Region Classification with Markov Field Aspect Models

Jakob Verbeek & Bill Triggs

LEAR Team, INRIA Grenoble, France

Summary

• **Models to segment images into semantic classes require**
  • **local interactions to capture spatial regularity**
  • **image-wide interaction to suppress improbable classes**

• **MRF has only local interaction, PLSA only image-wide feed-back**

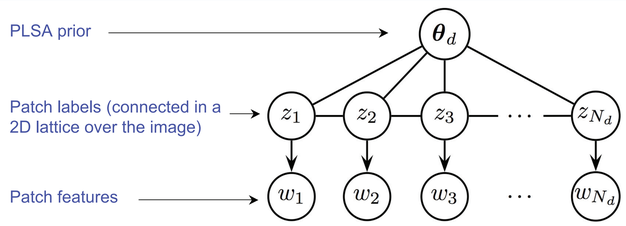
• **Our Markov Field Aspect (MFA) model incorporates both**

• **Model learned from image-wide labels, not pixel-level segmentation**

• **On the 9-class MSRC data set MFA learned from image-wide labels outperforms PLSA learned from pixel-level segmentations**

• **On the 21-class MSRC data set MFA performs comparably to TextonBoost but is about 50x faster in testing and training**

Graphical representation of the Markov Field Aspect Model



Learning a multi-modal PLSA model from image-level keywords

Image representation
• dense extraction of 20x20 pixel patches on 10x10 pixel grid
• each patch described by discretized features
  • texture: SIFT (1000 visual words, k-means)
  • color: robust HUE (100 visual words, k-means)
  • position: patch location indicated by cell in a 5x5 grid over image

Multi-modal version of PLSA
• single topic per patch, drawn from image-specific topic mixing weights
• texture, color and position independent *given the patch topic*

$$p(\{w_i\}_{i=1}^{N_d}|\theta_d) = \prod_{i=1}^{N_d} \sum_{t=1}^{T} \theta_{dt} \prod_{m=1}^{M} p(w_i^m|t).$$

patches ← → topics → modalities

Learning a PLSA model from image-level keywords
• set mixing weight to zero for classes not among the keywords
• allow any non-negative sum-to-one values for remaining mixing weights
• start EM with uniform assignments of patches to possible classes

Our experimental results show that
• even such weak supervision allows good topic models to be learned
• combining the modalities leads to significantly better recognition

Two different spatial extensions of PLSA

Forest of spanning trees

• Model only maximum spanning subtree of the 4 or 8 neighborhood graph
  • Edges have weight 1 if patches share observation, 0 otherwise

$$p(\{z_i\}_{i=1}^{N_d}|\theta_d) \propto \exp\left(\sum_i \rho[z_i = z_{\pi(i)}] + \log \theta_{dz_i}\right)$$

patches ← same as parent? → mixing weight

• To reduce dependence on arbitrary choices, we average the inference results over a forest of 10 randomly selected trees

• Using forests maintains tractable inference, while capturing some of the local dependencies between patch labels

• Many spatial neighboring patches end up being well separated in the trees, as there are only N_d-1 edges among the patches

Markov Field Aspect (MFA) model

• Markov Random Field (MRF) includes all edges between neighbors

$$p(\{z_i\}_{i=1}^{N_d}|\theta_d) \propto \exp\left(\sum_i \log \theta_{dz_i} + \sigma \sum_{i\sim j} [z_i = z_j]\right)$$

mixing weight at each patch ← → neighbors the same?

• Exact inference in MRF is intractable, we used EP approximate inference
  • term decomposition: MRF = (vertical edges) x (horizontal edges)
  • approximate each term in turn, projecting to fully factorized marginals
  • leads efficiently to Loopy Belief Propagation fixed points

• Experimentally, using PLSA-based mixing weights for the topics gives performance comparable to weights estimated using the full model

• Using cross-validation the coupling parameter was set to s=2

Comparison between PLSA and its spatial extensions

• MSRC data set, 240 images 9 semantic classes, ~30% unlabelled pixels

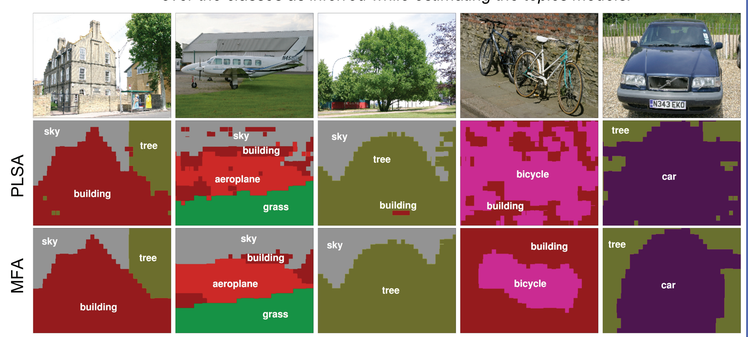
• Pixels in test images were classified using PLSA, MFA, and the tree model
• Class models p(w|t) were learned in three ways
  • using the pixel-level segmentation
  • using image-level keywords + PLSA
  • using image-level keywords + MFA

• MFA consistently outperforms PLSA by about 4%
• Spanning tree models are only slightly worse than MFA
• For different classification methods, using the MFA learned class models leads to ~2% better results than using class models learned using PLSA
• MFA learned from image-level keywords yields 80.2% accuracy, while PLSA model learned detailed pixel-level labels achieves only 78.5%

Topics \ Classification	PLSA	MFA	TREE
Pixel level labels	78.5	82.3	81.7
Keywords + MFA	76.6	80.2	79.5
Keywords + PLSA	74.0	78.1	77.5

Classification accuracies averaged over the 9 classes and random divisions of the images over train (90%) and test (10%) sets. Per class results can be found in the paper.

A training image labeled {Building, Grass, Sky, Tree}, and its decomposition over the classes as inferred while estimating the topics models.



Example images (top) segmented using PLSA (middle) and Markov field Aspect model (bottom)

Comparison between Markov Field Aspect Models and TextonBoost

• Extended MSRC data set: 591 images, 21 semantic classes
• Performance of the methods is comparable
• Markov Field Aspect model is much faster (x50) in training and testing
  • training: uses standard quantized SIFT and HUE features
  • testing: it operates on patch level

	Average per class	Average per pixel
TextonBoost from segmentation	58	72
MFA from segmentation	64	74
MFA from keywords	50	61

Classification accuracies averaged over classes and pixels, 55% train and 45% test data.

Example images (top) segmented using Markov field Aspect model (bottom)