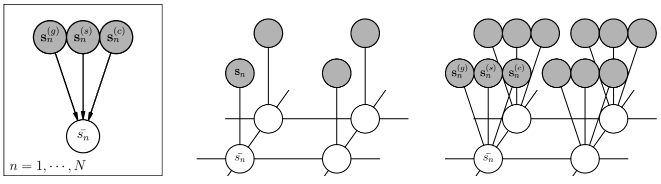
# Joint Multi-feature Spatial Context for Scene Recognition in the Semantic Manifold

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.



(a) (b) (c)

Figure 1: Contextual models: (a) multi-feature combination, (b) 4-connected spatial grid model, and (c) multi-feature spatial grid model.

In the semantic multinomial[2, 3] 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 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.

The common space is referred to as the semantic manifold[1] based on *semantic multinomial* (SMN)[2]. 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 SMN of the patch $\mathbf{x}_n$, and it lies on the (semantic) simplex $\Delta^{M-1}$. To exploit the spatial context, we consider the relations between neighboring patches. We firstly formulate the problem as denoising patch SMNs using a Markov Random Field (MRF), with a 4-connectivity grid (see Figure 1b). Considering a single feature, the objective is to maximize the joint probability over the observed SMNs and the denoised SMNs set defined as $P(\bar{\mathbf{s}}_1, \ldots \bar{\mathbf{s}}_N, \mathbf{s}_1, \ldots \mathbf{s}_N) = \frac{1}{Z} exp(-E(\bar{\mathbf{s}}_1, \ldots \bar{\mathbf{s}}_N, \mathbf{s}_1, \ldots \mathbf{s}_N))$, 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

$$E(\bar{\mathbf{s}}_1, \ldots \bar{\mathbf{s}}_N, \mathbf{s}_1, \ldots \mathbf{s}_N) = \sum_n g(\bar{\mathbf{s}}_n, \mathbf{s}_n) + \alpha \sum_{\{n,n'\}} g(\bar{\mathbf{s}}_n, \bar{\mathbf{s}}_{n'}) \qquad (1)$$

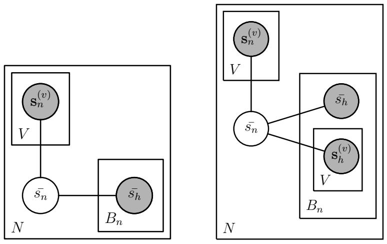
where $\bar{s}_n$ is the unknown denoised SMN of patch $n$ (in contrast to the original $s_n$) and $\{n, n'\}$ represents pairs of connected patches. We model the energy as distance between SMNs. A suitable choice for probability simplices is the geodesic distance $g(\mathbf{s}, \mathbf{s}')$[4].

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 1c. The corresponding energy is

$$E\left(\bar{\mathbf{s}}_1, \ldots \bar{\mathbf{s}}_N, \mathbf{s}_1^{(|V|)}, \ldots \mathbf{s}_N^{(|V|)}\right) = \sum_n \sum_{v \in V} g\left(\bar{\mathbf{s}}_n, \mathbf{s}_n^{(v)}\right) + \alpha \sum_{\{n,n'\}} g(\bar{\mathbf{s}}_n, \bar{\mathbf{s}}_{n'}) \qquad (2)$$

To solve the optimization problem, we resort to the Iterative Conditional Modes (ICM) algorithm, which loops over the different patches minimizing the energy related with one variable keeping the other variable nodes fixed.

The ICM algorithm updates the value of each patch by minimizing locally the related energy, keeping fixed the value of other patch variables.



(a) (b)

Figure 2: 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.

Now we can define the neighborhood $B_n$ as the set of neighbors of the patch $n$. In the case of Figure 1b, $B_n$ contains four neighbors. Now we can reformulate the model as $N$ independent patch-centred subgraphs (see Figure 2a, where all $\bar{\mathbf{s}}_h$ ($h \neq n$) are consider observed for a particular patch $n$) , and

$$E(\bar{\mathbf{s}}_n; \phi_n) = \frac{1}{|V|} \sum_{v \in V} g\left(\bar{\mathbf{s}}_n, \mathbf{s}_n^{(v)}\right) + \alpha \frac{1}{|B_n|} \sum_{\{n,h\}, h \in B_n} g(\bar{\mathbf{s}}_n, \bar{\mathbf{s}}_h) \qquad (3)$$

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 $|V|$. 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 2b).

$$E(\bar{\mathbf{s}}_n; \phi_n) = \frac{1}{|V|} \sum_{v \in V} g(\bar{\mathbf{s}}_n, \mathbf{s}_n^{(v)}) +$$
$$\alpha \frac{1}{|B_n|} \sum_{\{n,h\}, h \in B_n} g(\bar{\mathbf{s}}_n, \bar{\mathbf{s}}_h) + \beta \frac{1}{|B_n||V|} \sum_{\{n,h\}, h \in B_n} \sum_{v \in V} g(\bar{\mathbf{s}}_n, \mathbf{s}_h^{(v)}) \qquad (4)$$

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:

$$E'(\bar{\mathbf{s}}_n; \phi_n) = E(\bar{\mathbf{s}}_n; \phi_n) + \lambda H(\bar{\mathbf{s}}_n) \qquad (5)$$

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

Following the same idea of the ICM algorithm, we loop over the patches minimizing (5) for each patch $n$. This problem can be solved using gradient descent. The gradient corresponding to the patch $n$ is

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

where

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

[1] Roland Kwitt, Nuno Vasconcelos, and Nikhil Rasiwasia. Scene recognition on the semantic manifold. In *ECCV*, 2012.

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

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

[4] Dell Zhang, Xi Chen, and Wee Sun Lee. Text classification with kernels on the multinomial manifold. In *RDIR*, pages 266–273, 2005. ISBN 1-59593-034-5. doi: 10.1145/1076034.1076081.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.