Food Recognition Using Consensus Vocabularies

Giovanni Maria Farinella, Marco Moltisanti(✉), and Sebastiano Battiato

Image Processing LAB, Department of Mathematics and Computer Science,
Universitá degli Studi di Catania, Catania, Italy
{gfarinella,moltisanti,battiato}@dmi.unict.it

Abstract. Food recognition is an interesting and challenging problem with applications in medical, social and anthropological research areas. The high variability of food images makes the recognition task difficult for current state-of-the-art methods. It has been proved that the exploitation of multiple features to capture complementary aspects of the image contents is useful to improve the discrimination of different food items. In this paper we exploit an image representation based on the consensus among visual vocabularies built on different feature spaces. Starting from a set of visual codebooks, a consensus clustering technique is used to build a consensus vocabulary used to represent food pictures with a Bag-of-Visual-Words paradigm. This new representation is employed together with a SVM for recognition purpose.

1 Introduction and Motivation

People love food. This fact, coupled with the great diffusion of low cost imaging devices (e.g., wearable cameras, smartphones), makes the food one of the most photographed objects. The analysis of food images can drive to a wider comprehension of the relationship between people and their meals. For instance, automatic food recognition can be useful to build diet monitoring systems to combat obesity, by providing to the experts (e.g., nutritionists) objective measures to assess patients' food intake. Also, food recognition is a challenging and exciting task for computer vision researchers due its both high intra-class and low inter-class variabilities of visual content [16].

Several works have been published in last years addressing the problem of food classification [2–4,7,8,14–18]. Jimenez *et al.* [3] proposed an automatic recognition 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.* [4] 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 considering by 50 different food categories. Matsuda *et al.* [7,8] introduced a new dataset with food belonging to 100 classes. In [7] they employed Bag-of-SIFT on Spatial Pyramid, Histograms of Gradient, color histogram and Gabor filters to train a MKL-SVM after a detection of candidate regions based on Deformable Part Models. The trained models were used to classify multiple instances of

V. Murino et al. (Eds.): ICIAP 2015 Workshops, LNCS 9281, pp. 384–392, 2015.
DOI: 10.1007/978-3-319-23222-5_47

Fig. 1. Three different instances of the same food in the PFID dataset [2].

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

food. In [8] they extended their previous work including a ranking algorithm to be used for image retrieval purpose.

Most of the aforementioned approaches propose, along with a classification engine, a new dataset. However, it is difficult to find papers where different techniques are compared on the same dataset [16]. This makes not easy to fully understand which are the peculiarities of the different techniques and which is the best method for food recognition so far. We consider the Pittsburgh Fast-food Image Dataset (PFID) [2]. This dataset is composed by 1098 food images belonging to 61 different categories. Each class contains 3 different instances of the food (i.e., same food 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 [2] is the dataset itself. As baseline tests, the authors provide the food recognition results by employing Color Histograms and Bag-of-SIFT and a linear SVM classifier. This dataset, as well as the experimental protocol and the results, can be used to properly compare different algorithms.

Considering the aforementioned PFID dataset, Yang *et al.* [14] outperformed the baseline results of [2] using statistics of pairwise local feature in order to encode spatial relationship between different ingredients. As first step, a Semantic Textons Forest (STF) [11] is used to assign a soft label to represent the distribution over ingredients (beef, chicken, pork, bread, vegetable, tomato and tomato sauce, cheese and butter, egg/other). Then the food images are represented with the OM features, which are computed taking into account orientation (O) and midpoint (M) of the segment that connects two previously labelled pixels. These features are hence used with a SVM for classification (using a χ^2 kernel). A successive work considering the PFID dataset for testing purposes [15] pointed out that a classic Bag of Textons approach for texture discrimination outperforms the methods presented in [2] and [14]. The aforementioned approaches proposed in [14] have been considered in the experimental phase of this work for comparison purposes.

It is worth noting that some of the aforementioned food recognition techniques use combination of different features [4,7,8,14]. By exploiting multiple features it is possible to capture different aspects of food appearance (e.g., color,

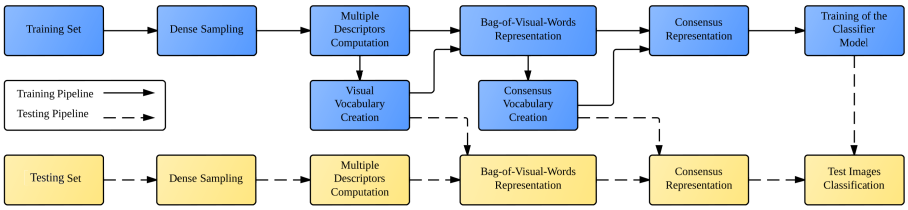


Fig. 3. Proposed image representation pipeline based on Consensus Vocabularies.

shape, texture, spatial relationships) and hence improve the recognition accuracy. On the other hand, image representations based on the alignment of multiple visual vocabularies (Bag-of-Visual-Phrases) built on different feature spaces have been used to address other Computer Vision problems (e.g., near duplicate image retrieval [1,19]).

In this paper we propose a new way to encode different and complementary aspects of local regions, which we call *Consensus Vocabularies*. Starting from vocabularies built on classic SIFT [6] and SPIN [5] features, we use consensus clustering [12] to build a the final vocabulary to be used for the image representation. The proposed representation is coupled with a Support Vector Machine classifier for food recognition. Our method is compared with respect to other state-of-the-art approaches on the PFID dataset [2,14,15].At the best of our knowledge, although consensus clustering is a well known topic in Pattern Recognition, it has not been previously exploited to build visual vocabularies for image representation and recognition purposes.

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

2 Consensus Vocabularies

In this work, we propose to augment the classic Bag-of-Features approach by exploiting a representation based on consensus vocabularies obtained from different visual codebooks computed on different and complementary descriptors. In our representation pipeline, we start by generating different partitions of SIFT [17] and SPIN [5] feature spaces employing a K-means clustering (as in the classic visual vocabularies in the Bag-of-Feature method). The generated partitions are then used to feed the consensus clustering algorithm. In this work we employ the technique proposed in [12]. As output, the consensus clustering gives a vocabulary which is then used to represent each image belonging to both training and testing sets. We use SIFT and SPIN features because they are able to capture different local information. SIFT is useful to capture local gradients [6], whereas SPIN are more powerful in encoding textures [20]. By employing the different descriptors and different runs of the clustering algorithms (i.e., input partitions

for the consensus clustering), we can tune the diversity of the partitions to build the final vocabulary. Of course, different consensus clustering strategies can be used for the proposed method [10,12]. We chosen the algorithm proposed in [12] since it is based on categorical clustering (i.e., can be applied straightforward on pre-built visual vocabularies) and because it does not need to solve the clusters correspondence problem to build the final consensus partition. The consensus partition is obtained as the solution of a maximum likelihood problem using an Expectation-Maximization approach. Figure 3 shows the general pipeline of both training and testing pipelines.

Let $\mathbf{I} = \{I_1, \ldots, I_T\}$ be the training set. Through a dense sampling of each image I_i, we obtain a set of interest points $\mathbf{X}^{(i)} = \{x_1^{(i)}, \ldots, x_{N_i}^{(i)}\}$ on which each feature descriptor is computed. Let $\mathbf{F} = \{F_1, \ldots, F_M\}$ be a set of descriptors. For each descriptor $F_m \in \mathbf{F}$ we compute the set of features $\mathbf{Y}_m^{(i)} = \left\{y_{1,m}^{(i)} = F_m\left(x_1^{(i)}\right), \ldots, y_{N_i,m}^{(i)} = F_m\left(x_{N_i}^{(i)}\right)\right\}$ related to the image $I_i \in \mathbf{I}$. Hence, $\mathbf{Y}_m = \{\mathbf{Y}_m^{(1)} \ldots \mathbf{Y}_m^{(T)}\}$ is the set of all feature descriptors of type m computed over the image set $\mathbf{I}$. Let $\mathbf{C} = \{C_1, \ldots, C_P\}$ be the set of the considered clustering algorithms and let $\boldsymbol{\Theta}^{(p)} = \{\theta_1^{(p)}, \ldots, \theta_{Q_p}^{(p)}\}$ be the set of the parameters considered for the clustering algorithm $C_p \in \mathbf{C}$ (i.e., the parameters corresponding to different runs of a specific algorithm). Each set $\mathbf{Y}_m$ is clustered taking into account a specific clustering algorithm $C_p \in \mathbf{C}$ and considering the set of parameters $\boldsymbol{\Theta}^{(p)}$. This gives as output a set of visual vocabularies for each feature of type m. Let $\mathbf{V} = \{V_1 \ldots V_H\}$ be the set of all vocabularies obtained with the aforementioned procedure. Each set $\mathbf{Y}_m$ is therefore represented with the *ids* related to the visual words belonging to the obtained vocabularies of type m. In this way each interest point $x_n^{(i)} \in I_{(i)}$ is projected into the vocabulary spaces:

$$\begin{array}{c|ccc} Input & V_1 & \ldots & V_H \\ \hline x_1^{(1)} & V_1\left(x_1^{(1)}\right) & \ldots & V_H\left(x_1^{(1)}\right) \\ \vdots & \vdots & \ddots & \vdots \\ x_{N_1}^{(1)} & V_1\left(x_{N_1}^{(1)}\right) & \ldots & V_H\left(x_{N_1}^{(1)}\right) \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(T)} & V_1\left(x_1^{(T)}\right) & \ldots & V_H\left(x_1^{(T)}\right) \\ \vdots & \vdots & \ddots & \vdots \\ x_{N_T}^{(T)} & V_1\left(x_{N_T}^{(T)}\right) & \ldots & V_H\left(x_{N_T}^{(T)}\right) \end{array} \tag{1}$$

At this point we employ the consensus clustering [12] to build a vocabulary. We define $\mathbf{v}_n^{(i)} = \left(V_1\left(x_n^{(i)}\right), \ldots, V_H\left(x_n^{(i)}\right)\right)$ as the vector that contains all the *ids* labels for the interesting point x_n^i. Considering the set of all vectors $\mathbf{v}_n^{(i)}$, the consensus clustering algorithm is used to find a consensus partition V_c called the *Consensus Vocabulary*.

The original formulation of the consensus clustering assigns each vector $\mathbf{v}_n^{(i)}$ to the most likely cluster of the consensus partition in a hard way. Taking into account possible visual words ambiguities [9,13], we use a soft assignment. Specifically, we employ the probability vector $\mathbf{z}_n^{(i)}$ given by the consensus algorithm to establish the membership degree of each vector $\mathbf{v}_n^{(i)}$ to the different consensus clusters. Every image I_i is hence represented as the normalized sum of all the $\mathbf{z}_n^{(i)}$:

$$\mathbf{S}_{I_i} = \frac{1}{N_i} \sum_{n=1}^{N_i} \mathbf{z}_n^{(i)} \tag{2}$$

To represent test images $\overline{I}_i$, we first project their interesting points in the set of vocabulary $\mathbf{V}$, and then the consensus vocabulary is used to compute the final signature in the same way as for the training images (Eq. 2).

To perform the classification, a multiclass SVM with a pre-computed kernel and cosine distance is used. Given two signature vectors $\mathbf{S}_{I_i}, \mathbf{S}_{I_j}$, the cosine distance $d_{\cos}$ is calculated as follows:

$$d_{\cos}\left(\mathbf{S}_{I_i}, \mathbf{S}_{I_j}\right) = 1 - \frac{\mathbf{S}_{I_i}\mathbf{S}'_{I_j}}{\sqrt{\left(\mathbf{S}_{I_i}\mathbf{S}'_{I_i}\right)\left(\mathbf{S}_{I_j}\mathbf{S}'_{I_j}\right)}}. \tag{3}$$

The kernel is defined as:

$$k_{\cos}\left(\mathbf{S}_{I_i}, \mathbf{S}_{I_j}\right) = e^{-d_{\cos}\left(\mathbf{S}_{I_i}, \mathbf{S}_{I_j}\right)}. \tag{4}$$

3 Experimental Results

To assess the proposed approach we have used the PFID dataset [2] described in Section 1. Our method has been compared against the two baseline methods reported in [2] as well as with respect to the different methods proposed in [14]. As in [14], we followed the experimental protocol of [2] by performing a 3-fold cross-validation for our experiments. We used 12 images from two instances of each class for training and the 6 remaining images of the third instance for testing.

A dense sampling procedure to extract the descriptors has been considered by using a spatial grid with steps of 8 pixels in both horizontal and vertical directions. The descriptors are computed on patches of 24×24 pixels centered on each point of the spatial grid. The visual vocabularies to be used as input for the consensus clustering have been obtained considering three different runs of the K-means clustering for each descriptor. We have used $K = 200$ on each run with a random initialization. So, each point into the spacial grid of the dense sampling has been projected into the 6-dimensional feature space of the computed visual vocabularies (3 on the SIFT features and 3 on the SPIN features). For the final consensus vocabulary, we chose a size of 300 consensus words. This means that the final food image is represented with a very small vector.

Table 1. Per-Class accuracy of the different methods on the 7 Major Classes of the PFID dataset. In each row, the two highest values are underlined, while the maximum is reported in **bold**.

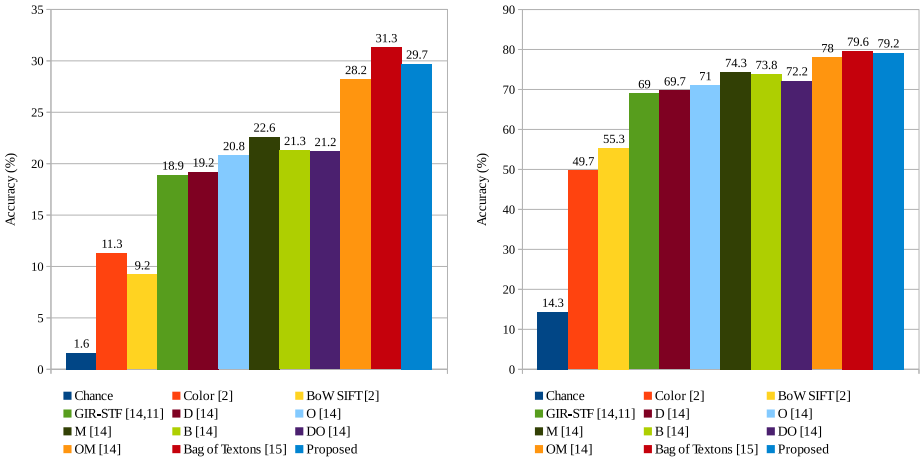
Class	Per-Class Accuracy % (# of images)					
	Color [2]	BoW SIFT [2]	GIR-STF [11,14]	OM [14]	BoW Textons [15]	Proposed
Sandwich	69.0 (157.3)	75.0 (171)	79.0 (180.1)	86.0 (196.1)	87.6 (199.7)	**89.0 (203)**
Salad & Sides	16.0 (5.8)	45.0 (16.2)	79.0 (28.4)	**93.0 (33.5)**	84.3 (30.3)	69.4 (25)
Bagel	13.0 (3.1)	15.0 (3.6)	33.0 (7.9)	40.0 (9.6)	**70.8 (17)**	62.5 (15)
Donut	0.0 (0)	18.0 (4.3)	14.0 (3.4)	17.0 (4.1)	**43.1 (10.3)**	29.2 (7)
Chicken	49.0 (11.8)	36.0 (8.6)	73.0 (17.5)	82.0 (19.7)	66.7 (16)	**91.7 (22)**
Taco	39.0 (4.7)	24.0 (2.9)	40.0 (4.8)	65.0 (7.8)	**69.4 (8.3)**	50.0 (6)
Bread & Pastry	8.0 (1.4)	3.0 (0.5)	47.0 (8.5)	**67.0 (12.1)**	53.7 (9.7)	66.7 (12)
Average	27.7 (26.3)	30.9 (29.6)	52.1 (35.8)	64.3 (40.4)	**67.9 (41.6)**	65.5 (41.4)

After representing images as described in Section 2, we trained the SVM classifier, using the training images and pre-computed kernel with cosine distance. The trained classifier has been then employed on the test images. The classification accuracy achieved employing consensus vocabularies on the 61 classes is reported in Figure 4a, along with the accuracies of the compared state-of-the-art approaches. The low accuracy in discriminating among the 61 different classes is mainly due to foods items of the PFID dataset have very similar appearances (and similar ingredients) despite they belong to different classes [15].

It is important to note that our method, differently than [14], does not need any manual labeling of the different ingredients composing the food items to be employed to produce the representation. 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 [14] difficult to be applied.

As in [14], we have also performed tests re-organizing the 61 PFID food categories into seven major groups (e.g. sandwiches, salads and sides, chicken, breads and pastries, donuts, bagels, tacos). Results obtained by the different approaches are reported in Fig. 4b. In Table 1 the per-class accuracies of the results of the different methods on the seven major classes of the PFID dataset are reported. Since the number of images belonging to the different classes is not balanced, for a better understanding of the results, the number of images is reported together with the per-class accuracy. Also in this case, our approach obtains better performances with respect to the best performing one proposed in [14].

We want also to underline that, despite the approach in [15] has better results in terms of accuracy, the proposed method is valuable under a theoretical perspective. In fact, it shows that the results obtained using a combination of different features are almost as good as standard techniques, but it captures different aspects of the image, such as local gradient and textures. Note that the representation proposed in [15] can be exploited together with SIFT and SPIN to build a novel Consensus vocabulary which takes into account the power of Textons in



(a) (b)

Fig. 4. Classification accuracy on the 61 categories (4a) and on the 7 major classes (4b) of the PFID dataset [2].

representing patterns. Moreover, the final vocabulary size used in this approach is much lower than the one used in [15].

4 Conclusions and Future Works

This paper introduces an augmented version of the Bag of Words approach useful to combine different feature descriptors. We propose to build a consensus vocabulary starting from different visual codebooks. The images are represented as a Bag of Consensus Words taking into account different aspects of local regions. The proposed image representation has been assessed by considering the problem of food recognition, obtaining results which closely match the state of the art. Future works could be devoted to the exploitation of consensus vocabularies on different computer vision applications, and by considering other type of features (as well as the spatial relationship among them) to build the final consensus codebook. Also, an in-depth analysis of the performances at varying of the involved parameters and considering bigger datasets (e.g., the UNICTFD-889 [16]) should be performed. Moreover the impact of the size of the consensus vocabulary both in terms of recognition rate and computational time should be investigated.

References

1. Battiato, S., Farinella, G.M., Puglisi, G., Ravì, D.: Aligning codebooks for near duplicate image detection. Multimedia Tools and Applications, 1–24 (2013)
2. Chen, M., Dhingra, K., Wu, W., Yang, L., Sukthankar, R., Yang, J.: Pfid: Pittsburgh fast-food image dataset. IEEE International Conference on Image Processing, 289–292 (2009)
3. Jiménez, A.R., Jain, A.K., Ceres, R., Pons, J.: Automatic fruit recognition: a survey and new results using range/attenuation images. Pattern recognition **32**(10), 1719–1736 (1999)
4. Joutou, T., Yanai, K.: A food image recognition system with multiple kernel learning. IEEE International Conference on Image Processing, 285–288 (2009)
5. Lazebnik, S., Schmid, C., Ponce, J.: A sparse texture representation using local affine regions. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(8), 1265–1278 (2005)
6. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2), 91–110 (2004)
7. Matsuda, Y., Hoashi, H., Yanai, K.: Recognition of multiple-food images by detecting candidate regions. IEEE International Conference on Multimedia and Expo, 25–30 (2012)
8. Matsuda, Y., Yanai, K.: Multiple-food recognition considering co-occurrence employing manifold ranking. In: International Conference on Pattern Recognition, pp. 2017–2020 (2012)
9. Perronnin, F.: Universal and adapted vocabularies for generic visual categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(7), 1243–1256 (2008)
10. Saffari, A., Bischof, H.: Clustering in a boosting framework, pp. 75–82. Computer Vision Winter Workshop (2007)
11. Shotton, J., Johnson, M., Cipolla, R.: Semantic texton forests for image categorization and segmentation. IEEE Conference on Computer Vision and Pattern Recognition, 1–8 (2008)
12. Topchy, A., Jain, A.K., Punch, W.: Clustering ensembles: Models of consensus and weak partitions. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(12), 1866–1881 (2005)
13. van Gemert, J.C., Veenman, C.J., Smeulders, A.W., Geusebroek, J.-M.: Visual word ambiguity. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(7), 1271–1283 (2010)
14. Yang, S., Chen, M., Pomerleau, D., Sukthankar, R.: Food recognition using statistics of pairwise local features. IEEE Conference on Computer Vision and Pattern Recognition, 2249–2256 (2010)
15. Farinella, G.M., Moltisanti, M., Battiato, S.: Classifying Food Images Represented as Bag of Textons. IEEE International Conference on Image Processing, 5212–5216 (2014)
16. Farinella, G.M., Allegra, D., Stanco, F.: A benchmark dataset to study the representation of food images. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014 Workshops. LNCS, vol. 8927, pp. 584–599. Springer, Heidelberg (2015)
17. Anthimopoulos, M.M., Gianola, L., Scarnato, L., Diem, P., Mougiakakou, S.G.: A Food Recognition System for Diabetic Patients Based on an Optimized Bag-of-Features Model. IEEE Journal of Biomedical and Health Informatics **18**(4), 1261–1271 (2014)

18. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101 – mining discriminative components with random forests. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part VI. LNCS, vol. 8694, pp. 446–461. Springer, Heidelberg (2014)
19. Hu, Y., Cheng, X., Chia, L.-T., Xie, X., Rajan, D., Tan, A.-H.: Coherent Phrase Model for Efficient Image Near-Duplicate Retrieval. IEEE Transactions on Multimedia **11**(8), 1434–1445 (2009)
20. Varma, M., Zisserman, A.: A Statistical Approach to Texture Classication from Single Images. International Journal of Computer Vision **62**(1-2), 61–81 (2005)