Unsupervised Learning of Stereo Vision with Monocular Cues

David McAllester and Hoang Trinh
Toyota Technological Institute at Chicago (TTI-Chicago)
Scene Understanding
Conditional Random Fields (Lafferty et al. 2001)

\[ P_\beta(y|x) = \frac{1}{Z(x)} e^{-E_\beta(x,y)} \]

\[ Z(x) = \sum_y e^{-E_\beta(x,y)} \]

\[ \beta^* = \arg\min_{\beta} \left( \sum_i \ln \frac{1}{P_\beta(y_i|x_i)} \right) + \frac{1}{2} \lambda \|\beta\|^2 \]
Hidden CRFs (Quattoni et al. 2007)

\[
P_\beta(y, z | x) = \frac{1}{Z(x)} e^{-E_\beta(x, y, z)}
\]

\[
Z(x) = \sum_{y, z} e^{-E_\beta(x, y, z)}
\]

\[
P(y | x) = \sum_{z} P(y, z | x)
\]

\[
\beta^* = \arg\min_{\beta} \left( \sum_{i} \ln \frac{1}{P_\beta(y_i | x_i)} \right) + \frac{1}{2} \lambda \|\beta\|^2
\]
## Related Formalisms

There are 8 combinations of structured vs. binary, CRF vs. SVM, and fully labeled vs. latent.

<table>
<thead>
<tr>
<th></th>
<th>SVM</th>
<th>CRF</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>binary output</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fully labeled</td>
<td>SVM</td>
<td>Logistic Regression</td>
</tr>
<tr>
<td>latent</td>
<td>L SVM</td>
<td>Latent Logistic Regression</td>
</tr>
<tr>
<td><strong>structured output</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fully labeled</td>
<td>Structured SVM</td>
<td>CRF</td>
</tr>
<tr>
<td>latent</td>
<td>Structured L SVM</td>
<td>Hidden CRF</td>
</tr>
</tbody>
</table>
Indirect CRFs

Training pairs \((x_1, y_1), \ldots, (x_N, y_N)\) are used to train a CRF \(P_{\beta_z}(z|x)\) for \(z\) hidden.

\[
P_{\beta_z,\beta_y}(y|x) = \sum_z P_{\beta_z}(z|x) P_{\beta_y}(y|x, z)
\]

\(P_{\beta_z}(z|x)\) and \(P_{\beta_y}(y|z, x)\) are CRFs.

\[
\beta^*_z, \beta^*_y = \arg \min_{\beta_z, \beta_y} \left( \sum_i \ln \frac{1}{P_{\beta_z,\beta_y}(y_i|x_i)} \right) + \frac{1}{2} \lambda \|\beta\|^2
\]
Classical Stereo Vision as an Indirect CRF

\[ E_\lambda(z) = \sum_{p,q \in N(p)} \lambda |z(p) - z(q)| \]

\[ y(p) = x(p + z(p)) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma) \]

\[ E_\sigma(x, y, z) = \sum_p \frac{(y(p) - x(p + z(p)))^2}{2\sigma^2} \]

\[ z^* = \arg\min_z E_\lambda(z) + E_\sigma(x, y, z) \]
Unsupervised Parameter Tuning

Zhang and Seitz (CVPR 2005) used the following for a classical stereo algorithm.

$$\beta_i^* = \arg\max_{\beta} P_\beta(y_i|x_i)$$

Here $\beta$ is low dimensional (5 parameters for Zhang and Seitz) and we can tune to a single image pair.

However, we would like to train high dimensional monocular predictors.
HOG Features

- Take an 8 pixel \( \times \) 8 pixel square image region (a “cell”).
- At each pixel compute a gradient vector.
- Quantize the gradient orientation into 9 bins.
- In each bin sum the magnitude of the gradients in that bin.
- Normalize the magnitude of the resulting 9 dimensional vector.
HOG as a Surface Orientation Cue
Analysis of Isotropic Texture

\[ \frac{H_{\text{min}}}{H_{\text{max}}} = \cos^3 \Psi \]

\( \Psi \) is the angle between the surface normal and the ray to the camera.
A Slanted Plane Model

We use an over-segmentation (Felzenszwalb and Huttenlocher) with a latent slanted plane at each segment.

\[ d(x, y) = d_0 + Ax + By \]

\( d(x, y) \) is image pair disparity at image coordinates \( x, y \).

\[ B \approx -\frac{d(p) \tan \Psi(p)}{f} \]
The Energy Function

A match energy.

\[ E_M = \sum_p \sum_k \lambda_k \left( \Phi_k^x(p) - \Phi_k^y(p + d(p)) \right)^2 \]

A smoothness energy.

\[ E_S = \sum_{i,j} \min \left( \tau_S, \sum_{(p,q) \in B_{i,j}} \lambda_S |d(p) - d(q)| \right) \]

A HOG energy.

\[ E_T = \sum_p \min \left( \tau_T, \lambda_A (d(p)(\beta_A \cdot H(p)) - A_i)^2 \right. \\
\left. + \lambda_B (d(p)(\beta_B \cdot H(p)) - B_i)^2 \right) \]
## Results

|                                | RMS Disparity Error (pixels) | Average Error $|\log_{10} Z - \log_{10} \hat{Z}|$ |
|--------------------------------|-------------------------------|-----------------|
| Saxena et al.                  |                               | .074            |
| Unsuper., Notexture            | 1.158                         | .073            |
| Unsuper., Texture              | 1.081                         | .069            |
| Super., Notexture              | 1.071                         | .069            |
| Super., Texture                | 1.001                         | .063            |
Hard EM

\[ z_i = \arg\max_z P_{\beta z}(z|x_i)P_{\beta y}(y_i|x_i, z) \]

\[ = \arg\min_z E_{\beta z}(x_i, z) + E_{\beta y}(x_i, y_i, z) + \ln Z_{\beta y}(x_i, z) \]

\[ \approx \arg\min_z E_{\beta z}(x_i, z) + E_{\beta y}(x_i, y_i, z) \quad (1) \]

\[ \beta_z = \arg\max_{\beta_z} \prod_i P_{\beta z}(z_i|x_i) \quad (2) \]

\[ \beta_y = \arg\max_{\beta y} \prod_i P_{\beta y}(y_i|z_i, x_i) \quad (3) \]

We do (1) with max-product particle belief propagation.
We do (2) and (3) with contrastive divergence (Hinton 2002).
Max-Product Particle Belief Propagation

Consider an MRF with continuous values at each node.

• Initialize a value at each node (we need a slanted plain assigned to each segment).

• Until convergence:
  – Stochastically generate additional candidate values at each node near the currently assigned value (we use a Gaussian proposal distribution).
  – Run your favorite max-product algorithm over the resulting discrete MRF to find a new best value at each node.
Contrastive Divergence

\[
\ln \frac{1}{P_{\beta y}(y_i | x_i, z_i)} = E_{\beta y}(x_i, y_i, z_i) + \ln Z_{\beta y}(x_i, z_i)
\]

\[
\nabla_{\beta z} \ln \frac{1}{P_{\beta y}(y_i | x_i, z_i)} = \nabla_{\beta y} E_{\beta y}(x_i, y_i, z_i) - E_{y \sim P_{\beta y}(y | x_i, z_i)} \left[ \nabla_{\beta y} E_{\beta y}(x_i, y, z_i) \right]
\]

We can estimate the expectation with MCMC samples. In contrastive divergence we start the MCMC process at \( y_i \) and take only a few steps.

Contrastive divergence yields a consistent estimator for \( \beta \) whenever the expected update direction is the gradient of a convex potential function.
Contrastive Divergence for a Standard MRF Process

Consider an MRF $P_\beta(x)$.

Let $x_{-i}$ denote the restriction of $x$ to the coordinates other than $i$.

Consider an MCMC process which selects a node at random and stochastically selects a new value under the conditional distribution on other nodes.

Single step contrastive divergence for this process follows the gradient of (log) pseudo-likelihood:

$$\tilde{P}_\beta(x) = \prod_i P_\beta(x_i | x_{-i})$$
Contrastive Divergence for a Differential Metropolis Process

Consider a Metropolis process:

• Generate a neighbor $x'$ of the current point $x$ using a (symmetric) proposal distribution.

• If $E_{\beta}(x') \leq E_{\beta}(x)$ move to $x'$.

• If $E_{\beta}(x') > E_{\beta}(x)$ move to $x'$ with probability $e^{-\Delta E}$.

For a symmetric proposal distribution whose variance can be taken to zero we have that in the limit of a zero-variance proposal distribution the metropolis process defines a stochastic differential equation (A Langevin process) and the one-step contrastive divergence update follows the gradient of Hyvärinen’s score matching objective:

$$D_H(P, Q) = \mathbb{E}_{x \sim P} \left[ \frac{1}{2} \left\| \nabla_x \ln \frac{P(x)}{Q(x)} \right\|^2 \right]$$

Here $P$ is the true data source and $Q$ is the model distribution.
Warning

Under mild conditions minimizing log-pseudo-likelihood and minimizing log-likelihood both yield consistent parameter estimates.

However, when the data distribution is not in the parameterized class these estimates (models) are different even for infinite training data.
Summary

- We have formulated the notion of indirect CRF training for learning $P_\beta(z|x)$ from training data $(x_1, y_1), \ldots (x_N, y_N)$ with $z$ latent.

- We have demonstrated indirect CRF training of monocular depth cues from (unlabeled) stereo pairs.

- Experiments with shape from shading cues have not yet succeeded.