A review on food recognition technology for health applications

Dario Allegra,[1] Sebastiano Battiato,[1,2] Alessandro Ortis,[1,2] Salvatore Urso,[2] Riccardo Polosa[2]

[1]Department of Mathematics and Computer Science; [2]Center of Excellence for the Acceleration of Harm Reduction (CoEHAR), University of Catania, Catania, Italy

Abstract

Food understanding from digital media has become a challenge with important applications in many different domains. On the other hand, food is a crucial part of human life since the health is strictly affected by diet. The impact of food in people life led Computer Vision specialists to develop new methods for automatic food intake monitoring and food logging. In this review paper we provide an overview about automatic food intake monitoring, by focusing on technical aspects and Computer Vision works which solve the main involved tasks (*i.e.*, classification, recognitions, segmentation, etc.). Specifically, we conducted a systematic review on main scientific databases, including interdisciplinary databases (*i.e.*, Scopus) as well as academic databases in the field of computer science that focus on topics related to image understanding (*i.e.*, recognition, analysis, retrieval). The search queries were based on the following key words: “food recognition”, “food classification”, “food portion estimation”, “food logging” and “food image dataset”. A total of 434 papers have been retrieved. We excluded 329 works in the first screening and performed a new check for the remaining 105 papers. Then, we manually added 5 recent relevant studies. Our final selection includes 23 papers that present systems for automatic food intake monitoring, as well as 46 papers which addressed Computer Vision tasks related food images analysis which we consider essential for a comprehensive overview about this research topic. A discussion that highlights the limitations of this research field is reported in conclusions.

Correspondence: Alessandro Ortis, Department of Mathematics and Computer Science, University of Catania, Italy.
E-mail: ortis@dmi.unict.it

Key words: Food recognition; health technology; computer vision; food image classification; food image retrieval.

Contributions: conceptualization: DA, SB, AO, RP and SU; investigation, DA, AO; writing–original draft preparation, AO; writing–review and editing, SB, AO and RP; supervision SB and RP; project administration, SU.

Conflict of interest: The authors declare no conflict of interest.

Funding: Produced with the help of a grant from the Foundation for a Smoke-Free World, Inc.

Received for publication: 2 August 2020.
Revision received: 3 October 2020.
Accepted for publication: 7 November 2020.

Health Psychology Research 2020; 8:9297
doi:10.4081/hpr.2020.9297

Introduction

Food plays a key role in human life and even in global economy. There is a strong correlation between eating choices and people culture, economic situation, and health status (Nishida, Uauy, Kumanyika, & Shetty 2004). Some unhealthy eating is done with intent, for various reasons, but some of it is simply thoughtless. People would often make a healthier choice if they thought about it and be happy about doing so. They know what is healthier if asked and are not necessarily averse to choosing it. But they do not habitually engage in the assessment necessary to make that choice. Calling their attention to their dietary choices, without any attempt at persuasion or even provision of new information, can be enough to improve health and welfare (Nishida *et al.*, 2004; Suthumchai *et al.*, 2016). Bad eating habits are one of the main cause of many chronic diseases such as: obesity, diabetes, dental diseases, cancer, osteoporosis and cardiovascular ones (Nishida *et al.*, 2004; Suthumchai, Thongsukh, Yusuksataporn, & Tangsripairoj 2016), which may affect the financial status of a country, because of the direct medical costs, productivity costs and also human capital cost (Hammond & Levine, 2010); in 2002, a Joint WHO/FAO (World Health Organization /Food and Agriculture Organization) Expert Consultation has proved the growing epidemic of chronic disease which affect most of the countries in the world is caused to dietary and lifestyle changes. Even though minimum life standards have improved, the raised of food availability and the higher diversified, have led serious negative consequences in terms of multiple aspects: unhealthy dietary habits; decrease of physical activities; increase in food-related chronic diseases. In 2001, chronic diseases contributed about 60% of the 56.5 million reported deaths in the world and the 46% of the global disease. It has been estimated, this percentage is going to increase to 57% by 2020. Moreover, cardiovascular problems are the cause of about half of the total chronic disease deaths. Another alarming trend is that obesity and diabetes have even started to appear earlier in people’s life. In most of the countries of the WHO, deaths caused by chronic diseases dominate the mortality statistics (Nishida *et al.*, 2004) and led the governments to interest on food and nutrition policy, health promotion, and strategies for the control and prevention of chronic diseases.

How do humans perceive food?

Global obesity epidemic led a large number of researchers to study human perception of food, the relationship to food choices and amount of food intake and the role of the visual stimuli. Visual presentation of food often affects eating behaviour and perception, and so a research method that routinely records the presentation can add a valuable overlooked component to food diaries. For example, in (Killgore & Yurgelun-Todd, 2005; Medic *et al.*, 2016; Rosenbaum Sy, Pavlovich, Leibel, & Hirsch 2008) the authors studied the relationship between brain activity, eating habits and food visual perceptions. Killgore *et al.* (2005), correlated orbitofrontal and anterior cingulate cortex activity of 13 women to the view of high-calorie and low-calorie foods. They found out that Body Mass Index (BMI) is negatively correlated with both cingulate and orbitofrontal activity during high-calorie viewing, and just with the orbitofrontal activity during low-calorie viewing. This suggests a relationship between weight and responsiveness of the orbitofrontal cortex to images that depict rewarding food. In Rosenbaum, *et al.* (2008), the authors found that maintenance of a reduced body weight was associated with changes in brain activity elicited by food-related visual cues. They perform their test on 6 obese patients and proved that this kind of brain activity can be reduced through leptin administration. Medic *et al.* (2016) examined the food choice and Magnetic Resonance Imaging (MRI) of overweight and lean people during an unlimited buffet. Their aim was to assess the capability of the two groups (lean and overweight people) to evaluate the healthiness of food. Results shown that both can well distinguish healthy from unhealthy food. This suggests that obesity can be related on how the presence of food surpasses prior value-based decision-making. Delwiche (2012) described how visual cues can affect taste and flavour of food. For example, flavour, can be viewed not just a mere combination of raw materials or chemicals components, but also as a combination of different stimuli. Multiple factors, including visual appearance, can influence the interpretation of the primary stimuli and change the perception of taste, smell, and flavour. McCrickerd and Forde (2016) focused on how visual and smell cues lead food choice. Specifically, they described how the size of food and the amount of food served can affect the food intake. Simply splitting foods like cookies or chocolate bars, so they are viewed as smaller more numerous pieces, results in a reduction of intake of that food without changing palatability. Moreover, there are evidences that indicate that some adults and children choose and consume larger portions when served with larger dishware. By observing that people seem to give more importance on the expected pleasure from food than the actual food intake, Petit, Cheok, and Oullier (2016) discussed how food-related contents published in social media can help to choose of healthy meal. Seeing food presented in an appetizing and/or "ready to be eaten" way, gives the possibility to the viewer's brain to vividly imagine the consumption experience. Currently, the food industry uses social media to promote their products with good-looking food photos. Hence, the authors claimed that public health prevention and organizations could promote healthy lifestyles by exploiting the same food industry strategies.

Such works prove that would be interesting and technically possible to use Computer Vision and Machine Learning to extract information on how the food is presented and then try to find a correlation with health statistics.

Motivation

The presented scenario led the current imaging technologies like wearable devices, tablets or smartphones, to have a fundamental role in the food intake monitoring. Such technologies allow to develop automatic systems for assessing people's diet and increase society's awareness about life's quality. Food image retrieval and classification might substitute the unreliable manual dietary assessment, which is mainly based on self-reporting. Hence, food understanding systems for mobile devices could help the creation of food-logs to assist the experts like physicians and nutritionist; such systems allow them to accurately assess the behaviour, the food choices and the eating disorders of patients, especially the ones which suffer of chronic food-related diseases. In this paper we review the literature about automatic food intake monitoring and logging technologies in order to provide the readers a comprehensive overview of this research topic.

However, in order to properly introduce these technologies, it is essential to investigate the more general topic of food understanding in Computer Vision. We use the term "food understanding" for referring to a set methods and approaches to extract information about food through automatic visual contents analysis. In fact, food intake monitoring technologies employ such Computer Vision methods to face the main challenges to automatically monitor the food intake. Figure 1 shows two typical Computer Vision tasks that can be performed on food images. The aim of image classification is to assign the input image to one general category (*i.e.*, a class) among a set of pre-defined categories. The image segmentation task aims to detect and localize areas related to the same object within the input image at pixel level (segments). The step forward is represented by the semantic segmentation task, which aims to perform the images segmentation driven by a classification method able to assign a class to each pixel, hence a class to each detected segment (Figure 1).

An effective automatic food intake monitoring application should be able to automatically answer different queries about food images: i) In which part of the image the food is located? ii) What is the food in the image? iii) Which are the ingredients? iv) What is the volume? v) What are the nutritional values?

Although Computer Vision works can be answered the first question by achieving acceptable performance, the intrinsic food variability in colour and shapes, as well as the huge number of existing ingredients, makes very challenging the development of efficient and effective techniques to face the rest of the problems.

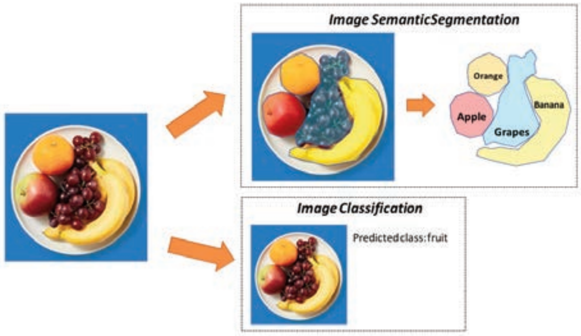

Figure 1. Difference between image semantic segmentation and classification tasks in Computer Vision.

Review method

We conducted a search of recent papers using the interdisciplinary research database Scopus, as well as academic databases specifically dedicated to computer science: ACM Digital Library, IEEE Xplore Digital Library and Springer Lecture Notes in Computer Science (LNCS). The search queries were based on the terms "food recognition", "food classification", "food portion estimation", "food logging" and "food image dataset".

We conducted literature searches from 2010 to 2020, and we selected the ones that resulted relevant with respect to the purposes of food understanding and health application. However, as we focused on automatic food logging and intake monitoring, we distinguished between papers that address specific Computer Vision tasks related to the food image analysis (*e.g.*, classification, recognition, segmentation, etc.) and papers that proposed a comprehensive food intake monitoring system. Given that methods that solve specific problems are normally employed in food logging and monitoring engine, we decided to also include them in this review, even if in a different section. In the first search in the aforementioned scientific databases we retrieved a total of 434 papers.

We excluded 329 works in the first screening, because most of them were related to medical, cultural and economic impact of food. Secondary screening was performed for the remaining 105 papers, by looking for the ones related to Computer Vision and Machine Learning. Then, we manually added 5 recent relevant studies. Finally, 23 papers that present systems for automatic food intake monitoring and logging were selected and described in the present work. In addition, we also selected 46 papers that addressed Computer Vision tasks related food images analysis which we consider essential for a comprehensive overview about this research topic. In Figure 2 it is reported a flowchart which summarizes the review strategy.

Computer Vision for food understanding

Even though food understanding has been largely addressed in the last years by Computer Vision specialists, it has a long history. From the beginning in 1977, it is possible to coarsely define four different areas: i) *Food detection and recognition for automatic harvesting:* automatic detection and recognition of vegetables are important to enhance the vision system of robots in order to improve the harvesting process in terms of quality and speed; ii) *Food quality assessment for industry:* in the 80s, industrial meals production knew a great scale expansion, especially in rich countries. Subsequently, the evaluation of food quality through vision systems became an important and worthy challenge; iii) *Food classification and retrieval:* the huge and fast spreading of mobile cameras and the diffusion of social networks, gave the chances to upload and share food's pictures. Hence, in the last few years, classification and retrieval of food images became a popular research topic. Although most of the solutions proposed in the mentioned areas overlap, the main goals of the developed systems may be different. In a nutshell, if a certain accuracy achieved by a system for the recognition of food for automatic harvesting can be acceptable in robotic industry, there is no guarantee which the same results were enough in systems for food intake monitoring, namely for patients with diseases like obesity or diabetes. For these reasons it has been chosen to categorize works about food in the aforementioned areas.

In the following subsections a detailed review the state-of-the-art works in the identified research areas is provided, in order to remark the importance of Computer Vision contribution.

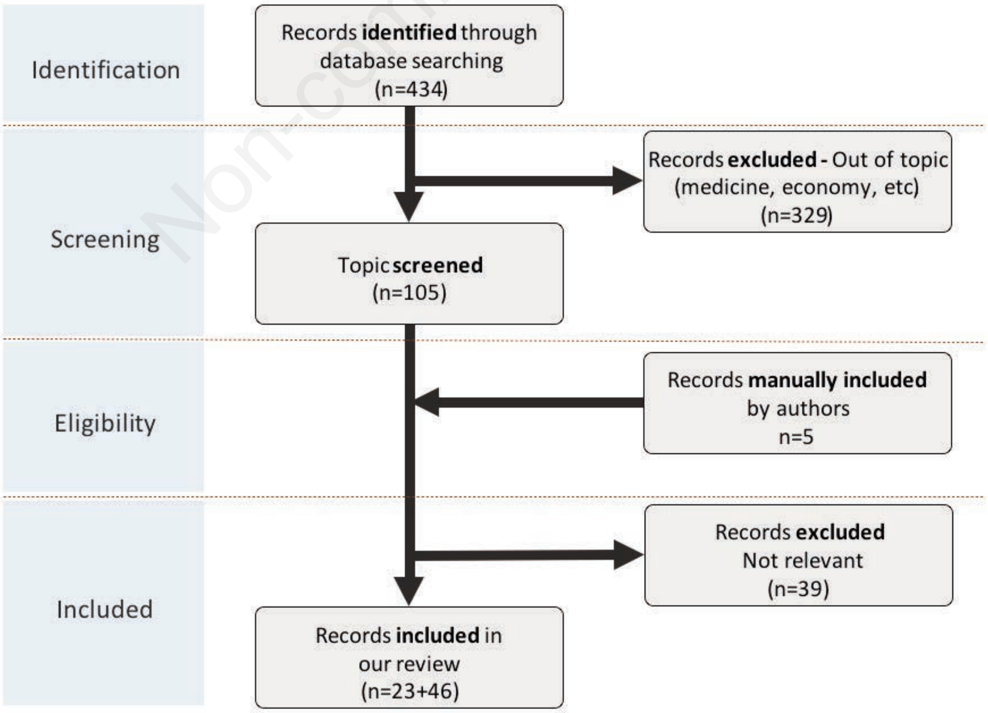
Identification
Records identified through database searching (n=434)
Screening
Records excluded - Out of topic (medicine, economy, etc) (n=329)
Topic screened (n=105)
Eligibility
Records manually included by authors n=5
Included
Records excluded Not relevant (n=39)
Records included in our review (n=23+46)

Figure 2. Flow diagram depicting the review strategy.

Food detection and recognition for automatic harvesting

Imaging techniques designed to aid in food harvesting are not necessarily much more relevant to assessing the content and appearance of a plate of food than are various other advances in imaging. But they have some common features and some examples of those provide examples of the available relevant technological development. Fruits harvesting has been addressed by several strategies; however, it is mandatory that they do not cause damages to the fruit and/or to the tree/branches. This means that accurate systems for fruits/vegetables detection and classification are useful in order to perform this task correctly. One of the first Computer Vision solutions has been proposed in 1977 by Parrish and Goksel (1997), and focuses on apples detection. The system they designed consists of a B/W camera and a red filter. First, the acquired image is binarized and then smoothed to reduce noise and artefacts. Then, the region roundness is estimated by evaluating the difference between the longest horizontal and vertical segments inside the region. Secondly, an area in the image is classified as apple through a density estimation procedure and a further thresholding step. Levi, Falla, and Pappalardo (1988) proposed a robot vision system for oranges recognition. In the first step, a pseudo-grey image is got by means of an electronic filter employed for image enhancement. The value of each of the pixels, coded using 6 bits, is proportional to the difference between the hue value and a reference hue value. Then, the edges are computed through a Sobel filter to get the magnitude and directions in two separate matrices. To correctly detect the oranges location, they perform a matching between the detected gradient and a predefined template. This strategy achieves an accuracy of 70% in oranges detection.

Another orange recognition approach is proposed in (Slaughter & Harrell 1987); it uses Hue and Saturation components of each pixel as coordinates of a two-dimensional feature space. Then, two thresholds based on the maximum and minimum values are used to define a certain region in the feature plane. Hence, each pixel inside this region is classified as orange. This method achieved 75% of the pixels correctly classified. The authors extend their study by using a Bayesian classifier (Slaughter *et al.*, 1989), and exploiting the RGB colour space rather than the Hue and Saturation components. The performed tests show again an accuracy of 75%.

Cardenas-Weber *et al.* proposed a machine video system for melon harvesting (Cardenas-Weber Hetzroni, & Miles 1991), developed at The Purdue University (USA) and The Volcani Center (Israel). The proposed method is able to analyse binary image to find the melons and measure their size. Operations like shape and textures analysis are performed in order to get different candidate regions from the original image. Then, thanks to prior knowledge on the domain, the candidate regions are evaluated to avoid false positive and multiple detections, with a true positive rate of 84%.

In 1995, Buemi *et al.* of the Italian institute CIRAA, proposed a robotic system called AGROBOT (Buemi Massa, & Sandini 1995). Their goal was automatizing greenhouse works. The images are acquired through a colour camera and are segmented through a thresholding on the Hue and Saturation histograms. Their method is able to extrapolate information about the 3D geometry of the scene by using stereo matching. The performances of AGROBOT shown 90% of correctly detected ripe tomatoes, where the main error causes in this system depends on the occlusions.

In general, research works on automatic harvesting led to improvements of the techniques for the estimation of food geometry. Such methods can be generalized and applied to other food recognition tasks (*i.e.*, food detection, segmentation, volume estimation, etc.).

Industrial food quality assessment

Although food quality inspection is not strictly related to domain of dietary food monitoring, it still concerns food image analysis. In recent years Computer Vision systems have been used for quality assessment, as this task plays a key-role in food industry that manufactures products that satisfy their customers. In Munkevik, Duckett, and Hall (2004) proposed an approach to assess the quality of industrial cooked meals.

They proposed to perform a segmentation on food images and then extract 18 different features from the obtained segments. Through these features, it is possible to represent different properties. Among them, the overlapping between different food items, the size of the food items, the shape of the food items. Secondly, a SOM (Self-Organizing Feature Map) (Kohonen, 1998) is employed to learn the model of a meal. The same authors extend their work in Munkevik, Duckett, and Hall (2005), by considering a larger number of food items and by exploiting an Artificial Neural Network (ANN) to improve classification performances.

In 2007, Kilic, Boyaci, Köksel, & Küsmenoglu (2007) addressed a beans quality estimation problem. For the experiments they use a dataset that consists of 511 images with a variable number of beans. Morphological operators are employed for image segmentation, then they compute the first 4 order statistic on the RGB channels as features. Beans quality is assessed by using a score based on three levels for both integrity and colour. Even though, 3×3=9 possible combinations can be defined for such score, the authors decided to exploit just 5 combinations; each of the combinations is considered as a different class. Finally, the classification is performed by using ANN and splitting the dataset in the following way: 69 beans images for training, 71 for validation and 371 for testing.

The quality of pizza production has been addressed in several works, as the ones of Du and Sun (Sun, 2000; Du & Sun, 2008). The proposed algorithms are intended to inspect three different pizza properties: shapes, toppings and sauce spread. While the approach in Sun (2000) faces only the evenness of the topping, the method described in Du and Sun (2008) is more complex and involves also other parts of the pizza to be evaluated. To perform quality assessment by exploiting shape, geometrical features such as the area ratio, aspect ratio, eccentricity, roundness and also some coefficients of the Fourier transform have been considered.

Concerning the topping and the sauces, HSV colour histograms and Principal Components Analysis (PCA) are used. Classification is performed by considering four quality levels for the shape and five for topping and sauce spread. To build the classification model a set of binary Support Vector Machine (SVM) organized in a Directed Acyclic Graph (DAG) has been employed. The dataset used for experiments includes 120 images for the shape, 120 images for the sauce and 120 images for the topping.

Finally, a review about the methods for food quality assessment is presented in Gunasekaran, (1996), Brosnan and Sun (2004), and Du and Sun, (2006). The authors address the different acquisition systems as well as the features that can be employed in different tasks. In particular, the overview presented in (Gunasekaran, 1996) revises vision techniques based on morphology and pattern recognition methods that were extensively exploited in industry up until '90s. In Brosnan and Sun (2004) the authors suggest that further developments on X-ray and 3D vision techniques would provide a great contribute for the future industry, whereas machine learning algorithms used to perform the decision for this task are highlighted in Du and Sun (2006).

The inspection of the food quality is usually addressed in constrained environment, with a few food classes and low variability. For this reason, very simple features such as colour or shape information are enough to face the problem and achieve very good results. This kind of scenario is different from the one where images of food are acquired during real meals of a patient or they are downloaded from a social network. A generic system for food intake monitoring has to be able to work in low constrained scenario without prior knowledge. Differently than an industrial factory where the ingredients, the quantity and the appearance of food are known in advance, in a generic food understanding problem there are many variables. High number of food classes and ingredients, the food mixing as well as illumination, orientation, different acquisition devices and so on, make this task very challenging.

Food identification: classification and retrieval

The scientific works described so far are related to specific tasks, *e.g.*, quality assessment, fruit recognition, food logging, etc. All the methods applied on different application fields have a common sub-task related to the recognition of the food depicted in an image. Exploiting the ever-growing availability of food images due to the diffusion of social media and image sharing platforms, the computer vision community investigated the task of food recognition in the last years. This allowed the definition of large-scale public dataset of human-labelled food images, with large variation in the number of samples, classes and type of labelling.

In the context of image identification, there are two main tasks. In both cases the algorithms are trained using a specific dataset of labelled images, also known as training set. Then, the algorithms are evaluated considering images that are not included in the training set. The image classification task aims to assign one predefined class to new instances of image depicting a food instance. During the training stage, 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 (Scale Invariant Feature Transform) or Textons features (Battiato Farinella, Gallo, & Ravì, 2010; Lazebnik, Schmid, & Ponce, 2005) whereas a learning mechanism is used to train a classifier (*e.g.*, a Support Vector Machine) to discriminate data belonging to different classes. After the training stage, new observations can be classified by considering the employed feature space and the trained classification model (*i.e.*, the training images are no longer needed). In the image retrieval problem, the input image is compared with a set of already known images (*i.e.*, training images) and the identification is performed comparing the images through similarity measures after their representation in the feature space.

The work by Yanai and Joutou (2009) proposes a food classification framework trained on a dataset of 50 Japanese food categories. The proposed approach extracts three features from the visual content: i) Bag of SIFT; ii) Colour Histograms; iii) Gabor Filters (Marĉelja, 1980).

The classification method exploits a Multiple Kernel Learning SVM (MKL-SVM) (Varma & Ray, 2007). The authors of Hoashi, Joutou, & Yanai (2010) further extended the dataset to 85 categories, and as well as 8 variants of Histogram of Oriented Gradients (HOG) (Dalal & Triggs, 2005) have been evaluated as new features. An extended version of the dataset, considering 100 food categories, has been proposed in Matsuda, Hoashi, & Yanai (2012), in this work candidate regions are identified using different methods: whole image, Deformable Part Model (DPM) (Felzenszwalb Girshick, Mcallester, & Ramanan 2010), a circle and the segmentation method proposed in Deng and Manjunath (2001). From each detected candidate region, the proposed approach extracts the following features: i) Bag of SIFT; ii) Bag of CSIFT (Abdel-Hakim & Farag, 2006); iii) HOG and Gabor Filters; iv) Spatial Pyramid Representation (Lazebnik, Schmid, & Ponce 2006). Experiments have been performed considering images containing either single and multiple food instances. In a successive work (Matsuda *et al.*, 2012) 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 extension of the dataset presented in Yanai and Joutou (2009) and Hoashi *et al.* (2010), and exploited in Matsuda *et al.* (2012), is called UEC FOOD 100. Thanks to its availability, it has been exploited considering different approaches, such as pretrained Convolutional Neural Networks (CNN) (Krizhevsky, Sutskever, & Hinton, 2012) are used in (Kawano *et al.*, 2014) for feature extraction. The CNN features are coded using the Fisher Vectors technique (Sánchez, Perronnin, Mensink, & Verbeek 2013), and then the classification is performed by means of SVM. Raví, Lo, & Yang (2015) exploited jointly different features in a hierarchy to obtain real-time food intake classification.

In particular, the Fisher Vector technique (Perronnin & Dance, 2007) is employed, and Principal Component Analysis (PCA) is applied as in Perronnin, Sánchez, & Mensink, 2010. In Kawano and Yanai (2014) the UEC FOOD100 has been extended to 256 classes (UEC FOOD 256) exploiting crowdsourced images and information. Yanai and Kawano (2015), exploited UEC FOOD 100 and UEC FOOD 256 and transfer learning to fine-tune a deep convolutional neural network previously trained on object detection.

The Pittsburgh Food Image Dataset (PFID) (Chen *et al.*, 2009) counts 4.545 images, 606 stereopairs, 303 306° video sequences for structure from motion, and 27 privacy-preserving video sequences of eating events. The images depict 3 instances of 101 food items, bought in 11 different fast food chains. A reference experimental baseline on PFID has been presented in (Chen *et al.*, 2009). The authors used colour histograms and Bag of SIFT features to build an SVM classifier. In Yang, Chen, Pomerleau, and Sukthankar 2010), an ingredient-based segmentation is performed using a Semantic Texton Forest (Shotton *et al.*, 2013). Hence, pairwise statistics of local features are computed on the segment connecting two points, and specifically: orientation, between-pair, midpoint, distance.

The PFID is also used for calories estimation in Wu and Yang, 2009. SIFT are extracted and a cosine-based distance function is used for matching. Rankings on food categories can be obtained in two ways: i) ranking-based matching, based on top T items of each frame-based rankings; ii) count-based matching based on sum of keypoint matching counts over all video frames.

Zong Nguyen, Ogunbona, & Li (2010) locate the keypoints using the SIFT detector, applying the Local Binary Pattern (LBP) (Ahonen Hadid, & Pietikäinen 2006). 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 (Belongie, Alik, & Puzicha, 2002) is calculated on the image space, considering the words as keypoints. Nguyen

Zong, and Ogunbona (2010) extended the previous mentioned approach introducing the Non-Redundant Local Binary Pattern (NRLBP) and propose two strategies to classify the images. The first exploits an SVM classifier, the second is based on a cost function. In 2014 and 2015 Farinella, Moltisanti, & Battiato, (2014) proposed two different approaches on the attempt to classify the images of PFID. The work in 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 approach proposed in (Farinella *et al.*, 2015), SIFT and SPIN (Lazebnik *et al.*, 2005) 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 (Topchy, Jain, & Punch, 2005). In both approaches, SVM is used as classifier. The method proposed in (Bettadapura, Thomaz, Parnami, Abowd, & Essa 2015) combines different descriptors calculated on patched centred on the keypoints detected by the Harris-Laplace detector. For each feature, a visual 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.* (2011) proposed a method for food identification based on a combination of global and local image features. As global features, they used: i) 1st and 2nd moment statistics computed on the colour channels of the image; ii) entropy statistics; iii) predominant colour statistics.

The following local features have been extracted from small image patches: i) local colour statistics; ii) local entropy colour; iii) Tamura features; iv) Gabor filters; v) SIFT descriptor; vi) Haar wavelets; vii) Steerable filters; viii) DAISY descriptor (Tola, Lepetit, & Fua 2009).

Global features have been used to train an RBF-SVM, whereas local ones have been used to train a Bag of Visual Words approach for image representation and K-Nearest Neighbour for classification. This approach was tested on a subset of the dataset presented in Bosh *et al.* (2011) and obtained by extending the USDA Food and Nutrient Database for Dietary Studies (FNDDS), with the aim 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, Pickering, Kerr, Boushey, & Delp, (2012) presented a dataset of 209 images acquired by using an iPhone 3, to be used for mobile image retrieval purposes. Another system for mobile food recognition has been proposed in Kawano and Yanai (2013). Colour histograms on the RGB space have been 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 (Vedaldi & Zisserman, 2012) is employed. It is interesting to note that the authors proposed a system able to suggest the direction to which the camera should be moved, in order to improve the classifier accuracy. Also, a dataset with 50 categories, containing 100 images each, is presented. A Computer Vision system for Chinese food identification has been proposed by Chen *et al.* (2012). The authors employed a dataset of ready-to-eat Chinese meals, with 50 classes and 100 images per category. From each image, the following features have been extracted: i) SIFT with sparse coding; ii) LBP with multi-resolution sparse coding; iii) colour histograms; iv) Gabor textures. An SVM is trained for each feature using 5-fold cross validation; the fusion is done using the Multi-Class AdaBoost algorithm. Marginally, the authors also proposed 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.* (2013). In this work, an imaging system composed by a matrix of optical fibres is properly set under an appropriately prepared chopping board. The sensors acquire the image and afterwards a 64-dimensional colour histogram and a 64-dimensional vector of Bag of SURF features are computed. The algorithms used to classify the images are K-Nearest Neighbour and SVM.

Random Forest (RF) (Ho, 1995) are used in Bossard, Guillaumin, and Van Gool (2014) for mining discriminative regions. Superpixels are generated from the images and dense SURF and colour histograms are computed and encoded using Fisher Vectors (Sánchez *et al.*, 2013). 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 by 1000 images for each one of the 101 most popular dishes on foodspotting.com.

In Xin, Kumar, Thome, Cord, and Precioso (2015) propose UPMC Food-101, a new dataset of 101000 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.

Other food datasets include images and related geocontext information, such as GPS coordinates, restaurant where the dish is cooked and so on. Herranz, Ruihan, and Shuqiang (2015) propose a probabilistic model to combine locations, restaurants, and visual features by exploiting a reduced set of the dataset collected by Ruihan *et al.* (2015) from Institute of Computing Technology, CAS. Each restaurant is associated with the related geographical coordinates to uniquely locate it and a menu that includes at least three dish categories. Then, for each of category, more than 15 images are included. Farinella, Allegra & Stanco (2015) propose UNICT-FD899. This dataset has been acquired by users with a smartphone in four years during meals (*i.e.*, iPhone 3-GS or iPhone 4) in unconstrained settings. Each dish has been acquired through the smartphone multiple times to introduce photometric and geometric variability (rotation, scale, point of view changes). The overall dataset contains 3.583. 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, cropping, etc. Other techniques such as those discussed in Battiato, Farinella, Puglisi, and Ravì (2014) and Hu *et al.* (2009) consider as near duplicate the images of the same scene but with different viewpoint and illumination. In Farinella *et al.* (2015) 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: i) Bag of Textons (Varma & Zisserman 2005), ii) PRICoLBP (Qi *et al.*, 2014) and iii) SIFT (Lowe, 2004). Results confirm that textures and colours are fundamental properties. The experiments performed point out that Bag of Textons representation is more accurate than the other two approaches for NDIR. UNICT-FD889 dataset is a collection of food images acquired by users in real cases of meals. Each plate of food has been acquired multiple times (four in the average) to guarantee the presence of geometric and photometric variability. It is designed to arouse research in this application domain with the aim of finding a good way to represent food images for recognition purposes. In 2016 Farinella, Allegra, Moltisanti, Stanco,and Battiato (2016) extend UNICT-FD899 by introducing new food classes and proposing UNICT-FD1200. In this work, new experiments have been conducted with the previous image descriptors and a novel representation called Anti-Textons is proposed. It exploits the co-occurrences between standard Textons to improve the effectiveness of texture description by outperforming the original methods. A comparative analysis on features and classifiers is the core of He, Xu, Khanna, Boushey, and Delp (2014). The authors test several features, basically related to three aspects (colour, texture, local regions) and two classifiers (kNN, Vocabulary Tree) on a novel dataset composed by 42 classes, with a total of 1453 images (Nistér & Stewénius, 2006). Most of the food recognition approaches classify food images considering a set of class labels that describe the whole recipe. Donadello and Dragon (2019) presented an ontology-based approach that models the knowledge of recipes, food categories and their relationship with chronic diseases by defining an ontology. FRIDa dataset has been proposed in Foroni, Pergola, Argiris, and Rumiati, (2013) and includes 877 images belonging to 8 different categories: natural-food, transformed-food (*e.g.*, cooked food), rotten-food (*e.g.*, mouldy fruits), natural-non-food items (*e.g.*, pinecone), artificial food-related objects (*e.g.*, fork, spoon), artificial objects, animals (*e.g.*, butterfly), and scenes (*e.g.*, mountains). This dataset has been validated on a sample of 73 standard variables (*e.g.*, ambiguity, familiarity, etc.) as well as variables related to food items (*e.g.*, distance from eatability, perceived calorie content, etc.).

Pouladzadeh, Yassine, and Shirmohammadi (2015) introduced FooDD. It is a dataset of 3000 images across a large variety of food photos taken from different devices and under different illumination conditions. The authors have used colour segmentation and k-mean clustering in order to perform food segmentation; then, they have employed Cloud SVM and deep neural network for recognition and calories estimation.

In Farinella, Allegro and Stanco (2015), the authors proposed to address the binary classification between food and non-food images by using One Class Classification paradigm and, specifically, One Class SVM method with Bag of Texton features. Ragusa, Tomaselli, Furnari, Battiato, and Farinella (2016) outperform the result of (Farinella, Allegro, & Stanco 2015), by employing deep neural network for features extraction before using One Class SVM. In Salvador *et al.* (2017) the authors presented a very large dataset named Recipe1M, consisting of over 1M written recipes and 800.000 related food images. The authors defined a model able to retrieve the textual recipe of a food dish by the analysis of an image of food. The system is able to infer both the ingredients and the cooking instructions for a given image. This work has been extended by Marin *et al.* (2019), in which the number of images in the dataset (named Recipe1M+) is about 13M.

In Fontanellaz, Christodoulidis, and Mougiakakou (2019), the authors proposed a method for the joint learning of meal images and recipe embedding, using a multi-path structure that incorporates natural language processing paths, as well as image analysis path. They used the dataset Recipe1M for training a testing. Table 1 details the main features of the above described datasets published in the last years.

Food logging, dietary management and food intake monitoring

The growth of the number of people affected by diseases caused by a non-healthy diet led the researchers to study the problem. From late 90s, the focus was moved to the usage of Computer Vision solutions to help food experts (*e.g.*, nutritionists) for the monitoring and understanding the relationships between patients and their meals. The first systems for food logging and intake monitoring were calculators for nutrition values that exploited standard food list (Rich, 1981; Wright, Shearing, Rich, & Johnston, 1978). Hence, they did not use the Computer Vision techniques. In the last decade Computer Vision researchers have put effort to propose reliable tools to improve the automatic detection and recognition of food images, as well as the nutritional merit evaluation. These

Table 1. Publicly available food datasets.

Dataset	Presented in	Classes	Images per class	N. of images
UEC FOOD 100	Matsuda *et al.*, 2012	100	≈100	9060
UEC FOOD 256	Kawano & Yanai, 2014	256	≈100	31651
PFID	Chen *et al.*, 2009	101	18	1818
FRIDa	Foroni *et al.*, 2013	8	ND	877
NTU-FOOD	Chen *et al.*, 2012	50	100	5000
ETHZ Food-101	Bossard *et al.*, 2014	101	1000	101000
UNICT-FD889	Farinella, Allegra *et al.*, 2015	899	3/4	3583
FooDD	Fouladzadeh *et al.*, 2015	23	ND	3000
UPMC Food-101	Xin *et al.*, 2015	101	1000	101000
CAS Dataset	Herranz *et al.*, 2015	ND	ND	117504
Recipe1M	Salvador *et al.*, 2017	≈1M	ND	≈800000
Recipe1M+	Marin *et al.*, 2019	≈1M	ND	≈13M

types of tools can increase self-awareness of eating habits, moreover, to add photographs to the written diary have a more effective impact on the patients. A discussion about the state-of-art systems for food intake monitoring a logging is given below.

FoodLog (http://www.foodlog.jp, Aizawa, Silva, Ogawa, & Sato 2010; De Silva *et al.*, 2010; Kitamura, Yamasaki, & Aizawa 2008; Kitamura, Yamasaki, & Aizawa 2009; Kitamura, De Silva, Yamasaki, & Aizawa 2010) is an Internet application that gives the possibility to acquire and store information regarding daily meals. The main aim of this system is to help the users to keep note of their meals and, above all, to correctly balance the main nutrients coming from different kinds of food (*e.g.*, carbohydrates, protein, etc.). The application enables the user to upload one or more pictures on a remote folder, where all information is stored.

Kitamura *et al.* proposed FoodLog in Kitamura *et al.* (2008). The images that include food items are detected by using colour features based on HSV and RGB, as well as the shape of the plate. Food detection is performed by training a SVM classifier according to the following strategy: the images are divided in 300 blocks and each block is classified as one of the five nutritional groups described in the "My Pyramid" official model (grains, vegetables, meat & beans, fruits, milk) or as "non-food". However, this model has been replaced in 2011 by the "MyPlate" model.

In 2009, Kitamura *et al.* (2009) extended their previous work by exploiting more local features. Colour information are coupled with SIFT descriptors (Lowe, 2004) by selecting keypoints with three different methods (Difference of Gaussians, centres of grid, centres of circles). Further improvements are proposed in Kitamura, *et al.* (2010), by including a pre-classification step and the customization of the food image estimator. Finally, in Maruyama, De Silva, Yamasaki, & Aizawa (2010) the Support Vector Machine classifier is replaced by a Naive Bayesian one. Shroff, Smailagic, and Siewiorek (2008) proposed a mobile phone-based calorie monitoring system to help people to follow their dietary rules. The authors employ two different kinds of features: objected-related features like colour, size, texture, shape; context features such as time of the day or user preferences. The authors use ANN classifier to prove that the context information led an improvement in the accuracy of the monitoring system. However, this technology, named DiaWear, requires the user to provide additional contextual information for better food recognition.

The work of Puri, Zhu, Yu, Divakaran, and Sawhney (2009) focuses on food recognition and 3D volume estimation. First, the photos, captured under different lighting conditions and poses, are normalized by colour and scale by using a particular calibration card placed besides the food items. For features selection they employ an Adaboost-based algorithm that combines colour (in RGB and LAB space) and texture information (Maximum Response filters). The goal is to perform a segmentation by classifying the different food items in a plate. The final classifier is obtained as a linear combination of several weak SVM classifiers, one for each feature. For 3D reconstruction they use RANSAC (Fischler & Bolles, 1981) to estimate pose and dense stereo matching for depth estimation.

Another work in which 3D reconstruction is exploited is the one by Dehais, Shevchik, Diem, & Mougiakakou (2013). The 3D model is used for food volume estimation. Stereo pairs are used to computer disparity map and then a dense points cloud is built and aligned with respect to the estimated table plane. This algorithm is designed to work by employing a specific marker placed on the table. By assuming the different food items in the plate are already segmented, each food segment is projected on the 3D model for volume computation. They define the volume 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). Allegra *et al.* (2017), proposed to exploit RGB-D images to learn a model for depth estimation. They performed semantic segmentation through U-Net and used a CNN for depth inference from monocular RGB image. Differently, Lu *et al.* (2018) proposed a Multi-Tasking Learning model to estimate volume of food dished from single RGB input. The proposed CNN consists of feature extraction module which uses ResNet50 and Feature Pyramid Network (FPN); then, a depth prediction net based on an autoencoder with skip connections is employed; a semantic segmentation step is performed by a Region Proposal Network (RPN) and then the volume is obtained through a CNN regressor. The work by Allegra *et al.* (2019), addressed food volume estimation as raking problem in a constrained scenario; the authors proposed to use Ranking SVM to sort food images according the food amount in the dishes (Figure 3).

In Chen, Lee, Rabb, and Schatz, (2010) the authors categorise food from video sequences taken in a supervised environment. The dishes are placed on a table covered with a black tablecloth. They considered an elliptical Region-of-Interest (ROI) and extract different kind of descriptors such as MSER (Matas, Chum, Urban, & Pajdla, 2004), SURF (Bay, Tuytelaars, & Van Gool 2006) and STAR (Agrawal, Konolige, & Blas 2008). Hence, the images are represented exploiting the Bag of Words paradigm and vocabulary with 10000 visual words built by using K-means clustering. Subsequently each data point is associated with the closest cluster using the Approximated Nearest Neighbour algorithm. To capture information about colour, histogram in the HSV space is computed inside the ROI and combined with the aforementioned descriptors. The final aim is to classify the dish in a specific frame of the sequence. In the proposed approach each unclassified frame is compared with frames that are already classified. To do this, a similarity score is computed for both, the Bag of Words representation and the colour histograms. The score for the first representation is computed by exploiting the term frequency-inverse document frequency (tf-idf) (Salton, 1989) technique, while for the colour similarity, the correlation coefficient between the $|L_1|$-norm of two histograms is used.

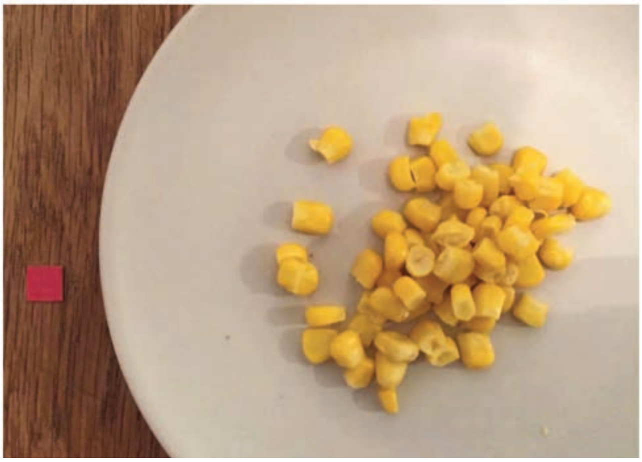

Figure 3. Example of food acquisition performed in McAllister *et al.*, (2015): food portion with a 1 cm^2 square next to the plate.

Finally, the two scores are linearly combined with different weights to obtain a global score for the considered frame. Moreover, since the calories for the reference dish are known, this score allow to roughly quantify the difference of them in the two frames. Food intake estimation is also studied in the work of Liu *et al.* (2012) where a wearable system equipped with a camera and a microphone is proposed. The microphone is used to detect chewing sounds, so that the Computer Vision part of the framework can be activated. To identify frames that contains food, they propose to use a simple approach based on ellipse detection and colour histograms. After the ellipse is found, it is split in four quadrants and, for each of them, the colour histogram is computed in the C-colour space (Burghouts & Geusebroek, 2009). Finally, the food consumption evaluation is performed by computing the difference between the histogram of subsequent frames.

The paper by Kong and Tan (2012) proposed a smartphone camera-based food intake monitoring system, named DietCam, aimed to help the user to assess the real food intake. The system requires to provide three images, or a video recorded during the meal. Before the use, the smartphone camera needs to be calibrated. To do so, users must take three pictures around the dish approximately every 120deg or record a short video of the plate with a credit card put beside. To perform the food classification, the authors used a Bayesian probabilistic approach.

In the last decade, a subfield of Machine Learning named Deep Learning, based on the extensive use of Artificial Neural Networks with high number of layers and artificial neurons led to significant improvements on several data analysis fields such as text analysis, audio and image processing and recognition. In particular, the rapid diffusion of Deep Learning techniques applied to Computer Vision techniques allowed the rise of performances on several visual tasks (Plebe & Grasso, 2019), including the food intake estimation problem. One of the first Deep Learning approach applied to this field has been presented by Meyers *et al.* (2015), that proposed a system named Im2Calories in 2015. The authors considered two different tasks.

The base task assumes that the food plate under analysis comes from restaurants where food menus were available. In this simplest case, the system is trained on food images taken from 23 different restaurant.

Then, at runtime, the classifier is used to predict which food are present on the plate and compute the corresponding calories. The second task consists of estimating the size of foods present in the scene. To do so, the system first perform a food segmentation followed by a volume estimation. This is obtained by combining two Convolutional Neural Networks (CNNs) for the meal detection and food recognition, with a food image segmentation technique and Google's Places API used to recognize the restaurant. Finally, the calories estimation is performed based on the U.S. Department of Agriculture (USDA) database. The proposed method has been evaluated on different 2D (*e.g.*, food vs. non-food, multi class, label per pixel, etc.) and 3D (*e.g.*, depth per level) tasks related to food intake estimation by using very large scale datasets (*i.e.*, from 10k to 150k images each) and different labelling schemes (*e.g.*, 2, 101, 201, or 2517 classes).

The work in McAllister, Zheng, Bond, and Moorhead (2015) consists of a semi-supervised technique to predict calories using a regression model. At the early stage, the user is required to use a polygon drawing tool to segment his/her food portion, the application then segments and save the segmented regions of food. The built dataset is then combined with a regression model to estimate calories of future portions without supervision. The estimation is based on a reference block of 1 cm^2 placed next to the plate. An example of portion acquisition is shown in Figure 3.

The Snap-n-eat system presented in (Zhang, Yu, Siddiquie, Divakaran, & Sawhney 2015) is able to estimate the calorific and nutritional content of the food items on the plate from only one image by counting the pixel assigned to the same kind of detected food. The image is first segmented into regions, then the detected regions are classified using a linear Support Vector Classifier (SVM). An advantage of such approach is that it needs only one input image and that is suitable to be applied on cluttered images. Figure 4 shows an example of detection of the food items in the scene performed by Snap-n-eat: first a saliency map is extracted from the original image, then the map is exploited to detect the location of the food in the image by applying a threshold on the saliency values.

The work by Zhu *et al.* (2010) presents a framework based on a client-server interaction between the patient's smartphone and a remote server. The user takes a picture of the meal before eating and sends it to the server through a proper smartphone app, together with additional meta-data (*e.g.*, date, time, GPS location). Algorithms running on the server perform an estimate on the kind of foods (*i.e.*, food identification) and volumes of each detected item. In particular, the system first performs an image segmentation on the input image to distinguish all the food items. Then, the segmented image and the original ones are used to classify each food item by a Linear SVM classifier. The same pair of input images are used to perform the volume estimation. To this aim, a set of feature points are extracted from the area related to the single food item, then the shape of the food is approximate by a spherical or prismatic 3D model. This process relies on a marker placed next to the plate, which dimensions and shape are known. The results are sent back to the smartphone app, requiring the user to confirm or adjust the inferred outputs. After this process, the system stores all the information related to food items and quantities that are correlated with the nutrient information of the Food and Nutrient Database for Dietary Studies (FNDSS) (Ahuja *et al.*, 2012). The FNDSS contains several information related to the most common foods consumed in the United States, their nutrient values and quantities for typical portions. The method proposed in (Fang, Liu, Zhu, Delp, & Boushey 2015) extends the work in (Zhu *et al.*, 2010) by further generalizing the shape estimation, adding the use of a geometric cylinder model for items such as liquid in a glass or in a bowl. The system performs the food volume estimation by exploiting contextual geometrical information extracted from the scene.

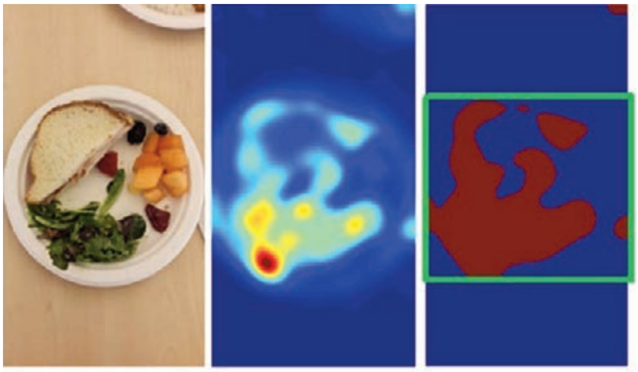

Figure 4. Example of saliency computation and food detection performed by Zhang *et al.* (2015).

Indeed, this approach first estimates the geometry of each object by approximating it with a cylinder or a prism. Then, the parameters of such known 3D models are inferred by relying on a well-known marker placed on the scene as in (Zhu *et al.*, 2010). The estimation of such 3D models allows the system to perform the volume estimation on each item in the scene (*i.e.*, food and beverage items). Figure 5 shows the experimental settings used in Zhu *et al.* (2010) and Fang *et al.*, (2015). In this scenario, the black background, the marker and the fact that the objects are not overlapping represent very strong constrains that are eligible only in a lab environment. The work in Sun *et al.* (2014) presents an overview of a wearable device called eButton, designed to perform daily living monitoring of the patient for several applications such as the evaluation of diet and activity, sedentary behaviour detection, blind and visually impaired assistance, and monitoring of elder people suffering from dementia. The food recognition system of eButton automatically detects and quantify the food consumed by the patients without his/her intervention. Compared to other methods, this approach is fully passive, as the system automatically take and analyse one frame every 2 seconds.

Figure 6 shows an example of acquisition, whereas Figure 7 shows an example of food detection and estimation. In particular, the inference is performed in three steps: detection of contextual objects such as plates or bowls, segmentation of food items and estimation of volumes (Figure 7).

The system proposed in Akpro Hippocrate Suwa, Arakawa, & Yasumoto (2016) exploits the presence of eating tools (*e.g.*, spoon, fork, chopsticks, etc.) as a reference to measure the volume of the food containers with a known shape (*e.g.*, plate, bowl, etc.) by which estimate the food quantity. This approach presents several issues that cause the over-estimation of food shown in the experiments. Indeed, in most of the cases, the food volume does not correspond to the container one. Moreover, this approach assumes that the cutlery items are always on the scene, and that they have standard and well-known shape dimension. The system proposed in (Hassannejad *et al.*, 2017) automatically extracts 6 frames from a short video of meal consumption. Then, the user is required to mark initial segments of food items in one frame.

Starting from this input, the system performs the food items segmentation and classification. The estimation of real volumes is based on the presence of a marker as done in Fang *et al.* (2015).

In Lu *et al.* (2019) the authors proposed an AI-based and fully automatic monitoring system for nutrient intake by hospitalised patients, which analyse the RGB-D image pairs captured before and after the meal. To train and evaluate the system they introduced a new dataset of food images with nutrient recipes. They used a Multi-Task Fully Convolutional Network for image seg-

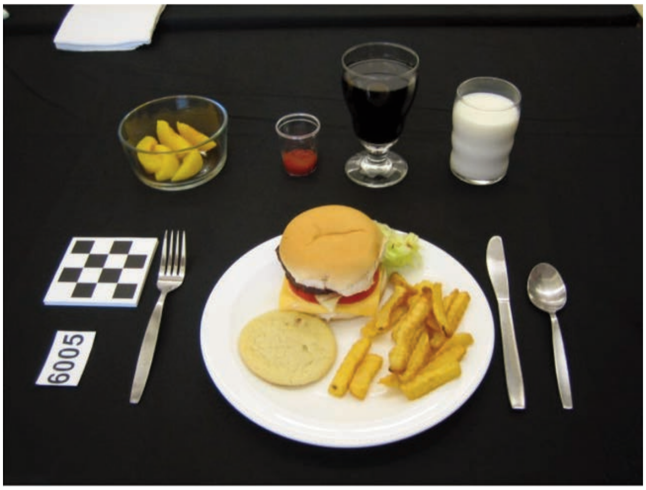

Figure 5. Example of input image analysed in Fang *et al.* (2015)

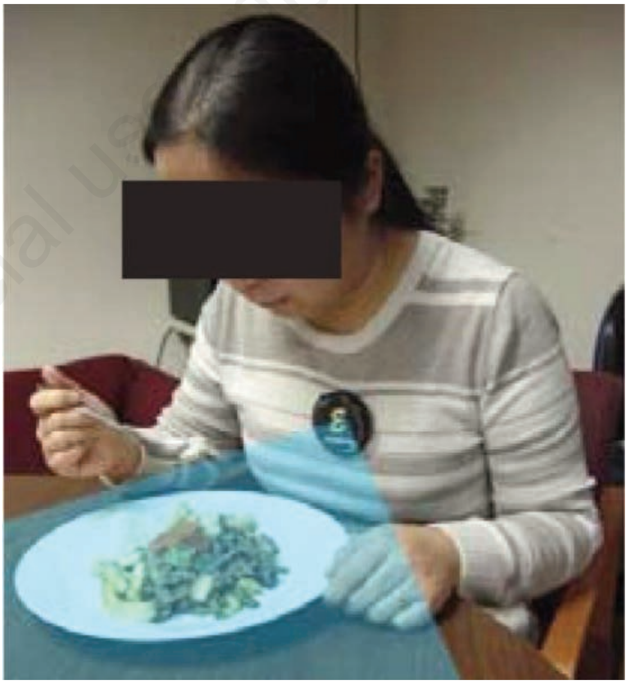

Figure 6. Automatic food detection performed by eButton (Sun *et al.*, 2014).

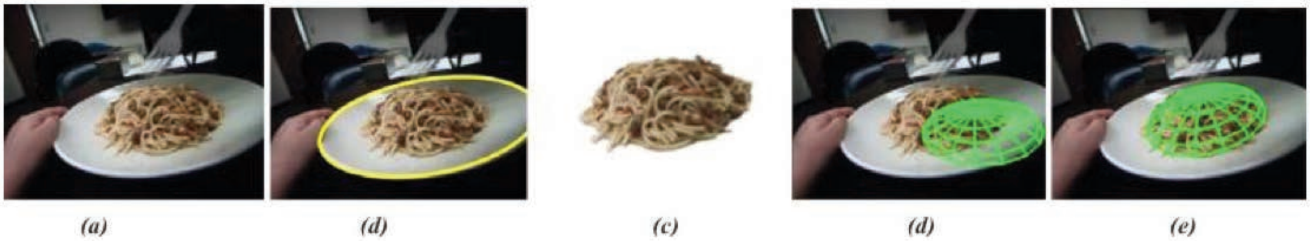

Figure 7. Image processing performed in Sun *et al.* (2014): a) input image, b) plate detection, c) food segmentation, d) shape modelling and e) shape fitting.

mentation, which takes RGB-D image pair as input, and outputs the types of food and plate in two different segmentation maps; the estimation of consumed food is derived by subtracting the food volumes before and after the meal, which are estimated by using 3D food surface and the plate surface combined with RANSAC. Although the recently developed methods of food intake estimation show promising results, this task still presents many challenges. Indeed, the quantitative results reported in the papers have been obtained considering limited number of food items and/or a small number of samples per meal often taken in constrained and controlled conditions available only in laboratory. Furthermore, most of them requires proper markers, multiple user intervention or a calibration stage. Hence, new approaches that correctly estimate the food intake over long time periods in free living settings from only a single pre-meal image. Table 2 summarizes the above described papers focusing on their feasibility. Indeed, the aimed application of these solutions is the automatic food intake monitoring in real-life scenarios (*e.g.*, they can be exploited in a scientific control experiment), however our study observed that most of the methods have an excessive need of user intervention (*e.g.*, several pictures, contextual information) or requires specific hardware or controlled settings. Therefore, systems that require only minimal intervention by the user (*i.e.*, one or two images or a short video of the plate) are marked as "Automatic", whereas systems that have been designed to work by exploiting a simple smartphone or a wearable device are marked as "Wearable". With only a few exception, such as Zhang *et al*. (2015) or Meyers *et al.* (2015), most of the listed approaches require more than two pictures by the user or are not designed for everyday practical usage. Moreover, several approaches marked either as "Automatic" and "Wearable" (*e.g.*, Zhu *et al.*, 2010; Fang *et al.*, 2015) still need a pre-defined marker to be placed next to the plate, to deal with the problem of automatic calibration of the scene without user intervention.

Discussion

Prediction systems on food images are of large interests among multidisciplinary communities. Indeed, in the last years several works on food computing applied to different applications and research fields. Due to its interdisciplinary nature, food computing can be applied to develop several applications and services in various fields such as health, culture, agriculture, medicine, and biology. In this paper, the food estimation techniques are studied in the specific context of health monitoring. For a broader knowledge of the applications of food computing on several fields, an extensive

Table 2. Summary of the revised methods for food intake monitoring.

Reference	Year	Authors	Main Task	Automatic	Wearable
Kitamura *et al.*, 2008	2008	Kitamura, K., Yamasaki, T., Aizawa, K.	Food balance estimation.	Yes	No
Aizawa *et al.*, 2010	2010	Aizawa, K., De Silva, G.C., Ogawa, M., Sato, Y.	Food balance estimation.	Yes	No
Kitamura *et al.*, 2009	2009	Kitamura, K., Yamasaki, T., Aizawa, K.	Food balance estimation.	Yes	No
Kitamura, De Silva *et al.*, 2010	2010	Kitamura, K., De Silva, C., Yamasaki, T., Aizawa, K.	Food balance estimation.	Yes	No
Maruyama *et al.*, 2010	2010	Maruyama, Y., De Silva, G.C., Yamasaki, T., Aizawa, K.	Food balance estimation.	Yes	No
Shroff *et al.*, 2008	2008	Shroff, G., Smailagic, A., & Siewiorek, D. P.	Food balance estimation.	No	Yes
Puri *et al.*, 2009	2009	Puri, M., Zhu, Z., Yu, Q., Divakaran, A., Sawhney, H.	Food classification.	No	Yes
Dehais *et al.*, 2013	2013	Dehais, J., Shevchik, S., Diem, P., Mougiakakou, S.G.	Volume estimation.	Yes	Yes
Allegra *et al.*, 2017	2017	Allegra, D., Anthimopoulos, M., Dehais, J., Lu, Y., Stanco, F., Farinella, G.M., Mougiakakou, S.	Volume estimation.	Yes	No
Lu *et al.*, 2018	2018	Lu, Y., Allegra, D., Anthimopoulos, M., Stanco, F., Farinella, G.M., Mougiakakou, S	Volume estimation.	Yes	No
Allegra *et al.*, 2019	2019	Allegra, D., Erba, D., Farinella, G.M., Grazioso, G., Maci, P.D., Stanco, F., Tomaselli, V.	Volume estimation.	Yes	No
Chen *et al.*, 2010	2010	Chen, N., Lee, Y. Y., Rabb, M., & Schatz, B	Food classification and calories estimation.	Yes	Yes
Liu *et al.*, 2012	2012	Liu, J., Johns, E., Atallah, L., Pettitt, C., Lo, B., Frost, G., Yang, G.Z.	Food classification.	Yes	Yes
Kong *et al.*, 2012	2012	Kong, F., Tan, J.	Food classification.	Yes	Yes
Meyers *et al.*, 2015	2015	Meyers, A., Johnston, N., Rathod, V., Korattikara, A., Gorban, A., Silberman, N., Guadarrama, S., Papandreou, G., Huang, J., Murphy, K.P.	Food classification and calories estimation.	Yes	Yes
McAllister *et al.*, 2015	2015	McAllister, P., Zheng, H., Bond, R., Moorhead	Food classification and calories estimation.	No	Yes
Zhang *et al.*, 2015	2015	Zhang, W., Yu, Q., Siddiquie, B., Divakaran, A., Sawhney, H.	Food classification and calories estimation.	Yes	Yes
Zhu *et al.*, 2010	2010	Zhu, F., Bosch, M., Woo, I., Kim, S., Boushey, C.J., Ebert, D.S., Delp, E.J	Food classification and calories estimation.	Yes	Yes
Fang *et al.*, 2015	2015	Fang, S., Liu, C., Zhu, F., Delp, E.J., Boushey	Food classification and calories estimation.	Yes	Yes
Sun *et al.*, 2014	2014	Sun, M., Burke, L.E., Mao, Z.H., Chen, Y., Chen, H.C., Bai, Y., Li, Y.; Li, C., Jia, W	Food classification and calories estimation.	Yes	Yes
Akpro Hippocrate *et al.*, 2016	2016	Akpro Hippocrate, E.A., Suwa, H., Arakawa, Y., Yasumoto, K.	Food classification and calories estimation.	Yes	Yes
Hassannejad *et al.*, 2017	2017	Hassannejad, H., Matrella, G., Ciampolini, P., Munari, I., Mordonini, M., Cagnoni, S.	Food classification and calories estimation.	No	Yes
Lu *et al.*, 2019	2019	Lu, Y., Allegra, D., Anthimopoulos, M., Stanco, F., Farinella, G.M., Mougiakakou, S	Food classification and calories estimation.	Yes	Yes

review on this topic has been published in Min, Jiang, Liu, Rui, and Jain, (2019). Although such systems analyse heterogeneous data for different applications and purposes, all of them are based on Computer Vision and Machine Learning techniques.

The reviewed methods are based on technological advantages in Computer Vision and the availability of smartphone, which facilitate the acquisition of food images related to all the eating episodes. Several works on portion volume estimation or automatic food classification have shown the high potential of image-based methods toward the objective categorization and quantification of food.

However, image-based methods are still affected by some issues mainly related to the high variability on food types, shapes, and appearance of meals that is often composed by mixed foods. For these reasons, the type of food is difficult to be estimated from images in real-case scenarios. Moreover, foods with different nutritional content may appear similar, especially for beverage. The containers of food bring further variance on the visual aspect of food, and in some cases, they can also occlude part of foods.

In this context, only prototype systems have been developed so far. These systems often rely on a fiducial marker to be depicted in the image. Moreover, the food containers used in the acquired images is the same (*i.e.*, the same shape, dimension, colour, etc.). This represents a technological issue.

As consequence, only laboratory settings allowed high performances and robustness of the systems proposed so far. In other scenario related to food healthcare, Surface-enhanced Raman scattering (SERS) (Xu, Zhou, Takei, & Hong 2019) is employed to detect specific substances like invisible pesticide (Xu, Gao, Han, & Zhao 2017). Even though these kind of technologies can be employed in automatic food analysis systems, they are expensive and cannot be used by common users in real-life context. Therefore, in future developments, it is important to design systems able to be tested on real-life conditions, in which the system analyses a variety of food images presented under different settings through mobile consumer devices. An important application of food recognition techniques described in this paper is related to food logging for dietary management and food intake monitoring. Indeed, diet monitoring results crucial for human health and modern technologies must be improved for supporting it. Smartphone applications for assisted food diaries mainly focus on food recognition, whereas the aspect of user engagement results marginal. Indeed, in such applications the user is required to interact continuously to report information about eaten food (*e.g.*, food picture, meta-information), weight progress, etc. In this context, methods for increasing participation and stimulate the user interaction are needed.

Considering experiments on data collection using mobile technologies (Wenz, Jäckle, & Couper 2019), the best way to obtain the user participation and ensure the quality of responses is to maintain a high level of engagement. In the context of social media platform and related applications, the engagement of users is an imperative aspect. Indeed, in the last years the research field of sentiment analysis applied on multi-media contents shared through social media platforms registered a rapid increment in terms of applications, algorithms, and public large-scale datasets (Ortis *et al.*, 2020). The growth of social media platforms furthered the development of several applications that take advantage from the automatic analysis of images published by users through social media platform every day to infer what content will be shared/liked most by users (Ortis *et al.,* 2020) or what parts of the post most contributed to the virality of the content (Ortis, Farinella, & Battiato 2019). These methods can be further specialized on the single individual (*i.e.*, user profiling). These techniques are hence useful to both increase and objectively assess the level of users' engagement.

Conclusions

Food recognition for health applications is an innovative technology that, once reached satisfactory performances for such specific health applications, will be applied on the dietary and calorific monitoring. Moreover, eating diary based on automatic food recognition could support the traditional self-reporting approach (*i.e.*, written diary). Then, as the technology improves, more sophisticated applications could be implemented. Future technologies and development will need huge amounts of labelled images. Considering that the main issues are related to the availability of data and labels, more efforts should be devoted to the collection of learning datasets with high quality annotations related to the type of food, areas, quantities and calories of each food item present in an image. However, this is still an in-progress technology, and more efforts are required to meet high level standards for feasible medical protocols. So far, state of the art works focused on specific tasks performed in controlled environment, with limited variability. The main limitation, which makes this research challenging, is related to the extreme variability which food can present. Many ingredients are not visible after some kind of preparation and part of them are naturally invisible (*e.g.*, oil, fats).

This is critical for ingredients detection and, consequently, for nutritional values estimation. Also, automatically finding the food volume is a hard task, as most of consumers' devices mount an RGB camera only. This drives the development of complex methods for volume estimation, which require multiples picture to properly compute 3D measures by 2D pictures. Nevertheless, these methods work better in supervised scenario and could need patients' training, so they cannot be successfully applied in general cases.

References

Abdel-Hakim, A.E., & Farag, A.A. (2006). CSIFT: A SIFT descriptor with color invariant characteristics. *Computer Vision and Pattern Recognition, 2*, 1978–1983. doi:10.1109/CVPR.2006.95.

Agrawal, M., Konolige, K., & Blas, M.R. (2008). CenSurE: Center Surround Extremas for Realtime Feature Detection and Matching. *Lecture Notes in Computer Science, 5305*, 102–115. doi:10.1007/978-3-540-88693-8_8.

Ahonen, T., Hadid, A., & Pietikäinen, M. (2006). Face description with local binary patterns: application to face recognition. *Transactions on Pattern Analysis and Machine Intelligence, 28*, 2037–2041. doi:10.1109/TPAMI.2006.244.

Ahuja, J., Montville, J.B., Omolewa-Tomobi, G., Heendeniya, K.Y., Martin, C.L., Steinfeldt, L.C., Anand, J., Adler, M.E., LaComb, R.P., & Moshfegh, A.J. (2012). USDA food and nutrient database for dietary studies, 5.0–documentation and user guide. US Department of Agriculture, Agricultural Research Service, Food Surveys Research Group: Beltsville, MD, USA.

Aizawa, K., Silva, G.C., Ogawa, M., & Sato, Y. (2010). Food Log by snapping and processing images. *2010 16th International Conference on Virtual Systems and Multimedia,* 71-74. doi:10.1109/VSMM.2010.5665963.

Akpro Hippocrate, E. A., Suwa, H., Arakawa, Y., & Yasumoto, K.

(2016). Food weight estimation using smartphone and cutlery. In: *Proceedings of the First Workshop on IoT-enabled Healthcare and Wellness Technologies and Systems (IoT of Health '16).* Association for Computing Machinery, New York, NY, USA, 9–14.

Allegra, D., Anthimopoulos, M., Dehais, J., Lu, Y., Stanco, F., Farinella, G.M., & Mougiakakou, S. (2017). A multimedia database for automatic meal assessment systems. *Lecture Notes in Computer Science, 10590*, 471-478. doi:10.1007/978-3-319-70742-6_46.

Allegra, D., Erba, D., Farinella, G.M., Grazioso, G., Maci, P.D., Stanco, F., & Tomaselli, V. (2019). Learning to rank food images. *Lecture Notes in Computer Science, 11752*, pp. 629–639. doi:10.1007/978-3-030-30645-8_57.

Battiato, S., Farinella, G.M., Gallo, G., & Ravì, D. (2010). Exploiting textons distributions on spatial hierarchy for scene classification. *Journal on Image and Video Processing, 2010*, 919367. doi:10.1155/2010/919367.

Battiato, S., Farinella, G.M., Puglisi, G., & Ravì, D. (2014). Aligning codebooks for near duplicate image detection. *Multimedia Tools and Applications, 72*, 1483–1506. doi:10.1007/s11042-013-1470-4.

Bay, H., Tuytelaars, T., & Van Gool, L. (2006). SURF: Speeded Up Robust Features. *Lecture Notes in Computer Science, 3951*, 404–417. doi:10.1007/11744023_32.

Belongie, S., Alik, J., & Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *Transactions on Pattern Analysis and Machine Intelligence, 24*, 509–522. doi:10.1109/34.993558.

Bettadapura, V., Thomaz, E., Parnami, A., Abowd, G.D., & Essa, I. (2015). leveraging context to support automated food recognition in restaurants. *IEEE Winter Conference on Applications of Computer Vision*, 580–587. doi:10.1109/WACV.2015.83.

Bosch, M., Schap, T., Zhu, F., Khanna, N., Boushey, C.J., Delp, E.J. (2011). Integrated database system for mobile dietary assessment and analysis. *International Conference on Multimedia and Expo*, 1–6. doi:10.1109/ICME.2011.6012202.

Bosch, M., Zhu, F., Khanna, N., & Boushey, C.J., (2011). Delp, E.J. Combining global and local features for food identification in dietary assessment. *International Conference on Image Processing,* pp. 1789–1792. doi:10.1109/ICIP.2011.6115809.

Bossard, L., Guillaumin, M., & Van Gool, L. (2014). Food-101 – Mining discriminative components with random forests. *Lecture Notes in Computer Science, 8694*, 446–461. doi:10.1007/978-3-319-10599-4_29.

Brosnan, T., & Sun, D.W. (2004). Improving quality inspection of food products by computer vision - A review. *Journal of Food Engineering, 61*, 3–16. doi:10.1016/S0260-8774(03)00183-3.

Buemi, F., Massa, M., & Sandini, G. (1995). Agrobot: a robotic system for greenhouse operations. *Workshop on Robotics in Agriculture and the Food Industry*, pp. 172–184.

Burghouts, G.J., & Geusebroek, J.M. (2009). Performance evaluation of local colour invariants. *Computer Vision and Image Understanding, 113,* 48–62. doi:10.1016/j.cviu.2008.07.003.

Cardenas-Weber, M., Hetzroni, A., & Miles, G.E. (1991). Machine vision to locate melons and guide robotic harvesting. *American Society of Agricultural Engineers*, p. 21.

Chen, M., Dhingra, K., Wu, W., Yang, L., Sukthankar, R., & Yang, J. (2009). PFID: Pittsburgh fast-food image dataset. *International Conference on Image Processing*, pp. 289–292. doi:10.1109/ICIP.2009.5413511.

Chen, M.Y., Yang, Y.H., Ho, C.J., Wang, S.H., Liu, S.M., Chang, E., Yeh, C.H., & Ouhyoung, M. (2012). Automatic Chinese food identification and quantity estimation. *SIGGRAPH Asia 2012 Technical Briefs,* 1–4. doi:10.1145/2407746.2407775.

Chen, N., Lee, Y. Y., Rabb, M., & Schatz, B. (2010). Toward dietary assessment via mobile phone video cameras. *AMIA . Annual Symposium proceedings.* pp. 106–110.

Dalal, N., & Triggs, B. Histograms of oriented gradients for human detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 1*, pp. 886-893. doi:10.1109/CVPR.2005.177.

Dehais, J., Shevchik, S., Diem, P., & Mougiakakou, S.G. (2013). Food volume computation for self dietary assessment applications. *International Conference on Bioinformatics and Bioengineering.* doi:10.1109/BIBE.2013.6701615.

Delwiche, J.F. (2012). You eat with your eyes first. *Physiology & Behavior, 107*, 502–504. doi:10.1016/j.physbeh.2012.07.007.

Deng, Y., & Manjunath, B.S. (2001). Unsupervised segmentation of color-texture regions in images and video. *Transactions on Pattern Analysis and Machine Intelligence, 23*, 800–810. doi:10.1109/34.946985.

Donadello, I., Dragoni, M. (2019). Ontology-driven food category classification in images. *International Conference on Image Analysis and Processing*. pp. 607–617.

Du, C.J., & Sun, D.W. (2006). Learning techniques used in computer vision for food quality evaluation: a review. *Journal of Food Engineering, 72*, 39–55. doi:10.1016/j.jfoodeng.2004.11.017.

Du, C.J.; & Sun, D.W. (2008). Multi-classification of pizza using computer vision and support vector machine. *Journal of Food Engineering, 86*, 232–242. doi:10.1016/j.jfoodeng.2007.10.001.

Fang, S., Liu, C., Zhu, F., Delp, E.J., & Boushey, C.J. (2015). Single-view food portion estimation based on geometric models. *2015 IEEE International Symposium on Multimedia (ISM).* pp. 385–390.

Farinella, G.M., Allegra, D., Moltisanti, M., Stanco, F., & Battiato, S. (2016). Retrieval and classification of food images. *Computers in Biology and Medicine, 77*, 23–39. doi:10.1016/j.compbiomed.2016.07.006.

Farinella, G.M., Allegra, D., & Stanco, F. (2015). A benchmark dataset to study the representation of food images. *Lecture Notes in Computer Science, 8927*, pp. 584–599. doi:10.1007/978-3-319-16199-0_41.

Farinella, G.M., Allegra, D., Stanco, F., & Battiato, S. (2015). On the exploitation of one class classification to distinguish food vs non-food images. *Lecture Notes in Computer Science, 9281*, 375–383. doi:10.1007/978-3-319-23222-5_46.

Farinella, G.M., Moltisanti, M., & Battiato, S. (2014). classifying food images represented as bag of textons. *International Conference on Image Processing.* pp. 5212–5216. doi:10.1109/ICIP.2014.7026055.

Farinella, G.M., Moltisanti, M., & Battiato, S. (2015). Food recognition using consensus vocabularies. *Lecture Notes in Computer Science, 9281*, 384–392. doi:10.1007/978-3-319-23222-5_47.

Felzenszwalb, P.F., Girshick, R.B., Mcallester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *Transactions on Pattern Analysis and Machine Intelligence, 32*, 1627–1645. doi:10.1109/TPAMI.2009.167.

Fischler, M.A., & Bolles, R.C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM, 24*, 381–395. doi:10.1145/358669.358692.

Fontanellaz, M., Christodoulidis, S., & Mougiakakou, S. (2019). Self-attention and ingredient-attention based model for recipe retrieval from image queries. *International Workshop on Multimedia Assisted Dietary Management*. pp. 25–31. doi:10.1145/3347448.3357163.

Foroni, F., Pergola, G., Argiris, G., & Rumiati, R.I. (2013). The FoodCast research image database (FRIDa). *Frontiers in Human Neuroscience, 7*. doi:10.3389/fnhum.2013.00051.

Gunasekaran, S. (1996). Computer vision technology for food quality assurance. *Trends in Food Science & Technology, 7*, 245–256. doi:10.1016/0924-2244(96)10028-5.

Hammond, R.A., Levine, R. (2010). The economic impact of obesity in the United States. *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy, 3*, 285–295. doi:10.2147/DMSOTT.S738.

Hassannejad, H., Matrella, G., Ciampolini, P., Munari, I., Mordonini, M., & Cagnoni, S. (2017). A new approach to image-based estimation of food volume. *Algorithms, 10*, 66.

He, Y., Xu, C., Khanna, N., Boushey, C.J., & Delp, E.J. (2014). Analysis of food images: Features and classification. *International Conference on Image Processing*. pp. 2744–2748. doi:10.1109/ICIP.2014.7025555.

Herranz, L., Ruihan, X., & Shuqiang, J. (2015). A probabilistic model for food image recognition in restaurants. *International Conference on Multimedia and Expo*. pp. 1–6. doi:10.1109/ICME.2015.7177464.

Ho, T.K. (1995). Random decision forests. *International Conference on Document Analysis and Recognition, 1*, pp. 278–282. doi:10.1109/ICDAR.1995.598994.

Hoashi, H., Joutou, T., & Yanai, K. (2010). Image recognition of 85 food categories by feature fusion. *International Symposium on Multimedia,* pp. 296–301. doi:10.1109/ISM.2010.51.

Hu, Y., Cheng, X., Chia, L.T., Xie, X., Rajan, D., & Tan, A.H. (2009). Coherent phrase model for efficient image near-duplicate retrieval. *Transactions on Multimedia, 11*, 1434–1445. doi:10.1109/TMM.2009.2032676.

Kawano, Y., & Yanai, K. (2014). Automatic expansion of a food image dataset leveraging existing categories with domain adaptation. *Lecture Notes in Computer Science, 8927*, pp. 3–17. doi:10.1007/978-3-319-16199-0_1.

Kawano, Y., & Yanai, K. (2014). Food image recognition with deep convolutional features. *International Joint Conference on Pervasive and Ubiquitous Computing*. pp. 589–593. doi:10.1145/2638728.2641339.

Kawano, Y., & Yanai, K. (2013). Real-time mobile food recognition system. *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. pp. 1–7. doi:10.1109/CVPRW.2013.5.

Kiliç, K., Boyaci, I.H., Köksel, H., & Küsmenoglu, I. (2007). A classification system for beans using computer vision system and artificial neural networks. *Journal of Food Engineering, 78*, 897–904. doi:10.1016/j.jfoodeng.2005.11.030.

Killgore, W.D., & Yurgelun-Todd, D.A. (2005). Body mass predicts orbitofrontal activity during visual presentations of high-calorie foods. *Neuroreport, 16,* 859–863. doi:10.1097/00001756-200505310-00016.

Kitamura, K., De Silva, C., Yamasaki, T., & Aizawa, K. (2010). Image processing based approach to food balance analysis for personal food logging. *International Conference on Multimedia and Expo.* pp. 625–630. doi:10.1109/ICME.2010.5583021.

Kitamura, K., Yamasaki, T., & Aizawa, K. (2008). Food log by analyzing food images. *International Conference on Multimedia*. pp. 999–1000. doi:10.1145/1459359.1459548.

Kitamura, K., Yamasaki, T., Aizawa, K. (2009). FoodLog: capture, analysis and retrieval of personal food images via web. *Workshop on Multimedia for cooking and eating activities*. pp. 23–30. doi:10.1145/1630995.1631001.

Kohonen, T. (1998). The self-organizing map. *Neurocomputing, 21*, 1–6. doi:10.1016/S0925-2312(98)00030-7.

Kong, F., & Tan, J. (2012). DietCam: Automatic dietary assessment with mobile camera phones. *Pervasive and Mobile Computing, 8*, 147–163.

Krizhevsky, A., Sutskever, I., & Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks. *Neural Information Processing Systems.* pp. 1097–1105.

Lazebnik, S., Schmid, C., & Ponce, J. (2005). A sparse texture representation using local affine regions. *Transactions on Pattern Analysis and Machine Intelligence, 27*, 1265–1278. doi:10.1109/TPAMI.2005.151.

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. *Conference onComputer Vision and Pattern Recognition, 2*, pp. 2169–2178. doi:10.1109/CVPR.2006.68.

Levi, P., Falla, A., & Pappalardo, R. (1988). Image controlled robotics applied to citrus fruit harvesting. *International Conference on Robot Vision and Sensory Controls*. pp. 2–4.

Liu, J., Johns, E., Atallah, L., Pettitt, C., Lo, B., Frost, G., & Yang, G.Z. (2012). An intelligent food-intake monitoring system using wearable sensors. *International Conference on Wearable and Implantable Body Sensor Networks.* doi:10.1109/BSN.2012.11.

Lowe, D.G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60*, 91–110. doi:10.1023/B:VISI.0000029664.99615.94.

Lu, Y., Allegra, D., Anthimopoulos, M., Stanco, F., Farinella, G.M., & Mougiakakou, S. (2018). A Multi-task learning approach for meal assessment. *Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management*. pp. 46–52. doi:10.1145/3230519.3230593.

Lu, Y., Stathopoulou, T., Vasiloglou, M.F., Christodoulidis, S., Blum, B., Walser, T., Meier, V., Stanga, Z., Mougiakakou, S. (2019). an artificial intelligence-based system for nutrient intake assessment of hospitalised patients. *International Conference of the IEEE Engineering in Medicine and Biology Society.* pp. 5696–5699. doi:10.1109/EMBC.2019.8856889.

Marĉelja, S. (1980). Mathematical description of the responses of simple cortical cells. *Journal of the Optical Society of America, 70*, 1297–1300. doi:10.1364/JOSA.70.001297.

Marin, J., Biswas, A., Ofli, F., Hynes, N., Salvador, A., Aytar, Y., Weber, I., Torralba, A. (2019). Recipe1M+: A dataset for learning cross-modal embeddings for cooking recipes and food Images. *IEEE transactions on pattern analysis and machine intelligence*.

Maruyama, Y., De Silva, G.C., Yamasaki, T., & Aizawa, K. (2010). Personalization of food image analysis. *International Conference on Virtual Systems and Multimedia*. pp. 75–78. doi:10.1109/VSMM.2010.5665964.

Matas, J., Chum, O., Urban, M., & Pajdla, T. (2004). Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing, 22*, 761–767. doi:10.1016/j.imavis.2004.02.006.

Matsuda, Y., Hoashi, H., & Yanai, K. (2012). Multiple-food recognition considering co-occurrence employing manifold ranking. *International Conference on Pattern Recognition*. pp. 2017–

2020.

McAllister, P., Zheng, H., Bond, R., & Moorhead, A. (2015). Semi-automated system for predicting calories in photographs of meals. *2015 IEEE International Conference on Engineering, Technology and Innovation/International Technology Management Conference (ICE/ITMC).* pp. 1–6.

McCrickerd, K., & Forde, C.G. (2016). Sensory influences on food intake control: moving beyond palatability. *Obesity Reviews, 17*, 18–29. doi:10.1111/obr.12340.

Medic, N., Ziauddeen, H., Forwood, S.E., Davies, K.M., Ahern, A.L., Jebb, S.A., Marteau, T.M., & Fletcher, P.C. (2016). The presence of real food usurps hypothetical health value judgment in overweight people. *eNeuro, 3*, 0025–16. doi:10.1523/ENEURO.0025-16.2016.

Meyers, A., Johnston, N., Rathod, V., Korattikara, A., Gorban, A., Silberman, N., Guadarrama, S., Papandreou, G., Huang, J., & Murphy, K.P. (2015). Im2Calories: towards an automated mobile vision food diary. Available from: https://static.googleusercontent.com/media/research.google.com/it//pubs/archive/44321.pdf.

Min, W., Jiang, S., Liu, L., Rui, Y., & Jain, R. (2019). A survey on food computing. *ACM Computing Surveys (CSUR), 52*, 92.

Munkevik, P., Duckett, T., & Hall, G. (2007). A computer vision system for appearance-based descriptive sensory evaluation of meals. *Journal of Food Engineering, 78*, 246–256. doi:10.1016/j.jfoodeng.2005.09.033.

Munkevik, P., Duckett, T., Hall, G. (2004). Vision system learning for ready meal characterisation. *International Conference on Engineering and Food.*

Nguyen, D.T., Zong, Z., Ogunbona, P., Li, W. (2010). Object detection using non-redundant local binary patterns. *International Conference on Image Processing*. pp. 4609–4612. doi:10.1109/ICIP.2010.5651633.

Nishida, C., Uauy, R., Kumanyika, S., & Shetty, P. (2004). The joint WHO/FAO expert consultation on diet, nutrition and the prevention of chronic diseases: process, product and policy implications. *Public health nutrition, 7*(1a), 245-250.

Nistér, D., & Stewénius, H. Scalable recognition with a vocabulary tree. (2006). *Computer Vision and Pattern Recognition, 2*, 2161–2168. doi:10.1109/CVPR.2006.264.

Ortis, A., Farinella, G. M., & Battiato, S. (2020). Survey on visual sentiment analysis. *IET Image Processing, 14*(8), 1440-1456.

Ortis, A., Farinella, G.M., & Battiato, S. (2019). Predicting social image popularity dynamics at time zero. *IEEE Access*, 1–1. doi:10.1109/ACCESS.2019.2953856.

Ortis, A., Farinella, G.M., Torrisi, G., & Battiato, S. (2020). Exploiting objective text description of images for visual sentiment analysis. *Multimedia Tools and Applications*. doi:10.1007/s11042-019-08312-7.

Parrish, E.A., & Goksel, K.A. (1977). Pictorial pattern recognition applied to fruit harvesting. *Transactions of the American Society of Agricultural and Biological Engineers, 20*, 822–827. doi:10.13031/2013.35657.

Perronnin, F., & Dance, C. (2007). Fisher Kernels on visual vocabularies for image categorization. *Computer Vision and Pattern Recognition,* pp. 1–8. doi:10.1109/CVPR.2007.383266.

Perronnin, F., Sánchez, J., & Mensink, T. (2010). Improving the Fisher Kernel for large-scale image classification. *Lecture Notes in Computer Science, 6314*, pp. 143–156. doi:10.1007/978-3-642-15561-1_11.

Petit, O., Cheok, A.D., Oullier, O. (2016). Can food porn make us slim? how brains of consumers react to food in digital environments. *Integrative Food, Nutrition and Metabolism, 3*, 251–255. doi:10.15761/IFNM.1000138.

Pham, C., Jackson, D., Schöning, J., Bartindale, T., Plotz, T., & Olivier, P. (2013). FoodBoard: surface contact imaging for food recognition. *International Joint Conference on Pervasive and Ubiquitous Computing*. pp. 749–752. doi:10.1145/2493432.2493522.

Plebe, A., & Grasso, G. (2019). The unbearable shallow understanding of deep learning. *Minds and Machines, 29*(4), 515-553.

Pouladzadeh, P., Yassine, A., & Shirmohammadi, S. (2015). FooDD: Food Detection Dataset for Calorie Measurement Using Food Images. *Lecture Notes in Computer Science, 9281,* 441–448. doi:10.1007/978-3-319-23222-5_54.

Puri, M., Zhu, Z., Yu, Q., Divakaran, A., & Sawhney, H. (2009). Recognition and volume estimation of food intake using a mobile device. *Workshop on Applications of Computer Vision*. doi:10.1109/WACV.2009.5403087.

Qi, X., Xiao, R., Li, C.G., Qiao, Y., Guo, J., & Tang, X. (2014). Pairwise rotation invariant co-occurrence local binary pattern. *Transactions on Pattern Analysis and Machine Intelligence, 36*, 2199–2213. doi:10.1109/TPAMI.2014.2316826.

Ragusa, F., Tomaselli, V., Furnari, A., Battiato, S., & Farinella, G.M. (2016). Food vs non-food classification. *International Workshop on Multimedia Assisted Dietary Management*. pp. 77–81. doi:10.1145/2986035.2986041.

Rahmana, M.H., Pickering, M.R., Kerr, D., Boushey, C.J., & Delp, E.J. (2012). a new texture feature for improved food recognition accuracy in a mobile phone based dietary assessment system. *International Conference on Multimedia and Expo Workshops*. pp. 418–423. doi:10.1109/ICMEW.2012.79.

Ravì, D., Lo, B., & Yang, G.Z. (2015). Real-time food intake classification and energy expenditure estimation on a mobile device. *International Conference on Wearable and Implantable Body Sensor Networks*. doi:10.1109/BSN.2015.7299410.

Rich, A.J. (1981). A programmable calculator system for the estimation of nutritional intake of hospital patients. *The American Journal of Clinical Nutrition, 34*, 2276–2279.

Rosenbaum, M., Sy, M., Pavlovich, K., Leibel, R.L., & Hirsch, J. (2008). Leptin reverses weight loss–induced changes in regional neural activity responses to visual food stimuli. *The Journal of Clinical Investigation, 118*, 2583–2591. doi:10.1172/JCI35055.

Ruihan, X., Herranz, L., Shuqiang, J., Shuang, W., Xinhang, S., & Jain, R. (2015). Geolocalized modeling for dish recognition. *Transactions on Multimedia, 17*, 1187–1199. doi:10.1109/TMM.2015.2438717.

Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.* Addison-Wesley Longman Publishing Co., Inc.

Salvador, A., Hynes, N., Aytar, Y., Marin, J., Ofli, F., Weber, I., & Torralba, A. (2017). Learning cross-modal embeddings for cooking recipes and food images. *Proceedings of the IEEE conference on computer vision and pattern recognition.* pp. 3020–3028.

Sánchez, J., Perronnin, F., Mensink, T., & Verbeek, J. (2013). Image classification with the fisher vector: theory and practice. *International Journal of Computer Vision*, *105*, 222–245. doi:10.1007/s11263-013-0636-x.

Shotton, J., Johnson, M., Cipolla, R. (2013). Semantic texton forests for image categorization and segmentation. *Conference in Advances in Computer Vision and Pattern Recognition.* pp. 211–227. doi:10.1007/978-1-4471-4929-3_15.

Shroff, G., Smailagic, A., & Siewiorek, D. P. (2008). Wearable context-aware food recognition for calorie monitoring. 12th IEEE International Symposium on Wearable Computers. pp. 119-120. IEEE.

Slaughter, D.C, & Harrell, R.C. (1989). Discriminating fruit for robotic harvest using color in natural outdoor scenes. *Transactions of the American Society of Agricultural and Biological Engineers, 32*, 757–763. doi:10.13031/2013.31066.

Sun, D.W. (2000). Inspecting pizza topping percentage and distribution by a computer vision method. *Journal of Food Engineering, 44*, 245–249. doi:10.1016/S0260-8774(00)00024-8.

Sun, M., Burke, L.E., Mao, Z.H., Chen, Y., Chen, H.C., Bai, Y., Li, Y., Li, C., & Jia, W. (2014). eButton: a wearable computer for health monitoring and personal assistance. *Proceedings of the 51st Annual Design Automation Conference. ACM.* pp. 1–6.

Suthumchai, N., Thongsukh, S., Yusuksataporn, P., & Tangsripairoj, S. (2016). FoodForCare: An Android application for self-care with healthy food. *International Student Project Conference (ICT-ISPC).* pp. 89–92. doi:10.1109/ICT-ISPC.2016.7519243.

Tola, E., Lepetit, V., & Fua, P. (2009). DAISY: An efficient dense descriptor applied to wide-baseline stereo. *Transactions on Pattern Analysis and Machine Intelligence, 32*, 815–830. doi:10.1109/TPAMI.2009.77.

Topchy, A., Jain, A.K., & Punch, W. (2005). Clustering ensembles: models of consensus and weak partitions. *Transactions on Pattern Analysis and Machine Intelligence, 27*, 1866–1881. doi:10.1109/TPAMI.2005.237.

Varma, M., & Ray, D. (2007). Learning The Discriminative Power-Invariance Trade-Off. *International Conference on Computer Vision.* pp. 1–8. doi:10.1109/ICCV.2007.4408875.

Varma, M., & Zisserman, A. (2005). A Statistical Approach to Texture Classification from Single Images. *International Journal of Computer Vision, 62*, 61–81. doi:10.1023/B:VISI.0000046589.39864.ee.

Vedaldi, A., & Zisserman, A. (2012). Efficient additive kernels via explicit feature maps. *Transactions on Pattern Analysis and Machine Intelligence, 34*, 480–492. doi:10.1109/TPAMI.2011.153.

Wenz, A., Jäckle, A., & Couper, M.P. (2019). Willingness to use mobile technologies for data collection in a probability household panel. *Survey Research Methods, 13*, 1–22. doi:10.18148/srm/2019.v13i1.7298.

Wright, P.D., Shearing, G., Rich, A.J., & Johnston, I. (1978) The role of a computer in the management of clinical parenteral nutrition. *Journal of Parenteral and Enteral Nutrition, 2*, 652–657. doi:10.1177/014860717800200506.

Wu, W., & Yang, J. (2009). Food recognition using statistics of pairwise local features. *International Conference on Multimedia and Expo.* pp. 1210–1213. doi:10.1109/ICME.2009.5202718.

Xin, W., Kumar, D., Thome, N., Cord, M., & Precioso, F. (2015). Recipe recognition with large multimodal food dataset. *International Conference on Multimedia Expo Workshops*. pp. 1–6. doi:10.1109/ICMEW.2015.7169757.

Xu, K., Zhou, R., Takei, K., & Hong, M. (2019). Toward Flexible Surface-Enhanced Raman Scattering (SERS) Sensors for Point-of-Care Diagnostics. *Advanced Science, 6*. doi:10.1002/advs.201900925.

Xu, M.L., Gao, Y., Han, X.X. & Zhao, B. (2017). detection of pesticide residues in food using surface-enhanced raman spectroscopy: a review. *Journal of Agricultural and Food Chemistry, 65*, 6719–6726. doi:10.1021/acs.jafc.7b02504.

Yanai, K., & Joutou, T. (2009). SURF: Speeded Up Robust Features. *International Conference on Image Processing*. pp. 285–288. doi:10.1109/ICIP.2009.5413400.

Yanai, K., & Kawano, Y. (2015). Food image recognition using deep convolutional network with pre-training and fine-tuning. *International Conference on Multimedia & Expo Workshops*. pp. 1–6. doi:10.1109/ICMEW.2015.7169816.

Yang, S., Chen, M., Pomerleau, D., & Sukthankar, R. (2010). Food recognition using statistics of pairwise local features. *Conference on Computer Vision and Pattern Recognition*. pp. 2249–2256. doi:10.1109/CVPR.2010.5539907.

Zhang, W., Yu, Q., Siddiquie, B., Divakaran, A., & Sawhney, H. (2015). "Snap-n-Eat" food recognition and nutrition estimation on a smartphone. *Journal of diabetes science and technology, 9*, 525–533.

Zhu, F., Bosch, M., Woo, I., Kim, S., Boushey, C.J., Ebert, D.S., & Delp, E.J. (2010). The use of mobile devices in aiding dietary assessment and evaluation. *European Journal of Clinical Nutrition, 4*, 756–766. doi:10.1109/JSTSP.2010.2051471.

Zong, Z., Nguyen, D.T., Ogunbona, P., & Li, W. (2010). On the combination of local texture and global structure for food classification. *International Symposium on Multimedia.* pp. 204–211. doi:10.1109/ISM.2010.37.