Large-Scale Food Image Datasets and Food Recognition Systems: A Survey

Wenze Chen and Ruizhuo Song

Abstract—**With the development of the Internet, more and more people like to share their daily food, which makes the Internet rich in food image data. Many researchers collect food images on the Internet, label them, and combine machine learning and other methods to achieve food recognition systems. These food recognition systems can quickly help users record their daily diets and analyze their eating habits. The advanced food identification system can currently accurately identify hundreds of foods. At the same time, it can explore the food nutrition content and give food reference suggestions by accessing the food nutrition database. In this paper, we collect and evaluate large-scale food image datasets and food recognition systems.**

Index Terms—**Food recognition, deep learning, computer vision, object detection, mobile applications**

I. Introduction

Improper eating habits are the main cause of many diseases, including peripheral tissue diseases such as obesity, hypertension, and diabetes, and central tissue diseases such as emotional disorders and cognitive disorders [1, 2].

When patients suffer from these diseases, good eating habits are the key to helping them get rid of these diseases. However, in the traditional diagnosis and treatment process, patients must recall what food they have eaten recently, which may lead to errors. This will increase the difficulty of doctors' diagnosis and treatment. On the other hand, it is also difficult for doctors to correct the eating habits of patients. Patients must have enough information about food and its impact on their values (how specific nutrients affect their metabolism, blood status, body composition, etc.). Individuals can cultivate a confident attitude towards food by acquiring and mastering information and knowledge. This enables them to adopt optimal eating behaviour when hungry. Dietary advice and counselling effectively change individuals' attitudes towards food by enhancing their understanding of food. Nutritionists usually provide dietary advice and habits. While this is a great way to establish proper eating habits, developing multimedia tools and accessing to much online information calls for a more cost-effective and automated on-demand system for the public. Initially, the nutrition analysis program mainly records the food category and weight manually, and then sends them to the nutrition analysis module to calculate the nutrient content.

In recent years, with the development of deep learning and computer vision, the system can automatically identify food categories by training large-scale food image datasets.

The core of food recognition system is food feature extraction. The stronger the ability of food feature extraction, the higher the accuracy of food recognition. Traditional food feature extraction methods include scale invariant feature transform (SIFT) [3, 4], histogram of oriented gradient (HoG) [5], scale invariant local term pattern (SILTP) [6], local binary pattern (LBP) [7], and speed up robust feature (SURF) [8]. The extracted feature points are expressed into word bags, such as Fisher vector word bags [9]. Then, principal component analysis (PCA) [10] can be used to reduce the feature dimensions. Finally, classifiers such as support vector machine (SVM) [11] classify the images. Kitamura et al. [12] obtained features by combining color histograms with discrete cosine transform. After training the SVM classifier, the recognition accuracy of food pictures can reach 88.00%. However, traditional food recognition methods only extract single features, which is not practical for food images with small diversity and multi-categories.

In addition to the traditional methods, the research of convolutional neural networks (CNNs) in food image classification is also increasing. In the implementation based on food recognition, Kagaya et al. [13] used the plain convolution network to classify 10 types of food images and achieved an accuracy rate of 73.70%. It provides reference and direction for the later development of food recognition based on deep learning.

Later, object detection methods such as you only look once (YOLO) [14] and fully convolutional one-stage object detector (FCOS) [15] appeared. Many scholars expanded these methods to recognize multiple foods in the same image. For example, Mao et al. [16] realized food recognition for three public datasets based on region-based convolutional neural networks (RCNNs), as shown in Fig. 1. This method can give food category and location information at the same time.

This paper provides a detailed introduction to five publicly available large-scale food datasets and a summary and evaluation of the food recognition algorithms developed using these datasets. In addition, we introduce three food recognition systems in detail and describe the technical information used in the system.

Manuscript received: 30 October 2023; revised: 16 February 2024; accepted: 26 April 2024. *(Corresponding author: Ruizhuo Song.)*

Citation: W. Chen and R. Song, Large-scale food image datasets and food recognition systems: A survey, *IJICS*, 2024, 29(2), 95–100.

Wenze Chen and Ruizhuo Song are with School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China (e-mail: g20209452@xs.ustb.edu.cn; ruizhuosong@163.com).

Digital Object Identifier 10.62678/IJICS202403.10113

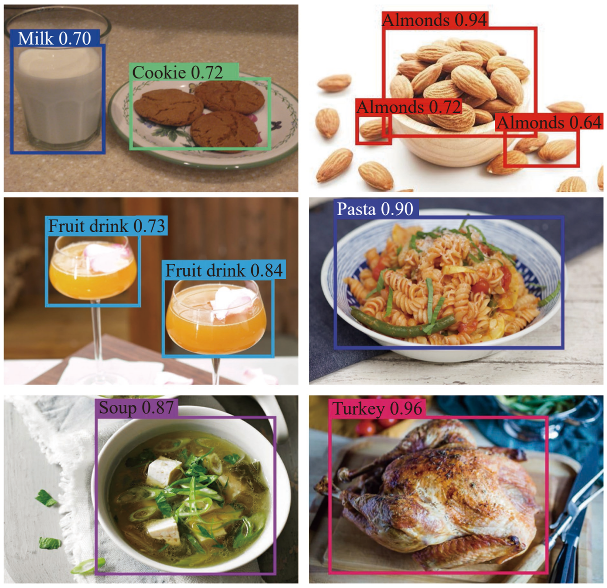


Figure 1 Detection examples of object detection networks. Numbers represent confidence scores, with higher values indicating more reliable prediction results.

II. Food Image Dataset

A. Food-101 Dataset

In the research of food data images, some universities and relevant research institutions have established some food datasets.

Bossard et al. [17] published a significant paper in the field of food identification. The paper introduces two main contributions. First, it proposes a food recognition model based on the random forest algorithm. This model improves classification efficiency by using color block alignment and enhances model learning efficiency by sharing parameters. The method achieves a mean average precision (mAP) of 50.76% in Food-101 dataset, which is better than other classification methods except for the convolutional neural network. When compared with SVM classification algorithm using Fisher vector, the image detection and recognition results are improved by 11.88%. The second contribution is the proposal of the Food-101 dataset, which consists of 101 kinds of food images, totaling 101,000 images. Most of the pictures in the dataset are of western food, and only the food category is marked, without including the food location information in the image. The dataset is primarily focused on classifying food images, with each class divided into a testing set and a training set. It is considered a critical dataset in food image recognition and an important indicator for evaluating the effectiveness of algorithm models.

Yanai and Kawano [18] used the deep convolutional neural network (DCNN) to extract food features, fine-tuned and pre-trained the DCNN, and finally obtained top-1 accuracy of 70.41% on the Food-101 dataset. In addition, it takes only 0.03 s to classify one food image with GPU.

Wu et al. [19] introduced a visual food recognition framework incorporating intrinsic semantic associations among fine-grained classes. The framework is based on joint depth feature learning and semantic tag reasoning and uses hierarchical semantics for food recognition on Food-101 dataset. The framework obtained a top-1 accuracy rate of 72.11%. The framework not only improved the accuracy of food recognition but also produced more semantic coherent predictions. Compared with the previous method of food recognition using deep neural networks alone, it has obvious advantages.

Pandey et al. [20] developed a multilayered CNN to recognize food. They combined three convolutional neural networks: AlexNet, GoogleNet, and ResNet. They superimposed the features extracted by different networks. Finally, the effect of the ensemble network they proposed was better than that of the CNN model of a single subnet. On Food-101 dataset, the method obtained top-1 accuracy of 72.12% and top-5 accuracy of 91.61%, respectively.

Martinel et al. [21] proposed a specific convolution structure for food recognition called WISeR, as shown in Fig. 2. This structure includes two backbone networks. One contains a convolution module and a maximum pooling layer, mainly extracting shallow features in food. The other backbone network contains multi-layer convolutions and an average pooling layer used to extract deep features in food. Finally, the features extracted from various branches are connected and fed to the fully connected layer that sends the classification prediction.

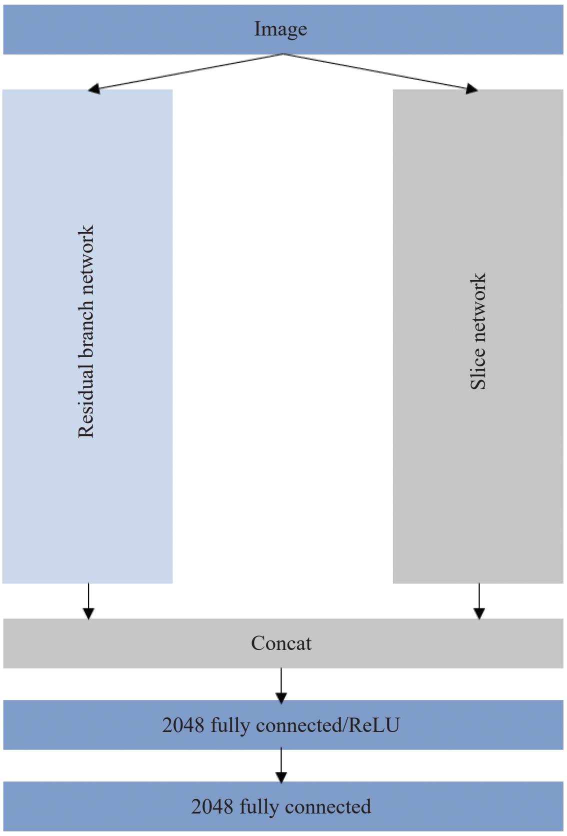


Figure 2 WISeR network structure diagram [21].

B. VireoFood-172 Dataset

Chen and Ngo [22] established a new food image dataset called VireoFood-172, which contains 110,241 images of 172

Chinese dishes. Like Food-101, this dataset only carries out category labeling, so it can only identify the overall classification of the image and one food in an image. 60% of the images in this dataset are used for training and 40% for testing. Their work mainly includes component identification and zero-shot recipe retrieval. The former is based on deep architecture (ArchD), which uses the reciprocal relationship between food and ingredient labels through multi-task learning. The latter extends ArchD's knowledge of non-seasonal scenes by learning the context of ingredients from many cookbook text corpora. After that, Jiang et al. [23] proposed a multi-scale multi-view feature aggregation (MSMVFA) method utilizing two two-dimensional CNN models. This approach involves fine-tuning the two-dimensional multi-scale components and the category CNN independently, and each scale's intermediate attribute features, high-level semantic features, and in-depth visual features are extracted using component information and food category as supervision label. Then, the fusion method of normalization and simple cascade is used to fuse three different features into multi-scale representations. Finally, three different types of multi-scale features are further aggregated into the final model using the multi-view feature aggregation method of z-score normalization and simple concatenation.

C. UECFood-100 Dataset

Matsuda et al. [24] constructed a new food image dataset named UECFood-100, including 100 categories, in which each food has category information and location information. The dataset contains about 100 images per category, totaling more than 14,000 images. It includes images of multiple foods and images of a single food. In food recognition, Matsuda et al. [24] first detected several candidate regions by integrating the outputs of multiple region detectors, including Felzenszwalb's deformable part model (DPM), circle detector, and joint space-color edge-based graph-cut (JSEG) region segmentation. In the second step, they applied the food recognition method, which is based on feature fusion, to the boundary boxes of candidate regions, utilizing various visual features. These features encompassed SIFT, color scale invariant feature transform (CSIFT), bag-of-features (BoFs) representations combined with the spatial pyramid (SP) model, and HoG. The combination of these features ensured a comprehensive and robust approach to food recognition. The experiment shows that they achieved a classification rate of 55.80% for the multi-food image dataset, which improves the baseline result with DPM only by 14.30%.

Liu et al. [25] developed a practical food recognition system based on deep learning for food evaluation in the edge computing service infrastructure. They obtained 76.30% top-1 accuracy and 94.60% top-5 accuracy on the UECFood-100 dataset. In addition to high recognition accuracy, the system can run in real-time on edge computing systems (such as mobile terminals).

People are increasingly interested in the wide application of smartphones in food image recognition. Among the existing methods, the method based on the middle image part shows promising performance because it is suitable for modeling the deformable food part (FP). However, the achievable accuracy is limited by the FP representation based on low-level features. The depth learning method has achieved the most advanced performance in some food image recognition problems thanks to the ability to learn powerful features from tag data. The middle layer based method and DCNN method have their advantages, but the most important thing is that they can be complementary. Therefore, Zheng et al. [26] introduced a novel framework that optimized the utilization of DCNN features for food image recognition by leveraging the combined strengths of middle layer based method and traditional DCNN approach. Furthermore, they tackled the difficulty of training DCNN models with unlabeled intermediate part data. They devised a clustering-based FP tag mining strategy to address this challenge of generating partial-level tags from unlabeled data. On UECFood-100 dataset, they finally obtained a top-1 accuracy of 86.51%.

D. UECFood-256 Dataset

Kawano and Yanai [27] built a new dataset named UECFood-256, encompassing 256 different types of foods. Each category in this dataset contains over 100 images. It was made by extending UECFood-100. Kawano and Yanai [27] used UECFood-256 as the training set to train a large-scale food recognition system. To deploy an image recognition system utilizing high-dimensional features on mobile devices, they employed a linear weight compression technique to optimize memory usage, achieving a top-5 accuracy of 74.40%.

Hassannejad et al. [28] classified food images based on Google image recognition architecture inception. The architecture is a 54-layer deep convolutional neural network. Then, they fine-tuned the architecture and obtained 76.17% top-1 accuracy and 92.58% top-5 accuracy on UECFood-256 dataset, respectively.

Later, some researchers used a lighter network structure for food recognition, which had a faster detection speed than previous methods [29].

E. VFN Dataset

Mao et al. [16] constructed a more challenging food dataset, which included 14,991 food images and 82 food categories, each of which may consist of many types of foods. The same food in this dataset may have significant differences in appearance, and the same food may have different morphological and color characteristics due to other raw materials and practices, which increases the difficulty of online learning. Mao et al. [16] used faster RCNN based methods for model training and ultimately achieved an average accuracy of 40.06% using this two-stage object detection algorithm.

III. Food Recognition System

A. FoodCam

Ravì et al. [30] implemented the augmented reality (AR) nutrition system of augmented reality based on FoodCam, using technologies such as artificial intelligence and computer vision. This system works similarly to a scanner. Compared with other systems in this category, in addition to food recognition, FoodCam also aims to identify user activities (walking, running, standing, leisure sports, etc.). Food images are recorded in a scanner manner. As shown in Fig. 3, the user will receive a list of the top five most likely foods after the initial identification process. Still, the system will not display the nutritional information in food.

Image acquisition and user input The food images are captured in a scanning manner. Following the initial recognition process, the system presents the user with a top five list of the most probable food items.

Feature extraction Three distinct types of features are investigated: the histogram of oriented gradients, the local binary pattern descriptor, and the red green blue (RGB) color features. Subsequently, these local descriptors are aggregated into a Fisher vector representation.

Image classification A linear support vector machine is employed for the classification task. Initially, classification is performed using individual features, and later, combinations of multiple features are utilized to enhance classification performance.

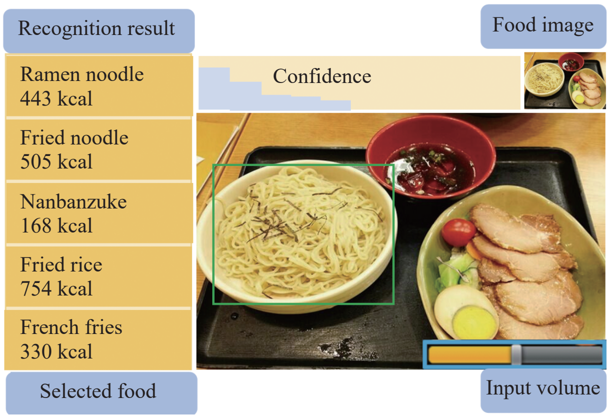


Figure 3 Top five classification list provided after identification process [30].

B. DietLens

Ming et al. [31] introduced and implemented a neural network powered food object detection system called DietLens. The overarching goal of this system is to promote health improvement by connecting clients with health professionals, friends, and support groups. Leveraging a client-server architecture with a vast database enables a more extensive neural network to analyze food instances. Furthermore, the authors proposed an innovative approach for portion size estimation.

Image acquisition and user input The system requires three distinct types of inputs. Initially, a photograph of food is captured. Subsequently, the user selects from a result list and optionally provides an estimate of the portion size. The authors presented a novel selection process for portion size estimation.

Feature extraction A large ResNet-50 convolutional neural network comprising 50 convolutional layers is utilized for feature extraction.

Image classification and dataset The ResNet-50 CNN is employed for the classification task, where feature extraction and classification procedures are executed concurrently. The system identified over 249 food categories during testing, encompassing nearly 88,000 images.

Volume estimation This approach achieves good results without relying on complex reconstruction algorithms. Following the classification procedure, the user is presented with a list of portion sizes corresponding to the identified food items. The user selects one of the images to finalize the portion size estimation. The authors compared their food object recognition and portion estimation approach with other manual input systems, such as myFitnessPal and Fatsecrets.

The proposed DietLens system outperforms all other systems in speed while maintaining exceptional classification and portion size estimation accuracy. In real-life testing environments, the entire process of logging a meal takes an average of 11.58 s.

C. Food Recognition System on Google Glass

The system devised by Jiang et al. [32] stands unique in its approach to object recognition, utilizing Google glass as the mobile device. This method aligns more closely with augmented reality, integrating artificial intelligence into our intuitive perception and daily routines. Notably, the primary goal of this system is to provide consumers with informative data before purchasing food items. The system is structured using a client-server architecture.

Image acquisition and user input The image acquisition process starts with setting a reference frame. Then, multiple food frames are captured and sent to the backend for recognition. After the system identifies the food, it lists food items and their nutrition information. The user selects the correct item to see detailed nutrition data.

Image processing No preliminary manipulations are performed on images before they are forwarded to the subsequent classification stages. As for image segmentation, a reference frame chosen by the user serves as the foundation for determining which food item is being tracked and analyzed.

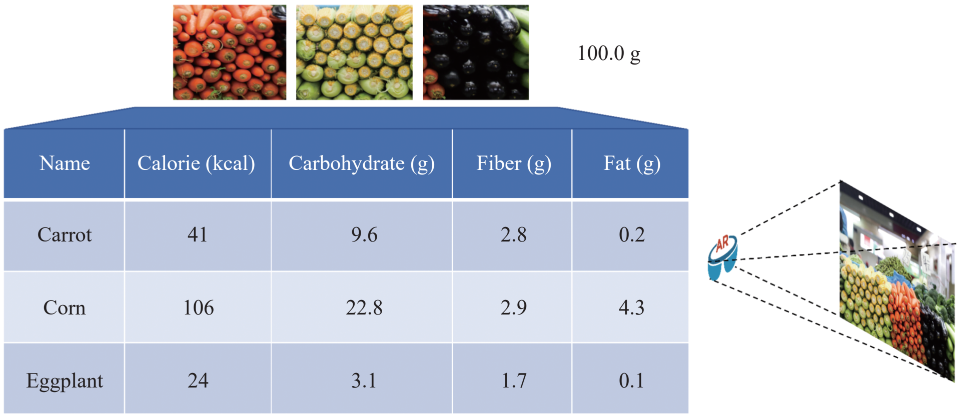
Feature extraction Captured reference images are uploaded to Google's hash server for classification. Each image is represented as a feature vector, encompassing colour, points, lines, textures, and SIFT.

Image classification and dataset Images undergo classification using the reverse image search (RIS) algorithm and text mining techniques. The system leverages Google's extensive image database as dataset.

Volume estimation As shown in Fig. 4, the system presents nutritional information based on a standard measurement of 100.0 g of food. These nutritional data are sourced from the comprehensive dataset of the United States Department of Agriculture (USDA).

IV. CONCLUSION

This paper introduced five large-scale food datasets in detail, including Food-101, VireoFood-172, UECFood-100, UECFood-256, and VFN, respectively, and introduced more than ten food recognition methods using these datasets as



Name	Calorie (kcal)	Carbohydrate (g)	Fiber (g)	Fat (g)
Carrot	41	9.6	2.8	0.2
Corn	106	22.8	2.9	4.3
Eggplant	24	3.1	1.7	0.1

Figure 4 Example of food identification on Google glass, identifying food categories and giving reference nutrient content [32].

evaluation criteria. We also introduced three food recognition systems based on intelligent terminals (such as smartphones), which are more integrated and open to people. We hope to provide some introductory guidance for developers and researchers committed to developing food recognition systems.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (Nos. 62273036 and 61873300) and the Interdisciplinary Research Project for Young Teachers of USTB (Fundamental Research Funds for the Central Universities) (No. FRF-IDRY-20-030).

REFERENCES

[1] T. Psaltopoulou, T. N. Sergentanis, D. B. Panagiotakos, I. N. Sergentanis, R. Kosti, and N. Scarmeas, Mediterranean diet, stroke, cognitive impairment, and depression: a meta-analysis, *Ann. Neurol.*, 2013, 74, 580–591.

[2] S. Knez and L. Šajn, Food object recognition using a mobile device: Evaluation of currently implemented systems, *Trends Food Sci. Technol.*, 2020, 99, 460–471.

[3] D. G. Lowe, Object recognition from local scale-invariant features, in *Proc. Seventh IEEE International Conference on Computer Vision*, Kerkyra, Greece, 1999, 1150–1157.

[4] D. G. Lowe, Distinctive image features from scale-invariant keypoints, *Int. J. Comput. Vis.*, 2004, 60(2), 91–110.

[5] N. Dalal and B. Triggs, Histograms of oriented gradients for human detection, in *Proc. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 2005, 886–893.

[6] S. C. Liao, G. Y. Zhao, V. Kellokumpu, M. Pietikäinen, and S. Z. Li, Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes, in *Proc. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, 2010, 1301–1306.

[7] T. Ojala, M. Pietikäinen, and T. Mäenpää, Multiresolution gray-scale and rotation invariant texture classification with local binary pattern, *IEEE Trans. Pattern Anal. Machine Intell.*, 2002, 24(7), 971–987.

[8] H. Bay, T. Tuytelaars, and L. Van Gool, SURF: Speeded up robust features, in *Proc. 9th European Conference on Computer Vision*, Graz, Austria, 2006, 404–417.

[9] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, Image classification with the Fisher vector: Theory and practice, *Int. J. Comput. Vis.*, 2013, 105(3), 222–245.

[10] J. Shlens, A tutorial on principal component analysis [Online], https://arXiv.org/abs/1404.1100, 4 December 2022.

[11] C. J. C. Burges, A tutorial on support vector machines for pattern recognition, *Data Min. Knowl. Discovery*, 1998, 2(2), 121–167.

[12] K. Kitamura, T. Yamasaki, and K. Aizawa, Food log by analyzing food images, in *Proc. 16th ACM International Conference on Multimedia*, Vancouver, Canada, 2008, 999–1000.

[13] H. Kagaya, K. Aizawa, and M. Ogawa, Food detection and recognition using convolutional neural network, in *Proc. 22nd ACM International Conference on Multimedia*, Orlando, FL, USA, 2014, 1085–1088.

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, You only look once: Unified, real-time object detection, in *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, 779–788.

[15] Z. Tian, C. H. Shen, H. Chen, and T. He, FCOS: Fully convolutional one-stage object detection, in *Proc. 2019 IEEE/CVF International Conference on Computer Vision*, Seoul, Korea, 2019, 9626–9635.

[16] R. Y. Mao, J. P. He, Z. M. Shao, S. K. Yarlagadda, and F. Q. Zhu, Visual aware hierarchy based food recognition, in *Proc. Pattern Recognition. ICPR International Workshops and Challenges*. 2021, 571–598.

[17] L. Bossard, M. Guillaumin, and L. Van Gool, Food-101—Mining discriminative components with random forests, in *Proc. 13th European Conference on Computer Vision*, Zurich, Switzerland, 2014, 446–461.

[18] K. Yanai and Y. Kawano, Food image recognition using deep convolutional network with pre-training and fine-tuning, in *Proc. 2015 IEEE International Conference on Multimedia & Expo Workshops*, Turin, Italy, 2015, 1–6.

[19] H. Wu, M. Merler, R. Uceda-Sosa, and J. R. Smith, Learning to make better mistakes: Semantics-aware visual food recognition, in *Proc. 24th ACM International Conference on Multimedia*, Amsterdam, The Netherlands, 2016, 172–176.

[20] P. Pandey, A. Deepthi, B. Mandal, and N. B. Puhan, FoodNet: Recognizing foods using ensemble of deep networks, *IEEE Signal Process. Lett.*, 2017, 24(12), 1758–1762.

[21] N. Martinel, G. L. Foresti, and C. Micheloni, Wide-slice residual networks for food recognition, in *Proc. 2018 IEEE Winter Conference on Applications of Computer Vision*, Lake Tahoe, NV, USA, 2018, 567–576.

[22] J. Chen and C. W. Ngo, Deep-based ingredient recognition for cooking recipe retrieval, in *Proc. 24th ACM International Conference on Multimedia*, Amsterdam, The Netherlands, 2016, 32–41.

[23] S. Jiang, W. Min, L. Liu, and Z. Luo, Multi-scale multi-view deep feature aggregation for food recognition, *IEEE Trans. Image Process.*, 2020, 29, 265–276.

[24] Y. Matsuda, H. Hoashi, and K. Yanai, Recognition of multiple-food images by detecting candidate regions, in *Proc. 2012 IEEE*

International Conference on Multimedia and Expo, Melbourne, Australia, 2012, 25–30.

[25] C. Liu, Y. Cao, Y. Luo, G. L. Chen, V. Vokkarane, M. Yunsheng, S. Q. Chen, and P. Hou, A new deep learning-based food recognition system for dietary assessment on an edge computing service infrastructure, *IEEE Trans. Serv. Comput*., 2018, 11(2), 249–261.

[26] J. N. Zheng, L. Zou, and Z. J. Wang, Mid-level deep food part mining for food image recognition, *IET Comput. Vis*., 2018, 12(3), 298–304.

[27] 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 International Conference on Multimedia*, Orlando, FL, USA, 2014, 761–762.

[28] H. Hassannejad, G. Matrella, P. Ciampolini, I. De Munari, M. Mordonini, and S. Cagnoni, Food image recognition using very deep convolutional networks, in *Proc. 2nd International Workshop on Multimedia Assisted Dietary Management*, Amsterdam, The Netherlands, 2016, 41–49.

[29] G. Ciocca, P. Napoletano, and R. Schettini, Food recognition: a new dataset, experiments, and results, *IEEE J. Biomed. Health Inform*., 2017, 21(3), 588–598.

[30] D. Ravì, B. Lo, and G. Z. Yang, Real-time food intake classification and energy expenditure estimation on a mobile device, in *Proc. 2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks*, Cambridge, MA, USA, 2015, 1–6.

[31] Z. Y. Ming, J. J. Chen, Y. Cao, C. Forde, C. W. Ngo, and T. S. Chua, Food photo recognition for dietary tracking: System and experiment, in *Proc. 24th International Conference*, Bangkok, Thailand, 2018, 129–141.

[32] H. T. Jiang, J. Starkman, M. H. Liu, and M. C. Huang, Food nutrition visualization on Google glass: Design tradeoff and field evaluation, *IEEE Consum. Electron. Mag*., 2018, 7(3), 21–31.

Wenze Chen received the BS degree from Harbin University of Science and Technology, China, in 2020. He is currently pursuing the MS degree at School of Automation and Electrical Engineering, University of Science and Technology Beijing, China. His research interests include deep learning and food recognition.

Ruizhuo Song received the PhD degree in control theory and control engineering from Northeastern University, Shenyang, China, in 2012. From 2013 to 2014, she was a visiting scholar at Department of Electrical Engineering, University of Texas at Arlington, Arlington, TX, USA. From January 2018 to February 2018, she was a visiting scholar at Department of Electrical, Computer, and Biomedical Engineering, University of Rhode Island, Kingston, RI, USA. She was a postdoctoral fellow at University of Science and Technology Beijing, Beijing, China, in 2012. She is currently a professor at School of Automation and Electrical Engineering, University of Science and Technology Beijing, China. She has authored or coauthored more than 100 journals and conference articles and coauthored 3 monographs. Her research interests include intelligent perception and computing, optimal decision making and gaming for virtual/real complex systems, and the results are applied to human body signal measurement and control, target positioning and recognition, energy internet, etc.