# Discovering Scene Categories by Information Projection and Cluster Sampling

Dengxin Dai[⋆,‡], Tianfu Wu[†,⋆] and Song-Chun Zhu[†,⋆]
[⋆]Lotus Hill Research Institute (LHI), Ezhou, China
[†]Department of Statistics, University of California, Los Angeles (UCLA), USA
[‡]Signal Processing Lab, School of Electronic Information, Wuhan University, China
dxdai@mail.whu.edu.cn, {tfwu, sczhu}@stat.ucla.edu

## Abstract

*This paper presents a method for unsupervised scene categorization. Our method aims at two objectives: (1) automatic feature selection for different scene categories. We represent images in a heterogeneous feature space to account for the large variabilities of different scene categories. Then, we use the information projection strategy to pursue features which are both informative and discriminative, and simultaneously learn a generative model for each category. (2) automatic cluster number selection for the whole image set to be categorized. By treating each image as a vertex in a graph, we formulate unsupervised scene categorization as a graph partition problem under the Bayesian framework. Then, we use a cluster sampling strategy to do the partition (i.e. categorization) in which the cluster number is selected automatically for the globally optimal clustering in terms of maximizing a Bayesian posterior probability. In experiments, we test two datasets, LHI 8 scene categories and MIT 8 scene categories, and obtain state-of-the-art results.*

## 1. Introduction

In this paper, we present a method for unsupervised scene categorization which is posed as a graph partition problem with each image being a vertex in the graph. We consider scene categories which include outdoor scenes, indoor scenes and object categories (such as highway, library, motorcycle, respectively). Unsupervised scene categorization is an important research topic with a wide range of applications such as image retrieval, web-based image search and top-down context in object detection and recognition [4, 15, 14, 19]. This paper addresses the following two main problems in unsupervised scene categorization:

(1) Automatic feature selection for different scene categories. As illustrated in the top panel of Fig.1, different scene categories are formed by class-specific appearance features (such as textures, textons, sketches and colors) and specific geometric configurations ranging from stochastic ones (modeled by bag-of-words) to regular ones (modeled by deformable template or active basis). In the literature of unsupervised scene categorization, most work do not address the feature selection problem. Instead, they use a predefined long vector of features for all categories [15, 21, 9, 20]. Recently, the information projection strategy has shown promising results on pursuing features under an information-theoretic framework [12, 23].

(2) Automatic cluster number selection. In the existing work of unsupervised scene categorization, the cluster number is often predefined or searched exhaustively in a given range [9, 19]. The key for automatic cluster number selection lies in whether an algorithm can explore the solution space efficiently instead of exhaustively. Automatic cluster number selection is a very important aspect of an unsupervised algorithm and affects performance largely. It entails the ability of generating new clusters, removing existing clusters and switching between two different clusters, as illustrated in the bottom panel of Fig.1. Generally, there are two kinds of methods which are capable of this task, one is the hierarchical Dirichlet process based on the stick-breaking constructions [5, 16] (which, however, has no obvious support for the feature selection) and the other is the Markov chain Monte Carlo (MCMC) strategy with death, birth and diffusion dynamics [18, 1] (which can incorporate the feature selection, but has not yet been addressed for unsupervised scene categorization).

Motivated by the two problems, we present a method for unsupervised scene categorization with automatic feature selection and cluster number selection. As illustrated in Fig.1, the basic idea is to represent images of different categories in a heterogeneous feature space. Then, we discover scene categories by information projection and cluster sampling. Concretely, by treating each image as a vertex in a graph, we formulate unsupervised scene categorization as a graph partition problem under the Bayesian framework, which has the following two key components:

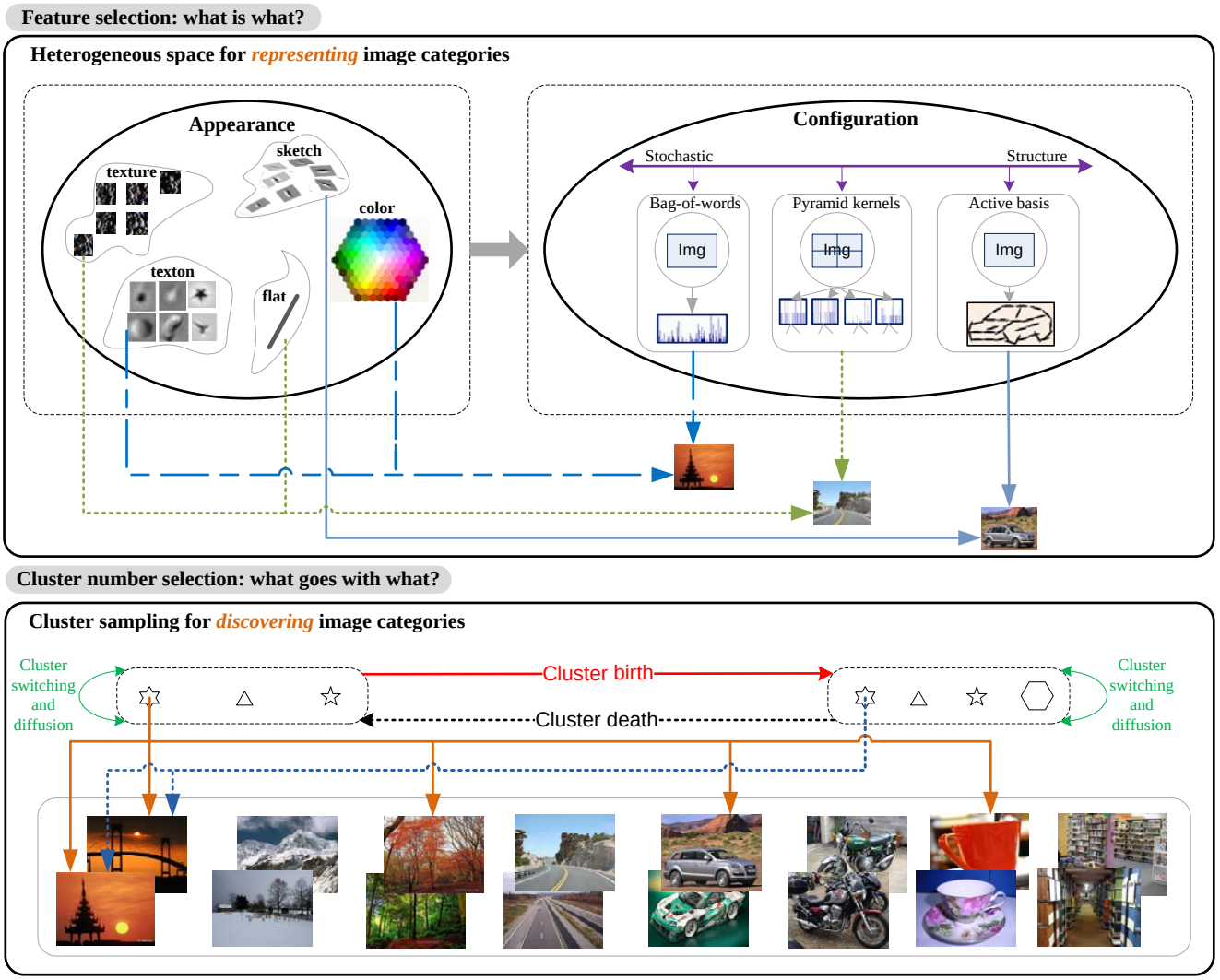


Figure 1. Illustration of the proposed method for unsupervised scene categorization. Our method addresses the automatic feature selection and the automatic cluster number selection simultaneously in a unified framework. *The top panel* illustrates the heterogeneous feature space in which we represent images of different categories to account for their large variabilities. The automatic feature selection is facilitated by the information projection strategy in the heterogeneous feature space. *The bottom panel* illustrate the cluster sampling strategy for automatically selecting cluster number. (see texts for details)

(1) The "what is what" component. We pursue features which are both informative and discriminative by using the information projection strategy [12, 23] and simultaneously learn a generative model for each cluster. The learning starts from a reference model pooled over the whole image set to be categorized. Then, the learned models are used as the likelihood in cluster sampling.

(2) The "what goes with what" component. We explore the partition space efficiently by using the cluster sampling algorithm [1] which has cluster birth, death and switching dynamics. The importance of the cluster sampling strategy is that it can jump in the solution space by flipping the label of many coupled images simultaneously (as a connected component in the graph). The cluster number is selected automatically for globally optimal clustering in terms of maximizing a Bayesian posterior probability. We use the Swendsen-Wang Cuts (SWC) algorithm [1] for the inference.

In the literature, most of the scene categorization methods are based on supervised learning [3, 13]. Unsupervised scene categorization is often formulated as the problem of searching and matching re-occurrences of some visual patterns among different images. Usually, visual patterns are represented by some low level holistic features (such as global color, histogram of filter response, Gabor filter and the gist), and scene categories are described as

bag-of-words [8], pyramid kernels [6, 22], subgraphs in a hierarchical structure [17] and active basis [23]. The categorization methods used in the literature are Probabilistic Latent Semantic Analysis (PLSA) [7], Latent Dirichlet Allocation (LDA) [2] and SUM-MAX algorithm [23].

The rest of the paper is organized as follows. Sec.2 introduces the graph partition formulation of unsupervised scene categorization as well as the information projection strategy and the cluster sampling strategy. Sec.3 presents the implementation by using the SWC algorithm. Sec.4 introduces the methods for performance evaluation. Sec.5 shows results of our experiments. Sec.6 concludes this paper.

## 2. Problem formulation

By treating each image as a vertex in a graph, we formulate unsupervised scene categorization as a graph partition problem under the Bayesian framework.

### 2.1. Unsupervised scene categorization

Let $\mathcal{D}$ be a set of input images containing a scene (outdoor or indoor) or object,

$$\mathcal{D} = \{I_1, I_2, \cdots, I_n\} \tag{1}$$

The objective of unsupervised scene categorization is to seek a partition of $\mathcal{D}$ with $K$ inferred clusters ($K$ is an unknown number to be inferred and $1 \leq K \leq n$). We have,

$$\mathcal{D} = \cup_{k=1}^{K} D_k \tag{2}$$

where $D_k \neq \emptyset$ and $D_i \cap_{\forall i \neq j} D_j = \emptyset$ $(i, j \in [1, K])$.

Denote $D_k$ as,

$$D_k = \{I_1^{(k)}, \cdots, I_{n_k}^{(k)}\} \tag{3}$$

where $n_k$ is the number of images in $D_k$ and $1 \leq n_k < n$. We have $n = \sum_{k=1}^{K} n_k$.

Let $\mathcal{F}$ be the feature pool in which we select features and simultaneously learn a model for $D_k$ $(1 \leq k \leq K)$,

$$\mathcal{F} = \{f_1, f_2, \cdots, f_m\} \tag{4}$$

where $f_i$ can be vector valued functions of different filter responses such as the intensity, colors, Gabor filters and SIFT descriptors extracted at certain locations in the image lattice and scales.

Let $F_k \subset \mathcal{F}$ be the selected features for $D_k$ and $\Theta_k$ be the parameters of the model built upon $F_k$ to describe images in $D_k$.

Given $\mathcal{D}$ and $\mathcal{F}$, the unsupervised image categorization seeks the output,

$$W = (K, \{D_k, F_k, \Theta_k\}_{k=1}^{K}) \tag{5}$$

### 2.2. Bayesian formulation

Under the Bayesian framework, we seek the optimal $W^*$ (Eqn.5) by maximizing a Bayesian posterior probability,

$$W^* = \arg\max_{w \in \Omega} p(W|\mathcal{D}) \tag{6}$$

where $\Omega$ is the solution space, and we have

$$\begin{aligned} p(W|\mathcal{D}) &\propto p(W)p(\mathcal{D}|W) \\ &= p(W) \prod_{k=1}^{K} [\prod_{i=1}^{n_k} \mathcal{P}_k(I_i^{(k)}; F_k, \Theta_k)] \end{aligned} \tag{7}$$

Before specifying the prior probability $p(W)$ and the likelihood $\mathcal{P}_k$, we first anatomize the solution space $\Omega$ and specify the graph partition formulation.

#### 2.2.1 Anatomy of the solution space $\Omega$

The solution space $\Omega$ is a product of the partition space (specifying "what goes with what") and the model space (specifying "what is what").

*Partition space.* In Eqn.2, when the image set $\mathcal{D}$ is categorized into $K$ disjoint clusters, we call it a $K$-partition, denoted by $\pi_K$,

$$\pi_K = \{D_1, D_2, \cdots, D_K\} \tag{8}$$

Let $\Pi_K$ be the set of all possible $K$-partitions and we have,

$$\Pi_K = \{\pi_K; |D_k| > 0, k = 1, \cdots, K\} \tag{9}$$

Then, the partition space, denoted by $\Omega_\Pi$, is,

$$\Omega_\Pi = \cup_{K=1}^{n} \Pi_K \tag{10}$$

*Model space.* For each $D_k \in \pi_K$, we need specify a model $\mathcal{P}_k$ which comes from a model space $\Omega_{\mathcal{P}_k}$. We have,

$$\Omega_{\mathcal{P}_k} = \{F_k, \Theta_k; F_k \in \mathcal{F}\} \tag{11}$$

*Solution space.* The solution space $\Omega$ is a union of subspaces $\Omega_K$ and each $\Omega_K$ is the product of one partition space $\Pi_K \in \Omega_\Pi$ and $K$ model spaces $\Omega_{\mathcal{P}_k}$, that is,

$$\begin{aligned} \Omega &= \cup_{K=1}^{n} \Omega_K \\ &= \cup_{K=1}^{n} \{\Pi_K \times \Omega_{\mathcal{P}_1} \times \cdots \times \Omega_{\mathcal{P}_K}\} \end{aligned} \tag{12}$$

In terms of the anatomy of $\Omega$, Eqn.5 can be rewritten as,

$$W = (K, \pi_K, \{F_k, \Theta_k\}_{k=1}^{K}) \tag{13}$$

So, in Eqn.7, the prior probability $p(W)$ is,

$$p(W) = p(K)p(\pi_K) \prod_{k=1}^{K} p(F_k)p(\Theta_k|F_k) \tag{14}$$

where $p(K)$ is the prior model for the number of cluster which is often assumed as an exponential family model,

$$p(K) \propto \exp\{-\beta K\} \tag{15}$$

$\beta$ is a tuning parameter, $p(\pi_K)$ is the uniform distribution in the partition space $\Pi_K$, and both $p(F_k)$ and $p(\Theta_k|F_k)$ are the uniform distribution in the model space.

#### 2.2.2 The graph partition specification

Based on the anatomy of the solution space $\Omega$, we present the graph partition formulation which includes the specifications of the adjacency graph and the edge probability.

*The adjacency graph.* This graph represents a $K$-partition $\pi_K$ by treating each image in $\mathcal{D}$ as a vertex. Denote $G =< V, E >$ as the adjacency graph. We know that $V = \mathcal{D}$. $E$ is the set of edges linking images in $V$. The initial adjacency graph $G_0 =< V, E_0 >$ is a fully connected graph so the initial edge set $E_0$ could be very large ($|E_0| = n(n-1)/2$). Each edge $e \in E_0$ is augmented with a Bernoulli random variable $\mu_e \in \{\text{on, off}\}$, which indicates whether the edge $e$ is turned on or off. We define the edge turn-on probability to reflect the image-to-image similarity based on the generative model learned by the information projection strategy (see Sec.2.3). Then, by turning off the edges probabilistically based on the edge probability, we obtain a relatively sparse adjacency graph $G =< V, E >$ ($E \subset E_0$).

*The edge probability.* Let $I_s$ and $I_t$ be the two images linked by edge $e \in E$. Denote $q_e$ as the edge probability,

$$q_e = p(\mu_e = \text{on } |I_s, I_t) \tag{16}$$

where $q_e$ indicates the probability of images $I_s$ and $I_t$ being from the same category.

By treating the image $I_s$ and $I_t$ as a cluster $D_e = \{I_s, I_t\}$, we can learn a probability model $\mathcal{P}_e$ based on the method described in Sec.2.3. Then, we define $q_e$ as

$$q_e \propto \mathcal{P}_e(I_s) \times \mathcal{P}_e(I_t) \tag{17}$$

which will be specified in Eqn.21.

In next section, we present the information projection strategy to do automatic feature selection and learn a generative probability model $\mathcal{P}_k$ for each inferred cluster $k$.

### 2.3. Feature selection by information projection

For each $D_k$ in a $K$-partition $\pi_K$, the goal of the feature selection is to pursue a subset of features $F_k \subset \mathcal{F}$ which are both informative and discriminative for $D_k$, and simultaneously learn a generative probability model $\mathcal{P}_k$. To achieve that end, we use the information projection strategy [12].

In information projection, we start from a reference model for $D_k$, denoted by $\mathbf{q}_k(I)$. It means that we specify a null hypothesis, $H_0 : I \sim \mathbf{q}_k(I)$, for image $I \in D_k$ (for simplicity, we omit the subscript in $I_i^{(k)}$ when there is no confusion). In this paper, in order to make different $\mathcal{P}_k$'s comparable, we use the same reference model for them, so we denote the reference model as $\mathbf{q}(I)$ (which could be the uniform distribution or Gaussian white noise distribution). Furthermore, information projection tells us that we do not need to specify $\mathbf{q}(I)$ explicitly except that we need specify the marginal distribution $q_i(I)$ of $\mathbf{q}(I)$ projected onto each feature dimension $f_i$ ($i \in [1, m]$) to do the feature selection. $q_i(I)$ is obtained by pooling the transformed response $h(f_i(I))$ of each feature $f_i \in \mathcal{F}$ over the whole image set $\mathcal{D}$ where $h(r) = \zeta[\frac{2}{1+e^{-2r/\zeta}} - 1]$ is a sigmoid transformation function with $\zeta$ being the saturation level ($\zeta = 6$ in our experiments).

The learning procedure in information projection tells us that we can modify $\mathbf{q}(I)$ to the model $\mathcal{P}_k$ by identifying test statistics of feature responses of $F_k$ on $D_k$ (say, the projected marginal distributions) to reject the null hypothesis $H_0$. For simplicity, we can design the feature set $\mathcal{F}$ to make all the $f_i$'s independent (in a spatial or frequency domain sense). Let $p_i^{(k)}(I)$ be the marginal distribution of $h(f_i(I))$ pooled on $D_k$. We parameterize $p_i^{(k)}(I)$ by the exponential family [12],

$$p_i^{(k)}(I) = q_i(I) \frac{1}{Z(\lambda_i^{(k)})} \exp\{\lambda_i^{(k)} h(f_i(I))\} \tag{18}$$

where $\lambda_i^{(k)}$ is estimated by MLE and then $Z(\lambda_i^{(k)})$ can be calculated (as in the active basis model [23]).

The information gain of a feature $f_i$ for the inferred category $k$, denoted by $\Delta^{(k)}(f_i)$, is defined as,

$$\begin{aligned} \Delta^{(k)}(f_i) &= \log \frac{p_i^{(k)}(I)}{q_i(I)} \\ &= \lambda_i^{(k)} h(f_i(I)) - \log Z(\lambda_i^{(k)}) \end{aligned} \tag{19}$$

which guides us to pursue features sequentially up to a certain threshold of the information gain (zero in our experiments) and then we obtain the selected feature subset $F_k$ for $D_k$. Denote $m_k$ as the number of features in $F_k$ which may vary for different categories. Here, because $q_i$ is pooled from the whole image set $\mathcal{D}$ and $p_i^k$ is pooled from $D_k \subset \mathcal{D}$, we can see that features selected according to the information gain criteria in Eqn.19 are both informative and discriminative to $D_k$. Then, we learn a generative probability model for $D_k$,

$$\begin{aligned} \mathcal{P}_k(I) &= \mathbf{q}(I) \prod_{i=1}^{m_k} \frac{p_i^{(k)}(I)}{q_i(I)} \\ &= \mathbf{q}(I) \prod_{i=1}^{m_k} \frac{1}{Z(\lambda_i^{(k)})} \exp\{\lambda_i^{(k)} h(f_i^{(k)}(I))\} \end{aligned} \tag{20}$$

where $f_i^{(k)} \in F_k$.

Now, we specify the edge probability $q_e$ in Eqn.17. We have $\mathcal{P}_e(I) = \mathbf{q}(I)\prod_{i=1}^{m_e}\frac{1}{Z(\lambda_i^{(e)})}\exp\{\lambda_i^{(e)}h(f_i^{(e)}(I))\}$. Let $\Delta(e) = \log\frac{\mathcal{P}_e(I_s)}{\mathbf{q}(I)} + \log\frac{\mathcal{P}_e(I_t)}{\mathbf{q}(I)}$. The edge probability is defined as the sigmoid transformation of $\Delta(e)$,

$$q_e = \frac{2}{1+\exp\{-\frac{\Delta(e)}{T}\}} - 1 \tag{21}$$

where $T$ is the temperature factor for simulated annealing, and all $\Delta(e)$ are calculated offline and stored in a look-up table for cluster sampling.

**Algorithm 1**: Discovering scene categories by information projection and cluster sampling

**Input** : A set of images $\mathcal{D}$ (Eqn.1) to be categorized, a feature pool $\mathcal{F}$ (Eqn.4), and the posterior probability $p(W|\mathcal{D})$ (Eqn.7)

**Output**: The unsupervised categorization results $W^* = (K^*, \{D_k^*, F_k^*, \Theta_k^*\}_{k=1}^K)$ (Eqn.5).

**Initialization**: create the initial adjacency graph $G_0 =< V, E_0 >$ and compute $q_e$ (Eqn.21) for all $e \in E_0$, then obtain the adjacency graph $G =< V, E >$ by sampling the edges in $E_0$ and random clustering for $CP$s in $G$.

**Repeat:** for a current partition $\pi_K$ with the solution state $W$,

. **Graph clustering:** create a new set of $CP$s by turning off $e \in E(\pi_K)$ with probability $1 - q_e$.

. **Graph flipping:** select a $CP$ $V_0$ with probability $q(V_O|CP)$ and suppose $V_0 \subset D_k$, and then flip $V_0$ as a whole to cluster $k^{'}$ which is sampled from $q(k^{'}|V_0, \pi_K)$. Accept the flipping with probability $\alpha(W \to W^{'})$ (Eqn.23)

## 3. Implementation

We adopt the SWC algorithm [1] for the inference. By recalling the posterior probability in Eqn.7, the prior probability in Eqn.14 and the factorized likelihood in Eqn.20, we can rewrite Eqn.6 as,

$$W^* = \arg\max_{W\in\Omega}\exp\{-\beta K\} \tag{22}$$
$$\times\prod_{k=1}^{K}[\prod_{i=1}^{n_k}\mathbf{q}(I_i^{(k)})\prod_{j=1}^{m_k}\frac{1}{Z(\lambda_j^{(k)})}\exp\{\lambda_j^{(k)}h(f_j^{(k)}(I_i^{(k)}))\}]$$

Since we do not specify $\mathbf{q}(I_i^{(k)})$ explicitly beyond the marginal distributions on the selected feature dimensions, the energy (say, $-\log p(W|\mathcal{D})$) could take on negative values (see the energy plot in the left bottom panel in Fig.2).

Next, we specify the clustering sampling procedure with the SWC algorithm to solve Eqn.22.

### 3.1. Inference by Swendsen-Wang Cut

The SWC algorithm is very effective for sampling arbitrary posterior probability or energy functions [1] with a graph partition formulation. To solve Eqn.22 by SWC, we need to specify the following three steps.

(1) *Initialization.* Given an initial adjacency graph $G_0 =< V, E_0 >$, we compute the edge probability $q_e$ according to the Eqn.21. We turn off the edges in $E_0$ independently and probabilistically according to $q_e$ ($\forall e \in E_0$), so that we can obtain a relatively sparse adjacency graph $G =< V, E >$ ($E \subset E_0$) with probability $q(E) = \prod_{e\in E} q_e \prod_{e\in E_0\setminus E}(1-q_e)$. Then, $G$ consists of a set of connected components ($CP$). We can assign the same initial category index for all $CP$s or can draw category indexes based on the prior probability in Eqn.14 and assign it to $CP$s. We pursue the features and probability models for each clusters, and we obtain an initial solution state. Then, the following two steps are iterated.

(2) *Graph clustering.* Given a current partition $\pi_K$ with $K$ hypothesized categories, the corresponding solution state is $W = \{K, \{D_k, F_k, \Theta_k\}_{k=1}^K\}$ with the posterior probability defined in the Eqn.7. It removes all the edges between images of different categories, and turns off the remaining edges according to their edge probabilities. Then, we get a new set of $CP$s.

(3) *Graph flipping.* It selects a connect component $V_0 \in CP$ with probability $q(V_0|CP)$ (in this paper, $q(V_0|CP) = \frac{1}{|CP|}$ is uniform) and suppose we get $V_0 \subset D_k$), and then flips, accepted with a probability (Eqn.23), all the images in $V_0$ to another image category $k^{'}$ with probability $q(k^{'}|V_0, \pi_K)$ ($k^{'} \in [1, K+1]$, in this paper $q(k^{'}|V_0, \pi_K) = \frac{1}{1+K}$ is uniform). If it is accepted, we get a new solution state $W^{'}$, and we know,

- If $k^{'} = K+1$, it generates a new category (cluster birth) $D_{k^{'}} = V_0$ and we pursue the features $F_{k^{'}}$ and learn the model $\mathcal{P}_{k^{'}}$. Then, we update the model for $D_k = D_k \setminus V_0$ (In the case when $V_0 = D_k$, we give death to the cluster $k$ actually at the same time).
- If $k^{'} \in [1, K]$ and $k^{'} \neq k$, it does regrouping and model diffusion by updating $D_k = D_k \setminus V_0$ (cluster death happens when $V_0 = D_k$) and $D_{k^{'}} = D_{k^{'}} \cup V_0$ and pursuing the feature sets $F_k$ and $F_{k^{'}}$ and the models $\mathcal{P}_k$ and $\mathcal{P}_{k^{'}}$.
- If $k^{'} = k$, we have $W^{'} = W$.

The acceptance probability, denoted by $\alpha(W \to W^{'})$, is,

$$\alpha(W \to W^{'}) = \min(1, \frac{Q(W^{'} \to W)}{Q(W \to W^{'})}\frac{p(W^{'}|\mathcal{D})}{p(W|\mathcal{D})}) \tag{23}$$

where $Q(W^{'} \rightarrow W)$ and $Q(W \rightarrow W^{'})$ are proposal probabilities for the state jumping and their ratio $\frac{Q(W^{'} \rightarrow W)}{Q(W \rightarrow W^{'})}$ can be derived as in [1].

Algorithm.1 summarizes our unsupervised image categorization method.

## 4. Performance evaluation

For the input image set $\mathcal{D}$, suppose there are $L$ categories and the corresponding ground truth category label set is,

$$\mathcal{C} = \{c_1, c_2, \cdots, c_n\} \tag{24}$$

where $c_i \in \{1, \cdots, L\}$, and both $L$ and $\mathcal{C}$ are unknown to the categorization algorithm.

Given the categorization result $W = (K, \{D_k, F_k, \Theta_k\}_{k=1}^{K})$, we have the inferred category label set,

$$\hat{\mathcal{C}} = \{\hat{c}_1, \hat{c}_2, \cdots, \hat{c}_n\} \tag{25}$$

where $\hat{c}_i \in [1, K]$ and we know that it is possible that $K \leq L$ or $K > L$

For each $D_k \subset \mathcal{D}$ with the inferred cluster label $k$, we also know the corresponding groundtruth label set $C_k$,

$$C_k = \{c_1^{(k)}, \cdots, c_{n_k}^{(k)}\} \subset \mathcal{C} \tag{26}$$

As noted in [19], we adopt the purity (larger value means better performance) and the conditional entropy (smaller value means better performance) as the evaluation criteria, which make more sense for evaluating performance of unsupervised image categorization.

The purity is defined as,

$$\text{Purity}(\hat{\mathcal{C}}|\mathcal{C}) = \sum_{k \in \hat{\mathcal{C}}} p(k) \max_{l \in \mathcal{C}} p(l|k) \tag{27}$$

where $p(k) = \frac{|D_k|}{n}$ and $p(l|k)$ is estimated from $C_k$.

The conditional entropy is defined as,

$$\mathcal{H}(\hat{\mathcal{C}}|\mathcal{C}) = \sum_{k \in \hat{\mathcal{C}}} p(k) \sum_{l \in \mathcal{C}} p(l|k) \log \frac{1}{p(l|k)} \tag{28}$$

## 5. Experiments

We test two datasets, LHI 8 scene categories (4 outdoor scene categories, 1 indoor scene category and 3 object categories) [24] and MIT 8 scene categories [11].

*Feature pool $\mathcal{F}$.* In our current experiments, we use four types of features: the "Gist" feature [11], the SIFT feature [10], the PACT feature [22] and the RGB color features. For the "Gist" feature, we compute it on the size of $256 \times 256$ pixels. The other features are computed using the original image size. For the SIFT features, we use the bag-of-words configuration with 100 visual words. For the PACT feature, we use a 2-layer pyramid with 4 cells in the second layer. For the color feature, we use 39 bins histogram with 13 bins for each of the red, green and blue channels. Note that more features can be added as illustrated in the top panel of Fig.1, such as the active basis model [23] and the constellation model [4], and we will explore them in our future work.

| | $k$-mean ($k$=8) | | | | pLSA | **Ours** |
|---|---|---|---|---|---|---|
| | Gist | SIFT | PACT | Color | SIFT | |
| Purity | 0.393 | 0.471 | 0.408 | 0.375 | 0.374 | **0.645** |
| Entropy | 1.619 | 1.412 | 1.502 | 1.601 | 1.592 | **1.052** |

Table 1. Results on the LHI 8 scene categories.

| | $k$-mean ($k$=8) | | | | pLSA | **Ours** |
|---|---|---|---|---|---|---|
| | Gist | SIFT | PACT | Color | SIFT | |
| Purity | 0.509 | 0.410 | 0.458 | 0.267 | 0.352 | **0.635** |
| Entropy | 1.285 | 1.510 | 1.463 | 2.401 | 1.652 | **1.114** |

Table 2. Results on the MIT 8 scene categories.

*Experimental setting.* In the experiments, we set $\beta = 6$ in the prior probability (Eqn.14) and the temperature $T$ in the edge probability (Eqn.21) starts from 60 and goes down in a logarithmic manner to 0.1 in 1000 sweeps. In addition, for each method we test on the two datasets to compare peformance, we run 10 times and use the average performance to do comparisons. For comparisons, we use the $k$-mean algorithm with correct cluster number (say, 8 in both datasets) specified as the baseline method. We also compare with the pLSA method (with the SIFT features and correct category number specified).

*Experiment 1: LHI 8 scene categories*[1]. As shown in Fig.2, the top panel shows the feature selection results for each category which is very perceptually intuitive, and the bottom panel shows the energy curve for the cluster sampling in which we pick up three points to show the cluster birth, cluster death and cluster switching and diffusion, respectively. The calculated purities and conditional entropies are in Table.1. The average cluster number we obtain in 10 times is 11.2.

*Experiment 2: MIT 8 scene image categories*[2]. The result is shown in Table.2. The average cluster number from the 10 times is 10.5 clusters.

**Discussion on cluster number selection.** In our experiments, the $\beta$ in the prior probability (Eqn.14) constrain the cluster number, and as illustrated in Fig.3, when $\beta$ is small (the bottom panel) we can divide a category (see the top panel) into a set of sub-clusters to fit the likelihood better.

[1] http://www.imageparsing.com/ImageCategory8.html

[2] http://people.csail.mit.edu/torralba/code/spatialenvelope/

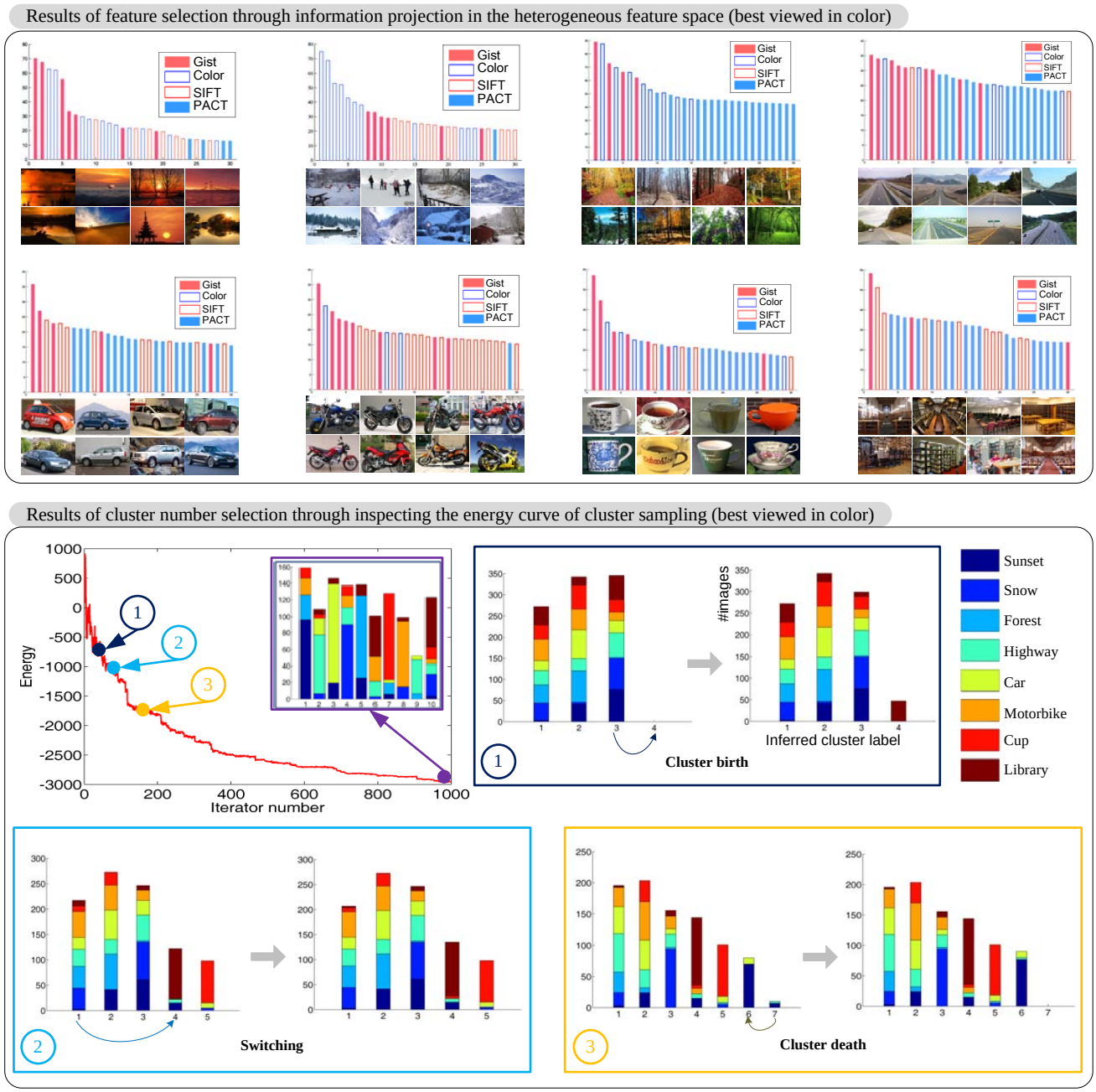


Figure 2. Demonstrations and results of feature selection and cluster number selection of our method on the LHI 8 scene categories. For the feature selection, we first choose $k_l = \arg\max_k p(l|C_k)$ (see $C_k$ in Eqn.26) to connect the obtained cluster label and the ground truth category label, and then plot the features $F_{k_l}$. The horizontal coordinate in the cluster sub-figures is the assigned cluster label $k$ and the bins with different colors are the ground truth category label $l$. Since we do not specify the reference model $\mathbf{q}(I)$ explicitly in Eqn.20 beyond the marginal distributions on those selected feature dimensions, the energy value can take on negative values. (Best viewed in color)

Here, the cluster number $K$ indeed depends on parameter $\beta$ which control the level of details. What we meant by "automatically" computing the number $K$ in this paper is that our cluster sampling algorithm can decide the optimal $K$ by its MCMC dynamics in the process of optimizing the Bayesian posterior probability with arbitrary initialization of $K$.

## 6. Conclusion

This paper presents a method for unsupervised scene categorization. We address the automatic feature selection and the automatic cluster number selection simultaneously by posing unsupervised scene categorization as graph partition

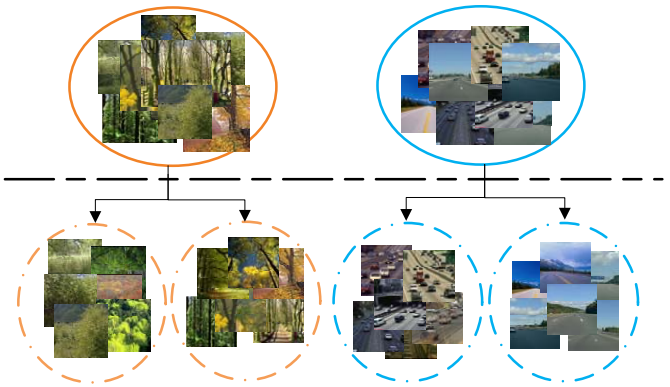

Figure 3. Examples of cluster number selection guided by the prior: $\beta = 6$ for the top panel and $\beta = 2$ for the bottom panel. Left: the forest is divided into two subsets with different scales. Right: the highway is divided into two subsets with different road environments.

problem (in which each image is treated as a vertex) under the Bayesian framework. In implementation, we use information projection to pursue features which are both informative and discriminative for each category and adopt the Swendsen-Wang Cut algorithm for effective inference with cluster number selected automatically for globally optimal clustering. In experiments, we test two datasets, LHI 8 scene categories and MIT 8 scene categories, and obtain state-of-the-art results.

**Acknowledgement.** The work at UCLA was supported by NSF grants IIS-0713652, ONR grant N00014-07-M-0287 and the work at LHI was supported by China 863 project 2008AA01Z126 and 2009AA01Z331, and NSF China grants 60728203. The authors are thankful to the anonymous reviews for their constructive comments, to Dr. Yingnian Wu, Zhenyu Yao and Brandon Rothrock at UCLA for their insightful discussions.

## References

[1] A. Barbu and S. C. Zhu. Generalizing swendsen-wang to sampling arbitrary posterior probabilities. *PAMI*, 27(8):1239–1253, 2005.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[3] A. Bosch, A. Zisserman, and X. Muńoz. Scene classification using a hybrid generative/discriminative approach. *PAMI*, 30(4):712–727, 2008.

[4] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, volume 2, pages 264–271, 2003.

[5] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky. An hdp-hmm for systems with state persistence. In *ICML*, pages 312–319, 2008.

[6] K. Grauman and T. Darrell. The pyramid match kernel: discriminative classification with sets of image features. In *ICCV*, volume 2, pages 1458–1465, 2005.

[7] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.*, 42(1-2):177–196, 2001.

[8] F.-F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, pages 524–531, 2005.

[9] D. Liu and T. Chen. Unsupervised image categorization and object localization using topic models and correspondences between images. In *ICCV*, pages 1–7, 2007.

[10] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[11] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[12] S. D. Pietra, V. J. D. Pietra, and J. D. Lafferty. Inducing features of random fields. *PAMI*, 19(4):380–393, 1997.

[13] Z. Si, H. Gong, Y. N. Wu, and S.-C. Zhu. Learning mixed image templates for object recognition. In *CVPR*, 2009.

[14] I. Simon, N. Snavely, and S. M. Seitz. Scene summarization for online image collections. In *ICCV*, pages 1–8, 2007.

[15] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering object categories in image collections. In *ICCV*, 2005.

[16] E. B. Sudderth, A. B. Torralba, W. T. Freeman, and A. S. Willsky. Describing visual scenes using transformed objects and parts. *IJCV*, 77(1-3):291–330, 2008.

[17] S. Todorovic and N. Ahuja. Unsupervised category modeling, recognition, and segmentation in images. *PAMI*, 30(12):2158–2174, 2008.

[18] Z. Tu, X. Chen, A. L. Yuille, and S. C. Zhu. Image parsing: Unifying segmentation, detection, and recognition. *IJCV*, 63(2):113–140, 2005.

[19] T. Tuytelaars, C. H. Lampert, M. B. Blaschko, and W. Buntine. Unsupervised object discovery: A comparison. *IJCV*, 2009.

[20] S. Waydo and C. Koch. Unsupervised learning of individuals and categories from images. *Neural Comput.*, 20(5):1165–1178, 2008.

[21] M. Weber, M. Welling, and P. Perona. Towards automatic discovery of object categories. *CVPR*, 2:2101, 2000.

[22] J. Wu and J. M. Rehg. Where am i: Place instance and category recognition using spatial pact. In *CVPR*, 2008.

[23] Y. Wu, Z. Si, H. Gong, and S.-C. Zhu. Learning active basis model for object detection and recognition. *IJCV*, Epub ahead, 2009.

[24] B. Yao, X. Yang, and S. C. Zhu. Introduction to a large scale general purpose ground truth dataset: methodology, annotation tool, and benchmarks. In *EMMCVPR*, 2007.