Computers in Biology and Medicine 77 (2016) 23–39

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Retrieval and classification of food images

Giovanni Maria Farinella *, Dario Allegra *, Marco Moltisanti, Filippo Stanco, Sebastiano Battiato

Dipartimento di Matematica e Informatica, Viale A. Doria 6, 95125 Catania, Italy

ARTICLE INFO

Article history:
Received 3 November 2015
Received in revised form
8 July 2016
Accepted 11 July 2016

Keywords:
Food retrieval
Food classification
Food representation
Textons
Anti-Textons

ABSTRACT

Automatic food understanding from images is an interesting challenge with applications in different domains. In particular, food intake monitoring is becoming more and more important because of the key role that it plays in health and market economies. In this paper, we address the study of food image processing from the perspective of Computer Vision. As first contribution we present a survey of the studies in the context of food image processing from the early attempts to the current state-of-the-art methods. Since retrieval and classification engines able to work on food images are required to build automatic systems for diet monitoring (e.g., to be embedded in wearable cameras), we focus our attention on the aspect of the representation of the food images because it plays a fundamental role in the understanding engines. The food retrieval and classification is a challenging task since the food presents high variableness and an intrinsic deformability. To properly study the peculiarities of different image representations we propose the UNICT-FD1200 dataset. It was composed of 4754 food images of 1200 distinct dishes acquired during real meals. Each food plate is acquired multiple times and the overall dataset presents both geometric and photometric variabilities. The images of the dataset have been manually labeled considering 8 categories: *Appetizer, Main Course, Second Course, Single Course, Side Dish, Dessert, Breakfast, Fruit*. We have performed tests employing different representations of the state-of-the-art to assess the related performances on the UNICT-FD1200 dataset. Finally, we propose a new representation based on the perceptual concept of Anti-Textons which is able to encode spatial information between Textons outperforming other representations in the context of food retrieval and Classification.

1. Introduction and motivations

It is well-known that a non-healthy diet can cause health problems such as obesity and diabetes, as well as risks for people with food allergy. The current mobile imaging technologies (e.g., smartphones and wearable cameras) give the opportunity of building advanced systems for food intake monitoring in order to assess the patients' diet [5,6,36,49,56,62,78,88,113,118]. Related assistive technologies can also be useful to increase the awareness of the society with respect to the quality of life. In this context the ability to automatically recognize images of food acquired with a mobile camera is fundamental to assist patients during their daily meals. Automatic food image retrieval and classification could replace the traditional dietary assessment based on self-reporting that is often inaccurate. As pointed out in different works [4,10,30,31,33,50–53,67], food understanding engines embedded in mobile or wearable cameras can create food-logs of the daily intake of a patient; these information help the experts (e.g., nutritionists, psychologists) to understand the behavior, habits and/or eating disorders of a patient.

However, food has a high variability in appearance and it is intrinsically deformable. This makes classification and retrieval of food images difficult tasks for current state-of-the-art methods [22,116,32], and hence an interesting challenge for Computer Vision researchers. The image representation used to automatically understand food images plays the most important role. Despite many approaches have been published, it is difficult to find works where different techniques are compared on the same dataset. This makes difficult to figure out peculiarities of the different representations, as well as to understand which is the best representation method for food retrieval and classification.

To find a suitable representation of food images it is important to have representative datasets with a high variety of dishes. Although different retrieval and classification methods have been proposed in the literature, most of the datasets used so far have not been designed having in mind the study of a proper image representation for food images. Many food datasets are composed

* Corresponding authors.
E-mail addresses: gfarinella@dmi.unict.it (G.M. Farinella), allegra@dmi.unict.it (D. Allegra), moltisanti@dmi.unict.it (M. Moltisanti), fstanco@dmi.unict.it (F. Stanco), battiato@dmi.unict.it (S. Battiato).

http://dx.doi.org/10.1016/j.compbiomed.2016.07.006
0010-4825/

of images collected through the Internet (e.g., downloaded from Social Networks), where a specific plate is present just once; there is no way to understand if a specific type of image representation is useful for the classification and retrieval of a specific dish acquired under different points of view, scales or rotation angles. Also the food images collected through the Internet have usually a low resolution and have been processed by the users with artistic or enhancement filters.

The automatic analysis of food images has a long history. The article by Parrish et al. [79], which is probably the first using Computer Vision techniques for a food analysis tasks, dates back to 1977. Looking at the literature in this context, it is quite evident that between 1980s and 2000s the interest on food image understanding was mainly for engineering applications related to the production chain and the assessment of the quality of the marketed food. From the beginning of the new century, with the proliferation of high performance mobile devices, the research has focused more and more on aspects which are strictly related to everyday life, and hence on problems and applications for food intake monitoring.

In this paper we consider the problem of food image representation for retrieval and classification purposes. After an in-depth review of the literature related to food image analysis, a new dataset designed for the study of the representation of images is introduced. The proposed dataset, called UNICT-FD1200, is composed of 1200 different food plates acquired by users during real meals. Each food plate has been acquired multiple times and in different light conditions to guarantee both high geometric and photometric variability. Building on top of previous works [30,32] we employ a bag of words like representation based on Texton features [105] to represent the images of food. We present an in-depth analysis of the main "ingredients" to be used into the bag of Textons representation pipeline to point-out which color domain, bank of filters and vocabulary size are more suitable to tackle the retrieval and classification in this specific domain. We also propose a new image representation building on the perceptual concept of Anti-Textons discussed in [110] by Williams and Julesz. The proposed Anti-Texton features extend Textons by encoding spatial information during feature extraction.

The contribution of this paper is three-fold:

(i) a deep review of the state-of-the-art approaches and datasets;
(ii) the introduction of a new public available dataset of food images;
(iii) a new method for texture-based representation of food images.

The reminder of the paper is organized as follows. In Section 2 we present a survey of the state-of-the-art in the field of food image analysis. Section 3 introduces the proposed UNICT-FD1200 food dataset, whereas Section 4 discusses the image representations used in this paper for experimental purposes and defines Anti-Texton features. Section 5 reports the experimental settings and the results. Finally, we draw our conclusions.

2. Food image analysis

Food image analysis has a long history. With the aim of giving a survey of the main works in the literature, we have identified four application areas:

- Detection and recognition of food for automatic harvesting.
- Quality assessment of meals produced by industry.
- Food logging, dietary management and food intake monitoring.
- Food classification and retrieval.

Despite most of the "ingredients" involved in the solutions proposed in the different application areas overlap, the main aims of the final systems are different. For instance, if a certain accuracy obtained by a system for the detection and recognition of food for automatic harvesting could be acceptable by a robotic industry, the same accuracy could be not sufficient in systems dedicated to the diet monitoring for patients with diabetes or food allergy. This motivated us in grouping the works in the literature by considering the four aforementioned areas.

Automatic detection and recognition of fruits and vegetables are useful to enhance robots affordable and reliable vision systems in order to improve the harvesting procedures both in terms of quality and speed. In the late 1980s, industrial meals production knew a large scale expansion, so the evaluation of the quality of the produced food with vision systems became an interesting and valuable challenge. From the late 1990s, the growth of the number of people affected by diseases caused by a non-healthy diet moved the focus to the usage of Computer Vision techniques to help experts (e.g., nutritionists) for the monitoring and understanding the relationships between patients and their meals. This particular researches can take advantage of the huge diffusion on low-cost imaging devices, such as the current smartphones and wearable devices. The large and fast growth of mobile cameras, together with the birth and diffusion of social network services – such as Facebook, Instagram, Pinterest – opened the possibility to upload and share pictures of food. For these reasons, in the past few years, classification and retrieval of food images become more and more popular.

In the following section we will review the state-of-the-art in the field in order to give to the reader an overview of what have been done in the four application domains mentioned above.

2.1. Detection and recognition for automatic harvesting

Among the several techniques used for the harvesting of fruits, the more desirable are the ones which do not cause damages to the fruit and/or to the tree. Thus, accurate systems for fruits detection and recognition from images are needed in order to perform the task correctly. One of the first Computer Vision approaches has been designed by Parrish et al. [79] and focuses on apples detection task. The vision system is composed of a B/W camera and an optical red filter. The image is binarized through thresholding operation, then smoothed to suppress noise and artifacts. At the end, the roundness of the region is estimated by measuring the difference between the longest horizontal and vertical segments inside the region itself. To accept a region as an apple, a density estimation and thresholding is performed.

In [61], a robot vision system called AID is implemented for oranges recognition. A pseudo-gray image is obtained by means of an electronic filter used to enhance the image. Every pixel is coded using 6 bits. The value is proportional to the distance between the pixel hue value and a reference hue value. Then, the gradient is computed using a classic Sobel filter to obtain the magnitude and directions in two separate images. Using a gradient template previously computed, a matching is performed to interpret the scene correctly. With this approach the 70% of the fruits were detected.

An orange recognition method, based on color images, is proposed in [96]. Here, Hue and Saturation components of each pixel are employed to form a two-dimensional feature space. Then, two thresholds based on the maximum and minimum values for each component are used as linear classifiers to define a region in the feature plane. Approximately 75% of the pixels were correctly classified. In [97], the same authors extend their study employing a Bayesian classifier, using the RGB values instead of the Hue and Saturation components, with the goal of segmenting the fruit

pixels from the background pixels. The tests show that 75% of the pixels are correctly classified.

The Purdue University (USA) and The Volcani Center (Israel) developed a vision system for melon harvesting [21]. A B/W image is analyzed to locate the melon and estimate its size. On this image, basic operations – involving also shape and textures analysis – are performed in order to obtain multiple candidate regions. Then, using prior knowledge on the domain, the candidates are evaluated to discard noisy and multiple detections by achieving a true positive rate of 84%.

The Italian institute CIRAA developed a robotic system named AGROBOT [18]. The goal was automatizing greenhouse operations. The scene is acquired through a color camera and is segmented using the Hue and Saturation histograms via thresholding. Then information about the 3-dimensional geometry of the scene is retrieved using stereo matching. The performances of the AGROBOT were pretty good – about 90% of correctly detected ripe tomatoes. The occlusions were the most frequent causes of errors.

The 3D information, obtained with a laser scanner, is employed also by Jiménez et al. [43] to perform automatic harvesting of spherical fruits. The laser scanner maps the points of each scene in the 3D world using spherical coordinates, and associates a distance to each point estimating the laser energy attenuation value. Combining the extracted features, the scene is mapped onto four images, representing the azimuth, the elevation angles, the distance from the sensor (i.e. a depth map) and the attenuation values. Exploiting the sensor model, these images are processed, and taking advantage of the information retrieved by the scanner, four images are produced in output. Of these four, three are actually used for the orange recognition: one is an enhancement of the previous image representing the distance from the sensor, the others encode respectively the apparent reflectance and the reflectance of the surfaces. The image analysis focuses on the last two images. The apparent reflectance image is thresholded to separate the background from the foreground and then the remaining pixels are clustered using the Euclidean distance. The detected clusters without a minimum number of pixels belonging to it are rejected as valid fruit in order to eliminate the possibility of random small areas of a highly reflective non-fruit object. This method, though, is not able to detect fruits whose reflectance is under 0.3. To cope with these kinds of items, the Circular Hough Transform is employed on the distance image to detect fruits.

Many other methods have been developed over the years: for an accurate review of this technique, the reader should refer to [43].

2.2. Quality assessment of meals produced by industry

The assessment of the food quality produced by an industry is a crucial task needed to guarantee a good experience to the final customer. Alongside with human control of the product chain, Computer Vision systems can be used to perform the quality assessment through the automatic inspection of images.

In [17,28,38], a review of methods for food quality assessment is presented. The authors consider different acquisition systems, the features that can be employed in different tasks, as well as the machine learning algorithms used to perform the decision among the quality of the food items.

In a typical Computer Vision based pipeline for quality assessment, an image preprocessing, a feature extraction process, and a classification are performed.

Munkevik et al. [72] propose a method to check the validity of industrial cooked meals. As first step, the images of the food are segmented. Then 18 features are extracted from the segmented image, in order to capture different aspects. Specifically, the features are related to the size of the food items on the plate, to the overlapping between different food items, to the shape of the food and to the colors. Eventually, the extracted features are used to train a Self-Organizing Feature Map [55], which is employed to learn the model of a meal. In [73] the approach is refined and extended by considering more food items and employing an Artificial Neural Network (ANN) for classification purposes.

A beans quality classification system was proposed by Kiliç et al. [48] in 2007. For testing purposes, they considered a dataset of images with variable number of beans. After a segmentation stage using morphological operators, the 1st to 4th order statistics on the RGB channels of the image are computed. Three quality levels for both color and integrity of the sample were defined, but only 5 out of the 9 possible combinations were used to better separate top quality beans from medium and low quality ones. In other words, given a rating from A to C for both colors and integrity, the considered classes are AA, BB, BC, CB, CC. The classification was performed using an ANN, using 69 samples for training and 71 for validation, while the testing set is composed of 371 beans images.

The quality of pizza production has been explored by different researchers. In [29,98] methods for inspecting shapes, toppings and sauce spread in pizza production are proposed. Different features were computed for the shape, sauce and topping inspection. Specifically, to assess the quality with respect to the shape, the area ratio, aspect ratio, eccentricity, roundness have been considered. For sauce and topping the Principal Component Analysis (PCA) on the histograms computed in the HSV color space has been exploited. The food items are classified considering 5 quality levels concerning the sauce spread and topping, and in 4 quality levels with respect to the shape. The quality classification task was performed by using a set of binary Support Vector Machine (SVM) classifier (one-vs.-all) organized in a Directed Acyclic Graph (DAG). The system is trained using 120 images for the shape, 120 images for the sauce and 120 images for the topping.

Despite the quality assessment and inspection of food is not strictly related to the application domain of dietary food monitoring, we have decided to include information on this application domains such that the reader can have a better overview of what has been done in the context of food image analysis. The inspection of the food quality is usually performed in constrained settings tackling with a small number of food classes and low variabilities. Usually, simple approaches (e.g., very simple features such as shape measurement) are enough to address the problem and the results claimed by the authors are very good. This scenario is very different from the one where images of food are acquired during meals of a patient or they are downloaded from a social network. The systems for generic food intake monitoring have to deal with a higher number of food classes, mixed food, and a number of image variabilities, such as different environment illumination, different point of view in the acquisition, and different acquisition devices (i.e., different resolution, compression factors, etc.). Moreover, usually these systems have to be able to work without prior knowledge. For instance, differently than an industrial production chain where the different ingredients (e.g., to make a pizza) are known in advance, in a generic food image understanding problem there are not a priori assumptions by making the task more challenging.

2.3. Food logging, dietary management and food intake monitoring

Diet monitoring has a keyrole for the human health and can help to reduce disease risks such as diabetes. For this reason, since the 1970s, the computers have been employed to help the medical teams for dietary assessment of the patients. However, the primordial systems for food logging and intake monitoring did not use the Computer Vision; they were calculators for nutrition

factors from a predefined food list [89,111].

During the last century, despite the great steps forward in the knowledge of nutrition, there has been a dramatic increase of food-related illnesses [117]. It has been proved that food diaries are effective instrument to boost self-awareness of eating habits, and augmenting written diaries with photographs have a more effective impact on the patients. Hence, Computer Vision researchers have put effort to provide reliable tools to make the automatic detection and recognition of meals images more accurate.

Among these systems, FoodLog[1] [4,50,51,53,67] is a multimedia Internet application that enables easy capture and archival of information regarding daily meals. The goal of this framework is to assist the user to keep note of their meals and balance the nutritional values coming from different kinds of food (e.g., carbohydrates, fats, etc.). The user uploads the pictures on a remote folder, where the archive is maintained. In [51], the images containing food items are identified by exploiting features related to the HSV and RGB color domains, as well as the shape of the plate. A SVM classifier is trained to detect food images. More specifically, the images are divided in 300 blocks and each block is classified as one of the five nutritional groups defined in the "My Pyramid" model[2] (grains, vegetables, meat & beans, fruits, milk) or as "non-food". In [53] more local features are considered. Color statistics were coupled with SIFT descriptor [64] obtained with three different keypoint selection methods (difference of Gaussians, centers of grid, centers of circles). In [50] the approach has been extended, adding also a pre-classification step and the personalization of the food image estimator. In [67] the Support Vector Machine is replaced by a Naive Bayesian Classifier.

The goal of the approach proposed in [95] is to help people affected by diabetes in following their dietary prescriptions. The authors used object-related features (color, size, texture and shape) and context-related features (time of the day and user preference). Using an ANN as a classifier, the authors proved that the context information can be exploited to improve the accuracy of the monitoring system.

Food recognition and 3D volume estimation is the goal of the work by Puri et al. [85]. The images, taken under different lighting conditions and poses, are normalized by color and scale, by means of dedicated calibration patterns placed besides the food items. They use an Adaboost-based feature selection method to combine color (RGB and LAB neighborhood) and texture (Maximum Response filters) information, in order to perform a segmentation by classification of the different food items in a dish. The final classifier is obtained as a linear combination of many weak SVM classifiers, one for each feature. Moreover, they reconstruct the 3D shape of the meal using dense stereo matching, after a pose estimation step performed using RANSAC [35].

Chen et al. [24] aim to categorize food from video sequences taken in a laboratory setting. The dishes are placed on a turntable covered with a black tablecloth. They consider an elliptical Region-of-Interest (ROI), inside which they first extracted MSER [68], SURF [9] and STAR [2] features. Since these detectors work on monochrome images, a color histogram in the HSV color space is computed inside the ROI, in addition to the aforementioned detectors, in order to capture the richness of food images in terms of colors. The images are then represented using the Bag of Words paradigm; they create a vocabulary with 10,000 visual words using K-means clustering and subsequently each data point is associated with the closest cluster using the Approximated Nearest Neighbor algorithm. For each image, hence, a Bag of Word representation and the color histogram in the HSV color space are provided. The goal is to classify the dish in a specific frame. The authors propose to compare the frame under examination with a frame already classified, in a retrieval-like fashion. To do so, a similarity score is computed separately for the Bag of Words representation and for the color histograms. For the first representation, the term frequency-inverse document frequency (tf-idf) technique is employed; for the latter one, the correlation coefficient between the $|L_1|$-norm of the histograms is computed. The two scores are then combined with different weights to obtain the global score for the considered frame. Since the calories for the reference dish are known, the similarity is able to coarsely quantify the difference of food in the two frames.

The 3D reconstruction is used in [26] for volume computation. A disparity map is computed from stereo pairs, and hence a dense 3D points cloud is computed and aligned with respect to the estimated table plane using a specific designed marker. The different food items present in the image are assumed to be already segmented. Each food segment is then projected on the 3D model, in order to compute its volume, which can be defined as the integral of the distance between the surface of each segment and either the plate (identified by its rim and reconstructed shape), or the table (identified by the reference pattern).

Food consumption estimation is also addressed in [62]. The authors propose a wearable system equipped with a camera and a microphone. When the microphone detects a chewing sounds, the Computer Vision part of the framework is activated. The algorithm tries to identify keyframes containing food by using simple features such as ellipse detection and color histograms. The first step is the ellipse detection. When the ellipse is found, it is split into four quadrants and, for each quadrant, the color histogram is computed in the C-color space [20]. Then, the difference between the histograms computed over subsequent frames is computed to evaluate the food consumption.

2.4. Food classification and retrieval

The approaches we have reviewed so far aim to solve specific food-related task, such as fruit recognition, quality assessment or food logging for dietary management. All of these application domains share a key component related to the recognition of the food. In the last years, this aspect has been considered by many computer vision researchers thanks to the increasing availability of large quantity of image data in Internet and the explosion of posts portraying food in social media. This led to the proliferation of datasets with a consistently increasing number of classes and samples. Table 1 summarizes the main features of the publicly available datasets reported in the state-of-the-art works in the last

Table 1
Food image datasets. C, classification; R, retrieval; CE, calorie estimation.

Dataset	Related works	Classes	Img per class	# of img	Task
UEC FOOD 100	[41,46,47,69,70,88,114,115]	100	≈ 100	9060	C
PFID	[12,22,32,74,108,116,119]	101	18	1818	C/R
FRIDa	[37]	8	ND	877	CE
NTU-FOOD	[23]	50	100	5000	C
ETHZ Food-101	[16]	101	1000	101,000	C
UNICT-FD889	[30,31,66]	899	3/4	3583	R
FooDD	[84]	23	ND	3000	CE
UPMC Food-101	[112]	101	1000	101,000	C
CAS dataset	[91,40]	ND	ND	117,504	C

[1] http://www.foodlog.jp
[2] http://www.mypyramid.gov/

years.[3]

In order to recognize food depicted in images, two computation strategies can be usually considered: classification and retrieval. In both cases the task is to identify the category of a new food image observation on the basis of a training set of data. The main difference between the two approaches stays in the mechanism used to perform the task. In the case of classification the training set is used just to learn the decision function by considering the representation space of the images. Hence, the training images are represented as vectors in a feature space through a transformation function (e.g., Bag of Visual Word approach by considering SIFT or Texton features [7,58]) whereas a learning mechanism is used to train a classifier (e.g., a Support Vector Machine) to discriminate data belonging to different classes. After that, the training dataset is discarded and a new observation can be classified by considering the employed feature space and the trained classification model. In the case of retrieval, the training set is maintained and the identification is performed comparing the images through similarity measures (e.g., Bhattacharyya distance) [13] after their representation in the feature space.

In [114], a framework for food classification of Japanese food is proposed. The approach is trained and tested on a dataset with 50 classes. Three kinds of features are extracted and used: (a) Bag of SIFT; (b) Color Histograms; (c) Gabor Filters [65]. The keypoint sampling strategy on which the SIFT descriptor has been computed is implemented with three different ways: using the DoG approach, by random sampling and using a regular grid. To compute Color Histograms, the images are first divided in 2×2 regions, and for each region a 64-bin RGB histogram is calculated. The region-based histograms are then concatenated into a 256-bin. In a similar way, the images are split into 3×3 and 4×4 blocks to compute Gabor Filters responses. The employed Gabor filters take into account four different scales and six orientations, so for the whole image a 216 or 384-dimensional vector arises as result of the extraction step. While Color Histograms and Gabor Filters provide a representation of the images by themselves, SIFT keypoints are clustered by generating two different vocabularies with 1000 and 2000 codewords and the images are represented using the Bag of Words paradigm. Summing up, for each image 9 different representations are provided, one coming from the Color Histograms, two from the Gabor Filters with different blocking schemes and six from the combination of sampling strategies and vocabulary size for SIFT features. Classification is performed using a Multiple Kernel Learning SVM (MKL-SVM) [104]. In [41] the dataset is extended up to 85 classes, and 8 variants of Histogram of Oriented Gradients (HOG) [25] are introduced as new features. Moreover, the χ^2 kernel is employed as a kernel function in the MKL-SVM. An extended version of the dataset, containing 100 food items, has been used in [70] where candidate regions are identified using different methods (whole image, Deformable Part Model (DPM) [34], a circle and the segmentation method proposed in [27]). The final segmentation arises by integration of the results of the aforementioned techniques. For each candidate region, four sets of features are computed: Bag of SIFT and Bag of CSIFT [1], Spatial Pyramid Representation [59], HOG and Gabor Filters. Then a MKL-SVM is trained for each category, and a score is assigned to every candidate region. The experiments are conducted on images containing both single and multiple food-items. In successive work [69] the same approach is used, but the scores assigned by the classification algorithm are re-arranged applying a manifold learning technique to the candidate regions. The dataset used in [69,70] is called UEC FOOD 100 and is an extension of the dataset presented in [41,114]. On this dataset, other approaches have been tested. For instance, pre-trained Convolutional Neural Networks (CNN) [57] are used in [47] for feature extraction. The CNN features are coded using the Fisher Vectors technique [92], and then the classification is performed by means of SVM. Ravì et al. [88] exploited jointly different features in a hierarchy to obtain real-time food intake classification. The hierarchy of features encodes, in some way, the complexity of the images: on simple classes, the classification will rely on the features at the first level, while on more complex classes more features will be used. To represent the images, the Fisher Vector [80] technique is employed, and PCA is applied as in [81]. To perform classification, a linear SVM is trained using the one-vs.-rest strategy. The UEC FOOD 100 has been extended to 256 categories (UEC FOOD 256) in [46] using a so-called "foodness classifier" and transfer learning on images coming from crowdsourcing. UEC FOOD 100 and UEC FOOD 256 have been employed by Yanai et al. [115] to fine tune a pre-trained deep convolutional neural network (pre-trained with 2000 categories in the ImageNet).

Another dataset used in the literature is the Pittsburgh Food Image Dataset (PFID) [22]. This dataset is composed of 4545 still images, 606 stereo pairs, 303 videos for structure from motion (360° videos), and 27 privacy-preserving videos of eating events of volunteers. The images portray 3 instances of 101 food items, bought in 11 different fast food chains. In [22], a baseline for future experiments is provided. The authors use color histograms and Bag of SIFT features to train a multi-class SVM. In [116], an ingredient based segmentation is performed using a Semantic Texton Forest [94]. Hence, pairwise statistics of local features are computed on the segment connecting two points, and specifically: (a) orientation; (b) midpoint; (c) between-pair; (d) distance. Moreover, two joint features are considered (Distance + Orientation and Orientation + Midpoint). A SVM with a χ^2 kernel is employed for classification purpose. The PFID is also used for calories estimation in [108]. SIFT are extracted and a cosine-based distance function is used for matching. Rankings on food categories can be obtained in two ways: (1) a ranking based matching, based on top T items of each frame-based rankings; (2) a count-based matching based on sum of keypoint matching counts over all video frames. Zong et al. [119] locate the keypoints using the SIFT detector, applying the Local Binary Pattern (LBP) [3]. Then they employ a BoW model, using a codeword filtering function to select the most discriminative words in the vocabulary. Dictionary creation is performed in a class-based manner. To provide spatiality, the shape context descriptor [11] is calculated on the image space, considering the words as keypoints. The images are classified by means a cost function which takes into account the Bhattacharyya distance and the shape context matching cost. Nguyen et al. extended the previous mentioned approach introducing the Non-Redundant Local Binary Pattern (NRLBP) [74] and propose two strategies to classify the images: the first makes use of a SVM, the second is based on a cost function. Farinella et al. propose two different approaches on the PFID: one [32] is based on the representation of food images as Bag of Textons. Textons are computed using the responses of MR4 filters, then clustered in a class-based fashion obtaining a visual vocabulary. In the other approach [33] SIFT and SPIN [58] features are computed over a dense grid, and multiple runs of the K-means algorithm are performed separately for SIFT and SPIN. The vocabularies obtained in output are used as input for an Expectation–Maximization based consensus clustering technique [102]. In both approaches, SVM is used as a classifier. The method proposed in [12] combines different descriptors calculated on patched centered on the keypoints detected by the Harris–Laplace detector. For each feature, a visual

[3] Some other datasets have proposed in the literature [5,14,26,39]. However these datasets have been not included in Table 1 because they are not publicly available. More information on these datasets can be found at URLs http://www.tadaproject.org and http://gocarb.eu.

codebook with 1000 words is built, and for each set a Gaussian kernel is computed. The resulting kernels are used as input to train a Sequential Minimal Optimization (SMO) MKL-SVM.

Bosch et al. propose a method for food identification based on global and local features [15]. As global features, they use (1) 1st and 2nd moment statistics computed on the color channels of the image; (2) entropy statistics; (3) predominant color statistics. As local features, they consider small patches and calculate the following features: (1) local color statistics; (2) local entropy color; (3) Tamura features; (4) Gabor filters; (5) SIFT descriptor; (6) Haar wavelets; (7) Steerable filters; (8) DAISY descriptor [101]. While the global features are used as input for a SVM with a RBF kernel, the Bag of Words approach is used with local features. Classification, in this case, is done using a Nearest Neighbor algorithm. This approach was tested on a subset of the dataset created at Purdue University [14]. The Purdue Food Dataset is an extension of the USDA Food and Nutrient Database for Dietary Studies (FNDDS), created having in mind the goal of augmenting *"an existing critical food database with the types of information needed for dietary assessment from the analysis of food images and other metadata"*.

Rahmana et al. in [87] present a dataset with 209 acquired using a iPhone3, to be used for retrieval purposes. They propose, as a baseline, Gabor filter variants to ensure scale and rotation invariance to their algorithm. However, they also perform a classification task, grouping the categories in 5 groups (Bread, Cereal, Veg, Fruit, Fast).

Another system for mobile food recognition is proposed in [45]. Here, color histograms on the RGB space are computed on 3×3 blocks and a dictionary with 500 visual words is built on SURF descriptors, to enclose local features in the general description of the image. To classify the images, a linear SVM with explicit embedding [107] is employed. It is interesting to note that the authors propose a system able to suggest the direction to which the camera should be moved, in order to improve classifier accuracy. Also, a dataset with 50 categories containing 100 images each is presented.

A Computer Vision system for Chinese food identification is presented in [23]. The authors work on a database composed of 50 categories of ready-to-eat Chinese meals, with 100 images per category. On each image, the following features are extracted: (1) SIFT with sparse coding; (2) LBP with multi-resolution sparse coding; (3) color histograms; (4) Gabor textures. A SVM is trained for each feature using 5-fold cross validation; the fusion is done using the Multi-Class AdaBoost algorithm. Marginally, the authors propose also a quantity estimation technique using Microsoft Kinect, but this approach has been tested only on a single item of "hot & sour soup".

A food recognition system integrated on a chopping board is the topic of the work by Pham et al. [82]. In this work, an imaging system composed of a matrix of optical fibers is placed under an appropriately prepared chopping board. The sensor acquires the image and afterwards a 64-dimensional color histogram and a 64-dimensional vector of Bag of SURF features are computed. The algorithms used to classify the images are kNN and SVM. The training and testing phases make use of a dataset composed of 1800 pictures of 12 food ingredients.

Random Forest (RF) [100] is used in [16] for mining discriminative regions. Superpixels are generated from the images and dense SURF and color histograms are computed and encoded using Fisher Vectors [92]. These descriptors are supplied to the RF for training. Once the RF has been trained, the leaves constitute the set of candidates for the components. Using a probability-based distinctiveness function, the most discriminative leaves are selected. Hence, a linear binary SVM is trained for each class, using the samples lying in the most discriminative leaves as positive samples and hard negative samples to speed up the learning process. Alongside with the algorithm, the authors present a novel dataset, called Food-101, composed of 1000 images for each one of the 101 most popular dishes on http://www.foodspotting.com. In [112] Xin et al. propose UPMC Food-101, a new dataset of 101,000 images to address the recipe recognition problem. This dataset includes the same 101 categories of Food-101 and 1000 new images for each one. Google Image Search engine is exploited to retrieve 1000 images for each of the categories, moreover for all the images the related HTML textual description is collected. To benchmark the dataset, Bag of Words and CNN approaches are employed and textual information are embodied to improve classification performance.

Other food dataset include images and related geocontext information, such as GPS coordinates, restaurant where the dish is cooked and so on. Herranz et al. [40] propose a probabilistic model to combine locations, restaurants and visual features by exploiting a reduced set of the dataset collected by Ruihan et al. [91] from Institute of Computing Technology, CAS. To each of the restaurants are associated the related geographical coordinates to uniquely locate it and a menu which includes at least 3 dish categories. Then, for each of these categories, more than 15 images are included.

The UNICT-FD899 [30] has been acquired by users with a smartphone in the last four years during meals (i.e., iPhone 3GS or iPhone 4) in unconstrained settings (e.g., different backgrounds and light environmental conditions). Each dish has been acquired through 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 the authors 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 since it depends on "how much" variability (both geometric and photometric) the system can tolerate. For instance, some approaches define the near duplicate of an image as the images obtained transforming the original by means of slight common editing, such as contrast equalization, scaling, and cropping. Other techniques (e.g., [8,42]) consider as near duplicate the images of the same scene but with different viewpoints and illumination. In [30], the authors consider this last definition of near duplicate food images to test different image representations on the proposed dataset. Then, they benchmark the proposed dataset in the context of NDIR by using three standard image descriptors: Bag of Textons [105], PRICoLBP [86] and SIFT [64]. Results confirm that textures and colors are fundamental properties. The experiments performed point out that Bag of Textons representation is more accurate than the other two approaches for NDIR.

A comparative analysis on features and classifiers is the core of [39]. The authors test several features, basically related to three aspects (color, texture, local regions) and two classifiers (kNN, Vocabulary Tree [75]) on a novel dataset composed of 42 classes, with a total of 1453 images.

3. Proposed dataset

The research in the field of Computer Vision needs a large amount of organized data in order to test the algorithms for task such as detection, recognition and so on. Unfortunately it is not always easy to collect meaningful data for the different tasks. In particular, in the case of food classification and retrieval for food intake monitoring, it can be very difficult to build a representative

dataset. Actually food comes in many forms and it is naturally deformable, so a representative dataset should contain different variabilities. Moreover it is important whether the data are acquired in real meal scenario rather than collected from the web, where images of food are usually posted to show the best aspect of a dish and, some time, are post processed for this scope. As discussed in previous section, different datasets have been proposed in the literature. However, most of them are build by collecting images downloaded from Internet [16,69,70,112], contain food images acquired with constrained laboratory settings [22,84] (e.g., variabilities related to light conditions and background are not considered), consider very simple food plates [37,84], or include only food from one nationality [23].

Considering the aforementioned limitations of the datasets currently available for testing purposes, in our previous work we have introduced the UNICT-FD889 dataset [30], which is a collection of food images acquired during real meals, useful for the study of the image representation to be used for food image retrieval purposes. This dataset is available online at the URL http://www.iplab.dmi.unict.it/UNICT-FD889/.

In this paper we extend the UNICT-FD889 in two different aspect. Specifically, we include more dishes as well as the labels related to the following 8 categories: *Appetizer, Main Course, Second Course, Single Course, Side Dish, Dessert, Breakfast, Fruit*. Images depicting mixed food (e.g., fish with salad) are labeled with multiple labels (e.g., Second Course and Side Dish). The proposed dataset is composed of 4754 images related to 1200 distinct dishes of food of different nationalities (e.g., English, Japanese, Indian, Italian, Thai, etc.). Each plate has been acquired multiple times (four in the average) to guarantee the presence of geometric and photometric variabilities. All the food photos have been taken in the last five years during real meals by using a mobile camera in unconstrained settings, such as different backgrounds and light conditions. This is a significant characteristic which is mandatory to test food understanding algorithms on real scenario data. At the best of our knowledge, all the other state-of-the-art datasets, except UNICT-FOOD889, include photos retrieved by web in semi-automatic way or acquired under laboratory settings. The mobile cameras used for the acquisition are iPhone 3GS, iPhone 4 and iPhone 5 with a max resolution (e.g., equals to 2448 × 3264 for the iPhone 5). The UNICT-FD1200 dataset is thought to help research in the field of food-understanding with the aim to study the best representation to use for food images. It can be used to test food image retrieval as well as food classification by considering the aforementioned classes. Fig. 1(a) shows image samples randomly selected from the UNICT-FD1200 dataset, whereas Fig. 1(b) can be useful to assess the multi-view acquisition as well as geometric and photometric variabilities. The UNICT-FD1200 dataset is available for research purposes at the URL http://www.iplab.dmi.unict.it/UNICT-FD1200/.

4. Image representation

To benchmark the proposed dataset we employed three different types of hand-crafted features: SIFT [63], PRICoLPB [86], Texton [44,60,105]. We decided to include SIFT features because of the good results obtained for Computer Vision tasks in the last years. We exploit SIFT to represent the food images as a set of features to be used together with a matching scheme during classification and retrieval, as well as to build a representation based on the bag of words paradigm as recently proposed in [5]. The PRICoLBP features have been included into the comparison since they have been recently proposed and tested on food dataset [86]. Building on our previous works [30,32] we considered Bag of Textons representation because of its capability to describe texture information. Despite the simplicity of the Bag of Textons representation, it has obtained good results in the context of food classification and retrieval [30,32]. Finally, we propose a new image representation based on the perceptual concept of Anti-Textons [103,109,110] to encode spaces between Textons. The proposed image representation outperforms all the others approaches. It is important to note that all the aforementioned representation methods are invariant or partially invariant to the illumination. Texton-based representations perform two normalization steps to strongly reduce the illumination effect (pre-processing normalization at mean 0 and variance 1 and post-processing normalization according with Weber's law [105]). PRICoLBP is a variation of LBP, which is invariant to global illumination changes [77]. Concerning SIFT, in the extraction process,

Fig. 1. (a) A sample of the UNICT-FD1200 dataset. (b) Three elements for 24 classes of the UNICT-FD1200 dataset. The variabilities for each class are evident.

the last normalization step employed to build the descriptor guarantees linear and non-linear illumination invariance [63]. Considering descriptors with illumination invariance property is mandatory because images into the proposed dataset have been acquired under different light conditions. In the following subsections we detail all the aforementioned image representation.

4.1. SIFT

SIFT algorithm allows us to detect visual interest points and describes them such that the final descriptor results invariant to scale, rotation, illumination changes and partially invariant to affine distortion [63,64]. SIFT is usually extracted sparsely or densely [59,76] from gray-scale or color images [5]. After SIFT extraction the set of descriptors can be used for matching purposes or to build an image representation based on the Bag of Words (BoW) paradigm. We tested both representation approaches in our experiments. To build the BoW SIFT representation we use a dense regular grid to compute the SIFT descriptor of a patch. At this point, a clustering algorithm is used to quantize descriptors space extracted on training images to create a visual words vocabulary. To represent image, each point of the regular grid is associated to the nearest visual word. When the visual vocabulary is computed, each image in the training and test set can be represented as a distribution of visual words. In our experiments a grid with spacing of 8 pixel and a patch of 16×16 is used during dense sampling on the three RGB channels. *K*-means clustering is exploited to compute the visual words vocabulary with different sizes. The SIFT descriptors are computed independently for each color channel. A Bag of SIFT is obtained for each color channel and the three visual word distributions are concatenated in a unique descriptor.

SIFT has been also tested for matching purposes. In this case the SIFT of a query image are matched to the keypoints of all the images in the training set. The query image is associated to the image of the training dataset with the highest number of matchings. Since the SIFT matching algorithm assigns a score to each matched point based on the quality of the match, we also consider to inversely weigh each matched keypoints by taking into account the similarity between the SIFT descriptors of the matched keypoints. We consider both gray and color domains. In the RGB domain the SIFT features are extracted and matched independently on each color channel, then the sum of the matchings for the three channels is considered to compute the similarity index. In our experiments VLFeat [106] library has been used to extract SIFT keypoints.

4.2. PRICoLBP

Pairwise Rotation Invariant Co-occurrence LBP (PRICoLBP) descriptor focuses on encoding spatial co-occurrences and pairwise orientations of the well-known Local Binary Pattern (LBP) features [77]. It preserves the relative orientations of LBP features pairs in order to obtain rotational invariance. To compute the PRICoLBP descriptor, we employed the original implementation provided by the authors which is available online.[4] We exploited PRICoLBP on both gray and color domains. In our experiments we set the radius 2, neighbor points equal to 8 and the template equals to 2. This results in two kinds of PRICoLBP descriptors of 1180 and 3540 components to represent gray and color images respectively.

[4] http://qixianbiao.github.io/

4.3. Bag of Textons

Textons have been introduced by Julesz as the putative unit for the visual perception during pre-attention processing. A computational model for Textons can be obtained through the responses of the gray or the color image to a bank of filters [60]. Filter responses of the training images are quantized through clustering procedure. Hence, each cluster centroid can be considered a Texton and a set of them compose a visual codebook [105]. To represent images each filtered pixel is associated with one of the Textons in the codebook considering a similarity metric (in this paper we use L^2 distance). Finally, the histogram of the distribution over the different Textons of an image is built. We considered different configurations involved in the Textons extraction pipeline to highlight which bank of filters, color domain, normalization procedure and size of the vocabulary are the most appropriate in the application context discussed in this paper. As similarity measure between two Texton distributions, we use the χ^2 distance. In the following subsections we detail the different 49×49 filter banks tested in this paper (LM, MR8, MR4, Schmid) and LINC normalization strategy.

4.3.1. Standard filter banks

The Leung–Malik (LM) filters bank [60] consists of 48 filters (Fig. 2(a)), among which smoothing filters, edge detectors and bar detectors. There are 4 Gaussian filters, first and second derivatives of Gaussian at 6 orientations and 3 scales, 8 Laplacian of Gaussian filters. The scale σ of the Gaussian functions is between 1 and 10.

The Maximum Response 8 (MR8) filters are derived by the Root Filter Set (RFS) which consists of 38 filters similar to the LM filters [60]. After the convolution with the 38 filters only 8 responses are selected. As in LM filter bank, MR8 contains filters with different scales and orientations. However, only the maximum response is selected across orientations of a specific filter (e.g., edge filter) in order to achieve rotation invariance. The 38 filters consist of a Gaussian filter and a Laplacian of Gaussian filter with scale $\sigma = 10$, first derivative of Gaussian filters at 3 scales and 6 orientations, second derivative of Gaussian filters with the same scales and orientations of the first derivative of Gaussian filters.

The Maximum Response 4 (MR4) is a subset of the MR8 filters which is built considering a single scale for the edge filters and bar filters [60]. Hence the filter bank to be applied contains 14 filters but 4 responses only are selected.

The Schmid filter bank [93] consists of 13 isotropic filters described by the equation:

$$F(r, \sigma, \tau) = F_0(\sigma, \tau) + \cos\left(\frac{\pi\tau r}{\sigma}\right) e^{-\frac{r^2}{2\sigma^2}} \quad (1)$$

where σ is the filter scale in pixel and τ a value which is proportional to the number of concentric rings in the kernel. $F_0(\sigma, \tau)$ is added to obtain a zero DC component for the filter with (σ, τ) pair taking values (2,1), (4,1), (4,2), (6,1), (6,2), (6,3), (8,1), (8,2), (8,3), (10,1), (10,2), (10,3) and (10,4). Those filters are shown in Fig. 2(b).

4.3.2. Local Intensity-normalized Colors (LINC) filter banks

To achieve invariance to local intensity changes the Local Intensity-normalized Color procedure has been proposed in [19]. The authors proposed to use opponent color space and a normalization of the filter responses. Specifically, for each filter response the Gaussian filter response for first channel at the same scale σ is exploited in order to obtain local intensity normalization. Despite LINC normalization has been proposed for MR8 filter bank, we have employed the procedure considering both MR8 and Schmid bank of filters.

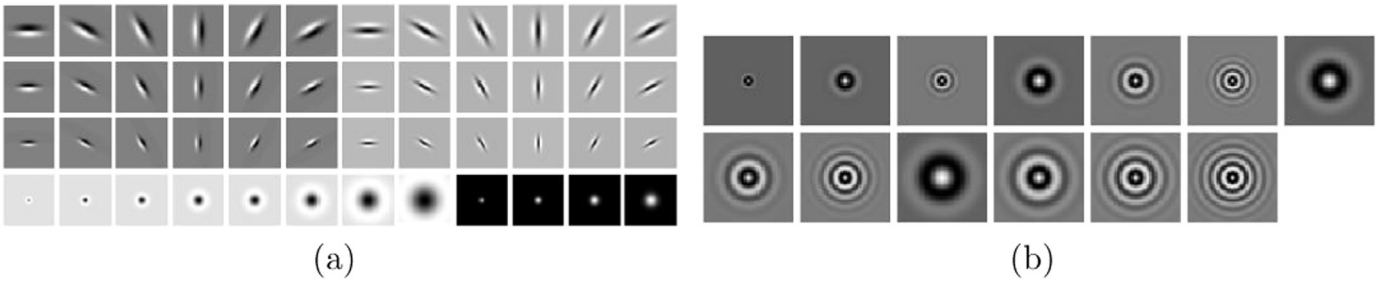

Fig. 2. (a) The 48 filters of Leung–Malik filter bank; (b) the 13 Schmid isotropic filters.

4.4. Anti-Textons representation

Bag of Textons representation has shown good performances in the context of food classification and retrieval [32,30]. However, this representation does not take into account the spatial relation between the visual words. This is because, in the bag of words paradigm, only the first order statistics of the visual words are used as image descriptor. We propose to exploit the spatial information around each Texton to build a more discriminative representation of food images. This idea is supported by the study presented in [103] where the authors defined the concept of Anti-Textons as the space between two Textons. Anti-Textons concept has been introduced in the literature by Williams and Julesz in [109,110] for the purpose of texture segregation (i.e., segmentation). At the best of our knowledge there is only a single attempt to find a suitable computational procedure to compute Anti-Textons for texture segmentation [103]. Differently from previous works we introduce a computational approach to compute Anti-Textons distribution for the purpose of image representation. The proposed method assumes that a textons vocabulary with N codewords has been obtained from the set of training images. Once the visual vocabulary is obtained, the Anti-Textons computation pipeline shown in Fig. 3 is applied to represent an image. The Anti-Textons representation is computed considering the following steps:

- The Textons map for an image I is computed. For each pixel the Textons map stores the corresponding Texton ID.
- For each Texton with ID i ($i = 1, \ldots, N$) a binary map is produced. The binary map B_i for the Texton i contains 1 in the position where the Texton i occurs and 0 in all the other positions. At this stage, N binary maps are computed.
- The Distance Transform [71,90] for each map B_i is computed. This results in a "saliency" map where the points close to the Texton i are less salient than the further ones. We use this saliency map to establish how much each Textons into the Textons map can be considered Anti-Textons with respect the Texton i. Each saliency map is normalized by dividing by its max value. We refer to the normalized map for Texton i with the symbol D_i. The maps D_i are inverted by computing $E_i = \mathbf{1} - D_i$. This is the way we encode the space between two Textons of the same class i.
- As next step each map E_i is used to weight the original Texton map to obtain the final Anti-Textons distribution for Texton i. In particular, we compute the histogram H_i as follows: $H_i(k) = \sum_{\mathbf{x}} E_i(\mathbf{x}) B_k(\mathbf{x})$, where $\mathbf{x}$ is the coordinate in the Texton Map. The normalized histogram $\widetilde{H}_i$ (at sum 1) represents the Anti-Textons distribution for the Texton i.
- Finally, we average all the N computed histograms $\widetilde{H}_i$ in order to produce the Anti-Textons representation for the image I.

The experiments confirm that the proposed Anti-Textons

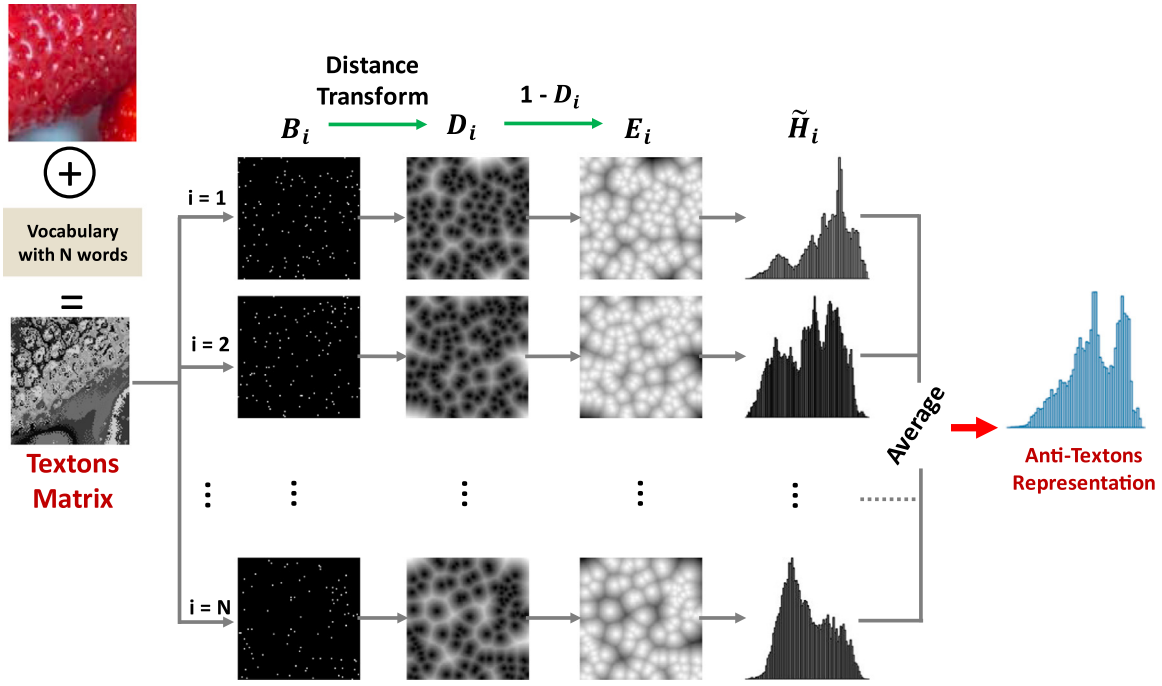

Fig. 3. Anti-Textons representation pipeline (see text for details).

representation outperforms the other representation.

5. Experimental settings and results

In this section we describe the settings and the quality measures used to compare the image representations presented in the previous section and tested on the dataset proposed in Section 3. We performed both, retrieval and classification tests. For retrieval purpose, all the 4754 images of the UNICT-FD1200 dataset have been resized to 320×240 pixels. For a proper evaluation of the representation methods, we performed the experiments three times with different training sets and test sets. All results are obtained by averaging among the three different tests. All the representation approaches are compared by using the same training set and test sets. To build a training set we selected a single image for each of the 1200 dishes. Hence, a training set is composed of 1200 different images. The intersection between the three training sets is empty. For each test set, we used the rest of the images. The dataset as well as details useful to properly replicate the experiments with the considered training and test sets are available at URL http://www.iplab.dmi.unict.it/UNICT-FD1200/. For each image representation, the results are obtained by averaging over the three runs. In the case of the retrieval a run consists in a group of queries composed of the test images for which we need to find the corresponding image in the training set. The retrieval performances are measured using the quality metric $P(n)$ which is based on the top-n criterion:

$$P(n) = \frac{Q_n}{Q} \tag{2}$$

where Q is the number of queries (test images) and Q_n the number of correct queries among the first n retrieved images. In this case $P(1)$ results in the classification accuracy measure of the system. As index to describe the whole retrieval result we decided to use the Mean Average Precision (MAP) described in [83].

For classification purposes, we consider the same three training and test sets employed for retrieval purpose and a $1-NN$ classifier with χ^2 distance. Because the images can have multiple labels (up to 2 labels), two performance metrics have been considered: as a first measure we considered intersection between the labels of the query image and the labels of the nearest retrieved images (according to the $1-NN$ criteria). If the intersection is not empty, we count a positive match. Only the overall accuracy is computed in this case. For the second classification test, we removed all the multi-labeled images. In this way, the training set is reduced from 1200 to about 965 images and the test set from 3479 to 2799. With a single label, we are able to build a standard confusion matrix for evaluation purpose. In the following subsections we detail both the performed experiments and the obtained results.

5.1. Global Textons vs. Class-Based Textons

Bag of Textons representation obtained in two modalities has been tested: class-based and global. For the class-based representation we consider each image in the training set as a class because it is related to a specific plate. Then, 10 Textons per image have been extracted by using K-means algorithm to quantize the space related to the considered categories. Hence, the vocabulary can be build by collecting all the extracted Textons. Since our training set is composed of 1200 images, the vocabulary contains 12,000 visual words. In the global approach all the filter responses of the training set are considered to build the final vocabulary through K-means clustering with $K=12{,}000$. We have performed several test by using MR4 filter banks in gray domain and different vocabulary sizes for the global approach. The results (Table 2) show that there is no meaningful difference between the class-based approach and the global one. Since the construction with the global approach allows us to perform tests at varying of the final vocabulary in a simple way, we have chosen this modality to build the visual codebook for the all other experiments presented in the paper.

5.2. Gray Textons vs. Color Textons

As a next experiment, we decided to compare Textons representation in gray domain with respect to the one obtained considering RGB domain. To this aim, we choose to apply the MR4 filter bank to each color channel and then concatenate the responses obtained for different channels. Hence, considering the MR4 filters we obtained features in 4-dimensional space for the gray domain and features in a 12-dimensional space for the color domain. The $P(n)$ graph in Fig. 4 shows that a great improvement has been achieved by using color information. For this reason, we guess that the color information is critical for a good representation of food images. Considering $P(1)$, which correspond to the recognition accuracy of the system, the gray representation obtains 28.94% whereas considering color domain an accuracy of 68.14% is obtained.

5.3. SIFT based representation

We test SIFT descriptor in both, gray and RGB color domains. To retrieve images, we have used two similarity measures. The first one is based on the number of matched points, while in the second one each matching is weighted by taking into account the matching quality score. The approach with weighted measure outperforms the one where only the number of matched points is considered. Also in this case, the plots in Fig. 5 and Table 3 show that the descriptors in color domain outperform the gray ones for both the SIFT measures employed. Considering the weighted measure in color domain, we obtained the best accuracy for SIFT based representation, that is 63.52%. Nevertheless, this result does not outperform the previous results obtained with MR4 filter bank in color domain.

5.4. PRICoLBP based representation

These descriptors can be described as a histogram of CoLBP pattern to encode textures in a rotational invariant way. Since PRICoLBP has been used for food classification with promising results [86], we take into account it in the comparison. Result is presented in Fig. 5. PRICoLBP in color domain is better than PRICoLBP in gray domain. However once again the best results are still obtained using Bag of Textons approach with MR4 filter bank and 12,000 visual words in RGB color domain. Hence, we decided to focus on Bag of Textons representation for the next experiments.

Table 2
Accuracy and Mean Average Precision for Global Textons vs. Class-Based Textons by using MR4 filters in gray domain.

Representation		Accuracy (%)	mAP (%)
12,000	Textons (MR4) – Gray – Class Based	**29.16**	**36.93**
12,000	Textons (MR4) – Gray – Global	28.94	36.56
6000	Textons (MR4) – Gray – Global	28.56	36.39
3000	Textons (MR4) – Gray – Global	28.30	36.06
1500	Textons (MR4) – Gray – Global	28.41	36.33
750	Textons (MR4) – Gray – Global	28.16	35.99
375	Textons (MR4) – Gray – Global	27.73	35.58

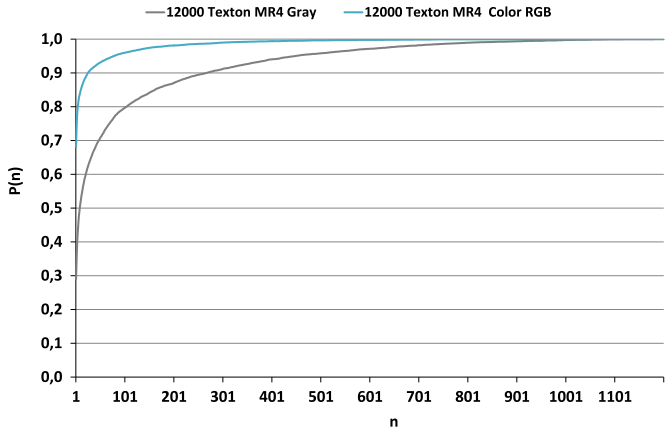

Fig. 4. $P(n)$ curves related to Gray Textons and RGB Color Textons. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

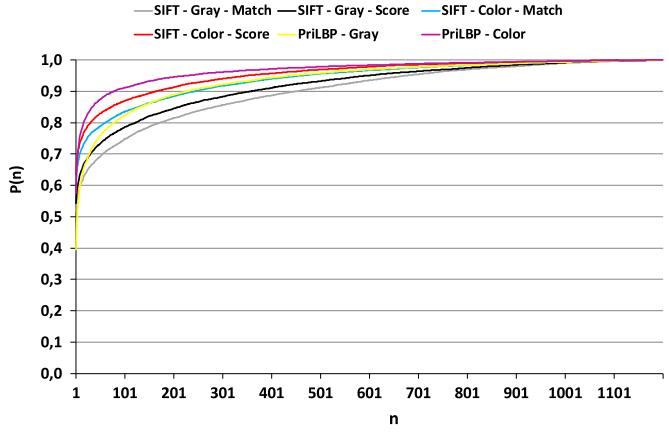

Fig. 5. $P(n)$ curves for SIFT matching approaches, SIFT matching with weighted scheme approach and PRICoLBP features in gray and RGB color domains. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

Table 3
Accuracy ($P(1)$) and Mean Average Precision for SIFT matching approach and SIFT matching with weighted scheme approach.

Representation	Accuracy (%)	mAP (%)
SIFT – RGB – Score	**63.52**	**67.30**
SIFT – RGB – Match	61.15	64.46
SIFT – Gray – Score	54.26	57.73
SIFT – Gray – Match	51.67	54.78

5.5. Dimension of the visual vocabulary

The vocabulary size is one of the parameters of the retrieval system to consider to better understand the retrieval performances when the number of visual words used to represent the food images is reduced. For this purpose, we performed tests by using MR4 filter bank in RGB color domain with different numbers of visual words: 12,000, 6000, 3000, 1500, 750, 375. In Table 4 are reported the performances of the tests where the number of visual words is reduced. Despite the retrieval accuracy decrease, no high drops are observed. This is reasonable because when the vocabulary is reduced, some discriminative visual words could be lost. Nevertheless a smaller vocabulary results in a better use of the resources (e.g., memory, CPU). However for the next comparisons, we decided to use the vocabulary size that guarantee the best performance (12,000 words).

Table 4
Accuracy ($P(1)$) and Mean Average Precision for different vocabulary sizes for the Bag of Textons representation considering MR4 filters and color domain.

Representation		Accuracy (%)	mAP (%)
12000	Textons (MR4) – RGB – Global	**68.14**	**73.99**
6000	Textons (MR4) – RGB – Global	66.60	72.65
3000	Textons (MR4) – RGB – Global	65.48	71.62
1500	Textons (MR4) – RGB – Global	63.03	69.69
750	Textons (MR4) – RGB – Global	60.53	67.50
375	Textons (MR4) – RGB – Global	56.58	63.91

5.6. Filter banks

We have performed tests considering Bag of Textons representation in RGB domain by using three more filters banks: MR8, LM and Schmid (see Section 4 for details). Tables 5 and 6 report an improvement for MR8 and Schmid filters banks with respect to MR4. On the other hand the LM filter bank has shown the worst performances. We guess this is because Leung–Malik set is not rotationally invariant. This idea is coherent with the best performance obtained with the Schmid set, which consists of 13 symmetric filters. The retrieval system employing Schmid bank of filters in RGB color domain obtained an accuracy of 75.74% and a MAP of 80.43%.

5.7. Bag of SIFT vs. Bag of Textons

For a proper comparison between Textons features and SIFT features we decided to test the Bag of Words paradigm using SIFT descriptors. Considering the work [5] where Bag of SIFT has been used for food classification purpose, to build Bag of SIFT representation we used a dense sampling on a grid with a spacing of 8 pixels. A 16×16 patch is extracted and SIFT descriptor is computed considering the three RGB channels as described [5]. To make more fair the comparison with respect to the Bag of Textons representation we repeated the Bag of Textons tests by using MR8 bank of filters, color domain, 12,000 visual words but considering the same 8×8 sampling used for SIFT descriptors. The results in Fig. 6 and Table 7 show Bag of Textons approach without spatial sampling Bag of SIFT representation. It is interesting to notice that Bag of Textons approach outperforms with a large margin Bag of SIFT also when spatial sampling is used.

5.8. Color space

Finally, we consider to change the color space used into the Bag of Textons representation to achieve further improvements in the retrieval performances. To this aim, we exploited the L*a*b* color space and the opponent color space. In the first case, we simply transform the pixel value of an image from the RGB color space to the L*a*b* one. The Textons are computed in the standard way, as described in Section 5.2. In the second case, we use the opponent color space and the normalization procedure described in [19]. In particular, it has been considered the procedure called Local Intensity-normalized Colors (LINC). The normalization is made by dividing each filter response by the Gaussian filter response (with the same σ value). We tested the MR8-LINC method proposed in [19]. Moreover we have adapted that the algorithm has been adapted to extract the LINC version of the Schmid filter banks (Schmid-LINC). As shown in Table 8, the best performance is achieved using Schmid filter banks computed in the L*a*b* color space with an accuracy of 87.44% and a MAP equal to 90.06.

Table 5
First $P(n)$ values ($n = 1…10$) related to Bag of Textons representation obtained with different filter banks in RGB domain.

Representation		$P(1)$ (%)	$P(2)$ (%)	$P(3)$ (%)	$P(4)$ (%)	$P(5)$ (%)	$P(6)$ (%)	$P(7)$ (%)	$P(8)$ (%)	$P(9)$ (%)	$P(10)$ (%)
12,000	Textons (MR4) – Color – Global	68.14	74.16	77.17	79.30	80.70	81.80	82.79	83.41	84.02	84.70
12,000	Textons (MR8) – Color – Global	71.55	77.41	80.20	81.81	83.11	84.21	85.07	85.77	86.33	86.84
12,000	Textons (Schmidt) – Color – Global	**75.74**	**80.79**	**83.16**	**84.43**	**85.68**	**86.68**	**87.49**	**88.10**	**88.68**	**89.20**
12,000	Textons (LM) – Color – Global	61.69	68.24	71.59	73.69	75.35	76.63	77.79	78.93	79.73	80.56

Table 6
Accuracy and Mean Average Precision related to Bag of Textons representation obtained with different filter banks in RGB domain.

Representation		Accuracy (%)	mAP (%)
12,000	Textons (MR4) – RGB – Global	68.14	73.99
12,000	Textons (MR8) – RGB – Global	71.55	77.00
12,000	Textons (Schmidt) – RGB – Global	**75.74**	**80.43**
12,000	Textons (LM) – RGB – Global	61.69	68.22

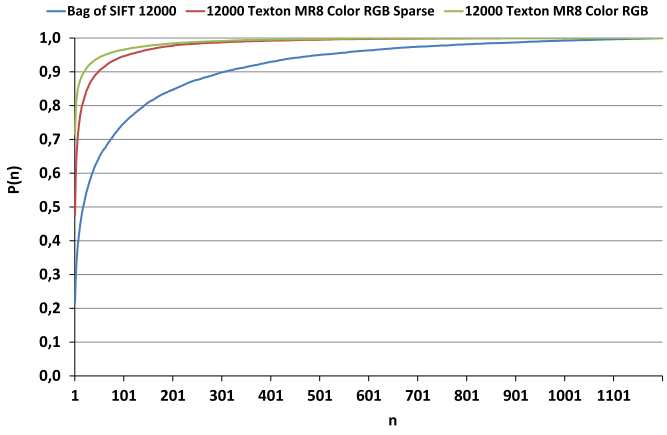

Fig. 6. $P(n)$ curves for Bag of Textons and Bag of SIFT representations in RGB domain.

Table 7
Accuracy and Mean Average Precision for Bag of Textons and Bag of SIFT representations in RGB domain.

Representation		Accuracy (%)	mAP (%)
12,000	Textons (MR8) – RGB – Global	**71.55**	**77.00**
12,000	Bag of SIFT	21.81	29.14
12,000	Textons (MR8) – RGB – Global – 8×8	47.45	57.00

Table 8
Accuracy and Mean Average Precision of Bag of Textons representation with different color spaces.

Representation		Accuracy (%)	mAP (%)
12,000	Textons (MR8) – Lab – Global	85.04	88.39
12,000	Textons (MR8) – LINC – Global	83.10	86.93
12,000	Textons (MR8) – RGB – Global	71.55	77.00
12,000	Textons (Schmidt) – Lab – Global	**87.44**	**90.06**
12,000	Textons (Schmidt) – LINC – Global	84.32	87.84
12,000	Textons (Schmidt) – RGB – Global	75.74	80.43

5.9. Visual analysis

In order to understand different discriminative capabilities among the employed representations, we performed a visual analysis of the results. For this purpose, we have included 5 representations in the analysis: Bag of Textons computed with Schmid filters in L*a*b* color space and 12,000 visual words, MR8 filters in L*a*b* with 120,00 visual words, MR8 in RGB space with 12,000 visual words and sparse sampling with step 8, Bag of SIFT, and SIFT based representation with matching scheme. Here we have reported some interesting result of one of the three tests for all the 5 representations. The complete visual comparison is available at the URL http://www.iplab.dmi.unict.it/UNICT-FD1200/. In Fig. 7 we show two queries where all the representations have a positive match. On the contrary, in Fig. 8, are shown queries where all the representations fail. Since we find out that the Schmid based representation outperforms all the other ones, we selected some queries where this representation had a positive match but all the other ones fail (Fig. 9). In Fig. 10 are shown the only 2 queries where the Schmid based representation fails whereas all the other ones have a correct match.

5.10. Result on the UNICT-FD889

To compare the results reported in [30] with the ones of this paper, we decide to perform an experiment on the UNICT-FD889 dataset using the representation which have obtained the best results on the UNICT-FD1200 (i.e., Bag of Textons with Schmid filter bank in L*a*b* color space and codebook of 12,000 words). The results in [30] are outperformed with an improvement of at least 26% for the accuracy and more than 20% for the MAP score as reported in Table 9. Recently in [66], the authors propose a Random Forest classification algorithm on the UNICT-FD889. The proposed representation outperforms also the results reported in [66].

5.11. Anti-Texton results

So far we have presented different experiments which have pointed out that Bag of Textons representation, obtained considering Schmid filters on L*a*b* domain, obtains the best performances on the UNICT-FD1200 dataset. One contribution of this paper is the introduction of a novel representation based on the concept of Anti-Textons, in order to encode spatial information in the classic Bag of Textons representation. To demonstrate the performances of Anti-Textons representation, we have compared the different filter banks to compute the Bag of Textons representation on L*a*b* space with a very small number of visual words equal to 375. As confirmed by the results reported in Table 10, Anti-Textons representation involved the best results in all of the configurations. Moreover it is interesting to note that the results obtained considering only 375 visual words with Anti-Textons representation and Schmid filters (85.01%) are close to the one obtained when 12,000 visual words are employed (87.44% – see Table 8) which has a higher cost in terms of representation storage and similarity computational time during retrieval. On the other hand, the computation of Anti-Textons representation is more expensive with respect to the original Textons based representation since it has to encode the spatial information among textons.

Fig. 7. A visual comparison where all the considered representations have a positive match.

Fig. 8. A visual comparison where all the considered representations fail.

Fig. 9. A visual comparison where only the Schmid L*a*b* representation gives a correct match.

5.12. Food classification

In previous sections we have presented different tests to assess the performances of a retrieval system at varying of features and parameters. As pointed out by our experiments, an accuracy of 87.44% and a MAP of 90.06% can be achieved on the UNICT-FD1200 dataset by exploiting Schmid Textons computed on the L*a*b* domain with a large vocabulary of 12,000 visual words. Moreover tests pointed out that the Anti-Textons representation improves the results in every configuration used. Another task we can consider in the UNICT-FD1200 dataset is classification. As detailed in Section 3 each image of the UNICT-FD1200 is labeled with one or two of the following classes: Appetizer, Main Course, Second Course, Single Course, Side Dish, Dessert, Breakfast, Fruit. To perform the classification test we have considered the best Bag of Textons representation mentioned above. For a proper evaluation, we have performed two kinds of experiments by using $1-NN$ classifier and χ^2 distance. First, to consider the fact that images can have multiple labels (e.g., Second Course and Side Dish) as evaluation criteria we count a positive match for the query i when $T_i \cap P_i$ is not empty. Let T_i be the set of the true labels for the query image i, and $_{Pi}$ is the set of the predicted labels. The average classification accuracy obtained by using Bag of Textons was 93.04%. Despite this strategy could produce too much positive match, we remark that the multi-labeled images of UNICT-FD1200 have no more than 2 labels. As second evaluation, the training sets

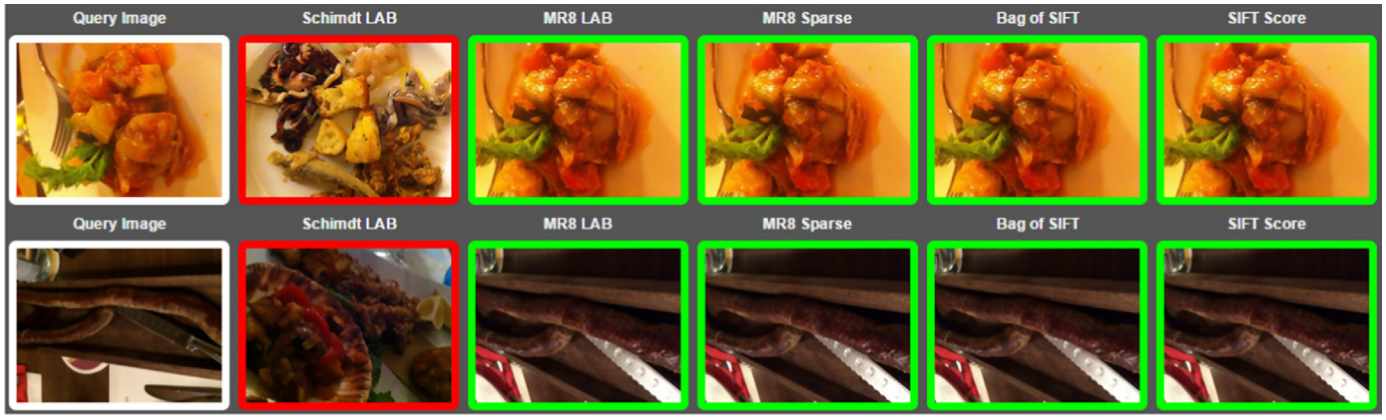

Fig. 10. The only 2 queries where the Schmid L*a*b* representation fails.

Table 9
Accuracy and Mean Average Precision of the representation used in [30] and the Bag of Textons representation with Schmid filters proposed in this paper.

Representation		Accuracy (%)	mAP (%)
8890	Textons – Gray – Global	27.70	35.98
1100	Textons – RGB – Global	60.17	67.46
	SIFT – RGB – Score	58.12	62.74
	PriCoLBP – RGB	56.33	63.52
12,000	Textons (Schmidt) – Lab – Global	**86.17**	**89.21**

Table 10
Accuracy and Mean Average Precision of the Bag of Texton and Anti-Textons representations with 375 visual words in L*a*b domain.

Representation		Accuracy (%)	mAP (%)
375	Textons – LM Lab	74.75	80.15
375	Anti-Textons – LM Lab	76.23	81.39
375	Textons – MR4 Lab	77.18	82.11
375	Anti-Textons – MR4 Lab	78.40	83.05
375	Textons – MR8 Lab	80.83	85.12
375	Anti-Textons – MR8 Lab	82.21	86.17
375	Textons – Schmid Lab	83.77	87.30
375	Anti-Textons – Schmid Lab	**85.01**	**88.22**

and test sets have been reduced by removing the images with multiple labels. Classification results for this test are reported in the confusion matrix in Fig. 11. In this case, the accuracy was 92.60%. Confusion matrix is a common statistical tool to report performance of a classification system [54]. Each of the columns of the confusion matrix shows the predicted class for a classification query, while each of the rows represents the actual (or true) class for a classification query [54].

We have also performed classification tests by using the proposed Anti-Textons representation (Schimd filters, L*a*b* color space) with a codebook of 375 elements. In order to compare properly the standard Bag of Textons approach, with the respect to Anti-Textons representation, the same test using Bag of Textons has been repeated by using a vocabulary of 375 visual words. The accuracy obtained with Bag of Textons was 90.42% whereas Anti-Textons representation has got an accuracy of 91.21% confirming its effectiveness. In Fig. 12(a) and (b) note that the Anti-Textons representation, with only 375 visual words, is able to reach an accuracy very close to the Bag of Textons with a vocabulary of 12,000 Textons (91.21% vs. 93.04%). For a proper evaluation of the proposed representations we have performed the classification experiments by employing a CNN-based method. Specifically, to perform tests we fine tuned GoogleNet [99]. Results show an

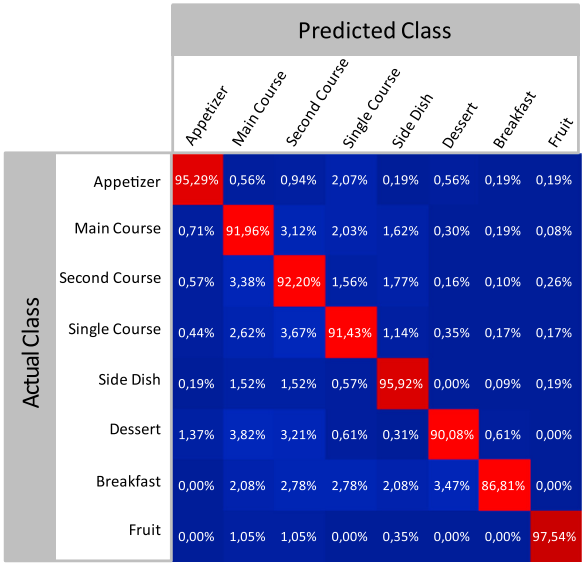

Fig. 11. The confusion matrix for the classification the tests related to food. The image representation used is the Bag of Textons with Schmid filter bank in L*a*b* color space and codebook of 12,000 words.

accuracy for the CNN method of 51.41% which is much lower than the accuracy obtained with the representations proposed in this paper. This is not a surprise, because the CNN-based methods usually need a huge amount of data for a proper training. Note that there are real cases (retrieval of a food offered of a canteen of a company) in which the construction of a big dataset is not possible (or at least has a high cost). In such cases the proposed approach for food classification can be suitable since CNN cannot be successfully applied.

5.13. *Experiments on the menu-match dataset*

To properly test the proposed approaches we performed experiments by employing another food dataset. Specifically we have considered the Menu-Match dataset introduced in [10] and we have compared the proposed approach with respect the approach described in [10]. The authors of [10] proposed a system which provides automatic classification by using priors about the provenance of food plate depicted in the acquired images. Specifically, the system is able to recognize food plates which are served in a predetermined set of restaurants. Thanks to GPS coordinates stored in the query image metadata, most of the restaurants can be discarded. The Menu-Match dataset contains 646 multi-labeled food images across 41 food categories, which have been acquired

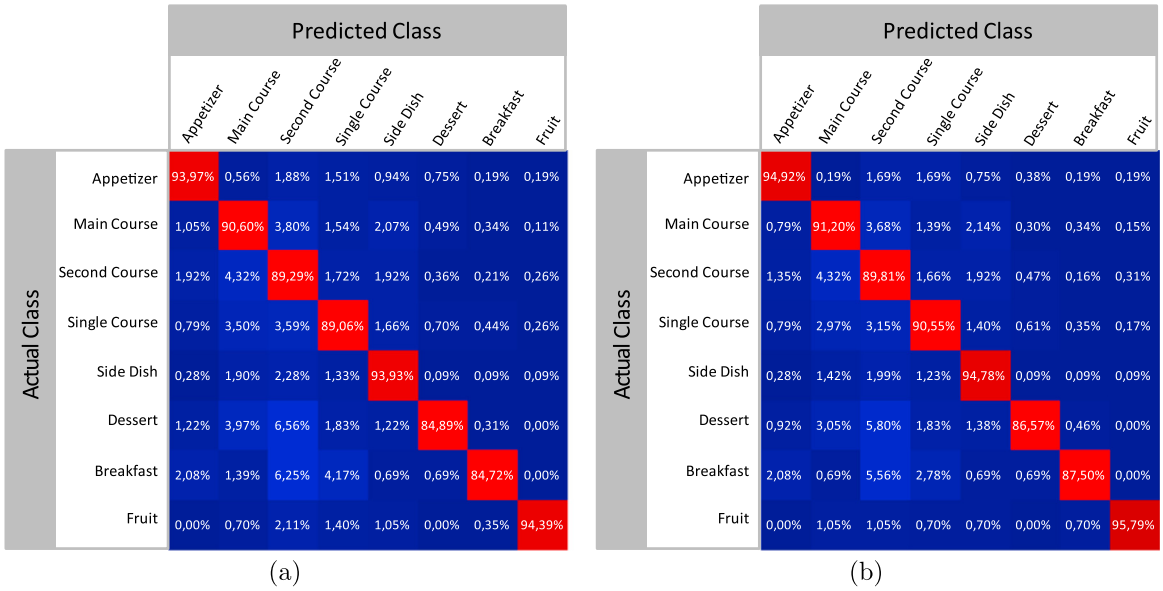

Fig. 12. The confusion matrices for the classification the tests related to food. The employed image representations are Bag of Textons (a) and Anti-Textons (b) with Schmid filter bank in L*a*b* color space and codebook of 375 visual words. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.).

with six different mobile devices by five photographers in three different restaurants, in order to guarantee a considerable photometric variability. We evaluated the approaches proposed in this paper on the Menu-Match dataset (GPS coordinates have not been used) to compare the performances with respect to the one obtained in [10]. In the original work, the acquired image was represented by employing Bag of Words paradigm and six different kinds of features, among which: color features, Histogram of Oriented Gradients (HOG), Scale-Invariant Features Transform (SIFT), Local Binary Pattern (LBP) and Textons with MR8 filters bank. All the aforementioned features were encoded through locally constrained linear encoding method (LLC) and finally joint in a unique feature vector. Since, Menu-Match dataset contains multi-labeled images, the top-5 average recall has been proposed by the authors as evaluation metric, while a one-vs.-all SVM has been employed for training and classification. The experiments in [10] report a top-5 average recall of 83.00% with a 30,720-dimensional feature vector. We performed tests employing the same Training–Testing protocol proposed by the authors, using the proposed Bag of Textons and Anti-Textons with a vocabulary of 1024 visual words. The experiments pointed out that the proposed Bag of Textons representation in L*a*b* domain and Schmid filters bank outperforms the representation suggested in [10] obtaining a top-5 recall of 84.05%. A further boost in the performances has been obtained with the proposed Anti-Textons representation (85.82%).

6. Conclusion and future perspective

In this paper the problem of Food Image Analysis has been taken into account. After a review of the literature we have focused on the problem of food image retrieval and classification. The new dataset UNICT-FD1200 has been introduced for the study of food image representation and different tests have been done to compare state-of-the-art representation approach. Another contribution of the paper is the introduction of a computational approach to encode the perceptual concept of Anti-Texton in order to encode spatial information into the Bag of Textons approach. Experiments have pointed out that Textons based representation computed in a L*a*b* domain considering the Schmid filter banks achieve good performances on both retrieval and classification tests. Finally, we have demonstrated that the proposed Anti-Textons representation is able to improve the results based on the Bag of Textons paradigm. Future works can consider the exploitation of more complex representation (e.g., deep learning) as well as a different level of classification (e.g., ingredients) to better describe a food plate. Moreover, considering the results achieved in this paper, systems based on retrieval mechanisms can also be built to deal with the problem of food intake monitoring and calories estimation. Finally, food understanding has become more and more of interest for both research community and society. There is a general consensus that multimedia assisted dietary management systems can be useful to improve the quality of life. To this aim will be important to build systems able to automatic answer different questions from food images: (1) which kind of food is in the image?; (2) what are the ingredients of the detected food? (3) does it contain allergic ingredients (e.g. nuts); (4) which is the volume of the food?; (5) how many calories I will assume with this plate?

The above questions pose many challenges. As first, it will be important to build and share benchmark labelled datasets in order to test and compare the different solutions. Common evaluation methods on the benchmark datasets should be proposed to better assess the performances of the systems with respect to different tasks (e.g., is a classification score of 99% acceptable in case of detection of allergic ingredient classification?). Studies on pixel-wise semantic segmentation of the food images are still needed to better deal with ingredients identification. An in-depth analysis of the volume estimation methods from single food images, as well as from multiple images is still missing in the literature.

References

[1] A.E. Abdel-Hakim, A.A. Farag, CSIFT: a SIFT descriptor with color invariant characteristics, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, 2006, pp. 1978–1983.
[2] M. Agrawal, K. Konolige, M.R. Blas, CenSurE: center surround extremas for realtime feature detection and matching, in: Lecture Notes in Computer Science, vol. 5305, 2008, pp. 102–115.
[3] T. Ahonen, A. Hadid, M. Pietikäinen, Face description with local binary patterns: application to face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 28 (12) (2006) 2037–2041.
[4] K. Aizawa, G.C. De Silva, M. Ogawa, Y. Sato, Food log by snapping and

processing images, in: 16th International Conference on Virtual Systems and Multimedia, 2010, pp. 71–74.
[5] M.M. Anthimopoulos, L. Gianola, L. Scarnato, P. Diem, S.G. Mougiakakou, A food recognition system for diabetic patients based on an optimized Bag-of-Features model, IEEE J. Biomed. Health Inform. 18 (4) (2014) 1261–1271.
[6] L. Arab, D. Estrin, D.H. Kim, J. Burke, J. Goldman, Feasibility testing of an automated image-capture method to aid dietary recall, European Journal of Clinical Nutrition 65 (10) (2011), 1156–1162.
[7] S. Battiato, G.M. Farinella, G. Gallo, D. Ravì, Exploiting textons distributions on spatial hierarchy for scene classification, J. Image Video Process. 2010 (7) (2010).
[8] S. Battiato, G.M. Farinella, G. Puglisi, D. Ravì, Aligning codebooks for near duplicate image detection, Multimed. Tools Appl. 72 (2) (2014) 1483–1506.
[9] H. Bay, T. Tuytelaars, L. Van Gool, SURF: speeded up robust features, in: Lecture Notes in Computer Science, vol. 3951, 2006, pp. 404–417.
[10] O. Beijbom, N. Joshi, D. Morris, S. Saponas, S. Khullar, Menu-match: restaurant-specific food logging from images, in: IEEE Winter Conference on Applications of Computer Vision, 2015, pp. 844–851.
[11] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Trans. Pattern Anal. Mach. Intell. 24 (4) (2002) 509–522.
[12] V. Bettadapura, E. Thomaz, A. Parnami, G.D. Abowd, I. Essa, Leveraging context to support automated food recognition in restaurants, in: IEEE Winter Conference on Applications of Computer Vision. IEEE, Waikoloa, 2015, pp. 580–587.
[13] A. Bhattacharyya, On a measure of divergence between two multinomial populations, Sankhyā: Indian J. Stat. (1946) 401–406.
[14] M. Bosch, T. Schap, F. Zhu, N. Khanna, C.J. Boushey, E.J. Delp, Integrated database system for mobile dietary assessment and analysis, in: 2011 IEEE International Conference on Multimedia and Expo, 2011, pp. 1–6.
[15] M. Bosch, F. Zhu, N. Khanna, C.J. Boushey, E.J. Delp, Combining global and local features for food identification in dietary assessment, in: International Conference on Image Processing, 2011, pp. 1789–1792.
[16] L. Bossard, M. Guillaumin, L. Van Gool, Food-101 mining discriminative components with random forests, in: European Conference on Computer Vision. Springer International Publishing, Zurich 2014, pp. 446–461.
[17] T. Brosnan, D.W. Sun, Improving quality inspection of food products by computer vision—a review, J. Food Eng. 61 (1) (2004) 3–16.
[18] F. Buemi, M. Massa, G. Sandini, Agrobot: a robotic system for greenhouse operations, in: Workshop on Robotics in Agriculture, 1995, pp. 172–184.
[19] G. Burghouts, J. Geusebroek, Material-specific adaptation of color invariant features, Pattern Recognit. Lett. 30 (3) (2009) 306–313.
[20] G.J. Burghouts, J.-M. Geusebroek, Performance evaluation of local colour invariants, Comput. Vision. Image Underst. 113 (1) (2009) 48–62.
[21] M. Cardenas-Weber, G. Miles, A. Hetzroni, of Agricultural Engineers. Summer Meeting, A.S., Machine Vision to Locate Melons and Guide Robotic Harvesting, American Society of Agricultural and Biological Engineers, 1991.
[22] M. Chen, K. Dhingra, W. Wu, L. Yang, R. Sukthankar, J. Yang, PFID: Pittsburgh Fast-food Image Dataset. International Conference on Image Processing, 2009, pp. 289–292.
[23] M.-Y. Chen, Y.-H. Yang, C.-J. Ho, S.-H. Wang, S.-M. Liu, E. Chang, C.-H. Yeh, M. Ouhyoung, Automatic Chinese food identification and quantity estimation, in: SIGGRAPH Asia 2012 Technical Briefs, 2012, pp. 1–4.
[24] N. Chen, Y.Y. Lee, M. Rabb, B. Schatz, Toward dietary assessment via mobile phone video cameras, in: Annual Symposium 2010, 2010, 106–110.
[25] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, 2005, pp. 886–893.
[26] J. Dehais, S. Shevchik, P. Diem, S.G. Mougiakakou, Food volume computation for self dietary assessment applications, in: IEEE International Conference on Bioinformatics and Bioengineering, IEEE, Chania, 2013, pp. 1–4.
[27] Y. Deng, B.S. Manjunath, Unsupervised segmentation of color-texture regions in images and video, IEEE Trans. Pattern Anal. Mach. Intell. 23 (8) (2001) 800–810.
[28] C.-J. Du, D.-W. Sun, Learning techniques used in computer vision for food quality evaluation: a review, J. Food Eng. 72 (1) (2006) 39–55.
[29] C.J. Du, D.W. Sun, Multi-classification of pizza using computer vision and support vector machine, J. Food Eng. 86 (2) (2008) 234–242.
[30] G.M. Farinella, D. Allegra, F. Stanco, A benchmark dataset to study the representation of food images, in: Workshop on Assistive Computer Vision and Robotics, vol. 1, Springer, Zurich 2014, pp. 584–599.
[31] G.M. Farinella, D. Allegra, F. Stanco, S. Battiato, On the exploitation of one class classification to distinguish food vs non-food images, in: New Trends in Image Analysis and Processing—MADiMa, Lecture Notes on Computer Science, vol. 9281, 2015, pp. 375–383.
[32] G.M. Farinella, M. Moltisanti, S. Battiato, Classifying food images represented as bag of textons, in: IEEE International Conference on Image Processing, 2014, pp. 5212–5216.
[33] G.M. Farinella, M. Moltisanti, S. Battiato, Food recognition using consensus vocabularies, in: Multimedia Assisted Dietary Management, Lecture Notes on Computer Science, vol. 9281, 2015.
[34] P.F. Felzenszwalb, R.B. Girshick, D. Mcallester, D. Ramanan, Object detection with discriminatively trained part based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1–20.
[35] M.a. Fischler, R.C. Bolles, Random sample consensus: a paradigm for model fitting with, Commun. ACM 24 (1981) 381–395.
[36] J.M. Fontana, E. Sazonov, Detection and characterization of food intake by wearable sensors. In: Sazonov, E., Neuman, M.R. (Eds.), Wearable Sensors. Academic Press, Oxford, 2014, pp. 591–616 (Chapter 7.4).
[37] F. Foroni, G. Pergola, G. Argiris, R.I. Rumiati, The FoodCast research image database, Front. Human Neurosci. 7 (March) (2013) 51.
[38] S. Gunasekaran, Computer vision technology for food quality assurance, Trends Food Sci. Technol. 7 (8) (1996) 245–256.
[39] Y. He, C. Xu, N. Khanna, C.J. Boushey, E.J. Delp, Analysis of food images: features and classification, in: IEEE International Conference on Image Processing, IEEE, Paris, 2014, pp. 2744–2748.
[40] L. Herranz, X. Ruihan, J. Shuqiang, A probabilistic model for food image recognition in restaurants, in: IEEE International Conference on Multimedia and Expo (ICME), June 2015, pp. 1–6.
[41] H. Hoashi, T. Joutou, K. Yanai, Image recognition of 85 food categories by feature fusion, in: IEEE International Symposium on Multimedia, 2010 pp. 296–301.
[42] Y. Hu, X. Cheng, L.-T. Chia, X. Xie, D. Rajan, A.-H. Tan, Coherent phrase model for efficient image near-duplicate retrieval, IEEE Trans. Multimed. 11 (8) (2009) 1434–1445.
[43] A.R. Jiménez, A.K. Jain, R. Ceres, J. Pons, Automatic fruit recognition: a survey and new results using Range/Attenuation images, Pattern Recognit. 32 (10) (1999) 1719–1736.
[44] B. Julesz, Textons, the elements of texture perception, and their interactions, Nature 290 (5802) (1981) 91–97.
[45] Y. Kawano, K. Yanai, Real-time mobile food recognition system, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2013, pp. 1–7.
[46] Y. Kawano, K. Yanai, Automatic expansion of a food image dataset leveraging existing categories with domain adaptation, in: Proceedings of ECCV Workshop on Transferring and Adapting Source Knowledge in Computer Vision, 2014.
[47] Y. Kawano, K. Yanai, Food image recognition with deep convolutional features, in: ACM International Joint Conference on Pervasive and Ubiquitous Computing, 2014, pp. 589–593.
[48] K. Kiliç, I.H. Boyaci, H. Köksel, I. Küsmenoglu, A classification system for beans using computer vision system and artificial neural networks, J. Food Eng. 78 (3) (2007) 897–904.
[49] S. Kim, T. Schap, M. Bosch, R. Maciejewski, E.J. Delp, D.S. Ebert, C.J. Boushey, Development of a mobile user interface for image-based dietary assessment, in: International Conference on Mobile and Ubiquitous Multimedia, 2010, pp. 13:1–13:7.
[50] K. Kitamura, C.D. Silva, T. Yamasaki, K. Aizawa, Image processing based approach to food balance analysis for personal food logging, in: IEEE International Conference on Multimedia and Expo, 2010, pp. 625–630.
[51] K. Kitamura, T. Yamasaki, K. Aizawa, Food Log by Analyzing Food Images, 2008, p. 999.
[52] K. Kitamura, T. Yamasaki, K. Aizawa, Foodlog: capture, analysis and retrieval of personal food images via web, in: Proceedings of the Workshop on Multimedia for Cooking and Eating Activities, 2009, pp. 23–30.
[53] K. Kitamura, T. Yamasaki, K. Aizawa, Foodlog: capture, analysis and retrieval of personal food images via web, in: Proceedings of the ACM Multimedia 2009 Workshop on Multimedia for Cooking and Eating Activities, 2009, p. 23.
[54] R. Kohavi, P. Foster, Glossary of terms, Mach. Learn. 30 (2) (1998) 271–274.
[55] T. Kohonen, The self-organizing map, Neurocomputing 21 (1) (1998) 1–6.
[56] F. Kong, J. Tan, Dietcam: automatic dietary assessment with mobile camera phones, Pervasive Mob. Comput. 8 (1) (2012) 147–163.
[57] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. (2012) 1–9.
[58] S. Lazebnik, C. Schmid, J. Ponce, A sparse texture representation using local affine regions, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27 (8) (2005) 1265–1278.
[59] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, 2006, pp. 2169–2178.
[60] T. Leung, J. Malik, Representing and recognizing the visual appearance of materials using three-dimensional textons, Int. J. Comput. Vision. 43 (1) (2001) 29–44.
[61] P. Levi, A. Falla, R. Pappalardo, Image controlled robotics applied to citrus fruit harvesting, in: International Conference on Robot Vision and Sensory Controls, IFS Publications, Zurich (Switzerland), 1988.
[62] J. Liu, E. Johns, L. Atallah, C. Pettitt, B. Lo, G. Frost, G.-Z. Yang, An intelligent food-intake monitoring system using wearable sensors, in: 2012 Ninth International Conference on Wearable and Implantable Body Sensor Networks, 2012, pp. 154–160.
[63] D.G. Lowe, Object recognition from local scale-invariant features, in: IEEE International Conference on Computer Vision, vol. 2, 1999, pp. 1150–1157.
[64] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vision. 60 (2) (2004) 91–110.
[65] S. Marčelja, Mathematical description of the responses of simple cortical cells*, J. Opt. Soc. Am. 70 (11) (1980) 1297–1300.
[66] N. Martinel, C. Piciarelli, C. Micheloni, G.L. Foresti, On filter banks of texture features for mobile food classification, in: Proceedings of the 9th International Conference on Distributed Smart Cameras, ACM, Seville, Spain, 2015, pp. 14–19.
[67] Y. Maruyama, G.C. De Silva, T. Yamasaki, K. Aizawa, Personalization of food

image analysis, in: 2010 16th International Conference on Virtual Systems and Multimedia, 2010, pp. 75–78.
[68] J. Matas, O. Chum, M. Urban, T. Pajdla, Robust wide-baseline stereo from maximally stable extremal regions, in: Image and Vision Computing (10 SPEC. ISS.), vol. 22, 2004, pp. 761–767.
[69] Y. Matsuda, H. Hoashi, K. Yanai, Multiple-food recognition considering co-occurrence employing manifold ranking, in: International Conference on Pattern Recognition, 2012, pp. 2017–2020.
[70] Y. Matsuda, H. Hoashi, K. Yanai, Recognition of multiple-food images by detecting candidate regions, in: IEEE International Conference on Multimedia and Expo, IEEE, Melbourne, Australia, 2012, pp. 25–30.
[71] C.R.J. Maurer, Q. Rensheng, V. Raghavan, A linear time algorithm for computing exact Euclidean distance transforms of binary images in arbitrary dimensions, IEEE Trans. Pattern Anal. Mach. Intell. 25 (2) (2003) 265–270.
[72] P. Munkevik, T. Duckett, G. Hall, Vision system learning for ready meal characterisation, in: International Conference on Engineering and Food, no. 1, 2004.
[73] P. Munkevik, G. Hall, T. Duckett, A computer vision system for appearance-based descriptive sensory evaluation of meals, J. Food Eng. 78 (1) (2007) 246–256.
[74] D.D.T. Nguyen, Z. Zong, P. Ogunbona, W. Li, Object detection using non-redundant local binary patterns, in: IEEE International Conference on Image Processing, 2010, pp. 4609–4612.
[75] D. Nistér, H. Stewénius, Scalable recognition with a vocabulary tree, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, 2006, pp. 2161–2168.
[76] E. Nowak, F. Jurie, B. Triggs, Sampling strategies for bag-of-features image classification, in: Proceedings of the 9th European Conference on Computer Vision, Springer-Verlag, Graz, Austria, 2006, pp. 490–503.
[77] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.
[78] G. O'Loughlin, S.J. Cullen, A. McGoldrick, S. O'Connor, R. Blain, S. O'Malley, G. D. Warrington, Using a wearable camera to increase the accuracy of dietary analysis, Am. J. Prev. Med. 44 (3) (2013) 297–301.
[79] E. Parrish, A.K. Goksel, Pictorial pattern recognition applied to fruit harvesting, in: Transactions of the ASAE, no. 20, 1977, pp. 822–827.
[80] F. Perronnin, C. Dance, Fisher kernels on visual vocabularies for image categorization, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2007, pp. 1–8.
[81] F. Perronnin, J. Sánchez, T. Mensink, Improving the fisher kernel for large-scale image classification, in: European Conference on Computer Vision, Springer, Crete, Greece, 2010, pp. 143–156.
[82] C. Pham, D. Jackson, J. Schöning, T. Bartindale, T. Plotz, P. Olivier, FoodBoard: surface contact imaging for food recognition, in: UbiComp, 2013, pp. 749–752.
[83] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, Object retrieval with large vocabularies and fast spatial matching, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007.
[84] P. Pouladzadeh, A. Yassine, S. Shirmohammadi, Foodd: food detection dataset for calorie measurement using food images, in: New Trends in Image Analysis and Processing—MADiMa 2015, Workshop in Conjunction with the 18th International Conference on Image Analysis and Processing, Lecture Notes in Computer Science, vol. 9281, Springer International Publishing, Genova, Italy, 2015, pp. 441–448.
[85] M. Puri, Z. Zhu, Q. Yu, A. Divakaran, H. Sawhney, Recognition and volume estimation of food intake using a mobile device. In: IEEE Workshop on Applications of Computer Vision. IEEE, Snowbird, USA, 2009, pp. 1–8.
[86] X. Qi, R. Xiao, J. Guo, L. Zhang, Pairwise rotation invariant co-occurrence local binary pattern, in: European Conference on Computer Vision, vol. 7577, 2012, pp. 158–171.
[87] M.H. Rahmana, M.R. Pickering, D. Kerr, C.J. Boushey, E.J. Delp, A new texture feature for improved food recognition accuracy in a mobile phone based dietary assessment system, in: Proceedings of the 2012 IEEE International Conference on Multimedia and Expo Workshops, 2012, pp. 418–423.
[88] D. Ravì, B. Lo, G.-z. Yang, Real-time food intake classification and energy expenditure estimation on a mobile device, in: Body Sensor Networks, 2015.
[89] A.J. Rich, A programmable calculator system for the estimation of nutritional intake of hospital patients, Am. J. Clin. Nutr. (1981) 34.
[90] A. Rosenfeld, J.L. Pfaltz, Distance functions on digital pictures, Pattern Recognit. 1 (1) (1968) 33–61.
[91] X. Ruihan, L. Herranz, J. Shuqiang, W. Shuang, S. Xinhang, R. Jain, Geolocalized modeling for dish recognition, IEEE Trans. Multimed. 17 (Auguest (8)) (2015) 1187–1199.
[92] J. Sánchez, F. Perronnin, T. Mensink, J. Verbeek, Image classification with the fisher vector: theory and practice, Int. J. Comput. Vision. 105 (3) (2013) 222–245.
[93] C. Schmid, Constructing models for content-based image retrieval, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, 2001, pp. II-39–II-45.
[94] J. Shotton, M. Johnson, R. Cipolla, Semantic texton forests for image categorization and segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008.
[95] G. Shroff, A. Smailagic, D.P. Siewiorek, Wearable context-aware food recognition for calorie monitoring, in: Proceedings of International Symposium on Wearable Computers, 2008, pp. 119–120.
[96] D.C. Slaughter, R.C. Harrell, Color vision in robotic fruit harvesting, Trans. Am. Soc. Agric. Biol. Eng. 30 (4) (1987) 1144–1148.
[97] D.C. Slaughter, R.C. Harrell, Discriminating fruit for robotic harvest using color in natural outdoor scenes, Trans. Am. Soc. Agric. Biol. Eng. 32 (2) (1989) 757–763.
[98] D.W. Sun, Inspecting pizza topping percentage and distribution by a computer vision method, J. Food Eng. 44 (4) (2000) 245–249.
[99] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.
[100] Tin Kam Ho, Random decision forests, in: Proceedings of the International Conference on Document Analysis and Recognition, vol. 1, 1995, pp. 278–282.
[101] E. Tola, V. Lepetit, P. Fua, DAISY: an efficient dense descriptor applied to wide-baseline stereo, IEEE Trans. Pattern Anal. Mach. Intell. 32 (5) (2010) 815–830.
[102] A. Topchy, A.K. Jain, W. Punch, Clustering ensembles: models of consensus and weak partitions, IEEE Trans. Pattern Anal. Mach. Intell. 27 (12) (2005) 1866–1881.
[103] G.J. Van Tonder, Y. Ejima, From image segregation to anti-textons, Perception 29 (2000) 1231–1247.
[104] M. Varma, D. Ray, Learning the discriminative power-invariance trade-off, in: Proceedings of the IEEE International Conference on Computer Vision, 2007.
[105] M. Varma, A. Zisserman, A statistical approach to texture classification from single images, Int. J. Comput. Vision. 62 (1–2) (2005) 61–81.
[106] A. Vedaldi, B. Fulkerson, VLFeat: An Open and Portable Library of Computer Vision Algorithms, 2008, ⟨http://www.vlfeat.org/⟩.
[107] A. Vedaldi, A. Zisserman, Efficient additive kernels via explicit feature maps, IEEE Trans. Pattern Anal. Mach. Intell. 34 (3) (2012) 480–492.
[108] W. Wen, Y. Jie, Fast food recognition from videos of eating for calorie estimation, in: IEEE International Conference on Multimedia and Expo, 2009, pp. 1210–1213.
[109] D. Williams, B. Julesz, Filters versus textons in human and machine texture discrimination, in: Neural Networks for Perception—Human and Machine Perception, vol. 1, 1991, pp. 145–175.
[110] D. Williams, B. Julesz, Perceptual asymmetry in texture perception, in: Proceedings of the National Academy of Sciences of the United States of America, vol. 89, 1992, pp. 6531–6534.
[111] P.D. Wright, G. Shearing, A.J. Rich, I. Johnston, The role of a computer in the management of clinical parenteral nutrition, J. Parenter. Enter. Nutr. 2 (1978) 652–657.
[112] W. Xin, D. Kumar, N. Thome, M. Cord, F. Precioso, Recipe recognition with large multimodal food dataset, in: IEEE International Conference on Multimedia Expo Workshops (ICMEW), June 2015, pp. 1–6.
[113] C. Xu, Y. He, N. Khannan, A. Parra, C. Boushey, E. Delp, Image-based food volume estimation, in: International Workshop on Multimedia for Cooking and Eating Activities, 2013, pp. 75–80.
[114] K. Yanai, T. Joutou, A food image recognition system with multiple kernel learning, in: IEEE International Conference on Image Processing, 2009, pp. 285–288.
[115] K. Yanai, Y. Kawano, Food image recognition using deep convolutional network with pre-training and fine-tuning, in: IEEE International Conference on Multimedia Expo Workshops (ICMEW), June 2015, pp. 1–6.
[116] S. Yang, M. Chen, D. Pomerleau, R. Sukthankar, Food recognition using statistics of pairwise local features, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 2249–2256.
[117] L. Zepeda, D. Deal, Think before you eat: photographic food diaries as intervention tools to change dietary decision making and attitudes, Int. J. Consum. Stud. 32 (6) (2008) 692–698.
[118] F. Zhu, M. Bosch, I. Woo, S. Kim, C.J. Boushey, D.S. Ebert, E.J. Delp, The use of mobile devices in aiding dietary assessment and evaluation, J. Sel. Top. Signal Process. 4 (4) (2010) 756–766.
[119] Z. Zong, D.T. Nguyen, P. Ogunbona, W. Li, On the combination of local texture and global structure for food classification, in: IEEE International Symposium on Multimedia, 2010, pp. 204–211.