Food Classification from Images Using Convolutional Neural Networks

David J. Attokaren, Ian G. Fernandes, A. Sriram, Y.V. Srinivasa Murthy, and Shashidhar G. Koolagudi
Deapartment of CSE, National Institute of Technology Karnataka,
Surathkal, Mangalore, India - 575 025.
davidjattokaren@gmail.com, iangfernandes96@gmail.com, anirudh.sriram96@gmail.com,yvsm@nitk.ac.in, and koolagudi@nitk.ac.in

***Abstract*—The process of identifying food items from an image is quite an interesting field with various applications. Since food monitoring plays a leading role in health-related problems, it is becoming more essential in our day-to-day lives. In this paper, an approach has been presented to classify images of food using convolutional neural networks. Unlike the traditional artificial neural networks, convolutional neural networks have the capability of estimating the score function directly from image pixels. A 2D convolution layer has been utilised which creates a convolution kernel that is convolved with the layer input to produce a tensor of outputs. There are multiple such layers, and the outputs are concatenated at parts to form the final tensor of outputs. We also use the Max-Pooling function for the data, and the features extracted from this function are used to train the network. An accuracy of 86.97% for the classes of the FOOD-101 dataset is recognised using the proposed implementation.**

***Index Terms*—Convolution filters; Convolution layer; Convolutional neural networks; Food-101 dataset; Food classification; Image recognition; MAX pooling.**

I. Introduction

In the current age, people are more conscious about their food and diet to avoid either upcoming or existing diseases. Since people are dependent on smart technologies, provision of an application to automatically monitor the individuals diet, helps in many aspects. It increases the awareness of people in their food habits and diet [1]–[5]. Over the last two decades, research has been focused on automatically recognising the food and their nutritional information from images captured using computer vision and machine learning techniques. In order to properly assess dietary intake, accurate estimation of calorie value of food is of paramount importance. A majority of the people are overeating and not being active enough. Given how busy and stressed people are today, it's effortless to forget to keep track of the food that they eat. This only increases the importance of proper classification of food.

Recently, smart applications for mobile devices such as Android phones and iPhone, have increased tremendously. They are capable of balancing the food habits of users and also warn them about unhealthy food. Due to the advances in various technologies used in smartphones, their computational power has also increased. They are capable of processing real-time multi-media information with their computational power, whereas traditional mobiles are incapable and hence, used to send the images to high processing servers that increase the cost of communication and delay. Since the present smartphones can handle the high-quality images too, research on food classification is focused on developing real-time applications which capture images and train the machine learning models instantly. It helps to take prevention to avoids diseases such as diabetes, blood pressure and so on.

Some of the methods currently in use for dietary assessment involve self-reporting and manually recorded instruments. The issue with such methods of assessment is that the evaluation of calorie consumption by a participant is prone to bias [6], i.e. underestimating and under reporting of food intake. In order to increase the accuracy and reduce the bias, enhancements to the current methods are required. One such potential solution is a mobile cloud computing system, which makes use of devices such as smartphones to capture dietary and calorie information. The next step is to automatically analyse the dietary and calorie information employing the computing capacity of the cloud for an objective assessment. However, users still have to enter the information manually. Over the last few years, plenty of research and development efforts have been made in the field of visual-based dietary and calorie information analysis. However, the efficient extraction of information from food images remains a challenging issue.

In this paper, an effort has been made to classify the images of food for further diet monitoring applications using convolutional neural networks (CNNs). Since the CNNs are capable of handling a large amount of data and can estimate the features automatically, they have been utilised for the task of food classification. The standard Food-101 dataset has been selected as the working database for this approach.

The rest of the paper is organised as follows. Section II details the related works in the field of food classification with their merits and demerits. Section III explains the proposed methodology including the database selected, and provides a description of the CNN. Section IV discusses the results and the observations. Finally, section V concludes the work with some future directions.

II. Related Work

The task of the food detection system is first initiated with four fast-food classes namely fries, apple pies, hamburgers and chicken burgers [7]. The images were segmented initially to form the feature vector with size, shape, texture, color (normalised RGB), and other context-based features. With this motivation, a minimised feature vector with the Gabor filter responses (texture), pixel intensity, and color components

978-1-5090-1134-6/17/$31.00 ©2017 IEEE

is used to categorise the 19 classes of foods. However, the performance is good for food replicas, and a less efficient performance is observed with real images [8]. The size of images and their variations in capturing could be the reason for the performance degradation. Based on this, scale invariant feature transform (SIFT) features have been extracted and experimented on homemade foods, fast-food, and fruits [9]. With this, the better performance is found with less number of classes, although the images of each class are more.

The term bag of features (BoF) which is derived from the bag of words (BoW) is the emerging trend in recent days. It is highly influenced to process the natural language. It is designed to catch frequently appearing words by ignoring the order in which they appear [10], [11]. Similarly, images contain some common visual patterns that are useful in recognising the category of food. This process reduces the complexity issues raised by the direct image matching techniques. Based on this, some works are found using the BoF approach.

The existing literature has utilised the variety of classifiers available. The notable ones and better performance giving classifiers among them are artificial neural networks (ANNs), support vector machines (SVMs), Naive Bayes, and Adaboost. Further, the pairwise classification framework has been proposed to enhance the recognition rate of food classification [12]. Texton histograms have been utilised to resemble BoF models. It is found that they can carry less information and failed to deal with high-resolution images. Moreover, the performance is not good for varying light conditions. Hence, a checker-board which is colored is captured to utilise the system for varying light conditions. However, the performance accuracy reduces from 95% to 80% when there is an increase in the number of classes. The database is a paramount attribute for the task of food classification, and the system should handle this scenario in a real-time fashion. Hence, a real-time database with fast-food images has been created and made as a benchmark by Chen *et al.* [13]. Later on, experiments have been conducted on the same dataset by considering it as a benchmark. At the initial stage, SIFT features have been used and tested with seven classes. The SVMs are considered as a classifier and provide an accuracy of 47%. Further, the usage of combined bag of SIFT along with Gabor filter responses, and color histogram based features are considered [14]. Using the k-means clustering technique and with 50 food classes, a recognition rate of 61% is recorded.

Later the research has focused on collecting the varieties of dataset and Gianluigi Ciocca *et al.* [15] proposed a new dataset to evaluate algorithms to recognize food which helps to monitor diets. The database has been built which contains more than 3,500 instances of food using images of canteen trays with food. The set of local and global features are also extracted and experimented on using the classifiers, namely SVMs [16], k-NN classifier [8], random forest (RF) [17], neural networks [18] that give knowledge on deciding the suitable classifier [19]. A medium size dataset has been created to develop a mobile-based log system, and for 85 food items, an accuracy of 62% is reported [8]. The three-dimensional properties of food shapes are used to reconstruct and further extracted the feature values [20].

Deep Convolutional Neural Networks have been used for food recognition recently [21], which have used the UEC-100 and UEC-256 datasets for testing, along with ImageNet and ILSVRC for training, which use a combination of baseline feature extraction and neural network fine-tuning. Another approach [22] uses Convolutional Neural Networks along with a Global Average Pooling layer, which generates Food Activation Maps (heat maps of food probability). Fine tuning is done for FAM generation, which includes adding a convolutional layer with stride, and setting a softmax layer. Additionally, via thresholding, bounding boxes are generated.

The present work aims to combine some of the above methodologies together, that creates a food classification system, that predicts the class of food the image is in, and also gives the calorie count based on the portion size visible. This concept has a high scope in the health sector, as people want to keep track of what and how much they eat and simplifying the process into the form of this implementation increases usage and awareness of health-related factors. Since CNNs are less focused in the literature, they have been utilized due to their inherent capabilities in computing features automatically.

III. Proposed Methodology

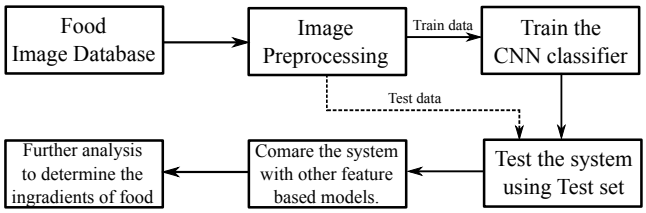

Fig. 1. The proposed framework to recognize the food items from images.

The proposed flow diagram is depicted in Fig. 1. Each block of the proposed flow diagram is clearly explained in this section. The steps in image processing are pre-processing and neural network training. From this, the trained model can be obtained which will classify any supplied image based on the trained dataset.

A. The Food-101 Dataset

The Food-101 dataset which contains 101,000 images and 101 categories is considered for this work. The reason for considering the dataset mentioned above, is to make the system more realistic. In the dataset, each food category contains 750 training and 250 test clips [5]. The majority of the training and testing clips are filled with noise, intense color, and some images are with wrong labels. Care has been taken to label the training and testing images properly. Further, the images have been rescaled to a unique size of 299x299 dimensions.

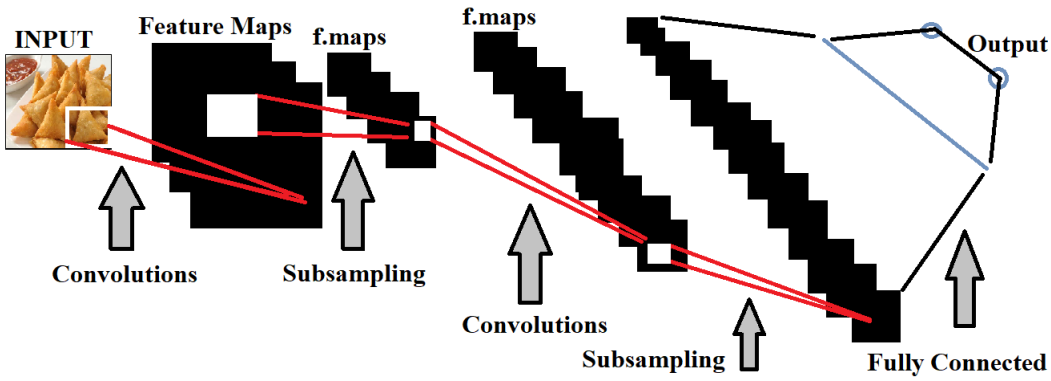

Fig. 2. The role of convolutional neural networks in the proposed food classification system.

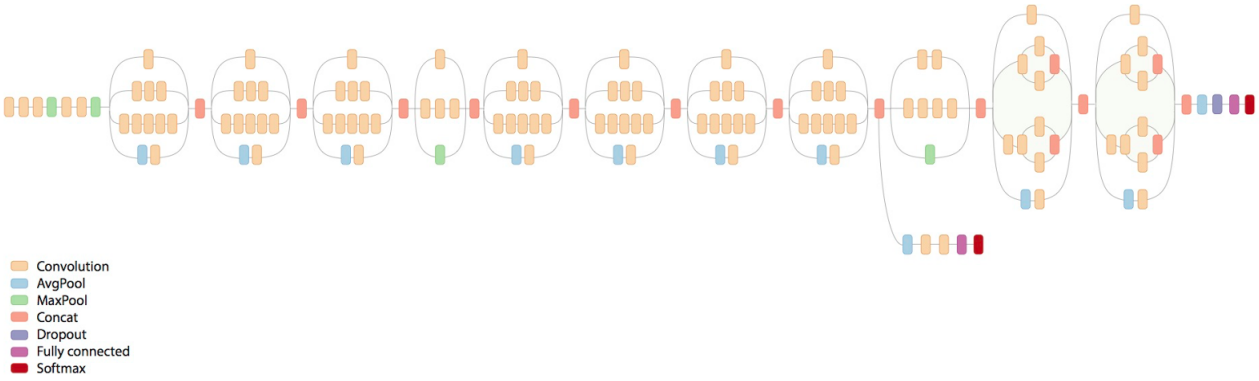

Fig. 3. The configuration of the CNN, which retrains a Google Inception V3 model.

B. Neural Network Configuration

A Google Inception V3 model (pretrained on ImageNet) is retrained. The architecture of the neural network is shown in Fig. 3. The layers are:

1. AvgPool: An AveragePooling2D function is used (pool size (8,8)). This reduces variance in the data and reduces computational complexity as well. This layer flows the output into the next layer.

2. Convolution: A Convolution2D function is used (input size (299,299,3). This layer creates feature maps by convolving input data to create feature maps.

3. MaxPool: A MaxPooling2D function is used. The pooling function reduces variance in the data and reduces computational complexity. Max pooling extracts the most important features like edges whereas, average pooling extracts features smoothly.

4. Concat: The Concat layer is a utility layer that concatenates its multiple input blobs to one single output blob. It takes as input a list of tensors, all of the same shape expect for the concatenation axis, and returns a single tensor, the concatenation of all inputs.

5. Dropout: Dropout is a regularization technique for reducing overfitting in neural networks by preventing complex co-adaptations on training data. It is a very efficient way of performing model averaging with neural networks. The term ”dropout” refers to dropping out units (both hidden and visible) in a neural network. We have defined a dropout scale of 0.4.

6. Fully connected: Fully connected layers connect every neuron in one layer to every neuron in another layer. It is in principle the same as the traditional multi-layer perceptron neural network (MLP).

7. Softmax: Using Softmax function as an output function almost works like Max layer as well as it is differentiable to train by gradient descent. Exponential function will increase the probability of maximum value of the previous layer compare to other value. Also, summation of all output will be equal to 1.0 always.

C. Image preprocessing

There are a few image preprocessing techniques used to ensure maximum efficiency from the proposed system. These preprocessing techniques ensure that any image taken from any angle will be able to get classified.

1) Image preprocessing parameters: The following are the parameters that are considered for image preprocessing.

- *Rotation_range = 45:* Images are randomly rotated by a degree of 45. This ensures that images taken at any angle can be predicted correctly, and that diversity of the patterns obtained (feature maps) is maintained.
- *Width_shift_range = 0.2:* Images are shifted horizontally by this fraction. This allows for "incomplete" or "half" images to be predicted, and patterns obtained will differ.
- *Height_shift_range = 0.2:* Images are shifted vertically by this fraction of the total width. The purpose is same as mentioned in horizontal shift.
- *Horizontal_flip = True:* Images are flipped horizontally. Random flipping of images helps in identifying different patterns and for "upside down" images to be predicted accurately as well.
- *Fill_mode = reflect:* Points outside the boundaries of the images are filled according to this mode.
- *Train_datagen.config['random_crop_size']:* Assigns the crop size for the images that are fed to the network, in this case 299x299x3. All images are forced to be cropped to this resolution which ensures the compatibility and linearity of input to the neural network.

D. Image Processing to CNN

The present work utilized the Google InceptionV3 model [13] which is pre-trained on ImageNet. Prior to that, all the images are reshaped to 299x299x3 size. The global average pooling function is applied on the dataset which takes the average of all features of an image. The dimensionality of output space is defined by using the $dense()$ function. The dropout fraction rate on input units with 0.5 is considered to avoid overfitting issues. Further, to determine the actual class from $n-$ number of classes softmax activation function is defined. It identifies the class based on the maximum probability obtained at output for that class and ignores the rest.

E. Neural Network Training

The simple CNN used for the proposed work is depicted in Fig. 2. The Stochastic Gradient Descent with a quickly decreasing learning schedule has been used to achieve better performance. The model is trained for 32 epochs and has three callbacks defined which record the progress into a log file. A learning rate scheduler is defined which takes the epoch index as input and returns a new learning rate as output. Model checkpoints are made via the check pointer callback. These are saved in the form of .hdf5 files. The best score is considered to save only best learned models.

F. Usage of Neural Networks and Web Scraping

This subsections describes the use of neural networks and web scraping for the task of food classification.

1) Classification of Food: The first step in tracking calorie intake using images is to identify the food being consumed. The difficulty arises when one considers the various assortments of cuisines and dishes that exist in the real world. Given the size and variety of the food items in the dataset, this has proven to be quite a formidable task. The use of neural networks seems to be better to deal with the issue of scaling primarily because of their ability to learn patterns that are not linearly separable, along with the concepts of dealing with other factors such as noise in the images.

2) Calorific Value Estimation: The remaining task after the process of classification is mapping the food names to a calorific value. This can be achieved relatively easily by scraping the web for the average calorie value of food items per unit weight. The average calorie values are considered for the different classes food, per 100g of serving.

3) CNN specifications: The most popular readily available dataset for image classification is the ImageNet database, which has been used to train the Google Inception CNN. It also has multiple existing classification categories. To generalize the system, another 101 categories are added to this model by training it on the Food-101 dataset. The specifications of the model are as follows. The input sensor has a size of 299x299x3, with a Max pooling downscale of 2 in each spatial dimension with a dropout rate of 0.4 and the softmax activation function. The optimization function used for this task is stochastic gradient descent, which basically finds the maxima and minima through optimal number of iterations.

4) Image Augmentation: In this step, one-hot encode is used to get a set of binary features from each label. This is better than one feature that can take on any value from $n-$ classes. An image augmentation pipeline has been used that comes with cropping tools and the inception image preprocessor. Using a multiprocessing tool allows for GPU usage to be maximized.

IV. Results and Analysis

This section discusses the results, and the observations found while experimenting starting from the performance measurement techniques for food classification.

A. Evaluation of Models

Now having multiple saved models, it is possible to evaluate them and to load the models with lowest loss/highest accuracy. Further, a confusion matrix has been considered based on the outputs obtain by CNNs. A confusion matrix will plot each class label, and how many times it was correctly labeled vs. the other times it was incorrectly labeled as a different class. To evaluate the test set, multiple crops have been used instead of single value. This increases accuracy as compared to a single crop evaluation scheme. The output is the $top - N$ predictions for each crop, which in turn is used to process the $top - 5$ predictions. Crops are created for every item in the test set. These are used to get the predictions. Hence, predictions for each image has been obtained from this stage. Mapping technique is used to map the test item index to the top predictions.

B. Obtained Results

The confusion matrix discussed earlier will show the correctly versus incorrectly labeled classes. From results, it is found that the CNNs are more appropriate for image classification. They provide features such as filtering and Max-pooling that gives better recognition rate for image classification than traditional neural networks. Convolving the image allows feature extraction, regardless of its orientation and position in the image.

TABLE I
A COMPARISON OF TOP ACCURACIES FOR DIFFERENT MODELS AND DATASETS.

Sl.No.	Model	Dataset	Accuracy (in %)
1.	SVM	Food-101	50.76
2.	Neural Networks	Food-101	56.40
3.	RFDC-based Approach	Food-101	56.76
4.	Resnet18	Food-101	67.23
5.	CNN	UEC-FOOD100	78.77
6.	CNN (ILSVRC)	Food-101	79.20
7.	CNN (Food-101)	EgocentricFood	90.90
8.	Proposed Approach	Food-101	**86.97**

The results obtained for various classifiers used in literature on the standard Food-101 dataset are shown in Table I. The present system is developed using CNNs on Food-101 dataset which gives 86.97% accuracy at $top-1$ prediction. The same is giving better performance at $top-5$ prediction, and the accuracy is 97.42%. However, the $top-1$ prediction alone is considered to compare with the other models. The same Food-101 dataset is considered to compare with the state-of-art systems.

Initially, the SVMs and neural networks are considered as they are designed to capture the patterns which are highly non-linear. Random forest decision classifier (RFDC) is used since it is the most acceptable classifier for non-linear patterns in present trend. As assumed, the RFDCs had given better accuracy when compared to other two classifiers mentioned above. However, the features based on SIFT invariant technique, BoF models, and other useful features have been computed for above three classifiers. Later, recent work using the Resnet18 model has been used to test with the Food-101 dataset. The Resnet18 model has given better accuracy when 10 classes are used from the Cifar10 dataset which is around 86%. The same model offers less performance with the Food-101 dataset. The performance is around 67.23%. Since the CNNs are capable of estimating the features automatically and highly capable of mapping the non-linear relations, a better accuracy of 86.97% is obtained with CNNs. However, experimentation is yet to be done on realistic images and all kinds of food.

V. CONCLUSION AND FUTURE WORK

The performance of the system is high, and is considered acceptable from a usage point of view. However, the CNNs need high-performance computing machines in order to experiment on the huge multi-media datasets. The CNN is capable of train highly non-linear data, and for that in contrast, it takes more computational time to train the network. However, the performance matters a lot, and once the system is properly trained, the system can produce the results in less time. The images are properly preprocessed and all kinds of images are tested with CNN. From this, it is concluded that CNNs are more suitable for classifying the images when the number of classes are more.

The task of image classification can be extended using prominent features that can categorize food images. Since the CNNs are consuming high computational time, the feature-based approach is highly appreciable. A multi-level classification approach (hierarchical approach) is suitable to avoid mis-classifications when the number of classes is more. Moreover, a dataset containing all food categories is also not available in the literature yet.

REFERENCES

[1] W. Wu and J. Yang, "Fast food recognition from videos of eating for calorie estimation," in *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*. IEEE, 2009, pp. 1210–1213.

[2] N. Yao, R. J. Sclabassi, Q. Liu, J. Yang, J. D. Fernstrom, M. H. Fernstrom, and M. Sun, "A video processing approach to the study of obesity," in *Multimedia and Expo, 2007 IEEE International Conference on*. IEEE, 2007, pp. 1727–1730.

[3] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar, "Food recognition using statistics of pairwise local features," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2249–2256.

[4] M. Bosch, F. Zhu, N. Khanna, C. J. Boushey, and E. J. Delp, "Combining global and local features for food identification in dietary assessment," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*. IEEE, 2011, pp. 1789–1792.

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

[6] P. Pouladzadeh, S. Shirmohammadi, and R. Al-Maghrabi, "Measuring calorie and nutrition from food image," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 8, pp. 1947–1956, 2014.

[7] G. Shroff, A. Smailagic, and D. P. Siewiorek, "Wearable context-aware food recognition for calorie monitoring," in *Wearable Computers, 2008. ISWC 2008. 12th IEEE International Symposium on*. IEEE, 2008, pp. 119–120.

[8] F. Zhu, M. Bosch, I. Woo, S. Kim, C. J. Boushey, D. S. Ebert, and E. J. Delp, "The use of mobile devices in aiding dietary assessment and evaluation," *IEEE journal of selected topics in signal processing*, vol. 4, no. 4, pp. 756–766, 2010.

[9] F. Kong and J. Tan, "Dietcam: Automatic dietary assessment with mobile camera phones," *Pervasive and Mobile Computing*, vol. 8, no. 1, pp. 147–163, 2012.

[10] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2. IEEE, 2005, pp. 524–531.

[11] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," *Machine learning: ECML-98*, pp. 137–142, 1998.

[12] M. Puri, Z. Zhu, Q. Yu, A. Divakaran, and H. Sawhney, "Recognition and volume estimation of food intake using a mobile device," in *Applications of Computer Vision (WACV), 2009 Workshop on*. IEEE, 2009, pp. 1–8.

[13] M. Chen, K. Dhingra, W. Wu, L. Yang, R. Sukthankar, and J. Yang, "Pfid: Pittsburgh fast-food image dataset," in *Image Processing (ICIP), 2009 16th IEEE International Conference on*. IEEE, 2009, pp. 289–292.

[14] T. Joutou and K. Yanai, "A food image recognition system with multiple kernel learning," in *Image Processing (ICIP), 2009 16th IEEE International Conference on*. IEEE, 2009, pp. 285–288.

[15] G. Ciocca, P. Napoletano, and R. Schettini, "Food recognition: A new dataset, experiments, and results," *IEEE journal of biomedical and health informatics*, vol. 21, no. 3, pp. 588–598, 2017.
[16] G. M. Farinella, M. Moltisanti, and S. Battiato, "Classifying food images represented as bag of textons," in *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 5212–5216.
[17] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–mining discriminative components with random forests," in *Lecture Notes in Computer Science*, vol. 8694. Springer, 2014, pp. 446–461.
[18] H. Kagaya, K. Aizawa, and M. Ogawa, "Food detection and recognition using convolutional neural network," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 1085–1088.
[19] Y. He, C. Xu, N. Khanna, C. J. Boushey, and E. J. Delp, "Analysis of food images: Features and classification," in *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 2744–2748.
[20] J. Chae, I. Woo, S. Kim, R. Maciejewski, F. Zhu, E. J. Delp, C. J. Boushey, and D. S. Ebert, "Volume estimation using food specific shape templates in mobile image-based dietary assessment," in *Proceedings of SPIE*, vol. 7873. NIH Public Access, 2011, p. 78730K.
[21] Y. K. Keiji Yanai, "Food image recognition using deep convolutional network with pre-training and fine-tuning," *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2015.
[22] P. R. Marc Bolanos, "Simultaneous food localization and recognition," *arXiv:1604.07953v2 [cs.CV]*, 2017.