Stereo Matching Using Belief Propagation

Jian Sun, Nan-Ning Zheng, *Senior Member*, *IEEE*, and
Heung-Yeung Shum, *Senior Member*, *IEEE*

Abstract—In this paper, we formulate the stereo matching problem as a Markov network and solve it using Bayesian belief propagation. The stereo Markov network consists of three coupled Markov random fields that model the following: a smooth field for depth/disparity, a line process for depth discontinuity, and a binary process for occlusion. After eliminating the line process and the binary process by introducing two robust functions, we apply the belief propagation algorithm to obtain the maximum a posteriori (MAP) estimation in the Markov network. Other low-level visual cues (e.g., image segmentation) can also be easily incorporated in our stereo model to obtain better stereo results. Experiments demonstrate that our methods are comparable to the state-of-the-art stereo algorithms for many test cases.

Index Terms—Stereoscopic vision, belief propagation, Markov network, Bayesian inference.

1 Introduction

Stereo vision infers 3D scene geometry from two images with different viewpoints. This fundamental problem has been investigated for many years not only in computer vision but also in cognitive science and psychophysiology. Recent applications such as view synthesis and image-based rendering make stereo vision again an active research topic in computer vision.

Classical dense two-frame stereo matching computes a dense disparity or depth map from a pair of images under known camera configuration. In general, the scene is assumed Lambertian or intensity-consistent from different viewpoints, without specularities, reflective surfaces, or transparency. The known camera configuration can provide a powerful epipolar geometry constraint for matching. Stereo matching remains a difficult vision problem for the following reasons.

- **Noise.** There are always unavoidable light variations, image blurring, and sensor noise in image formation. A practical stereo algorithm must be robust.
- **Textureless regions.** This is also called the *aperture problem*. The intensity-consistency constraint is useless in textureless regions. Thus, information from highly textured regions needs to be propagated into textureless regions for stereo matching, e.g., by using spatial smoothness constraint.
- **Depth discontinuities.** The spatial smoothness constraint should be broken at object (depth) boundaries. In other words, information propagation should stop at depth discontinuities.
- **Occlusions.** Occluded pixels in one view should not be matched with pixels in the other view.

- *J. Sun and N.-N. Zheng are with Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China. E-mail: sj@aiar.xjtu.edu.cn, nnzheng@xjtu.edu.cn.*
- *H.-Y. Shum is with Microsoft Research Asia, Beijing 100080, China. E-mail: hshum@microsoft.com.*

Manuscript received 24 July 2002; revised 25 Feb. 2003; accepted 3 Mar. 2003.
Recommended for acceptance by J. Rehg.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 118202.

Clearly, stereo matching is an ill-posed problem with inherent ambiguities. The Bayesian approach provides a promising way for such ill-posed problems because it treats a task as an inference problem or finding the "best guess" solution. For stereo matching, we want to infer scene structure S given images I. The output from the Bayesian approach is not only a single solution but also a posterior probability distribution $P(S|I)$. By Bayes law, $P(S|I) \propto P(I|S)P(S)$, where $P(I|S)$ is the likelihood that encodes the process of forward image formation and $P(S)$ is the prior that encodes our assumptions on scene structure.

The Bayesian approach has many advantages when applied to stereo vision. It can encode various prior constraints, e.g., spatial smoothness, uniqueness, and the ordering constraint. It can also deal with uncertainties in stereo matching. Because the Bayesian approach states explicitly what assumptions are made, the strengths and the weaknesses of the proposed algorithm can be clearly examined. In addition to stereoscopic vision, people also use other cues to infer scene structure, e.g., shape from shading, shape from shadows, shape from focus, shape from silhouette, and shape from texture. The Bayesian approach provides a natural way to integrate the information from multiple sensors.

There are two contributions in this paper. First, we formulate stereo matching using three MRF's and subsequently estimate the optimal solution by a Bayesian Belief Propagation algorithm. Second, we propose a probabilistic framework to integrate additional information (e.g., segmentation) into the stereo algorithm.

The rest of paper is organized as follows: After reviewing related work in Section 2, we propose in Section 3 a novel stereo matching approach to explicitly model discontinuities, occlusions, and the disparity field in the Bayesian framework. In Section 4, Bayesian Belief Propagation is applied to infer the stereo matching. The basic stereo model is then extended in Section 5 to integrate other cues such as region similarity. The experimental results shown in Section 6 demonstrate that our model is effective and efficient. In Section 7, we adapt the stereo model for multiview stereo. Finally, we discuss in Section 8 why our stereo matching with belief propagation can produce results that are comparable to the state-of-the-art stereo algorithms.

2 Related Work

In this section, we review related stereo algorithms, especially those using the Bayesian approach. We refer the reader to a more detailed and updated taxonomy of dense, two-frame stereo correspondence algorithms by Scharstein and Szeliski [30]. A testbed for quantitative evaluation of stereo algorithms is also given in [30].

A stereo algorithm is called a global method if there is a global objective function to be optimized. Otherwise, it is called a local method. The central problem of local or window-based stereo matching methods is to determine the optimal size, shape, and weight distribution of aggregation support for each pixel. An ideal support region should be bigger in textureless regions and should be suspended at depth discontinuities. The central problem of global algorithms is not only to define a good objective function but also to provide an effective computing method to find local or global minimum. In the taxonomy of Scharstein and Szeliski [30], a local method consists of matching cost computation, aggregation of cost, and disparity computation; a global method consists of matching cost computation and disparity optimization. From the Bayesian point of view, matching cost computation is a measurement or observation. The most common matching costs, e.g., squared intensity difference(SD), absolute intensity difference [20], normalized-cross correlation [28], [7], binary matching cost [25], rank transform [35], shifted absolute difference [3], are ways of computing the likelihood function. Different aggregation methods reflect different priors assumed on scene structure. For example, a fixed-window method implies a frontal-plane scene, and a 3D window method limits the disparity gradient. Obviously, the fixed window is invalid at depth discontinuities. Some improved window-based methods, such as adaptive windows [20] and shiftable windows [6], [33], [21] try to avoid windows that span depth discontinuities.

Bayesian methods (e.g., [13], [18], [2], [10], [6]) are global methods that model discontinuities and occlusion. Bayesian methods can be classified into two categories: dynamic programming-based or MRFs-based, depending on the computation model. Geiger et al. [13] and Ishikawa and Geiger [18] derived an occlusion process and a disparity field from a matching process. Assuming an "order constraint" and "uniqueness constraint," the matching process becomes a "path-finding" problem where the global optimum is obtained by dynamic programming. Belhumeur [2] defined a set of priors from a simple scene to a complex scene. A simplified relationship between disparity and occlusion is used to solve scanline matching by dynamic programming. Unlike Geiger and Belhumeur who enforced a piecewise-smooth constraint, Cox et al. [10] and Bobick and Intille [6] did not require the smoothing prior. Assuming corresponding features are normally distributed and a fixed cost for occlusion, Cox proposed a dynamic programming solution using only the occlusion constraint and ordering constraints. Bobick and Intille incorporated the Ground Control Points constraint to reduce the sensitivity to occlusion cost and the computation complexity of Cox's method. These dynamic programming methods assume that the occlusion cost is the same in each scanline.

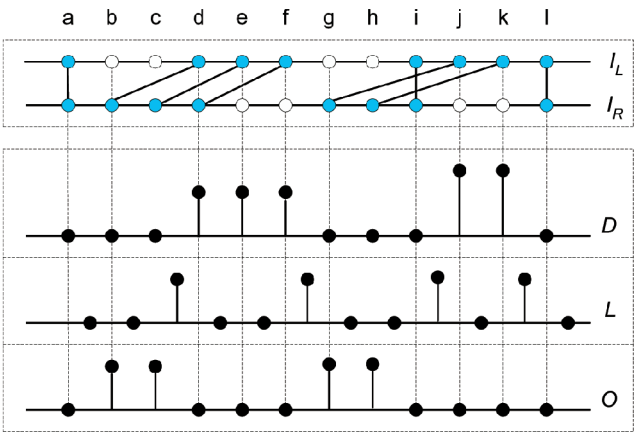


Fig. 1. A scene illustrates the geometric relationship among depth process D, discontinuity process L, and occlusion process O. Matched points between I_L (the reference view) and the right image I_R are connected by thick lines.

Ignoring the dependence between scanlines results in the characteristic "streaking" in the disparity maps.

Markov Random Fields (MRF) is a powerful tool to model spatial interaction. Bayesian stereo matching can be formulated as a maximum a posteriori MRF (MAP-MRF) problem. There are several methods to solve the MAP-MRF problem: simulated annealing [14], Mean-Field annealing [12], the Graduated Non-Convexity algorithm (GNC) [5], and Variational approximation [17]. Finding a solution by simulated annealing can often take an unacceptably long time although global optimization is achievable in theory. Mean-Field annealing is a deterministic approximation to simulated annealing by attempting to average over the statistics of the annealing process. It reduces execution time at the expense of solution quality. GNC can only be applied to some special energy functions. Variational approximation converges to a local minimum. Recently, the Graph Cut (GC) method [8] has been proposed based on the max flow algorithm in graph theory. This method is a fast efficient algorithm to find a local minimum for a MAP-MRF whose energy function is Potts or Generalized Potts.

The absence of an efficient stochastic computing method has made probabilistic models less attractive. In this paper, we formulate a probabilistic stereo model that can be efficiently solved by a Bayesian Belief Propagation algorithm.

3 Basic Stereo Model

We model stereo matching by three coupled MRF's: D is the smooth disparity field defined on the image lattice of the reference view, L is a spatial line process located on the dual of the image lattice and represents explicitly the presence or absence of depth discontinuities in the reference view, and O is a spatial binary process to indicate occlusion regions in the reference view. Fig. 1 illustrates these processes in the 1D case.

Using Bayes' rule, the joint posterior probability over D, L, and O given a pair of stereo images $I = \{I_L, I_R\}$, where I_L, I_R is the left (reference) and right images, respectively, is:

$$P(D, L, O|I) = \frac{P(I|D, L, O)P(D, L, O)}{P(I)}. \quad (1)$$

Without occlusion, $\{D, L\}$ are coupled MRF's proposed by [14] to model a piecewise-smooth surface with two random fields: one representing the variable required to estimate, the other representing its discontinuities. Similar models such as the "weak membrane" model [5] in surface reconstruction and the "Mumford-Shah" model in image segmentation [26] have also been studied in computer vision. However, in image formation of stereo pairs, the piecewise-smooth scene is projected on a pair of stereo images. Some regions are only visible in one image. Each pixel in the occlusion region has no matching pixel in the other view. For example, in Fig. 1, points b, c, g, h from I_L cannot be matched in I_R. Adding occlusion process O into the piecewise-smooth model $\{D, L\}$ is therefore necessary.

3.1 Likelihood

We assume that the likelihood $P(I|D, O, L)$ is independent of L,

$$P(I|D, O, L) = P(I|D, O) \tag{2}$$

because the observation (I) is pixel-based. Assuming that the observation noise follows an independent identical distribution (i.i.d.), we can define the likelihood $P(I|D, O)$ as:

$$P(I|D, O) \propto \prod_{s \notin O} \exp(-F(s, d_s, I)), \tag{3}$$

where $F(s, d_s, I)$ is the matching cost function of pixel s with disparity d_s given observation I. Our likelihood considers the pixels only in nonoccluded areas $\{s \notin O\}$ because likelihood in occluded areas cannot be well defined.

For the matching cost, we use Birchfield and Tomasi's pixel dissimilarity, which is provably insensitive to image sampling [3]:

$$F(s, d_s, I) = \min\{\overline{d}(s, s', I)/\sigma_f, \overline{d}(s', s, I)/\sigma_f\},$$

where

$$\overline{d}(s, s', I) = \min\{|I_L(s) - I_R^-(s')|, |I_L(s) - I_R(s')|, |I_L(s) - I_R^+(s')|\},$$

s' is the matching pixel of s in the right view with disparity d_s, $I_R^-(s')$ is the linearly interpolated intensity halfway between s' and its neighboring pixel to the left, $I_R^+(s')$ is the linearly interpolated intensity halfway between s' and its neighboring pixel to the right, $\overline{d}(s', s, I)$ is the symmetric version of $\overline{d}(s, s', I)$, and σ_f is the image noise variance to be estimated.

3.2 Prior

There is no simple statistical relationship between coupled fields $\{D, L\}$ and field O. The ordering constraint [1] assumes that the order of neighboring correspondences is always preserved. This ordering allows the construction of a dynamic programming scheme. However, this constraint may not always be true. For instance, this constraint is violated when a thin object is close to the viewer. As shown in Fig. 1, a thin object $\{j, k\}$ causes the order of points i and j in I_L to be different from that of their matched points in I_R.

In this paper, we ignore the statistical dependence between O and $\{D, L\}$ and assume that:

$$P(D, O, L) = P(D, L)P(O). \tag{4}$$

The Markov property asserts that the conditional probability of a site in the field depends only on its neighboring sites. Assuming D, L, and O follow the Markov property, by specifying the first order neighborhood system $G(s)$ and $N(s) = \{t | t > s, t \in G(s)\}$ of site s, the prior (4) can be expanded as:

$$P(D, L, O) \propto \prod_s \prod_{t \in N(s)} \exp(-\varphi_c(d_s, d_t, l_{s,t})) \prod_s \exp(-\eta_c(o_s)), \tag{5}$$

where $\varphi_c(d_s, d_t, l_{s,t})$ is the joint clique potential function of sites d_s, d_t (neighbor of d_s) and $l_{s,t}$. $l_{s,t}$ is the line variable between d_s and d_t, and $\eta_c(o_s)$ is the clique potential function of o_s. $\varphi_c(d_s, d_t, l_{s,t})$ and $\eta_c(o_s)$ are user-customized functions to enforce the contextual constraints for stereo matching. To enforce spatial interactions between d_s and $l_{s,t}$, we define $\varphi_c(d_s, d_t, l_{s,t})$ as follows:

$$\varphi_c(d_s, d_t, l_{s,t}) = \varphi(d_s, d_t)(1 - l_{s,t}) + \gamma(l_{s,t}), \tag{6}$$

where $\varphi(d_s, d_t)$ penalizes the different assignments of neighboring sites when no discontinuity exists between them and $\gamma(l_{s,t})$ penalizes the occurrence of a discontinuity between sites s and t. Typically, $\gamma(0) = 0$.

By combining (3), (5), and (6), our basic stereo model (1) becomes:

$$(P(D, O, L|I) \propto \prod_{s \notin O} \exp(-F(s, d_s, I)) \prod_s \exp(-\eta_c(o_s)) \prod_s \prod_{t \in N(s)} \exp(-(\varphi(d_s, d_t)(1 - l_{s,t}) + \gamma(l_{s,t}))). \tag{7}$$

4 Approximate Inference by Belief Propagation

To find the MAP solution of (7), we need to:

- determine the forms and parameters of $\varphi(d_s, d_t)$, $\gamma(l_{s,t})$, and $\eta_c(o_s)$ and
- provide a tractable inference algorithm.

It is, however, nontrivial to specify or to learn appropriate forms and parameters of $\varphi(d_s, d_t)$, $\gamma(l_{s,t})$, and, especially, $\eta_c(o_s)$. Even if the forms and parameters are given, it is still difficult to find the MAP of a composition of a continuous MRFs D and two binary MRFs L and O. Although the Markov Chain Monte Carlo (MCMC) [14], [15] approach provides an effective way to explore a posterior distribution, the computational requirement makes MCMC impractical for stereo matching. The solution space of our model is $\Omega = \Omega_d \times \Omega_l \times \Omega_o$, where Ω_d, Ω_l, and Ω_o are the solution spaces of depth, discontinuity, and occlusion, respectively.

This is why we need to make some approximations on both the model and algorithm. In Section 4.1, the unification of line process and robust statistics [4] provides us a way to eliminate the binary random variable from our MAP problem. In Section 4.2, after converting MRFs to the corresponding Markov network, the approximate inference algorithm, a loopy belief propagation algorithm can be used to approximate the posterior probability for stereo matching.

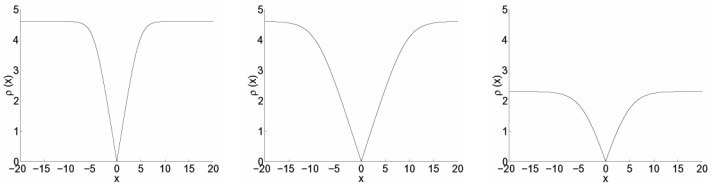


Fig. 2. (a) $e = 0.01, \sigma = 1.0$. (b) $e = 0.01, \sigma = 2.0$. (c) $e = 0.1, \sigma = 2.0$. Robust function $\rho(x) = -\ln((1-e)\exp(-\frac{|x|}{\sigma}) + e)$ derived from TV model. Parameters σ and e, respectively, control the sharpness and the upper-bound of the function.

4.1 Model Approximation: From Line Process to Outlier Process

Maximization of the posterior (7) can be rewritten as

$$\max_{D,L,O} P(D,L,O|I) = \max_{D}\left\{\max_{O}\prod_{s}\exp(-(F(s,d_s,I)(1-o_s)+\eta_c(o_s)o_s)) \max_{L}\prod_{s}\prod_{t\in N(s)}\exp(-(\varphi(d_s,d_t)(1-l_{s,t})+\gamma(l_{s,t})))\right\} \quad (8)$$

because the first two factors on the r.h.s of (7) are independent of L and the last factor on the r.h.s of (7) is independent of O.

Now, we relax the binary processes $l_{s,t}$ and o_s to analog processes $l^a_{s,t}$ and o^a_s ("outlier process" [4]) by allowing $0 \le l^a_{s,t} \le 1$ and $0 \le o^a_s \le 1$. For the first term in (8),

$$\begin{aligned} &\max_{O}\prod_{s}\exp(-(F(s,d_s,I)(1-o^a_s)+\eta_c(o^a_s)o^a_s)) \\ &= \exp(-\min_{O}\sum\nolimits_{s}(F(s,d_s,I)(1-o^a_s)+\eta_c(o^a_s)o^a_s)), \end{aligned} \quad (9)$$

where $\min_{O}\sum_s (F(s,d_s,I)(1-o^a_s)+\eta_c(o^a_s)o^a_s)$ is the objective function of a robust estimator. The robust function of this robust estimator [4] is

$$\rho_d(d_s) = \min_{o^a_s}(F(s,d_s,I)(1-o^a_s)+\eta_c(o^a_s)o^a_s). \quad (10)$$

For the second term in (8), we also have a robust function $\rho_p(d_s,d_t)$:

$$\rho_p(d_s,d_t) = \min_{l^a_{s,t}}(\varphi(d_s,d_t)(1-l^a_{s,t})+\gamma(l^a_{s,t})). \quad (11)$$

We get the posterior probability over D defined by two robust functions:

$$P(D|I) \propto \prod_{s}\exp(-\rho_d(d_s))\prod_{s}\prod_{t\in N(s)}\exp(-\rho_p(d_s,d_t)). \quad (12)$$

Thus, we not only eliminate two analog line processes via the outlier process but also model outliers in measurements. We convert the task of modeling the prior terms $\{\eta_c(o_s), \varphi(d_s,d_t), \gamma(l_{s,t})\}$ explicitly into defining two robust functions $\rho_d(d_s)$ and $\rho_p(d_s,d_t)$ that model occlusion and discontinuity implicitly.

In this paper, our robust functions are derived from the Total Variance (TV) model [23] with the potential function $\rho(x) = |x|$ because of its discontinuity preserving property. We truncate this potential function as our robust function:

$$\rho_d(d_s) = -\ln\left((1-e_d)\exp\left(-\frac{|F(s,d_s,I)|}{\sigma_d}\right)+e_d\right), \quad (13)$$

$$\rho_p(d_s,d_t) = -\ln\left((1-e_p)\exp\left(-\frac{|d_s-d_t|}{\sigma_p}\right)+e_p\right). \quad (14)$$

Fig. 2 shows different shapes of our robust functions. By varying parameters e and σ, we control the shape of the robust function and, therefore, the posterior probability.

After approximating the model, the next task is to provide an effective and efficient inference algorithm. We describe below how the belief propagation algorithm is used to compute the MAP of the posterior distribution (12).

4.2 Algorithm Approximation: Loopy Belief Propagation

In the literature of probabilistic graph models [19], a Markov network is an undirected graph as shown in Fig. 3. Nodes $\{x_s\}$ are hidden variables and nodes $\{y_s\}$ are observed variables. By denoting $X = \{x_s\}$ and $Y = \{y_s\}$, the posterior $P(X|Y)$ can be factorized as:

$$P(X|Y) \propto \prod_{s}\psi_s(x_s,y_s)\prod_{s}\prod_{t\in N(s)}\psi_{st}(x_s,x_t), \quad (15)$$

where $\psi_{st}(x_s,x_t)$ is called the compatibility matrix between nodes x_s and x_t, and $\psi_s(x_s,y_t)$ is called the local evidence for node x_s. In fact, $\psi_s(x_s,y_s)$ is the observation probability $p(y_s|x_s)$. If the number of discrete states of x_s is L, $\psi_{st}(x_s,x_t)$ is an $L \times L$ matrix and $\psi_s(x_s,y_s)$ is a vector with L elements.

It can be observed that the form of our posterior (12) is same as the form of (15). If we define

$$\psi_{st}(x_s,x_t) = \exp(-\rho_p(x_s,x_t)), \quad (16)$$

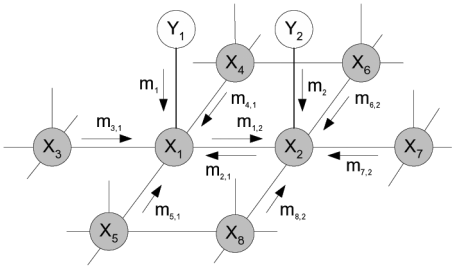


Fig. 3. Local message passing in a Markov Network. Gray nodes are hidden variables. White nodes are observable variables. In the "max-product" algorithm, the new message sent from node x_1 to x_2 is: $m^{new}_{1,2} \leftarrow \kappa \max_{x_1}\psi_{12}(x_1,x_2)m_1m_{3,1}m_{4,1}m_{5,1}$. The belief at node x_1 is computed as: $b_1 \leftarrow \kappa m_1 m_{2,1} m_{3,1} m_{4,1} m_{5,1}$.

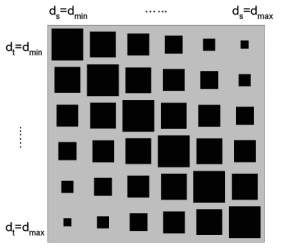


Fig. 4. Compatibility matrix $\psi_{st}(x_s, x_t)$. The range of disparity is $[d_{min}, d_{max}]$. A larger box represents a bigger value.

$$\psi_s(x_s, y_s) \propto \exp(-\rho_d(x_s)), \tag{17}$$

our posterior (12) is exactly the posterior of a Markov network. Fig. 4 gives an illustration of $\psi_{st}(x_s, x_t)$ for our stereo model. Thus, finding the MAP of (12) is equal to finding the MAP of a Markov network.

For this Markov network, exact inference such as variable elimination is obviously intractable due to the large state space of D. Approximation methods include variational methods, sampling methods, bounded cutset conditioning, and parametric approximation methods [19]. In particular, loopy belief propagation is a linear time algorithm proportional to the number of hidden nodes. Loopy belief propagation applies Pearl's algorithm [27] to the graph that has loops. For graphs without loops, Pearl's algorithm is an exact inference method. For graph with loops, such as our Markov network for stereo matching, the belief propagation algorithm cannot guarantee the global optimal solution. Despite loops in the network, however, belief propagation has been applied successfully to some vision [11] and communication [34] problems recently.

Belief propagation (BP) is an iterative inference algorithm that propagates messages in the network. Let $m_{st}(x_s, x_t)$ be the message that node x_s sends to x_t, $m_s(x_s, y_s)$ be the message that observed node y_s sends to node x_s (in fact, $m_s(x_s, y_s) = \psi_s(x_s, y_s)$), and $b_s(x_s)$ be the belief at node x_s. Note that $m_{st}(x_s, x_t)$, $m_s(x_s, y_s)$, and $b_s(x_s)$ are all vectors with L elements. We simplify $m_{st}(x_s, x_t)$ as $m_{st}(x_t)$, and $m_s(x_s, y_s)$ as $m_s(x_s)$. There are two kinds of BP algorithms with different message update rules: "max-product" and "sum-product" which maximize the joint posterior $P(X|Y)$ of the network and the marginal posterior of each node $P(x_s|Y)$, respectively. The standard "max-product" algorithm is shown below:

1. Initialize all messages $m_{st}(x_t)$ as uniform distributions and messages $m_s(x_s) = \psi_s(x_s, y_s)$.

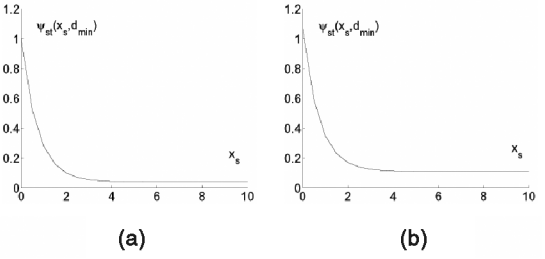


Fig. 5. (a) $seg(s) = seg(t)$. The left shows the first row of $\psi_{st}(x_s, x_t)$ when node x_s and x_t are in the same region. (b) $seg(s) \neq seg(t)$. The right shows the first row of $\psi_{st}(x_s, x_t)$ when node x_s and x_t are in different regions. ($\{e_p = 0.01, \sigma_p = 1.0, \lambda_{seg} = 0.05$).

2. Update messages $m_{st}(x_t)$ iteratively for i = 1:T

$$m_{st}^{i+1}(x_t) \leftarrow \kappa \max_{x_s} \psi_{st}(x_s, x_t) m_s^i(x_s) \prod_{x_k \in N(x_s) \backslash x_k} m_{ks}^i(x_s).$$

3. Compute beliefs

$$b_s(x_s) \leftarrow \kappa m_s(x_s) \prod_{x_k \in N(x_s)} m_{ks}(x_s)$$

$$x_s^{MAP} = \arg\max_{x_k} b_s(x_k).$$

For example, in Fig. 3, the new message sent from node x_1 to x_2 is updated as: $m_{1,2}^{new} \leftarrow \kappa \max_{x_1} \psi_{12}(x_1, x_2) m_1 m_{3,1} m_{4,1} m_{5,1}$. The belief at node x_1 is computed as: $b_1 \leftarrow \kappa m_1 m_{2,1} m_{3,1} m_{4,1} m_{5,1}$ (the product of two messages is the component-wise product); κ is the normalization constant.

The computational complexity of a standard "max-product" BP algorithm is $O(TNL^2)$, where N is the number of pixels and T is the number of iterations. Most of the computation focuses on the multiplication of matrix $\psi_{st}(x_s, x_t)$ and vector $m_s(x_s) \prod_{x_k \in N(x_s) \backslash x_s} m_{ks}(x_s)$. However, in our experiments, some statistical properties of messages can be used to speed up belief propagation.

Propagation Speedup. It can be observed that each row of $\psi_{st}(x_s, x_t)$ is a unique peak distribution in our stereo model. In our experiments, most messages have unique peaks. We can exploit this property to identify unnecessary computation during iterations. We simplify matrix $\psi_{st}(x_s, x_t)$ as $[a_1^T, \cdots, a_L^T]^T$, $m_s(x_s) \prod_{x_k \in N(x_s) \backslash x_s} m_{ks}(x_s)$ as b and $m_{st}^{i+1}(x_t)$ as c. The message update at one iteration is:

$$c(i) = \arg\max_j a_i(j) \cdot b(j). \tag{18}$$

TABLE 1
Quantitative Statistics Based on Known Ground Truth Data

Percentage of bad matching pixels in non-occlusion regions $\bar{\mathtt{O}}$	$B_{\bar{\mathtt{O}}} = \frac{1}{N} \sum_{s \in \bar{\mathtt{O}}} (\lvert d(s) - d_T(s) \rvert > \delta_d)$
Percentage of bad matching pixels in textureless regions $\bar{\mathtt{T}}$	$B_{\bar{\mathtt{T}}} = \frac{1}{N} \sum_{s \in \bar{\mathtt{T}}} (\lvert d(s) - d_T(s) \rvert > \delta_d)$
Percentage of bad matching pixels in depth discontinuity regions $\mathtt{D}$	$B_{\mathtt{D}} = \frac{1}{N} \sum_{s \notin \mathtt{D}} (\lvert d(s) - d_T(s) \rvert > \delta_d)$

$\delta_d = 1$ *in all our experiments.*

TABLE 2
The Performance of Different Stereo Algorithms with Fixed Parameters on Four Test Image Pairs

Algorithms	Tsukuba			Sawtooth			Venus			Map	
	$B_{\bar{0}}$	$B_{\bar{T}}$	$B_{\bar{D}}$	$B_{\bar{0}}$	$B_{\bar{T}}$	$B_{\bar{D}}$	$B_{\bar{0}}$	$B_{\bar{T}}$	$B_{\bar{D}}$	$B_{\bar{0}}$	$B_{\bar{D}}$
Belief prop. (seg)	1.15	0.42	6.31	0.98	0.30	4.83	1.00	0.76	9.13	0.84	5.27
Belief prop.	1.61	0.66	9.17	0.85	0.37	7.92	1.17	1.00	12.87	0.67	3.42
Graph cuts [30]	1.94	1.09	9.49	1.30	0.06	6.34	1.79	2.61	6.91	0.31	3.88
GC+occl. [22]	1.27	0.43	6.90	0.36	0.00	3.65	2.79	5.39	2.54	1.79	10.08
Graph cuts [8]	1.86	1.00	9.35	0.42	0.14	3.76	1.69	2.30	5.40	2.39	9.35
Realtime SAD [16]	4.25	4.47	15.05	1.32	0.35	9.21	1.53	1.80	12.33	0.81	11.35
Bay. diff. [30]	6.49	11.62	12.29	1.43	0.69	9.29	3.89	7.15	18.17	0.20	2.49
SSD+MF [30]	5.23	3.80	24.66	2.21	0.72	13.97	3.74	6.82	12.94	0.66	9.35
Scanl. opt. [30]	5.08	6.78	11.94	4.06	2.64	11.90	9.44	14.59	18.20	1.84	10.22
Dyn. prog. [30]	4.12	4.63	12.34	4.84	3.71	13.26	10.10	15.01	17.12	3.33	14.04

The underlined number is the best in its category.

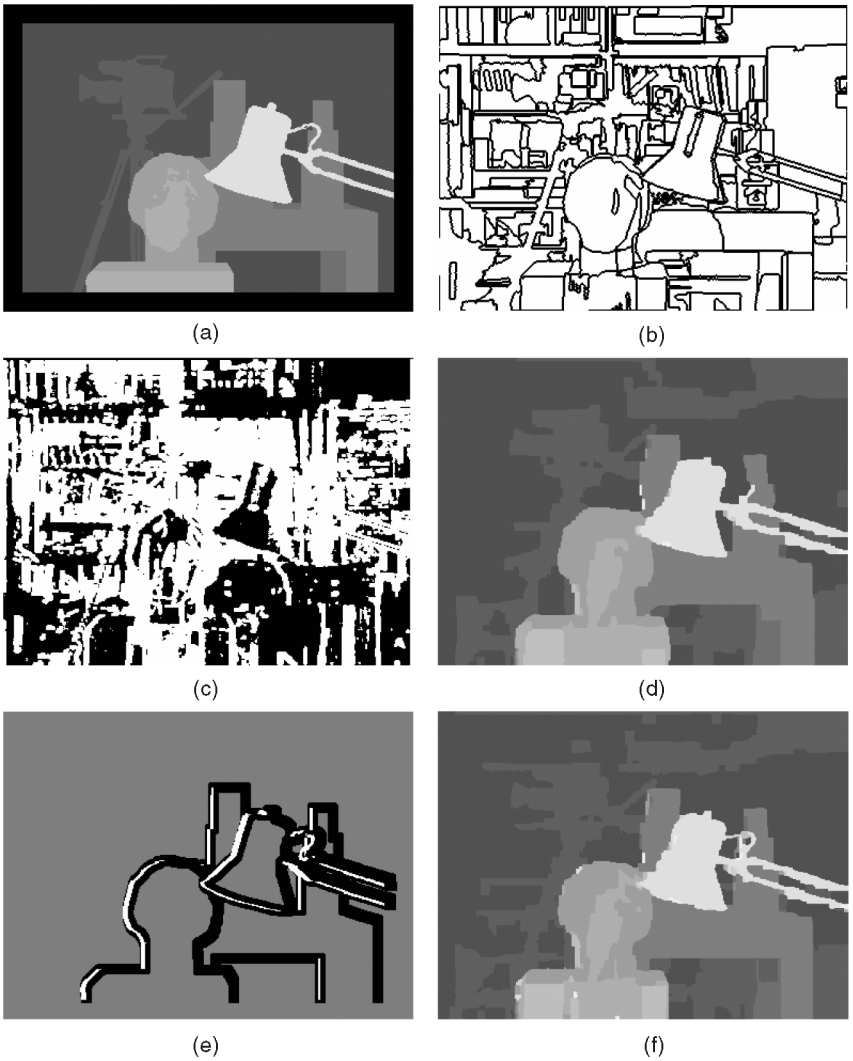


Fig. 6. The results on the Tsukuba data set. (a) Ground truth. (b) Image segmentation result. (c) Textureless regions. (d) Max-product result without segmentation. (e) Discontinuity (white) and occlusion (black) regions. (e) Max-product result with segmentation.

We denote the peak positions of a_i and b by $j_{a_i}^{max}$ and j_b^{max} separately. If both a_i and b are unique peak distributions, the position of c's peak j_c^{max} must lie between $j_{a_i}^{max}$ and j_b^{max}. Thus, we can avoid unnecessary multiplications for the messages with unique peaks. This simple accelerating technique can improve the efficiency about 30-60 percent in our experiments.

5 Integrating Multiple Cues

More low-level visual cues (e.g., segmentation, edges, corners) can be incorporated into the intensity constraint to improve stereo matching. Recently, a segmentation-based stereo algorithm [32] has been proposed based on the assumption that the depth discontinuities occur on the boundary of the segmented regions. In [32], the segmentation

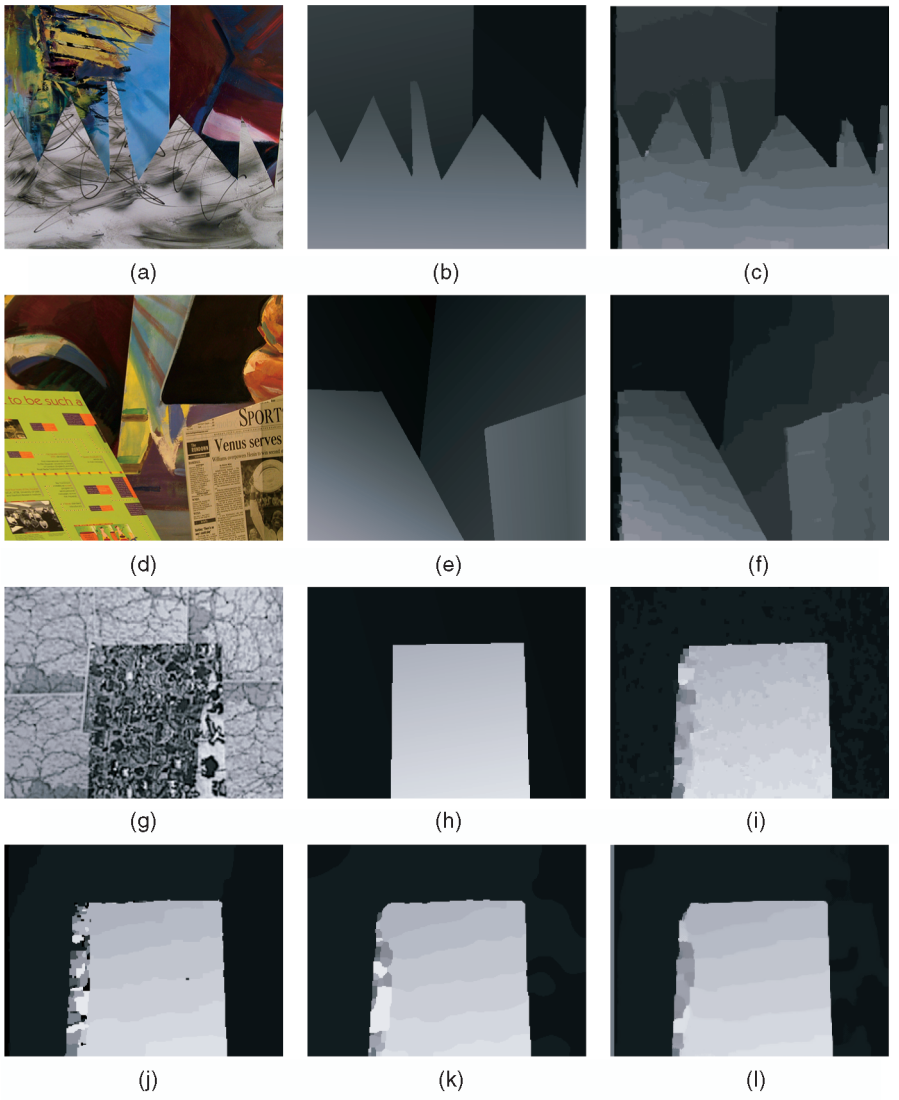


Fig. 7. (a) Sawtooth: image. (b) Ground truth. (c) Max-product result. (d) Venus: image. (e) Ground truth. (f) Max-product result. (g) Map: image. (h) Ground truth. (i) Max-product result. (j) Graph cut result. (k) Bayes diffusion result. (l) Sum-product result. The results of Sawtooth, Venus, and Map based on the "max-product" algorithm are shown in (c), (f), and (i). Graph Cut ($B_{\bar{O}} = 0.31$, $B_D = 3.88$) and Bayesian diffusion results ($B_{\bar{O}} = 0.20$, $B_D = 2.49$) are shown in (j) and (k), while the "sum-product" result ($B_{\bar{O}} = 0.16$, $B_D = 2.11$) is shown in (l).

results are used as hard constraints. In our work, we incorporate segmentation results into our basic stereo model as soft constraints (priors) under a probabilistic framework.

With additional cues, we extend the basic stereo model (12) to:

$$P(D,O,L|I) \propto \prod_s \exp(-\rho_d(d_s)) \prod_s \prod_{t\in N(s)} \exp(-\varphi_c(d_s,d_t,l_{s,t})) \exp(-\rho_{pcue}(d_s,d_t)), \tag{19}$$

where $\rho_{pcue}(d_s, d_t)$ encodes some constraints between sites. To integrate region similarities from image segmentation, we define $\rho_{pcue}(d_s, d_t)$ as:

$$\rho_{pcue}(d_s,d_t) = \rho_{seg}(d_s,d_t) = \begin{cases} 0 & seg(s) = seg(t) \\ \lambda_{seg} & seg(s) \neq seg(t), \end{cases} \tag{20}$$

where $seg(s)$ is the label of the segmentation result at site s. The larger the λ_{seg}, the more difficult passing the message between neighbor sites becomes. In other words, the influence from neighbors becomes smaller as λ_{seg} increases. In our experiments, the segmentation labels are produced by the Mean-Shift algorithm [9]. It takes just a few seconds for each image used in our experiments.

With the introduction of $\rho_{pcue}(d_s, d_t)$, the compatibility matrix $\psi_{st}(x_s, x_t)$ becomes:

$$\psi_{st}(x_s,x_t) = \exp(-\rho_p(x_s,x_t))) \exp(-\rho_{pcue}(x_s,x_t))). \tag{21}$$

Fig. 5 shows the first row of $\psi_{st}(x_s, x_t)$ when x_s and x_t are in the same region and in different regions.

When the scene consists of several 3D planes, layers extracted can also be treated as a cue. We can define $\rho_{pcue}(d_s, d_t)$ as:

$$\rho_{pcue}(d_s,d_t) = \rho_{layer}(d_s,d_t) = \begin{cases} 0 & layer(s) = layer(t) \\ \lambda_{layer} & layer(s) \neq layer(t). \end{cases} \tag{22}$$

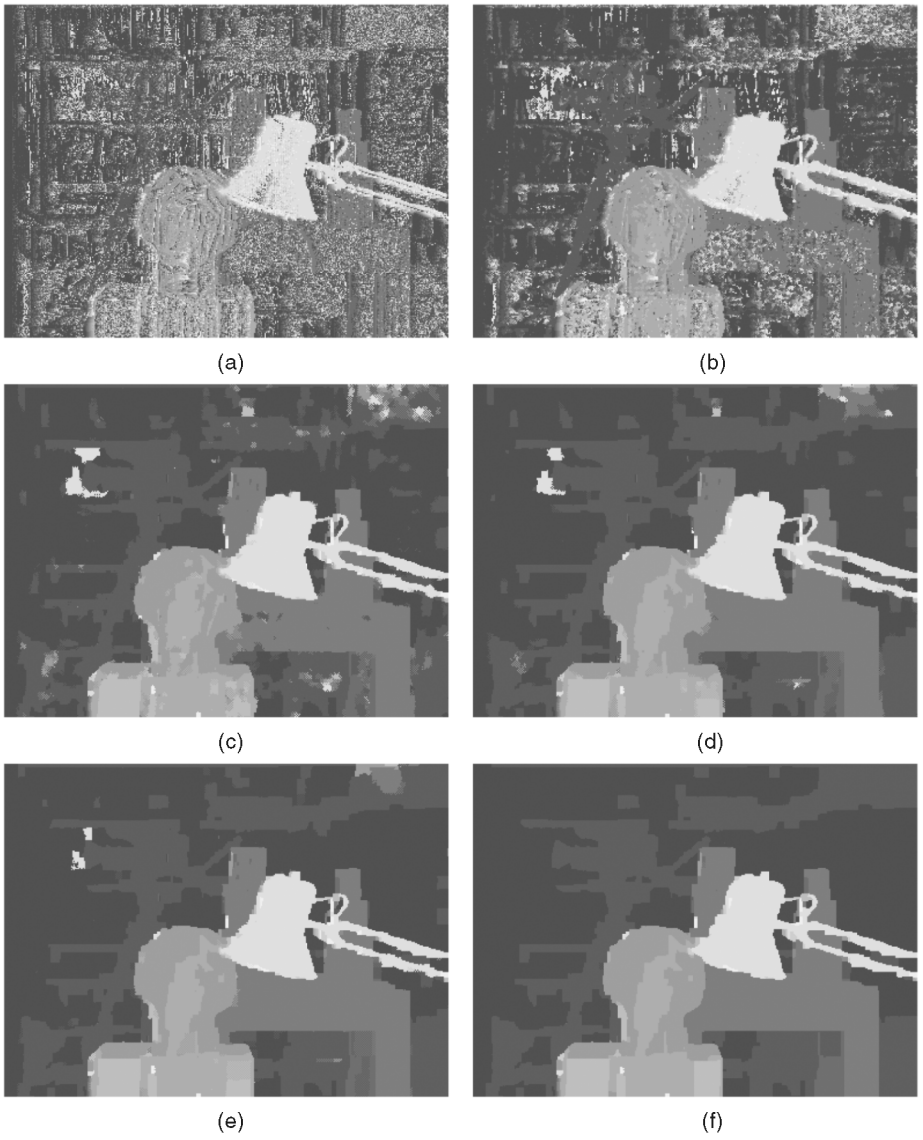


Fig. 8. (a) Iteration (0). (b) Iteration (1). (c) Iteration (8). (d) Iteration (16). (e) Iteration (32). (f) Iteration (64). Intermediate results on Tsukuba data at different iterations.

6 Experimental Results

In this paper, we evaluate the performance of our stereo algorithm using the quality measures proposed in [30] based on known ground truth data listed in Table 1. In particular, $B_{\bar{O}}$ represents the overall performance of a stereo algorithm.

TABLE 3
Running Time of BP Algorithm on Tsukuba Data

Iterations	Time(seconds)		Tsukuba
	standard	speedup	$B_{\bar{0}}$
1	1.4	1.1	19.67
8	10.2	7.9	4.24
16	21.6	15.7	2.53
32	40.4	28.4	1.69
64	82.8	51.6	1.15
64 (Dual CPU)	43.1	27.4	1.15

The test data set consists of four pairs of images: "Map," "Tsukuba," "Sawtooth," and "Venus" [30]. "Tsukuba" is a complicated indoor environment with frontal surfaces and contains a number of integer-valued disparities. Other pairs consist of mainly slanted planes.

Table 2 shows the results of applying our BP algorithm to all four pairs of images. It also lists the results of other stereo algorithms. This table is courtesy of Scharstein and Szeliski (see http://www.middlebury.edu/stereo/results.html for more details). Our results with and without image segmentation incorporated into stereo matching are shown in the first and the second row, respectively.

For a complicated environment like "Tsukuba," incorporating image segmentation improves stereo matching significantly, with a 40 percent error reduction in $B_{\bar{O}}$. In fact, our algorithm ranks as the best for "Tsukuba" and outperforms Graph Cut (with occlusion) [22] which was widely regarded as one of the best current stereo matching algorithms. Our algorithm compares well with other stereo algorithms for the three other data sets, "Sawtooth,"

"Venus," and "Map." It is interesting to note that, for these three data sets with simple slanted surfaces, incorporating image segmentation does not necessarily improve stereo matching, as seen from the first and second rows.

Figs. 6 and 7 show the results obtained by our algorithm. The segmentation map is obtained by the Mean-Shift algorithm with default parameters suggested by [9]. Note that a fixed set of parameters $\{e_d = 0.01, \sigma_d = 8, e_p = 0.05, \sigma_p = 0.6\}$ is used in our BP algorithm for all four image pairs. Obviously, this set of parameters is not the optimal for "Map" because the disparity range of this data is almost twice that of "Tsukuba."

In our experiments on the "sum-product" BP algorithm, most results are overly smooth because the objective function of the "sum-product" BP algorithm is the marginal posterior of each node. However, the best result (Fig. 7l) is obtained for "Map" data by "sum-product" BP algorithm. Figs. 7j and 7k are the best two results given by Graph-cut and Bayesian Diffusion.

To evaluate the efficiency of the BP algorithm, we present in Fig. 8 and Table 3 the intermediate results and running time for "Tsukuba" data on a Pentium IV 1.7GHz PC at different iterations. Two characteristics of our BP algorithm can be observed from this experiment. First, most disparity computation is completed in the first several iterations. Second, the speedup method becomes more effective in later iterations.

The BP algorithm is very suitable for hardware implementation because the message update at each iteration of BP algorithm is parallelizable. The last row of Table 3 is the running time of a parallel version of the BP algorithm on a dual CPU Pentium IV 1.7GHz PC. The parallel efficiency $EP = \frac{82.8}{2\times 43.1} \approx 1$ demonstrates potential for real-time high-performance stereo.

The local oscillation phenomena of the BP algorithm also occurred in our experiments. A time average operation is executed after a fixed number of iterations: $m^t_{st}(x_t) = m^{t-1}_{st}(x_t) + m^t_{st}(x_t)$. This heuristic worked well in our experiments.

7 Multiview Stereo

In multiview stereo, the observation is a collection of images $\{I_k, k = 0 \ldots K\}$ with camera intrinsic parameters $\{A_k\}$ and camera extrinsic parameters $\{R_k, t_k\}$. The likelihood (3) in our basic stereo model is modified as follows:

$$P(I|D,O) \propto \prod_{s \notin O} \exp\left(-\sum_{k}^{K} w(s,d_s,k) F(s,d_s,I_r,I_k)\right), \quad (23)$$

where r is an index of the reference view, $F(s,d_s,I_r,I_k)$ is the matching cost function of pixel s with disparity d_s between I_r and I_k, and $w(s,d_s,k)$ is the convolution kernel. In multiview stereo, $w(s,d_s,k)$ plays a role for visible view selecton.

7.1 Matching Cost Function

We define s^k as the matching point of s in image I_k with disparity d_s. In generalized multiview stereo, the image coordinate of s^k is:

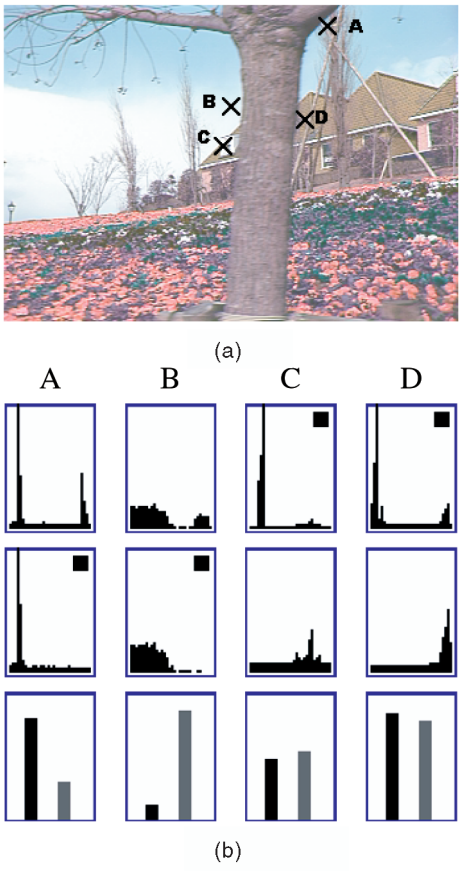


Fig. 9. Local evidence with different convolution kernels. (a) Four points (A, B, C, and D) in the sixth frame of the "garden" sequence (11 frames) are semiocclusion points. (b) Each column corresponds A, B, C, and D, respectively. The first row shows local evidence $\Psi_s(x_s, y_s)$ using convolution kernel $w_d(s,d_s,k)$. The horizonal and vertical coordinates are disparity and matching probabilities, respectively. The second row shows local evidence $\Psi_s(x_s, y_s)$ using convolution kernel $w_F(s,d_s,k)$. The last row shows F^- (black) and F^+ (gray). The evidences marked with a black rectangle at the top right corner are computed by our convolution kernel $w_a(s,d_s,k)$.

$$\begin{bmatrix} s^k_x \\ s^k_y \\ 1 \end{bmatrix} \cong [\, A_k \quad 0\,] \begin{bmatrix} R_k & t_k \\ 0 & 1 \end{bmatrix} \begin{bmatrix} R_r & t_r \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} A_r^{-1} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} s_x \\ s_y \\ 1 \\ d_s \end{bmatrix}. \quad (24)$$

This mapping can be represented by a forward warp function: $s^k = H_k(s,d_s)$. For an arbitrary camera configuration, we generalized Birchfield and Tomasi's [3] shift absolute difference along the epipolar line:

$$F(s,d_s,I_r,I_k) = \min\{\overline{d_1}(s,s^k,I_r,I_k)/\sigma_f, \overline{d_2}(s,s^k,I_r,I_k)/\sigma_f\}$$

$$\overline{d_1}(s,s^k,I_r,I_k) = \min\{|I_r(s) - I_k^-(s^k)|, |I_r(s) - I_k(s^k)|, |I_r(s) - I_k^+(s^k)|\}$$

$$\overline{d_2}(s,s^k,I_r,I_k) = \min\{|I_r^-(s) - I_k(s^k)|, |I_r(s) - I_k(s^k)|, |I_r^+(s) - I_k(s^k)|\},$$

where $I_k^-(s^k)$ is the linearly interpolated intensity halfway between s^k and $H_k(s,d_s-1)$. Similarly, $I_k^+(s^k)$ is between s^k and $H_k(s,d_s+1)$, $I_r^-(s)$ is between s and $H_k^{-1}(s^k,d_s-1)$, and $I_r^+(s)$ is between s and $H_k^{-1}(s^k,d_s+1)$.

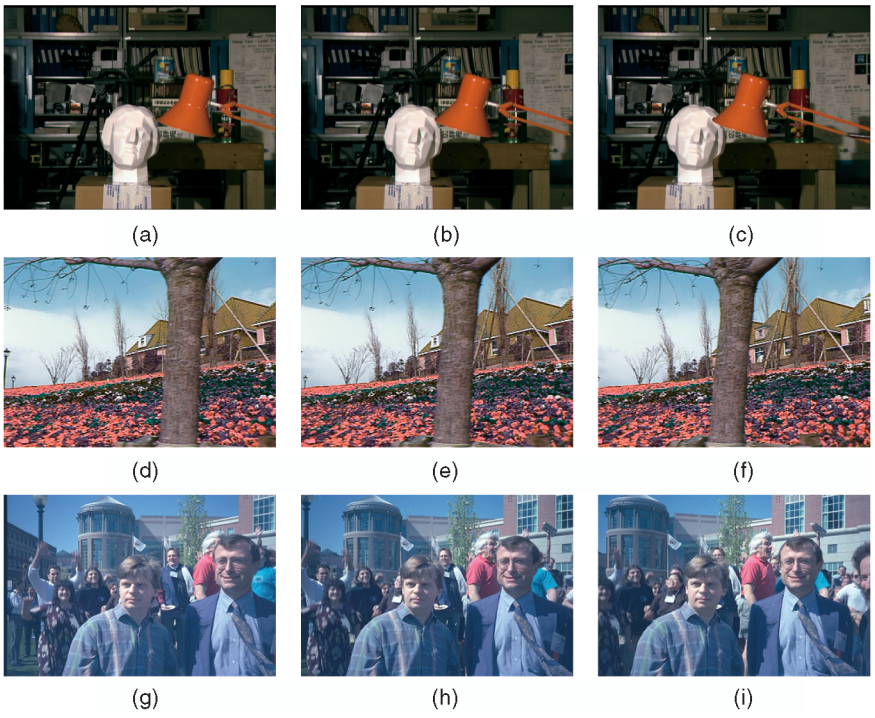


Fig. 10. "Tsukuba" (1st, 3rd, 5th frame), "Garden" (1st, 6th, 11th frame), and "Dayton" (1st, 3rd, 5th frame) data.

7.2 Convolution Kernel

The crux of multiview stereo is to find an optimal convolution kernel $w(s, d_s, k)$. To handle occlusion, $w(s, d_s, k)$ should be zero when the 3D point (s, d_s) cannot be seen in I_k. In [21], a temporal selection method is based on the assumption that the occlusion region in the reference view will be seen in the left views $\{I_k, k = 0, \cdots, r-1\}$ or the right views $\{I_k, k = r+1, \cdots, K\}$. Under this assumption, we define two kinds of convolution kernels as:

$$w_d(s, d_s, k) = \begin{cases} \frac{\delta(F^-_{d_s} \le F^+_{d_s})}{r-1} & k < r \\ 0 & k = r \\ \frac{\delta(F^-_{d_s} \ge F^+_{d_s})}{K-r} & k > r, \end{cases} \tag{25}$$

where $F^-_{d_s} = \sum_{k=0}^{r-1} F(s, d_s, I_r, I_k)$ and $F^+_{d_s} = \sum_{k=r+1}^{K} F(s, d_s, I_r, I_k)$, and

$$w_F(s, d_s, k) = \begin{cases} \frac{\delta(F^- \le F^+)}{r-1} & k < r \\ 0 & k = r, \\ \frac{\delta(F^- \ge F^+)}{K-r} & k > r \end{cases} \tag{26}$$

where $F^- = \sum_{k=0}^{r-1} \sum_{d_s} F(s, d_s, I_r, I_k)$ and $F^+ = \sum_{k=r+1}^{K} \sum_{d_s} F(s, d_s, I_r, I_k)$.

The convolution kernel $w_d(s, d_s, k)$ is dependent on the disparity d_s. It keeps more information in measurements than $w_F(s, d_s, k)$. However, $w_d(s, d_s, k)$ contains more ambiguities. To reach a balance, an adaptive convolution kernel is defined as follows:

$$w_a(s, d_s, k) = \begin{cases} w_d(s, d_s, k) & \min\{F^-, F^+\} \ge m \max\{F^-, F^+\} \\ w_F(s, d_s, k) & \min\{F^-, F^+\} < m \max\{F^-, F^+\}, \end{cases} \tag{27}$$

where m $(0 < m < 1)$ is a winner threshold. When there is an obvious winner between F^- and F^+, we take $w_F(s, d_s, k)$ to reduce ambiguity. Otherwise, we take $w_d(s, d_s, k)$ to keep more information.

The advantage of our adaptive convolution kernel can be illustrated by considering the local evidence distribution $\Psi_s(x_s, y_s)$ depicted in Fig. 9. Points A, B, C, and D are all semiocclusion points. For A and B, we prefer kernel $w_F(s, d_s, k)$ to kernel $w_d(s, d_s, k)$ for less ambiguity because of the big difference between F^- and F^+. For C and D, $w_F(s, d_s, k)$ is a more risky choice than $w_d(s, d_s, k)$ because there is no obvious winner between F^- and F^+.

7.3 Multiview Stereo Experiments

For multiview stereo, three sequences "Tsukuba" (5 frames), "Garden" (11 frames) and "Dayton" (5 frames) are used as our test data, as shown in Fig. 10. Fig. 11 shows the results of applying our BP algorithm. Figs. 11a, 11c, and 11e are depth maps of the third frame in "Tsukuba," sixth frame in "Garden" and third frame in "Dayton," respectively. Table 4 gives the quantitative performance improvement on "Tsukuba" data. Obviously, the depths of occlusion regions in two-view stereo are recovered very well. Because there is no ground truth for "Garden" and "Dayton," we present results of new view synthesis shown in Figs. 11b, 12d, and 13f. The reference view is forward warped to a new viewpoint using a computed depth map by a two-pass algorithm [31]. Some large textureless regions, such as the sky in "Garden" and "Dayton," are still hard to handle.

8 Discussion

8.1 Assumptions

In Bayesian approaches, all assumptions must be made explicitly. In order to apply the BP algorithm, the assumption

Fig. 11. The results of multiview stereo. (a), (c), and (e) are depth maps. (b), (d), and (f) are novel views rendered. (a) Our result (3rd frame). (b) Novel view. (c) Our result (6th frame). (d) Novel view. (e) Our result (3rd frame). (f) Novel view. (Yellow regions are disocclusion regions. Please refer to the electronic version for better quality viewing.)

made in our priors is (2): occlusion is independent of discontinuity. In fact, an occlusion region doesn't need to exist given a discontinuity.

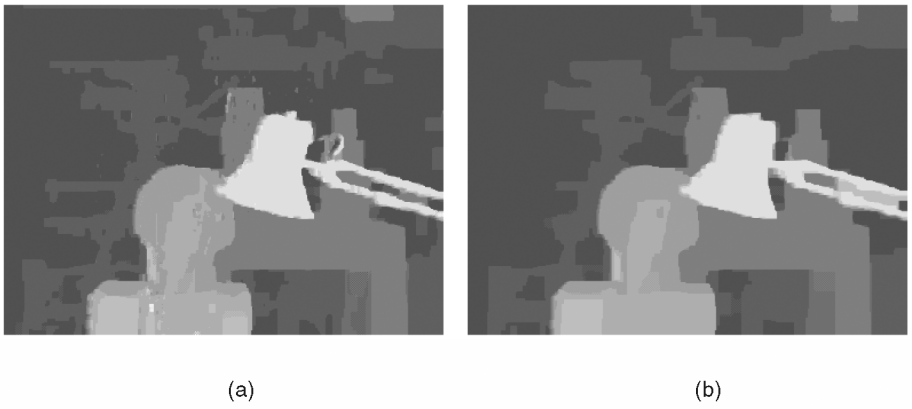
However, modeling the conditional probability $P(O|D, L)$ needs longer distance pixel interaction that is beyond the ability of our first-order MRF system. On the other hand, the shortage of efficient inference algorithms prevents us from using higher order MRFs.

Although good experimental results are obtained with the independence assumption, further investigations on this issue would be useful. One possible approach is to enforce the uniqueness constraint, such as the method in [22]. However, the uniqueness constraint is hard to impose as in [22] because we use a probabilistic distribution as the final solution. Another possible approach is to resort to a region-based method, such as neighborhood depth hypothesis [32] to infer occlusions. A more promising approach to handle occlusion for two-frame stereo matching is Left Right Check(LRC) [24].

TABLE 4
The Performance of Our BP Multiview Stereo Algorithm on "Tsukuba" Sequence

Algorithms	Tsukuba		
	$B_{\bar{\mathtt{O}}}$	$B_{\bar{\mathtt{T}}}$	$B_{\mathtt{D}}$
Belief prop. (multi-view)	0.95	0.25	5.40
Belief prop. (two-view)	1.15	0.42	6.31

In Section 4.2, we simplify the basic stereo model from (7) to (12) by introducing two robust functions. The model that is most similar to our posterior probability (12) is Scharstein and Szeliski's [29]. Unlike Scharstein and Szeliski's contaminated Gaussian cost function, we used an absolute difference cost in the robust functions $\rho_d(d_s)$ and $\rho_p(d_s, d_t)$. The energy function corresponding to our prior of D is a Total Variance (TV) energy. In many applications, such as image restoration and denoising, it has been shown that the TV model is more successful than the Gaussian model for edge preservation. To illustrate the intrinsic characteristics of the TV model, Fig. 12a gives the result using the TV model $\rho_p(d_s, d_t) = \frac{|d_s - d_t|}{\sigma_p}$ and Fig. 12b gives the result using the



(a) (b)

Fig. 12. (a) Result of TV model ($B_O = 2.16$, $B_T = 1.01$, $B_{DD} = 11.67$). (b) Result of Gaussian model ($B_O = 2.82$, $B_T = 1.58$, $B_{DD} = 15.14$). The results obtained by TV model $\rho_p(d_s, d_t) = \frac{|d_s - d_t|}{\sigma_p}$ and Gaussian model $\rho_p(d_s, d_t) = \frac{|d_s - d_t|^2}{\sigma_p}$. The parameter σ_p is chosen experimentally to produce the best result for each model.

Gaussian model $\rho_p(d_s, d_t) = \frac{|d_s - d_t|^2}{\sigma_p}$. We observe that the result using the Gaussian model is overly smooth. This experiment also demonstrates that the TV model is more robust with regard to the robust parameter e_d than the Gaussian model.

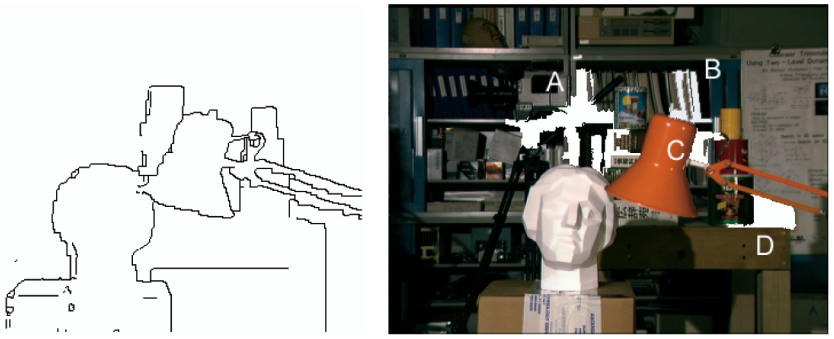
The eliminated discontinuity process L can be recovered from the depth map through the robust function $\rho_p(d_s, d_t)$. We identify a discontinuity between nodes s and t when $\rho_p(d_s, d_t)$ reaches an upper bound. Fig. 13a shows the recovered discontinuity map for the Tsukuba data. For the occlusion process O, the pixels that do not have a low matching cost can also be recovered from $\rho_s(d_s)$ similarly. But, our model cannot identify occluded pixels that have a low matching cost. This shows the ability and limitation of our robust function for occlusion handling.

8.2 Why Does BP Work?

The magic of the BP algorithm lies in its powerful message passing. A message presents the probability that the receiver should be at a disparity according to all information from the sender up to the current iteration. Message passing has two important properties. First, it is asymmetric. The entropy of the messages from high-confidence nodes to low-confidence nodes is smaller than the entropy of the messages from low-confidence nodes to high-confidence nodes. Second, it is adaptive. The influence of a message between a pair of nodes with larger divergence would be weakened more.

Therefore, BP's message passing provides a time-varying adaptive support region for stereo matching to deal with textureless regions and depth discontinuities elegantly. In textureless regions, for example, the influence of a message can be passed far away. On the other hand, the influence in discontinuous regions will fall off quickly. Fig. 14 shows an example of this adaptive smoothing procedure. In Fig. 14, the image pair is modified from that used in [20] and [29]. A linear ramp in the direction of the baseline is used as the underlying intensity pattern. The disparities of the background and the foreground are 2 and 5, respectively. Unlike [20] or [29], a smaller pure textureless square is overlapped in the center of the foreground of original stereo pair. This modification makes the original example harder.

We use entropy $H(b) = -\sum_i b_i log b_i$ to measure the confidence of the belief, and the symmetric version of the Kullback-Leiber (KL) divergence $KL_s(b^1 \| b^2) = \sum_i (b_i^1 - b_i^2)\ log(\frac{b_i^1}{b_i^2})$ to measure the difference between belief b^1 and b^2. Smaller entropy represents higher confidence of a belief. Larger divergence represents larger dissimilarity between beliefs. As shown in the Fig. 14, the entropy map of a belief represents the confidence of disparity estimation for each node. Clearly, the confidence of each



(a) (b)

Fig. 13. (a) Recovered discontinuity map. (b) White regions A, B, C, and D are incorrect segments.

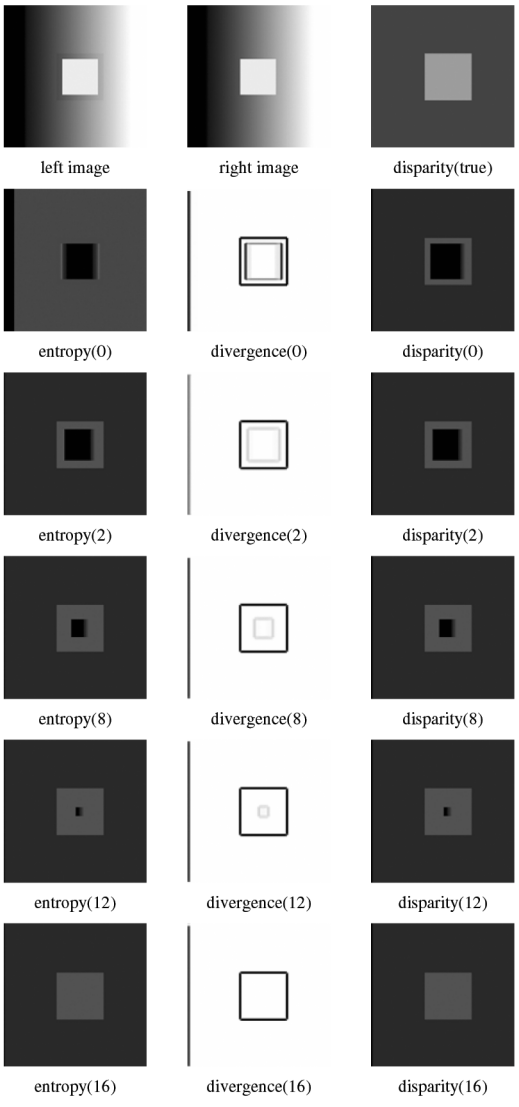


Fig. 14. Time-varying adaptive smoothing mechanism of the BP algorithm in stereo matching is illustrated from row 2 to row 6. The input image pair and the ground truth are shown in the first row. The number of the iterations is shown in the parentheses.

node increases with each iteration. Note that the confidence in occlusion regions and corners is lower than that of other regions. The probabilistic method outputs not only a solution, but also its certainty. The divergence map of a belief shows where message-passing is stopped. The divergence map after convergence illustrates the ideal support regions.

8.3 Image Segmentation and Multiview Stereo

A segmentation-based method like [32] assumes the depth discontinuities must appear at the boundaries of segmentation. There are two differences on the use of image segmentation between our method and [32]. First, segmentation results are treated as a prior but not a hard constraint. Our method is more robust to incorrect oversegmentation results because there is still message passing between regions. Second, discontinuity is preserved in the interior of a segment. In [32], the depth representation of each segment is a frontal plane, or a 3D plane, or a plane-plus-parallax. This prohibits depth discontinuity in a segment. Oversegmentation cannot guarantee that there must be a boundary corresponding to a discontinuity, such as a low texture sphere surface. Our method can still identify a discontinuity in an undersegmented segment. Fig. 13b gives the illustrations. In segments A, B, C, and D, our method finds correct depth discontinuities.

In our experiments, the larger textureless regions require more BP iterations that propagate belief from outside to inside. This inspires a two-step method. First, we estimate a 3D model parameter for large segments and compute the initial depth for the pixels in these segments. Second, we prune the evidence of the pixels in these segments and run the BP algorithm. One possible pruning method is to convolute the evidence distribution $\psi_s(x_s, y_s)$ with a Gaussian kernel centered at the initial depth for each pixel. The key is that the results obtained in Step 1 are again used as a soft constraint. This reduces the chance of falling into a local minimum introduced in Step 1.

In multiview stereo, most occlusions can be handled well by a temporal support region. In Figs. 11a, 11c, 11e, sharp depth discontinuities nearby occlusion regions are recovered. There is some performance improvement on the Tsukuba data obtained from multiview stereo in comparison to two-view stereo (see Table 4). This demonstrates that degradation of performance caused by our simplified model (12) is small in two-frame stereo. In other words, the posterior distribution $P(O|I)$ is approximated well by our simplified model (12).

9 Summary

In this paper, stereo matching is formulated as a Bayesian inference problem with three coupled MRF's that is solved efficiently by the Belief Propagation algorithm. Image segmentation is also integrated into our basic stereo model as a soft constraint. We further extend the two-view stereo model to multiview stereo. Excellent experimental results demonstrate the power of probabilistic models and approximate inference algorithms. For future work, we plan to investigate how to improve stereo matching using Generalized Belief Propagation [34].

Acknowledgments

The authors would like to thank Dr. Sing Bing Kang of Microsoft Research for his constructive comments and multiview stereo data. The anonymous reviewers gave very useful comments and suggestions. This work waspartially supported by the National Natural Science Foundation of China under grant 60024301. This work is performed when J. Sun visited Microsoft Research Asia.

References

[1] H. Baker and T. Binford, "Depth from Edge and Intensity Based Stereo," *Proc. Int'l Joint Conf. Artificial Intelligence*, 1981.
[2] P.N. Belhumeur, "A Bayesian-Approach to Binocular Stereopsis," *Int'l J. Computer Vision*, vol. 19, no. 3, pp. 237-260, 1996.
[3] S. Birchfield and C. Tomasi, "A Pixel Dissimilarity Measure that Is Insensitive to Image Sampling," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 4, pp. 401-406, Apr. 1998.

[4] M.J. Black and A. Rangarajan, "On the Unification of Line Processes, Outlier Rejection, and Robust Statistics with Applications in Early Vision," *Int'l J. Computer Vision,* vol. 19, no. 1, pp. 57-91, 1996.
[5] A. Blake and A. Zisserman, *Visual Reconstruction.* MIT Press 1987.
[6] A.F. Bobick and S.S. Intille, "Large Occlusion Stereo," *Int'l J. Computer Vision,* vol. 33, no. 3, pp. 1-20, 1999.
[7] R.C. Bolles, H.H. Baker, and M.J. Hannah, "The Jisct Stereo Evaluation," *Proc. DARPA Image Understanding Workshop,* 1993.
[8] Y. Boykov, O. Veksler, and R. Zabih, "Fast Approximate Energy Minimization via Graph Cuts," *Proc. Int'l Conf. Computer Vision,* 1999.
[9] D. Comaniciu and P. Meer, "Robust Analysis of Feature Spaces: Color Image Segmentation," *Proc. Computer Vision and Pattern Recognition,* 1997.
[10] I.J. Cox, S.L. Hingorani, S.B. Rao, and B.M. Maggs, "A Maximum-Likelihood Stereo Algorithm," *Computer Vision and Image Understanding,* vol. 63, no. 3, pp. 542-567, 1996.
[11] W.T. Freeman, E.C. Pasztor, and O.T. Carmichael, "Learning Low-Level Vision," *Int'l J. Computer Vision,* vol. 40, no. 1, pp. 25-47, 2000.
[12] D. Geiger and F. Girosi, "Parallel and Deterministic Algorithms from MRFs: Surface Reconstruction," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 13, no. 5, pp. 401-412, May 1991.
[13] D. Geiger, B. Ladendorf, and A. Yuille, "Occlusions and Binocular Stereo," *Int'l J. Computer Vision,* vol. 14, no. 3, pp. 211-226, 1995.
[14] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 6, no. 6, pp. 721-741, 1984.
[15] W. Gilks, S. Richardson, and D. Spiegelhalter, *Markov Chain Monte Carlo in Practice.* Chapman and Hall, 1996.
[16] H. Hirschmueller, " Improvements in Real-Time Correlation-Based Stereo Vision," *Proc. IEEE Workshop Stereo and Multi-Baseline Vision,* 2001.
[17] B.K.P. Horn and M.J. Brooks, "The Variational Approach to Shape from Shading," *Computer Vision, Graphics, and Image Processing,* vol. 33, no. 2, pp. 174-208, 1986.
[18] H. Ishikawa and D. Geiger, "Occlusions, Discontinuities, and Epipolar Lines in Stereo," *Proc. European Conf. Computer Vision,* 1998.
[19] M.I. Jordan, *Learning in Graphical Models.* MIT Press, 1998.
[20] T. Kanade and M. Okutomi, "A Stereo Matching Algorithm with an Adaptive Window: Theory and Experiment," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 16, no. 9, pp. 920-932, Sept. 1994.
[21] S.B. Kang, R. Szeliski, and J. Chai, "Handling Occlusions in Dense Multi-View Stereo," *Proc. Computer Vision and Pattern Recognition,* 2001.
[22] V. Kolmogorov and R. Zabih, "Computing Visual Correspondence with Occlusions via Graph Cuts," *Proc. Int'l Conf. Computer Vision,* 2001.
[23] S. Osher, L.I. Rudin, and E. Fatemi, "Nonlinear Total Variation Based Noise Removal Algorithms," *Physica D,* vol. 27, no. 60, pp. 259-268, 1992.
[24] A. Luo and H. Burkhardt, "An Intensity-Based Cooperative Bidirectional Stereo Matching with Simultaneous Detection of Discontinuities and Occlusions," *Int'l J. Computer Vision,* vol. 15, no. 3, pp. 171-188, 1995.
[25] D. Marr and T.A. Poggio, "Cooperative Computation of Stereo Disparity," *Science,* vol. 194, no. 4262, pp. 283-287, 1976.
[26] D. Mumford and J. Shah, "Optimal Approximations by Piecewise Smooth Functions and Variational Problems," *Comm. Pure and Applied Math.,* 1988.
[27] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Mateo, Calif.: Morgan Kaufmann Publishers, 1988.
[28] T.W. Ryan, R.T. Gray, and B.R. Hunt, "Prediction of Correlation Errors in Stereo-Pair Images," *Optical Eng.,* vol. 19, no. 3, pp. 312-322, 1980.
[29] D. Scharstein and R. Szeliski, "Stereo Matching with Nonlinear Diffusion," *Int'l J. Computer Vision,* vol. 28, no. 2, pp. 155-174, 1998.
[30] D. Scharstein and R. Szeliski, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms," *Int'l J. Computer Vision,* vol. 47, no. 1, pp. 7-42, 2002.
[31] J.W. Shade, S.J. Gortler, L.W. He, and R. Szeliski, "Layered Depth Images," *Proc. SIGGRAPH,* 1998.
[32] H. Tao, H.S. Sawhney, and R. Kumar, "A Global Matching Framework for Stereo Computation," *Proc. Int'l Conf. Computer Vision,* 2001.
[33] O. Veksler, "Stereo Matching by Compact Windows via Minimum Ratio Cycle," *Proc. Int'l Conf. Computer Vision,* 2001.
[34] J.S. Yedidia, W.T. Freeman, and Y. Weiss, "Bethe Free Energy, Kikuchi Approximations, and Belief Propagation Algorithms," Technical Report TR-2001-16, Mitsubishi Electric Reseach, 2001.
[35] R. Zabih and J. Woodfill, "Non-Parametric Local Transforms for Computing Visual Correspondence," *Proc. European Conf. Computer Vision,* 1994.

Jian Sun received the BS degree from the College of Electrical and Communication Engineering, Xi'an Jiaotong University in 1997 and the MS degree from the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University in 2000. He is a PhD student in joint program between Xi'an Jiaotong University and Microsoft Research Asia. His research interests include computer vision and machine learning.

Nan-Ning Zheng (SM'93) graduated in 1975 from the Department of Electrical Engineering, Xi'an Jiaotong University, Xi'an, China, received the ME degree in information and control engineering from Xi'an Jiaotong University, Xi'an, China in 1981, and the PhD degree in electrical engineering from Keio University, Japan, in1985. He is currently a professor and the director of the Institute of Artificial Intelligence and Robotics at Xi'an Jiaotong University. His research interests include computer vision, pattern recognition, computational intelligence, image processing, and hardware implementation of intelligent systems. He served as the general chair for the International Symposium on Information Theory and Its Applications in 2002, and the general co-chair for the International Symposium on Nonlinear Theory and Its Applications in 2002. Since 2000, he has been the Chinese representative on the Governing Board of the International Association for Pattern Recognition. He presently serves as executive editor of *Chinese Science Bulletin.* He became a member of the Chinese Academy Engineering in 1999. He is a senior member of IEEE.

Heung-Yeung Shum received the PhD degree in robotics from the School of Computer Science, Carnegie Mellon University in 1996. He worked as a researcher for three years in the vision technology group at Microsoft Research Redmond. In 1999, he moved to Microsoft Research Asia where he is currently a senior researcher and the assistant managing director. His research interests include computer vision, computer graphics, human computer interaction, pattern recognition, statistical learning and robotics. He is the General Co-Chair of Ninth International Conference on Computer Vision (ICCV 2003 Beijing). He is a senior member of the IEEE.