Pattern Recognition 46 (2013) 2437–2449

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

A region-centered topic model for object discovery and category-based image segmentation

Iván González-Díaz*, Fernando Díaz-de-María

Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Leganés 28911, Madrid

ARTICLE INFO

Article history:
Received 7 September 2012
Received in revised form
7 January 2013
Accepted 27 January 2013
Available online 12 February 2013

Keywords:
Latent topic models
Topic discovery
Category-based image segmentation
Kernel Logistic Regression
Context

ABSTRACT

Latent topic models have become a popular paradigm in many computer vision applications due to their capability to unsupervisely discover semantics in visual content. Relying on the Bag-of-Words representation, they consider images as mixtures of latent topics that generate visual words according to some specific distributions. However, the performance of these methods is still limited by the way in which they take into account the spatial distribution of visual words and, what is even more important, the currently used appearance distributions. In this paper, we propose a novel region-centered latent topic model that introduces two main contributions: first, an improved spatial context model that allows for considering inter-topic inter-region influences; and second, an advanced region-based appearance distribution built on the Kernel Logistic Regressor. It is worth highlighting that the proposed contributions have been seamlessly integrated in the model, so that all the parameters are concurrently estimated using a unified inference process. Furthermore, the proposed model has been extended to work in both unsupervised and supervised modes. Our results for unsupervised mode improve 30% those of previous latent topic models. For supervised mode, where discriminative approaches are preponderant, our results are quite close to those of discriminative state-of-the-art methods.

1. Introduction

During the last years, a significant amount of research effort has been devoted to the *category-based image segmentation* problem since it has become an essential part of contemporary scene understanding systems, which have emerged as a natural extension of the classical image classification and recognition systems. The category-based image segmentation (also known as object class image segmentation) differs from standard image segmentation in that it not only divides the image into a set of coherent regions, but also assigns a category to each region. Several methods have been proposed to address this problem. Most of them are discriminative solutions using conditional random fields (CRF), such as those in [1–4], but generative approaches can be also found in the literature [5,6].

In this paper we focus on latent topic models (LTM), a generative paradigm that explains the data of a corpus as a mixture of latent topics that represent semantic entities. In particular, probabilistic latent semantic analysis (PLSA) [7] and latent Dirichlet allocation (LDA) [8] are the most outstanding examples of this type of models. Although both PLSA and LDA were originally conceived as unsupervised models, their formulation has been extended to the supervised case (the interested reader is referred to [9] for an excellent example of supervised topic models), thus providing a unified framework to work in both modes. However, traditional approximations to supervised scenarios suffer from one drawback that we tackle in this paper. In particular, in previous approaches, labels for supervised training were usually applied at a granularity level that does not fit with topics. Therefore, these approaches to supervised topic models were not able to take full advantage of ground-truth pixel-wise segmentations typical of category-based segmentation tasks.

This paper complements and extends our previous work described in [10]. Specifically, the model proposed in this paper has been built on LDA instead of PLSA; we have moved from an *intra-topic* to an *inter-topic* influence model, improving the modeling of the spatial arrangement of the topics; we have introduced a novel KLR-based appearance model; and, finally, the experimental evaluation has been significantly extended.

In summary, this paper makes a couple of significant contributions. First, the proposed model extends LDA to take into account the spatial arrangement of topics in an image. This is achieved by modeling not only the typical spatial location of a topic, but also its context. In particular, the proposed model

* Corresponding author. Tel.: +34 916246262. fax: +34 916248749.
E-mail addresses: igonzalez@tsc.uc3m.es (I. González-Díaz), fdiaz@tsc.uc3m.es (F. Díaz-de-María).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.01.034

allows for a flexible management of inter-region inter-topic influences, outperforming the conventional approaches found in the latent topic literature. Second, the appearance model usually employed by latent topic models (a multinomial distribution over visual words) has been notably enhanced by means of the use of a Kernel Logistic Regression (KLR), which takes into account the relations among descriptors within a region. The inclusion of a KLR is not a simple plug-in in the model, since one needs to develop inference methods that concurrently optimize all the variables involved in the generative process, while keeping the computational complexity low enough to make the optimization feasible. Furthermore, we also demonstrate how our model is able to work in both unsupervised and supervised modes, a key differentiating factor with respect to most of the (discriminative) approaches found in the literature. Specifically, a soft-labeling technique has been proposed that keeps the latent nature of the topics unaltered and improves the results in supervised tasks when compared to the customary hard-labeling approach.

The paper is organized as follows: Section 2 summarizes related work. Section 3 provides an overview of the proposed generative latent topic model. Sections 4 and 5 describe the two main contributions of this work: the context model and the appearance distribution model, respectively. Section 6 puts forward the required extensions for the model to work in supervised mode. Section 7 describes the proposed inference algorithm. Section 8 describes the experiments and discusses the results; and, finally, Section 9 summarizes our conclusions.

2. Related work

This section focuses on the existing models for the spatial distribution of visual words and the appearance in LTMs, which are the two areas where this paper contributes.

2.1. Modeling the spatial distribution of visual words in LTMs

Undoubtedly, the most important limitation of the original formulation of PLSA and LDA for computer vision is that they do not take into account the spatial distribution of visual words in the images. The potential benefits of this spatial modeling are twofold: first, an improved performance of latent topic models in tasks such as image classification or topic discovery; and second, an enrichment of such models with the capability of generating robust image segmentations. Nevertheless, modeling the spatial location of visual words is no longer straightforward in this framework since both appearance and spatial models must be jointly trained using the same learning algorithm that infers the latent topics.

Some early approaches considering simple geometric modeling deserve to be mentioned. In [11], the use of doublets of visual words over PLSA allowed the authors to add simple geometric considerations, achieving notable improvements in object localization. In [12], the authors modeled the joint distribution of visual features and their locations using a translation- and scale-invariant approach for unsupervised category discovering. And in [13], Gaussian and uniform spatial distributions were used to model foreground and background topics, respectively. Furthermore, in [14,15], LDA and PLSA were extended, respectively, to model the spatial distribution of words using fixed grid cells. In other kind of approaches, such as that described in [16], the geometric information was encoded using what is known as part models, in which the objects are assumed to be made up of constituent parts.

Other proposals went a bit further and incorporated a blind segmentation of the images into the latent topic models. In [17], a new version of PLSA was proposed that considered topics at region level (where the regions come from a previous segmentation) for an image retrieval task. In [18], a novel approach to deal with under- and over-segmentations was proposed; specifically, segmentations were generated at different levels, then PLSA was used to unsupervisely detect categories, and finally the best segmentation level was chosen according to the distance between the proposed regions and the detected categories. In [19], an extension of LDA was proposed that considered topics at an intermediate level (regions); these topics produced two kinds of visual words, one related to the color of the whole region and the other related to the texture descriptors of the local patches within the region, so that the algorithm started from an over-segmented version of the image to end up with a more realistic segmentation, where regions were (hopefully) associated with semantic concepts. Similar approaches have been successfully applied to image classification and annotation [20], as well as to scene understanding [21].

Nevertheless, in all of these models, the regions were considered as independent entities that did not interact with each other. Other methods, such as [22,23], imposed certain spatial coherence by allowing interactions among regions; specifically, Markov Random Fields (MRFs) were used to drive spatially connected regions toward the same topic. We refer to these models as *inter-region intra-topic* context models since a region pushes other surrounding regions to belong to the same topic. The model proposed in this paper goes beyond by defining an *inter-region inter-topic* context model, which allows for inter-topic interactions as described later. A similar idea using MRFs was proposed in [24].

2.2. Appearance model in LTMs

Traditionally, the appearance model in LTMs follows a multinomial distribution over each visual word. Although assigning topics at visual word level might seem appealing for simplicity reasons, many authors have preferred to work at region level in order to provide more stable representations than those directly derived from individual visual words (see [2,17]). In the LDA formulation, this region-based granularity level has been customarily handled by considering the probability associated with the appearance of a region as the product of the probabilities (multiplicative model) of the visual words that lie within that region (the interested reader is referred to [19,20] for more information). Nevertheless, the multiplicative model may become overly dependent on a particular visual word when estimating the probability associated with a whole region; furthermore, this multiplicative appearance model actually considers local patches as individual entities so that, given the topic of the region, their appearances are conditionally independent.

Our proposal differs from these approaches in that a descriptor for the whole region is computed and used in the appearance model. Furthermore, the appearance of a region is modeled through a Kernel Logistic Regressor (KLR), so that the appearance model takes into account the relations among visual words within a region. Although the KLR has been already used in discriminative models, as far as we know it is the first time that has been included in a latent topic model. As mentioned in the Introduction, it is not straightforward at all since the incorporation of the KLR involves developing inference methods for all the variables while keeping a moderate complexity level.

3. Model overview

In this section we provide an overview of the proposed generative model, which is built on LDA. For a detailed description of LDA, the interested reader is referred to [8].