Characterizing Layouts of Outdoor Scenes Using Spatial Topic Processes

Dahua Lin
TTI Chicago
dhlin@ttic.edu

Jianxiong Xiao
Princeton University
xj@princeton.edu

Abstract

In this paper, we develop a generative model to describe the layouts of outdoor scenes – the spatial configuration of regions. Specifically, the layout of an image is represented as a composite of regions, each associated with a semantic topic. At the heart of this model is a novel stochastic process called Spatial Topic Process, which generates a spatial map of topics from a set of coupled Gaussian processes, thus allowing the distributions of topics to vary continuously across the image plane. A key aspect that distinguishes this model from previous ones consists in its capability of capturing dependencies across both locations and topics while allowing substantial variations in the layouts. We demonstrate the practical utility of the proposed model by testing it on scene classification, semantic segmentation, and layout hallucination.

1. Introduction

Spatial configurations of regions, often referred to as the *layouts*, are a key to scene understanding. As illustrated in Figure 1, layouts convey significant information for both semantic interpretation (*e.g.* scene classification) and low-level tasks (*e.g.* segmentation). Therefore, a good model of layouts is of fundamental importance. Our primary goal here is to develop a layout model that can capture the common structures of outdoor scenes while allowing flexible variations. We believe that such a model not only advances the frontier of traditional problems, but also creates opportunities for new applications (*e.g.* scene extrapolation).

With this goal in mind, we revisit previous approaches, which roughly fall into three categories: *descriptive*, *discriminative*, and *generative*. Descriptive methods, including GIST [17] and Spatial Pyramid matching [10, 13] describe a scene through a holistic representation. In spite of their simplicity, these methods work very well in scene classification. However, they lack the capability of expressing fine grained variations, which are useful in other tasks such as segmentation and annotation. Recently, nonparametric methods using dense map of local descriptors, such as SIFT-Flow [15] and SuperParsing [24], has gained a lot of interest. While they are generally more expressive, the reliance on large set of examples leads to limited generalizability and substantially increased computational cost.

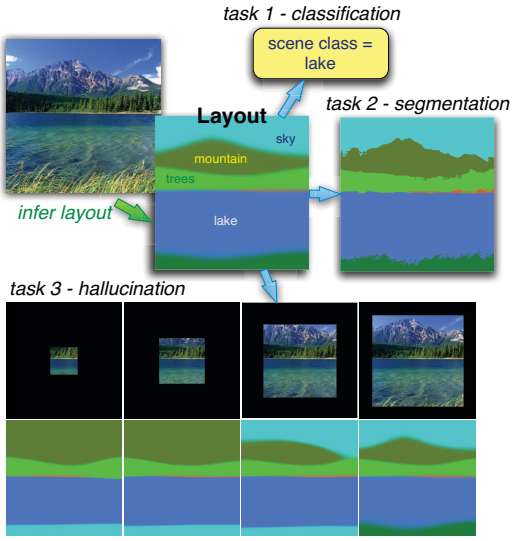


Figure 1. This figure shows an image of a lake and its layout (using different colors for different regions). In this work, we develop a generative model of layouts, which can be used in various vision tasks, including scene classification and semantic segmentation. Moreover, leveraging the scene structures captured by the model, one can extrapolate the scenes beyond the visible scope.

Discriminative models [9,11,25,28] are often formulated as Conditional Random Fields (CRFs). Instead of modeling layouts explicitly, these models typically utilize spatial relations via potentials that couple semantic labels at different sites. Another popular approach is to consider a scene as composed of deformable parts [18, 19], and uses Latent SVM [7] to train the model. Discriminative methods are devised for a specific task (usually classification or labeling), and do not provide a generic characterization of layouts, which is our major goal.

Generative models, unlike discriminative ones, often resort to hierarchical Bayesian models to describe a scene [1, 5, 14, 23]. Taking advantage of the flexibility of graphi-

cal models, they are able to express various relations in a complex scene, such as the ones between a scene and its parts [1, 23] and those between concurrent objects [5, 14]. Since the introduction of Latent Dirichlet Allocation [3] to scene categorization [6], topic models have also been widely used in scene understanding [4, 16, 19, 26, 29]. Whereas some of them take into account spatial relations, their treatment is often simplified – focusing only on pairwise relations between objects or making assumptions that ignore important spatial dependencies. Hence, the resultant models are generally not the most appropriate choices for characterizing the layouts of outdoor scenes.

Towards the goal of providing an effective layout model, we develop the *Spatial topic process*, a new formulation that builds upon topic models and goes beyond by allowing distributions of topics to vary continuously across the image plane. Specifically, to capture the statistical dependencies across both spatial locations and visual categories, we introduce a set of Gaussian processes (GPs) to generate a map of topic distributions. These GPs are coupled via a latent representation that encodes the global scene structure. This model provides a rich representation that can express layout variations through pixel-dependent topic distributions, and on the other hand ensures both local coherence and global structural consistency via the use of coupled GPs.

This new layout model is useful for a variety of vision problems. We demonstrate its practical utility on three applications: (1) scene classification using the layout representation, (2) semantic segmentation based on spatially varying topic distributions, and (3) layout hallucination, a task trying to extrapolate beyond the visible part of a scene.

2. Related Work

This work is related to several models developed in recent years that try to incorporate spatial relations into topic models. Wang and Grimson proposed *Spatial LDA* [26], where each pixel is assigned a topic chosen from a local document. This model enables spatial variation of topics, but ignores the dependencies between topic assignments by assuming that they are independently chosen. The *Topic Random Field* proposed by Zhao *et al*. [29] goes one step further by introducing an MRF to encourage coherent topic assignment. However, such local regularization techniques do not capture long range correlation, which is crucial to modeling global layouts. Recently, Parizi *et al*. [19] proposed a *reconfigurable model* for scene recognition, which treats a scene as a composite of a fixed number of rectangular regions, each governed by a topic. While allowing flexible topic-region association, it does not take into account the dependencies between topic assignments either.

There has been other work that combines latent GPs for spatially coherent segmentation [8, 21, 22]. Sudderth and Jordan [22] proposed a formulation of *dependent Pitman-Yor processes (DPY)*, where spatial dependencies are induced via thresholded GPs. It is, however, important to note that there is a fundamental aspect that distinguishes our work from this paper: we aim to learn a generative model that is able to capture the prior structure of outdoor scenes, such that one can sample new scenes from it or infers missing parts of a scene. Their work focuses on producing accurate segmentation instead of learning the underlying structures of scenes.

To sum up, two key aspects distinguish our work: (1) the aim to a learn a *prior* model of layouts that captures common structures; (2) a novel design coupling GPs across layers, which leads to not only the capability of capturing *long range spatial dependencies* and *cross-topic relations*, but also a vector representation that expresses the global structure in a compact yet flexible way.

3. Generative Model of Layouts

Following the paradigm of topic models, we characterize an image by a set of visual worlds: $S = \{(x_i, y_i, w_i)\}_{i=1}^n$. Here, x_i and y_i are the pixel coordinates of the i-th visual word, and w_i is the quantized label. We aim to develop a generative model to explain the spatial configuration of S. As shown in Figure 2, This model considers a scene as a composition of several regions, each associated with a topic, *i.e*. a distribution of visual words. Each pixel location (x_i, y_i) is attached an indicator z_i that assigns it to a particular region. Given z_i, one can draw the visual word w_i from the corresponding topic.

Generally, the distribution of z_i, which we denote by θ_i, is location-dependent. More importantly, the values of θ_i at different locations are strongly correlated. For example, a pixel is very likely in an ocean if so are the surrounding pixels. Therefore, it is desirable to jointly model the distributions of z_i over the entire image so as to capture the correlations between them. In particular, we develop a probabilistic model called *Spatial Topic Process* that can generate a continuous map of topic distributions based on a set of coupled Gaussian processes.

Given $\boldsymbol{\lambda}$, the parameters of the Spatial Topic Process, and $\beta = (\beta_1, \ldots, \beta_K)$, a set of word distributions, each corresponding to a topic, the overall procedure to generate S is summarized below. **(1)** Generate a continuous map of topic distributions as $\theta \sim \theta|\boldsymbol{\lambda}$. Here, $\theta(x, y)$ is the predicted distribution of topics at (x, y). **(2)** Randomly sample a set of locations $\{(x_i, y_i)\}_{i=1}^n$. While one can use all pixels of an image, we will show empirically that this is unnecessary, and a much smaller subset usually suffices. **(3)** Draw the topic indicator z_i at each location from $\theta_i \triangleq \Theta(x_i, y_i)$. **(4)** Draw the visual word w_i from the corresponding topic β_{z_i}.

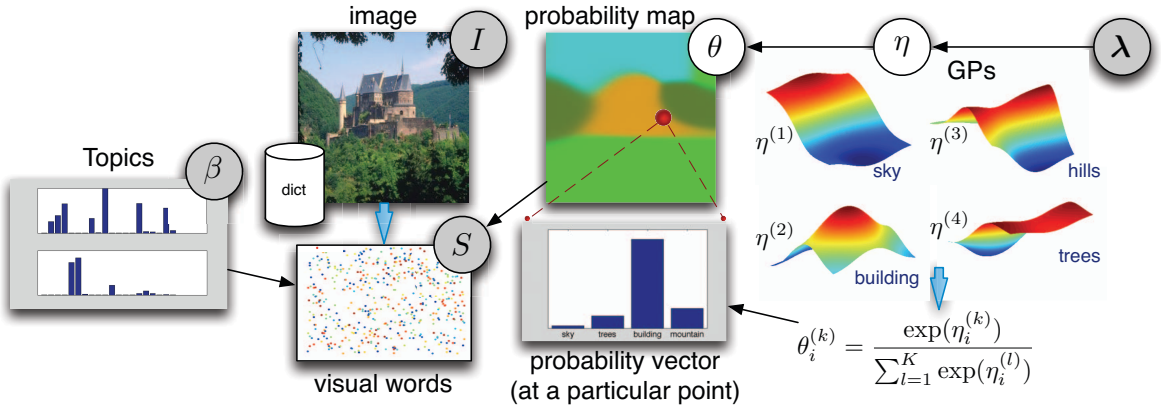


Figure 2. This model generates a set of visual words (with coordinates) for each image. First, K real-valued smooth maps η are generated from a set of coupled Gaussian processes (with parameter $\boldsymbol{\lambda}$), where each map corresponds to a topic. Then a continuous map of probability vectors θ is derived by applying the softmax function at each point. Finally, at each sample point (x_i, y_i), a topic is chosen according to θ_i, and then a visual word w_i is drawn from the word distribution of the corresponding topic, *i.e.* β_{z_i}.

3.1. Spatial Topic Process

We derive the *Spatial Topic Process*, a stochastic process that can generate spatially continuous maps of discrete distributions, based on GPs. To begin with, we first consider a simpler problem – devising a joint distribution to incorporate correlations between a finite set of discrete distributions $\theta_1, \ldots, \theta_n$. This can be accomplished by mapping them to real vectors that are jointly Gaussian distributed, as

$$(\eta_1^{(k)}, \ldots, \eta_n^{(k)}) \sim \mathcal{N}(\boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}), \quad k = 1, \ldots, L; \quad (1)$$

$$\theta_i^{(k)} = \exp(\eta_i^{(k)}) / \sum_{l=1}^{K} \exp(\eta_i^{(l)}). \quad (2)$$

Here, Eq.(2) converts K real values into a probability vector by *softmax*. The Gaussian distribution at Eq.(1) captures the correlation between the underlying real values, thus implicitly introducing statistical dependency between the probability vectors. As we shall see, the use of softmax here makes it a convex optimization problem to estimate η.

By further extending the finite dimensional Gaussian distributions in Eq.(1) to Gaussian processes, we obtain a stochastic process as follows, which can generate continuous maps of discrete distributions.

$$\eta^{(k)} \sim \mathrm{GP}(\mu^{(k)}, \Sigma^{(k)}); \quad (3)$$

$$\theta^{(k)}(x, y) = \exp(\eta^{(k)}(x, y)) / \sum_{l=1}^{k} \exp(\eta^{(l)}(x, y)). \quad (4)$$

Here, $\eta^{(k)} : \Omega \to \mathbb{R}$ is a real-valued function defined on the image plane Ω. $\boldsymbol{\mu}^{(k)}$ and $\boldsymbol{\Sigma}^{(k)}$ are respectively extended to a mean function and a covariance function. In particular, $\mu^{(k)}(x)$ is the mean of $\eta^{(k)}(x, y)$, and $\Sigma^{(k)}((x, y), (x', y'))$ is the covariance between $\eta^{(k)}(x, y)$ and $\eta^{(k)}(x', y')$.

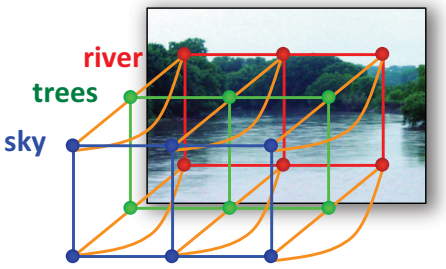


Figure 3. This figure illustrates part of the Gaussian MRF for coupling GPs. Within each topic are links between values at neighboring grid points. There are also links (depicted in orange color) between values for different topics at corresponding grid points.

3.2. Coupling Grid-based Gaussian Processes

Gaussian processes are often formulated in a translation invariant form in practical use. For example, the radial basis function is a popular choice to define the covariance. However, this is not an appropriate design in the context of scene layout modeling, where both the mean and the variance are location dependent. Here, we propose a more flexible way to define the mean and covariance functions.

We first define a Gaussian distribution over a finite grid and then extend it to a Gaussian process via smooth interpolation. Let $s_1, \ldots, s_m$ be a set of grid points, and $(g_1, \ldots, g_m) \sim \mathcal{N}(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ be jointly Gaussian distributed variables, each for a grid point. Then we can extend this finite vector to a continuous map η over the image domain Ω as follows. Let $v = (x, y) \in \Omega$, then

$$\eta(v) = \sum_{j=1}^{m} c_j(v) g_j, \quad c_j(v) = w_j(v) / \sum_{j'=1}^{m} w_{j'}(v). \quad (5)$$

Here, $w_j(v) = \exp(-d(v, s_j)^2/\sigma_g^2)$ is a weight value that reflects the influence of the j-th seed to v. It is not difficult to show that η as defined above is a Gaussian process. In particular, the mean function is $\mu(v) = \mathbf{c}_v^T \boldsymbol{\mu}_g$ and the covariance function is $\Sigma(u, v) = \mathbf{c}_u^T \boldsymbol{\Sigma}_g \mathbf{c}_v$. Here, $\mathbf{c}_v = (c_1(v), \ldots, c_m(v))$.

Eq.(3) and (4) introduce K Gaussian processes – each can be characterized by a finite dimensional Gaussian distributions using the grid-based parametrization as above. In this design, the smooth interpolation as in Eq.(5) ensures local coherence, while Gaussian distributions over the grid capture long range spatial relations. Also, it is important to note that spatial configurations of different visual categories (*i.e*. topics) are also related. For example, a river often comes with hills along its side. To capture such relations, it is desirable to further couple all GPs – this can be achieved through a joint distributions over the grid values for all topics. Let m be the number of grid points, then the joint dimension is $d_g = mK$. Empirical testings showed that a 6-by-6 grid suffices to express most variations in the layout of natural scenes, and regions roughly fall into 20 to 30 categories (*e.g*. skys, trees, and sea). Thus the joint dimension d_g is about 1000.

Reliable estimation of such a Gaussian distribution requires a very large number of samples if a full covariance matrix is used. To address this difficulty, we consider a Gaussian Markov random field (GMRF) as below:

$$p(\mathbf{g}|\boldsymbol{\lambda}) \propto \exp\left(E_{int} + E_{ext}\right); \tag{6}$$

$$E_{int} = \sum_{i,k} \lambda_i^{(k)} (g_i^{(k)})^2 + \sum_{i,j,k\,:\, i\sim j} \lambda_{i,j}^{(k)} g_i^{(k)} g_j^{(k)}; \tag{7}$$

$$E_{ext} = \sum_{i,k,l} \lambda^{(k,l)} g_i^{(k)} g_i^{(l)}. \tag{8}$$

Here, $\mathbf{g}$ is an mK-dimensional vector that contains all values at grid points, which we call the *latent layout representation*, and $g_i^{(k)}$ is the value for the k-th topic at the i-th grid point. $\boldsymbol{\lambda}$ is the parameter vector. As shown in Figure 3, this GMRF comprises two types of links: the ones between values for the same topic at neighboring sites ($i \sim j$ indicates i and j are neighbors), and those between values for different topics at the same site. In this way, the parameter dimension is substantially reduced.

3.3. Joint Formulation

Combining all components, we derive the joint formulation as follows. Suppose there are n visual words from an image. Given β (the word distributions) and $\boldsymbol{\lambda}$ (the parameter of the Spatial Topic Process), the joint probability of these visual words and their associated topic indicators is

$$p(\mathbf{g}|\boldsymbol{\lambda}) \prod_{j=1}^{n} p(z_j|x_j, y_j, \mathbf{g}) p(w_j|z_j; \beta). \tag{9}$$

Here, $p(\mathbf{g}|\boldsymbol{\lambda})$, which is defined in Eq.(6), is the prior of the latent layout representation. $p(z_j|x_j, y_j, \mathbf{g})$ is the topic probability at (x_j, y_j), which is defined by Eq.(4) as

$$p(z_j = k|x_j, y_j) = \theta^{(k)}(x_j, y_j) \propto \exp(\eta^{(k)}(x, y)). \tag{10}$$

Here $\eta^{(k)}$ is determined by $\mathbf{g}$ as in Eq.(5). $p(w_j|z_j; \beta)$ is the probability of choosing visual word w_j from the topic β_{z_j}.

4. Inference and Learning Algorithms

This section presents algorithms to infer layouts of images and to learn model parameters.

4.1. Inferring Layouts

Given the model parameters, including $\boldsymbol{\lambda}$ and β, we can derive the latent layout representation $\mathbf{g}$ of a new image as follows. Specifically, we first extract a set of local features, and quantize them into visual words. Each word is represented by a triple (x_j, y_j, w_j), and is associated with a hidden variable z_j that assigns it to a topic. Then the MAP estimator, which we denote by $\hat{\mathbf{g}}$, is given by

$$\hat{\mathbf{g}} = \operatorname*{argmax}_{\mathbf{g}} p(\mathbf{g}|\boldsymbol{\lambda}) \prod_{j=1}^{n} p(w_j|x_j, y_j, \mathbf{g}). \tag{11}$$

Here, $p(w_j|x_j, y_j, \mathbf{g}) = \sum_{z=1}^{K} p(w_j|z) p(z|x_j, y_j, \mathbf{g})$. This problem can be readily solved using an EM procedure that alternates between two updating steps as below.

$$q_j^{(t)}(k) \propto \exp(\eta_j^{(t-1)}(k)) \cdot \beta_k(w_j), \tag{12}$$

$$\mathbf{g}^{(t)} \leftarrow \operatorname*{argmax}_{\mathbf{g}} \sum_{j=1}^{n} L_j(q_j^{(t)}; \mathbf{g}) - \frac{1}{2}\mathbf{g}^T \boldsymbol{\Lambda} \mathbf{g}. \tag{13}$$

Here, $\eta_j^{(t-1)}$ depends on $\mathbf{g}^{(t-1)}$ as in Eq.(5). $L_j(q; \mathbf{g})$ is the expectation of $\log p(z_j|x_j, y_j; \mathbf{g})$ w.r.t. q, which is given by $\sum_{k=1}^{K} q(k) \log p(k|x_j, y_j; \mathbf{g})$. In addition, Eq.(13) is a convex optimization problem. To bootstrap the EM procedure, one can initialize $\mathbf{g}^{(0)}$ to a zero vector.

4.2. Learning Model Parameters

The goal of learning is to estimate the word distribution β_k of each topic, and the GP parameter $\boldsymbol{\lambda}$ that governs the spatially varying topic distribution. We first consider a supervised learning algorithm and then extend it to a semi-supervised setting. As a preceding step, we extract local descriptors densely from each image and quantized them (using K-means) into visual words. Suppose pixel-wise topic labeling is provided for each training image. Then, each word is represented by a 4-tuple as (x_j, y_j, w_j, z_j). Here, (x_j, y_j) is the coordinate, w_j is the word label, and z_j is the topic label. Then, the MAP estimator of β_k is

$$\beta_k(w) = \frac{\alpha + \#\text{occurrences of } w \text{ in } k\text{-th topic}}{\alpha V + \#\text{pixels belong to } k\text{-th topic}}. \tag{14}$$

Here, V is the vocabulary size. We use a prior count α to regulate the estimated distributions. This is equivalent to placing a Dirichlet prior with parameter $\alpha + 1$.

The parameter $\boldsymbol{\lambda}$ can be estimated by maximizing the following objective function w.r.t. both the model parameter $\boldsymbol{\lambda}$ and the latent layout representations $\mathbf{g}_1, \ldots, \mathbf{g}_N$.

$$\sum_{i=1}^{N} \left(\log p(\mathbf{g}_i|\boldsymbol{\lambda}) + \sum_{j=1}^{n} \log p(z_{ij}|x_{ij}, y_{ij}, \mathbf{g}_i) \right). \quad (15)$$

This problem can be solved using an coordinate ascent procedure that alternately updates $\boldsymbol{\lambda}$ and $(\mathbf{g}_i)_{i=1}^N$, as

$$\mathbf{g}_i^{(t)} \leftarrow \operatorname*{argmax}_{\mathbf{g}} \left(\sum_{j=1}^{n} \log p(z_{ij}|x_{ij}, y_{ij}, \mathbf{g}) - \frac{1}{2}\mathbf{g}^T \boldsymbol{\Lambda}^{(t-1)} \mathbf{g} \right);$$

$$\boldsymbol{\lambda}^{(t)} \leftarrow \operatorname*{argmax}_{\boldsymbol{\lambda}} \sum_{i=1}^{N} \log p(\mathbf{g}_i^{(t)}|\boldsymbol{\lambda}). \quad (16)$$

Here, $\boldsymbol{\Lambda}^{(t-1)}$ is the precision matrix determined by $\boldsymbol{\lambda}^{(t-1)}$. Note that the probability defined in Eq.(4) is log-concave w.r.t. η (and thus $\mathbf{g}$). Hence, both steps in Eq.(16) are convex optimization problems that can be readily solved. In practice, one can improve the numerical stability using L2 regularization of $\boldsymbol{\lambda}$ (*i.e.* add a term $-r\|\boldsymbol{\lambda}\|_2^2$ to Eq.(15)).

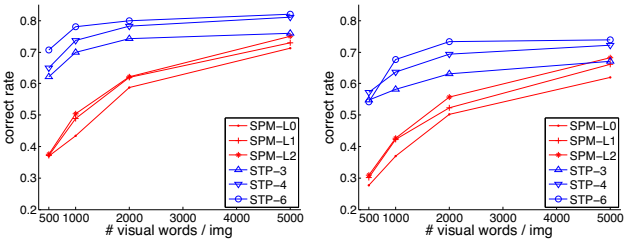
It is straightforward to extend the learning algorithm to leverage unsegmented images as part of the training data. The basic idea is to treat the topic indicators for such images as hidden variables, and use E-steps to infer the expected probabilities of their values, as in Eq.(12) and Eq.(13).

5. Applications and Experiments

We conducted experiments on three applications – *scene classification*, *semantic segmentation*, and *layout hallucination* – to test the practical utility of the proposed model.

We used two datasets: (1) *MSRC (v2)* [20], which contains 591 images in 20 scene categories and 23 object classes. Pixel-wise labeling are provided for each image. (2) *SUN* [27], a large database with 908 scene classes. However, many classes have too few annotated images for reliable training. Therefore, we selected a subset of 18 classes according to two criteria: (a) natural outdoor scenes, and (b) containing over 50 annotated images. This subset contains 8,952 images, which, we believe, is large enough to obtain statistically significant results. Annotations in the SUN dataset were noisy – regions (or objects) of the same types are often tagged with different names. Merging tags of the same meanings results in 28 distinct region (object) classes. Each dataset was randomly partitioned into two disjoint halves: *training* and *testing*.

Feature extraction and quantization were performed as a preprocessing step to obtain a set of visual words for each image. First, a bank of 25 filters (24 Gabor filters of 3 scales and 8 orientations plus a Gaussian smooth filter) is applied to three color channels (RGB) respectively. Combining the filter responses results in a 75-dimensional feature vector at each pixel. We found empirically that this feature tends to achieve better performance than dense SIFT in outdoor scenes, as significant parts of such scenes are textured regions instead of objects. Extracted features were whitened and then quantized using K-means ($K = 1000$).



(a) Results on MSRC (b) Results on SUN

Figure 4. Classification results obtained on two datasets. In the legend, SPM-Lk refers to spatial pyramid matching with k-levels, STP-k refers to spatial topic process on a $k \times k$ grid.

We learned the layout models from the training set following the procedure described in section 4. In specific, we set the prior count α to 10^{-4} in estimating the word distributions of each topic. We learned the *spatial topic processes* on three grid sizes 3×3, 4×4 and 6×6 over a standard image size 256×256, and set σ_g to 80, 60, and 40 respectively. We used supervised learning for *MSRC*, which provides pixel-wise labeling for all images and semi-supervised learning for *SUN*, where labeling is partially available.

5.1. Scene Classification

Given an image I, one can infer the *latent layout representation* $\mathbf{g}$ using the optimization algorithm presented in section 4.1. Here, $\mathbf{g}$ is a finite-dimensional vector and thus can be used as a holistic descriptor of the scene. We trained a set of linear SVMs based on these vectors, each for a scene category. Each testing image was classified to the class that yields highest prediction score. For comparison, we also implemented *Spatial Pyramid Matching (SPM)* [13], a popular discriminative method for scene recognition. We varied the number of visual words extracted from each image and studied how it influences performance.

Figure 4 compares the classification correct rates on both datasets. We observe: (1) For the proposed method *(STP)*, the classification accuracy increases when using finer grids, which suggests that local scale variations in the layouts convey useful information for classification. (2) STP outperforms SPM when using a 4×4 or 6×6 grid, which indicates that discriminative information is effectively captured by the layout representation. (3) It is a general trend for both methods that performance increases as more visual words

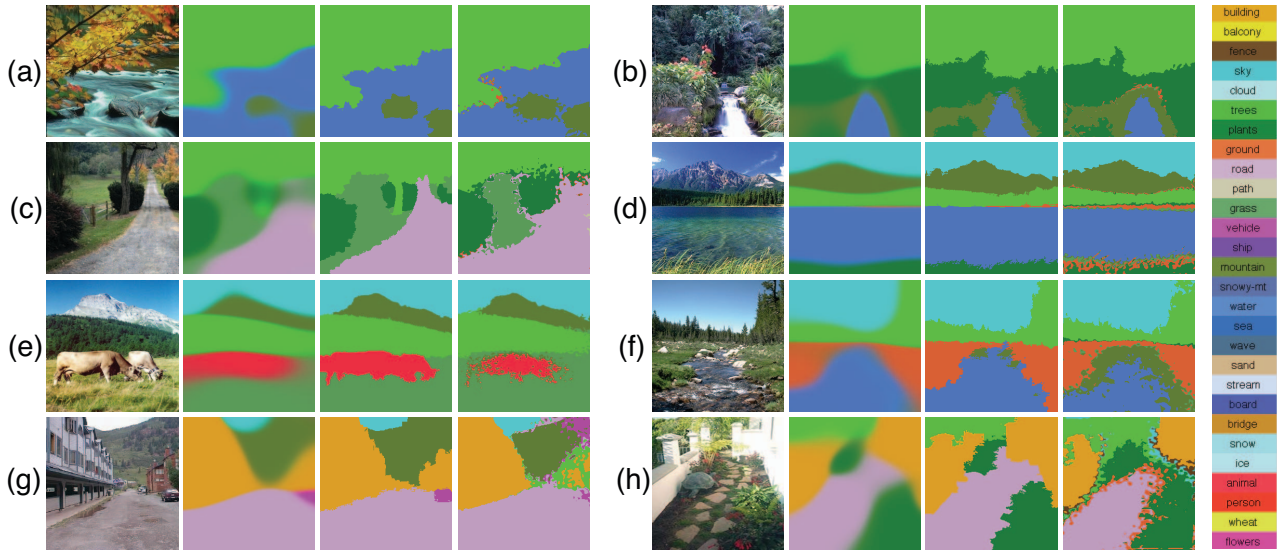


Figure 6. Eight groups of semantic segmentation results on the SUN dataset. Each group has four images. From left to right are the input image, the inferred layout (using a 4×4 grid), the result by our method (based on the inferred layout), and the result by SLDA. Particularly, the image to visualize the inferred layout is generated by mixing the colors of different topics using the probabilities $\theta(x, y)$ as weights.

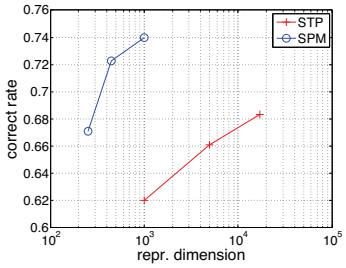


Figure 5. Correct rates vs representation dimensions, based on the results obtained on SUN (using 5000 visual words/image).

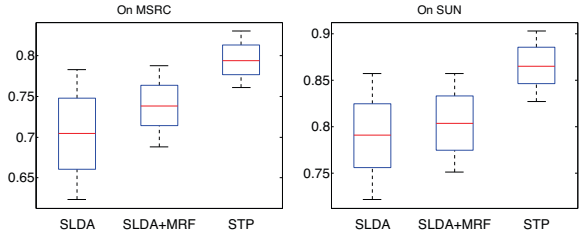


Figure 7. Quantitative comparison of segmentation performances.

are used. However, it is interesting to notice that such increase is much faster for STP than for others – a small subset of visual words is sufficient to estimate the layout reliably.

Generally, one may get better performance with higher representation dimension (*e.g.* increasing spatial resolution). A good method should be able to achieve high accuracy with low dimension. Figure 5 compares both methods using a dimension-accuracy diagram. Clearly, STP yields superior performance with much lower representation dimension as opposed to the other two, manifesting its effectiveness in capturing statistical dependencies in the layouts.

5.2. Semantic Segmentation

Semantic segmentation is to assign each pixel a semantic label (*i.e.* a topic in our model). The pixel-wise labeling divides an image into several regions with well-defined meanings (*e.g.* sky, hills). Given an image I, we first oversegment it into super-pixels using SLIC [2], and then obtain a semantic segmentation by assigning a label to each super pixel. The use of super-pixels not only reduces the computational cost but also improves the coherence of labeling.

Note that one can derive a continuous map θ of topic distributions from the layout representation $\mathbf{g}$ using Eq.(5) and (4), which provides a prior over the topics. We can then combine this prior with the visual words within a super pixel to infer its topic label. Specifically, let z_s denote the label of a super pixel s, then its posterior distribution is given by

$$p(z_s|s;\theta) \propto \prod_{i \in s} p(z_s|\theta(x_i, y_i))p(w_i|z_s;\beta). \qquad (17)$$

Here, we use $i \in s$ to indicate the i-th visual word is within the super-pixel s. Then, the optimal value for z_s is

$$\hat{z}_s = \operatorname*{argmax}_k \sum_{i \in s} \eta^{(k)}(x_i, y_i) + \log \beta_k(w_i). \qquad (18)$$

Here, we use the relation: $\theta^{(k)}(x_i, y_i) \propto \exp(\eta^{(k)}(x_i, y_i))$. For comparison, we also implemented a variant of spatial LDA [26, 29], which incorporates an MRF to enforce coherence between topics allocated to neighboring pixels.

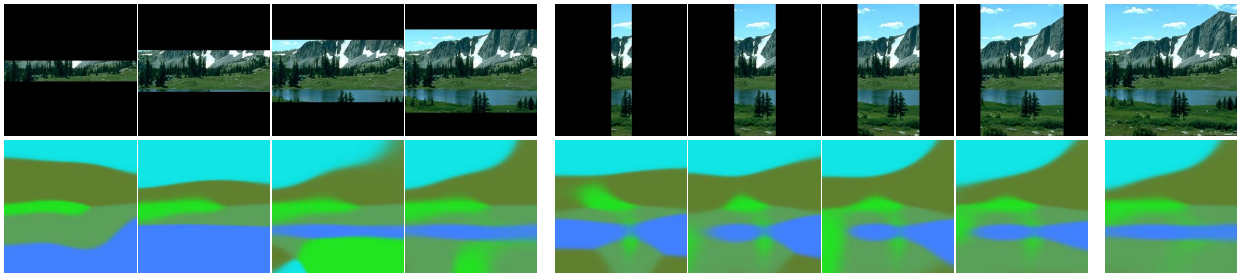


Figure 8. The first row are the inputs, and the second row are the inferred layouts. Columns (a) - (d) show the case where only middle rows are visible, (e) - (h) show the case where only middle columns are visible, and (i) shows the original image and the layout inferred thereon.

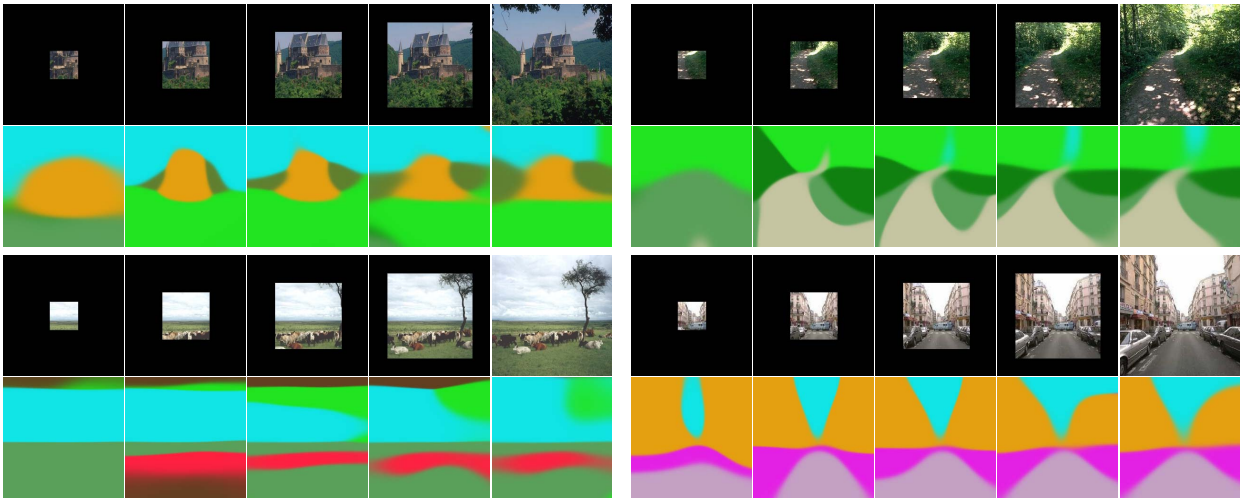


Figure 9. More results of layout hallucination. The algorithm infers the layout over the image plane based on a visible block at the center.

Figure 6 shows part of the segmentation results obtained on the SUN dataset, which accurately reflect the scene structures and have very good spatial coherence. Whereas SLDA was able to recover the major structures, the results are noisy especially in ambiguous regions. The improvement achieved by our method is, to a large extent, ascribed to the strong prior provided by the layouts. As we can see, the inferred layouts capture the spatial structures very well, thus substantially reducing the ambiguities of labeling. Empirically, the entropy of $\theta(x, y)$ is below 0.5 on average, implying that most labels have already been filtered out, leaving only one or two labels for each pixel to choose from.

We also perform quantitative comparison, where the performance is measured by the correct rate of labeling. Figure 7 reports the average performance over the testing sets. The results clearly show that our method consistently outperforms spatial LDA (+MRF) on both *MSRC* and *SUN*.

5.3. Layout Hallucination

It is an interesting phenomenon in human vision system that people often remember seeing a surrounding region of a scene that was not visible in the view. This is referred to as *boundary extension* [12] in cognitive science. The false memory here actually reflects the person's prior knowledge about visual scenes, and is a good prediction of the world that did exist beyond the original scope. These findings lead us to the belief that *a model that effectively captures the visual structures of a scene category should be able to extrapolate beyond the input images.* We devised experiments to verify this. In this experiment, only part of an image was visible, and we used the proposed method to infer the invisible parts. Specifically, we solve the optimal layout representation $\mathbf{g}$ based on a subset of visual words extracted from the visible part, and use it to generate the entire layout.

We first consider cases where the algorithm only sees the middle rows or columns. Figure 8 shows the results. Initially, seeing only a very small part of the image, the algorithm is able to produce a reasonable layout, which, however, does not necessary conform to the "ground-truth". As more regions are revealed, the true layout is gradually recovered. Generally, the predictions made based on the middle columns are more accurate than those based on the middle rows, since columns tend to contain more structural information than rows. Figure 9 shows more results obtained

on cases where only a block at the center is visible to the algorithm. These results demonstrates our model's capability of extrapolating layouts beyond the visible part.

6. Conclusions

We presented a novel approach to layout modeling. At the heart of this model is a spatial topic process which uses a set of coupled Gaussian processes to generate topic distributions that vary continuously across the image plane. Using the grid-based parameterization, we further derived a finite dimensional representation of layouts that captures the correlations across both locations and topics. The experiments on both scene classification and semantic segmentation showed that the proposed methods achieve considerable improvement over state-of-the-art, which is owning to the strong structural prior provided by the layout model. We also performed experiments on layout hallucination, which demonstrates that our model is able to extrapolate scene layouts beyond the visible part.

References

[1] Towards Total Scene Understanding: Classification, Annotation and Segmentation in an Automatic Framework. In *Proc. of CVPR'09*, 2009.

[2] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Ssstrunk. SLIC Superpixels Compared to State-of-the-art Superpixel Methods. In *IEEE Trans. on PAMI*, 2012.

[3] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003.

[4] L. Cao and L. Fei-Fei. Spatially Coherent Latent Topic Model for Concurrent Object Segmentation and Classification. In *Proc. of ICCV'07*, 2007.

[5] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky. Exploiting Hierarchical Context on a Large Database of Object Categories. In *Proc. of CVPR'10*, pages 129–136, 2010.

[6] L. Fei-fei and P. Perona. A Bayesian Hierarchical Model for Learning Natural Scene Categories. In *Proc. of CVPR'05*, 2005.

[7] P. Felzenszwalb, D. McAllester, and D. Ramanan. A Discriminatively Trained, Multiscale, Deformable Part Model. In *Prof. of CVPR'08*, 2008.

[8] M. A. T. Figueiredo. Bayesian Image Segmentation using Gaussian Field Priors. In *CVPR Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, 2005.

[9] S. Gould, R. Fulton, and D. Koller. Decomposing a Scene into Geometric and Semantically Consistent Regions. In *Proc. of ICCV'09*, 2009.

[10] T. Harada, Y. Ushiku, Y. Yamashita, and Y. Kuniyoshi. Discriminative Spatial Pyramid. In *Proc. of CVPR'11*, 2011.

[11] Q. Huang, M. Han, B. Wu, and S. Ioffe. A Hierarchical Conditional Random Field Model for Labeling and Segmenting images of Street Scenes. In *Proc. of CVPR'11*, 2011.

[12] H. Intraub and M. Richardson. Wide-angle memories of close-up scenes. *Journal of experimental psychology. Learning, memory, and cognition*, 15(2):179–187, Mar. 1989.

[13] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *Proc. of CVPR'06*, 2006.

[14] C. Li, D. Parikh, and T. Chen. Automatic Discovery of Groups of Objects for Scene Understanding. In *Proc. of CVPR'12*, volume 1, 2012.

[15] C. Liu, J. Yuen, and A. Torralba. SIFT Flow: Dense Correspondence across Scenes and Its Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):978–94, 2011.

[16] Z. Niu, G. Hua, X. Gao, and Q. Tian. Context Aware Topic Model for Scene Recognition. In *Proc. of CVPR'12*, 2012.

[17] A. Oliva and A. Torralba. Modeling the Shape of the scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[18] M. Pandey and S. Lazebnik. Scene Recognition and Weakly Supervised Object Localization with Deformable Part-based Models. In *Prof. of ICCV'11*, 2011.

[19] S. N. Parizi, J. G. Oberlin, and P. F. Felzenszwalb. Reconfigurable Models for Scene Recognition. In *Proc. of CVPR'12*, pages 2775–2782, 2012.

[20] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *IJCV*, 81(1), 2009.

[21] A. Shyr, T. Darrell, M. Jordan, and R. Urtasun. Supervised Hierarchical Pitman-Yor Process for Natural Scene Segmentation. In *Proc. of CVPR'11*, 2011.

[22] E. Sudderth and M. Jordan. Shared segmentation of natural scenes using dependent pitman-yor processes. In *Proc. of NIPS'08*, 2008.

[23] E. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky. Learning Hierarchical Models of Scenes, Objects, and Parts. In *Proc. of ICCV'05*, 2005.

[24] J. Tighe and S. Lazebnik. SuperParsing : Scalable Nonparametric Image Parsing with Superpixels. In *Proc. of ECCV'10*, 2010.

[25] J. Tighe and S. Lazebnik. Understanding Scenes on Many Levels. In *Proc. of ICCV '11*, 2011.

[26] X. Wang and E. Grimson. Spatial Latent Dirichlet Allocation. In *Proc. of NIPS'07*, 2007.

[27] J. Xiao, J. Hays, K. a. Ehinger, A. Oliva, and A. Torralba. SUN Database: Large-scale Scene Recognition from Abbey to Zoo. In *Proc. of CVPR'10*, 2010.

[28] R. S. Zemel and A. Miguel. Multiscale Conditional Random Fields for Image Labeling. In *Proc. of CVPR'04*, 2004.

[29] B. Zhao, L. Fei-Fei, and E. Xing. Image segmentation with topic random field. In *Proc. of ECCV'2010*, 2010.