Pattern Recognition 45 (2012) 2962–2979

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Leveraging social media for scalable object detection

E. Chatzilari [a,b,1], S. Nikolopoulos [a,c,*,1], I. Patras [c,2], I. Kompatsiaris [a,1]

[a] Centre for Research and Technology Hellas, Informatics and Telematics Institute, 6th Km Charilaou-Thermi Road, Thermi-Thessaloniki, GR-57001 Thessaloniki, Greece
[b] Centre for Vision, Speech and Signal Processing University of Surrey Guildford, GU2 7XH, UK
[c] School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS London, UK

ARTICLE INFO

Article history:
Received 8 May 2010
Received in revised form
8 December 2011
Accepted 9 February 2012
Available online 18 February 2012

Keywords:
Social media
Collaborative tagging
Flickr
Object detection
Weak annotations
Scalable learning
Effortless learning

ABSTRACT

In this manuscript we present a method that leverages social media for the effortless learning of object detectors. We are motivated by the fact that the increased training cost of methods demanding manual annotation, limits their ability to easily scale in different types of objects and domains. At the same time, the rapidly growing social media applications have made available a tremendous volume of tagged images, which could serve as a solution for this problem. However, the nature of annotations (i.e. global level) and the noise existing in the associated information (due to lack of structure, ambiguity, redundancy, and emotional tagging), prevents them from being readily compatible (i.e. accurate region level annotations) with the existing methods for training object detectors. We present a novel approach to overcome this deficiency using the collective knowledge aggregated in social sites to automatically determine a set of image regions that can be associated with a certain object. We study theoretically and experimentally when the prevailing trends (in terms of appearance frequency) in visual and tag information space converge into the same object, and how this convergence is influenced by the number of utilized images and the accuracy of the visual analysis algorithms. Evaluation results show that although the models trained using leveraged social media are inferior to the ones trained manually, there are cases where the user contributed content can be successfully used to facilitate scalable and effortless learning of object detectors.

1. Introduction

Semantic object detection is considered one of the most useful operations performed by the human visual system and constitutes an exciting problem for computer vision scientists. Many researchers in the field have focused on trying to discover a scalable (in terms of the number of concepts) and effortless (in terms of the necessary annotation) way to teach the machine how to recognize visual objects the way a human does. The authors of [1] make the hypothesis that once a few categories have been learned with significant cost, some information may be abstracted from the process to make learning further categories more efficient. Similarly in [2] when images of new concepts are added to the visual analysis model, the computer only needs to learn from the new images, since profiling models are used to store the information learned from previous concepts. In the same lines the need to efficiently handle the huge amounts of data generated on the Web, has prompted many researchers to investigate the use of online learning algorithms [3] for exploiting those data. Motivated by the same need but relying on a non-parametric approach, the authors of [4] claim that with the availability of overwhelming amounts of data many problems can be solved without the need for sophisticated algorithms. The authors present a visual analog to Google's "Did you mean" tool, which corrects errors in search queries by memorizing billions of query–answer pairs and suggesting the one closest to the user query. Additionally, the authors of [5] employ multiple instance learning [6] to learn models from globally annotated images, while in [7] object recognition is viewed as machine translation that uses expectation maximization in order to learn how to map visual objects (blobs) to concept labels. The approaches relying on human computation such as Google Image Labeler [8] and Peekaboom [9] for image global and regional annotation respectively, also belong to the category of methods that aim at scalable and effortless learning. Motivated by the same objective, in this work we investigate whether the knowledge aggregated in social tagging systems by the collaboration of web users, can help in the process of teaching the machine to recognize objects.

Machine learning algorithms for object detection fall in two main categories in terms of the annotation granularity characterizing their

* Corresponding author at: Centre for Research and Technology Hellas, Informatics and Telematics institute, 6th Km Charilaou-Thermi Road, Thermi-Thessaloniki, GR-57001 Thessaloniki, Greece. Tel.: +30 2311257752; fax: +30 2310474128.
E-mail addresses: ehatzi@iti.gr (E. Chatzilari), nikolopo@iti.gr (S. Nikolopoulos), i.patras@eecs.qmul.ac.uk (I. Patras), ikom@iti.gr (I. Kompatsiaris).
[1] Tel.: +30 2311 257701–3; fax: +30 2310 474128.
[2] Tel.: +44 20 7882 7523; fax: +44 20 7882 7997.

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2012.02.006

learning samples. The algorithms that are designed to learn from strongly annotated samples [10–12] (i.e. samples in which the exact location of an object within an image is known) and the algorithms that learn from weakly annotated samples [13,7,5,14] (i.e. samples in which it is known that an object is depicted in the image, but its location is unknown). In the first case, the goal is to learn a mapping from visual features f_i to semantic labels c_i (e.g. a face [10,12] or a car [11]) given a training set made of pairs (f_i,c_i). New images are annotated using the learned mapping to derive the semantic labels that correspond to the visual features of the new image. On the other hand, in the case of weakly annotated training samples the goal is to estimate the joint probability distribution between the visual features f_i and the semantic labels c_i given a training set made of pairs between sets $\{(f_1,\ldots,f_n),(c_1,\ldots,c_m)\}$. New images are annotated by choosing the semantic labels that maximize the learned joint probability distribution given the visual features of the new image. Some indicative works that fall within the weakly supervised framework include the ones relying on aspect models like probabilistic Latent Semantic Analysis (pLSA) [13,15] and Latent Dirichlet Allocation (LDA) [16,17] that are typically used for estimating the necessary joint probability distribution.

While model parameters can be estimated more efficiently from strongly annotated samples, such samples are very expensive to obtain raising scalability problems. On the contrary, weakly annotated samples can be easily obtained in large quantities from social networks. Motivated by this fact our work aims at combining the advantages of both strongly supervised (learn model parameters more efficiently) and weakly supervised (learn from samples obtained at low cost) methods, by allowing the strongly supervised methods to learn from training samples that can be mined from collaborative tagging environments. The problem we consider is essentially a multiple-instance learning problem in noisy context, where we try to exploit the noise reduction properties that characterize massive user contributions, given that they encode the collective knowledge of multiple users. Indeed, flickr hosts a series of implicit links between images that can be mined using criteria such as geo-location information, temporal proximity between the timestamps of images uploaded by the same user, or images associated with the same event. The goal of this work is to exploit the social aspect of the contributed content at the level of tags. More specifically, given that in collaborative tagging environments the generated annotations may be considered to be the result of the collaboration among individuals, we can reasonably expect that tag assignments are filtered by the collaborative effort of the users, yielding more consistent annotations. In this context, drawing from a large pool of weakly annotated images, our goal is to benefit from the knowledge aggregated in social tagging systems in order to automatically determine a set of image regions that can be associated with a certain object.

In order to achieve this goal, we consider that if the set of weakly annotated images is properly selected, the most populated tag-"term" and the most populated visual-"term" will be two different representations (i.e. textual and visual) of the same object. We define tag-"terms" to be sets of tag instances grouped based on their semantic affinity (e.g. synonyms, derivatives, etc.). Respectively, we define visual-"terms" to be sets of region instances grouped based on their visual similarity (e.g. clustering using the regions' visual features). The most populated tag-"term" (i.e. the most frequently appearing tag, counting also its synonyms, derivatives, etc.) is used to provide the semantic label of the object that the developed classifier is trained to recognize, while the most populated visual-"term" (i.e. the most populated cluster of image regions) is used to provide the set of positive samples for training the classifier in a strongly supervised manner. Our method relies on the fact that due to the common background that most users share, the majority of them tend to contribute relevant tags when faced with similar type of visual content [18]. Given this fact, it is expected that as the pool of the weakly annotated images grows, the most frequently appearing "term" in both tag and visual information space will converge into the same object.

In this context, the contribution of our work is on studying theoretically and experimentally the conditions under which the most frequently appearing "terms" in tag and visual information space are expected to converge into the same object. This is evident in the ideal case where tags are accurate and free of ambiguity, and no error is introduced by the visual analysis algorithms. However, considering that this is rarely the case, we expect that the use of a large size dataset favors convergence since a statistically significant amount of samples can compensate for the error introduced by noisy tagging. On the contrary, the amount of error introduced by the visual analysis algorithms (i.e. segmentation accuracy and clustering efficiency) hinders convergence since the formulated clusters of image regions may not be consistent in a semantic sense. Our purpose in this work is to examine how these two aforementioned factors influence the convergence level between the most frequently appearing "terms" in visual and tag information space.

Preliminary versions of this work include [19,20]. The main difference with [19] is that in this early work we have followed a different methodological approach for selecting the set of regions that can be associated with a certain object. More specifically, the full set of image regions was split in two clusters and the cluster with the smallest population was selected to provide the training samples for the object detection model. Although successful for the objects that appeared frequently in social context, it was observed that our framework performed poorly for a non-negligible number of cases. This was the reason for turning into the methodological approach presented in this work, an early version of which was included in [20]. However, while the focus of [20] has been mostly on experimenting with various feature spaces and tuning the clustering algorithm, in this manuscript we provide a solid theoretical ground for gaining insight into the functionality of the proposed approach and deriving some conclusions about its success or failure. Moreover, we experimentally examine the ability of our method in scaling to various types of objects, allowing us to derive useful conclusions about the learning efficiency of the resulting object detection models.

The rest of the manuscript is organized as follows. Section 2 reviews the related literature. Section 3 describes the general architecture of the framework we propose for leveraging social media and provides technical details for the analysis components that are employed by our framework. Section 4 investigates theoretically the relation between the size of the dataset, the visual analysis error and the convergence level of the most frequently appearing tag and visual "terms". Our experimental study is presented in Section 5, while Section 6 discusses the results and provides some directions for future work.

2. Related work

Lately there has been considerable interest on weakly labeled data and their potential to serve as the training samples for various computer vision tasks. The common objective of these approaches is to compensate for the loss in learning from weakly annotated and noisy training data, by exploiting the arbitrary large amount of available samples. Web 2.0 and collaborative tagging environments have further boosted this idea by making available plentiful user tagged data.