A Benchmark Dataset to Study the Representation of Food Images

Giovanni Maria Farinella, Dario Allegra, Filippo Stanco
{gfarinella, allegra, fstanco}@dmi.unict.it, Image Processing Laboratory, University of Catania

IMAGE PROCESSING LABORATORY

Abstract

Food is an essential component of human life and it is well-known that people love food. Nevertheless, an insane diet can cause problems in the general health of the people. Automatic recognition of food images (e.g., acquired with mobile/wearable cameras) has a key role in building monitoring systems to assess the daily food intake. A food recognition system could replace the traditionally dietary assessment based on self-reporting in a food diary that is often inaccurate. It could be important when a patient (e.g., with obesity, diabetes, or food allergy) has to be assisted during his daily meals. Moreover, experts (e.g., nutritionists) could use the food intake monitoring system to study the daily diet of patients to better understand their habits and/or eating disorders.

However, food recognition is a challenging task since the food is intrinsically deformable and presents high variability in appearance. The image representation employed in a food recognition engine plays the most important role. To properly study the peculiarities of the image representation in the food application context, a benchmark dataset is needed.

We introduce a food dataset composed by 889 distinct plates of food of different nationalities (e.g., Italian, English, Thai, Indian, Japanese, etc.). Each dish has been acquired multiple times by users (with a smartphone) in real cases of meals and in unconstrained settings (e.g., background, light environment conditions, etc). The dataset presents both photometric (e.g., flash vs no flash) and geometric variabilities (rotation, scale, point of view changes). The dataset is designed to push research in this application domain with the aim of finding a good way to represent food images for recognition purposes.

The first question we try to answer is the following: are we able to perform a near duplicate image retrieval (NDIR) in case of food images?

The UNICT-FD889 dataset

The overall dataset contains 3583 images related to 889 distinct plates of food. In the image on the left are shown examples of 96 dishes of the proposed dataset. In the image on the right are shown three instances of each dish for 32 different plates of the UNICT-FD889 Dataset.

Representation of food images

To benchmark the proposed dataset for Near Duplicate Image Retrieval (NDIR) purpose, we explore three standard state-of-the-art image representations in our tests: Bag of Textons [1], PRICoLBP [2] and SIFT features [3]. We decided to use Bag of Textons model for its power in representing textures and because have been obtained the best results so far on the PFID dataset [4]. We tested both class-based and global-based Bag of Textons representation with different vocabulary sizes. PRICoLBP descriptor has been chosen since it encodes spatial co-occurrence of local LBP features which are useful in representing textures. Finally SIFT features have been considered due their good performances in the context of near duplicate image retrieval [5].

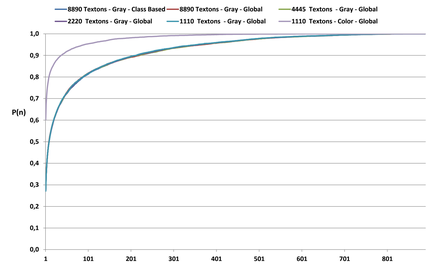
References

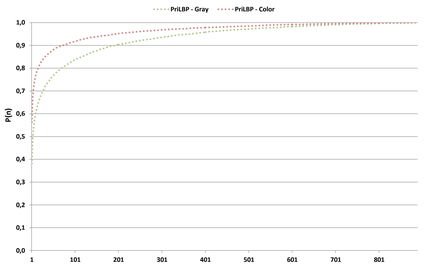
[1] Varma M. et al., A Statistical Approach to Texture Classification from Single Images, International Journal of Computer Vision, 62(1-2),61-81, 2005
[2] Qi X. et al., Pairwise rotation invariant co-occurrence local binary pattern, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014
[3] Lowe D.G., Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision, 60(2), 91-110, 2004
[4] Farinella G.M. et al., Classifying food images represented as bag of textons, IEEE International Conference on Image Processing, 2014
[5] Nistér D. et al. Scalable recognition with a vocabulary tree, IEEE Conference on Computer Vision and Pattern Recognition, 2006

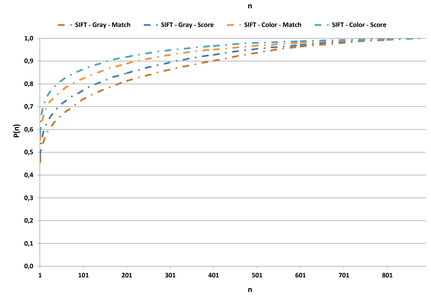
Experimental settings and results

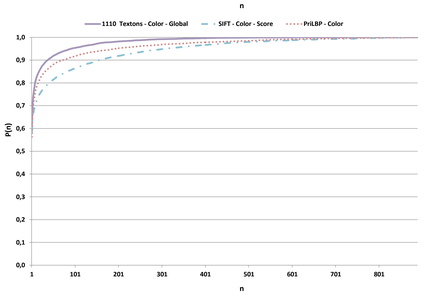
For testing purposes images have been resized to 320 x 240 pixels. We have employed the χ^2 distance to measure the similarity between two images represented as Bags of Textons [1,4] or PRICoLBP [2]. The similarity measure tested with SIFT [3] is based on the number of matchings. Moreover, the similarity measure in which the SIFT matchings are inversely weighted taking into account matching distances have been also taken into account. All the considered local descriptors are rotationally invariant. SIFT is also scale invariant. The representation have been considered in both grayscale and color domains.

To properly evaluate the different representation methods, the experiments have been repeated three times. At each run different approaches are executed on the same training and test sets. To this purpose, at each run we have built a training set composed by 889 images, by selecting one image of the UNICT-FD889 dataset per dish, whereas the rest of images have been used for testing purposes. The images considered for the three training sets are different. At each run, test images are used to perform queries on the corresponding training dataset used for that test. Given an image representation, the final results are obtained by averaging over the three tests.









The retrieval performances on each run have been evaluated with the probability of the successful retrieval $P(n)$ in a number of test queries:

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

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