Pattern Recognition (2013). 46(9), 2437-2449.
DOI: http://dx.doi.org/10.1016/j.patcog.2013.01.034

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

Iván González-Díaz[a], Fernando Díaz-de-María[a]

[a]*Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Leganés, 28911, Madrid e-mail: {igonzalez,fdiaz}@tsc.uc3m.es*

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.

Keywords: Latent Topic Models, Topic Discovery, Category-based Image segmentation, Kernel Logistic Regression, Context

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, 2, 3, 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 an 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 in-

troduced 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 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] and [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 (MRF) 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

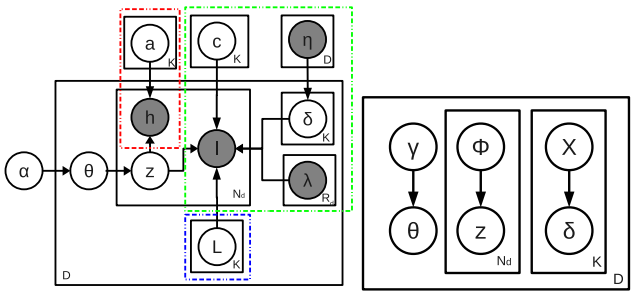
Traditionally, the appearance model in LTMs follows a multinomial distribution over each visual word. Although assigning topics at visual word level might seem appealing for simplicity reasons, many authors have preferred to work at region level in order to provide more stable representations than those directly derived from individual visual words (see [2, 17]). In the LDA for-

(a) (b) (c) (d)

Figure 1: Example of the steps involved in the proposed generative model. (a) Image to be processed. (b) Image segmentation that is used as the geometric layout of the image. (c) Ground truth segmentation (desired output of the algorithm), where each color is associated with a particular semantic concept (green denotes 'grass' and blue 'sheep'). (d) Outcome of the proposed method.

mulation, 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.



(a) (b)

Figure 2: (a) Graphical model of the proposed approach. The new variables of the model are drawn within dotted boxes: context model (green), appearance model (red), and extensions for supervised mode (blue). (b) Graphical model of the variational distribution used to approximate the posterior in the proposed model. Nodes represent random variables (observed-shaded, latent-unshaded), edges show dependencies among variables, and boxes refer to different instances of the same variable.

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].

First, it is important to mention that our model relies on a previous blind (over) segmentation of the image. This segmentation encodes the spatial geometry of the scene so that a sample in our method is associated with a region instead of a local patch (that was what happened in traditional latent topic approaches). In particular, we have generated image partitions of about 20-40 regions using the algorithm described in [25] (see Fig. 1(b) for an example).

Fig. 2(a) shows the graphical model representation of the proposed latent topic model. As shown in the figure, given a corpus D of documents (images), each image $d \in D$ is represented by means of a set of N_d samples and R_d regions in the segmentation. The objetive of the model is to explain each image as a mixture of K latent topics, each of them showing specific appearance and geometric properties. The components of the graphical model have been grouped into three subsystems identified by means of dashed boxes in the fig-

ure; namely: the context model, the appearance model, and the extensions for supervised mode. These subsystems will be explained in detail along the next sections.

Each variable in the model has been proposed in accordance with a particular assumption regarding the image formation process. Specifically, the following assumptions have been considered:

a) Images are generated by means of a mixture of latent topics z that are in turn associated with semantic concepts (such as 'grass', 'sky', or 'road'). Hence, the topic is the key variable in the generative model and will serve to provide category-based image segmentations.

b) Each topic produces samples whose appearance is encoded by means of a single descriptor h that is made-up from the descriptors of the local patches within that region. We firmly believe that local descriptors cannot be considered as independent variables, and that the relations among descriptors within the same region are of great importance to decide on the associated topic. This element is represented inside the dotted red-box in Fig. 2(a).

c) The geometric layout l of an image can be modeled by means of an over-segmentation based on low-level features. This segmentation produces a set of R_d regions, each of them belonging to just one topic (then, under-segmentation is not suitable). Of course, some heterogeneous connected regions may belong to the same object and thus be associated with the same topic. The generative model is then in charge of bridging the gap between the initial over-segmentation and the final representation of the image, in which regions are associated with semantic concepts (see Fig. 1 (b) and (d) for an illustrative example).

d) Some topics exhibit a strong spatial correlation with each other, appearing in neighboring areas of the image (e.g. 'car'/'road', 'aeroplane'/'sky', etc.). This observation motivates our concept of what is referred to as 'spatial context'. Traditional context models in LTMs, such as the Markov Random Field presented in [23], just provide intra-topic influences among regions (i.e., a region pushes neighboring regions to belong to the same topic); however, such a model is not expressive enough to handle the important inter-topic correlations that

actually occur in typical scenes showing several semantic concepts. The context model is represented inside the green dotted box in Fig. 2(a).

Next we provide an overall description of the corresponding generative process in unsupervised mode (when no labels are available to train the model):

1. Consider a K-dimensional Dirichlet parameter $\boldsymbol{\alpha}$, with $\alpha_k > 0$, that defines a parametric distribution of the topics in the corpus (the topic proportions at the corpus level).
2. For each image d,
   (a) Generate a blind over-segmentation of the image into R_d regions.
   (b) Sample a Dirichlet random variable $\boldsymbol{\theta}|\boldsymbol{\alpha} \sim Dir(\boldsymbol{\alpha})$ that defines the particular probability distribution over the K latent topics for this image.
3. For each sample $n \in \{1, 2, \ldots, N_d\}$:
   (a) Choose a topic z_n according to the probability distribution defined by $\boldsymbol{\theta}$: $z_n|\boldsymbol{\theta} \sim Mult(\boldsymbol{\theta})$, where $Mult(.)$ stands for a Multinomial Distribution.
   (b) Draw an appearance h_n, as will be explained in Section 5.
   (c) Choose a topic location l_n by selecting a region $r \in 1, 2 \ldots R_d$ from the initial segmentation, as will be explained in Section 4.

At this point, it is worth clarifying the conceptual difference between samples (indexed by $n = 1, 2, \ldots, N_d$) and regions (indexed by $r = 1, 2, \ldots, R_d$). The samples are those inherent to the generative process, i.e., in order to generate an image according to our model, a set of samples are generated, and for each one, a topic is chosen, then an appearance, and then a location. The regions are those resulting from the previous over-segmentation that is used as a geometric layout on which to build our image representation. Therefore, in general, it would be possible (from the generative model point of view) to have either more than one sample associated with an actual region, or even empty regions. For practical purposes, in the image representation used in this paper the correspondence is

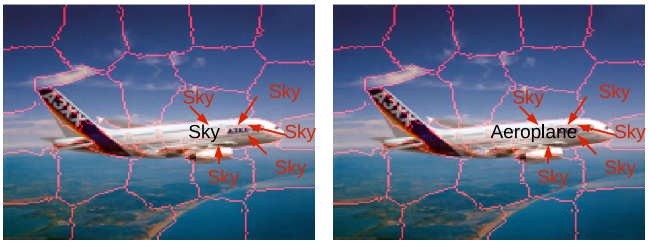


Figure 3: Illustrative example of two context models: (Left) Intra-topic inter-region context model from MRF-LDA [23] (Right) Proposed inter-topic inter-region context model.

one-to-one, i.e., a sample is generated for each actual region. Nevertheless, in order to provide a general formulation, we keep both N_d and R_d further on.

4. The Context-based Location model

This part of the proposal gives meaning to the variables of our model that lie inside the green dashed box in Fig. 2(a). The aim is to select a location given a topic, i.e., to choose the most appropriate region given a topic. To this purpose, a *context-based spatial location* distribution is proposed.

The proposed context model incorporates inter-region inter-topic relations to the generative process while keeping it simple enough to allow for closed expressions in the inference process. As mentioned before, the objective of this model is to set the basis for *inter-region inter-topic cooperation*. This means that regions belonging to a particular topic A may push other regions towards belonging to other topic B when both topics are spatially correlated (they tend to appear together). Fig. 3 compares our inter-topic inter-region context model with the intra-topic inter-region cooperation model used in MRF-LDA [23].

Intuitively speaking, the generative process of the context model is as follows: once we have selected a topic z_n, we look for its best location in the image, i.e., we look for the particular region that best fits our context model. To that end, the proposed context model, illustrated in Fig. 4, relies on three variables:

- $\boldsymbol{\lambda}$ represents what we call the *geometric context*, which is a measurement of the influence of a region on the others according to a relative measurement

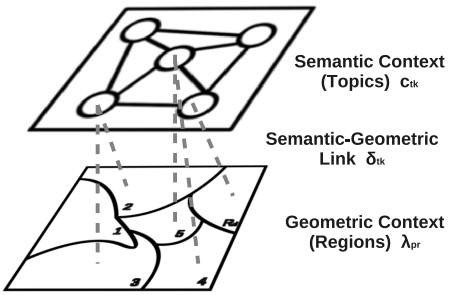


Figure 4: A graphical representation of the proposed context model involving the semantic and geometric spaces, and the links between them.

of their common boundary. Defining l_{pr} as the common boundary of two regions p and r, and l_r as the perimeter of the region r, the influence λ_{pr} of the region p on the region r is computed as $\lambda_{pr} = \frac{l_{pr}}{l_r}$ if $p \neq r$, and zero otherwise. Hence, it measures the portion of the total perimeter of the influenced region that is shared with the influencing region. Additionally, $\boldsymbol{\lambda}$ is further normalized to obey $\sum_{r=1}^{R_d} \lambda_{pr} = 1$. As can be easily noticed, the influences are not symmetric: $l_{pr} \neq l_{rp}$. In fact, our definition favors the influence of larger regions over smaller ones, what, from our point of view, makes sense. $\boldsymbol{\lambda}$ values are pre-computed and remain fixed during inference.

- $\mathbf{c}$ represents what we call the *semantic context*, which takes into consideration the spatial correlation among topics. In particular, this variable is a collection of K K-dimensional multinomial parameters $\mathbf{c}_t$ shared by all the documents in the corpus. In particular, c_{tk} estimates the probability of co-occurrence of topics t and k in spatially adjoining regions. These probabilities satisfy $\sum_{t=1}^{K} c_{tk} = 1$ and, again, are not symmetric, i.e. $c_{tk} \neq c_{kt}$.

- $\boldsymbol{\delta}$, called the *semantic-geometric link*, provides a link between the topic space and the geometric layout of each image. It is a document-dependent collection of K R_d-dimensional multinomial parameters $\boldsymbol{\delta}_t = [\delta_{t1} \dots \delta_{tR_d}]$, with $\sum_{p=1}^{R_d} \delta_{tp} = 1$. Each component of the vector δ_{tp} intends to capture

the importance of a region p given a topic t, and must be inferred during the inference process.

Putting these three components together, and with the aim of limiting the complexity of the solution, we propose a context model as a product of discrete distributions:

$$p(l_n|z_n, \boldsymbol{\delta}, \mathbf{c}, \boldsymbol{\lambda}) = \sum_{t=1}^{K} \sum_{p=1}^{R_d} 1[l_n = r] c_{tz_n} \delta_{tp} \lambda_{pr} \tag{1}$$

where the expression $1[l_n = r]$ is a simple indicator variable that means that the location l_n of a sample n points to the region r in the previous over-segmentation.

That is, for each potential location r, with $r \in \{1, 2, \ldots, R_d\}$, we consider the influence of each neighboring location p (λ_{pr}), the influence of each correlated topic t (c_{tz_n}), and the link between both (δ_{tp}). It is worth mentioning that this model could remind a Random Field that connects neighboring regions with a pairwise potential, which is scaled as a function of the influence between regions and is dependent on the combination of classes associated with each region. However, the proposed formulation has been specifically developed for a topic model and leads to a simpler optimization process.

It is easy to notice that $\sum_r p(l_n = r|z_n, \boldsymbol{\delta}, \mathbf{c}, \boldsymbol{\lambda}) = 1$, so that l_n lives in a R_d-simplex of regions coming from the initial over-segmentation of the image. Furthermore, for regularization purposes, we have also used a prior Dirichlet hyperparameter $\boldsymbol{\eta}$ over the semantic-geometric links $\boldsymbol{\delta}$.

It is also worth noting that regions have to be in contact in order to generate positive geometric influences (since they depends on the common boundary). Consequently only those topics that actually have some contact are considered as correlated topics, thus removing relations between topics that, although appear in the same image, are located at disconnected areas. The rationale behind this approach is related to the computational complexity of the model, which is initially quadratic with both the number of topics $o(K^2)$ and the number of regions $o(R^2)$ in the image. Thus, taking advantage of this limited number of positive influences, the computational complexity is dramatically reduced. In

Figure 5: A comparison of the intra-topic and inter-topic empirical context distributions: a) original image, b) grass intra-topic context, c) sheep intra-topic context, d) grass inter-topic context, and e) sheep inter-topic context. Lighter colors represent higher probabilities.

practice, we have found that the actual complexity is linear with the number of regions $o(nR)$, with $n \approx 5$ in our experiments. Furthermore, we have noticed that less constrained influence models do not achieve significantly better results while notably increasing the computational cost of the inference.

Fig. 5 shows an illustrative example comparing the proposed context model to the intra-topic inter-region model for an image of the MSRC dataset.

In summary, the process followed to set-up the context model is as follows: each image d of the dataset is segmented into R_d regions. Then, the geometric context of the image, parametrized through the influences between regions $\boldsymbol{\lambda}$, is computed from their common boundaries. These influences remain constant during the learning phase of the generative model, in which the other variables in the context model ($\mathbf{c}$ and $deltav$) are jointly optimized with the rest of the elements in the whole generative model.

5. Improving the appearance model using a Kernel Logistic Regressor

In the proposed approach the appearance model relies on region-level descriptors, and is implemented by means of a nonlinear probabilistic machine learning approach known as the Kernel Logistic Regressor (KLR). This subsystem of our model is depicted within a red dashed box in Fig 2. In the following subsections, we describe both the parametrization and the learning approach of our appearance distribution.

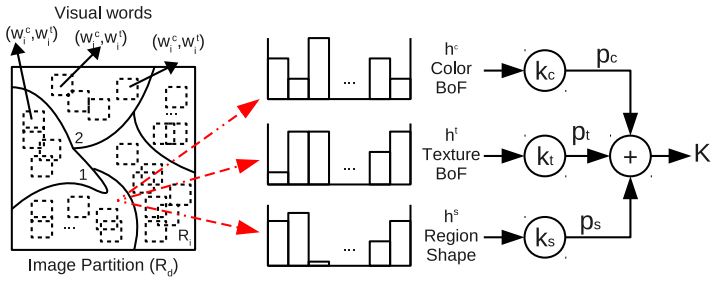


Figure 6: Illustration of the process for generating the region-level descriptors. Each region contains several local patches, where color and texture descriptors are computed. Region level descriptors (h^t, h^c) are then computed using a BoW approach and a shape descriptor h^s is added. Then, independent kernel values using KLR are computed for each feature (k_c, k_t, and k_s) and a final unique output (K) is computed by weighting the individual kernels.

5.1. Obtaining features at region level

For each region in the image, three descriptors are computed: color, texture, and shape descriptors. Color and texture descriptors are computed by means of a Bag-of-Words (BoW) approach at region level. Specifically, we have computed these descriptors over a multi-scale dense grid of local circular patches, as described in [26]. Color descriptors are 36-dimensional Robust Hue Histograms [27] whereas texture descriptors are 128-dimensional SIFT features [28]. Then, independent visual vocabularies for color and texture are calculated, with $V^c = 1000$ and $V^t = 4000$ for color and texture, respectively. Once each local descriptor is assigned to the closest visual word in each vocabulary (w_i^t, w_i^c), color and texture word occurrence histograms (h_n^c, h_n^t) are computed at region level. This approach, illustrated in Fig. 6, allows the generative model to learn the relations between descriptors within the same region what, to the authors best knowledge, has not been handled by any latent topic model yet.

In addition, a simple shape descriptor h_n^s has been also included by computing a 8-orientation histogram from the Freeman chain code of the region boundary [29].

5.2. Proposed generative appearance distribution based on the KLR

The appearance of a region is computed using a Kernel Logistic Regressor (KLR) that takes into account the nonlinear relations among visual words within the region. As shown in [30], the negative log-likelihood cost function of the KLR exhibits a similar shape to that of the Support Vector Machine (SVM) [31] except for the well-classified samples (which still influence the KLR, but no the SVM). Consequently, the KLR keeps the outstanding discriminative power of the SVM.

Although the KLR provides an estimate of the (discriminative) probability $p(z_n|h_n)$ of a topic z_n given the visual descriptor h_n of a region, in this work a modified version of the KLR has been used as part of a generative model, what represents in itself a novel approach. In particular, given a sample n and an associated topic z_n, we propose the use of the following distribution:

$$p(h_n|z_n, \mathbf{a}) = \frac{n_{z_n}}{1 + e^{-f_{z_n}(h_n)}} \tag{2}$$

where h_n represents the input features for the region n; n_{z_n} is a normalization term that ensures that $p(h_n|z_n, \mathbf{a})$ is a probability density function over the potential values of h_n; and $f_{z_n}(h_n)$ is a function whose optimal form, using the representer theorem, is as follows:

$$f_{z_n}(h_n) = \sum_{s=1}^{S} a_{z_n s} K(n, s) \tag{3}$$

where S is the training dataset where each sample s is called a *Reference Point* in S; $K(n, s)$ denotes the Kernel function between a sample n and a reference point s; and the parameters a_{z_n} represent the weights of the KLR associated with each reference point. For simplicity, the bias term has been omitted.

In practice, since some of the weights a_{z_n} are zero and, even more, many of them can be set to zero without significant loss of performance, the complexity of the KLR can be reduced significantly by selecting only those data that have a significant influence on the final result (so that the set S in eq. (3) is, in practice, just a subset of the database). It should also be noted that, since S

does not depend on z_n, we use the same set of reference points for every KLR (every topic in our model).

The normalization factor n_{z_n} in eq. (2) deserves a few words. Since the combination of different words in a region leads to a huge number of potential values for h_n, providing an exact normalization that ensures a unit integral for the appearance distribution becomes unfeasible. However, assuming that our features are limited to a set of finite volume in the feature space, we propose an approximate normalization. In particular, the normalization has been chosen to satisfy $\sum_{n \in TrainingSet} p(h_n|z_k, \mathbf{a}) = 1$ for each topic z_k, with $k = 1, 2, \ldots, K$. Since the proposed normalization considers just a limited combination of words, two comments are in order: first, the larger the training set, the better the approximation; and second, for normalization purposes, during the test, each sample should be converted into its nearest neighbor in the training set, so that the unseen samples do not break the normalization.

5.3. Feature fusion: a linear combination of kernels

As mentioned in section 5.1, three types of features are extracted for every region: color BoW, texture BoW, and region shape. In order to combine these features in a unique KLR output, a simple multiple kernel learning strategy [32] has been followed: for each feature h, a specific kernel K_h is computed; then, as illustrated in Fig. 6, a global kernel function is computed as a linear combination of the individual kernels:

$$K = \frac{1}{\sum_{h=1}^{3} p_h} \sum_{h=1}^{3} p_h K_h \tag{4}$$

In our case, histogram intersection kernels have been used due to the nature of the features. However, other features might be used that lead to other type of kernels (Linear, RBF, etc.). In our experiments, the weights values p_h have been selected by cross validation as described later in the experimental section.

5.4. Selecting the reference points

As we mentioned before, S is a subset of the training database, so that only samples showing notable influence on the results are included. The selection of

those samples that are taken as reference points in each KLR plays an important role in terms of both quality and efficiency. The first option is to use the whole training dataset, so that every region in every document is taken as reference. However, it makes more sense to look for a sparse representation that requires less computations and minimizes the over-fitting. Several authors have investigated this issue and both, unsupervised (such as [33]) and supervised ([34]) methods, have been proposed.

In our proposal we simply consider the likelihood of each sample, which implicitly considers label information if available. Hence, we select an initial set of reference points S_0 using a k-means-based clustering stage. Then, at each iteration, we add a new set of points whose appearance has not been properly modeled yet (samples with low likelihood). Although this approach is optimal when the number of reference points added at each iteration is $S_i^{new} = 1$ (otherwise, some of the samples might be highly correlated), the value of this parameter must be selected as a trade-off between performance and computational complexity. We have used $S_i^{new} = 50$ in our experiments.

6. Model extensions for supervised mode

In this section we describe the extensions of the model that make it suitable to work in supervised mode, in which a set of labeled training images is used to learn the distributions that are used later on a test set of unlabeled images. Specifically, in our case, these labels are given by ground truth pixel-wise segmentations.

The variables that support the supervised mode are represented inside the dotted blue box in Fig. 2 and described in detail in the following subsections.

6.1. Soft-labeling vs. hard-labeling

In supervised mode, we use the image labels to align the latent topics with the actual semantic concepts. Since we work with latent topic models, preserving its latent nature during the training phase becomes an important prerequisite.

The benefits of this approach are diverse: a) the model can successfully handle approximate annotations (see, for example, that illustrated in Fig. 1(c)); b) it may overcome under-segmentation of the training images; and c) it might handle datasets where some classes are pixel-wise annotated but others not.

A hard labeling strategy entails introducing multinomial distributions $z_n|L_n$ that depend on the label L_n associated with a region. In doing so, the topics are no longer latent since the topic associated with every region is actually being imposed (we would be closer to discriminative approaches than to generative ones). Therefore, we suggest a soft-labeling approach. First, we add to each image an artificial new region that is meant to be located outside the image and, consequently, neither contains local patches nor influences any other region. This new region is called the *non-image region*. Then, we propose to use a discrete distribution over the spatial location of the topics $l_n|z_n, L_r$, which is estimated in accordance with the ground truth segmentation available (training dataset). In our experiments, this estimation has been performed as follows: for every region (except for the non-image region), given a topic z, we set the spatial location l to 1 if the topic appears in the region, or to ϵ if not. While for the non-image region, $l = 1$ when the topic is not in the image and $l = \epsilon$ otherwise. Finally, the distribution is normalized by the sum over all regions. The value of ϵ is supposed to be very low ($\epsilon = 1e-4$ in our experiments), but not zero, in order to reach softer solutions.

Finally, one just have to compute the final location distribution $l_n|z_n, \delta, \lambda, \mathbf{c}, L_r$ as the product of the context-based and supervised distributions and normalize it to ensure that $\sum_{r=1}^{R_d} p(l_n = r|z_n, \delta, \lambda, \mathbf{c}, L_r) = 1$. In summary, the proposed approach actually sets a spatial distribution that depends on the selected topic rather than setting the topics themselves, thus preserving the latent nature of topics.

6.2. Extending the KLR-based appearance: taking into account negative samples

Following our graphical model, the appearance distribution in eq. (2) is only computed once a topic z_n has been chosen as the one that generates the region,

thus lacking of negative samples. This becomes a critical issue in supervised mode since only positives samples are available to train the regressor of the appearance model.

To overcome this issue we propose the following appearance distribution:

$$p(h_n|z_n, \mathbf{a}) = n_{z_n}\left(\frac{1}{1+e^{-f_{z_n}(h_n)}}\right)^{z_n}\left(\frac{1}{1+e^{f_{z_n}(h_n)}}\right)^{\bar{z}_n} \tag{5}$$

where $\bar{z}_n$ represents a new variable such that $p(\bar{z}_n) = 1 - p(z_n)$. In this manner, we ensure that both positive and negative samples are properly taken into account by the appearance distribution of the topics. Obviously, for test purposes, the term depending on $\bar{z}_n$ should be removed, the appearance models remain unchanged, and eq. (2) is used.

7. Inference

This Section describes the inference process. Given the set of model parameters $\Theta_p = \{\mathbf{a}, \mathbf{c}, \alpha, \delta, \lambda, \eta, \mathbf{g}, L\}$, exact inference is not possible due to the coupling between the variables θ and $\mathbf{z}$, what prevents from inferring the posterior distribution of the parameters given the data. Therefore, we propose to use a simplified variational distribution q (that is tractable) and mean-field variational inference, so that the Kullback-Leibler divergence between the variational distribution q and the posterior distribution of the parameters given the data $p(\Theta_p|\mathbf{h}, \mathbf{l}, \mathbf{g})$ is minimized. The new variational distribution q is represented in Fig. 2(b), where it is easy to notice how some links have been removed so that the independence among variables allows for an analytic solution. The variational distribution q can be written as follows:

$$q(\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\delta}|\Theta_v) = q(\boldsymbol{\theta}|\boldsymbol{\gamma}) \prod_{n=1}^{N_d} q(z_n|\boldsymbol{\phi}_n) \prod_{k=1}^{K} q(\boldsymbol{\delta}_k|\boldsymbol{\chi}_k) \tag{6}$$

where $\Theta_v = \{\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{\chi}\}$ are the variational parameters; $q(\boldsymbol{\theta}|\boldsymbol{\gamma})$ and $q(\boldsymbol{\delta}|\boldsymbol{\chi})$ are Dirichlet distributions; and $q(\mathbf{z}|\boldsymbol{\phi})$ is a multinomial distribution.

Hence, the log-likelihood of the data can be lower bounded as:

$$\log p(\mathbf{h},\mathbf{l},\mathbf{g}|\Theta_p) \geq E_q[\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})] + \sum_{n=1}^{N_d}\Big(E_q[\log p(z_n|\boldsymbol{\theta})] + E_q[\log p(h_n|z_n,\mathbf{a})] + E_q[\log p(l_n|z_n,\boldsymbol{\delta},\boldsymbol{\lambda},L_n)]\Big) + \sum_{k=1}^{K} E_q[\log p(\boldsymbol{\delta}|\boldsymbol{\eta})] + H(q) \tag{7}$$

where $E_q[\cdot]$ denotes the expectation over the variational distribution q, and $H(\cdot)$ the entropy of a distribution.

7.1. Obtaining a lower bound of the context term

The term of the log-likelihood that is associated with the context of a region requires computing a lower bound to make it tractable. To this end, we introduce a new variational parameter r_{tkpr}, such that $\sum_{t=1}^{K}\sum_{p=1}^{R_d} r_{tkpr} = 1$, that aims to capture the whole normalized relation, coming from both the geometric context and the semantic link between two regions p and r, given that the regions p and r belong to the topics t and k, respectively. Once this new variational parameter has been defined, the Jensen's inequality can be applied to determine the lower bound:

$$E_q[\log p(l_n|z_n,\boldsymbol{\delta},\boldsymbol{\lambda})] \geq \sum_{k=1}^{K}\sum_{t=1}^{K}\sum_{p=1}^{R_d}\phi_{nk}1[l_n=r]r_{tkpr}\left[\log\frac{c_{tk}\lambda_{pr}}{r_{tkpr}} + \Psi(\chi_{tp}) - \Psi\Big(\sum_{m=1}^{R_d}\chi_{tm}\Big)\right] \tag{8}$$

where $\Psi(\cdot)$ is the first derivative of the $\log\Gamma$ function and we have additionally introduced the variational parameter $\boldsymbol{\chi}$.

7.2. Reducing the complexity of the appearance term

Before starting and in order to shorten the notation, hereafter, we will indistinctly use f_{nz_n} instead of $f_{z_n}(h_n)$ for referring to the KLR expression in eq. (3).

With the purpose of reducing the complexity of the appearance term, the logistic function can be symmetrized as follows: $\log f(x) = -log(1+e^{-x}) =$

$\frac{x}{2} - \log(e^{x/2} + e^{-x/2})$ [35]. Then, working out (5) produces:

$$E_q[\log p(h_n|z_n, \mathbf{a})] = E_q[\log n_k] + E_q\left[(z_n - \bar{z}_n)\left(\frac{f_{nk}}{2} - \log g_{nk}\right)\right] \tag{9}$$

where $g_{nk} = e^{\frac{1}{2}f_{nk}} + e^{-\frac{1}{2}f_{nk}}$. Since g_{nk} is convex over the variable f_k^2, the last term can be lower bounded using a first-order Taylor expansion. This process involves a new variational parameter ξ and leads to the following expression:

$$\begin{aligned} E_q[\log p(h_n|z_n, \mathbf{a})] \geq \sum_{k=1}^{K} \Bigg\{ & \phi_{nk} \log n_k + \left(\phi_{nk} - \frac{1}{2}\right) f_{nk} \\ & -\frac{\xi}{2} - log(1 + e^{-\xi_{nk}}) - A(\xi_{nk})\left(f_k^2(h_n) - \xi_{nk}^2\right) \Bigg\} \end{aligned} \tag{10}$$

with $A(\xi_{nk}) = \frac{1}{4\xi_{nk}} \tanh\left(\frac{\xi_{nk}}{2}\right)$. Note that this lower bound is exact when $\xi^2 = f_k^2(h_n)$. Moreover, the regression function f is now outside the logarithm, thus allowing for a much simpler optimization.

To update the regressor, a L2-norm regularized function has to be maximized, namely:

$$L_{f_k} = \sum_{n=1}^{N_d} \sum_{k=1}^{K} C_{nk}^{(1)} f_{nk} - C_{nk}^{(2)} f_k^2(h_n) - \frac{\mu}{2}\|f\|_{\mathcal{H}_k}^2 \tag{11}$$

where the parameters C^1, C^2 are:

$$C_{nk}^{(1)} = \phi_{nk} - \frac{1}{2} \tag{12}$$

$$C_{nk}^{(2)} = \frac{1}{4\xi_{nk}} \tanh\left(\frac{\xi_{nk}}{2}\right) \tag{13}$$

Thus, in order to obtain the optimal values of the regressors $\mathbf{a}_k$, an iterative Newton-Raphson method can be used, so that at iteration t:

$$\mathbf{a}_k^{(t+1)} = \mathbf{a}_k^{(t)} - H_k^{-1}\nabla_k \tag{14}$$

where the values of the gradient ∇_k and the Hessian H_k obey:

$$\nabla_k = K_k^T C^{(1)} - 2K_k^T (C^{(2)} \circ f_k) - \frac{\mu}{2} K_k' \mathbf{a}_k \tag{15}$$

$$H_k = -2K_k^T \operatorname{diag}(C^{(2)}) K_k - \frac{\mu}{2} K_k' \tag{16}$$

where K and K' denote the data kernel matrix and the regularization matrix, respectively, and $\circ$ represents the Hadamard product (element-wise) of matrices. Finally, the normalization term is computed as:

$$n_k^{-1} = \sum_{d=1}^{D} \sum_{n=1}^{N_d} \frac{1}{1 + e^{-f_k(h_{dn})}} \tag{17}$$

7.3. Parameter updating equations

To learn the values of the model parameters, we use a variational EM approach. The updating equations that govern the variational parameters in the E-step of the proposed algorithm are:

$$\xi_{nk} = \pm f_{nk} \tag{18}$$

$$r_{tkpr} \propto c_{tk} \lambda_{pr} \exp\left[\Psi(\chi_{tp}) - \Psi\left(\sum_{m=1}^{R_d} \chi_{tm}\right)\right] \tag{19}$$

$$\chi_{tp} = \eta_p + \sum_{n=1}^{N_d} \sum_{t=1}^{K} \phi_{nk} 1[l_n = r] r_{tkpr} \tag{20}$$

$$\phi_{nk} \propto \exp\left\{\Psi(\gamma_k) + \log 1[l_n = r] L_{rk} + \xi_{nk} + \sum_{t=1}^{K} \sum_{p=1}^{R_d} 1[l_n = r] r_{tkpr} \left[\log \frac{c_{tk} \lambda_{pr}}{r_{tkpr}} + D_\Psi(\chi_{tp})\right]\right\} \tag{21}$$

$$\gamma_k = \alpha_k + \sum_{n=1}^{N_d} \phi_{nk} \tag{22}$$

with $D_\Psi(\chi_{tp}) = \Psi(\chi_{tp}) - \Psi\left(\sum_{m=1}^{R_d} \chi_{tm}\right)$.

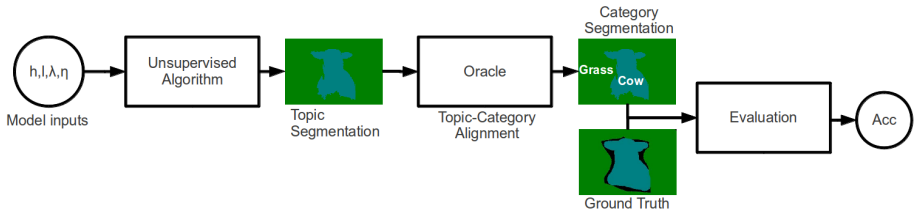
Furthermore, in the ϕ_{nk} update equation, Multinomials L_k associated with the region labels should be included only in the training phase in supervised mode.

In the M-step, the optimal values of the model parameters are computed. In particular:

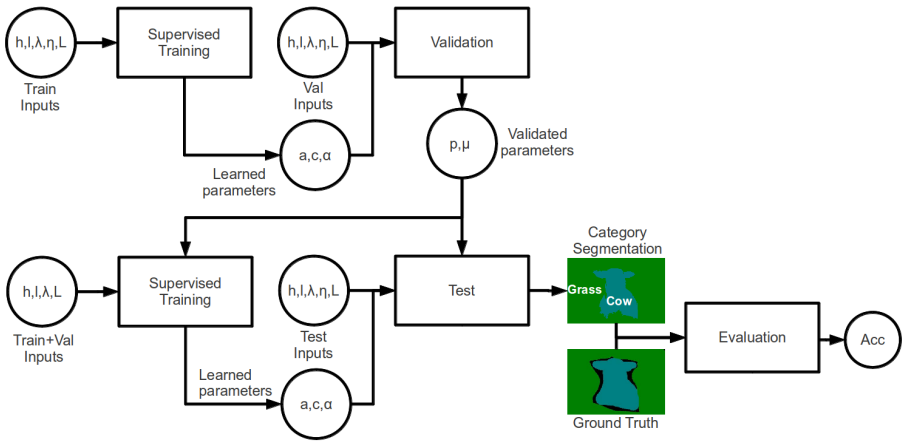
$$c_{tk} \propto \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{p=1}^{R_d} \phi_{drk} 1[l_{dn} = r] r_{tkpr} \tag{23}$$

n_k, a_k as in eq. (17) and (14), respectively.

$\boldsymbol{\alpha}$ as in the original LDA formulation [8]



(a)



(b)

Figure 7: Processing pipelines for (a) unsupervised experiments and (b) supervised experiments. In both cases, each step of the process is identified together with the involved variables (inputs, learned variables, and outputs).

Finally, a note on the convergence speed of the algorithm is in order. We have found that the convergence speed basically depends on when the KLR-based appearance model reaches a large enough number of reference points. Once this number of reference points is reached, the model expressiveness is enough to solve the problem and the segmentation accuracy stops improving. This number of reference points turns out to be quite low in the unsupervised case (about 200 reference points) and, consequently, the convergence is really fast (4-5 iterations); whereas it is higher for the supervised case (up to 1500 reference points), which requires 25-30 iterations.

8. Experimental Results

We have assessed the proposed model in two different scenarios for category-based image segmentation: unsupervised and supervised. The unsupervised scenario, in which we do not have labeled data, is the main one. The goal in this case is to produce unsupervised segmentations and to discover semantic concepts. In contrast, the supervised scenario, in which labeled data are provided to train our model, is a complementary scenario that allows us to highlight that the proposed method is capable of producing competitive results (with respect to state-of-the-art methods) in both scenarios and, consequently, is an outstanding alternative for scene analysis tasks where different degrees of supervision are present.

In any case, although not described in the paper, we have used location prior distributions as in [2].

8.1. Unsupervised category-based image segmentation and topic discovery

These experiments have been conducted on the Microsoft Cambridge Segmentation database (MSRC) [36]. MSRC database contains 591 images of 23 object classes, two of which ('horse' and 'mountain') have been removed from the evaluation due to their low number of positives, as suggested in the evaluation protocol proposed in [36]. Obviously, each image may contain more than one class, and there is a significant degree of both intra-class variation and inter-class overlapping.

In this case we aim to unsupervisely produce category-based segmentations or, in other words, to unsupervisely detect topics that correspond to semantic categories in the database. The interested reader is referred to [26] for an excellent survey that compares the performance of several methods in a topic discovery task, for which latent topic models become one of the most prevalent approaches. Although the underlying task is essentially the same, our assessment method is different since we are not only interested in topic detection, but also in category-based image segmentation. In particular, given a topic discovery task in a multiclass problem, Tuytelaars et al. evaluate a set of independent

binary detection systems, whereas we focus on how good the category-based segmentation is (each image contains one or more categories and each pixel in the image belongs to one and just one semantic category).

The processing pipeline used in this experiment is illustrated in Fig. 7(a), where every step of the process is identified together with the variables involved. In the unsupervised mode, our algorithm uses input variables (observed variables in the model) in order to produce topic-based segmentations. Since the algorithm is actually not aware of the categories in the database and it simply produces image segmentations according to the latent topics, we need to use an oracle that provides an association of each topic with the most likely class in the database. Obviously, this alignment is one-to-one so that every topic represents one and only one category in the dataset. From our point of view, this approach produces a more realistic assessment since the assignment of more than one topic to the same semantic category is actually penalized, what did not happen in [26]. Once this previous alignment has been generated, the segmentation accuracy can be measured in the same manner as in a supervised mode.

In the conducted experiments we have compared our generative model with several state-of-the-art latent topic models found in the computer vision literature, and with two versions of our proposal where some components have been removed:

a) **LDA** [8], as the baseline topic model;

b) **SP-LTM** [19], that introduces segmentations considering independent regions;

c) **MRF-LDA** [23], that extends the previous one by modeling an intra-class inter-region MRF-based geometric context.

d) **Proposed-Mult**, in which the KLR-based appearance model has been substituted by a simple multiplicative model with multinomial distributions;

e) **Proposed-w/o Context**, in which we have removed the context model.

The compared algorithms were run over the whole MSRC dataset (591 im-

Table 1: Experimental results for unsupervised category-based image segmentation on the MSRC database.

Algorithm	Overall Accuracy
LDA	24.36%
SP-LTM	24.56%
MRF-LDA	25.41%
Proposed-Mult	26.04%
Proposed-w/o Context	30.42%
Proposed	**33.25**%

ages). Table 1 shows the segmentation accuracy achieved by all of them. As can be seen, our proposal clearly outperforms the rest of the latent topic models: 31% improvement with respect to MRF-LDA, which provides the best reference result. In order to analyse how the proposed context and appearance models contribute to the performance of the proposed method, we have compared the proposed method performance to that of incomplete versions of it. The larger part of the performance improvement is due to the KLR-based appearance model (an improvement of 27.7% when comparing 'Proposed' to 'Proposed-Mult'), but there is also a relevant improvement coming from the inclusion of the context model (9.3% when comparing 'Proposed' to 'Proposed-w/o Context'). In summary, the proposed appearance model provides much more expressiveness than the multinomial distributions used in the rest of the latent topic models, and the inter-region inter-topic context model clearly outperforms previous context models (as the one proposed in MRF-LDA).

In order to gain more insight into the manner the proposed method works, it is worthwhile to discuss its strengths and weaknesses. To this end, a selection of both good and wrong visual segmentation results are provided in Fig. 8. Examples of good segmentations are given in the first two rows, while wrong segmentations are given in the last two rows. As can be observed, in general, the proposed method tended to assign just one or two topics to the whole image,

Figure 8: An illustration of visual category-based segmentation results achieved by our proposal in unsupervised mode on the MSRC database. Top rows: correctly segmented samples. Bottom rows: segmentation errors

thus producing segmentations in which the main category was usually spread along the whole image. This resulted in good segmentations for images showing either just one semantic category (see 'flower' or 'sign' examples in the top row) or various large objects belonging to different categories (see 'tree-water' example in the top row).

On the other hand, we find two main causes of error that allows us to explain most of the cases where the proposed algorithm achieved poor performance; namely: a) object-oriented categories were absorbed by other scene-oriented categories that tend to appear in their surroundings (see the first three examples in the last rows of the figure, where 'sign', 'cows' and 'bicycles' have been absorbed by 'sky', 'grass' and 'road', respectively); and b) the same topic was associated with two 'visually similar' classes (e.g. 'dog/cow' and 'sign/book' in the last row of the figure).

Therefore, we can conclude that the unsupervised version of our latent topic model successfully discovers semantic categories. Likewise, it generates segmentations and categorizations. The category-based image segmentations produced

by the proposed method turn out to be better for large regions (in both scene- or object-oriented categories) than for small instances of object-oriented categories. This fact is a direct consequence of the clustering property inherent to the latent topic models, which tend to divide the data space into equally-sized topics, thus favoring topics that occupy large regions.

8.2. Supervised category-based image segmentation

The experiments on supervised category-based image segmentation have been conducted on two different databases: MSRC, described in the previous subsection, and PASCAL VOC 2010 Segmentation Database [37].

PASCAL VOC Segmentation is a challenging segmentation dataset with 20 object categories. In order to provide a meaningful assessment of every model element, we have used the *segmentation 'trainval' set*, divided into a 'train set' (964 images) for training and validation, and a 'val set' (964 images) for test.

It should be noticed that *segmentation accuracy* is computed differently depending on the database: for the MSRC database, it is computed as the percentage of pixels correctly classified within the considered 21 class labels. Therefore, pixels belonging either to the 2 discarded classes or to the non-defined class (see black pixels in Fig. 1(c), where the black regions are those belonging to the non-defined class) are not taken into account. For the PASCAL VOC database, in contrast, background pixels are also taken into account. Furthermore, whereas MSRC considers only a global accuracy measure, in PASCAL individual accuracies for each class are firstly computed and then averaged to provide the global measure.

It is worth mentioning that following other approaches, such as [38], we have also included the outputs of SVM-based classifiers using a Spatial Pyramid Representation of images [39].

As in the unsupervised experiments, several different versions of our proposal were included in the evaluation to provide some insight into the performance improvement that comes from each of the proposed extensions. In addition to the ones considered in the unsupervised environments, a version that follows a

Table 2: Experimental results on the MSRC database.

Algorithm	**Accuracy**	**Algorithm**	**Accuracy**
TextonBoost [2]	72%	Yang et al. [40]	75%
Auto-context [41]	75%	Zhang et al. [42]	75%
Verbeek et al. [24]	74%	Krähenbühl et al. [43]	**86**%
Ladickỳ et al. [44]	**87**%	Proposed-Mult	56%
Proposed-w/o Context	81%	Proposed-Hard	84%
Proposed	**85**%		

hard-labeling strategy ('Proposed-Hard').

8.2.1. Results on the MSRC database

For the MSRC database, the evaluation procedure follows the one described in [36], which divides the complete dataset into a train set (276 images), a validation set (59 images), and a test set (256 images).

The complete processing pipeline is illustrated in Fig. 7(b), where we have identified the variables learned in each step of the process. We have used the validation set to optimize the values of several parameters that cannot be automatically optimized; namely: the weights p_h of the weighted kernel strategy described in Section 5.3; and the regularization term μ in the KLR. To that end, we have followed a two-step approach: 1) For each combination of the parameters $(\mathbf{p}, \mu)$, where $\mathbf{p}$ is a vector grouping the weights p_h, we trained our generative models on the train set to learn parameters of the model ($\mathbf{a}$, $\mathbf{c}$, α) and obtain the corresponding segmentation accuracies on the validation set; and 2) keeping the parameter values that led to the highest accuracy on the validation set, we trained the final model using the train+validation set and obtained the segmentation accuracy on the test set. As a result of this process, the weights of the weighted kernel were set to $p^t = 1.00$ (texture), $p^c = 0.25$ (color), and $p^s = 0.01$ (shape); and μ was set to $\mu = \frac{Nr}{1000}$, where Nr is the total number of regions in the training dataset.

Figure 9: Some illustrative examples of segmentation outputs in MSRC provided by our proposal in supervised mode.

Table 2 shows the segmentation accuracy achieved by the compared algorithms on the MSRC dataset. In particular, we compare the proposed method to several state-of-the-art methods for which results have been reported on the same database and following the same evaluation protocol. As can be observed, although all the state-of-the-art references are discriminative (CRFs or SVMs), the proposed method achieved reasonably competitive results when compared to them: only two of the compared methods provides slightly better results. These results demonstrates that generative models are able to reach the same performance level than discriminative approaches when ground-truth pixel-wise segmentations are used in the training phase.

In what concerns to our proposal, it is easy to notice how the KLR-based appearance model is in charge of most of the improvement, emphasizing the relevance of this subsystem. Furthermore, an in-depth analysis of the results allows us to draw the following conclusions: a) the soft-labeling improves the hard-labeling approach by 1.2% ('Proposed' vs. 'Proposed-Hard'), what is a nice consequence of giving some degree of freedom to the topics in order to keep their latent nature unaltered; and b) including context information (again) provides a notable improvement of the results (up to 5%).

Some illustrative visual examples in which our proposal achieved good performance are shown in Fig. 9. As can be seen, the algorithm not only provided

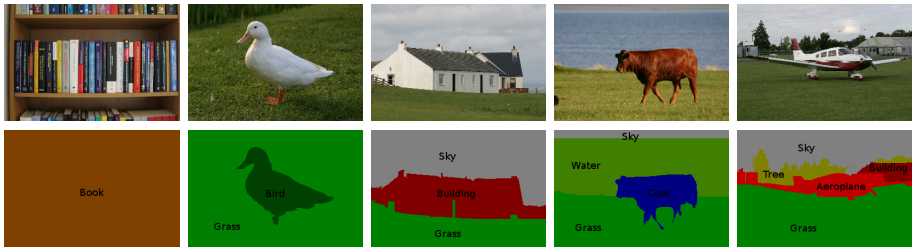


Figure 10: Highest ranked images corresponding to scenes containing just 1, 2, 3, 4 or 5 semantic categories. These are the images for which our proposal is more confident.

suitable image partitions, but also successfully assigned each region to its associated semantic category.

Furthermore, in order to provide a more realistic visual assessment, we selected those images with the highest levels of confidence, that is, those ones with the highest values of the lower bound of the posterior in eq. (7); however, we found that these maximum values were always detected in images showing only one category. Then, we repeated the experiment aiming to find the highest ranked images for scenes containing just 1, 2, 3, 4 or 5 semantic categories. The results are shown in Fig. 10. As can be seen, the algorithm provided proper segmentations for almost all cases. In general, segmentation errors were either associated with small regions that were absorbed by larger regions in their surroundings (e.g. the legs of the 'cow' in the third example have been absorbed by the 'grass' region), or to confusable objects that appeared together in the image (e.g. 'aeroplane' and 'building' in the fifth example).

8.2.2. Results on the PASCAL VOC database

For the PASCAL VOC dataset, following the approach described in [43], we have also included, as additional features, the responses provided by the bounding box object detectors [45] for each object class (with a weight $p^o = 0.10$). Furthermore, we have used the same parameter values validated for MSRC.

Regarding the results on this dataset, we have compared our proposal with

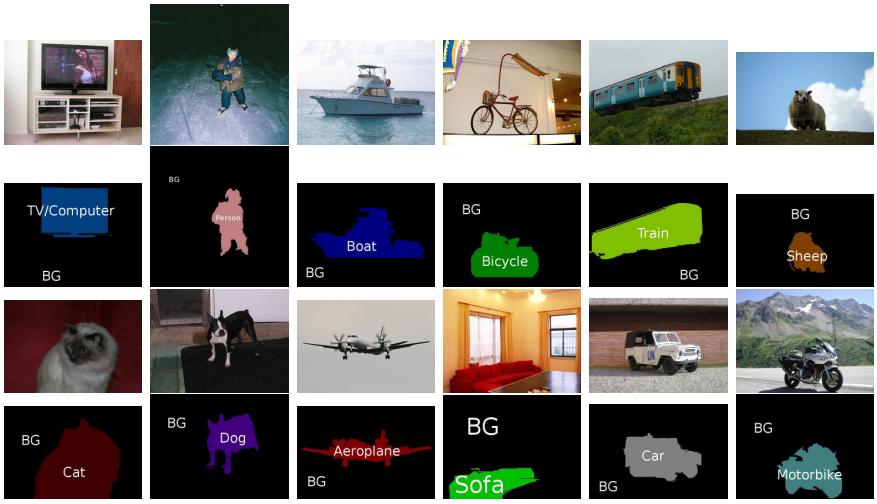


Figure 11: Some illustrative examples of segmentation outputs in PASCAL VOC 2010 provided by our proposal in supervised mode.

the state-of-the-art method proposed in [43] where the authors reported results using the same 'trainval' dataset. For this particular dataset, we have obtained a Segmentation Accuracy of **32.7**%, which compares favorably with the 30.2% reported in [43]. In addition, some visual examples are provided in Fig. 11.

9. Discussion

In this paper we have presented a latent topic model for category-based image segmentation. Two are the main contributions of the model with respect to the state-of-the-art in the latent topic literature: 1) an inter-topic inter-region context model that successfully takes into account the spatial neighborhood of a region to decide which topic is the most appropriate for that region; and 2) a novel KLR-based appearance distribution that allows for considering the non-linear relations among local descriptors within the same region, while keeping the computational complexity low enough to reach a practical solution. Furthermore, it is worth emphasizing how these contributions have been designed within an unified inference framework, what is not easily found in systems alike.

In addition to these two contributions, a set of extensions of the model have been proposed to allow it to work in supervised mode. Some of these extensions are related to the KLR-based appearance (handling negative samples during training), and other to a soft-labeling strategy that keeps unaltered the latent nature of topics. This is in itself a valuable contribution since the proposed model is able to work in both unsupervised and supervised modes, in contrast to other (usually discriminative) alternatives.

All the contributions have been experimentally assessed in both unsupervised and supervised category-based image segmentation tasks. We have also shown to what extent each specific proposal contributes to the whole system performance. Furthermore, our experimental results prove that the algorithm not only outperforms several techniques found in the latent topic literature, but also compares reasonably well to discriminative state-of-the-art methods in a supervised scenario.

10. Acknowledgments

This work has been partially supported by the project AFICUS, co-funded by the Spanish Ministry of Industry, Trade and Tourism, and the European Fund for Regional Development, with Ref.: TSI-020110-2009-103, and the National Grant TEC2011-26807 of the Spanish Ministry of Science and Innovation

References

[1] C. Galleguillos, A. Rabinovich, S. Belongie, Object categorization using co-occurrence, location and appearance, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, Anchorage, Alaska, USA, pp. 1–8.

[2] J. Shotton, J. Winn, C. Rother, A. Criminisi, Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context, International Journal of Computer Vision 81 (2009) 2–23.

[3] S. Gould, T. Gao, D. Koller, Region-based segmentation and object detection, in: Advances in Neural Information Processing Systems 22, 2009, Vancouver, B.C., Canada, pp. 655–663.

[4] M. Kim, Large margin cost-sensitive learning of conditional random fields, Pattern Recognition 43 (2010) 3683 – 3692.

[5] D. Larlus, J. Verbeek, F. Jurie, Category level object segmentation by combining bag-of-words models with dirichlet processes and random fields, International Journal of Computer Vision 88 (2010) 238–253.

[6] Y. J. Lee, K. Grauman, Collect-cut: Segmentation with top-down cues discovered in multi-object images, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, San Francisco, USA, pp. 3185 –3192.

[7] T. Hofmann, Unsupervised learning by Probabilistic Latent Semantic Analysis, Machine Learning 42 (2001) 177–196.

[8] D. M. Blei, A. Y. Ng, M. I. Jordan, J. Lafferty, Latent Dirichlet Allocation, Journal of Machine Learning Research 3 (2003) 993–1022.

[9] D. Blei, J. McAuliffe, Supervised topic models, in: Advances in Neural Information Processing Systems 20, 2007, MIT Press, Vancouver, B.C., Canada, 2007, pp. 121–128.

[10] I. González-Díaz, D. García-García, F. Díaz-de María, A spatially aware generative model for image classification, topic discovery and segmentation, in: International Conference on Image Processing, 2009, Cairo, Egypt, pp. 781–784.

[11] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, W. T. Freeman, Discovering objects and their location in images, in: IEEE International Conference on Computer Vision, 2005, volume 1, San Diego, CA, USA, pp. 370–377.

[12] R. Fergus, L. Fei-Fei, P. Perona, A. Zisserman, Learning object categories from google's image search, in: IEEE International Conference on Computer Vision, 2005, volume 2, San Diego, CA, USA, pp. 1816 –1823.

[13] D. Liu, T. Chen, Semantic-shift for unsupervised object detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006, Washington, DC, USA, pp. 16–22.

[14] X. Wang, E. Grimson, Spatial Latent Dirichlet Allocation, in: Advances in Neural Information Processing Systems, 2007, volume 20, Vancouver, B.C., Canada, pp. 1577–1584.

[15] A. Bosch, A. Zisserman, X. Muoz, Scene classification using a hybrid generative/discriminative approach, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2008) 712–727.

[16] L. Lin, T. Wu, J. Porway, Z. Xu, A stochastic graph grammar for compositional object representation and recognition, Pattern Recognition 42 (2009) 1297 – 1307.

[17] R. Zhang, Z. Zhang, Hidden semantic concept discovery in region based image retrieval, in: IEEE Conference on Computer Vision and Pattern Recognition, 2004, volume 2, Washington, DC, USA, pp. 996 – 1001.

[18] B. Russell, W. Freeman, A. Efros, J. Sivic, A. Zisserman, Using multiple segmentations to discover objects and their extent in image collections, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006, volume 2, New York, NY, USA, pp. 1605 – 1614.

[19] L. Cao, L. Fei-Fei, Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes, in: International Conference on Computer Vision, 2007, Rio de Janeiro, Brazil, pp. 1–8.

[20] C. Wang, D. M. Blei, F. fei Li, Simultaneous image classification and annotation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, Miami, FL, USA, pp. 1903–1910.

[21] L.-J. Li, R. Socher, L. Fei-Fei, Towards total scene understanding:classification, annotation and segmentation in an automatic framework,

in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, Miami, FL, USA, pp. 2036–2043.

[22] D. Larlus, F. Jurie, Combining Appearance Models and Markov Random Fields for Category Level Object Segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, Anchorage, Alaska, USA, pp. 1–7.

[23] B. Zhao, L. Fei-Fei, E. P. Xing, Image segmentation with topic random field, in: European Conference on Computer Vision, 2010, Crete, Greece, pp. 785–798.

[24] J. Verbeek, B. Triggs, Region classification with markov field aspect models, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007, Minneapolis, Minnesota, USA, pp. 1 –8.

[25] P. F. Felzenszwalb, D. P. Huttenlocher, Efficient graph-based image segmentation, International Journal of Computer Vision 59 (2004) 167–181.

[26] T. Tuytelaars, C. Lampert, M. Blaschko, W. Buntine, Unsupervised object discovery: A comparison, International Journal of Computer Vision 88 (2010) 284–302.

[27] J. van de Weijer, C. Schmid, Coloring local feature extraction, in: European Conference on Computer Vision, 2006, Graz, Austria, pp. 334–348.

[28] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2004) 91–110.

[29] H. Freeman, On the encoding of arbitrary geometric configurations, Institute of Radio Engineers, Transactions on Electronic Computers 10 (1961) 260–268.

[30] J. Zhu, T. Hastie, Kernel logistic regression and the import vector machine, in: Advances in Neural Information Processing Systems 14, 2001, Vancouver, BC, Canada, pp. 1081–1088.

[31] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (1995) 273–297.

[32] F. R. Bach, G. R. G. Lanckriet, M. I. Jordan, Multiple kernel learning, conic duality, and the smo algorithm, in: International Conference on Machine Learning, 2004, Banff, Alberta, Canada, pp. 6–13.

[33] A. J. Smola, B. Schökopf, Sparse greedy matrix approximation for machine learning, in: International Conference on Machine Learning, 2000, Stanford, CA, USA, pp. 911–918.

[34] J. Lafferty, X. Zhu, Y. Liu, Kernel conditional random fields: representation and clique selection, in: International conference on Machine learning, 2004, Banff, Alberta, Canada, pp. 64–71.

[35] T. S. Jaakkola, M. I. Jordan, Bayesian parameter estimation via variational methods, Statistics and Computing 10 (2000) 25–37.

[36] J. Shotton, J. Winn, C. Rother, A. Criminisi, TextonBoost: Joint Appearance, Shape and Context Modeling for Multi-Class Object Recognition and Segmentation, in: European Conference on Computer Vision, 2006, Graz, Austria, pp. 1–15.

[37] M. Everingham, L. Gool, C. K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, International Journal of Computer Vision 88 (2010) 303–338.

[38] G. Csurka, F. Perronnin, A simple high performance approach to semantic segmentation., in: British Machine Vision Conference, 2008, Leeds, UK, pp. 22.1–22.10.

[39] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006, volume 2, New York, NY, USA, pp. 2169–2178.

[40] L. Yang, P. Meer, D. Foran, Multiple class segmentation using a unified framework over mean-shift patches, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007, Minneapolis, Minnesota, USA, pp. 1 –8.

[41] Z. Tu, X. Bai, Auto-context and its application to high-level vision tasks and 3d brain image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (2010) 1744–1757.

[42] L. Zhang, Q. Ji, Image segmentation with a unified graphical model, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (2010) 1406–1425.

[43] P. Krähenbühl, V. Koltun, Efficient inference in fully connected crfs with gaussian edge potentials, in: Advances in Neural Information Processing Systems 24, 2011, Sierra Nevada,Spain, pp. 109–117.

[44] L. Ladicky, C. Russell, P. Kohli, P. Torr, Inference methods for crfs with co-occurrence statistics, International Journal in Computer Vision, [Preprint] (2012).

[45] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Transactions on Pattern Analyisis and Machine Intelligence 32 (2010) 1627–1645.