A Study of Food Recognition Techniques

Suvarna Pansambal[1] , Yamini Tawde[2], Chetali Surti[3] , Chhaya Patil[4], Dhara Patel[5]
[1,2,3,4,5] Department of Computer Engineering,
Atharva College of Engineering,
University of Mumbai, Mumbai-400095, India.

***Abstract*—Now a days sharing food related photos on social media has become a trend and people are looking for the interested food dishes and the restaurants. So detecting the food items, classifying them and analyzing have been the topic of in-depth studies for various applications related to food recognition, eating habits and dietary assessment. This paper gives a broad study of food recognition techniques. It also focuses on the various feature extraction methods as well as the classification techniques. This paper also gives a brief survey of the datasets available for the food.**

Keywords— Multiple-food image, region detection, feature extraction.

I. INTRODUCTION

Food-related photos have become popular, due to social networks, food recommendation and dietary assessment systems. Social networking sites are nowadays flooded with Food related photos. For instance, new trend is sharing dining-out experiences on social networks. In fact, people are increasingly interested in discovering and sharing new cuisines, and knowing more about different aspects of the food they consume. Many works on food recognition have been put forward in recent years based on different visual representations [1] [2] [3] [4] [5] [6] [7] [8] [9] [10] most of them are limited to a few food classes in controlled settings. Accurate food recognition from only visual information is still a troublesome task. In contrast to objects, food items are deformable and with high intra-class variability, e.g. diverse cooking styles and seasonings will lead to different appearances of the same food. Moreover, different foods share many ingredients and often differences between some food classes are difficult to detect. Also the difference in appearance and presentation of same dish at various restaurants add to the complexity of recognizing the dish.

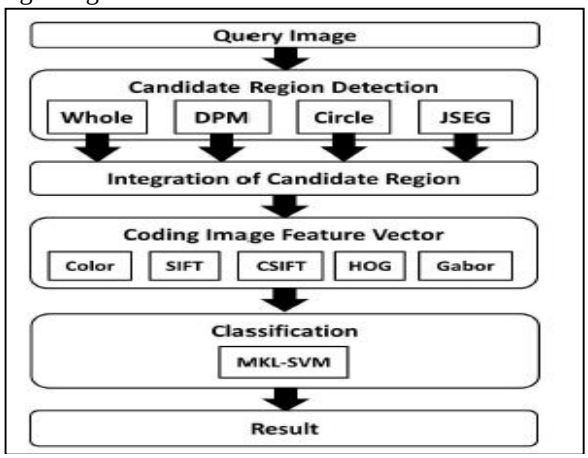

Fig. 1. Block Diagram [4]

Given an input image, first, the framework detects candidate regions of dishes. We use four types of detectors including the deformable part model (DPM) [15], a circle detector, the JSEG region segmentation [16], and whole image. Next, it integrates bounding boxes of the candidate regions detected by the four methods. Then, we check the aspect ratio of width and height of the bounding boxes, and exclude irrelevant bounding boxes regarding their shapes from the candidate set. The framework extracts various kinds of image features from the selected regions using Histogram of Oriented Gradients (HoG), and Gabor texture features ,Bag-of-Features (BoF) of SIFT and color SIFT with spatial pyramid from the selected regions,and calculate SVM scores by multiple-kernel learning (MKL) [17] Finally, we obtain the name of the dish.

A. Candidate Region Detection

There can be cases where food image may contain more than two food items; also it may happen that along with the food item there are other objects like spoons and some part of table. To eliminate all these, input food image is proceeded for candidate region detection.

The four kinds of candidate region detection methods include whole image, the deformable part Model method (DPM) [15], a circle detector, and the JSEG region segmentation [16]

1) Whole Image

Whole Image candidate region is same as the existing food recognition systems [1][6][7][15] which assume that one image contains only one food item. This candidate probably works well for an image containing one large food item,, but does not suit for an image containing multiple food items.

2) Deformable Part Model (DPM)

DPM uses the sliding window approach. The DPM is a two layered hierarchical model, which consists of a global "root" filter and several part models. Each part model specifies a spatial model and a part filter. The spatial model defines a setoff permitted placements for a part relative to a detection window, and a deformation cost for each placement. The total rating of a detection window is the value of the root filter on the window plus the sum over parts, of the maximum over placements of that part, of the part filter score on the resulting sub window minus the deformation cost. Both root and part filters are scored by computing the dot product between a set of weights and HoG features [18] within a window. In addition, the DPM is defined at a fixed scale, and we detect objects by examining over an image pyramid. Therefore, to reduce computational cost, linear SVM are used in the DPM method.

3) Circle Detector

A circle detector identifies regions of dishes by extracting circular contours from a food image. First, it converts a given image to a gray-scale image. Then, it extracts contours by the Canny Edge Detector. Finally, it detects circles by the Hough transform from extracted contours.

4) Region Segmentation

JSEG divides an image by color space quantization and color class map. JSEG uses the number of segmented regions as a parameter.

B. Feature Extraction

To obtain best results we integrate various kinds of image features [13][14] in the same way as Joutou et al.'s work [6]. In this subsection, we describe the image features including *Bag-of-features of SIFT and CSIFT,* histogram of oriented gradient (HoG), Gabor texture features, and color histograms.

1) Bag-of-features of SIFT and CSIFT

In the scheme of BoF, first, a set of local image points is sampled and visual descriptors are extracted by the Scale Invariant Feature Transform (SIFT) descriptor [13] on every point. In addition to SIFT, we also extract CSIFT [14] which is extracted SIFT from a RGB color space. CSIFT is proved to be adaptable against illumination changes [14]

2) Histogram of Oriented Gradients

Histogram of Oriented Gradients (HoG) was proposed by N. Dalal et al. [18]. It is analogous to SIFT in terms of how to define local patterns which is based on gradient histogram. The difference between HoG and BoF is that BoF completely overlooks location information of key points, while HoG keeps rough location information by constructing histograms for each dense grid and concatenating them as one long feature vector .In short, HoG and BoF have diverse characteristics while both are composed of many local gradient histograms.

3) Gabor texture feature

A Gabor texture feature characterizes texture patterns of local regions with numerous scales and orientations. Before applying the Gabor filters, the given region is divided into 8×8 blocks. 24 Gabor filters are applied to each block, then average filter responses within the block, and obtain a 24-dim Gabor feature vector for each block. At the end simply concatenate all the extracted 24-dim vectors into one 1536-dim vector for each region.

C. Classification for Candidate Region

After extraction of feature vectors from each candidate region, we calculate evaluation values of the candidate region regarding each of all the given categories using support vector machines (SVM) which are trained by multiple kernel learning(MKL).

We use the multiple kernel learning(MKL) [17] to incorporate various kinds of image features. With MKL, we can train a SVM with an adaptively-weighted combined kernel which merges various kinds of image features. By applying trained models for each candidate regions regarding all the categories, we obtain evaluation values for each candidate region. We arrange the evaluation values over all the candidate regions and all the categories in the descending order, and output the top N categories in terms of the evaluation values so that single food category is included in the output food name list only once.

II. LITERATURE REVIEW

Yang et al.[1] proposed a method that calculates pairwise statistics between local features computed over a soft pixel level segmentation of the image into eight ingredient types. These statistics in a multi-dimensional histogram, which are then used as a feature vector for a discriminative classifier. The image is represented as the statistics of pairwise local features, known as pairwise feature distribution (PFD). Pairwise local feature distribution includes Soft labeling of pixels, Global Ingredient Representation (GIR) , Pairwise Features, Histogram representation for pairwise feature distribution, Histogram normalization. The method states that exploiting the spatial characteristics of food, in combination with statistical methods for pixel-level image labeling will enable to develop practical systems for food recognition.

Zong et al.[2] proposes a food image classification method by means of local textural patterns and their global structure to describe the food image. The method uses a visual codebook of local textural patterns is created by employing Scale Invariant Feature Transformation (SIFT) interest point detector using the Local Binary Pattern (LBP) feature. The global structure of the food object is represented as the spatial distribution of the local textural structures and encoded using shape context. By using shape context to represent the relative spatial relationship between codewords, the proposed method can accommodate deformations and transformations in the shape of food objects. But this technique does not incorporate view invariant texture feature.

Kong et al.[3] developed an automatic camera phone based multi-view food classifier named DietCam. DietCam uses probabilistic method to separates every food from multiple images.The recognition accuracy is increased through an enhanced joint distribution from every viewpoint. First for classifying food items from the images, they detect and extract local feature points in every image and classify these features based on an existing feature database .They then increase the recognition accuracy through result verifications from multiple viewpoints. It considers the images are taken by three cameras at a synchronized time. A new technique has been introduced as perspective distance, which reflects the geometric relation between two features concerning their appearances in all the possible perspectives. It shows an accuracy of 84% for regular shape food items

Matsuda et al.[4] proposed a two-step method for recognizing multiple food images by detection region of interest i.e. candidate region using various method and classifying them according to the features extracted. They

detected several candidate region by integrating several results acquired from the region detected using (DPM), a circle detector and the JSEG region segmentation. And then applying feature extraction method like including bag-of-features of SIFT and CSIFT with spatial pyramid (SP-BoF), histogram of oriented gradient(HoG), and Gabor texture features and finally classify them according to SVM score. They estimated ten food candidates for multiple-food images in the descending order of the SVM scores. They we have achieved the 55.8% classification rate.

Y. Kawano et al.[5] built interactive and real time food recognition and recording them on user smart phone. First, the user draws bounding box according to the region of interest for more accurate results they segmented the food image by Grub Cut, extracted a color histogram and SURF based bag-of-features. And finally classify them using a liner SVM with a fast $\chi 2$ kernel for fast and accurate food recognition. They have achieved 81.55% classification rate for the top 5 candidates when ground-truth bounding boxes are given.

Joutou et al.[6] proposed food recognition systems which are 50 kinds of common food items in Japan. They have proposed a method for recognizing food images by integrating various kinds of image features including SIFT-based bag-of-features, Gabor, and color histograms and classify it into one of the given food categories with the trained MKL-SVM and they have achieved the 61.34% classification rate.

H. Hoashi et al.[7] discussed new image features and food categories . As new image features added are, gradient histogram which can be regarded as a simple version of Histogram of Oriented Gradient (HoG), and added 35 new food categories to the existing 50 categories in the existing system. They have also described the proposed method based on feature fusion of various kinds of image features with MKL

Y. Maruyama et al.[8] consists of work based on "FoodLog" system, where the user takes photos of the foods and uploads them to the system, and the system performs image processing to detect food images and determine the food balance. The foodlog system allows the user to correct the results of the system. They proposed a method to make use of the corrections made by the user by Naive Bayes which is one of the Bayesian networks It also investigates how to improve the accuracy by using user feedbacks. First, we compare the accuracy of the performance between SVM and Naive Bayes Then, making use of the user's corrections as feedback, the Bayesian network is updated to improve the performance.

Kawano et al.[9] proposed to extension of an existing image dataset automatically leveraging existing categories and crowdsourcing. There are uncountable food categories, since foods are different from a place to a place. Dataset of one cultured food cannot be used for Food detection system of another culture. This enables not only to build other -cultured food datasets based on an original food image dataset automatically, but also to save as much crowd -sourcing costs as possible. Basically, they focused on expansion on food image data set to build food dataset irrespective of food culture.

M. M. Zhang et al.[10] used the idea of attribute-based classification, they classified plates of food to the exact cuisine by the country, using the ingredients as attributes for a plate of food. First recognized the ingredients, gave each ingredient a probability marker, and then used pairwise local features among the ingredients to find out the food category, by calculating the orientation, distance, and other properties between each pair of ingredients. To identify the cuisine using ingredient, attribute-based classification is used. First level includes use of Earth Mover's Distance as the low-level feature. EMD turns out to be problematic for dishes where only one ingredient is present. In the last level, to determine the cuisine category from the attributes, we use the attribute vectors to train the final classifier, with the cuisine category ID as ground truth which is the image's area ratio for the definite ingredient classifier, as extracted from the image's attribute vector & determine at intermediate level. Hence the final categorization of the cuisines is separate from the raw images, as it uses the intermediate high-level attributes layer to predict the results.

Rother et al.[11] extends the graph-cut approach in three respects. First, developed a more powerful, iterative version of the optimization. Second, the power of the iterative algorithm is used to simplify considerably the user interaction needed for a given value of result. Third, a robust algorithm for "border matting" has been developed to estimate and at the same time the alpha-matte around an object boundary and the colors of foreground pixels. Graph-cut is used to accomplish robust segmentation even in camouflage, when foreground and background colors distributions are not well separated. Graph-cut method is failed in terms of user interactions, can occur in three cases: (i) regions of low contrast at the transition from foreground to background (ii) hide, in which the true foreground and background distributions overlap to some extent in colour space (iii) background material inside the user rectangle happens not to be effectively represented in the background region. Graph-cut is successful by where the bounding rectangle on your own is a sufficient user interaction to enable foreground extraction to be accomplished automatically by GrabCut.

Y. Matsuda et al.[12] proposed a method to identify multiple food items from one food image considering co-occurrence statistics with the manifold ranking method. As in traditional method, first candidate regions detected after that image features extraction technique is applied to classify models trained by MKL & obtain the names of the top N food item candidates over the given image. Some familiar combinations of food items such as "hamburger and french-fries" and "rice and miso-soup" & improbable combinations exist such as "sushi and hamburger" or "sashimi and french-fries" are observed. From these observations, co-occurrence statistics is try to enhance the performance of multiple-food image recognition. By considering co-occurrence statistics, we can reduce improbable combinations which are present in the higher ranked candidates. For multi-food recognition with co-

occurrence statistics, we use manifold ranking which is actually a re-ranking method to consider similarities between items.

III. DATASET

The dataset surveyed can be used in the techniques discussed in this paper. Pittsburgh fast-food image data sets contain 61 categories of food items. The dataset UEC FOOD 100 contains 100-kind food photographs. Every food photo has a bounding box indicating the location of the food item in the photo. Most of the food categories in this dataset are popular foods in Japan.

IV. CONCLUSION

In this paper we present the study of recent Techniques development in Food recognition. For feature extraction the most commonly used methods include SIFT, Bag of Features, HOG and Gabor . To classify and recognize the food name different types of SVM and MKL are used. The techniques discussed in this paper have achieved accuracy for various categories of food and can be used for wide variety of food recognition and dietary assessment applications.

V. REFERENCES

[1] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar, “Food recognition using statistics of pairwise local features,” in International Conference on Computer Vision and Pattern Recognition, 2010, pp. 2249–2256.

[2] Z. Zong, D. T. Nguyen, P. Ogunbona, and W. Li, “On the combination of local texture and global structure for food classification,” in International Symposium on Multimedia, 2010, pp. 204–211.

[3] F. Kong and J. Tan, “Dietcam: Regular shape food recognition with a camera phone,” in International Conference on Body Sensor Networks, 2011, pp. 127–132.

[4] Y. Matsuda, H. Hoashi, and K. Yanai, “Recognition of multiple-food images by detecting candidate regions,” in International Conference on Multimedia and Expo, 2012, pp. 25–30.

[5] Y. Kawano and K. Yanai, “Real-time mobile food recognition system,” in International Conference on Computer Vision and Pattern Recognition Workshops, 2013, pp. 1–7.

[6] T. Joutou and K. Yanai, “A food image recognition system with multiple kernel learning,” in International Conference on Image Processing, 2009, pp. 285–288.

[7] H. Hoashi, T. Joutou, and K. Yanai, “Image recognition of 85 food categories by feature fusion,” in International Symposium on Multimedia, 2010, pp. 296–301.

[8] Y. Maruyama, G. C. de Silva, T. Yamasaki, and K. Aizawa, “Personalization of food image analysis,” in International Conference on Virtual Systems and Multimedia, 2010, pp. 75–78.

[9] ——, “Automatic expansion of a food image dataset leveraging existing categories with domain adaptation,” in European Conference on Computer Vision Workshops, 2014.

[10] M. M. Zhang, “Identifying the cuisine of a plate of food,” University of California San Diego, Tech. Rep., 2011.

[11] Rother, V. Kolmogorov, and A. Blake, “GrabCut: Interactive foreground extraction using iterated graph cuts,” ACM Transactions on Graphics, vol. 23, no. 3, pp. 309–314, 2004.

[12] Y. Matsuda and K. Yanai, “Multiple-food recognition considering co-occurrence employing manifold ranking,” in International Conference on Pattern Recognition, 2012, pp. 2017–2020.

[13] Shirke S, Pawar S, Shah K (2014) Literature review: model free human gait recognition. In: 2014 fourth international conference on communication systems and network technologies (CSNT). IEEE, pp 891–895.

[14] Swati, Shirke, and Suvarna Pansambal. "Enhancement of IRIS recognition using Gabor over FFBPANN." Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on. IEEE, 2015.

[15] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan,“Object detection with discriminatively trained part-based models,IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, pp. 1627–1645, 2010.

[16] Y. Deng and B. S. Manjunath, “Unsupervised segmentation of colortexture regions in images and video,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 8, pp. 800–810, 2001.

[17] S. Sonnenburg, G. R¨atsch, C. Sch¨afer, and B. Sch¨olkopf, “Large Scale Multiple Kernel Learning,” The Journal of Machine Learning Research,vol. 7, pp. 1531–1565, 2006.

[18] N. Dalal, B. Triggs, I. Rhone-Alps, and F. Montbonnot, “Histograms of oriented gradients for human detection,” in Proc. of IEEE Computer Vision and Pattern Recognition, pp. 886–893, 2005.