A Review of Image-Based Food Recognition and Volume Estimation Artificial Intelligence Systems

Fotios S. Konstantakopoulos, Eleni I. Georga, *Member, IEEE*, and Dimitrios I. Fotiadis, *Fellow, IEEE*

(Methodological Review)

***Abstract*—The daily healthy diet and balanced intake of essential nutrients play an important role in modern lifestyle. The estimation of a meal's nutrient content is an integral component of significant diseases, such as diabetes, obesity and cardiovascular disease. Lately, there has been an increasing interest towards the development and utilization of smartphone applications with the aim of promoting healthy behaviours. The semi – automatic or automatic, precise and in real-time estimation of the nutrients of daily consumed meals is approached in relevant literature as a computer vision problem using food images which are taken via a user's smartphone. Herein, we present the state-of-the-art on automatic food recognition and food volume estimation methods starting from their basis, i.e., the food image databases. First, by methodically organizing the extracted information from the reviewed studies, this review study enables the comprehensive fair assessment of the methods and techniques applied for segmenting food images, classifying their food content and computing the food volume, associating their results with the characteristics of the used datasets. Second, by unbiasedly reporting the strengths and limitations of these methods and proposing pragmatic solutions to the latter, this review can inspire future directions in the field of dietary assessment systems.**

***Index Terms*—Dietary assessment system, food databases, food segmentation, food recognition, food classification, food volume estimation, nutrient information, computer vision, machine learning, deep learning, artificial intelligence.**

I. Introduction

The global incidence of chronic diet-related diseases, such as obesity, diabetes, and cardiovascular diseases, shows an ever –increasing trend, which tends to take on epidemic proportions. The number of obese people has nearly tripled since 1975. In 2016, more than 1.9 billion adults were overweight, out of which over 650 million were obese. Moreover, in 2019, 38 million children under the age of five were overweight or obese [1]. Diabetes is considered as a major cause for blindness, kidney failure, heart attacks, stroke, and lower limb amputation. The World Health Organization (WHO) estimated that 1.5 million deaths were directly caused by diabetes and that diabetes was the seventh leading cause of death in 2019 [2]. According to the International Diabetes Federation, 463 million people (adults 20-79 years) suffer from diabetes worldwide nowadays [3]. As far as cardiovascular diseases (CVDs) are concerned, they are a group of disorders of the heart and blood vessels that include coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions. CVDs are the number one cause of death globally while, in 2016, 17.9 million people died from CVDs representing 31% of all global deaths [4]. The above-mentioned diseases are inextricably linked. Healthy diet has been shown to be the common denominator that can either positively or negatively affect the aforementioned diseases. A healthy lifestyle, which includes a balanced diet, maintaining a healthy weight and regular exercise can significantly reduce the percentage of individuals suffering from these diseases.

Daily diet monitoring by experts is definitely the most appropriate way to achieve a healthy and balanced diet, which includes daily recording of the type and the estimated amount of food consumed [5]. However, since daily diet monitoring by specialists is almost impossible, patients are advised to record their daily eating habits themselves. Although these methods are widely used, their accuracy remains questioned, especially for children and adolescents who lack motivation and the required skills [6], with the average error in estimating the amount of food consumed being more than 20% [7]. Even well-trained individuals with diabetes have difficulty in calculating, with a relative accuracy, the amount of carbohydrates of their meal [8]. The rapid increase in the use of smartphones and their advanced computing capabilities during the last decade, have led to the development of smartphone applications [9] that can detect food, recognize its type and calculate its nutritional value, by estimating its quantity, via the analysis of food images [10]. In a typical scenario, the user is asked to take one or more photos or even videotape their meal, and then, the application computes the corresponding nutritional information.

Manuscript received 13 January 2021; revised 14 February 2022, 3 December 2022, and 7 April 2023; accepted 12 May 2023. Date of publication 5 June 2023; date of current version 15 January 2024. This work was supported by the GlucoseML project and has been co-financed by the European Regional Development Fund of the European Union and Greek National funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE under Project: T1EDK-03990. *(Corresponding author: Dimitrios I. Fotiadis.)*

The authors are with the Unit of Medical Technology and Intelligent Information Systems, Materials Science and Engineering Department, University of Ioannina, 45110 Ioannina, Greece, and also with the Biomedical Research Institute, FORTH, University of Ioannina, 45110 Ioannina, Greece (e-mail: fotkonstan@uoi.gr; egeorga@uoi.gr; fotiadis@uoi.gr).

Digital Object Identifier 10.1109/RBME.2023.3283149

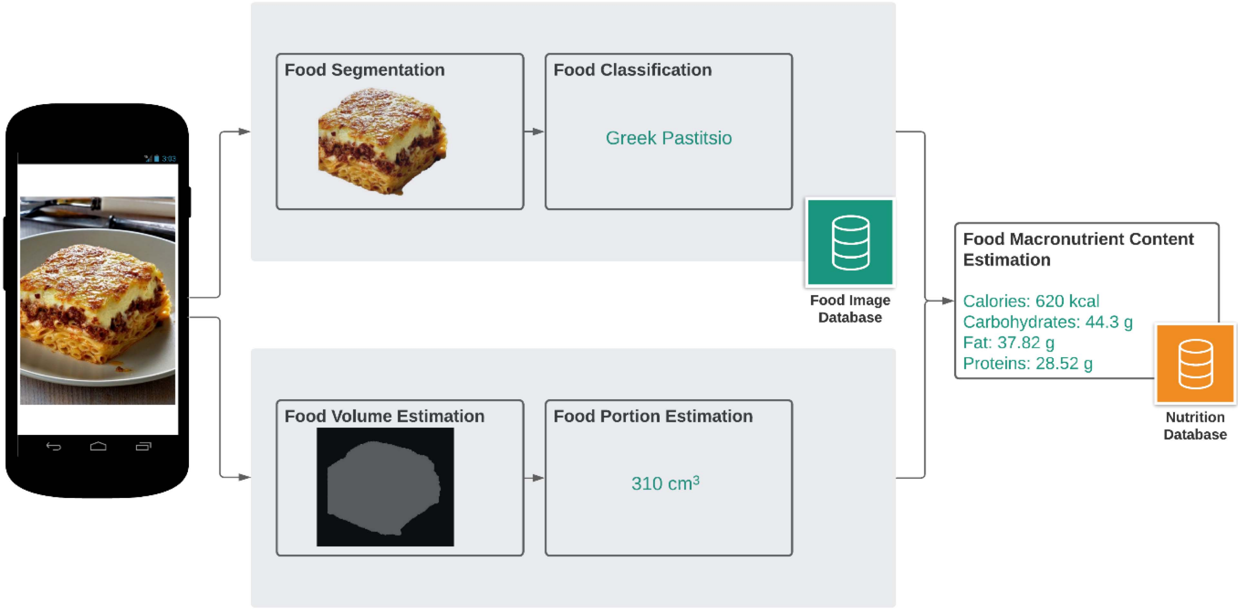

Fig. 1. An automated vision-based dietary assessment system.

Nowadays, the advances in the field of computer vision and Artificial Intelligence (AI) provide users with the possibility to monitor their health every day through appropriate applications [11]. Recent studies have shown that AI-based applications are more popular among users, compared to traditional dietary recording methods, for recording the nutritional composition of food [12]. AI-based methods can be divided into semi-automatic which require user participation, and automatic that do not require any human participation. These applications do not aim to replace dieticians, on the contrary their goal is to provide them with an additional tool in the monitoring patients' diet. The performance and accuracy of these applications depend to a large extend on various factors, such as the food image databases used for training of the system and extraction of nutritional composition, the food segmentation techniques, the food recognition methods and the volume estimation techniques.

The quality and the quantity of images of a food database mainly affects the performance of the food recognition step [13]. Food classification, which consists of food segmentation and food recognition steps is next. Food segmentation is the process of partitioning a food image into multiple segments (sets of pixels) [14]. Food recognition comprises the identification of the foods which are present in the food image through the application of machine and deep learning techniques [15], [16]. The final step is the volume estimation for each food item which is present in the food image. This step depends directly on the previous steps of segmentation and recognition. Volume calculation of each identified segment, in combination with a food nutritional database, is used for the extraction of the nutritional composition [17]. A typical procedure of an automated vision-based dietary assessment system is shown in Fig. 1.

In this article, we present a review of the literature over the past 10 years (2012 - 2021) in the field of food images segmentation, food classification, food volume estimation and food macronutrient content estimation based on smartphone-captured food images, assessing, in parallel, the main characteristics of the employed food image databases. The in-depth analysis of the methods used in each of the above components of a dietary assessment system comprises the main distinguishing characteristic of this review in comparison with existing reviews in the specified research topic [18], [19], [20], [21], [22]. This analysis led to the categorization of the employed methods as: (i) semi-automatic and automatic food image segmentation methods, (ii) traditional machine learning (ML) - based and deep learning-based methods for food image classification, and (iii) 3D reconstruction, pre-build shape templates, perspective transformation, depth camera and deep learning methods for food volume estimation (Table I). The algorithms and techniques pertaining to each of these categories are identified per investigated study, and their performance, strengths and limitations are presented and contrasted. Importantly, we suggest pragmatic solutions to deal with the identified limitations starting from the construction of relevant datasets to the computation of the food nutrient value. This manuscript is hereunder organized in six sections, with Sections II-V presenting the review of the methods and techniques used in each of the components of a dietary assessment system, and Sections VI and VII being devoted to the discussion of the outcomes and conclusions derived by this review study.

II. Food Image Databases

The process of collecting food images, which can be used in the food classification model, is crucial and it directly affects the performance of the classification models. A comprehensive collection of food images is the key to a classifier's performance. Large food image databases, such as Food-101 [23], UEC-Food100 [14], VIREO Food-172 [24], and UEC-Food256 [25],

TABLE I
MAIN TECHNIQUES, METHODS AND PERFORMANCE METRICS FOR EACH STEP IN DIETARY ASSESSMENT SYSTEM

Step	Methods and Techniques			Performance Metrics
Food segmentation	Semi – automatic approaches (GrabCut algorithm)			Intersection over union (IoU), Pixel accuracy, Panoptic Quality (PQ)
	Automatic ML approaches with handcrafted feature extraction (HOG, JSEG)			
	Automatic ML approaches using deep learning for feature extraction (CNNs, Instance and Semantic segmentation)			
Food classification	Traditional approaches	Feature extraction	SIFT, SURF, HOG, Gabor, LBP	Recall, Precision, F1-Score, Top-1 accuracy, Top-5 accuracy
		Feature representation	BoF, Fisher vectors	
		Classification	SVM, kNN, RF	
	Deep learning approaches	CNN and DCNN		
Food volume estimation	3D reconstruction, Pre-build shape template, Perspective transformation, Depth camera, Deep learning			Mean absolute error (MAE) Mean absolute percentage error (MAPE) Root mean square error (RMSE)

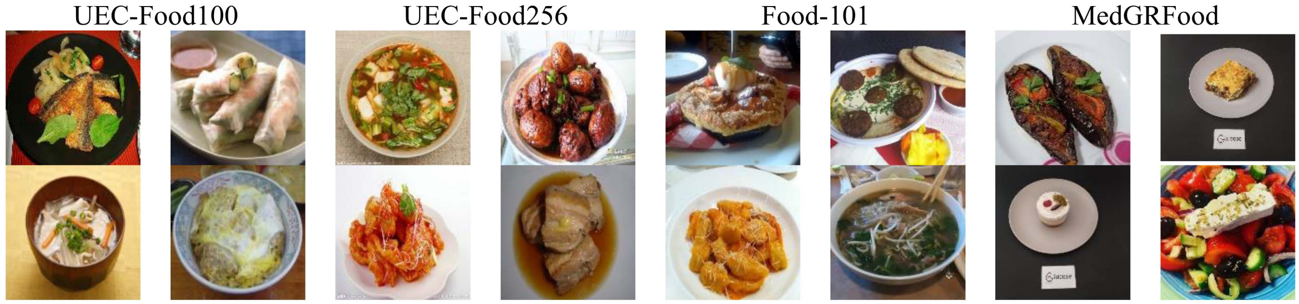

Fig. 2. Food images from UEC-Food100, UEC-Food256, Food-101 and MedGRFood databases.

are benchmark food databases and are typically used to evaluate machine learning models. Existing databases are distinguished by the different characteristics they have, such as cuisine type, the number of images, the number of food classes, the food categories, the way of acquisition, the task of use (classification or segmentation task) as well as by how many different food items are included in each photo. For instance, Diabetes [26] has 11 classes with a total of 5420 pictures out of which 3800 images are downloaded from the web and 1620 are captured in a controlled environment. A few food databases have been created by compiling images of existing food databases. For instance, the database Food524DB [27] were created from existing publicly available food image databases: Food-101, UEC-Food256 and VIREO Food-172. Moreover, there are several food image databases that have collected food images from specific types of cuisines. For example, Chen [28] and ChineseFoodNet [29] represent the Chinese cuisine, FFoCat [30] and MedGRFood [31] refer to Mediterranean food, Indian food database [32] contains images with local food dishes, while [33], [34], [35] present databases with images of fruits and vegetables. FLD-469 [36] refers to Japanese food, while FoodX-251 [37], Menu-Match [38], UPMC Food-101 [39], NutriNet [40] and UNICT-FD889 [41] consist of a mix of eastern and western food images. Moreover, a critical feature of the food image database is whether it is used for classification [42], [43], [44], [45] or segmentation tasks [46], [47], [48], [49], [50], [51]. For example, Food201-Segmented [52] contains segmented images from Food-101 dataset for the USA cuisine. Also, an important element for the classifier is the way the pictures were acquired, namely whether they were taken in a controlled environment (in terms of lighting conditions and the food's image background) or in a free environment. In addition, with the increasing use of deep learning methods for image classification, the food image databases must contain a large number of images per class to support training of a deep learning model. Furthermore, the diversity of the images contained in a class leads to a more advanced model, which can classify food even if it has been cooked in a similar way. Fig. 2 presents sample images from four food image databases.

The techniques used in the later stages of food image-based analysis nutrition systems, emphasize the need to create databases that contain a large number of images for each food class. It may be easier nowadays to collect the images for a large food image database, due to the tendency to capture food images using smartphones and to the existence of many images in social networks. Although, there is a plethora of food image databases, we note that there are no food image databases related to healthy diet patterns. In addition, there exist a few annotated databases, mainly referring to the Japanese cuisine, which could be used in the segmentation and classification tasks (Fig. 3). Fig. 4 illustrates the size (number of images) of existing food image databases for different types of cuisine annotated by the associated method of constructions. We observe that the majority of databases belong to generic and Asian cuisine, while a large number of them are either collected from the web or created using other databases. Finally, it is worth mentioning that there is no benchmark food image database for general classification purposes. As food has no borders and we live in

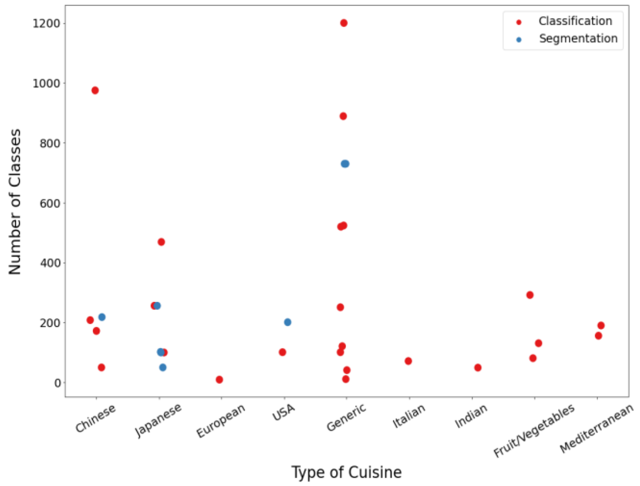

Fig. 3. Type of cuisine distribution according to the number of classes and how they are used.

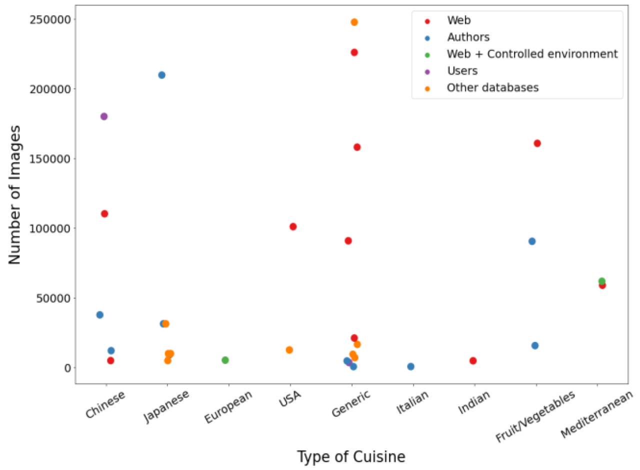

Fig. 4. Size of existing databases for different types of cuisine annotated by the means of food image collection.

multicultural societies, it is needed to create a large food image database, that will include different types of cuisines, to allow the development of systems and applications that will be able to detect and calculate the amount of as many foods as possible. Therefore, the creation of an annotated food image dataset that would take into account the type of cuisine could include foods with the same name but from different regions. For example, it is possible for an annotated food image database to contain the same food name and characterize it additionally by its cuisine or its region. Therefore, the creation of an annotated food image dataset that would take into account the type of cuisine could include foods with the same name but from different regions. For example, it is possible for an annotated food image database to contain the same food name and characterize it additionally by its cuisine or its region. Table II summarizes the most representative food image databases, and their most significant features.

III. Food Image Segmentation

Segmentation is the initial step required to identify food and refers to the process of localization and extracting regions that have different colour and texture features. The purpose of food image segmentation is to localize a food item or the food items (if there is more than one) present in an image, and to separate them from the background or other food items [24]. When the image contains more than one food, food segmentation is considered a necessary step in dietary assessment systems. It is a challenging task to segment foods that overlap each other, or foods that have an indeterminate shape, or foods that do not have strong colour or texture features in contrast with the other food items in a plate. In addition, the lighting conditions, under which an image is taken, can affect the segmentation step by creating shadows and reflections [17]. Although segmentation is a difficult process, the accuracy of segmentation directly affects the effectiveness of the subsequent steps, such as the classification and volume estimation. The main metrics for assessing food image segmentation are the Intersection over Union – IoU:

$$IoU = \frac{Y_{true} \, \cap \, Y_{pred}}{Y_{true} \cup Y_{pred}}, \tag{1}$$

where Y_{true} is the ground truth of the food image and Y_{pred} is the prediction mask; the meanIoU for multiclass segmentation:

$$meanIoU = \frac{1}{N} \sum_{i\,=\,1}^{N} IoU_i, \tag{2}$$

where N is the number of food classes; and the pixel accuracy:

$$Pixel_{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{3}$$

where *True Positive (TP)* represents a pixel that is correctly predicted to belong to the given class, *True Negative (TN)* represents a pixel that is correctly identified as not belonging to the given class, *False Positive (FP)* represents a pixel that is wrongly predicted to belong to the given class and False *Negative (FN)* represents a pixel that is wrongly identified as not belonging to the given class.

Several methods have been proposed to address issues in food image segmentation. An initial classification of methods is: (i) semi-automatic food segmentation, (ii) automatic ML with handcrafted feature extraction, and (iii) automatic ML with deep learning feature extraction.

In several studies, the use of semi-automatic techniques for food segmentation is preferred, where the user is asked to select regions of interest in the image, the foreground and the background (Fig. 5). The results of semi-automatic techniques are highly accurate, distinguishing details of each food item in the image, as the user knows the exact boundaries of food items contained in the image/tray [53], [54], [55], [56]. Hassannejad et al. [57], used a customized interactive graph cut algorithm. Initially, the user imposes a number of hard constraints to segmentation, by marking some pixels. Then they use the Gaussian mixture model and K-Means to generate image clusters and initialize the graph. Finally, an iterative graph cut algorithm is used to

TABLE II
FOOD IMAGES DATABASES

Authors	Database Name	Food Category	Database Use	# classes/ # images	Image Source
Chen 2012, [28]	Chen	Chinese	Classification	50/5,000	Downloaded from the web
Matsuda et al. 2012, [14]	UEC-Food100	Japanese	Classification	100/9,132	Captured by authors
Anthimopoulos et al. 2014, [26]	Diabetes	European	Classification	11/5,420	Downloaded from the web + controlled environment
Bossard et al. 2014, [23]	Food-101	USA	Classification	101/101,000	Downloaded from the web
Kawano et al., 2014, [25]	UEC-Food256	Japanese	Classification	256/31,397	Captured by authors
Farinella et al. 2014, [41]	UNICT-FD889	Generic	Classification	889/3,583	Captured by users
Meyers et al. 2015, [52]	Food201-Segmented	USA	Segmentation	201/12,625	Acquired from other databases
Beijbom et al. 2015, [38]	Menu-Match	Generic	Classification	41/646	Captured by authors
Wang et al. 2015, [39]	UPMC Food-101	Generic	Classification	101/90,840	Downloaded from the web
Zhou and Lin, 2016, [42]	Food-975	Chinese	Classification	975/37,785	Captured by authors
Chen and Ngo 2016, [24]	Vireo Food-172	Chinese	Classification	172/110,241	Downloaded from the web
Ciocca et al. 2016, [13]	UNIMIB 2016	Italian	Classification	73/1,027	Captured by authors
Singla et al. 2016, [43]	Food-11	Generic	Classification	11/16,643	Acquired from other databases
Farinella et al. 2016, [44]	UNICT-FD1200	Generic	Classification	1,200/4,754	Captured by authors
Ciocca et al. 2017, [27]	Food524DB	Generic	Classification	524/247,636	Acquired from other databases
Chen et al. 2017, [29]	ChineseFood Net	Chinese	Classification	208/180,000	Captured by users
Pandey et al. 2017, [32]	Indian Food Database	Indian	Classification	50/5,000	Downloaded from the web
Mezgec et al., 2017, [40]	NutriNet	Generic	Classification	520*/225,953	Downloaded from the web
Hou et al. 2017, [33]	VegFru	Fruit and Vegetables	Classification	292/160,731	Downloaded from the web
Waltner et al. 2017, [34]	FruitVeg-81	Fruit and Vegetables	Classification	81/15,737	Captured by authors
Qing Yu et al. 2018 [36]	FLD-469	Japanese	Classification	469/209,700	Captured by authors
Muresan et al. 2018, [35]	Fruits-360	Fruits	Classification	131/90,483	Captured by authors
Aguilar et al. 2019, [45]	MAFood-121	Generic	Classification	121/21,175	Downloaded from the web
Donadello et al., 2019, [30]	FfoCat	Mediterranean	Classification	156/58,962	Downloaded from the web
Kaur et al. 2019, [37]	FoodX-251	Generic	Classification	251/158,000	Food-101 + downloaded from the web
Gao et al. 2019, [46]	SUEC Food	Japanese	Segmentation	256/31,395	Acquired from other databases
Ege et al. 2019, [47]	UECFoodPix	Japanese	Segmentation	100/10,000	Acquired from other databases
Wang et al. 2019, [51]	Mixed dishes	Chinese	Segmentation	218/12,105	Captured by authors
Aslan et al. 2020, [50]	Food50Seg	Japanese	Segmentation	50/5,000	Acquired from other databases
Konstantakopoulos et al. 2021, [31]	MedGRFood	Mediterranean	Classification	160/51,840 & 190/5,000	Downloaded from the web + controlled environment
Okamoto et al. 2021, [48]	UECFoodPix Complete	Japanese	Segmentation	102/10,000	Acquired from other databases
Wu et al. 2021, [49]	FoodSeg103/ FoodSeg154	Generic	Segmentation	730/7,118 730/9,490	Acquired from other databases

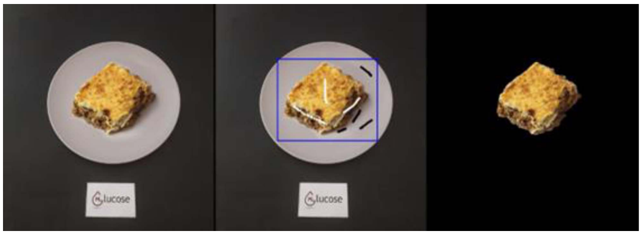

Fig. 5. Example of food image segmentation using the GrabCut algorithm. The blue rectangle represents the region of interest, the white lines represent the foreground and the black lines represents the background.

segment the food image. The users who were familiar with the application achieved up to 93% accuracy (images with less than 5% of false segmented pixels), while the users who were not familiar achieved 88% accuracy.

In automatic food segmentation methods with handcrafted feature extraction, the user only needs to capture the image. Then, existing image processing techniques are employed to solve the segmentation problem by making assumptions about the shape, colour and number of food items in the plate. These approaches use algorithms and techniques to extract texture, shape and colour features, such as the J measure-based segmentation (JSEG), the Normalize cuts (NCut) [58], or region

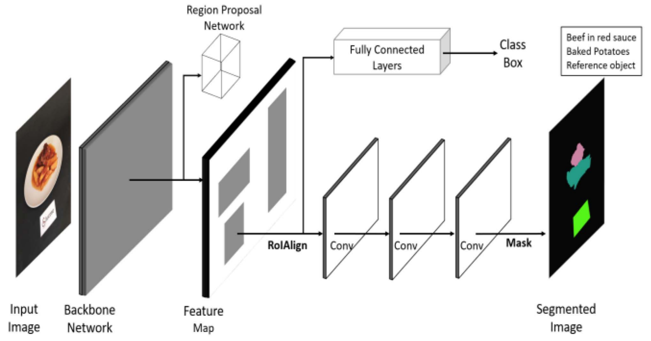

Fig. 6. An instance segmentation model.

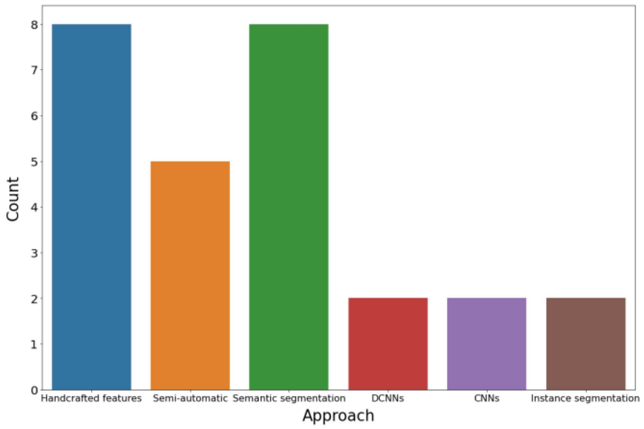

Fig. 7. The counts of segmentation approaches in dietary assessment systems.

merging and growing [59]. For example, Anthimopoulos et al. [60] suggested the use of a five-step food segmentation algorithm based on colour information: CIELAB conversion, pyramidal mean-shift filtering, region growing, region merging and plate detection/background subtraction. The proposed method achieves an 88.5% segmentation accuracy.

In recent years, deep learning approaches [61], [62], [63], [64] and Convolutional Neural Networks (CNNs) [65] in some cases have shown state of the art performance in computer vision tasks, allowing the use of automated food image segmentation methods. In these approaches the segmentation models consist of two main parts: (i) the first part, acts as an encoder by extracting a large number of features from the image, while (ii) the second part act as decoder and is responsible for image segmentation (Fig. 6). Several popular CNNs models, such as ResNet50 [66], [67] and InceptionV3 [68] are used as the backbone network in the encoder, while well-known architectures, such as Fully Convolutional Network (FCN) [69] and DeepLab [70], are used as a decoder. Shimoda and Yanai [71], presented a method to make consistency between a food segmentation model and a plate segmentation model. More specifically, they used Class Activation Mapping (CAM), which is one of the basic visualization techniques of CNNs. A food category classifier can highlight food regions containing no plate regions, while a food/non-food category classifier can highlight food regions including plate regions. They demonstrated that they boosted the accuracy of weakly-supervised food segmentation. In a recent study, Wu et al. [49] proposed a novel fully automatic semantic segmentation method consisting of a recipe learning module and an image segmentation module. They used a Long short-term memory (LSTM) network as the encoder and the vision transformer architecture as the decoder and they achieved 0.439 mIoU in the FoodSeg103 database. In a new study, Nguyen and Ngo [72] presented an instance segmentation model for multiclass segmentation, using the terrace representation for food items. They employed the panoptic quality metric, a combination of IoU and pixel accuracy metrics, which achieved a score 0.693. Although the segmentation step is not necessary in several dietary assessment systems, we observe that the studies using the semi-automated segmentation method result in better performance. However, this leads to a delay in calculating the nutritional composition, as it requires interaction with user of the system. In automated food segmentation, the use of deep learning techniques has resulted in better performance compared to handcrafted techniques. Instance segmentation is a technique that has been used on a small scale (Fig. 7) in food image segmentation and could further improve the segmentation performance of dietary assessment systems. Moreover, it can be used to segment multiple foods in an image, allowing the development of more realistic applications, as each dish tends to have more than one food items. This presupposes the use of annotated food image databases, as it is a requisite to build segmentation models based on deep learning techniques. In recent studies, the food image segmentation step is omitted and in some others the performance is not reported. In other studies, although the performance of the methods used to segment food images is high and improves the classification accuracy, there are still open issues related to cases where mixed or overlapping foods exist. In these cases, the use of state-of-the-art segmentation techniques, such as semantic and instance segmentation, can be used to improve performance and increase accuracy in the classification step. In Table III, the main segmentation techniques are summarized.

IV. Food Image Classification

Food image classification is a complex process that may be affected by many factors. For instance, the way food is cooked or if other food items, like sauce, covering the main food are present. Provided that the results of classification highly affect the effectiveness of next steps (the food volume estimation step and the food nutritional composition step), researchers have developed various techniques and methods to improve classification accuracy. The training of the classifier is affected by the number and quality of images used in the training phase, so the food database plays a crucial role in this process. Moreover, the techniques used to extract the features of the images, through which the images are recognized, greatly affect the accuracy of the classifier. The most basic metrics used for classification models are top-1 and top-5 accuracy. Top-1 accuracy is the accuracy where true class matches with the most probable classes

TABLE III
FOOD IMAGE SEGMENTATION APPROACHES

Authors	Approach	Performance
Kawano and Yanai 2013, [53]	Bounding Box (bbox) and GrabCut segmentation algorithm	Classification accuracy is improved when the ground-truth bounding boxes are given
Shimoda and Yanai 2015, [55]	Bbox using CNNs and GrabCut	It can detect bounding boxes around food items with a mean average precision of 49.9%
Hassannejad et al. 2015, [57]	Customized interactive version of the graph-cut algorithm	93.1% accuracy for familiar users with the application and 88% for users who were not familiar
Inunganbi et al. 2018, [54]	Interactive segmentation. Boundary detection/ filling and Gappy Principal Component Analysis methods are applied to restore the missing information	Outperforms the existing methods
Fang et al. 2018, [56]	Manual design of bbox, manual selection of food tag and GrabCut	Performs efficiently when used on a large image database
Matsuda et al. 2012, [14]	JSEG segmentation, circle detector and DPM	Overall accuracy 21%
Anthimopoulos et al. 2013, [60]	CIELAB conversion, pyramidal mean-shift filtering, region growing, region merging and plate detection/ background subtraction	Accuracy 88.5%
Pouladzadeh et al. 2014, [17]	Graph Cut segmentation	Accuracy of 95%
Zhu et al. 2014, [15]	Multiple segmentation hypotheses by selecting segmentations using confidence scores assigned to each segment.	Outperforms normalized cut method and improves the classification accuracy
Meyers 2015, [52]	DeepLab model	Classification accuracy is improved
Wang et al. 2016, [58]	Normalized cut and superpixels	Outperforms some widely used segmentation methods
Ciocca et al. 2016, [13]	A combination of saturation, binarization, JSEG segmentation and morphological operations	Achieves better segmentation accuracy in contrast to JSEG approach
Zheng et al. 2018, [59]	Adaptive K-means image segmentation	The segmentation accuracy is improved, compared with other traditional methods
Minija and Emmanuel 2020, [16]	Salient region detection, multi-scale segmentation and fast rejection	Classification accuracy is improved
Dehais et al. 2016, [61]	DCNN and region growing/merging techniques	The automatic and semi-automatic segmentation methods reached average accuracies of 88% and 92%, respectively
Bolanos et al. 2016, [64]	Employed a DCNN to simultaneously perform food localization	Outperforms the existing methods
Aguilar et al. 2018, [69]	Fully convolutional network (FCN) and bounding box	IoU over 0.96
Aslan et al. 2018, [70]	DeepLab-v2 for semantic segmentation	mIoU: 0.433 in UNIMIB 2016
Ciocca et al. 2019, [62]	DCNN to discriminate food regions from the background in different illumination conditions	IoU 0.79
Pfisterer et al. 2019, [67]	DCNN for semantic segmentation of food on a plate using monocular RGB images	IoU 0.912
Shimoda and Yanai 2020, [71]	Class Activation Mapping (CAM)	The segmentation accuracy is improved, compared with existing -supervised segmentation methods
Yarlagadda et al. 2021, [65]	Finding salient missing objects before and after eating images	AUC (Area Under the Curve): 0954
Okamoto et al. 2021, [48]	DeepLab V3+	Mean IoU: 0555
Poply et al. 2021, [63]	Semantic segmentation - RefineNet	$IoU_{0.75}$: 0.962 in UNIMIB 2016
Wu et al. 2021, [49]	ReLeM semantic segmentation model	Mean IoU 0.439 in FoodSeg103
Park et al. 2021, [66]	Mask R-CNN pretrained on synthetic data	Average precision: 0.522
Nguyen and Ngo 2021, [72]	Terrace-based instance segmentation	MAE:0.45, PQ:0.693 in Mixed dishes dataset

predicted by the model, defined as:

$$Classification_{accuracy} = \frac{number\ of\ correct\ predictions}{number\ of\ all\ predictions}, \tag{4}$$

Top-5 accuracy is the accuracy where true class matches with any one of the 5 most probable classes predicted by the model. Other known metrics for classification task are:

$$Precision\ = TP/\left(TP + FP\right), \tag{5}$$

$$Recall\ = TP/\left(TP + FN\right), \tag{6}$$

$$F1 - Score = \frac{2 \times (Precision \times Recall)}{Precision + Recall}. \tag{7}$$

The task of food image recognition can be divided into two categories: traditional machine learning approach with hand-crafted features and deep learning approach using convolutional neural networks (Fig. 8).

A. Traditional Machine Learning Approaches

Approaches that fall into this category are differentiated based on the technique chosen to extract the image features and, on the classifier selected for their classification. Feature extraction is the process in which the most representative features of an image are extracted, creating the corresponding feature vector. There are several feature extraction algorithms, such as speeded-up robust features (SURF), scale invariant feature transform (SIFT), local binary patterns (LBP) [73], Gabor filter [74] and

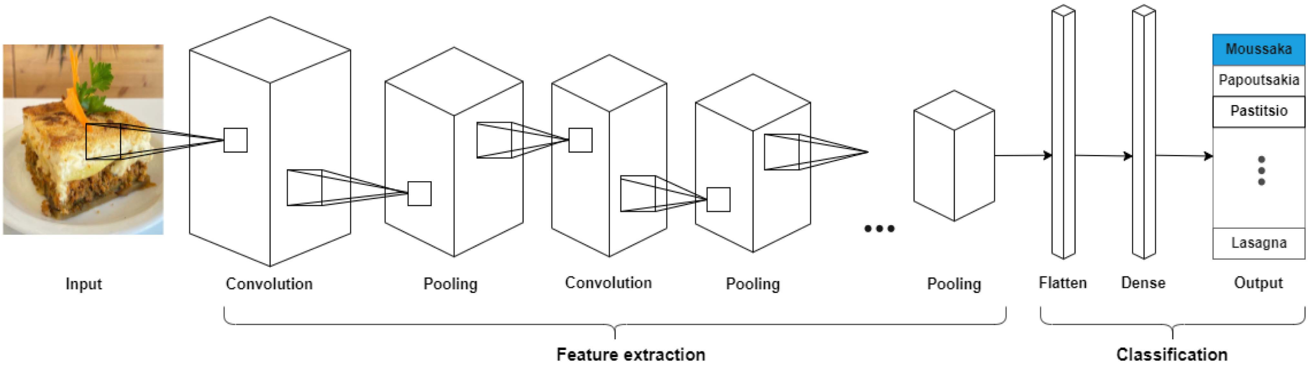

Fig. 8. A deep learning classification model of food images.

TABLE IV
TRADITIONAL CLASSIFICATION APPROACHES

Authors	Features	Classifier	Database	Top 1 accuracy	Top 5 accuracy
Chen et al. 2012, [28]	SIFT, LBP, Gabor and colour	SVM & Adaboost	Chen	68.3%	90.9%
Matsuda et al. 2012, [14]	BoF of SIFT and CSIFT, HOG and Gabor	MKL-SVM	UEC-Food100	55.8%	n/a
Kawano and Yanai 2013, [53]	Colour histogram and Bag-of-SURF	Linear SVM	6,781 images/50 food	53.5%	81.6%
Kawano and Yanai, 2014, [76]	Fisher Vector and RootHoG	One-vs-rest linear classifier	UEC-Food256	50.1%	74.4%
Anthimopoulos et al. 2014, [26]	BoF, hsvSIFT and colour moment invariant	SVM	Diabetes	78.0%	n/a
He et al. 2014, [75]	DCD, MDSIFT, SCD and SIFT	KNN	1,453 images/42 classes	64.5%	84.2% top-4 accuracy
Bossard et al. 2014, [23]	SURF and Lab colour	Random Forests	Food-101	50.8%	n/a
Pouladzadech et al. 2015, [74]	Gabor and colour	Cloud-based SVM	6,000 images/ 30 classes	94,5%	n/a
Beijbom et al. 2015, [38]	HOG, SIFT, LBP and MR8	One-vs-rest linear SVM	Menu-Match	77.4%	96.2%
Christodoulidis et al. 2015, [73]	Colour Histograms, LBP	SVM	Own database	82.2%	n/a

histogram of oriented gradients (HOG). In numerous approaches the feature extraction is performed by a combination of the above algorithms, improving the classification accuracy. The exported features then, feed a classifier for training the prediction model, based on machine learning algorithms, such as support vector machine (SVM), bag of features (BoF), random forests (RF), k-nearest neighbours (kNN) [75] and multiple kernel learning (MKL). For example, Bossard et al. [23] introduced a method to mine discriminative parts using RF. To improve effectiveness of mining and classification, they consider patches that are adjusted with image superpixels. For each superpixel, they extracted Dense SURF and L∗a∗b colour features. Then, they train a multi-class SVM for final classification, with an average accuracy 50.8% in Food-101 image dataset. In another study, Kawano and Yanai [76] proposed a food recognition system that can identify 256 food categories using the food image database UEC-Food256. They applied RootHoG and colour features and coded them into a Fisher Vector to train one-vs-all linear classifier, with top-1 accuracy 50.1% and 74.4% top-5 accuracy. Pouladzadech et al. [74], classified 30 food classes using a cloud-based SVM classifier, achieving 94.5% accuracy. They used a combination of features, including colour, texture, size and shape, while most prevailing methods use only colour and shape features. Table IV summarizes traditional food classification approaches and their main characteristics.

B. Deep Learning Approaches

The CNN is a class of deep neural networks (DNNs); it constitutes the state-of-the-art method in image recognition. They are most used to analyse visual imagery and are frequently working behind the scenes (hidden layers) in image classification. A CNN convolves learned features with input data and uses 2D convolutional layers. This means that this type of network is ideal for processing 2D images. Compared to other image classification algorithms, CNNs actually use very little pre-processing. A CNN works by extracting features from images. This eliminates the need for manual feature extraction. The features are not trained but they are learned while the network is trained on a set of images. This makes deep learning models extremely accurate for computer vision tasks. CNNs learn feature detection through tens or hundreds of hidden layers. Each layer increases the complexity of the learned features.

Several studies use pre-trained CNN models [77], [78], [79], [80], [81], [82] to classify food images, such as Inception V3 [83], [84] and EfficientNet [85], [86]. Moreover, fine-tuning

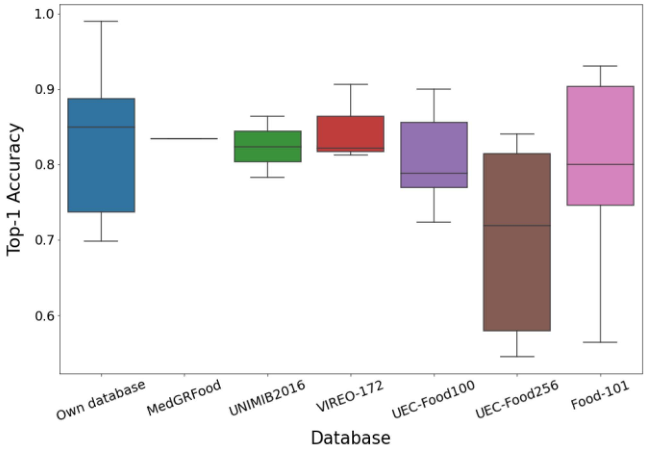

Fig. 9. Boxplot distribution of top-1 accuracy of deep learning-based food recognition algorithms for different food image databases.

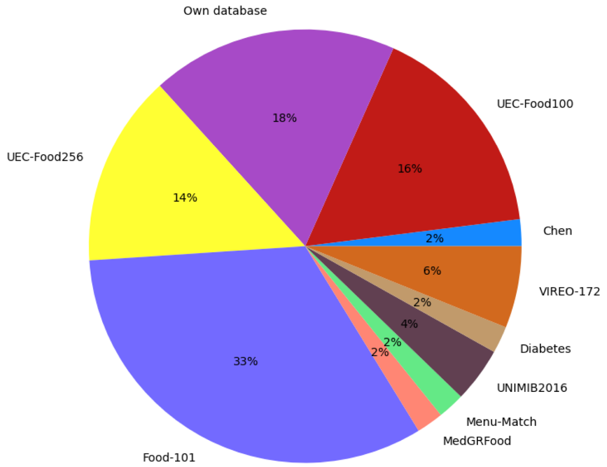

Fig. 10. Percentage use of food image databases in food recognition-related studies.

[87], transfer learning [88] and data augmentation techniques are applied to improve the accuracy of classification models. Definitely, the last years, deep learning is the state of the art for food image classification [89]. Hassannejad et al. [90], evaluated a fine-tuned version of Inception V3 model, increasing the accuracy and decreasing the computational cost. In particular, they achieved 81.5%, 76.2% and 88.3% top-1 accuracy, on UEC-Food100, UEC-Food256 and Food-101 databases, respectively. In addition, they achieved 97.3%, 92.6% and 96.9% top-5 accuracy on UEC-Food100, UEC-Food256 and Food-101 databases, respectively. In another study, they have built a DNN model consisting of two stages: The first stage is a residual network, encoding generic visual depictions of food images, while the second stage is a slice network with a slice convolutional layer capturing the vertical food features. The extracted features are linked and fed to the fully connected layers that give out the classification prediction. Tan and Le [91], proposed a new CNN scaling architecture, the EfficientNet. They scaled up the depth, width and resolution of the network, outperforming the state-of-the-art deep learning studies. EfficientNet-B7 achieves 93% accuracy in the Food-101 dataset. In several deep learning-based studies for food recognition, it is observed that the evaluation of the models is performed in the databases of food images: UEC-Food100 [92], UEC-Food256 [93], Food-101 [94] and VIREO-172 [95].

Fig. 9 shows the box plots of top-1 accuracy achieved by deep learning approaches for existing food image databases. We observe the top-1 accuracy features a high interquartile range for the UEC-Food256 and Food-101 databases; this is an indication of the complexity characterising multi-class problems. On the other hand, a higher and less spread top-1 accuracy obtained for databases with a small number of classes or focused on specific tasks. Fig. 10 presents the percentage usage of existing food image databases as development datasets in food recognition, where databases with a large number of classes being used more often. In addition, a considerable amount of studies (18%) do not refer any information about the used databases, diminishing their replicability potential. We observe that the Food-101 is the database with the highest percentage, while newer databases have used very little. Table V presents the main characteristics of deep learning approaches applied in food image classification. We can observe that the accuracy of conventional classification models can be improved by combining feature extraction algorithms. Moreover, the combination of different classifiers seems to work better than using only one classifier. In addition, we notice that the traditional approaches are used on small food datasets where deep learning techniques cannot be applied, and it is obvious that deep learning techniques for food image recognition outperform the traditional ones [19]. Although CNNs were firstly used to extract features that feed a classifier, in recent years only deep learning models have been used to classify food images. Furthermore, we note that there is a tendency to use deeper learning networks to train food image classification models (for example, the EfficientNet B-7 consists of 813 layers). However, the need of computing power seems to limit the possibilities of such an approach. In the future, with the ever-increasing computing power to train deep learning models (e.g., deep learning cloud servers) and to build deeper networks, combined with training in larger datasets, their performance can be further improved.

V. Food Volume Estimation

The last step in food nutritional composition systems comprises the estimation of foods quantity and the analysis of their nutritional composition, such as carbohydrates, proteins, fat and total calories. Accurate estimation of the amount of food, assumes that the previous stages of the segmentation and recognition of the food have been accomplished correctly. Then, using appropriate approaches, such as 3D reconstruction, pre-build shape templates, perspective transformation, depth camera and deep learning techniques, the volume of food is estimated. This is a demanding process which in most cases requires a specific number of photos and a specific way of taking them, a controlled environment and in many cases dedicated cameras for capturing food images. In fact, calculating the nutritional composition of

TABLE V
DEEP LEARNING CLASSIFICATION APPROACHES

Authors	Techniques	Database	Top-1 accuracy	Top-5 accuracy
Kagaya et al. 2014, [77]	CNN	Own database	73.7%	n/a
Christodoulidis et al. 2015, [73]	Patch-wise CNN		84.9%	n/a
Pouladzadeh et al. 2016, [80]	CNN		99%	n/a
Termritthikun et al. 2017, [81]	NU-InNet1.0		69.8%	92.3%
He et al. 2020, [78]	CNN		88.7	
Konstantakopoulos et al. 2021, [85]	DCNN	MedGRFood	83.4%	97.8%
Ciocca et al. 2016, [13]	CNN	UNIMIB2016	78.3%	n/a
Mezgec and Seljak 2017, [40]	NutriNet		86.4%	n/a
Chen and Ngo 2016, [24]	Arch–D	VIREO-172	82.1%	95.9%
Min et al. 2019, [94]	IG-CMAN		90.6%	98.4%
Metwalli et al. 2020, [95]	DenseFood		81.2%	95.4%
Kawano and Yanai 2014, [82]	Pre-trained DCNN	UEC-Food100	72.3%	92.0%
Yanai and Kawano 2015, [87]	DCNN-Food		78.8%	95.2%
Hassannejad et al. 2016, [90]	Inception V3		81.5%	97.3%
Liu et al. 2016, [89]	DeepFood		76.3%	94.6%
Liu et al. 2017, [83]	Inception Module		77.5%	94.6%
Martinel et al. 2018, [79]	WISeR		89.6%	99.2%
Arslan et al. 2021 [92]	ResNeXt101 & DenseNet161		90.0%	-
Yanai and Kawano 2015, [87]	DCNN-Food	UEC-Food256	67.6%	89.0%
Hassannejad et al. 2016, [90]	Inception V3		76.2%	92.6%
Liu et al. 2016, [89]	DeepFood		54.7%	81.5%
Liu et al. 2017, [83]	Inception Module		54.5%	87.0%
Martinel et al. 2018, [79]	WISeR		83.2%	93.4%
Zhao et al. 2020, [93]	JDNet		84.0%	96.2%
Bossard et al. 2014, [23]	CNN	Food-101	56.4%	n/a
Yanai and Kawano 2015, [87]	DCNN-Food		70.4%	n/a
Meyers 2015, [52]	GoogLeNet		79.0%	n/a
Hassannejad et al. 2016, [90]	Inception V3		88.3%	96.9%
Liu et al. 2016, [89]	DeepFood		77.4%	93.7%
Chen and Ngo 2016, [24]	Arch–D		82.1%	97.3%
Pandey et al. 2017, [32]	Ensemple Net		72.1%	91.6%
Liu et al. 2017, [83]	Inception Module		77.0%	94.0%
Cui et al. 2018, [88]	DSTL		90.4%	n/a
Martinel et al. 2018, [79]	WISeR		90.3%	98.7%
Tan and Le 2019, [91]	EfficientNetB7		93.0%	n/a
Merchant and Pande 2019, [84]	ConvFood		70.0%	n/a
Min et al. 2019, [94]	IG-CMAN		90.4%	98.4%
Zhao et al. 2020, [93]	JDNet		91.2%	98.8%
VijayaKumari et al.2022, [86]	EfficientNetB0		80.0%	-

The highlighted significance refers to the food image database used in each classification approach.

a food is a challenging task, even for nutritionists. This is why in many nutritional estimation systems; it is considered appropriate to have a reference object to determine the depth of the image. The metrics which are used to evaluate the volume of food are: the mean absolute error (MAE):

$$MAE = \frac{1}{n}\sum_{j=1}^{n} \left|V_{real} - V_{est}\right|, \tag{8}$$

the mean absolute percentage error (MAPE):

$$MAPE_i = \frac{1}{n}\sum_{j=1}^{n} \left|\frac{V_{real} - V_{est}}{V_{real}}\right| * 100, \tag{9}$$

and the root mean square error (RMSE):

$$RMSE = \frac{1}{n}\sqrt{\sum_{j=1}^{n} (V_{real} - V_{est})^2}, \tag{10}$$

where V_{real} is the real volume of food, V_{est} is the estimated volume and n is the total number of foods. Having estimated the amount of food, using local food composition databases, its nutritional composition can be calculated.

Several studies require taking two or more images of the food for its 3D reconstruction [96], [97]. The first step in these studies is the feature points extraction, using appropriate feature extraction algorithms, among others SIFT and SURF.

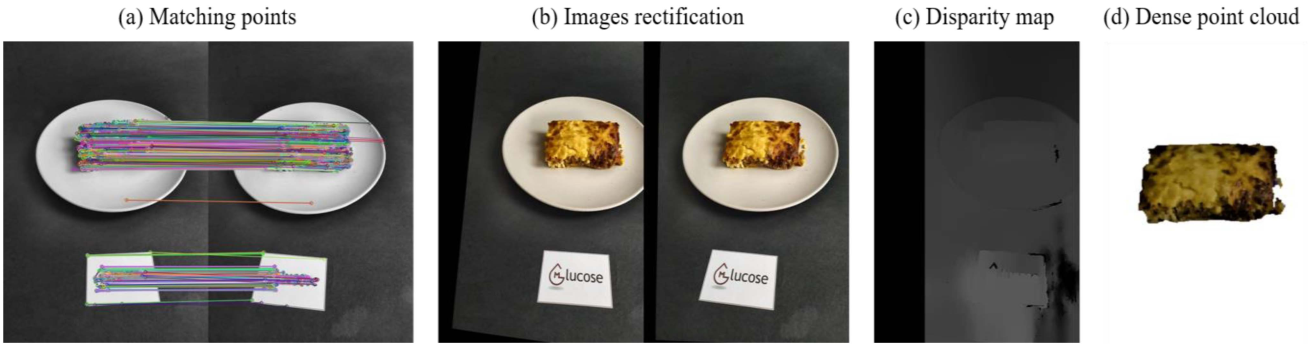

Fig. 11. Dense reconstruction steps of two captured images.

Then, the relative camera pose is estimated between the captured images. Furthermore, reference objects, with known dimensions, are used to estimate the scale of the image, for instance a reference card. Consequently, dense stereo matching is utilized for 3D food reconstruction, projecting the image coordinate system to the world coordinate. The next step is to estimate the volume of the food by removing the background from the image and keeping only the food in it. Finally, the nutritional composition of the food is analysed using the relevant nutrient database, such as the USDA Food and Nutrient Database for Dietary Assessment (FNDDS) [98]. Dehais et al. [99], estimated the volume of multi-food meals by capturing two images, with the food placed inside an elliptical plate and a reference card placed next to it. The proposed system comprised of three stages. The first stage is extrinsic calibration (computation of camera rotation and translation matrices) which is performed in three steps: salient point matching, relative pose extraction and scale extraction. The second stage is dense reconstruction, which also consists of three steps: rectification of the images, stereo matching and point cloud generation (Fig. 11). Volume estimation is the final stage, which consists of the following steps: food surface extraction, dish surface extraction and volume calculation. The system was evaluated on 77 food dishes of known volume, and achieved MAPE from 8.2 - 9.8% in two different datasets. It is worth mentioning that the researchers in order to extract the relative pose, modified the classical Random sample consensus (RANSAC) algorithm by including local optimization and an adaptive threshold estimation method. 3D food reconstruction is a methodology that can be used in a food of any shape and in capturing food images in a non-controlled environment. However, the need to capture at least two images, as well as to extract the features using image processing algorithms, such as SIFT or SURF, makes the methodology sensitive to the acquisition of images and make the process significantly slower, affecting food volume estimation accuracy.

Some studies suggest the use of specific geometrical shapes or templates (for example spherical and cylindrical objects) to reconstruct the food image from the 2D space into the 3D space from a single image [100], [101], [102]. Moreover, they utilize a fiducial marker (a checkboard pattern or a reference card) to obtain the camera parameters and provide a reference for the object scale and pose of each food item. The requirement for predefined geometrical shapes or templates for the 3D reconstruction of food, renders these methods extremely difficult to use in systems for daily dietary monitoring, because of the different and irregular shapes that food items present. For instance, in [103], the dimensions of the reference object used by the user must be pre-registered, to be able to calculate the real size of the food region. They assume that the food portion height is correlated with the food size, and they estimate calories of food items directly from the food size. For this purpose, they utilize quadratic curve estimation of food calories based on their 2D size. The quadratic curve of each food is calculated based on data annotated with real food calories. This approach gives good results in foods that have a regular shape, such as lasagna and cheesecake. Otherwise, the calculation of the amount of food is inaccurate and must be used in conjunction with methodologies for volume estimation of food having irregular shape. For food items that have irregular 3D shapes, researchers suggest using area-based volume estimation methods from a single image [104], [105]. The pinhole camera model provides a perspective transformation from the 3D plane to the 2D plane [106]. Perspective transformation is a linear projection where 3D objects are projected on a picture plane. This causes distant objects to appear smaller the nearest ones and also means that lines which are parallel appear to intersect in the projected image. In order to accurately determine the food region, the 2D image should be rectified, so that the projective distortion may be removed. In this case, the existence of a reference object in the 2D image is a prerequisite [107]. In, [108] they have proposed a system which requires the user's thumb placed beside the dish when capturing the picture. Then the system, which already knows the dimensions of user's thumb, can calculate the food area of each food item, and multiplies the total area of food (TA) by the depth (d) of the image to estimate its volume. The advantage of perspective transformation methodology is that it can handle irregular food shapes based on a single image. Its disadvantages are that it requires a special capture of food images and that the distance cannot be computed accurately.

In order to obtain the depth of the food image, the use of special devices and sensors is suggested in some studies. In [109], new generation smartphone cameras (Time of Flight

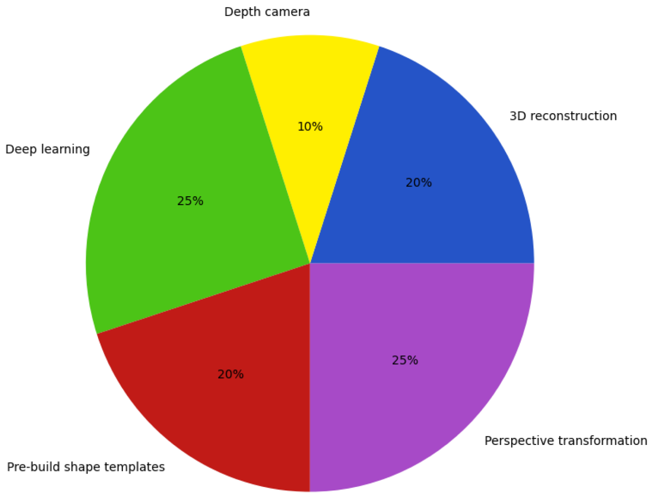

Fig. 12. Percentage use of each volume estimation approach.

(ToF) sensor or depth-sensing camera) were utilized to estimate depth and distance, where a pair of rear cameras can create the depth map in real time. The use of an additional depth camera to calculate the depth makes this approach less popular. However, with the development of technology which captures 3D images using smartphones, the depth camera methodology is expected to dominate in the next years. At the moment, the high cost of these smartphones prohibits the use of such technology.

In recent years with the ever-increasing use of deep learning networks in computer vision problems [110], they have been used in food volume estimation problems. Moreover, the ever-increasing computing power has allowed the use of Generative Adversarial Networks (GANs) to estimate the amount of food [111], providing a new dimension in the solution of this problem. In [112], a CNN is employed to deduce the depth from RGB food images to be used in Bread Units (BU) regression. This is why they have created a large- scale dataset of around 9 K different RGB-D images of 60 western dishes taken using a Microsoft Kinect v2 sensor. They have proved that depth maps from RGB images can replace RGB-D input data at high importance for the BU regression task. In another study [113], GANs are utilized to estimate food energy distribution. For the GANs training, they have created a food image dataset, which consists of 1875 paired images, based on ground truth food labels and segmentation masks for each food image including energy information correlated with the food image. The average energy estimation error is 10.89%. In Fig. 12 we can observe a quasi-even use of different food volume estimation approaches, except for depth-camera -based ones, with deep learning and perspective transformation covering each 25% of the studies. Table VI summarizes the main food volume estimation approaches, along with the techniques used to estimate the amount of food and their performance.

VI. Discussion

The 21st century is characterized as the century of data explosion. With the AI and the Internet of Things (IoT) becoming omnipresent technologies, we now have a huge amount of data being created. Since the enormous volume of image data we receive is not structured, we rely on advanced techniques, such as machine learning for efficient image analysis. Food image database, food image segmentation, food classification and food volume estimation are parts of image analysis and can be used to dietary assessment systems as part of mobile health (mHealth) applications, capturing images through a smartphone. This is what today is used and it is easy to use by most of the people and of all ages to capture photos and more specifically food images, that will offer the possibility of continuous recording of health data in real time. The use of mobile devices and cloud technology to monitor health data and sharing it with physicians, can lead to faster and less misdiagnosis of diseases, such as diabetes and CVDs. In vision-based dietary assessment systems, all stages are important towards building a reliable integrated system for food nutrition analysis. Although the dietary assessment systems have been researched for many years, several challenges remain to be explored.

The way food images are captured plays an important role in the individual steps of these systems. For both the creation of the databases and their input in the food analysis systems, the way the images are taken affects the performance of segmentation, classification, and volume estimation. In the database creation, similar foods must be captured in a way that emphasizes their different features. To input food images in the dietary assessment system, many applications require capturing images from specific shooting angles [99] and with specific objects placed next to them [108]. These prerequisites make it difficult to use these applications and prevent users from employing them, which renders it imperative to create simpler systems.

In food image databases, the use of deep learning techniques for food recognition tends to create databases with the largest possible number of images for each food class. However, the existing databases are limited to the number of food classes, depending on the dietary habits of the database constructor. Thus, there is a necessity to create a generic food image database which covers as many food categories as possible and represents the types of food from all cuisines. The collection of food images and the creation of food image databases is an easier task nowadays, due to the habit of capturing and posting images on social media. However, creating a database that will additionally include the ingredients of the food or its weight, is still a demanding task. Furthermore, creating an annotated database of food images using their weight in addition to the type of food, will help build better and more accurate models for the next steps of nutritional analysis systems. Also, one possible way to increase the number of images per food class is to use GANS models. Finally, it is worth mentioning that the acquisition of databases remains difficult, and the creation of a unified food image database cannot be achieved.

In several recent studies, the step of food image segmentation is omitted and in some others the performance of this step is not reported. In other studies, although the performance of the methods used to segment food images is high and improves the classification accuracy, there are still open issues related to cases where there are mixed foods. There are also open issues in cases where lighting conditions can create shadows or reflections in the

TABLE VI
FOOD VOLUME ESTIMATION APPROACHES

Authors	Approach	Techniques	Performance
Rahman et al. 2012, [96]	A pair of stereo images for dense reconstruction	• Feature matching & stereo rectification • Camera calibration • Disparity and depth map generation • 3D reconstruction & 3D volume estimation	The average error is 7.7%, for 6 fruits
Dehais et al. 2017, [99]	Two-view dense 3D reconstruction	• Extrinsic calibration • Dense reconstruction • Volume estimation	Mean absolute percentage error ranging from 8.2% to 9.8% in two different datasets
Gao et al. 2018, [97]	Make use of a monocular wearable camera based on SLAM system for food volume estimation	• Sparse-map generation by SLAM method • Apply convex-hull algorithm to form a 3D mesh object • Volume estimation based on 3D mesh object	Mean volume estimation error is 15.98% and 20.54% on static food and on during food consumption respectively
Konstantakopoulos et al. 2021, [31]	Structure from motion 3D reconstruction with a reference card	• Feature matching and pose estimation • Stereo matching and 3D reconstruction • Scale determination and volume estimation	MAPE from 4.6% to 11.1% for seven types of food
Xu et al. 2013, [100]	Make use of a pre-built 3D model of food items and then compute the pose estimation	• 3D model generation • Pose initialization • Pose finalization	The average error is 10%, for 5 types of food
Jia et al. 2014, [101]	An electronic device (eButton) captures images every 2 to 4 sec.	• A self-developed image undistortion algorithm applied to the image with the best quality • A virtual shape method used to measure the portion size	85 out of the 100 food items have less than 30% error
Fang et al. 2015, [102]	Singe-view 3D reconstruction using the geometric contextual information from the scene	• Points of interest estimation • Area-based volume estimation through the prism model • Weight estimation through the food density	Achieve less than 6% error in energy estimation
Okamoto and Yanai 2016, [103]	Food calorie estimation by a single image	• A user needs to register a known size reference object • Quadratic curve is estimated from the 2D size of foods to their calories • Quadratic curve is trained based on the training data annotated with real food calories independently	Relative average error on calorie estimation is 21.3%, for 60 food images
Jia et al. 2012, [104]	Utilize the plate and LED methods	• Object location and orientation using the plate method • Object location and orientation using the LED method	The average error is 12.01% for the plate method and 29.01%, for the LED method
Yue et al.2012, [106],	Single image with a known size circular container	• Perspective transformation • Orientation estimation • Dimension estimation	Average length and thickness error estimation 3.41%
He et al.2013, [105],	Use the methods: Shape template 3D reconstruction and area-based weight estimation for foods	• Camera calibration • Camera pose information • Shape template method to reconstruct a 3D food item • Food area and weight estimation	The average error is 11%, for beverage images using cylinder shape, and 10% for area-based weight estimation
Pouladzadeh et al. 2014, [108]	Captured two photos (over and side of the food), with the user's thumb as reference	• Food area measurement from top view captured image • Depth estimation from the side-view captured image	For 5 types of food, 10% error in the worst case and less than 1% error in the best case
Yang et al. 2019, [107]	A fiducial marker free method, making use of the smartphone motion sensor	• A smartphone motion sensor determines camera orientation • The length or the width of the smartphone determines the location of any visible point on the tabletop • The food image captures with a special way	The average absolute error is 16.65% for ten types of food
Chen et al. 2012, [28]	Depth camera (Microsoft Kinect)	• Calculate the area of food container • Calculate the depth value of the contained food	The system performance is not reported
Ando et al. 2019, [109]	Food volume and calorie estimation using depth camera	• Take a RGB-D food image • Estimate volumes of food on the dish • Calculate foods calories using the pre-registered calorie density of each food category.	The proposed system achieves higher accuracy than CalorieCam and AR CalorieCam V2 applications
Meyers et al. 2015, [52]	Use a CNN model to estimate the 3D volume	• Use a CNN architecture to predict the depth map • Convert the depth map to voxel representation	Until 400ml absolute error across the 11 meals in the NFood-3d dataset
Christ at al. 2017, [112]	State of the art deep learning methods	• Predict the depth map using a CNN • Estimate the bread units using ResNet-50	RMSE for bread units is 1.53. Food categories were evaluated from Diabetes60 dataset
Fang et al. 2018, [113]	Use of generative model for food energy estimation	• GAN is trained on paired images to map a food image to its equivalent energy distribution image	10.89% energy estimation error. 2095 paired images were used for the generative network.
Fang et al. 2019, [111]	Estimate food energy based on learned energy distribution images	• Use GAN to estimate the image to energy mappings • A CNN-regression model estimates the energy value based the learned energy distribution images	Average food energy estimation error 209 kcal for 347 food images
Lo et al. 2019, [110]	Vision-based method using real-time 3D reconstruction and deep learning view synthesis	• A mobile phone with depth sensors is captured a single depth image • A fine-tuned Mask R-CNN are segmented the food items • The depth image is converted from image to camera coordinate • The partial point cloud is directed to the point completion network to perform 3-D reconstruction • The portion size of food items is estimated	The average error ranging from 15 to 79 cm3 for eleven types of food

TABLE VII
COMPARISON OF EXISTING REVIEW STUDIES

	V. Bruno et al., *J. Health Med. Inform.*, 2017	M. C. Archundia Herrera et al., *Nutrients*, 2018	W. Min, *et al.*, *ACM Comput. Surv.*, 2019	F.P. Wen Lo, *et al.*, *IEEE JBHI*, 2020	Wang, *et. al.*, *Trends Food Sci. Technol.*, 2022	This review
Analysis of Food Image Databases						
Reporting general database information	✓	-	✓	-	✓	✓
Reporting # of images and # of classes	✓	-	✓	-	✓	✓
Reporting the image acquisition process	✓	-	✓	-	-	✓
Reporting database use	-	-	✓	-	✓	✓
Reporting pros and cons	-	-	✓	-	✓	✓
Reporting future directions	✓	-	✓	-	✓	✓
Analysis of Food Image Segmentation Techniques						
Categorisation of techniques to:						
Semi – automatic approaches	-	-	-	-	-	✓
Automatic ML-based approaches with handcrafted feature extraction	-	-	-	-	-	✓
Automatic ML-based approaches using deep learning for feature extraction	-	-	-	-	✓	✓
Reporting the description of the approach followed for each of the reviewed study	✓	-	-	-	✓	✓
Reporting of performance	✓	-	-	-	✓	✓
Reporting performance metrics	-	-	-	-	✓	✓
Reporting pros and cons	✓	-	-	-	-	✓
Reporting future directions	-	-	-	-	-	✓
Analysis of Food Image Classification Techniques						
Categorisation of techniques to ML & DL approaches	✓	-	✓	✓	✓	✓
Reporting the database used	✓	-	-	✓	✓	✓
Reporting of performance	✓	-	✓	✓	✓	✓
Reporting performance metrics	-	-	-	-	✓	✓
Reporting pros and cons	✓	-	✓	✓	-	✓
Reporting future directions	-	-	✓	✓	✓	✓
Analysis of Food Volume Estimation Techniques						
Categorisation of techniques to:						
3D reconstruction	-	-	-	✓	✓	✓
Pre-build shape template	-	-	-	✓	-	✓
Perspective transformation	-	-	-	✓	-	✓
Depth camera	-	-	-	✓	-	✓
Deep learning	-	-	-	✓	-	✓
Reporting the description of the approach followed for each of the reviewed study	✓	✓	-	✓	-	✓
Reporting of performance	✓	✓	-	✓	✓	✓
Reporting performance metrics	-	-	-	-	✓	✓
Reporting pros and cons	✓	✓	-	✓	✓	✓
Reporting future directions	-	✓	-	✓	✓	✓
Reporting food intake monitoring devices and apps	-	✓	✓	-	-	-

image or blurring the food items contained in the image. In these cases, the use of state-of-the-art segmentation techniques, such as semantic and instance segmentation, can be used to improve the performance of this step and improve the efficiency to the classification step.

Studies have shown that deep learning techniques perform better than traditional food image classification techniques and that is the reason why they are considered the state-of-the-art methods for food image classification. To classify food images, as mentioned above, databases with a large number of food images are required. This requirement becomes even bigger for deep learning techniques, where the number of images in the database affects the performance of the food image classification system. In addition, blurred images, inadequate lighting conditions when capturing them and the different ways of cooking the same food, can lead to misidentification of the food. The use of deeper classification models and the application of transfer learning, fine tuning, and data augmentation techniques, could improve the accuracy of deep learning classification models. The use of pre-trained DNNs in existing food image databases could lead to the construction of models with better accuracy and even lower loss.

Volume and nutrient estimation are the most challenging task in automated vision-based dietary assessment systems. The controlled environment for capturing food images, taking multiple photos, the inability to estimate the volume of food with weak texture features, for instance yogurt, and the creation of databases according to the techniques used in each study, render the estimation of the amount of food through images the most demanding stage for nutrient analysis systems. In addition, the need to use a reference object or the use of a depth camera to calculate the scale and quantity of food, limits their possibility for extensive use. Moreover, food estimation techniques based on geometric patterns allow volume estimation to be calculated in only few foods which have a specific shape. Finally, although the recent use of deep learning techniques in food volume estimation was a very promising approach, studies have shown that they do not outperform the existing techniques. In the 3D reconstruction approach, CNNs could be used instead of image processing algorithms to extract the features, significantly increasing the number of matched features and improving the reconstruction of food 3D point cloud. One possible approach that would solve many problems regarding the way images are captured, the number of images required and the depth sensors needed would be to build a machine learning model on an annotated food image database with regard to the weight of the food items.

Considering the continuous technological development and the techniques of recording data, the use of alternative ways to enter data and information related to the food consumed (for example via speech or text), could help optimize the performance of nutritional analysis systems. In particular, combining traditional food recognition and quantity estimation techniques with voice and text input and processing techniques could further improve the performance of nutritional assessment systems. In addition, using advanced deep learning techniques and algorithms, such as reinforcement learning, it is possible to build dietary assessment systems based on personalized nutrition, providing dynamic dietary recommendations by monitoring the user's environment and aiming to optimize a reward function.

Table VII provides a comparative assessment of existing review studies including our work with respect to the elements of dietary assessment systems that are reviewed and assessed therein. Considering the level of information (quality, quantity, and granularity) provided by the existing reviews, herein, we aimed at improving the completeness of the information by reviewing all the elements of such a system (Sections II-V) and unbiasedly capturing all the different classes of methods/techniques/algorithms that have been proposed over the last 10 years in the specified research topic. In this direction, the above discussion of both the strengths and limitations of the existing approaches alongside the identification of solutions to their shortcomings aimed at strengthening future research works.

VII. Conclusion

This review study assessed and contrasted the methods constituting the intelligence logic of a dietary assessment system aiming at providing to the reader the potentialities of the existing approaches. First, we highlighted the need for annotated food image databases including meals from multiple cuisines and with adequate size per class in view of their use as training/test sets in image segmentation or image classification tasks. Second, we stressed the potential of instance and semantic image segmentation approaches to augment the performance of food classification models orchestrated under the same pipeline. Third, we verified, as it was expected, the superiority of deep learning architectures in classifying the content of food images over conventional machine learning algorithms, and the tendency of increasing the number of hidden layers towards increasing the accuracy of predictions. Finally, further annotation of food images (e.g., with respect to their weight) could complement the current functionality of food volume estimation approaches.

References

[1] World Health Organization, *Obesity and overweight*. Accessed: Jun. 9, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

[2] World Health Organization, *Diabetes*. Accessed: Sep. 16, 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes

[3] International Diabetes Federation, *Diabetes facts & figures*. Accessed: Dec. 19, 2021. [Online]. Available: https://www.idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html

[4] World Health Organization, *Cardiovascular diseases (CVDs)*. Accessed: Jun. 11, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[5] M. Rusin et al., "Functionalities and input methods for recording food intake: A systematic review," *Int. J. Med. Inform.*, vol. 82, no. 8, pp. 653–664, 2013.

[6] M. B. E. Livingstone et al., "Issues in dietary intake assessment of children and adolescents," *Brit. J. Nutr.*, vol. 92, no. S2, pp. S213–S222, 2004.

[7] T. Hernández et al., "Portion size estimation and expectation of accuracy," *J. Food Comp. Anal.*, vol. 19, pp. S14–S21, 2006.

[8] M. R. Graff et al., "How well are individuals on intensive insulin therapy counting carbohydrates?," *Diabetes Res. Clin. Pract.*, vol. 50, pp. 238–239, 2000.

[9] F. S. Konstantakopoulos et al., "GlucoseML Mobile application for automated dietary assessment of mediterranean food," in *Proc. 44th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2022, pp. 1432–1435.

[10] J. Ngo et al., "A review of the use of information and communication technologies for dietary assessment," *Brit. J. Nutr.*, vol. 101, no. S2, pp. S102–S112, 2009.

[11] F. Jiang et al., "Artificial intelligence in healthcare: Past, present and future," *Stroke Vasc. Neurol.*, vol. 2, no. 4, pp. 230–243, 2017.

[12] M. C. Carter et al., "Adherence to a smartphone application for weight loss compared to website and paper diary: Pilot randomized controlled trial," *J. Med. Internet Res.*, vol. 15, no. 4, 2013, Art. no. e32.

[13] G. Ciocca et al., "Food recognition: A new dataset, experiments, and results," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 3, pp. 588–598, May 2017.

[14] Y. Matsuda et al., "Recognition of multiple-food images by detecting candidate regions," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2012, pp. 25–30.

[15] F. Zhu et al., "Multiple hypotheses image segmentation and classification with application to dietary assessment," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 1, pp. 377–388, Jan. 2015.

[16] S. J. Minija and W. S. Emmanuel, "Food recognition using neural network classifier and multiple hypotheses image segmentation," *Imag. Sci. J.*, vol. 68, no. 2, pp. 100–113, 2020.

[17] P. Pouladzadeh et al., "Using graph cut segmentation for food calorie measurement," in *Proc. IEEE Int. Symp. Med. Meas. Appl.*, 2014, pp. 1–6.

[18] V. Bruno and C. J. Silva Resende, "A survey on automated food monitoring and dietary management systems," *J. Health Med. Inform.*, vol. 8, no. 3, 2017.

[19] F. P. W. Lo et al., "Image-based food classification and volume estimation for dietary assessment: A review," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 7, pp. 1926–1939, Jul. 2020.

[20] M. C. Archundia Herrera and C. B. Chan, "Narrative review of new methods for assessing food and energy intake," *Nutrients*, vol. 10, no. 8, 2018, Art. no. 1064.

[21] W. Min et al., "A survey on food computing," *J. ACM Comput. Surv.*, vol. 52, no. 5, pp. 1–36, 2019.

[22] W. Wang et al., "A review on vision-based analysis for automatic dietary assessment," *Trends Food Sci. Technol.*, vol. 122, pp. 223–237, 2022.

[23] L. Bossard et al., "Food-101–mining discriminative components with random forests," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 446–461.

[24] J. Chen and C.-W. Ngo, "Deep-based ingredient recognition for cooking recipe retrieval," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 32–41.

[25] Y. Kawano and K. Yanai, "Automatic expansion of a food image dataset leveraging existing categories with domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 3–17.

[26] M. M. Anthimopoulos et al., "A food recognition system for diabetic patients based on an optimized bag-of-features model," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 4, pp. 1261–1271, Jul. 2014.

[27] G. Ciocca et al., "Learning CNN-based features for retrieval of food images," in *Proc. Int. Conf. Image Anal. Process.*, 2017, pp. 426–434.

[28] M.-Y. Chen et al., "Automatic Chinese food identification and quantity estimation," in *Proc. SIGGRAPH Asia Tech. Briefs*, 2012, pp. 1–4.

[29] X. Chen et al., "Chinesefoodnet: A large-scale image dataset for Chinese food recognition," 2017, *arXiv:1705.02743*.

[30] I. Donadello and M. Dragoni, "Ontology-driven food category classification in images," in *Proc. Int. Conf. Image Anal. Process.*, 2019, pp. 607–617.

[31] F. Konstantakopoulos et al., "3D reconstruction and volume estimation of food using stereo vision techniques," in *Proc. IEEE 21st Int. Conf. Bioinf. Bioeng.*, 2021, pp. 1–4.

[32] P. Pandey et al., "FoodNet: Recognizing foods using ensemble of deep networks," *IEEE Signal Process. Lett.*, vol. 24, no. 12, pp. 1758–1762, Dec. 2017.

[33] S. Hou et al., "VegFru: A domain-specific dataset for fine-grained visual categorization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 541–549.

[34] G. Waltner et al., "Personalized dietary self-management using mobile vision-based assistance," in *Proc. Int. Conf. Image Anal. Process.*, 2017, pp. 385–393.

[35] H. Muresan and M. Oltean, "Fruit recognition from images using deep learning," *J. Acta Univ. Sapientiae*, vol. 10, no. 1, pp. 26–42, 2018.

[36] Q. Yu et al., "Food image recognition by personalized classifier," in *Proc. IEEE 25th Int. Conf. Image Process.*, 2018, pp. 171–175.

[37] P. Kaur et al., "Foodx-251: A dataset for fine-grained food classification," 2019, *arXiv:1907.06167*.

[38] O. Beijbom et al., "Menu-match: Restaurant-specific food logging from images," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2015, pp. 844–851.

[39] X. Wang et al., "Recipe recognition with large multimodal food dataset," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2015, pp. 1–6.

[40] S. Mezgec and B. Koroušić Seljak, "NutriNet: A deep learning food and drink image recognition system for dietary assessment," *Nutrients*, vol. 9, no. 7, 2017, Art. no. 657.

[41] G. M. Farinella et al., "A benchmark dataset to study the representation of food images," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 584–599.

[42] F. Zhou and Y. Lin, "Fine-grained image classification by exploring bipartite-graph labels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1124–1133.

[43] A. Singla et al., "Food/non-food image classification and food categorization using pre-trained googlenet model," in *Proc. 2nd Int. Workshop Multimedia Assist. Dietary Manage.*, 2016, pp. 3–11.

[44] G. M. Farinella et al., "Retrieval and classification of food images," *Comput. Biol. Med.*, vol. 77, pp. 23–39, 2016.

[45] E. Aguilar et al., "Regularized uncertainty-based multi-task learning model for food analysis," *J. Vis. Commun. Image Representation*, vol. 60, pp. 360–370, 2019.

[46] J. Gao et al., "MUSEFood: Multi-sensor-based food volume estimation on smartphones," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov.*, 2019, pp. 899–906.

[47] T. Ege et al., "A new large-scale food image segmentation dataset and its application to food calorie estimation based on grains of rice," in *Proc. 5th Int. Workshop Multimedia Assist. Dietary Manage.*, 2019, pp. 82–87.

[48] K. Okamoto and K. Yanai, "UEC-FoodPIX complete: A large-scale food image segmentation dataset," in *Proc. Int. Conf. Pattern Recognit.*, 2021, pp. 647–659.

[49] X. Wu et al., "A large-scale benchmark for food image segmentation," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 506–515.

[50] S. Aslan et al., "Benchmarking algorithms for food localization and semantic segmentation," *Int. J. Mach. Learn. Cybern.*, vol. 11, no. 12, pp. 2827–2847, 2020.

[51] Y. Wang et al., "Mixed dish recognition through multi-label learning," in *Proc. 11th Workshop Multimedia Cooking Eating Activities*, 2019, pp. 1–8.

[52] A. Meyers et al., "Im2Calories: Towards an automated mobile vision food diary," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1233–1241.

[53] Y. Kawano and K. Yanai, "Real-time mobile food recognition system," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2013, pp. 1–7.

[54] S. Inunganbi et al., "Classification of food images through interactive image segmentation," in *Proc. Asian Conf. Intell. Inf. Database Syst.*, 2018, pp. 519–528.

[55] W. Shimoda and K. Yanai, "CNN-based food image segmentation without pixel-wise annotation," in *Proc. Int. Conf. Image Anal. Process.*, 2015, pp. 449–457.

[56] S. Fang, C. Liu et al., "cTADA: The design of a crowdsourcing tool for online food image identification and segmentation," in *Proc. IEEE Southwest Symp. Image Anal. Interpretation*, 2018, pp. 25–28.

[57] H. Hassannejad et al., "A mobile app for food detection: New approach to interactive segmentation," in *Proc. FORITAAL Conf.*, 2015, pp. 19–22.

[58] Y. Wang et al., "Efficient superpixel based segmentation for food image analysis," in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 2544–2548.

[59] X. Zheng et al., "Image segmentation based on adaptive K-means algorithm," *EURASIP J. Image Video Process.*, vol. 2018, no. 1, 2018, Art. no. 68.

[60] M. Anthimopoulos et al., "Segmentation and recognition of multi-food meal images for carbohydrate counting," in *Proc. 13th IEEE Int. Conf. Bioinf. Bioeng.*, 2013, pp. 1–4.

[61] J. Dehais et al., "Food image segmentation for dietary assessment," in *Proc. 2nd Int. Workshop Multimedia Assist. Dietary Manage.*, 2016, pp. 23–28.

[62] G. Ciocca et al., "Evaluating CNN-based semantic food segmentation across illuminants," in *Proc. Int. Workshop Comput. Color Imag.*, 2019, pp. 247–259.

[63] P. Poply and J. A. A. Jothi, "Refined image segmentation for calorie estimation of multiple-dish food items," in *Proc. Int. Conf. Comput., Commun., Intell. Syst.*, 2021, pp. 682–687.

[64] M. Bolaños and P. Radeva, "Simultaneous food localization and recognition," in *Proc. IEEE 23rd Int. Conf. Pattern Recognit.*, 2016, pp. 3140–3145.

[65] S. K. Yarlagadda et al., "Saliency-aware class-agnostic food image segmentation," *ACM Trans. Comput. Healthcare*, vol. 2, no. 3, pp. 1–17, 2021.

[66] D. Park et al., "Deep learning based food instance segmentation using synthetic data," in *Proc. 18th Int. Conf. Ubiquitous Robots*, 2021, pp. 499–505.

[67] K. J. Pfisterer et al., "When Segmentation is Not Enough: Rectifying Visual-Volume Discordance Through Multisensor Depth-Refined Semantic Segmentation for Food Intake Tracking in Long-Term Care," 2019, *arXiv:1910.11250*.

[68] C. Szegedy et al., "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.

[69] E. Aguilar et al., "Grab, pay, and eat: Semantic food detection for smart restaurants," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3266–3275, Dec. 2018.

[70] S. Aslan et al., "Semantic food segmentation for automatic dietary monitoring," in *Proc. IEEE 8th Int. Conf. Consum. Electron.-Berlin*, 2018, pp. 1–6.

[71] W. Shimoda and K. Yanai, "Weakly-supervised plate and food region segmentation," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2020, pp. 1–6.

[72] H.-T. Nguyen and C.-W. Ngo, "Terrace-based food counting and segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, pp. 2364–2372.

[73] S. Christodoulidis et al., "Food recognition for dietary assessment using deep convolutional neural networks," in *Proc. Int. Conf. Image Anal. Process.*, 2015, pp. 458–465.

[74] P. Pouladzadeh et al., "Cloud-based SVM for food categorization," *Multimedia Tools Appl.*, vol. 74, no. 14, pp. 5243–5260, 2015.

[75] Y. He et al., "Analysis of food images: Features and classification," in *Proc. IEEE Int. Conf. Image Process.*, 2014, pp. 2744–2748.

[76] Y. Kawano and K. Yanai, "Foodcam-256: A large-scale real-time mobile food recognitionsystem employing high-dimensional features and compression of classifier weights," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 761–762.

[77] H. Kagaya et al., "Food detection and recognition using convolutional neural network," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 1085–1088.

[78] J. He et al., "Multi-task image-based dietary assessment for food recognition and portion size estimation," in *Proc. IEEE Conf. Multimedia Inf. Process. Retrieval*, 2020, pp. 49–54.

[79] N. Martinel et al., "Wide-slice residual networks for food recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 567–576.

[80] P. Pouladzadeh et al., “Food calorie measurement using deep learning neural network,” in *Proc. IEEE Int. Instrum. Meas. Technol. Conf.*, 2016, pp. 1–6.

[81] C. Termritthikun et al., “NU-InNet: Thai food image recognition using convolutional neural networks on smartphone,” *J. Telecommun. Electron. Comput. Eng.*, vol. 9, no. 2/6, pp. 63–67, 2017.

[82] Y. Kawano and K. Yanai, “Food image recognition with deep convolutional features,” in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.: Adjunct Pub.*, 2014, pp. 589–593.

[83] C. Liu et al., “A new deep learning-based food recognition system for dietary assessment on an edge computing service infrastructure,” *IEEE Trans. Serv. Comput.*, vol. 11, no. 2, pp. 249–261, Mar./Apr. 2018.

[84] K. Merchant and Y. Pande, “ConvFood: A CNN-based food recognition mobile application for obese and diabetic patients,” in *Emerging Research in Computing, Information, Communication and Applications*. Berlin, Germany: Springer, 2019, pp. 493–502.

[85] F. S. Konstantakopoulos et al., “Mediterranean food image recognition using deep convolutional networks,” in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2021, pp. 1740–1743.

[86] G. VijayaKumari et al., “Food classification using transfer learning technique,” *Glob. Transitions Proc.*, vol. 3, pp. 225–229, 2022.

[87] K. Yanai and Y. Kawano, “Food image recognition using deep convolutional network with pre-training and fine-tuning,” in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2015, pp. 1–6.

[88] Y. Cui et al., “Large scale fine-grained categorization and domain-specific transfer learning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4109–4118.

[89] C. Liu et al., “DeepFood: Deep learning-based food image recognition for computer-aided dietary assessment,” in *Proc. Int. Conf. Smart Homes Health Telematics*, 2016, pp. 37–48.

[90] H. Hassannejad et al., “Food image recognition using very deep convolutional networks,” in *Proc. 2nd Int. Workshop Multimedia Assist. Dietary Manage.*, 2016, pp. 41–49.

[91] M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[92] B. Arslan et al., “Fine-grained food classification methods on the UEC food-100 database,” *IEEE Trans. Artif. Intell.*, vol. 3, no. 2, pp. 238–243, Apr. 2022.

[93] H. Zhao et al., “JDNet: A joint-learning distilled network for mobile visual food recognition,” *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 4, pp. 665–675, May 2020.

[94] W. Min et al., “Ingredient-guided cascaded multi-attention network for food recognition,” in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 1331–1339.

[95] A.-S. Metwalli et al., “Food image recognition based on densely connected convolutional neural networks,” in *Proc. IEEE Int. Conf. Artif. Intell. Inf. Commun.*, 2020, pp. 027–032.

[96] M. H. Rahman et al., “Food volume estimation in a mobile phone based dietary assessment system,” in *Proc. 8th Int. Conf. Signal Image Technol. Internet Based Syst.*, 2012, pp. 988–995.

[97] A. Gao et al., “Food volume estimation for quantifying dietary intake with a wearable camera,” in *Proc. IEEE 15th Int. Conf. Wearable Implantable Body Sensor Netw.*, 2018, pp. 110–113.

[98] N. K. Fukagawa et al., “USDA’s FoodData central: What is it and why is it needed today?,” *Amer. J. Clin. Nutr.*, vol. 115, no. 3, pp. 619–624, 2022.

[99] J. Dehais et al., “Two-view 3D reconstruction for food volume estimation,” *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 1090–1099, May 2017.

[100] C. Xu et al., “Model-based food volume estimation using 3D pose,” in *Proc. IEEE Int. Conf. Image Process.*, 2013, pp. 2534–2538.

[101] W. Jia et al., “Accuracy of food portion size estimation from digital pictures acquired by a chest-worn camera,” *Public Health Nutr.*, vol. 17, no. 8, pp. 1671–1681, 2014.

[102] S. Fang et al., “Single-view food portion estimation based on geometric models,” in *Proc. IEEE Int. Symp. Multimedia*, 2015, pp. 385–390.

[103] K. Okamoto and K. Yanai, “An automatic calorie estimation system of food images on a smartphone,” in *Proc. 2nd Int. Workshop Multimedia Assist. Dietary Manage.*, 2016, pp. 63–70.

[104] W. Jia et al., “Imaged based estimation of food volume using circular referents in dietary assessment,” *J. Food Eng.*, vol. 109, no. 1, pp. 76–86, 2012.

[105] Y. He et al., “Food image analysis: Segmentation, identification and weight estimation,” in *Proc. IEEE Int. Conf. Multimedia Expo*, 2013, pp. 1–6.

[106] Y. Yue et al., “Measurement of food volume based on single 2-D image without conventional camera calibration,” in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2012, pp. 2166–2169.

[107] Y. Yang et al., “Image-based food portion size estimation using a smartphone without a fiducial marker,” *Public Health Nutr.*, vol. 22, no. 7, pp. 1180–1192, 2019.

[108] P. Pouladzadeh et al., “Measuring calorie and nutrition from food image,” *IEEE Trans. Instrum. Meas.*, vol. 63, no. 8, pp. 1947–1956, Aug. 2014.

[109] Y. Ando et al., “Depthcaloriecam: A mobile application for volume-based foodcalorie estimation using depth cameras,” in *Proc. 5th Int. Workshop Multimedia Assist. Dietary Manage.*, 2019, pp. 76–81.

[110] F. P.-W. Lo et al., “Point2Volume: A vision-based dietary assessment approach using view synthesis,” *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 577–586, Jan. 2020.

[111] S. Fang et al., “An end-to-end image-based automatic food energy estimation technique based on learned energy distribution images: Protocol and methodology,” *Nutrients*, vol. 11, no. 4, 2019, Art. no. 877.

[112] P. Ferdinand Christ et al., “Diabetes 60-inferring bread units from food images using fully convolutional neural networks,” in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2017, pp. 1526–1535.

[113] S. Fang et al., “Single-view food portion estimation: Learning image-to-energy mappings using generative adversarial networks,” in *Proc. IEEE 25th Int. Conf. Image Process.*, 2018, pp. 251–255.