are served while taking meals such as food buffets, shared large plates and hot pots. In this study, we propose a mobile meal recording system that estimates food calories during eating in the real-time way by eating action recognition with RGB-D images obtained by a front-mounted depth sensor on a smartphone. In the experiments with the mobile app implemented for an iPhone, in the situation of eating grilled meat, the proposed system improved the accuracy of calorie estimation by up to 28% and recognized the correct meal category with 6.67 times higher accuracy in eating action recognition compared to the baseline system.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; • **Human-centered computing** → **Interactive systems and tools**.

KEYWORDS

eating action recognition, calorie estimation, food segmentation, RGB-D image

ACM Reference Format:
Kento Adachi Keiji Yanai. 2022. DepthGrillCam: A Mobile Application for Real-time Eating Action Recording Using RGB-D Images. In *Proceedings of the 7th International Workshop on Multimedia Assisted Dietary Management (MADiMa '22), October 10, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3552484.3555752

1 INTRODUCTION

An automatic meal recording is one of typical applications of image recognition technology. The existing food recognition applications create food records by taking a photo of a single or multiple food items before eating. However, this approach cannot be used for the meal the amount of which is unknown beforehand, for instance, food buffets, shared large plates and hot pots. For this problem, Okamoto *et al.* [11, 13] proposed a system that allows users to estimate their calorie intake in a real time way by placing a smartphone on a tabletop and using the front camera to capture images of themselves while they are eating, thereby recognizing the type of meal and the approximate amount of calories. However, this method uses a fixed amount of calories based on the food category without considering the portion size of each meal, which is far from the actual caloric intake. In this study, we aim to estimate the amount of calories in real time considering the amount of each food instance by obtaining the depth information of the meal image using the depth information that can be obtained by smartphones used for face recognition, and estimating the gram-based weight of the food. We implemented the proposed system on iPhone, the name of which is "DepthGrillCam." Figure 1 shows an example of usage of "DepthGrillCam."

Figure 1: An example of the proposed system usage.

2 RELATED WORK

Current research on calorie estimation using food images is mainly based on a combination of food category classification and volume estimation. In particular, a variety of approaches have been proposed for estimating the portion size of a meal. Okamoto *et al.* [12] calculated the actual size of the food region by calculating the actual size per pixel of the food image using reference objects with known areas such as credit cards and long wallets. Similarly, some research [1, 4] also use the reference objects in food scene such as rice grain or chopsticks. Our method also uses an approach to estimate the amount of calories based on the relationship between the surface area equivalent of the food and the amount of calories.

MADiMa '22, October 10, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9502-1/22/10…$15.00
https://doi.org/10.1145/3552484.3555752

In recent years, several methods have been proposed to estimate calories by estimating the 3D volume of food from images. Fang *et al.* [5] define the geometry of a meal and project it onto a food image to estimate the volume of the meal. Myers *et al.* [9] estimate the amount of calories by calculating the volume of a meal region from a voxel representation using depth estimation networks. Graikos *et al.* [7] similarly use depth estimation networks, but with point cloud-based volume estimation. Lu *et al.* [8] also use depth estimation networks to estimate food calories from a single image. Naritomi *et al.* [10] proposed a method for precisely estimating the volume of a food by reconstructing the three-dimensional shapes of a meal including dishes and a dish itself from a meal image. Fang *et al.* [6] and Shao *et al.* [14] estimated the caloric content by estimating the energy distribution of each pixel in a food image.

Ando *et al.* [2] proposed a method for estimating the amount of calories in a food image using the depth image obtained from the depth sensor mounted on the rear camera of a smartphone. This method calculates the meal volume by accumulating the height from the depth of the reference surface, such as a table, to the meal depth. Thames *et al.* [15] use a large dataset of 5000 meals, Nutrition5k, to estimate the weight and nutritional content of meals from images with depth. However, these methods cannot be used in the situations where the amount of food to be eaten is not determined in advance, such as food buffets, shared Chinese plates and hot pot meals. In addition, since it is not possible to monitor calorie intake in real time during a meal, it is difficult to provide feedback on eating behavior such as stopping eating according to calorie intake.

Okamoto *et al.* [11] proposed a system that enables users to check their calorie intake in real time by placing a smartphone on a table and taking a picture of themselves while eating. This method uses a pixel histogram around the tip of the chopsticks to classify the meal being grasped, and then calculates the total calorie amount by adding the calories corresponding to the food class. However, the portion sizes of individual food instances are not considered. Our method uses depth information to estimate the caloric content based on the portion size of a meal.

3 DEPTHGRILLCAM

In this study, we propose the method combining calorie estimation and eating action recognition to provide a real-time system that allows users to monitor their calorie intake using a smart phone placed on a tabletop. The overview of our method is shown in Figure 2. In this section, we describe the two parts of the proposed method, i.e., food calorie estimation and eating action recognition.

3.1 Food Calorie Estimation

In food calorie estimation, the food area is divided by food region segmentation, the food weight is estimated using image with depth information, and then converted into calorie amount by multiplying the calorie density per weight of each food category obtained in advance. We used DeepLabV3 [3] as a segmentation model. To estimate the food weight, we calculate the average depth, the number of the pixels and food category from estimated food mask and depth image. Then we estimate the food weight from these data as input using the other neural network for regression.

3.2 Eating Action Recognition

For eating action recognition, food region masks obtained by food region segmentation and mouth key-point coordinates obtained by facial key-point detection are used to estimate meal and mouth bounding boxes, respectively. Then, using these two bounding boxes and depth information, the system detects the three-dimensional overlap between the mouth and the food, determines if the food has been "eaten," and records the average amount of calories calculated up to that time. This is expected to improve the accuracy of calorie content estimation by resetting the coordinate gaps that accumulate as object tracking continues, as well as by performing multiple calorie content estimations for the same food instance from different aspects. After detecting eating action, the system returns to the segmentation of the meal region again after a certain time interval to prevent the re-detection of once-eaten meal instances.

3.3 Implementation on iPhone

In the experiments, we implemented the proposed method on the iPhone using Swift, CoreML, and Vision Framework. The iPhone X or above has a TrueDepth camera sensor for face recognition in the top bezel, and through APIs, depth data corresponding to each pixel obtained from the camera with a resolution of 600×400. We trained the food region segmentation model and weight regression model on a PC using Pytorch, converted these models to MLModel format that can be used with CoreML using CoreMLTools, and implemented them on a device using Vision Framework. During training of the segmentation model, random cropping and random horizontal flipping were applied as preprocessing to the images. The bounding box of the mouth region is calculated from the region surrounded by the six feature points corresponding to the position of the inner lips that can be obtained by the "VNDetectFaceLandmarksRequest" API, which is provided in the Vision Framework, an image recognition framework available on iOS.

3.4 User Interface

The user interface of the proposed method application is shown in Figure 3. In the application interface, the following information is displayed on the screen so that the user can know the amount of calories taken in real time.

(1) Food area bounding box
(2) Mouth area bounding box
(3) Food mask
(4) Detected food categories
(5) Estimated calorie content
(6) Total amount of calories currently taken
(7) The number of food instances currently taken

In addition, to make it easier for a user to use, a visual guide is displayed to show where the user fits in the screen. When an eating action is detected, the system notifies the user by changing the colors of the food area bounding box and mouth area bounding box displayed on the screen. By eliminating direct manipulation of the device, such as inputting numerical values and pressing buttons, users can check the amount of calories they have consumed in real time through the interface described above by simply setting the device on the table before a meal and running the application. This

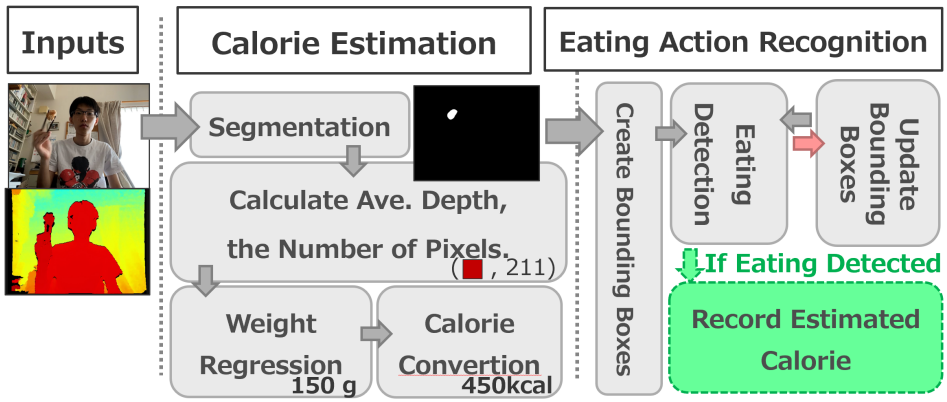

Figure 2: Overview of the proposed architecture.

Figure 3: User interface of the proposed method (1. Food area bounding box, 2. Mouth area bounding box, 3. Food mask, 4. Detected food categories, 5. Estimated calorie content, 6. Total amount of calories currently taken, 7. Number of food instances currently taken)

is expected to reduce the burden on users compared to manually inputting food records or repeatedly taking pictures of the meal each time a new meal is served.

4 EXPERIMENTS

4.1 Evaluation of Calorie Estimation

To evaluate how accurate the proposed method is in estimating caloric content, we conducted a comparison experiment with the existing work, GrillCam [11, 13]. Since GrillCam accounts for a fixed amount of calories based on the type of food eaten and does not estimate the amount of calories based on the portion size of the meal, we defined the baseline as the average amount of calories for each food category in the dataset. We used (B) Mean Absolute Error (MAE) of estimated calorie and (C) Relative Error(RE) as evaluation metrics calculated by the following equations:

$$MAE(Y, Y')_{(kcal)} = \frac{1}{N} \sum_{y \in Y, y' \in Y'} |y - y'| \quad (1)$$

$$RE(Y, Y')_{(kcal)} = \frac{1}{N} \sum_{y \in Y, y' \in Y'} \frac{|y - y'|}{y}, \quad (2)$$

where N is the number of data, Y is the set of Ground-Truth calorie amounts, and Y' is the set of estimated calorie amounts. We also show (A) the gram based MAE of the absolute weights in the result table.

The result is shown in Table 1. The proposed method achieved better accuracy than the baseline for all dietary categories except for green pepper. Especially in the category of grilled meat, the accuracy was improved by 28% compared to the baseline. On the other hand, the accuracy of the proposed method was lower than the baseline for categories with less weight variation, such as bell peppers. The reason for this is that the information obtained from the number of pixels and the average depth corresponds to the actual size of the food region as viewed from the camera. However, in the situation of a dining scene, where the orientation of a meal is more flexible than in the case of a meal placed on a flat surface, it was difficult to formulate the relationship between the surface area of the meal and the amount of calories because the surface area as viewed from the camera fluctuates greatly when the orientation of the food instance changes. As a solution to this problem, it is possible to estimate the three-dimensional shape of the meal using depth information, and then calculate the volume to estimate the caloric content, but this is a subject for future work. The experimental results show that the calorie estimation by the proposed method is especially effective for categories with a wide range of food weights.

4.2 Evaluation of Eating Action Recognition

To evaluate the performance of the proposed method for eating action recognition, we conducted a comparison experiment using an actual smartphone. We compared the behavior of both applications when pretending to eat a food sample using the proposed method implemented on an iPhone 11 and GrillCam [11] implemented on a Nexus 5 as a baseline.

The experiment was conducted according to the following procedure.

(1) A device that implements the proposed method and a device that implements the baseline are placed side by side on a

tabletop, and a food instance is prepared on a plate placed outside the view angle of the device front camera.
(2) The user raises the food instance with chopsticks and brings it to the mouth while running the both application.
(3) Wait for the application to operate and check if an eating action is detected or not.
(4) When an eating action is successfully detected, the system matches the detected food category and records the CORRECT if it is correct or the INCORRECT if it is the wrong category.
(5) If eating action is not detected within 5 seconds, FAILURE is recorded.

The above procedure is conducted 10 times for each food category per each participant, and the accuracy is calculated based on the evaluation metrics. The experiment was conducted with the participation of seven volunteers from our laboratory.

As evaluation metrics, we calculated correct ratio (CR) and success ratio (SR) as shown in the equation 3, 4.

$$correct\ ratio(CR) = \frac{CORRECT}{CORRECT + INCORRECT} \tag{3}$$

$$success\ ratio(SR) = \frac{CORRECT}{CORRECT + INCORRECT + FAILURE} \tag{4}$$

The results of the experiments are shown in Table 2 and 3.

With the proposed method, we achieved better accuracy than the baseline for all food categories in both metrics. Since SR depends on not only the performance of methodology but also the performance of device, the data is for reference. Comparing CR for the entire meal category, the proposed method can recognize the correct meal category with 6.67 times higher accuracy than the baseline.

In GrillCam [11], the pixel histogram around the chopstick tips is used to classify food. However, when eating action is detected too close to the mouth, the skin color element around the mouth may affect the histogram. In addition, the Hough transform is used to detect the chopstick in GrillCam [11], when the linear component of the background such as joint between wall and ceiling, beam, rack, window and fluorescent light existed in the vicinity of the face, the linear element of the background was mistakenly detected as chopsticks, and the feeding motion was mistakenly detected, resulting in the recording of incorrect answers in some cases.

In the proposed method, categories such as green pepper and pumpkin, where there are no similar colors in the environment, the eating action is correctly recognized in most cases. In contrast, the detection accuracy of grilled meat and carrots was lower. These foods are close in color to the skin tone, and when the food is brought close to the mouth, it is assimilated with the skin tone, which is considered to be the reason why the food could not be detected correctly. To deal with this problem, in this experiment, only RGB images were used for detecting food instances for implementation reasons. However, we believe that learning a food region segmentation model with depth information will improve the accuracy.

5 CONCLUSION

In this study, we proposed a mobile food recognition system that can check calories during a meal in real time by combining calorie estimation and eating action recognition with depth information that can be obtained by a depth sensor mounted on the front of a smartphone used for face recognition. The experiments show that in the situation of eating grilled meat, the proposed method can improve the accuracy of calorie estimation by up to 28% compared to the conventional method, GrillCam [11], and can recognize the correct meal category with 6.67 times higher accuracy in eating action recognition. In the future, we will expand the dataset and aim to improve the accuracy of calorie estimation by considering the three-dimensional shape of the food.

The prototype mobile application of "DepthGrillCam" for iPhone is available at https://apps.apple.com/us/app/depthgrillcam/id1612911438.

Table 1: Comparision of calorie estimation by proposed method and GrillCam [11]

category	ours			GrillCam [11]		
	A	B	C	A	B	C
grilled meat	**2.967**	**13.381**	**31.8**	3.214	14.495	43.9
greenpepper	2.449	1.322	16.23	**2.199**	**1.188**	**15.8**
pumpkin	**1.172**	**1.231**	**24.7**	1.217	1.278	27.8
carrot	**3.268**	**3.366**	**30.4**	3.507	3.612	38.5
rice	**3.794**	**5.918**	**22.2**	4.290	6.693	26.4
fried chicken	**4.291**	**10.684**	**20.5**	4.572	11.384	23.6

Table 2: Results of eating action recognition for each food category of the proposed method

category	data	correct	incorrect	failure	CR	SR
grilled meat	70	6	9	55	0.4000	0.0857
greenpepper	70	65	1	4	0.9848	0.9286
pumpkin	70	53	11	6	0.8281	0.7571
carrot	70	18	1	51	0.9474	0.2571
rice	70	43	1	26	0.9773	0.6143
overall	350	185	23	142	0.8894	0.5286

Table 3: Results of eating action recognition for each food category of the GrillCam [11]

category	data	correct	uncorrect	failure	CR	SR
grilled meat	70	4	13	53	0.2353	0.0571
greenpepper	70	2	18	50	0.1000	0.0286
pumpkin	70	0	3	67	0.0000	0.0000
carrot	70	2	7	61	0.2222	0.0286
rice	70	0	11	59	0.0000	0.0000
overall	350	8	52	290	0.1333	0.0229

Acknowledgment This work was supported by JSPS KAKENHI Grant Numbers, 21H05812, 22H00540, 22H00548, and 22K19808.

REFERENCES

[1] Elder Akpro Hippocrate Akpa, Hirohiko Suwa, Yutaka Arakawa, and Keiichi Yasumoto. 2017. Smartphone-Based Food Weight and Calorie Estimation Method for Effective Food Journaling. *SICE Journal of Control, Measurement, and System Integration* 10, 5 (2017), 360–369. https://doi.org/10.9746/jcmsi.10.360
[2] Yoshikazu Ando, Takumi Ege, Jaehyeong Cho, and Keiji Yanai. 2019. DepthCalorieCam: A Mobile Application for Volume-Based Food Calorie Estimation using Depth Cameras. In *Proc. of ACM Multimedia Workshop on Multimedia Assisted Dietary Management.*
[3] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. arXiv:1706.05587, 2017. Rethinking Atrous Convolution for Semantic Image Segmentation. arXiv:1706.05587 http://arxiv.org/abs/1706.05587
[4] Takumi Ege, Wataru Shimoda, and Keiji 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 the 5th International Workshop on Multimedia Assisted Dietary Management.*

[5] Shaobo Fang, Chang Liu, Fengqing Zhu, Edward J. Delp, and Carol J. Boushey. 2015. Single-View Food Portion Estimation Based on Geometric Models. In *Proc. of IEEE International Symposium on Multimedia*. 385–390. https://doi.org/10.1109/ISM.2015.67

[6] Shaobo Fang, Zeman Shao, Runyu Mao, Chichen Fu, Deborah A. Kerr, Carol J. Boushey, Edward J. Delp, and Fengqing Zhu. arXiv:1802.09670, 2018. Single-View Food Portion Estimation: Learning Image-to-Energy Mappings Using Generative Adversarial Networks. arXiv:1706.05587 http://arxiv.org/abs/1706.05587

[7] Alexandros Graikos, Vasileios Charisis, Dimitrios Iakovakis, Stelios Hadjidimitriou, and Leontios Hadjileontiadis. 2020. Single Image-Based Food Volume Estimation Using Monocular Depth-Prediction Networks. In *Proc. of International Conference on Universal Access in Human-Computer Interaction. Applications and Practice*. 532–543.

[8] Ya Lu, T. Stathopoulou, and S. Mougiakakou. 2021. Partially Supervised Multi-Task Network for Single-View Dietary Assessment. In *Proc. of International Conference on Pattern Recognition (ICPR)*. 8156–8163.

[9] Austin Myers, Nick Johnston, Vivek Rathod, Anoop Korattikara, Alex Gorban, Nathan Silberman, Sergio Guadarrama, George Papandreou, Jonathan Huang, and Kevin Murphy. 2015. Im2Calories: towards an automated mobile vision food diary. In *Proc.of IEEE International Conference on Computer Vision*.

[10] Shu Naritomi and Keiji Yanai. 2020. Hungry Networks: 3D mesh reconstruction of a dish and a plate from a single dish image for estimating food volume. In *Proc. of ACM Multimedia Asia*.

[11] Koichi Okamoto and Keiji Yanai. 2014. Realtime Eating Action Recognition System on a Smartphone. In *Proc. of IEEE International Conference on Multimedia and Expo Workshops*.

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

[13] K. Okamoto and K. Yanai. 2016. GrillCam: A Real-time Eating Action Recognition System. In *Proc. of International Conference on Multimedia Modelling (MMM)*.

[14] Zeman Shao, Shaobo Fang, Runyu Mao, Jiangpeng He, Janine Wright, Deborah Kerr, Carol Jo Boushey, and Fengqing Zhu. arXiv:2103.07562, 2021. Towards Learning Food Portion From Monocular Images With Cross-Domain Feature Adaptation. arXiv:2103.07562 [cs.CV]

[15] Quin Thames, Arjun Karpur, Wade Norris, Fangting Xia, Liviu Panait, Tobias Weyand, and Jack Sim. 2021. Nutrition5k: Towards Automatic Nutritional Understanding of Generic Food. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8903–8911.