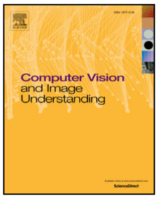
Computer Vision and Image Understanding 148 (2016) 67–86

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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



A supervised extreme learning committee for food recognition

Niki Martinel*, Claudio Piciarelli, Christian Micheloni

Department of Mathematics and Computer Science, University of Udine, Udine 33100, Italy

ARTICLE INFO

Article history:
Received 15 April 2015
Accepted 23 January 2016

Keywords:
Food recognition
Extreme Learning Machines
Structural SVM

ABSTRACT

Food recognition is an emerging topic in computer vision. The problem is being addressed especially in health-oriented systems where it is used as a support for food diary applications. The goal is to improve current food diaries, where the users have to manually insert their daily food intake, with an automatic recognition of the food type, quantity and consequent calories intake estimation. In addition to the classical recognition challenges, the food recognition problem is characterized by the absence of a rigid structure of the food and by large intra-class variations. To tackle such challenges, a food recognition system based on a committee classification is proposed. The aim is to provide a system capable of automatically choosing the optimal features for food recognition out of the existing plethora of available ones (e.g., color, texture, etc.). Following this idea, each committee member, i.e., an Extreme Learning Machine, is trained to specialize on a single feature type. Then, a Structural Support Vector Machine is exploited to produce the final ranking of possible matches by filtering out the irrelevant features and thus merging only the relevant ones. Experimental results show that the proposed system outperforms state-of-the-art works on four publicly available benchmark datasets.

1. Introduction

According to the World Health Organization, in the last years there has been a rapid increase of diseases related to excessive or wrong food intake, most notably obesity and derived issues such as diabetes, cardiovascular diseases, musculoskeletal disorders and some types of cancers. In particular, it is estimated that in 2014 about 39% of the world's adult population were overweight, including a 13% of obese people, whose number more than doubled between 1980 and 2014. Contrary to popular belief, the problem also affects many low- and middle-income countries, particularly in urban settings [1].

Despite obesity being a complex disease involving many factors, from genetics to life styles, proper actions against it necessarily include a strict control over the daily food intake. Obese people should constantly take note of their daily meals, both for self-monitoring and to acquire useful statistics for dietitians. This justifies the large amount of food diary applications for mobile devices that have recently been developed [2–4]. However, these apps typically require a manual annotation of the food intake, a tedious task that often discourages the potential users. To face this problem, many food recognition works have been recently proposed, whose aim is to automatically classify food (and possibly its amount) directly from smartphone-acquired pictures.

Apart from the main health-oriented task, food recognition techniques can be applied in several other contexts as well. The recent rise in popularity of food-related TV shows, food blogs, etc. has lead to the production and sharing of a large amount of food-based multimedia (a trend sometimes referred to as “food porn” [5]). This information deserves proper tools for automatic search and classification, e.g. for image retrieval, user profiling, targeted advertising applications and so on. For example, it has become quite common for people to share pictures of their own meals on social networks, either on generic-purpose networks such as Facebook or image-oriented ones, such as Instagram or Pinterest. Automatic food recognition could help to identify the personal tastes of the users, in order to deliver finely-tuned advertising such as the best restaurants in the nearby that match the user's taste.

Regardless of the specific application, automatic food recognition is a tough problem with many specific challenges. Differing from other common image classification tasks, in food recognition there is no spatial layout information to be exploited. While for example human body recognition can benefit from prior knowledge on the spatial relationships between the parts to be detected (e.g. the head being always over the torso [6–8]) this is rarely the case when considering food. More generally, food is typically non-rigid, and thus no structure information can be easily exploited. Intra-class variation is another source of uncertainty, since the recipe itself for the same food can vary depending on the location, the

* Corresponding author. Tel.: +39 0432 558423.
E-mail addresses: niki.martinel@uniud.it (N. Martinel), claudio.piciarelli@uniud.it (C. Piciarelli), christian.micheloni@uniud.it (C. Micheloni).

http://dx.doi.org/10.1016/j.cviu.2016.01.012
1077-3142/© 2016 Elsevier Inc. All rights reserved.

available ingredients and, last but not least, the personal taste of the cook. Finally, inter-class confusion is a source of potential problems too. Different foods may look very similar, as in many soups where the main ingredients may be hidden below the liquid level. On the other hand, food images often have distinctive properties, especially in terms of colors and textures, which humans are able to exploit to recognize foods even from a single example, thus the task is still tractable, despite the non-trivial challenges.

A possible solution to sidestep the aforementioned problems might be a system that uses as many different features as possible but exploits only a subset of those to perform the food classification task. Following this idea, a food classification system based on a supervised learning committee is introduced.

As demonstrated in [9–11], learning with a committee has two main benefits: (i) a committee might exhibit performance unobtainable by an individual committee member on its own. This is due to the fact that individual errors made by the committee members cancel out to some degree when their predictions are combined; (ii) a committee of learning machines has modularity properties. Since different members can focus on a particular region in the input space, the mapping from input to target is not approximated by one estimator but by several estimators. Despite these benefits, since many different possible visual features can be used to address the task, they cannot be just integrated in a single feature vector of very high dimensionality. Indeed, this might yield to intractable computational loads as well as to the curse of dimensionality problem. To address such issues, the Supervised Extreme Learning Committee (SELC) approach is introduced.

SELC relies on a committee of Extreme Learning Machines (ELM) [12], where each ELM is trained with a specific feature type only. In this way, each member specializes on classifying a food only by using a certain feature type. This has the advantage of both reducing computational loads as well as to keep the committee learning benefits. Among all the possible neural-based learning systems [13], Extreme Learning Machines have been chosen for their excellent performances in terms of computational burden while maintaining a classification accuracy comparable to similar learning tools.

Committee-based approaches require the selection of a supervisor to fuse the discordant members' classifications. The typical output of the supervisor is a class. However, when classification results must be presented to users, a rank could be more appropriate. While ranking information can be obtained from single committee members, none of the existing works have adopted a supervisor considering it. Motivated by this, we introduce a Structural Support Vector Machine [14] as supervisor. It automatically selects the ranking produced by the members and combines them to obtain optimal classification performance as well as an optimal ranking.

The rest of the papers is organized as follows: in Section 2 we review the main state-of-the-art works in food recognition. Section 3 describes the proposed approach, explaining how ELMs can be applied to food classification and how the committee outputs can be merged into a final rank by means of a Structured Support Vector Machine. In Section 4 we give some comparative experimental results, showing how our system performs with respect to state-of-the-art methods. Finally, conclusions are drawn in Section 5.

2. Related work

The topic of automatic food recognition has not been deeply investigated until recent years. The first works date back to late '90s, but they are limited to very specific contexts. For example the work by Jiménez et al. [15] focuses on automatic spherical fruit detection by means of a laser range-finder and image-based color and shape analysis to operate a robotic arm for fruit picking.

Most of the modern works on automatic classification of generic food images, typically for health-oriented applications, have been proposed since 2010. The work by Chen et al. [16] introduced a system exploiting different classifiers trained on multiple features. The authors compute both SIFT and LBP features with sparse coding for each image, as well as color histograms and Gabor filter responses to model the image colors and textures. A Support Vector Machine is trained for each texture separately, and the results are fused to form a single classifier using a multiclass AdaBoost algorithm. However, no details are given on the algorithm used for results fusion. A preliminary technique to estimate the amount of food using depth information computed via stereo matching techniques is also proposed. While being similar to our work, the results are fused to obtain optimal classification performance and not a "plausibility-rank" as in this work. Farinella et al. [17] exploit the texture information by applying a bank of rotation and scale invariant filters to each class of food images, in order to extract texture-oriented features known as Textons. The feature space is then quantized via K-means to create a codebook of textons for each class. All the textons prototypes are collected in a single visual dictionary which is used to represent each image as visual words distributions, effectively implementing a Bag of Textons approach. Finally, a Support Vector Machine is used in the classification stage. In [18], Yang et al. claim that spatial relationships between different ingredients could be exploited in the recognition of some types of food, as in a sandwich, where the meat is always between the bread slices. They perform a soft pixel-level segmentation of the image into eight ingredient types using a Semantic Texton Forest. Then, they compute pairwise statistics over the detected local ingredients, such as distance, orientation, etc. The statistics are accumulated in a multi-dimensional histogram, which is then used as the input feature vector for a χ^2 kernel Support Vector Machine. The algorithm has been evaluated on the PFID dataset (see Section 4). Bossard et al. [19] believe that local information is crucial in food recognition. They introduced a weakly-supervised mining method which relies on Random Forests to extract relevant image patches (components) that are typical of specific foods. Recognition is performed by scoring image superpixels according to their similarity with the mined components, with a final multi-class SVM-based classification step. Their work is also notable for introducing the *Food-101* dataset. Also the nowadays popular deep learning techniques have been applied to food recognition tasks. For example, Kagaya et al. [20] trained a Convolutional Neural Network on the food images acquired by the FoodLog web service. They tuned the CNN parameters such as kernel size, number of layers etc. to achieve experimental results that outperformed traditional techniques such as SVM-based classification. By analyzing the resulting convolution kernels, they also observed that color seems to be a predominant feature in the specific task of food recognition. The authors also used the same approach to train a food detector, although this required the nontrivial creation of a non-food training set.

Many works are explicitly tuned for food diary applications on smartphones and other mobile devices [21–23]. Kawano and Yanai [21], for instance, are particularly concerned with real-time performances on an Android-based smartphone. To speed up the process, the user is asked to manually select a proper bounding box delimiting the food to be recognized. The bounding box is then adjusted based on the segmentation result by the GrabCut algorithm. The user also receives hints on how to move the camera to better acquire the food pictures. Then, the system extracts both color histograms and SURF-based Bag of Features, and uses them to assign the acquired image to one of 15 possible classes using a Support Vector Machine. Kong et al. [22] have developed DietCam,