Unsupervised Learning of Stereo Vision with Monocular Cues

Hoang Trinh and David McAllester
{ntrinh,mcallester}@tti-c.org
The Toyota Technological Institute at Chicago, Chicago IL 60637, USA

The 20th British Machine Vision Conference 2009
London, UK

I. Overview

We demonstrate unsupervised learning of a 62 parameter slanted plane stereo vision model involving HOG features as surface orientation cues.

- Unsupervised learning is based on maximizing conditional likelihood.
- We implemented a Hard conditional EM algorithm for training:
  - For the hard E step, we use a form of Max-Product Particle Belief Propagation.
  - For the hard M step, we implement gradient descent using a contrastive divergence approximation
- The performance achieved with unsupervised learning is close to that achieved with supervised learning for this model.

II. Slanted-Plane Stereo Model

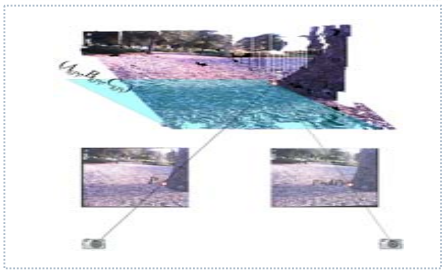
Pixel disparity: $d(p) = A_{i(p)}x_p + B_{i(p)}y_p + C_{i(p)}$

Smoothness cost: $E_S = \sum_{i,j} \min\left(\tau_S, \sum_{(p,q)\in B_{i,j}} \lambda_S |d(p) - d(q)|\right)$

Matching cost: $E_M = \sum_p \sum_k \lambda_k (\Phi_k(p) - \Phi_k(p + d(p)))^2$

Texture cost:

$$E_T = \sum_p \min\left(\tau_T, \begin{array}{l} \lambda_A \left(d(p)(\beta_A \cdot H(p)) - A_{i(p)}\right)^2 \\ + \lambda_B \left(d(p)(\beta_B \cdot H(p)) - B_{i(p)}\right)^2 \end{array}\right)$$



Here:

$A_{i(P)}, B_{i(P)}, C_{i(P)}$	*Disparity plane parameters*
x_p, y_p	*Image coordinate of pixel p*
$d(p)$	*Disparity vaue of p*
$\phi(p)$	*K-dimensional Image feature vector at pixel p*
$H(p)$	*HOG feature vector at pixel p*
$\tau_S, \lambda_S, \lambda_k, \lambda_A, \lambda_B, \beta_A, \beta_B$	*Model parameters*

III. HOG as Surface Orientation Cues

Perpective camera: $x' = (fx)/z \quad y' = (fy)/z$

Assume a coordinate system on a surface patch. Mapping from surface coordinates to 3-D coordinates gives:

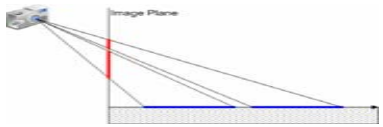
$$x = x_s \quad y = y_s \cos\Psi \quad z = z_0 + y_s \sin\Psi$$
$$\Rightarrow z = z_0 + y \tan\Psi$$

Let: $d = bf/z$

We have: $d = d_0 + By'$

This gives the general relation between the disparity plane parameter and the surface orientation:

$$B = -\frac{d(p)\tan\Psi}{f}$$



Here:

Ψ	*Angle between camera ray and surface normal*
B	*Y coefficient of the disparity plane*
b	*Distance between the foci of 2 cameras*

IV. Parameter Training Using Hard Conditional EM

Hard Conditional EM

$$\beta^* = \operatorname*{argmax}_\beta \sum_{i=1}^N \max_z \ln P(y_i, z | x_i, \beta)$$

Hard E Step

$$z_i := \operatorname*{argmax}_z P(y_i, z | x_i, \beta)$$

Solved by a Stereo Inference Algorithm using Particle BP

Hard M Step

$$\beta := \operatorname*{argmax}_\beta \sum_{i=1}^N \ln P(y_i, z_i | x_i, \beta)$$

$$P(y, z | x, \beta_y, \beta_z) = P(z | x, \beta_z) P(y | x, z, \beta_y)$$

$$P(z | x, \beta_z) = \frac{\exp(-E_z(x, z, \beta_z))}{Z_z(x, \beta_z)}$$

$$E_i(z) = E_z(x_i, z_i, \beta)$$

$$L = \sum_i \ln P(z_i | x_i, \beta_z)$$

$$\beta_z := \operatorname*{argmax}_{\beta_z} \sum_i \ln P(z_i | x_i, \beta_z)$$

$$\beta_y := \operatorname*{argmax}_{\beta_y} \sum_i \ln P(y_i | x_i, z_i, \beta_y)$$

$$\nabla_{\beta_z} L = \sum_{i=1}^N \left(\mathrm{E}_{z \sim P_z(z|x_i,\beta)} [\nabla_{\beta_z} E_i(z)] - \nabla_{\beta_z} E_i(z_i) \right)$$

Solved using Contrastive Divergence

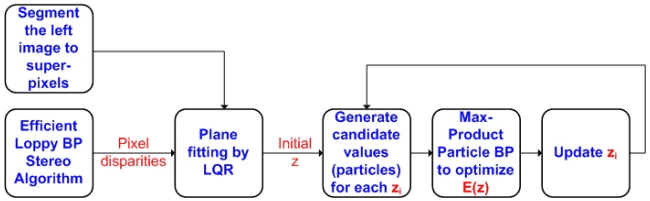
Main References

[1] Jamie Schulte Ashutosh Saxena and Andrew Y. Ng. Depth estimation using monocular and stereo cues. In IJCAI, 2007.
[2] Min Sun Ashutosh Saxena and Andrew Y. Ng. 3-d depth reconstruction from a single still image. IJCV, 2007.
[4] Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. Neural Computation, 14(8):1771–1800, 2002.
[5] Dan Kong and Hai Tao. A method for learning matching errors in stereo computation. In BMVC, 2004.
[6] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proc. 18th International Conf. on Machine Learning (ICML), pages 282–289. Morgan Kaufmann, San Francisco, CA, 2001.
[7] Daniel Scharstein and Chris Pal. Learning conditional random fields for stereo. In CVPR, 2007.
[8] Li Zhang and Steven M. Seitz. Estimating optimal parameters for mrf stereo from a single image pair. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 29(2), 2007. based on "Parameter Estimation for MRF Stereo", CVPR 2005.

V. Stereo Inference Algorithm using Particle BP

Minimize $E(z) = \sum_i (E_T(z_i) + E_M(z_i)) + \sum_{i,j \in N(i)} E_S(z_i, z_j)$

where $z_i = (A_i, B_i, C_i)$



VI. Experiment 1: Middlebury dataset

System parameters:

- 4 image pairs were used as unsupervised training data
- 6 iterations of Hard Conditional EM
- For Particle BP inference: 15 candidate planes (particles) for each segment.
- For gradient descent with contrastive divergence: 8 parameters updates with constant learning rate

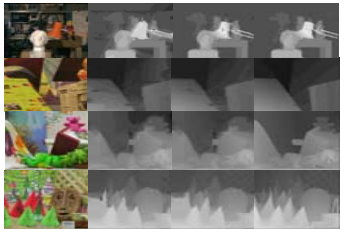


Figure: Improvement with training on the Middlebury dataset

Avg. Rank	Tsukuba nonocc	Tsukuba all	Tsukuba disc	Venus nonocc	Venus all	Venus disc	Teddy nonocc	Teddy all	Teddy disc	Cones nonocc	Cones all	Cones disc	Avg. bad
31.4	3.12	5.22 44	13.9	1.03	1.17 28	11.5	7.08	7.30 3	16.1	6.90	10.7 26	16.0	8.33%

Table: Performance on the Middlebury stereo evaluation. The number shown are for unsupervised learning with texture features.

VII. Experiment 2: Stanford dataset

System parameters:

- 200 rectified stereo image pairs (with ground truth depth) of outdoor scenes (buildings, grass, forests, trees, bushes, etc.) and some indoor scenes: 180 for training + 20 for testing
- We used groud truth depth for supervised training, and unlabeled stereo pairs for unsupervised training
- For Particle BP inference: 15 candidate planes (particles) for each segment.
- For gradient descent with contrastive divergence: 8 parameters updates with constant learning rate

	RMS Disparity Error (pixels)	Average Error $\left\lvert \log_{10} d - \log_{10} \hat{d} \right\rvert$
Saxena et al. [3]		.074
Unsuper., Notexture	1.158	.073
Unsuper., Texture	1.081	.069
Super., Notexture	1.071	.069
Super., Texture	1.001	.063

Table: RMS disparity error and average error (average base 10 logarithm of the multiplicative error) on Stanford dataset. Note that texture information helps imrove the performance in both supervised and unsupervised cases.