1408 COMBINED CONTOUR DETECTION AND POINT CLOUD OF RGB-DEPTH IMAGE FOR FOOD VOLUME ESTIMATION

By Yuita Arum Sari

WORD COUNT 3341 TIME SUBMITTED 31-DEC-2024 12:18PM

PAPER ID 113798237

COMBINED CONTOUR DETECTION AND POINT CLOUD OF RGB-DEPTH IMAGE FOR FOOD VOLUME ESTIMATION

Yuita Arum Sari[1]

Abstract

Assessing nutritional consumption entails a procedure that enables nutritionists and dietitians to track the eating habits of patients within healthcare settingsTraditionally, this measurement relies on manual observations by specialists utilizing visual analysis. However, this approach is prone to subjectivity due to the risk of expert fatigue, which can result in inaccuracies. Furthermore, the evaluations may differ among experts based on varying viewpoints. In a decision support system, a more objective analysis is necessary. Previous research has utilized the area captured in a food image to estimate the weight of food on a plate. Nonetheless, this technique still results in numerous prediction errors. To tackle this issue, we propose a novel method to calculate the volume of food from a camera image, which aims to provide a more accurate weight prediction. In this paper, we introduce a new approach that combines contour detection with a point cloud derived from RGB depth images to capture height information. The Root Mean Square Error (RMSE) for height prediction is 1.04 and 0.69 when viewed from the first and second sides, respectively, while the volume prediction reaches an RMSE of 45.08. This suggests that the differences between the predicted and actual values for volume and height are suitable for practical applications.

Keywords : *food volume estimation, RGB Depth Image, point cloud, food volume, contour detection*

Abstrak

Mengukur asupan nutrisi adalah sebuah proses yang memungkinkan ahli gizi dan ahli diet untuk memantau kebiasaan makan pasien di fasilitas medis. Secara tradisional, pengukuran ini dilakukan dengan observasi manual oleh para ahli menggunakan analisis visual. Namun, metode ini rentan terhadap subjektivitas karena potensi kelelahan ahli yang dapat menyebabkan kesalahan. Selain itu, penilaian dapat bervariasi antara para ahli karena perspektif yang berbeda. Dalam konteks sistem pendukung keputusan, analisis objektif diperlukan. Penelitian sebelumnya telah menggunakan area dalam gambar makanan untuk mengukur berat makanan di piring. Akan tetapi, metode ini masih menyebabkan banyak kesalahan prediksi. Sehingga, untuk mengatasi hal ini, metode baru diusulkan untuk menghitung volume makanan dalam gambar kamera untuk memprediksi berat makanan dengan lebih akurat. Dalam penelitian ini, kami mengusulkan kombinasi baru deteksi kontur dan *point cloud* dari gambar RGB-depth untuk mendapatkan informasi ketinggian. Untuk evaluasi *height* prediction nilai Root Mean Square Error (RMSE) masing-masing adalah 1,04 dan 0,69 dilihat dari sisi satu dan sisi dua, dan RMSE 45,08 untuk estimasi volume. Hal ini menunjukkan bahwa perbedaan antara estimasi volume dan ketinggian object yang diprediksi dengan nilai sebenarnya cukup akurat untuk aplikasi praktis.

Kata kunci : *estimasi volume citra makanan, gambar RGB-depth, point cloud, deteksi kontur*

1. INTRODUCTION

There are many applications under development to evaluate the nutritional value of foods through photography. [1], [2], [3], [4]. One such application focuses on monitoring the dietary habits of patients [5], [6], [7], [8]. Traditionally, nutritionists would analyze the remains of patients' meals to gain further insight into their health status. The conventional approach to measuring food weight with digital scales has been deemed inefficient because it requires extensive measurements. As a solution, the Comstock method was developed, where leftover food is weighed on scales that have three to seven tiers. This method has received commendations for being easy to implement, employing trained observers, and providing quick results. However, it is still prone to subjectivity when multiple observers are involved, which could impact measurement precision due to the necessity for trained and experienced personnel. Furthermore, observer fatigue during the estimation process can result in inaccuracies. Consequently, there is an increasing demand for applications that utilize computer vision for objective visual analysis to reliably assess food mass..

In a previous study [9], the calculation of remaining food was confined to the surface area of images, thereby prompting a need for further research on incorporating height information to achieve enhanced accuracy. Relying exclusively on area measurements poses significant challenges in the assessment of leftovers and the analysis of nutritional intake through computer vision, as it fails to account for the actual weight of the food [10]. Consequently, enhancing the precision of food weight estimation through the integration of volume data emerges as a pivotal area of research. The proposed algorithm for volume estimation is versatile and can be applied to a range of food items, facilitating the use of volume data for accurate weight prediction.

In order to obtain volumetric information from a photograph, it is essential to utilize a camera sensor that is capable of detecting the depth of objects within the image. Numerous methods exist for calculating volume; however, the most effective approach depends on the specific task at hand.For instance, when photographing food, a 360-degree video approach is recommended. This approach involves capturing a comprehensive view of the food item, followed by uploading the footage to the DDRS cloud database. Through advanced 3D modeling techniques, the database can estimate the volume of the meal with a high degree of accuracy. While this process is efficient and accurate, it is imperative to exercise caution as lasers, which are sometimes utilized, can pose risks to human eyes [11].

A safer alternative involves the use of two cameras (Android) or a stereo camera (iPhoneX), set at angles of 90 and 75 degrees, respectively. Video recording occurs while adjusting the smartphone to capture various perspectives.This method is user-friendly as it does not require additional sensors; however, challenges may arise in the adjustment of angles. The variability among users in shooting from different viewpoints necessitates the development of a unique method for achieving angles of 90 and 75 degrees, respectively [12].

Other solutions for mobile phones include depth sensors and depth cameras (Intel RealSense/Microsoft Kinect).This technique also attempts to address self-occlusion in food items, which is one of the most challenging problems in volume estimation.The dataset is restricted to a few 3D objects with irregular shapes [13].

According to previous research, 3D height information may significantly improve prediction capacity [14]. This research proposes a method for determining volume utilizing information about the object's area multiplied by height. Area is calculated using contour detection based on the reference, and height is predicted using 3D point cloud information from the RGB depth picture.

This paper is organized as follows. The next section details the research methodology, beginning with data collection and concluding with the evaluation metrics. Subsequently, section 3 presents the results and analysis. Lastly, the conclusion summarizes the main findings and outlines future developments, particularly regarding the establishment of an improved environment for dataset collection.

2. RESEARCH METHOD

2.1 Setting the Prototype

The acquisition of food images is facilitated by the Intel Realsense D435i depth camera, which captures images from two perspectives: from the side and above. This approach disregards ambient illumination, as the method outlined in this system remains applicable under diverse lighting conditions, enabling the volumetric measurement without the necessity for calibration. The prototype's design is illustrated in Figure 1. The food images were captured from

perpendicular positions to project the area of the food image and two lateral views of the food image to predict the height .From the lateral views, the image is captured from the anterior and posterior sides, so the object is rotated 180 degrees to obtain both views. It is hypothesized that both views will represent height information from the object. The height of the depth camera sensor is 140 mm, while the focal length is 19.33 mm [15]. These values are required to measure the real height of the food image.

Verification of the accuracy of volume computation necessitates ground truth, i.e., an independent measurement of the true volume of the object. Contemporary practice involves the use of a measuring cup that is straightforward and accessible to all users. The volume of an object is determined through the application of Archimedes' principle [16]. The object is placed into a measuring cup of a designated size, and subsequently, the object is introduced until it is fully immersed, thereby causing an increase in the water level within the measuring cup. The discrepancy between the initial and final levels of water in the measuring cup is indicative of the true volume of the object under consideration is depicted in the Figure 2.

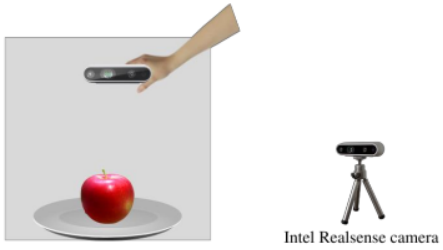

Figure 1. The prototype of proposed method using depth camera

Figure 2. Measuring volume using Archimedes approach

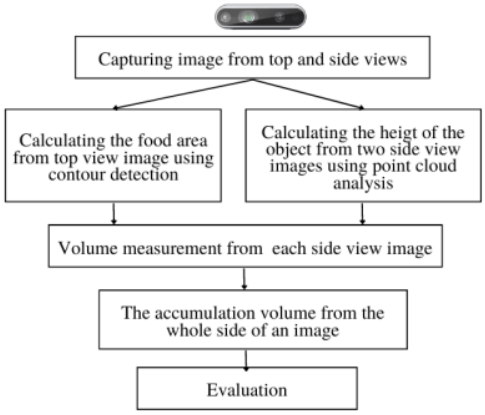

Figure 3. General phase.

2.2 General Phase of Proposed Method

The proposed method is outlined in Figure 3. The technique's sequence of operations comprises the following: contour recognition to identify the food area, point cloud analysis using RGB picture segmentation and depth image to provide height predictions, and volume calculation by combining area and height information in each side view using (1). Following the acquisition of volume predictions from each side, the volumes of both sides are aggregated. To validate the experimental outcome, the RMSE calculation is employed.

$$v_{\{s\}} = a \times h \tag{1}$$

where $v_{\{s\}}$ is volume prediction from each side view of two images, a represents the area prediction and h is height prediction of object.

2.3 Dataset

The prototype detailed in section 2.1 is used to acquire two categories of food images: one taken from above and the other from the side. The overhead image serves as a reference to ascertain the true area of the pixel region through contour detection, whereas the side image is used to estimate the height of the food item by applying point cloud analysis. This research explores six varieties of individual food data, with every dataset containing three photographs: one from the top and two from the side. Figure 4 illustrates an example of a dataset created with a depth camera.

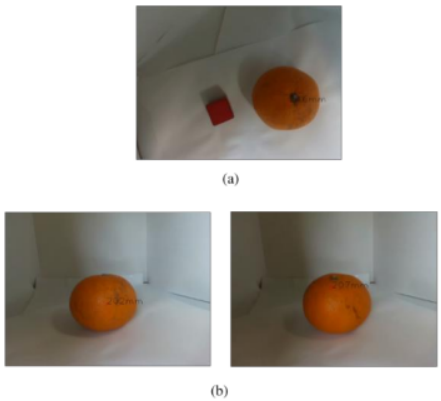

Figure 4. Dataset were taken from depth camera. (a) The food object is captured from the top view with a red block as a reference. (b) The object is captured from side view without using reference

2.4 Contour Detection for Predicting the Area of Food Image

The initial step in the contour detection and area prediction process entails the identification of optimal segmentation outcomes, with the objective of differentiating the background from the food items in the image. In this study, a clustering-based image segmentation algorithm that has been utilized in previous research was employed [17]. Subsequent to segmentation, the subsequent stage involves the detection of the food area and the rectangle. The outcomes of image segmentation and contour detection are demonstrated in Figure 5. The calculation of the area is based on contour detection of the reference, which is a rectangle, and the food object being analyzed.

Figure 6 illustrates the technique for obtaining the expected area of the food object described above. The following processes are taken to compute the food area in pixels: picture filtering from RGB photos, contour recognition of the food area, and calculation of the area of the food. Concurrently, the area of the reference in pixels is determined. Given the rectilinear configuration of the reference, the subsequent procedures are implemented: conversion of the RGB picture to a gray image, followed by adaptive thresholding, contour analysis to obtain an approximation of shape equal to 4, and final calculation of the area of the rectangle. Subsequent to acquiring the information regarding the area of the food and the reference in pixels, the formula (2) is employed to obtain the actual area in centimeters (cm).

$$a = \frac{F_{px}}{R_{px}} \times s \qquad (2)$$

where Fpx is prediction of food area in pixel, Rpx represents the reference area in pixel, and s is side of rectangle in real measurement using cm.

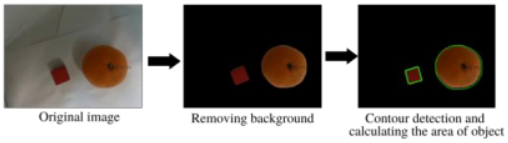

Figure 5. The result of rectangle and food object detection

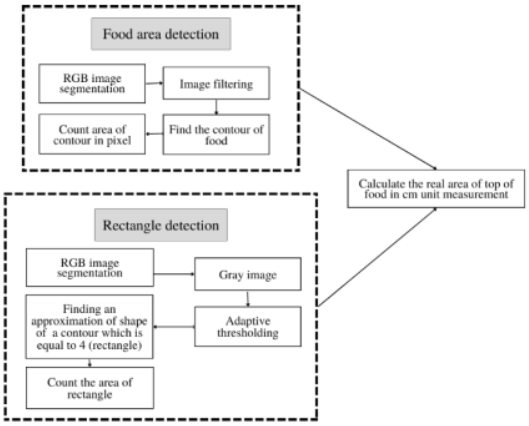

Figure 6. The detailed process of contour detection to recognize the reference and object

2.5 Point Cloud Analysis for Heigh Prediction

The Intel Realsense D435i depth camera generates RGB and depth images; however, a segmentation procedure is required to forecast a food object's height, which removes the background so that only information from the food object is extracted. To produce a point cloud from the food object, a depth picture that has been segmented is also required by comparing the pixel values of the segmented RGB image. As illustrated in Figure 7, this segmentation process enables the acquisition of a segmented point cloud, which can be utilized to delineate objects in a given scene. The segmentation process is designed to mitigate the impact of exposure variations on the captured images, which can occur due to fluctuations in illumination during image capture. However, it should be noted that predictions may become less precise in scenarios where exposure levels are either under- or overexposed. To address the issue of exposure in image capture, the background should be removed from the point cloud. Consequently, the versatility of this prototype extends to its applicability in diverse lighting environments.

Following the estimation of the food area, the subsequent step entails the determination of its

height. Figure 8 illustrates a methodology for measuring the height of a food product by leveraging information from a point cloud, in conjunction with two-dimensional images that generate a distance value from the camera to the object (in millimeters). The depth camera quantifies the depth values of *Zoc* and *Zod*, as well as *ZOC* and *ZOD*. It is noteworthy that *Zoc* and *Zod* do not possess units; however, *ZOC* and *ZOD* are camera distances measured in millimeters.

The vertical line positioned at the center of the x-axis signifies the measured height. To ascertain the midpoint, it is necessary to identify the minimum and maximum values of *X*, subsequently dividing by 2. The location of the *Y* ordinate is derived from the minimum value of *Y* and the maximum value of *Y*. Subsequent to obtaining the center line, which serves as an estimate of the height, the subsequent step involves identifying the depth value, or z, of an image. The *z* information is acquired from the distance captured by the depth camera with the object.In this case, two types of depth values are present: *z* in the form of coordinates called Zoc and *Zod*, and the distance between the object and camera in millimeters, called ZOC and *ZOD*. Each *X* and *Y* value will have a *Z* value, so that a point can be represented as a vector (*X, Y, Z*).

Therefore, the distance between the two points can be calculated using the formula for the Euclidean distance.The formulas for calculating the distance between *Zoc* and *Zod* (formula 3) and between *ZOC* and *ZOD* (formula 4) are provided.Because the results of Ecd and ECD have different units, the formula 5 and 6 must be used to find the actual height of an image object. Figure 9 shows a complete illustration of the actual height measurement of an object.

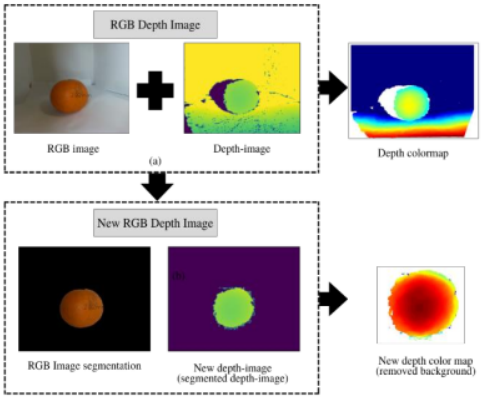

Figure 7. Point cloud segmentation

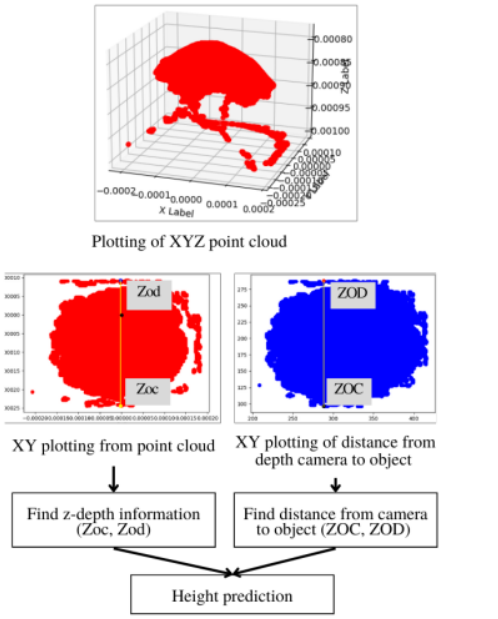

Figure 8. Point cloud of 3D and 2 D image for getting distance information

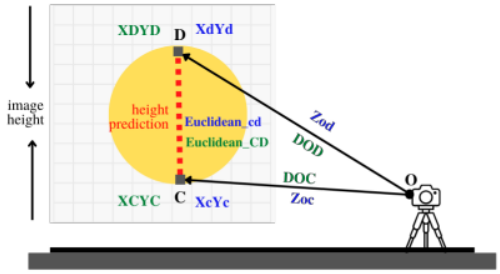

Figure 8. Height prediction

It is imperative to note that, in scenarios where high measurements are attained, the D435i depth-camera sensor must be equipped with crucial information regarding its height and focal length.

$$E_{cd} = \sqrt{(X_c - X_d)^2 + (Y_c - Y_d)^2 + (Z_c - Z_d)^2} \quad (3)$$

$$E_{CD} = \sqrt{(X_C - X_D)^2 + (Y_C - Y_D)^2 + (Z_C - Z_D)^2} (4)$$

where Ecd is Euclidean distance between vector (Xc, Yc, Zc) and $(Xd, Y\,d, Zd)$, while E_{CD} is Euclidean distance between vector (X_C, Y_C, Z_C) and (X_D, Y_D, Z_D)

$$k = \left(\frac{E_{cd} \times sh}{fl \times h_i}\right) \times \frac{1}{4} \quad (5)$$

$$h = k \times E_{CD} \quad (6)$$

where k is weight value of height, sh is sensor height of depth camera equals to 140 mm, fl means the focal length which is equal to 19.33 mm, hi represents the height of the image (in pixel) and in this case is 480, and h is the real height prediction.

2.6 Volume Prediction

The volume of a food object is predicted from the volume calculations on each side of the food object. Since this study uses two side view images, the sum of the two side views is calculated. Formula 7 is used to calculate height predictions.

$$v_o = \sum_{i=1}^{2} v_s \quad (7)$$

where v_o is volume prediction from two side view images.

2.7 Root Mean Square Error (RMSE)

The standard deviation of prediction errors is represented by the RMSE. It is a common way to calculate a model's error in predicting quantitative data. The lower the RMSE errors, the better the proposed model. The presence or absence of outliers in the final forecast also affects this error figure. The equation illustrates the RMSE formula (8).

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(g_v - p_v)^2}{n}} \quad (8)$$

where g_v is ground truth volume getting from measuring cup observation, p_v represents the predicted volume and n is the number of experimented object.

3.RESULT AND ANALYSIS

The results of the experiment were obtained by comparing the height prediction results with the ground truth measured with a ruler and the volume measured with a measuring cup using the Archimedes' Law technique.

3.1 Height Evaluation

The Table 1 the evaluation findings from the computation of the object's height prediction on both sides, compared to the ground truth data. Overall, the RMSE created by side 1 is 1.04, whereas the RMSE generated by the side 2 is 0.60. The RMSE value provided by this assessment is minimal, ranging from 0.60 to 1.04 when comparing the estimated height to the real high value.

Based on the results of this high prediction, it can be concluded that the proposed algorithm can handle this problem well. Problems arise when the same side of the object has a different height. The method in this paper is still limited if, on the same side, the height used for calculations is the height in the middle of the object.

Table 1. The height prediction from 2 side view images comparing to the ground truth

Food name	Height		
	Ground truth	Side 1	Side 2
Orange1	6.5	6.57	5.62
Orange2	3.8	2.47	2.49
Apple1	7.3	7.93	7.94
Apple2	7.5	6.08	5.11
Apple3	7.5	6.08	5.11
Tomato1	5	5.5	5.29
RMSE	**1.04**	**0.69**	

3.1Volume Estimation

Table 2 summarizes the results of the volume prediction. Each item has a ground truth volume, which is the volume predicted by the research and obtained using the Archimedes principle. According to these results, one prediction, Apple2, is an outlier among the others, with a ground truth of 200 ml and a predicted result of 119 ml. This significant gap is due to incorrect segmentation. The Figure 10 shows how bad segmentation results can affect the estimated food area value. This is also the reason why the segmentation results are not ideal, since the camera's field of view does not include the entire top view of the image.

Table 2. The volume between ground truth (Gv) and prediction (Pv)

Food name	Gv (ml)	Pv (ml)
Orange1	200	194.3
Orange2	60	53.38
Apple1	260	233.1
Apple2	200	119
Apple3	200	130.6
Tomato1	105	10

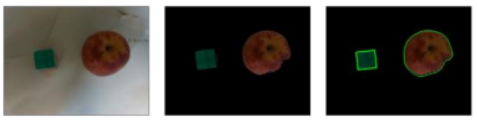

Figure 10. The result of area prediction of Apple2

4. CONCLUSION

Volume estimation of food image based on the multiplication of area and height from contour detection and a point cloud of RGB depth image is presented in this paper for a new framework in prediction with less reference. Contour detection is applied to images taken from the top view to get the food area with a reference, while the height information is measured from the two side view of the images without any reference. Both measure the area, and height prediction requires good segmentation to separate the object and background properly. Based on the experiment implemented on the data, the RMSE from the volume measurement reaches 45.08, while the RMSE from the height prediction is 1.04 and 0.69, from the side 1 and side 2 view, respectively, which means that the proposed method is adequate for projecting the food volume. However, a volume prediction with a significant error affects the overall result of RMSE. The outlier of the volume prediction is caused by the prediction of the area in the contour detection resulting from the image segmentation.

Therefore, future work needs to improve the quality of image segmentation. In addition, the complexity of food items in the single image must also be considered, because in the real world, the different food items become a problem. Food image identification algorithms are acquired to facilitate the recognition of food items. A further step is related to weight measurement by listing the density obtained from the volume and the actual mass of the food.

ACKNOWLDGMENT

I would like to express my sincere gratitude to Faculty of Computer Science, Brawijaya University and Okayama University for their invaluable guidance and support throughout this research.

1408 COMBINED CONTOUR DETECTION AND POINT CLOUD OF RGB-DEPTH IMAGE FOR FOOD VOLUME ESTIMATION

ORIGINALITY REPORT

3%
SIMILARITY INDEX

PRIMARY SOURCES

1	www.ncbi.nlm.nih.gov Internet	17 words — < 1%
2	hdl.handle.net Internet	16 words — < 1%
3	jurnal.syntaxliterate.co.id Internet	15 words — < 1%
4	journal.unpas.ac.id Internet	12 words — < 1%
5	Nicole Janotte, Benedikt Kölsch, Eckhard Lüpfert, Johannes Pernpeintner et al. "Application of a combination of innovative non-destructive measurement techniques for structural, energetic and safety analysis of buildings", Journal of Building Engineering, 2024 Crossref	9 words — < 1%
6	siba-ese.unile.it Internet	9 words — < 1%
7	Shiwei Zhang, Zhengzheng Wang, Wei Ke. "One point is all you need for weakly supervised object detection", Pattern Recognition, 2024 Crossref	8 words — < 1%

8	Wenyan Jia, Boyang Li, Qi Xu, Guangzong Chen et al. "Image-based volume estimation for food in a bowl", Journal of Food Engineering, 2024 Crossref	8 words — < 1%
9	media.neliti.com Internet	8 words — < 1%
10	ul.qucosa.de Internet	8 words — < 1%

EXCLUDE QUOTES	OFF	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	OFF	EXCLUDE MATCHES	OFF