Joint Multi-feature Spatial Context for Scene Recognition in the Semantic Manifold

Xinhang Song, Shuqiang Jiang, Luis Herranz
Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS)
Institute of Computer Technology, CAS, Beijing, 100190, China
{xinhang.song, shuqiang.jiang, luis.herranz}@vipl.ict.ac.cn

Abstract

In the semantic multinomial framework patches and images are modeled as points in a semantic probability simplex. Patch theme models are learned resorting to weak supervision via image labels, which leads the problem of scene categories co-occurring in this semantic space. Fortunately, each category has its own co-occurrence patterns that are consistent across the images in that category. Thus, discovering and modeling these patterns is critical to improve the recognition performance in this representation. In this paper, we observe that not only global co-occurrences at the image-level are important, but also different regions have different category co-occurrence patterns. We exploit local contextual relations to address the problem of discovering consistent co-occurrence patterns and removing noisy ones. Our hypothesis is that a less noisy semantic representation, would greatly help the classifier to model consistent co-occurrences and discriminate better between scene categories. An important advantage of modeling features in a semantic space is that this space is feature independent. Thus, we can combine multiple features and spatial neighbors in the same common space, and formulate the problem as minimizing a context-dependent energy. Experimental results show that exploiting different types of contextual relations consistently improves the recognition accuracy. In particular, larger datasets benefit more from the proposed method, leading to very competitive performance.

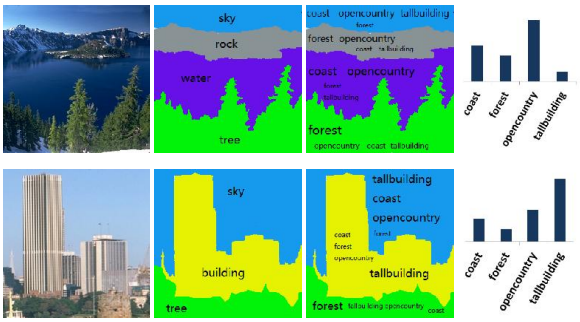
1. Introduction

Typically, a scene is a very abstract representation composed of many less abstract semantic entities localized in regions (*e.g. sky, rock, table, car*). Accurate scene recognition remains a challenge because it implies reasoning from low-level visual features to high-level scene categories. Scene categories can be modeled directly from low descriptors[36, 31, 26]. However, the required statistical knowledge to infer scene categories (*e.g. coast, mountain, office*) is difficult to obtain directly from low-level visual descriptors, due to a large semantic gap.

A more plausible approach is to split the reasoning in a two (or more) steps with smaller semantic gap (e.g. features to themes, themes to scenes). This intermediate representation is typically localized to regions in the image, and defined over a vocabulary of mid-level concepts or themes. Figure 1a-b shows an example of two images and their regions with corresponding mid-level themes. This vocabulary can be defined explicitly, but that requires labeling regions with the corresponding themes for training specific themes classifiers. Instead, themes can be modeled as hidden topics in a latent space to be discovered during learning[7, 32, 30, 17, 16]. Topics capture co-occurrences of low-level visual features, and scene categories are modeled from co-occurring topics in one image.

An alternative to (predefined or hidden) mid-level vocabularies is directly learning mid-level themes using scene category labels. Note that themes are still local, but referred to the same vocabulary as scene categories. In this paper, we focus on the *semantic multinomial (SMN)* representation[21] and its extensions[22, 23, 8]. The semantic multinomial represents the probability that a given patch (or image) belongs to each scene category. As a probability, it lies on a probability simplex (semantic simplex or semantic space). As no local annotations are available, all the patches in one image share the same label, but they correspond to different regions with different intermediate concepts. This weakly-supervised learning induces that related scene categories, sharing regions with the same mid-level concept (e.g. *sky*, *road*, *trees*) would show certain probability in the SMN, leading to categories co-occurring in the representation (see Figure 1c). We refer to this as *(scene) category co-occurrences*[1]. Rasiwasia

[1]In [23], the authors use the term *contextual co-occurrences* to refer to consistent and thus desirable co-occurrence patterns. Here, we refer to them as *(scene) category co-occurrences* to emphasize that they are high-level categories rather than low or mid-level co-occurrences. We also want



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

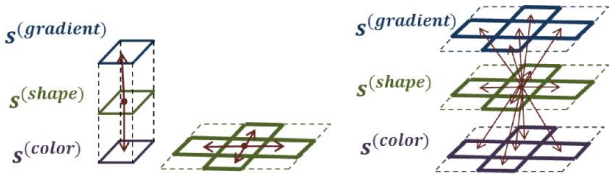
Figure 1. Types of co-occurrences in scene recognition: (a) images from the *opencountry* (top row) and *tallbuilding* (bottom row) categories of the *15 scenes* dataset, (b) regions with their corresponding mid-level themes and vocabulary, (c) scene category co-occurrences in different regions resulting from weakly-supervised learning through image category labels, and (d) the corresponding image semantic multinomial.

and Vasconcelos[23] showed that these co-occurrence patterns are consistent across the images in the same category, so they can be modeled and separated from accidental co-occurrences (i.e. noise in the semantic representation) with a suitable classifier (e.g. Dirichlet mixtures[23], SVM[8]). They also point out that patch SMNs are too noisy to model reliable co-occurrence models, so multiple patch SMNs are aggregated into a single image SMN with some caution to preserve co-occurrence patterns (see Figure 1d). Thus, these works only model *global* co-occurrence patterns in the image-level.

In contrast, we want to focus on *local* category co-occurrences in patch SMNs, as many category co-occurrences depend on the particular region (see Figure 1c). Our motivation is to exploit (in an unsupervised way) contextual relations to reinforce consistent co-occurrence patterns and remove accidental ones (i.e. noise). We consider two types of contextual relations: spatial relations between neighboring patches and multi-feature relations obtained from complementary low-level visual features (e.g. color, gradient, shape). Note that it is not possible to combine directly these two types of contextual relations, as each visual feature lies in a different low-level feature space, and so are the different feature-specific spatial contexts (see Figure 2a and b). In general they can be processed independently and then combined.

Note that SMN representations all lie in the same semantic space, independently of which visual feature they come from. This *common space* allows us to combine easily spatial relations and multi-feature relations in a single *multi-feature spatial context*. Thus, in this paper we propose a joint context model to reinforce consistent co-occurrence patterns and filter out accidental ones. We show that delivering cleaner SMNs to the classifier can help to discover intrinsic co-occurrence patterns that can model a scene category, thus improving the recognition performance.

to avoid confusion other type of context, such as the spatial neighborhood or inter-feature relations.



(a) (b) (c)

Figure 2. Contextual relations in a 4-connected neighborhood: (a) multi-feature context, (b) spatial context, and (c) joint multi-feature spatial context.

The rest of the paper is organized as follows. Sections 2 and 3 review related works and the semantic multinomial framework. The proposed joint context model is described in Section 4. Experiments are presented in Section 5 and Section 6 draws the conclusions.

2. Related work

A number of methods have proposed mid-level representations using explicit classifiers. Vogel and Schiele[29] proposed a vocabulary with nine local concepts to model natural scenes. Object bank[13, 38] is a semantic representation that encodes the response at different spatial locations of a number of pretrained object classifiers. Classemes[1] are intermediate semantic representations based on a set of 2659 basis classes. These methods require explicit intermediate level training and often exploit large amounts of external training data (e.g. ImageNet) to learn these mid-level classifiers.

Closely related to this paper, Vicept[14] and methods extending the SMN framework[21] only require to label the scene category of the image. Rasiwasia and Vasconcelos[23] propose a contextual model based on Dirichlet mixtures to model contextual co-occurrences. Kwitt et al[8] propose a discriminative version which uses SVM combined with a suitable kernel for the semantic space (i.e. the negative geodesic kernel[37]). In this case, the semantic simplex is described as a *semantic manifold* (SM). This approach is extended with rough spatial context via spatial pyramids[10] and an approximate embedding of the NGD kernel for large scale recognition is proposed. In contrast, we encode explicitly SMN relations between neighboring patches and multiple features. These works only model global co-occurrences in the image SMNs, while most of our proposed techniques focus on local co-occurrences at patch level.

Latent topics models are often modeled using Latent

Dirichlet Allocation (LDA)[7, 24]. However, most LDA have been shown to capture irrelevant general regularities rather than the semantic regularities of interest, due to poor supervision[24]. Spatial context can be included to model the global layout and enforce local coherence in the topics[32, 17]. Recently, Li and Guo[16] proposed a patch-based latent framework which jointly learns the contextual representation and the classification model. Most latent topic models are generative, and usually do not scale well to large scale datasets. Compared to topic models, our approach has two main differences. First, the vocabulary of patch SMNs is still the (high-level) scene categories, while topic models are mid-level representations. Second, the objective is to encourage contextual co-occurrences and then let the classifier disambiguated them a posteriori.

Co-occurrences at different levels are in the core of many scene understanding systems. Lang et al[9] propose a feature co-occurrence matrix useful for scene classification. Topics in computer vision model essentially low-level visual co-occurrences. Li and Guo[15] propose to segment the image into superpixels, classify them into object classes and then exploit the object co-occurrences to predict the scene category. In contrast to these types of co-occurrences, weak supervision in learning SMN representations induces a very special type of co-occurrences (i.e. category co-occurrences), which are at the highest-level of abstraction.

3. Framework overview

3.1. Semantic manifold

The semantic manifold is also based on the semantic multinomial[23] for the mid-level theme representation. The probability distribution of each category is estimated from local visual descriptors defined in some visual space X. Images are represented as a bag of local visual descriptors $I = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, $\mathbf{x}_n \in X$, densely sampled in a grid with N local patches. Given a vocabulary of scene categories $\{w_1, \ldots, w_M\}$, each image is labeled with one of those M categories. As patch labels are not available, theme conditional distributions $P_{\mathbf{X}|W}(\mathbf{x}_n|w)$ are learned using weak supervision via image labels. All the patches in a given image share the same label (i.e. scene category), which we showed that induces category co-occurrences. Themes conditional distributions are modeled as mixtures of Gaussians (GMM), one model per scene category.

For a given patch, we can obtain the vector of posterior probabilities $\mathbf{s} = (s_1, ..., s_M)^T$ with $s_w = P_{W|\mathbf{X}}(w|\mathbf{x}_n)$, which can be referred to as the *semantic multinomial* (SMN)[21] of the patch $\mathbf{x}_n$, and it lies on the (semantic) simplex Δ^{M-1}.

Multiple patch SMNs are combined into a single image SMN using a voting-based method. First, the most probable category is assigned to each patch as $w_n^* = \max_w s_{nw}$. Then a histogram is obtained by counting the occurrences of each category in the image as $o_w = |\{w_n : w_n^* = w\}|$. The image SMN $\mathbf{s}$ is obtained as

$$s_w = \Omega_w^{\text{vot}}(I) = \frac{o_w + \beta - 1}{\sum_{w=1}^{M}(o_w + \beta - 1)} \tag{1}$$

where β is a regularization parameter.

Category co-occurrences in the image SMNs are modeled using an SVM. Note that the more consistent the co-occurrence patterns are in the images in the training set, the better the classifier discriminate between categories. Rather than using conventional kernels (e.g. polynomial, RBF), a kernel designed for the particular geometry of the semantic simplex is used, based on the geodesic distance $g(\mathbf{s}, \mathbf{s}') = 2\arccos\left(\left\langle\sqrt{\mathbf{s}}, \sqrt{\mathbf{s}'}\right\rangle\right)$ where $\sqrt{\mathbf{s}}$ denotes element-wise square root. A negative geodesic distance (NGD) kernel can be defined from this distance as $k_{NGD}(\mathbf{s}, \mathbf{s}') = -g(\mathbf{s}, \mathbf{s}')$[37]. Finally, a spatial pyramid representation is used to rough encode the spatial context.

Note that using kernels limits the application of SVM classifiers to large datasets, due to computational cost. Kwitt et al[8] also propose an approximate mapping of the NGD kernel, so the same framework can be used for large scale scene recognition combined with a linear SVM.

Figure 3 shows our scene recognition framework, build upon the semantic manifold framework with some differences. First, our framework includes multiple features, for which we learn feature-specific theme models and then combine them in the semantic space. Second, prior to combining patch SMNs into a single image SMN, we process patches SMNs using a joint multi-feature spatial context model. Finally, we use a different approximation of the NGD kernel based on projecting on a lower dimensional feature space[4].

3.2. Multi-feature combination in the semantic space

Instead of a single type of visual feature, we now consider a set V of complementary ones (in our experiments $V = \{\text{gradient, shape, color}\}$). Each feature $v \in V$ generates a set of local visual descriptors $I^{(v)} = \{\mathbf{x}_1^{(v)}, \ldots, \mathbf{x}_N^{(v)}\}$, $\mathbf{x}_n^{(v)} \in \mathbf{X}^{(v)}$, and $I = \{I^{(1)}, \ldots, I^{(|V|)}\}$ represents all the features in the image. Now we assume that we learn feature-specific theme models $P_{\mathbf{X}^{(v)}|W}\left(\mathbf{x}_n^{(v)}|w^{(v)}\right)$, learned independently in the same way as in the single feature case. Thus, we can define the feature-specific patch SMN of the patch n and the feature v as $\mathbf{s}_n^{(v)} = (s_{n1}^{(v)}, ..., s_{nM}^{(v)})^T$. Figure 4 shows an example with three feature-dependent patch SMNs. In this figure we can observe how certain regions are noisier than others in some features. We can also observe certain patterns across

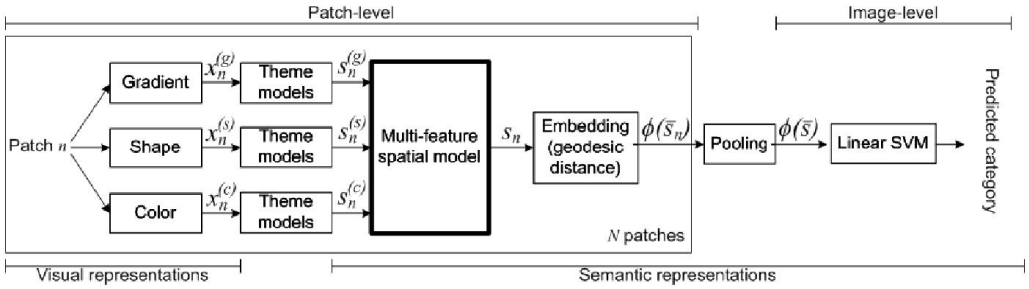


Figure 3. Overview of the recognition framework with the proposed methods highlighted.

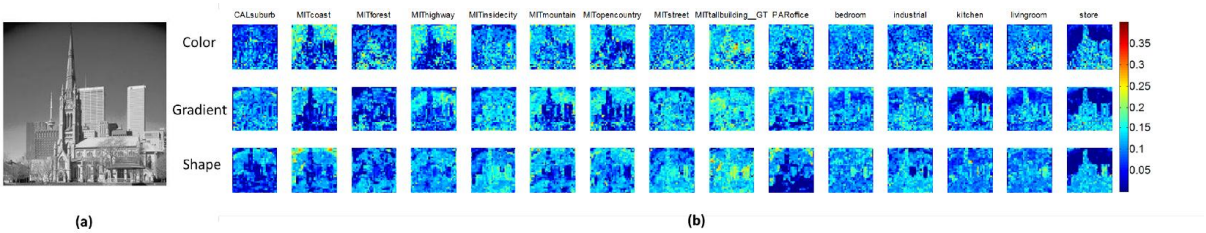


Figure 4. Patch SMNs: (a) image of the *15 scenes* dataset (category: *MITtallbuilding*), and (b) probability maps illustrating each component of the patch SMNs. Each row corresponds to SMNs obtained for a different visual descriptor.

categories (category co-occurrences), across features (inter-feature relations) and between neighboring patches (spatial relations).

Note that all SMNs lie in the same (semantic) space and all represent a probability, so we can combine them using probability models. In particular, we obtain the multi-feature SMN from several feature-dependent SMNs as a representative SMN closer to all of them. A suitable choice for probability distributions is minimizing the Kullback-Leibler (KL) divergence[23] as

$$\mathbf{s}_n = \underset{\hat{\mathbf{s}}_n}{argmin} \sum_{v \in V} KL(\mathbf{s}_n^{(v)} || \hat{\mathbf{s}}_n) \quad (2)$$

which results in

$$s_{nw} = \frac{\exp(\frac{1}{|V|}\sum_{v\in V}\log(s_{nw}^{(v)}))}{\sum_{w\in W}\exp(\frac{1}{|V|}\sum_{v\in V}\log(s_{nw}^{(v)}))} \quad (3)$$

4. Multi-feature Spatial Context Model

4.1. Global models

To exploit the spatial context, we consider the relations between neighboring patches. In contrast to the feature-dependent SMNs, we could use a similar approach as the one used, but here each patch is not independent. So here we resort to undirected models.

We first formulate the problem as denoising patches SMNs using a Markov Random Field (MRF), with a 4-connectivity grid (see Figure 5b). Considering

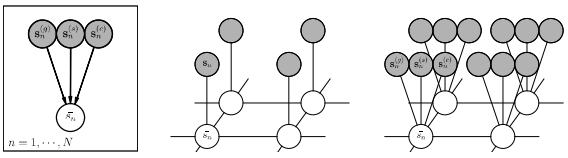
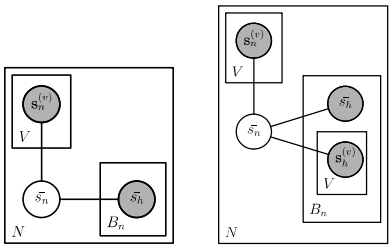


Figure 5. Contextual models: (a) multi-feature combination, (b) 4-connected spatial grid model, and (c) multi-feature spatial grid model.

a single feature, the objective is to maximize the joint probability over the observed SMNs and the denoised SMNs set defined as $P(\bar{\mathbf{s}_1}, \ldots \bar{\mathbf{s}_N}, \mathbf{s}_1, \ldots \mathbf{s}_N) = \frac{1}{Z} exp\left(-E\left(\bar{\mathbf{s}_1}, \ldots \bar{\mathbf{s}_N}, \mathbf{s}_1, \ldots \mathbf{s}_N\right)\right)$, where Z is the partition function to normalize the probability. Thus, the problem is equivalent to minimizing the global energy of the network modeled as

$$E(\bar{\mathbf{s}_1}, \ldots \bar{\mathbf{s}_N}, \mathbf{s}_1, \ldots \mathbf{s}_N) = \sum_n g(\bar{\mathbf{s}_n}, \mathbf{s}_n) + \alpha \sum_{\{n,n'\}} g(\bar{\mathbf{s}_n}, \bar{\mathbf{s}_{n'}}) \quad (4)$$

where $\bar{s_n}$ is the unknown denoised SMN of patch n (in contrast to the original s_n) and $\{n, n'\}$ represents pairs of connected patches. We model the energy as distance between SMNs. A suitable choice for probability simplices is the geodesic distance $g(\mathbf{s}, \mathbf{s}')$[37]. We chose it over the KL divergence used in (2) because KL divergence is asymmetric, and in the semantic manifold framework has been proved effective[8].



(a) (b)

Figure 6. Local patch contextual models (joint multi-feature spatial): (a) only features from the target patch, and (b) features from all the patches in the neighborhoods of target patch

As both feature-dependent SMNs and the denoised SMNs are in the same space, this model can be easily extended to multiple features using the model in Figure 5c. The corresponding energy is

$$E\left(\bar{\mathbf{s}_1},\dots\bar{\mathbf{s}_N},\mathbf{s}_1^{(1)},\dots\mathbf{s}_N^{(1)},\mathbf{s}_1^{(|V|)},\dots\mathbf{s}_N^{(|V|)}\right) = \sum_n\sum_{v\in V} g\left(\bar{\mathbf{s}_n},\mathbf{s}_n^{(v)}\right) + \alpha\sum_{\{n,n'\}} g\left(\bar{\mathbf{s}_n},\bar{\mathbf{s}_{n'}}\right) \quad (5)$$

To solve the optimization problem, we resort to the Iterative Conditional Modes (ICM) algorithm[2], which loops over the different patches minimizing the energy related with one variable keeping the other variable nodes fixed. It can be seen as coordinate-wise gradient descent. This algorithm converges to a local maximum of the probability. Other algorithms can be used, such as graph cuts, but their extension to larger neighborhoods without pairwise cliques is difficult, computationally more expensive and they do not lead to the local formulation of ICM.

4.2. Local models

The ICM algorithm updates the value of each patch by minimizing locally the related energy, keeping fixed the value of other patch variables. Now we can define the neighborhood B_n as the set of neighbors of the patch n. In the case of Figure 5b, B_n contains four neighbors. Now we can reformulate the model as N independent patch-centred subgraphs (see Figure 6a, where all $\bar{\mathbf{s}_h}$ $(h \neq n)$ are consider observed for a particular patch n) , and

$$E\left(\bar{\mathbf{s}_n};\phi_n\right) = \frac{1}{|V|}\sum_{v\in V} g\left(\bar{\mathbf{s}_n},\mathbf{s}_n^{(v)}\right) + \alpha\frac{1}{|B_n|}\sum_{\{n,h\},h\in B_n} g\left(\bar{\mathbf{s}_n},\bar{\mathbf{s}_h}\right) \quad (6)$$

where $\phi_n = \left\{\mathbf{s}_n^{(v)}|\forall v\in V\right\}\bigcup\{\mathbf{s}_h|\forall h\in B_n\}$ is the set of SMNs in the multi-feature spatial neighborhood of the patch n. For convenience we also normalize by the size of the neighborhood $|B_n|$ and the number of features $|V|$. Including larger neighborhoods with the global models of the previous section would lead to some factors being no longer pairwise, and the complexity of the problem increases significantly. However, by using this local approximation we can easily include larger neighborhoods.

We also consider an extended context, which not only considers feature-dependent SMNs from the target patch, but also from the neighbors (the graphical model is shown in Figure 6b).

$$E\left(\bar{\mathbf{s}_n};\phi_n\right) = \frac{1}{|V|}\sum_{v\in V} g(\bar{\mathbf{s}_n},\mathbf{s}_n^{(v)}) + \alpha\frac{1}{|B_n|}\sum_{\{n,h\},h\in B_n} g(\bar{\mathbf{s}_n},\bar{\mathbf{s}_h}) + \beta\frac{1}{|B_n||V|}\sum_{\{n,h\},h\in B_n}\sum_{v\in V} g(\bar{\mathbf{s}_n},\mathbf{s}_h^{(v)}) \quad (7)$$

now with $\phi_n = \left\{\mathbf{s}_n^{(v)}|\forall v\in V\right\}\bigcup\left\{\mathbf{s}_h^{(v)}|\forall h\in B_n,\forall v\in V\right\}$.

Finally, we include an additional term in the energy to penalize too flat SMNs, which would lead to uninformative patches:

$$E'\left(\bar{\mathbf{s}_n};\phi_n\right) = E\left(\bar{\mathbf{s}_n};\phi_n\right) + \lambda H\left(\bar{\mathbf{s}_n}\right) \quad (8)$$

where $H\left(\mathbf{s}\right) = -\sum_{w=1}^{M} s_w\log\left(s_w\right)$ is the entropy of $\mathbf{s}$.

Following the same idea of the ICM algorithm, we loop over the patches minimizing (8) for each patch n. This problem can be solved using gradient descent. The gradient corresponding to the patch n is

$$\frac{\partial E'\left(\bar{\mathbf{s}_n};\phi_n\right)}{\partial s_{nw}} = \frac{1}{|V|}\sum_{v\in V} f\left(\bar{s_{nw}},s_{nw}^{(v)}\right) + \alpha\frac{1}{|B_n|}\sum_{\{n,h\},h\in B_n} f\left(\bar{s_{nw}},\bar{s_{hw}}\right) + \beta\frac{1}{|B_n||V|}\sum_{\{n,h\},h\in B_n}\sum_{v\in V} f\left(\bar{s_{nw}},s_{hw}^{(v)}\right) - \gamma(1+\log(\bar{s_{nw_n}})) \quad (9)$$

where

$$f\left(x,y\right) = \frac{\partial g\left(x,y\right)}{\partial x} = -\frac{\sqrt{y}}{2\sqrt{x}\sqrt{1-(\sqrt{x}\sqrt{y})^2}}$$

5. Experiments

In this section we evaluate the different context models described previously over different dataset, comparing with the related works.

5.1. Experimental setup

Datasets. The proposed methods are evaluated on three small datasets. *15 scenes* [7, 10] contains 4485 images across 15 scene categories. *LabelMe*[18] consists of 8 outdoor scene categories, with a total of 2600 images. *UIUC-Sports*[12] consists of 1585 images labeled into 8 complex sport scene categories. Following settings in previous works, we use 100, 100 and 70 images for training, respectively. We also evaluate the proposed methods on larger scale datasets, including MIT67[20] and SUN397[34]. MIT67 contains 15620 images of 67 indoor scene classes. SUN397[34] consists of 397 categories, with 108762 images in total. In the case of MIT67 Indoor and SUN397, the training/testing configurations are provided by the original authors.

Visual and semantic features. We use three kinds of kernel descriptors[3] as the local descriptors, including gradient, shape (LBP) and color. All local visual descriptors are extracted on a regular 16×16 pixel dense grid (step 8 pixels). For themes, we train GMMs with 512 mixtures for each theme model. We also extend the descriptor using a spatial pyramid[10] with four levels (1×1, 2×2, 3×3, 4×4) for SVM classification.

Baselines. We compare our approach with the same framework without context model, which is equivalent to the spatial pyramid semantic manifold (SPMSM)[8], just using a different approximate embedding[4]. We evaluate it independently over the same visual features.

Variations of the proposed methods. We evaluate four variations of the proposed context models:

- *Multi-feature context* (MF): multiple features are combined in the semantic space using (3), corresponding to the context in Figure 2a and the model in Figure 5a.
- *Spatial context* (S): single feature neighboring relations (see Figure 2b). Obtained by minimizing (6) when only one feature is used.
- *Multi-feature spatial context* (MFS): combines multiple-features of the target patch and neighboring relations (i.e. the combination of Figure 2a and b). Obtained by minimizing (6) in the multi-feature case.
- *Extended multi-feature spatial context* (EMFS): also includes multiple-features from patches in the neighborhood (see Figure 2c). Obtained by minimizing (6) and corresponding to the model in Figure 6b.

5.2. Impact of the neighborhood size and entropy regularization

Two critical parameters are the size of the spatial neighborhood and the entropy regularization, as it has impact on the co-occurrence patterns. We evaluate them on the 15 scenes dataset, using the EMFS setting and fixing α and β to 1. The results are illustrated in Figure 8.

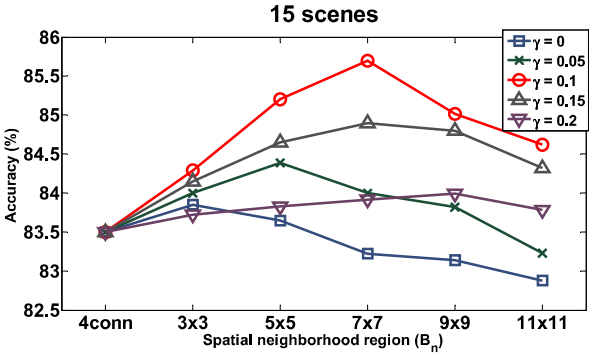


Figure 8. Region size and sparse parameter evaluation

We evaluate different neighborhoods, including the 4-connectivity spatial neighborhood shown in Figure 2b and c, and other dense neighborhoods of size LxL patches (3x3 corresponds to 8 neighbors). We can observe that larger neighborhoods can effectively reinforce consistent patterns and filter accidental ones. However, too large neighborhoods cannot capture properly local co-occurrence patterns. From our experiments, a good trade-off is 7x7 patches.

We also evaluate the impact of the entropy regularization varying γ from 0 to 0.2, with a step of 0.05. In general, the performance increases with γ with maximum around 0.1 and then decreases. Figure 7 illustrates the effect of entropy on the patch SMNs. Without penalizing the entropy ($\gamma = 0$) we obtain too flat patch SMNs (i.e. high entropy) which are not suitable for co-occurrence modeling. Too low entropy in the patch SMNs is not useful either ($\gamma = 0.2$), as one category may dominate with a too high probability, there are few consistent co-occurrence patterns. The best results in this experiment were achieved for $\gamma = 0.1$ and $L = 7$. For the rest of the experiments we will use this configuration, although specific parameters may improve the performance.

5.3. Context models

We evaluate the different variations of the proposed methods on the three small scale datasets to show how different types of context models improve the accuracy. Table 1 shows that the classification accuracy increases consistently when we model different types of context. Combining multiple features help with a gain around 1.1-2.5% over the best single feature. Spatial context is more variable and varies from no gain to modest gains around 1%. However, combining both can increase an additional 0.5-1% over only multi-feature context. The extended multi-feature spatial context contributes with an additional 0.5-1% gain by incorporating multiple features from the neighboring patches. The total gain with the extended context model over the baseline with no context is around 2.6-5.7%.

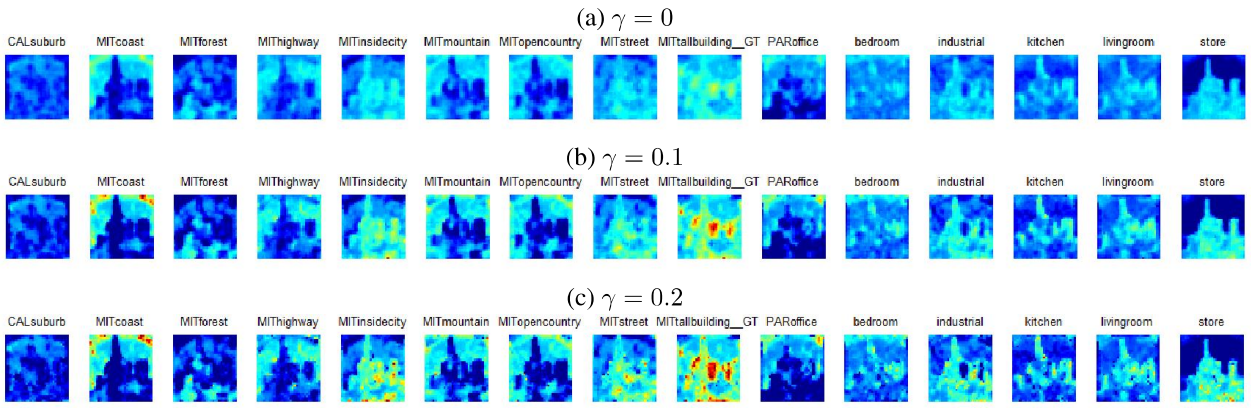


Figure 7. Effect of entropy regularization on the patch SMNs. The spatial neighborhood is 3x3 patches.

Table 1. Accuracy (%) for different context models.

Method (feature)	15 scenes	LabelMe	Sports
No context model			
Baseline (gradient)	78.9	86.5	83.9
Baseline (shape)	80.0	85.0	84.3
Baseline (color)	75.4	72.4	72.8
Spatial context (7x7 patches)			
Spatial (Gradient)	81.0	86.7	83.7
Spatial (Shape)	81.4	84.9	83.9
Spatial (Color)	76.6	72.9	73.1
Multiple feature context			
Multi-feature	82.5	88.3	85.4
Joint multi-feature spatial context (7x7 patches, $\gamma = 0.1$)			
MFS	83.5	88.9	85.9
Extended MFS	**85.7**	**89.3**	**86.9**

Table 2. Comparison with related works.

Dataset	Method	Accuracy (%)
15 scenes	SMN[23]	71.7
	LDA[24]	76.6
	CMN[23]	77.2
	ObjectBank[16]	80.9
	Kernel descriptor[3]*	82.2
	SPMSM[8]	82.5
	SR-LSR[16]	85.7
	Proposed (EMFS)	**85.7**
	Object-to-Class kernels[38]	88.8
LabelMe	Wang et al[30]	76.0
	SPMSM[8]	87.5
	Kernel descriptor[3]*	87.3
	Proposed (EMFS)	**89.3**
	SR-LSR[16]	89.8
Sports	Li and Fei-Fei[12]	73.4
	ObjectBank[13]	76.3
	SPMSM[8]	83.0
	SR-LSR[16]	83.9
	Kernel descriptor[3]*	85.2
	Object-to-Class kernels[38]	86.0
	Proposed (EMFS)	**86.9**

* Results are based on our own implementation using the code available from the authors.

5.4. Comparison with related works

We compare with recent works using mid-level semantic representations, such as latent topic models[12, 30, 15, 24, 16], extensions of the SMN framework[23, 8] and others such as extensions of Object bank[13, 38] and Classemes[28, 1]. Most of these approaches cannot be used in large scale datasets, so we separate comparisons for small datasets and larger datasets.

5.4.1 Small scale datasets

Table 1 compares the results reported by the authors in their corresponding references. Although a completely fair comparison is not possible, due to different implementations, features and other parameters, our framework at least seems to be very competitive in the three evaluated datasets. Comparing with methods based on SMNs is of particular interest. Note that the contextual multinomial (CMN) exploits CCO at the image level SMN using a generative model, Dirichlet mixture models (DMM). SPMSM exploits discriminative classification and rough spatial context, achieving better performance. The proposed method, which also exploits multiple features and local context in patch level achieves better performance than those methods. We also compare with modeling categories directly from the same low-level kernel descriptors (concatenated to combine them), with and a SVM and spatial pyramid. We observe that our method, which uses a mid-level representation achieves better results.

Table 3. Comparison on MIT67 dataset.

MIT67	Method	Acc (%)
Proposed	Baseline (gradient)	34.7
	Baseline (shape)	36.9
	Baseline (color)	26.8
	Proposed (MF)	42.4
	Proposed (MFS)	44.7
	Proposed (EMFS)	**48.2**
State-of-the-art	ObjectBank[16]	37.6
	Object-to-Class kernels[38]	39.6
	Deformable Part Models[19]	43.1
	SPMSM[8]	44.0
	Sparse Spatial Coding[11]	44.4
	Geometric Phrase Pooling[35]	46.4
	Linear Distance Coding[33]	46.7
	IFV[27]	60.8
	Discriminative parts[5]	64.0
	Places-CNN[39]	68.2
	CNNaug-SVM [25]	**69.0**

Table 4. Comparison on SUN397 dataset.

SUN397	Method	Accuracy (%)
Proposed	Baseline (gradient)	25.4
	Baseline (shape)	23.2
	Baseline (color)	18.2
	Proposed (MF)	30.4
	Proposed (MFS)	34.9
	Proposed (EMFS)	**40.7**
State-of-the-art	SUN (HOG)[34]	27.2
	SPMSM[8]	28.2
	Meta-classes[1]	36.8
	SUN(MKL)[34]	38.0
	CNN (Decaf)[6]	40.9
	IFV[27]	47.2
	Places-CNN**[39]**	**54.3**

5.4.2 Large scale datasets

We evaluate the proposed methods on the medium scale dataset MIT67 and the much larger SUN397. The results are shown in Tables 3 and 4, respectively. The gains due to incorporating different contexts are much higher than in smaller datasets, for significant gains of 11% and 15% over the best single feature baseline for the MIT67 and SUN397 datasets, respectively. This suggests that contextual relations become much more important important as the number of scene categories increases, resulting in co-occurrence patterns much noisier in these larger datasets. Exploiting the context to emphasize representative category co-occurrence patterns can greatly help to improve the recognition performance. Other mid-level semantic representations, such as Object-bank and Meta-classes exploit larger amounts of external data (e.g.. ImageNet) to model the mid-level classifiers. The proposed method outperforms them without resorting to external data.

As the number of mid-level representation approaches that can be trained for these datasets is limited, as a reference we also compare with recent works based on coding in the bag-of-words framework[33, 11, 35], Fisher vector[27], mining discriminative parts[5] and convolutional neural networks (CNN)[6, 39, 25]. We achieve slightly better performance than LLC but not Fisher vector. Note however than the result in [27] uses a much denser grid for sampling local features resulting in a much higher dimensional feature. Mining discriminative parts also achieves better performance in MIT67. This dataset contains indoor scenes with many objects where part-based representations can achieve good performance. Even being a purely scene-level representation, our approach still achieves competitive performance in both datasets. We achieve comparable performance to CNN features learned on ImageNet[6] but not on Places[39], since this dataset is scene-centric and thus more suitable. Note that, in contrast to CNN, we do not make use of any external data.

6. Conclusions

Intermediate semantic spaces are very helpful to recognize complex scenes. In contrast to topic models exploiting low and mid-level feature co-occurrences, we focus on a special type of pattern resulting from learning local theme models with weak supervision. Exploiting these patterns (i.e. scene category co-occurrences) properly can boost the recognition performance.

We extend the semantic manifold framework[8] by including a context model integrating multiple features and neighboring patches. We exploit the property that the semantic simplex is a common space where multiple features and neighboring patches can be naturally integrated. A joint context model exploiting these relations is critical to improve the performance in this framework. In particular, large datasets benefit more from the proposed context models, as the number of classes is higher and useful category co-occurrence patterns are more subtle and hidden in noisy patterns. Exploiting local spatial and multi-feature relations can help to discover consistent patterns and filter out noisy patterns, making things easier to the classifier which can focus on modeling these patterns.

Acknowledgment. This work was supported in part by National Basic Research Program of China (973 Program): 2012CB316400, in part by the National Natural Science Foundation of China: 61322212 and 61450110446, in part by National Hi-Tech Development Program (863 Program) of China: 2014AA015202, in part by the Key Technologies R&D Program of China under Grant no. 2012BAH18B02, and in part by the CAS President's International Fellowship Initiative: 2011Y1GB05 . This work is also funded by

Lenovo Outstanding Young Scientists Program (LOYS).

References

[1] A. Bergamo and L. Torresani. Classemes and other classifier-based features for efficient object categorization. In *IEEE Trans. on Pattern Anal. and Mach. Intell.*, 2014.

[2] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B*, 48(3):259–302, 1986.

[3] L. Bo, X. Ren, and D. Fox. Kernel descriptors for visual recognition. In *NIPS*, 2010.

[4] L. Bo and C. Sminchisescu. Efficient match kernel between sets of features for visual recognition. In *NIPS*, 2009.

[5] C. Doersch, A. Gupta, and A. A. Efros. Mid-level visual element discovery as discriminative mode seeking. In *NIPS*, pages 494–502. 2013.

[6] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. In *ICML*, 2014.

[7] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.

[8] R. Kwitt, N. Vasconcelos, and N. Rasiwasia. Scene recognition on the semantic manifold. In *ECCV*, 2012.

[9] H. Lang, Y. Xi, J. Hu, L. Du, and H. Ling. Scene classification by feature co-occurrence matrix. In *Workshop on Scene Understanding for Autonomous System, ACCV*, 2014.

[10] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[11] G. Leivas Oliveira, E. Nascimento, A. Wilson Vieira, and M. Montenegro Campos. Sparse spatial coding: A novel approach to visual recognition. *IEEE Trans. on Image Process.*, 23(6):2719–2731, June 2014.

[12] L. Li and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. In *ICCV*, 2007.

[13] L. Li, H.Su, E. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *NIPS*, 2010.

[14] L. Li, S. Jiang, and Q. Huang. Learning hierarchical semantic description via mixed-norm regularization for image understanding. *IEEE Trans. on Multimedia*, 14(5), 2012.

[15] X. Li and Y. Guo. An object co-occurrence assisted hierarchical model for scene understanding. In *British Machine Vision Conference*, pages 1–11, 2012.

[16] X. Li and Y. Guo. Latent semantic representation learning for scene classification. In *ICML*, 2014.

[17] Z. Niu, G. Hua, X. Gao, and Q. Tian. Context aware topic model for scene recognition. In *CVPR*, 2012.

[18] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, 42(3):145–175, May 2001.

[19] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *ICCV*, 2011.

[20] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR*, 2009.

[21] N. Rasiwasia and N. Vasconcelos. Bridging the gap: Query by semantic example. *IEEE Trans. on Multimedia*, 9(5):923–938, 2007.

[22] N. Rasiwasia and N. Vasconcelos. Holistic context modeling using semantic co-occurrences. In *CVPR*, pages 1889–1895, 2009.

[23] N. Rasiwasia and N. Vasconcelos. Holistic context models for visual recognition. *IEEE Trans. on Pattern Anal. and Mach. Intell.*, 34(5):902–917, 2012.

[24] N. Rasiwasia and N. Vasconcelos. Latent dirichlet allocation models for image classification. *IEEE Trans. on Pattern Anal. and Mach. Intell.*, 35(11):2665–2679, 2013.

[25] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *CVPR*, 2014.

[26] J. Sanchez and F. Perronnin. High-dimensional signature compression for large-scale image classification. In *Neural Comput.*, 2011.

[27] J. Sanchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *Int. J. Comput. Vision*, 105(3):222–245, 2013.

[28] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. In *ECCV*, 2010.

[29] J. Vogel and B. Schiele. Semantic modeling of natural scenes for content-based image retrieval. *Int. J. Comput. Vision*, 72(2):133–157, Apr. 2007.

[30] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, pages 1903–1910, 2009.

[31] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010.

[32] X. Wang and E. Grimson. Spatial latent dirichlet allocation. In *NIPS*, 2007.

[33] Z. Wang, J. Feng, S. Yan, and H. Xi. Linear distance coding for image classification. *IEEE Trans. on Image Process.*, 22(2):537–548, Feb 2013.

[34] J. Xiao, J. Hayes, K. Ehringer, A. Olivia, and A. Torralba. SUN database: Largescale scene recognition from abbey to zoo. In *CVPR*, 2010.

[35] L. Xie, Q. Tian, M. Wang, and B. Zhang. Spatial pooling of heterogeneous features for image classification. *IEEE Trans. on Image Process.*, 23(5):1994–2008, May 2014.

[36] J. Yang, K. Yu, Y. Gong, and T. S. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.

[37] D. Zhang, X. Chen, and W. S. Lee. Text classification with kernels on the multinomial manifold. In *RDIR*, pages 266–273, 2005.

[38] L. Zhang, X. Zhen, and L. Shao. Learning object-to-class kernels for scene classification. *IEEE Trans. on Image Process.*, 23(8):3241–3253, Aug 2014.

[39] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *NIPS*, pages 487–495, 2014.