Sim-food
GIORNATA di STUDI MIMOS
sulla SIMULAZIONE nel FOOD & BEVERAGE
22 aprile 2015 Parma
Centro Congressi S. Elisabetta - Campus Universitario
a cura di: MIMOS UNIVERSITÀ DEGLI STUDI DI PARMA DASSAULT SYSTEMES
con il patrocinio di: EXPO MILANO 2015

Food Understanding from Digital Images

Giovanni Maria Farinella, Dario Allegra, Marco Moltisanti, Filippo Stanco, Sebastiano Battiato
{gfarinella, allegra, moltisanti, fstanco, battiato}@dmi.unict.it

Image Processing Laboratory
Department of Mathematics and Computer Science
University of Catania
Viale A. Doria 6, 95125, Catania, Italy

Abstract — There is a general consensus on the fact that people love food. Due to the great diffusion of low cost image acquisition devices (e.g., smartphones and wearable cameras), the food is nowadays one of the most photographed objects. The number of food images on the web is increasing and novel social networks for food lovers (e.g., foodspotting) are more and more popular. The automatic analysis and classification of food items from images can provide a wider comprehension of the relationship between people and their meals (e.g., preferences of subjects with respect to the dishes of a fast food) and can be useful to build diet monitoring systems (e.g., to combat obesity), as well as recommendation systems for food advertising purposes. On the other hand, automatic food understanding from images is a challenging and exciting task since the high variability of the content depicted in images. In this paper, we address the problem of understanding food from images by proposing a framework for automatic food classification.

1 Introduction and Motivation

Understanding food in everyday life (e.g., the recognition of dishes and the related ingredients, the estimation of quantity, etc.) is a problem which has been considered in different research areas due its important impact under the medical, social and anthropological aspects. For instance, an insane diet can cause problems in the general health of the people. Since health is strictly linked to the diet, advanced computer vision tools to recognize food images (e.g., acquired with mobile/wearable cameras), as well as their properties (e.g., calories, volume), can help the diet monitoring by providing useful information to the experts (e.g., nutritionists) to assess the food intake of patients (e.g., to combat obesity). On the other hand, the great diffusion of low cost image acquisition devices embedded in smartphones allows people to take pictures of food and share them on internet (e.g., on social media); the automatic analysis of the posted images could provide information on the relationship between people and their meals and can be exploited by food retailer to better understand the preferences of a person for further recommendations of food and related products.

Food understanding from images is a challenging computer vision task since the food is intrinsically deformable and presents high variability in appearance. Image representation plays a fundamental role [1] for the automatic understanding of a food image. Moreover, benchmark datasets are needed [2] to properly study the peculiarities of the image representation in this context. This paper presents a computational framework for automatic classification of food images. The proposed method have been compared with respect to other approaches on two benchmark datasets. The results confirm the effectiveness of the proposed approach which outperforms previous methods.

2 Image Representation Model

The Bag-of-Visual-Word model (BoW) is one of the most popular method to represent images for classification purpose. This model has been recently applied in the context of food understanding by obtaining good performances with respect to other state-of-the-art models for the problem of food classification [1][2]. The main steps of the BoW model are the following: i) a sparse or dense keypoint detector used to collect local features on the images; ii) a description of the detected local features; iii) a quantization of the feature space; iv) the final representation of the image as collection of local features by taking into account the quantized feature space. In the proposed framework we employ a dense pixel-wise sampling by representing the local patch surrounding each pixel as a vector of responses to a bank of filters [1][2]. The feature space of the filters responses is then quantised with a k-means clustering. The final representation of a food image is obtained as a normalised histograms on the quantised feature space. This representation model is called Bag-of-Textons.

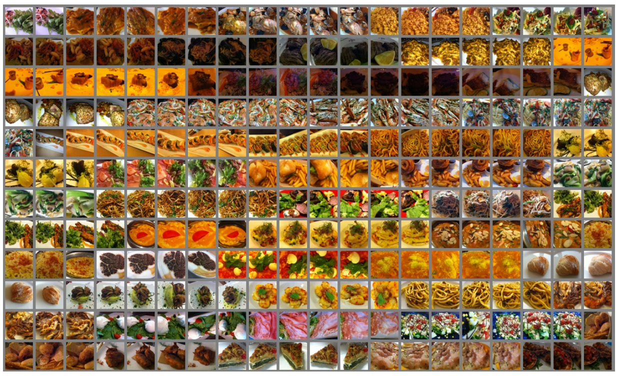

Fig. 1 - Examples of dishes of the UNICT-FD889 Dataset [2].

Fig. 2 - Examples of dishes of the PFID Dataset [3].

3 Food Datasets

To assess the classification performances of the described image representation method we used two different dataset in our experiments: the UNICT-FD889 [2] and the PFID [3] datasets. The UNICT-FD889 is a food dataset composed by 889 distinct plates of food[a]. Each dish has been acquired with iPhone devices multiple times to introduce photometric (e.g., flash vs no flash) and geometric variability (rotation, scale, point of view changes). The overall dataset contains 3583 images (Fig. 1). The dataset is designed to push research on automatic food understanding with the aim of finding a good way to represent food images for recognition purposes. The PFID dataset (Fig. 2) is composed by 1098 food images belonging to 7 major food categories: Sandwich, Salad & Sides, Bagel, Donut, Chicken, Taco, Bread & Pastry. Each food plate is present with 3 different instances (i.e., same food plate but acquired in different days and restaurants), and 6 images of different viewpoints.

4 Food Classification Results

We have compared the representation described in Section 2 with respect to other state of the art approaches [1][2][3][4]. To properly compare the different methods, the experiments have been repeated three times. At each run the different approaches are used on the same training and test sets. Training images have been used to perform the quantization of the feature space and as food image models to be recognized, whereas test images have been used to evaluate the classification performances. The final results have been obtained by averaging over the three executions. When image representations are obtained, a classifier can be used to recognize the class of the food images. In our experiments we have employed the K-Nearest Neighbours (KNN) and the Support Vector Machine (SVM) for classification purpose [1][2]. More specifically, we have used the UNICT-FD889 Dataset with the described image representation and the 1NN (with a χ^2 distance) to classify every test image with the same class of the closest training image. The SVM classifier has been used to test the classification performances on the PFID dataset. In Fig. 3 and Fig. 4 are reported the results obtained on the two considered datasets of food. The proposed Bag of Textons representation obtains the best results in discriminating the different classes of food from images.

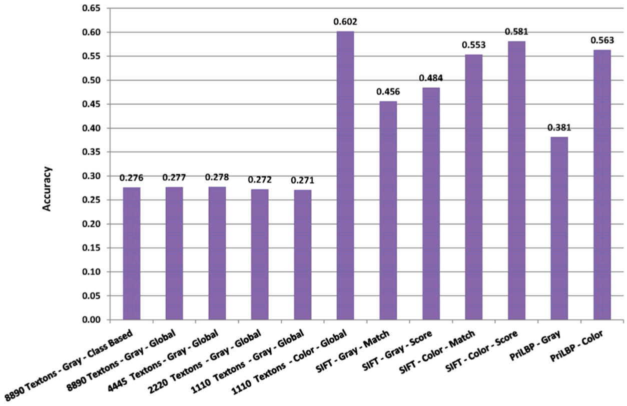

Fig. 3 – Classification results on the UNICT-FD889 Dataset [2].

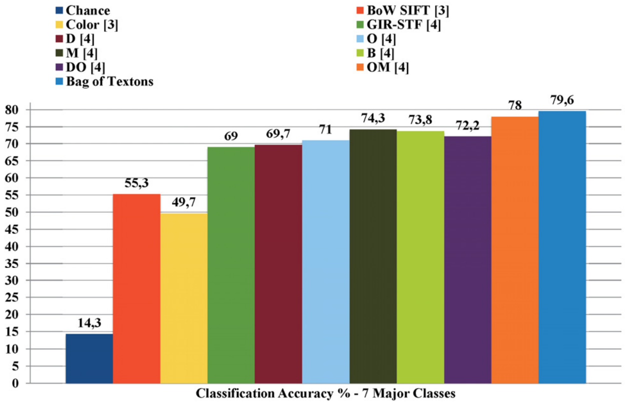

Fig. 4 – Classification results on the PFID Dataset [1].

References

[1] G. M. Farinella, M. Moltisanti, S. Battiato, Classifying Food Images Represented as Bag of Textons, IEEE International Conference on Image Processing, 2014

[2] G. M. Farinella, D. Allegra, F. Stanco, A Benchmark Dataset to Study the Representation of Food Images, International Workshop on Assistive Computer Vision and Robotics (ACVR), 2014

[3] M. Chen, K. Dhingra, W. Wu, L. Yang, R. Sukthankar, J. Yang, Pfid: Pittsburgh fast-food image dataset, IEEE International Conference Image Processing, 2009

[4] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar, Food recognition using statistics of pairwise local features, IEEE Conference on Computer Vision and Pattern Recognition, 2010

[a] The UNICT-FD889 is publicly available at the following URL: http://iplab.dmi.unict.it/UNICT-FD889 .