Special Course in Information Science IV:
Unsupervised Learning

Learning Mixture Models of Spacial Coherence

by Suzanna Becker and Geoffrey E. Hinton

Paulo Esquef
November 23^{th}, 2000

Introduction

Main Objective

- The use of Unsupervised Learning to discover spatially coherent properties of image scenes.

Strategy

- Maximization of the mutual information that extracted parameters from different parts of a sensory input convey about some common cause.

Application

- Depth estimation from random dot stereograms of curved surfaces with depth discontinuities.

Propositions

- A model to detect and reject depth discontinuities in curved surfaces.
- A multiple model scheme in which a set of expert interpolators are developed to be invoked according to the location of the depth discontinuity.

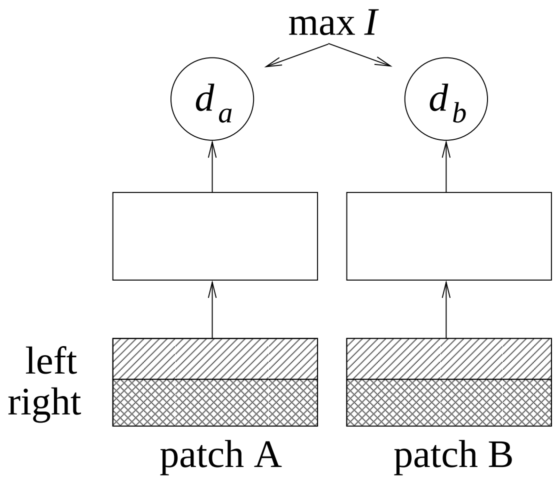
Learning Spatially Coherent Features in Images

Basic Stereo Net

- Basic Assumption: parameters vary smoothly over space or time so that adjacent sensoring modules can used to extract mutually predictable features from neighboring image patches.
- Maximizing

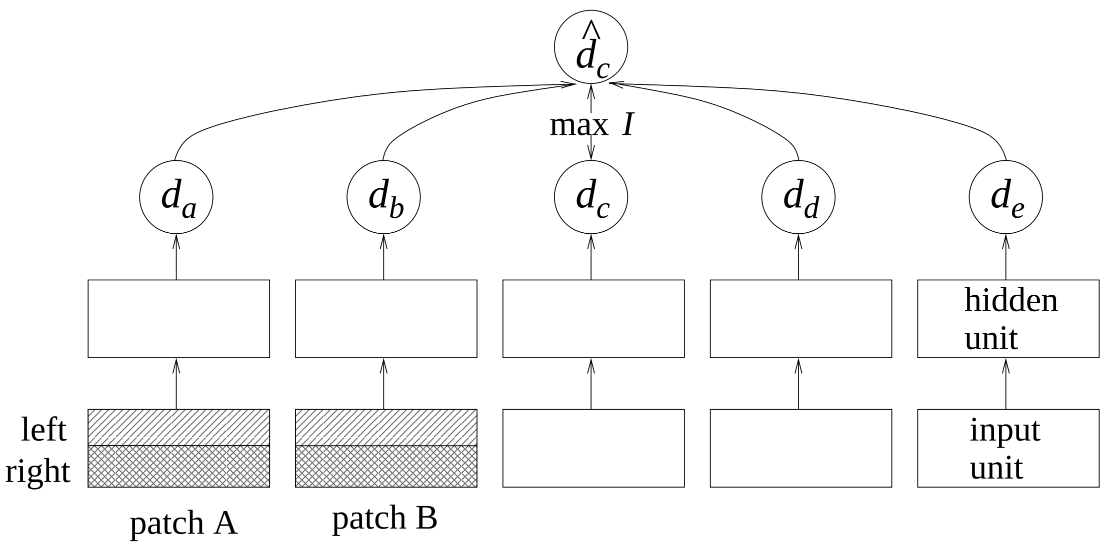
$$I = 0.5 \log \frac{V(d_a + d_b)}{V(d_a - d_b)},$$

where d_a and d_b are scalar outputs of the modules and V is the variance over the training, one gets as pure a version as possible of the underlying signal.



The Interpolating Net

- The model of coherence assumed in the basic stereo net was that an underlying parameter at one location is approximately equal to the parameter at a neighboring location.
- The previous assumption is fine for front parallel surfaces but not for slanted or curved surfaces.
- A more general model assumes the parameter at one location as a linear function of the parameters at nearby locations.
- In this case, the contextual $\hat{d}_c$ was computed as a weighted sum of the outputs of *two* adjacent modules on either side.



Continued

- To speed up the learning of the interpolating weights a two-stage training scheme was used, in which the agreement between neighboring locally computed outputs were maximized first, then followed by information maximization between the top units and back-propagation of gradients to the hidden units.
- The two-stage training procedure also increases the quality of the solution.

Mixture Models of Coherence

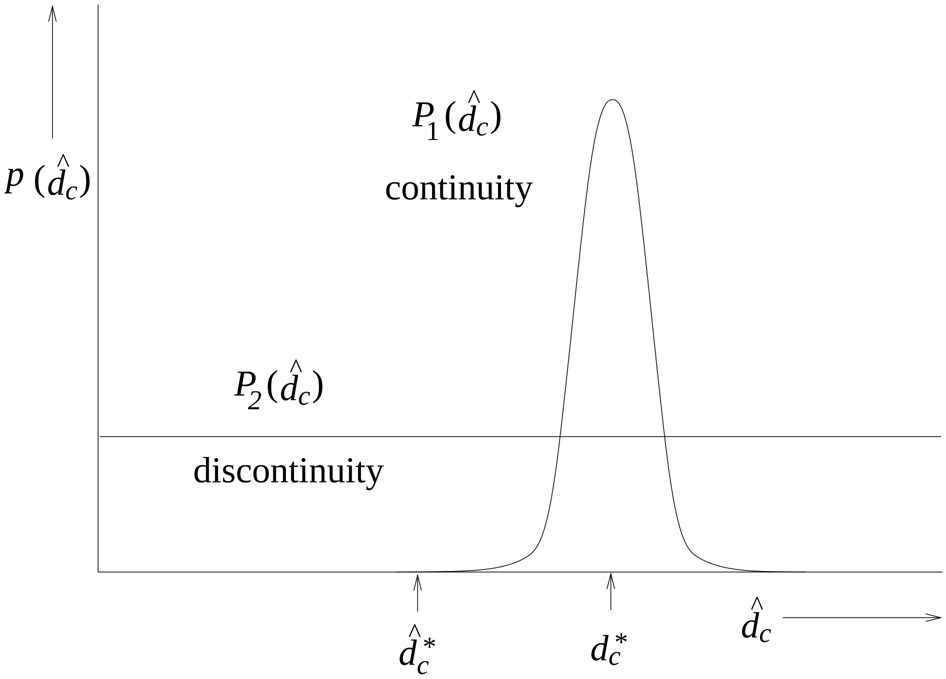
Objectives

- Deal with more than one simple type of coherence in images.
- Consider coherent structures at several scales.
- Deal with the presence of depth discontinuity.
- Develop multiple models of coherence by having a mixture of distributions.

Throwing out Discontinuities

- Training data containing depth discontinuities will contribute considerable noise to the interpolating weights and to the depth estimates.
- Discontinuity cases can be treated as outliers and discarded by means of a detection scheme.
- A soft decision can be performed by using a mixture model in which $d_c - \hat{d}_c$ is assumed to be drawn from a zero-mean Gaussian if it is a *continuity* case and from an uniform distribution if it is a *discontinuity* case.

Throwing out Discontinuities (Cont. I)



- For each training case the probability of a continuity case is estimated by

$$p_{\text{cont}}(d_c - \hat{d}_c) = \frac{N[d_c - \hat{d}_c, 0, V_{\text{cont}}(d_c - \hat{d}_c)]}{N[d_c - \hat{d}_c, 0, V_{\text{cont}}(d_c - \hat{d}_c)] + k_{\text{discont}}}$$

  where N is a Gaussian and k_{discont} is a constant representing a uniform density.

- Assuming that no information is transmitted in discontinuity cases, the *average* information transmitted by d_c and $\hat{d}_c$ can be optimized as:

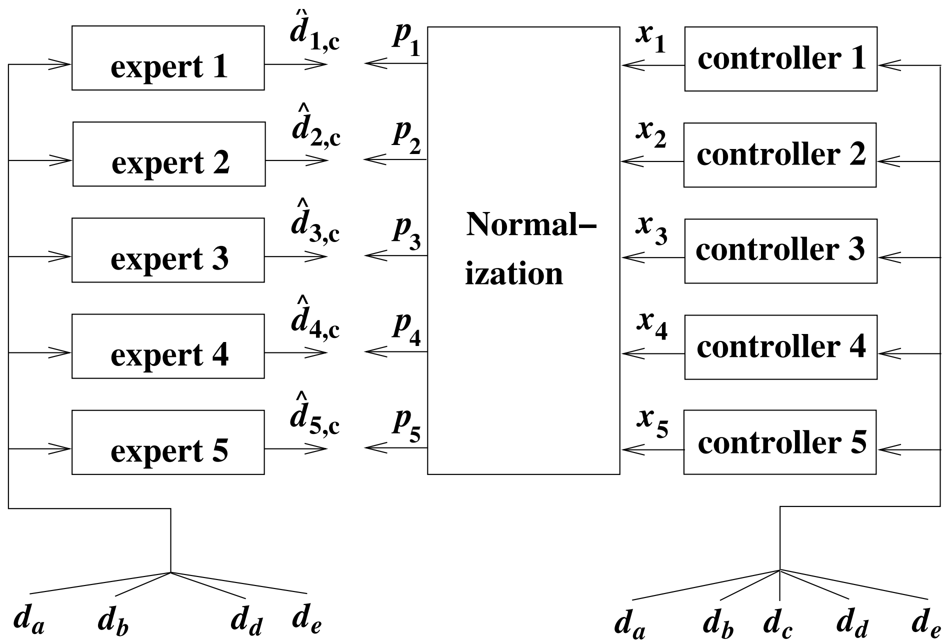
$$I^* = 0.5 \, \langle p_{\text{cont}}(d_c - \hat{d}_c) \rangle \log \frac{V_{\text{cont}}(d_c + \hat{d}_c)}{V_{\text{cont}}(d_c - \hat{d}_c)}$$

Throwing out Discontinuities (Cont. II)

- In this scheme, the choice of the initial value of $\hat{V}_{\text{cont}}(d_c - \hat{d}_c)$ is more critical than the choice of k_{discont} to the learning algorithm's performance.
- Many cases are treated as discontinuous if a small variance is chosen and *vice versa* but in the latter as the variance is shrunk during the learning, some discontinuous cases are detected.
- This scheme is reported to be good at rejecting discontinuity cases but does not produce good interpolators when there is a depth discontinuity between d_a and d_b or d_d and d_e.

Learning a Mixture of Interpolators

- The goal is to allow the network to detect the location of the discontinuities and use this information as a basis to choosing a well suited model to apply on a given case.
- This can be attained by using a set of competing interpolators.



- Each $\hat{d}_{i,c}$ is treated as the mean of a Gaussian distribution and p_i is treated as the mixing proportion of that Gaussian.

Learning a Mixture of Interpolators (Cont. I)

- Each interpolator will try to maximize the agreement between its output and the locally extracted depth as well as the probability that its model holds.
- The objective function maximized by the network is:

$$I^{**} = \sum_i \langle p_i \rangle \log \frac{V^i(\hat{d}_{i,c} + d_c)}{V^i(\hat{d}_{ic} - d_c)}$$

- The probability that each interpolator's model holds is computed by the controller units.
- The weights of both the controllers and the interpolating experts can be learned simultaneously through maximizing I^{**}.
- The outputs of the controllers are normalized as to represent the probability with which the system selects a particular expert.

Learning a Mixture of Interpolators (Cont. II)

- The normalization, computed as

$$p_i = \frac{e^{x_i^2/T\hat{\sigma}(x_i)^2}}{\sum_j e^{x_j^2/T\hat{\sigma}(x_j)^2}},$$

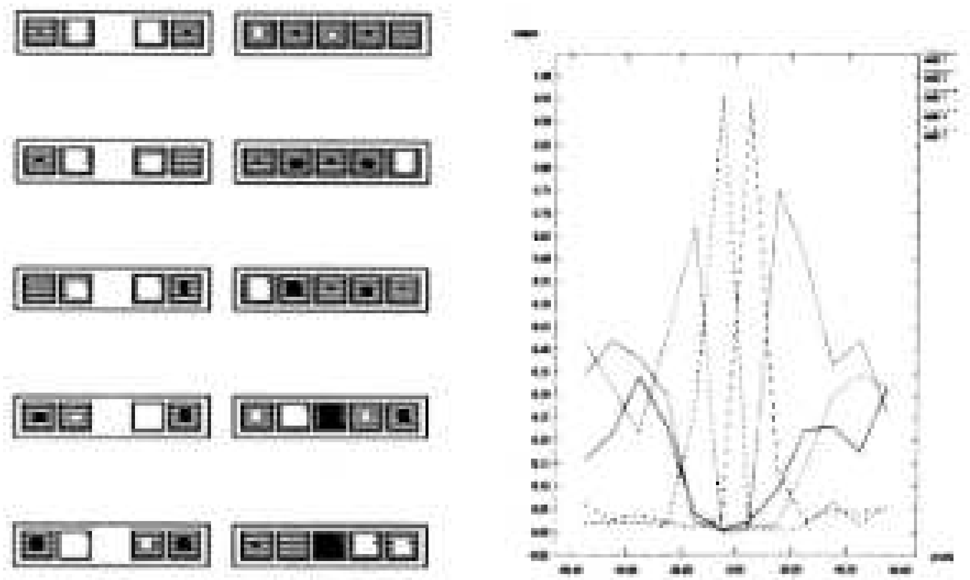
  allows the detection of a depth independently of the direction of contrast change and also discourage any one unit from trying to model all of the cases simply by having huge weights.

The Learning Procedure

- The lower layers are trained first.
- Then, the interpolators are pretrained using the continuity model (all of them learn similar weights).
- The weights in the lower layers are frozen and a small amount of noise is added to the interpolators weights. After that, the mixture model is used to improve the interpolators and train the controller units.

Results

- Four of the controller units are tuned to discontinuities at different locations.
- Each controller's corresponding interpolator uses the depths on the opposite side of the discontinuity to predict d_c.



Conclusions

- Two ways of modeling spatially coherent features in images were presented. Both of them were based on maximizing the information conveyed by nearby sensor units due to a common underlying cause.
- The first case tries to distinguish between patterns with and without depth discontinuities.
- The second one uses multiple models of coherence to represent and detect depth discontinuities at different locations.
- The presented scheme can be useful to applications such as scene segmentation and boundary detection in two-dimensional stereo images.