International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. #, No.# (Mon-20..)

3D Reconstruction with SFM-MVS Method for Food Volume Estimation

Nurdzakirah Amir[1], Zahir Zainuddin[1*], and Zulkifli Tahir[1]

[1] *Departement of Informatics, Hasanuddin University, Makassar, Indonesia*

E-mail address: amirn21d@student.unhas.ac.id, zahir@unhas.ac.id, zulkifli@unhas.ac.id

Received ## Mon. 20##, Revised ## Mon. 20##, Accepted ## Mon. 20##, Published ## Mon. 20##

Abstract: The growing need for accurately understanding calorie and nutrient intake highlights the limitations of traditional methods for assessing food portions. Advanced computer vision technologies, mainly 3D reconstruction methods, provide a more precise and automated approach to estimating food volume. This study focuses on developing a cutting-edge system that creates Three-Dimensional (3D) food images using the Structure From Motion-Multi View Stereo (SFM-MVS) method. Detailed volume estimation is performed after constructing the 3D model to ensure accurate measurements. Using a sophisticated mobile application to capture images from multiple angles, the system undertakes a comprehensive 3D reconstruction of food items. This complex process is enhanced by subsequent slicing and segmentation, allowing for detailed extraction and precise volume calculation of each food component. The system has undergone rigorous testing on various food types, consistently showing a volume estimation error rate below 10%, thus significantly improving the accuracy of food volume estimation. This research significantly advances automatic diet monitoring and calorie consumption management. By leveraging state-of-the-art 3D reconstruction techniques, the system effectively overcomes the limitations of traditional methods, providing a reliable, efficient, and user-friendly approach to dietary assessment. Consequently, it supports more effective nutritional planning and health management, meeting the growing demand for precise and automated dietary monitoring solutions. The impact of this research is extensive, offering considerable benefits for health professionals, nutritionists, and individuals seeking accurate nutritional information to support better health outcomes.

Keywords: 3D reconstruction, Structure From Motion, Multi View Stereo, Computer Vision, Food Volume Estimation, Diet Management

1. INTRODUCTION

In recent years, there has been an increase in chronic diseases caused by diet, prompting people to raise awareness of healthy thinking patterns and strictly regulate food intake [1]. Many people use mobile apps to record their daily food intake in response to this trend. The result of this recording is an image of the food, which is then used to analyze the type, nutrient intake, or calories contained in the food to be consumed [2]. Knowing the amount of nutrients or calories in a food plays a vital role as it can help regulate the amount of food consumption and prevent potential problems due to over- or under-consumption.

The rapid development of computer vision technology paves the way for researchers to analyze food images [3]. In a typical situation, a user takes one or more images of their food, perhaps even a video and the system reports the associated nutritional or caloric information. Several steps must be taken to obtain nutritional information, including food detection or segmentation, volume estimation, and calorie content assessment. A more profound knowledge of food volume is essential in the whole process to get a more accurate and precise analysis of nutritional value or calories.

This research proposes a method that can solve the challenge of food calorie estimation: volume estimation. The proposed method utilizes multiple parameters to achieve good accuracy of different types of food with various shapes and textures in a reasonable time. It is more precise than traditional method that rely only on 2D images and can handle the complexity of irregular food shapes, which are often challenging in traditional volume estimation methods. Several recent attempts have been made to automatically estimate food volume using

E-mail:author's email

smartphone applications. To achieve this, the proposed system must first reconstruct the shape of the food's three-dimensional (3D) image by taking pictures from different angle. Volume estimation using 3D images can simplify the steps and make it more realistic to get more accurate results. There are 3 stages of volume estimation in this study on research:

1) **3D Reconstruction**: In the shape of the 3D food image is reconstructed using one or more images (views). The SFM-MVS method was used to obtain the 3D model.
2) **3D Slicing**: After the 3D model is obtained, slicing is performed using Blender until several parts are obtained. The slicing results are segmented and then continued by counting the number of pixels on the object using the Bware method, where each pixel generated will be used to determine the surface area based on the existing model.
3) **Volume Estimation**: Calculate the volume of each sliced section using the existing equation.

2. RELATED WORK

Puri et al. created a system to improve the accuracy of food intake assessment by estimating food volume using a 3D volume estimation algorithm, Multi-view Dense Stereo Reconstruction. The evaluation results show that the food recognition module can recognize food types with an accuracy of about 85 to 90% and produce an average volume estimation error ranging from 2.0% to 9.5%. There are still some areas for improvement in this paper, namely the inaccuracy in determining volume estimation on foods with less texture due to difficulties in dense stereo matching. This system also requires user input through voice to help the food recognition process; this dependence on user input can reduce the efficiency of automation and increase the burden on users [4].

Xu et al. proposed a method to estimate food volume from images taken with mobile devices. The approach uses 3D reconstruction and camera calibration to create 3D graphical models of each food type, which are then projected back onto the image plane for volume estimation. Essential steps include the creation of 3D models from various viewpoints during the training stage, the determination of translational and elevational parameters relative to the camera coordinates through calibration, and the projection of the 3D models back onto the image plane. Experimental results show that this method can calculate the volume of food with good accuracy and reliability. However, there are some drawbacks: this method relies heavily on the accuracy of camera calibration to determine geometric parameters, and calibration errors may reduce the accuracy of volume estimation. In addition, the segmentation quality of the resulting image still affects this method, although it is more resistant to noise than template-based methods. The 3D reconstruction and volume estimation processes involved are also more complex and require more excellent computational resources compared to traditional methods [5].

Dehais et al. introduced a system specifically designed to measure food portions. This research utilizes a three-stage system and applies SURF and RANSAC methods to calculate portion sizes using images captured from mobile devices. The first stage involves understanding the configuration of two pictures taken from different viewpoints; the next stage includes creating a solid 3D model of the two images, and the third stage is extracting the volume of food items on the 3D model. The performance evaluation of the system shows an average error below 10%, with an execution time of about 5.5 seconds per dish antenna. This method has some drawbacks that need to be noted, namely that it relies heavily on the firm and varied texture of the scene to detect critical points and perform dense reconstruction. Foods that have little or no texture cannot be reconstructed accurately [6].

In this study, a technique was used to measure food intake using a wearable camera. This technique combines Simultaneous Localization and Mapping (SLAM), a modified convex husk algorithm, and a 3D mesh object reconstruction technique to measure food volume accurately. The research evaluation results show that the average volume estimation errors range from 11.7% to 19.2% statically and 16.4% to 27.9% in real-time measurements. Several drawbacks need to be considered, namely limiting the accuracy of the system in measuring the volume of food with irregular or asymmetrical shapes, reliance on the convex hull reconstruction method, which tends to overestimate the volume of objects due to its sensitivity to noise and outliers and the Simultaneous Localization and Mapping (SLAM) method used produces a sparse map that often loses essential information due to the limited viewing angle during image capture. This can lead to errors in food volume estimation. [7].

Recently, a researcher faced challenges in estimating food volume due to the diverse nature of food and its multiscale characteristics. This approach requires only a front view of the reference for food volume estimation. The strategy involves optimizing the bounding box and converting the height, width, and area of the food from pixel-level to absolute values with high precision. Experimental results demonstrate the effectiveness of the proposed method in predicting food volume, with the average absolute error of each food type being less than 4.5%. This result shows that the model is robust in estimating the volume of irregularly shaped food. However, there are some drawbacks in that the method relies heavily on using a Rubik's cube as a reference in the image to measure food volume. This reliance limits the flexibility of the technique as the user must always have a Rubik's cube for accurate measurement, which is only practical in some situations. The experiments conducted in

this paper are limited to five types of food, so the external validity of the method needs to be further tested with various other types of food to ensure the generalization of the results [1].

Cai et al. 3D structure from 2D images is a highly complex process that requires expertise with often limited results. Therefore, this study performs 3D reconstruction on objects such as sculptures using SFM, PMVS, and PSR methods, where high-quality and textured 3D models can be recovered automatically. The results show that the proposed system outperforms state-of-the-art approaches regarding accuracy and completeness [8].

This research's essential contribution lies in applying the Structure from Motion - Multi-View Stereo (SFM-MVS) method to 3D reconstruction. This method allows the system to build 3D food models from multiple viewpoints, reducing the volume estimation errors that often occur with 2D image-based methods and making it should be in handling irregularly shaped foods. SFM-MVS also excels in addressing the issue of low texture on food and does not heavily rely on the accuracy of camera calibration. This reduces the likelihood of errors caused by calibration inaccuracies and makes the method more flexible and easier to apply in various imaging conditions. Another advantage of the SFM-MVS method is its ability to overcome the limitations of convex-based approaches like those used in Simultaneous Localization and Mapping (SLAM) techniques. The SLAM method generates sparse maps and often misses critical information because of the limited viewpoints during capture. Conversely, the SFM-MVS method can create denser and more detailed 3D models or maps by utilizing camera poses from multiple perspectives.

3. METODHOLOGY

The system proposed in this study consists of two main stages: 3D image reconstruction and volume estimation, as shown in Figure 1. Testing was done using Python programming language and slicing process using the Sketchup application. The system began with the process of taking a series of images of the food from different angles. For each disk analyzed, images were taken from various vertical viewing angles using an iPhone Xr smartphone. For each disk analyzed, photos were taken from various vertical viewing angles using an iPhone Xr smartphone. The smartphone is equipped with a 12-megapixel rear camera featuring a wide-angle lens, 1080p/30 or 60 fps (3024 x 4032 pixels) optical image stabilization, and advanced HDR capabilities. This step is essential to ensure that the data obtained is consistent and accurate. Consistency in the distance and angle at which the images were taken ensures that every detail of the food is captured, which is crucial for accurate 3D image reconstruction.

This research meticulously ensures the data collected is sufficiently high quality to facilitate the 3D image reconstruction by maintaining a stable shooting distance and consistent angle. This methodological consistency is crucial as it minimizes unwanted variations in the dataset, which could otherwise compromise the accuracy and integrity of the resulting 3D model. Additionally, this rigorous approach significantly simplifies the feature-matching process inherent in the reconstruction method. By capturing images from carefully controlled and consistent angles, these images algorithmic analysis and processing become more straightforward and reliable. Consequently, the reconstruction algorithm can more efficiently identify and match features across different viewpoints, leading to more accurate and coherent 3D models. This enhanced precision improves the overall quality of the 3D reconstruction and ensures that the models generated are robust and dependable for subsequent analysis and applications.

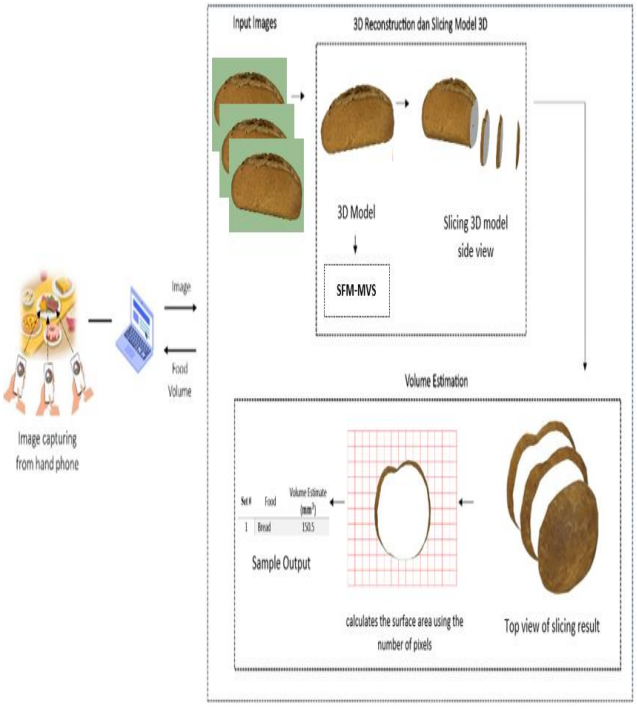


Figure 1. System Design

This research for 3D image reconstruction uses the Structure From Motion-Multi View Stereo (SFM-MVS) method, with several stages. These stages include feature extraction and matching, outlier removal, camera pose estimation and triangulation, bundle adjustment, and Multi-View Stereo (MVS). After obtaining the 3D model, volume estimation is carried out by slicing the 3D model and then segmenting it, then calculating the volume in the part of each slicing that has been segmented. The last step is to sum the volumes of each acquired slice to get the overall volume of the food item. Figure 2 offers further details on each stage of the process.

International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. #, No.# (Mon-20..)

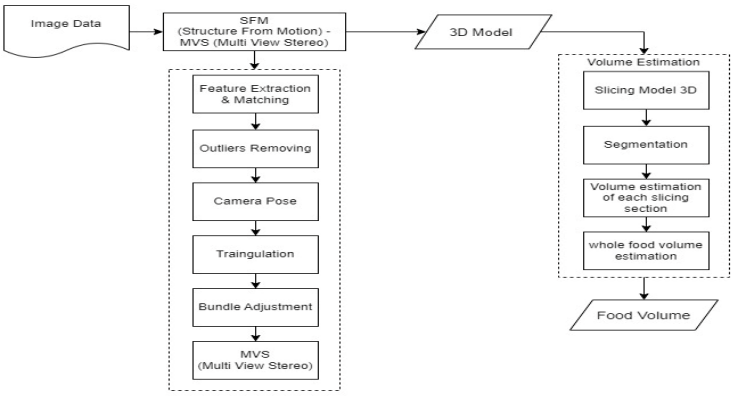


Figure 2. Block Diagram of Volume Estimation

A. 3D Reconstruction

1) Feature Extraction and Feature Matching

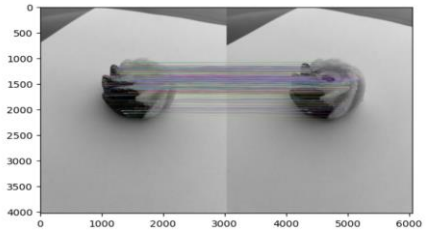
Extracting features from each image, these features are used to identify the same features in different images using the SIFT Algorithm [9]. SIFT was first introduced in 1999 [10][8] and has become famous for its uniqueness and invariance to scale, rotation, and lighting changes. Unlike the Harris Algorithm [11], which relies on image parameters such as viewing angle, depth, and scale in the image, SIFT can detect features independently by converting image data into invariant coordinates. This makes SIFT very reliable under various image capture conditions.

The SIFT algorithm consists of four main steps: building a scale space, extracting key points, assigning principal directions, and generating feature point descriptors [12]. First, it creates a scale space by filtering images at various scales to find extreme points in the scale space. This step is essential for detecting stable feature points at multiple scales and extracting critical points from the image. These key points location are in the image that show significant local variations and tend to remain stable despite changes in scale or rotation of the image. Third, assign a principal direction to each key point based on the local gradient around it. This principle direction allows the feature descriptor to be rotationally invariant. Fourth, it generates feature point descriptors, vectors that describe the local neighbourhood around each key point. These descriptors then match features extracted from pairs of interconnected images. The main idea is to

Filter out the extreme points in the scale space, thus finding stable feature points. Finally, local features from the pictures are extracted around each stable feature point to form a local descriptor and use it in future matching [13]. It can be seen in Figure 3 (a) the feature detection on the Yellow Cake and Figure 3 (b) the feature matching of the two Yellow Cake images.

The feature-matching process is to matches the descriptors extracted from each pair of images. This is done to identify the same features in different images, establishing relationships between different viewpoints of the same object. Accurate feature matching is essential to ensure that the resulting 3D model accurately represents the original object. The SIFT algorithm has the advantage of finding and matching stable features, which is the basis for many applications in computer vision, including 3D reconstruction. It can be seen in Figure 3 (a) the feature detection on the Yellow Cake and Figure 3 (b) the feature matching of the two Yellow Cake images.

(a)



(b)

Figure 3. (a) Feature Detection, (b) Feature Matching

E-mail:author's email

2) Outlier Removal

Appropriate refinement steps are essential to remove ambiguities in the point cloud recovered by Structure From Motion (SFM). At this stage, we use the Random Sample Consensus (RANSAC) method to identify outliers in the data set and remove them from feature matching. The RANSAC-based method serves as the basis [14]; as a hypothesis and verification framework, RANSAC randomly takes a small sample from the initial matching and repeatedly estimates the fundamental matrix.

The RANSAC method is often used in image processing and computer vision to identify and remove outliers from data sets. Outlier removal with RANSAC is an iterative process that focuses on identifying and using data that fits the expected model. This is especially useful when the data contains a lot of noise or outliers that can affect the analysis or modelling performed. By applying RANSAC, we can minimize the influence of inappropriate data, thereby improving the accuracy and reliability of the resulting model. The RANSAC process involves several key steps:

1) It randomly selects a small subset of the data and uses this subset to estimate an initial model.

2) It tests the model against the entire data set to identify points that fit the model (inliers) and those that do not (outliers). This process is repeated several times with different subsets to find the best model with the most significant number of inliers.

3) The best model is optimized by using all identified inliers.

However, choosing the RANSAC parameters carefully, such as the fit threshold and the number of iterations, is essential. The threshold determines how closely the data must fit the model to be considered an inlier. In contrast, the number of iterations determines how often the subset selection and testing process is repeated. Proper parameter selection is essential to ensure the best results are obtained, and the resulting model is accurate and reliable.

The Random Sample Consensus (RANSAC) method is highly effective in scenarios where the data is replete with outliers, and its application can markedly enhance the quality of the resulting model. To understand the outlier removal process using RANSAC, refer to the illustration in Figure 4, which delineates the steps involved in identifying and excising outliers from the matching feature dataset. By integrating the RANSAC method into the Structure From Motion (SFM) process, we can significantly refine the point cloud, rendering it cleaner and more precise. This improvement in data quality directly translates to a higher fidelity in the 3D reconstruction outcomes. The application of RANSAC ensures that erroneous data points are effectively filtered out, thereby optimizing the accuracy and reliability of the reconstructed 3D models. This method's robustness in handling noisy datasets makes it an indispensable tool in computer vision and 3D modelling, ultimately leading to superior model quality and enhanced interpretability of the reconstructed scenes.

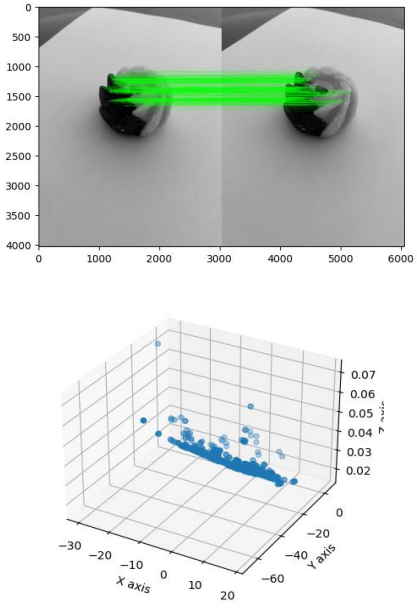


Figure 4. Outlier Identification

3) Camera Pose Estimation and Triangulation

Camera pose estimation and Triangulation are two important steps in the Structure From Motion (SFM) process to reconstruct 3D objects from 2D images. By using feature matching information, Pose Estimation [15] is used to predict and track the location of objects in the image by looking at specific combinations of pose and object orientation. Triangulation process is finding the intersection of two lines in a space. The triangulation process [16] in SFM determines each feature match's point (coordinates) in 3D space.

Camera position estimation can be done using various techniques, including feature matching methods between different images to determine the relative movement of the camera from one image to another. This allows us to understand how the camera motion and changes orientation when taking images from different angles, which is important for building an accurate 3D model of the observed object.

Camera position triangulation is calculating the 3D positions of points on the observed object based on the information obtained from the images taken by the camera. The goal is to determine the actual position of those points in 3D space corresponding to the observed 2D image. Triangulation involves matching features between images to assess the viewing angles and relative distances between the observed points and then using this information to calculate the exact 3D coordinates of the points. With the combination of Pose Estimation and Triangulation, we can understand how SFM reconstructs 3D objects from 2D

images. Post Estimation helps track the movement or transformation of objects in an image or video, while Triangulation helps determine the accurate 3D position of the observed points. These two processes are work together to produce a precise 3D model of the natural world based on the visual information obtained from 2D images.

Pose Estimation involves analyzing sequential images to predict changes in the position and orientation of objects, while Triangulation uses viewpoint and distance information to determine precise 3D positions. Combining these two techniques allows SFM to produce accurate and detailed 3D reconstructions of observed objects, enabling more in-depth analysis and visualization. Camera position estimation and Triangulation are critical components in SFM that will allow the reconstruction of 3D objects from 2D images, providing a powerful tool for various applications in computer vision and image analysis.

4) Bundle Adjustment

Bundle Adjustment [17] is the last process in the Structure from Motion (SFM) workflow and plays a vital role in fine-tuning the 3D reconstruction. The main objective of this process is to fine-tune the camera pose and sparse point cloud by minimizing the re-projection errors that result from an improper matching process. This helps to improve the quality of the generated 3D structure and optimize the camera parameters within the SFM framework.

Bundle Adjustment can be formulated as a non-linear least squares problem to minimize the re-projection error. Re-projection occurs when a 3D point projected back onto an image is not aligned with the actual position of the point on the original image, thus causing errors. By reducing these errors, the quality and accuracy of 3D reconstruction can be significantly improved.

This process handles a set of corresponding 3D-2D points, expressed as $\{X_i, x_{ij}\}$ where X_i represents the coordinates of a point in 3D space, while x_{ij} is the projection of that point on the image taken by the *j*th camera [8]. Bundle Adjustment optimizes camera parameters that include camera position and orientation (camera pose) and intrinsic parameters such as focal length and lens distortion. In addition, the process also fine-tunes the position of the sparse point cloud, which is a 3D representation of matching feature points found in the input images, to ensure that they are optimally aligned with their re-projection on all pictures.

As with the Levenberg-Marquardt method, this process is performed iteratively to solve most non-linear minor square problems. This iteration continues until the re-projection error reaches a local minimum or no longer changes significantly. By minimizing the re-projection error, Bundle Adjustment improves the overall accuracy of the 3D reconstruction, ensuring that the resulting 3D model is a highly accurate representation of the real world. In addition, this process also helps to maintain geometric consistency between the various images, ensuring that all photos are optimally aligned with the 3D model. By optimizing the camera parameters, Bundle Adjustment ensures that all cameras in the dataset contribute optimally to the 3D reconstruction, thus improving the quality and accuracy of the final model. Figure 5 shows a flowchart of the Bundle Adjustment process, demonstrating how each component interacts with each other to achieve optimal 3D reconstruction.

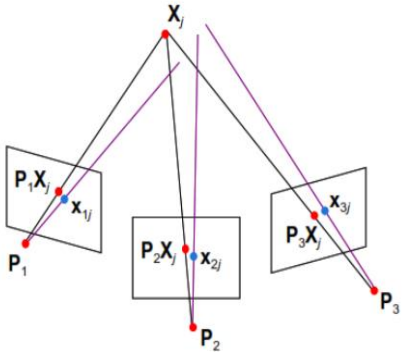


Figure 5. Bundle Adjustment Process

5) Multi-View Stereo (MVS)

The Multi-View Stereo (MVS) technique effectively solves the limitation of sparse feature matching points in Structure from Motion (SFM) in 3D reconstruction. When SFM suffers from sparse point clouds and unsatisfactory rebuilding due to the lack of matching points, MVS can be an instrumental next step [18].

One of the advantages of MVS is its ability to acquire dense point cloud data. MVS can produce a more complete and detailed 3D representation because it has richer information than the camera pose parameters obtained from SFM. In other words, MVS utilizes the camera pose information from various viewpoints to identify corresponding points in different images more effectively.

The main purpose of using MVS is to increase the density of objects in 3D representations to produce higher-quality reconstructions, especially on objects with complex shapes and good details. By combining stereo and multi-view methods, MVS can create more detailed and precise 3D models, overcoming the constraint of sparse matching points often occurring in SFM.

In this research, MVS becomes an essential step in the 3D reconstruction process, especially when dealing with objects with a high level of complexity. We can obtain more satisfactory and accurate reconstruction results by utilizing the advantages of both SFM and MVS. 3D reconstruction results and the process can be seen in Figure 6.

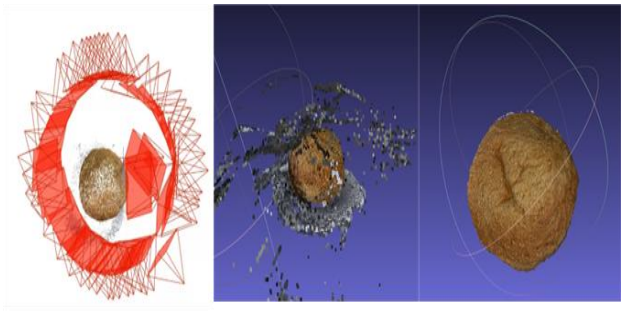


Figure 6. 3D Reconstruction

B. Volume Estimation

1) Slicing 3D Model of Food Object

Slicing is the process of cutting or separating an object into smaller parts. This process facilitates further analysis and processing of the sliced parts. Slicing techniques can be used in various ways, physically using cutting tools or digitally using computer software to cut 3D models into thin layers. In this study, slicing is done using Sketchup software.

After slicing, the next step is segmentation. Segmentation involves grouping pixels or other minor elements in an image or object based on specific characteristics, such as color, brightness, texture, or shape. The primary purpose of segmentation is to identify and separate different parts or components in an image or model. By performing segmentation, different parts of an object can be analyzed independently, allowing for more precise examination and processing. Segmentation is essential in image analysis, object recognition, and computer vision.

When the segmentation process is complete, the next step is calculating the number of pixels in each segmented object using the Bware method. The Bware method, or "Binary Area," is often used in image processing to extract objects based on their specific pixel values. This technique helps to identify and quantify areas of interest in an image by counting pixels that meet specific criteria.

The pixel count results obtained using the Bware method are then used to determine the object's surface area. Surface area is an essential parameter in 3D object analysis, providing detailed information about the object's size, shape, and overall dimensions. Understanding surface area is essential for various applications, including quality control, material usage estimation, and physical properties analysis. More details can be seen in Figure 7 to understand this process visually. This figure illustrates the steps involved in slicing and segmentation, providing an overview of the workflow and highlighting the importance of each step in the context of 3D object analysis. Figure 7 shows an image of a whole loaf of bread that has been sliced and then segmented, and the number of pixels is determined. Through this visualization, we can more easily understand how each stage contributes to the desired result.

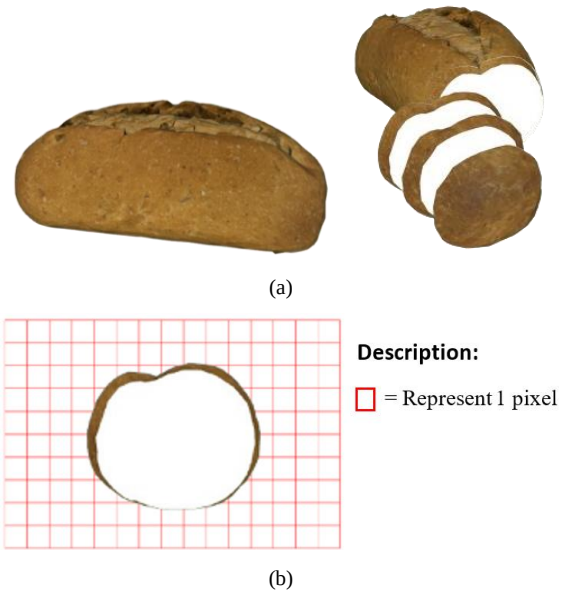


Figure 7. (a) Slicing Process, (b) The number of pixels in an image

2) Calculating the Volume of Each Slicing Section

Calculate the volume of each part of the slicing result using equation (1). If there are five parts, calculate the volume of the five parts using equation (1).

$$V_{Bn} = A(px) \times q \quad (1)$$

V_B = Is the volume of each part, q= Height/thickness based on pixels, and $A(px)$= Surface area based on pixel function.

To get $A(px)$, use equation (2).

$$A(px) = \text{Total number of pixels in the object} \times \text{length of 1 pixel} \quad (2)$$

The unit used is mm^2, so each pixel must also have its length determined in *mm*. To determine the length of 1 pixel using equation (3).

$$Length\ of\ 1\ pixel\ (mm) = \frac{Object\ length\ (mm)}{number\ of\ pixels\ along\ the\ object} \quad (3)$$

Thus, calculating the volume of each part involves:

- Measuring the length of 1 pixel.
- Calculating the surface area based on the number of pixels.
- Using the height or thickness value based on the pixels.

These steps can be applied iteratively for each part resulting from the slicing process, as mentioned in the condition where there are five parts.

3) Calculating the Food Volume

Calculating the volume is the last step in the process, combining the results of the previous sections. The volume of each part has been obtained, and then the estimated volume of the entire food object is to add up all the volumes of each part that have been received. Here is the formula for calculating the volume of the whole object:

$$V = V_{B1} + V_{B2} + V_{B3} + V_{B4} + \ldots\ldots + V_{Bn}$$

$$V = \sum_1^i V_{Bi} \qquad (4)$$

Where V = is the volume of the whole food object, V_B = the volume of each slicing section, and i= the number of pieces.

$V = V_{B1} + V_{B2} + V_{B3} + V_{B4} + \ldots\ldots + V_{Bn}$, for V_{B1} entered is the volume value of the first slicing section, then summed with V_{B2} the volume value of the next slicing section, summed up to the volume of the desired slicing section or V_{Bn} . From the description of the volume equation, the volume equation $\sum_1^i V_{Bi}$ is obtained, which is the Sigma of V_{Bi} from i=1 to i. "Sigma" refers to the summation operation of all slicing section volumes (V_{Bi}) from i=1 to i, the total number of pieces.

4. RESULT AND DISCUSSION

The performance evaluation of the 3D reconstruction method to assist the proposed food volume estimation was conducted through a series of experiments on ten different food dishes. Each primary stage of the process, namely 3D reconstruction, cropping, and volume estimation, is given a subsection to ensure a comprehensive evaluation. First, in the 3D reconstruction stage, images of ten food objects were taken using the integrated camera by using iPhone Xr, with a resolution of 3024 x 4032 pixels per image, and multiple images were taken from different angles for each object. This step is essential to obtain enough data to perform 3D reconstruction with sufficient detail. The time taken to perform 3D reconstruction of each object varies depending on the number of images inputted. The greater the number of images inserted for each object, the longer it takes to generate the 3D model. More detailed information can be found in Table 1 for a more in-depth understanding.

After the process 3D reconstruction, cuts are made on each reconstructed food object to obtain small parts that can be measured for volume separately. This is necessary to enable more accurate and detailed volume estimation of each part of the object. After the cutting stage, the proposed method performs volume estimation for each food part. This volume estimation is based on the pre-constructed 3D data and the information obtained from the cutting process.

Evaluating process of each stage it complite this method's performance is carried out carefully and systematically, taking into account various aspects such as the number of images used, the time required, the accuracy of reconstruction, and the accuracy of volume estimation. This aims to ensure that the proposed method can produce accurate and reliable food volume estimation for various kinds of food.

TABLE I. 3D RECONSTRUCTION TIME BASED ON THE NUMBER OF IMAGES

Food Item	Number of Images	Time (s)
Yellow Cake	55	3420
Risol	38	1320
Panada	56	3550
Fried tofu	46	2820
Fried tempeh	38	1428
Fried chicken breast	50	3102
Hard-boiled egg	45	2760
Milkfish	48	3215
Nugget	42	2415
Burger	40	1530

Table 2 shows the volume estimation results for the ten food items in millimeters. These volumes were compared with direct measurements using measuring instruments. To determine the accuracy of the volume estimation, the estimation error equation in equation (5).

$$\text{Error rate} = \frac{|V_e - V_g|}{V_g} \qquad (5)$$

where V_e is the estimated volume, and V_g is the ground truth volume or the actual volume obtained from direct measurement. Using this equation, we can determine the significant difference between the estimated and actual volumes to evaluate the reliability and accuracy of the method used [19].

By comparing volume estimation results with direct measurements allows us to assess the extent to which the developed volume estimation method can produce values close to reality. Accurate volume estimation is crucial to ensure that further analysis, such as determining calorie or nutrient content, can be conducted confidently. By using high-precision measuring instruments as a benchmark provides a solid foundation for evaluating the performance of the proposed estimation method. Additionally, calculating estimation errors using the provided equations enables the identification of areas where the technique may require further refinement to improve its accuracy.

This assessment of reliability and accuracy provides valuable insights into how well the volume estimation method performs under real-world conditions. The results of this analysis not only indicate the existing error level but also help identify factors that influence the estimation's accuracy. Thus, this research significantly contributes to developing better and more reliable methods for food volume estimation, which can be applied in various

contexts, including automated diet monitoring and effectively calorie consumption management.

TABLE II. COMPARISON OF THE ACCURACY OF MANUAL VOLUME ESTIMATION AND VOLUME ESTIMATION USING THE SYSTEM

Food Item	Volume (mm^3)		Error
	System	The Ground Truth	
Yellow Cake	98.2	92	6.7%
Risol	163.6	155	5.5%
Panada	130.7	124	5.4%
Fried tofu	80.7	84	4.0%
Fried tempeh	95.3	100	4.7%
Fried chicken breast	153.2	143	7.1%
Hard-boiled egg	110.3	116	4.9%
Milkfish	170.3	159	7.1%
Nugget	82.5	87	4.1%
Burger	172.5	162	6.5%

In table 2 shows detailed data comparing the volumes measured using the system with the actual volumes in the field for different types of food and the associated estimation errors. These results demonstrate the system's performance system's performance in various contexts and provide a deep insight into its effectiveness.

The volume estimated by the developed method is close to the ground truth volume for most foods tested. The new method generally shows good accuracy with an error range between 4.0% to 7.1%, which means that the method is reliable for most types of food. However, there is still some variation depending on the type of food being measured. One example of the estimated volume of a yellow cake is 98.2 mm³ compared to its ground truth volume of 92 mm³, resulting in an estimation error of 6.7%; this estimation error shows that the method is quite accurate in measuring the volume of foods that have relatively simple shapes and textures.

The method performed very well on some food items, such as fried tofu and nuggets, with estimation errors of 4.0% and 4.1%, respectively. These results show that the method is very reliable for these items, possibly due to the more homogeneous shape and texture of the food, which facilitates the volume measurement process.

There is some variation in the level of accuracy achieved by this method. As with the other food types, the risol, panada, hard-boiled egg, and burger show moderate estimation errors, ranging from 5.4% to 6.5%. These errors are still within acceptable limits for many practical applications, indicating that the method is quite flexible and adaptable to different shapes and sizes of food.

However, some foods show higher estimation errors. Fried chicken breast, for example, has an estimation error of 7.1%, and milkfish has an estimation error of 7.1%. These higher errors may be due to the more irregular shapes and varying densities of these foods, which make the volume estimation process more complex.

Overall, our proposed method showed promising performance relatively low estimation errors for most of the tested foods. The accuracy of the method demonstrated its potential as a more reliable and accurate tool compared to traditional dietary assessment methods. The proposed new 3D model-based method achieved an average volume estimation error of 7.1%. Given that the estimation error in traditional dietary assessment methods can exceed 50% [20][21], this developed method offers a significant improvement.

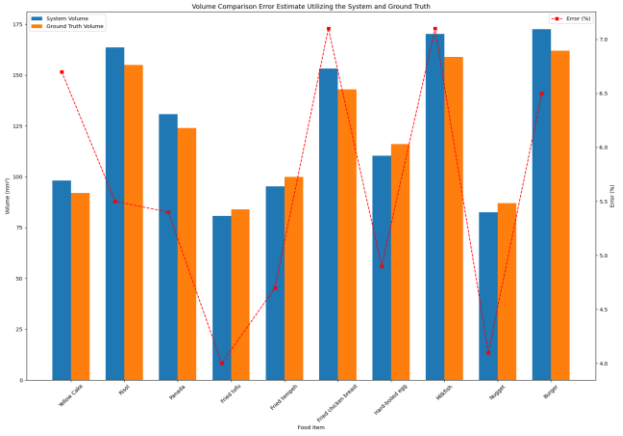


Figure 8. Accuracy graph of the comparison of volume estimation using the system and manual volume estimation

The graphical representation of this data reinforces these findings by clearly visualizing the comparison between the estimated volume, the ground truth volume, and the percentage error. The graph in Figure 8 shows that, although there are variations in the level of accuracy, the developed method generally gives better results.

CONCLUSION

This study successfully develops and evaluates a 3D reconstruction method using the SFM-MVS technique for food volume estimation. The proposed method has been extensively tested on ten different types of food, demonstrating its effectiveness in providing accurate volume estimation.

The results show that the estimated volume is generally close to the ground truth volume, with estimation errors ranging from 4.0% to 7.1%. This means that the developed method is reliable for most types of food.

The 3D model-based method developed in this study offers significant improvements over traditional dietary assessment techniques, providing a more accurate and reliable tool for food volume estimation. Future work could focus on refining the method to further reduce estimation errors for more complex food items and explore

its application in real-time diet monitoring and assessment systems.

REFERENCES

[1] Y. Liu *et al.*, "Food Volume Estimation Based on Reference," *ACM Int. Conf. Proceeding Ser.*, pp. 84–89, 2020, doi: 10.1145/3390557.3394123.

[2] B. Amoutzopoulos *et al.*, "Portion size estimation in dietary assessment: a systematic review of existing tools, their strengths and limitations," *Nutr. Rev.*, vol. 78, no. 11, pp. 885–900, 2020, doi: 10.1093/nutrit/nuz107.

[3] C. K. Martin, S. Kaya, and B. K. Gunturk, "Quantification of food intake using food image analysis," *Proc. 31st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. Eng. Futur. Biomed. EMBC 2009*, pp. 6869–6872, 2009, doi: 10.1109/IEMBS.2009.5333123.

[4] M. Puri, Z. Zhu, Q. Yu, A. Divakaran, and H. Sawhney, "Recognition and volume estimation of food intake using a mobile device," *2009 Work. Appl. Comput. Vision, WACV 2009*, 2009, doi: 10.1109/WACV.2009.5403087.

[5] C. Xu, Y. He, A. Parra, E. Delp, N. Khanna, and C. Boushey, "Image-based food volume estimation," *CEA 2013 - Proc. 5th Int. Work. Multimed. Cook. Eat. Act.*, pp. 75–80, 2013, doi: 10.1145/2506023.2506037.

[6] J. Dehais, M. Anthimopoulos, S. Shevchik, and S. Mougiakakou, "Two-view 3D reconstruction for food volume estimation," *IEEE Trans. Multimed.*, vol. 19, no. 5, pp. 1090–1099, 2017, doi: 10.1109/TMM.2016.2642792.

[7] A. Gao, F. P. W. Lo, and B. Lo, "Food volume estimation for quantifying dietary intake with a wearable camera," *2018 IEEE 15th Int. Conf. Wearable Implant. Body Sens. Networks, BSN 2018*, vol. 2018-Janua, no. March, pp. 110–113, 2018, doi: 10.1109/BSN.2018.8329671.

[8] Y. Cai, M. Cao, L. Li, and X. Liu, "An End-to-End Approach to Reconstructing 3D Model from Image Set," *IEEE Access*, vol. 8, pp. 193268–193284, 2020, doi: 10.1109/ACCESS.2020.3032169.

[9] F. Guo, J. Yang, Y. Chen, and B. Yao, "Research on image detection and matching based on SIFT features," *2018 3rd Int. Conf. Control Robot. Eng. ICCRE 2018*, pp. 130–134, 2018, doi: 10.1109/ICCRE.2018.8376448.

[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004, doi: 10.1023/B:VISI.0000029664.99615.94.

[11] J. Feng, C. Ai, Z. An, Z. Zhou, and Y. Shi, "A feature detection and matching algorithm based on Harris algorithm," *Proc. - 2019 Int. Conf. Commun. Inf. Syst. Comput. Eng. CISCE 2019*, pp. 616–621, 2019, doi: 10.1109/CISCE.2019.00144.

[12] V. K. Mali, P. Venu, M. K. Nagaraj, and S. N. Kuiry, "Demonstration of structure-from-motion (SfM) and multi-view stereo (MVS) close range photogrammetry technique for scour hole analysis," *Sadhana - Acad. Proc. Eng. Sci.*, vol. 46, no. 4, 2021, doi: 10.1007/s12046-021-01758-2.

[13] Q. Yu, C. Yang, and H. Wei, "Part-Wise AtlasNet for 3D point cloud reconstruction from a single image," *Knowledge-Based Syst.*, vol. 242, p. 108395, 2022, doi: 10.1016/j.knosys.2022.108395.

[14] M. A. Fischler and R. C. Bolles, "Random sample consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981, doi: 10.1145/358669.358692.

[15] S. Mills, "Four- and seven-point relative camera pose from oriented features," *Proc. - 2018 Int. Conf. 3D Vision, 3DV 2018*, pp. 218–227, 2018, doi: 10.1109/3DV.2018.00034.

[16] J. Chen, D. Wu, P. Song, F. Deng, Y. He, and S. Pang, "Multi-View Triangulation: Systematic Comparison and an Improved Method," *IEEE Access*, vol. 8, pp. 21017–21027, 2020, doi: 10.1109/ACCESS.2020.2969082.

[17] L. Zhou, Z. Zhang, H. Jiang, H. Sun, H. Bao, and G. Zhang, "Dp-mvs: Detail preserving multi-view surface reconstruction of large-scale scenes," *Remote Sens.*, vol. 13, no. 22, pp. 1–20, 2021, doi: 10.3390/rs13224569.

[18] A. C. Review, "Large-Scale 3D Reconstruction from Multi-View Imagery :," pp. 1–38, 2024.

[19] C. Xu, Y. He, N. Khanna, C. J. Boushey, and E. J. Delp, "MODEL-BASED FOOD VOLUME ESTIMATION USING 3D POSE School of Electrical and Computer Engineering , Purdue University Department of Electronics and Communication Engineering , Graphic Era University , Dehradun , India Cancer Epidemiology Program , University," pp. 2534–2538, 2013.

[20] T. E. Schap, B. L. Six, E. J. Delp, D. S. Ebert, D. A. Kerr, and C. J. Boushey, "Adolescents in the United States can identify familiar foods at the time of consumption and when prompted with an image 14 h postprandial, but poorly estimate portions," *Public Health Nutr.*, vol. 14, no. 7, pp. 1184–1191, 2011, doi: 10.1017/S1368980010003794.

[21] C. J. Boushey, D. A. Kerr, J. Wright, K. D. Lutes, D. S. Ebert, and E. J. Delp, "Use of technology in children's dietary assessment," *Eur. J. Clin. Nutr.*, vol. 63, no. March, pp. S50–S57, 2009, doi: 10.1038/ejcn.2008.65.

Nurdzakirah Amir is a student at Hasanuddin University of Makassar, Indonesia. He received his bachelor's degree in informatics engineering from PLN Institute of Technology, West Jakarta, Indonesia in 2020. His research interests focus on Computer Vision, 3D Reconstruction.

Zahir Zainuddin holds a Doctor of Computer Engineering from Bandung Institute of Technology, Indonesia, in 2004. He also received his B.Sc. in Electrical Engineering Department, Hasanuddin University, Indonesia, in 1988 and his M.Sc. (Computer Engineering) from Florida Institute of Technology USA in 1995. He is an associate professor at the Department of Informatics at Hasanuddin University in Indonesia. His research includes Computer Systems, intelligent systems, computer vision, and smart cities. He has published over 60 papers in international journals and conferences. In 1989, he was a JSPS research fellow at the Tokyo Institute of Technology.

Zulkifli Tahir received B. Eng. from Institut Teknologi Telkom in 2006, Master degree in industrial computing from Universiti Teknikal Malaysia and doctoral degree in computing engineering from Ehime University. Published more than fifty publications in form of book chapter, research reports, journals and international conference papers. Currently active as lecturer at the Hasanuddin University. Research interest include decision support system, artificial intelligence, distributed computer network and applied multimedia.