Scene Segmentation with Conditional Random Fields Learned from Partially Labeled Images

Jakob Verbeek & **Bill Triggs**

LEAR Team, INRIA Rhône-Alpes, Grenoble, France

Oral presentation at NIPS 2007, Vancouver, Canada

CENTRE NATIONAL
DE LA RECHERCHE
SCIENTIFIQUE

INRIA
RHÔNE-ALPES

Overview

- Introduction
- Image representation & features
- Segmentation model & learning
- Experimental results

Visual Recognition

- Recognition of visual categories is performed at different levels of detail
  - categorization: presence/absence of category in image
  - localization: mark category instances with enclosing bounding-box
  - segmentation: give flexible outline of (instances of) category in image

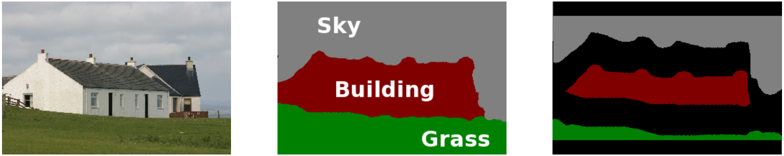
- Training data also comes in these different forms
  - in general pairs $\{\text{image}_n, \text{annotation}_n\}_{n=1}^{N}$

- Training data and recognition task may use different levels of detail
  - e.g. classification annotation to learn segmentation model [Verbeek & Triggs 2007]

Some images and annotations from the PASCAL Visual Object Classes Challenge 2008

Learning to Segment from Partially Labeled Images

- Goal: joint recognition and segmentation
- Training data: images with semantic segmentation
- Question: how (good) can we do using partially labeled images?
  - full manual labeling is tedious to produce
  - labeling near category borders error prone
  - full segmentation not critical for learning?



An example image, its full labeling, and partial labeling: black pixels remain unlabeled.

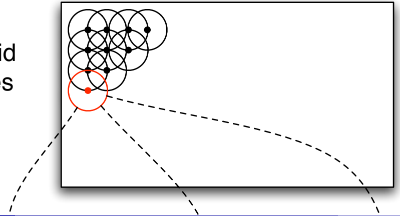
Overview

- Introduction
- Image representation & features
- Segmentation model & learning
- Experimental results

Modeling Images as Collections of Local Patches

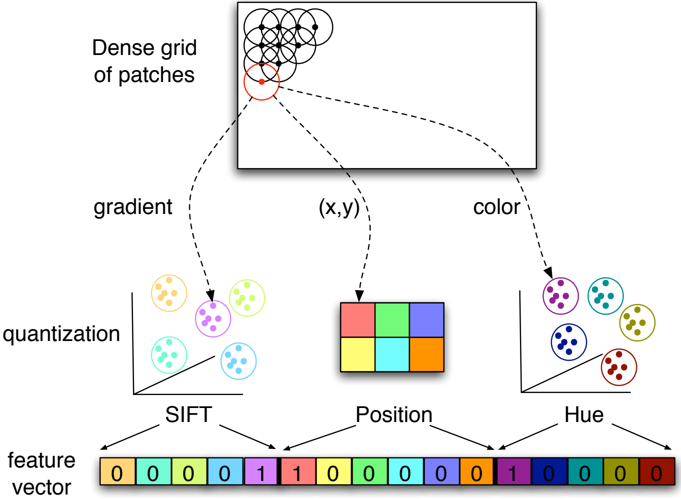
- Dense sampling of image patches on regular grid
- Feature vector associated with each patch
- Class label associated with each patch
  - e.g. *grass*, *building*, *sky*, ...

Dense grid of patches

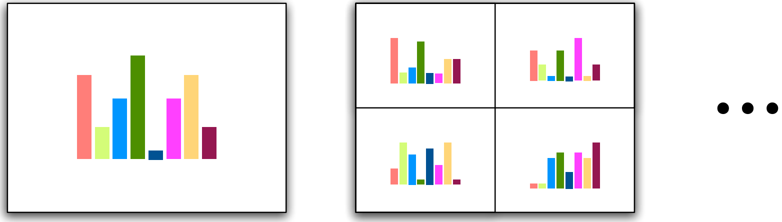


Local Image Descriptors

- Quantization of feature space (regular grid, or k-means)
- Each patch represented by corresponding "visual words"
- Patch described with bit-vector using concatenated one-of-k coding



Region Level Context Using Aggregate Features



- **Accumulate a local feature histogram** ("bag of visual words") in each cell of a coarse grid covering the image (1×1, 2×2, ...)
- **Histogram used as feature by every patch in the cell**

Overview

- Introduction
- Image representation & features
- Segmentation model & learning
- Experimental results

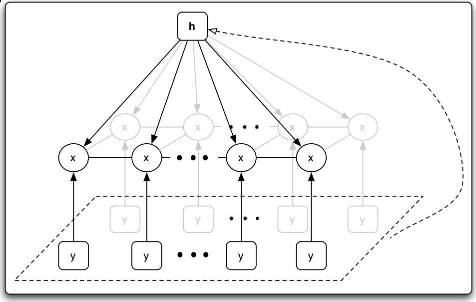
Conditional Random Field Model

- **Random field models spatial contiguity of labeling X**

$$
\begin{aligned}
p(X|Y) &= \frac{1}{Z} \exp -E(X|Y) \\
Z &= \sum_{X} \exp -E(X|Y)
\end{aligned}
$$

- **Partition function Z generally intractable to compute**

- **CRF energy function combines**
  - local image features
  - aggregate features
  - neighboring labels



Energy Function using Single Aggregate Feature

- Let n index the N image patches, $X = \{\mathbf{x}_n\}$ and $Y = \{\mathbf{y}_n\}$
  - $\mathbf{x}_n \in \{0,1\}^C$ is a one-of-C coding for the C class labels

- Let $\mathbf{h}$ denote the average of the feature vectors $\mathbf{h} = \frac{1}{N}\sum_n \mathbf{y}_n$

$$E(X|Y) = \sum_n \mathbf{x}_n^\top A \mathbf{y}_n + \sum_n \mathbf{x}_n^\top B \mathbf{h} + \sum_{n \sim m} \phi_{nm}(\mathbf{x}_n, \mathbf{x}_m)$$

- Matrices A and B are $C \times D$ (with D dimension of feature vector)

- Pairwise potential:
  - Potts-model (with contrast term): $\phi_{nm}(\mathbf{x}_n, \mathbf{x}_m) = (\sigma + \tau d_{nm}) \cdot \mathbf{x}_n^\top \mathbf{x}_m$
  - Class dependent potential: $\phi_{nm}(\mathbf{x}_n, \mathbf{x}_m) = \mathbf{x}_n^\top C \mathbf{x}_m$

- Trivial to obtain derivative of $\partial E(X|Y)/\partial\theta$ for an image Y and a labeling X.

Learning from Partially Labelled Images

- **Usual likelihood maximization of complete label field not possible**
  - Deleting unlabeled patches from model could remove all label transitions
- **Partial labeling defines a set of compatible complete labelings S**
  - unlabeled sites that can have any label, e.g. near object boundaries
  - allows more general constraints: e.g. force some sites to have the same label
- **Maximize the probability to get a labeling in S**

$$L = \log p(X \in S|Y) = \log \sum_{X \in S} p(X|Y)$$

- **Intractable sum over exponential nr. of label completions $X \in S$**

Learning from Partially Labelled Images

- **Recall the partition function:**

$$Z = \sum_X \exp -E(X|Y)$$

- **Situation is not much worse than the complete labeling case**

$$\begin{aligned} L &= \log \sum_{X \in S} p(X|Y) = \log \sum_{X \in S} \frac{1}{Z} exp - E(X|Y) \\ &= -\log \left(\sum_X exp - E(X|Y) \right) + \log \left(\sum_{X \in S} exp - E(X|Y) \right) \end{aligned}$$

- **Gradient of log-likelihood for a parameter θ**

$$\frac{\partial L}{\partial \theta} = \left\langle \frac{\partial E}{\partial \theta} \right\rangle_{p(X|Y)} - \left\langle \frac{\partial E}{\partial \theta} \right\rangle_{p(X|Y, X \in S)}$$

Learning from Partially Labelled Images

- **Gradient of log-likelihood for a parameter θ**

$$\frac{\partial L}{\partial \theta} = \left\langle \frac{\partial E}{\partial \theta} \right\rangle_{p(X|Y)} - \left\langle \frac{\partial E}{\partial \theta} \right\rangle_{p(X|Y, X \in S)}$$

- **To compute expectations of gradient of energy we need**
  - unary terms: marginal label distribution for single sites
  - pairwise potential: marginal label distribution for neighboring sites

- **We run Loopy Belief Propagation twice**
  - for prediction $p(X|Y)$ & for label completion $p(X|Y, X \in S)$

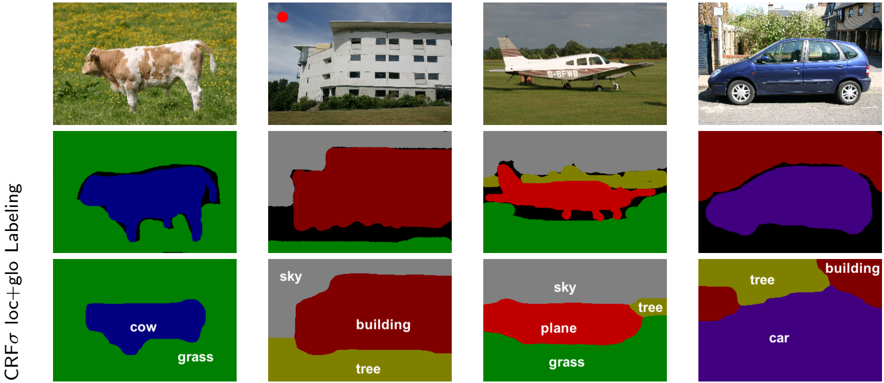
- **Log-likelihood given by difference of log-partition functions**
  - Use LBP marginals to compute the Bethe free-energy approximations

$$L = \log \sum_{X \in S} p(X|Y) = -\log Z_{p(X|Y)} + \log Z_{p(X|Y, X \in S)}$$

Overview

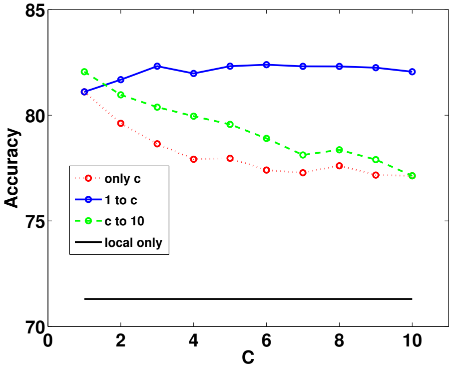
- Introduction
- Image representation & features
- Segmentation model & learning
- Experimental results

Data Set and Experimental Setup



- **MSRC data set:** 240 images of 320×213 pixels, 70% of pixels labeled
- **9 classes:** *building, grass, tree, cow, sky, plane, face, car, bike.*
- **120 images to train**, 120 to evaluate, average over 20 trials

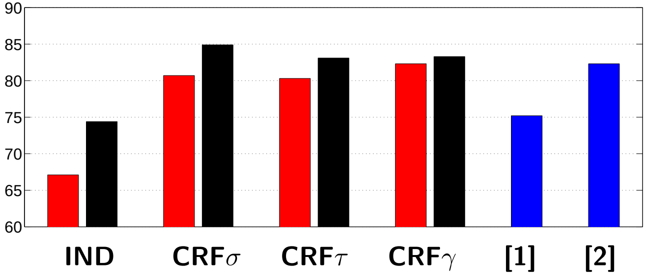
Performance of Local & Aggregate Features



- **Performance without CRF neighbor coupling**
  - no aggregate features, at single scale, or at multiple scales
- **Result: Large-scale aggregates are most informative**
  - including additional aggregate scales improves results slightly

The Pairwise Potential of the CRF

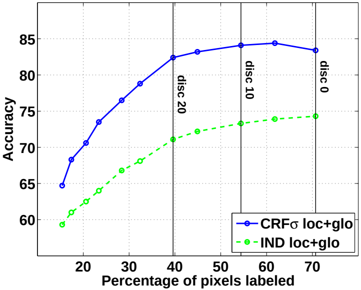
- **Both random field spatial coupling and image-wide context are useful**
- **Exact choice of pairwise potential is less important**



- **IND:** no coupling, **CRF**σ**:** Potts, **CRF**τ**:** contrast Potts, **CRF**γ**:** class based
- local features only (**red**); including global aggregate (**black**)
- **[1]** Schroff et al. ICVGIP'06: optimized aggregation window, no coupling
- **[2]** our PLSA-MRF model CVPR'07: generative, cross-validation for σ

Recognition as a function of the amount of labeling

- Decimate training labels using morphological erosion filters of increasing size



- **Good performance with CRF when only** 40–70% **of labels available**
- **Applying small erosion improves the model – due to label errors**

Summary

- **Good CRFs can be learned from partially labelled training images**
  - marginalize over all possible label completions
  - works if label transitions are completely unobserved
- **Including aggregate features significantly improves performance**
  - image-wide aggregates are the most informative
- **Pairwise potential is crucial for good segmentations**
  - but different forms yield comparable performance