# Inferring Depth from Single Images in Natural Scenes

**Byron Boots**
Department of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
beb@cs.cmu.edu

## Abstract

The inverse optics problem is one of the oldest and most well known problems in visual perception. Inferring the underlying sources of visual images has no analytic solution. Recent work on brightness, color, and form has suggested that visual percepts represent the probable sources of visual stimuli and not the stimuli as such, suggesting an empirical theory of visual perception. Here I explore this idea by framing the perception of depth as a machine learning problem. I apply two algorithms with varying levels of model complexity, compare their ability to infer depth, both with each other and with the best previous solutions.

## 1 Introduction

It has long been recognized that sources of visual stimuli cannot be uniquely specified by the energy that reaches sensory receptors; the same pattern of light projected onto the retina may arise from different combinations of illumination, reflectance and transmittance, and from objects of different sizes, at different distances and in different orientations (Figure 1). Nevertheless, visual agents must respond to real-world events. The inevitably uncertain sources of visual stimuli thus present a quandary: although the physical properties of a stimulus cannot uniquely specify its provenance, success depends on behavioral responses that are appropriate to the stimulus source. This dilemma is referred to as the inverse optics problem.

For more than a century now, investigators have surmised that the basis of successful biological vision in the face of the inverse optics problem is the inclusion of prior experience in visual processing, presumably derived from both evolution and individual development. This empirical influence on visual perception, first suggested by George Berkeley in 1709[1]. has been variously considered in terms of Helmholtz's "unconscious inferences,"[2] the "organizational principles" advocated by gestalt psychology[3], and the framework of "ecological optics" developed by Gibson[4]. More recently, these broad interpretations have been bolstered by a wealth of evidence suggesting that many visual percepts can be predicted according to the real-world sources to which an animal has always been exposed[5,6,7]. In fact, many of the anomalous percepts that humans see in response to simple visual stimuli may be rationalized in this way[6,7].

In the present work, I have explored the notion of an empirical approach to visual perception by framing the inverse optics problem as a machine learning problem. Specifically, I have asked how depth to surfaces in natural scenes may be inferred from monocular images. I look at previous approaches to the problem and suggest novel alternatives. My results demonstrate the feasibility of solving the inverse problem from two perspectives: a naive linear regression perspective and from a more complex graphical modeling perspective.

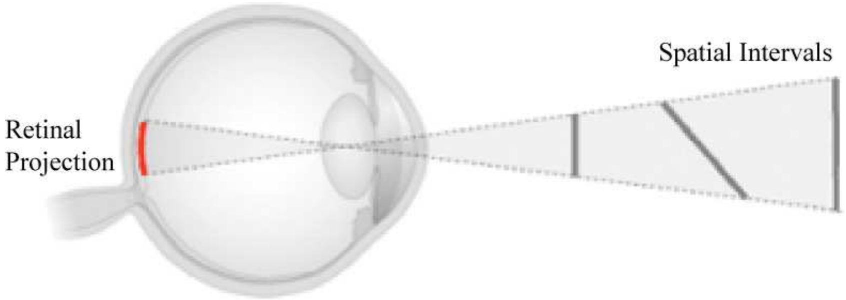

Figure 1: The inverse optics problem with respect to geometry. Objects of different sizes at different distances and in different orientations may project the same image on the retia.

## 2 Related Work

### 2.1 Traditional approaches to computer vision

Geometrical aspects of the inverse optics problem are frequently encountered in computer vision in the form of recovering three-dimensional structure from two-dimensional images. Most work in this area has focused on stereopsis [8], structure from motion [9], or depth from defocus [10]; all of which rely on a differential comparison between multiple images. Animals, however, are able to judge spatial geometry from a monocular image, and it is thought that this ability lies at the heart of the perception of geometrical space[11]. Despite this fact, the only well-known method of inferring depth from a single image is the ”shape from shading” algorithm[12], a technique that devises models of image formation based on the physics of light interaction and then inverts the models to solve for depth. These inverted models are highly underconstrained, requiring many simplifying assumptions (e.g., Lambertian surface reflectance) that seldom hold in images of natural scenes[14]. Recently researchers have begun to recover geometrical structure from two dimensional images empirically using learning-based techniques.

### 2.2 Learning-based methods in computer vision

Despite the large quantity of evidence suggesting the importance of empirical data in vision, there have been surprisingly few attempts to leverage machine learning techniques to infer scene geometry from monocular images. I am only aware of two different methods. Andrew Ng’s group at Stanford University is using discriminatively trained Markov random fields to infer depth from monocular images collected from a mobile platform [13]. This approach is quite successful and has the advantage of directly learning depth-maps based on statistics of images and their underlying sources. Potentially, such an approach could be tied to vision studies which have similarly used images and depth maps to explain perceptual phenomena [5, 14].

Aloysha Efros’ group at Carnegie Mellon University is using a completely different technique where subjects hand-label the possible orientation of surfaces in images[15]. Their algorithm learns geometric classes defined by simple orientations such as sky, ground, and vertical surfaces in a scene. The labels are then used to ”cut and fold” the image providing a simple ”pop-up” model of a visual scene. This method performs surprisingly well for a wide range of images and is visually appealing, but is highly inaccurate and not directly related to the the true statistics of underlying depths in visual images.

### 2.3 Filtering in visual inference

In previous attempts at monocular inference it has been suggested that a variety cues are essential for judging depth. In particular, convolutional filters such as Laws' masks for texture energy and oriented edge detectors have been used to develop complex feature vectors describing local information in image patches [13,15]. Additionally, the local patches themselves are augmented with information from multi-scale decompositions of the image in order to provide additional scene context[13]. This latter point is extremely important as local information is insufficient to determine depth [6,7].

Despite their extensive use, linear filters are problematic. Features derived in this way introduce a priori assumptions about the importance of particular patterns and spatial frequencies on the image plane for inferring depth. Studies of projected image statistics and their underlying range statistics show that each have different properties [14, 17, 18]. For example much of the high frequency variation found in images and picked up by linear filters consist of shadow and surface markings that are not observed in the underlying range images [14].

## 3 Methods

Surprisingly, no one has attempted to infer depth from monocular images without first extracting features through convolutional filters. In the present work I develop a simple yet principled strategy for inferring depth that does not rely on a priori assumptions concerning filters or multi-scale decompositions. Given the goal of inferring the depth to surfaces underlying a single monocular image (Figure 2), use *every* pixel in the image to infer depth at a single point. This has the advantage of using all of the available global contextual information in the image to determine depth.

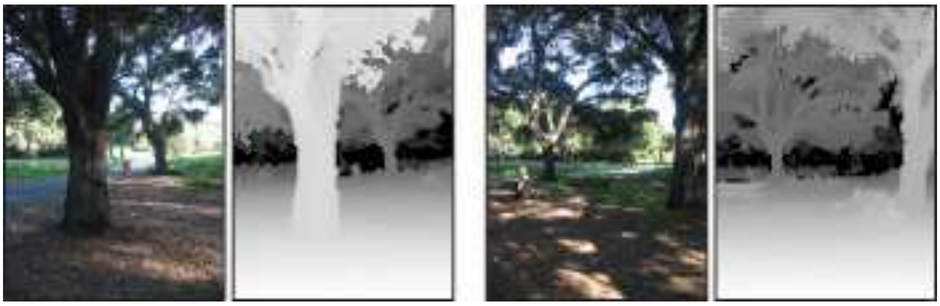

Figure 2: Examples of images and correlated depth maps. Lighter color indicates closer surfaces.

### 3.1 Linear regression

I will begin by using linear regression to infer depth from a single image. More formally, linear regression is used to derive weights $w_i$ for each depth $i$ over $N$ image-depthmap pairs:

$$w_i = arg\min_{w_i} \frac{1}{2} \sum_{j=1}^{N} (w_i^T x_j - ln(d_{ij}))^2$$

where $x_j$, is the array of features in image $j$ and $d_{ij}$ is the distance to the surface underlying position $i$ in the depth map corresponding to image $j$. The depths are transformed to a log scale to emphasize multiplicative rather than additive errors in training [13,14]. Each feature in the feature vector $x_j$ consists of an unprocessed pixel value. The linear regression problem is easily solved via the normal equations.

Due to the small number of training examples used in the present experiments, I additionally applied ridge regression with a range of different Tikhonov factors. This did not provide any significant improvement over linear regression. Thus, the results described below are given for simple linear regression.

### 3.2 A discriminatively trained Markov random field

In natural environments, the depth at a particular location is highly correlated with other local depths[18]. This fact suggests that inference may be improved by accounting for the relationship between a predicted depth value and the predicted values at neighboring points. In previous work, discriminatively trained Markov random fields were used to model this relation. Here I model the probability of depth $d$ as a jointly Gaussian Markov random field, following (with some modification) Saxena et al. [13]:

$$P(d|X;\theta,\sigma) = \frac{1}{Z} \exp\{- \sum_{i=1}^{M} \frac{(d_i - x_i^T \theta_i)^2}{2\sigma_1^2} - \sum_{i=1}^{M} \sum_{j \in N_s(i)} \frac{(d_i - d_j)^2}{2\sigma_{ij}^2}\}$$

where $M$ is the total number of depths in a depth map, $x_i$ is the absolute depth feature vector consisting of unprocessed pixel values, $N_s(i)$ are the four neighbors (up, down, left and right) of the depth $d_i$, $Z$ is the normalization constant and $\theta$ and $\sigma$ are parameters of the model. The first term in the exponent models the depth at a single point as a function of the entire image. The second term places a data-dependent constraint on the depths modeling the interaction between neighboring sites. This second term is thus similar to the "interaction potential" in a discriminative random field [19].

### 3.3 Parameter estimation

To find the parameters $\theta_i$ one can maximize the conditional likelihood $p(d|X;\theta_i)$ by solving a linear least squares problem. The variance parameter $\sigma_1^2$ is a fixed constant making the first term roughly equivalent to linear regression. If the "variance" $\sigma_{ij}^2$ in the second term was also modeled as a constant, then the second term would simply smooth the depth estimate. In practice dependencies are not the same everywhere, thus the "variance" $\sigma_{ij}^2$ is modeled as an estimate of patch $i$ and $j$'s expected difference [13,19]. In detail, $\sigma_{ij}^2 = u_r^T y_{ij}$ where $y_{ij}$ are the features(image pixes) underlying the positions of $d_i$ and $d_j$ and the parameter $u_r$ is learned via linear regression on the test set to fit $\sigma_{ij}^2$ to the expected value of $(d_i - d_j)^2$ for each row $r$, with the constraint that $u_r \geq 0$ (to keep the estimated $\sigma_{ij}^2$ positive)[13].

### 3.4 Inference

Given a new test image, the aim is to find the optimal estimate of depths to surfaces underlying the image. After learning the parameters, this was accomplished by finding the maximum a posteriori estimate by maximizing $P(d|X;\theta,\sigma)$ with respect to $d$. To accomplish this, one can take the log of the Gaussian:

$$\ln P(d|X;\theta,\sigma) = -\sum_{i=1}^{M} \frac{(d_i - x_i^T\theta_i)^2}{2\sigma_1^2} - \sum_{i=1}^{M} \sum_{j\in N_s(i)} \frac{(d_i - d_j)^2}{2\sigma_{ij}^2}$$

which is quadratic in $d$. The maximum may be found by taking the derivative, setting it to zero, and solving a system of linear equations where we have an equation for each $d_i$:

$$0 = \frac{\partial}{\partial d_i} = \frac{1}{\sigma_1^2}(d_i - x_i^T\theta_i) - \sum_{j\in N_s(i)} \frac{1}{\sigma_{ij}^2}(d_i - d_j)$$

Thus, the maximum may be found in closed form. This step takes less than 1 second per image to compute in Matlab.

## 4 Experiments

### 4.1 Data

To investigate the feasibility of depth inference from monocular images, I employed a database of 350 images of unstructured outdoor environments and their corresponding depth maps collected in and around the Stanford University campus. The images were acquired with a 3-dimensional laser range finder with a maximum range of 81 meters and resolution of 86x107 pixels and a camera with image resolution of 1704x2272 pixels. In the experiments reported here, 80% of the images were used for training and the remaining 20 percent for hold-out testing. Due to motor noise, reflections, limited laser range, and lack of accounting for the camera properties such as visual angle, alignment, and calibration, the images and laser scans are often misaligned and sometimes exhibit rather significant errors.

Each of the images in the database was convolved with a Gaussian kernel and sub-sampled to form smaller 173x130 pixel images. The depthmaps were also sub-sampled without convolution, reducing the size of the depthmaps to 54x43 pixels. Convolution was not necessary for the depthmaps because range images typically lack high frequency variation found in images [18]. Once reduced, each image was converted into a feature vector consisting of every pixel in each of 3 (RGB) color channels and a bias term. Thus, the length of each feature vector was 173*130*3+1 or 67471 pixels. A target vector containing the range values corresponding to a given image was built out of the 54x43 or 2322 pixels in a depthmap.

After segregating the data into training and test sets, linear regression and Markov random field inference were performed on the data.

### 4.2 Results

The two algorithms were trained and tested on a very small set of images taken in particularly challenging unstructured outdoor environments that contained many nebulous structures and high frequency variation caused by shadows and jagged edges. This being said, each performed well enough to be competitive with the best previous methods.

#### 4.2.1 Linear regression

Recall that each pixel in the target vector was independently estimated based on every pixel in the input image. This strategy foregoes the idea of preprocessing the input image

Table 1: Results: Average absolute error of the different approaches to inference (on a log base 10 scale). The first two algorithms were implemented in the present work, and the second two are from previous studies with the same dataset (although it should be noted that the dataset used in this work is incomplete; it consists of $25\%$ fewer images than the dataset used in Saxena et al. [13]) .

| ALGORITHM | ERROR |
|---|---|
| Linear Regression | 0.156 |
| Markov Random Field | 0.148 |
| Saxena et al. Baseline MRF | 0.343 |
| Saxena et al. Laplacian MRF | 0.142 |

with linear filters and prevents arbitrary assumptions about the joint statistical structure of color images and depthmaps. The results (Figure 3 and Table 1) indicate that, despite the simplicity of this proposed paradigm, highly accurate estimations of depth are possible. In fact, linear regression performed nearly as well as the much more complex Laplacian Markov random field (the best performing previous method) which uses a combination of linear convolutional filters, summary statistics, multi-scale and column features on the same data-set (Table 1)[13]. One obvious problem with this technique is the independent estimations of depth. While each estimate is fairly accurate, error accumulates when measuring pairwise relative depths creating high frequency variation not typically found in depthmaps (Figure 3). The discriminatively trained Markov random field was employed to help overcome this problem by adding constraints on the depth of neighboring patches.

### 4.2.2 The Markov random field

Parameters were estimated with the same set of training data as linear regression and inference was performed with the same test set. The results indicate a small improvement over linear regression (Figure 3 and Table 1). In particular, the data driven constrains smoothed the images in many areas overcoming the limitations of independent inference at each point. The gains over linear regression, however, were not substantial. One possible explanation is that linear regression captures the majority of the information in the image, given that the entire image is used to determine depth at each point.

## 5 Conclusion

The results described here demonstrate that the naive approach to the inference of depth in single monocular images, that is, training linear regressors over every pixel in the image at each estimated depth position, performs nearly as well as the most complex state-of-the-art algorithms. This is interesting because linear regression is fast and easy to implement, making it an attractive choice for practical applications. The results also demonstrate that while Markov random fields can offer some improvement over linear regression in this domain, the improvement is not dramatic. In future work, the role of Markov random fields for depth estimation could be explored more deeply by expanding the size of each depth value's neighborhood thereby further constraining the values at a particular point. Finally, these experiments make clear that a variety of machine learning techniques, each of varying complexity, may be employed to overcome the famous inverse optics problem in a principled and disciplined way.

## References

[1] Berkeley, G. (1709/1975) *Philosophical Works Including Works on Vision.* (Ayers MR, Ed.) London: Everyman/j.M. Dent.

[2] Helmholtz, H. von (1924-25) *Helmholtz's Treatise on Physiological Optics.* Rochester, New York: The Optical Society of America, .

[3] Wertheimer, M. (1938) Laws of organization in perceptual forms. In: *A sourcebook of Gestalt psychology* (Ellis WD, Transl. and Ed.), pp. 71-88. New York: Humanities Press.

[4] Gibson, J.J. (1979) *The Ecological Approach to Visual Perception.* Hillsdale, NJ: Lawrence Erlbaum.

[5] Yang, Z. & Purves, D. [2003] A statistical explanation of visual space. *Nature Neuroscience*, **6**, 632-640.

[6] Purves, D. & Lotto, B. (2003) *Why We See What We Do: An Empirical Theory of Vision*. Sunderland, MA: Sinaur Associates.

[7] Howe, C. & Purves, D. (2005) *Perceiving Geometry: Geometrical Illusions Explained by Natural Scene Statistics*. New York: Springer.

[8] Scharstein, D. & Sezeliski, R. (2002) A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, **47**:7-42.

[9] Tomasi, C. & Kanade, T. (1993) Shape and motion from image streams under orthography. *Proc. Natl. Acad. Sci.* USA, **90**(21):9795-9802.

[10] Das, S. & Ahuja, N. (1995) Performance analysis of stereo, vergence, and focus as depth cues for active vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17:1213-1219.

[11] Loomis, J., Da Silva, J., Philbeck, J., & Fukusima, S. (1996) Visual perception of location and distance. *J. of Psychological Science*, **5**(3):72-77.

[12] Horn, B.K.B. (1975) Obtaining shape from shading information. *The Psychology of Computer Vision*. New York: McGraw Hill, pp. 115-155

[13] Saxena, A., Chung, H., & Ng. A. (2006) Learning depth from single monocular images. In *NIPS 18*

[14] Potetz, B., & Lee, T.S. (2003) Statistical Correlations between 2D Images and 3D Structures in Natural Scenes. *Journal of Optical Society of America*, A. **20**(7):1292-1303.

[15] Hoiem, D., Efros, a.a., & Hebert, M. (2005) Automatic Photo pop-up, in *ACM SIGGRAPH*.

[16] Forsyth, A., & Ponce, J. (2003) *Computer Vision: A Modern Approach.* NJ: Prentice Hall.

[17] Huang, J., & Mumford D. (1999) Statistics of natural images and models, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp.541-547

[18] Huang, J., Lee, A.B., & Mumford D. (2000) Statistics of range images, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp.324-331

[19] Kuma, S., & Hebert, M. (2003) Discriminative random fields: A discriminative framework for contextual interaction in classification. *Proc. IEEE ICCV* (ICCV '03), Vol. 2, 2003, pp. 1150-1157

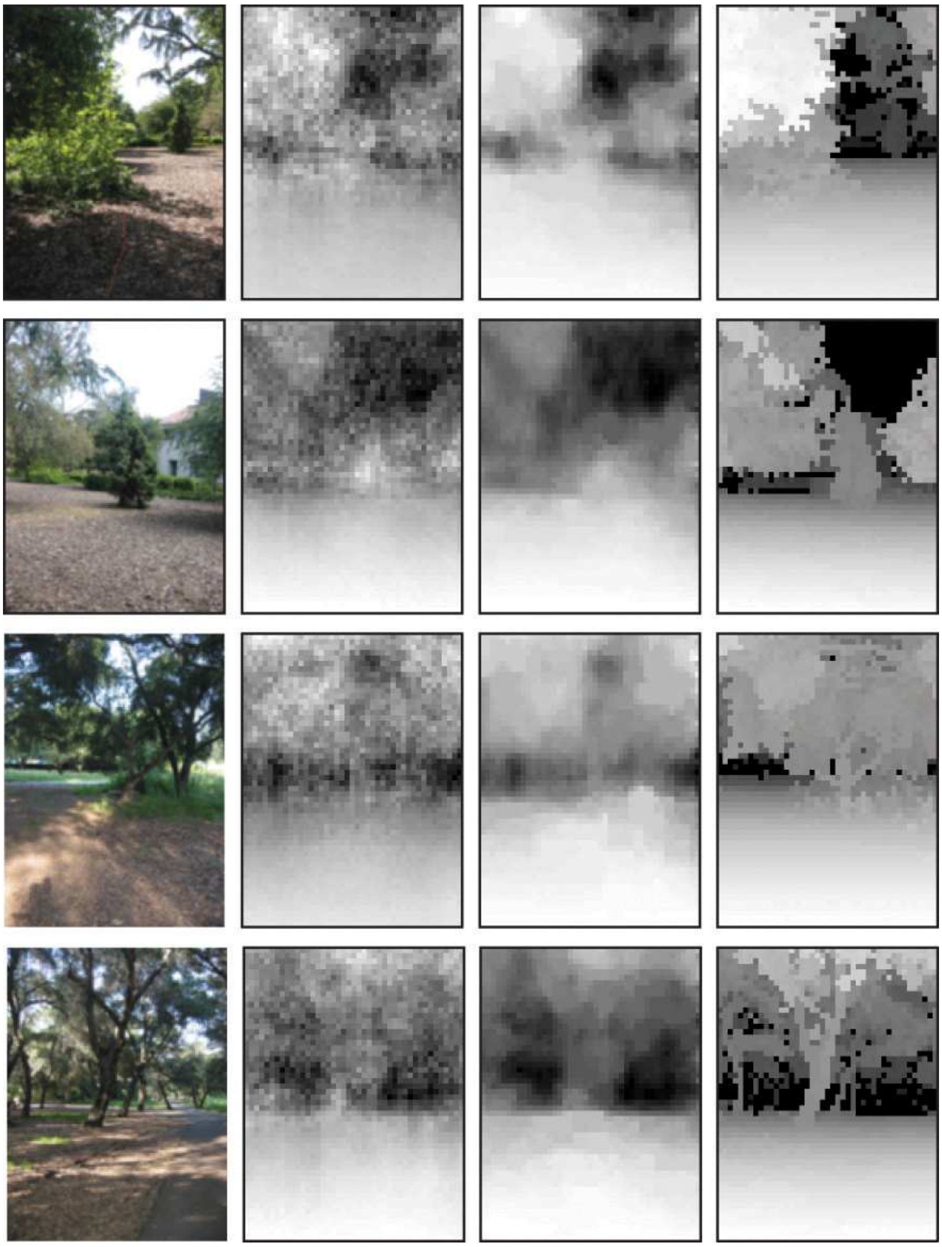

Figure 3: Linear regression and Markov random field estimates of depth given monocular images from the test set. The first column is the actual image, the second column is the linear regression estimate, the third column is the Markov random field estimate, and the fourth column is the true depth. Lighter is closer while darker is farther away. Note how the second term in the Markov random field smooths local depth estimates.