A Benchmark Dataset to Study the Representation of Food Images

Giovanni Maria Farinella, Dario Allegra, Filippo Stanco

{gfarinella, allegra, fstanco}@dmi.unict.it
Image Processing Laboratory
Department of Mathematics and Computer Science
University of Catania, Italy

Abstract. It is well-known that people love food. However, an insane diet can cause problems in the general health of the people. Since health is strictly linked to the diet, advanced computer vision tools to recognize food images (e.g. acquired with mobile/wearable cameras), as well as their properties (e.g., calories), can help the diet monitoring by providing useful information to the experts (e.g., nutritionists) to assess the food intake of patients (e.g., to combat obesity). The food recognition is a challenging task since the food is intrinsically deformable and presents high variability in appearance. Image representation plays a fundamental role. To properly study the peculiarities of the image representation in the food application context, a benchmark dataset is needed. These facts motivate the work presented in this paper. In this work we introduce the UNICT-FD889 dataset. It is the first food image dataset composed by over 800 distinct plates of food which can be used as benchmark to design and compare representation models of food images. We exploit the UNICT-FD889 dataset for Near Duplicate Image Retrieval (NDIR) purposes by comparing three standard state-of-the-art image descriptors: Bag of Textons, PRICoLBP and SIFT. Results confirm that both textures and colors are fundamental properties in food representation. Moreover the experiments point out that the Bag of Textons representation obtained considering the color domain is more accurate than the other two approaches for NDIR.

Keywords: Food Dataset, Food Recognition, Near Duplicate Image Retrieval, Textons, PRICoLBP, SIFT

1 Introduction

Food is an essential component of human life. Nowadays, there are researches in different fields (e.g., social, ethical and medical science) where the food has a key role. In particular, automatic recognition of food images can provide the ability to build monitoring systems to be embedded in wearable cameras in order to assess the patients' diet [1, 2]. The food intake monitoring can be relevant especially when patients (e.g., old people with obesity and/or diabetes, people

with food allergy) have to be assisted during their daily meals. Moreover, experts (e.g., nutritionists, psychologists) could use these monitoring applications to study the daily diet of patients to better understand their habits and/or eating disorders. Automatic food recognition could replace the traditional dietary assessment based on self-reporting in a food diary that is often inaccurate. Recent works discuss the possibility of dietary assessment through images acquired from mobile and wearable cameras [1–6]. Hence, considering the wide diffusion of mobile devices equipped with a camera, as well as the forthcoming consumer wearable cameras (e.g., Google glass), automatic food recognition is a useful resource for the assistive technology domain. Of course, the recognition of food images can let imagine many other applications (e.g., finding restaurants which serve a dish previously photographed, retrieving the list of ingredients to cook a specific dish, etc.).

Nevertheless, food has a high variability in appearance (e.g., shape, colors) due to the great assortment of existent ingredients and it is intrinsically deformable. This makes food recognition a difficult task for current state-of-the-art classification methods [7–9], and hence an important challenge for Computer Vision researchers. The image representation employed in a food recognition engine plays the most important role. Moreover, to study a representation for food it is essential to get a huge number of food images, with a high variety of dishes. In sum, representative datasets are indispensable. Although many approaches have been published, the datasets used for the tests have a limited number of classes/images. Moreover, the current datasets have not been designed with the aim of properly studying an image representation for food images; they are usually composed by images collected through the Internet, where a specific dish is present only once. For instance, considering the current food datasets, there is no way to understand if a specific type of feature is repeatable in different images of the same dish acquired under different points of view, scales or rotation angles. Hence, 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 recognition so far.

In this paper we introduce the first food dataset composed by 889 distinct plates of food. Each dish has been acquired with a smartphone multiple times to introduce photometric (e.g., flash vs no flash) and geometric variability (rotation, scale, point of view changes). The overall dataset contains 3583 images acquired with smartphones. The dataset is designed to push research in this application domain with the aim of finding a good way to represent food images for recognition purposes. The first question we try to answer is the following: are we able to perform a near duplicate image retrieval (NDIR) in case of food images? Note that there is no agreement on the technical definition of near-duplicates (see [10] for an in-depth discussion). The definition of near duplicate depends on the degree of variability (photometric and geometric) that is considered acceptable for each particular application. Some approaches (e.g. [11]) consider as near duplicate images the ones obtained by slightly modifying the original

ones through common transformations such as changing contrast or saturation, scaling, cropping, etc. Other techniques (e.g. [12]) consider as near duplicate the images of the same scene but with different viewpoint and illumination. In this paper we consider this last definition of near duplicate food images to test different image representations on the proposed dataset. We benchmark the proposed dataset in the context of NDIR by using three standard state-of-the-art image descriptors: Bag of Textons [13], PRICoLBP [14] and SIFT [15]. Results confirm that both textures and colors are fundamental properties. The experiments performed point out that the Bag of Textons representation is more accurate than the other two approaches for NDIR.

The paper is organized as following. In Section 2 the related works on food classification are presented. A brief description of the food dataset used so far is also given. In Section 3, we describe the proposed dataset. The image representation methods used to perform Near Duplicate Food Retrieval are detailed in Section 4, whereas Section 5 reports the experimental settings and results. Finally, Section 6 concludes the paper with hints for further researches.

2 Related Works

In literature there are several works related to food recognition. Most of them address the challenging problem of health monitoring by proposing solutions for dietary assessment [1–6]. Kim et al. [3] presented a mobile user interface which provides a front-end to a client-server image recognition and portion estimation software. Chen et al. [16] proposed a three steps pipeline for calories estimation in dietary assessment: in the first step the food base plane is localized, then the food image is segmented and a 3D model is obtained; finally calories of the food are inferred using information related to the volume of the estimated 3D model. Kong et al. [1] proposed a system called "DietCam" to verify the diet assessment for obesity monitoring purposes.

In all of the aforementioned assistive technologies, a robust and effective recognition engine of food dishes is important. It can help the users in self-reporting the daily food intake. Moreover, the automatically collected information (e.g., through the system implemented in wearable smart glasses) can be also used by the experts to monitor patients over time to better understand eating disorders. A key role in food recognition engine is given to the representation models used to describe the visual content of food images. Jimnez et al. [17] proposed a method to recognize spherical fruits in natural environments by considering variabilities such as shadows, bright areas, occlusions and overlapping fruits. To this aim a laser scanner has been used to obtain a representation based on 3D information. Joutou et al. [18] exploited multiple features together with a Multiple Kernel Learning (MKL) for food classification. Specifically they used the bag of visual words paradigm on SIFT features, color histogram, and responses to Gabor filters to encode the images. Yang et al. [8] proposed to exploit spatial relationships between the different ingredients composing a dish after a soft-labeling performed through semantic segmentation. The direction of

Fig. 1. Examples of 96 dishes of the proposed UNICT-FD889 dataset.

the spatial ingredients co-occurrences and the information of the soft-labeling of the midpoint among them have been also included in the representation.

To properly evaluate the performances of a food recognition method a dataset has to be used for testing purposes. Chen et al. [7] proposed the "Pittsburgh Fast-food Image Dataset" (PFID). This dataset is composed by 1098 food images belonging to 61 different categories of fast-food dishes mainly acquired in laboratory. Food pixels are labeled in order to discard background for experimental purposes. This dataset has been used in recent works to compare the performances of different food image representations [8, 9]. Another dataset that can be used for food classification purposes is the one proposed by Matsuda et al. [19]. This dataset, called "UECFood100", is composed by typical asian food images collected through the internet and belonging to 100 food categories.

3 Proposed Dataset

Considering the aforementioned works it can be summarised that the main representation for food images have been obtained considering SIFT features, Textures information (e.g., obtained through Gabor filters) and color information. The bag of visual word paradigm is usually used to obtain the final feature vector to be used for food classification purpose. Moreover spatial relationship among features are employed. However, to properly study the peculiarities of the image representation in the food application context, a representative benchmark

Fig. 2. An example of 32 dishes of the proposed UNICT-FD889 dataset. Three different instances for each dish are shown. The images of a dish present both geometric and photometric variabilities (e.g., see images of the first, second and third columns in the second row).

datasets is needed. These facts motivated our work. We introduce the first food images dataset composed by 3583 images related to 889 distinct dishes of food of different nationalities (e.g., Italian, English, Thai, Indian, Japanese, etc.). In Fig. 1 some instances of the different dishes within the proposed dataset are shown. We refer to this dataset as "UNICT-FD889" (UNICT Food Dataset 889) from here on. Food images have been acquired by users in the last four years during meals with a smartphone (i.e., iPhone 3GS or iPhone 4) in unconstrain settings (e.g., background, light environment conditions, etc). Hence, the UNICT-FD889 dataset is a collection of food images acquired by users in real cases of meals. Each plate of food has been acquired multiple times (four in the average) to guarantee the presence of geometric and photometric variabilities (see Figs. 1 and 2). UNICT-FD889 dataset is designed to arouse research in this application domain with the aim of finding a good way to represent food images for recognition purposes. The UNICT-FD889 differs from the PFID [7] and UECFood100 [19] dataset not only because it contains a larger number of distinct food dishes. The main difference is related to how the dataset has been acquired, i.e., by users during meals, which allows to test real cases. Moreover, differently than the other two datasets, each dish is present multiple times (with variabilities) allowing to perform a more accurate study in building a represen-

tation model for food recognition. The complete set of images composing the UNICT-FD889 dataset can be visually assessed (and downloaded) at the following URI: `http://iplab.dmi.unict.it/UNICT-FD889`

4 Representation Methods

In this paper we benchmark the proposed dataset considering it for Near Duplicate Image Retrieval (NDIR) purpose. We employ three standard state-of-the-art image descriptors as baseline in our tests: Bag of Textons [13], PRICoLBP [14] and SIFT features [15]. We decided to use Textons because they are powerful in representing textures and have been obtained the best results so far on the PFID dataset [8, 9]. PRICoLBP descriptor has been chosen since it encodes spatial co-occurrence of local LBP features which are useful to represent textures. Finally SIFT features have been considered due their good performances in the context of near duplicate image retrieval. All the considered local descriptors are rotationally invariant. The SIFT is also scale invariant. The representation have been considered in both grayscale and color domains. The experiments reported in Section 5 pointed out that the Bag of Textons representation in color domain is more accurate than the other approaches for NDIR. In the following subsections the three baseline representations used in our experiments are briefly summarized.

4.1 Bag of Textons

Textons have been introduced by Julesz in 1981 [20] as the elementary structure for the visual perception and in particular as key atoms of textures. Textons have been used in computer vision studies in the context of texture analysis and image classification [13, 21–23]. From a computational point of view, Textons are obtained as responses of the gray or color image to a bank of filters. To represent an image the filter responses of the images of a training dataset are collected and quantised through clustering procedure (e.g., K-means). Each cluster prototype is hence considered as a Textons and the collection of Textons compose the final codebook. The pixel-wise filter responses of an image (from training and testing sets) are hence associated to the different clusters (i.e., Texton prototype) considering a similarity measure (e.g., Euclidean distance). The distribution over the different Textons is hence used to represent an image. In [9] Textons have been exploited for food classification purposes obtaining good results. To obtain Textons, the feature space of the filter responses related to the training images can be clustered taking into account the classes or in a global way. When classes are considered, the quantization procedure is performed considering the responses of the different classes separately and then the Textons obtained for the different classes are taken all together to compose the final codebook. On the other hand, clustering all the filter responses independently from the different classes allows to obtain a single global Textons vocabulary to be used for classification purposes. Following the experiments performed in [9] we

consider both method (class-based and global) to obtain Textons and compare them considering the UNICT-FD889 dataset for NDIR purposes. For the class-based Textons representation each dish is considered a class. So considering a quantization to obtain K Textons per dish we have a final codebook composed by $889 \times K$ Textons. In our experiments we use the rotational invariant MR4 bank of filters to extract Textons [13]. All the details on how to properly obtain the Bag of Textons representation (e.g., image intensities normalization and filter responses normalization) are available in [13] and [9]. Since the number of clusters (i.e., Textons) to be used for quantizing the filter responses space is a parameter of this approach, we have performed tests at varying of it. In our tests we have employed the χ^2 distance to measure the similarity between the Bags of Textons related to two different images.

4.2 PRICoLBP: Pairwise Rotation Invariant Co-occurrence Local Binary Pattern

PRICoLBP [14] is a descriptor to encode texture information. This descriptor considers spatial co-occurrence and pairwise orientations of the well-known LBP local features [24]. The descriptor preserves the relative angles between the orientations of LBP features pairs by obtaining rotational invariance. Given two pairs A and B, the orientation of the point A is computed and then the uniform pattern of B is obtained taking into account the orientation of A. For the co-occurrence pattern, the authors use the gradient magnitude of two points to weight the co-pattern. PRICoLBP can be extracted on both grayscale and color domain. This descriptor has been recently employed in the context of food recognition obtaining the best results with respect to others state-of-the-art methods [25]. To compute the PRICoLBP descriptor, we have used the original implementation provided by the authors which is available at the following page: `http://qixianbiao.github.io/`. Also in this case, we have employed the χ^2 distance to measure the similarity between two different images represented with PRICoLBP descriptor.

4.3 SIFT Features

Scale-Invariant Feature Transform (SIFT) [15] is one of the most popular descriptor used in computer vision. This descriptor has been tested in different application contexts, such as object recognition [26], image stitching [27] and near duplicate image retrieval [28, 29] and food recognition [8]. SIFT is a transform able to detect keypoints invariant to scale and rotation. Moreover, it is robust to affine transformations and changes in illumination. For near duplicate image retrieval purpose the SIFT features extracted from a query image are matched to the SIFT features of the image in the training dataset. The matching is done through Euclidean-distance based nearest neighbor approach. A rejection procedure based on the ratio of the nearest neighbor distance to the second nearest neighbor distance is used to increase robustness. The query image is associated to the image of the training dataset with the highest number of matches for

retrieval purposes. In our test we also tested a weighted matching score in which each match is inversely weighted taking into account the similarity between the SIFT descriptors of the matched keypoints. In our experiments we use VLFeat [30] to extract SIFT features considering both grayscale and color domain. In the case of color images the SIFT features are extracted and matched independently on each color channel. Then, given an image query, the sum of the matching over the three channels are considered to compute the most similar image on the training set.

5 Experimental Settings and Results

In this section we discuss the settings and the evaluation measures used to compare the three considered image representations on the UNICT-FD889 dataset. For testing purposes images have been resized to 320 x 240 pixels. To properly evaluate the different representation methods, the experiments have been repeated three times. At each run different approaches are executed on the same training and test sets. To this purpose, at each run we have built a training set composed by 889 images, by selecting one image of the UNICT-FD889 dataset per dish, whereas the rest of images have been used for testing purposes. The images considered for the three training sets are different. At each run, test images are used to perform queries on the corresponding training dataset used for that test. Given an image representation, the final results are obtained by averaging over the three tests. As in [12, 29], the retrieval performances on each run have been evaluated with the probability of the successful retrieval $P(n)$ in a number of test queries:

$$P(n) = \frac{Q_n}{Q} \tag{1}$$

where Q_n is the number of successful queries according to $top - n$ criterion, i.e., the correct near duplicate image is among the first n retrieved images, and Q is the total number of queries. We also consider the precision/recall values at $top - n = 1$. Note that the precision and recall for $top - n = 1$ are equivalent because there is only one correct match for each query in the training set. Finally the retrieval results are evaluated through the mean average precision (mAP) measure, i.e., the area under the precision-recall curve (see [31] for further details).

As first we tested the Bag of Textons representation obtained in two modalities [9, 13]: class-based and global-based. For the class-based representation each dish image within the training set has been considered as a class. Then, 10 Textons per image have been extracted by quantizing through K-means clustering the filter responses space related to the considered dish image. The final Textons vocabulary to be used for the representation has been obtained by collecting all the Textons extracted on each dish image. The size of the Textons vocabulary is hence equal to 8890. For the global-based Textons, all the filters responses related the 889 dish images of the training set have been considered in a single

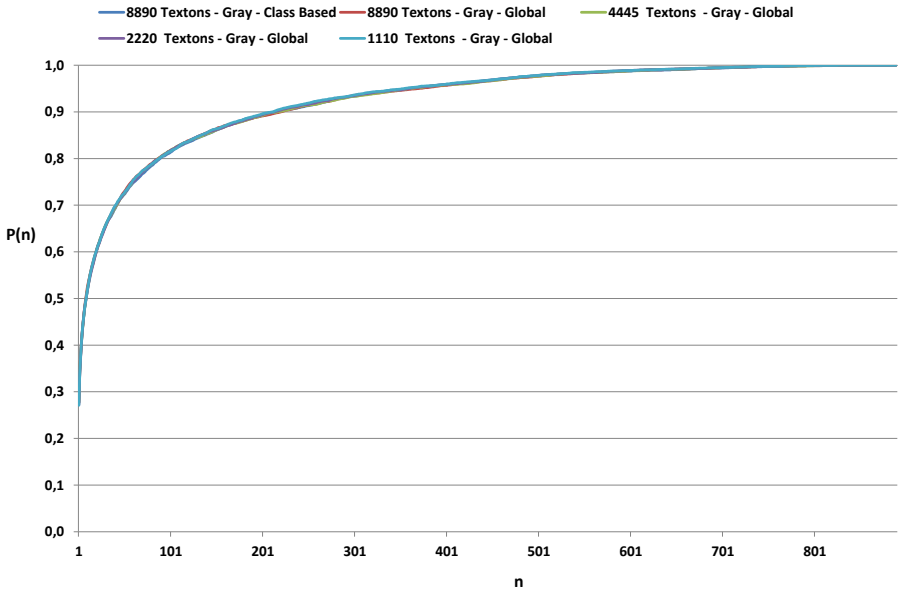

Fig. 3. Global Textons vs Class-Based Textons.

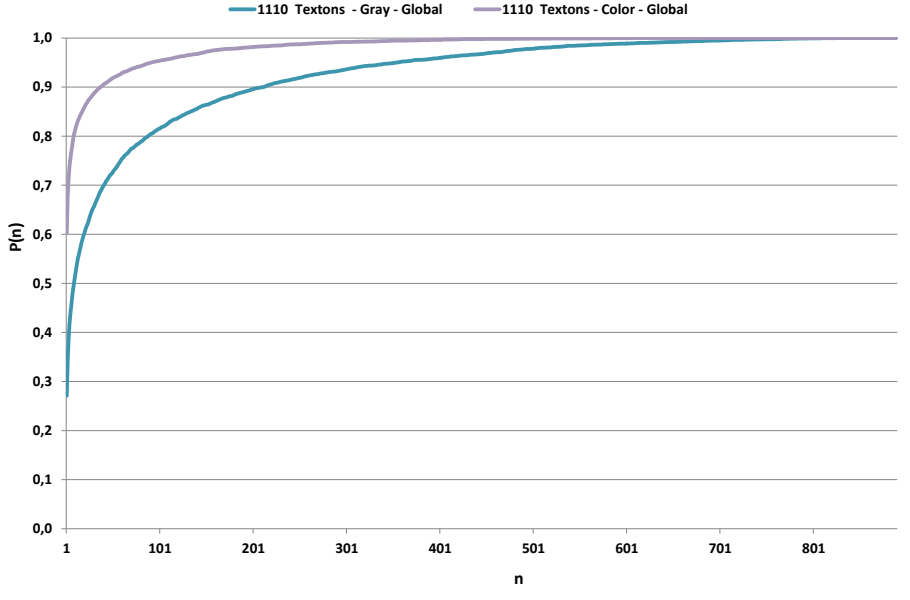

Fig. 4. Gray Textons vs Color Textons.

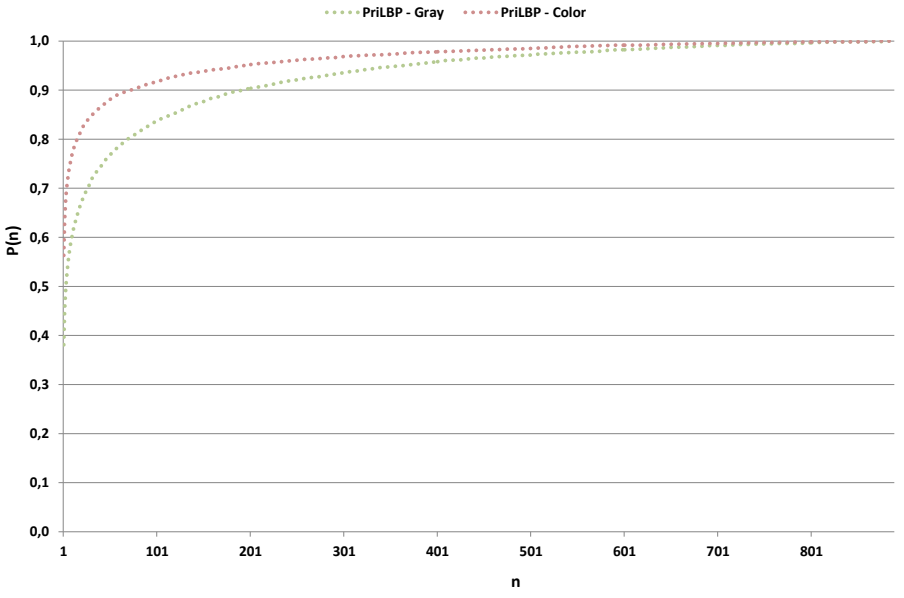

Fig. 5. Gray PRICoLBP vs Color PRICoLBP.

run of the Kmeans clustering with K=8890. Moreover the global-based representation has been considered by reducing the number of Textons composing the final vocabulary. These tests have been performed by considering gray scale version of the images. Fig. 3 shows $P(n)$ curves related to the aforementioned tests. It can be seen that all the curves are overlapped. This means that the two Bag of Textons representation modalities (class-based and global) obtain similar performances. Moreover reducing the global vocabulary to 1110 does not affect the results. This differs from the results obtained by the authors of [9], where class-based Bag of Textons outperformed global one. Motivation can be due to the fact that the UNICT-FD889 dataset is bigger than the dataset used in [9]; the guess is that by increasing the number of images and classes the two representation modalities converge in performances. Finally it is interesting to note that by strongly reducing the number of the global Textons in the final vocabulary from 8890 to 1110 the performances are maintained with high reduction of both time and space resources. Note that the accuracy obtained considering a global vocabulary with 1110 Textons at $top-n = 1$ is less than 30%. To understand the influence of color information in representing food, we have compared Bag of Textons obtained in a global way in grayscale and color domains (i.e., responses of filters on the three RGB channels). The results of the comparison are reported in Fig. 4. Considering $top - n = 1$ the representation obtained in the color domain achieves more than 30% of improvement (i.e., 60.20%).

Since the results obtained with the Bag of Textons approach were promising, we considered the recent image descriptor PRICoLBP [14] for comparison pur-

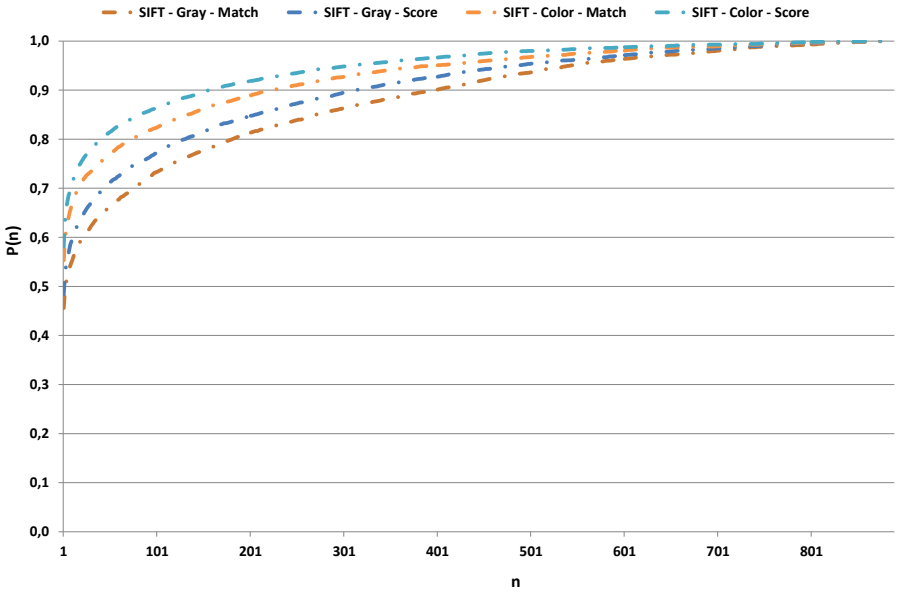

Fig. 6. Gray SIFT vs Color SIFT. The words "match" and "score" are used to identify respectively the similarity measure based on the number of matching and the weighted one.

poses. This descriptor is able to encode textures and also spatial co-occurrence of local features in a rotational invariant way. It has been recently used for food classification obtaining the best results [25]. However it has never been compared with respect to Bag of Textons representation in the context of food recognition. The results obtained by PRICoLBP descriptor (both gray and color domain) are reported through $P(n)$ curves in Fig. 5. Despite the PRICoLBP descriptor achieves better performances than Bag of Textons in gray scale domain, it does not outperforms Bag of Textons in the color domain.

Finally we have tested SIFT which are the most popular features used for near duplicate image matching. The results obtained with the SIFT descriptor are reported in Fig. 6. In this case we have used the similarity measures based on number of matchings (with Lowe's rejection procedure) and also the one in which the matchings are inversely weighted taking into account matching distances (called score in Fig. 6). The approach with weighted similarity performs better than considering only the number of matchings. However, Bag of Textons in color domain wins the comparison again.

The best results obtained with the Bag of Textons representation, PRICoLBP and SIFT descriptors are reported in Fig. 7. By zooming the $P(n)$ curves (see Fig. 8) it can be observed that at $top - n = 1$, Bag of Textons representation obtain the best performances and SIFT matching outperform PRICoLBP. In Fig. 9 we report Precision/Recall (i.e., $top - n = 1$ since we have only one

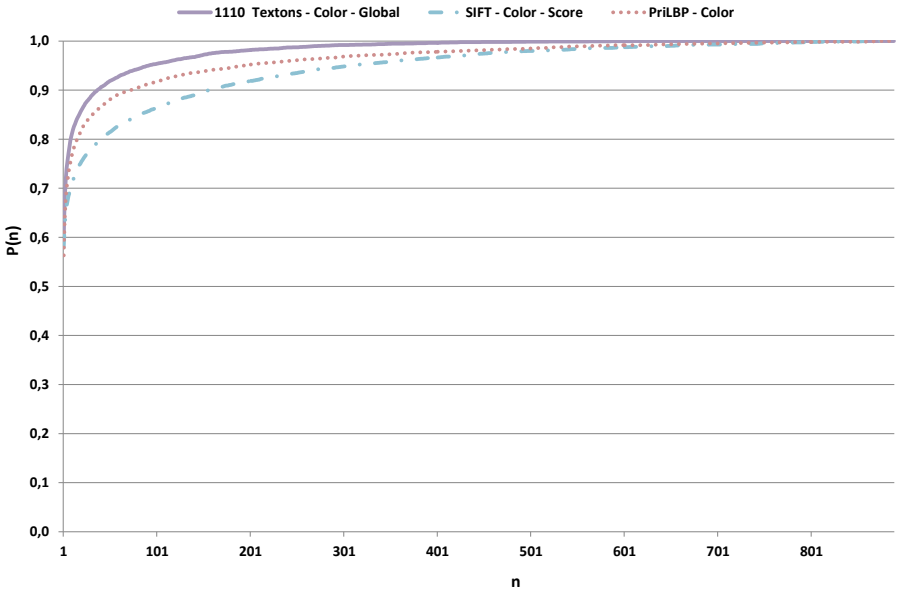

Fig. 7. Best results obtained with Bag of Textons, PRICoLBP and SIFT.

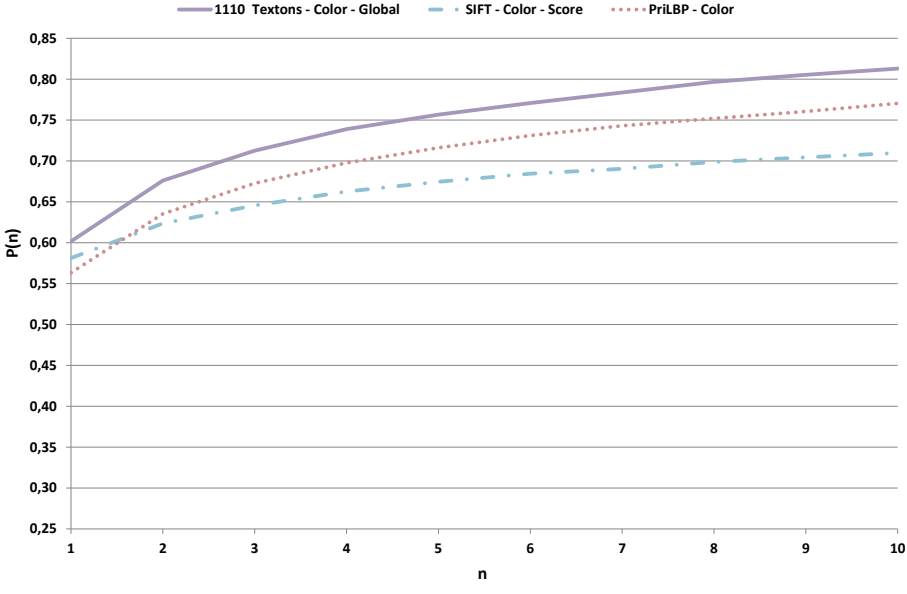

Fig. 8. Best results obtained with Bag of Textons, PRICoLBP and SIFT.

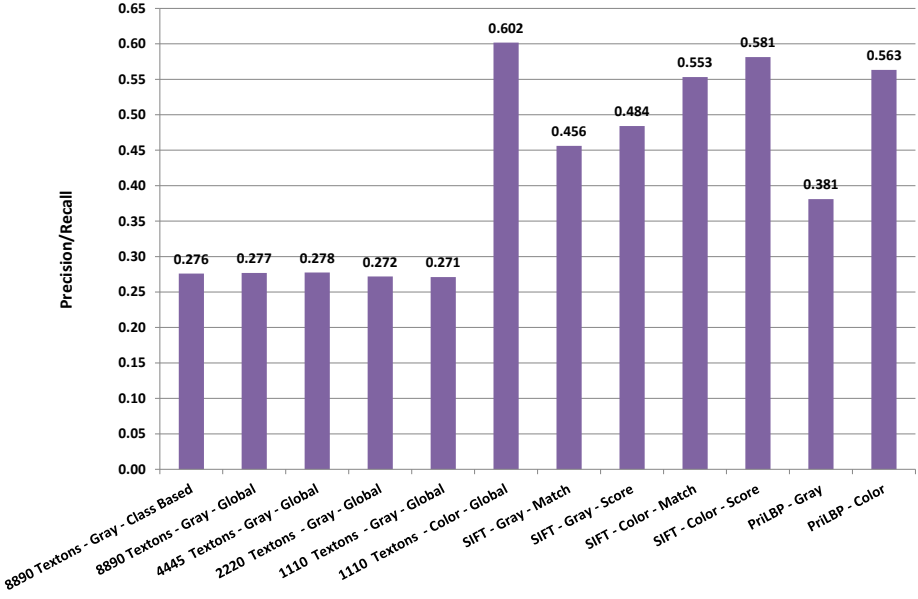

Fig. 9. Precision/Recall results (i.e., $top - n = 1$)

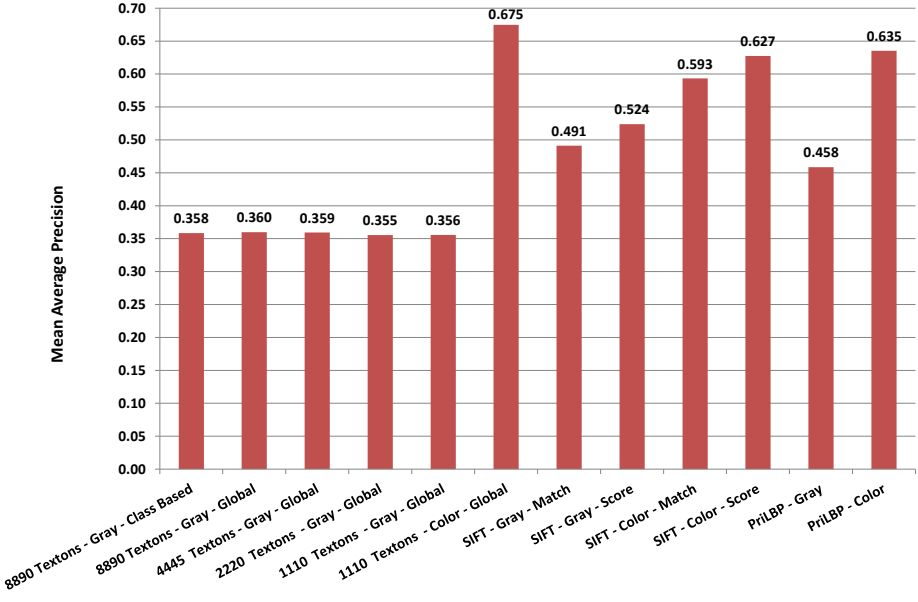

Fig. 10. mAP comparison.

image per dish in the training set) of all tested representation. Finally in Fig. 10 the mAP results are shown.

6 Conclusion and Future Works

In this paper we considered the problem of representation and distinguishing among different food dishes. The automatic recognition of food images is a challenging area for computer vision researchers and is fundamental for assistive technologies concerning diet monitoring. Image representation plays an important role in food recognition and, for this reason, we proposed a representative dataset called UNICT-FD889 to evaluate image representation models. The dataset contains over 800 dishes, and can be employed to study the peculiarities and weaknesses of different representation techniques. In this work we have tested Bag of Textons, PRICoLBP and SIFT representations to benchmark the dataset considering the problem of near duplicate image retrieval. Experiments pointed out that Bag of Textons representation gives the best result. Future works could be devoted to extend the proposed dataset in order to include other samples of dishes, to organize the images of the dataset in categories with different levels (e.g., main courses vs second courses, pasta vs pizza), and to label the main ingredients of each dish. Moreover classification experiments (e.g., food vs non-food, main courses vs second courses, pasta vs pizza) to assess the performances of different representation models coupled with state-of-the-art classifiers (e.g., SVM) could be done. By taking into account the results obtained in this paper, others image descriptors which consider spatial co-occurrence of Textons (e.g., Correlatons [32]) can be considered and eventually revised to address the problem of food recognition. Finally combination of different descriptors could be tested to exploit their peculiarities.

References

1. Kong, F., Tan, J.: Dietcam: Automatic dietary assessment with mobile camera phones. Pervasive and Mobile Computing **8**(1) (2012) 147–163
2. Xu, C., He, Y., Khannan, N., Parra, A., Boushey, C., Delp, E.: Image-based food volume estimation. In: International Workshop on Multimedia for Cooking and Eating Activities. (2013) 75–80
3. Kim, S., Schap, T.R., Bosch, M., Maciejewski, R., Delp, E.J., Ebert, D.S., Boushey, C.J.: Development of a mobile user interface for image-based dietary assessment. In: International Conference on Mobile and Ubiquitous Multimedia. (2010) 1–13
4. Arab, L., Estrin, D., Kim, D.H., Burke, J., Goldman, J.: Feasibility testing of an automated image-capture method to aid dietary recall (2011)
5. Zhu, F., Bosch, M., Woo, I., Kim, S., Boushey, C.J., Ebert, D.S., Delp, E.J.: The use of mobile devices in aiding dietary assessment and evaluation. Journal of Selected Topics in Signal Processing **4**(4) (2010) 756–766
6. O'Loughlin, G., Cullen, S.J., McGoldrick, A., O'Connor, S., Blain, R., O'Malley, S., Warrington, G.D.: Using a wearable camera to increase the accuracy of dietary analysis. American Journal of Preventive Medicine **44**(3) (2013) 297–301

7. Chen, M., Dhingra, K., Wu, W., Yang, L., Sukthankar, R., Yang, J.: Pfid: Pittsburgh fast-food image dataset. In: IEEE International Conference Image Processing. (2009) 289–292
8. Yang, S., Chen, M., Pomerleau, D., Sukthankar, R.: Food recognition using statistics of pairwise local features. In: IEEE Computer Vision and Pattern Recognition. (2010) 2249–2256
9. Farinella, G.M., Moltisanti, M., Battiato, S.: Classifying food images represented as bag of textons. In: IEEE International Conference on Image Processing. (2014)
10. Oliveira, R.D., Cherubini, M., Oliver, N.: Looking at near-duplicate videos from a human-centric perspective. ACM Transaction on Multimedia Comput. Commun. Appl. **6**(3) (2010) 15:1–15:22
11. Ke, Y., Sukthankar, R., Huston, L.: Efficient near-duplicate detection and subimage retrieval. In: ACM International Conference on Multimedia. (2004) 869–876
12. 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) (2009) 1434–1445
13. Varma, M., Zisserman, A.: A Statistical Approach to Texture Classification from Single Images. International Journal of Computer Vision **62**(1-2) (2005) 61–81
14. Qi, X., Xiao, R., Guo, J., Zhang, L.: Pairwise rotation invariant co-occurrence local binary pattern. In: European Converence on Computer Vision. (2012) 158–171
15. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision **60**(2) (2004) 91–110
16. Chen, H.C., Jia, W., Yue, Y., Li, Z., Sun, Y.N., Fernstrom, J.D., Sun, M.: Model-based measurement of food portion size for image-based dietary assessment using 3d/2d registration. (2013)
17. Jimnez, A.R., Jain, A.K., Ruz, R.C., Rovira, J.L.P.: Automatic fruit recognition: a survey and new results using range/attenuation images. Pattern Recognition **32**(10) (1999) 1719–1736
18. Joutou, T., Yanai, K.: A food image recognition system with multiple kernel learning. In: IEEE International Conference on Image Processing. (2009) 285–288
19. Matsuda, Y., Hoashi, H., Yanai, K.: Recognition of multiple-food images by detecting candidate regions. In: IEEE International Conference on Multimedia and Expo. (2012) 25–30
20. Julesz, B.: Textons, the elements of texture perception, and their interactions. Nature **290**(5802) (1981) 91–97
21. Malik, J., Belongie, S., Leung, T., Shi, J.: Contour and Texture Analysis for Image Segmentation. International Journal of Computer Vision **43**(1) (2001) 7–27
22. Leung, T., Malik, J.: Representing and Recognizing the Visual Appearance of Materials using Three-dimensional Textons. Int. J. Comput. Vision **43**(1) (2001) 29–44
23. Battiato, S., Farinella, G.M., Gallo, G., Ravì, D.: Exploiting textons distributions on spatial hierarchy for scene classification. Eurasip Journal on Image and Video Processing (2010) 1–13
24. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(7) (2002) 971–987
25. Qi, X., Xiao, R., Li, C., Qiao, Y., Guo, J., Tang, X.: Pairwise rotation invariant co-occurrence local binary pattern. IEEE Transactions on Pattern Analysis and Machine Intelligence (2014)
26. Lowe, D.G.: Object recognition from local scale-invariant features. In: IEEE International Conference on Computer Vision. (1999) 1150–1157

27. Brown, M., Lowe, D.: Automatic panoramic image stitching using invariant features. International Journal of Computer Vision **74**(1) (2007) 59–73
28. Chum, O., Philbin, J., Zisserman, A.: Near duplicate image detection: min-hash and tf-idf weighting. In: British Machine Vision Conference. (2008) 1–10
29. Battiato, S., Farinella, G.M., Puglisi, G., Ravì, D.: Aligning codebooks for near duplicate image detection. Multimedia Tools and Applications (2013)
30. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable libraryof computer vision algorithms. `http://www.vlfeat.org/` (2008)
31. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: Conference on Computer Vision and Pattern Recognition. (2007)
32. Savarese, S., Winn, J., Criminisi, A.: Discriminative object class models of appearance and shape by correlatons. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. (2006) 2033–2040