MUSEFood: Multi-sensor-based Food Volume Estimation on Smartphones

Junyi Gao[†§‖**], Weihao Tan[†‖**], Liantao Ma[†‖], Yasha Wang[†‡§*] and Wen Tang[¶]

[†]Key Laboratory of High Confidence Software Technologies, Ministry of Education, Beijing 100871, China
[‡]National Engineering Research Center for Software Engineering, Peking University, Beijing 100871, China
[§]Peking University Information Technology Institute (Tianjin Binhai), Tianjin 300450, China
[‖]School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China
[¶]Department of Nephrology, Peking University Third Hospital, Beijing 100191, China
[**]Same contribution
[*]Corresponding to wangyasha@pku.edu.cn

Abstract—Researches have shown that diet recording can help people increase awareness of food intake and improve nutrition management, and thereby maintain a healthier life. Recently, researchers have been working on smartphone-based diet recording methods and applications that help users accomplish two tasks: record what they eat and how much they eat. Although the former task has made great progress through adopting image recognition technology, it is still a challenge to estimate the volume of foods accurately and conveniently. In this paper, we propose a novel method, named *MUSEFood*, for food volume estimation. *MUSEFood* uses the camera to capture photos of the food, but unlike existing volume measurement methods, *MUSEFood* requires neither training images with volume information nor placing a reference object of known size while taking photos. In addition, considering the impact of different containers on the contour shape of foods, *MUSEFood* uses a multi-task learning framework to improve the accuracy of food segmentation, and uses a differential model applicable for various containers to further reduce the negative impact of container differences on volume estimation accuracy. Furthermore, *MUSEFood* uses the microphone and the speaker to accurately measure the vertical distance from the camera to the food in a noisy environment, thus scaling the size of food in the image to its actual size. The experiments on real foods indicate that *MUSEFood* outperforms state-of-the-art approaches, and highly improves the speed of food volume estimation.

Index Terms—food volume estimation, diet management, image segmentation, smartphone sensing

I. INTRODUCTION

People's food intake has been proven to have a major impact on health. According to medical researches, many chronic diseases such as diabetes and kidney disease have been attributed to dietary factors [1]. Diet recording can help people to maintain a healthy diet and thus improve their health status [2].

With the popularity of smartphones, a number of apps for diet recording (e.g. MyFitnessPal[1], LoseIt[2], etc) have been developed to help users assess their food intake. A number of studies have demonstrated that smartphone-based diet record APPs can significantly increase the user's awareness of food

intake, improve their nutritional management, and thereby improve their health [3].

Smartphone-based diet recording methods and applications help users accomplish two tasks: record what they eat and how much they eat. For the former task, researchers have proposed a series of methods based on image recognition technology in recent years [4], [5], which can accurately identify the food categories and greatly simplify the manual workload of users. However, for the latter task, existing APPs do not provide much support. Users still have to estimate foods' weight or volume according to their own personal experience and then input data into the APPs manually. Studies have shown that the average of the error in the manual estimation is above 20% [6], [7], leading to the estimation of food volume becoming the bottleneck of smartphone-based dietary recording methods. To solve this problem, researchers have started developing affordable smartphone aided food volume estimation solutions in recent years [8]–[11]. However, existing works are limited in terms of user convenience and accuracy, mainly because they cannot effectively address the following two major challenges.
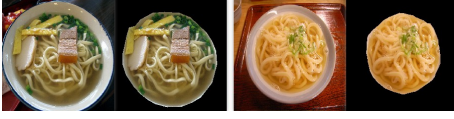
- C_1: **Converting the relative size of the food in an image to its actual size conveniently and accurately.** From the image, we can see the relative size of the food comparing to other items such as containers, chopsticks, forks, etc., which are also present in the same image. However, since the actual physical size corresponding to each pixel is unknown, the actual volume of the food cannot be obtained from the image alone. In order to solve this problem, some existing works require the user to place some reference objects of known size (e.g. credit card [10], chopsticks [8] or fingers [12]) together with the food into the same photo. However, these methods require the user to always carry the reference objects, which causes inconvenience. Some other methods use images with volume information (e.g. depth images taken by special device [13] or up to 16 images of the same food taken from different heights and angles [14]) to train their models. These methods do not require the placement of reference objects, however, they can only be used for the same type of food as the food in the training set. And

---

because of the high cost of expanding the training set, the scope of their application is quite limited.

- C_2**: Segmenting the food in the image from the background accurately and efficiently.** The segmentation of food from the background is an important basis and prerequisite for constructing a food volume model. The accuracy and efficiency of segmentation has a critical impact on the overall accuracy and efficiency of food volume estimation. Recent researchers combine *Convolutional Neural Networks*(CNN) and *Grabcut* [9], [15] or use point matching [10] to segmented food items. In order to achieve better segmentation accuracy, these methods have strict requirements on image capture methods and image quality (e.g. good camera exposure and focus, strong, variable texture in the scene and limited secular reflections and even specified shooting angles), which pose significant inconvenience and robustness issues in actual use [10].

(a) The foods contained in round bowls are usually rounded at the edges

(b) The shapes of edges of the foods contained in plates may vary greatly

Fig. 1. The shapes of food contours are often related with the shapes of containers

With the objective of taking full consideration of the above mentioned challenges, we have the following observations: 1). The existing methods mainly use cameras. However, other sensors are generally installed on smartphones, and the use of multiple sensors may improve the accuracy of the food volume estimation. 2). Compared with the conventional computer vision-based methods, using deep learning-based methods to directly segment food items will greatly improve the accuracy and speed of segmentation. Furthermore, we notice that the shapes of food contours are often related with the shapes of containers. Figure 1 shows an example. By utilizing the relationship between the shapes of food contours and the categories of the food containers, the accuracy of food segmentation can be improved.

Based on the above observations, we propose a **Mu**lti-**Se**nsor-based **Food** volume estimation method on smartphones, *MUSEFood*. *MUSEFood* collects data from multiple sensors on smartphones, processes and calculates sensing data, and aggregates the processing results to calculate the actual food volume. Our main contributions are summarized as follows:

- We propose a method for food volume estimation, *MUSEFood*, which involves three stages: sensing, data processing

and data aggregation. To the best of our knowledge, our method is the first to require neither training images with volume information nor setting a reference object of known size while taking photos of foods.

- We introduce and further extend two technical mechanisms to calculate food volume more accurately, efficiently and versatilely. Specifically, 1).We utilize echo ranging method to accurately measure the vertical distance between the camera and the food by using the speaker to emit a *Maximum Length Sequence* sound wave and using the microphone to sense the reflected wave. Based on this distance, we accurately calculate the actual volume of food without using an object of known size as a reference (i.e., addressing C_1). 2).We adopt a multi-task learning framework based on *Fully Convolutional Networks* to segement food items. The proposed deep learning framework makes full use of the relationship between the shapes of food contours and the shapes of containers, resulting in accurate and efficient segmentation. This framework helps to achieve more accurate food volume modeling (i.e., addressing C_2).

- We use real foods to verify and evaluate *MUSEFood*. Compared with the state-of-the-art models, *MUSEFood* is more accurate and faster in estimating the food volume with no need of using reference objects. Besides, *MUSEFood* is applicable for both bowls and plates and also for different foods with irregular shapes, which is universal for more situations.

- Along with this paper, we release the food dataset, SUEC Food[3], which contains 600 segmented food images annotated manually and 31395 images annotated using *Grabcut* method. All the images are from the UEC Food-256 dataset [16]. The foods in SUEC Food are mainly Asian, and the containers of foods are varied(e.g. plates, bowls and cups). Prior to this work, there is only one public food image dataset with segmentation labels, which only contains western food(e.g. pizzas and burgers) served on plates [5].

II. RELATED WORK

There have been several recent attempts to automatically estimate food volume using smartphones. To achieve this, the proposed systems have to involve three modules: food item segmentation, actual size scaling and food modeling .

Food Item Segmentation: The objective of the food item segmentation step is to detect the exact location of the food items in the images. Researchers used to use traditional computer vision methods such as *Grabcut* [17] and *Hough Transform* [18] to segment food items. Recently, researchers start to combine deep-based methods and conventional computer vision systems to solve this problem. Some researchers [9], [15] use CNN to draw bounding boxes around food items, and use *Grabcut* method to segment food items from bounding boxes. Another researchers [10], [19] use point matching methods to locate food items in images, which require two images taken

---

[3]SUEC Food dataset and the source code in this paper are available at https://github.com/MUSEFood/MUSEFood

from different angles to match similar objects, and then get segmented food items. However, these methods are not robust and convenient enough due to the requirement of shooting angle and the background of images. Most importantly, for both methods, the accuracy of the segmentation still needs to be improved, especially for foods with complex and irregular shapes. Besides, these algorithms require a large amount of computation, which will take a lot of time and affect the user experience in practice.

Actual Size Scaling: The actual size scaling module is used to obtain the actual size of each pixel of the image in order to estimate the actual size of the food in the image. Many researchers choose to place a reference object of known size next to the food and compare the relative size of the food to the reference object in the image to calculate the actual volume of food. There are various reference objects used to estimate food volume, such as chessboard-like marker [20], credit card [15], chopsticks [8], finger [21] and so on. However, different kinds of reference objects bring different problems. For reference objects such as credit card, users need to carry them at any time, which is inconvenient in actual use. Though, indeed, plates, chopsticks and fingers are reference objects that do not have to be carried by users. However, we can't guarantee that everyone uses the same length of chopsticks or have the same size of fingers, which may cause huge errors. Other researchers use training images with volume information (e.g. depth map [13], [14] or images of the same food taken from from different heights and angles [14]) to avoid using reference objects. These methods require special devices to collect food data and build their own dataset, and the foods must be placed on plates, so that they are greatly limited in application.

Food Modeling: Food modeling is the module that builds a 3D model of food based on 2D images. The volume of food can be estimated using this 3D model. In some early food modeling methods [21], food items are modeled as regular geometries. Since the shapes of foods are not always regular, these methods may cause huge errors. Some researches use deep neural networks to reconstruct depth maps of foods [13]. However, these methods require special training images as mentioned above. Recently, some researchers use stereo matching based modeling methods to build the point cloud of foods [10], [22]. Other researchers build different food models based on different containers [8]. These methods require users to take 2-3 photos for each food. In this paper, we mainly focus on the improvements for food item segmentation and actual size scaling. For food modeling, based on the previous researches of 3D reconstruction, we propose a simplified method to quickly build food models using two food images, which is applicable to different containers and different foods with irregular shapes.

III. METHODS

A. Overview

In *MUSEFood*, the whole task of food volume estimation is divided into three steps: Sensing, Data Processing and Data Aggregation. First, the user uses smartphone camera to take

food photos. The speaker of the smartphone emits *Maximum Length Sequence* of a specific length while the user is taking photos. Then the microphone receives the echo of the signal. In the second step, Data Processing, the food in the image is segmented from background. Meanwhile, we analyze the received echo signal to calculates the vertical distance from the smartphone lens to the food. In the third step, Data Aggregation, we combine the processed data from images and sound waves to build food model and finally estimate food volume.

B. Sensing

Different from other food volume estimation methods [8]–[11] which only use the camera to collect data, *MUSEFood* uses multiple sensors on smartphone including camera and microphone. This section is divided into two steps: Food Image Sensing and Audio Sensing.

1) Food Image Sensing: In order to accurately calculate the volume of the food, the user is required to take two food photos. First, the user needs to take a top-down photo of the food while ensuring that the phone and the surface of the food container are parallel to each other. The user then needs to take a side photo of the food while ensuring that the phone is perpendicular to the surface of the container.

2) Audio Sensing: We aim to develop a method that uses the distance from the smartphone to desktop as the reference instead of placing an object of known size. A non-invasive method for distance measurement is echo ranging. We use the Maximum Length Sequence (MLS) as the audio signal for ranging.

The MLS measurement technique is widely used in the field of acoustics for measuring concert hall acoustics [23] as well as loud-speaker transfer functions. The MLS measurement is particularly attractive because of the higher computational efficiency of processing and distortion and noise immunity compared with other audio signals such as chirp signal and sine signal [24].
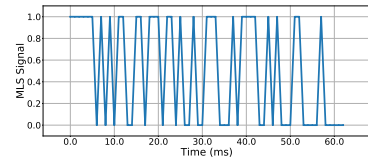

Fig. 2. MLS signal

The MLS is a pseudorandom binary sequence (PRBS) composed of a sequence of 1 and 0 (Fig. 2), generated recursively using a series of digital shift registers with selected XOR feedback. The length L_s of the MLS given by $L_s = 2^n - 1$ with n denoting the order of the sequence and also the number of digital shift registers.

While the user is taking a top-down photo, the microphone sensor is activated to start recording. Then the phone speaker emits a specific length L_s of MLS audio signal s_o. When the user finishes photoing, the microphone stops recording.

This recorded audio signal s_r contains the original MLS signal played by the speaker and the audio signal reflected by obstacles (i.e. body, desktop and wall). These two signals will be used to calculate the vertical distance from the smartphone lens to the desktop while taking the photo.

C. Data Processing

In this section, we will process the collected sensing data and extract information from raw data to calculate the food volume. It is divided into two steps: processing the audio signal to obtain the vertical distance from the phone to the desktop, and segmenting the images to obtain segmented food items in the images.
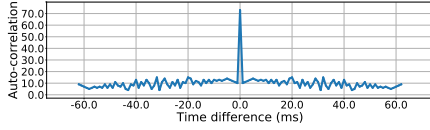
Fig. 3. Auto-correlation of MLS

1) Audio Signal Processing For Distance Measurement: As mentioned above, MLS is essentially pseudo-random binary sequences which yield a unit impulse upon circular autocorrelation (Fig. 3). Its periodic autocorrelation function $\phi_{nn}(l)$ is the two valued Kronecker function $\delta(l)$:

$$\phi_{nn}(l) = \frac{L+1}{L}\delta(l) - \frac{1}{L} \quad (1)$$

with

$$\delta(l) = \begin{cases} 1, & \text{for } l = 0 \\ 0, & \text{for } l \neq 0 \end{cases} \quad (2)$$

where l is calculated modulo L_s. A longer MLS sequence produces a smaller error in the autocorrelation function.

The time when the microphone receives the MLS signals can be retrieved from the cross-correlation function between the recorded signal s_r and the original signal s_o.
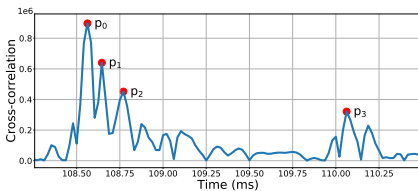
Fig. 4. Cross-correlation of the recorded signal and the original MLS signal

Since the recorded audio signal s_r contains the original MLS signal and the audio signal reflected by obstacles, there are multiple peaks in the cross-correlation results. The highest peak represents the recorded original MLS signal, and each of the remaining peaks represents an echo that is reflected by an obstacle. As shown in Figure 4, p_0 represents the recorded original MLS signal, and p_3 represents the MLS signal reflected by desktop. p_1 and p_2 represents the signal reflected by other closer obstacles such as the user's body. t_0, t_1, t_2, t_3 are the times corresponding to the peaks, which

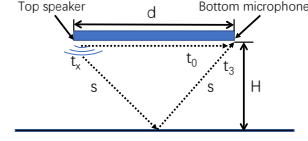indicate the moments when the microphone recorded the reflected echos.

Fig. 5. Echo ranging on the smartphone using MLS

As shown in Figure 5, the vertical distance of the phone from the desktop while shooting can be caluculated as:

$$\begin{cases} (t_0 - t_x)v & = d \\ (t_3 - t_x)v & = 2s \\ H^2 + (\frac{d}{2})^2 & = s^2 \end{cases} \rightarrow H = \frac{\sqrt{(vt_3 - vt_0 + d)^2 - d^2}}{2} \quad (3)$$

where t_x is the moment when the speaker emits the MLS signal, $2s$ is the distance traveled by echo p_3, d is the distance from the phone speaker to the microphone, v is the sonic speed and H is the distance from the phone to the desktop to be measured.

2) Food Item Segmentation: Fully Convolutional Networks (FCN) [25] is an end-to-end system with input as a image, and output as pixel-wise predictions for the image. However, to the best of our knowledge, it has not been used in food volume estimation so far. In this work, we use FCN for food image segmentation instead of the CNN and *Grabcut* methods used in traditional food volume estimation methods [9], [15].

First, we use deep convolutional networks to extract convolution features of the image, and then input the extracted features into FCN for segmentation. Considering that the appearance of food in different containers may vary, sharing shape information of containers (e.g. bowls or plates) during the segmentation process may increase the accuracy of segmentation. To this end, we propose to combine the food segmentation task with the container classification task to jointly train the FCN.

We formulate food item segmentation and container classification as a multi-task deep learning problem and modify the architecture of the FCN for our purpose. Both tasks influence each other through updating the shared intermediate layers. The modification is not straightforward for involvement of a design issue, which is about the degree in which the intermediate layers should be shared. Ideally, each task should have its own private layer(s) given that the nature of both tasks, binary classification versus segmentation, is different. In such a way, the updating of parameters can be done more freely for optimization of individual performance.

Inspired by [4], we derive three different deep architectures as depicted in Figure 6, respectively name as MFCN-A to MFCN-C.

- MFCN-A allows classification network and FCN to privately own their intermediate layers on top of the convolutional layers for parameter learning.
- MFCN-B considers stacked architecture by adding the intermediate layers for container classification before
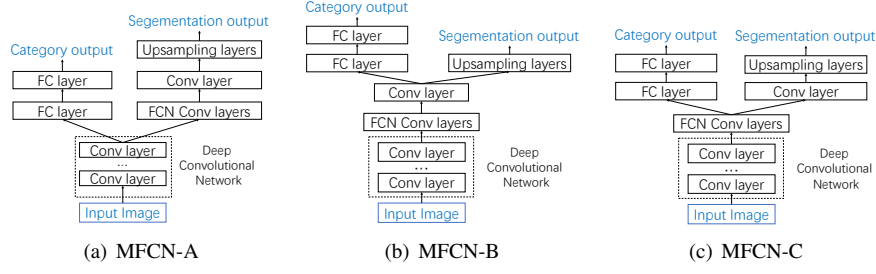
Fig. 6. Three different deep architectures for multi-task learning of container classification and image segmentation

upsampling the FCN extracted convolution features. In this setting, both tasks share all the convolution layers of FCN. Both MFCN-A and MFCN-B designs are relatively straightforward to implement by adding additional layers to FCN.

- MFCN-C considers the decoupling of some intermediate layers. It is a compromise version between the first and second architectures, by having shared convolution layers and private layers for each task.

We adopt two cross-entropy functions L_1, L_2 as the loss functions for both tasks. Denote N as the total number of training images, the overall loss function L is as following:

$$L = -\frac{1}{N} \sum_{n=1}^{N} (L_1 + \lambda L_2) \quad (4)$$

where λ is a hyper-parameter trading off the loss terms.

D. Data Aggregation

In this section, we build the food model and calculate food volume using segmented food items and the distance from the smartphone to the table obtained in the former section.

1) Actual Size Scaling: To get the actual size of the food in the image, we need to know the actual size corresponding to each pixel in the image. As shown in Figure 7, according to the principle of scaling, the actual width of the image photoed at the vertical height H can be calculated as: $M = \frac{mH}{f}$, where H is the vertical distance from smartphone to table calculated in the former step, f is the focal length of the smartphone lens, m is the width of the smartphone image sensor(i.e. CCD or CMOS). For a certain smartphone, both f and m are fixed values.
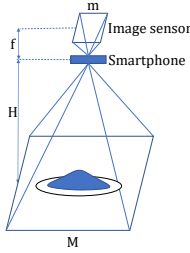

Fig. 7. Calculating the actual size of the image

Knowing the actual width of the image, we can easily get the actual size corresponding to each pixel in the image. Then the actual size of segmented food items in the top view image can be calculated by simply calculate how many pixels the food item has.
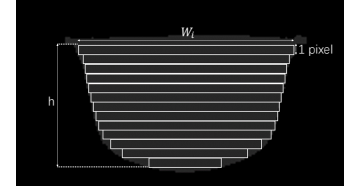

Fig. 8. Dividing food into columns with a height of 1 pixel

2) Food Modeling And Volume Calculating: After obtaining the actual food size in the top view image, we calculate food volume using the side view of the food based on a differential model. After segmenting the side image as well, We can obtain the side shape of the food (e.g. height and shape outlines) from the segmented food item. We assume that the food is uniformly distributed in the container (when the container is a bowl or a cup, the height of the food equals to the height of the container). As shown in Figure 8, the food model can be regarded as the accumulation of columns with a height of 1 pixel, and the cross-sectional shape of each column is the shape of the food in the top view image. Assuming that the width of the segmented food item in the top view is w_f, the calculated actual size of food items in the top view is S_{Real}, the number of columns is h (i.e. the height of the segmented food item in the side view image), and the width of the cross-section of the column i is w_i.

The food volume V can be calculated as:

$$V = \sum_{i=1}^{h} S_{Real} (\frac{w_i}{w_f})^2 \quad (5)$$

IV. EVALUATION

We split the experiments into three parts. The first part aims to evaluate the accuracy of the MLS ranging method at different distances and the robustness under varying degrees of noise. The second part aims to evaluate different deep architectures of multi-task learning in comparison to single-task FCN and conventional CNN and *Grabcut* method. The last part is designed to evaluate the final food volume estimation performance.

TABLE I
THE MLS RANGING ACCURACY AT DIFFERENT DISTANCES

	Distance				
	10cm	**20cm**	**30cm**	**40cm**	**50cm**
Measurement (cm)	10.08 ± 0.00	19.86 ± 0.09	29.74 ± 0.12	39.83 ± 0.00	49.68 ± 0.08
Relative error	$(0.80 \pm 0.00)\%$	$(0.70 \pm 0.49)\%$	$(0.88 \pm 0.41)\%$	$(0.42 \pm 0.00)\%$	$(0.64 \pm 0.15)\%$

TABLE II
THE LONGEST RETRY TIME AT DIFFERENT NOISE LEVEL

	Dining room (23db)	**Dining room (38db)**	**Restaurant (50db)**	**Cafeteria (68db)**
The longest retry time (s)	0.0253 ± 0.0038	0.0265 ± 0.0021	0.0537 ± 0.0343	0.1048 ± 0.0524

A. MLS ranging

The MLS we use has a length of 1023 samples (order $n = 10$) and is generated using a sample rate of 48kHz. The 48kHz clock frequency is limited by the data acquisition hardware, and the length of the sequence is a compromise between resolution and the sequence period (approximately 20ms). The smartphone we use is an iPhone 6 Plus. Due to the hardware limitations of the phone speaker, playing with the bottom speaker can cause severe signal distortion, so the MLS signal must be played using the top speaker. We use the bottom microphone of the phone to record.

1) Ranging accuracy: In order to evaluate the accuracy of the MLS ranging method, we perform measurements between 10cm and 50cm. This is also the possible distance range between the phone and the desktop while the user is taking photos in actual use. We repeat the measurement 5 times at each distance. Table I shows the results of our measurements using the MLS method at different distances. The results show that the MLS ranging method achieves high accuracy within the range of 10cm to 50cm. In addition, the distance itself has little effect on the accuracy of the MLS method.

2) Ranging robustness: Environmental noise can have an impact on sound related methods. In order to evaluate the effect of different degrees of noise on the MLS method, we test the longest retry time of the MLS method ranging in environments of different level of noise. When the MLS signal is disturbed by noise, the ranging result value is usually far from the normal value. We repeat the MLS signal every 50ms until the correct distance is obtained, and we define the time spent in this process as the longest retry time. We test in four different situations, dining room (23db), dining room (38db), restaurant (50db) and cafeteria (68db). At each situation, we repeat the test 5 times.

Table II shows the longest retry time of the MLS method at different noise level. The results show that although noise has a certain impact on the MLS method, this method still shows high robustness. In a quiet room, the MLS method can get an accurate result with just one measurement. In a relatively noisy environment, the MLS method may need to measure 2-3 times to get results. But the time spent on this process is very short and does not affect user experience. Furthermore, we find that in general situations, ranging results will not be affected by other objects on the desktop. However, if there are large objects on the table that are very close to the food, they may affect ranging results.

B. Food image segmentation

1) Dataset: We use our SUEC Food dataset as training images for food image segmentation. The images in SUEC Food are from the UEC Food-256 dataset [16], which contains 256-kind food images with only food category labels and bounding boxes. Most of the food categories in UEC Food-256 are popular foods in Japan and other Asian countries. FCN requires training images with segmentation labels. In order to obtain the segmentation labels for UEC Food-256 images, we first segment the images in bounding boxes using the *Grabcut* method. Then we manually segment 500 food images. We pre-train our models with *Grabcut* generated segmentation images, then fine-tune the models with manually segmented images. In addition, we manually segmented another 100 food images as the test dataset. We have published this dataset, SUEC Food, including 31395 segmented images annotated by *Grabcut* method and 600 segmented images annotated manually. For the container classification task, there are many types of containers. In this work, we divide the food containers into plates and bowls and label all the images manually.

2) Experiment Setup: Several state-of-the-art food volume estimation methods (e.g. [9], [15]) use Fast R-CNN and *Grabcut* to segment food images, which are taken as our baseline approaches. The *Grabcut* method requires repeated iterations for each image to get the segmentation results. The more iterations, the longer it takes, and the more accurate the segmentation result. The iterations of *Grabcut* are set to 1,2,3 and 5. In addition, in order to evaluate the multi-task models we proposed, we also use the original single-task FCN as one of our baselines.

We use VGG-16 [26] as the architecture of the deep convolutional neural networks. In the proposed MFCN, the hidden dimensions of the fully connect layers for container classification are set to 128 and 2. The convolutional layer before upsampling layers has 151 filters, and the kernel size is set to 1. The hyper-parameter λ is set to 1. The learning rate is set to $1e - 5$ with Adam optimizer.

3) Evaluation Metric: All the deep convolutional neural networks we use are pre-trained with 2000 categories in

ImageNet [27], including 1000 food-related categories. The networks are then trained with training images in SUEC Food. 90% of the images are randomly picked for training, and 10% for validation. We finally test all the models on the 100 test images with manually annotated ground truth labels.

Similar to other researches of image segmentation, we assess segmentation performance using the mean Intersection over Union ($mIoU$)

Besides, in order to evaluate the speed improvement brought by FCN, we segment 100 images using each method and record their time consuming.

TABLE III
RESULTS OF FOOD IMAGE SEGMENTATION ON TEST DATASET

	Method	mIoU	Time (s)
Baseline	**Grabcut@1**	0.7428	27.14
	Grabcut@2	0.7665	36.56
	Grabcut@3	0.7720	39.38
	Grabcut@5	0.7730	54.27
	FCN	0.9143	11.17
Our approach	**MFCN-A**	0.9160	**10.25**
	MFCN-B	**0.9210**	12.62
	MFCN-C	0.9166	13.30

4) Performance: Table III lists the results for food image segmentation. The $mIoU$ of FCN significantly outperformed all the *Grabcut*-based methods. And the $mIoU$ of all three structures of MFCN are higher than the single-task FCN, which confirms our observation that the shapes of food contours are related with the shapes of containers. By sharing this information with food segmentation using multi-task learning, we can obtain a more accurate segmentation model. Specifically, among the three structures of MFCN, the MFCN-B structure performs best, followed by MFCN-C. The performance of MFCN-A is slightly lower than the other two structures, but still better than the single-task FCN. This is because there may be a close relationship between the shapes of food contours and the shapes of containers, and a large number of identical features are shared between the two tasks. Therefore, the more convolutional layers shared by the two tasks, the more accurate the final segmentation result. Besides, for the FCN and MFCN models, it takes only about 10 seconds to segment 100 images, while the fastest *Grabcut*

method (*Grabcut@1*) takes 27 seconds. Different multi-task learning structures have no significant differences in time. The accuracy of the *Grabcut* method is related to the time it takes. However, even if it takes more than 50 seconds, the accuracy of the *Grabcut* method is only about 77%. Thus, using FCN-based methods will significantly improve the efficiency of the whole food volume estimation process.

C. Food Volume Estimation

We selected three foods for volume estimation, chicken drumstick(Fig. 9(a)), fried pork(Fig. 9(b)) and congee(Fig. 9(c)), which are common home cooking. Chicken drumstick and fried pork are on the plates, and congee is in the bowl. The shapes of these foods vary, where chicken drumstick has a more regular shape and fried pork has a complex stacked shape. We use the water displacement method to get the actual volume of all foods.

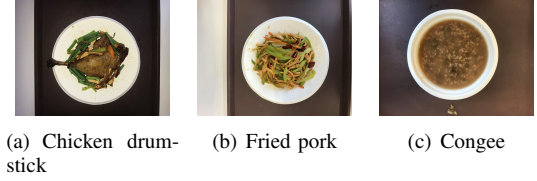(a) Chicken drumstick   (b) Fried pork   (c) Congee

Fig. 9. Three foods for volume estimation testing

We select three state-of-the-art reference objects-based food volume estimation models for comparison(i.e, Okamoto et al. 2016 [15], Akpa et al. 2017 [8], Pouladzadeh et al. 2016 [28]). For those methods which use chopsticks or fingers as reference objects, the sizes of chopsticks or fingers are usually very different. To be fair, we conduct two experiments using these methods. First, we use chopsticks or fingers of known length as the reference object and then use chopsticks or fingers of average length as the reference object (i.e. the average length of 10 different sizes of chopsticks or fingers). In addition, we also asked 10 people to visually estimate the food volume. We use the average of their estimation as one of our baselines. Table IV lists the results of final food volume estimation.

The results show that *MUSEFood* outperforms all the state-of-the-art methods for foods with different shapes in different containers. For foods with regular shapes and foods in bowls,

TABLE IV
RESULTS OF FOOD VOLUME ESTIMATION

Container shape	Plate				Bowl	
Food	**Chicken drumstick**		**Fried pork**		**Congee**	
	Estimation(ml)	**Relative Error(%)**	**Estimation(ml)**	**Relative Error(%)**	**Estimation(ml)**	**Relative Error(%)**
Real value	285.5	0	372.0	0	334.0	0
Eye-measurement	246.0	-13.84	307.5	-17.34	410.4	22.87
Prepaid card	151.7	-46.85	247.8	-33.40	213.8	-35.98
Chopsticks(Known size)	313.5	9.80	446.9	20.13	356.1	6.62
Chopsticks(Average size)	359.9	26.04	513.0	37.90	408.8	22.40
Finger(Known size)	536.9	88.05	598.3	60.83	498.4	49.22
Finger(Average size)	671.9	135.06	747.9	101.04	623.0	86.52
Our approach	**293.2**	**2.70**	**418.0**	**12.37**	**333.1**	**-0.27**

the error of *MUSEFood* is very low. For foods with irregular shapes, the error of *MUSEFood* is slightly larger, but the accuracy is still far superior to all the baseline methods.

V. CONCLUSION AND DISCUSSION

In this paper, we have introduced *MUSEFood* to calculate food volume using data collected from multiple sensors on smartphones. *MUSEFood* uses FCN and utilizes shape information of food containers through multi-task learning structures, resulting in more accurate and fast food image segmentation. We use the MLS ranging instead of using reference objects, which improves the convenience of use and achieves higher food volume estimation accuracy.

MUSEFood shows sufficient robustness and versatility in our experiments. The shape of the different food containers does not introduce errors in the results of the food volume estimation. *MUSEFood* can handle food without regular shapes, which makes our models applicable to more types of foods. Though some smartphones have ToF cameras, which can directly obtain the target distance, smartphones with ToF cameras are not very popular. Requiring user to buy the special equipment will increase the cost inevitably. Furthermore, the develop interfaces of these ToF cameras are hardly public to developers. After obtaining the food volume, we can conveniently estimate the detailed calories and nutrients intake using food nutrients database. We will work on this function in the future.

REFERENCES

[1] J. Who and F. E. Consultation, "Diet, nutrition and the prevention of chronic diseases," *World Health Organ Tech Rep Ser*, vol. 916, no. i-viii, 2003.

[2] R. Ortega, A. Requejo, A. Lopez-Sobaler, M. Quintas, P. Andres, M. Redondo, B. Navia, M. Lopez-Bonilla, and T. Rivas, "Difference in the breakfast habits of overweight/obese and normal weight schoolchildren." *International journal for vitamin and nutrition research. Internationale Zeitschrift fur Vitamin-und Ernahrungsforschung. Journal international de vitaminologie et de nutrition*, vol. 68, no. 2, pp. 125–132, 1998.

[3] S. S. Coughlin, M. Whitehead, J. Q. Sheats, J. Mastromonico, D. Hardy, and S. A. Smith, "Smartphone applications for promoting healthy diet and nutrition: a literature review," *Jacobs journal of food and nutrition*, vol. 2, no. 3, p. 021, 2015.

[4] J. Chen and C.-W. Ngo, "Deep-based ingredient recognition for cooking recipe retrieval," in *Proceedings of the 24th ACM international conference on Multimedia*. ACM, 2016, pp. 32–41.

[5] G. Ciocca, P. Napoletano, and R. Schettini, "Food recognition: a new dataset, experiments, and results," *IEEE journal of biomedical and health informatics*, vol. 21, no. 3, pp. 588–598, 2017.

[6] S. L. Godwin, E. Chambers Iv, and L. Cleveland, "Accuracy of reporting dietary intake using various portion-size aids in-person and via telephone," *Journal of the American Dietetic Association*, vol. 104, no. 4, pp. 585–594, 2004.

[7] T. Hernández, L. Wilder, D. Kuehn, K. Rubotzky, P. Moser-Veillon, S. Godwin, C. Thompson, and C. Wang, "Portion size estimation and expectation of accuracy," *Journal of food composition and analysis*, vol. 19, pp. S14–S21, 2006.

[8] E. A. H. Akpa, H. Suwa, Y. Arakawa, and K. Yasumoto, "Smartphone-based food weight and calorie estimation method for effective food journaling," *SICE Journal of Control, Measurement, and System Integration*, vol. 10, no. 5, pp. 360–369, 2017.

[9] Y. Liang and J. Li, "Deep learning-based food calorie estimation method in dietary assessment," *arXiv preprint arXiv:1706.04062*, 2017.

[10] J. Dehais, M. Anthimopoulos, S. Shevchik, and S. Mougiakakou, "Two-view 3d reconstruction for food volume estimation," *IEEE transactions on multimedia*, vol. 19, no. 5, pp. 1090–1099, 2017.

[11] J. Chung, J. Chung, W. Oh, Y. Yoo, W. G. Lee, and H. Bang, "A glasses-type wearable device for monitoring the patterns of food intake and facial activity," *Scientific reports*, vol. 7, p. 41690, 2017.

[12] P. Pouladzadeh, A. Yassine, and S. Shirmohammadi, "Foodd: food detection dataset for calorie measurement using food images," in *International Conference on Image Analysis and Processing*. Springer, 2015, pp. 441–448.

[13] A. Meyers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and K. P. Murphy, "Im2calories: towards an automated mobile vision food diary," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1233–1241.

[14] D. Allegra, M. Anthimopoulos, J. Dehais, Y. Lu, F. Stanco, G. M. Farinella, and S. Mougiakakou, "A multimedia database for automatic meal assessment systems," in *International Conference on Image Analysis and Processing*. Springer, 2017, pp. 471–478.

[15] K. Okamoto and K. Yanai, "An automatic calorie estimation system of food images on a smartphone," in *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*. ACM, 2016, pp. 63–70.

[16] Y. Kawano and K. Yanai, "Automatic expansion of a food image dataset leveraging existing categories with domain adaptation," in *European Conference on Computer Vision*. Springer, 2014, pp. 3–17.

[17] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," in *ACM transactions on graphics (TOG)*, vol. 23, no. 3. ACM, 2004, pp. 309–314.

[18] R. O. Duda and P. E. Hart, "Use of the hough transformation to detect lines and curves in pictures," SRI INTERNATIONAL MENLO PARK CA ARTIFICIAL INTELLIGENCE CENTER, Tech. Rep., 1971.

[19] M. H. Rahman, Q. Li, M. Pickering, M. Frater, D. Kerr, C. Bouchey, and E. Delp, "Food volume estimation in a mobile phone based dietary assessment system," in *2012 Eighth International Conference on Signal Image Technology and Internet Based Systems*. IEEE, 2012, pp. 988–995.

[20] I. Woo, K. Otsmo, S. Kim, D. S. Ebert, E. J. Delp, and C. J. Boushey, "Automatic portion estimation and visual refinement in mobile dietary assessment," in *Computational Imaging VIII*, vol. 7533. International Society for Optics and Photonics, 2010, p. 75330O.

[21] R. Almaghrabi, G. Villalobos, P. Pouladzadeh, and S. Shirmohammadi, "A novel method for measuring nutrition intake based on food image," in *2012 IEEE International Instrumentation and Measurement Technology Conference Proceedings*. IEEE, 2012, pp. 366–370.

[22] H. Hassannejad, G. Matrella, P. Ciampolini, I. Munari, M. Mordonini, and S. Cagnoni, "A new approach to image-based estimation of food volume," *Algorithms*, vol. 10, no. 2, p. 66, 2017.

[23] Y. Ando, *Concert hall acoustics*. Springer Science & Business Media, 2012, vol. 17.

[24] C. Dunn and M. J. Hawksford, "Distortion immunity of mls-derived impulse response measurements," *Journal of the Audio Engineering Society*, vol. 41, no. 5, pp. 314–335, 1993.

[25] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," 2009.

[28] P. Pouladzadeh, P. Kuhad, S. V. B. Peddi, A. Yassine, and S. Shirmohammadi, "Food calorie measurement using deep learning neural network," in *2016 IEEE International Instrumentation and Measurement Technology Conference Proceedings*. IEEE, 2016, pp. 1–6.