CLASSIFYING FOOD IMAGES REPRESENTED AS BAG OF TEXTONS

Giovanni Maria Farinella *Marco Moltisanti* *Sebastiano Battiato*

Department of Mathematics and Computer Science
Image Processing Laboratory - University of Catania
{gfarinella, moltisanti, battiato}@dmi.unict.it

ABSTRACT

The classification of food images is an interesting and challenging problem since the high variability of the image content which makes the task difficult for current state-of-the-art classification methods. The image representation to be employed in the classification engine plays an important role. We believe that texture features have been not properly considered in this application domain. This paper points out, through a set of experiments, that textures are fundamental to properly recognize different food items. For this purpose the bag of visual words model (BoW) is employed. Images are processed with a bank of rotation and scale invariant filters and then a small codebook of Textons is built for each food class. The learned class-based Textons are hence collected in a single visual dictionary. The food images are represented as visual words distributions (Bag of Textons) and a Support Vector Machine is used for the classification stage. The experiments demonstrate that the image representation based on Bag of Textons is more accurate than existing (and more complex) approaches in classifying the 61 classes of the Pittsburgh Fast-Food Image Dataset.

***Index Terms*—** Food Classification, Bag of Words, Textons

1. INTRODUCTION AND MOTIVATIONS

There is a general consensus on the fact that people love food. Thanks to the great diffusion of low cost image acquisition devices (e.g., smartphones), the food is nowadays one of the most photographed objects; the number of food images on the web is increasing and novel social networks for food lovers are more and more popular.

Automatic food classification is an emerging research topic, not only to recognize food images for the web and social networks application domain (e.g., for advertising purposes). Indeed, researchers (in different fields) study food because of its importance under medical, social and anthropological point of view. Food images can provide a wider comprehension of the relationship between people and their meals. Hence, automatic food classification can be useful to build diet monitoring systems to combat obesity, by providing to the experts (e.g., nutritionists) objective measures to assess the food intake of patients [1, 2]. On the other hand, food classification is a difficult task for vision systems, and offer an exciting challenge for computer vision researchers. Food is intrinsically deformable and presents high variability in appearance; classic approaches used to classify images perform very poorly on food images [3].

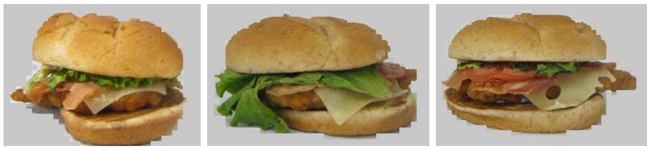


Fig. 1. Three different instances of the same food in the PFID dataset [3].

Several works have addressed the problem of food classification [2, 3, 4, 5, 6, 7]. As any emerging research topic, most of the works propose, along with the classification algorithm, a new dataset composed by various food classes. So, despite many approaches have been published, it is difficult to find papers where different techniques are compared on the same dataset. This makes difficult to understand which are the peculiarities of the different techniques and which is the best method for food classification so far. For this reason, we have tested our method on an existing and public food dataset with clear testing protocol (i.e., the Pittsburgh Fast-Food Image Dataset - PFID [3]) on which different state-of-the-art approaches have been tested [2, 3].

One of the first food classification method have been proposed by Jimenez *et al.* [4]. The authors proposed a method able to detect spherical fruits (e.g., oranges) in natural environment. To this purposes they used range images, obtained via a 3D laser scanner. Joutou *et al.* [5] used a Multiple Kernel Learning SVM (MKL-SVM) to exploit different kinds of features. They combined Bag-of-SIFT with Color Histograms and Gabor Filters to discriminate between images of a dataset composed by 50 different food categories. Matsuda *et al.* [6, 7] introduced a new dataset with food images belonging to 100 classes. In [6] they employed Bag-of-SIFT on Spatial Pyramid, Histograms of Gradient, Color Histogram and Gabor Filters to train a MKL-SVM after the detection of candidate regions based on Deformable Part Models. In [7] they extended their previous work including a ranking algorithm to be used for image retrieval purpose.

As aforementioned, a public available benchmark dataset for food classification is the Pittsburgh Fast-food Image Dataset (PFID) [3]. This dataset is composed by 1098 food images belonging to 61 different categories. Each food class contains 3 different instances of the food (i.e., same food

978-1-4799-5751-4/14/$31.00 ©2014 IEEE

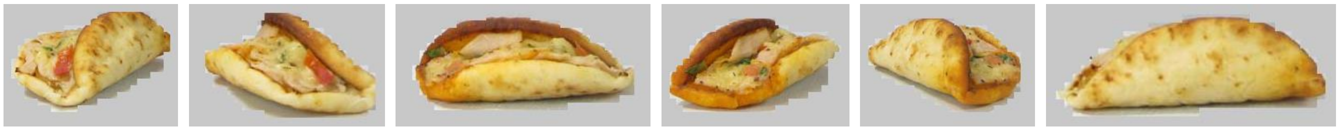


Fig. 2. Six different point of view of one instance of food in the PFID dataset [3].

class but acquired in different days and in different restaurants - see Fig. 1), and 6 images of different viewpoints for each instance (see Fig. 2). The main contribution of [3] is the dataset itself. The authors provided both the benchmark dataset and the evaluation protocol for classification and comparison purposes. As a baseline, in [3] are reported the food classification results by employing representations based on Color Histograms and Bag-of-SIFT, coupled with linear SVM.

Considering the PFID dataset, Yang *et al.* [2] outperformed the baseline results using the statistics of pairwise local feature in order to encode spatial relationship between different ingredients. As first step, the Semantic Textons Forest (STF) [8] approach is used to assign a soft label (distribution over ingredients) to each pixel in the image. Eight basic ingredients categories have been considered: beef, chicken, pork, bread, vegetable, tomato/tomato sauce, cheese/butter, egg/other. Starting from the semantic segmentation of the image, the authors computed and tested several features to demonstrate the usefulness of encoding spatial relationships of ingredients. Among the tested features, the best results have been obtained by employing the so called OM features. Employing these features the authors of [2] outperformed both the baseline results presented in [3], as well as the global ingredient representation (GIR) approach based on statistics of food ingredients collected after semantic segmentation with STF [2, 8]. The OM features encode the information of the soft labeling (obtained with the STF) considering two spatial positions of the food images. Moreover, this local descriptor encodes the direction of the spatial co-occurrences and the information of the soft labeling of the midpoint among them.

Many of the aforementioned food recognition approaches use a combination of different features [2, 5, 7, 6]. By exploiting multiple features it is possible to capture different aspects of food appearance (e.g., color, shape, spatial relationships) and hence improve the recognition accuracy. Although a number of food classification techniques have been presented in literature, we believe that texture features have been not properly considered in this application domain. Looking at images of food (see Fig. 1 and 2) it is straightforward the association of food classification to a problem of texture discrimination. Differently than one can expect, classic approaches for texture classification haven't been taken into account as a baseline for comparison purpose with respect to novel food classification techniques.

In this paper we demonstrate that textures are fundamental to properly classify different food items. Bag of Textons model [9, 10] is employed to this aim. Images are processed with the Maximum Response Filter Banks (MR) [9]. The maximum response is taken on both orientations and scale of the different filters to achieve invariance to these transformations. Hence, a small codebook of Textons [10, 11, 12] is built for each food class. Then, the learned class-based Textons are collected in a single visual dictionary and the food images are represented as visual words distributions (Bag of Textons). Finally, a Support Vector Machine is used for classification purpose. To the best of our knowledge, Textons have never been exploited for food classification. The experiments reported in Section 3 point out that the Bag of Textons representation is more accurate in recognizing food classes than existing (and more complex) approaches [2, 3].

The remainder of this paper is structured as follows: Section 2 presents the proposed approach to build the representation of food images, whereas in Section 3 the experimental settings and the results are described. Finally, Section 4 concludes the paper with hints for further works.

2. BAG OF TEXTONS BASED CLASSIFICATION

The Bag-of-Visual-Word paradigm (BoW) [13] is one of the most used method to represent images for classification purpose. Four main steps are involved in representing images: feature detection, feature description, codebook generation and image representation. Each of these four steps introduces a variability on the final model used to represent the images, and influences the overall pipeline as well as the results of the classification. Different local feature descriptors can be exploited to generate the codebook. For instance, in [3] SIFT has been used to test BoW paradigm on the PFID dataset. Among the other descriptors, Textons [11] have been employed when the content of the images is rich of textures [9, 10, 12, 14]. Since textures are one of the most important aspects of food images, here we treat the classification of food as a texture classification problem. In the learning stage, training images are convolved with a filter bank to compute filter responses. This feature space is quantised via K-Means clustering and the obtained clusters prototypes (i.e., the visual vocabulary) are used to label each filter response (i.e., each pixel) of the training images. The distribution of Textons is then used to feed the SVM classifier and hence to build the model to be used for classification purpose. During classification phase, test images are represented as distribution on the pre-learned Textons vocabulary after filter bank processing. Each test image, represented as Bag of Textons, is then classified accordingly with the previous learned SVM model. In our experiments we use the Maximum Response filter bank [9] which is composed by filters (Gaussian, first and second derivative of Gaussian and Laplacian of Gaussian) computed at multiple orientation and scales. To achieve rotational and scale invariance, the responses of the anisotropic filters are recorded at the maximum response on both scales and orientations (MRS4 filters). In this way, a very compact 4-dimensional vector for each color channel is associated to every pixel of the food images. As suggested in [9], filters are L_1 normalised so that the filter responses lie approximately in the same range. To achieve invariance to the global affine transformation of the illumination, the intensity of the images is normalised (i.e., zero mean and unit standard deviation on each color channel) before the convolution with the MRS4 filter bank. Finally, the filter response **r** at each pixel is contrast normalised as formalised in the following:

$$\mathbf{r}_{final} = \frac{\mathbf{r}\left[log\left(1 + \frac{||\mathbf{r}||_2}{0.03}\right)\right]}{||\mathbf{r}||_2} \tag{1}$$

Regarding the Textons vocabulary generation, differently than the classic procedure where the feature descriptors extracted from all training images of the different classes are quantized all together, here we consider a class-based quantization [9]. First, a small codebook D_c with K_c Textons is built for each food class c. Then, the learned class-based Textons vocabularies are collected in a single visual dictionary $D = \bigcup_c D_c$ of cardinality $K = \sum_c K_c$, and the food images are represented as visual words distributions considering the vocabulary D. The rationale beyond this codebook generation is similar to the one presented in [15]. Each class-based Textons vocabulary is considered suitable to encode textures of a specific class of food and not suitable to encode the textures of the other classes; this is reflected in the image representation in which all the class-based vocabularies are collected in a single codebook D. Intuitively, when an image of class c is encoded as Textons distribution considering the final vocabulary D, the bins of the sub-vocabulary D_c are more expressed than the bins related to the other sub-vocabularies $D_{c'}$, $c' \neq c$, making the representation more discriminative. The experiments reported in Section 3 show that, considering the PFID dataset, the class-based Textons representation achieve better results than the one learned without considering the different food classes during the codebook generation.

For classification purpose, we use a multiclass SVM with a pre-computed kernel by considering the cosine distance. Given two Bag of Textons signatures S_{I_i}, S_{I_j}, the cosine distance d_{cos} is calculated as following:

$$d_{cos}\left(S_{I_i}, S_{I_j}\right) = 1 - \frac{S_{I_i} S_{I_j}^{'}}{\sqrt{\left(S_{I_i} S_{I_i}^{'}\right)\left(S_{I_j} S_{I_j}^{'}\right)}}. \tag{2}$$

The kernel is defined as:

$$k_{cos}\left(S_{I_i}, S_{I_j}\right) = e^{-d_{cos}\left(S_{I_i}, S_{I_j}\right)}. \tag{3}$$

3. EXPERIMENTAL SETTINGS AND RESULTS

The proposed method have been compared against the techniques reported in [2, 3] on the PFID dataset [3]. As in [2, 3], we follow the experimental protocol defined for the PFID dataset [3]: 3-fold cross-validation using 12 images from two instances of each class for training, and the 6 remaining images of the third instance of each class for testing. We employed the libSVM library [16] to assess the class-based Bag of Textons representation described in the previous section.

Table 1. Class-based vs Global Textons Vocabularies. In all settings class-based vocabulary achieve better results.

Vocabulary Size	610	1220	1830	2440
Class-Based Textons	27.9 %	29.1 %	29.4 %	31.3%
Global Textons	23.1%	25.3%	26.0%	26.2%

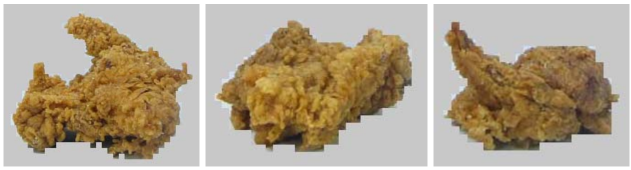
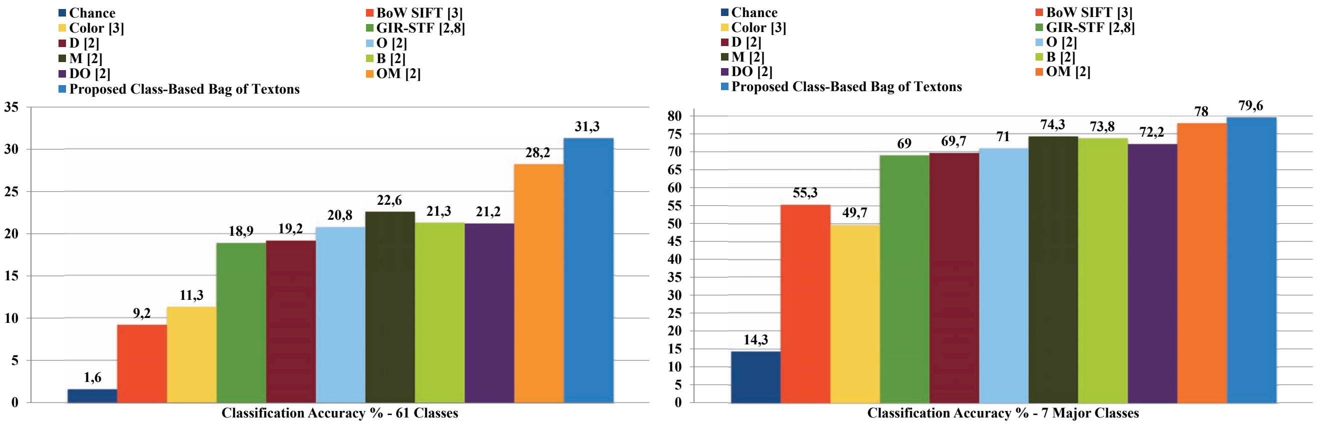


Fig. 3. Three different classes of the PFID dataset [3]. Left: Crispy Chicken Breasts. Middle: Crispy Chicken Thighs. Right: Crispy Whole Chicken Wing

As pointed out in [2], many foods items of the PFID dataset have very similar appearances despite they belong to different classes. For instance, in Fig. 3 different type of chicken are considered as belonging to different classes, but their discrimination is very difficult even for humans. Following the testing protocol in [2], we have also performed tests by re-organizing the 61 PFID food categories into seven major groups: Sandwiches, Salads & Sides, Chicken, Breads & Pastries, Donuts, Bagels, and Tacos.

As first test, we have compared the class-based Textons vocabulary with respect to the global one, i.e., the one obtained considering all the feature descriptors of the different classes all together during quantization. Table 1 reports the results in terms of accuracy at varying of the vocabulary size for the classification of the 61 classes of the PFID dataset. The size of the vocabulary has been fixed by considering the number of class-based Textons K_c to be learned for each food class. We have considered $K_c \in \{10, 20, 30, 40\}$ Textons for each class c, corresponding to a final vocabulary size of $K \in \{610, 1220, 1830, 2440\}$. As expected, increasing the number of Textons, the classification accuracy improve. Nevertheless, we do not have further improvements by considering more than 40 Textons per class. Note that the class-based vocabulary achieve better results in all cases.

The comparison of the class-based Bag of Textons representation (with $K_c = 40$) against to the others state-of-the-art methods [2, 3] is shown in Fig. 4(a) and Fig. 4(b) for both the 61 classes and the 7 major classes respectively. The names of the different methods are related to the original name used by the authors in their papers [2, 3]. The chance recognition rate is also indicated. The classification accuracy of the class-based Bag of Textons representation was 31.3% for the 61 classes and 79.6% for the 7 major classes. Although its simplicity, the class-based Bag of Textons representation achieve much better results ($> 20\%$) than the global BoW considering SIFT descriptor. It also outperforms the method proposed in [2] where OM features encoding spatial informa-



(a) 61 Classes

(b) 7 Major Classes

Fig. 4. Comparison of the different approaches on the of the PFID dataset [3].

Table 2. Accuracy of the different methods on the 7 Major Classes of the PFID dataset [3] (i.e., diagonal of the obtained confusion matrices). Since the number of images belonging to the different classes are not balanced, for each class we report the per class accuracy percentage together with the corresponding number of images (within parenthesis).

Images per class	**Sandwich**	**Salad & Sides**	**Bagel**	**Donut**	**Chicken**	**Taco**	**Bread & Pastry**	**Average**
On each test run	228	36	24	24	24	12	18	52.3
Per class accuracy % (Number of Images)	**Sandwich**	**Salad & Sides**	**Bagel**	**Donut**	**Chicken**	**Taco**	**Bread & Pastry**	**Average**
Color [3]	69.0 (157.3)	16.0 (5.8)	13.0 (3.1)	0.0 (0)	49.0 (11.8)	39.0 (4.7)	8.0 (1.4)	27.7 (26.3)
BoW SIFT [3]	75.0 (171)	45.0 (16.2)	15.0 (3.6)	18.0 (4.3)	36.0 (8.6)	24.0 (2.9)	3.0 (0.5)	30.9 (29.6)
GIR-STF [2, 8]	79.0 (180.1)	79.0 (28.4)	33.0 (7.9)	14.0 (3.4)	73.0 (17.5)	40.0 (4.8)	47.0 (8.5)	52.1 (35.8)
OM [2]	86.0 (196.1)	**93.0 (33.5)**	40.0 (9.6)	17.0 (4.1)	**82.0 (19.7)**	65.0 (7.8)	**67.0 (12.1)**	64.3 (40.4)
Class-Based Bag of Textons	**87.6 (199.7)**	84.3 (30.3)	**70.8 (17)**	**43.1 (10.3)**	66.7 (16)	**69.4 (8.3)**	53.7 (9.7)	**67.9 (41.6)**

tion are used after a semantic segmentation performed trough STF [8]. It is important to note that, differently than [2], Textons based representation does not require any manual labeling of the different ingredients composing the food items to be employed. Although the labeling of the different food ingredients is possible for a small set of plates, the up-scaling to a huge number of categories (composed by many ingredients) became not feasible, making the approach described in [2] difficult to be applied. The experiments point out that a proper encoding of textures play an important role for food classification. Note that, even considering only a few Textons per class (i.e., 10 Textons for a total of 610 visual word - see Table 1 and Fig. 4(a)) the accuracy obtained by the proposed method on the 61 classes (27.9%) outperforms the ones achieved by other methods and is very close to a more complex food classification pipeline described in [2] (28.2%).The proposed representation outperform all the others methods with a number of class-based Textons $K_c \geq 30$. In Table 2 are reported the accuracies of the different methods on the 7 major classes of the PFID dataset. Since the number of images belonging to the different classes are not balanced, for a better understanding of the results, the number of images is reported together with the per-class accuracy. Also in the case of 7 major classes the average per-class accuracy is in favour of the Textons based representation.

4. CONCLUSIONS AND FUTURE WORK

This paper evaluates the class-based Bag of Textons representation in the context of food classification. The MRS4 filter banks are used to build class-based Textons vocabularies. The image representation is coupled with a Support Vector Machine for classification purpose. This representation is compared with respect to other state-of-the-art methods on the public available Pittsburgh Fast-Food Image Dataset (PFID). The class-based Bag of Textons representation obtained better results with respect to all the other methods. Future works could be devoted to the exploitation of Textons (and/or other types of texture-like feature, such as CLBP [17]) in joint with other kind of features [18, 19], as well as in encoding spatial information between local Textons (e.g., through correlograms of textons [20]) to better discriminate food items. Moreover, could be important to test the Textons based representation (both Global and Class-Based) on bigger food image datasets for both classification and retrieval purposes.

5. ACKNOWLEDGEMENTS

We would like to thank the authors of [2] who have provided information on the testing protocol of the PFID dataset and the labeling of the seven major category.

6. REFERENCES

[1] R. Spector, "Science and pseudoscience in adult nutrition research and practice," in *Skeptical Inquirer*, 2009.

[2] Shulin Yang, Mei Chen, Dean Pomerleau, and Rahul Sukthankar, "Food recognition using statistics of pairwise local features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2249–2256.

[3] Mei Chen, Kapil Dhingra, Wen Wu, Lei Yang, Rahul Sukthankar, and Jie Yang, "Pfid: Pittsburgh fast-food image dataset," in *IEEE International Conference on Image Processing*, 2009, pp. 289–292.

[4] Antonio Ramón Jiménez, Anil K Jain, R Ceres, and JL Pons, "Automatic fruit recognition: a survey and new results using range/attenuation images," *Pattern recognition*, vol. 32, no. 10, pp. 1719–1736, 1999.

[5] Taichi Joutou and Keiji Yanai, "A food image recognition system with multiple kernel learning," in *IEEE International Conference on Image Processing*, 2009, pp. 285–288.

[6] Yuji Matsuda, Hajime Hoashi, and Keiji Yanai, "Recognition of multiple-food images by detecting candidate regions," in *IEEE International Conference on Multimedia and Expo*, 2012, pp. 25–30.

[7] Yuji Matsuda and Keiji Yanai, "Multiple-food recognition considering co-occurrence employing manifold ranking," in *International Conference on Pattern Recognition*, 2012.

[8] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[9] Manik Varma and Andrew Zisserman, "A statistical approach to texture classification from single images," *International Journal of Compututer Vision*, vol. 62, no. 1-2, pp. 61–81, 2005.

[10] S. Battiato, G. M. Farinella, G. Gallo, and D. Ravì, "Exploiting textons distributions on spatial hierarchy for scene classification," *Eurasip Journal on Image and Video Processing*, pp. 1–13, 2010.

[11] B. Julesz, "Textons, the elements of texture perception, and their interactions," *Nature*, vol. 290, pp. 91–97, 1981.

[12] L. W. Renninger and J. Malik, "When is scene recognition just texture recognition?," *Vision Research*, vol. 44, pp. 2301–2311, 2004.

[13] G. Csurka, C. Bray, C. Dance, and L. Fan, "Visual categorization with bags of keypoints," 2004.

[14] S. Battiato, G.M. Farinella, G. Gallo, and D. Ravì, "Scene categorization using bag of textons on spatial hierarchy," in *IEEE International Conference on Image Processing*, 2008, pp. 2536–2539.

[15] Florent Perronnin, "Universal and adapted vocabularies for generic visual categorization," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 30, no. 7, 2008.

[16] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: a library for support vector machines," 2001.

[17] Z. Guo, L. Zhang, and D. Zhang, "A completed modeling of local binary pattern operator for texture classification," *IEEE Transactions on Imgage Processing*, vol. 19, no. 6, pp. 1657–1663, 2010.

[18] S. Battiato, G. M. Farinella, G. Puglisi, and D. Ravì, "Aligning codebooks for near duplicate image detection," *Multimedia Tools and Applications*, 2013.

[19] S. Battiato, G. M. Farinella, G. Giuffrida, C. Sismeiro, and G. Tribulato, "Using visual and text features for direct marketing on multimedia messaging services domain," *Multimedia Tools and Applications*, vol. 42, no. 1, pp. 5–30, 2009.

[20] S. Savarese, J. Winn, and A. Criminisi, "Discriminative object class models of appearance and shape by correlatons," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 2033–2040.