2024 IEEE International Conference on E-health Networking, Application & Services (HealthCom) | 979-8-3503-5054-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/HEALTHCOM60970.2024.10880732

An automatic nutrition estimation framework based on food images from diabetic patients

Jiarui Chen
Tongji University
Shanghai, China
2153273@tongji.edu.cn

Qinpei Zhao(✉)
Tongji University
Shanghai, China
qinpeizhao@tongji.edu.cn

Weixiong Rao
Tongji University
Shanghai, China
wxrao@tongji.edu.cn

Yi Wang
Tongji University
Shanghai, China
2231544@tongji.edu.cn

Xinyuan Lu
Tongji University
Shanghai, China
2333098@tongji.edu.cn

Quanquan Ge
Jinshan Branch of Shanghai
Sixth People's Hospital
Shanghai, China
nemogqq@163.com

Ming Li
Shanghai DianJi University
Shanghai, China
lim420@sdju.edu.cn

Congrong Wang
Department of
Endocrinology & Metabolism,
Shanghai Fourth People's Hospital
School of Medicine, Tongji University
Shanghai, China
crwang@tongji.edu.cn

***Abstract*—With the burgeoning prevalence of diabetes and the urgent imperative for effective dietary management, harnessing the power of image analysis and computer vision technology presents a promising solution to the automatic nutrition estimation challenge. However, constructing a comprehensive framework for this endeavor presents its own set of challenges, encompassing multiple intricate procedures, including data collection, image segmentation, food recognition, and volume estimation. In this paper, we introduce a sophisticated and automated framework for Chinese food nutrition estimation, leveraging food images sourced from diabetic patients. In addition to developing models for the various procedures, the framework necessitates label images for model training. These label images encompass ground truth masks delineating precise object boundaries for image segmentation, food type labels for food recognition, and comprehensive nutrition tables for each food item. Through rigorous experimentation and validation, our framework has demonstrated its efficacy, offering a convenient and practical tool for managing dietary requirements in diabetes care.**

***Index Terms*—diabetes, nutrition estimation, food recognition, image segmentation, depth estimation**

I. Introduction

Diabetes has emerged as one of the most prevalent chronic diseases globally. According to the latest research report from the International Diabetes Federation, the number of individuals living with diabetes worldwide has reached 463 million, reflecting an 11% increase since 2019 [1]. Furthermore, the white paper of China's blood glucose health management industry in 2023 indicates that the number of diabetes patients in the country has risen to 141 million.

A study conducted by Jing Tiantian et al. [2] underscores the significant impact of dietary patterns on diabetes management. Adopting appropriate dietary practices can aid in regulating blood sugar levels, mitigating the risk of diabetes-related complications, and ensuring the body receives essential nutrients. However, this endeavor poses substantial challenges due to its prolonged and complex nature. Many individuals with diabetes often exhibit poor dietary habits and lack comprehensive knowledge about chronic diseases, further complicating dietary regulation. The introduction of advanced automated tools, as proposed by Carter et al. [3], offers a promising solution to this issue. Such tools could facilitate dietary tracking and nutritional assessment, streamlining processes and improving methodologies. As a result, this could enhance patient adherence to medical recommendations, leading to better health outcomes.

Despite considerable advancements in image processing and computer vision techniques in recent years, the segmentation of Chinese food images remains a significant challenge. The distinctive cooking methods prevalent in Chinese cuisine, such as frying, differ markedly from those commonly found in Western cuisine. Moreover, the characteristic feature of Chinese dishes often involves the amalgamation of diverse ingredients into a single dish. The extensive variety of Chinese dishes further complicates food segmentation and recognition. Currently, there is a notable lack of comprehensive Chinese food datasets and corresponding model validation studies. Furthermore, methods like binocular depth estimation, which require patients to take multiple images, can increase their workload. In the context of managing diabetic patients, it is crucial to simplify this process by enhancing the accuracy of monocular depth estimation, thereby optimizing results while utilizing only a single image.

Moreover, there is a significant lack of comprehensive life-cycle analysis frameworks in the realm of diabetes treatment, which can integrate nutritional assessment of food with Chinese cuisine. Existing research primarily emphasizes isolated tasks, such as image segmentation, food recognition, or depth estimation, rather than adopting a holistic approach.

In this paper, a comprehensive framework for food nutritional assessment is proposed to address these challenges. By segmenting, recognizing, and estimating the depth of food images, the type, quantity, and volume of each food item are accurately obtained, allowing calculation of nutritional

979-8-3503-5054-8/24/$31.00 ©2024 IEEE

components such as energy, protein, and carbohydrates. In collaboration with Shanghai Eastern Hospital and Shanghai Fourth People's Hospital, the DChiFood dataset has been established, containing over 220,000 fully labeled images across 287 food categories. The framework has been extensively validated on this dataset, demonstrating its effectiveness in food nutritional assessment tasks.

II. Related work

Recently, there has been growing interest in developing and utilizing smartphone applications to promote healthy behaviors. The semi-automatic or automatic, precise, and real-time estimation of nutrients in daily consumed meals is addressed in the relevant literature as a computer vision problem using food images captured via a user's smartphone [4]. To address this issue, it generally necessitates the support of algorithms from fields such as image segmentation, image recognition, and depth estimation.

The presence of diverse food categories often intermixed within food images, renders the application of image segmentation techniques critically important. Many studies have used traditional image processing methods such as mean shift filtering [5], combined method with SLIC and Ncut [6], and K-means clustering algorithm [7]. With the rise of deep neural networks, deep learning methods have also been gradually applied to food image segmentation [8] [9]. A food calorie and nutrition analysis system was developed in [10] based on instance segmentation network Mask R-CNN [11], which can segment and analyze the composition of foods based on the provided images.

In the realm of food image recognition, the adoption of deep learning methods has steadily gained traction. For instance, in the work by Min et al. [12], a deep progressive region enhancement network (PRENet) was proposed for food recognition. Similarly, a method for food image recognition utilizing deep convolutional networks was introduced by K. Yanai et al. [13]. Furthermore, deep network models such as MSMVFA [14] have been developed to incorporate ingredient information, thereby enhancing the accuracy of food recognition.

Depth estimation, essential for measuring food volume and quality, can be categorized into two types: multi-image-based (such as binocular depth estimation) and single-image-based (such as monocular depth estimation). In multi-image-based depth estimation, Joachim Dehais et al. [15] explored the 3D reconstruction of binocular images, while H.Hirschmuller [16] utilized images from varying perspectives to compute pixel-wise depth information. However, multi-image-based methods necessitate at least two input images to adhere to human eye imaging principles, often requiring specialized binocular cameras or cumbersome calibration, rendering them less user-friendly. Consequently, our focus in this paper lies on monocular depth estimation. In monocular depth estimation, depth information of objects is inferred from a single image [17] by analyzing visual cues such as relative size, texture, and perspective. Google [18] proposed employing a three-scale neural network for depth map estimation, initially converting it into voxel representation, and subsequently estimating food volume in conjunction with the segmentation mask.

The integration of food nutrition and computer vision technology in diabetes treatment is also examined in our work. An automatic food recognition methodology, based on the bag of features (BoF) model, was proposed in [19]. A two-level image classification scheme utilizing convolutional neural networks (CNN) was introduced by K. Kogias et al. [20]. In [21], a computer vision-based approach was proposed, which combined image processing and machine learning, based on a dataset of Greek food images they developed.

Most existing literature overlooks the Chinese food scenario and rarely integrates with diabetes treatment. There is a lack of a comprehensive nutrition assessment framework and a sufficiently labeled dataset for the discussed scenarios. Additionally, improvements are needed in system integrity and the application of new technologies.

III. The automatic nutrition estimation framework

A typical procedure of an automated vision-based dietary assessment system is illustrated in Fig. 1. The process of collecting food images is also very important, as it will affect whether the dataset can accurately reflect the patient's dietary habits and impact the performance of the assessment.

Since 2019, a registry study on Diabetes Data Registry and Individualized Lifestyle Intervention (DiaDRIL) has been initiated at Shanghai East Hospital and Shanghai Fourth People's Hospital affiliated to Tongji University. The aims of this project are to provide evidence for personalized lifestyle recommendations and to optimize glycemic control. In this study, patients were recruited at Shanghai East Hospital (from September 2019 to March 2021) and at Shanghai Fourth People's Hospital (from June 2021 to November 2021).

In proposed system, the food images are first segmented to obtain the precise food area and type via a food recognition model. The depth of each pixel in the image is then estimated using a depth estimation model. By combining the area and depth information, the volume and weight of the food can be calculated, allowing for nutritional assessment. In food image segmentation, images with masks delineating the precise boundaries of objects are provided for model training. A substantial collection of food images featuring Chinese food types is gathered for the food recognition models. To evaluate the nutrition estimation model, a dataset containing food images and the corresponding weights from diabetic patients is also provided.

The input of the algorithm is a single monocular food image, which is required to take the food image vertically downward. For segmentation, the ChineseFoodSeg algorithm and the Two-Path Global Local Network algorithm [22] are employed for the image segmentation and recognition. The LapDepth algorithm is used for the depth prediction.

ChineseFoodSeg is an algorithm meticulously tailored for the diabetes scenario, uniquely equipped to handle food images featuring multiple food items on a plate. Consequently, the

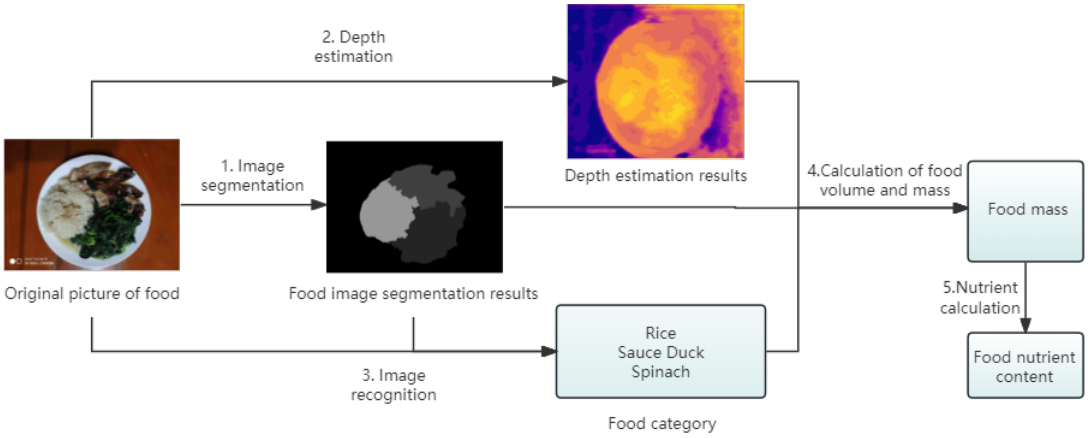

Fig. 1: The automatic nutrition estimation framework

algorithm necessitates a parameter to specify the number of food types present on the plate. Comprising three integral components-food region extraction, super pixel detection, and segmentation generation-ChineseFoodSeg offers superior accuracy and efficiency compared to conventional segmentation algorithms.

Food region extraction is transformed into a binary semantic segmentation problem, where each pixel in the image is classified as either foreground or background. For this purpose, the U-net network model is selected for binary semantic segmentation. Following this, the SLIC super pixel detection algorithm is employed to cluster pixels into groups of super pixels with similar color and texture features, facilitating image segmentation. The process is further refined using the Region Growing and Merging based on Color and Texture for Superpixels algorithm (RGM-CTS) [23], which generates hyperpixels by leveraging color and texture information. Through iterative region growing and merging, image segmentation is progressively completed. After super pixel detection, the image is segmented into K super pixels. Ultimately, food masks can be obtained.

The Two-Path Global Local Network (TPGLNet) is a deep learning architecture designed to effectively capture both global and local information from images. Its primary concept is to comprehensively leverage global and local features via distinct global and local paths.

The global path is responsible for extracting the overall features of the image. Utilizing a convolutional neural network (CNN) structure, it gradually extracts global features through multiple convolutional and pooling layers. These global features capture the overall structure and contextual information within the image. The local path, conversely, focuses on extracting local features by dividing the image into multiple small regions or blocks and processing each independently. Each region or block passes through a sub-CNN network, sharing some convolutional layers with the global path. This shared layer design allows the local paths to benefit from global features while capturing details of different regions. Ultimately, the features from the global and local paths are fused to form the final feature representation, which can then be input into the classifier for image classification or other computer vision tasks.

By integrating global and local information, TPGLNet is able to capture both the overall structure and local details of the image, thereby enhancing the performance of image classification.

The core idea of the LapDepth algorithm [24] is to accurately elucidate the relationship between the encoding features and the final output using a decoder structure based on the Laplacian pyramid for binocular depth estimation. Laplacian pyramids are extensively employed in various areas of scene understanding due to their ability to retain local information within the data.

Specifically, encoding features are fed into stacked convolutional blocks that produce subband depth residuals at each pyramid layer. By combining the depth residuals of each pyramid layer, depth maps are gradually recovered from coarse to fine scales. This recovery process enhances the prediction performance of depth boundaries. Rather than merely repeating the upsampling operation to restore the original resolution, the algorithm uses the residues of the input color image from different levels of the Laplacian pyramid to guide the decoding process. By combining these prediction results (the depth residuals), the final depth map is reconstructed from coarse to fine. This multi-layer depth residual decoding scheme allows encoded features to be used more effectively for estimating depth information in complex scenarios.

IV. Experiments and Results

A. DChiFood datasets

The DChiFood dataset encompasses several subsets, including DChiFood-seg, DChiFood-reg, and DChiFood-nutrition, offering extensive Chinese food nutrition information from various perspectives. An example is shown in Fig. 2. The

Food category and mass	Rice	150g
	Braised Shrimp in chili oil	50g
	Fried cabbage	200g
	Steamed spare ribs in black bean sauce	70g
	Scallion oil winter melon	100g
Carbohydrate	59.69g	
Fat	20.63g	
Protein	22.86g	
Cellulose	3.56g	
Heat	500.69kcal	

Fig. 2: A sample example on the result of the automatic nutrition estimation

dataset aims to provide a broad range of samples and abundant resources for Chinese nutrition analysis and related tasks.

The DChiFood-seg dataset is designed for image segmentation tasks, featuring images captured from 30-40 cm above the food to prevent overlap. It includes various diabetic-friendly foods, such as buckwheat steamed buns and sea cucumbers, with both single-food and multi-food images. A deep learning-based segmentation model is used, requiring extensive pixel-level manual annotation via the Labelme tool, which saves annotations in JSON files with masks represented by polygonal points. Masks are defined by image features like contours and shapes, with segmentation masks assigned to each class.

Encompassing 205,978 images spanning 287 Chinese food categories, the DChiFood-reg dataset is characterized by each image being annotated with food category labels, facilitating precise identification of individual food items. This dataset comprises a wide range of Chinese food items, such as rice, noodles, vegetables, fruits, meats, and seafood, among others, thereby providing a comprehensive and diverse collection of Chinese food images.

The DChiFood-nutrition dataset contains 774 images across 87 food categories, with each food item accompanied by its carbohydrate, fat, protein, fiber, and total calorie content. Captured by diabetes patients, the images feature food placed on a circular plate with a fixed radius of 10 cm to prevent overlap. Photographs were taken from a distance of 30-40 cm above the food, showcasing a wide range of portion sizes and dish complexities, from 100+ to 1500+ calories, and from single-food items to combinations of up to five different foods.

By leveraging the DChiFood dataset, researchers and developers can tackle a range of tasks related to nutritional analysis. These encompass computing nutrient contents in foods, evaluating dietary balance, and crafting intelligent health management applications. The broad utility of this dataset bolsters the precision and trustworthiness of food image recognition and nutritional analysis technologies, thereby driving progress in research and practices related to healthy eating and nutrition.

B. Image segmentation and food recognition experiments

For food image segmentation, a monocular food image is required as input, and the mask for each food image is obtained using the ChineseFoodSeg algorithm. The segmentation algorithm extracts the segmentation area and bounding box for each food, which are then utilized as inputs to the TPGLNet model to predict the species of each food. In the experiments, eight images are selected. The original images and their segmentation results are illustrated in Fig. 3.

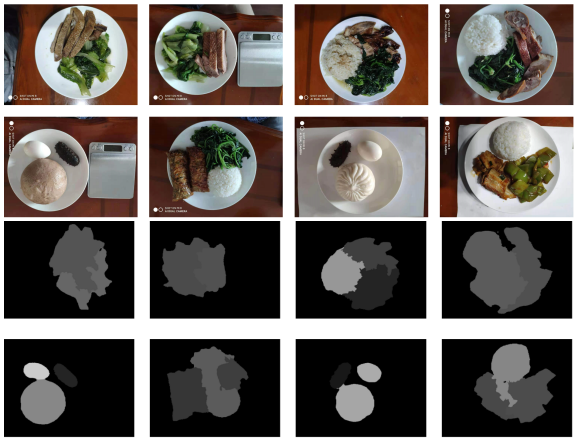

Fig. 3: The original images and the segmentation results

The food recognition results are presented in Table I, where the category number corresponds to the serial number of the food in the dataset. In this experiment, the prediction accuracy of the maximum possible food type (top_1) for the 10 foods was 37.5%, while in the top five possible food types (top_5), the accuracy of containing the ground truth label reached 70.8%.

The low identification accuracy observed in this experiment may be attributed to several factors. Firstly, the selected food images encompass a variety of foods, with an average of two foods per image in the first to eighth groups, which challenges the algorithm's ability to accurately identify images containing multiple foods. Furthermore, factors such as image resolution and variations in light and shade contrast, caused by lighting conditions, contribute to the diminished accuracy of food species recognition. Moreover, misidentification may also be closely related to the high similarity between certain

TABLE I: Results of the food recognition

Group	Food type	Groundtruth category	Top five categories
1	Duck in brown sauce	4	12,2,33,18,183
	Sauted lettuce	2	273,7,18,2,12
2	Duck in brown sauce	4	273,7,284,66,209
	Sauted lettuce	2	2,33,18,3,218
3	Rice	0	0,271,272,222,282
	Duck in brown sauce	4	12,166,267,273,170
	Sauteed spainch	3	109,3,8,46,63
4	Rice	0	8,3,63,0,46
	Sauteed spainch	3	8,273,252,3,12
	Duck in brown sauce	4	283,13,102,106,94
5	Buckwheat bread	5	22,5,6,260,141
	Egg	6	106,6,283,284,94
	Sea cucumber	7	7,95,65,273,0
6	Spring rolls	159	273,65,162,241,104
	Sauteed spainch	3	8,3,284,18,63
	Rice	0	8,12,55,175,43
7	Green vegetable bun	9	9,25,254,260,151
	Egg	6	6,284,106,0,283
	Sea cucumber	7	7,6,0,170,273
8	Rice	0	0,284,79,170,283
	Double cooked pork slices	11	287,273,241,12,51
	Green pepper	12	67,95,12,43,103
9	Pork and vegetable wonton	10	10,70,6,106,108
10	Pork and vegetable wonton	10	10,106,108,70,6

categories of Chinese food images. Some food categories exhibit significant resemblance in visual appearance and texture. Even humans may find it challenging to distinguish between these food categories. It may be necessary to devise more fine-grained visual feature learning methods to classify these food categories.

C. LapDepth algorithm depth estimation

The training set comprises the Nutrition5k dataset [25] and KITTI dataset [26]. Nutrition5k is a dataset containing visual and nutritional data of approximately 5000 realistic food dishes captured from the Google cafe using a custom scanning device. The device captures images vertically from a distance of 0.4 meters directly above the food, providing depth images and RGB image data. On the other hand, the KITTI dataset is collected using various sensors mounted on vehicles, recording real traffic scenes for up to 6 hours. The distance between the recorded objects and the sensors ranges from one meter to 80 meters. Due to the differing scenarios, KITTI is solely used for pre-training the model in our study.

The LapDepth algorithm model is initially pre-trained on KITTI and subsequently fine-tuned on Nutrition5k. The depth information estimated from the LapDepth algorithm for the tested images is visualized in Fig. 4.

D. Food mass and nutrition estimation

Using the Canny operator for edge detection, all the edges in the image can be obtained. Subsequently, we detect the edges of elliptical shapes to identify the edge of the plate. By calculating the number of pixels within the ellipse, the total number of pixels T for the plate in the image can be determined.

TABLE II: Results of food weight estimation in gram

Group	Food type	Estimate volume (cm^3)	Estimate mass (g)	Actual mass (g)	Differ (g)	Differential rate (%)
1	Duck in brown sauce	139.3	121.2	90	31.2	34.67
	Sauted lettuce	107.6	37.6	110	-72.3	**65.73**
2	Duck in brown sauce	130.4	113.4	70	43.4	62.00
	Sauted lettuce	307.7	107.7	100	7.7	7.70
3	Rice	104.8	76.4	100	-23.6	23.60
	Duck in brown sauce	97.5	84.8	85	-0.2	**0.24**
	Spinacia oleracea	151.1	52.9	100	-47.1	47.10
4	Rice	134.0	97.9	100	-2.2	2.20
	Sauteed spainch	139.7	48.9	100	-51.1	51.10
	Duck in brown sauce	145.3	126.4	80	46.4	58.00
5	Buckwheat bread	309.5	134.0	110	24.0	21.82
	Egg	70.6	42.8	43	-0.6	1.40
	Sea cucumber	40.3	24.2	24	0.2	0.83
6	Spring rolls	207.6	110.1	70	40.1	57.29
	Sauteed spainch	235.3	82.4	90	-7.6	8.40
	Rice	89.2	65.1	100	-34.9	34.90
7	Green vegetable bun	161.0	112.7	92	20.7	22.50
	Egg	60.6	36.4	43	-6.6	15.35
	Sea cucumber	52.3	31.4	22	9.4	42.73
8	Rice	76.0	55.4	100	-44.5	44.50
	Double cooked pork slices	79.3	82.5	50	32.5	65.00
	Green pepper	84.7	42.4	75	-32.6	43.47

Through the food mask obtained by image segmentation and food identification experiment steps, the total number of pixels corresponding to each food is N. The percentage of the plate area of the food τ is calculated by $\tau = \frac{N}{T}$. The diameter r of the plate is 15cm, then the area S of the plate is calculated by $S = \pi r^2$, we can obtain that $S = 176.71cm^2$. Then the projection area F of food is calculated by $F = S\tau$. The volume of food is represented by V. Through the previous depth estimation, we have obtained the average depth H of each food. So the volume of food can be represented by $V = FH$. The experimental results are presented in Table II.

In the table, the error in food quality ranges from 0.24% to 65.73%, while the overall deviation remains within an acceptable range. Analysis of the segmentation, recognition, and depth estimation results reveals that errors accumulate during image segmentation and depth estimation. Challenges such as variations in lighting and shadows, viewing angles,

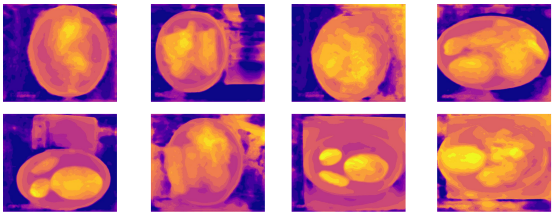

Fig. 4: The depth information of the images

and similarities between adjacent foods inevitably lead to misclassification of areas belonging to one type of food as another adjacent type. When combined with the errors in depth estimation, this amplification effect can result in significant deviations. This issue warrants further investigation in future studies.

Finally, we obtained the nutrient content of food, including carbohydrates, fat, protein, fiber, calories, and other indicators, by utilizing an auxiliary table constructed by us. This table comprises information on the volume density and mass density of nutrients for 287 types of Chinese food, derived from the Chinese food composition table. An example of nutrition assessment is illustrated in Fig. 2.

V. Conclusions

This study delves into the critical realm of dietary nutrition assessment for diabetic patients, introducing a novel amalgamation of image segmentation, recognition, and monocular depth estimation techniques. Herein, we unveil a comprehensive automated framework meticulously crafted to appraise the nutritional content of food items. At the core of our endeavor lies a large-scale food image dataset, incorporating a rich variety of Chinese cuisine. This dataset serves as the bedrock for our innovative approach, facilitating the seamless execution of image segmentation, recognition, and depth estimation processes, culminating in precise measurements of food volume and nutritional composition. Through methodical experimentation, our findings underscore the efficacy of our approach, heralding a promising avenue for tailored nutritional assessment tailored specifically for diabetic patients.

References

[1] I. D. Federation. (Accessed Dec. 19, 2021) Diabetes facts & figures. [Online]. Available: https://www.idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html

[2] T. Jing, S. Zhang, Y. Bai, Z. Chen, S. Gao, S. Li, and J. Zhang, "Effect of dietary approaches on glycemic control in patients with type 2 diabetes: A systematic review with network meta-analysis of randomized trials," *Nutrients*, vol. 15, 07 2023.

[3] M. Carter, V. J. Burley, C. Nykjaer, and J. E. Cade, "Adherence to a smartphone application for weight loss compared to website and paper diary: Pilot randomized controlled trial," *Journal of Medical Internet Research*, vol. 15, 2013.

[4] F. S. Konstantakopoulos, E. I. Georga, and D. I. Fotiadis, "A review of image-based food recognition and volume estimation artificial intelligence systems," *IEEE Reviews in Biomedical Engineering*, vol. 17, pp. 136–152, 2024.

[5] M. Anthimopoulos, J. Dehais, S. Shevchik, B. Ransford, D. Duke, P. Diem, and S. Mougiakakou, "Computer vision-based carbohydrate estimation for type 1 patients with diabetes using smartphones," *Journal of diabetes science and technology*, vol. 9, 04 2015.

[6] Y. Wang, C. Liu, F. Zhu, C. J. Boushey, and E. J. Delp, "Efficient superpixel based segmentation for food image analysis," *Proceedings. International Conference on Image Processing*, pp. 2544–2548, September 2016.

[7] S. K. Abdulateef, S. R. A. AHMED, and M. D. Salman, "A novel food image segmentation based on homogeneity test of k-means clustering," *IOP Conference Series: Materials Science and Engineering*, vol. 928, no. 3, p. 032059, nov 2020.

[8] G. Ciocca, D. Mazzini, and R. Schettini, "Evaluating cnn-based semantic food segmentation across illuminants," in *Computational Color Imaging*, S. Tominaga, R. Schettini, A. Trémeau, and T. Horiuchi, Eds. Springer International Publishing, 2019, pp. 247–259.

[9] P. Poply and J. A. Arul Jothi, "Refined image segmentation for calorie estimation of multiple-dish food items," in *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, 2021, pp. 682–687.

[10] M.-L. Chiang, C.-A. Wu, J.-K. Feng, C.-Y. Fang, and S.-W. Chen, "Food calorie and nutrition analysis system based on mask r-cnn," in *2019 IEEE 5th International Conference on Computer and Communications (ICCC)*, 2019, pp. 1721–1728.

[11] K. He, G. Gkioxari, P. DollÃąr, and R. Girshick, "Mask r-cnn," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[12] W. Min, Z. Wang, Y. Liu, M. Luo, L. Kang, X. Wei, X. Wei, and S. Jiang, "Large scale visual food recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 9932–9949, 2023.

[13] K. Yanai and Y. Kawano, "Food image recognition using deep convolutional network with pre-training and fine-tuning," in *2015 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, 2015, pp. 1–6.

[14] S. Jiang, W. Min, L. Liu, and Z. Luo, "Multi-scale multi-view deep feature aggregation for food recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 265–276, 2020.

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

[16] H. Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, 2005, pp. 807–814 vol. 2.

[17] C. Godard, O. M. Aodha, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3827–3837, 2018.

[18] A. Myers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and K. Murphy, "Im2calories: Towards an automated mobile vision food diary," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1233–1241.

[19] M. M. Anthimopoulos, L. Gianola, L. Scarnato, P. Diem, and S. G. Mougiakakou, "A food recognition system for diabetic patients based on an optimized bag-of-features model," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 4, pp. 1261–1271, 2014.

[20] K. Kogias, I. Andreadis, K. Dalakleidi, and K. S. Nikita, "A two-level food classification system for people with diabetes mellitus using convolutional neural networks," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 2603–2606.

[21] F. Konstantakopoulos, E. I. Georga, K. Klampanas, D. Rouvalis, N. Ioannou, and D. I. Fotiadis, "Automatic estimation of the nutritional composition of foods as part of the glucoseml type 1 diabetes self-management system," in *IEEE 19th International Conference on Bioinformatics and Bioengineering*, 2019, pp. 470–473.

[22] Y. Liang, J. Li, Q. Zhao, W. Rao, C. Zhang, and C. Wang, "Image segmentation and recognition for multi-class chinese food," in *2022 IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 3938–3942.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[24] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *ArXiv*, vol. abs/1706.05587, 2017.

[25] Q. Thames, A. Karpur, W. Norris, F. Xia, L. Panait, T. Weyand, and J. Sim, "Nutrition5k: Towards automatic nutritional understanding of generic food," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8899–8907.

[26] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361.