Computer Vision and Image Understanding 131 (2015) 56–71

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Semantically-driven automatic creation of training sets for object recognition

Dong Seon Cheng [a], Francesco Setti [b,*], Nicola Zeni [b], Roberta Ferrario [b], Marco Cristani [c]

[a] Hankuk University of Foreign Studies, Yongin Global Campus, 449-791, South Korea
[b] ISTC–CNR, via alla Cascata 56/C, I-38123 Povo, Trento, Italy
[c] Università degli Studi di Verona, Strada Le Grazie 15, I-37134 Verona, Italy

ARTICLE INFO

Article history:
Received 7 February 2014
Accepted 8 July 2014

Keywords:
Object recognition
Training dataset
Semantics
WordNet
Internet search

ABSTRACT

In the object recognition community, much effort has been spent on devising expressive object representations and powerful learning strategies for designing effective classifiers, capable of achieving high accuracy and generalization. In this scenario, the focus on the training sets has been historically weak; by and large, training sets have been generated with a substantial human intervention, requiring considerable time. In this paper, we present a strategy for automatic training set generation. The strategy uses semantic knowledge coming from WordNet, coupled with the statistical power provided by Google Ngram, to select a set of meaningful text strings related to the text class-label (*e.g.*, "cat"), that are subsequently fed into the Google Images search engine, producing sets of images with high training value. Focusing on the classes of different object recognition benchmarks (PASCAL VOC 2012, Caltech-256, ImageNet, GRAZ and OxfordPet), our approach collects novel training images, compared to the ones obtained by exploiting Google Images with the simple text class-label. In particular, we show that the gathered images are better able to capture the different visual facets of a concept, thus encoding in a more successful manner the intra-class variance. As a consequence, training standard classifiers with this data produces performances not too distant from those obtained from the classical hand-crafted training sets. In addition, our datasets generalize well and are stable, that is, they provide similar performances on diverse test datasets. This process does not require manual intervention and is completed in a few hours.

1. Introduction

Object recognition has been since its beginnings and still is one of the main and most studied topics in computer vision and its applications are many and varied, ranging from image indexing and retrieval, to video surveillance, robotics and medicine.

Even though at a first glance one may think that what is being recognized is an object that is given "out there in the world", at a closest look one may see that what is detected and then assigned to a certain class of objects is something that is constructed out from an aggregation of features that a classifier has been trained to recognize as that particular kind of object [1]. As a consequence, the fact that a certain aggregation of features is recognized as a dog or as a building, strongly depends on the images that have been chosen to be part of the training set [2].

Traditionally, classifiers have been and often are still trained with datasets that were created *ad hoc* by computer vision scientists, whose expertise drives the choice towards images with certain characteristics (being class-prototypical instances or making the recognition particularly challenging, see [3]); important examples are the Caltech-101/256 [4,5], MSRC [6], the PASCAL VOC series [7], LabelMe [8] and Lotus Hill [9]. Of course such choice is not arbitrary, but the criteria of choice are left implicit and so are the criteria of identity of the target object which is detected (or, better, constructed). As long as we are only concerned with object recognition tasks, probably this is not such a big issue, but when such tasks are part of more complex processes that include visual inference, this could constitute a drawback. Another relevant drawback is that building object recognition datasets is costly and thus the number of images that are collected is limited.

To overcome the disadvantage of having few training images per class and, in general, few object classes, in the last years projects have emerged, which exploit the so called "wisdom of crowd" to populate object recognition datasets, through *web-based* data collection methods. The idea is to employ web-based annotation tools that provide a way of building large annotated datasets by relying on the collaborative effort of a large population of users [10,11]. The outcome consists of millions of tagged images, but

* Corresponding author.

http://dx.doi.org/10.1016/j.cviu.2014.07.005
1077-3142/

Fig. 1. First 20 images obtained by searching "cat" in Google Images. The order (row-major) follows the ranking given by the search engine.

usually of these only few are accessible, and they are not organized into classes by a proper taxonomy.

Differently, one of the most important web-based projects which focuses on the concept of class is ImageNet [12]. ImageNet takes the tree-like structure in which words are arranged in WordNet [13] and assigns to each word (or, better, to each *synset* of WordNet) a set of images that are taken to be instantiations of the class corresponding to the synset. The candidate images to be assigned to a class are quality-controlled and human-annotated through the service of the Amazon Mechanical Turk (AMT), an online platform on which everyone can put up tasks for users, to be completed in order for them to get paid. Nowadays, ImageNet is the largest clean image dataset available to the vision research community, in terms of the total number of images, number of images per category, as well as the number of categories (80 K synsets).

Apart from these advantages, an important fact that should be discussed is *where the images come from*. In ImageNet, the source of data is Internet, so that the ImageNet project *partially* falls in the category of those approaches which build training sets by performing *automatic retrieval* of the images [14,15]. In very general terms, the idea consists in using a term denoting the class of a target object as keyword for an image search engine and forming with the images retrieved in this way the training set. Search engines index images on the basis of the texts that accompany them and of users' tags, when they are present.

The obvious advantage of these approaches is that they can use a great amount of images to form the training set; on the other hand, the training set obtained in this way depends on the ranking of the images, that is, the first images provided by a search engine (say Google Images) are those which rank high in its indexing system. This is not beneficial for our purpose, since we would like to obtain a set of images covering the visual heterogeneity of a visual concept, and not only prototypical instances. As an example, we can take a look to the first 20 images retrieved by Google Images when using the keyword "cat" (Fig. 1). As visible, in most of the cases the cat is frontal, on a synthetic background, focusing on the snout.

These considerations suggest that, starting from the simple image search of Google, many steps ahead could be taken towards the creation of an expressive dataset.

So, the challenging question we will try to answer is: how is it possible to exploit the big amount of images that are available on the web and to automatize the search, providing a training set of pictures which mostly represent the variety of a given concept?

Our proposal is to refine the web search by adding to the standard keyword denoting the class of objects to be detected some other related terms, in order to make the search more expressive. However, we would like these terms to be added not to be arbitrarily chosen, but rather selected with a criterion that has to be explicit and meaningful. More specifically, we would like such accompanying keywords to have three important features:

1. to be frequently associated with the word denoting the target object (otherwise, too few images would be retrieved by the association of the two keywords);
2. to be meaningful from a visual point of view (as usually people tag pictures on the basis of what is depicted in them);
3. to capture the maximum possible level of variability of the addressed class.

Our approach can be summarized as follows: in the first step, we consider a large textual dataset (Google Ngram[1]), containing 930 Gigabytes of text material; from Google Ngram we extract bi-grams containing the word denoting the target object (for simplicity, let's call it "target word") plus other terms, associated with their frequency in the dataset. In the second step, this input is filtered in various ways, distilling information useful for capturing the visual variability of the object of interest. To this aim, WordNet will be exploited. More specifically, among the most frequent *nouns* that accompany the target word in the bi-grams, *hyponyms* will be kept, thus capturing entities which belong to subclasses of the object of interest. *Adjectives* denoting visual properties will be also kept, that is, adjectives which characterize visible aspects of objects (their color, their patterns). Finally, among *verbs*, present participles are kept, in order to capture actions that can be performed or are performed by the entity of interest.

In the third step, these aspects will be fused together following two different criteria: in the first "frequency based" one we choose, among all selected words, those that, coupled with the target word, have the highest score in terms of frequency (disregarding whether they are visual adjectives, verbs or hyponyms). The final result of such process will be a list of pairs of words, composed by the target word plus an accompanying word, chosen with explicit and semantic criteria that, fed into image search engines, will provide semantically rich shots for training the object classifiers.

In the second strategy, we build three separate image sets, including bi-grams formed by target word + visual properties, by target word + hyponyms, and by target word + verbs, respectively. These are then fed into three separate classifiers, whose classification decisions on a given test sample are subsequently fused using standard fusion rules. In addition, a "grounding" operation is adopted to reduce polysemy issues: it is assumed that, at the moment of the definition of a target word, a more generic term is also given (an *hypernym*). This term is added to all the strings created so far. Experimentally, this ensures a semantically more coherent image collection.

The aim of the experiments is to validate the goodness of the training datasets automatically built by our method, under different respects. We take inspiration from the ImageNet paper [12], following some of its experimental protocols. In first instance, we analyze the object classification accuracy derived from our data, mainly focusing on the PASCAL VOC 2012 "comp2" competition. This is car-

[1] https://books.google.com/ngrams/.