Pattern Recognition Letters 74 (2016) 17–23

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Towards generic image classification using tree-based learning: An extensive empirical study☆

Raphaël Marée*, Pierre Geurts, Louis Wehenkel

Systems and Modeling, Department of Electrical Engineering and Computer Science and GIGA-Research, Building B28, Quartier Polytech 1, 10 Allée de la découverte, Université de Liège, 4000 Liège, Belgium

ARTICLE INFO

Article history:
Received 7 September 2015
Available online 8 February 2016

Keywords:
Image classification
Decision trees
Extremely randomized trees
Random forests
Empirical evaluation

ABSTRACT

This paper considers the general problem of image classification without using any prior knowledge about image classes. We study variants of a method based on supervised learning whose common steps are the extraction of random subwindows described by raw pixel intensity values and the use of ensemble of extremely randomized trees to directly classify images or to learn image features. The influence of method parameters and variants is thoroughly evaluated so as to provide baselines and guidelines for future studies. Detailed results are provided on 80 publicly available datasets that depict very diverse types of images (more than 3800 image classes and over 1.5 million images).

1. Introduction

The aim of supervised image classification is to automatically build computerized models able to predict accurately the class (among predefined ones) of new images, once trained from a set of labeled images. In the real world, this generic problem encompasses well-known tasks such as the automatic recognition of images of handwritten characters, faces, cells, and road signs, to name but a few.

Since the early days of computer vision practice, when a researcher approaches a new image classification task, he or she often develops a dedicated algorithm to implement human prior knowledge as a sequence of specific operations, also known as a hand-crafted approach. Such an approach often involves the design and calculation of tailored filters and features capturing expected invariant image characteristics. In our preferred field of application, life science imaging, although several specific works have proved effective, the design choices are rarely straightforward hence such a strategy requires a lot of research and development efforts for each specific problem, and it might require major adjustments when parameters of the problem vary (e.g. sample preparation protocols, imaging modality, phenotypes to recognize, etc.). In other words, this engineering approach does not scale well as there are hundreds of thousands of biological entities that can be screened using many different sample preparation techniques and imaging modalities. Hence, scientific studies are often limited in scale, or still partially performed by hand (e.g. 50 millions of galaxies were manually labeled into morphological classes by almost 150, 000 humans within one year through the GalaxyZoo web-based project [13]), while others required very large computing infrastructures because they relied on dense feature computations (e.g. computers of the members of the Help Conquer Cancer project have contributed over 100CPU-millenia for the automated classification of tens of millions of protein crystallization-trial images at a rate of 55CPU-years per day [10]).

1.1. This work

Following and extending previous works [14–17], we consider the generic problem of supervised image classification without any preconception about image classes, i.e. it encompasses the recognition of numerous types of images under various image acquisition conditions. Indeed, with the design of a general-purpose yet simple and easily applicable image classifier in mind, we proposed earlier an appearance-based, learning method, relying on dense random subwindow extraction in images, their description by raw pixel values, and the use of ensembles of extremely randomized trees to classify these subwindows hence images. Despite its conceptual simplicity and its rather low run-time complexity, it yielded interesting results on a few datasets. Subsequently, variants of the method were proposed in [4,18,20,28] for object categorization, image segmentation, interest point detection, and content-based image retrieval.

In this paper, we extend and thoroughly evaluate our generic framework for image classification. Our contributions are as follows:

☆ This paper has been recommended for acceptance by Maria De Marsico.

* Corresponding author. Tel.: +32 4 366 2644; fax: +32 4 366 2988.
E-mail address: raphael.maree@ulg.ac.be (R. Marée).

http://dx.doi.org/10.1016/j.patrec.2016.01.006
0167-8655/

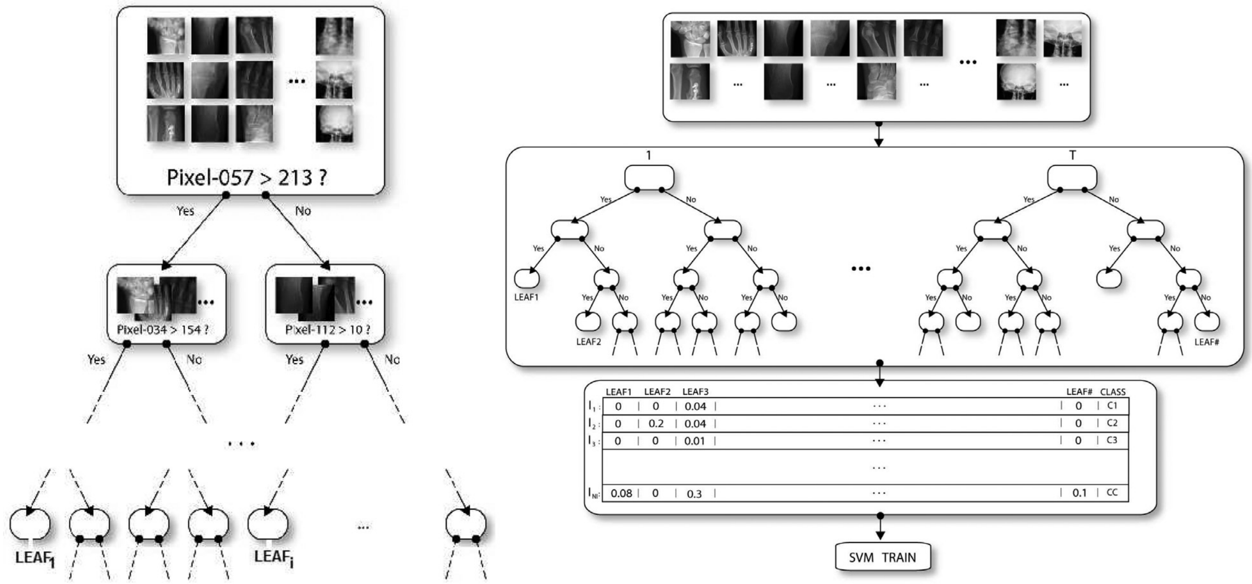

Fig. 1. Left: A single tree induced from a training set of random subwindows, using node tests with single pixel thresholding, for the ET-FL scheme. Right: An ensemble of T trees, the derived, quantitative frequency global representation for training images, and training of a final linear SVM classifier in ET-FL mode.

- While the main building blocks of the framework, subwindows extraction and extremely randomized trees, have been proposed in our earlier research, several algorithmic variants have not yet been considered and deserve to be tested. In particular, extending the work of [20], we explore in this paper several novel variants of the feature learning approach, corresponding to different ways to derive features from trees. We also consider yet unexplored parameter ranges (e.g., subwindow size intervals) and several simple pre-processing strategies (e.g., filters), which both turned out to be very beneficial on several datasets. These new algorithmic variants therefore greatly extend the range of image classification tasks that can be addressed by our framework and improve its generality.
- To assess our framework, we perform an extensive, systematic study of its performances on 80 publicly-available datasets (among which 25 bioimaging datasets). By conducting such a large-scale study, we are able to characterize the performances of the method and its recent variants, to study rigorously the influence of its parameters and classification schemes, to bring out the most influential design choices, and to draw general guidelines for future use so as to speed its application on new problems.
- To the best of our knowledge, no other image classification method has been evaluated so extensively. We deeply believe that generic methods can only be fully and fairly assessed by confronting them to several representative tasks and by extensively studying the influence of their parameters. By summarizing publicly available databases and by providing our positive and negative results, our hope is thus also to foster research in generic methods, by encouraging other researchers to evaluate and compare their methods on a wide range of imagery.

2. Experimental setup

We work with a large variety of datasets from many application domains. Our hypothesis is that by considering the image classification problem as a whole, it will possible to derive trends that are generally valuable, i.e. applicable in several areas. For example, observations derived from experiments related to the recognition of traffic signs (captured with onboard cameras) or galaxies (captured during wide-field sky surveys) might be helpful for the recognition of cells (captured by microscopes) as these datasets are sharing some essential characteristics (they consist in different classes of shapes and they exhibit illumination and noise variations due to the acquisition process). Similarly, observations derived from material classification datasets might be of interest for biological tissue recognition (as their images have textured patterns).

2.1. Datasets and evaluation criteria

Our experimental setup comprises 80 image datasets that were previously published and are publicly and freely available. They sum up roughly to 1.5 million images depicting approximately 3850 distinct classes. The choice of datasets was made a priori and independently of the results obtained with our method. More details about these datasets are given in Supplementary material. In particular, a summary of their characteristics is given in Supplementary Table I, and an overview of image classes for all datasets is given in Supplementary Figs. 1–4. Images were acquired worldwide, in controlled or uncontrolled conditions, using professional equipments in laboratory settings, individuals' digital camera in the real-world, various biomedical imaging equipments (fluorescence or brightfield microscopes, plain film radiography, etc.), robotic telescopes, synthetic aperture radars, etc. For a given dataset, image classes possibly exhibit subtle or prominent changes in their appearance due to various sources and levels of variations including possible changes in position, illumination, scale, and viewpoint, and/or presence of background clutter, occlusions, and noise. Moreover, either significant intra-class variations or high similarity between distinct classes could be present. Several of these datasets are synthetic and therefore variations are controlled (e.g. backgrounds are uniform) and well characterized, while many others contains real-world images so variations are mixed. Note that we only included in our experiments two widely used face datasets among tens of existing ones, given that face databases were recently summarized and evaluated thoroughly [8,23,27]. Also, we did not include the Pascal VOC challenge datasets [5] whose evaluation criteria (precision/recall curves for each object class) does not fit well into our evaluation framework (see below).

Our evaluation protocols are summarized in Supplementary Table I. Our evaluation metric is the misclassification error rate evaluated on independent test images. If a precise dataset protocol was