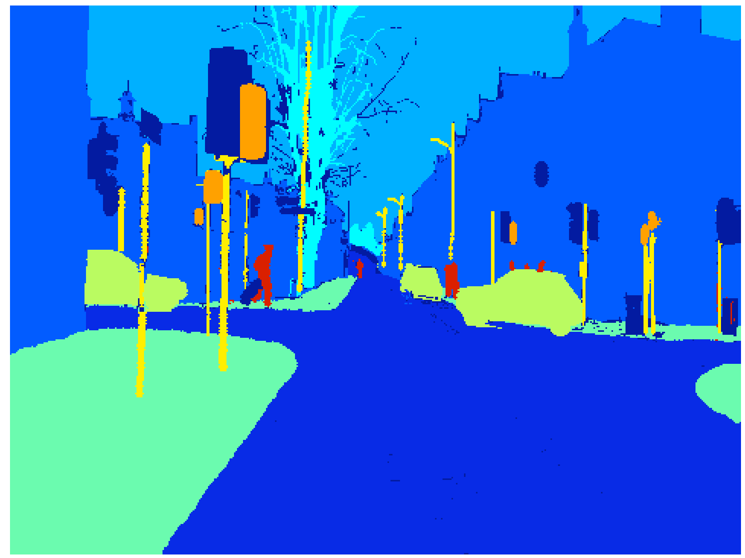
Segmentation

Bottom up Segmentation

Semantic Segmentation

Semantic Labeling of Street Scenes

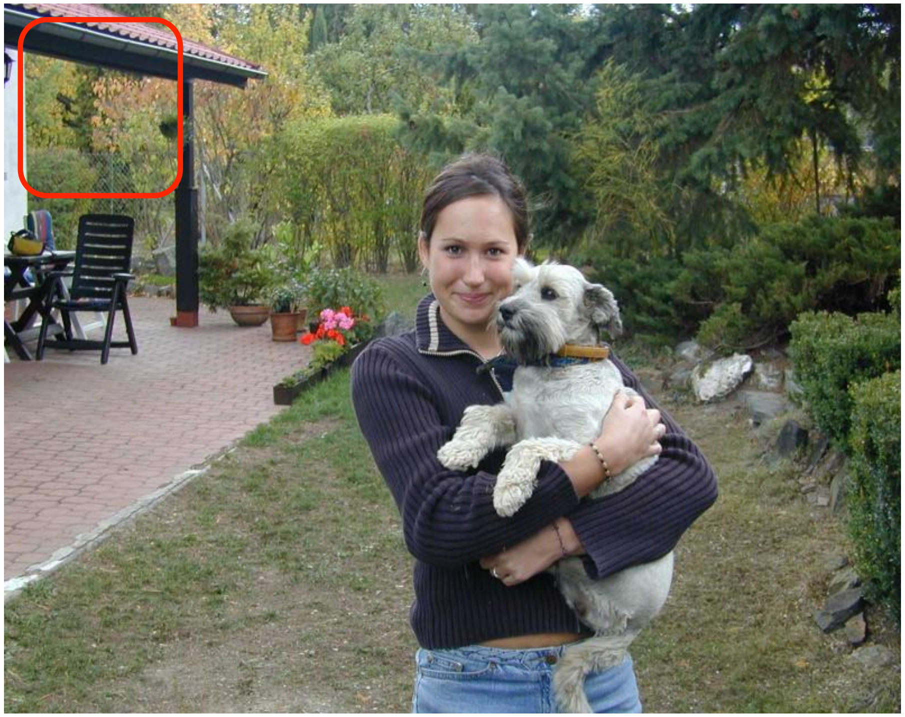
Ground Truth Labels



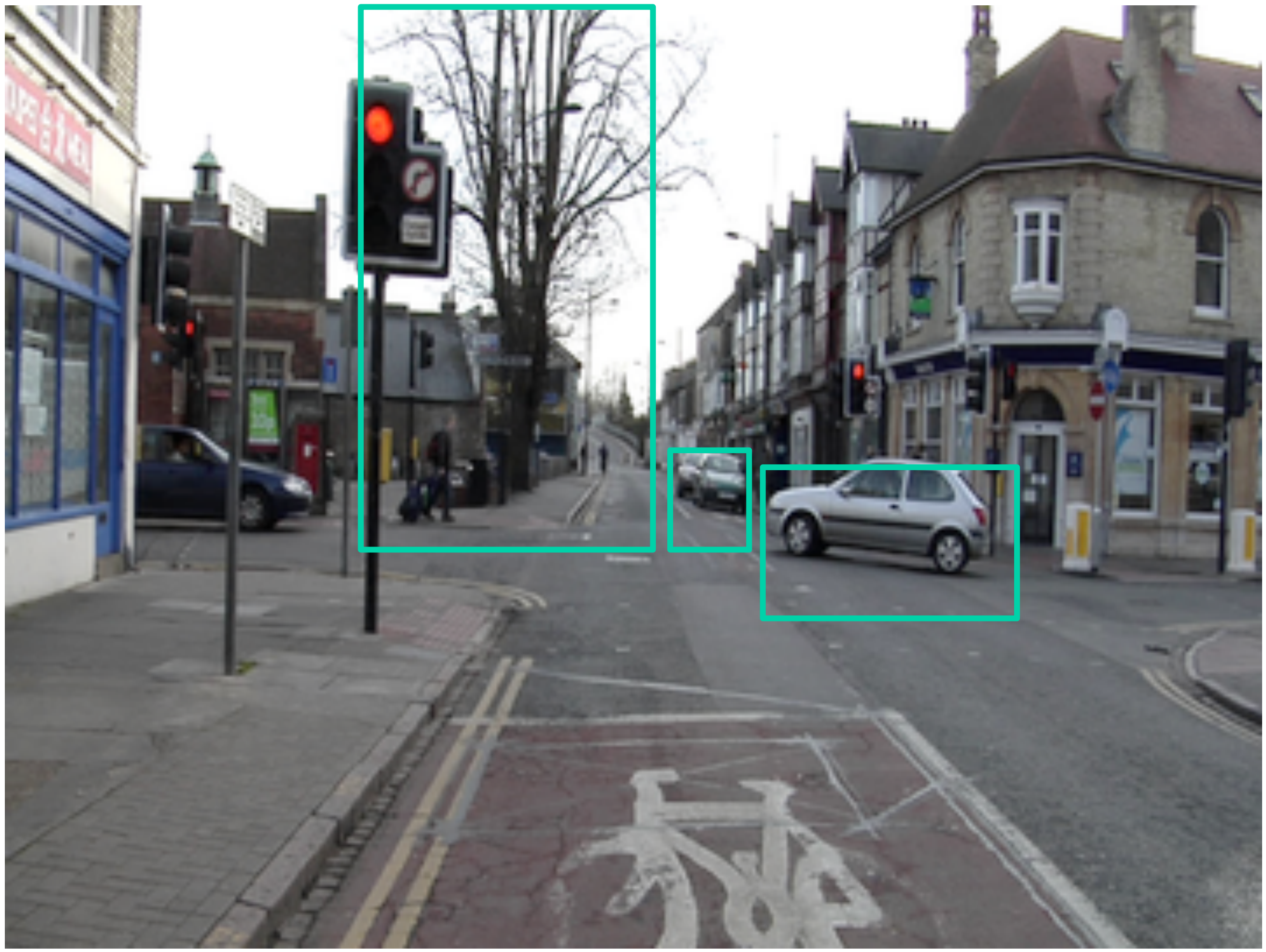
- 11 classes, almost all occur simultaneously, large changes in viewpoint, scale
- sky, road, sidewalk,tree,fence,pole,traffic sign,car,pedestrian,
- bicyclist

Object detection

- Basic idea: slide a window across image and evaluate a (object) face model at every location



Object detection



- Sliding window approach does not scale well
- Top pedestrian [HOG] detector only 15.3% with 85% FP

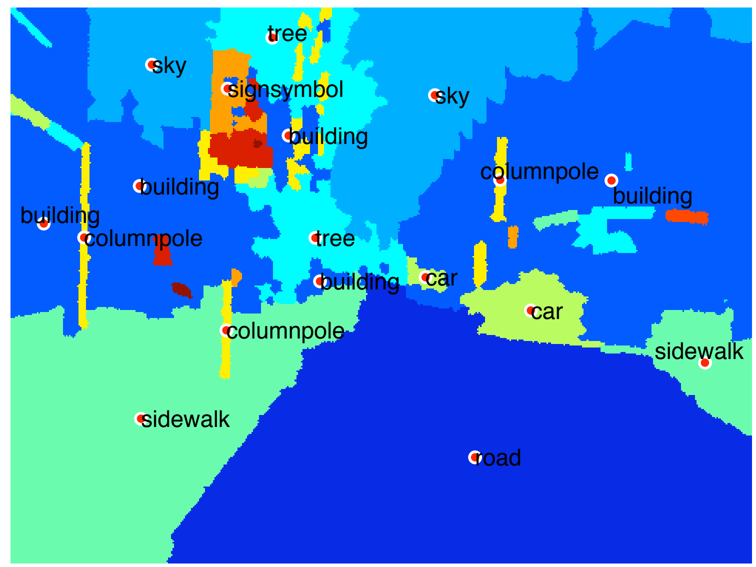
Ingredients

- What are the elementary regions for classification ?
- How to spatially integrate the evidence from neighbouring regions ?
- What is the role of geometric features for classification ?
- How to exploit street scene context ?
- Video database – 3 daytime one at dusk sequence

Semantic Labeling

- Problem formulation, compute optimal assignment of labels to regions given appearance and geometry features

$$P(\mathbf{L}|\mathtt{A},\mathtt{G}) = \frac{P(\mathtt{A},\mathtt{G}|\mathbf{L})\, P(\mathbf{L})}{P(\mathtt{A},\mathtt{G})}$$



$$\mathbf{L} = (l_1, l_2, \ldots l_S)^\top$$

Semantic Segmentation

Ingredients

- simultaneous segmentation and recognition
- choice of elementary regions
  (pixels, super-pixels, rectangular regions)
- choice of features to describe the regions
  (color edge statistics, area/shape/moments …)
- computation of the likelihoods of features given labels
- context modeling
  - spatial co-occurrence between class labels
  - modeling relative location of object classes
  - large support regions for training category classifier
- semantic labeling formulated in MRF/CRF framework

Image Labelling Problems

In general - Assign a label to each image pixel

Geometry Estimation | **Image Denoising** | **Object Segmentation** | **Depth Estimation**

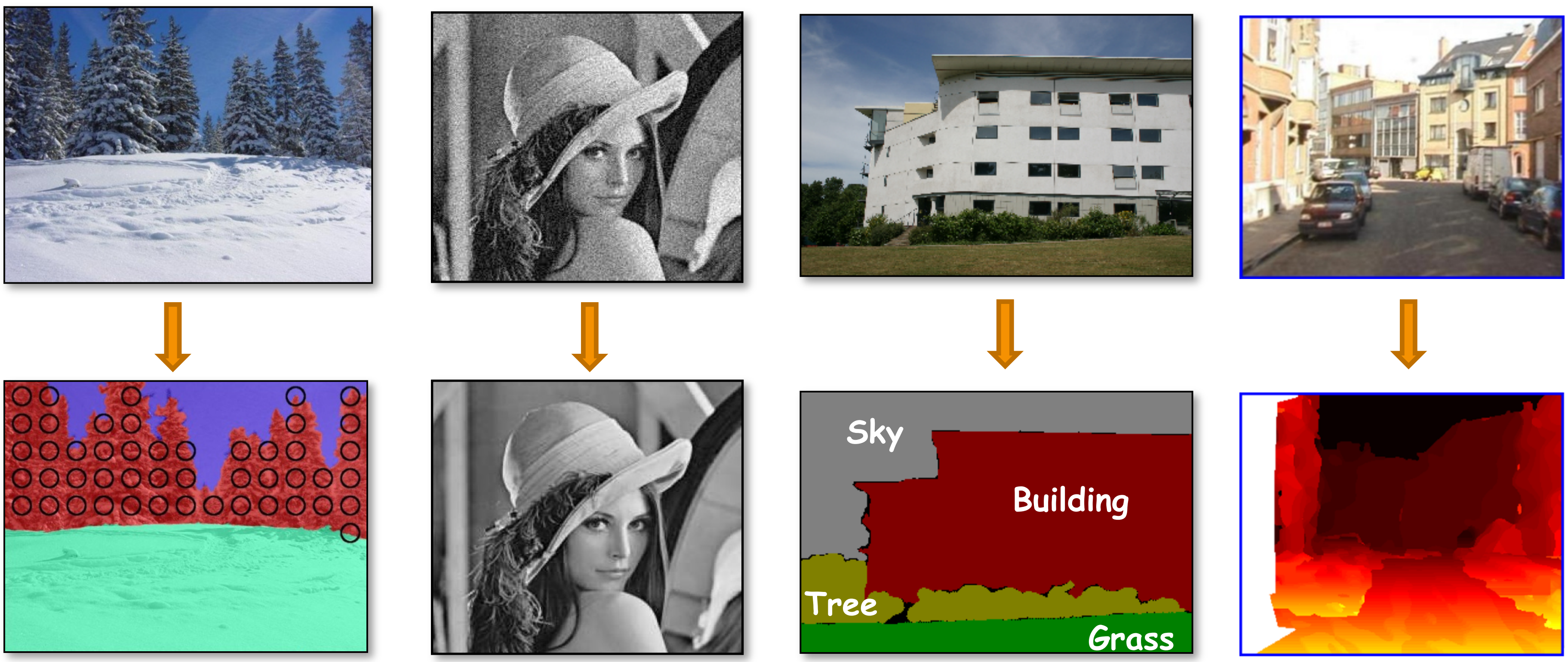


Image Labelling

- MRF/CRF framework
- Typical pair-wise MRF – probabilistic semantics

$$P(\mathbf{x}|\mathbf{y}) = P(\mathbf{y}|\mathbf{x})P(\mathbf{x}) \qquad P(\mathbf{y}|\mathbf{x}) = \prod_{i \in S} P(\mathbf{y}_i|x_i)$$

$P(\mathbf{x})$ Prior distribution over labels

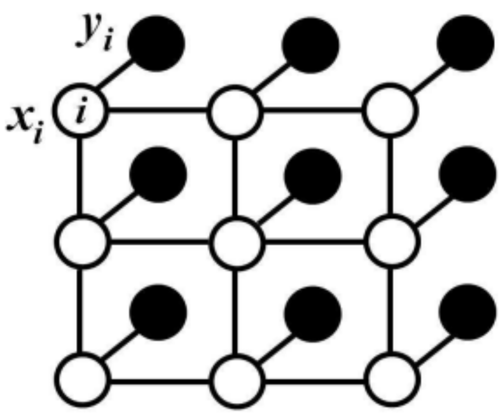
- set of labels $\mathbf{x} = \{x_1, \ldots, x_n\}$

- Commonly used priors for semantic labeling
  - spatial co-occurrence between class labels
  - modeling relative location of object classes
  - large support regions for training category classifier

Markov Random Field Framework

- Typical pair-wise MRF in vision

$$P(\mathbf{x}|\mathbf{y}) = P(\mathbf{y}|\mathbf{x})P(\mathbf{x})$$



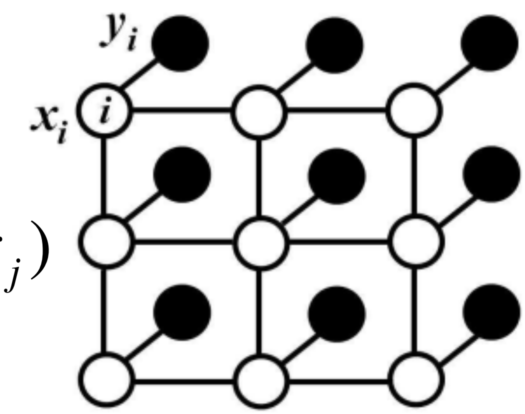
$$P(\mathbf{y}|\mathbf{x}) = \prod_{i \in S} P(\mathbf{y}_i|x_i)$$

- If data likelihood is independent given the labels and prior is homegenous Ising prior with pairwise non zero potentials, we can write posterior as

$$P(x \mid y) = \frac{1}{Z_m} \exp\Big(\sum_{i \in S} \log p(y_i \mid x_i) + \sum_{i \in S} \sum_{j \in N_i} \beta_m x_i x_j$$

Markov Random Field Framework

- Typical pair-wise MRF in vision



$$P(x \mid y) = \frac{1}{Z_m} \exp\Big(\sum_{i \in S} \log p(y_i \mid x_i) + \sum_{i \in S} \sum_{j \in N_i} \beta_m x_i x_j\Big)$$

- Can be expressed in terms of an energy function

$$P(x) = \frac{1}{Z_m} \exp(-E(x, \theta))$$

- Z_m is a partition function – important for learning Parameters, not critical for inference

- Typical energy function $E(x) = \sum_{i \in S} D(x_i) + \sum_{(i,j) \in E} V(x_i, x_j)$

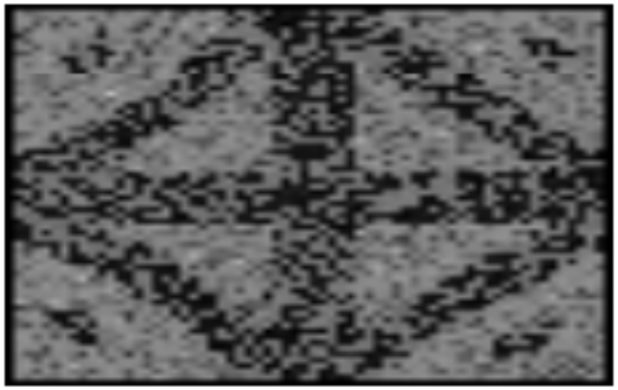
Toy Problem

- Binary image denoising
- Given a noisy image get most likely binary image
- Set of labels {white, black}
- Commonly used prior: Isotropic Ising Prior
- Goal find image *x,* such that *p(x|y)* is maximized
- i.e. find image that minimizes energy *E(x)*

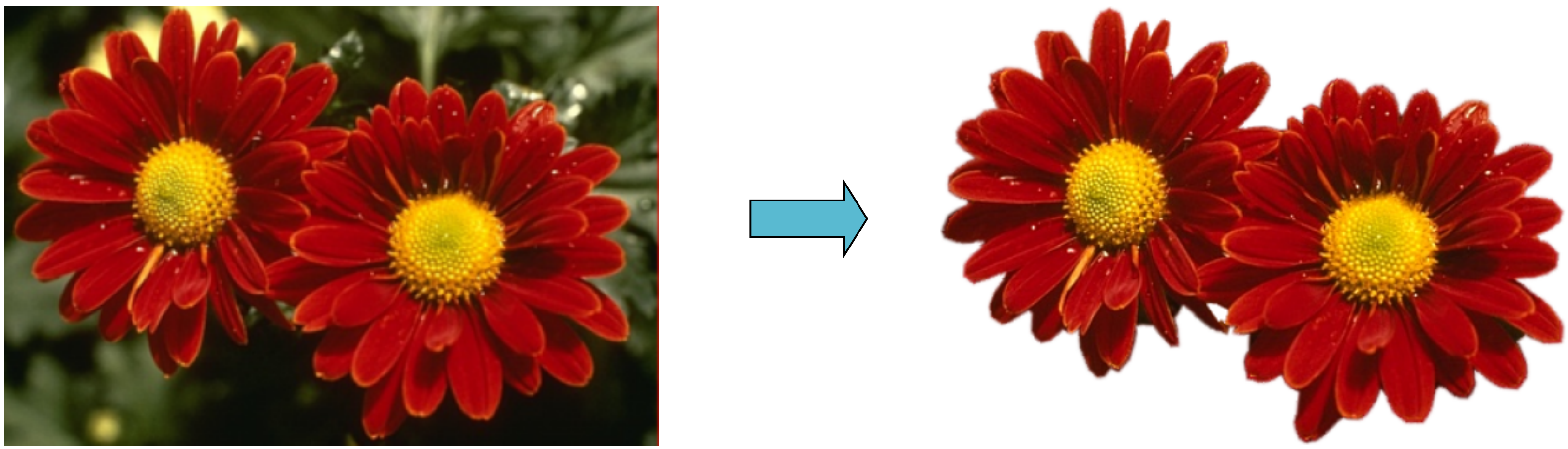
$$E(x) = \sum_{i \in S} D(x_i) + \sum_{(i,j) \in E} V(x_i, x_j) \quad V(x_i, x_j) = \beta \,|\, x_i - x_j \,|, \beta > 0$$

$$D(x_i) = -\log(1-\theta) \; for \; x_i = y_i$$

$$D(x_i) = -\log(\theta) \; for \; x_i \neq y_i$$



Foreground/Background Estimation



Rother et al. SIGGRAPH04

Foreground/Background Estimation

$$E(\mathbf{x}) = \boxed{\sum_{i \in \mathcal{V}} \psi_i(x_i)} + \boxed{\sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi_{ij}(x_i, x_j)}$$

Data term **Smoothness term**

Data term

$$\psi_i(x_i = 0) = -\log(p(x_i \notin FG))$$
$$\psi_i(x_i = 1) = -\log(p(x_i \in FG))$$

Estimated using FG / BG colour models

Smoothness term

$$\psi_{ij}(x_i, x_j) = K_{ij}\delta(x_i \neq x_j)$$

where $K_{ij} = \lambda_1 + \lambda_2 \exp(-\beta(I_i - I_j)^2)$

Intensity dependent smoothness

Foreground/Background Estimation

$$E(\mathbf{x}) = \boxed{\sum_{i \in \mathcal{V}} \psi_i(x_i)} + \boxed{\sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi_{ij}(x_i, x_j)}$$

Data term **Smoothness term**

$$\mathbf{x}^* = \arg\min_{\mathbf{x} \in \mathbf{L}} E(\mathbf{x})$$

Foreground/Background Estimation

$$E(\mathbf{x}) = \underbrace{\boxed{\sum_{i \in \mathcal{V}} \psi_i(x_i)}}_{\text{Data term}} + \underbrace{\boxed{\sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi_{ij}(x_i, x_j)}}_{\text{Smoothness term}}$$

$$\mathbf{x}^* = \arg\min_{\mathbf{x} \in \mathbf{L}} E(\mathbf{x})$$

How to solve this optimisation problem?

Foreground/Background Estimation

$$E(\mathbf{x}) = \underbrace{\boxed{\sum_{i \in \mathcal{V}} \psi_i(x_i)}}_{\text{Data term}} + \underbrace{\boxed{\sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi_{ij}(x_i, x_j)}}_{\text{Smoothness term}}$$

$$\mathbf{x}^* = \arg\min_{\mathbf{x} \in \mathbf{L}} E(\mathbf{x})$$

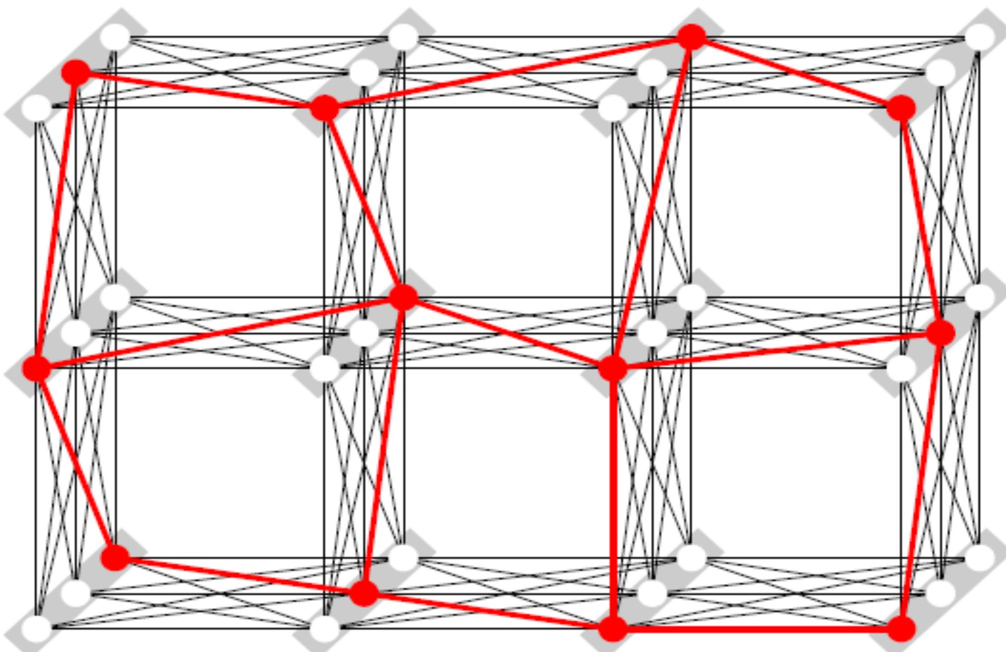
How to solve this optimisation problem?

- Transform into min-cut / max-flow problem
- Solve it using min-cut / max-flow algorithm
- Loopy belief propagation, tree-weigted message passing etc
- Gibbs Sampling, ICM, Simulated Annealing

More general multi-label problems

- MAP of MRF - solve optimal labelling problem
- Labels semantic categories
- Nodes: pixels, superpixels, patches etc

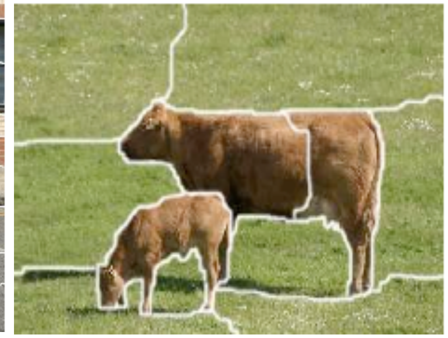
Toy example:



12 superpixels, 3 labels each

✓ available solvers: Kolmogorov PAMI'06, Werner PAMI'07

MRSC 21 class dataset



Class co-occurrence

Typically 2-5 classes per image, small scale and viewpoint variati

Context modeling

- spatial co-occurrence between class labels
- modeling relative location of object classes
- Best performing methods are discriminative (CRF framework)

Semantic Labeling

$$P(\mathbf{L}|\mathtt{A},\mathtt{G}) = \frac{P(\mathtt{A},\mathtt{G}|\mathbf{L})\,P(\mathbf{L})}{P(\mathtt{A},\mathtt{G})}$$

- Appearance is captured using scale invariant features organized in visual vocabulary trees, discrete set of visual words $$\mathbf{V} = (v_1, v_2, \ldots, v_S)^\top$$
- Maximize likelihood of the labels using appearance and geometry cues

appearance likelihood geometry likelihood

$$\operatorname*{argmax}_{\mathbf{L}} P(\mathbf{L}|\mathbf{V},\mathtt{G}) = \operatorname*{argmax}_{\mathbf{L}} P(\mathbf{V}|\mathbf{L})\,P(\mathtt{G}|\mathbf{L})\,P(\mathbf{L})$$

$$\operatorname*{argmin}_{\mathbf{L}} \left(\sum_{i=1}^{S} E_{app} + \lambda_g \sum_{i=1}^{S} E_{geom} + \lambda_s \sum_{(i,j)\in\mathcal{E}} E_{smooth} \right)$$

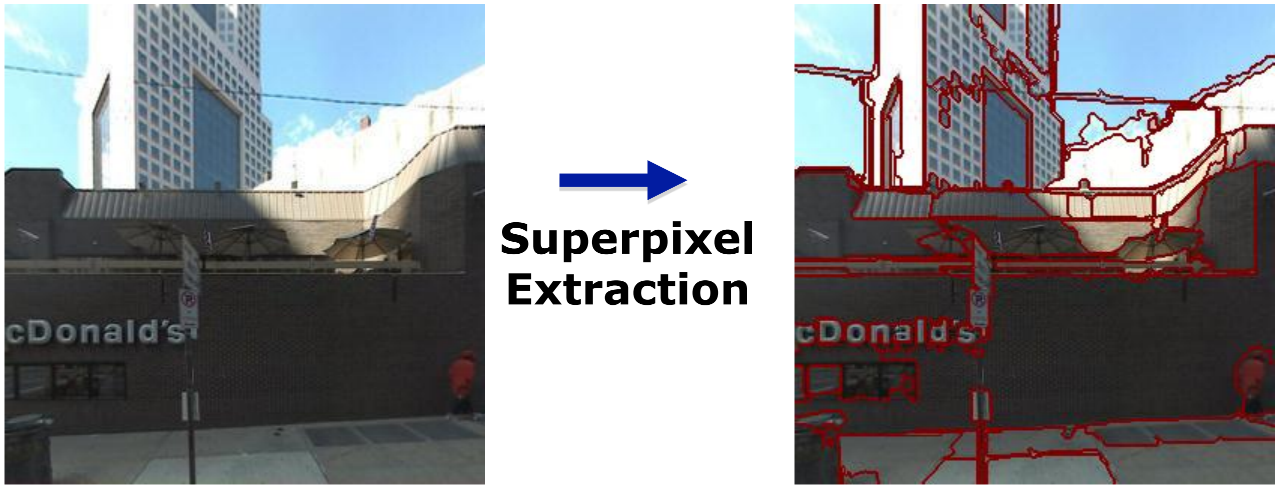
Semantic Segmentation

Image

Ground Truth Annotation

- Semantic categories considered- buildings, roads, sky, cars and trees
- Fully annotated datasets
  - 320 side views dataset
  - 90 frontal views dataset

Semantic Segmentation



- Images sites for semantic labeling - superpixels
  - superpixels obtained from graph based segmentation method of [Felzenszwalb 2004]
- Features – color, texture, location, perspective cues
- 194 dimensional feature computed for each superpixel of image

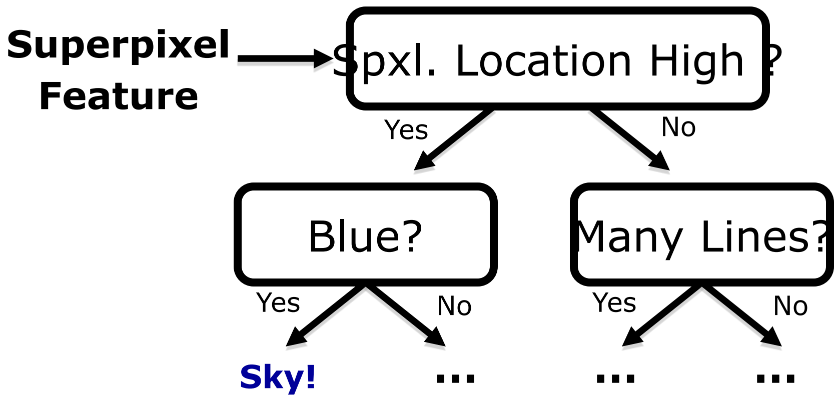
P.F. Felzenszwalb and D.P. Huttenlocher. *Efficient graph-based image segmentation*, IJCV 2004

D. Hoiem, A.A. Efros and M. Hebert. *Recovering surface layout from an image*, IJCV 2007

Semantic Segmentation

- Observation likelihood - $P(a_i | I_i)$ – computed using boosting classifier

- Boosting classifiers
  - Learn ensemble of weak learners
  - Decision Trees as weak learner

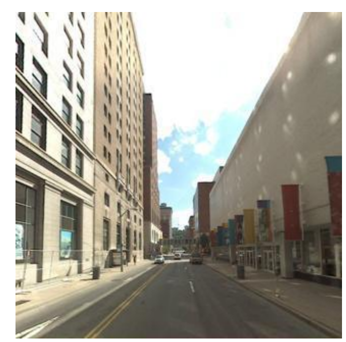


Sample decision tree for sky class

  - Learn one vs. all boosting classifiers
  - Superpixel semantic label determined by classifier with maximum score

- Two separate models trained
  - 320 side views dataset
  - 90 frontal views dataset

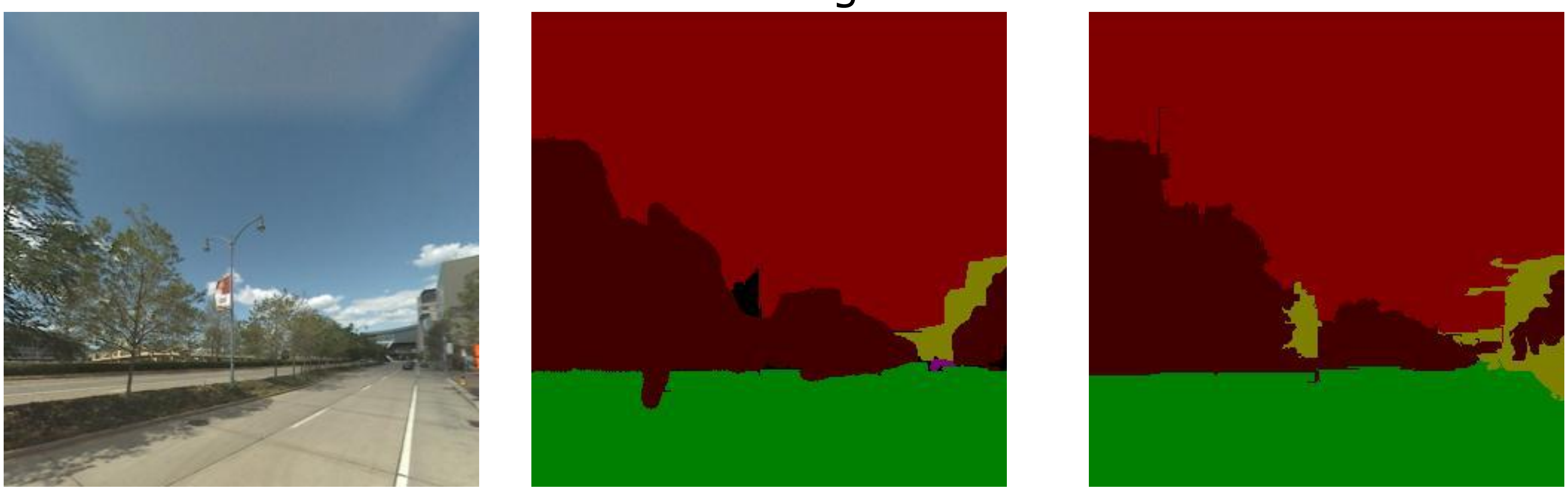
Side View



Frontal View

Example Results

Image | Ground Truth Labeling | Predicted Labeling



Nonparametric Methods

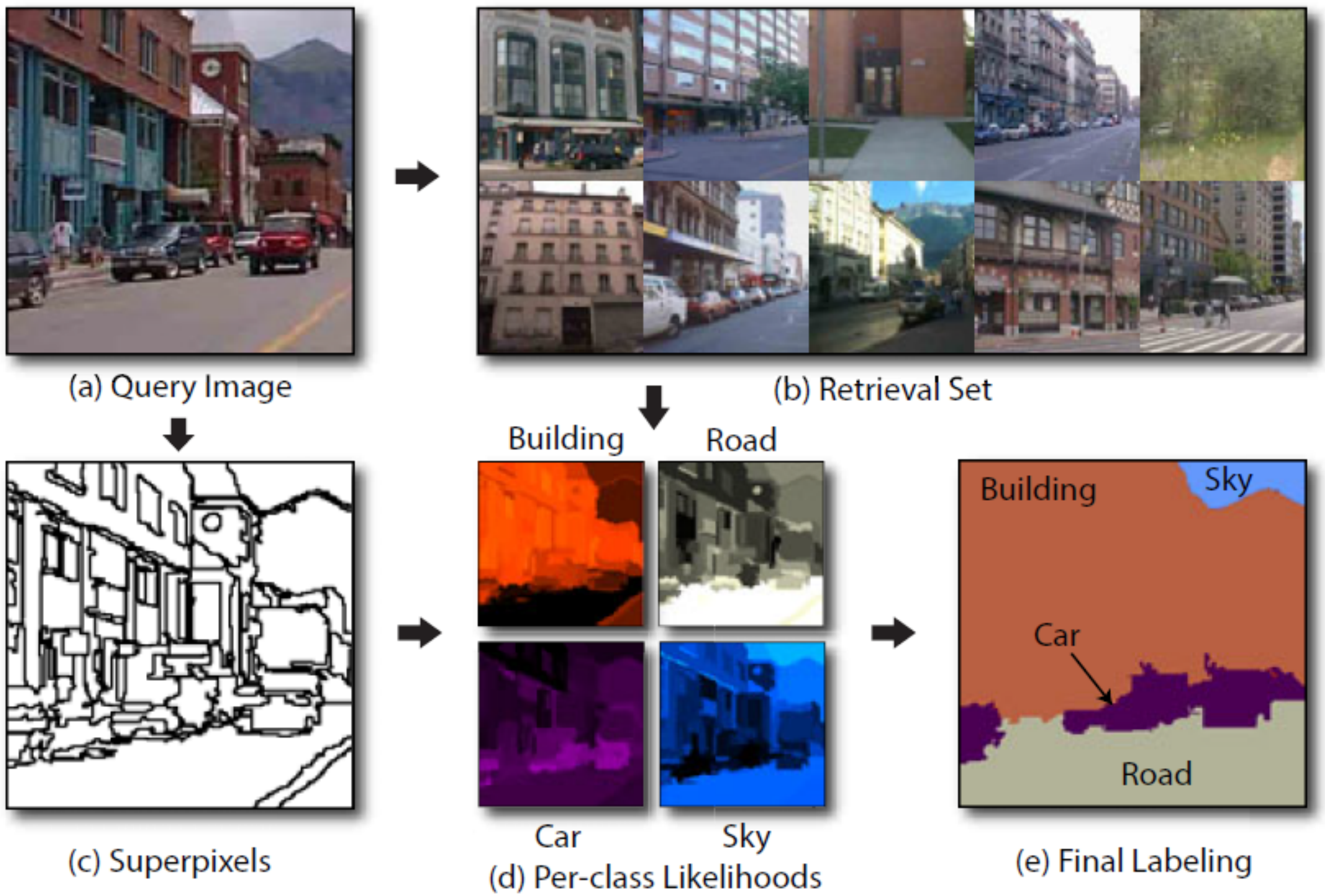


Figure credit: J. Tighe and S. Lazebnik. Superparsing: *Scalable nonparametric parising with superpixels*, ECCV 2010

References

- J. Tighe and S. Lazebnik. *Superparsing: Scalable nonparametric image parsing with superpixels*. ECCV 2010
- D. Eigen and R. Fergus. *Nonparametric image parsing using adaptive neighbor sets*, CVPR 2012
- P. Sturgess, L. Ladicky, N. Crook, and P. Torr. *Scalable cascade inference for semantic image segmentation*. BMVC 2012
- G. Singh and J.Kosecka: Nonparametric Scene Parsing with Adaptive Feature Relevance and Scene Context, CVPr 2013

Conditional Random Fields

- MAP of MRF - solve optimal labelling problem
- Posterior = product of likelihood and prior
- Sometimes difficult to obtain – generative model

$$P(x \mid y) = \frac{1}{Z_m} \exp\left(\sum_{i \in S} \log p(y_i \mid x_i) + \sum_{i \in S} \sum_{j \in N_i} \beta_m x_i x_j\right)$$

- Conditional Random Fields
- Model the posterior directly

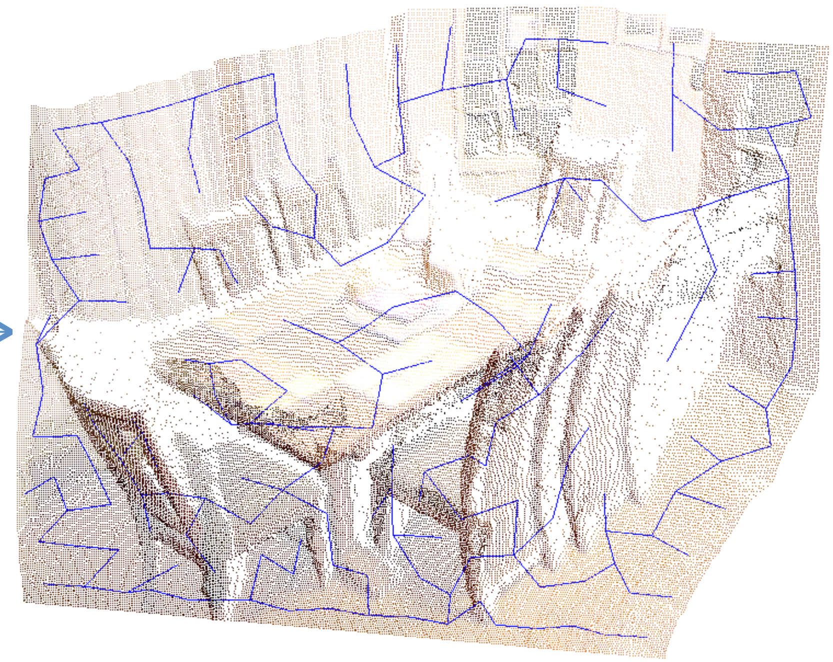
$$P(x \mid y) = \frac{1}{Z_m} \exp\left(\sum_{i \in S} f(x_i, y) + \sum_{i \in S} \sum_{j \in N_i} g(x_{,i} x_j, y)\right)$$

Conditional Random Fields

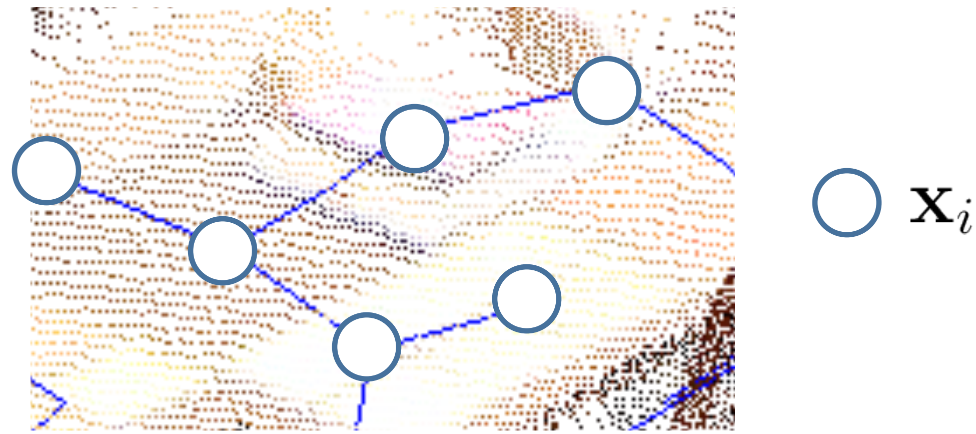
- Semantic Parsing of Indoors scenes
- Tree Graph Structure
- Informative Features using RGB-D

Graph Structure: Our choice

Minimum Spanning Tree
Over 3D

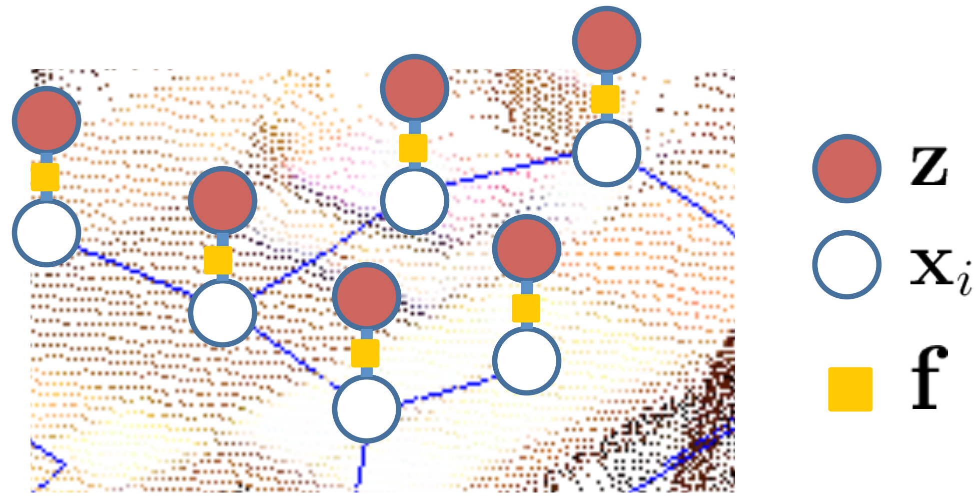


Potentials: Pairwise CRFs



$$p(\mathbf{x}|\mathbf{z}) = \frac{1}{Z(\mathbf{z})} \exp\left(\mathbf{w}_1 \sum_{i \in \mathcal{N}} \mathbf{f}(\mathbf{x}_i, \mathbf{z}) + \mathbf{w}_2 \sum_{i,j \in \mathcal{E}} \mathbf{g}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{z}) \right)$$

Potentials: Pairwise CRFs



$$p(\mathbf{x}|\mathbf{z}) = \frac{1}{Z(\mathbf{z})} \exp\left(\mathbf{w}_1 \sum_{i \in \mathcal{N}} \mathbf{f}(\mathbf{x}_i, \mathbf{z}) + \mathbf{w}_2 \sum_{i,j \in \mathcal{E}} \mathbf{g}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{z}) \right)$$