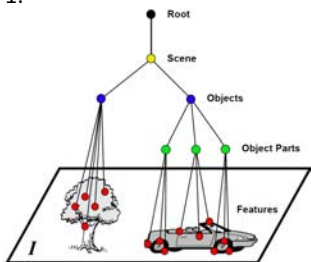# Unsupervised Learning of Hierarchical Spatial Structures in Images

Devi Parikh (Carnegie Mellon University)   Larry Zitnick (Microsoft Research)   Tsuhan Chen (Cornell University)
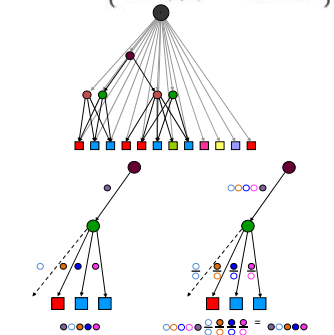
1. Overview

Learning: Unsupervised EM style
Inference: Minimum cost Steiner tree

2. Hierarchical model

○ Image part
□ Feature / Visual word

3. Inference

$$\Pr(I \mid T, R) = \prod_{v}^{V} \prod_{x}^{Ch(r_v)} \rho(x, v)$$

$$\rho(x, v) = \begin{cases} \Pr(x \mid r_v)\,\Pr(l_x \mid r_v) & x \in Ch(v) \\ 1 - \Pr(x \mid r_v) & \text{otherwise} \end{cases}$$
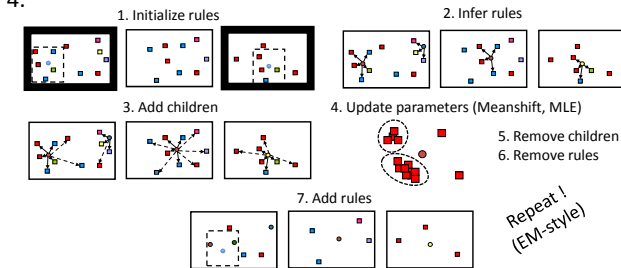
Minimum Spanning Tree   Minimum Steiner Tree

Charikar et al. 1998

Candidate locations via a voting scheme

4. Unsupervised learning
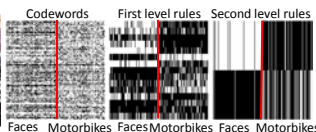
1. Initialize rules
2. Infer rules
3. Add children
4. Update parameters (Meanshift, MLE)
5. Remove children
6. Remove rules
7. Add rules

Repeat !
(EM-style)

5. Example rules learnt

Localization behavior

Category-specific and shared rules

Categorization behavior

Codewords   First level rules   Second level rules

Faces   Motorbikes   Faces   Motorbikes   Faces   Motorbikes

6. Unsupervised object categorization

Bag-of-words: 94%
Proposed: 100%
SIFT

Bag-of-words: 55%
Proposed: 82%
Edges

Bag-of-words: 87%
Proposed: 95% (Kim et al. 2008)
61 L1 (9), 12 L2 (3)
SIFT

Words
Rules   Tree
Kmeans, PLSA, SVM

7. Outdoor scenes

Image   Mean color   Codeword

Color segments

8. Street scenes

SIFT

— irrelevant
— relevant

Number of images
Number of parts