Global Stereo Reconstruction under Second-Order Smoothness Priors

Oliver Woodford, Philip Torr, *Senior Member*, *IEEE*, Ian Reid, *Member*, *IEEE*, and Andrew Fitzgibbon, *Member*, *IEEE*

Abstract—Second-order priors on the smoothness of 3D surfaces are a better model of typical scenes than first-order priors. However, stereo reconstruction using global inference algorithms, such as graph cuts, has not been able to incorporate second-order priors because the triple cliques needed to express them yield intractable (nonsubmodular) optimization problems. This paper shows that inference with triple cliques can be effectively performed. Our optimization strategy is a development of recent extensions to α-expansion, based on the "QPBO" algorithm. The strategy is to repeatedly merge *proposal* depth maps using a novel extension of QPBO. Proposal depth maps can come from any source, for example, frontoparallel planes as in α-expansion, or indeed any existing stereo algorithm, with arbitrary parameter settings.

Index Terms—Stereo, second-order prior, discrete optimization, graph cuts.

1 Introduction

Dense stereo has made considerable progress in recent years, in part because the problem can be cast in an energy minimization framework for which there exist inference algorithms that can efficiently find good (if not always global) minima. Algorithms based on graph cuts, in particular, can incorporate *visibility reasoning* as well as *smoothness priors* into the estimation of depth maps. However, the smoothness priors used in graph-cuts-based estimates have to date been first-order priors, which favor low-curvature frontoparallel surfaces—indeed, the prior is maximized by frontoparallel planes. Even in man-made scenes, this is far from accurate, as illustrated in Fig. 1, and leads to inaccurate depth estimates. It has long been known [1], [2] that a second-order smoothness prior can better model the real world, but it has not yet been possible to combine visibility reasoning and second-order smoothness in an optimization framework which finds good optima.

The contributions discussed in this paper have been introduced in two previous works by the authors [4], [5], with a more in-depth discussion of these, further experimentation into objective energies and additional results introduced here. The main contribution is the development of an effective optimization strategy for stereo reconstruction with triple cliques. A further significant contribution is the development of an accurate asymmetrical occlusion model. This means that visibility reasoning and second-order priors can be combined for the first time in a powerful inference framework. We show that this algorithm produces excellent results both on the Middlebury test set [6] and on real-world examples with both planar *and* curved surfaces.

The areas of smoothness priors, visibility reasoning, and optimization are all relevant to this work, so we will review the relevant literature in these areas.

1.1 Smoothness Priors

The "order" of a smoothness prior is given by the order of the derivative of depth or disparity which the prior regularizes. First-order priors penalize nonzero first derivatives, encouraging frontoparallel depth maps, while second-order priors penalize nonzero second derivatives, encouraging planar depth maps. Second-order smoothness priors are not new in stereo. Indeed, Grimson [1] and Terzopoulos [2] both proposed a second-order prior for *surface reconstruction* in the early 1980s, in the form of the *thin plate* model, while Horn [7] and Gennert [8] used second-order priors in *dense stereo*[1] later that decade. However, they have all but disappeared from the more recent stereo literature, usurped by the now far more common first-order prior [10], [11], [12], [13], [14], [15].

A significant reason for this shift to first-order priors lies in the development of powerful optimization techniques such as graph cuts [10], [16], [17], [18] and belief propagation (BP) [12], [13]. The use of truncated linear and Potts model kernels has been commonplace with these methods. This stems in part from the fact that graph-cuts-based stereo algorithms [10], [15], [17] tend to use α-expansion, which cannot optimize convex kernels [10] (though Veksler [19] recently showed how this could be overcome).

Attempts have been made to model surfaces more accurately, while using the powerful graph cuts and BP optimizers, but these attempts have all employed pairwise

- *O. Woodford and I. Reid are with the Department of Engineering Science, University of Oxford, Parks Road, Oxford OX1 3PJ, United Kingdom. E-mail: {ojw, ian}@robots.ox.ac.uk.*
- *P. Torr is with the Department of Computing, Oxford Brookes University, Wheatley, Oxford OX33 1HX, United Kingdom. E-mail: philiptorr@brookes.ac.uk.*
- *A. Fitzgibbon is with Microsoft Research Ltd., 7 JJ Thomson Avenue, Cambridge, CB3 0FB, United Kingdom. E-mail: awf@microsoft.com.*

Manuscript received 5 Dec. 2008; revised 21 Apr. 2009; accepted 17 June 2009; published online 19 June 2009.
Recommended for acceptance by K. Boyer, M. Shah, and T. Syeda-Mahmood. For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMISI-2008-12-0838.
Digital Object Identifier no. 10.1109/TPAMI.2009.91.

1. While the aim of dense stereo and surface reconstruction is the same, the former computes data costs based on photoconsistency, while the latter computes data costs based on the distance from points (either dense or sparse) precomputed using stereopsis [9].

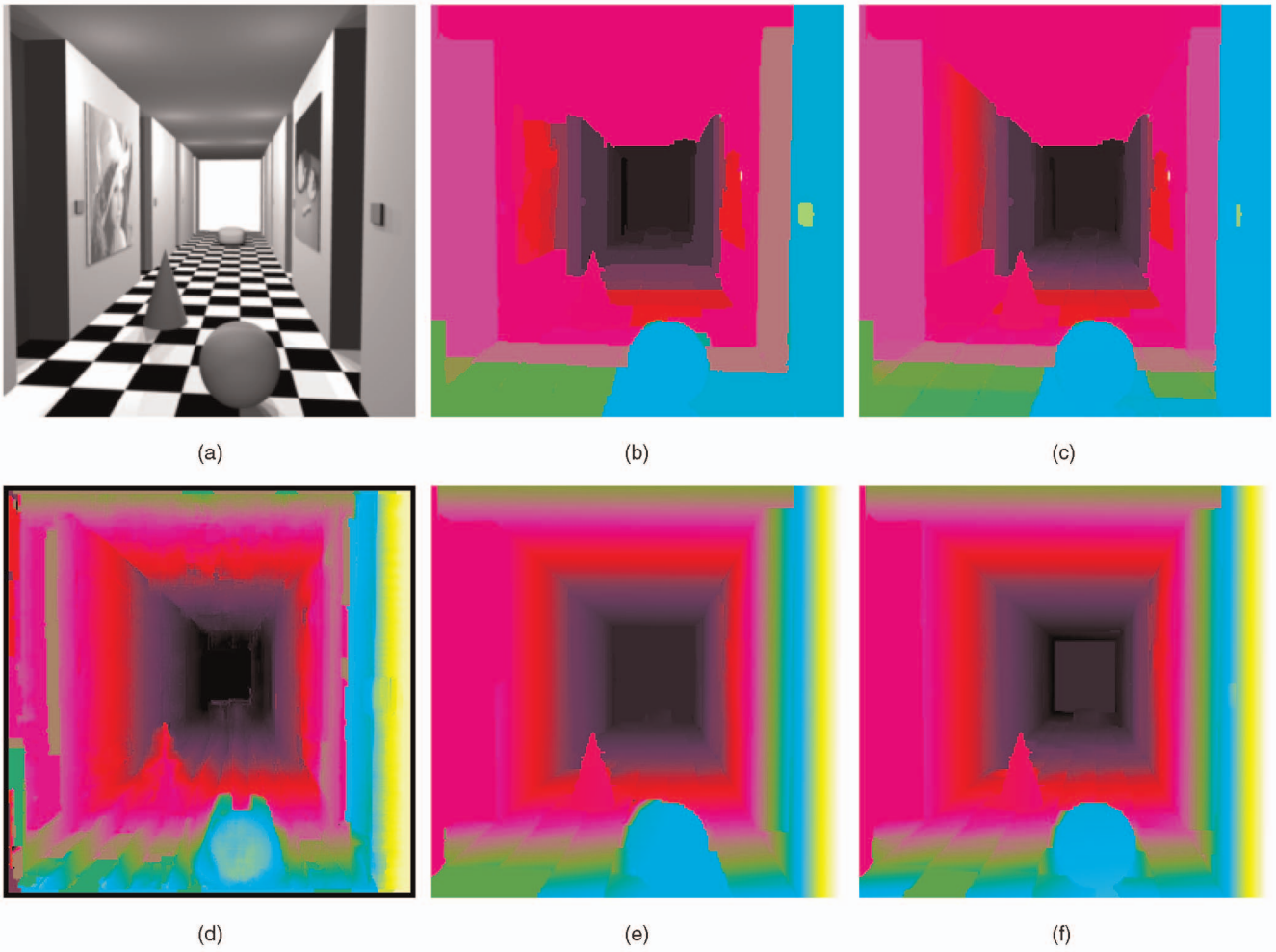

Fig. 1. The impact of different smoothness priors on stereo. (a) Reference image. (b) Li and Zucker's result [3]. (c) 1op, linear kernel. (d) 1op, quadratic kernel. (e) 2op, linear kernel. (f) 2op, quadratic kernel.

cliques rather than the higher order cliques required for a true second-order prior. Such attempts include *layered* [20], [21], [22] and *segment-based* [23], [24], [25], [26], [27], [28] approaches, which segment the reference image and enforce the constraint that such regions be planar. The pairwise regularization of the latter algorithms encourages neighboring regions to be coplanar, while the former algorithms achieve the same simply by iterating the segmentation and plane fitting processes. Li and Zucker [3] retain the pixel-based model while incorporating both second- and third-order priors, therefore merely encouraging planarity when there is ambiguity rather than enforcing it across entire regions. However, their algorithm precomputes local surface normals and in fact optimizes a first-order prior on the normals, which is an approximation to the true problem. The reason for the current absence of true higher order priors, despite their improved scene modeling capability, can be found in the literature. Li and Zucker [3] say, on using triple cliques, that

> such an endeavor quickly makes the problem computationally infeasible,

and Bykov and Veksler [29] comment, regarding graph cuts, that

> it is not clear if [triple cliques] can be used to encode a higher order smoothness.

Indeed, it has recently been shown [30] that a second-order smoothness prior generates nonsubmodular terms, precluding optimization using graph cuts.

Most recently, Bhusnurmath and Taylor [31] used an interior point method to find the optimal solution to an objective function that included a second-order prior, but this required that both data costs and smoothness costs be made convex, thus approximating the former and constraining the choice of kernels in the latter.

1.2 Visibility Reasoning

An important aspect of stereo frameworks, affecting the data likelihood term rather than the prior, is the use of an occlusion model to determine when a correspondence is not visible in an input image. The data term is typically based on the assumption that a point on a surface will be the same color when viewed from all angles. However, such a point may be visible in some views, but not in others, because it is occluded by another part of the scene, in which case the color sampled at the location of the correspondence will be that of an entirely different scene point. In this case, the assumption leads to the correct depth being given a low likelihood.

One solution to this problem is to consider these occluding samples to be outliers, and to construct a data likelihood distribution that models the outlier process [32] and hence

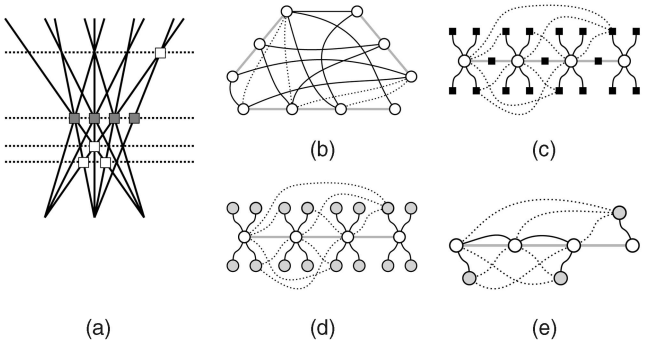

Fig. 2. (a) A simple stereo problem and (b) and (c) the resulting graph constructions for different approaches and (d) and (e) ours. See text for details of each.

provides a degree of robustness. A more successful approach to modeling occlusions has been to make explicit the visibility of each correspondence, and determine this either geometrically, by warping the depth map into each input view [11], [15], [17], [28], [33], [34], [35] or by constructing a generative color model of occluding surfaces [14]. Optimization of visibility is achieved either synchronously with depth [11], [15], [17], [28], [33], [34] or by iterating between optimizing visibility and optimizing depth [14], [35]. While the former approach is less prone to fall into local minima, the optimization techniques that can be used to successfully optimize the resulting objective energy are limited: Dynamic programming can be used [33], [34], but this enforces the ordering constraint, a constraint which is often violated in real scenes, especially those with thin foreground objects [35]; some formulations without the ordering constraint can be optimized using α-expansion graph cuts [11], [15], [17], [28], but it has been shown [36] that other optimizers, such as BP, do not perform well on these problems.

There are various forms of graph construction for visibility reasoning among the graph-cuts-based approaches, as shown in Fig. 2. Kolmogorov and Zabih give two different constructions, for two [17] and multiview [11] data sets respectively, but both approaches treat all images symmetrically, generating a node per input view pixel. This increases complexity significantly for multiple input views. A different, but also symmetrical construction is used in a segment-based framework in [28]. Wei and Quan [15] introduce an asymmetrical construction which generates a node per reference view pixel, considerably reducing complexity over symmetrical approaches. However, their construction generates both higher order cliques and nonsubmodular energies, which leads them to make approximations to avoid these cases. A contribution of our work, introduced in [4] and detailed more fully here, is to remove the higher order cliques with an equivalent pairwise graph construction and to use an optimizer which can handle nonsubmodular edges.

Modeling occlusions, and other outliers, in this principled way not only improves the quality of results, but has also been shown to reduce the need for a surface smoothness prior [15].

1.3 Optimization

The ability to optimize an objective function well is as important as the objective function itself in finding a good solution in stereo, a fact borne out not only by the high concentration of frameworks using graph cuts and BP at the top of the Middlebury stereo evaluation table,[2] but also by various comparisons of optimizers on the same stereo problems [36], [37].

While recent developments of BP have lead to continuous representations of the state space of variables [38], [39] and also linear[3] increases in computational cost with clique size for certain classes of clique functional [40] (including those based on derivatives), the fact remains that BP performs very poorly on the highly connected graphs that result from including geometrical occlusion reasoning in the stereo problem [36]. In contrast, graph cuts has been shown to handle occlusion reasoning well [11], [15], [17], [28], [36] and so is the preferred optimizer for this problem, but its use poses a number of problems. First, it is limited to binary (a.k.a. boolean) labeling problems, where each variable has only two possible values. Second, it can only solve problems whose objective function can be reparameterized as a sum of unary and pairwise terms; thus,

$$E(\mathbf{X}) = \sum_i \phi_i(X_i) + \sum_{i<j} \phi_{ij}(X_i, X_j), \quad X_i \in \{0,1\}\ \forall i, \quad (1)$$

with all pairwise terms satisfying the submodularity constraint [41]:

$$\phi_{ij}(0,0) + \phi_{ij}(1,1) \leq \phi_{ij}(0,1) + \phi_{ij}(1,0), \quad (2)$$

in which case the globally optimal solution is found. Nonsubmodular problems are those which cannot satisfy this constraint.

For optimization using graph cuts, higher order cliques must, therefore, be decomposed into a set of pairwise terms. Kolmogorov and Zabih [41] show how this is possible for triple cliques, noting specifically that for graph-cuts-based solutions the decomposition is valid if all pairwise projections of the variables are submodular.

Graph cuts has been extended to multilabel problems in two ways: A multilabel problem can be converted to an equivalent binary problem through a transformation of the graph [42], [43], then solved; alternatively, the multilabel problem can be solved through a sequence of graph cuts on pairs of labelings [10]. The latter approach is a specific example of a more general class of methods, recently given the name "fusion move" approaches [44], which solve problems through a sequence of binary optimizations (not necessarily graph-cuts based, e.g., [45]). The submodularity constraint poses constraints on the multilabel pairwise clique functionals, ϕ_{ij}, of both these solutions, requiring those of the former solution to be convex, while those of the latter to be either metric or semimetric [10] depending on whether the α-expansion or $\alpha\beta$-swap approach is used. Generally, this has limited applications to which graph cuts can be applied, predominantly to those where the value of any given label, α, is consistent across nodes, though there

2. http://vision.middlebury.edu/stereo/.

3. The time to compute update messages is $\mathcal{O}(NM^2)$, reduced from $\mathcal{O}(M^N)$, where M is the number of labels and N the clique size.

are some notable exceptions which generate submodular graphs with values for label α being inconsistent across the variables, e.g., [46].

Several approaches also exist for dealing with nonsubmodular problems. Truncation [47] is a method of converting a submodular binary problem to a nonsubmodular one which will not necessarily find the optimal solution, but which guarantees not to increase the cost of the solution. More recently, an extension of graph cuts [48], introduced to the field of Computer Vision as *Quadratic Pseudo-Boolean Optimization* (QPBO) by Kolmogorov and Rother [49], is able to optimize both submodular and nonsubmodular graphs optimally, though potentially only for a subset of nodes in the latter case (the rest of the pixels being unlabeled). Two extensions to QPBO which improve the output solution by labeling unlabeled nodes have been proposed [50], [51], and are considered in Section 3.2. Importantly, both truncation and QPBO approaches can be used in a fusion move framework to generate a convergent, multilabel optimizer. This has allowed graph-cuts-based multilabel optimization to be extended to both consistently labeled problems with nonsubmodular terms [47], [51], [52], [53] and problems with inconsistent labels [5], [44], [54].

Continuous valued variables can be optimized in a fusion move framework by optimizing over either a suitably large [44] or a judiciously chosen [5], [45], [54] discrete set of labelings.

1.4 Outline

In the next section we derive the objective function for our stereo problem, with the second-order prior. The following section details how the objective function is minimized, detailing the contributions made in this work. The penultimate section contains information on the experiments we performed in evaluating our proposed framework, before concluding.

2 Problem Statement

In this section we outline the objective function containing our second-order prior. All aspects of the objective function other than the prior are kept as standard as possible, as they are not the focus of this research.

The aim of dense stereo can be posed as that of finding the most likely scene model, given a suitable Bayesian posterior distribution—the *maximum a posteriori* (MAP) solution—where the scene model is $\{\mathrm{D}, \mathbf{I}_0^*\}$, D being the dense disparity map for the given reference view, $\mathbf{I}_0$, and $\mathbf{I}_0^*$ being the true, noiseless version of the same image.

As input we are given a set of $N+1$ images, $\{\mathbf{I}_i\}_{i=0}^N$. A 2D vector, $\mathbf{x}$, denotes a pixel location in the reference view, the color of which is written as $I_0(\mathbf{x})$, and the corresponding disparity is $D(\mathbf{x})$. We are also given projection functions $\{\pi_i(\mathbf{x}, d) : \mathbb{R}^2 \mapsto \mathbb{R}^2\}_{i=1}^N$, where $\pi_i(\mathbf{x}, d)$ is the projection into the ith image of the 3D point corresponding to disparity (1/depth) d in front of pixel $\mathbf{x}$ in the reference view. For a rectified stereo pair, $N = 1$ and only π_1 is required, with the simple definition $\pi_1(\mathbf{x}, d) = \mathbf{x} + [d, 0]$. The abbreviation $I_i^\pi(\mathbf{x}, d) = I_i(\pi_i(\mathbf{x}, d))$ will be used to reduce clutter and may be read as "the color of the pixel corresponding to $\mathbf{x}$ in image i if the disparity at $\mathbf{x}$ is d."

The posterior distribution can therefore be written as

$$p(\mathrm{D}, \mathbf{I}_0^* | \mathbf{I}_0, \ldots, \mathbf{I}_N) = \frac{p(\mathbf{I}_0, \ldots, \mathbf{I}_N | \mathrm{D}, \mathbf{I}_0^*) p(\mathrm{D}, \mathbf{I}_0^*)}{p(\mathbf{I}_0, \ldots, \mathbf{I}_N)}, \quad (3)$$

where $p(\mathbf{I}_0, \ldots, \mathbf{I}_N | \mathrm{D}, \mathbf{I}_0^*)$ is the data likelihood term which models noise from the camera sensor, and $p(\mathrm{D}, \mathbf{I}_0^*)$ is the prior probability of the output variables. $p(\mathbf{I}_0, \ldots, \mathbf{I}_N)$ is the prior probability of the input data, but, being a constant, it can be ignored in the optimization. Note that it has been assumed that the input projection functions, $\{\pi_i\}_{i=1}^N$, are noiseless; hence, they do not appear in the above formulation. A more advanced approach is to combine the estimation of D and $\{\pi_i\}_{i=1}^N$ into a single framework, e.g., [55].

While generating $\mathbf{I}_0^*$ as well as D is the correct way of accounting for sensor noise, this approach is used only rarely, e.g., [14]. The simpler (in both data likelihood and prior terms) and more common approach is to assume that $\mathbf{I}_0^* = \mathbf{I}_0$, giving the following approximate posterior:

$$p(\mathrm{D} | \mathbf{I}_0, \ldots, \mathbf{I}_N) \propto p(\mathbf{I}_1, \ldots, \mathbf{I}_N | \mathrm{D}, \mathbf{I}_0) p(\mathrm{D} | \mathbf{I}_0). \quad (4)$$

Rather than maximize the posterior probability we seek to minimize its negative log, called the *energy*. While equivalent, this can simplify the problem by removing exponentials, e.g., in Gaussian noise models, and turning products into summations. The energy can be written as

$$E(\mathrm{D} | \mathbf{I}_0, \ldots, \mathbf{I}_N) = \underbrace{E_{\mathrm{photo}}(\mathbf{I}_1, \ldots, \mathbf{I}_N | \mathrm{D}, \mathbf{I}_0)}_{\text{data likelihood}} + \underbrace{E_{\mathrm{smooth}}(\mathrm{D} | \mathbf{I}_0)}_{\text{smoothness prior}}. \quad (5)$$

The components of the energy will now be described.

2.1 Data Likelihood

The data likelihood term is constructed based on the assumptions that $\mathbf{I}_0$ and $\{\pi_i\}_{i=1}^N$ are noiseless and that the other input images are corrupted with i.i.d. noise. It can therefore be evaluated independently over each of the pixels in $\{\mathbf{I}_i\}_{i=1}^N$, and summed. An approximation adopted by most stereo algorithms, and indeed here, is to sum over the reference image pixels,[4] as this suits the form of scene model better. Our data likelihood term is therefore written as

$$E_{\mathrm{photo}}(\mathbf{I}_1, \ldots, \mathbf{I}_N | \mathrm{D}, \mathbf{I}_0) = \sum_{\mathbf{x} \in \mathcal{X}} \sum_{i=1}^{N} f\big(I_i^\pi(\mathbf{x}, D(\mathbf{x})) - I_0(\mathbf{x}), V_{\mathbf{x}}^i\big), \quad (6)$$

where $\mathcal{X}$ are the set of reference image pixels and $V_{\mathbf{x}}^i$ is a *visibility flag*, to be discussed below, indicating whether the 3D point defined by $(\mathbf{x}, D(\mathbf{x}))$ is visible in $\mathbf{I}_i$. Given $V_{\mathbf{x}}^i$, the consistency metric f is defined as

$$f(\Delta I, V) = \begin{cases} \rho_{\mathrm{d}}(\Delta I), & \text{if } V = 1, \\ \nu, & \text{if } V = 0. \end{cases} \quad (7)$$

When an input sample is occluded in the reference view by another part of the scene there is assumed to be no dependence between the colors of the input sample and

4. It is known that this approach miscounts the contribution of each input pixel to the overall probability [56], leading to errors in wide baseline situations. We employ it for simplicity, and stick to narrow baseline sequences.

its projection into the reference image. Rather, $I_i^{\pi}(\mathbf{x}, D(\mathbf{x}))$ is assumed to be drawn from a uniform distribution, generating a constant penalty cost, ν, which is paid by occluded pixels. When an input sample is visible in the reference view its likelihood is computed from the noise model, ρ_{d}, assuming Lambertian reflectance of the surface being viewed. The value of ν needs to be greater than the largest possible value of $\rho_{\mathrm{d}}(\Delta I)$ in order to avoid encouraging self-occlusions, but, aside from this, the noise model could take any form. Our noise model is based on a contaminated Gaussian [32], and is therefore a robust measure of color difference defined by

$$\rho_{\mathrm{d}}(\Delta I) = -\log\left(1 + \exp(-\|\Delta I\|^2/\sigma_{\mathrm{d}})\right), \tag{8}$$

where σ_{d} is set from the noise level in the sequence. The measure doesn't incorporate any means of accounting for sampling issues, e.g., [57], [58], but it could.

The value of $V_{\mathbf{x}}^i$ is computed using the asymmetrical occlusion model of Wei and Quan [15]—if there is another reference view pixel, $\mathbf{p}$, which projects to the same point[5] in $\mathbf{I}_i$ as pixel $\mathbf{x}$, and for which the projected depth is less than that of $\mathbf{x}$ then $V_{\mathbf{x}}^i = 0$, otherwise $V_{\mathbf{x}}^i = 1$. $V_{\mathbf{x}}^i$ adds nonlocal terms to the energy, making optimization of this energy difficult, even before priors are incorporated. It is, therefore, more correctly written $V_{\mathbf{x}}^i(\mathrm{D})$, indicating the dependence on many entries of the disparity map D.

2.2 Surface Smoothness

The smoothness prior regularizes the disparity map by placing a cost on unlikely geometry. Following the standard stereo prior approach, it is assumed that the prior likelihood of disparity values are only dependent on those of their close neighbors, creating a neighborhood, $\mathcal{N}$. The *Hammersley-Clifford theorem* [59], states that the energy can be decomposed into a sum of functionals over the cliques defined by the set of neighborhoods, $\mathbb{N}$, generating a Markov Random Field. E_{smooth} is, therefore, written as

$$E_{\mathrm{smooth}}(\mathrm{D}|\mathbf{I}_0) = \sum_{\mathcal{N}\in\mathrm{N}} W(\mathcal{N})\rho_{\mathrm{s}}(S(\mathcal{N}, \mathrm{D})). \tag{9}$$

The first term, $W(\mathcal{N})$, modulates the smoothness term according to some function of the reference image, conditioning smoothness on $\mathbf{I}_0$ and making this formulation a Conditional Random Field (CRF). This term is discussed further below. The function $S : \mathbb{R}^{|\mathcal{N}|} \mapsto \mathbb{R}$ is generally a derivative of disparity. The commonly used first derivative is given by

$$S(\{\mathbf{p}, \mathbf{q}\}, \mathrm{D}) = D(\mathbf{p}) - D(\mathbf{q}), \tag{10}$$

$\mathbb{N}$ being the set of all 2×1 and 1×2 patches in the image. This derivative permits frontoparallel surfaces without penalty.

We do not make the assumption that surfaces in the scene are generally frontoparallel (impossible when one considers a shift of reference viewpoint immediately changes the assumption). Instead, we wish to permit all *planar* surfaces without penalty, which can be achieved by using the second derivative of disparity [8]. The full second derivative consists of the derivatives d_{xx}, d_{xy}, and d_{yy}. While numerical computation of the derivatives d_{xx} and d_{yy} leads to triple cliques, d_{xy} leads to a quadruple clique and is therefore ignored, with the effect that the larger class of all harmonic functions is unpenalized [8]. We therefore define our second-order prior as

$$S(\{\mathbf{p}, \mathbf{q}, \mathbf{r}\}, \mathrm{D}) = D(\mathbf{p}) - 2D(\mathbf{q}) + D(\mathbf{r}), \tag{11}$$

where the neighborhoods, $\mathcal{N} = \{\mathbf{p}, \mathbf{q}, \mathbf{r}\}$, are from the set of all 3×1 and 1×3 patches in the reference image.

The kernel placed on the derivative response is generally given by

$$\rho_{\mathrm{s}}(s) = \sigma_{\mathrm{s}} \cdot \left(\min\left(\frac{|s|}{\sigma_{\mathrm{s}}}, 1\right)\right)^{\gamma}, \tag{12}$$

where $\gamma = 1$ or 2 and σ_{s} is a discontinuity preserving threshold, creating the truncated linear and truncated quadratic kernels respectively.

2.3 CRF Weights

The CRF weights $W(\cdot)$ are set to encourage disparity edges to align with edges in the reference image, $\mathbf{I}_0$. The commonest image feature used in $W(\cdot)$ is the magnitude of the local image gradient of that neighborhood, parallel to the neighborhood [10], [11], [14], [21], [60], [61], [62]. The form of $W(\cdot)$ is generally handpicked, though Scharstein and Pal [62] learn maximum likelihood weights for a range of image gradient magnitudes, given a Potts smoothness model, using a gradient ascent approach. Other image features have also been used, such as the output of the Canny edge detector [63] or an image oversegmentation [12]. The latter model's smoothness constraint is strengthened if the pixels in $\mathcal{N}$ are part of the same segment, encouraging discontinuities to align with segment boundaries. Note that this contrasts with the segment-based stereo methods [23], [24], [25], [26], [27], [28], which *force* discontinuities to align with segment boundaries.

We use the oversegmentation approach, generating a single segmentation of the reference image (we use mean-shift segmentation [64], $h_s = 4$ and $h_r = 5$), and assign one of two weights to each neighborhood, depending on whether or not it overlaps a segmentation boundary. Precisely, if L is the map which assigns to each pixel its segmentation label, then

$$W(\mathcal{N}) = \begin{cases} \lambda_{\mathrm{h}}, & \text{if } L(\mathbf{p}) = L(\mathbf{q}) \;\forall\; \mathbf{p},\ \mathbf{q} \in \mathcal{N}, \\ \lambda_{\mathrm{l}}, & \text{otherwise}, \end{cases} \tag{13}$$

where $\lambda_{\mathrm{h}} > \lambda_{\mathrm{l}}$ (exact values given in Appendix B, which can be found in the IEEE Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2009.91.) to discourage disparity edges from cutting through segments, where they are less likely. However, our optimization framework permits any form of CRF weight to be used.

5. We define "same point" to mean within half a pixel in both horizontal and vertical directions. This measure is an approximation, as a pixel's projected footprint will vary according to its position and disparity. While a more accurate definition could be employed, this one was found to work suitably well. More details on evaluating $V_{\mathbf{x}}^i$ are given in Appendix B, which can be found in the IEEE Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2009.91.

3 Optimization

The above defines $E(\mathrm{D}|\mathbf{I}_0,\ldots,\mathbf{I}_N)$ as a function of a real-valued disparity image D. In this section, we describe how we solve the following optimization problem:

$$\mathrm{D} = \underset{\mathrm{D}}{\operatorname{argmin}}\, E(\mathrm{D}|\mathbf{I}_0,\ldots,\mathbf{I}_N). \tag{14}$$

In order to optimize the energy over the real-valued space we follow the fusion move approach, reducing it to a sequence of binary problems as follows: Suppose we have a current estimate of the disparity, D_t, and a *proposal* depth map D^{p}. In the α-expansion method, for example, the proposal depth at each step is a frontoparallel plane [10]; in this paper, we shall use more complex proposals (see Section 3.3). The goal is to optimally combine ("fuse") the proposal and current depth maps to generate a new depth map D_{t+1} for which the energy $E(\mathrm{D}_{t+1}|\mathbf{I}_0,\ldots,\mathbf{I}_N)$ is lower than D_t. This fusion move is achieved by taking each pixel in D_{t+1} from one of $(\mathrm{D}_t, \mathrm{D}^{\mathrm{p}})$, as controlled by a binary indicator image B with elements $B(\mathbf{x})$:

$$\mathrm{D}^b(\mathrm{B}) = (1-\mathrm{B})\cdot \mathrm{D}_t + \mathrm{B}\cdot \mathrm{D}^{\mathrm{p}}, \tag{15}$$

where dot indicates elementwise multiplication. Thus, B may be read as "copy the disparity from the proposal $\mathrm{D}^{\mathrm{p}}(\mathbf{p})$ if $B(\mathbf{p}) = 1$, otherwise keep the current estimate D_t." Then, the energy $E(\mathrm{D}|\mathbf{I}_0,\ldots,\mathbf{I}_N)$ is a function only of the indicator image B, so we may define a boolean optimization problem

$$\mathrm{D}_{t+1} = \mathrm{D}^b(\mathrm{B}^*), \quad \text{for } \mathrm{B}^* = \underset{\mathrm{B}}{\operatorname{argmin}}\, E(\mathrm{D}^b(\mathrm{B})|\mathbf{I}_0,\ldots,\mathbf{I}_N). \tag{16}$$

In the next three sections, we discuss three important contributions which enable and improve the algorithm: 1) the pairwise graph construction which allows each binary subproblem to be effectively solved using a graph-cuts-based optimizer; 2) a variety of alternative fusion methods for finding the best possible B^*; and 3) the selection of proposal depth maps.

3.1 Graph Construction

The challenges to creating a binary pairwise graph to solve (16) are twofold: the existence of nonsubmodular triple cliques [30] from the smoothness term and including the geometrical occlusion model.

We tackle the problem of triple cliques using the decomposition of [41]. Each triple clique is decomposed into unary and pairwise terms via the addition of a latent variable, attached to each of the three variables in the clique but to no others in the graph. This idea was originally proposed only for submodular energies in the case of minimization via graph cuts. However, the smoothness cost of (12) can generate nonsubmodular triple cliques, as shown in Appendix A, which can be found in the IEEE Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2009.91. Fortunately, the decomposition remains valid for nonsubmodular energies (also shown in Appendix A), though some of the resulting pairwise cliques are nonsubmodular, ruling out graph cuts to optimize the energy.

To demonstrate the graph construction required to model visibility, it is instructive to have a simple example of the problem given by (16). Such a problem is given in Fig. 2a. The problem consists of a four pixel, 1D reference image (center) and two other input images of three pixels. The current disparity map, D_t, is indicated by the white squares, and a proposal disparity map, D^{p} (in this case frontoparallel), is indicated by the gray squares. All of the potential occlusion interactions for the binary problem can be computed prior to constructing the graph, conferring the advantage (over non-fusion-based optimization strategies) that the number of potentially occluding pixels is relatively small for each such subproblem, so the cost of including visibility is relatively low.

Fig. 2b shows the submodular pairwise graph generated by the approach of [11]; solid gray lines indicate pairwise cliques resulting from the smoothness prior (assumed to be first order across the figure to simplify matters), black lines indicate data likelihood terms and dashed lines represent terms resulting from visibility reasoning. Images are treated symmetrically, and a graph node is generated for every pixel in every image, leading to highly complex graphs, even for just a few input images. The asymmetrical model of [15], shown in Fig. 2c in the form of a factor graph, has nodes only for each pixel in the reference view. However, the construction has two drawbacks: First, it generates nonsubmodular edges; second, it can potentially generate cliques larger than size three (as shown). Approximations to the objective energy were made in [15] to overcome these problems; here, we show how these approximations can be avoided by improving the graph construction and optimizing the nonsubmodular energy as it stands.

To address the second of these problems, we introduce the graph construction shown in Fig. 2d (first described by us in [4]), which reduces the maximum clique size of the asymmetrical model to two. The simplification comes from adding nodes which explicitly represent the binary visibility variables (shown as gray circles). In fact, two nodes are required per visibility variable, $V^i_{\mathbf{x}}$, one for $B(\mathbf{x}) = 1$ and one for $B(\mathbf{x}) = 0$, denoted $V^{i0}_{\mathbf{x}}$ and $V^{i1}_{\mathbf{x}}$, respectively. Pairwise edges between the visibility nodes and their associated disparity nodes (black lines) are then used to model the data costs of (7) as follows:

$$\phi\big(B(\mathbf{x}), V^{i\alpha}_{\mathbf{x}}\big) = \begin{cases} 0, & \text{if } B(\mathbf{x}) \neq \alpha, \\ \rho_{\mathrm{d}}(\Delta I), & \text{else if } V^{i\alpha}_{\mathbf{x}} = 1, \\ \nu, & \text{otherwise.} \end{cases} \tag{17}$$

Given a node $\mathbf{p}$ which occludes the input sample of $V^{i0}_{\mathbf{x}}$ when $B(\mathbf{p}) = 1$, the visibility node can be set to 0 (i.e., occluded) in this case by creating a pairwise edge for which $\phi(B(\mathbf{p}) = 1, V^{i0}_{\mathbf{x}} = 1) = \infty$ and all other states carry zero cost.[6] The visibility node will be set to 0 when one or more nodes occlude the associated input sample in order to avoid the higher infinite cost(s), producing the desired effect. This construction is very similar to the higher order Potts model construction of [65]. Note that, in the example given, the occlusion edge is nonsubmodular. These infinite cost occlusion edges generate the dashed lines in the graph. While the construction may seem to add complexity by introducing $2nN$ nodes, where $n = |\mathcal{X}|$, the visibility nodes are in fact only necessary when two or more pixels potentially occlude an input sample; the data costs for

6. In fact, in order to ensure the correct cost is paid for occluded pixels it is sufficient that $\phi(B(\mathbf{p}) = 1, V^{i0}_{\mathbf{x}} = 1) \geq \nu - \rho_{\mathrm{d}}(I^{\pi}_i(\mathbf{x}, D^{\mathrm{p}}(\mathbf{x})) - I_0(\mathbf{x}))$.

input samples of pixel $\mathbf{x}$ that cannot be occluded can be incorporated into the unary costs of $B(\mathbf{x})$, while those that can only be occluded by a single pixel, say $\mathbf{p}$, can be incorporated into the edge $(B(\mathbf{x}), B(\mathbf{p}))$. This simplifies the graph greatly, generating the construction shown in Fig. 2e.

The final graph construction is a combination of the data and occlusion edges from Fig. 2e, which implement the data costs of (6), and the triple clique decomposition which is used to model the second-order smoothness costs of (9). A first-order smoothness prior uses $2n$ edges and generates four incident edges per pixel node, ignoring boundary effects. Our second-order prior uses $10n$ edges, nominally six edges per clique, two of which are shared with neighboring cliques making an average of five edges per clique, and generates 14 incident edges per pixel node. While a second-order prior therefore generates a substantial increase in complexity, the problem remains tractable.

3.2 Fusion Strategies

The section above described the binary pairwise graph constructed to solve the optimization problem of (16). This graph contains some nonsubmodular edges in both the smoothness edges and the data and occlusion edges, precluding optimization using the standard graph-cuts algorithm. As a result, we use QPBO to solve the graph. Unlike the submodular case, where the globally optimal B is guaranteed, QPBO returns a solution B and an associated mask M with the guarantee that at pixels $\mathbf{x}$ where $M(\mathbf{x}) = 1$, the value $B(\mathbf{x})$ is at the value it would have at the global minimum,[7] but pixels where $M(\mathbf{x}) = 0$ have "unlabeled" values. These unlabeled nodes must be set to 0 or 1 in a postprocessing step in order to generate D_{t+1} using (16), and furthermore, in order to ensure convergence of the algorithm, the set labels should ensure that $E(\mathrm{D}_{t+1}|\mathbf{I}_1, \ldots, \mathbf{I}_N) \leq E(\mathrm{D}_t|\mathbf{I}_1, \ldots, \mathbf{I}_N)$. This section discusses the ways in which this label fixing can be done. Note that, while there are many more nodes (e.g., visibility nodes) involved in the QPBO optimization, the aim is only to find the values of B, i.e., the labels for the nodes corresponding to pixel disparities, so these are the only nodes whose labels are fixed. When computing the energy of a labeling, the values of the visibility variables can be computed directly from the disparity.

Several approaches to fixing labels have already been proposed in the literature:

QPBO-F. *Fix to current* [4]: fix unlabeled nodes to 0, the current best labeling (D_t).

QPBO-L. *Lowest energy label* [44]: fix unlabeled nodes collectively to whichever of 0 or 1 gives the lower energy.

QPBOP. *Probe*: probe the graph, as described in [50], [51], in order to find the labels of more nodes, that form part of an optimal solution.

QPBOI-F. *Fix to current and improve*: fix unlabeled nodes to 0, and transform this labeling using QPBOI [51].

We introduce two new approaches to label fixing which are based on the optimal splice technique of [45]. That technique split the two labelings into independent regions, and independently selected the label, 0 or 1, which gave the lower energy for each region. The unlabeled nodes of B can similarly be split into independent regions, growing each region from a seed unlabeled node by adding all nodes that share a clique with the seed node, then repeating the process for all new unlabeled nodes in the region, and so on. Given an ordered list of cliques containing ordered node indexes this process can be achieved in $\mathcal{O}(|\mathrm{B}|)$ (i.e., linear) time. A looser constraint than regions being independent is regions being *strongly connected*. Nodes in two different strongly connected regions (SCRs) can share a clique, but the dependence between the two regions can only be unidirectional; in practice (in this application), SCRs are almost always independent. This is relevant because the SCRs can be computed in $\mathcal{O}(|\mathrm{B}|)$ time [66] *without* the ordering preprocess, making them an efficient approximation to independent regions. The two new approaches to fixing labels are therefore:

QPBO-R. *Lowest cost label per region*: split unlabeled nodes into SCRs, as per [66]. For each SCR, independently select the labeling, 0 or 1, which gives the lower total energy for cliques connected to that region.

QPBOI-R. *Improve lowest cost label per region*: Label nodes as per QPBO-R, then use QPBOI to transform this labeling.

All of the described methods ensure convergence because they are all guaranteed[8] to have an energy equal to or lower than the output of QPBO-F, which is itself guaranteed not to increase the energy as a result of the "autarky" property of QPBO [51, p. 2]. In Section 4.2, we empirically compare the various fusion strategies in the context of our problem.

3.3 Proposal Generation

The final component of the algorithm to be defined is the choice of proposals. In previous work [4], [11], [15], the proposals have just been frontoparallel planes (denoted "SameUni" below). As shown in [10], repeated fusion of these proposals leads to a strong local optimum in the case of a first-order prior. In the case of a second-order prior, the nature of these proposal disparity maps has a much larger effect on the generated disparity map, as we show empirically in Section 4. We use the following schemes for generating the jth proposal disparity map $\mathrm{D}_j^{\mathrm{p}}$:

SameUni. Draw d_j from a uniform distribution, and set $D_j^{\mathrm{p}}(\mathbf{x}) = d_j$ for all $\mathbf{x}$.

SegPln. Proposals are piecewise-planar disparity maps generated using the ad hoc approach of segmentation-based methods [26], [27], with a number of segmentation algorithms and parameters used to generate a wide range of sizes of planar region. Further details are given in Appendix B, which can be found in the IEEE Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2009.91.

Smooth. $D_j^{\mathrm{p}}(\mathbf{x}) = (D_j(\mathbf{x} + \Delta) + D_j(\mathbf{x} - \Delta))/2$, where $\Delta = [0, 1]$ when j is odd and $\Delta = [1, 0]$ when j is even. The set is prefixed with the two disparity maps generated from the fusion of the other proposal sets, and this set is repeated every six iterations.

These proposal methods represent the different approaches used by the main types of stereo algorithms: The frontoparallel proposals of SameUni are essentially those used at each iteration of an α-expansion-based stereo

7. More correctly, *a* global optimum, as there may be several labelings with the same energy.

8. In fact, the approximation of independent regions with SCRs used here can theoretically lead to an increase in energy, but only a small one, and this is extremely rare.

algorithm (except drawn from a continuous, rather than discrete, space), SegPln proposals are those used by segment-based algorithms, and Smooth proposals, generated by a smoothing operation on the current disparity map, can be viewed as a proxy for local methods such as gradient descent. With QPBO-based fusion, we gain the benefits of all these algorithms—indeed, any stereo algorithm available—without affecting the global optimum. For example, the SegPln proposals, the main workhorse of our algorithm, are produced with a range of algorithms and parameter settings; in general, we expect these disparity maps to be correct in some parts of the image, and for some parameter settings, but that no settings can be found for which any algorithm works best. By fusing the proposals in a well-defined energy minimization framework, the parameter sensitivity of these methods is turned into an advantage: we can select the best parts from each proposal, at the pixel (as opposed to segment) level.

Some further implementation notes, which will allow the reader to replicate our method more accurately, are available in Appendix B, which can be found in the IEEE Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2009.91.

4 Experiments

In this section, we describe the experiments carried out to evaluate the efficacy of QPBO in optimizing our non-submodular energy, the trade-offs of each of the QPBO labeling methods, the effect of using different disparity proposals, and to compare our method, with its second-order prior, to the same method with a first-order prior, and other, competing approaches to stereo.

The optimization method used in each experiment is characterized by the order of the prior ("1op" for first-order prior, etc.), the smoothness kernel ("linear" for $\gamma = 1$, "quadratic" for $\gamma = 2$), the set of proposals, and the fusion strategy, e.g., "2op, linear, SameUni, QPBOI-R," or "1op, quadratic, SegPln, QPBOP."

4.1 Number of Unlabeled Nodes

The first experiment was to determine whether optimization of the binary, pairwise, nonsubmodular graph described in Section 3.1 (an NP-hard problem) was feasible using QPBO. The proportion of pixels that are labeled by QPBO has a direct impact on the quality of the solution found—trivially, if no nodes are labeled then (using QPBO-F) the final solution will be the same as the initial solution. It is, therefore, important to have as many nodes labeled as possible.

We used QPBO-F in these experiments, but varied the proposal schemes, order of prior and prior kernel to see what effect these had on the number of unlabeled nodes. The experiments were carried out on both the Teddy and Cones sequences of the Middlebury evaluation framework, and results were averaged across both sequences and the binary optimizations within each category.

Table 1 shows the results of these experiments using the default objective function parameters.[9] For both first-order priors, the vast majority of pixels are labeled optimally, more so with the SameUni proposals and fewest with the Smooth proposals. The second-order priors generate more unlabeled nodes, the truncated quadratic kernel performing markedly worse in this respect, with the SegPln proposals generating the most unlabeled nodes. This potentially makes good optimization difficult.

TABLE 1
Mean Number of Unlabeled Nodes Per Iteration (Percent), Averaged over the Teddy and Cones Sequences, for the Various Priors and Proposal Schemes, Using QPBO-F

	SameUni	SegPln	Smooth
1op, linear	0.52	1.3	4.0
1op, quadratic	0.31	2.0	3.5
2op, linear	1.1	21	4.8
2op, quadratic	3.8	62	30

4.2 Comparison of Fusion Strategies

With a relatively high number of unlabeled nodes when using a second-order prior, it is clearly important to try to fix them as effectively as possible. In the following experiments, we tested the six post-QPBO labeling strategies described in Section 3.2. We used 2op, linear, and SegPln settings as these give a level of unlabeled nodes that is high, but not prohibitively so (in the case of the more costly strategies).

The first experiment involved trying all of the fusion strategies at each iteration, on exactly the same binary optimization (the optimal labeling given by QPBOP was used to update D), and this was carried out on the Middlebury Teddy, Cones, and Cloth3 sequences and the results concatenated. The speed of the fusion strategy is important. Fig. 3a shows that QPBOP rapidly becomes several orders of magnitude slower as the number of unlabeled pixels rises, while other methods show a more modest increase over the same range; of these there is only a fractional difference in speed, though order of fastest to slowest is consistently QPBO-F, QPBO-L, QPBO-R, QPBOI-F, QPBOI-R. Also important is the energy reduction performance of each strategy—QPBOP, which gives an optimal solution, performs best, while QPBO-F, with the simplest labeling strategy, is guaranteed to perform worst. Fig. 3b shows how the other strategies perform relative to these two, by normalizing the energy reduction between 0, representing the performance of QPBO-F, and 1, representing the performance of QPBOP. The normalized energy reduction of the four remaining strategies were discretized into 20 equally sized bins, which (except the bin for 0-0.05) are shown in the stacked bar graph of Fig. 3. The graph indicates that QPBOI-R achieves the largest energy reduction after QPBOP, based on it having the largest mass toward the right of the graph.

The second experiment tested the performance of each strategy over an entire iterative optimization, by running them individually until convergence on the same set of proposals. Table 2 shows the quantitative results of this experiment on the Teddy sequence. In terms of performance, QPBOI-R registers the lowest energy after QPBOP, inline with the previous experiment. In terms of time per fusion, QPBOP is the slowest method by two orders of

9. Additional results showing the effect of the weight of the prior on the number of unlabeled pixels are given in Appendix C, which can be found in the IEEE Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2009.91.

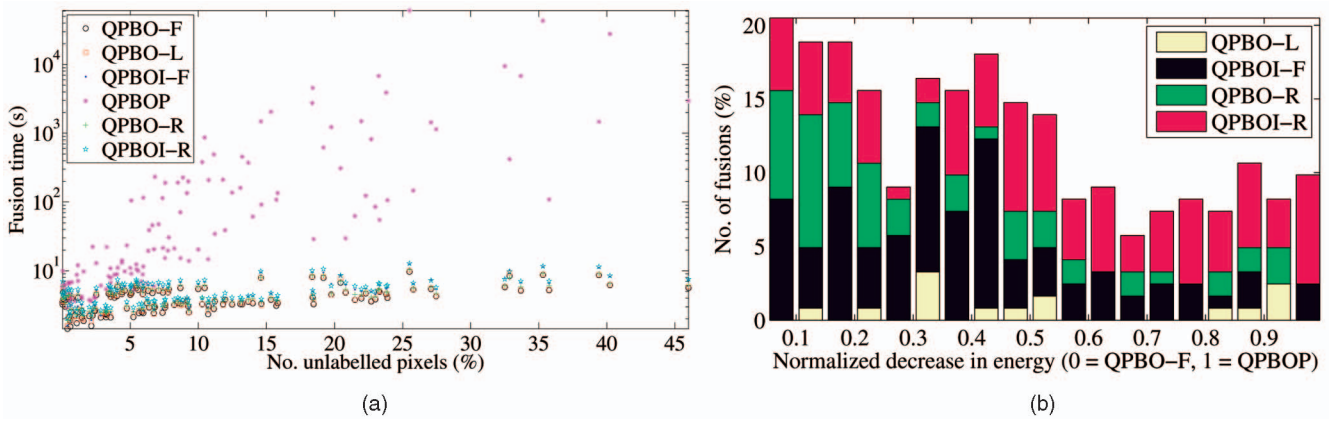

Fig. 3. Effect of fusion strategies on (a) time and (b) energy.

magnitude, but converges in the smallest number of iterations and with the lowest average number of unlabeled nodes. The QPBOI methods are slower than the remaining methods, due to their costly graph resolving; however, QPBOI-F is more than twice as slow as QPBOI-R, which was not predicated by the results of the previous experiment. This is most likely due to its larger number of unlabeled nodes per fusion, which has a linear effect on the time taken to fix unlabeled nodes—there is a trend for the methods which fix nodes better (i.e., generate lower energies) to also generate fewer unlabeled nodes in successive iterations. QPBO-R is competitive with the other non-QPBOI methods in terms of speed, while outputting a lower energy.

Considering the trade-off between time and efficacy, we felt QPBOI-R to be the most suitable method for our problem, and used this in all further experiments save those involving a 2op quadratic prior—the potentially high number of unlabeled pixels involved in the latter optimizations can make the QPBOI method prohibitively expensive also, so we used QPBO-R in this case instead.

4.3 Proposals

In Section 3.3 we introduced three classes of proposal. Fig. 4 demonstrates the effect of using these proposals on the Venus sequence, under the various smoothness priors. The number in the bottom left corner of each image is the number of fusion iterations used to generate that image (when the convergence criterion was met). From these we see that many more SameUni proposals (drawn from an infinite set) are required before convergence, compared to the other approaches, making this a slower approach. We also see that the output from the SameUni proposals is always piecewise frontoparallel, regardless of the prior used, in spite of the fact that the lower energy final output of the Smooth proposals (which incorporates the other two outputs) are only piecewise frontoparallel with the first-order linear prior. Since the disparity converges on a local minimum with respect to the fusion moves, the moves themselves must create a convergence basin which includes only piecewise-frontoparallel solutions. Indeed, this can be seen to be the case when one considers the simple problem of Fig. 5—given a frontoparallel current solution and planar optimal solution, any intermediate solutions increase the E_{smooth} cost, thereby creating an energy hump which cannot be overcome by any single fusion. This suggests that the SameUni proposals are only suitable for use with a first-order linear prior.

The output from SegPln proposals, with their planar segments which contain many small changes in disparity instead of a few large ones, is forced to be as frontoparallel

TABLE 2
Results of the Various Fusion Strategies Applied to the Teddy Sequence Using "2op, linear, SegPln, QPBOI-R"

	QPBO-F	QPBO-L	QPBOI-F	QPBOP	QPBO-R	QPBOI-R
Energy (% > QPBOP)	1.13	0.666	0.374	0	0.571	0.279
Fusion time (avg. secs.)	8.33	8.24	50.9	1680	9.77	17.3
No. iterations	42	44	45	37	42	42
Unlabelled (avg. %)	27.8	12.8	16.1	11.0	12.8	13.1

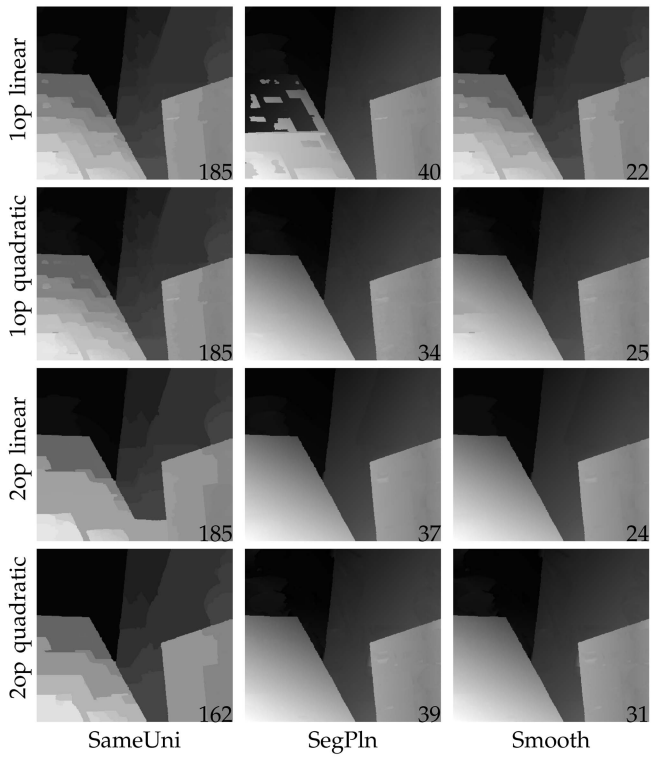

Fig. 4. Effect of proposals and the form of prior.

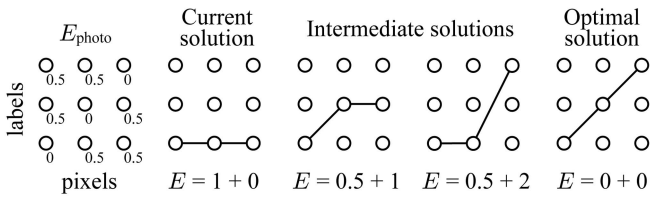

Fig. 5. The frontoparallel proposal energy hump.

as possible by the first-order linear prior, generating an output that is far from accurate. However, these proposals are favored by both the truncated quadratic first-order prior and the two forms of second-order prior, generating plausible results that are incorporated into the output of the Smooth proposals also.

As well as combining the outputs from the SameUni and SegPln proposals, the Smooth proposals also allow disparity gradient discontinuities to become smoother. This effect can be seen by comparing the output from the Smooth proposals (Fig. 6e) to presmoothed disparity, i.e., fusion of SegPln and SameUni outputs (Fig. 6d) for the linear second-order prior—gradient discontinuities (e.g., top right and bottom center of the image) become more curved. What this shows is that the linear second-order prior, like the quadratic first- and second-order priors, does allow curved surfaces where the data supports this, unlike the linear first-order prior.

The success of both the SegPln and Smooth proposal schemes, in spite of sometimes high numbers of unlabeled nodes, suggests that the optimization framework proposed here will work well with any arbitrary proposals, and not just the three schemes put forward here as examples.

4.4 Comparison of Priors

To evaluate the performance of second-order priors we compare their results with those of first-order priors generated using the same stereo framework presented here. We use four quite different sequences, each containing two rectified views. The Corridor sequence (Fig. 1) is a synthetic gray-scale sequence, Venus (Fig. 4), a Middlebury evaluation sequence, is of a highly planar scene, Teddy (Fig. 7), another Middlebury evaluation sequence, is of a cluttered scene of curved and planar textured and textureless objects, and Cloth3 (Fig. 6), an additional Middlebury sequence, is of a highly curved and textured surface.

The first thing to notice is that all results using a 1op linear prior are highly piecewise frontoparallel. This results from the facts that the prior permits only frontoparallel surfaces with zero cost, and that the concave kernel prefers a large jump in disparity over many small ones. This results in a highly unnatural reconstruction.

The convex center of the quadratic kernel overcomes this piecewise nature, preferring several smaller jumps in

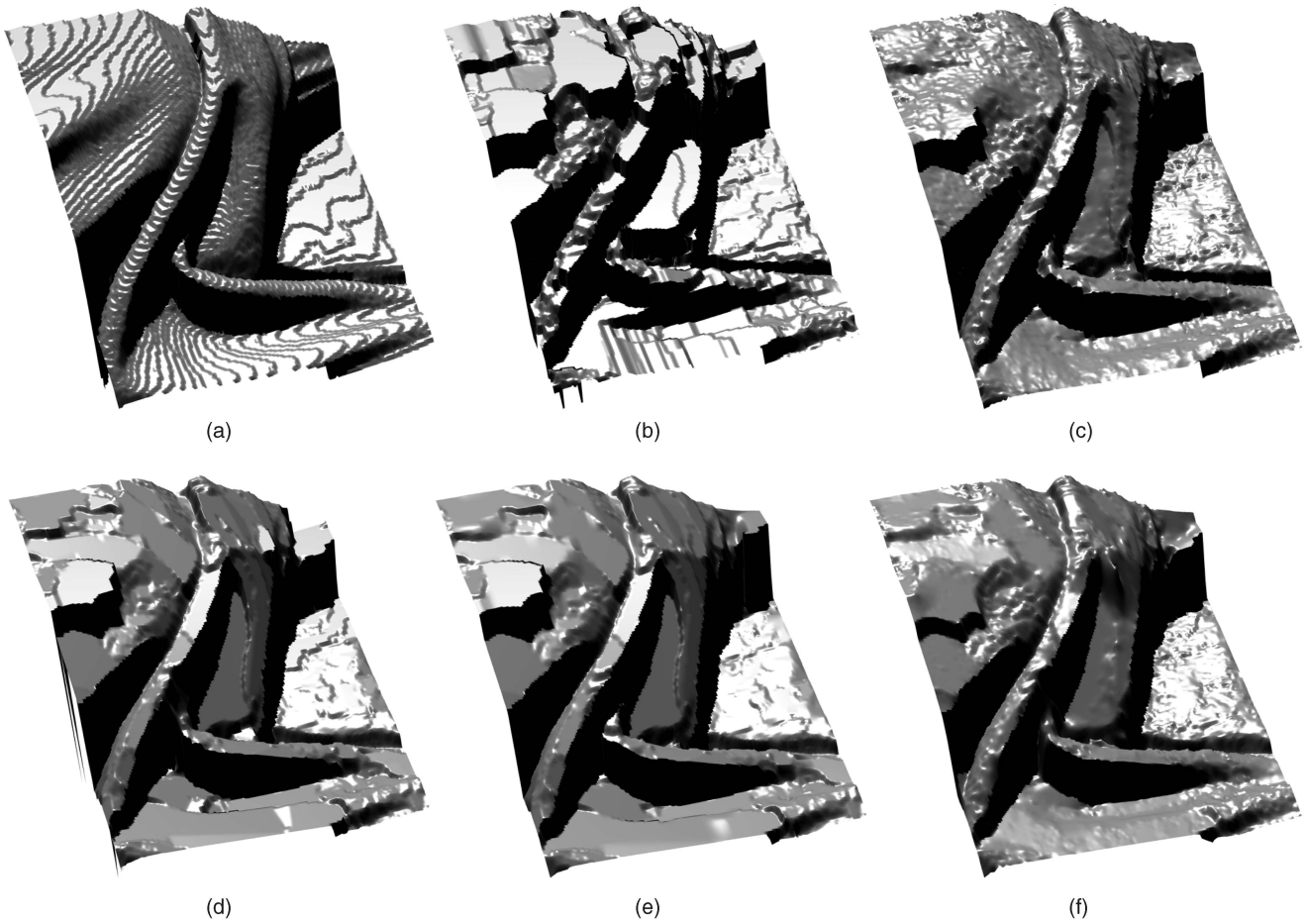

Fig. 6. Results for a region of the Middlebury Cloth3 sequence, displayed as a shaded 3D disparity surface. (a) Ground truth (discretized). (b) 1op linear. (c) 1op quadratic. (d) 2op linear presmooth. (e) 2op linear. (f) 2op quadratic.

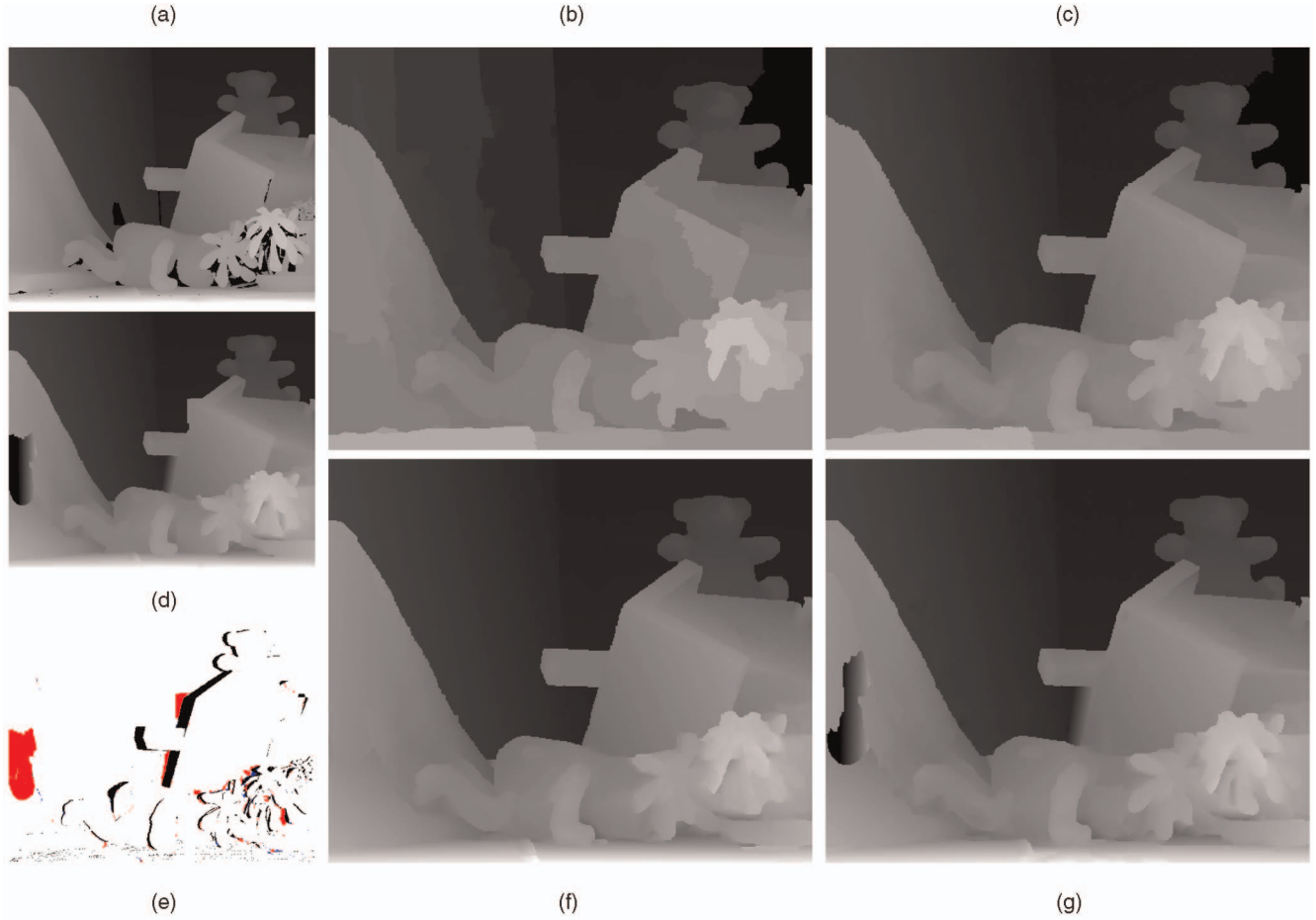

Fig. 7. Results for the Middlebury Teddy sequence. (a) Ground truth. (b) 1op linear. (c) 1op quadratic. (d) 2op linear no-vis. (e) Occlusions. (f) 2op linear. (g) 2op quadratic.

disparity over a larger one. In the case of the first-order prior this allows the generation of both planar (Fig. 4) and curved surfaces (Fig. 6c). However, it should be noted that these nonfrontoparallel surfaces must pay a smoothness cost, and this cost must be outweighed by the savings in data cost of the surface over a frontoparallel one. In the cases where the surfaces aren't textured enough (Fig. 1d) or the surface gradient is too great (bottom of Fig. 7c), the reconstruction inevitably reverts to piecewise-frontoparallel planes.

The second-order priors can equally generate curved surfaces where the data costs favor this, but additionally allow planar surfaces to be reconstructed in low texture and steep gradient regions, as seen in Figs. 1 and 7, respectively, improving the results in these regions. The choice of kernel has a lesser impact on second-order priors (see Fig. 6), but the truncated linear kernel generates a more piecewise-planar output than the quadratic kernel, which allows for slightly rougher surfaces at a fine scale. What is noticeable with a truncated quadratic second-order prior are that large-scale artifacts can occur, e.g., to the left of Fig. 7g, due to problems in the optimization caused by the high number of unlabeled pixels with this prior.

Fig. 8 shows the quantitative results from the Middlebury evaluation framework for all combinations of smoothness prior order and kernel. Error rates at various error thresholds (*left*) show that the 1op linear prior consistently performs worst, while the 2op linear prior performs best at all error thresholds save the lowest, at which 1op quadratic performs best—the performance of this latter prior drops behind at higher thresholds due to the greater number of gross errors caused by the frontoparallel preference of the prior. The 2op quadratic prior is always a fixed distance behind the 1op linear prior, a result of the extra gross errors caused by the optimization. Performance compared with other algorithms (*right*) shows exactly the same in terms of the relative performance of the priors, but it also shows that, while all priors perform worst at an error threshold of 1 (the default error threshold, which many algorithms are tuned

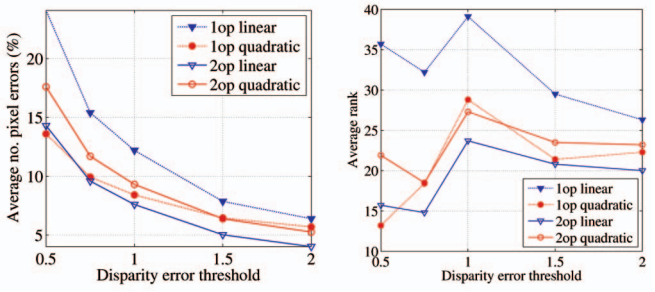

Fig. 8. Scores for different priors in the Middlebury stereo evaluation framework.

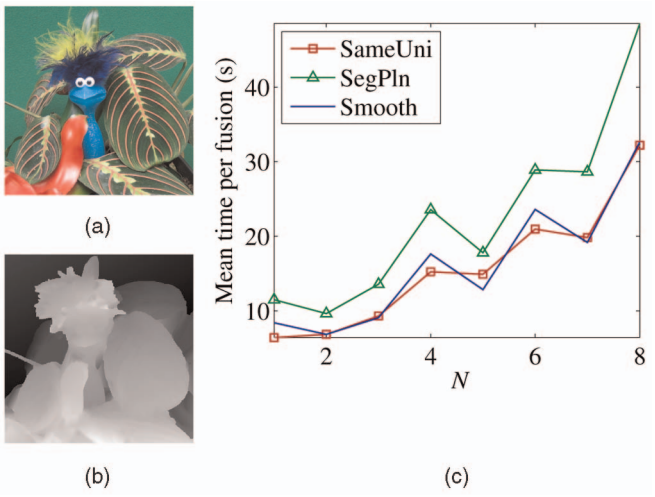

Fig. 9. **Multiple, arbitrary views**. (a) $\mathbf{I}_0$ for the Plant and toy sequence, which has arbitrary input views. (b) Output disparity using "2op, Smooth, QPBOI-R," for $N = 2$. (c) Graph of mean fusion times for each proposal scheme as a function of N.

to perform well at), the performance of all priors but 1op linear improves more at lower error thresholds than at higher ones, indicating their better subpixel accuracy in comparison to other methods.

4.5 Visibility

Fig. 7e highlights the benefits of a visibility constraint. It shows the visibility map for $\mathbf{I}_1$ of Teddy—pixels deemed occluded according to the following disparity maps are painted (covering the previous color) in the following order: 2op linear prior without visibility constraint (d), red; 2op linear prior with visibility constraint (f), blue; ground truth (a), black. Comparing numbers of red and blue pixels one can see that the visibility constraint reduces the number of falsely occluded pixels—it essentially encourages uniqueness of correspondences between input images. As unique correspondence is a constraint on real-world scenes, incorporating such a constraint in a stereo framework produces better results.

4.6 Multiple and Arbitrary Views

The formulation of our objective function allows for any number of input images to be used, and for those images to have arbitrary viewpoints. Fig. 9 shows results for such a data set—the Plant and toy sequence from [4]—for the 2op linear prior. We found little or no qualitative improvement between $N = 2$ (three views) and $N > 2$, something we believe can be attributed to the fact that three views are sufficient (in this case) to ensure that each pixel of $\mathbf{I}_0$ is visible in at least one other view. However, should more views be required, Fig. 9c shows that, in practice, the time per fusion iteration rises approximately linearly with N.

5 Conclusion

In this work, we have introduced a powerful framework for optimizing a second-order smoothness prior in stereo with geometrical visibility reasoning. In doing so, we introduce these key contributions which make this possible: providing a means to combine arbitrary, ad hoc disparity proposals in a reasoned way, minimizing a single objective energy; proving that binary nonsubmodular triple cliques can be decomposed into sets of pairwise cliques, enabling the use of a second-order prior; developing an asymmetrical occlusion model that consists of only pairwise cliques; proposing two new and effective methods for fixing the labels of nodes left unlabeled by QPBO.

We have compared the performance of four different priors within this framework and demonstrated that the second-order prior, with a linear truncated kernel and its complementary optimization framework, produces depth maps that more accurately reconstruct the scene, especially in low texture or highly slanted regions. We have shown that the algorithm can equally be applied to multiview stereo with arbitrary camera viewpoints, and does so for a computational cost roughly linear in N.

This work, in concentrating on the optimization of a second-order prior, has paid scant attention to the form of the data likelihood term, the use of a sampling-insensitive measure of photoconsistency, the form of the contrast dependent weighting of the smoothness term (i.e., the form of the CRF), the learning of optimal model parameters and the quality of the ad hoc proposals. We expect that improvements in these areas will bring about significant increases in performance. In addition, the optimizer we use, QPBO, is relatively new to the field of Computer Vision, and therefore we can expect its performance to increase significantly in the future with the development of further techniques to improve label fixing, to the extent that using a 2op quadratic prior may become viable.

Acknowledgments

The authors thank Vladimir Kolmogorov for making available his QPBO software and also for discussing graph-cut stereo with us. They also thank Yuri Boykov, Pushmeet Kohli, Carsten Rother, and Ali Shahrokni for their helpful comments. Research funded by EPSRC grants EP/C007220/1 and EP/C006631/1(P). Oliver Woodford is sponsored by the EPSRC CASE studentship with Sharp. Philip Torr is sponsored by the Royal Society Wolfson Merit Award.

References

[1] W.E.L. Grimson, *From Images to Surfaces: A Computational Study of the Human Early Visual System*. MIT Press, 1981.
[2] D. Terzopoulos, "Multilevel Computational Processes for Visual Surface Reconstruction," *Computer Vision, Graphics, and Image Processing*, vol. 24, no. 1, pp. 52-96, Oct. 1983.
[3] G. Li and S.W. Zucker, "Surface Geometric Constraints for Stereo in Belief Propagation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 2355-2362, 2006.
[4] O.J. Woodford, I.D. Reid, P.H.S. Torr, and A.W. Fitzgibbon, "On New View Synthesis Using Multiview Stereo," *Proc. British Machine Vision Conf.*, vol. 2, pp. 1120-1129, 2007.
[5] O.J. Woodford, P.H.S. Torr, I.D. Reid, and A.W. Fitzgibbon, "Global Stereo Reconstruction under Second Order Smoothness Priors," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2008.
[6] 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.
[7] B.K.P. Horn, *Robot Vision*. MIT Press, 1986.

[8] M.A. Gennert, "Brightness-Based Stereo Matching," *Proc. Int'l Conf. Computer Vision,* pp. 139-143, 1988.
[9] D. Marr and T. Poggio, "A Computational Theory of Human Stereo Vision," *Philosophical Trans. Royal Soc. London A,* vol. 204, pp. 301-328, 1979.
[10] Y. Boykov, O. Veksler, and R. Zabih, "Fast Approximate Energy Minimization via Graph Cuts," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, no. 11, pp. 1222-1239, Nov. 2001.
[11] V. Kolmogorov and R. Zabih, "Multi-Camera Scene Reconstruction via Graph Cuts," *Proc. European Conf. Computer Vision,* vol. 3, pp. 82-96, 2002.
[12] J. Sun, N.-N. Zheng, and H.-Y. Shum, "Stereo Matching Using Belief Propagation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 7, pp. 787-800, July 2003.
[13] P.F. Felzenszwalb and D.P. Huttenlocher, "Efficient Belief Propagation for Early Vision," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 261-268, 2004.
[14] C. Strecha, R. Fransens, and L. Van Gool, "Wide Baseline Stereo from Multiple Views: A Probabilistic Account," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 552-559, June 2004.
[15] Y. Wei and L. Quan, "Asymmetrical Occlusion Handling Using Graph Cut for Multi-View Stereo," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 902-909, 2005.
[16] H. Ishikawa and D. Geiger, "Occlusions, Discontinuities, and Epipolar Lines in Stereo," *Proc. European Conf. Computer Vision,* pp. 232-248, 1998.
[17] V. Kolmogorov and R. Zabih, "Computing Visual Correspondence with Occlusions via Graph Cuts," *Proc. Int'l Conf. Computer Vision,* pp. 508-515, 2001.
[18] S. Roy and I.J. Cox, "A Maximum-Flow Formulation of the N-Camera Stereo Correspondence Problem," *Proc. Int'l Conf. Computer Vision,* pp. 492-502, 1998.
[19] O. Veksler, "Graph Cut Based Optimization for MRFs with Truncated Convex Priors," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2007.
[20] S. Baker, R. Szeliski, and P. Anandan, "A Layered Approach to Stereo Reconstruction," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 434-441, 1998.
[21] S. Birchfield and C. Tomasi, "Multiway Cut for Stereo and Motion with Slanted Surfaces," *Proc. Int'l Conf. Computer Vision,* pp. 489-495, Sept. 1999.
[22] P.H.S. Torr, R. Szeliski, and P. Anandan, "An Integrated Bayesian Approach to Layer Extraction from Image Sequences," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, no. 3, pp. 297-304, Mar. 2001.
[23] H. Tao, H.S. Sawhney, and R. Kumar, "A Global Matching Framework for Stereo Computation," *Proc. Int'l Conf. Computer Vision,* pp. 532-539, 2001.
[24] M. Bleyer and M. Gelautz, "A Layered Stereo Algorithm Using Image Segmentation and Global Visibility Constraints," *Proc. Int'l Conf. Image Processing,* vol. 5, pp. 2997-3000, Oct. 2004.
[25] L. Hong and G. Chen, "Segment-Based Stereo Matching Using Graph Cuts," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 74-81, 2004.
[26] A. Klaus, M. Sormann, and K. Karner, "Segment-Based Stereo Matching Using Belief Propagation and a Self-Adapting Dissimilarity Measure," *Proc. Int'l Conf. Pattern Recognition,* pp. 15-18, 2006.
[27] Q. Yang, L. Wang, R. Yang, H. Stewénius, and D. Nistér, "Stereo Matching with Color-Weighted Correlation, Hierachical Belief Propagation and Occlusion Handling," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 2347-2354, 2006.
[28] M. Bleyer and M. Gelautz, "Graph-Cut-Based Stereo Matching Using Image Segmentation with Symmetrical Treatment of Occlusions," *Signal Processing: Image Comm.,* vol. 2, pp. 127-143, Feb. 2007.
[29] Y. Boykov and O. Veksler, "Graph Cuts in Vision and Graphics: Theories and Applications," *The Handbook of Math. Models in Computer Vision,* Springer, 2006.
[30] P. Kohli, "Minimizing Dynamic and Higher Order Energy Functions Using Graph Cuts," PhD dissertation, Oxford Brookes Univ., Nov. 2007.
[31] A. Bhusnurmath and C.J. Taylor, "Solving Stereo Matching Problems Using Interior Point Methods" *Proc. Fourth Int'l Symp. 3D Data Processing, Visualization, and Transmission,* pp. 321-329, June 2008.
[32] R. Szeliski, "A Multi-View Approach to Motion and Stereo," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 157-163, 1999.
[33] D. Geiger, B. Ladendorf, and A. Yuille, "Occlusions and Binocular Stereo," *Int'l J. Computer Vision,* vol. 14, pp. 211-226, 1995.
[34] P.N. Belhumeur, "A Bayesian Approach to Binocular Steropsis," *Int'l J. Computer Vision,* vol. 19, no. 3, pp. 237-260, Aug. 1996.
[35] J. Sun, Y. Li, S.B. Kang, and H. Shum, "Symmetric Stereo Matching for Occlusion Handling," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2005.
[36] V. Kolmogorov and C. Rother, "Comparison of Energy Minimization Algorithms for Highly Connected Graphs," *Proc. European Conf. Computer Vision,* vol. 2, pp. 1-15, 2006.
[37] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother, "A Comparative Study of Energy Minimization Methods for Markov Random Fields," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 30, no. 6, pp. 1068-1080, June 2008.
[38] M. Isard, "PAMPAS: Real-Valued Graphical Models for Computer Vision," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 613-620, 2003.
[39] E.B. Sudderth, A.T. Ihler, W.T. Freeman, and A.S. Willsky, "Nonparametric Belief Propagation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 605-612, 2003.
[40] B. Potetz, "Efficient Belief Propagation for Vision Using Linear Constraint Nodes," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2007.
[41] V. Kolmogorov and R. Zabih, "What Energy Functions Can be Minimized via Graph Cuts?" *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, no. 2, pp. 147-159, Feb. 2004.
[42] H. Ishikawa, "Exact Optimization for Markov Random Fields with Convex Priors," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 10, pp. 1333-1336, Oct. 2003.
[43] D. Schlesinger and B. Flach, "Transforming an Arbitrary Minsum Problem into a Binary One," Technical Report TUD-FI06-01, Dresden Univ. of Technology, 2006.
[44] V. Lempitsky, C. Rother, and A. Blake, "LogCut-Efficient Graph Cut Optimization for Markov Random Fields," *Proc. Int'l Conf. Computer Vision,* 2007.
[45] O. Woodford, I.D. Reid, P.H.S. Torr, and A.W. Fitzgibbon, "Fields of Experts for Image-Based Rendering," *Proc. British Machine Vision Conf.,* vol. 3, pp. 1109-1108, 2006.
[46] J. Winn and J. Shotton, "The Layout Consistent Random Field for Recognizing and Segmenting Partially Occluded Objects," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2006.
[47] C. Rother, S. Kumar, V. Kolmogorov, V. Lempitsky, and A. Blake, "Digital Tapestry," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 589-596, 2005.
[48] P.L. Hammer, P. Hansen, and B. Simeone, "Roof Duality, Complementation and Persistency in Quadratic 0-1 Optimization," *Math. Programming,* vol. 28, pp. 121-155, 1984.
[49] V. Kolmogorov and C. Rother, "Minimizing Non-Submodular Functions with Graph Cuts—A Review," Technical Report MSR-TR-2006-100, Microsoft Research, 2006.
[50] E. Boros, P.L. Hammer, and G. Tavares, "PreProcessing of Unconstrained Quadratic Binary Optimization," Technical Report RRR 10-2006, Rutgers Center for Operations Research, Apr. 2006.
[51] C. Rother, V. Kolmogorov, V. Lempitsky, and M. Szummer, "Optimizing Binary MRFs via Extended Roof Duality," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2007.
[52] A. Raj, G. Singh, and R. Zabih, "MRF's for MRI's: Bayesian Reconstruction of MR Images via Graph Cuts," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1061-1068, 2006.
[53] O.J. Woodford, I.D. Reid, and A.W. Fitzgibbon, "Efficient New View Synthesis Using Pairwise Dictionary Priors," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2007.
[54] V. Lempitsky, S. Roth, and C. Rother, "FusionFlow: Discrete-Continuous Optimization for Optical Flow Estimation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2008.
[55] M. Pollefeys, R. Koch, and L. Van Gool, "Self Calibration and Metric Reconstruction in spite of Varying and Unknown Internal Camera Parameters," *Proc. Int'l Conf. Computer Vision,* pp. 90-96, 1998.
[56] P. Gargallo and P. Sturm, "Bayesian 3D Modeling from Images Using Multiple Depth Maps," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 885-891, June 2005.

[57] 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.
[58] R. Szeliski and D. Scharstein, "Sampling the Disparity Space Image," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, no. 3, pp. 419-425, Mar. 2004.
[59] J. Besag, "Spatial Interaction and the Statistical Analysis of Lattice Systems," *J. Royal Statistical Soc., Series B,* vol. 36, no. 2, pp. 192-236, 1974.
[60] P. Fua, "A Parallel Stereo Algorithm that Produces Dense Depth Maps and Preserves Image Features," *Machine Vision and Applications,* vol. 6, no. 1, pp. 35-49, 1993.
[61] A. Alvarez, R. Deriche, J. Sánchez, and J. Weickert, "Dense Disparity Map Estimation Respecting Image Discontinuities: A PDE and Scale-Space Based Approach," *J. Visual Comm. and Image Representation,* vol. 13, no. 1, pp. 3-21, Mar. 2002.
[62] D. Scharstein and C. Pal, "Learning Conditional Random Fields for Stereo," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2007.
[63] E. Gamble and T. Poggio, "Visual Integration and Detection of Discontinuities: The Key Role of Intensity Edges," Technical Report AI Memo No. 970, MIT Artificial Intelligence Lab., Oct. 1987.
[64] D. Comaniciu and P. Meer, "Mean Shift: A Robust Approach toward Feature Space Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 5, pp. 603-619, May 2002.
[65] P. Kohli, M.P. Kumar, and P.H.S. Torr, "P^3 & Beyond: Solving Energies with Higher Order Cliques," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2007.
[66] A. Billionnet and B. Jaumard, "A Decomposition Method for Minimizing Quadratic Pseudoboolean Functions," *Operations Research Letters,* vol. 8, no. 3, pp. 161-163, June 1989.

Oliver Woodford received the MEng degree in engineering from Cambridge University in 2002, before spending some time in the embedded processor industry. He is currently studying for the DPhil degree in engineering at Oxford University, where he specializes in computer vision and, in particular, prior models for new-view synthesis.

Philip Torr received the DPhil degree from the University of Oxford under Professor David Murray, working there as a research fellow for a further three years, and remains a visiting fellow there. He then worked for six years at Microsoft Research, first in Redmond, Washington, then in Cambridge, United Kingdom, founding the vision side of the Machine Learning and Perception group. He is now a professor in computer vision and machine learning at Oxford Brookes University. He has won several awards, including the Marr prize (the highest honor in vision) in 1998, and is a Royal Society Wolfson Research Merit Award holder. He was involved in the algorithm design for Boujou, released by 2D3, which has won a number of industry awards. He continues to work closely with this Oxford-based company, as well as other companies such as Sony and Sharp. Recent SIGGRAPH work on VideoTrace with the University of Adelaide has been featured extensively on the Internet, including slashdot. He is a senior member of the IEEE.

Ian Reid received the BSc degree from the University of Western Australia in 1987 and came to Oxford University on a Rhodes Scholarship in 1988, where he completed the DPhil degree in 1991. He is a reader in engineering science and a fellow of the Exeter College at the University of Oxford, where he jointly heads the Active Vision Group. His research has touched on many aspects of computer vision, concentrating on algorithms for visual tracking, control of active head/eye robotic platforms (for surveillance and navigation), SLAM, visual geometry, novel view synthesis, and human motion capture. He has published more than 100 papers on these and related topics. He serves on the editorial boards of *Image and Vision Computing Journal* and *IPSJ Transactions on Computer Vision*. He is a member of the IEEE and the IEEE Computer Society.

Andrew Fitzgibbon studied mathematics and computer science at the University College Cork and received the PhD degree from Edinburgh University in 1997. Until June 2005, he held a Royal Society University Research Fellowship at Oxford University's Department of Engineering Science. He is a senior researcher at Microsoft Research, Cambridge, United Kingdom. His research interests are in the intersection of computer vision and computer graphics, with excursions into neuroscience. Recent papers have been on the recovery of 3D geometry from 2D images, general-purpose camera calibration, human 3D perception, and the application of natural image statistics to problems of figure/ground separation and new-view synthesis. He has twice received IEEE's Marr Prize, the highest in computer vision, and software he wrote won the Engineering Emmy Award in 2002 for significant contributions to the creation of complex visual effects. He is a member of the IEEE.