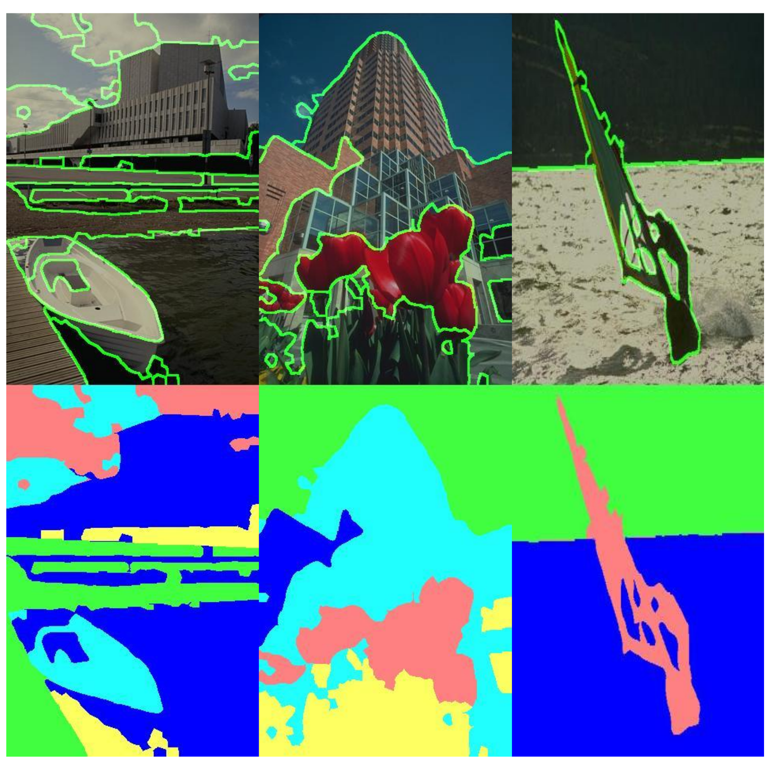
Spatial distance dependent Chinese restaurant processes for image segmentation

Soumya Ghosh[1], Andrei B. Ungureanu[2], Erik B. Sudderth[1], David M. Blei[3]
1: Department of Computer Science, Brown University , 2: Morgan Stanley, 3: Department of Computer Science, Princeton University

Image Segmentation



Goals

- Split images into “homogeneous” regions/segments/clusters.
- Develop a statistical model that automatically infers an appropriate *number* of segments for each image, and handles segments of widely varying sizes.

Contributions

- We explore the *spatial distance dependent Chinese restaurant process* as a consistent prior over spatial image partitions.
- We develop a *hierarchical* version, and demonstrate its ability to model *human-like* segmentations.
- We perform controlled comparisons against other recent Bayesian nonparametric models.

Statistical Model

- A mixture model with a ***spatial*** Bayesian nonparametric prior**.**

Image Representation

- An image is a collection of ≈1000 pre-computed ***super-pixels***.
- Super-pixels are described by stacked color and texture histograms of constituent pixels. $x_i = (x_i^t, x_i^c)$
- Color is represented by a 120-bin ***HSV*** color space, and texture by a 128-bin ***texton*** histogram.

Likelihood

- Mixture components are associated with multinomial distributions over the color and texture histograms:

$$p(x_i^t, x_i^c \mid z_i, \theta) = \mathrm{Mult}(x_i^t \mid \theta_{z_i}^t)\,\mathrm{Mult}(x_i^c \mid \theta_{z_i}^c)$$

$$\theta \sim \mathrm{Dir}(\lambda)$$

Distance dependent Chinese restaurant process (ddCRP)

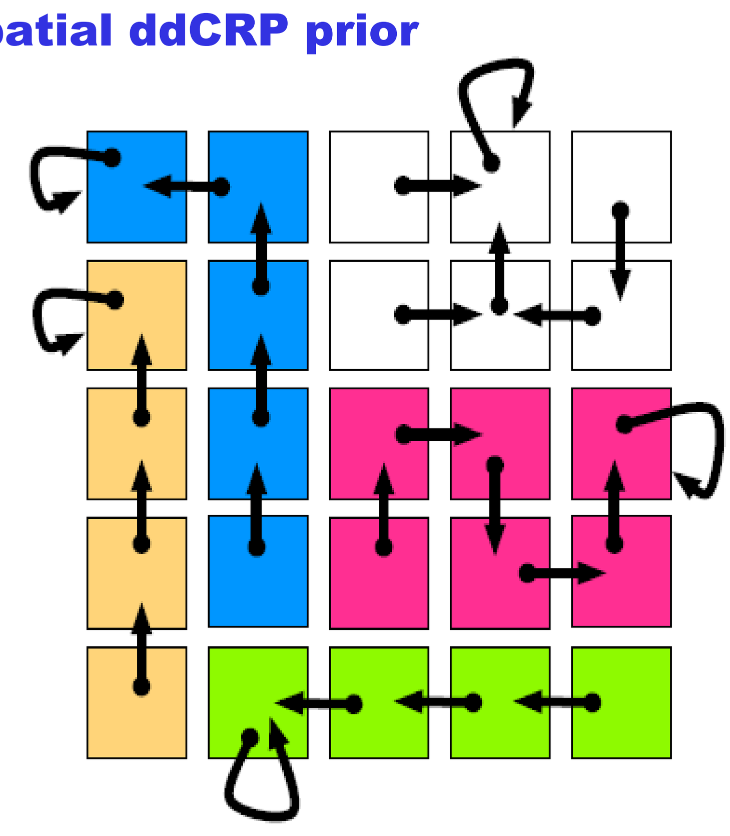
- The ddCRP extends the traditional Chinese restaurant process (CRP). It prefers placing data instances closer in an “external”, sense in the same cluster.
- Each customer (data instance) links to others with probability proportional to the distance between them:

$$p(c_i = j \mid D, f, \alpha) \propto \begin{cases} f(d_{ij}) & j \neq i \\ \alpha & j = i \end{cases}$$

Distance matrix *Decay function*

- The links determine the partition. Two customers belong to the same component if they are reachable.
- If each customer is allowed to connect to all preceding customers in some order, the Chinese restaurant process is recovered.

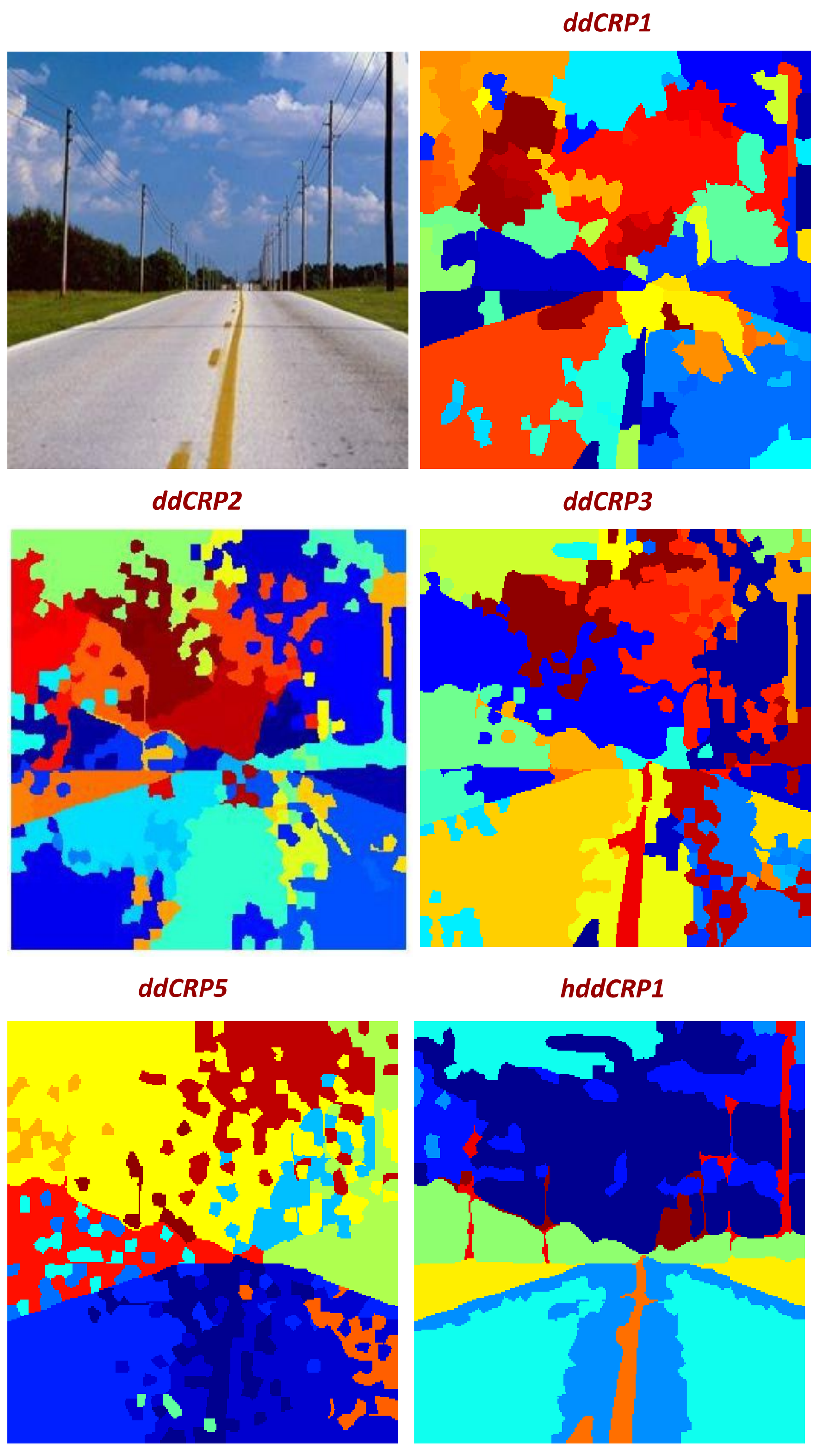
Spatial ddCRP prior



- Spatial distance between super-pixels is used to define a prior over image partitions.
- Distance between two super-pixels is the number of hops needed to reach one from the other.
- The decay function used is $f(d) = 1[d \leq a]$
- Setting a = 1, super-pixels can only directly connect to neighboring super-pixels. This *guarantees spatially connected* segments.

Hierarchical region level ddCRP

- Human segmentations contain regions larger than those produced by a ddCRP with a = 1.
- Two alternatives: increase *a,* or introduce a hierarchy, which groups regions into larger ones.



- The hierarchical model produces more human-like partitions, by avoiding isolated super-pixels.
- It extends the traditional Chinese restaurant franchise representation of the HDP by modeling each restaurant with a ddCRP instead of a CRP.

Summary of generative model

- For each customer, sample customer assignments

$$c_i \sim \mathrm{ddCRP}(\alpha, f, D)$$

This determines the table assignments $t_{1:N}$

- For each table *t*, sample region assignments

$$k_t \sim \mathrm{CRP}(\gamma)$$

- For each region, sample parameters $\phi_k \sim G_0$
- For each super-pixel, independently sample observed data

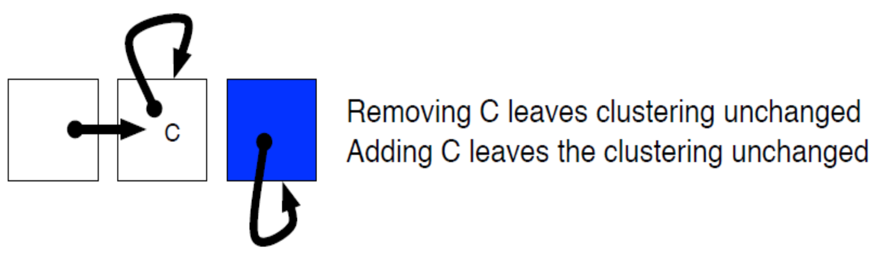
$$x_i \sim p(.\,|\,\phi_{z_i}), \quad z_i = k_{t_i}$$

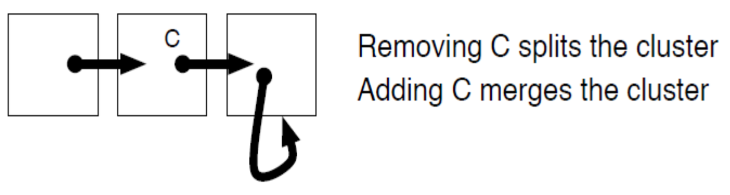
Inference: Gibbs Sampling

Distance dependent CRP inference

$$p(c_i \mid c_{-i}, x_{1:N}, D, \alpha, \lambda) \propto p(c_i \mid D, \alpha)\, p(x_{1:N} \mid z(c_{1:N}), \lambda)$$

- The *sampler* does not assume *exchangeability*.
- Split and merge behavior leads to *fast* mixing:





- Likelihood decomposes as:

$$p(x_{1:N} \mid z(c_{1:N}), \lambda) = \prod_{k=1}^{K(c_{1:N})} p(x_{z(c_{1:N})=k} \mid z(c_{1:N}), \lambda)$$

- Update equation:

$$p(c_i \mid c_{-i}, x_{1:N}, D, \alpha, \lambda) \propto \begin{cases} p(c_i \mid D, \alpha)\Gamma(x, z, \lambda) & \text{if } c_i \text{ joins } l \text{ and } m \\ p(c_i \mid D, \alpha) & \text{otherwise} \end{cases}$$

where

$$\Gamma(x, z, \lambda) = \frac{p(x_{z(c_{1:N})=k} \mid \lambda)}{p(x_{z(c_{1:N})=l} \mid \lambda)\, p(x_{z(c_{1:N})=m} \mid \lambda)}$$

Region level ddCRP inference

- Likelihood depends on all super-pixels in the same region, not just the same segment.
- Region assignments need to be re-sampled according to:

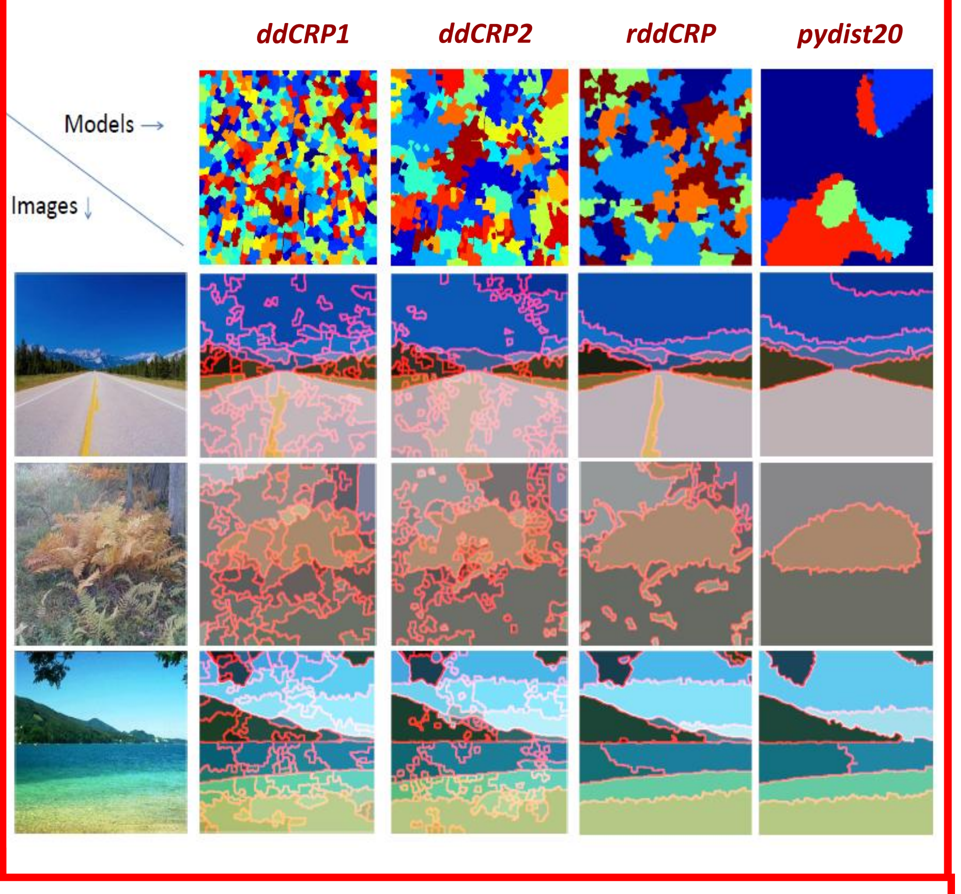
$$p(k_t = l \mid k_{-t}, x_{1:N}, t(c_{1:N}), \gamma, \lambda) \propto \begin{cases} m_l^{-t}\, p(x_t \mid x_{-t}, \lambda) & \text{old } l \\ \gamma p(x_t \mid \lambda) & \text{new } l \end{cases}$$

Results

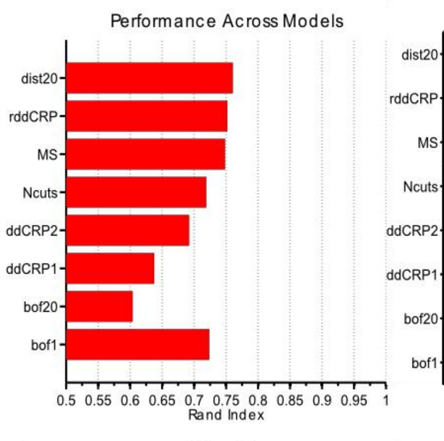
- Benchmarked on a subset of Oliva and Torralba’s natural scene category dataset. ***100 images*** were chosen at random from each of the ***eight scene categories.***

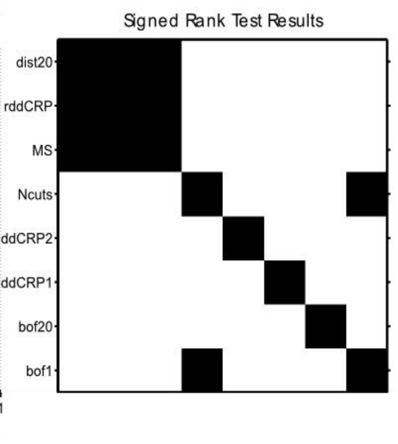
- Compared various proposed ddCRP models to normalized cuts (spectral clustering), mean shift (classical kernel density estimate), and spatially dependent Pitman-Yor processes (via Gaussian processes, pydist20).

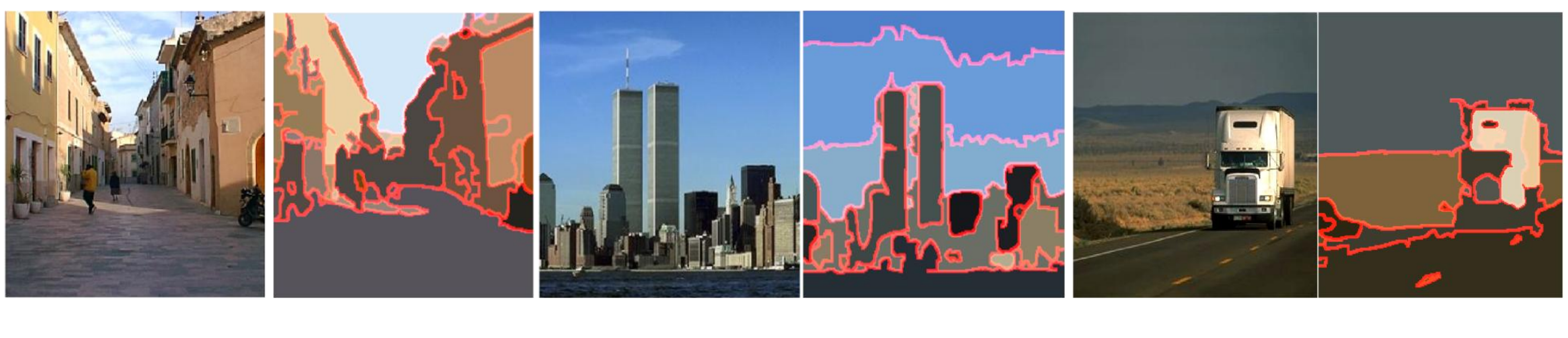
Qualitative model comparison

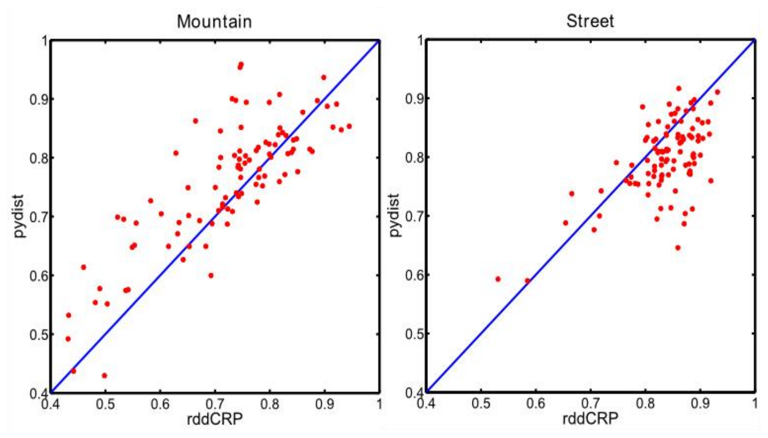


Quantitative model comparisons, and example segmentations by the hierarchical region-level ddCRP









Top left: Average segmentation performance across the eight categories. Right: Dark pixels indicate pairs that are statistically indistinguishable.
Bottom left: Scatter plots comparing the pydist20 and rddCRP methods on the Mountain and Street scene categories. Right: Example segmentations produced by the rddCRP.