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

Xinhang Song, Shuqiang Jiang, Luis Herranz

IEEE 2015 Conference on Computer Vision and Pattern Recognition

Overview:

➢ Two steps scene recognition

- Images are represented with local intermediate representations, which are defined over a vocabulary of mid-level concepts or themes.
- Global scene categories are modeled from these local intermediate representations.

➢ Semantic multinomial (SMN) representation

- Idea: instead of using a different mid-level vocabulary for patches, we use directly the same **scene categories**.
- Patches are **weakly-labeled** with scene categories
  - No need for additional labels for regions
- Compared with topic models: no need to discover topics, two independent steps (simpler and more scalable)

➢ Scene category co-occurrences

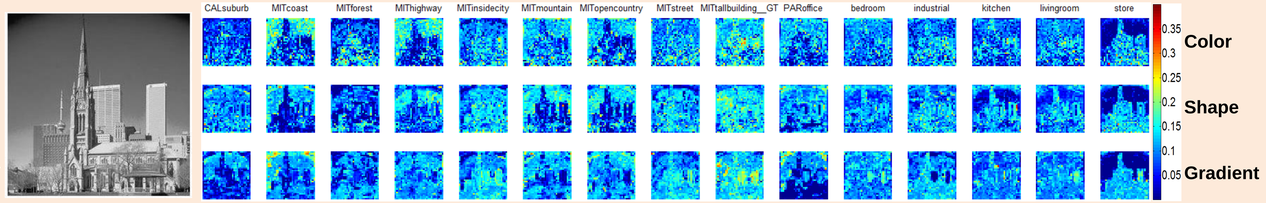
- Weak-labeling patches with image labels causes the problem of **scene category co-occurrences**

- Observation 1: co-occurrence patterns are **consistent** across the images in the same category;

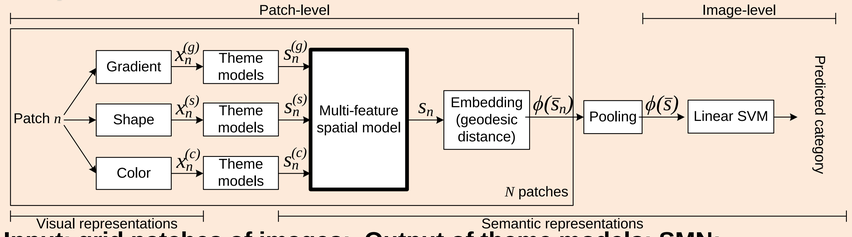
➢ Goal: model consistent co-occurrences

- Previous works [1,2]: global category co-occurrences
- Observation 2: the SMN space is **common** for all representations (independently of the original visual feature)
- Our goal: exploit this common space to model local and multi-feature category co-occurrences

Multiple feature patch SMN maps: 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.



Proposed Framework



Input: grid patches of images; Output of theme models: SMN;

Multi-feature and Spatial Context Model:

➢ Multi-feature context

- Motivation: SMNs of same patches learned from different visual features should be similar (*inter-feature co-occurrences*).
- Context model: multi-feature SMN combination.

➢ Spatial context

- Motivation: Neighboring patches often represent similar concepts (*local co-occurrences*).
- Context model: Markov Random Field on patch SMNs;

Joint Context Model:

➢ Joint Context Model:

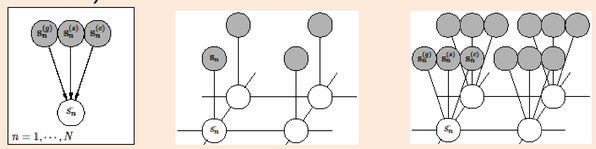
$$E(\bar{s_n};\emptyset_n)=\frac{1}{|V|}\sum_{v\in V} g(\bar{s_n},s_n^v)+\alpha\frac{1}{|B_n|}\sum_{\{n,h\},h\in B_n} g(\bar{s_n},\bar{s_h})$$

E is the energy function. s_n^v is the SMN learned from visual feature *v*. *g(x,y)* is the geodesic distance [2]. $|B_n|$ is the neighbors of patch *n*.

➢ Extended model

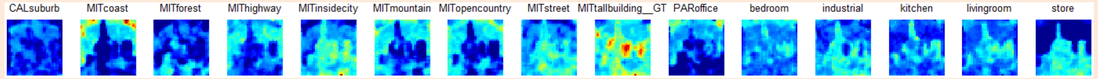
$$E(\bar{s_n};\emptyset_n)=\frac{1}{|V|}\sum_{v\in V} g(\bar{s_n},s_n^v)+\alpha\frac{1}{|B_n|}\sum_{\{n,h\},h\in B_n} g(\bar{s_n},\bar{s_h})+\beta\frac{1}{|B_n||V|}\sum_{\{n,h\},h\in B_n}\sum_{v\in V} g(\bar{s_n},s_h^v)$$

We extend the neighborhood $|B_n|$ to the $|B_n||V|$ (the neighbors of all different features)



Multi-feature Spatial Joint multi-feature and spatial

Experiments and References:



[1] N. Rasiwasia and N. Vasconcelos. Holistic context models for visual recognition. *IEEE Trans. on Pattern Anal. and Mach.Intell*,34(5):902–917, 2012.
[2] R. Kwitt, N. Vasconcelos, and N. Rasiwasia. Scene recognition on the semantic manifold. In *ECCV*, 2012.

No context (**NC**), spatial context (S), multi-feature context (**MF**), multi-feature spatial context (**MFS**) and extended multi-feature spatial context (**EMFS**).

	Scene15	LabelMe	Sports	MIT67	SUN397
NC	78.9	86.5	83.9	34.7	25.4
S	81.0	86.7	84.3	-	-
MF	82.5	85.4	72.8	42.4	30.4
MFS	83.5	85.9	85.9	44.7	34.9
EMFS	**85.7**	**89.3**	**86.9**	**48.2**	**40.7**