Food recognition and calorie extraction using Bag-of-SURF and Spatial Pyramid Matching methods

Nayan Kumar Konaje
Computer Science department
Stanford University
450 Serra Mall, Stanford, CA 94305
Email: nayankk@stanford.edu

***Abstract*—In this paper we propose methods to recognize food and estimate calorie using Computer Vision algorithms. Specifically, we propose SURF based bag-of-features and spatial pyramid approaches to recognize the food items. In our experiments, we have achieved upto 86% classification rate on a smaller image dataset of 6 categories. We also experimented with larger PFID dataset containing around 111 food item categories and obtained around 18% classification rate. This trained model can be ported to mobiles platforms such as Android or iOS for real-time recognition purpose.**

I. Introduction

According to a study conducted by Centers for Disease Control and Prevention [11], around 1/3rd of American adults are seriously overweight or obese. Key factor in obesity control is diet management. It is important to understand the calorie and nutritional facts of the food items that we consume for proper diet management and to help us keep physically fit. During the recent years, fitness related mobile applications are getting popular among the people, signifying the increase in consciousness about fitness. Most of these mobile applications automatically record physical activities carried out and calories burnt each day, however, they lack capability to automatically record calorie intake each day.

In this paper, we propose methods to automatically detect food item and estimate calorie from a given image of food item. Such model can then be ported to mobile devices, therefore it can serve as a way to automatically record calorie intake. Also, such system can be used in health-care industry to monitor the patients diet habits. Integrating such system in wearable devices such as Google Glass would further ease recognition and recoding of food items.

Food recognition is a challenging task. First, there are large number of food categories. Building a dataset of all categories of food by itself is a challenging task. Second, there can be significant intra-class variations in the observed food items. Same found food can have multiple visual appearance. Finally, presence of occlusions around food items adds extra complexity for its recognition, same food might be served on a bowl or wrapped within a paper cover.

We propose two methods to recognize food items: Speeded Up Robust Features[1] (SURF) descriptor based Bag-of-features model and Spatial Pyramid Matching[2] (SPM) based method. In bag-of-SURF method, we first build a dictionary of codewords, then generate a histogram of codewords for all training images and use linear kernel classification scheme. Spatial pyramid matching method tries to account for the spatial information by dividing and subdividing the given food image and constructing the histogram of codewords of individual regions. We then train a classifier with spatial pyramid kernel using libsvm[3] package. Recognition using proposed system is shown in figure 1.

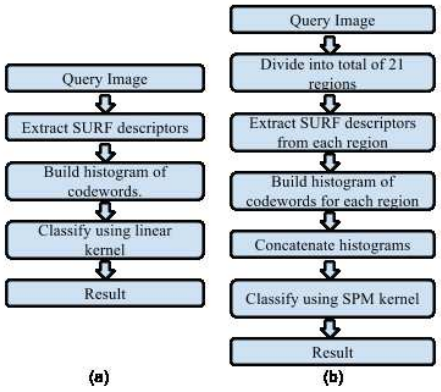

Fig. 1: Recognition using (a) Bag-of-SURF model and (b) Spatial Pyramid Matching method

The rest of this paper is organized as follows: Section 2 describes the related work. Section 3 describes proposed method in detail. Section 4 presents food recognition results on smaller dataset and on a larger Pittsburgh Fast-food Image Dataset[4] (PFID). Finally, we conclude this paper in section 5.

II. Related work

Several papers have been presented to solve the problem of food recognition. Chen et al[4] introduced the PFID dataset for food recognition along with benchmarks using two baseline algorithms: color histogram and bag-of-SIFT. Yang et al[5] proposed food recognition using statistics of pairwise local feature matching approach. Kawano[6] proposed real-time food recognition system using bag-of-SURF method. Matsuda[7] proposed method for recognition multiple-food images, in

which food regions are detected using several detectors and recognition is carried out using multiple kernel learning(MKL) by extracting multiple features such as color, texture, gradient, and SIFT. Shroff[8] proposed a wearable computing system to recognize food for calorie monitoring.

Our goal was to provide benchmark for computer vision researchers who are working on this area rather than to propose such system as state of the art. Baseline results as published in PFID[4] is 9.2% accuracy using bag-of-SIFT method and 11.3% using color histogram approach. In our experiments, we achieved 86% of classification rate on smaller dataset and 18.04% of classification rate on larger PFID dataset. Food recognition method using statistics of pairwise local feature matching[5] achieves upto 28.2% classification accuracy, so far the best on PFID dataset.

III. Bag-of-SURF and Spatial Pyramid Matching

This section provides the technical details of the proposed methods. We first describe SURF, local feature that we use in our methods and then propose two methods to classify food-items: bag-of-SURF and spatial pyramid matching approach.

A. SURF

As local features, we use SURF[1] descriptors in this experiment. Partly inspired by SIFT detector, SURF is a robust local feature descriptors which is used in many computer vision tasks such as object recognition, 3D reconstruction etc. Like SIFT, SURF is scale and rotation invariant. SURF features vector can be either 64 or 128 dimension in length, 64 dimension descriptor is widely used in real-time application as it is known to perform faster with less memory overhead. Figure 2 shows the key points obtained using SURF detector.

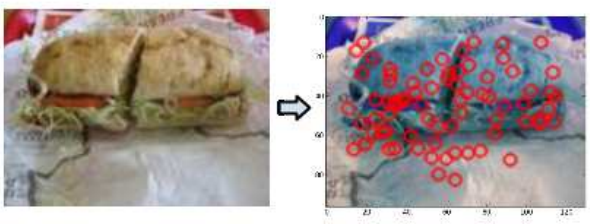

Fig. 2: SURF keypoints extracted from a sample image. Original image is shown on the left. Image on the right shows the extracted SURF keypoints

B. Bag-of-SURF method

First step in bag-of-SURF method is to build a dictionary of codewords. We use a subset of training images to prepare the codeword dictionary. For each of the image of this subset, we extract 64 dimension SURF descriptors. Descriptors from all images belonging to this subset are then plotted on a high-dimensional space and clustered using k-means algorithm into 200 clusters, which is the size of our dictionary. We also experimented using dictionary of size 100 and 300, dictionary size of 200 performs the best in our experiment. Next, we train a classifier on a set of training images. For this, we extract 64 dimensional SURF descriptors from each of the training images. By means of euclidean distance metric, we then compute the nearest neighbor of each of the feature vector in dictionary of codewords and build a histogram of codewords of length 200. This histogram is normalized so that it is probability density. Finally, we train a classifier using libsvm package using linear kernel. We also experimented using chi-square kernel, result of which are documented in section 5.

Testing procedure is also very similar. For all the test images, we get the SURF descriptors, build a histogram of codewords, normalize the histogram and predict the output using the model learned during training section. This approach is shown in figure 3.

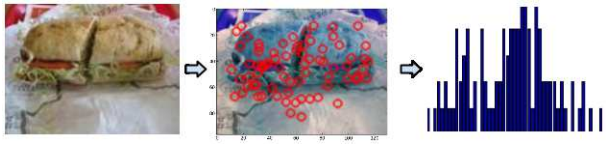

Fig. 3: Food recognition using Bag-of-SURF method

C. Spatial Pyramid Matching method

Spatial Pyramid Matching (SPM) technique works by partitioning the image into increasingly smaller sub-regions and computing histograms of local features found inside each sub-region. By doing this, we embed the spatial information of scene in our histogram. Spatial pyramid matching method is depicted in figure 4. Similar to bag-of-SURF method, we begin

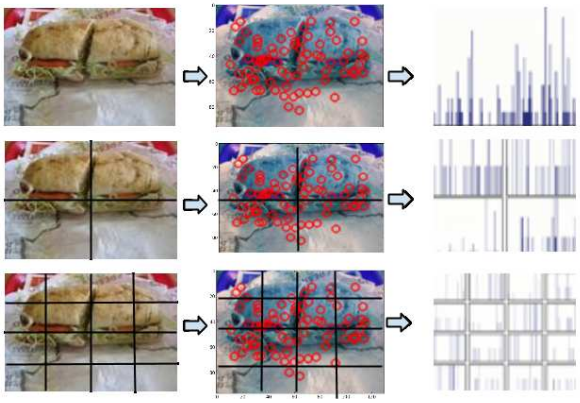

Fig. 4: Spatial Pyramid Matching approach. Image is partitioned into increasingly smaller sub-regions and histogram of codewords is computed for each region

by building dictionary of codewords by extracting the SURF feature vectors from a subset of training images, then we plot the descriptors in 64 dimensional space and cluster them into 200 clusters using k-means algorithm. In our experiment, we employed 3 level SPM: level 0, level 1 and level 2. Level 0 represents the whole image, we get level 1 by partitioning the image into 4 sub regions and finally level 2 is constructed by partitioning into 16 sub regions. In total we will have 21 regions. We then extract SURF descriptors for each of the regions and then a histogram of codewords. Histograms of all sub-regions are concatenated together to form a long feature

vector of dimension 21*200 = 4200. Finally this histogram is normalized so that it is probability density.

We use Spatial Pyramid Matching Kernel[2] to train a classifier using libsvm package. SPM kernel is defined as follows. Let X and Y be the two histogram of codewords. Histogram intersection kernel is defined as,

$$K^L(X,Y) = \frac{1}{2^L} I^0 + \sum_{l=1}^{L} \frac{1}{2^{L-l+1}} I^l$$

where L is the highest level of SPM levels and histogram intersection kernel I at each level is defined as,

$$I(h_1, h_2) = \sum_{i=1}^{N} min(h_1(i), h_2(i))$$

In our experiments, L=2.

Similarly, during testing we divide the test image into 21 sub regions, SURF feature vectors from each of the sub-region is extracted. Histogram for each of the sub-regions is computed, then concatenated and normalized. We then predict the output using learnt model.

IV. Experiments

We experimented proposed food recognition system on two datasets: a smaller dataset of 6 food categories collected by us and larger PFID dataset of 111 food categories. Results on each of the datasets are presented in this section.

A. Smaller dataset

Smaller dataset contained around 6 food categories. These categories are shown in figure 5. Note that food categories are grouped into one of the 6 major food categories, for e.g, Subway Veggie Sub and Subway Chicken Sub are grouped into category Sub in this dataset. This dataset contained total of 53 training images and 15 test images.

Fig. 5: Food images of smaller dataset

Classification accuracy is presented in figure 6. Bag-of-SURF method with linear kernel resulted in classification accuracy of 80%. Both Bag-of-SURF using chi-squared kernel and SPM based methods resulted in classification accuracy of 86%. We experimented with dictionary size of 100, 200 and 300, dictionary size of 100 performs the best in our experiment on this dataset.

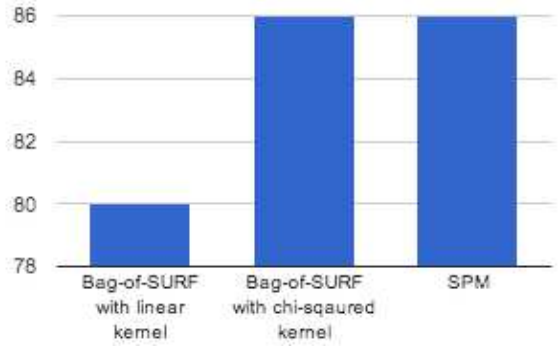

Fig. 6: Classification accuracy on small dataset

B. Larger PFID dataset

We evaluated our algorithms on larger Pittsburgh Fast-food Image dataset. Current version of PFID dataset has total of around 2982 images of fast food belonging to 111 categories. These images were captured under lab and realistic conditions. Images from 13 restaurants chains were included in this dataset.

Each category of this dataset had 3 instances of test images. 3/4th of number of images in all 3 instances per category of images were used for training purpose and remaining images were used for testing purpose. Specifically, we used 2145 images for training and 837 images for testing. Total of 290 images from a subset of training images were used for building dictionary of codewords. Also, we experimented with dictionary of size 100, 200, 256 and 300, dictionary of size 256 performed the best in our experiment. Classification accuracy obtained using our algorithms along with result published using two standard baseline algorithms: Bag-of-SIFT and Color histogram is presented in figure 7.

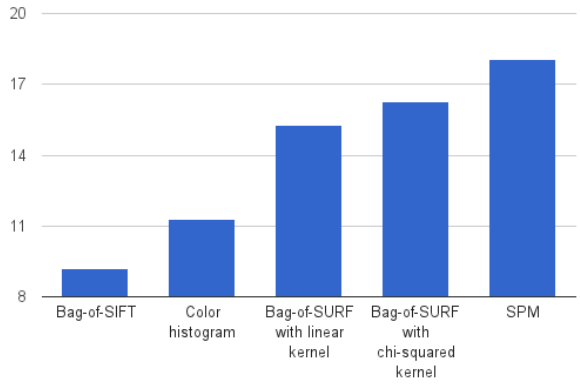

Fig. 7: Classification accuracy on PFID dataset

To summarize, bag-of-SURF method with linear kernel classification scheme resulted in 15.29% of classification rate. If we replace linear kernel with chi-squared kernel, accuracy of bag-of-SURF boosts to 16.25%. Spatial Pyramid Matching based approach shows the best accuracy of 18.04%. Baseline accuracy as published in PFID is 9.2% using bag-of-SIFT

descriptors and 11.3% using color histogram. Note that it might not be accurate to compare the baseline results with our results, as baseline results were published for 61 categories compared to 111 categories that we use in our experiments.

One of the factor attributing to the low accuracy of recognition rate is due to the intrinsic nature of food items in which many food items with similar visual appearances are assigned to different categories in the PFID food database, as shown in figure 8. With the advancement of computer vision algorithms, we hope to achieve better classification results.

Fig. 8: Similar food items in PFID: Visual appearance of subway veggie delight sandwich(left) and subway turkey breast sandwich(right) are same, however they are labelled different in PFID

V. Conclusion

Given the number of categories of food items and intra-class variations within each class, food recognition is a challenging task. With the increase in popularity of fitness applications and advancements of wearable devices such as Google Glass, exploration of food recognition methods are growing. While we do not claim that the methods presented here are state of the art, we achieved significant improvement over the baseline methods. However, food recognition using statistics of pairwise local features[5] remains the best approach so far with the classification rate of 28.2% on PFID dataset.

In future work, we plan to extend our work to: (1) use object bank approach for food item classification, (2) port the trained model to mobile devices for real-time recognition purpose.

References

[1] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. "SURF: Speeded Up Robust Features". In ECCV, 2006.

[2] S Lazebnik, C Schmid, J Ponce. "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories". Proceedings of CVPR, 2006.

[3] Chih-Chung Chang and Chih-Jen Lin. "LIBSVM: A Library for Support Vector Machines". ACM Transactions on Intelligent Systems and Technology, Volume 2 Issue 3, April 2011.

[4] Mei Chen, Kapil Dhingra , Wen Wu, Lei Yang, Rahul Sukthankar, Jie Yang. "PFID: Pittsburgh Fast-food Image Dataset". Proceedings of International Conference on Image Processing, 2009.

[5] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar. "Food recognition using statistics of pairwise local features". In Proc. of IEEE Computer Vision and Pattern Recognition, 2010.

[6] Yoshiyuki Kawano and Keiji Yanai. "Real-time Mobile Food Recognition System". In CPVR 2013.

[7] Y. Matsuda, H. Hoashi, and K. Yanai. "Recognition of multiple-food images by detecting candidate regions". In Proc. of IEEE International Conference on Multimedia and Expo, pages 15541564, 2012.

[8] G. Shroff, A. Smailagic, and D. Siewiorek. "Wearable context-aware food recognition for calorie monitoring". In Proceedings of International Symposium on Wear- able Computing, 2008.

[9] L.J. Li, H. Su, E. Xing, and L. Fei-Fei, Object bank: A high-level image representation for scene classication and semantic feature sparsication *Advances in Neural Information Processing Systems*, 2010.

[10] D. G. Lowe, Distinctive image features from scale- invariant keypoints, *Int. J. Computer Vision*, vol. 60, no.2, pp. 91110, 2004.

[11] Cynthia L. Ogden, Margaret D. Carroll, Brian K. Kit and Katherine M. Flegal. "Prevalence of Obesity in the United States". Internet:http://www.cdc.gov/nchs/data/databriefs/db82.pdf