Contents

List of Figures

Chapter 1

Mechanisms for Propagating Surface Information in Three-Dimensional Reconstruction

Abstract Bayesian and other related statistical techniques have emerged as a dominant paradigm in computer vision for estimating 3-D surfaces in the presence of noisy and sparse depth cues. In particular, for 3-D reconstruction problems these techniques have been implemented using Markov random fields (MRFs), which are probabilistic models that express how variables arranged in a spatial structure jointly vary, such as a rectangular grid of disparity variables aligned to a pixel lattice in a stereo model. Such MRF models incorporate powerful priors on surface geometry, which allow 3-D estimates to combine evidence from depth cues with prior knowledge of smoothness constraints. They embody a natural mechanism for propagating surface information from regions with highly informative depth cues to neighboring regions with unreliable or missing depth cues, without crossing over depth discontinuities. Recent advances in inference algorithms have enlarged the range of statistical models that are tractable for computer implementation, enabling the use of increasingly realistic and expressive models and leading to some of the most robust and accurate 3-D estimation algorithms to date. Finally, a "belief propagation" framework allows these models to be implemented on a massively parallel computer architecture, raising the possibility that they may be realized in a biologically plausible form.

1.1 Introduction

Three-dimensional reconstruction is a major theme in computer vision, with techniques for estimating shape from a variety of cues, including shading [1], texture [2] and multi-view stereo [3]. A fundamental challenge of 3-D reconstruction is estimating depth information everywhere in a scene despite the fact that depth cues are noisy and sparse. These properties of depth cues mean that depth information must be propagated from regions of greater certainty to regions of lesser certainty.

The 3-D reconstruction problem that is perhaps the most mature in computer vision is stereo using two or more calibrated cameras with known epipolar geometry. The most successful stereo algorithms use Markov random fields (MRFs) [4], which are probabilistic models that express the joint distribution of variables arranged in a network structure, with the state of any variable exerting a direct influence on the states of its neighbors. In the case of stereo (in this paper the discussion is restricted to two-view stereo), a basic MRF model consists of a lattice (grid) of disparity variables, one for each pixel in the image. For any disparity hypothesized at a given pixel there is (usually ambiguous) evidence from the corresponding match between the left and right images that this disparity implies. Nearest-neighbor connections enforce a prior smoothness constraint on disparities, which equates to a related smoothness constraint on depths in the scene. Inference is performed using a global optimization method such as graph cuts [5] or belief propagation [6], which determines a near-optimal assignment of disparities to each pixel given the image data. The MRF framework embodies a natural mechanism for propagating surface information in the presence of noisy and sparse data.

This paper is intended to present the basic principles of this framework (which generalize to any 3-D reconstruction problem, not just two-frame stereo) to an audience of vision researchers who are not computer vision specialists. I discuss recent extensions incorporating more realistic modeling of surface geometry, and point to recent work suggesting the possibility that belief propagation (or something close to it) may be realized in a biologically plausible form. Finally, possible directions for future research are explored.

1.2 Markov Random Fields for Stereo

In this section I outline a simple MRF (Markov random field) formulation of stereo, described using a Bayesian model. (Other statistical variants such as conditional random fields [7, 8], mentioned below, are popular alternatives that are very similar.) Two grayscale images L and R (left and right) are taken of the scene, which are assumed rectified so that a pixel in one image is guaranteed to match a pixel in the same row in the other image. The unknown disparity field is represented by D, with D_r representing the disparity at pixel location r. A particular disparity value D_r, where $r = (x, y)$ specifies the pixel coordinates, has the following interpretation: $(x + D_r, y)$ in the left image corresponds to (x, y) in the right image.

The prior on the disparity field D enforces smoothness as follows:

$$P(D) = \frac{1}{Z} e^{-\beta V(D)} \tag{1.1}$$

where Z is a normalizing constant ensuring that $P(D)$ sums to 1 over all

possible values of D, β is a positive constant that controls the peakedness of the probability distribution (which in turn determines the importance of the prior relative to the likelihood, discussed below), and

$$V(D) = \sum_{<rs>} f(D_r, D_s) \tag{1.2}$$

where the sum is over all neighboring pairs of pixels r and s. Here $f(D_r, D_s)$ is an energy function that penalizes differences between disparities in neighboring pixels, with higher energy values corresponding to more severe penalties. (In other words, non-zero values of the first derivative of the disparity field in the x and y directions are penalized.) One possible choice for the function is $f(D_r, D_s) = |D_r - D_s|$. A popular variation [9] is $f(D_r, D_s) = \min(|D_r - D_s|, \tau)$, which ensures that the penalty can be no larger than τ; this is appropriate in scenes with depth discontinuities, where a large difference between disparities on either side of a depth edge may be no less probable than a moderate difference.

Note that a prior of this form enforces a bias towards fronto-parallel surfaces, over which the disparity is constant; I will discuss ways of relaxing this bias later.

Next I define a likelihood function, which defines how the left and right images provide evidence supporting particular disparity values:

$$P(m|D) = \prod_r P(m_r(D_r)|D_r) \tag{1.3}$$

where the product is over all pixels in the image, and m is the matching error across the entire image. Specifically, $m_r(D_r)$ is the matching error between the left and right images assuming disparity D_r, defined as $m_r(D_r) = |L(x + D_r, y) - R(x, y)|$ (again $r = (x, y)$). (The product form assumes that the matching errors are conditionally independent given the disparity field.) A simple model for the matching error is given by:

$$P(m_r(D_r)|D_r) = \frac{1}{Z'} e^{-\mu m_r(D_r)} \tag{1.4}$$

which assigns a higher penalty (lower probability) to higher matching errors.

The Bayesian formulation defines a posterior distribution of disparities given both images, given by Bayes theorem:

$$P(D|m) = P(D)P(m|D)/P(m) \tag{1.5}$$

To perform inference with the model, one typically finds the MAP (maximum a posterior) estimate of the disparity field, i.e. the value of D that maximizes the posterior. Note that, since $P(m)$ is independent of D, one can write the MAP estimate of D, denoted D^*, as:

$$D^* = \arg\max_D P(D)P(m|D) \tag{1.6}$$

Since maximizing any function is equivalent to maximizing the log of the function, the MAP estimate can be re-expressed as:

$$D^* = \arg\max_D \{-\beta \sum_{<rs>} f(D_r, D_s) - \mu \sum_r m_r(D_r)\} \tag{1.7}$$

where constants independent of D have been removed. This is equivalent to:

$$D^* = \arg\min_D \{\sum_{<rs>} f(D_r, D_s) + \gamma \sum_r m_r(D_r)\} \tag{1.8}$$

where $\gamma = \mu/\beta$ expresses the relative weight of the prior and likelihood energies.

Methods for estimating the MAP are discussed in the next section.

1.2.1 Model improvements and refinements

This MRF is a particularly simple model of stereo, and many improvements and refinements are commonly added, such as the following.

(1.) Considerable performance improvements have been attained by exploiting the tendency for disparity discontinuities to be accompanied by intensity edges [10, 8]. To accomplish this, the disparity smoothness function is modulated by a measure of the image gradient, so that large disparity differences between neighboring pixels are penalized less severely when there is a strong intensity difference between the pixels. (Alternatively, a general-purpose monocular segmentation algorithm may be run to determine the likely locations of edges between regions of different intensity or color, instead of relying on a purely local measure of the image gradient.) Thus, surface information is naturally propagated from regions with highly informative depth cues to neighboring regions with unreliable or missing depth cues, without crossing over depth discontinuities.

(2.) The matching function used in the likelihood model can be based on comparisons of image properties that are richer than grayscale intensity – for example, color, intensity gradient (magnitude and direction) or higher-level descriptors incorporating neighboring image structure (such as DAISY [11]).

(3) MRF models may be multi-scale [12], with a pyramid structure coupling the original disparity lattice with sub-sampled (coarsened) versions of it.

(4.) The conditional independence assumption in Eq. 1.3 can be relaxed with the use of conditional random fields [7, 8], resulting in a more realistic posterior distribution.

1.2.2 How MRFs propagate information

This section briefly explains how the MRF model propagates noisy and sparse disparity cues throughout the image. The general principle behind

this process is that the prior and likelihood distributions compete for various disparity hypotheses, and the relative strength of these two sources of information automatically varies across the image according to which source is more reliable.

Fig. 1.1 shows a simple example of a scene consisting only of a flat surface oriented fronto-parallel to the cameras. In this figure, only the right image of a stereo image pair is shown, and the fronto-parallel orientation of the scene implies that the correct disparity field is uniform across the image, with value d_0. The entire surface contains a highly textured, non-periodic pattern, except for a central region which is textureless. Everywhere in the textured part of the image, the likelihood model provides strong evidence for the correct disparity d_0 and very low evidence for anything but d_0 (in practice the likelihood model will be less discriminating, but we assume near-perfect disparity discrimination for the sake of argument). By contrast, in the central region there will be no direct evidence favoring one disparity over another.

Any attempt to perform inference using the likelihood without the prior will obviously fail in the central region. However, by incorporating the prior with the likelihood, it is easy to see that a disparity field that has a uniform value of d_0 over the entire image will have a higher posterior probability than any other possible disparity field, because any disparity field that deviates from d_0 will be less smooth than the correct disparity field (while having the same support from the likelihood model).

More realistic examples can be analyzed exhibiting similar behavior, in which the MRF prior model propagates surface information even when the disparity cues are noisy or non-existent.

1.3 Tractable Inference with MRFs

A significant challenge posed by the stereo MRF model (and many other Bayesian models) is the difficulty in estimating the MAP, which is equivalent to minimizing the energy function in Eq. 1.8. There are no techniques that are guaranteed to find the exact MAP in general cases, aside from exhaustive search, which is impractical (e.g. S^N possible disparity fields must be evaluated, where typical values are $S = 50$ for the number of possible disparities and $N = 10000$ for the number of pixels in the image). Typical energy minimization techniques such as gradient descent [13] work well only when the energy function has one dominant, global minimum, or when the gradient descent procedure can be initialized sufficiently close to a suitable local minimum. However, the type of energy function defined by Eq. 1.8 has a much more irregularly shaped energy landscape, with many local minima, and gradient descent-type techniques will typically converge to a local minimum that

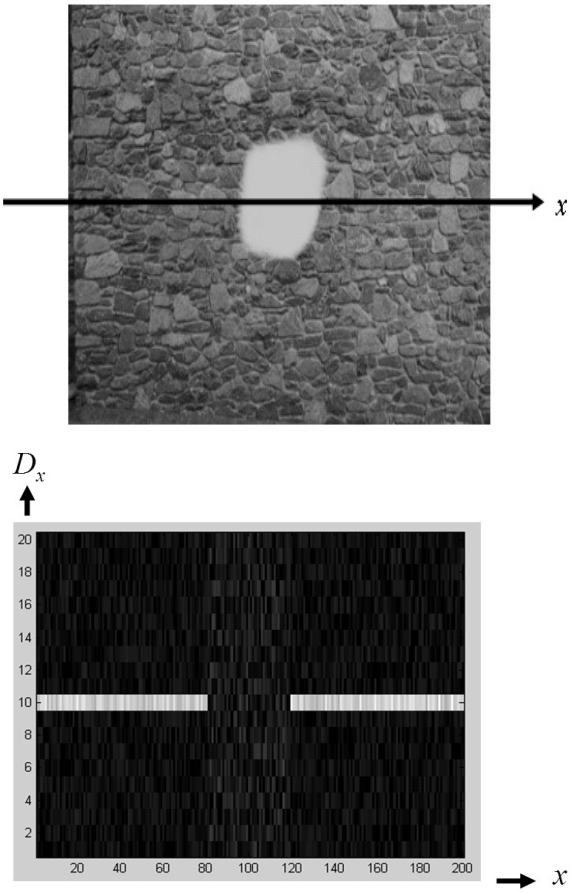

FIGURE 1.1: Idealized example of how disparity information is propagated across an image. Top: image of a fronto-parallel surface (stone wall), with x-axis slice superimposed on image. Bottom: support for different disparities D_x for all possible values of x, with brightness proportional to degree of support. In textured parts of slice, the disparity value $d_0 = 10$ is strongly supported; in the textureless part of slice towards the center, all disparity values have approximately equal support. See text for explanation of how MRF model propagates $d_0 = 10$ solution across the untextured region.

is far from the global minimum.

A variety of approximation techniques are available for estimating the MAP. Until fairly recently, one of the best available techniques was simulated annealing [13], which is essentially a form of a gradient descent perturbed by stochastic noise. While simulated annealing can succeed in locating a "good" local minimum (if not the global minimum), it is an extremely slow technique. In fact, its slowness meant that most MRF models in computer vision were extremely difficult to use, and as a result progress in MRF model development was hindered.

However, approximately ten years ago, two efficient energy minimization techniques were introduced to computer vision (imported from other fields), graph cuts [5] and belief propagation [6]. Both energy minimization techniques find approximate solutions in general, only attaining exact solutions in certain special cases. While both techniques are popular in a variety of computer vision MRF models, I will only discuss belief propagation (BP), which is more intuitive than graph cuts (and perhaps more related to biologically plausible mechanisms), and will provide additional insight into how MRFs propagate information. (The techniques perform minimization of energy functions with *discrete* variables, which means that the disparity values must be quantized into a finite set of values, e.g. integer pixel values or sub-pixel values within a finite range.)

BP is a fast iterative procedure for estimating the minimum energy (maximum probability) joint configuration of all the variables in an MRF, i.e. the joint estimate of the most likely states of all the variables. (It is also used to solve the related problem of estimating marginal probabilities of individual variables, which specify the probabilities of all possible states for each variable.) The main idea behind BP is that neighboring variables "talk" to each other at each iteration, with each variable passing messages to its neighbors with their estimates of the neighbors' likely states. After enough iterations, this series of "conversations" is likely to converge to a consensus specifying which state is most likely for each variable.

The messages exchanged between variables in BP pass information throughout the MRF. Information can only be passed from variables to their immediate neighbors in one iteration of BP, but after enough iterations it is possible for information to be passed between all variables. At the start of BP messages express no preference for one state over another, but after enough iterations the messages become more "opinionated," expressing strong preferences for certain states.

Finally, it is important to note that stereo inference using BP (or graph cuts) is still slow relative to simpler (but less accurate) stereo algorithms, requiring on the order of seconds (or even minutes) for each pair of images. However, ongoing work on boosting the efficiency of these algorithms, as well as steadily improving computer hardware, is continually increasing execution speed (see, for instance, work on a real-time stereo BP implementation using a graphics processing unit [14]).

1.4 Learning MRF parameters

The MRF model in Sec. 1.2 has a number of free parameters (such as β, τ and μ) that must be set correctly for the model to be realistic and accurate enough to make good inferences. There are well-established procedures [8] for learning MRF parameters from "labeled" data samples, in this case left/right image pairs and true ("ground truth") disparities. However, until recently few datasets included ground truth disparity fields, which made it difficult to learn the MRF parameters. Fortunately this obstacle is being removed now that there are an increasing number of datasets that include ground truth, which is determined using tools such as laser range finders (used to measure the precise depth, and hence disparity, of nearly every pixel in a scene). In addition, more advanced techniques have been developed [9] which allow MRF parameters to be estimated directly from left/right image pairs, without the need for ground truth disparity data.

1.5 More Realistic Priors

An important limitation of the MRF prior in Eq. 1.1 is that it penalizes disparity differences in neighboring pixels, which implies a bias in favor of fronto-parallel surfaces. Such a bias is inappropriate for many real-world scenes with slanted surfaces. Even the toy example considered in Sec. 1.2.2 is likely to fail if the surface is slanted: the prior may have trouble propagating the linearly changing disparity beyond the textured region of the image. In such cases, while the first x and y derivatives of disparity may be non-zero, the second derivatives are zero. (Any planar surface has an associated disparity field $D_r = ax + by + c$, where $r = (x, y)$, i.e. the disparity is linear in the x and y image coordinates.)

Ongoing research in my laboratory seeks to overcome this fronto-parallel bias in the context of a specific application: terrain analysis for visually impaired wheelchair users. In this application [15], a stereo camera is pointed at the ground, such that the optical axis makes an angle of approximately 45° with the ground surface. The goal is to detect terrain irregularities such as obstacles, holes in the ground and curbs, and to convey this information to the wheelchair user.

My colleagues and I have designed a real-time algorithm for detecting and reporting terrain irregularities using a fast, commercially available stereo algorithm that is integrated with the stereo camera hardware. The stereo algorithm is based on simple window correlation rather than an MRF model and is therefore very fast, processing many frames per second. The disadvan-

tage of using such a fast algorithm is that it produces sparse, noisy disparity estimates, and smooths over depth discontinuities. However, the quality of the disparity estimates suffices for detecting large terrain irregularities such as trees and other obstacles. When the algorithm fails to detect any significant deviations from the dominant ground plane (e.g. sidewalk surface) in the scene, it seems sensible to apply a more sophisticated stereo algorithm such as an MRF model to examine the scene in more detail. A second algorithm such as this may reveal the presence of a curb or other subtle depth discontinuity that was missed by the first algorithm.

The slant of the ground plane means that the disparity of the ground changes appreciably from one image row to the next, violating the fronto-parallel assumption. One possible solution to this problem, originally proposed in [16], is to warp one of the images so as to remove the disparity corresponding to the ground plane. Thus, only scene points that lie off the ground plane will have non-zero disparity, and planes parallel to the ground plane (e.g. the road bordering the sidewalk) will have roughly *uniform* disparities. In this way the image data is transformed so that the fronto-parallel bias is appropriate.

Such a transformation may prove valuable for our application, but a more general solution is to impose a prior that assumes that locally planar surfaces *with arbitrary slant and tilt* are common. One way to enforce such a prior is to penalize deviations of the second derivatives of the disparity field from zero. At a minimum, such a prior must evaluate the relationship among *three* consecutive pixel disparities, since a second derivative requires three consecutive samples to be estimated. (A second derivative of zero implies that the three points are collinear in 3-D.) This measure is beyond the capability of the *pairwise* MRF presented in this paper, and a straightforward implementation using a more powerful MRF with ternary (triplet) interactions would be extremely computationally demanding. Recent work [17] replaces BP for such an implementation with another energy minimization algorithm that is much more efficient for this problem. The result is a tractable stereo algorithm with superior performance, particularly in its ability to propagate surface information on non-fronto-parallel surfaces.

1.6 Biological Plausibility

Several neuroscientists and psychophysicists have asked me if MRF models such as the ones described in this paper have anything to do with biological vision systems. While I am not an expert on biological vision, I would like to point to work by others arguing that the MRF-BP framework (perhaps extended to incorporate multiple depth cues) may be biologically plausible.

From a biological perspective, perhaps the most important property of models cast in this framework is that they are fully parallelizable: one can implement BP in a parallel hardware system with one computing node for each variable in the MRF, with directed connections between neighboring variables to represent BP messages. In each iteration of BP, messages flow along these connections from each variable node to neighboring nodes. Lee and Mumford [18] have argued that BP may be a model for how information is passed top-down and bottom-up in the brain. Recent research [19] has established that BP for MRFs with binary-valued variables (i.e. each variable can assume only two possible states) can be formulated with continuous time updates (rather than discrete time updates), resulting in behavior that closely matches the dynamics of a Hopfield network. Other work [20] relaxes the assumption of binary-valued variables and relates BP to a spiking network model.

1.7 Discussion

In this paper I have described a standard MRF framework for propagating surface information in 3-D reconstruction in the presence of noisy and sparse depth cues. In addition to automatically weighing prior and likelihood information according to their reliability, the framework is the basis for many of the top-performing stereo algorithms in computer vision (see [10] and the website associated with it, vision.middlebury.edu/stereo, which maintains up-to-date performance rankings of state-of-the-art stereo algorithms). While the standard prior used in MRF stereo algorithms imposes an unnatural fronto-parallel bias, promising recent work demonstrates the value of using a more realistic prior that accommodates the frequent occurrence of locally planar surfaces with arbitrary slant and tilt.

Although 3-D reconstruction algorithms have improved a lot in recent years, much work remains. Despite the recent emphasis on learning model parameters from training data, the images used for training and testing often contain more highly textured, colorful objects than commonly occur in real-world scenes, which casts doubt on the ability of even the top-performing algorithms to generalize to the real-world domain. Additional performance measures may need to be developed to reward algorithms that minimize the kinds of *catastrophic* inference errors that are all too common at present, in which the disparities of some points are estimated incorrectly by tens of pixels.

More realistic priors will be also needed for algorithms to improve further. Such priors will be higher-level than the ones described here, and may need to represent *coherent surfaces*, such as planar and cylindrical patches with explicit boundaries, rather than pixel-based depth or disparity fields.

Another avenue for improvement will be to integrate multiple depth cues,

including monocular cues such as shading and texture, in addition to standard disparity cues. (Indeed, impressive work by Saxena et al [21] estimates a depth field from a *single* color image using such cues.) It will also be important to integrate information over time (i.e. multiple video frames).

Finally, it is worth pointing out that improvements in optimization techniques such as BP will be required to realize many of the proposed extensions above, and may well influence the direction of future research.

1.8 Acknowledgments

The author was supported by National Science Foundation grant no. IIS0415310. I would like to thank Dr. Volodymyr Ivanchenko and Dr. Ender Tekin for helpful feedback on this manuscript.

References

[1] Haines, T.S., Wilson, R.C.: Belief propagation with directional statistics for solving the shape-from-shading problem. In: Proc. European Conference on Computer Vision (ECCV). (2008)

[2] White, R., Forsyth, D.: Combining cues: Shape from shading and texture. In: Proc. Computer Vision and Pattern Recognition (CVPR). (2006)

[3] Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: Proc. Computer Vision and Pattern Recognition (CVPR). (2006)

[4] Chellappa, R., A. Jain, e.: Markov Random Fields: Theory and Application. (Boston: Academic Press)

[5] Y. Boykov, O.V., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(11) (2001)

[6] Yedidia, J., Freeman, W., Weiss, Y.: Understanding belief propagation and its generalizations. MERL Cambridge Research Technical Report **TR 2001-16** (2001)

[7] Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the International Conference on Machine Learning. Volume 18. (2001)

[8] Scharstein, D., Pal, C.: Learning conditional random fields for stereo. In: Proceedings of Computer Vision and Pattern Recognition (CVPR). Volume 2. (2005) 838–845

[9] Zhang, L., Seitz, S.: Parameter estimation for mrf stereo. In: Computer Vision and Pattern Recognition (CVPR). (2005)

[10] Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision **47**(1) (2002)

[11] Tola, E., Lepetit, V., Fua, P.: A fast local descriptor for dense matching. In: Proceedings of Computer Vision and Pattern Recognition (CVPR). (2008)

[12] Felzenszwalb, P., Huttenlocher, D.: Efficient belief propagation for early vision. International Journal of Computer Vision **70**(1) (2006)

[13] Gershenfeld, N.: The Nature of Mathematical Modeling. Cambridge University Press (1998)

[14] Yang, Q., Wang, L., Yang, R.: Real-time global stereo matching using hierarchical belief propagation. In: British Machine Vision Conference (BMVC). (2006) III:989

[15] Ivanchenko, V., Coughlan, J., Gerrey, B., Shen, H.: Computer vision-based clear path guidance for blind wheelchair users. In: 10th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2008). (2008)

[16] Burt, P., Wixson, L., Salgian, G.: Electronically directed "focal" stereo. In: ICCV 95. (1995) 94–101

[17] Woodford, O.J., Torr, P.H.S., Reid, I.D., Fitzgibbon, A.W.: Global stereo reconstruction under second order smoothness priors. In: Computer Vision and Pattern Recognition (CVPR). (2008)

[18] Lee, T., Mumford, D.: Hierarchical bayesian inference in the visual cortex. JOSA A (**20**(7)) 1434–1448

[19] Ott, T., Stoop, R.: The neurodynamics of belief propagation on binary markov random fields. In: Neural Information Processing Systems (NIPS). (2006)

[20] Doya, K., Ishii, S., Pouget, A., R. Rao, e.: The Bayesian brain: Probabilistic Approaches to Neural Coding. MIT Press (2006)

[21] Saxena, A., Sun, M., Ng, A.: Learning 3-d scene structure from a single still image. In: ICCV workshop on 3D Representation for Recognition (3dRR-07). (2007)