electronics

Article

Automatic Estimation of Food Intake Amount Using Visual and Ultrasonic Signals

Ki-Seung Lee

Department of Electrical and Electronics Engineering, Konkuk University, 120 Neungdong-ro, Gwangjin-gu, Seoul 05029, Korea; kseung@konkuk.ac.kr; Tel.: +82-02-450-3489

Abstract: The continuous monitoring and recording of food intake amount without user intervention is very useful in the prevention of obesity and metabolic diseases. I adopted a technique that automatically recognizes food intake amount by combining the identification of food types through image recognition and a technique that uses acoustic modality to recognize chewing events. The accuracy of using audio signal to detect eating activity is seriously degraded in a noisy environment. To alleviate this problem, contact sensing methods have conventionally been adopted, wherein sensors are attached to the face or neck region to reduce external noise. Such sensing methods, however, cause dermatological discomfort and a feeling of cosmetic unnaturalness for most users. In this study, a noise-robust and non-contact sensing method was employed, wherein ultrasonic Doppler shifts were used to detect chewing events. The experimental results showed that the mean absolute percentage errors (MAPEs) of an ultrasonic-based method were comparable with those of the audio-based method (15.3 vs. 14.6) when 30 food items were used for experiments. The food intake amounts were estimated for eight subjects in several noisy environments (cafeterias, restaurants, and home dining rooms). For all subjects, the estimation accuracy of the ultrasonic method was not degraded (the average MAPE was 15.02) even under noisy conditions. These results show that the proposed method has the potential to replace the manual logging method.

Keywords: food intake estimation; chewing sound detection; food image recognition

Citation: Lee, K.-S. Automatic Estimation of Food Intake Amount Using Visual and Ultrasonic Signals. *Electronics* **2021**, *10*, 2153. https://doi.org/10.3390/electronics10172153

Academic Editors: Alfredo Arcos Jiménez, Fausto Pedro García Márquez and Caroline Leonore König

Received: 29 July 2021
Accepted: 31 August 2021
Published: 3 September 2021

1. Introduction

Increases in the prevalence of obesity and associated metabolic diseases is undesirable in terms of a deterioration of individual life quality and an increase in social health care costs. The development of an effective method for the monitoring of caloric intake is very important in order to reduce the prevalence of people being overweight and/or obese [1]. The "manual logging" method, where the type and amount of foods are recorded by users themselves, is known as a straightforward way of monitoring the intake of calories [2,3]. At present, there are many kinds of manual logging methods available in the form of smartphone apps. Since it is cumbersome for most users to record the type of food each mealtime, a method using a camera mounted on a smartphone to automatically recognize types of food has also been developed [4]. Although high accuracy in the recognition of food type helps to partially reduce user discomfort, there is a problem with the lack of reliability due to human errors that include user laziness and erroneous record-keeping. Such problems are mainly due to the fact that some user intervention is necessary for the manual-logging methods.

Recent advances in sensor technology, signal processing techniques, detection, and recognition algorithms make it possible to continuously monitor the intake of calories with minimal user intervention [5–13]. Visual cues have been used for the classification of food types and the estimation of food amount [14–20]. A classical pattern recognition approach has also been adopted in visual-based food type recognition methods, where segmentation, feature extraction, and classification were sequentially carried out on the incoming images. Convolutional neural networks (CNN) have successfully been applied to various pattern recognition problems, and a CNN-based food recognition approach has also been proposed [18]. In this approach, the relationship between the food image and its food type is explained by a nonlinear mapping rule that is constructed using a supervised learning algorithm. A higher level of recognition accuracy (typically, more than 95% [17,18]) has been achieved via the visual modality, by comparison with the other modalities (e.g., audio [8] and ultrasound [21]). It is known that the accuracy of the supervised learning-based classifier is more or less affected by the size of the training dataset. Hence, the performance would further be improved by using several food image datasets for food recognition, such as UEC-FOOD100 [22], FOOD-101 [23], VIREO Food-172 [24], and other food image datasets [25,26]. The visual information has also been very useful in estimating the amount of food for the measurement of caloric intake. In this approach, the problem of measuring the amount of food is formulated as the volume estimation of the food via the given food image. The single-view approach [15–19] is attractive in that it can easily be applied to existing mobile devices (e.g., smartphones with a single camera [16]) without hardware modification. However, volume estimation using only a single view is very challenging, and has resulted in inaccurate caloric estimates. Accordingly, the multi-view approaches [20,27] that typically employ a stereo camera have generally been used for the volume estimation of food. There are inherent limitations with the image-based caloric estimation methods; first, the users must take a picture of the food eaten at every meal. Then, the measurement of caloric intake is achieved based on the assumption that all the food in the picture is eaten. Finally, food that is left behind creates a large error between the true caloric intake and the measured caloric intake.

The act of eating can be described by a series of specific movements or gestures. Based on this principle, the sequence of the signals derived from the motion sensors can be represented by a well-known sequence model, such as a hidden Markov model (HMM) [28] or via long short-term memory (LSTM) networks [29,30]. Food intake events were detected with an accuracy of 80% for 10 subjects [29] when LSTM was adopted. Since motion sensors are generally very small in size, motion-based methods have the advantage of being easily applied to many wearable devices. The amount of food consumed is assumed to be highly correlated with the number of the ingestive actions (chewing or swallowing). Hence, caloric intake can be estimated by counting the number of chewing events or swallowing events that are recognized using an ingestive motion recognizer. The caloric density, which is essential for calculating caloric intake from the amount of food consumed, depends on the type of food. Since it is very difficult to recognize the type of food using ingestive motion, a high level of estimation accuracy in caloric intake cannot be achieved using only motion-based approaches.

Sound from eating foods is useful for recognizing the type of food being consumed and chewing/swallowing events. The underlying principle is that chewing sounds vary depending on the shape, hardness, and moisture content of the food. Acoustic-based estimation of the ingestive behaviors has been proposed by several researchers, and these can be grouped into one of two broad categories: detection of chewing/swallowing [5,6,9–11,13,21,31] and classification of food type [8,32]. Caloric intake can be estimated by using caloric density and the amount of food consumed derived from a food-type recognizer and chewing/swallowing events detector, respectively. An average absolute error of 3.0 calories per food-type was obtained from the 10 subjects and four types of food using a neck-worn sensor equipped with acoustic, proximity, and inertial sensors [33]. Although acoustic modality provides useful clues for estimating food intake amount, there are some drawbacks associated with the usage of the audio signal. Some degree of audible noise is always present in most eating environments (e.g., cafeteria and restaurants). Therefore, performance degradation due to background noise cannot be avoided in the audio-based methods. A contact-sensing method was employed to alleviate this problem, in which external noise was shielded by attaching the acoustic sensor to the skin of the neck (e.g., neck- or throat-microphones [7,33]) or placing it inside the ear (e.g., in-ear microphone [12]). This approach can reduce noise effects to some extent, but the use of contact sensors can be uncomfortable for users and could exacerbate skin allergies. A Doppler-based approach has also been used to detect food ingestive behaviors [21,32]. The underlying principle is that Doppler frequency shifts are caused by movements of the chin and neck during meal time. The experimental results showed that the accuracy of recognizing chewing and swallowing events is comparable to that of the audio-based methods [21]. Moreover, the performance was not seriously degraded even under severe noise conditions [32].

In this study, ultrasonic Doppler signals (UDS) were used to detect chewing events during meal time, and the amount of food intake was estimated by using the number of chewing events during intervals of swallowing. The usefulness of the UDS in terms of detecting chew and swallow events was previously evaluated [21]. In the present study, however, the feasibility tests were carried out using UDS to evaluate the accuracy of estimations of food-intake amounts. Although a reasonable performance in terms of food recognition accuracy ($\geq$80%) was obtained via UDS alone [32], it was difficult to discriminate the subtle differences between certain foods using only UDS. In this study, visual information was used to identify food items, where food images were recognized via CNN.

The remainder of this paper is organized as follows. An explanation of the developed sensor is presented in Section 2. The overall procedure of the method used to estimate the food amount is described in Section 3 which includes the audio-visual principle, feature extraction, a method for detecting chewing/swallowing events, and classifications of the types of food. The experimental results are given in Section 4. Finally, concluding remarks and directions for future study are provided in Section 5.

2. Hybrid Ultrasonic and Visual Sensor Device

In this study, a sensor device was developed to detect jaw movements and sounds associated with food ingestion. A photograph of the developed sensor device is shown in Figure 1. There are four modules in the sensor device: a sine wave generator, an ultrasound speaker, a wide band microphone, and an analog signal-processing module that appears in Figure 2. In a sine wave generator module, a square wave of 40 kHz is first generated by ATtiny402 microprocessor [34] and a 5th-order Butterworth low-pass filter (MAX7424CUA) [35] with a cutoff frequency of 40 kHz is then applied to obtain the 40 kHz sine wave. Such a sine wave generator has several advantages over an analog oscillator: a very stable frequency and the ability to change the frequency via software. The 40 kHz sinusoidal signal was inputted to an MA40H1S-R [36] ultrasonic speaker.

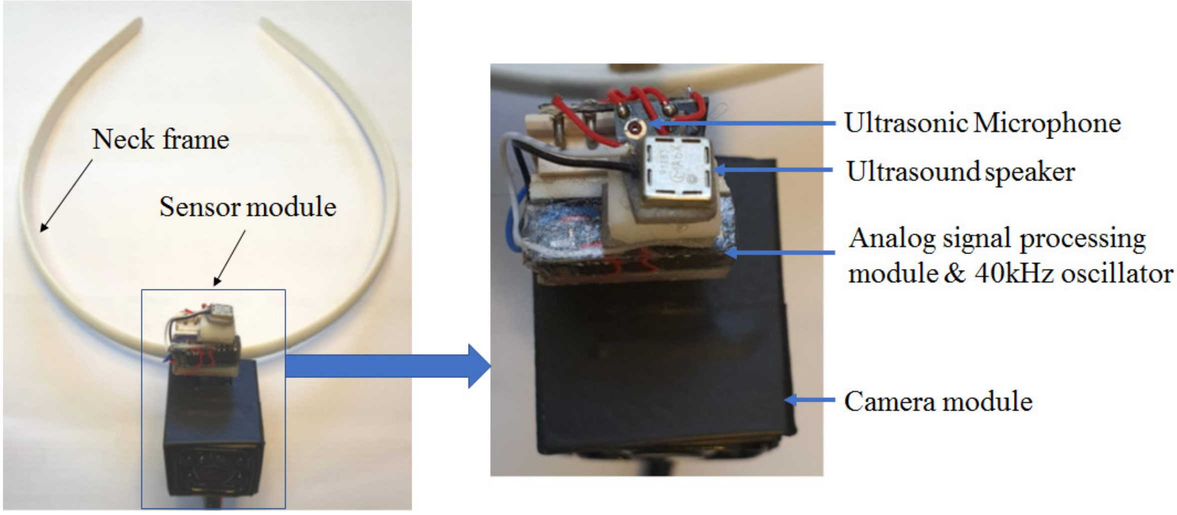

Figure 1. Developed sensor device.

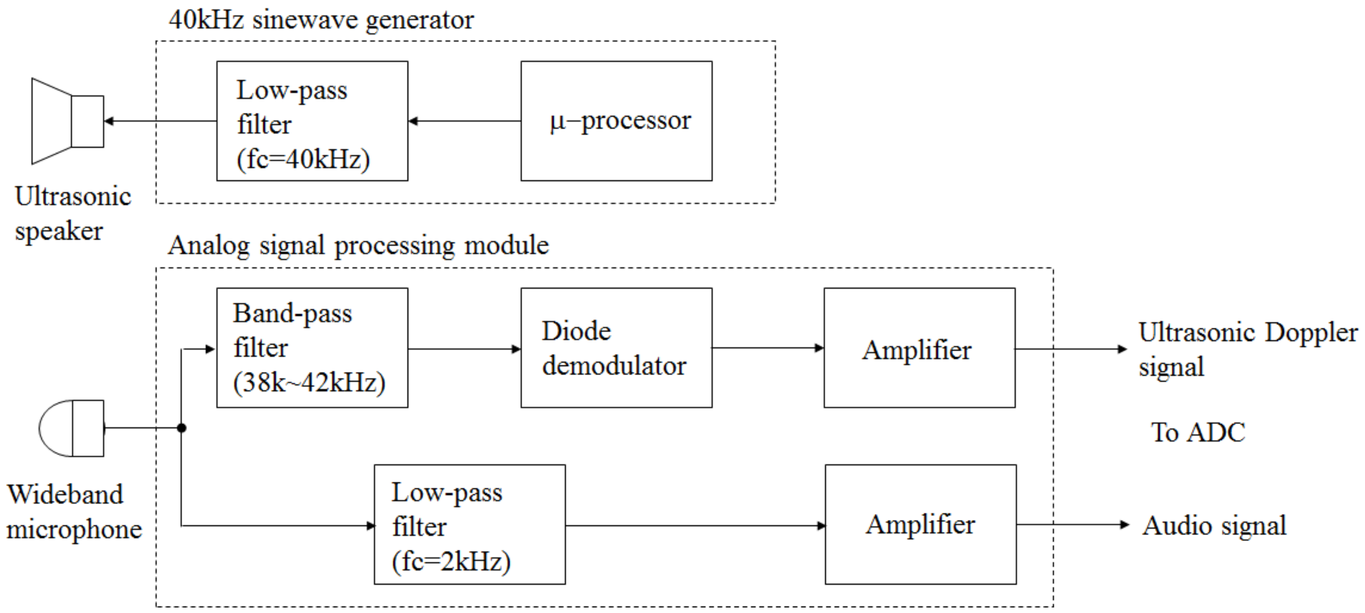
40kHz sinewave generator
Low-pass filter (fc=40kHz)
μ–processor
Ultrasonic speaker
Analog signal processing module
Band-pass filter (38k~42kHz)
Diode demodulator
Amplifier
Ultrasonic Doppler signal
To ADC
Wideband microphone
Low-pass filter (fc=2kHz)
Amplifier
Audio signal

Figure 2. Block diagram of the developed sensor device.

To record the sound and Doppler frequency shifts associated with food ingestion, the SPM0687LR5H-1 microphone [37] was mounted towards the lower jaw. The sensitivity of the employed microphone is greater than −5 dB in a frequency that ranges from 0∼42 kHz, which is sufficient for the detection of ingestive behaviors. A low-pass filter with a cutoff frequency of 2 kHz was employed to extract audio signals from the received signal. Such a cutoff frequency seemed to produce a relatively high detection of ingestive sounds. The experimental results, however, showed that setting the cutoff frequency as lower than 2 kHz yielded lower performance in terms of detecting ingestive behaviors. The ultrasonic components were obtained via band-pass filters (38∼42 kHz) and were demodulated by the diode detector. Two signals (audio and demodulated) were amplified so that the level of each was matched with the dynamic range of the subsequent analog-to-digital convertor (ADC). With the exceptions of the microphone and speaker, all components were assembled on a single printed circuit board and molded into epoxy resin to protect them from outside elements. Note that although the developed sensor module outputted both audio signals and UDS output, the audio signal was used only to compare the estimations of food amounts.

A photograph of a subject equipped with the developed sensor device is shown in Figure 3. In this photo, the wings of the sensor system are inserted into each collar and fixed to the neck when the subject wears a shirt with a collar. Such a method prevented the undesired displacement of the sensor device due to the movements of the neck and face that were not related to food intake.

Examples of the recorded signals using the developed device are shown in Figures 4 and 5 with the sampling frequency set at 96 kHz. The example signals presented in these figures were recorded under noise-free conditions, and, hence, it can be assumed that each location of the prominent peaks from audio signals correspond to chewing instances. Cashew nuts are an example of a hard-crunch food for which chewing patterns were clearly recognized in the audio signal compared with a food such as pizza. The envelope of the received ultrasonic signal takes a shape that is similar to the corresponding audio signal, and, hence, the demodulated ultrasonic signal is easily matched with the corresponding audio signal. For pizza, the amplitudes of the peaks were varied. This was because the chewing patterns depended on the hardness of each part of the pizza. (e.g., the dough portion is harder than the center). In this case, the positions of the peaks for the two signals (low-pass filtered audio signal and demodulated signal) matched, regardless of the amplitudes of the peaks. Although only two examples of food were presented, similar characteristics were commonly observed for other foods. This implies that when demodulated ultrasonic signals are used to detect chewing events, the accuracy can be as good as that of a noise-free audio signal. It is also noteworthy that the shapes of the demodulated signals remained similar to the noise-free audio signal when audible noises were presented. This indicates that by using an ultrasonic signal, it is possible to achieve a high degree of accuracy in the detection of ingestive behaviors, even under noisy conditions.

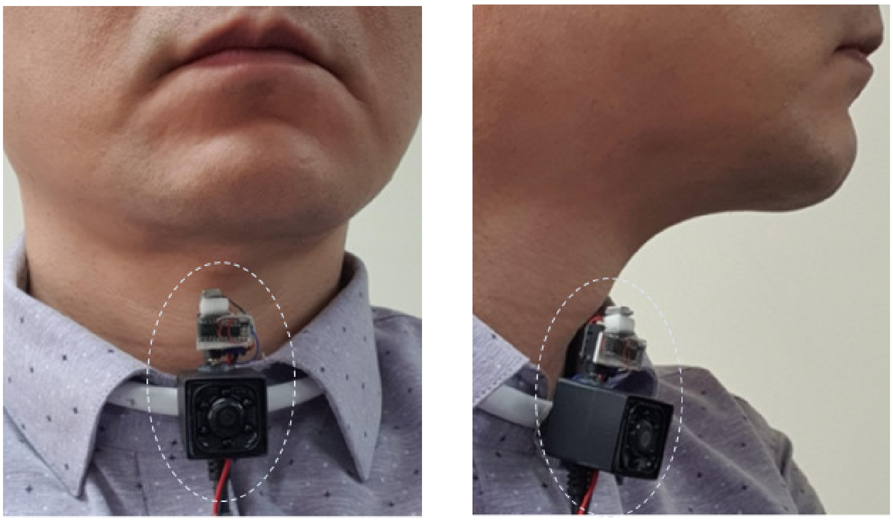

Figure 3. A photograph of a subject equipped with the developed sensor device.

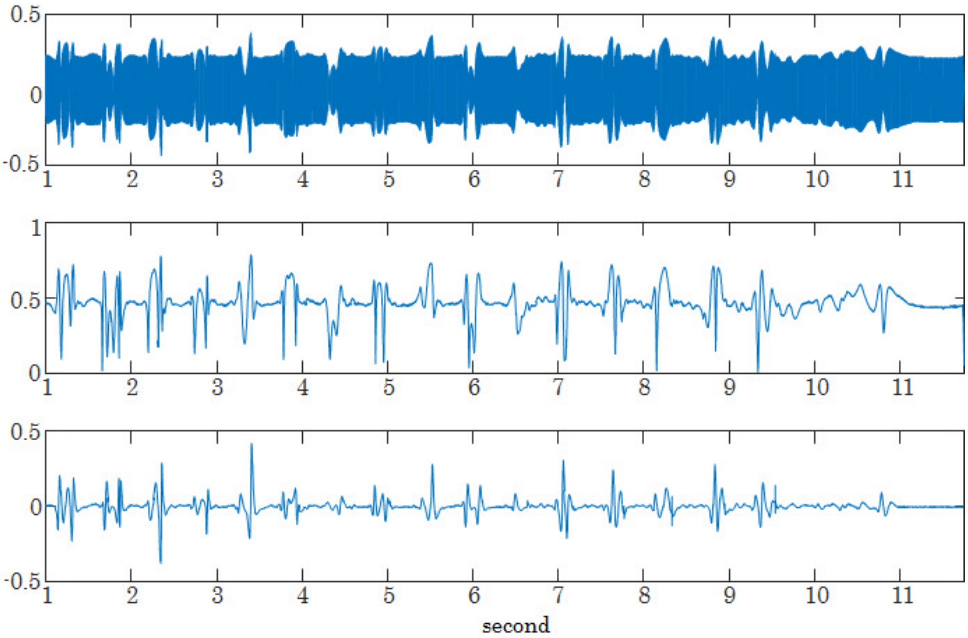

Figure 4. An example of signals received from the sensor device when cashew nuts were eaten. **Top**: The bandpass filtered signal. **Middle**: Demodulated signal. **Bottom**: Audio signal obtained by the low-pass filter.

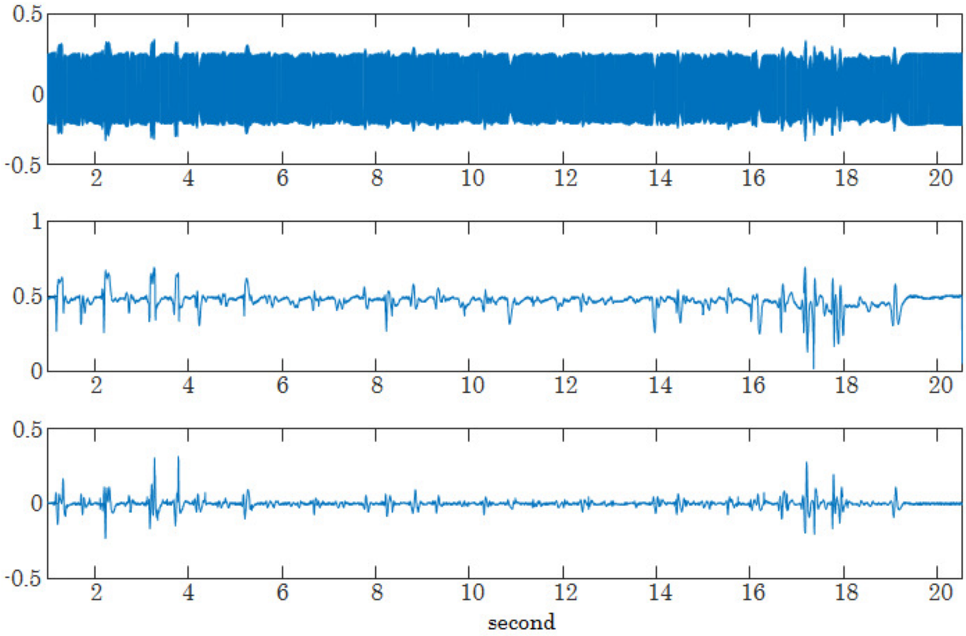

Figure 5. An example of signals received from the sensor device when pizza was eaten. **Top**: The bandpass filtered signal. **Middle**: Demodulated signal. **Bottom**: Audio signal obtained by the low-pass filter.

3. Dual-Modality Estimation of Food-Intake Amount

The proposed method of estimating food-intake amount is based on the assumptions that (1) the amount of food consumed (or, equivalent caloric intake) is positively proportional to the chewing counts before swallowing, (2) the relationship between the amount of food consumed and chewing counts depends on the type of food.

Such assumptions were empirically validated in this study, where the relationship between the number of chews and the food intake amount in mass (g) was analyzed for nine types of food (boiled chicken breasts, mocha bread, fried chicken, potato chips, cashew nuts, pizza, vegetable salad, rye bread, and plain yogurt). Ten subjects participated in the experiment and were asked to press a buzzer button at every instance of chewing. The subjects were also asked to press the buzzer button just before or after swallowing food. The amount of food eaten was measured each time before it was eaten. In this way, a true number of chewing events and swallowing instances were obtained and subsequently used for regression analysis. The correlation coefficients of the foods used in the experiment are presented in Table 1, which also includes the *p*-values, the root mean squared errors (RMSE) and the mean percentage errors. The results from the validation test are comparable to those of conventional image-based methods [27]. The relatively high correlation and low amount of estimation error ensured the validity of the food-amount estimation method when using the chewing counts. For pizza and vegetable salad, however, the number of chewing events was not strongly correlated with the amount of food intake, which resulted in a larger RMSE and a percentage error. The common characteristics of such foods is that they are mixed with several ingredients. Similarly, the firmness of the mocha bread depends on the outer and inner portions, and this is a possible explanation for the lower correlation value of the mocha bread. Consequently, the usefulness of chewing counts for the estimation of food-intake amount was confirmed by the validation test to some degree.

Table 1. The correlation coefficients, *p*-values, root mean square errors (RMSEs), and average percentage errors between the number of the chewing instances and food intake amount (g), for foods such as chicken breast (CH), mocha bread (MB), fried chicken (FC), potato chips (PC), cashew nuts (CN), pizza (PZ), vegetable salad (VS), rye bread (WB), and plain yogurt (YG).

Food Kind	**CH**	**MB**	**FC**	**PC**	**CN**	**PZ**	**VS**	**WB**	**YG**
Correlation coefficient	0.880	0.593	0.833	0.867	0.903	0.394	0.557	0.766	0.466
p-value	0.036	0.046	0.040	0.010	0.012	0.068	0.050	0.049	0.060
RMSE (g)	0.747	1.075	0.742	0.291	0.257	2.153	2.633	0.618	1.470
Average percentage error (%)	16.3	19.7	14.3	14.3	14.9	20.5	29.8	18.9	15.9

A block diagram of the proposed food intake estimation method is presented in Figure 6. The major parts of the proposed estimation scheme include detection of food ingestion behavior for chewing counts and recognition of food items. In the proposed method, the detection of chewing/swallowing events and the identification of food items were achieved using the UDS and visual modalities, respectively. A more detailed description of each part is explained in the following subsection.

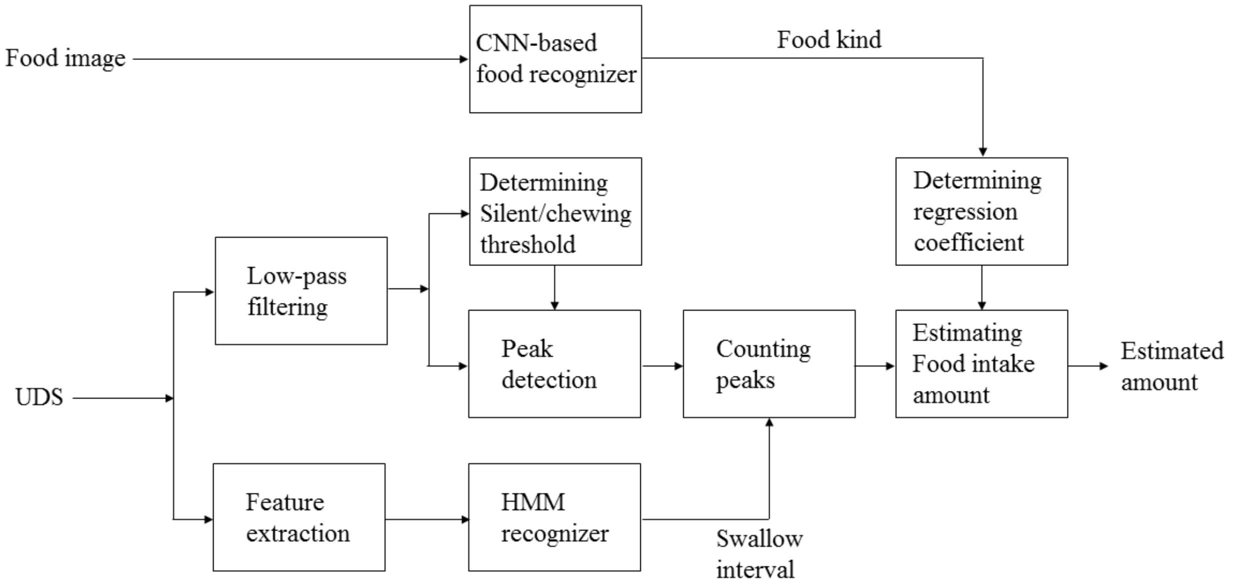

Figure 6. A block diagram of the proposed dual-modality food intake estimation method.

3.1. Detection of Food Ingestion Behaviors

The demodulated ultrasonic signal (from the sensor module) was first digitally sampled at 8 kHz with 16-bit precision. In this study, the number of chewing events during a swallow interval was practically estimated by counting the number of prominent peaks in UDS signals. Although UDS signals are more robust against environmental noises compared with audible signals, signals unrelated to food intake, which could be caused by vocalization and heartbeat, were often observed in the demodulated ultrasonic signals. The sampled signals revealed that the amplitudes of interference signals were generally smaller than those of food-intake-related signals. Accordingly, it was necessary to distinguish whether the signals were related to food intake according to signal amplitude. To this end, a soft clustering method was adopted, wherein the distribution of each group (food-intake and others) was represented by a Gaussian function. Assuming that the conditional probability density function of the feature, x, derived from UDS under each hypothesis (H0: null, H1: food-intake), is represented as a Gaussian distribution, the observation probability of x is given by

$$p(x) = \sum_{i=1}^{2} p(\lambda_i) \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left\{ -\frac{(x-\mu_i)^2}{2\sigma_i^2} \right\} \tag{1}$$

where λ_i, μ_i, and σ_i are the weight, mean, and standard deviation for the i-th Gaussian, which can be estimated using an expectation-maximization (EM) algorithm [38]. The optimum threshold was determined so that the overall misclassification errors were minimized, given by the root of the following equation:

$$x_{th} : \left\{ f(x) = Ax^2 + Bx + C \right\}|_{x=x_{th}} = 0 \tag{2}$$

where

$$\begin{aligned} A &= \sigma_1^2 - \sigma_2^2 \\ B &= 2(\mu_1\sigma_2^2 - \mu_2\sigma_1^2) \\ C &= \sigma_1^2\mu_2^2 - \sigma_2^2\mu_1^2 + 4\sigma_1^2\sigma_2^2 \log\left[\frac{\sigma_2 \cdot p(\lambda_1)}{\sigma_1 \cdot p(\lambda_2)}\right] \end{aligned}$$

A simple peak-picking method was adopted to obtain a set of candidate peaks (by taking the local maxima). The final peaks were then chosen so that the amplitude of the peak was not less than the optimum threshold. Counting of the number of chews was initiated and terminated by swallowing events. Accordingly, a reliable estimation of swallowing events is desirable to increase estimation accuracy. The specific patterns of muscle movements around the neck associated with swallowing were represented by means of a hidden Markov model (HMM). Hence, an HMM recognizer was employed to detect swallowing events. The HMM features were heuristically determined so that the accuracy of swallowing detection was maximized. The experimental results showed that the log-energy of each of the spectral bins, ranging from 2$\sim$50 Hz, yielded the highest accuracy of swallowing detection.

The relationship between chewing counts and the amount of food intake was represented by a linear regression model. Although the amount of food intake was expected to be more accurately estimated by employing a nonlinear regression model (e.g., the multi-layer perceptron), the experimental results showed that the accuracy of a linear model was slightly superior to that of a nonlinear model. The average percentage error of a linear model was 0.05 lower than that of a nonlinear model. The regression coefficients were obtained differently from food items using training data. This requires the identification of the food item currently being eaten. To this end, a visual modality was used to recognize food items with relatively high accuracy.

3.2. Recognition of Food Items

To date, there are many image recognition schemes that can be applied to automatic food recognition. Among them, a supervised learning approach that employs convolutional neural networks (CNN) was adopted in this study. The architecture of the CNN, including the number of layers and the kernel size of each layers was heuristically determined using a validation dataset that was 10% of the entire learning dataset. The resultant CNN was composed of two convolution/max pooling layers and a fully connected multi-layer perceptron (MLP) with a recognition output, and is presented in Figure 7. The kernel sizes of the 1st and the 2nd convolution layers were 11 × 11 and 7 × 7, respectively, while the window sizes of the max pooling layers were commonly 4 × 4. I tested the performance in terms of image classification accuracy according to different sizes of CNN input (input image sizes). The result showed that sizes of 64 × 64 or 128 × 128 yielded the highest classification accuracy. Lower resolution is more preferable when reducing computational loads of the small-size real-time hardware system. Accordingly, all images from the camera were reduced to 64 × 64 for input to the CNN.

There were three fully connected layers in the proposed CNN, which corresponded to input from the final convolution layer, and to hidden and output layers. The numbers of nodes for each of the layers were determined using the validation dataset as 128, 112, and 18, respectively. Note that the output nodes include a "no food" node, which is used to discriminate whether the incoming image is a valid food image. A rectified activation function was adopted for the the hidden layer. A soft-max function was employed for the output layer. All parameters of both convolutional and fully connected layers were initialized by random values. A back-propagation algorithm involved with the minimum cross entropy (MCE) criterion was used in the training procedure. The learning gain η was first set to 0.005 and changed to 0.001 after five epochs. A total of 100 epochs resulted in a trained CNN with sufficient performance in terms of food recognition accuracy. A dropout regularization with a keep probability of 0.75 was employed. Experimental results showed that the accuracy of food classification was highly affected by the mini-batch size and the decision of whether or not to apply a random batch shuffle. A high level of classification accuracy was achieved by applying a relatively small mini-batch size ($\leq$10) and a random batch shuffle. It is also interesting to note that the overall architecture of the CNN used in this study was simpler than others popularly used in image recognition tasks (e.g., ResNet-50, Inception-v3, Xception). This is mainly due to the smaller number of image items being recognized (31 vs. 1000). A lower level of accuracy was often observed for a test dataset when a relatively large number of convolution layers ($\geq$3) was adopted.

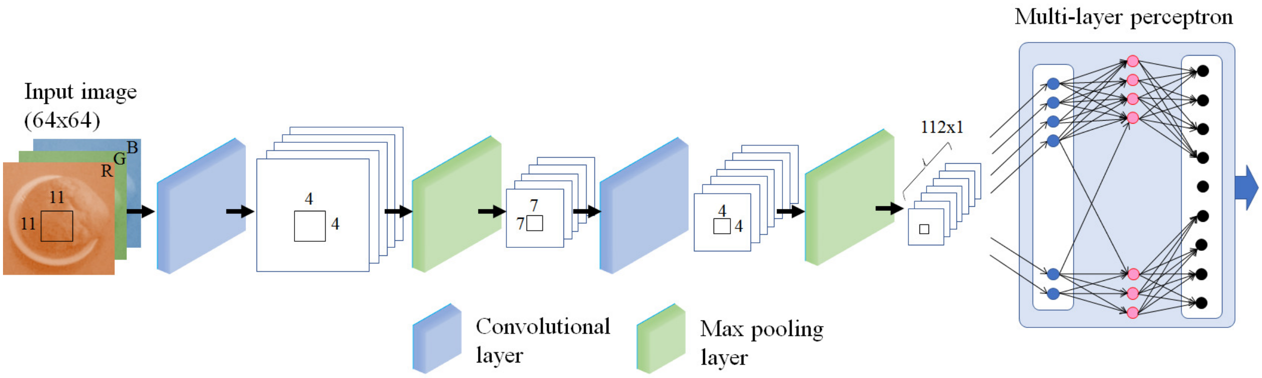

Figure 7. Architecture of the proposed CNN for the classification of food type.

3.3. Estimating the Food Intake Amount

In this study, the amount of food consumed was given by a linear combination of the chewing counts and the corresponding regression coefficients. The estimation was determined by the amount ingested between swallows. The period between two neighboring swallowing instances was referred to as "one food intake cycle" in this study. For the n-th food intake cycle, s_n, $N(s_n)$ denotes the chewing counts, and the estimated amount of food consumed $\hat{M}(s_n)$ is given by

$$\hat{M}(s_n) = c_1 N(s_n) + c_0 \tag{3}$$

where c_1, c_0 represent the regression coefficients. Let $M(s_n)$, and $F(s_n)$ denote the true amount of food consumed (in mass) and the index of food kind, respectively, and the regression coefficients are obtained by

$$c_1^{(i)}, c_0^{(i)} = \arg\min_{c_1, c_0} \sum_{s_n \in S_i} \{M(s_n) - \hat{M}(s_n)\}^2, 1 \leq i \leq N_f \tag{4}$$

where S_i is the set of the food intake cycles when the i-th food is consumed, i.e., $S_i = \{s_n | F(s_n) = i\}$ and N_f is the number of food types. To obtain the reference amount of food consumed, the food was cut into edible amounts each time and the weight was measured. Then, the subjects participating in the experiments were asked to eat all the given food pieces during one food intake cycle. Note that the regression coefficients were obtained separately for each food. Such a food-by-food basis regression analysis is beneficial for achieving a high level of estimation accuracy for each food. The drawback of this approach, however, is that the performance is more or less affected by the recognition accuracy of the adopted food classification scheme. The experimental results showed that the overall errors in food amount estimation were not significantly increased unless the classification accuracy was lower than 95%. Note that experimental results showed that the regression coefficients varied from subject to subject, even for the same food. This was because chewing sounds also varied depending on the person's age, health, social factors, and teeth condition. Accordingly, food intake estimation was implemented in a subject-dependent way.

The entire process of food intake estimation was designed to minimize user intervention and could be implemented using a small microprocessor board (Resberry pi zero W). An image was captured every 1.5 s. The captured image was inputted to the CNN. Detection of chewing and swallowing was simultaneously carried out on the incoming UDS. If swallowing was detected, the last 10 captured images were used to determine whether the period between the previous and current swallowing instances was a valid food-intake cycle. This was based on the assumption that food intake was completed (by swallowing) within 15 s of the food being exposed to the camera. If multiple food items were detected during this period, the recognition result nearest to the current swallowing instance was selected as the current food item. A period of 15 s was heuristically determined, which was a good compromise between the performance and the necessary memory size.

4. Experimental Results

4.1. Experimental Setup

Eight healthy subjects (5 males and 3 female with ages ranging from 20 to 52 years) participated in the evaluation tests. The average body mass index (BMI) of the subjects was 23.2 kg/m^2. The subjects had no medical complications that would have interfered with normal food intake. The food items were selected by considering the preferences of the subjects and by striking a balance between healthy and unhealthy foods, which included chicken breasts, mocha bread, fried chicken, potato chips, cashew nuts, pizza, vegetable salads, rye bread, plain yogurt, fried potatoes, strawberries, bananas, rice cakes, chocolate, ramen noodles (Korean style), fried rice, almonds, chocolate bars, dumplings, hot dogs, jelly, jerky, beef steak, spaghetti, sweet red-bean porridge, dried sweet potatoes, cereal (with milk), sandwiches, tangerines, and fried eggs. Such food items covered the various physical properties of everyday foods, such as hardness, cruchiness, and moisture content. The portions of the food served to each subject were divided into amounts they could easily chew and swallow. Accordingly, the subjects were requested to swallow only once if possible, but were allowed to swallow twice if necessary.

For all subjects, the total amount of food eaten was limited to alleviate the effect of subject fatigue and problems caused by the excessive intake of certain foods. This was achieved by dividing all acquisition times for each subject into five sessions, each of which consisted of seven sub-sessions. In each sub-session, the subjects were asked to eat certain foods repeatedly. One of 30 food items was selected in each sub-session. The number of repetitions for each session was determined according to the types of food and the eating patterns of the subjects. The subjects were allowed to drink water or cola between the sub-sessions, but the samples recorded for these moments were removed prior to the evaluation test. The average number of food intake cycles per food item was 25, and, hence, the average number of total food intake cycles was $25 \times 30 = 750$ per subject. The properties of the dataset are summarized in Table 2, where the values were obtained by averaging the results for all subjects.

Table 2. Dataset properties per food item.

Food Item	Total Duration (s)	Avg. Duration (s)	Total Weight (g)	Avg. Weight (g)
chicken breast	540.0	21.6	52.4	2.1
mocha bread	540.0	20.0	127.2	4.7
fried chicken	634.4	26.4	126.1	5.3
potato chips	450.8	18.0	69.0	2.8
cashew nuts	388.0	14.4	26.3	1.0
pizza	950.0	30.6	370.4	11.9
vegetable salad	633.5	25.3	228.4	9.1
rye bread	703.8	28.2	72.9	2.9
yogurt	140.2	5.6	241.3	9.7
fried potato	388.2	15.5	84.8	3.4
strawberry	202.7	8.4	209.9	8.7
banana	309.8	12.4	285.5	11.4
rice cake	483.7	18.6	164.2	6.3
chocolate	247.9	9.9	18.7	0.7
ramen	530.2	19.6	208.8	7.7
fried rice	580.5	23.2	257.3	10.3
fried egg	295.5	12.3	68.1	2.8
almond	436.6	17.5	85.4	3.4
chocolate bar	343.5	12.7	251.0	9.3
dumpling	254.2	10.2	279.5	11.2
hot dog	411.0	16.4	171.8	6.9
jelly	357.2	14.3	33.7	1.3
jerky	533.1	21.3	247.2	9.9
spaghetti	471.4	18.9	90.3	3.6
beefsteak	592.7	22.8	313.7	12.1
sweet red-bean porridge	417.0	16.0	276.9	10.7
dried sweet potato	313.8	12.1	57.6	2.2
cereal (with milk)	287.2	11.5	100.9	4.0
sandwiches	499.2	18.5	55.9	2.1
tangerine	446.7	17.9	106.2	4.2
Total	13,382.7	17.4	4681.4	6.1

In training the HMM parameters for swallowing detection, a typical left-to-right model was used. The distribution of the features was modelled using mixtures of diagonal Gaussians. The number of states for HMM was determined heuristically, at five. The number of Gaussians was determined in the same way, at three. The HMM parameters were derived from the training set (500 swallowing events). The longitudinal pharyngeal muscles are known to condense and expand the pharynx as well as help elevate the pharynx and larynx during swallowing. Accordingly, the surface electromyogram (sEMG) signals were acquired from the neck skin close to the inferior pharyngeal constrictors. The ground-truth swallowing events were obtained from the prominent peaks of the low-pass filtered sEMG signals. The instances of each of the swallowing events were annotated as "reference swallowing instances" to train the HMM.

The food images were captured before the subjects ate the foods, and were not captured during the meal. The average number of captured images for each food type was 877. For each food, various images were captured, which accounted for the view angle, the ambient light conditions, the food location, the types of trays, and the amount of food consumed. There are several food image datasets for food recognition such as UEC-FOOD100 [22], FOOD-101 [23], and VIREO Food-172 [24]. The food items employed in this study are also found in such datasets, which allowed for their use in the training of the image classifier. Although the food images captured from the camera used in this study were sometimes different from the food images in the datasets, the inclusion of the food image datasets in the training data were helpful for reliable food recognition. Therefore, 30% of the training/evaluation images were borrowed from the above-mentioned food image datasets. The ratio between the training data for learning CNN and test data for evaluation was set at 7:3. The validation dataset was randomly selected from the training dataset, which was used to determine the configuration of CNN. The set of the non-food images was composed of everyday images including people, animals, furniture, and landscapes.

The experimental results showed that the food classification accuracy of the visual-only estimator was 99.02% for 30 food items. The confusion matrix was shown in Table 3, obtained from a total of 7894 test images. Due to its high classification accuracy, it can be observed that diagonal components are prominent and other components are almost zero. Such results demonstrated that the overall performance was not seriously affected by the usage of recognized food items instead of true food items.

Table 3. The confusion matrix for 30 food items and 7894 food images. (pz:pizza, rb:rye bread, bn:banana, ch:milk chocolate, fp:fried potato, pc:potato chips, cn:cashew nuts, sb:strawberry, mb:mochabread, vs:vegetable salad, yg:yogurt, ck:chicken breast, fr:fried rice, fe:fried egg, rc:rice cake, rn:ramen noodle, fc:fried chicken, al:almond, cb:chocolate bar, cm:cereal with milk, dp:dumpling, hd:hotdog, je:jelly, jk:jerky, og:orange, rb:red-bean porridge, sw:sandwiches, sp:spaghetti, bs:beefsteak, ds:dried sweet potato).

		Predicted																													
		pz	**rb**	**bn**	**ch**	**fp**	**pc**	**cn**	**sb**	**mb**	**vs**	**yg**	**ck**	**fr**	**fe**	**rc**	**rn**	**fc**	**al**	**cb**	**cm**	**dp**	**hd**	**je**	**jk**	**og**	**rb**	**sw**	**sp**	**bs**	**ds**
Actual	**pz**	298	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	rb	0	297	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
	bn	0	0	300	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
	ch	0	0	0	312	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
	fp	0	0	1	0	297	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	pc	0	0	0	0	0	299	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	cn	0	0	0	0	0	0	299	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	sb	0	2	0	0	0	0	0	298	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	mb	0	1	0	0	0	0	0	0	299	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	vs	0	0	0	0	0	1	0	0	0	298	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	yg	0	0	0	0	0	0	0	0	0	0	312	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	ck	0	0	0	0	0	0	0	0	0	0	0	299	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	fr	1	0	0	0	1	0	0	0	0	0	0	0	295	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	fe	0	0	0	0	0	0	0	0	0	0	0	0	0	300	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	rc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	300	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	rn	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	300	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	fc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	360	0	0	0	0	0	0	0	0	0	0	0	0	0
	al	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	138	0	0	0	0	1	33	6	0	0	0	1	0
	cb	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	180	0	0	0	0	0	0	0	0	0	0	0
	cm	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	180	0	0	0	0	0	0	0	0	0	0
	dp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	43	8	120	0	0	0	0	0	0	0	0	9
	hd	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	171	0	0	0	0	4	0	0	5
	je	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	432	0	0	0	0	0	0	0
	jk	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	178	0	0	0	0	0	0
	og	0	0	0	2	0	0	0	2	0	0	4	0	0	0	0	0	0	1	0	0	0	0	0	0	169	0	0	2	0	0
	rb	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	3	0	174	0	0	0	0
	sw	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	178	0	0	1
	sp	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	296	0	0
	bs	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	179	0
	ds	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	180

4.2. Estimation Results of Food Intake Amount for Each of the Modalities

The estimations for food intake amounts were first evaluated for each modality (audio, ultrasound, and vision). The mean absolute percentage error [20] (MAPE) in mass was used as an evaluation metric. Eventually, the calories can be computed by the estimated weight and the nutrition facts for each food. For the k-th food item, the MAPE is given by

$$MAPE_k = \frac{1}{N_k} \sum_{n=1}^{N_k} \frac{|M_k(s_n) - \hat{M}_k(s_n)|}{M_k(s_n)} \tag{5}$$

where $M_k(s_n)$ and $\hat{M}_k(s_n)$ are the true and estimated weights for the n-th food intake cycle, respectively. N_k is the total number of samples for the k-th food. Both the audio and ultrasonic modalities estimated food intake amount by using the type of food that was recognized by the CNN-based classifier. Note that audio modality method used audible frequency range signal to detect chewing and swallowing instances, while demodulated ultrasonic signals were used in the ultrasonic modality method. The MAPEs for each modality are presented in Table 4, where the results were obtained under noise-free environments. The average MAPE of the ultrasonic modality estimator was comparable to that of the audio-modality estimator. Although for all subjects, the average MAPEs of the ultrasonic-modality estimator were commonly lower than those of the audio-modality estimator, the differences were not significant for all subjects. This was also confirmed by a significance test (a two-way ANOVA test), in which the p-value of the factor “modality” was 0.3023 ($df = 1$). Such results were somewhat expected, since it has been experimentally confirmed that the ultrasonic signals can be approximated as a modulated version of the corresponding audio signals. Visual inspection revealed that some details of the audio signals (mainly corresponding to the high-frequency components) were lost in the demodulated ultrasonic signals. This resulted in slight differences in terms of the MAPE from the audio-modality estimator.

The performance of the visual-modality estimator was also evaluated for comparison. A regression CNN was employed to estimate food amount [18,19]. The structure of the CNN was heuristically determined to minimize the overall MAPE for the test data set. The food images were captured using the same camera as that used for the audio- and ultrasonic-modality estimator. In the original studies [18,19], the caloric content of the foods was estimated, but the present study estimated the weights (amounts) of the foods. For each food, the average number of acquired images and the number of different amounts (in weight) was 1270 and 25, respectively. The average MAPE appears at the bottom of Table 4. The MAPE of the video-only estimator was higher than that of other estimators. These results indicate that although food types were correctly classified using the trained CNN, there is a limit to estimating the amount of food intake by visual information alone. That result shows that the average MAPE of the video-only estimator was somewhat lower than that in previous studies. (e.g., 8.2% in [20]). In the present study, however, the results were obtained from a single image (not a stereo image), and experiments were carried out on a wide variety of food amounts. Consequently, it is apparent that the collaborative use of audio and visual information (the types of food and the number of chewing events are estimated by using visual and audio information, respectively) is very helpful in achieving high accuracy in the estimations of food amount.

Table 4. The MAPEs for each modality and each subject (sub-1∼8).

Modality	sub-1	sub-2	sub-3	sub-4	sub-5	sub-6	sub-7	sub-8	Average
Audio + visual	13.5	15.2	13.8	14.7	15.9	14.3	15.5	15.8	14.6
Ultrasonic + visual	13.9	15.5	14.8	15.1	16.0	15.2	15.8	16.2	15.3
Visual only	19.45								

4.3. Estimation Results of Food Intake Amount for Each Food Item

Accuracy in estimating the food amounts was investigated for each food. The results are summarized in Figure 8. There was no clear difference in the MAPEs of the audio- and ultrasonic-modality estimator. This was confirmed by the fact that the correlation and the root mean squared error (RMSE) between the two MAPEs were 0.9727 and 1.14, respectively. For a total of 27 food items, the percentage error of the ultrasonic estimator was less than 10%. Such results indicate that, for various types of food, ultrasonic Doppler signals are as useful as audio signals in estimating the amount of consumed food.

The differences in the MAPEs between the specific food groups were also investigated. First, all food items used in the experiments were split into healthy and unhealthy food groups according to diet guidelines recommended by the world health organization (WHO) [39,40]. The accuracy of food intake estimation was then evaluated for two groups (healthy foods: bananas, chicken breast, cashew nuts, strawberries, vegetable salad, rye bread, yogurt, dried sweet potatoes, tangerines, and almonds, and, unhealthy foods: chocolate, fried potatoes, mocha bread, fried chicken, potato chips, pizza, ramen, dumplings, chocolate bar, and beefsteak). The average MAPEs for the healthy and unhealthy food groups were 15.62 and 14.23, respectively, when the audio-modality estimator was employed. Similar results were obtained when using the ultrasonic-modality estimator (16.17 and 15.04 for healthy and unhealthy, respectively). Considering that the average MAPE for the audio-modality estimator was 14.6 for all foods, the estimation accuracy for food amount was slightly reduced when the healthy foods were eaten. However, the 2-way ANOVA test showed that there was no significant difference between the two groups. Hence, the dual-modality (visual + audio or visual + ultrasonic)-based estimation method for food intake amount worked reasonably well for both healthy and unhealthy foods.

A similar trend was also observed when evaluation was separately carried out on hard/soft food groups. The hard food group includes chocolate, potato chips, cashew nuts, chocolate bars, cereal, almonds, and jerky. Fried eggs, fried rice, mocha bread, rice cakes, ramen, strawberries, vegetable salad, yogurt, sweet red-bean portage, spaghetti, dumplings, and jelly are included in the soft food group. The average MAPEs for the two groups were 14.09 (hard group) and 13.80 (soft group), respectively, when the audio-modality estimator was employed. Similar results were observed for the ultrasonic-modality estimation method where the average MAPEs were 14.98 (hard group) and 14.19 (soft group), respectively. Visual inspection of both audio waveforms and ultrasonic waveforms showed that the chewing pattern was less consistent over time when relatively hard foods were eaten. For example, the amplitude of the waveform was high at the beginning of chewing, then decreased over time. The decreasing rate of the amplitude was varied according to the amount of food, initial chewing force, and a subject's teeth condition. This resulted in changing the number of chewing events during a constant time interval. The inhomogeneity of food is another reason for performance degradation, clearly observed for pizza, chocolate bars, and sandwiches. In the case of such foods, there was a significant change in the number of chews, even for the same amount, which depended on the local properties (solidity, moisture content, and ingredients). A possible solution, alleviating this problem, is to use local visual information and recognize the local properties of foods.

A 2-way ANOVA test was performed to investigate the major factors affecting the performance of food intake estimation. In this experiment, the factors affecting the MAPEs included the subject ($df = 4$) and the type of food ($df = 29$). The results showed that the p-values for each factor were 0.002 (the type of food) and 0.654 (subject). This indicated that MAPEs are strongly affected by the type of food, while the variation in MAPEs for subjects was not significant. Such results were obtained when using the ultrasonic-modality estimator. Similar results were observed for the audio-modality estimator. These results were mainly due to the usage of the subject-dependent estimation rules.

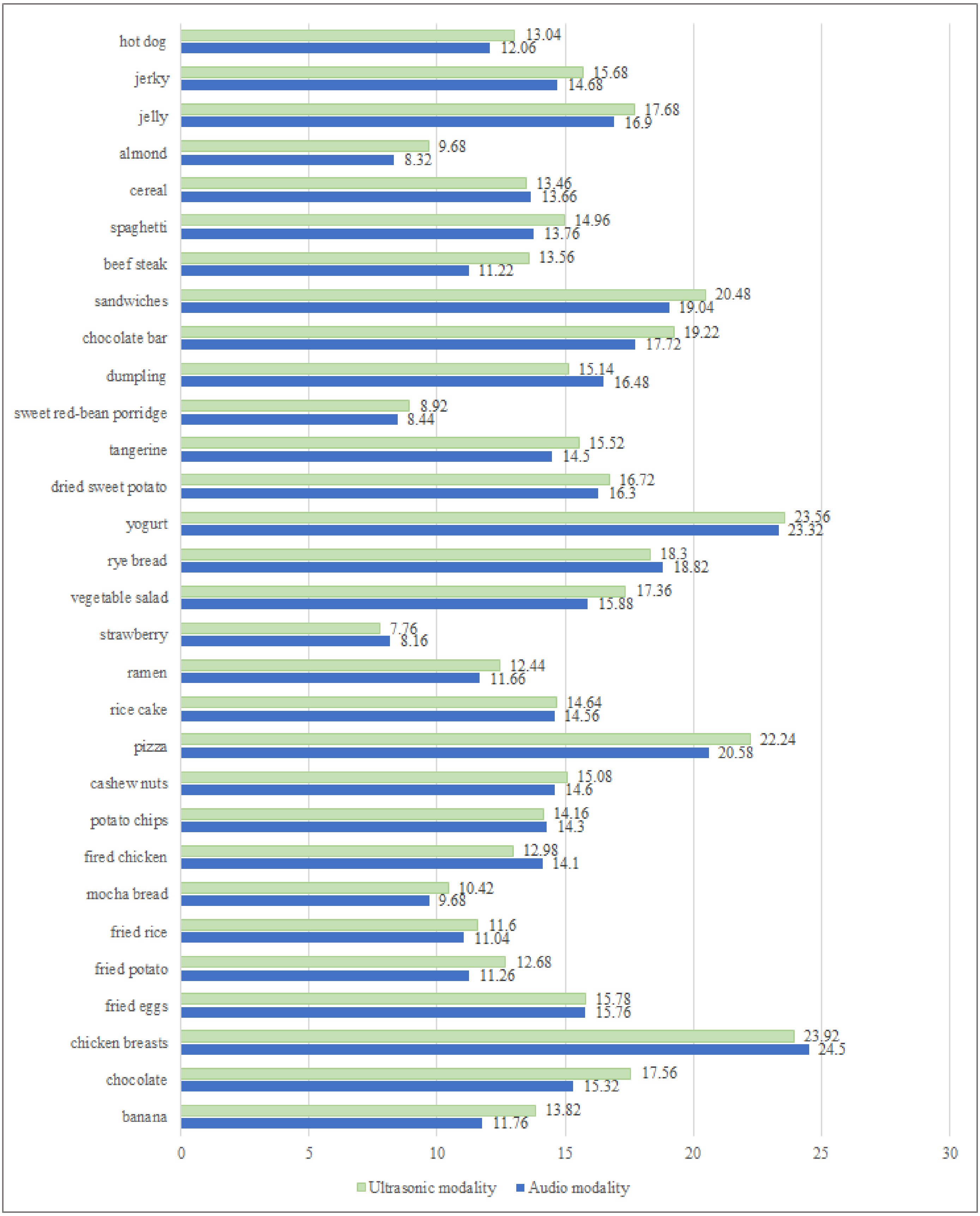

Figure 8. The MAPEs for each modality and each food item.

4.4. The Accuracy of Food Intake Amount Estimation under Noisy Environments

Thus far, all presented results were obtained under a noise-free environment. However, there are many kinds of interference in most meal situations. Such factors should be considered for practical application. For this section, the possibility of using the ultrasonic-based estimation method was investigated under actual meal situations. The noisy audio signals were obtained by digitally adding the noise signals to the clean audio signal. Note that environmental noises were mainly observed in the auditory frequency band and were very rarely found near the ultrasonic band used in this study. Hence, noise signals were only added to audio signals. The SNR was controlled by changing the scale factor of the noise signal before addition. The noise signals were separately recorded using the developed sensor in various eating environments that included cafeterias, restaurants, and home dining rooms.

To investigate the range of the signal-to-noise ratio (SNR) for most eating environments, the average levels of chewing sounds and noise signals were computed. Two food items (cashew nuts and bananas) were chosen since the difference in amplitude between the chewing sounds of the two foods was relatively large. Thus, a wider range of SNR was obtained by using these two foods. SNRs were computed by using chewing sounds from two subjects (subject-1 and subject-2) and the noise signals acquired from three different cafeterias, two different restaurants, and each subject's home dining room. A wide variety of environmental interferences (conversation, sounds caused by the use of cutlery such as spoons, forks, chopsticks, etc., and sounds from music and TV) were allowed for recording noise signals.

Examples of the SNRs measured for the various eating situations are presented in Table 5. For the same environment, the SNR varied slightly from subject to subject, since the level of the chewing sounds varied slightly according to the subject. The overall SNRs of cashew nuts were higher than that those of bananas. This was because the level of chewing sounds is relatively high when chewing cashew nuts. Compared with cafeterias and restaurants, the noise levels from the home dining rooms were relatively small. According to such results, the SNRs of ∞ (noise-free), −5.25 dB, −17.5 dB, and −19 dB were selected for the subsequent experiments. These SNRs were the most common values under each environment.

Table 5. An example of the measured SNR (in dB) for the various eating environments, for the two subjects (sub-1 and sub-2) and for the two food items (cashew nuts and banana).

Food Item	Subject	Cafe-1	Cafe-2	Rest-1	Rest-2	Rest-3	Home
Cashew nuts	sub-1	−10.09	−14.31	−14.49	−15.98	−16.42	−2.70
	sub-2	−9.21	−13.36	−12.90	−14.66	−14.61	−1.16
Banana	sub-1	−12.65	−16.87	−17.05	−18.54	−18.98	−5.26
	sub-2	−11.98	−16.12	−15.67	−17.43	−17.32	−3.93

The MAPEs of the clean audio signal, noisy audio signal, and ultrasonic signal are shown in Table 6. All results presented in this table were obtained by averaging all food items. The evaluation was carried out on the five subjects (subject 1–5). As expected, the performance of food amount estimation was generally degraded as the SNR decreased when the audio modality was used. The 2-way ANOVA test showed no clear difference in the MAPEs between a noisy signal at −5.25 dB SNR and an ultrasonic signal ($p = 0.575$). For low SNRs (−17.5 and −19 dB), however, significant differences were observed between the two modalities (audio/ultrasound) ($p = 0.018$ and $p = 0.011$, for −17.5 dB and −19 dB, respectively). The average MAPE for ultrasonic signals was slightly lower than that of the audio signal in the case of a noise-free and relatively high SNR (−5.25 dB). These results verified that the ultrasonic signal was as useful for food amount estimation as the noise-free audio signal. Moreover, a good performance for the ultrasonic signal was maintained even under noisy environments. The major reason for the noise-robustness of the ultrasonic signal against environmental noise is that the frequency range of the background noise did not overlap with that of the ultrasonic signal (38∼42 kHz).

Table 6. The MAPEs of the clean audio signal, noisy audio signal, and ultrasonic signal. All results were obtained by averaging all food items.

	sub-1	sub-2	sub-3	sub-4	sub-5	Avg.
Clean audio	13.5	15.2	13.8	14.7	15.9	14.62
Noisy audio (SNR = −5.25 dB)	13.7	15.2	14.5	14.8	15.5	14.74
Noisy audio (SNR = −17.5 dB)	15.2	17.5	16.3	17.6	17.1	16.74
Noisy audio (SNR = −19.0 dB)	15.4	17.8	16.6	17.0	17.2	16.80
Ultrasound	13.9	15.5	14.6	15.1	16.0	15.02

5. Conclusions

Automatic tracking of the types and amounts of consumed foods is very helpful for individuals who are engaged in a diet program. There are many techniques that are useful in the implementation of an automatic diet tracking system. In this study, acoustic and visual information was cooperatively used to automatically estimate food intake amount. It was assumed that specific patterns of ultrasonic Doppler would be observed when chewing foods. Hence, chewing events were detected using the ultrasonic signals received from the eating person's mouth region. The amount of consumed food was computed via linear regression analysis between the chewing count and the amount of food intake. Evaluation tests were carried out on 30 food items, commonly eaten in daily life, in order to verify the effectiveness of the proposed method. The results showed that the ultrasonic signal was a good alternative to the audio signal. Moreover, the performance of this food amount estimation method was maintained even in noisy environments. Since there are many kinds of ambient interferences in real-world environments, ultrasonic-based methods would be a good choice for practical use.

Although the sensor device developed was light and small in size, the wearing of the device can sometimes cause discomfort to users. The usage of wireless data transmission and further reducing size of the sensor/camera modules are a possible solution for this problem. Increasing the number of food items would be desirable to extend the usability of the proposed method. Our future study will focus on this issue.

Funding: This work is funded by the Basic Science Research Program through the National Research Foundation of Korea through the Ministry of Science and ICT under Grant S202103S00081.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data sharing not applicable.

Acknowledgments: The author is grateful to the members of the bio-signal processing lab at Konkuk University for participating in several experiments.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Moayyedi, P. The epidemiology of obesity and gastrointestinal and other diseases: An overview. *Dig. Dis. Sci.* **2008**, *9*, 2293–2299. [CrossRef] [PubMed]
2. Prentice, A.M.; Black, A.E.; Murgatroyd, P.R.; Goldberg, G.R.; Coward, W.A. Metabolism or appetite: Questions of energy balance with particular reference to obesity. *J. Hum. Nutr.* **1989**, *2*, 95–104. [CrossRef]
3. De Castro, J.M. Methodology, correlational analysis, and interpretation of diet diary records of the food and fluid intake of free-living humans. *Appetite* **1994**, *2*, 179–192. [CrossRef]
4. Kaczkowski, C.H.; Jones, P.J.H.; Feng, J.; Bayley, H.S. Four-day multimedia diet records underestimate energy needs in middle-aged and elderly women as determined by doubly-labeled water. *J. Nutr.* **2000**, *4*, 802–805. [CrossRef] [PubMed]
5. Sazonov, E.S.; Fontana, J.M. A sensor system for automatic detection of food intake through non-invasive monitoring of chewing. *IEEE Sens. J.* **2012**, *5*, 1340–1348. [CrossRef]

6. Sazonov, E.S.; Makeyev, O.; Schuckers, S.; Meyer, P.L.; Melanson, E.L.; Neuman, M.R. Automatic detection of swallowing events by acoustical means for applications of monitoring of ingestive behavior. *IEEE Trans. Biomed. Eng.* **2010**, *3*, 626–663. [CrossRef] [PubMed]
7. Alshurafa, N.; Kalantarian, H.; Pourhomayoun, M.; Liu, J.; Sarin, S.; Sarrafzadeh, M. Recognition of nutrition-intake using time-frequency decomposition in a wearable necklace using a piezoelectric sensor. *IEEE Sens. J.* **2015**, *7*, 3909–3916. [CrossRef]
8. Bi, Y.; Lv, M.; Song, C.; Xu, W.;Guan, N.; Yi, W. Autodietary: A wearable acoustic sensor system for food intake recognition in daily life. *IEEE Sens. J.* **2016**, *3*, 806–816. [CrossRef]
9. Päfiler, S.; Fischer, W. Food intake monitoring: Automated chew event detection in chewing sounds. *IEEE J. Biomed. Health Inf.* **2014**, *1*, 278–289.
10. Päfiler, S.; Wolff, M.; Fischer, W.-J. Food intake monitoring: An acoustical approach to automated food intake activity detection and classification of consumed food. *Physiol. Meas.* **2012**, *33*, 1073–1093.
11. Amft, O. A wearable earpad sensor for chewing monitoring. *IEEE Sens.* **2010**, *4*, 222–227.
12. Nishimura, J.; Kuroda, T. Eating habits monitoring using wireless wearable in-ear microphone. In Proceedings of the International Symposium on Wireless Pervasive Communication, Santorini, Greece, 7–9 May 2008; pp. 130–133.
13. Makeyev, O.; Meyer, P.L.; Schuckers, S.; Besio, W.; Sazonov, E. Automatic food intake detection based on swallowing sounds. *Biomed. Signal Process. Control* **2012**, *6*, 649–656. [CrossRef]
14. Weiss, R.; Stumbo, P.J.; Divakaran, A. Automatic food documentation and volume computation using digital imaging and electronic transmission. *J. Am. Diet. Assoc.* **2010**, *1*, 42–44. [CrossRef]
15. Sun, M.; Liu, Q.; Schmidt, K.; Yang, J.; Yao, N.; Fernstrom, J.; Fernstrom, M.; DeLany, J.; Sclabassi, R. Determination of food portion size by image processing. In Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Vancouver, BC, Canada, 21–24 August 2008; pp. 871–874.
16. Zhu, F.; Bosch, M.; Woo, I.; Kim, S.Y.; Boushey, C.J.; Ebert, D.S.; Delp, E.J. The use of mobile devices in aiding dietary assessment and evaluation. *IEEE J. Sel. Top. Signal Process.* **2010**, *4*, 756–766.
17. Pouladzadeh, P.; Shirmohammadi, S.; Al-Maghrabi, R. Measuring calorie and nutrition from food image. *IEEE Trans. Instrum. Meas.* **2014**, *8*, 1947–1956. [CrossRef]
18. Ege, T.; Yanai, K. Simultaneous estimation of food categories and calories with multi-task CNN. In Proceedings of the 15th International Conference on Machine Vision Applications, Nagoya, Japan, 8–12 May 2017; pp. 198–201.
19. Ege, T.; Ando, Y.; Tanno, R.; Shimoda, W.; Yanai, K. Image-based estimation of real food size for accurate food calorie estimation. In Proceedings of the IEEE conference on Multimeda Information Processing and Retrieval, San Jose, CA, USA, 28–30 May 2019; pp. 274–279.
20. Dehais, J.; Anthimopoulos, M.; Shevchik, S.; Mougiakakou, S. Two-view 3D reconstruction for food volume estimation. *IEEE Trans. Multimed.* **2017**, *5*, 1090–1099. [CrossRef]
21. Lee, K.-S. Food intake detection using ultrasonic doppler sonar. *IEEE Sens. J.* **2017**, *18*, 6056–6068. [CrossRef]
22. UECFOOD-100 Dataset. Available online: http://foodcam.mobi/dataset100.html (accessed on 21 August 2021).
23. The Food-101 Data Set. Available online: https://data.vision.ee.ethz.ch/cvl/datasets_extra/food-101 (accessed on 21 August 2021).
24. VireoFood-172 Dataset. Available online: http://vireo.cs.cityu.edu.hk/VireoFood172/ (accessed on 21 August 2021).
25. Chairi, K.; Pavlidis, G.; Markantonatou, S. Deep learning approaches in food recognition. In *Machine Learning Paradigms*; Springer: Cham, Switzerland, 2020; pp. 83–108.
26. Thames, Q.; Karpur, A.; Norris, W.; Xia, F.; Panait, L.; Weyand, T.; Sim, J. Nutrition5k: Towards Automatic Nutritional Understanding of Generic Food. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Online, 19–25 June 2021.
27. Subhi, M.A.; Ali, S.H.M.; Ismail, A.G.; Othman, M. Food volume estimation based on stereo image analysis. *IEEE Instrum. Meas. Mag.* **2018**, *6*, 36–43. [CrossRef]
28. Kyritsis, K.; Tatli, C.L.; Diou, C.; Delopoulos, A. Automated analysis of in meal eating behavior using a commercial wristband IMU sensor. In Proceedings of the 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Seogwipo, Korea, 11–15 July 2017; pp. 2843–2846.
29. Kyritsis, K.; Diou, C.; Delopoulos, A. Food intake detection from inertial sensors using LSTM networks. In Proceedings of the International Conference on Image Analysis and Processing, Catania, Italy, 11–15 September 2017; pp. 411–418.
30. Kyritsis, K.; Diou, C.; Delopoulos, A. Modeling Wrist Micromove-ments to Measure In-Meal Eating Behavior from Inertial Sensor Data. *IEEE J. Biomed. Health Inform.* **2019**, *6*, 2325–2334. [CrossRef] [PubMed]
31. Zhang, R.; Amft, O. Monitoring chewing and eating in free-living using smart eyeglasses. *IEEE J. Biomed. Health Inform.* **2018**, *1*, 23–32. [CrossRef]
32. Lee, K.-S. Joint Audio-ultrasonic food recognition using MMI-based decision fusion. *IEEE J. Biomed. Health Inform.* **2020**, *5*, 1477–1489. [CrossRef]
33. Zhang, S.; Nguyen, D.; Zhang, G.; Xu, R.; Maglaveras, N.; Alshurafa, N. Estimating Caloric Intake in Bedridden Hospital Patients with Audio and Neck-Worn Sensors. In Proceedings of the IEEE/ACM International Conference on Connected Health, Washington, DC, USA, 26–28 September 2018; p. 18474393.
34. Microchip Inc. ATTINY402. Available online: https://www.microchip.com/en-us/product/ATtiny402 (accessed on 1 August 2021).

35. Maxim Integrated Inc. MAX7424. Available online: https://www.maximintegrated.com/en/products/analog/analog-filters/MAX7424.html (accessed on 1 August 2021).
36. Murata Manufacturing, MA40H1S-R. Available online: http://www.murata.com/$\sim$/media/webrenewal/products/sensor/ultrasonic/open/datasheet_masmd.ashx?la=ja-jp (accessed on 1 November 2016).
37. Knowles Electronics Inc. SPW0442HR5H-1. Available online: https://www.knowles.com/docs/default-source/model-downloads/spw0442hr5h-1-datasheet-rev-g.pdf?Status=Master&sfvrsn=4 (accessed on 1 July 2021).
38. Dempster, A.; Laird, N.; Rubin, D. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc.* **1977**, *39*, 1–38.
39. World Health Organization. Healthy Diet. Available online: https://www.who.int/news-room/fact-sheets/detail/healthy-diet (accessed on 1 July 2021).
40. Glasbey, C.A. An analysis of histogram-based thresholding algorithms. *Comput. Vis. Graph. Image Process.* **1993**, *55*, 532–537. [CrossRef]