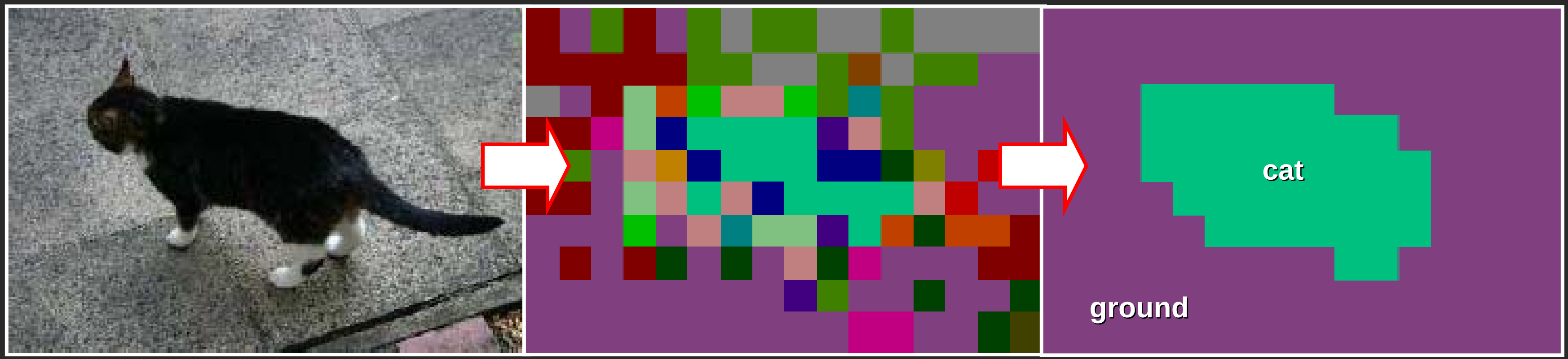


An Empirical Bayes Approach to Contextual Region Classification

Svetlana Lazebnik
University of North Carolina at Chapel Hill

Maxim Raginsky
Duke University

The problem

Image

Ground truth

Local model

Contextual model

- Our goal: Improving a purely local model without prior learning of contextual interactions

A minimalistic approach to context

- Key question: Can we get useful contextual information about the class labels from the unlabeled test data – with minimal prior assumptions?

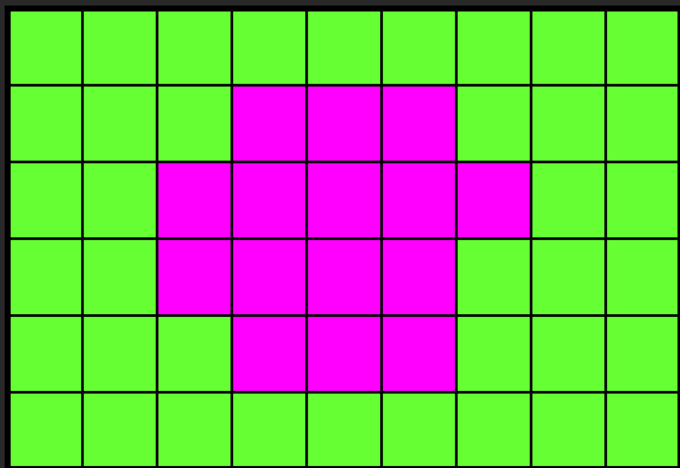
- Key insight: The structure of the unknown label sequence is indirectly revealed through the statistical redundancy of the observation sequence
  – A contextual model of the observations can be turned into a contextual model of the class labels

Methodology

- Empirical Bayes methods (Robbins 1956): obtain a prior directly from the data instead of committing to it in advance
  - No parametric contextual model
  - No need to learn context from training data
- Compound decision theory (Robbins 1951): solve a series of decision problems that share a common statistical structure
- Universal denoising (Weissman et al. 2005)

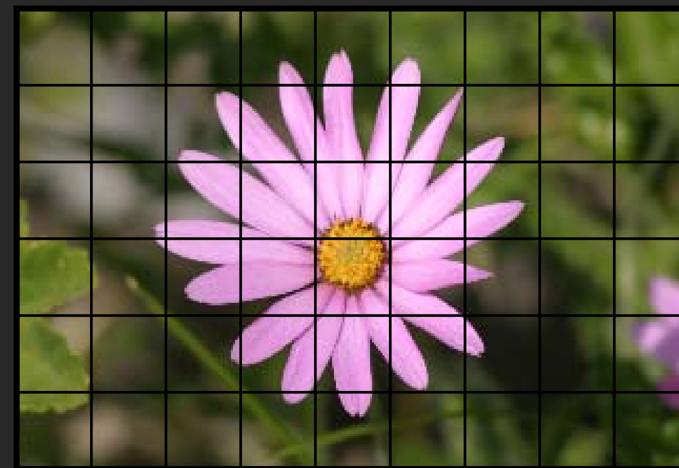
Region classification as denoising

Class labels



Noisy channel

Observations

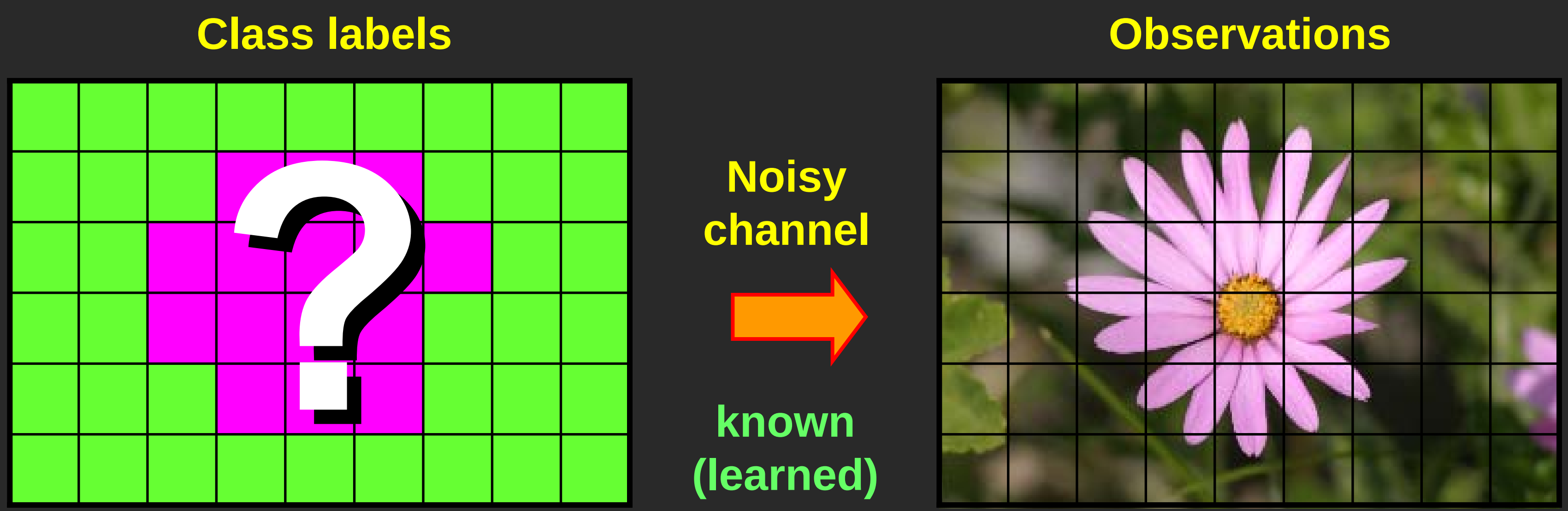


- The elements of the label sequence **x** are independently corrupted by the noisy channel *Q* to obtain the observation sequence **y**

$$P(\mathbf{Y}=\mathbf{y}\,|\,\mathbf{x}) = \prod_{i=1}^{n} Q(y_i \,|\, x_i)$$

Q: channel transition matrix

Region classification as denoising



- The elements of the label sequence **x** are independently corrupted by the noisy channel *Q* to obtain the observation sequence **y**
- Our goal is to design a *denoising procedure* to estimate **x** given *Q* and **y**

Compound decision approach

- Optimal decision rule: $\hat{x}_i = \arg\max_x P(X_i = x \mid \mathbf{y})$

$$\begin{aligned} P(x_i \mid \mathbf{y}) &= P(x_i \mid y_i, \mathbf{y}_{-i}) \\ &\propto P(y_i \mid x_i, \mathbf{y}_{-i}) P(x_i \mid \mathbf{y}_{-i}) \\ &= Q(y_i \mid x_i) P(x_i \mid \mathbf{y}_{-i}) \end{aligned}$$

The whole observation sequence excluding y_i

- Simplification 1: replace whole sequence with local neighborhood (sliding window rule)

$$\hat{x}_i = \arg\max_x Q(y_i \mid x) P(x \mid \mathbf{y}_{N(i)})$$

- Simplification 2: define a *context function* ξ

$$\hat{x}_i = \arg\max_x Q(y_i \mid x) P(x \mid \xi_i)$$

Function of $\mathbf{y}_{N(i)}$

Estimating the contextual prior

- Decision rule:

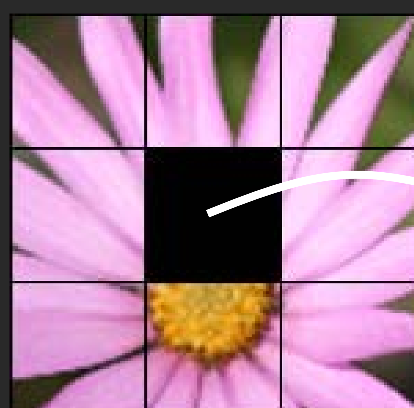
$$\hat{x}_i = \arg\max_x Q(y_i \mid x) P(x \mid \xi_i)$$

Channel transition matrix
Assumed known (i.e. learned
at training time)

Probability of unobserved
clean symbol given
observed context

- We need an estimate of $P(x|\xi_i)$, but we only have direct access to $P(y|\xi_i)$

context ξ_i



$P(\boxed{x}|\xi_i)$

Statistical inversion

- How to go from *output distribution* $P_y = P(y|\xi)$ to *input distribution* $P_x = P(x|\xi)$?
- We have $P(y|\xi) = \sum_x Q(y|X=x)P(X=x|\xi)$

  or $P_y = Q^T P_x$
- Estimating P_x: $\hat{P}_x = Q^{-T} P_y$
- More robust approach: find input distribution that minimizes KL-divergence between observed and predicted output distributions

$$\hat{P}_x = \arg\min_P D(P_y \| Q^T P)$$

Summary of algorithm

Training:

- Learn channel transition matrix *Q* from labeled data

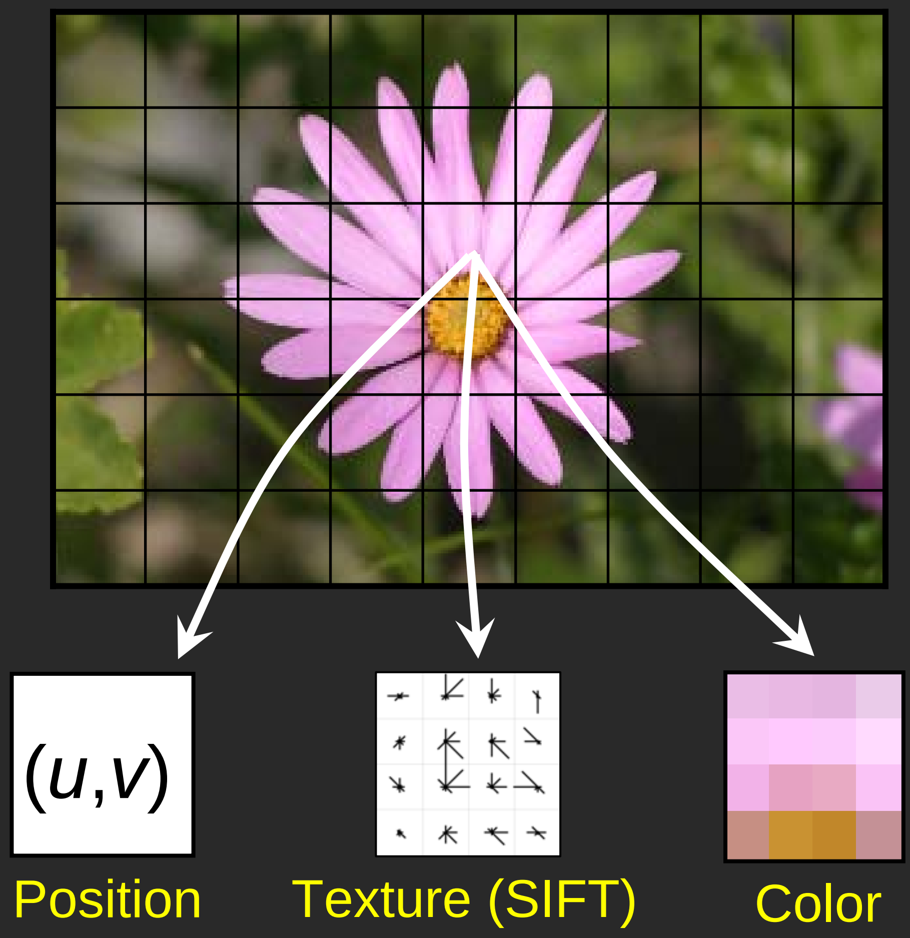
Testing:

- For each test patch *i*:
  - Estimate output distribution $P(y|\xi_i)$
  - Obtain *contextual prior* $P(x|\xi_i)$ by statistical inversion
  - Find x_i by MAP rule

$$\hat{x}_i = \arg\max_x Q(y_i \mid x) P(x \mid \xi_i)$$

Implementation: Feature extraction

- Three types of image features



Similar to Verbeek & Triggs (2007)

Implementation: Feature extraction

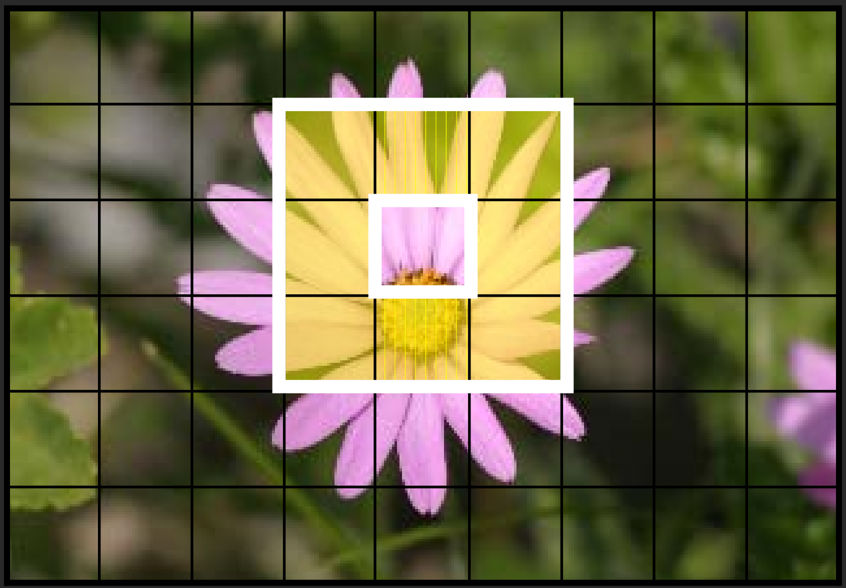
- Observation model 1: Quantizer
  - Observation *y* is a tuple of discrete quantizer labels for each feature
  - Channel transition matrix is estimated by Naive Bayes
- Observation model 2: Classifier
  - Observation *y* is the output of an SVM classifier
  - Channel transition matrix is the confusion matrix of the classifier on a validation dataset

Context representation

- Orderless context function: ξ_i is the histogram of observation labels in a neighborhood of region *i*
- Estimating P(y|ξ_i): *k* nearest neighbors (*k*=500)

Context representation

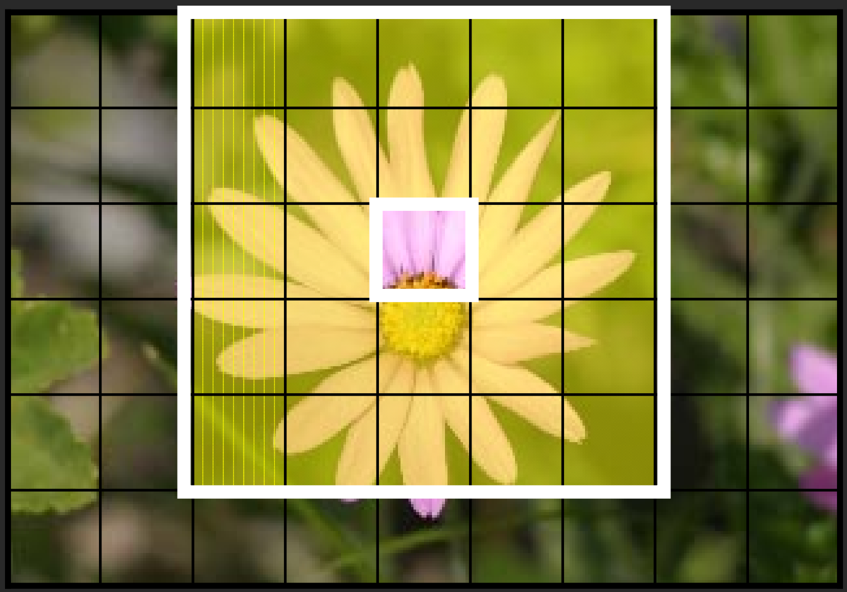
- Orderless context function: ξ_i is the histogram of observation labels in a neighborhood of region *i*
- Estimating $P(y|\xi_i)$: *k* nearest neighbors (*k*=500)
- Context size



Neighborhood size 1

Context representation

- Orderless context function: ξ_i is the histogram of observation labels in a neighborhood of region *i*
- Estimating $P(y|\xi_i)$: *k* nearest neighbors (*k*=500)
- Context size



Neighborhood size 2

Effect of context size

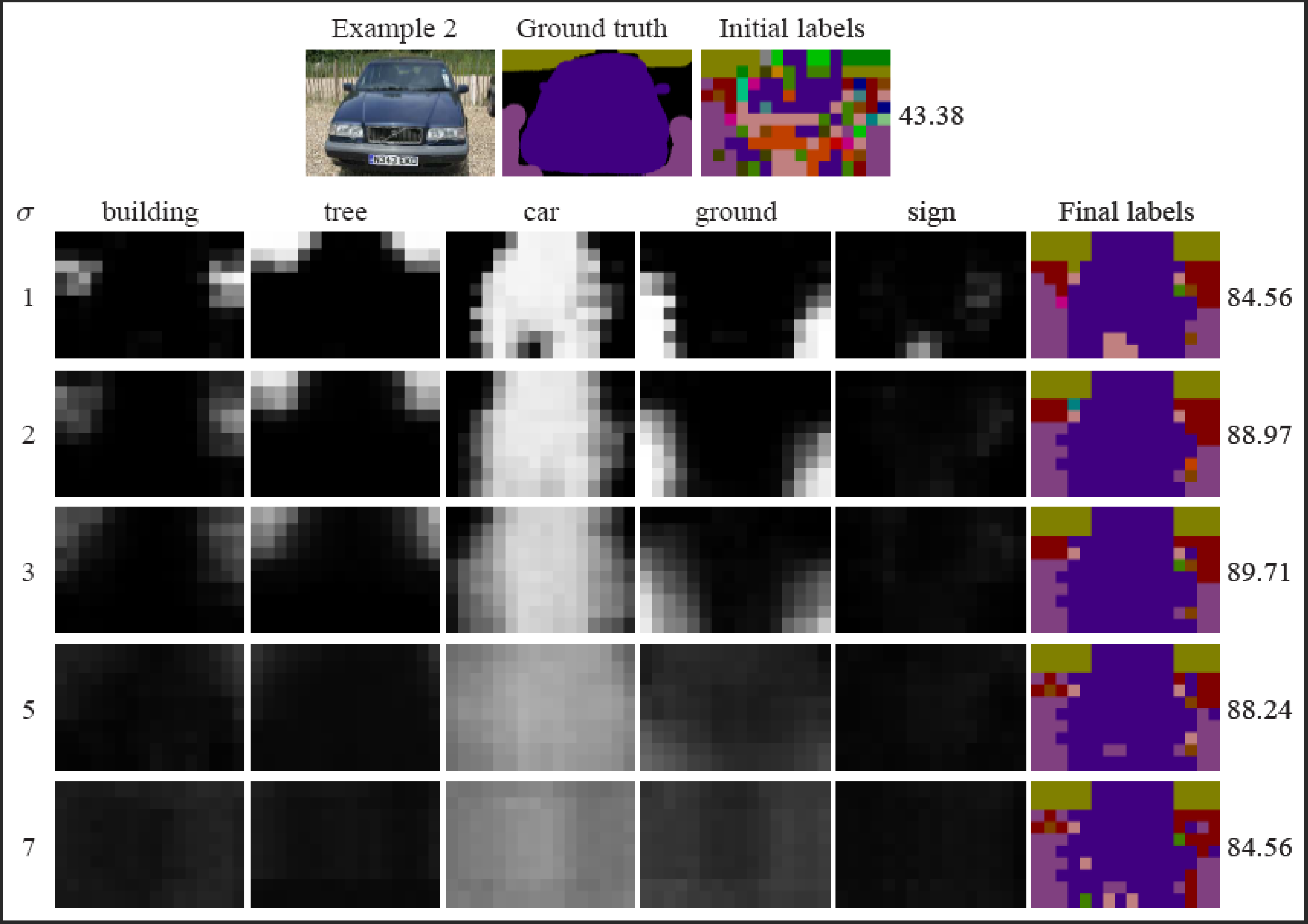


Image-level context

- When neighborhood radius becomes large enough to encompass the whole image, all regions in that image share the same context
- The estimate of P(y|ξ) is given by the histogram of observation labels in the image
- This reduces to pLSA!

$$P(y \mid \xi) = \sum_{x} Q(y \mid X = x) P(X = x \mid \xi)$$

Context (ξ) = document index

Label (x) = topic

Observation (y) = word

Enriching the context function

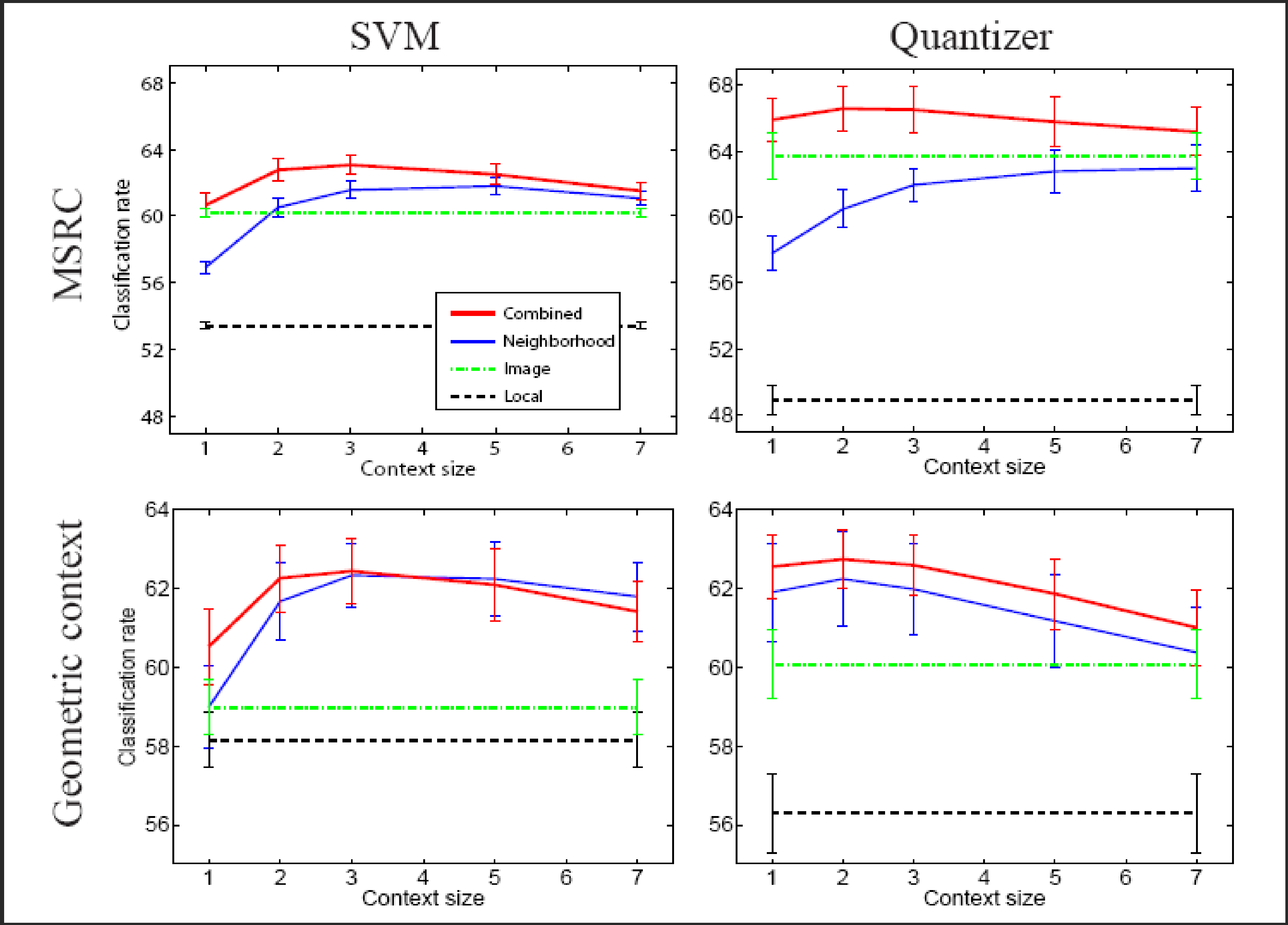
- Context ξ can depend not only on the observations in a local neighborhood, but also on estimated labels in that neighborhood
- An initial estimate of labels can come from the image-level context
- Denoising can be applied repeatedly with improved contextual estimates – similar to ICM

Datasets

- MSRC dataset (Shotton et al. 2006)
  - 594 images, 21 classes

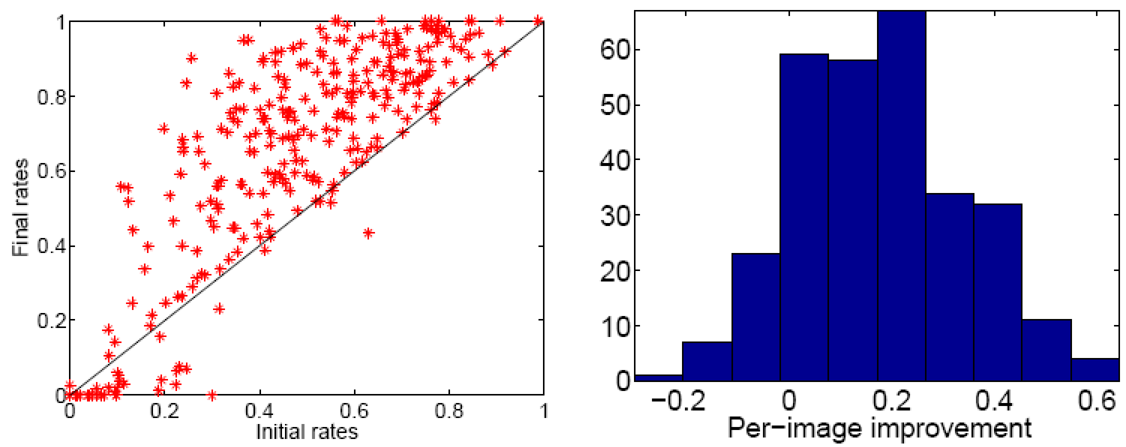
- Geometric context dataset (Hoiem et al. 2005)
  - 300 images, 7 classes

Context vs. local model

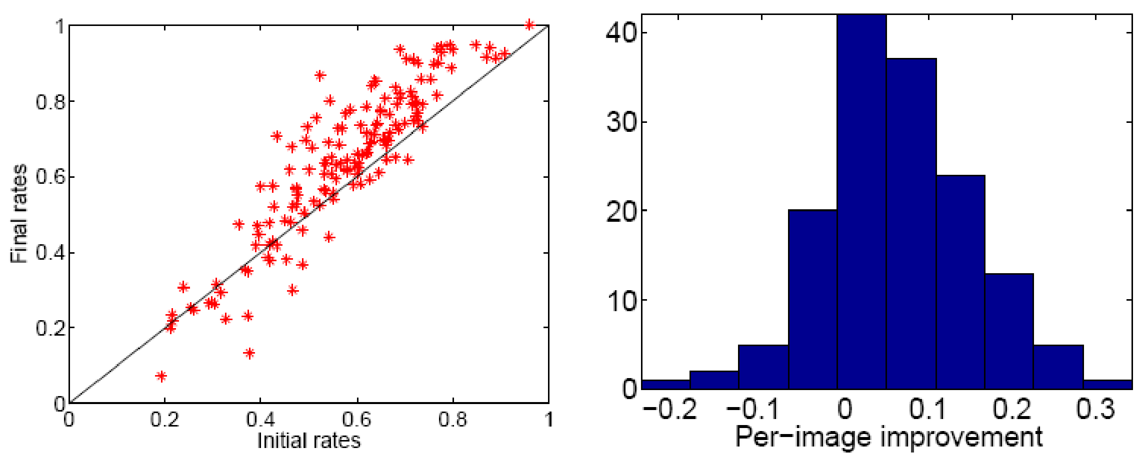


Per-image improvements

MSRC



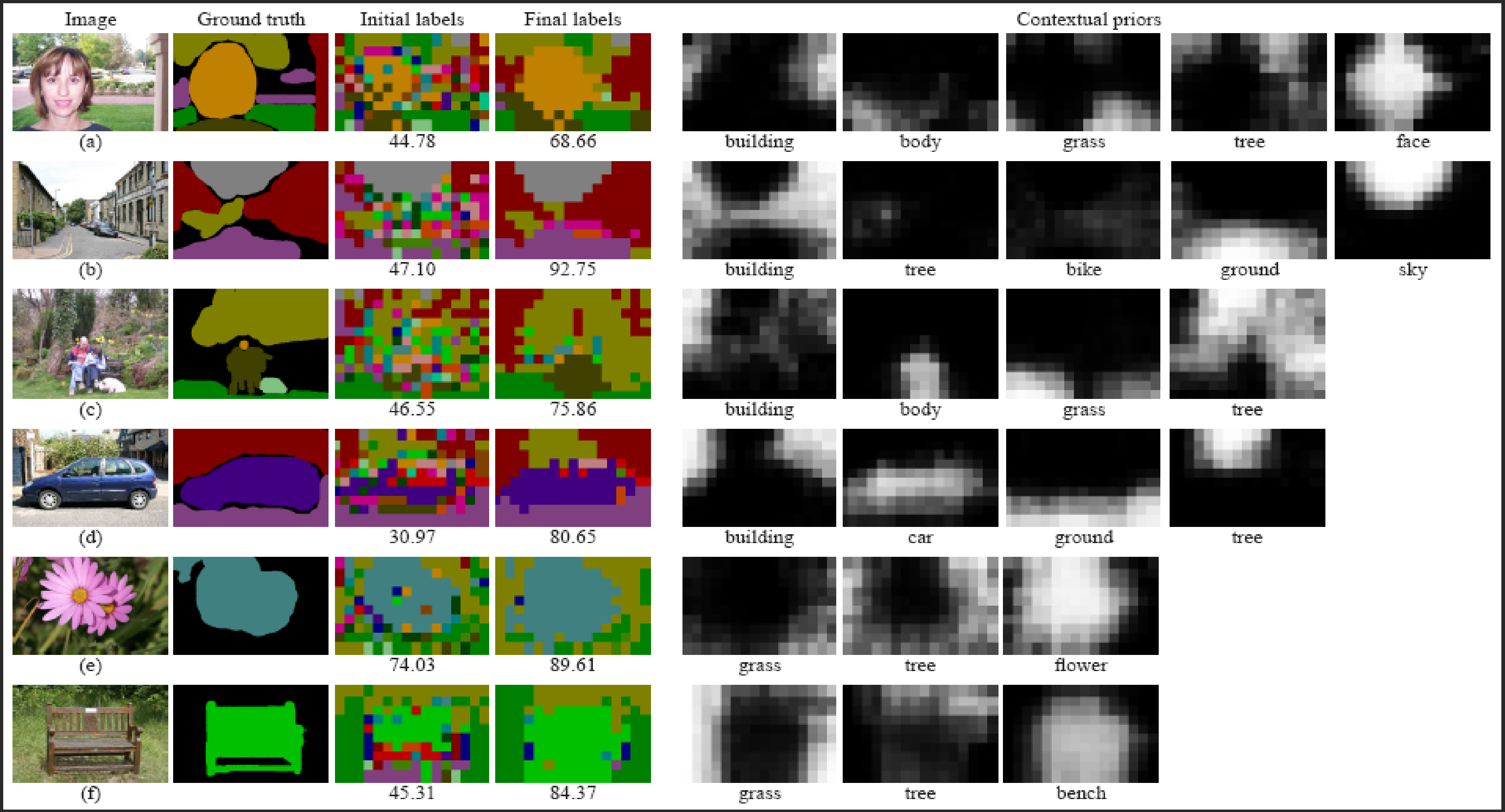
Geometric context



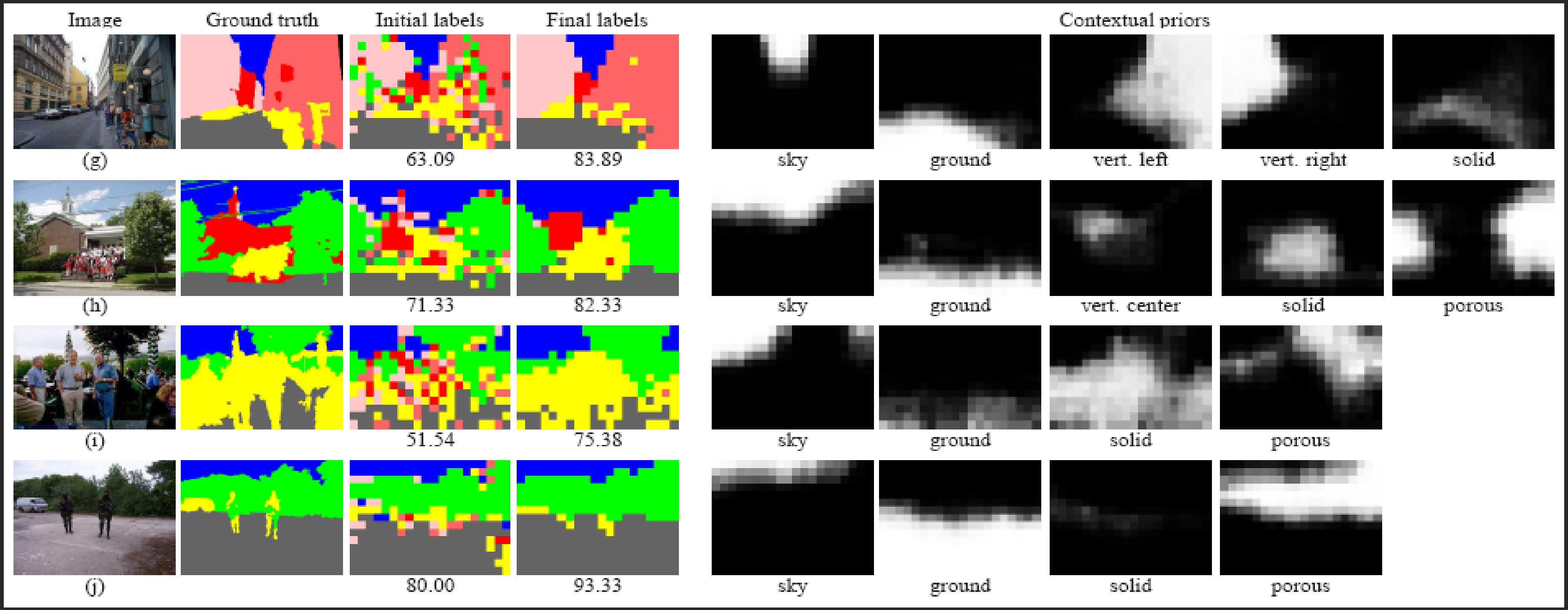
Initial rates: local quantizer model

Final rates: combined context, neighborhood size 2

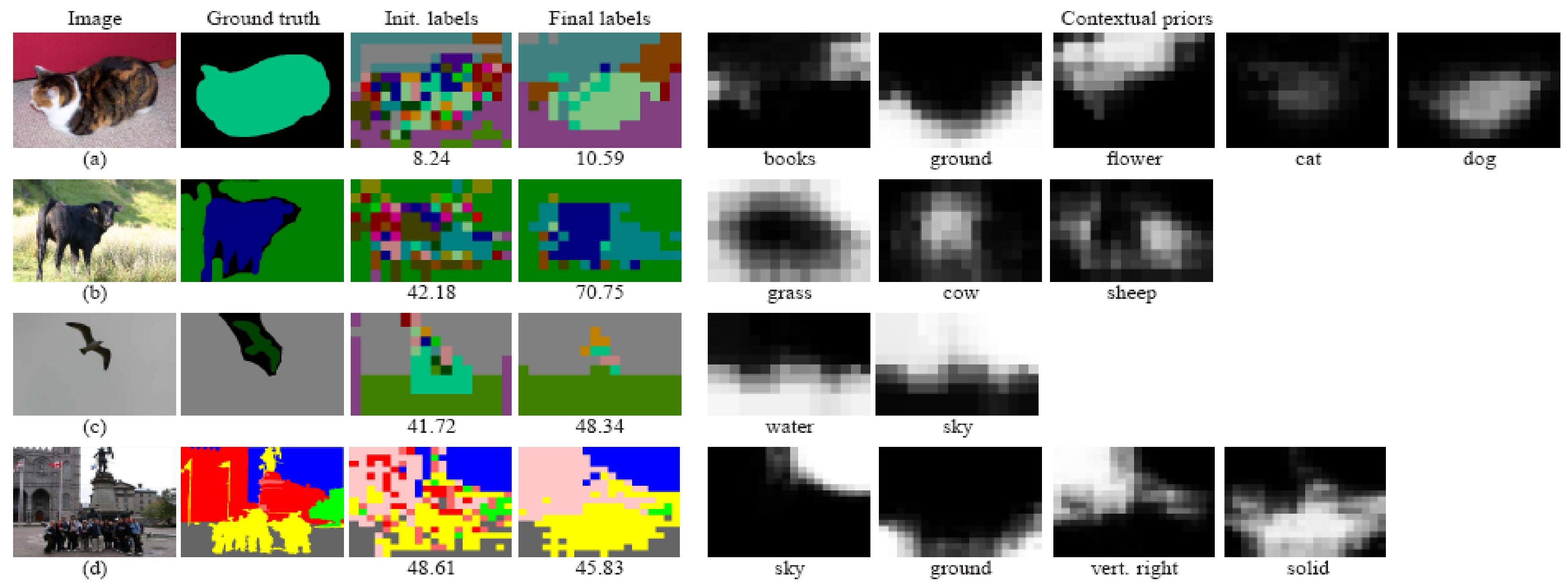
Examples on MSRC dataset



Examples on geometric context dataset



A few failures



Summary

- Contextual region classification as denoising
  - Image observations can be regarded as a systematically "corrupted" version of the underlying class labels
  - All we need to know is the mapping converting labels to observations (local likelihood)
  - Can denoise the output of any black-box local classifier provided we know its confusion matrix
- An empirical Bayes approach
  - A spatially varying prior over class labels is obtained from the unlabeled test data by statistical inversion
  - No specific assumptions about the distribution of the label sequence
  - No need to learn a contextual model from training data

Current limitations

- The transition matrix has to be estimated from labeled training data
  - Use EM to simultaneously estimate transition matrix and contextual prior?
- Estimation of contextual probabilities is very slow
  - Use fast approximate nearest neighbors or context hashing