Category co-occurrence modeling for large scale scene recognition

Xinhang Song[1], Shuqiang Jiang[1], Luis Herranz[1], Yan Kong[2], Kai Zheng[3]

[1]Institute of Computer Technology, Chinese Academy of Sciences (CAS) [2]Institute of Automation, CAS [3]School of Computer Science, Soochow University, China

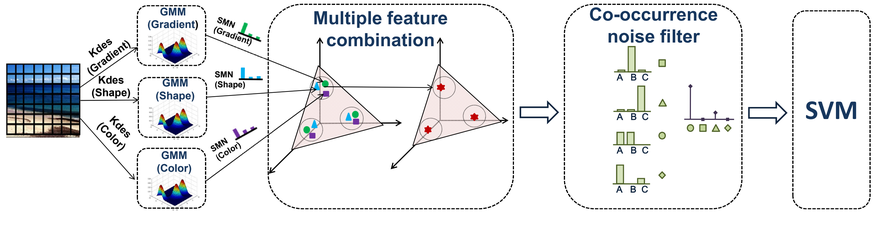
Published in Pattern Recognition 2016

SUNw: Scene Understanding Workshop

CVPR 2016

Introduction

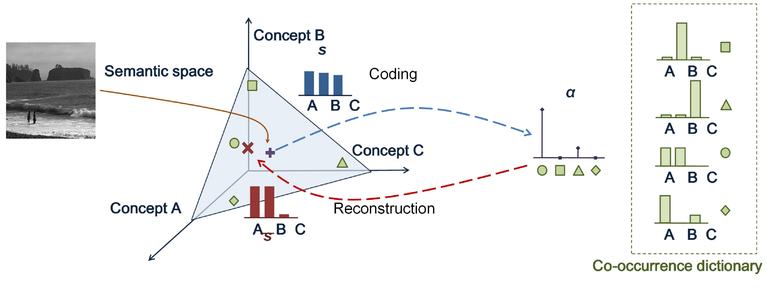
- **SMN-based scene recognition [1,2]**
  - Semantic multinomial (SMN) representation [1]: probability vector $\boldsymbol{s} = (s_1, \dots, s_w, \dots, s_W)$
    where s_w is the probability that an image belongs to category w
  - Patch models are learned using scene labels (weak supervision)
    - Benefits: image and patches share the same label, **not requiring patch level label** nor **discovering latent topics**
    - Problems: leads to **ambiguity** due to co-occurrences between scene categories [1].
- **Types of scene category co-occurrences**
  - **Consistent co-occurrences** (good): co-occurrences that are *consistent* across the images in the same category.
  - **Co-occurrence noise** (bad): *accidental* co-occurrences.
  - **Inter-features co-occurrences**: SMNs from different visual features still lie in the same common space, so we can exploit this to find consistent co-occurrences.



- **Co-occurrence modeling**
  - Idea: exploit consistent co-occurrences, while **filtering out co-occurrence noise**.
  - Limitation of previous models [1,2]: only focus on **global** co-occurrences, **single** visual feature and **supervised** modeling.
  - Contributions:
    - **Unsupervised modeling** to exploit **local** co-occurrences and remove co-occurrence noise
    - Integrate **multiple features** in the **common space**.
    - Three different **semantic representations** (*filtered SMNs*, *co-codes* and *KCNF embedding*)

Co-occurrence noise filtering (CNF)

- **Main idea**
  - **Co-occurrence dictionary** $Q = [q_1, .. q_K]$: consistent co-occurrences will be **sparse** and appear in clusters
  - **Projection-reconstruction** to recover *cleaner* SMNs $\bar{s}$ (filtered SMNs) $\bar{\alpha} = \arg\min_{\alpha} \|s - Q\alpha\| + \lambda\|\alpha\|_1$ (co-occurrence codes, or *co-codes*)
  - Reconstruction $\bar{s} = Q\bar{\alpha}$ (filtered SMNs)
  - SVMs trained with cleaner SMNs (or *co-codes*) will be more robust



- **Multiple features**

$$s_w = \prod_{n=1}^{N} \prod_{v \in V} P_{W^v|W}(w|w^v) s_{nw}^v$$

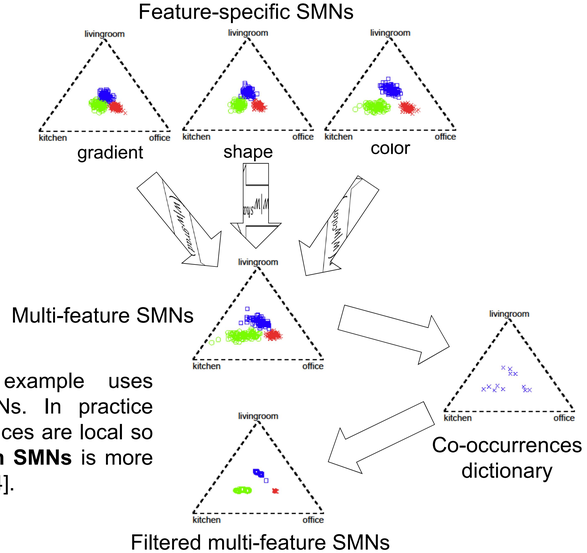
- **Kernelized co-occurrence noise filter (KCNF)**
  - SVM classification: requires suitable kernels (SMNs lie on simplices)
  - Suitable kernel[2]: Negative geodesic distance **(NGD) kernel** $k_{NGD}(s, s')$
  - Problem: computing $k_{NGD}(s, s')$ only possible in small datasets, and no mapping $\phi_{NGD}(s)$. [2] approximates $k_{NGD}(s, s') \approx \phi_{SM}(s)^T \phi_{SM}(s')$

Proposed solution: reformulating CNF in the kernel space we can by the way obtain an **exact embedding** (for $\lambda = 0$)

$$\phi_{KCNF}(I) = GK_{qs}(I) = \frac{1}{N} G \sum_{n=1}^{N} \sum_{i=1}^{K} k_{NGD}(s_n, q_i)$$

In contrast to $\phi_{SM}(s)$, the proposed embedding $\phi_{KCNF}(s)$ is exact.

Toy example



This toy example uses image SMNs. In practice co-occurrences are local so using **patch SMNs** is more effective[3,4].

Experiments and references

Method	K=2000				
	15scenes	Labelme	Sports	MIT67	SUN397
SPMSM[2]	78.9	85.9	81.3	37.6	30.0
Co-codes[3]	83.1	**89.7**	**92.3**	42.1	35.4
CNF[3]	82.9	89.6	89.4	42.6	33.2
KCNF[3]	**84.9**	89.6	87.5	**48.1**	**40.8**

[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, N. Rasiwasia, Scene recognition on the semantic manifold, *ECCV* 2012.

[3] X. Song, S. Jiang, L. Herranz, Y. Kong, K. Zheng, Category co-occurrence modeling for large scale scene recognition, *Pattern Recognition*, January 2016

[4] X. Song, S. Jiang, L. Herranz, Joint Multi-feature Spatial Context for Scene Recognition on the Semantic Manifold, *CVPR* 2015