Unsupervised Learning of Stereo Vision with Monocular Cues

Hoang Trinh
http://ttic.uchcago.edu/~trinh
David McAllester
http://ttic.uchicago.edu/~dmcallester

The Toyota Technological Institute at Chicago
6045 Kenwood Ave
Chicago IL 60637

We demonstrate unsupervised learning of a stereo vision model involving monocular depth cues (shape from texture cues). We formulate a conditional probability model defining the probability of the right image given the left. This conditional model does not model a probability distribution over images. Maximizing conditional liklihood rather than joint liklihood is similar using a CRF (Conditional Random Field, [6]) rather than an MRF (joint Markov Random Field).

The most closely related earlier work seems to be that of Zhang and Seitz [8] who give a method for adapting five parameters of a stereo vision model. In contrast we train highly parameterized monocular depth cues. Also, we avoid the need for independence assumptions through the use of contrastive divergence training — a general method for optimizing CRFs [4].

There is also related work by Saxena et al. on supervised learning of highly parameterized monocular depth cues [1, 2]. Unlike Saxena et al. we train monocular depth cues as part of unsupervised training of a stereo algorithm. Other related work includes that of Scharstein and Pal [7] and Kong and Tao [5] who perform supervised training of stereo algorithms using general CRF methods.

We focus on histogram of oriented gradient (HOG) features as a (texture) surface orientation cue. As a surface is tilted away from the camera the edges in the direction of the tilt become foreshortened while the edges orthogonal to the tilt are not. The effect on the edge distribution is shown in the image below where the average HOG feature is shown for regions of tree trunk and forest floor. The cylindrical shape of the tree trunk is clearly indicated by the warping of the HOG feature.

We use a slanted plane stereo model in which we first oversegment one of the images and model each image segment by a slanted plane. We set up a hidden CRF on the plane parameters for each segment which includes an energy term relating the plane slant to the HOG feature at each pixel in the segment. This CRF involves a latent depth map. Letting x denote the segmented left image, z denote the assignment of plane parameters to segments, and y denote the right image, we formulate a conditional probability model as follows.

$$P(y|x,\beta) = \sum_z P(y,z|x,\beta) \tag{1}$$

Given training data $(x_1,y_1),\ldots(x_N,y_N)$ we use hard EM to approximately train the parameter vector β with the following objective.

$$\beta^* = \operatorname*{argmax}_\beta \sum_{i=1}^N \ln P(y_i|x_i,\beta) \tag{2}$$

Hard EM, also known as Viterbi training, works with the single most likely (hard) value of z and alternates the following updates.

$$z_i := \operatorname*{argmax}_z P(y_i,z|x_i,\beta) \tag{3}$$

$$\beta := \operatorname*{argmax}_\beta \sum_{i=1}^N \ln P(y_i,z_i,|x_i,\beta) \tag{4}$$

We will call (3) the hard E step and (4) the hard M step. Our implementation of the hard M step relies on a factorization of the probability model into two conditional probability models each of which is defined by an energy functional. Unlike CRFs, we do not require the energy functional to be linear in the model parameters.

$$P(y,z|x,(\beta_y,\beta_z)) = P(z|x,\beta_z)P(y|x,z,\beta_y) \tag{5}$$

$$P(z|x,\beta_z) = \frac{\exp(-E_z(x,z,\beta_z))}{Z_z(x,\beta_z)} \tag{6}$$

$$P(y|x,z,\beta_y) = \frac{\exp(-E_y(x,y,z,\beta_y))}{Z_y(x,z,\beta_y)} \tag{7}$$

Given this factorization of the model, the hard M step (4) can be written as the following pair of updates.

$$\beta_z := \operatorname*{argmax}_{\beta_z} \sum_i \ln P(z_i|x_i,\beta_z) \tag{8}$$

$$\beta_y := \operatorname*{argmax}_{\beta_y} \sum_i \ln P(y_i|x_i,z_i,\beta_y) \tag{9}$$

For the experiments reported here we use contrastive divergence [4] to optimize (8) and least squares regression to optimize (9).

The table shows results on the Stanford color stereo dataset [1] which has been used to train monocular depth estimation [2]. Because of problems with image pair rectification we removed from the dataset all pairs for which the energy value achieved by loopy BP was above a specified threshold. This left 200 out of an original 250 stereo pairs. Each stereo pair in this dataset is associated with ground truth depth information from a laser range finder. We randomly divide the 200 properly rectified stereo pairs into 180 training pairs and 20 test pairs.

	RMS Disparity Error (pixels)	Average Error $\lvert\log_{10}Z-\log_{10}\hat{Z}\rvert$
Saxena et al. [1]		.074
Unsuper., Notexture	1.158	.073
Unsuper., Texture	1.081	.069
Super., Notexture	1.071	.069
Super., Texture	1.001	.063

[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.

[1] http://ai.stanford.edu/ asaxena/learningdepth/data