Review

A review on vision-based analysis for automatic dietary assessment

Wei Wang a,c,d, Weiqing Min b,e,* , Tianhao Li b,e, Xiaoxiao Dong a,c,d, Haisheng Li a,c,d,** , Shuqiang Jiang b,e

a School of Computer and Engineering, Beijing Technology and Business University, Beijing, 100048, China
b The Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China
c Beijing Key Laboratory of Big Data Technology for Food Safety, Beijing, 100048, China
d National Engineering Laboratory For Agri-product Quality Traceability, Beijing, 100048, China
e University of Chinese Academy of Sciences, Beijing, 100049, China

Abstract

Background: Maintaining a healthy diet is vital to avoid health-related issues, e.g., undernutrition, obesity and many non-communicable diseases. An indispensable part of the health diet is dietary assessment. Traditional manual recording methods are not only burdensome but time-consuming, and contain substantial biases and errors. Recent advances in Artificial Intelligence (AI), especially computer vision technologies, have made it possible to develop automatic dietary assessment solutions, which are more convenient, less time-consuming and even more accurate to monitor daily food intake.

Scope and Approach: This review presents Vision-Based Dietary Assessment (VBDA) architectures, including multi-stage architecture and end-to-end one. The multi-stage dietary assessment generally consists of three stages: food image analysis, volume estimation and nutrient derivation. The prosperity of deep learning makes VBDA gradually move to an end-to-end implementation, which applies food images to a single network to directly estimate the nutrition. The recently proposed end-to-end methods are also discussed. We further analyze existing dietary assessment datasets, indicating that one large-scale benchmark is urgently needed, and finally highlight critical challenges and future trends for VBDA.

Key Findings and Conclusions: After thorough exploration, we find that multi-task end-to-end deep learning approaches are one important trend of VBDA. Despite considerable research progress, many challenges remain for VBDA due to the meal complexity. We also provide the latest ideas for future development of VBDA, e.g., fine-grained food analysis and accurate volume estimation. This review aims to encourage researchers to propose more practical solutions for VBDA.

1. Introduction

The malnutrition such as undernutrition, overweight and obesity is now increasingly recognized as one of the greatest health and societal challenges (Ingram et al., 2020). In 2020, 39 million children under 5 years old were overweight or obese due to unhealthy diets, such as the intake of high-fat and high-energy food (World Health Organization, 2021). Recently, Cao et al. (2021) found that a high-fat diet can damage the mesenteric lymphatic vessels, and the leakage of lymphatic vessels can cause metabolic problems such as abdominal obesity and insulin resistance. It is worth mentioning that undernutrition and insulin resistance are major risk factors that may lead to many non-communicable diseases, such as cardiovascular disease, diabetes, and certain cancers (Lauby-Secretan et al., 2016). The previous International Agency for Research on Cancer (IARC) working group concluded that there was sufficient evidence for a cancer-preventive effect of avoidance of weight gain for cancers of the colon, kidney (renal cell), breast (postmenopausal), esophagus (adenocarcinoma), and corpus uteri (IARC, 2002). Fortunately, obesity and many chronic diseases can be prevented via dietary assessment, which can monitor daily food intake and control eating habits (Nordström, Coff, Jönnson, Nordenfelt, & Görmann, 2013). Furthermore, we can utilize these dietary data to analyze the relations between dietary patterns and some diseases (Chu, Yu, Chen, Tian, & Zhai, 2021), and...
support personalized nutrition (McDonald, Glusman, & Price, 2016). Therefore, dietary assessment has become the focus of widespread attention in various fields of computer vision, medicine, nutrition and health (Liu et al., 2018; McPherson, Hoelscher, Alexander, Scanlon, & Serdula, 2000; Poulaazadeh, Shirrmohammadi and Almaghrabi, 2014; Probst, Nguyen, Tran, & Li, 2015; Thames et al., 2021).

Over the years, researchers have explored various methods for dietary assessment, such as the 24-h dietary recall (24-HDR) (Foster et al., 2008; Kirkpatrick et al., 2014; Slimani et al., 2000), food frequency questionnaire (FFQ) (Forster et al., 2014; Kristal et al., 2014; Willett et al., 1985; Wong, Boushey, Novotny, & Gustafson, 2008), dietary record (DR) (Gersovitz, Madden, & Smicklas-Wright, 1978), and brief dietary assessment instruments or screeners (Illner et al., 2012). Among them, FFQ can be regarded as a long-term dietary assessment approach, while 24-HDR and dietary records are the primary subjective methods used in short-term evaluations (Shim, Oh, & Kim, 2014). The implementation of these methods mainly involves paper questionnaires and interview-based tools (Thompson et al., 2015), which have provided an important contribution to the development of nutrition research. Nevertheless, these traditional manual recording methods are time-consuming, inaccurate (e.g., energy intake under-reporting (Gibney, Allison, Bier, & Dwyer, 2020)), labor-intensive, and high-demanding for a certain level of literacy and communication skills. Therefore, traditional methods make the research process more difficult or even infeasible for special populations such as children, adolescents and the elderly. In addition, these issues make the collected dietary data unreliable and difficult to conduct large-scale assessment. To overcome these problems, some automatic methods have been developed, such as smart cooking systems and portable systems (Poulaazadeh, Shirrmohammadi and Yassine, 2016). The major limitations to these approaches are the limited or illogical usage. For example, the smart kitchen (Chi, Chen, Chu, & Lo, 2008) captures images of the food preparation process to measure all of the ingredients inside the kitchen, yet has the inability to be used outside the home. GoBe (2021) only calculates the calories after the user has already eaten the food via measuring the glucose of user's cells, which is late for patients who need to know the amount of calories before eating.

Recent advances in Artificial Intelligence (AI), especially computer vision and machine learning, have paved the way for more robust automatic dietary assessment. The widespread use of portable devices (e.g., smartphones) with enhanced capabilities together with the advancements in computer vision enabled the development of VBDA. It allows users to easily measure the nutrition and calories of food by taking pictures of food through their mobile or wearable devices. Compared with traditional methods, VBDA can provide a solution to eliminate subjectivity, get rid of time and space constraints, and enhance the comprehensiveness and accuracy of dietary intake assessment. It can not only reduce the burden of keeping food journaling, but also provide immediate dietary assessments, demonstrating great potentials in effective diet monitoring and control. Furthermore, driven by advances in deep learning (LeCun, Bengio, & Hinton, 2015), the higher capacity of automatic food visual analysis results in higher performance of dietary assessment. For these reasons, VBDA is becoming one mainstream method for dietary assessment. It should be emphasized that deep learning algorithms become computationally expensive when applied to high-dimensional data (e.g., food images), possibly due to the learning phase associated with a deep-layered hierarchy (Poulaazadeh, Shirrmohammadi et al., 2016).

Some surveys related to VBDA have been conducted. For example, Tay, Kaur, Quek, Lim, and Henry (2020) focused on highlighting various methods of food volume estimation. They pointed out that visual analysis and volume estimation are the basis of multi-stage dietary assessment. Liu, Pu, and Sun (2021) reviewed Convolutional Neural Networks (CNNs) for feature extraction from food images. We summarized both deep learning methods and traditional manual feature extraction methods (e.g., SIFT and LBP) for food visual analysis. Additionally, other articles focused on surveying various food recognition methods (Knez & Šajn, 2020; Lo, Sun, Qiu, & Lo, 2020a; Subhi, Ali, & Mohammed, 2019; Tahir & Loo, 2021), where most of these recognition works are not evaluated on the dietary assessment framework but only on the recognition task. Höchsmann and Martin (2020) illustrated only 12 different image-assisted food intake assessment methods. In contrast, we have listed more methods to analyze their effectiveness and feasibility comprehensively. In summary, different from their works, we proposed two VBDA architectures to more comprehensively and systematically review current dietary assessment methods, and discussed their advantages and limitations. In addition, most of our surveyed works are directly for VBDA. Finally, we discuss existing VBDA datasets and evaluation metrics, and the key challenges and future trends for VBDA are highlighted.

This review is organized into the following subtopics: Section 2 introduces the VBDA framework including multi-stage one and end-to-end one. Section 3 describes the multi-stage dietary assessment including vision-based food analysis, food portion estimation and nutrient derivation. The end-to-end approaches are summarized in Section 4. We present various dietary assessment datasets and evaluation metrics in Section 5. Section 6 discusses future development trends in this field. Finally, we conclude this article in Section 7.

2. Vision-based dietary assessment

VBDA refers to taking images of a meal as input and then using computer vision to automatically identify relevant dietary information as the output. We divide it into two architectures, as shown in Fig. 1. In the earlier years, the VBDA was conducted in a step-by-step way, namely multi-stage dietary assessment. In recent years, with the rapid development of deep learning, instead of training a pipeline of models to handle subtasks at different stages, the end-to-end deep learning solutions for dietary assessment are proposed to apply input data to a single network for direct nutrient derivation. Next, we will briefly introduce multi-stage and end-to-end solutions, respectively. At the end of this section, we summarize the representative works of the VBDA.

2.1. Multi-stage architecture

As shown in Fig. 1(a), the multi-stage VBDA generally consists of three stages, including food image analysis, portion estimation, and nutrient derivation. The performance of the first two stages highly depends on the used AI algorithms and the available food datasets, while the last stage depends on the nutritional composition database. Food image analysis. The key in the visual analysis is compact and expressive feature representations that summarize the information of food images. This phase is related to food recognition, detection and segmentation. The first column of Fig. 1(a) shows one example for food recognition, detection and segmentation. Food recognition mainly predicts the type or composition of food items from images (Min, Jiang, Liu, Rui, & Jain, 2019). Food detection aims to locate and classify each food item from the food image, where the localization is realized via estimating the bounding box of each dish. Compared with food recognition, food detection additionally provided the localization for the recognized food item. Food segmentation involves partitioning food images into multiple food items in a pixel-level way. It is the process of assigning a food label to every pixel in one food image. Therefore, it is more precise to localize food items with arbitrary shapes compared with rectangular shapes for food detection. Such more precise food area localization is generally more helpful for the following portion estimation of these food items.
Food portion estimation. The visual features are far from sufficient to estimate calorie content. In order to accurately estimate the nutrient content of the food from the image, it is necessary to identify the food item and estimate the weight. Earlier methods conduct rough food weight estimation based on food area (He, Xu, Khanna, Boushey, & Delp, 2013; Okamoto & Yanai, 2016a). As the technology evolves, calculating the food portion based on volume estimation in combination with food density databases becomes more popular.

Nutrient derivation. The conversion from food to meaningful nutritional information depends on the accuracy of the nutrition fact databases. There are several food databases in the public domain, such as USDA Food and Nutrient Database for Dietary Studies (FNDDS) (U.S. Department of Agriculture, Agricultural Research Service, 2020), Food Databanks National Capability extended dataset (FDNC) (Food Databanks National Capability, 2020), The Canadian Nutrient File (Health Canada, 2021) and Open Food Facts (Open Food Facts, 2020). Take FNDDS as an example, as shown in Table 1. The FNDDS contains the most common foods and beverages consumed in the United States, their nutritional values, and the weights of typical food portions. Nutritional value data for foods are stored in these tables, which are available from national and international health organizations. The information in the nutrition database for each food can be used to calculate the energy of the food consumed. The actual nutrient content $N$ can be calculated as:

$$N = N_F \times \frac{M}{M_T} \quad (1)$$

where the food mass $M$ can be obtained after portion estimation. $N_F$ and $M_T$ represent food nutrition and mass in the table, respectively. Therefore, the nutrient content of each food consumed is obtained by simply subtracting the nutrient content after the meal from the nutrient content before the meal. This means that the energy and nutrients of the food a person eats can be assessed based on the images obtained before and after the food is consumed (Woo et al., 2010).

2.2. End-to-end architecture

Although multi-stage VBDA improves a number of baseline methods for dietary assessment, they have some limitations. First, these methods are required to be defined and optimized individually at each stage, while their accuracy remains a challenge (Ruede et al., 2021). Since the error propagation path leads to the accumulation of errors, multi-stage leads to the accumulation of errors and has an impact on subsequent operations. The end-to-end architecture replaces these different stages with a single neural network, reducing error propagation and joint optimization paths. Second, the multi-stage architecture requires the definition of separate stages as well as inputs and outputs for each stage. It means that potentially helpful information from the early stages cannot be passed on or used later to improve predictions. In contrast, the end-to-end architecture emphasizes a single model instead of a pipeline of multi-stage methods. Only the original inputs and the
For dietary assessment, the former starts with relevant features manually extracted from images, and the features are then used to create a classification model to recognize the food class. In contrast, the latter performs end-to-end learning, where a network is given and a classification task will be performed, and it automatically extracts relevant features from food images and learns how to do this automatically.

### 3.1. Food recognition

Food recognition can be regarded as the most basic work in visual analysis, and food item labels can be obtained via food recognition. Existing research methods are mainly divided into manual feature-based recognition methods and deep learning-based recognition ones for dietary assessment. The former starts with relevant features manually extracted from images, and the features are then used to create a classification model to recognize the food class. In contrast, the latter performs end-to-end learning, where a network is given and a classification task will be performed, and it automatically extracts relevant features from food images and learns how to do this automatically.

#### 3.1.1. Food recognition

Food recognition is the key to food image recognition. Manual features vary from simple features, such as color, texture, and shape, to other complex features. Compared with features such as color and texture, the Scale Invariant Feature Transform (SIFT) descriptor has the advantages of scale, orientation and affine distortion invariance, which can effectively deal with the deformation characteristics of food images. As a result, it is widely used in food image recognition tasks.

W. Wang et al.
trend to improve the recognition performance. Puri et al. (2009) adopted manual features, more methods integrate different manual features to achieve acceptable recognition. Considering the complementarity between different knowledge of food types and food scales, calories can be estimated via multi-view food classifier in a health perception system. It calculated per image as their final representation for food recognition and food mass estimation. It is worth mentioning that the method was limited to the surface and background of the food images and can only predict the calorie content of a single food. Hand-engineered features are time-consuming, brittle, and not scalable in practice, especially when it comes to unstructured data such as food images and videos. In contrast, deep learning has become the mainstream method for automatically recognizing food types and maximum response features to achieve acceptable recognition rates on a large number of food types. Chen et al. (2012) combined SIFT, local binary patterns, color, Gabor filter descriptors for food recognition. They separately trained a Support Vector Machine (SVM) classifier for each type of features and fused them using the AdaBoost algorithm for classification. Chokr and Elbassuoni (2017) used the Mathworks Image Processing Toolbox to extract raw features from food images. Then they applied InfoGain, selected the top 6000 features followed by principal component analysis, and ended up with 23 features per image as their final representation for food recognition.

| Table 2 | Food image analysis approaches in multi-stage VBDA. |
|---------|-----------------------------------------------|
| References | Dataset# images (class) | Methods | Method evaluation | Nutrient estimation |
| **Recognition** | | | | |
| Wu and Yang (2009) | 100(-10) | SIFT feature | 52% | – |
| Kong and Tan (2012) | (-) | SIFT feature and nearest neighbor classifier | ~92% | – |
| Puri et al. (2009) | (-150) | Color neighborhood, maximum response features and SVM classifier | ~90% | 5.75±3.75% on volume estimation |
| Chen et al. (2012) | 100(50) | Multi-feature fusion, SVM classifier | 68.3% | – |
| Chokr and Elbassuoni (2017) | ~1000(6) | Random forest and SVM classifier | 99.1% | MAE 0.0993 on calorie estimation |
| Christodoulidis, Anthimopoulos, and Mougiasakalou (2015) | 246(573) | Own designed CNN | 84.9% | – |
| Mezgeb and Koroušić Seljak (2017) | 225,953(520) | Modification of the AlexNet | 86.72% | – |
| Ayon, Mashrafi, Yousuf, Hossain, and Hossain (2021) | 23,000(23) | InceptionV3 | 89.48% | – |
| Ma et al. (2021) | 10,047(100) | InceptionV3 | 78.26% | R² 0.73 on protein estimation |
| **Detection** | | | | |
| Gocca, Napolitano, and Schettini (2015) | 2000(15) | Circle Hough transform | – | – |
| Pouladzadeh and Shirmohammadi (2017) | 7000(30) | Selective search | 93.05% | Precision |
| Jiang, Qiu, Liu, Huang, and Lin (2020) | 12,740(100) | Faster R-CNN | 17.5% Map | – |
| Mao, He, Shao, Yarlagadda, and Zhu (2021) | 14,991(82) | Faster R-CNN | 0.7064 F1 | – |
| Okamoto and Yanai (2016b) | – | Background subtraction and the hough transform | – | – |
| Qiu, Lo, and Lo (2019) | 4200(14) | Mask R-CNN | – | – |
| Lei, Qiu, Lo, and Lo (2021) | 14(1) | Mask R-CNN | – | – |
| **Segmentation** | | | | |
| Zhu et al. (2008) | (-) | Threshold segmentation | – | – |
| Mariapann et al. (2009) | 50(~3) | Threshold segmentation | – | – |
| Eskin and Mikhailidis (2012) | 676(49) | Threshold segmentation with the morphological method | – | – |
| Zhu, Bosch, Bouhery, and Lo (2010) | 63(3) | Normalized cut segmentation | – | Percentage of misreported nutrient information: 1% ~10% on weight estimation |
| He et al. (2013) | 1453(96) | Local variation segmentation | – | – |
| Zhu, Bosch, Khanna, Bouhery, and Delp (2015) | (-63) | Multiple hypothesis segmentation | – | – |
| Pouladzadeh, Shirmohammadi and Yassine (2014) | (-15) | Graph cut segmentation | ~90% PA | – |
| Anthimopoulos et al. (2015) | 1620(248) | Region growing segmentation | 88.5% PA | 10% MAE on CHO estimation |
| Sudo, Muranaka, Shimamura, and Taniguchi (2014) | 2500(29) | Super-pixel segmentation | – | 31.8% average error on calorie estimation |
| Myers et al. (2015) | 12,625(201) | Finetuned DeepLab | 25% IoU | – |
| Pouladzadeh, Kuhad, Peddi, Yassine and Shirmohammadi (2016) | (-30) | Graph cut segmentation | – | – |
| Okamoto and Yanai (2016a) | ~1M(1K) | GrabCut segmentation | – | 0.213 relative average error on calorie estimation |
| Ege, Shimoda, and Yanai (2019) | 1000(6) | U-Net | 84.1% mIoU | – |
| Ando, Ege, Cho, and Yanai (2019) | 5301(1) | U-Net | 80% mIoU | – |
| Dehais, Anthimopoulos, and Mougiasakalou (2016) | 821(-) | CNN based border map generation, seeded region growing segmentation | 88% PA | – |
| Dehais, Anthimopoulos, and Mougiasakalou (2015) | (-) | Seeded region growing segmentation | 80% Fmax | – |
| Lu et al. (2020) | 21,807(80) | Mask-RCNN | 83.9% Fmax | Pearson correlations 0.69 on protein estimation |
| Lu et al. (2021) | 1281(521) | Pretrained dilated ResNet50 | 83.5% PA | R² 0.923 on calorie estimation |
discovering the discriminative feature representation from raw data. Recent studies have demonstrated that visual features learned by deep learning are more expressive and robust than handcrafted ones (Cai et al., 2020; Kamilaris & Prenafeta-Boldú, 2018). For this reason, recently proposed VBDA methods adopt deep-learning networks for food vision analysis.

Deep learning based recognition methods. Deep learning, as an exclusively superior framework of contemporary machine learning, was extensively explored and brought a new stage for food image recognition (Zhou, Zhang, Liu, Qiu, & He, 2019). The key element of deep learning lies in CNN and its derivative algorithms. Christodoulidis et al. (2015) used a six-layer CNN to classify food images. Image chunking was used to extract a set of overlapping square blocks. The voting determined the type of food in the area. In contrast, Mezgec and Koroulič Seljak (2017) proposed NutriNet for food and beverage recognition based on AlexNet. The input image was expanded from 256 × 256 to 512 × 512 pixels, and adding a convolutional layer to obtain hidden features in higher resolution images. The model has been used in practice as one part of a mobile application for dietary assessment of Parkinson’s disease patients. However, its limitation was that for images with multiple food items, not every item could be accurately identified. Ayon et al. (2021) proposed a food detection system based on multi-label recognition by setting a threshold. They also deployed trained models into the webpage to improve the user experience. Ma et al. (2021) trained their model by fine-tuning four deep architectures that were pre-trained on the ImageNet dataset. They modified the last fully connected layer of the architecture to output 100 classes. Because the food category has been defined and has a fixed nutritional configuration (e.g., calories, carbohydrates and proteins), the nutrient estimation mainly relies on the classification results. Besides, they carried out almost the same experiments on their another dataset ChinaMartFood-109 (Ma, Lau, Yu, Li, & Sheng, 2022). Unlike studies that summarized food recognition methods, which only recognized food items in their work, we focused on summarizing approaches that directly employed food recognition for the nutrition assessment task. For more works only for food image recognition, please refer to some surveys, such as Min et al. (2019).

3.1.2. Food detection

In dietary assessment, multi-food localization and recognition are the primary detection technique application. Ciocca et al. (2015) proposed an automatic dietary monitoring system for canteen customers based on multi-food detection. They added a plate localization method to reduce the area to be recognized. Several descriptors are evaluated, such as color-based, statistical, spatial-frequency or spectral, structural, and hybrid. Due to the weekly fixed menus and the control conditions of food image acquisition in the canteen, the canteen scenario is relatively simple. The cafeteria staff provides a fixed amount of food according to the nutrition table, and the system can evaluate the leftovers according to the degree of variability in adjacent regions. Zhu, Bosch, Boushey et al. (2010) applied the active contour model to segment the plate and food in succession. Different from these works, Zhu, Bosch, Boushey et al. (2010) applied the active contour model to segment food images. The model can deform an initial curve to the boundary of the object under some constraints from the image. They discussed the limitation of active contours and chose the normalized cut as their segmentation method because it can measure both the total dissimilarity between different groups and the total similarity within the groups. Normalized cut is a kind of graph-theoretic segmentation method that treats one pixel from the image as one node of a graph and considers segmentation a graph partitioning problem. There are many other graph-theoretic based segmentation methods. For example, He et al. (2013) used the image-based local variation method to segment food according to the degree of variability in adjacent regions. Zhu et al. (2015) proposed multiple segmentation hypotheses and adopted normalized cut to dynamically select the number of segments guided
by the classification results. The similarity between these two works lies in the iterative use of the classification feedback to refine the segmentation results. A series of works by Pouladzadeh et al. used the graph cut in the task of food calorie estimation. Pouladzadeh, Shirmohammadi and Yassine (2014) introduced a new segmentation approach that combined the graph cut with the texture segmentation method. It was widely used in their later works (Pouladzadeh, Shirmohammadi and Almaghbrabi, 2014). These works combined the segmentation approach with SVM to analyze food images. GoCARB (Anthimopoulos et al., 2015) is a mobile daily carbohydrate assessment system for patients with type 1 diabetes. The food was segmented using a region-growing algorithm after detecting the localized plate regions. It should be noted that the system has certain assumptions, including the use of flat-bottomed elliptical dishes, and food cannot be overlapped or blocked. It can also be extended to display other macronutrients, advancing the development of a healthy human diet.

However, all of the above are unsupervised segmentation methods. In contrast, in another study (Sudo et al., 2014), semantic segmentation completes image segmentation and labeling. Label histogram shows the frequency of occurrence of food ingredient labels, based on which regression analysis was performed, and nutritional data was estimated. However, in order to train the semantic segmentation model, a large amount of recipe data and corresponding food images and ingredient labels need to be collected.

Afterward, the methods with deep learning have become more and more popular. Im2Calories (Myers et al., 2015) was the first work to achieve semantic segmentation specifically for food images using deep CNN. They combined the segmentation network with a deep prediction network and converted voxels to estimate the food volume. However, their evaluation was limited to plastic food replicas for training nutritionists. Pouladzadeh, Kuhad, Peddi et al. (2016) took full advantage of the graph cut segmentation and CNNs for calorie measurement. However, they could only be applied to images with a single food label. Okamoto and Yanai (2016a) applied GrabCut with bounding boxes as a segmentation seed, and used a CNN-based recognition engine to classify them. In addition, food segmentation methods based on region growing and merging algorithms were subsequently improved via boundary detection networks. DepthCalorieCam (Ando et al., 2019) employed food segmentation based on U-Net (Ronneberger, Fischer, & Brox, 2015), the accuracy of segmentation was much higher than (Okamoto & Yanai, 2016a). Similarly, Ege et al. (2019) also used U-Net for segmentation. The difference is that the it estimated food calories based on the actual size of the food area calculated, while the DepthCalorieCam used a depth camera to obtain the food volume information and applied the estimated food volume to a regression equation of the calorie. Dehais et al. (2016) combined CNN-based border detection with an optimized region growing and merging framework to achieve high average accuracy for dish segmentation. It was improved from the previous work (Dehais et al., 2015), and yet the computational resources were costly. Following the GoCARB system, the team developed goFOOD™ (Lu et al., 2020). For semi-automatic segmentation, region growing and merging algorithms were still used. They also developed an automatic food segmentation method based on the Mask-RCNN (He et al., 2017). For the recognition module, an improved Inception V3 was used to achieve hierarchical food recognition. GoCARB focused on calculating carbohydrates, while goFOOD™ can be used to estimate the calories and nutrients in a meal. These methods are easy to implement. However, they are limited by the complexity of the food scene (e.g., background, lighting, inter-class variability of food). Recently, they were committed to studying the nutritional intake of hospitalized patients (Lu, Stathopoulou, Vasiloglou et al., 2021). Through the development of a series of novel AI algorithms, the goal of achieving high performance on a small amount of training data has been achieved. Lu, Stathopoulou, Vasiloglou et al. (2021) proposed the multi-task contextual network to simultaneously segmented plates and foods. In addition, the application of the contextual layer enhanced the relationship between food and plate types and thus resulted in improved segmentation accuracy. Due to the finite nature of food samples, they trained one few-shot learning-based classifier for food recognition in the meta-learning framework (Snell, Swersky, & Zemel, 2017). The core idea was to learn the transferred knowledge in many similar tasks and use it for new tasks. It was a fully automated pipeline system based on AI, which promoted further development in nutrient assessment. In addition to these works, Aslan, Ciocca, Mazzini, and Schettini (2020) benchmarked many deep learning-based food segmentation methods on their proposed dataset Food50Seg. Food/no-food segmentation and semantic segmentation methods were benchmarked from many different perspectives, which will guide the design and improvement of food analysis systems.

3.2. Food volume estimation

A common approach to estimate food quality is to evaluate its volume, which can be used in conjunction with a food density database to gauge the quality of food in the image. Therefore, this section mainly focuses on volume estimation. Area and depth estimation based on top and side views can be directly used to estimate the food volume (Pouladzadeh, Shirmohammadi and Almaghbrabi, 2014). However, this method requires the user’s thumb to be a reference object and is not accurate. Therefore, the 3D reconstruction methods are commonly used to estimate the volume. It relies on computer vision techniques to add missing dimensions from 2D images to create a 3D space, using visual cues to understand the third dimension of objects.

3D reconstruction starts with the calibration of camera parameters to perform volume reconstruction. Most methods use fiducial markers to facilitate the camera calibration process. Traditionally, spherical objects (Zhang, Wong, & Zhang, 2007) and the checkerboard patterns (Yu & Peng, 2006) are used as fiducial markers. 3D reconstruction includes methods that use a single view or multiple views as the input. Reconstruction based on a single view is an ill-posed problem, but the geometric model is conducive to recover the 3D parameters of the food in the scene. Zhu et al. (2010) used both the spherical approximation model and prismatic approximation model for 3D volume reconstruction. Extracting feature points from food images based on segmentation, constructing base surfaces for 3D volume estimation. Chen et al. (2013) constructed a 3D geometric shape model library, including regular and irregular models. The food contours are matched to the 3D shape model, and the food volume is estimated according to the size of the model. However, the shape model library requires heavy user input, and the results are unstable. There are some studies, which utilize the prior knowledge of “container shape” as the geometric context information to obtain food volume (Fang, Liu, Zhu, Delp, & Boushey, 2015; Xu et al., 2013). In comparison, the method proposed by Gao, Lo, and Lo (2018) does not require prior knowledge such as shape models. They adopted the simultaneous localization and mapping system based on the monocular vision for dynamic food volume measurement. The optimized multi-convex hull algorithm was applied to the sparse map to form a 3D mesh object and continuously measured the food volume. Based on multiple views, sparse reconstruction or dense reconstruction can be performed. The 3D models reconstructed using sparse feature points have lower resolution (Kong & Tan, 2012). In dense reconstruction, the stereo matching method is most commonly used. Puri et al. (2009) composed two food images into a stereo pair. According to the hierarchical pyramid matching scheme, the left and right images were matched to perform the 3D reconstruction. Based on the reconstructed 3D point cloud, the scale was further estimated, and the 3D volume of the food was calculated. This method takes full advantage of dense reconstruction. However, the whole process is labor-intensive. Rahman et al. (2012) proposed a volume estimation method based on a set of images taken from the left and right sides of the food. They converted the reconstructed 3D point cloud into a series of slices and added the volume of each slice to obtain the total volume of the
food. Although many attempts on volume estimation were made, most of them relied on user input as well as strong assumptions. Dehais, Anthimopoulos, Shevchik, and Mougiakakou (2017) proposed a dense two-view reconstruction method with a reference card. The method made minimal assumptions about the type or shape of the food and was highly automated. They added disparity range extraction to the stereo matching process. A disparity map is a depth map where the depth information is derived from offset images of the same scene. Once the disparity map is estimated, the disparities can be converted to depth according to the following:

$$\text{depth} = \frac{B \cdot f}{\text{disparity}}$$

(2)

where $B$ is the length of the baseline connecting the camera centers and $f$ is the focal length of the camera. Using the depth map and camera parameters, the 3D positions of all the points in the food image can be restored. In addition to the depth information from stereo-matched images, depth information can be easily obtained using a depth camera. Chen et al. (2012) introduced a quality estimation method using the depth information. Take the “sour and spicy soup” as an example. They used a depth camera Kinect to obtain the color and depth information of the noodle soup. The area of the bowl and the depth of the food is used for quantity estimation. However, the Kinect is a depth camera of structured infrared light, which affects the depth estimation due to the light reflection. Myers et al. (2015) adopted depth camera-based 3D volume estimation. The depth map was converted to a voxel representation using CNN and random sample consensus (RANSAC) reconstruction. With the development of technology, stereo images can be quickly acquired through devices with stereo cameras such as 3D smartphones, and thus becomes another type of popular methods.

Later, researchers have been exploring more convenient and accurate ways to estimate the volume. For example, Yang, Jia, Bucher, Zhang, and Sun (2019) argued that carrying fiducial markers was very inconvenient and affected subsequent image processing. Therefore, they proposed a novel virtual reality-based volume estimation method. The key idea of eliminating fiducial markers was to determine the camera orientation through the motion sensor inside the smartphone and determine the location of visible points based on the width or length of the phone. Lo, Sun, Qiu, and Lo (2020b) combined the strength of deep learning and 3D reconstruction and developed a point completion network to solve the problems of scale ambiguity and food occlusion.

As one typical multi-stage VBDA work in Fig. 3, in the image analysis phase, Lu, Stathopoulou, Vasiloglou et al. (2021) performed hyper food semantic segmentation and fine-grained food recognition. The next stage includes volume and weight estimation, as well as nutritional calculation. They used the Intel RealSense RGB-D Sensor to capture food images before and after a meal, which can output aligned RGB images and depth images simultaneously. Food consumption depends on the volume of food before and after a meal. To obtain the plate surface, the segmentation map allows to determine the plate’s position and then enables the RANSAC algorithm to estimate the orientation of the plate. In addition, the acquired food depth image was converted into a point cloud. The triangular surfaces were divided by the Delaunay triangulation method for the construction of food 3D surfaces. They combined the plate surface estimation with 3D food surface extraction to further calculate the food volume. To calculate nutrition, they converted the volume to weight. According to the calculated weight and food category, the nutrient content of each food before and after meals was calculated. It can be seen that each stage has its specific task and the multi-stage architecture relies on the input and output of each stage.

### 4. End-to-end approaches

The end-to-end approach focuses on using a single model to replace the pipeline of the multi-stage method. Considering that end-to-end methods have shown promising prospects in tasks such as image recognition (He, Zhang, Ren, & Sun, 2016), more and more end-to-end works for nutrition prediction are proposed. Fang et al. (2019) proposed one novel end-to-end food energy estimation method, which used the generative adversarial network to estimate the image-to-energy mapping. The energy distribution map (Fang et al., 2018) is a new method that replaces the “depth map” in Myers et al. (2015) for visualizing the position of food in the images and how much relative energy was presented in different food regions. He et al. (2021) improved this work (Fang et al., 2018) by adding a localization network to detect food items in the image. Bounding boxes were applied to the energy distribution map to reduce the estimation errors.

However, all of these works only achieved an estimate of the calories in the food. In dietary assessment, multiple nutrient estimates (e.g., different types of macronutrients) are often required. There are also associated representations among other VBDA subtasks. Focusing on a single task tends to neglect information that might help do better on the metrics we care about (Ruder, 2017). By realizing the shared representation between related tasks, the method of learning multiple related tasks together is called Multi-Task Learning (MTL). Compared with a single task, it can improve the generalization performance of the task to realize a more comprehensive VBDA architecture. MTL is typically implemented with either soft or hard parameter sharing of hidden layers.

In soft parameter sharing, each task has its model with its parameters. The distance between the parameters of the model is then regularized to encourage the parameters to be similar. He et al. (2020) performed food classification and portion size estimation in a multi-task framework with shared soft parameters. They used the L2-norm to regularize the parameters of the two models. Finally, the output of both models is concatenated and then normalized to regress the calorie estimation result.

Hard parameter sharing is generally applied by sharing the hidden layers among all tasks while reserving several task-specific output layers. Ege and Yanai used this architecture in a series of end-to-end VBDA works. Ege and Yanai (2017) designed a multi-task CNN based on VGG-16 to extract features. After that, a Fully-Connected (FC) layer is shared by all tasks. Another FC layer is branched to each task for the output unit, including predicting calories and food categories, ingredients, and cooking instructions. However, they assumed that food images only contain one dish. Their new work further used the improved YOLOv2 (Redmon, Divvala, Girshick, & Farhadi, 2016) to apply object detection to multiple dishes (Ege & Yanai, 2018). The network can estimate the boundaries, categories, and calories of dishes from images of multiple dishes. However, the output of this method
does not consider food amount and only corresponds to the calories of a dish. Considering there are no datasets on multiple-dish food images with both bounding boxes and food calories, two types of datasets were used alternately in Ege and Yanai (2019) to train a single CNN. The results showed that their multi-task approach achieved higher accuracy, higher speed, and smaller network size than a sequential model for food detection and calorie estimation. Similarly, Situju et al. (2019) collected two image datasets at different scales for calories and salinity assessment. They used the Inception (Chollet, 2017) model to extract features. The network branches to each task from the global average-pooling layer. Each task has a global average pooling layer, a FC layer with the dropout, and an output layer, respectively. Two-stage fine-tuning was applied to improve the accuracy. The results proved that the multi-task CNN outperformed the single-task CNN. Ruede et al. (2021) designed the network from a different perspective. They introduced a framework for retrieving nutritional information of recipes by matching ingredients and their mass to a nutrient database using phrase embedding. Pre-trained ResNet and DenseNet (Huang, Liu, van der Maaten, & Weinberger, 2017) architectures were leveraged as the backbone models. The last FC layer was replaced with a linear layer to regress the macronutrient information. A good distinction is made between high and low-calorie dishes.

The idea of multi-tasking learning had also been used to design some complex end-to-end VBDA networks. Lu, Allegra, Anthimopoulos, Stancu, Farinella, and Mougiakakou (2018) presented a MTL-based VBDA network, which simultaneously implemented food recognition, segmentation, and volume estimation. The feature extraction module is composed of ResNet50 and feature pyramid network. A Region Proposal Net (RPN) (Ren et al., 2017) along with the feature extraction module preliminary produced the bounding box for each candidate object. MTL is applied to recognize food, regress bounding-box, and predict binary mask from the output of RPN. In terms of accuracy, the performance of food segmentation and volume estimation outperformed the state-of-the-art at that time. Lu, Stathopoulou and Mougiakakou (2021) focused on reconstructing the 3D food model from a single RGB-image input without the use of depth information, and for the first time realized the entire pipeline single view diet assessment. They utilized the ResNet50 as the feature extraction module. The encoded features were decoded in several different branches to predict depth maps, semantic maps, and the camera pose between the consecutive frames.

Thames et al. (2021) carried out a series of VBDA experiments on their proposed Nutrition5k dataset, which covered the typical end-to-end works. The experiments were based on a MTL end-to-end network in Fig. 4. The Inception V2 backbone encoder was used to extract feature maps. Then average-pooling was applied, and two FC layers followed. They trained a separate multi-task head for each regression task (calorie, macronutrients, and optionally mass), including two FC layers of different sizes. The experiments included estimating nutrients per gram from RGB image, estimating nutrients directly from RGB image, estimating nutrients directly from RGB-D image, and estimating nutrients from RGB image and portion. The results showed that the complexity of VBDA substantially increased when portion estimation is required, and depth information is essential for estimating nutrients when considering food portion size. Besides, the network outperformed professional nutritionists at caloric and macronutrient estimation in the generic food setting.

5. Datasets and evaluation metrics

5.1. Datasets

In this section, we briefly summarize and compare various datasets for dietary assessment. Table 3 details the major nutrition-related datasets from different aspects, such as classes, total size of images, volume, calorie type and macronutrients (i.e., carbohydrate, protein and fat).

Most of the datasets related to dietary assessment consist of specific types of foods. For example, Menu-Match (Beijbom et al., 2015) focuses on the restaurant scenarios, and contains food images and associated nutritional information. FoodLog (Miyazaki et al., 2011) contains user-uploaded food images, text, and calorie information. It is mainly composed of Japanese food. Chen et al. (2012) collected 50 kinds of Chinese food for food identification and quantity estimation. However, only food recognition and quantity estimation was performed and no nutritional information was labeled. Later, Ma et al. (2021) built the first Chinese food image dataset ChinaFood-100 with comprehensive descriptions and nutritional information annotations (i.e., protein, fiber, vitamin C, calcium and iron). Recently, Nutrition5k (Thames et al., 2021) is released for the development of dietary assessment. It consists of food dishes with corresponding video streams, ingredient weights, depth images and high-precision nutrition annotations (i.e., carbohydrate, protein and fat).

As we all know, the depth image of food is beneficial for volume estimation. Lu, Stathopoulou, Vasiloglou et al. (2021) used the Intel RealSense RGB-D sensor to acquire both RGB images and depth images of food, and constructed the NIAD dataset. It should be emphasized that there are few datasets for food volume estimation. The annotations needed in such datasets are pixel-level food segmentation maps, image-level labels, and basic facts about food volume. MADI-Ma (Allegra et al., 2017) is a dataset that meets these requirements. It contains RGB images and depth maps of meals, segmentation and recognition...
maps, 3D models, weights and volumes. Therefore, many studies have conducted experiments on this dataset to provide baseline performance of food segmentation, depth, and volume estimation. For example, Lu et al. (2020) evaluated the performance of the entire process of dietary assessment based on this MADiMa dataset. Considering that MADiMa only contains meal images taken by a monocular camera, they created one fast food dataset from McDonald’s to support two-view images and stereo image pairs as input to comprehensively evaluate the performance of goFOOD™. Lu et al. (2018) also used the MADiMa dataset as the training and evaluation dataset. Lu, Stathopoulou and Mougiakakou (2021) used it to evaluate the performance of the proposed method. Furthermore, the Canteen dataset (Lu, Stathopoulou and Mougiakakou, 2021) was proposed as a real scenario-based evaluation compared with MADiMa, which was created in a laboratory environment.

In addition, many studies have proposed the dataset of eating occasion images for food energy datasets (Fang et al., 2019; He et al., 2021, 2020). The distribution of food energy in the eating scene can be represented as an energy distribution image. The purpose is to estimate the energy values of the images based on the estimated energy distribution. Such datasets usually contain annotations of food categories, bounding box information, and the food size. To verify the effectiveness of multi-task learning, Ege and Yanai (2017) constructed two datasets of Japanese and American with food categories, ingredients and cooking directions. In addition, due to the high cost of manually creating nutrition annotations, they chose to collect data from the web. Many researches were done using this approach. For example, Ruede et al. (2021) collected the dataset from a German recipe website, namely the pic2kcal dataset, which contains cooking instructions, ingredients and images. Situju et al. (2019) collected large datasets of food images with calorie and salinity annotations based on cooking food websites. The proliferation of these recipe-sharing websites has facilitated the collection of food datasets. For example, Ajinomoto Park includes food ingredients, cooking instructions and rich nutritional information. All recipes also contains the above information and adds a user comment module. They are usually provided by experts such as culinary researchers or cooks and thus has excellent credibility.

With the rapid development of technological infrastructure, web-based data collection methods will become more popular.

### 5.2. Evaluation metrics

One dietary assessment system should be evaluated in multiple respects, such as accuracy, run time, and resource requirements. However, most researchers have only focused on metrics for assessment accuracy. For the multi-stage method, the evaluation should be performed in each stage. In contrast, the end-to-end method should be evaluated as a whole, which is similar to the final stage of the multi-stage method. Due to the difference of each stage, evaluation metrics are classified and summarized by the output type.

For food image analysis, Top-1 Accuracy and Top-5 Accuracy are often used to evaluate the performance of recognition. Here, Top-N accuracy means the probability that the correct class of food presents within the Top-N recognition results. Detection procedures will mark out targets with bounding boxes. The area in the bounding box is a set of pixels. Segmentation procedures will output a mask of the input image. For one particular category in the image, its mask is also a set of pixels. As a result, the evaluation metrics of food detection and segmentation are similar, and they are evaluated using the Intersection over Union (IoU):

\[
\text{IoU} = J(A, B) = \frac{|A \cap B|}{|A \cup B|}
\]

where IoU is defined as the area of intersection between the predicted map $A$ and the ground truth map $B$, divided by the area of the union between the two maps, and ranges between 0 and 1. Mean-IoU (mIoU) is further defined as the average IoU over all classes (Minaee et al., 2021). Precision/Recall/F1 score can be defined for each class, as well as at the aggregate level, as follows:

\[
\text{Precision} = \frac{TP}{TP + FP} \quad \text{Recall} = \frac{TP}{TP + FN}
\]

where TP, FP, FN refer to the true positive fraction, the false positive fraction, and the false negative fraction, respectively. The Precision–Recall curve shows the tradeoff between the Precision and Recall values for different thresholds. This curve helps to select the best threshold to maximize both metrics. Compared to graphically deciding the best values for Precision and Recall, using the F1 score to measure their

### Table 3

| Dataset                  | Total # images/class | Data type                                           | References                             |
|--------------------------|----------------------|-----------------------------------------------------|----------------------------------------|
| American recipe dataset  | 2848(21)             | Images with categories and recipe                   | Ege and Yanai (2017)                   |
| Japanese recipe dataset  | 4877(15)             | Images with categories and recipe                   | Shirmohammadi (2015)                  |
| Situju                   | 3051(14)             | Images with calories and salinity values            | Situju et al. (2019)                  |
| Image-to-energy dataset  | -79                  | Images with categories, segmentation masks and calorie values | Fang et al. (2019)                    |
| Eating occasion image to food energy dataset | 834(21) | Images with categories and portion size groundtruth | He et al. (2020)                      |
| Eating occasion image to food energy dataset | 915(31)  | Images with bounding box, categories and calorie values | He et al. (2021)                      |
| pic2kcal                 | 306,000(70,000)      | Images with categories, recipe and nutrient composition | Ruede et al. (2021)                   |
| Nutrition5k              | 20k(5k)              | Multimedia data includes video streams and depth images with component weights, and nutrient composition | Thames and Shirmohammadi (2021)       |
| Canteen dataset          | 16k(80)              | Video frames with annotation                        | Lu, Stathopoulou and Mougiakakou (2021) |
balance is better. It is defined as the harmonic mean of Precision and Recall:

\[ F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \]  

(5)

When the value of F1 is high, this means both the Precision and Recall are high. In addition, the Average Precision (AP) is a way to summarize the Precision–Recall curve into a single value representing the average of all Precisions (Ahmed Fawzy Gad, 2021). It is defined as the weighted sum of Precisions at each threshold where the weight is the increase in Recall. The mean of the APs for all classes is the mAP.

Besides, Pixel Accuracy (PA) is also widely used in evaluating segmentation. Its definition is the ratio of correctly classified pixels divided by the total number.

\[ \text{PA} = \frac{\text{The number of correctly classified pixels}}{\text{The total number of pixels}} \]  

(6)

Considering the output type of portion estimation, nutrient estimation, and end-to-end methods is numeric, Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) are usually used in the evaluation.

\[ \text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |\hat{w}_i - \tilde{w}_i| \]  

(7)

\[ \text{MAPE} = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{\hat{w}_i - \tilde{w}_i}{\tilde{w}_i} \right| \]  

(8)

where \( \hat{w}_i \) is the estimated portion size of the \( i \)th image, \( \tilde{w}_i \) is the groundtruth value of \( i \)th image and \( N \) is the number of testing images (Shao et al., 2021).

6. Future works

According to comprehensive discussions on existing efforts, we now articulate key open challenges and give some prospective analysis on future works for VBDA.

6.1. Large-scale benchmark dataset construction

Currently, VBDA works are driven mainly by data and thus highly rely on the construction of datasets. However, most works tend to construct their own datasets for specific tasks rather than reusing existing datasets. Besides, accurate portion and nutrition annotation are extremely burdensome and costly to obtain, which limits the scale of VBDA datasets. These reasons lead to the lack of large-scale VBDA benchmarks. There are two representative datasets, namely MADiMa (Allegra et al., 2017) and recently released Nutrition5k (Thames et al., 2021). The MADiMa dataset contains food images, depth maps, segmentation maps, weight/volume measurements of served meals, nutrient content together with the corresponding annotations, labels and accelerometer data. It is constructed for multi-stage VBDA. Segmentation map plays a vital role in multi-stage VBDA. However, it is not necessary for an end-to-end approach. Nutrition5k (Thames et al., 2021) thus drops the segmentation maps and adds food ingredient information for an end-to-end VBDA. Ingredient information gives a new perspective to understand nutrients and enables the model to handle unseen food categories. However, these two datasets only contain western food with small scale. The standard benchmark dataset should contain not only western cuisines, but also eastern ones, and cover a wide range of food types from all over the world. Different cuisines differ in many aspects, such as the level of separation of ingredients, and cooking procedures. The ingredients from western cuisines are generally separate while ones from eastern ones are often mixed. The cooking procedure in the eastern cuisine is more various than western ones. Such differences make the VBDA solutions probably different.

And for each item, there should be corresponding images, depth map, weight, nutrition information, recipe, and ingredient distribution map.

Therefore, the VBDA dataset with various cuisine types should be constructed to enable targeted solutions for different cuisines, resulting in more widespread use and popularization of VBDA. Because of the data-driven nature of deep learning, a large-scale VBDA dataset is also urgently needed for developing advanced VBDA algorithms.

6.2. Fine-grained visual analysis

Compared with ordinary objects, many kinds of food do not display unique spatial configurations. They are generally non-rigid, and food appearance can be diverse depending on the used ingredients and the cooking styles. Additionally, hidden ingredients and distinctive cooking methods (e.g., stir-fry, steaming, boiling, roasting) present obstacles to assessing the nutrient composition of different dishes. In addition, external factors such as illumination are also a pressing issue in image analysis. In order to solve these problems, we should implement more fine-grained visual analysis and present three possible future directions in detail. First, more complex and advanced CNNs and Transformers have demonstrated powerful capabilities in a range of visually related tasks (Khan et al., 2021). Similarly, we believe that a new network structure designed for VBDA allows for better extraction of visual features and fine-grained visual analysis. Second, the performance of visual analysis can be improved using pre-trained models on the large-scale food analysis dataset (Min et al., 2020, 2021). For example, Min et al. (2021) verified the better generalization of their proposed large-scale dataset Food2K in various vision tasks, such as food recognition, detection, and segmentation. We can study transfer learning from more aspects in the future, such as cross-cuisine, cross-scenario, and cross super-class. Transfer learning expands the training data in a sense which makes the model more robust and enhances the generalization ability. Third, besides improving the performance of neural networks, utilizing food recipes can also help realize fine-grained visual analysis. For example, Fontanellaz, Christodoulidis, and Mougiakakou (2019) suggested that multimodal recipe retrieval could be the first step in a pipeline for nutrient assessment. Understanding the food composition is more scalable than identifying food categories. The ingredient information and cooking methods in food recipes can establish the nutritional connection among relative foods and help the model understand different visual forms of the same ingredient.

6.3. Accurate food volume estimation

Volume estimation is a commonly used method of portion estimation. However, the performance of current volume estimation portion is not yet satisfactory. Existing studies mainly focused on model-based or stereo-based methods. The first food volume estimation system constructed 3D geometric shape model libraries (Chen et al., 2013). They matched 2D food contours to the 3D shape model, and estimated the food volume based on the size of the model. However, the shape model library is not applicable to all food objects. To solve the problem, multi-view based approaches are developed to perceive depth according to more spatial information of food geometry (Dehais et al., 2017; Kong & Tan, 2012; Puri et al., 2009). However, it is inconvenient for users to take multiple food images from different angles. Although food volume estimation methods based on a single view are generally preferred, such methods cannot provide complete information about the 3D shape of the food (e.g., the back of the food is not visible). Later, the depth information of food images is explored to facilitate the food localization and 3D reconstruction. However, the depth images estimated based on food images are not accurate enough. Therefore, Lu, Stathopoulos, Vasiloglou et al. (2021) simultaneously acquired RGB and depth images of food using the RealSense RGB-D sensor. They converted the acquired depth images into the point cloud to further calculate the food volume. If more mobile devices with depth images acquisition sensors could be developed and popularized, it would simplify the volume estimation to a great extent. Additionally, Chen, Wang, Chen, and Zeng (2021)
found that the spatial structure information in the scene plays a crucial role in depth perception. They proposed 52R-DepthNet to remove the influence on the texture information and utilized the human visual system features for spatial structure grasping to enable the network the powerful generalization capability for depth estimation. This is because the human visual system can perceive depth through stereo vision of both eyes. Based on the accumulation of many natural scenes, humans can judge the distance between objects in the image when there is only one image. Therefore, integrating human visual observation features of structural information into volume estimation is probably one future direction.

6.4. Practical VBDA solution

There has been considerable effort to achieve fully automated, highly accurate nutrient assessments in the academic field. However, the applications available to end-users still mainly rely on the manual recording of food types and amounts, indicating a dichotomy between the academic and the industrial fields. There is no doubt that the accuracy of visual analysis and volume estimation affects the accuracy of the results. Besides, the scenario provides additional context relations and can further improve the assessment accuracy. In a scenario-specific dietary assessment, making full use of the knowledge of the scenario is the key to improve usability. For example, in the canteen environment, the type of food is fixed, specific containers will be used to hold specific types of food. In addition, specialized devices such as depth cameras can be deployed to gather more information. Thimes et al. (2021) proved the depth information improved the accuracy of nutrient assessment in the control experiment. However, it is not practical to require users to acquire such equipment as depth cameras. Considering smartphones are still the most widely used portable device in daily dietary assessments. We can also obtain rich multimodal information from the smartphone’s camera, positioning system, and inertial measurement sensors. Therefore, designing solutions specifically for portable devices, such as smartphones, is a hot topic for research. However, the solutions are confined to a laboratory environment and have many limitations in practice, such as the scarcity of analyzable food types and the lack of large-scale opening testing. In addition, complex eastern cuisines are more difficult so that existing VBDA methods cannot obtain satisfactory assessment performance. Therefore, toward realizing a practical VBDA solution, there is still a long way to go.

6.5. Intelligent dietary management

People generally understand the connection between diet and health. However, the concept cannot be applied to practice, and it is difficult for them to reasonably control their dietary intake in daily life. For example, excessive sodium intake over a long period can raise blood pressure and become a risk factor for hypertension. Therefore, dietary management can help people solve potential dietary problems and achieve personalized health needs. Intelligent food record and food recommendation are two critical aspects of intelligent dietary management.

For the intelligent food record, FoodLog (Kitamura, Yamasaki, & Aizawa, 2008) has been created to process images taken by users. We expect to conduct statistics on the practical and diversified food data spontaneously generated by users. Food nutrition information is obtained and recorded through VBDA, generating a weekly or monthly diet record sheet for users. By analyzing the diet record sheet, we can decide that the intake of a certain type of food should be reduced or increased, thus assisting users in dietary management. For the food recommendation, intelligent food record can obtain the user’s real-time eating habit information, and then learn the dynamics of user preferences based on VBDA (Delarue & Loescher, 2004), and finally combine user’s health information with the learned user preference to realize personalized dietary recommendations (Min, Jiang and Jain, 2020). Furthermore, VBDA-based recommendations can be used as an auxiliary tool for nutritionists. For example, with the popularity of the healthy and green living concept, more and more people prefer healthy weight loss. It is undeniable that healthy weight loss must not only ensure a balanced diet, but also meet the nutritional needs of the body and exercise. Based on the dietary recommendations provided by VBDA, nutritionists with rich professional knowledge further provide reasonable nutritional requirements for the body and exercise. In summary, VBDA has obvious potential and advantages in automatically monitoring diet and compliance with nutritional interventions for intelligent dietary management.

7. Conclusions

VBDA is currently a hot and challenging research field, which has gradually replaced traditional dietary assessment methods. In this review, we divide existing VBDA methods into two types of architectures. One is a multi-stage VBDA architecture, which mainly consists of three parts: food image analysis, portion estimation, and nutrient derivation. Each stage has its own specific task and is linked to each other for nourishment. The end-to-end VBDA architecture emphasizes specifying the original input and nutritional output without multiple steps. The food datasets and evaluation metrics related to VBDA have also been clarified in the article. Although VBDA has made good progress, it has great potential to further enhance the practicality of VBDA. For example, it is urgent to build a large-scale benchmark dataset to enable the faster development of VBDA. How to achieve fine-grained visual analysis and accurate volume estimation based on food characteristics are worthy of further exploration. The promising directions for VBDA are discussed in this paper, and we look forward to seeing a wide range of VBDA applications in the future.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Scientific Research Program of Beijing Municipal Education Commission (grant No. KZ202110011017), National Natural Science Foundation of China (grant No. 61877002, 61923738, U1936203, U19B2040) and Open Research Fund of Beijing Key Laboratory of Big Data Technology for Food Safety (grant No. BTBD-2020KF04).

References

Ahmed Fawzy Gad (2021). Evaluating object detection models using mean average precision (mAP). https://blog.paperspace.com/mean-average-precision. Accessed January 13, 2022.

Allegra, D., Anthimopoulos, M., Dehais, J., Lu, Y., Stanko, F., Farinella, G. M., et al. (2017). A multimedia database for automatic meal assessment systems. In International conference on image analysis and processing (pp. 471–478). Springer, http://dx.doi.org/10.1007/978-3-319-70742-6_46.

Ando, Y., Ege, T., Cho, J., & Yano, K. (2019). Depthcaloriecam: A mobile application for volume-based food calorie estimation using depth cameras. In Proceedings of the 5th international workshop on multimedia assisted dietary management (pp. 76–81). http://dx.doi.org/10.1145/3347448.3357172.

Anthimopoulos, M., Dehais, J., Shevchik, S., Ransford, B. H., Duke, D., Diem, P., et al. (2015). Computer vision-based carbohydrate estimation for type 1 patients with diabetes using smartphones. Journal of Diabetes Science and Technology, 9(3), S07-515. http://dx.doi.org/10.1177/19322968155808159.

Aylon, S., Ciocca, G., Mazzini, D., & Schettini, R. (2020). Benchmarking algorithms for food localization and semantic segmentation. International Journal of Machine Learning and Cybernetics, 11(12), 2827–2847. http://dx.doi.org/10.1007/s13304-020-01153-z.
