Green University
GUBJSE: ISSN: 2409-0476

Food Depth Estimation Using Low-Cost Mobile-Based System for Real-Time Dietary Assessment

D.M.S. Zaman, Md. Hasan Maruf, Md. Ashiqur Rahman, Jannatul Ferdousy, ASM Shihavuddin

***Abstract*—** Real time estimation of nutrition intake from regular food items using mobile-based applications could be a break-through in creating public awareness of threats in overeating or faulty food choices. The bottleneck in implementing such systems is to effectively estimate the depths of the food items which is essential to calculate the volumes of foods. Volumes and density of food items can be used to estimate the weights of food eaten and their corresponding nutrition contents. Without specific depth sensors, it is very difficult to estimate the depth of any object from a single camera. Such sensors are equipped only in very advanced and expensive mobile devices. This work investigates the possibilities of using regular cameras to calculate the same using a specific frame structure. We proposed a controlled camera setup to acquire overlapping images of the food from different positions already calibrated to estimate the depths. The results were compared with the Kinect device's depth measures to show the efficiency of the proposed method. We further investigated the optimum number of camera positions, their corresponding angles, and distances from the object to propose the best configuration for suchn a controlled system of image acquisition with regular mobile cameras. Overall the proposed method presents a low-cost solution to the depth estimation problem and opens up the possibilities for mobile-based apps for dietary assessment for various health-related problem-solving.

***Index Terms*—** Depth estimation, 3D reconstruction, nutrition analysis, 2D segmentation, 3D triangulation, image segmentation.

I. Introduction

GLOBAL food production per year has increased significantly in recent years meeting estimated demands however not necessarily equally distributed as hoped [1]. Foods with essential nutritional contents are easily available at the nearest supermarkets or groceries in most of the countries. However due to lack of awareness of exceeding sugar or fat contents in some of these easily available images, and together with aggressive marketing by some of the giant food processing companies, people tend to make wrong food choices increasing the probability of getting diseases such as obesity, diabetics, etc. People in modern first-world societies are faced with diet-related chronic diseases either caused by over-eating or simply having a homogeneous unhealthy diet. People have been increasingly suffering from these types of diseases since the last decade [2]. As a consequence overall public health condition is in critical condition that could be prevented having useful ICT tools at hand. Hence monitoring the consumption of food and its effect on our health is a crucial part of possible solutions to these problems.

Weighing, identifying, and calculating the nutritional values of food items can be time-consuming and complicated using traditional ways. Therefore scholars have sought to develop various methods that use common mobile-phones as a tool to estimate the nutritional value of a meal [3, 4, 5, 6, 7]. However, the main bottlenecks of such methods are effectively identifying food and estimating volumes from single images that can be used to further study of the nutrition intake. Dietary assessment is the procedure of monitoring the food intake of individual persons throughout a day/week/month depending on requirements. In this work, we focused on the development of dietary assessment by depth estimation method after analyzing the food images. As depth estimation from multiple images is already a well-studied field and widely considered to be computationally heavy. Depth can be determined through some typical dedicated devices but these types of sensors are immobile, energy-consuming [8] and expensive. However, it is convenient to estimate food volume from images captured by a depth camera or a mobile phone with a depth-sensor or a simple mobile camera. The proposed method deploys a control structure to acquire images with a single camera on multiple locations calibrated beforehand to an estimated depth of segmented food. For detection of the common interest points on the acquired optical images, HAHOG features [9, 10] are being used and described in detail in the method section. Though there had been plenty of research works addressing

This paper was received on 05 May 2020, revised on 01 October 2020, and accepted on 13 October 2020. This work was supported financially by GUB.

D. M. S. Zaman is with the Department of Electrical and Electronic Engineering, Green University of Bangladesh (GUB). E-mail: *saaduzzaman@eee.green.edu.bd.*

Md. Hasan Maruf is with the Department of Electrical and Electronic Engineering, GUB. E-mail: *maruf@eee.green.edu.bd.*

Md. Ashiqur Rahman is with the Department of Electrical and Electronic Engineering, GUB. E-mail: *ashiqur@eee.green.edu.bd.*

Jannatul Ferdousy is with the Department of Electrical and Electronic Engineering, GUB. E-mail: *ferdousy@eee.green.edu.bd.*

ASM Shihavuddin is with the Department of Electrical and Electronic Engineering, GUB. E-mail: *shihav@eee.green.edu.bd.*

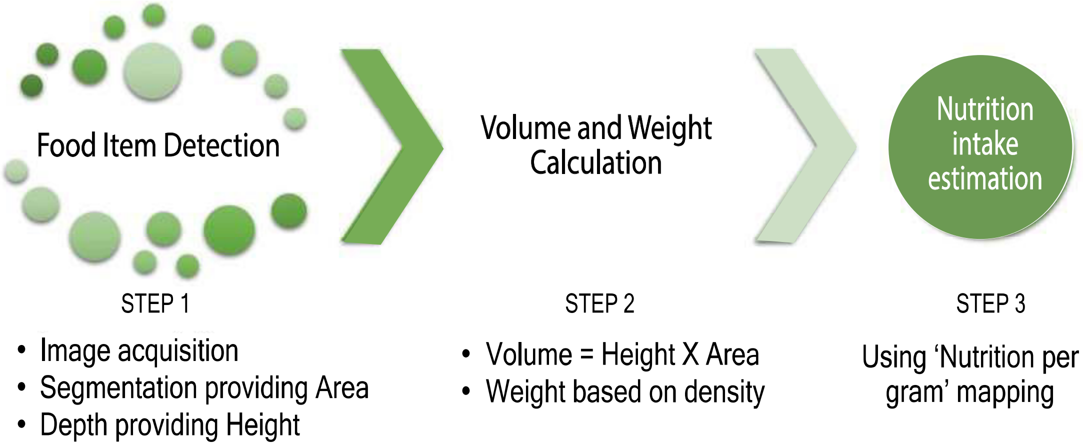
Food Item Detection
Volume and Weight Calculation
Nutrition intake estimation
STEP 1
• Image acquisition
• Segmentation providing Area
• Depth providing Height
STEP 2
• Volume = Height X Area
• Weight based on density
STEP 3
Using 'Nutrition per gram' mapping

Fig. 1: Schematic diagram of the steps from image acquisition to nutrition intake estimation deployable in various mobile applications. Here the Step 1 is the detection of food items on the plate, Step 2 is the the volume and weight estimation of each item and the step 3 is the final nutrition intake calculation from the earlier estimations.

the low-cost depth estimation methods, only a few or none have addressed the problem of finding the right camera angles and depth from objects to create good quality depth maps. To be able to use hand-held mobile cameras on our regular smartphones for depth estimation, it is crucial to know the best possible angles and distances enabling the extraction of good quality image features. In this work, we have investigated these parameters and compared the quality of produced depth maps to propose semi optimal values. These estimations could be used to construct the physical structure for image acquisition of mobile cameras.

From the available state of the segmentation methods as described earlier, it is possible to identify each food item on the plate with their corresponding area in terms of pixels. Together with the depth information from the proposed methods, we can get the height of each food item. From the height and area of each food item, we can estimate the volume in terms of pixels which can be converted into a real volume using the camera resolution information. Based on the density information of the food items, the volumes would represent corresponding weights that can be used to estimate the nutrition intake. This entire process is illustrated in Figure 1. In this work, we investigated and focused on the depth estimation part, assuming the other steps could be implemented following the corresponding state of the art methods available for other parts.

The main contributions of this work are as followings:

- Development of a novel low cost set up for depth estimation in a controlled environment. For this method, with a single camera changing the location on the structure, images can be acquired with a normal household mobile camera for approximate depth estimation.
- A investigation on identifying the workable number of camera positions, angles, and distance from the object to produce the best possible results within the proposed framework and constraints is performed.

II. RELATED WORKS

Depth-camera based approaches are considered as a handy and efficient method to estimate the food volume within reliable approximation. As the name implies it gets an accurate depth map from a single view and has the advantage of knowing the exact scale of the scene, excluding the need for a fiducial marker. This method is computationally fast and robust. The downside is that very few modern mobile phones are equipped with a depth-sensor [11]. However, it is worth noting that this technology is being introduced in 2019 by the world's most established camera phone producers. The Samsung Galaxy S10 G5, iPhone 2020, and Huawei P30 Pro are all equipped with a Time of Flight (ToF) sensor [12]. This sensor emits an infrared signal and measures how long it takes for the light to return and determines the depth from those signals.

Structure from Motion (SfM) [13] is a technique that utilizes a set of 2D images to estimate a 3D structure. Images from most mobile cameras are stored in an Exchangeable image file format. This format stores important information regarding the camera that was used to take the image and the image itself. From this information, it is possible to get a good estimate of the depth of the scene.

There had been few other works on estimating depth from single-camera using deep learning [14, 15, 16], where the model is trained on a large number of original and corresponding depth images. Using the trained model, the depth is estimated based on prior experience without sensing the physical depth of the image at hand. These methods rely heavily on the large number of annotated data-set, performed well on previously seen objects, and also require heavy-duty GPU accelerated hardware. More often than not in real applications, we need lightweight simple solutions to work fast and efficiently and also to be ready to encounter new cases regularly. In that regard, for monitoring foods in restaurants, hospitals, or on personal mobile apps, we proposed a solution framework that is practical, cost-effective, and works without prior training.

Garg *et al.* published a method on utilizing dual-pixel technology and deep learning to get monocular depth estimates [8]. In their approach, they needed a data-set consisting of dual pixel images paired with the depth of that image. Since dual-pixel technology is rather new, they needed to create their own data-set. They did that by using a tool called COLMAP, a general-purpose tool in Structure-from-Motion and Multi-View Stereo pipeline. For gathering images, they built a rig, where 5 mobile-camera phones could be mounted on at once and pictures taken on all phones simultaneously by pressing a button.

To estimate the accurate estimation of nutrition intake, we need to simultaneously segment the food items on the menu as well and map the nutrition estimated based on the corresponding class. 2D segmentation together with depth can provide the volume of each food item that can be converted into weights and sequentially to individual nutrition portions. The tasks of segmentation have been studied in details in [17, 18, 19, 20, 21, 22]. Most of these work deployed SVM [23] of visual features or mainly deep learning-based segmentation model with well established architectures like Alexnet [24], GoogleLeNet from [25], DeepLab from [26], UNet [27] and Mask R-CNN [28]. Final volume estimation to facilitate the approximation of nutritional contents is being only approached by Eigen et al. in [29]. In this work, they used a multi-scale deep network, mapping these artificial neurons into their respective voxel representations. In this way, they were able to combine corresponding segmentation masks to generate volume estimates of the dishes. Also there had been several state of the art datasets published [18, 22, 30, 31, 32] to further progress the research along this direction. In this work, we assumed our dataset to be similar to UNIMIB2016 [18].

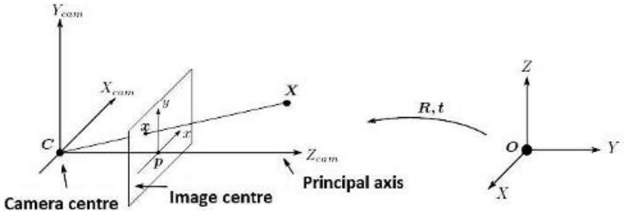

Fig. 2: Mechanism of pin-hole camera.

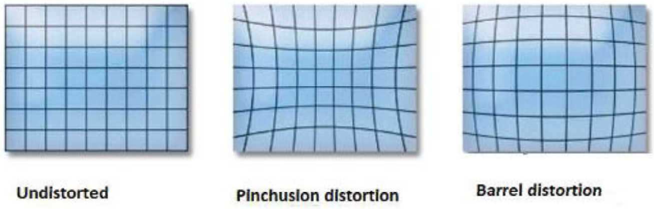

Fig. 3: Barrel and Pincussion distortion.

III. Theory and Methodology

The conventional pin-hole camera model is widely used in computer vision. It is perceived as the light being projected through a pinhole on the side of a box representing the camera center, light traveling through that pinhole is projected to the back of the box which represents the image plane. This would cause the image to be displayed upside down on the back of the box. In figure 6, the projection has been moved on the opposite side of the camera center and flipped horizontally for illustration purposes. This model eliminates the lens and therefore does not consider radial or tangential distortion.

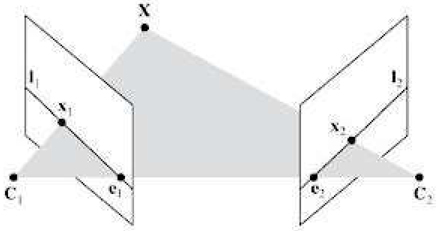

Fig. 4: Epipolar geometry.

When taking images of large sharp-edged objects such as tall buildings, the sharp edges can appear curved. This is due to so-called radial distortion which is illustrated in Figure 3. The radial distortion can be corrected by the two following equations.

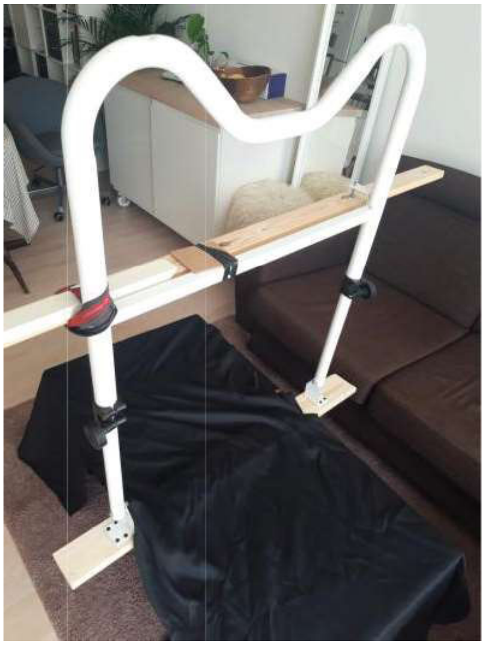

Fig. 5: The rig built to experiment with angles and distances.

$$x_{corrected} = x_c + L(r)(x - x_c) \quad (1)$$

$$y_{corrected} = y_c + L(r)(y - y_c), \quad (2)$$

Where $L(r)$ is an approximated function given by a Taylor expansion $L(r) = 1 + k_1 + K_2 + K_3$, with $r^2 = (x - x_c)^2 + (y - y_c)^2$.

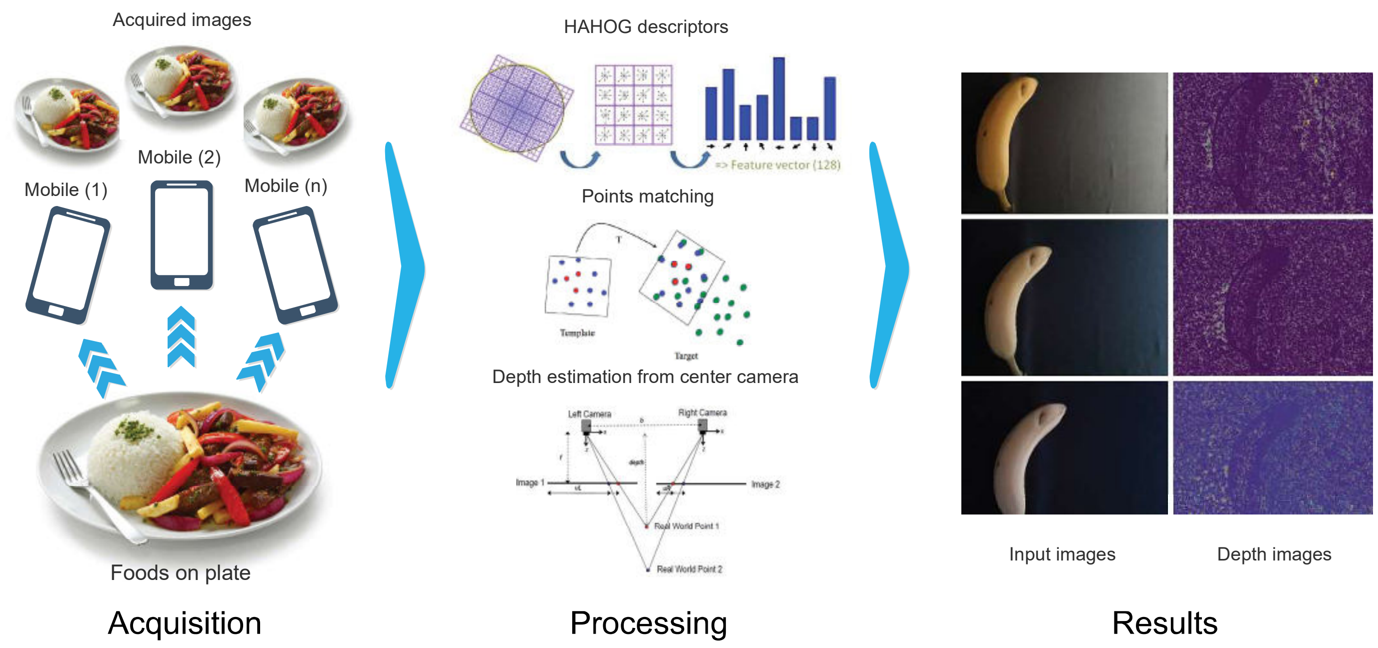

Fig. 6: Schematic of the proposed low cost depth estimation method.

A calibration matrix κ is a $3x3$ matrix containing the cameras intrinsic parameters. κ can be written as follows.

$$\kappa = \begin{bmatrix} \alpha f & s & p_x \\ 0 & f & p_y \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

Where f is the focal length, s is the skew factor. Which is only non-zero when the pixel elements on the camera sensor are skewed. In most cases, s can be set to zero. The aspect ratio α is used for none-square pixel elements and in most mobile cameras it is safe to assume that α = 1. Then the calibration matrix κ is

$$\kappa = \begin{bmatrix} f & s & p_x \\ 0 & f & p_y \\ 0 & 0 & 1 \end{bmatrix} \tag{4}$$

Now remains the focal length the coordinates of the principal point $p = [p_x p_y]$ which represents the optical center where the optical axis intersects the image plane. The camera matrix P, often referred to as the projection matrix is a $3x4$ matrix which can be decomposed as follows.

$$P = \kappa[R|t] \tag{5}$$

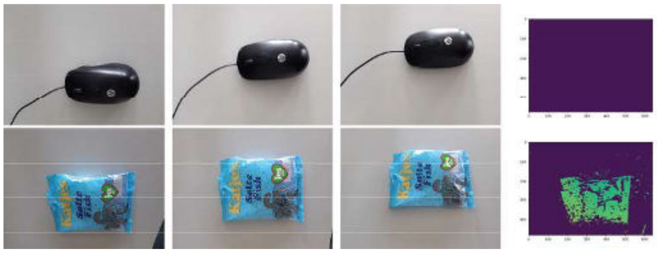

Fig. 7: Experiments on Computer mouse and Candy bag.

Where κ is the calibration matrix from Eq. 3. The 3x3 matrix R is the so-called rotational matrix describing the camera orientation and t is the translation vector describing the position of the camera center. R and t represent the camera's extrinsic parameters. From the camera matrix, we can transfer the point X from world coordinates into the camera coordinates x as follows

$$x = PX = \kappa[R|t]X =$$
$$\begin{bmatrix} f & s & p_x \\ 0 & f & p_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{6}$$

Let us consider two images of the same scene in the usual case we will have matching features and in those features for a given point x_1 in the first image we will have a corresponding point x_2 in the second image. The projection of x_1 in the first image, the position of x_2 in the second image is restricted to the epipolar line l_2 in the second image. All epipolar lines have a common point called the epipole denoted e_1 and e_2 in Figure 4. The epipoles e_1 and e_2 are defined by the intersections between the camera baseline and the image planes. A projection x_1 of a $3D$ point has its corresponding projection x_2 restricted to the epipolar line l_2 going through its epipole e_2. l_1 and l_2 span up an epipolar plane containing the $3D$ point X.

The epipolar constraint is formulated algebraically using the essential matrix E. Let us assume two pinhole cameras P_1 and P_2 and their separate Euclidean coordinate systems. Their coordinate systems can be aligned knowing from corresponding points X_1 in camera C_1 coordinate system is represented as X_2 in camera C_1 coordinate system.

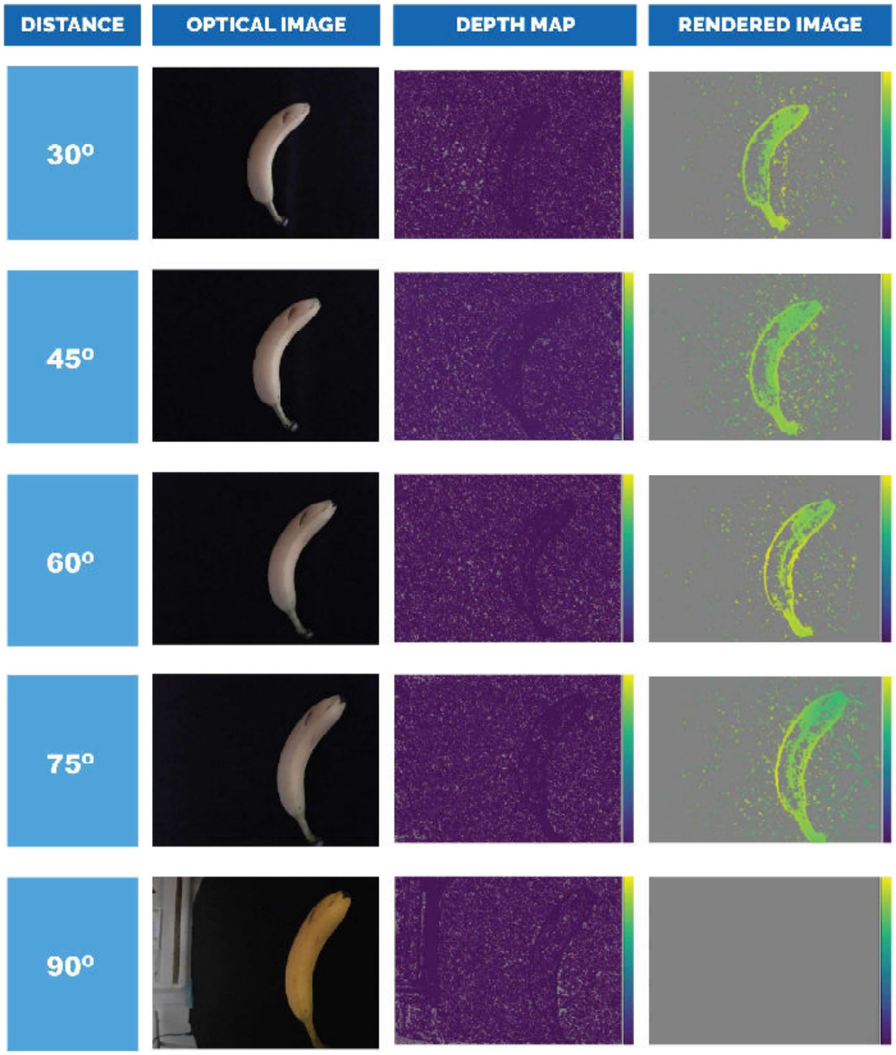

Fig. 8: Depth maps from the banana data set at 30 cm from the surface of the table top.

$$X_1 = RX_2 + t \tag{7}$$

Where R is the $3x3$ rotation matrix and t is the translation vector. By multiplying both sides with $X_1^T S_t$ gives

$$X_1^T S_t X_2 = X^T E X_2, \tag{8}$$

where S_t is the is the skew symmetric matrix of the translation.

$$S_t = \begin{bmatrix} 0 & -t_3 & t_2 \\ t_3 & 0 & -t_3 \\ -t_2 & t_1 & 0 \end{bmatrix} \tag{9}$$

And R is the $3x3$ rotation matrix. Equation 8 also holds for image points x_1 and x_2 which defines the epipolar constraint.

$$X_1^T E X_2 = 0, \tag{10}$$

From equation 14 image points x can be translated to pixel positions u by inversing the camera calibration matrix K.

$$x \sim K^{-1} u, \tag{11}$$

These new image coordinates also satisfy the fundamental equation of 11. Knowing this, we can rewrite the epipolar constraint in terms of pixel positions.

$$(\kappa_1^{-1} u_1)^T E (\kappa_1^{-1} u_2) = 0, \tag{12}$$

$$u_1^T (\kappa_1^{-1T} E \kappa_2^{-1}) u_2 = 0, \tag{13}$$

$$u^T F u_2 = 0 \tag{14}$$

where $F \sim \kappa_1^{-1T} E \kappa_2^{-} 1$ is the fundamental matrix. F is a $3x3$ matrix that has to rank 2 and can be estimated linearly given 8 or more corresponding points.

From equation 14, we see that each point correspondence, $u_i \sim [u_1^i v_1^i 1]^T$ and $[u_2^i v_2^i 1]^T$ generates one constraint on the elements of the fundamental matrix F:

$$\begin{bmatrix} u_2^i & v_2^i & 1 \end{bmatrix} \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{bmatrix} \begin{bmatrix} u_1^i \\ v_1^i \\ 1 \end{bmatrix} = 0 \tag{15}$$

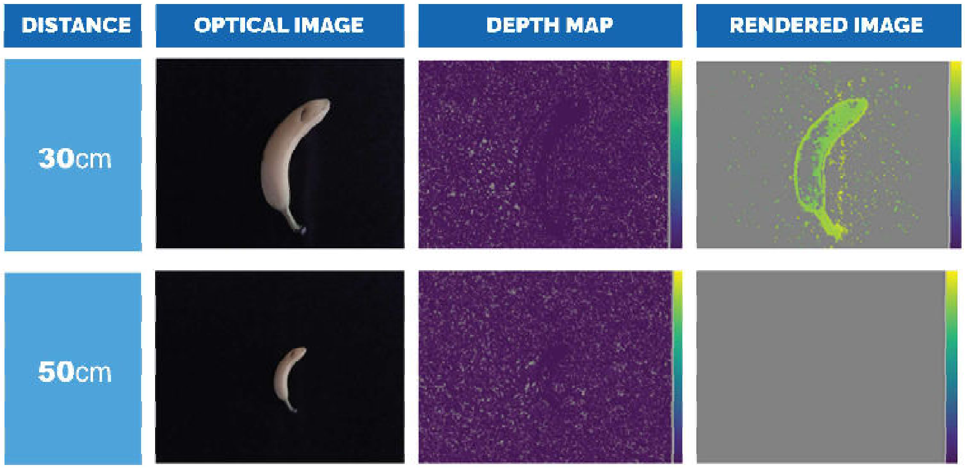

Fig. 9: Depth-maps of the banana data set at an fixed angle of 30° keeping 30 cm and 50 cm away from the surface of the tabletop.

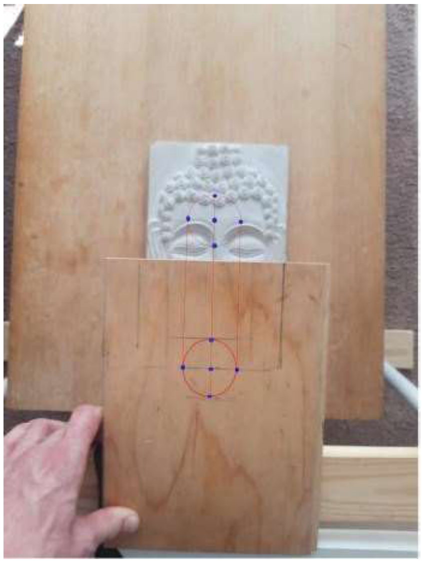

Fig. 10: Back of a rough rig for taking five images with a central image and 4 images 2.5cm from the center $90°$.

For n pairs of correspondences, the constraints can be rearranged as linear system in the 9 unknown elements of the fundamental matrix:

$$Af = \begin{pmatrix} u_1^2u_1^1 & u_1^2v_1^1 & u_1^2 & u_1^2u_1^1 & v_2^1 & u_1^1v_1^1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ u_1^nu_1^n & u_1^nv_1^n & u_1^n & u_1^nu_1^n & v_2^n & u_1^nv_1^n & 1 \end{pmatrix} = 0 \quad (16)$$

Where f is a vector representation of the 9 unknown coefficients of the fundamental matrix. Given 8 or more correspondences a least-squares solution can be found to solve for f.

Projection matrices of two separate views can be retrieved from a fundamental matrix by singular value decomposition. P_1 represents camera matrix 1 and P_2 represents camera matrix 2 for an initial pair of a reconstruction. $P_1 = \kappa_1[I|O]$ $P_2 = \kappa_2[R|t]$

Triangulation is a minimization problem that minimizes the sum of re-projection errors of X_j points in model coordinates to P_i views.

$$min \sum_{i=1}^{m} \sum_{j=1}^{n} d(x_{ij}, P_iX_j)^2 \quad (17)$$

To the depth estimation theory from the feature points triangulation, the proposed method required a fixed structure that enables image acquisition from control distance and angle. This structure needs to be rigid and should contain a mobile camera holding place facing the object of interest. Our experimental design of this structure is made of wood and illustrated in figure 5.

IV. Experiments and Implementation

The fastest way of computing depth estimates is not through multi-view reconstructions rather through single view depth cameras or through neural networks. This work focused both on the accuracy and the optimization of the computational time as well as resources. With future modifications, it would have the potential to be served as a part of a volume estimation software that needed fast computation. In this work, we proposed a method that can produce precise estimation to serve as a novel food image acquisition framework and consecutive food volume calculation technique. The proposed method can estimate depth from each camera position as illustrated in Figure 6 representing the Schematic of our proposed method.

The accuracy of the depth estimates is highly dependent on detected features. Feature detector HAHOG [33] is decided to be used in this work and experimentation on the performance of the proposed framework. HAHOG detector uses a combination of Hessian Affine feature point detection [9] and the HOG descriptors [10]. Single-use of feature detectors keeps the framework simple however also keeps the potential

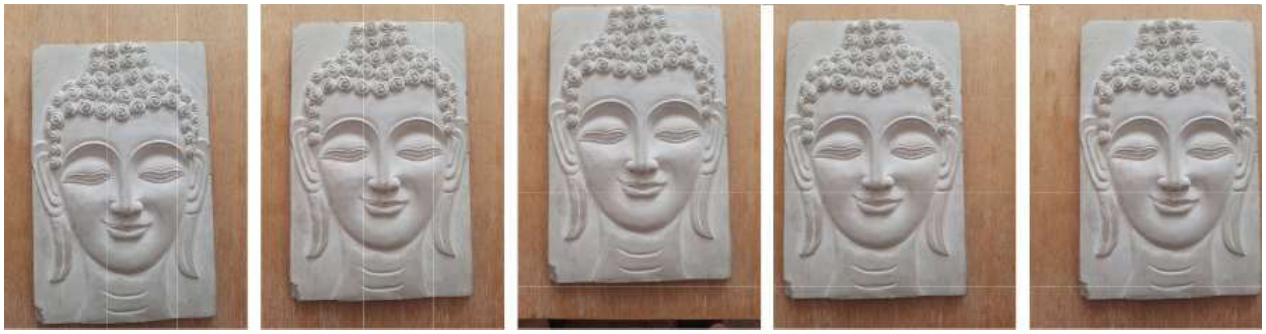

Fig. 11: Five images taken using the configuration as described in Figure 10.

for further improvements in terms of interest points detection and matching in overlapping images. It was realized through experimentation that to acquire accurate depth estimates, the angles between images need to be kept to a minimum. Narrow angles and strong features appear to be ideal in obtaining good density in the depth-map. Dense points help to estimates with transfer function and eventually depths with less uncertainty within the region of interest.

A. 3 *images from random angles*

Images of random objects were taken from 3-5 different angles, sometimes results were obtained and sometimes the pipeline did not give any results. Figure 7 shows how we get a depth estimate of the candy bag but fails to obtain any depth values for the computer mouse. This is presumed to be because of the lack of features in the mouse images. The depth estimate for the candy bag shown in this figure is the depth estimate for the image furthest to the left. It is well known that the SfM method is dependent on the detected features of the set of images as it depends on matching those features to get depth estimates. After trying multiple descriptors for feature detection such as AKAZE, ORB, and HAHOG, the vital role of features of an image in photogrammetry was come to know.

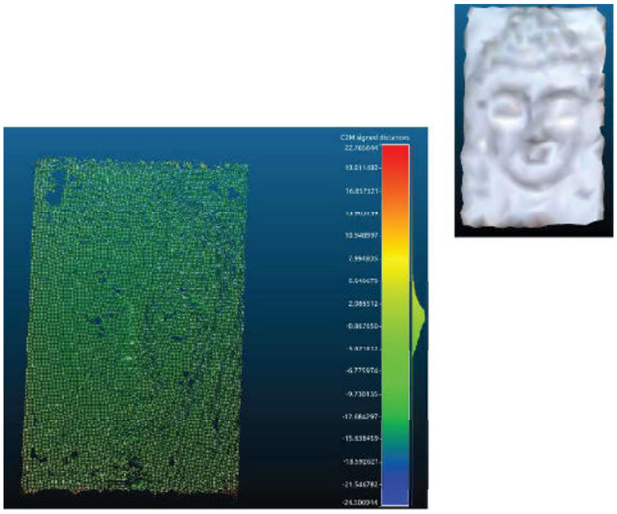

Fig. 12: Mesh surface created from the estimated depth map from images of Figure 11.

B. *Angle and distance variations*

This report investigated the distance from the object which would be ideal to get depth estimation. The idea was to take pictures from the same angles above the horizontal. A data set of images taken of a banana and paprika. Images were captured of each item separately at angles 30°; 45°; 60°; 75° and 90° as well as heights of 30 cm, 50 cm and 70 cm from the surface of the tabletop from each side.

From figure 5, it is seen that the legs can be rotated around the table frame. The camera was mounted in the center of the wooden plank between the two poles and the banana placed in the center of the table.

In figure 8, it is observed that for getting good depth estimates of an object from multiple views, the 15° difference in angle views is suitable for good depth estimates and gets worse as we get further from the object. When the camera gets further away from the object and is rotated at the same angle, the difference of view between images enhances, and matching features become harder to identify between neighboring views. At 90° it was impossible to find feature points from different angles, which results in a flat depth map and rendered image.

From the figure 9, it is worth noting that the only view - from a 50 cm distance to the banana got a cleansed depth map was the top view as none of the images had more than one neighboring view. Images taken at a 70 cm distance from the banana did not give any results. The paprika had to much glance for feature detection and therefore gave no results. After the first two experiments, it is apparent that to get a decent depth map of an object at close range from multiple images, small angles between the neighboring views and enough feature matches between images would be preferred to initialize a reconstruction.

Inspiring from the work of Garg *et al.*, a similar setup for testing a rough rig was built as seen in figure 10. A mat colored object with strong features was picked as well as mounted it on a wooden table with detectable features. The object is referred to as Cleopatra. The lines on the board are marked to align the phone with when each image is taken. Using this setup we get a set of five images of the object which is placed 30cm from the rig as shown in 11.

The Kinect v1 is a depth sensor developed by Microsoft for xBox360 in 2010. We compare our estimates by reconstructing a 3D point cloud of Cleopatra from our cleaned depth-map shown in figure13. To get a structured estimate of Cleopatra

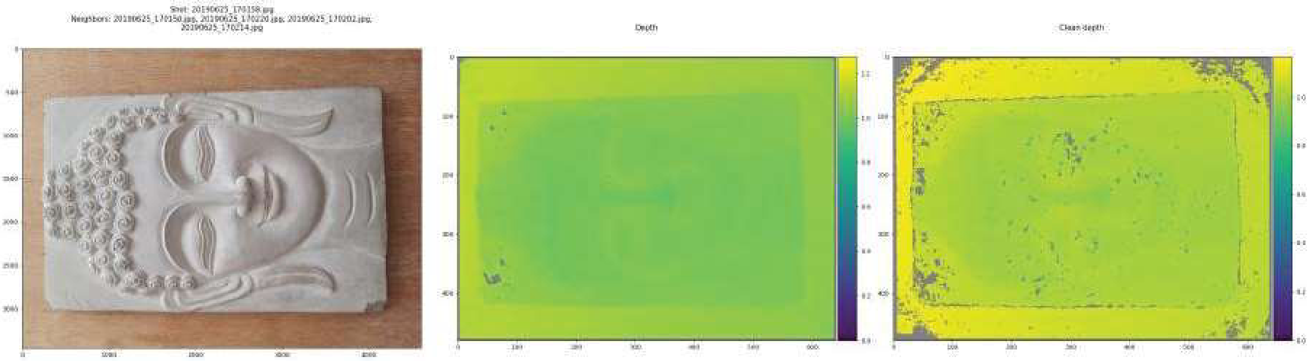

Fig. 13: Depth maps for the central image of Cleopatra. On the left is the optical image of the Cleopatra, in the center is the depth map generated using the proposed method and on the right is the depth map generated using kinect V1.

depth using the Kinect V1 sensor, we scanned the object using the sensor together with Skanect. This gave us a 3D textured mesh.

To compare the output from the Kinect V1 sensor and our reconstruction, CloudCompare as used. We wanted to see how well our reconstruction fitted with the mesh produced by the Kinect sensor and if our method produced a lot of noise and how much our 3D point reconstruction of Cleopatra deviated from the mesh. Focusing on the object we manually cut out the 2D rectangular outlines of Cleopatra and compared them. In order to compare them, we scaled our reconstruction to fit the mesh to an iterative closest point (ICP). Assuming that the scale of the mesh created with Kinect is accurate in global coordinates the average spatial difference of our reconstruction to the mesh being -0.43 mm. The largest errors of min -24.5 mm and max $22 : 77$ mm are at the edges of the area of comparison in our point cloud. The high deviation of spatial difference could be due to the surface of the mesh we are comparing to has evened out flat surfaces as opposed to our point cloud that has sharper edges.

Within the experimentation to identify ideal camera angles and distances from the object of interest to produce the best quality depth maps under the defined resources and constraints, we examined each parameter keeping other variables constant. In this work, the number of camera angles and distances was varied from 30°, 45°, 60°, 75°, and 90° and 30 cm, 50 cm, 70 cm respectively. Following the rules of eliminations, we chose the best option among the mentioned parameter ranges. To validate the performance we used the mean absolute pixel-wise differences between the obtained using the proposed methods and the depth map provided by Kinect v1. The performance is compared based on measured distances in mm scale. The proposed method outperforms the existing works in terms of cost but not in terms of accuracy.

V. Discussions

This report has experimented with an incremental reconstruction method using structure from motion to investigate how it can be used to estimate depth.

1) In IV-A, it is experienced how features are important in order to estimate the orientation and position of a camera.
2) In IV-A it is seen that rotating a camera around a singular axis with 15° difference between angles results in worse estimates of depth-maps as the cameras move further away from the object.
3) The final experiment in IV-B showed that when we take 5 images of a ”strongly featured” scene from close range, we can obtain dense depth-maps.

The depth-maps we produced in our final experiment compared with the ones obtained by a Kinect v1 depth sensor, showed that in an ideal setting, we can get good depth estimates with our method. The quality of a mobile camera enhances the depth of the images. Also better light provides us better depth images. Although a mobile camera providing better resolution images costs high, it will be suitable for better depth maps. In our future work, we will try to develop a mobile App by which we would be able to estimate the depth from the food images as well as the calorie intake in our everyday life. Using the Body Mass Index (BMI), the calorie requirements can be aligned with the calorie intake which alarms the users about overeating or dieting. The mobile app can actually monitor the temporal progress of any user based on the nutrition intake data, weight calculation, a heart condition, exercise information, pre-health condition, etc. and provide informed suggestions or warnings as appropriate. In recent years, there have been lots of demands for such mobile-based applications, however, still, the bottlenecks are AI-based accurate image segmentation and reliable depth estimation from hand-held cameras. In this work, we focused mainly on the 2nd mentioned challenge and came up with a low-cost solution in addition to a suggestion for workable camera angles and distances.

The main cost of a depth estimation system comes from expensive dual camera setup where in our proposed method we can do the same with our already existing handhold mobile camera. The only cost with occur in the physical structure setup as described in the method which generates a very small portion of cost compared to optical cameras. And also our mobile cameras are now a days equipped with high powered sensors which is very useful for making very accurate detection and depth estimation.

VI. Conclusion

The aim of this project was to create the depth-maps from multiple images taken from different angles of a scene and compare the results with a Kinect v1 sensor. Based on our experiments, we may say that that a low cost set up is developed for depth estimation in various camera positions, angles, and distances from the object which is a better way than the other photogrammetry methods. In comparing our results to the Kinect, we can only say that when we took five images from a close range of an object, called Cleopatra in 10, the results that our method created a shape that was similar to the shape that the Kinect created of the same object. Even though the developed method doesn't outperform other photogrammetry methods, it can definitely provide a low-cost solution assuming accessibility of a hand-held mobile camera in regular smartphones. For the proposed solution, the only hardware needed is a fixed structure to place the camera such as shown in figure 8, costing up to a maximum of 20-30 USD. In contrast, other methods need a dedicated dual camera setup such as Kinect v1 costing up to 10 times more than the proposed system.

Contributions

A.S. and M.H.M. conceived the method. D.Z., M.A.R., and J.F. did the experiments and acquired the images. A.S. and D.Z. wrote the manuscript. All authors edited the manuscript.

References

1. Xiaoyong Zhang, Hans Dagevos, Yuna He, Ivo Van der Lans, and Fengying Zhai. Consumption and corpulence in china: a consumer segmentation study based on the food perspective. *Food Policy*, 33(1):37–47, 2008.
2. Chang Xu, Ye He, Nitin Khannan, Albert Parra, Carol Boushey, and Edward Delp. Image-based food volume estimation. In *Proceedings of the 5th international workshop on Multimedia for cooking & eating activities*, pages 75–80, 2013.
3. Kaylen J Pfisterer, Robert Amelard, Audrey G Chung, Braeden Syrnyk, Alexander MacLean, and Alexander Wong. Fully-automatic semantic segmentation for food intake tracking in long-term care homes. *arXiv preprint arXiv:1910.11250*, 2019.
4. Parisa Pouladzadeh, Abdulsalam Yassine, and Shervin Shirmohammadi. Foodd: food detection dataset for calorie measurement using food images. In *International Conference on Image Analysis and Processing*, pages 441–448. Springer, 2015.
5. Shaobo Fang, Chang Liu, Khalid Tahboub, Fengqing Zhu, Edward J Delp, and Carol J Boushey. ctada: The design of a crowdsourcing tool for online food image identification and segmentation. In *2018 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)*, pages 25–28. IEEE, 2018.
6. Yu Wang, Ye He, Carol J Boushey, Fengqing Zhu, and Edward J Delp. Context based image analysis with application in dietary assessment and evaluation. *Multimedia tools and applications*, 77(15):19769–19794, 2018.
7. Joachim Dehais, Marios Anthimopoulos, and Stavroula Mougiakakou. Food image segmentation for dietary assessment. In *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*, pages 23–28, 2016.
8. Rahul Garg, Neal Wadhwa, Sameer Ansari, and Jonathan T Barron. Learning single camera depth estimation using dual-pixels. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7628–7637, 2019.
9. Guoshen Yu and Jean-Michel Morel. A fully affine invariant image comparison method. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1597–1600. IEEE, 2009.
10. Rui Hu and John Collomosse. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *Computer Vision and Image Understanding*, 117 (7):790–806, 2013.
11. Shahram Izadi, David Molyneaux, Otmar Hilliges, David Kim, Jamie Daniel Joseph Shotton, Stephen Edward Hodges, David Alexander Butler, Andrew Fitzgibbon, and Pushmeet Kohli. Reducing interference between multiple infra-red depth cameras, January 26 2016. US Patent 9,247,238.
12. Miles Hansard, Seungkyu Lee, Ouk Choi, and Radu Patrice Horaud. *Time-of-flight cameras: principles, methods and applications*. Springer Science & Business Media, 2012.
13. Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016.
14. Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European Conference on Computer Vision*, pages 740–756. Springer, 2016.
15. Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5162–5170, 2015.
16. Anirban Roy and Sinisa Todorovic. Monocular depth estimation using neural regression forest. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5506–5514, 2016.
17. Christopher J.C. Burges. A tutorial on support vector machines for pattern recognition. 2009. doi: 10.1.1.18.1083.
18. Gianluigi Ciocca, Paolo Napoletano, and Raimondo Schettini. Food recognition: a new dataset, experiments and results. *IEEE Journal of Biomedical and Health Informatics*, 2017. doi: 10.1109/JBHI.2016.2636441.
19. Yining Deng and B. S. Manjunath. Unsupervised segmentation of color-texture regions in images and video. *Ieee Transactions on Pattern Analysis and Machine Intelligence*, 23(8):800–810, 2001. ISSN 19393539, 01628828, 21609292. doi: 10.1109/34.946985.
20. Fengqing Zhu, Marc Bosch, Tusa Rebecca Schap, Nitin Khanna, David S. Ebert, Carol J. Boushey, and Edward J.

Delp. Segmentation assisted food classification for dietary assessment. *Proceedings of Spie - the International Society for Optical Engineering*, 7873(1):78730B, 2011. ISSN 1996756x, 0277786x. doi: 10.1117/12.877036.

21. Austin Myers, Nick Johnston, Vivek Rathod, Anoop Korattikara, Alex Gorban, Nathan Silberman, Sergio Guadarrama, George Papandreou, Jonathan Huang, and Kevin Murphy. Im2calories: Towards an automated mobile vision food diary. *Proceedings of the Ieee International Conference on Computer Vision*, 2015:7410503, 1233–1241, 2015. ISSN 15505499, 23807504. doi: 10.1109/ICCV.2015.146.
22. Oscar Beijbom, Neel Joshi, Dan Morris, Scott Saponas, and Siddharth Khullar. Menu-match: Restaurant-specific food logging from images. *Proceedings - 2015 Ieee Winter Conference on Applications of Computer Vision, Wacv 2015*, pages 7045971, 844–851, 2015. ISSN 24726737. doi: 10.1109/WACV.2015.117.
23. Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.
24. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 2:1097–1105, 2012. ISSN 10495258.
25. Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *Proceedings of the Ieee Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-:7298594, 1–9, 2015. ISSN 2332564x, 10636919. doi: 10.1109/CVPR.2015.7298594.
26. Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. page 14, 2016.
27. Dario Allegra, Marios Anthimopoulos, Joachim Dehais, Ya Lu, Filippo Stanco, Giovanni Maria Farinella, and Stavroula Mougiakakou. A multimedia database for automatic meal assessment systems. In *International Conference on Image Analysis and Processing*, pages 471–478. Springer, 2017.
28. Yang Yu, Kailiang Zhang, Li Yang, and Dongxing Zhang. Fruit detection for strawberry harvesting robot in non-structural environment based on mask-rcnn. *Computers and Electronics in Agriculture*, 163:104846, 2019.
29. David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. 2015.
30. Giovanni Maria Farinella, Dario Allegra, and Filippo Stanco. A benchmark dataset to study the representation of food images. *Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8927:584–599, 2015. ISSN 16113349, 03029743. doi: 10.1007/978-3-319-16199-0_41.
31. Giovanni Maria Farinella, Dario Allegra, Marco Moltisanti, Filippo Stanco, and Sebastiano Battiato. Retrieval and classification of food images. *Computers in Biology and Medicine*, 77:23–39, 2016. ISSN 18790534, 00104825. doi: 10.1016/j.compbiomed.2016.07.006.
32. Gianluigi Ciocca, Paolo Napoletano, and Raimondo Schettini. Learning cnn-based features for retrieval of food images. *Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10590:426–434, 2017. ISSN 16113349, 03029743. doi: 10.1007/978-3-319-70742-6_41.
33. Jhacson Meza, Andrés G Marrugo, Enrique Sierra, Milton Guerrero, Jaime Meneses, and Lenny A Romero. A structure-from-motion pipeline for topographic reconstructions using unmanned aerial vehicles and open source software. In *Colombian Conference on Computing*, pages 213–225. Springer, 2018.

D. M. S. Zaman received the B.Sc. (Hons.) and M.S. (Thesis) degrees in physics from the Department of Physics of Jahangirnagar University, Bangladesh. His M. S. thesis was mainly concerned with nuclear-acoustic waves in self-gravitating degenerate quantum plasmas. He was awarded “National Science and Technology Fellowship 2016-17” for his research work. He is currently working as a Lecturer of Physics with the Electrical and Electronic Engineering Department, GUB. He has already authored four research articles in prestigious international peer-reviewed journals like European Physical Journal Plus, Chinese Physics B, Journal of the Physical Society of Japan, and High Temperature. Mr. Zaman is also working as a member of the editorial board of the International Journal of Mathematical Physics. He is an honorary member of the research group ”SELF” (Self Education and Learning Forum). His research interests include Medical Physics and Biophysics, Plasma Physics, and Material Science.

Md. Hasan Maruf has been working as an Assistant Professor of the EEE department at Green University of Bangladesh (GUB) for six years. Besides his full-time academic teaching, he is also functioning as a Program Coordinator since 2015. His leadership skill in the academic field is very sturdy and stirring to others. In fact, he attended and completed more than 10 academic teaching and learning workshops.

Mr. Maruf has several publications in renowned journals. He has also presented his work at prestigious international conferences. His current research is focused on high-speed and low power CMOS design techniques, Memory, and Memristor design. He has vast knowledge on-chip designing and tape-out through his master’s study. He is an active member of IEEE from his academic life.

Mr. Maruf received his B.Sc. degree in Electrical and Electronic Engineering (EEE) from American International University-Bangladesh (AIUB) in 2010 and an M.Sc. degree in Electrical Engineering (Specification: Communication Electronics) from Linköping University (LiU), Sweden in 2013. Now he is working toward his Ph.D. degree in the Electrical Engineering department at the Islamic University of Technology (IUT), Bangladesh.

Md. Ashiqur Rahman has completed his graduation from the Department of Electrical and Electronic Engineering (EEE) of the Islamic University of Technology (IUT). He is currently working as the Lecturer of the Department of EEE of the Green University of Bangladesh. His research interests include biomedical signal processing, biomedical image processing, and biophysics.

Jannatul Ferdousy has been working as a lecturer of Physics in EEE Department at Green University of Bangladesh since September 2018. She received her B.Sc (Hons) and MS degree in Physics from the University of Dhaka. She has several publications in renowned journals. She has also presented her works at national and international conferences. Her current research fields are nanoparticles and material physics.

Dr. ASM Shihavuddin graduated from the EEE Department of IUT, is currently working as Chairperson of the Department of EEE at Green University of Bangladesh (GUB). He is renowned for his contributions in the fields of Computer Vision and Deep Learning. In 2017, he published his work in the optimum 2D projection of 3D microscopic data in Nature Communications, also the same year his contributions in analyzing ciliogenesis were published in Science. In 2018, he published another nature paper by successfully discovering cilia beating properties. Neuron, Development, Remote Sensing, Energies are among other top journals where his novel contributions in image processing and machine learning fields are already being published over the years. In fact, he has 38 research articles in different international journals and international conferences with 357 citations (12 H index and 14 i10 indexes).

Dr. ASM Shihavuddin had worked as Associate Professor at the CSE department of ULAB. Earlier he worked as an Assistant professor at DTU compute, Technical University of Denmark (DTU), Denmark. He previously worked in the same department as a postdoctoral researcher for 10 months on ‘Automated damage detection from wind turbine inspection images with deep learning’.

He completed his first postdoc at Ecole Normale Superieure (ENS), Paris with computational biology and bioinformatics group. His principal research area was to develop machine learning and image analysis algorithms in the study of Ependymal cell microscopy images.

He did his Ph.D. from the VICOROB group, Universitat De Girona (UDG) Spain. His main research Problem was “Automated Underwater Object classification using Optical Imagery”. He completed his MSc in ‘European Masters in Computer Vision and Robotics (VIBOT)’ under Erasmus Mundus Grants.