Food recognition: a new dataset, experiments and results

Gianluigi Ciocca, Paolo Napoletano, and Raimondo Schettini

***Abstract*—We propose a new dataset for the evaluation of food recognition algorithms designed for dietary monitoring. Each image depicts a real canteen tray with dishes and foods arranged in different ways. Each tray contains multiple instances of food classes. We collected a set of 1,027 canteen trays for a total of 3,616 food instances belonging to 73 food classes. The food on the tray images have been manually segmented using carefully drawn polygonal boundaries. We benchmark the dataset designing an automatic tray analysis pipeline that takes a tray image as input, finds the regions of interest, and predicts for each region the corresponding food class. We experimented three different classification strategies using also several visual descriptors. In the experiments, we have achieved about 79% of food and tray recognition accuracy using Convolutional-Neural-Networks-based features. The dataset, as well as the benchmark framework, are made available to the research community.**

***Index Terms*—Food dataset, Food recognition, Algorithm benchmarking, Convolutional Neural Networks (CNN), Dietary monitoring.**

I. Introduction

HEALTH care on food and good practices in dietary behavior are drawing people's attention recently. Nowadays technology can support the users in keep tracks of their food consumption, and to increase the awareness in their daily diet by monitoring their food habits. In the recent years many research works have demonstrated that machine learning and computer vision techniques can help to build systems to automatically recognize diverse foods and to estimate the food quantity [1], [2], [3], [4], [5]. To be useful for dietary monitoring, food recognition systems should also be able to operate on "wild" environments such as restaurants, canteens, and such. Obviously, the fair benchmarking of these systems, requires the availability of suitable datasets that actually pose the challenges of the food recognition task in unconstrained environments.

A. Food recognition systems

Researches in the literature have often focused on different aspects of the food recognition problem. Many works address the challenges in the recognition of food by developing recognition strategies that differ in terms of features and classification methodologies. With respect to the features, the work of He et al. [6] describes the food image by combining both global and local features, while the work of Farinella et al [7] uses a vocabulary built on textons. SIFT and local binary patterns are used in [8], while in [9], the context of where the pictures were taken is also exploited along with the visual features. With respect to the classification strategies, the most widely used are k-NN classifiers [6], [10], and Support Vector Machines [7], [8]. An evaluation of different classification methodologies is reported in [5] where SVM, Artificial Neural Networks and Random Forest classification methods are analyzed. Recently, Convolutional Neural Network (CNN) are also being used in the context of food recognition [11], [12], [13].

Other works in the literature focus on the design of a complete system for diet monitoring in real contexts. Often these systems exploit mobile application for food recognition, assessment, and logging. Examples of such systems are Food-Log [14], DietCam [15], Menu-Match [16], FoodCam [17], and those described in [18], [19], and [10].

Food quantity estimation is very important in the context of a dietary monitoring since on it depends the assessment of the food intakes. Works that tackle this problem are for example [20], [21], [22], [23], [24], [25], [26], [27]. All these works require a reference information to be able to estimate the quantity of food on the plate. This information may came from markers or token for camera calibration, the size of a reference objects (e.g. thumb, or eating tools), or from the specific location where the food is consumed (e.g. canteen). Other works, instead of estimating the amount of food from 2D images, use 3D techniques coupled with template matching or shape reconstruction algorithms [28], [29], [20].

Very few works specifically consider the problem of leftover estimation. Often the problem is theoretically treated as a special case of the problem of food recognition and quantity estimation [23], [18]. Only one work to date explicitly tackles the problem with assessment experiments on a dedicated dataset [10].

B. Food Datasets

Regardless of the objective, a dataset of food images is required to evaluate the performance of the different feature extraction and classification algorithms proposed. To this end, the above research works either used existing datasets or introduce new ones.

One of the first food dataset was introduced in [30]. It contains 50 food categories (mostly Japanese food) and the images, gathered from the Web, depict a close-up of a single food. Using MKL-based feature fusion, they obtained a recognition accuracy of 61.34%. This dataset was enlarged to

All the authors are with the Department of Informatics, System and Communication, University of Milano-Bicocca, Italy

TABLE I
LIST OF FOOD DATASETS USED IN THE LITERATURE.

Name	Year	#Images	#Classes	Type	Acquisition	Task	Annotation	Availability	Reference
Food50	2009	5,000	50	Single	Wild	Food Recognition	Label	Proprietary	[30]
PFID	2009	1,098[a]	61[a]	Single	Wild and Lab	Food Recognition	Label	Public	[31]
TADA	2009	50/256	-	Single and Multi	Lab	Food Recognition	-	Proprietary	[22]
Food85[b]	2010	8,500	85	Single	Wild	Food Recognition	Label	Proprietary	[32]
Chen	2012	5,000	50	Single	Wild	Food Recognition	Label	Public	[33]
UEC FOOD-100	2012	9,060	100	Single and Multi	Wild	Food Recognition	BBox	Public	[34], [35]
Food-101	2014	101,000	101	Single	Wild	Food Recognition	Label	Public	[36]
UEC FOOD-256[c]	2014	31,397	256	Single and Multi	Wild	Food Recognition	BBox	Public	[37], [38]
UNICT-FD889	2014	3,583	889	Single	Wild	Near Duplicate Food Retrieval	Label	Public	[39]
Diabetes	2014	4,868	11	Single	Wild	Food Recognition	Label	Public	[5]
UNIMIB2015	2015	1,000× 2	15	Multi	Wild/Canteen	Food Recognition and Leftover Estimation	Poly	Public[d]	[10]
UNIMIB2016	2016[d]	1,027	73	Multi	Wild/Canteen	Food Recognition	Poly	Public[d]	-

[a] Numbers refer to the baseline dataset.
[b] Includes Food50.
[c] Includes UECFOOD-100.
[d] http://www.ivl.disco.unimib.it/activities/food-recognition/

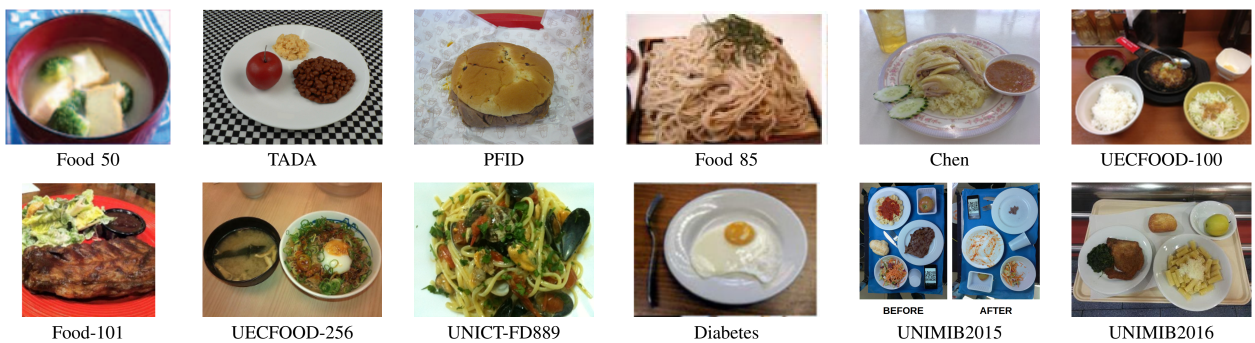


Fig. 1. Food dataset examples.

85 food categories in a subsequent work [32]. Using a similar approach to the previous work, the authors achieved a classification accuracy of 62.85%. These two datasets are proprietary. Other proprietary datasets are the ones introduced in [22] and [4]. These datasets have been acquired in a lab settings and use markers to help the recognition phase. Differently from the previous datasets the TADA dataset [22], contains images of real foods (256 images) as well as food replica (50 images). Also, the images can have multiple food depicted. This makes the dataset more challenging since it requires the segmentation of each food in the image. Another dataset that contains images with multiple foods is the UECFOOD-100 dataset [35]. It is public and contains more than 9,000 images of 100 food categories. For the recognition, SVM classifiers with color histogram and SURF features are used, achieving a classification rate of 81.55% for the top 5 category candidates when the ground-truth bounding boxes are given. The dataset was extended to 256 food categories in [38] and the classification rate in this case was 74.4% for the top 5 categories. Chen et al. [33] published a dataset of 5,000 images of 50 foods. Using multi-label SVMs trained on SIFT, LBP, color and and Gabor features, they achieved a food recognition overall accuracy of 68.3%.

Currently, the largest dataset available is Food-101 [36]. It contains 101,000 images divided into 101 food categories. Random forest are used to mine discriminant parts in the food images extracted from superpixels. These parts are then classified with SVM achieving an average accuracy of 50.76% on the 101 classes. If the Food-101 is the largest dataset available, the UNICT889 [7] is the dataset with most food categories. It contains 889 classes on a total of 3,583 images. Given these numbers, each class contains few instances of a given food. However, the goal of the authors is the near duplicate food retrieval, and not food recognition. Different features are tested and the best results for near duplicate retrieval was achieved by color Bag-of-Textons with a mean average precision of 67.5%.

Anthimopoulos et al. [5] uses a dataset of 4,868 food images organized into 11 classes to evaluate a food recognition system based on the Bag-of-Features model. The system is designed to help diabetic patients in controlling their carbohydrates daily consumption. Different visual features and classification strategies are tested and the best combination shows a classification accuracy of slightly less that 78% using a 10,000 words dictionary.

In [10] we presented a dataset used for testing a system that recognizes foods and estimates food leftovers. The dataset contains 2,000 images of 15 classes of foods placed on trays. The images were acquired in a real canteen location, and are paired with the corresponding leftover images acquired after the meals. The images are associated to a given canteen customer by using a QR code automatically generated by the dietary monitoring system on the customer's mobile. [10] is the first dataset explicitly designed for both food recognition and leftover estimation.

Table I, summarizes the characteristics of the different datasets of food images available in the literature. For each dataset we report its size and the number of food categories it contains. The datasets have been categorized according to the type of images considered (i.e. images containing a single food or a set of foods), the acquisition procedure (e.g. in-the-wild for unconstrained acquisitions, or in-the lab for constrained acquisitions), the task for which it is used or created, the annotation type (label only, bounding boxes, or polygonal areas), and the availability (i.e. either public, or proprietary). Figure 1 shows some examples of the images contained in each dataset.

As it can be seen from Table I, and Figure 1, most of the existing datasets depict single instance foods with only three dataset having multiple instance foods in the images. Not all the environments (and cultures) are characterized by a single food plate. For example, Asian food usually are placed in different small plates and are usually brought on the table at the same time (UECFOOD-100 is an example). Moreover in all the canteen environments different plates, for the first course, main course, side dishes and desserts, are placed on the same tray. In these cases, it is more convenient to take a single picture of the whole meal than separate pictures for each food. To date, only the UNIMIB2015 dataset is specifically designed for the canteen environments.

Canteens and cafeterias are very important in everyday life because they are often the preferred (or only) choice for workers, employees, or students. Cafeterias and canteens are receiving more and more attention with respect to health issues and wellness for the customers. The problem of healthy and balanced meal in schools is seriously tackled by the different health agencies with the aim at reducing obesity and unbalanced nutrition. For example, the Department of Health of the Italian Government promoted an extensive campaign for food and nutrition education[1]. The Department of Health of the Australian Government, compiled a very detailed report with guidelines for healthy foods in school canteens[2]. Similar actions can be found across many other countries (e.g. UK[3], USA[4], etc. . .).

Also corporations are addressing the dietary wellness of their employees. For example Google re-engineered its cafeterias to drive people towards healthier food choices by changing food disposition and using color coding to highlight food calories [5]. Other corporate dining services are following a similar approach to provide healthier food and to educate their employees to a correct diet[6].

For these reasons, we believe that datasets of food images acquired in canteen or cafeteria environments are very important for the problem of food recognition and dietary monitoring, and large and representative datasets are thus required.

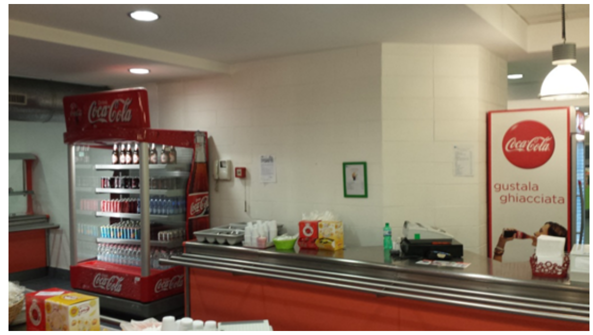


Fig. 2. The canteen situated within the University of Milano-Bicocca campus where we have acquired the images in the UNIMIB2016 dataset.

In this paper we introduce a new food dataset named UNIMIB2016. This dataset is similar to our previous dataset UNIMIB2015. Both contains images taken in a canteen environment where different foods are placed on a tray to be taken on the dining table. Differently from the UNIMIB2015 dataset, here we have much more classes and the dishes are more difficult to locate due to the similar color of plates, tray and placemats.

In fact, in UNIMIB2015, the placemat being dark blue is clearly distinguishable from the other items. In UNIMIB2016, the placemat is white as the plates. This could make it more difficult the location and segmentation of the plates. Moreover, the higher number of food classes with respect to UNIMIB2015 makes this dataset more representative of the typical foods found in canteens. As it can be seen form Figure 3, many food classes have a very similar appearance. For example, we have four different "Pasta al sugo", but with other main ingredients (e.g. fish, vegetables, or meat) added. Finally, on the tray there can be other "noisy" objects that must be ignored during the recognition. For example, we may find cell phones, wallets, id cards, and other personal items. For these reasons we need to design of a very accurate recognition algorithm.

These differences make this dataset more challenging than the previous one for the task of food recognition. Finally, as in the UNIMIB2015 dataset, here we have conducted a careful annotation of the food regions using polygonal shapes. This will allow design of food quantity estimation algorithms using a very precise ground truth. However the UNIMIB2015 dataset is the only dataset available that contains images and annotations of canteen trays taken before and after the meal (see Figure 1) and therefore can be used for leftover estimation. Also the two dataset are both publicly available for research purposes.

II. The UNIMIB2016 Food Dataset

The dataset has been collected in a real canteen environment. The particularities of this setting are that each image depicts different foods on a tray, and some foods (e.g. fruit, bread and dessert) are placed on the placemats rather than on plates. Sides are often served in the same plate as the main dish

[1] http://www.salute.gov.it/imgs/c_17_pubblicazioni_1248_allegato.pdf

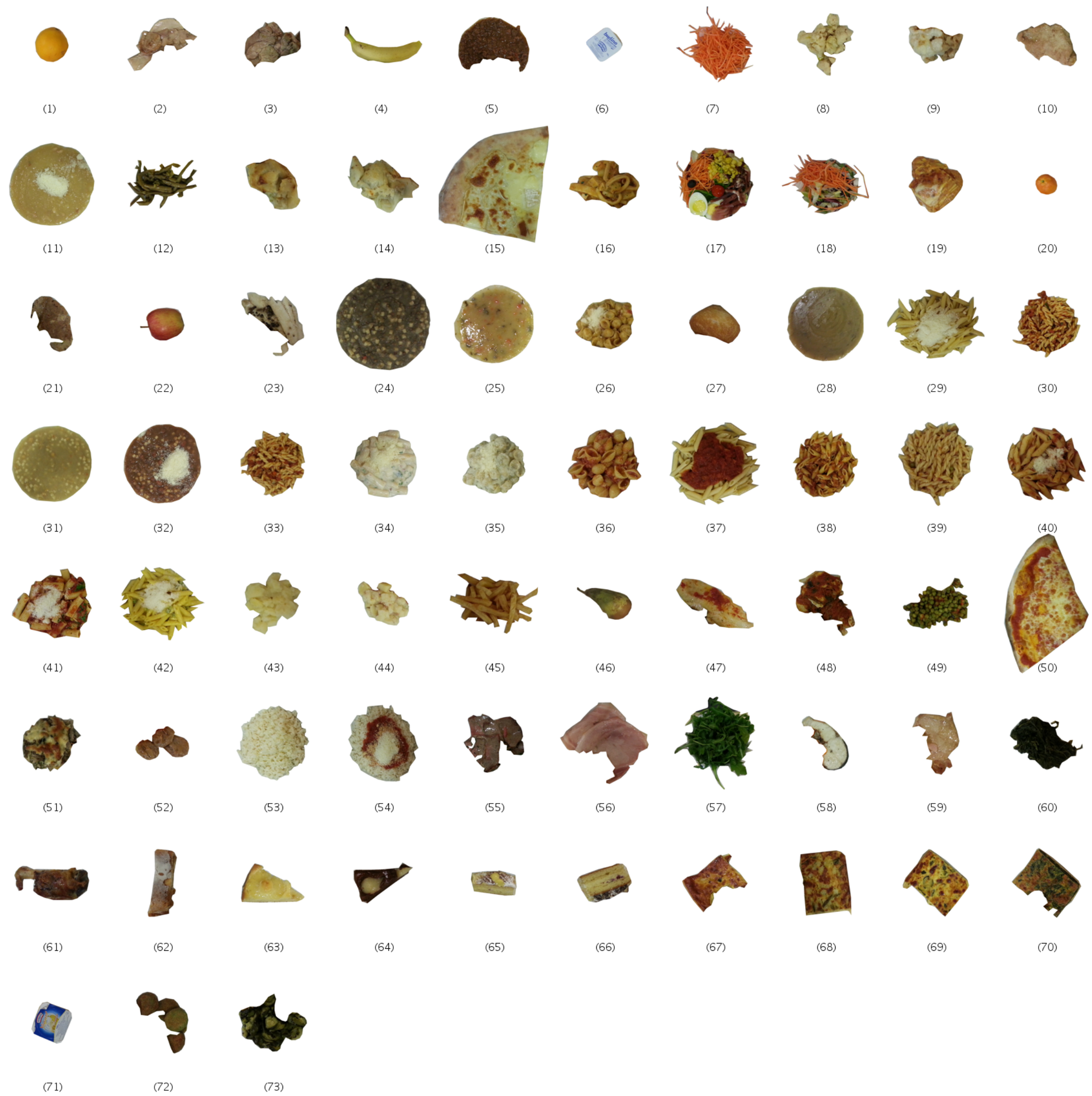
[2] https://education.nt.gov.au/policies/canteen-nutrition-and-healthy-eating

[3] http://www.schoolfoodplan.com/actions/school-food-standards/

[4] http://www.fns.usda.gov/school-meals/child-nutrition-programs

[5] http://www.fastcodesign.com/1669355/6-ways-google-hacks-its-cafeterias-so-googlers-eat-healthier

[6] http://www.timesfreepress.com/news/business/aroundregion/story/2015/jan/20/todays-company-cafeterias-offer-healthier-brighter-fare/283592/



01. Arancia
02. Arrosto
03. Arrosto di vitello
04. Banane
05. Bruscitti
06. Budino
07. Carote
08. Cavolfiore
09. Cavolfiore gratinato
10. Cotoletta
11. Crema zucca e fagioli
12. Fagiolini
13. Finocchi gratinati
14. Finocchi in umido
15. Focaccia bianca
16. Guazzetto di calamari
17. Insalata mista 2
18. Insalata mista 1
19. Lasagna alla bolognese
20. Mandarino
21. Medaglioni di carne
22. Mela
23. Merluzzo alle olive
24. Minestra
25. Minestra lombarda
26. Orecchiette al ragú
27. Pane
28. Passato alla piemontese
29. Pasta in bianco
30. Pasta cozze e vongole
31. Pasta e ceci
32. Pasta e fagioli
33. Pasta mare e monti
34. Pasta pancetta e zucchine
35. Pasta pesto besciamella e cornetti
36. Pasta ricotta e salsiccia
37. Pasta al sugo
38. Pasta al sugo pesce
39. Pasta al sugo vegetariano
40. Pasta tonno
41. Pasta tonno e piselli
42. Pasta zafferano e piselli
43. Patate/puré
44. Patate/puré prosciutto
45. Patatine fritte
46. Pera
47. Pesce 1
48. Pesce 2
49. Piselli
50. Pizza
51. Pizzoccheri
52. Polpette di carne
53. Riso in bianco
54. Riso sugo
55. Roast-beef 1
56. Roast-beef 2
57. Rucola
58. Trancio di merluzzo
59. Scaloppine
60. Spinaci
61. Stinco di maiale
62. Strudel
63. Torta ananas
64. Torta cioccolato e pera
65. Torta crema 1
66. Torta crema 2
67. Torta salata alla valdostana
68. Torta salata
69. Torta salata rustica
70. Torta salata spinaci e ricotta
71. Yogurt
72. Zucchine impanate
73. Zucchine in umido

Fig. 3. Segmented images of the 73 food categories in the proposed UNIMIB2016 dataset. On the right, the Italian names of the classes. Note that in some cases foods slightly differ in the ingredients, and thus are named as “FoodName 1”, “FoodName 2”, etc.

Fig. 4. Examples of acquired trays. The black polygon around the food represents the manual annotations.

making it difficulty to separate the two. Moreover, the acquisition of the images has been performed in a semi-controlled settings so the images present visual distortions as well as illumination changes due to shadows. These characteristics make this dataset challenging requiring both the segmentation of the trays for food localization, and a robust way to deal with multiple foods.

Figure 2 shows the location where the images have been acquired. It is a canteen situated within the University of Milano-Bicocca Campus that serves students and faculty members. Images have been acquired using a hand-held Samsung Galaxy S3 (GT-i9300) smart phone. The acquisition station is located at the end of the tray line after the cashier. Customers place the tray on the acquisition station and the images are taken by an operator. Unfortunately, due to privacy issues and the intense affluence of customers, we have been unable to take pictures of the trays after the meal. This has prevented us to include leftover information in this dataset as we have done in the UNIMIB2015 dataset.

We have collected a total of 1,442 images that went through a quality check phase were we removed excessively blurred images, and duplicated photos. After this phase we obtained a final dataset of 1,027 tray images, 73 food categories, and a total of 3,616 food instances. Figure 3 shows a sample of each food category of the UNIMIB2016 dataset, while Figure 4 shows some examples of the acquired images.

To create the ground truth, we have annotated the dataset using an improved version of our Image Annotation Tool (IAT) [40], [41]. The modifications include the support of touchscreens, the drawing of freehand shapes, and the automatic approximation these shapes to polygon using the Ramer-Douglas-Peucker algorithm [42], [43]. These modification allowed us a significant speed up in the annotation process with respect to the standard point and click mouse. Figure 4 shows some examples of annotations superimposed to the acquired images. Using our tool, to each image we have associated an annotation file containing the list of food identities, and the segmentation region of each food in terms of points of the polygon surrounding it.

Most of the existing food databases are characterized by images that contain a single food (often in a close-up setting), and in most of the cases the food annotations are provided in terms of bounding boxes around the food, see Table I. The UNIMIB2016 dataset is characterized by images that contain multiple foods and moreover the dataset includes accurate segmentation of foods (see Figure 4). These type of annotations allow researchers to work on methods for food segmentation, as well as food quantity estimation.

III. Tray analysis

In Figure 5 we show the schema of our tray analysis method. The segmentator module takes the tray image as input. The output of this module is a list of regions of interest. As benchmarking, we also consider the regions of interest obtained from the ground-truth annotations. The regions of interest are then processed by the food class predictor. The output of the predictor is a list of recognized foods. Given a region of interest we investigate the use of three different approaches for predicting the food class. The first approach is a global one that extracts the visual features from the whole region of interest. The second approach is a local one that extracts the visual features from local patches of the region of interest. The third approach combines the posterior probabilities computed by the global and local classifiers with the sum and product operators [44]. Given a region of interest r_i, the probability that a region is of class m is calculated in two ways:

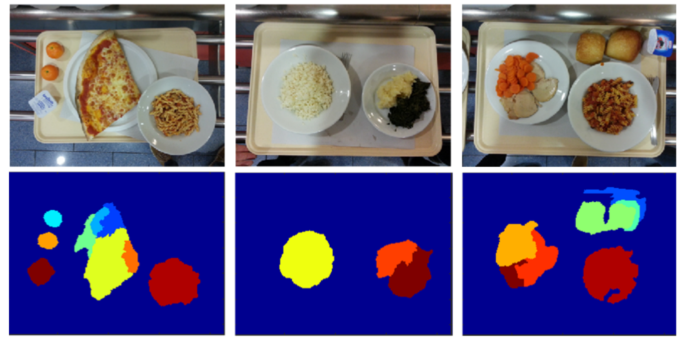


Fig. 7. Examples of segmentation results.

1) sum rule:$P(m|r_i) = P_G(m|r_i) + P_L(m|r_i)$;
2) product rule: $P(m|r_i) = P_G(m|r_i) \cdot P_L(m|r_i)$.

where $P_G(m|r_i)$ and $P_L((m|r_i)$ are the probability that a region of interest r_i is of class m with respect to the global and local approach respectively. The sum rule is expected to produce reliable results when the approaches catch information that is highly correlated, while the product rule is expected to be effective when the two approaches catch independent information.

A. Tray segmentation

Figure 6 shows the segmentation pipeline of the segmentator module used to detect the regions of interest on the tray that presumably will contain food samples. It is composed of four main steps. First, in order to speed up the computation without losing relevant information, the input RGB image is resized to an height of 320 pixels. The resized image undergoes two separate processing pipeline: a saturation-based one, and a color texture one. In the first one, the image is firstly gamma corrected and then the RGB values are converted to HSV to extract the saturation channel (step 1a of Figure 6). These values are automatically thresholded and morphological operations are applied to clean up the obtained binary image (step 1b). We have noticed that the saturation channel contain good cues for the localization of food regions since they have saturation values higher than the plate regions, the tray and the cutlery. Of course, other regions may have saturation values comparable to those of the food and thus we have introduced a second processing based on the segmentation algorithm JSEG [45] that works on both color and texture features (step 2a of Figure 6). We use the standard implementation of the authors with the default parameters (i.e. automatic segmentation) and

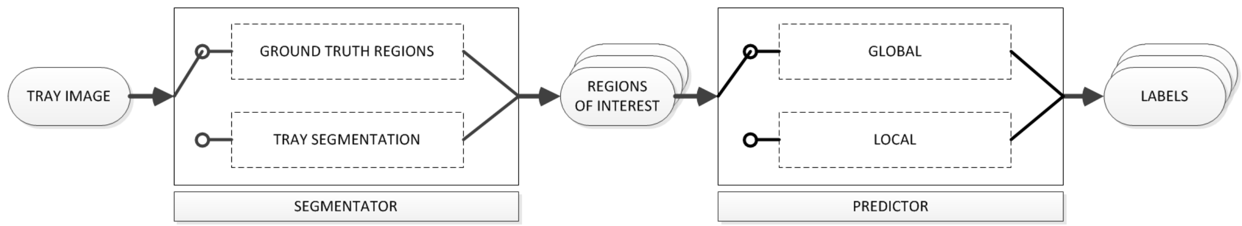


Fig. 5. Tray analysis pipeline.

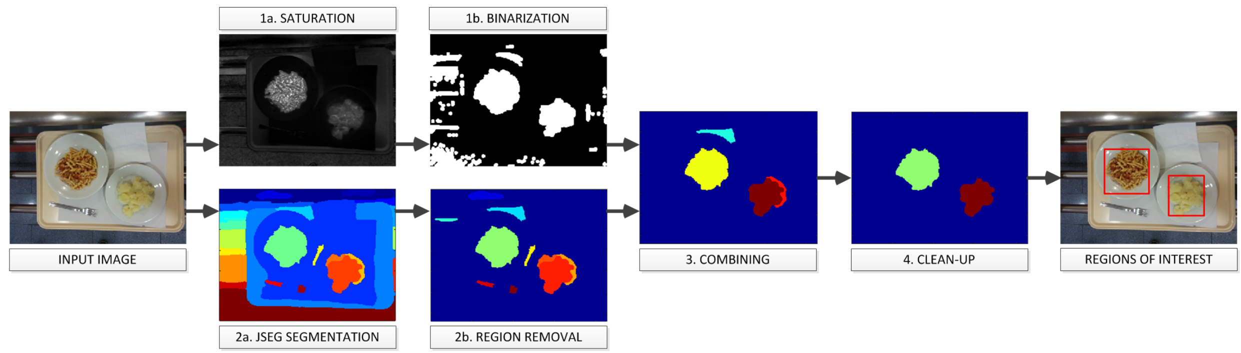


Fig. 6. Processing pipeline for the food segmentation.

TABLE II
REGION-BASED AND BOUNDARY-BASED SEGMENTATION PERFORMANCE RESULTS.

	Region-based			Boundary-based		
	Covering	PRI	VI	Recall	Precision	F-Measure
JSEG	0.385	0.389	3.106	0.870	0.198	0.323
Proposed	0.916	0.931	0.429	0.714	0.734	0.724

we found that if works well in most cases. The segmentation is able to detect the regions having similar visual characteristics.

The segmented image is then processed in order to remove non relevant regions (step 2b). For instance, the regions that touch the border of the image do not belong to the food regions and thus can be eliminated. Also, regions larger or smaller than predefined thresholds can be discarded as well (e.g. the placemat, the tray, highlights). The final segmented image contains with high probability the food regions and few non relevant ones. In order to retain only the food regions, the outputs of the saturation-based processing and the output of the color and texture processing are combined together (step 3). The combination performs a cross analysis between the two outputs with the aim to retain only the segmented regions that have a large percentage of saturated pixels. With this analysis we are able to remove most of the regions of the cutlery and the spurious ones while retaining the food regions. To further ensure that only few, relevant, regions are retained for the classification phase, geometric constraints are used to clean-up the output of the combining step (step 4). The bounding boxes of all the regions of interest are passed to the prediction phase.

In order to assess the proposed segmentation pipeline we applied the evaluation benchmarks suggested in [46]. Specifically we computed the following region-based measures: covering of ground-truth (Covering), the Probabilistic Rand Index (PRI), and the Variation of Information (VI). Moreover, following the same work, we also computed the boundary-based precision-recall measures. We compare the final results obtained by the proposed segmentation pipeline against the segmentation initially obtained by the JSEG algorithm. Results are reported in Table II. As it can be seen the proposed strategy obtains the best segmentation results by all the measures considered. The region-based measures shows the highest improvements: 0.916 against 0.385, and 0.931 against 0.389 for Covering and PRI respectively, while the obtained VI is 0.429 against the initial 3.106 (in this case the lower the better). With respect to the boundary-based measures, we see that the initial segmentations have a high recall but with a very low precision, while the proposed one has a more balanced precision-recall values. On the overall, the proposed segmentation pipeline outperforms the JSEG one with an F-Measure of 0.724 against 0.323. The results shows that the proposed segmentation strategy is able to effectively locate the food regions.

Figure 7 shows some results of out segmentation pipeline. As it can be seen, we are able to separate different food on the same plate. We still have some spurious regions that we hope to classify as non-food regions in the next phase. Moreover, the JSEG algorithm often over-segments foods that shows heterogeneous regions such as the pizza slice or very textured foods such as the salads and vegetables. Each one of these regions will be independently classified. Before being passed to the classification phase, the coordinates of the bounding boxes of the food regions are transformed back to match the image's original size.

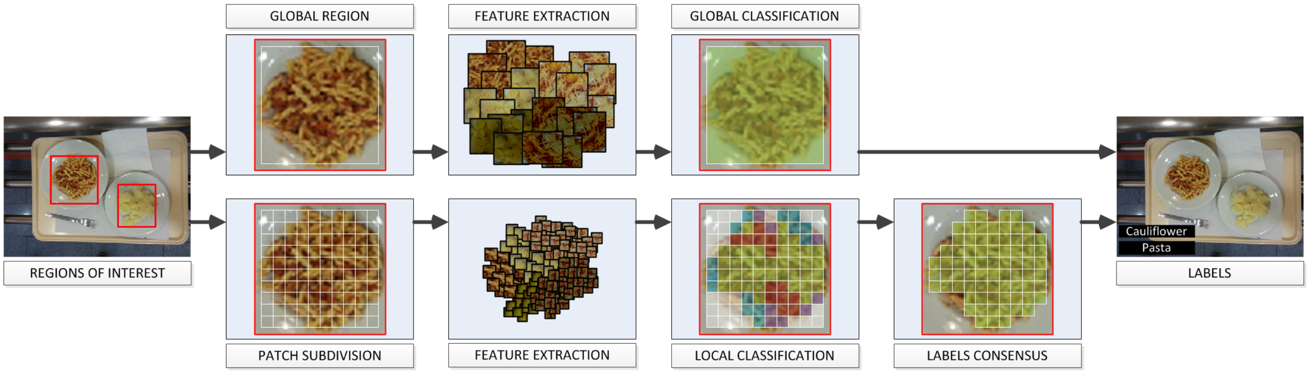


Fig. 8. Processing pipeline for the food classification.

B. Classification of regions of interest

Figure 8 shows the processing pipeline for the food classification used in the predictor module. As we discussed before, we compare three different classification strategies: a global strategy (top path in Figure 8), a local one (bottom path), and a combination of them. The classification module works as follows. Depending on the classification strategy, from each region of interest one sub-image (global strategy) or several, non-overlapping, image patches (local strategy) are extracted. These images are then fed to a feature extractor where several visual descriptors are computed. Specifically, we have evaluated the foollowing visual descriptors: color histogram (HIST) [47], Gabor features (Gabor) [48], Opponent Gabor features (OG) [49], Local Color Contrast (LCC) [50], [51], Chromaticity Moments (CM) [49], Complex Wavelet features (CWT) [49], [52], Color and Edge Directivity Descriptor (CEDD) [53], non-uniform invariant Local Binary Pattern on the RGB channels (LBP) [54] , Convolutional Neural Network (CNN) [55], [56], and Bag of Convolutional Filter Responses (BOCFR) [57], [58], [59].

The visual descriptors are independently evaluated by pre-trained classifiers for predicting the corresponding food label. We experimented the use of two classifiers as predictor: the k-Nearest Neighbour (k-NN) and Support Vector Machines (SVM). The training of the classifiers is done by considering a suitable split of the UNIMIB2016 that will be described in Sec. IV. In the case of the local classification strategy, for each region of interest, we have several food labels, one for each image patch. Thus it is necessary a post-processing phase to merge all these labels into a final classification decision. The local strategy is similar to the one presented in our previous work [10], and it should be useful when the food region contains part of different foods as often happens in the case of the side dishes.

IV. Experimental setup

For comparison, we evaluate the different visual features and classification strategies. In order to evaluate how much the segmentation process influences the classification process, we also present experiments considering the ideal food segmentation provided by the the ground-truth.

A. Visual descriptors

In this work we compare several visual descriptors. All feature vectors are L^2 normalized [7]:

- 768-dimensional RGB [47];
- 32-dimensional *Gabor* features composed of mean and standard deviation of six orientations extracted at four frequencies for each color channel [49];
- 264-dimensional *opponent Gabor* feature vector extracted as Gabor features from several inter/intra channel combinations: monochrome features extracted from each channel separately and opponent features extracted from couples of colors at different frequencies [49];
- 256-dimensional *Local Color Contrast* feature vector. The LCC vector is obtained by comparing the color at a given location with the average color in a surrounding neighborhood. The is computed in terms of the angular difference between the color vectors [50];
- 10-dimensional feature vector composed of normalized *chromaticity moments* as defined in [49];
- 8-dimensional *Dual Tree Complex Wavelet Transform* (CWT) features obtained considering four scales, mean and standard deviation, and three color channels [49], [52];
- 144-dimensional *Color and Edge Directivity Descriptor* (CEDD) features. This descriptor uses a fuzzy version of the five digital filters proposed by the MPEG-7 Edge Histogram Descriptor (EHD), forming 6 texture areas. CEDD uses 2 fuzzy systems that map the colors of the image in a 24-color custom palette [49], [52];
- 18-dimensional *Local Binary Patterns* (LBP) feature vector for each channel. We consider LBP applied to color images represented in RGB [60]. We select the LBP with a circular neighbourhood of radius 2 and 16 elements, and 18 uniform and rotation invariant patterns;
- 4096-dimensional *Convolutional Neural Networks* features (CNN4096). The CNN-based features are obtained as the intermediate representations of deep convolutional neural networks originally trained for ILSVRC 2012 [61]. The networks are used to generate a visual descriptor by removing the final softmax nonlinearity and the last fully-connected layer. This network is the BVLC AlexNet [61];

[7] The feature vector are divided by its L^2-norm.

- 128-dimensional *Convolutional Neural Networks* features (CNN128). Features are extracted in the same way as in the case of CNN4096. Here the network is the Vgg M [56] that is similar to the one presented in [62] with a reduced number of filters in the convolutional layer four. The last fully-connected layer is 128-dimensional. Even this network is trained for ILSVRC 2012;
- 1024-dimensional BoCFR: we consider the *Bag of Convolutional Filter Responses* (BoCFR) of the first convolutional layer of the BVLC AlexNet trained for ILSVRC 2012. We built a codebook of 1024 visual words by exploiting images from external sources.

B. Training process

Since we collected our dataset in a real canteen scenario, and with different daily menus, the number of occurrences of each food is highly variable. This number ranges from a maximum of 479 instances for the "Pane" class downto one for some other classes (e.g. "Strudel" and "Rucola"). We have removed from the dataset the images with foods having fewer than four instances. The final dataset used in the experiments thus contains 1,010 tray images, and 65 foods. We split the 1,010 tray images into a training set and a test set such that the sets contain about 70% and 30% of each food instances respectively. This resulted in a training set of 650 tray images, and a test set of 360 images.

For training the global and local food classifiers, we extracted the visual descriptors from the regions of interest provided by the ground-truth segmentation of the training trays.

Regarding the $k-$NN classifier, we have evaluated different values of k ranging from 1 to 11 and we have selected the value that gave the best results across visual descriptors and classification strategies, that is $k = 3$. For what concerns the SVM classifier, we have adopted the radial basis function kernel with width and regularization parameters found after a cross validation procedure.

During the prediction process in the case of the local classification approach, the region of interest is subdivided in patches of size 140×140. The resulting patches may contain both food and no-food classes. This is quite clear looking at the bottom part of the Figure 8. For both the global and local classifiers, during the training process, we added the class no-food to the classifier by choosing randomly samples from the portion of the tray images that do not overlap with the regions of interest. Once the prediction of each patch is obtained, the class with the maximum number of patches predicted is assigned to the region of interest.

C. Evaluation measures

To cope with the class imbalance problem of the test set we jointly used two assessment metrics for food recognition: the *Standard Accuracy (SA)* and the *Macro Average Accuracy (MAA)* [63]. Denoting NP_c the number of positives, i.e., the number of times the class c occurs in the dataset; TP_c the number of *true positives* for class c, i.e., the number of times that the system recognizes the dish c; C the number of classes, for each class, the metrics can be defined as follows:

$$SA = \frac{\sum_{c=1}^{C} TP_c}{\sum_{c=1}^{C} NP_c}; \qquad MAA = \frac{1}{C}\sum_{c=1}^{C} A_c = \frac{1}{C}\sum_{c=1}^{C} \frac{TP_c}{NP_c}.$$

The metric for the evaluation of the error in the tray analysis is the *Tray accuracy*. This is defined as the percentage of trays correctly analyzed. A tray is correctly analyzed when all the foods contained are correctly recognized.

V. RESULTS

Results are presented in Table III. Is quite clear that the CNN-based visual descriptors achieve better results than others in all the classification strategy. In particular, the CNN4096 features coupled with the combination of posterior probability strategy obtains the best performance. It is quite interesting to note that, apart some exceptions, the combination strategy, with both k-NN and SVM classifiers, reduces the performance with respect to the use of global and patch-based approaches. It happens in all cases when global and patch-based approaches are coupled with visual descriptors that are not good performing. It also interesting to note that, the patch-based approach outperforms the global approach only when it is coupled with the SVM classifier. This is due to the fact that the radial basis function used in the SVM classifier is more suitable than the linear k-NN to separate the food classes in the feature space when the number of samples increases. Moreover, the patch-based strategy greatly outperforms the global one when coupled with traditional visual descriptors (no CNN-based). This suggest that the lower discriminant power of the these features, compared to the CNN-based ones, is somewhat compensated by the larger amount of information obtained by aggregating the classification results from the local patches. For example, among the non CNN visual descriptors, the Hist RGB combined with the local classification approach achieves a performance that is very close to some of the CNN-based descriptors. This is due to the fact that the local approach in some way takes into account the spatial variability of the food. In fact, the local approach, when applied to the UNIMIB2015 dataset, has demonstrated to be very useful for food quantity estimation [10]: the number of patches labeled as food X suggests the quantity of the food X itself.

Overall, the SVM classifier performs slightly better than k-NN with a tray accuracy of 78.9% obtained using the sum of posteriors combination strategy. The Table III contains also the results achieved using the ideal ground-truth (GT) as a perfect segmentation algorithm. The differences between the results obtained using the proposed segmentation pipeline and GT, permit to evaluate the influences of the automatic segmentation on the classification performance of the entire pipeline. It is quite clear that when the ideal segmentation is used we achieve a gain of about 10% with a maximum of 86% accuracy for the food recognition.

As it can be seen from the results, the UNIMIB2016 dataset is indeed more challenging for the recognition task than the UNIMIB2015 dataset.

TABLE III

FOOD RECOGNITION RESULTS USING THE PROPOSED TRAY ANALYSIS PIPELINE AND $k-$NN OR SVM CLASSIFIER. PROPOSED: OUR AUTOMATIC SEGMENTATION. GT: GROUND-TRUTH SEGMENTATION. G: GLOBAL APPROACH. P: LOCAL, PATCH-BASED APPROACH. G $\bigoplus$ P, COMBINATION EXPLOITING THE SUM OF POSTERIORS. G $\bigotimes$ P, COMBINATION EXPLOITING THE PRODUCT OF POSTERIORS. FOR EACH ROW, THE BEST RESULT IS REPORTED IN BOLD.

Classifier	Segment.	Approach	Measure	LBP	CEDD	Hist	Gabor	OG	LCC	CM	CWT	CNN128	CNN4096	BoCFR
k-NN	Proposed	G	Food SA	0.343	0.423	0.555	0.397	0.463	0.320	0.439	0.276	0.656	**0.728**	0.689
			Food MAA	0.139	0.184	0.356	0.168	0.253	0.127	0.210	0.079	0.467	**0.585**	0.490
			Tray Accuracy	0.353	0.383	0.561	0.367	0.446	0.306	0.409	0.231	0.676	**0.732**	0.689
		P	Food SA	0.488	0.594	0.689	0.597	0.667	0.492	0.608	0.624	0.679	0.697	**0.697**
			Food MAA	0.202	0.315	0.474	0.318	0.443	0.201	0.326	0.387	0.453	0.473	**0.490**
			Tray Accuracy	0.438	0.560	0.685	0.563	0.673	0.433	0.573	0.621	0.674	0.692	**0.694**
		G $\bigoplus$ P	Food SA	0.490	0.608	0.673	0.612	0.684	0.489	0.593	0.591	0.742	**0.764**	0.729
			Food MAA	0.193	0.298	0.470	0.329	0.453	0.160	0.299	0.334	0.509	**0.561**	0.539
			Tray Accuracy	0.399	0.515	0.636	0.540	0.655	0.367	0.509	0.536	0.715	**0.738**	0.711
		G $\bigotimes$ P	Food SA	0.436	0.477	0.637	0.461	0.515	0.350	0.511	0.313	0.714	**0.763**	0.716
			Food MAA	0.198	0.235	0.428	0.238	0.331	0.150	0.285	0.137	0.504	**0.601**	0.554
			Tray Accuracy	0.360	0.402	0.592	0.398	0.497	0.301	0.454	0.274	0.696	**0.747**	0.709
	GT	G	Food SA	0.394	0.446	0.628	0.427	0.536	0.358	0.518	0.289	0.748	**0.820**	0.761
			Food MAA	0.171	0.219	0.380	0.192	0.299	0.151	0.255	0.085	0.555	**0.652**	0.559
			Tray Accuracy	0.434	0.492	0.662	0.470	0.570	0.408	0.534	0.313	0.783	**0.842**	0.782
		P	Food SA	0.543	0.656	0.719	0.682	0.719	0.557	0.651	0.723	0.745	**0.774**	0.734
			Food MAA	0.221	0.312	0.505	0.367	0.458	0.201	0.346	0.420	0.464	0.500	**0.510**
			Tray Accuracy	0.501	0.625	0.720	0.648	0.721	0.499	0.632	0.681	0.738	**0.762**	0.743
		G $\bigoplus$ P	Food SA	0.504	0.629	0.732	0.641	0.752	0.518	0.631	0.623	0.814	**0.855**	0.811
			Food MAA	0.210	0.313	0.493	0.377	0.492	0.176	0.332	0.360	0.586	**0.631**	0.577
			Tray Accuracy	0.431	0.529	0.686	0.580	0.701	0.391	0.557	0.565	0.787	**0.826**	0.777
		G $\bigotimes$ P	Food SA	0.437	0.536	0.705	0.518	0.619	0.412	0.586	0.330	0.805	**0.858**	0.791
			Food MAA	0.222	0.273	0.457	0.291	0.380	0.183	0.327	0.143	0.611	**0.685**	0.614
			Tray Accuracy	0.389	0.461	0.650	0.475	0.580	0.368	0.552	0.295	0.791	**0.840**	0.785
SVM	Proposed	G	Food Accuracy	0.398	0.465	0.610	0.396	0.434	0.320	0.432	0.297	0.694	**0.715**	0.666
			Food MAA	0.185	0.215	0.346	0.203	0.234	0.098	0.211	0.093	0.479	**0.546**	0.449
			Tray Accuracy	0.440	0.440	0.575	0.394	0.403	0.313	0.408	0.255	0.703	**0.738**	0.669
		P	Food SA	0.607	0.645	0.721	0.627	0.732	0.515	0.606	0.650	0.742	**0.783**	0.708
			Food MAA	0.332	0.356	0.483	0.377	0.498	0.168	0.330	0.428	0.496	**0.560**	0.479
			Tray Accuracy	0.585	0.605	0.705	0.630	0.729	0.421	0.570	0.655	0.720	**0.767**	0.708
		G $\bigoplus$ P	Food SA	0.640	0.628	0.703	0.670	0.713	0.382	0.612	0.646	0.777	**0.798**	0.702
			Food MAA	0.387	0.399	0.452	0.446	0.518	0.100	0.261	0.453	0.616	**0.632**	0.465
			Tray Accuracy	0.596	0.610	0.690	0.638	0.712	0.304	0.469	0.640	0.768	**0.789**	0.702
		G $\bigotimes$ P	Food SA	0.489	0.555	0.612	0.529	0.640	0.414	0.498	0.504	0.746	**0.789**	0.698
			Food MAA	0.281	0.354	0.367	0.322	0.443	0.114	0.277	0.228	0.626	**0.636**	0.442
			Tray Accuracy	0.465	0.513	0.580	0.499	0.630	0.322	0.461	0.441	0.756	**0.777**	0.689
	GT	G	Food SA	0.480	0.520	0.643	0.456	0.533	0.425	0.495	0.326	0.774	**0.825**	0.756
			Food MAA	0.231	0.249	0.375	0.234	0.298	0.134	0.274	0.106	0.552	**0.644**	0.489
			Tray Accuracy	0.525	0.562	0.667	0.502	0.560	0.416	0.538	0.347	0.798	**0.842**	0.753
		P	Food SA	0.646	0.718	0.759	0.711	0.795	0.609	0.650	0.718	0.816	**0.857**	0.763
			Food MAA	0.346	0.405	0.518	0.388	0.538	0.180	0.360	0.449	0.541	**0.575**	0.505
			Tray Accuracy	0.659	0.694	0.762	0.694	0.788	0.489	0.646	0.726	0.804	**0.838**	0.763
		G $\bigoplus$ P	Food SA	0.672	0.700	0.721	0.698	0.769	0.385	0.581	0.702	0.872	**0.891**	0.734
			Food MAA	0.419	0.444	0.505	0.470	0.545	0.092	0.263	0.454	0.677	**0.684**	0.508
			Tray Accuracy	0.641	0.658	0.723	0.665	0.745	0.279	0.459	0.670	0.845	**0.871**	0.702
		G $\bigotimes$ P	Food SA	0.565	0.619	0.642	0.576	0.711	0.418	0.528	0.551	0.816	**0.858**	0.722
			Food MAA	0.322	0.370	0.434	0.359	0.471	0.125	0.310	0.248	0.670	**0.687**	0.557
			Tray Accuracy	0.530	0.567	0.634	0.546	0.669	0.324	0.498	0.464	0.814	**0.843**	0.691

VI. CONCLUSION

In the recent years, it has been demonstrated that visual recognition and machine learning methods can be used to develop systems that keep tracks of human food consumption. The actual usefulness of these system heavily depends on the capability of recognizing foods in unconstrained environments. In this paper, we propose a new dataset for the evaluation of food recognition algorithms designed for dietary monitoring. The images have been acquired in a real canteen and depict a real canteen tray with foods arranged in different ways. Each tray contains multiple instances of food classes. We collected a set of 1,027 canteen trays for a total of 3,616 food instances belonging to 73 food classes. The tray images have been manually segmented using carefully drawn polygonal boundaries. We designed a suitable automatic tray analysis pipeline that takes a tray image as input, finds the regions of interest, and predicts for each region the corresponding food class. We evaluated three different classification strategies using several visual descriptors. The best performance has been obtained by using Convolutional-Neural-Networks-based features. The dataset, as well as the benchmark framework, are made available to the research community. Thanks to the way it has been annotated, this database along with the UNIMIB2015 can be used for food segmentation, recognition and quantity estimation.

Acknowledgment

The authors would like to thank Alessio Albé for his invaluable help in the collection of the images during the acquisition campaign.

References

[1] W. Wu and J. Yang, "Fast food recognition from videos of eating for calorie estimation," in *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*. IEEE, 2009, pp. 1210–1213.

[2] N. Yao, R. J. Sclabassi, Q. Liu, J. Yang, J. D. Fernstrom, M. H. Fernstrom, and M. Sun, "A video processing approach to the study of obesity," in *Multimedia and Expo, 2007 IEEE International Conference on*. IEEE, 2007, pp. 1727–1730.

[3] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar, "Food recognition using statistics of pairwise local features," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2249–2256.

[4] M. Bosch, F. Zhu, N. Khanna, C. Boushey, and E. Delp, "Combining global and local features for food identification in dietary assessment," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*, 2011, pp. 1789–1792.

[5] M. M. Anthimopoulos, L. Gianola, L. Scarnato, P. Diem, and S. G. Mougiakakou, "A food recognition system for diabetic patients based on an optimized bag-of-features model," *Biomedical and Health Informatics, IEEE Journal of*, vol. 18, no. 4, pp. 1261–1271, 2014.

[6] Y. He, C. Xu, N. Khanna, C. Boushey, and E. Delp, "Analysis of food images: Features and classification," in *Image Processing (ICIP), 2014 IEEE International Conference on*, 2014, pp. 2744–2748.

[7] G. Farinella, M. Moltisanti, and S. Battiato, "Classifying food images represented as bag of textons," in *Image Processing (ICIP), 2014 IEEE International Conference on*, 2014, pp. 5212–5216.

[8] D. T. Nguyen, Z. Zong, P. O. Ogunbona, Y. Probst, and W. Li, "Food image classification using local appearance and global structural information," *Neurocomputing*, vol. 140, pp. 242–251, 2014.

[9] V. Bettadapura, E. Thomaz, A. Parnami, G. Abowd, and I. Essa, "Leveraging context to support automated food recognition in restaurants," in *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, 2015, pp. 580–587.

[10] G. Ciocca, P. Napoletano, and R. Schettini, "Food recognition and leftover estimation for daily diet monitoring," in *New Trends in Image Analysis and Processing – ICIAP 2015 Workshops*, ser. Lecture Notes in Computer Science, V. Murino, E. Puppo, D. Sona, M. Cristani, and C. Sansone, Eds., vol. 9281. Springer International Publishing, 2015, pp. 334–341.

[11] H. Kagaya, K. Aizawa, and M. Ogawa, "Food detection and recognition using convolutional neural network," in *Proceedings of the ACM International Conference on Multimedia*, ser. MM '14, 2014, pp. 1085–1088.

[12] Y. Kawano and K. Yanai, "Food image recognition with deep convolutional features," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '14 Adjunct, 2014, pp. 589–593.

[13] W. Zhang, D. Zhao, W. Gong, Z. Li, Q. Lu, and S. Yang, "Food image recognition with convolutional neural networks," in *2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom)*, 2015, pp. 690–693.

[14] K. Kitamura, T. Yamasaki, and 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, pp. 23–30.

[15] F. Kong and J. Tan, "Dietcam: Automatic dietary assessment with mobile camera phones," *Pervasive and Mobile Computing*, vol. 8, no. 1, pp. 147–163, 2012.

[16] O. Beijbom, N. Joshi, D. Morris, S. Saponas, and S. Khullar, "Menu-match: Restaurant-specific food logging from images," in *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*. IEEE, 2015, pp. 844–851.

[17] Y. Kawano and K. Yanai, "Foodcam: A real-time food recognition system on a smartphone," *Multimedia Tools and Applications*, pp. 1–25, 2014.

[18] F. Zhu, M. Bosch, I. Woo, S. Kim, C. Boushey, D. Ebert, and E. Delp, "The use of mobile devices in aiding dietary assessment and evaluation," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 4, pp. 756–766, 2010.

[19] Z. Ahmad, N. Khanna, D. A. Kerr, C. J. Boushey, and E. J. Delp, "A mobile phone user interface for image-based dietary assessment," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2014, pp. 903 007–903 007.

[20] M. Puri, Z. Zhu, Q. Yu, A. Divakaran, and H. Sawhney, "Recognition and volume estimation of food intake using a mobile device," in *Applications of Computer Vision (WACV), 2009 Workshop on*, 2009, pp. 1–8.

[21] M. Sun, Q. Liu, K. Schmidt, J. Yang, N. Yao, J. Fernstrom, M. Fernstrom, J. P. DeLany, and R. Sclabassi, "Determination of food portion size by image processing," in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, 2008, pp. 871–874.

[22] A. Mariappan, M. Bosch, F. Zhu, C. J. Boushey, D. A. Kerr, D. S. Ebert, and E. J. Delp, "Personal dietary assessment using mobile devices," vol. 7246, 2009, pp. 72 460Z–72 460Z–12.

[23] P. Pouladzadeh, S. Shirmohammadi, and R. Al-Maghrabi, "Measuring calorie and nutrition from food image," *Instrumentation and Measurement, IEEE Transactions on*, vol. 63, no. 8, pp. 1947–1956, 2014.

[24] P. Pouladzadeh, G. Villalobos, R. Almaghrabi, and S. Shirmohammadi, "A novel svm based food recognition method for calorie measurement applications," in *Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on*, 2012, pp. 495–498.

[25] G. Villalobos, R. Almaghrabi, P. Pouladzadeh, and S. Shirmohammadi, "An image processing approach for calorie intake measurement," in *Medical Measurements and Applications Proceedings, 2012 IEEE International Symposium on*, 2012, pp. 1–5.

[26] E. A. Akpro Hippocrate, H. Suwa, Y. Arakawa, and K. Yasumoto, "Food weight estimation using smartphone and cutlery," in *Proceedings of the First Workshop on IoT-enabled Healthcare and Wellness Technologies and Systems*, ser. IoT of Health '16. ACM, 2016, pp. 9–14.

[27] P. Pouladzadeh, P. Kuhad, S. V. B. Peddi, A. Yassine, and S. Shirmohammadi, "Food calorie measurement using deep learning neural network," in *2016 IEEE International Instrumentation and Measurement Technology Conference Proceedings*, 2016, pp. 1–6.

[28] J. Chae, I. Woo, S. Kim, R. Maciejewski, F. Zhu, E. J. Delp, C. J. Boushey, and D. S. Ebert, "Volume estimation using food specific shape templates in mobile image-based dietary assessment," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2011, pp. 78 730K–78 730K.

[29] Y. He, C. Xu, N. Khanna, C. Boushey, and E. Delp, "Food image analysis: Segmentation, identification and weight estimation," in *Multimedia and Expo (ICME), 2013 IEEE International Conference on*, 2013, pp. 1–6.

[30] T. Joutou and K. Yanai, "A food image recognition system with multiple kernel learning," in *Image Processing (ICIP), 2009 16th IEEE International Conference on*. IEEE, 2009, pp. 285–288.

[31] M. Chen, K. Dhingra, W. Wu, L. Yang, R. Sukthankar, and J. Yang, "Pfid: Pittsburgh fast-food image dataset," in *Image Processing (ICIP), 2009 16th IEEE International Conference on*. IEEE, 2009, pp. 289–292.

[32] H. Hoashi, T. Joutou, and K. Yanai, "Image recognition of 85 food categories by feature fusion," in *Multimedia (ISM), 2010 IEEE International Symposium on*. IEEE, 2010, pp. 296–301.

[33] M.-Y. Chen, Y.-H. Yang, C.-J. Ho, S.-H. Wang, S.-M. Liu, E. Chang, C.-H. Yeh, and M. Ouhyoung, "Automatic chinese food identification and quantity estimation," in *SIGGRAPH Asia 2012 Technical Briefs*. ACM, 2012, p. 29.

[34] Y. Matsuda, H. Hoashi, and K. Yanai, "Recognition of multiple-food images by detecting candidate regions," in *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, 2012, pp. 25–30.

[35] Y. Kawano and K. Yanai, "Real-time mobile food recognition system," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, 2013, pp. 1–7.

[36] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–mining discriminative components with random forests," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 446–461.

[37] Y. Kawano and K. Yanai, "Foodcam-256: A large-scale real-time mobile food recognitionsystem employing high-dimensional features and com-

pression of classifier weights," in *Proceedings of the ACM International Conference on Multimedia*, ser. MM '14, 2014, pp. 761–762.

[38] ——, "Automatic expansion of a food image dataset leveraging existing categories with domain adaptation," in *Proc. of ECCV Workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV)*, 2014.

[39] G. M. Farinella, D. Allegra, and F. Stanco, "A benchmark dataset to study the representation of food images," in *ECCV European Conference in Computer Vision, Workshop Assistive Computer Vision and Robotics, Zurich*, 2014.

[40] G. Ciocca, P. Napoletano, and R. Schettini, "Iat-image annotation tool: Manual," *arXiv preprint arXiv:1502.05212*, 2015.

[41] S. Bianco, G. Ciocca, P. Napoletano, and R. Schettini, "An interactive tool for manual, semi-automatic and automatic video annotation," *Computer Vision and Image Understanding*, vol. 131, pp. 88–99, 2015.

[42] U. Ramer, "An iterative procedure for the polygonal approximation of plane curves," *Computer Graphics and Image Processing*, vol. 1, no. 3, pp. 244–256, 1972.

[43] D. H. Douglas and T. K. Peucker, "Algorithms for the reduction of the number of points required to represent a digitized line or its caricature," *Cartographica: The International Journal for Geographic Information and Geovisualization*, vol. 10, no. 2, pp. 112–122, 1973.

[44] J. Kittler, M. Hatef, R. P. Duin, and J. Matas, "On combining classifiers," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 3, pp. 226–239, 1998.

[45] Y. Deng and B. S. Manjunath, "Unsupervised segmentation of color-texture regions in images and video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 8, pp. 800–810, 2001.

[46] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 898–916, 2011.

[47] C. L. Novak, S. Shafer *et al.*, "Anatomy of a color histogram," in *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on*. IEEE, 1992, pp. 599–605.

[48] B. S. Manjunath and W.-Y. Ma, "Texture features for browsing and retrieval of image data," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 18, no. 8, pp. 837–842, 1996.

[49] F. Bianconi, R. Harvey, P. Southam, and A. Fernández, "Theoretical and experimental comparison of different approaches for color texture classification," *Journal of Electronic Imaging*, vol. 20, no. 4, 2011.

[50] C. Cusano, P. Napoletano, and R. Schettini, "Combining local binary patterns and local color contrast for texture classification under varying illumination," *JOSA A*, vol. 31, no. 7, pp. 1453–1461, 2014.

[51] ——, "Intensity and color descriptors for texture classification," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2013, pp. 866 113–866 113.

[52] M. Barilla and M. Spann, "Colour-based texture image classification using the complex wavelet transform," in *Electrical Engineering, Computing Science and Automatic Control, 2008. CCE 2008. 5th International Conference on*, nov. 2008, pp. 358 –363.

[53] S. A. Chatzichristofis and Y. S. Boutalis, "Cedd: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval," in *Computer Vision Systems*. Springer, 2008, pp. 312–322.

[54] T. Ojala, M. Pietikäinen, and T. Mänepää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, 2002.

[55] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, 2014, pp. 512–519.

[56] A. Vedaldi and K. Lenc, "Matconvnet – convolutional neural networks for matlab," *CoRR*, vol. abs/1412.4564, 2014.

[57] J. Ng, F. Yang, and L. Davis, "Exploiting local features from deep networks for image retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 53–61.

[58] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003, pp. 1470–1477.

[59] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. of the Int'l Conf. on Advances in Geographic Information Systems*, 2010, pp. 270–279.

[60] T. Mäenpää and M. Pietikäinen, "Classification with color and texture: jointly or separately?" *Pattern Recognition*, vol. 37, no. 8, pp. 1629–1640, 2004.

[61] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[62] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 818–833.

[63] H. He and Y. Ma, *Imbalanced Learning: Foundations, Algorithms, and Applications*. John Wiley & Sons, 2013.

Gianluigi Ciocca is currently associate professor in computer science at DISCo (Department of Informatics, Systems and Communication) of the University of Milano-Bicocca. He has been fellow at the Institute of Multimedia Information Technologies of the Italian National Research Council. His research focused on the development of systems for the management of image and video databases and the development of new methodologies and algorithms for automatic indexing. His current research interests focus on image and video analysis and abstraction.

Paolo Napoletano is currently a post-doc researcher at DISCo (Department of Informatics, Systems and Communication) of the University of Milano-Bicocca. In 2007, he received a PhD degree in Information Engineering from the University of Salerno with a thesis focused on Computational Vision and Pattern Recognition. In 2003, he received a Master's degree in Telecommunications Engineering from the University of Naples Federico II, with a thesis focused on Transmission of Electromagnetic Fields. His current research interests focus on image and video analysis and information retrieval.

Raimondo Schettini is Full Professor at the University of Milano-Bicocca. He is Vice-Director of the Department of Informatics, Systems and Communication, and head of Imaging and Vision Lab (www.ivl.disco.unimib.it). He has been associated with Italian National Research Council (CNR) since 1987 where he has leaded the Color Imaging Lab from 1990 to 2002. He has been team leader in several research projects and published more than 300 refereed papers and several patents about color reproduction, image processing, analysis and classification. Raimondo Schettini is Fellow of the International Association of Pattern Recognition (IAPR) for his contributions to pattern recognition research and color image analysis.