In 3DV 2017, Qingdao, China

Semi-Global Stereo Matching with Surface Orientation Priors

Daniel Scharstein
Middlebury College
schar@middlebury.edu

Tatsunori Taniai
RIKEN AIP
tatsunori.taniai@riken.jp

Sudipta N. Sinha
Microsoft Research
sudipta.sinha@microsoft.com

Abstract

Semi-Global Matching (SGM) is a widely-used efficient stereo matching technique. It works well for textured scenes, but fails on untextured slanted surfaces due to its fronto-parallel smoothness assumption. To remedy this problem, we propose a simple extension, termed SGM-P, to utilize precomputed surface orientation priors. Such priors favor different surface slants in different 2D image regions or 3D scene regions and can be derived in various ways. In this paper we evaluate plane orientation priors derived from stereo matching at a coarser resolution and show that such priors can yield significant performance gains for difficult weakly-textured scenes. We also explore surface normal priors derived from Manhattan-world assumptions, and we analyze the potential performance gains using oracle priors derived from ground-truth data. SGM-P only adds a minor computational overhead to SGM and is an attractive alternative to more complex methods employing higher-order smoothness terms.

1. Introduction

Semi-Global Matching (SGM) is a widely-used stereo matching technique introduced by Hirschmüller [24, 26]. It combines the efficiency of local methods with the accuracy of global methods by approximating a 2D MRF optimization problem with several 1D scanline optimizations, which can be solved efficiently via dynamic programming. It has been shown that SGM is a special case of message-passing algorithms such as belief propagation and TRW-T [12].

SGM has had significant impact, and the method is widely used in real-world applications, including 3D mapping, robot and UAV navigation, and autonomous driving [27, 35, 4]. SGM is also present in popular computer vision libraries such as OpenCV and has been implemented in hardware via FPGAs [19] and on GPUs [3].

While the method works well for aerial imagery and textured outdoor scenes, it works less well for indoor scenes with large untextured regions. The reason is that the algorithm employs a simple first-order smoothness assumption that favors fronto-parallel surfaces. It therefore tends to hallucinate fronto-parallel patches on weakly-textured slanted surfaces, which fail common consistency checks and result in large "holes" in the reconstructed surface, in particular when matching high-resolution images (Fig. 1d).

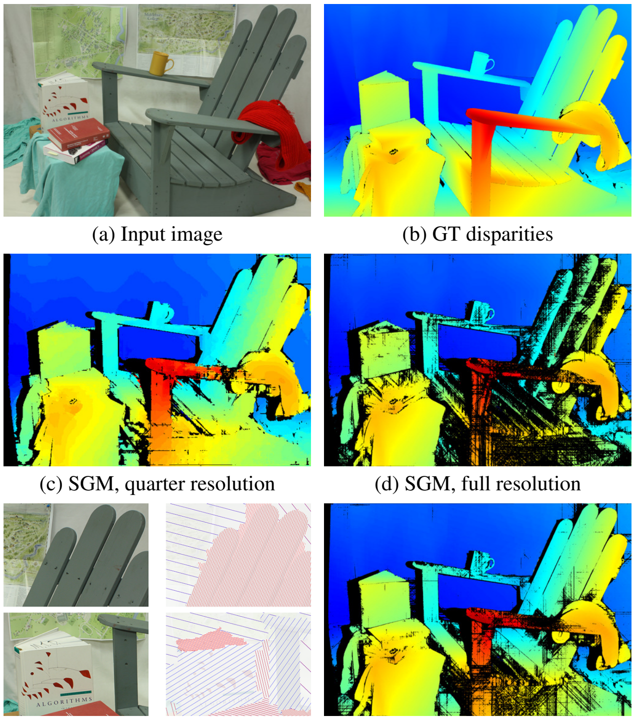

(a) Input image
(b) GT disparities
(c) SGM, quarter resolution
(d) SGM, full resolution
(e) Estimated orientation priors
(f) SGM-P, full resolution

Figure 1. (a, b) Adirondack input image and ground-truth disparities. (c, d) High-confidence standard SGM disparities at quarter and full resolution; slanted surfaces cause problems at full resolution. (e) Planar surface orientation priors derived from (c). (f) Our SGM-P method uses these priors, yielding high-confidence disparities with significantly fewer holes on slanted surfaces.

In this paper we propose a simple extension to the SGM algorithm, SGM-P, that utilizes precomputed surface orientation priors. Such priors favor different surface slants in different regions of the disparity space, and are implemented via local adjustments to SGM's transition penalties. The basic idea is to render the prior surface in the

3D disparity space and store the locations of discrete disparity steps in an *offset image* (Fig. 1e). SGM's smoothness term is then modified so that the zero-cost surface follows these steps and thus stays parallel to the prior surface. Arbitrary surfaces, not just planes, can be used as orientation priors, and our algorithm also supports multiple surface priors at different depths via an *offset volume*. Our method acts as a soft constraint during matching and only adds a minor computational overhead. In this paper we demonstrate that even simple surface priors can yield significant improvements in difficult indoor scenes containing slanted surfaces with weak texture (see Fig. 1f). Importantly, our experiments show that in the absence of such difficulties the performance never significantly decreases.

The SGM-P algorithm is agnostic as to the source of the priors, which could be computed in many different ways. For instance, surface priors can be derived from matched sparse feature points via triangulation [20] or plane fitting [42]. Planes (or other parametric surfaces) can also be fitted to disparities estimated at a lower resolution, which is one method we explore in this paper. Alternately, surface orientation priors could stem from domain knowledge (e.g., expected ground plane orientation in autonomous driving [28]), from semantic analysis [15, 16, 2], or from vanishing point analysis and Manhattan-world assumptions [41, 43, 17, 32, 36], which we also explore in this paper. Finally, surface priors could be derived from other sensors with lower resolution (e.g., commodity depth cameras) to aid high-resolution stereo matching in untextured regions.

2. Related work

Stereo matching is one of the oldest and most-thoroughly studied problems in computer vision [40, 10]. Methods can generally be categorized into local and global methods [40]. Both types of methods make smoothness assumptions about the observed world; the former implicitely (e.g., by aggregating a matching cost over a local window), and the latter explicitly via a smoothness term that imposes a prior on the surfaces in the world. The simplest and most common smoothness assumption is *first order* and states that two neighboring pixels most likely have the same depth. This is assumed in both simple window-based methods such as SSD and pixel-based global MRF approaches [9]. A first-order smoothness assumption introduces a fronto-parallel bias. This is not a problem when there is sufficient texture in the scene, but causes errors on untextured slanted surfaces, which is often problematic for indoor scenes (see Fig. 1d).

Many approaches have been proposed that can handle slanted surfaces. Woodford et al. [47] show how second-order smoothness terms can be efficiently optimized via QBPO. Li and Zucker [33] derive smoothness models for slanted and curved surfaces using differential geometry. Bleyer et al. propose surface stereo [7] and object stereo [8] algorithms in which the scene is modeled with planes or splines, and PatchMatch stereo [6], in which local estimates of disparities and surface slant are propagated to neighboring regions. Sinha et al. [42] run local plane sweeps around disparity planes estimated from sparse feature matches.

Plane-sweep stereo with a preferred plane orientation was proposed by Collins [11], and subsequently extended to multiple plane orientations, Manhattan-world scenes, and piece-wise planar scenes [18, 17, 43, 30].

A more recent trend is to formulate stereo matching using continuous MRF frameworks [48]. While PatchMatch stereo [6] was a greedy algorithm, follow-up work such as PMBP [5] incorporates regularization. Other work employs MRFs with continuous labels, using fusion moves for optimization [34, 44]. Several recent stereo algorithms [21, 22, 50] use surface normal priors derived from single images [15, 31, 16, 2] These methods utilize continuous optimization and require minimization techniques such as primal-dual methods or linear programming. These ideas cannot be directly incorporated into SGM or another discrete MAP inference framework. Our proposed algorithm offers a simple and efficient alternative to such complex approaches, and contributes a practical way of imposing surface orientation priors amenable to discrete optimization.

In the context of SGM, several improvements have been proposed. The CSGM method by Hirschmüller [25] estimates the disparities in untextured regions by fitting planes to adjacent textured pixels. In contrast to such post-processing, our method incorporates orientation priors during the matching. Hermann et al. [23] suggest an approximate second-order smoothness term for SGM but do not demonstrate a clear performance gain. A hierarchical approach to SGM [37, 46] aims to reduce ambiguities and runtime by restricting the search range based on SGM results computed at a lower resolution. We use a similar idea as one possible mechanism to derive surface orientation priors. While hard constraints such as search-range reduction can cause fine detail to be missed, in our case we only use the result from a coarser resolution to obtain a soft constraint on surface orientations.

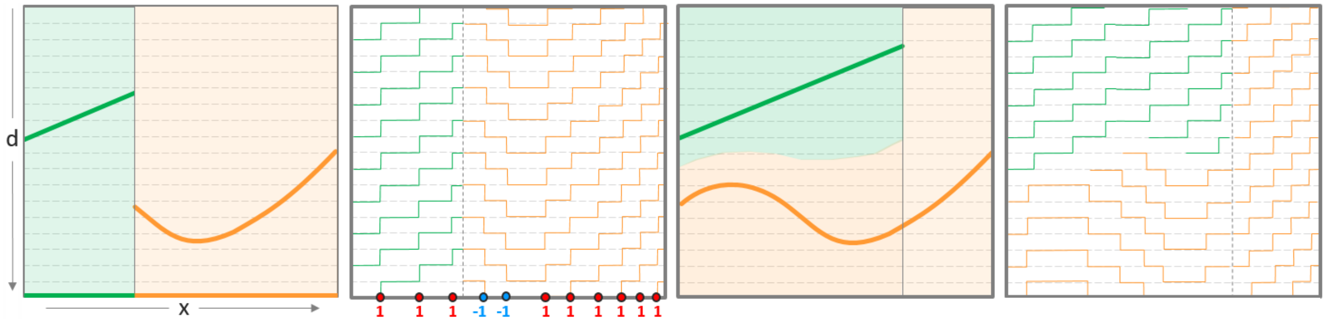
Finally, our priors could also be added to other recent modifications of SGM, for instance MGM [14], which integrates results from multiple directions. Similarly, our technique is orthogonal to recent advances in matching cost learning by CNNs [49] as we show in experiments below.

3. Algorithm

We first review the SGM algorithm, then describe our proposed extension SGM-P.

3.1. SGM

The Semi-Global Matching (SGM) algorithm [26] is an efficient technique for approximate energy minimization of



(a) (b) (c) (d)

Figure 2. Illustration of SGM-P's smoothness term: An x-d slice of the disparity volume with two prior surfaces (green and orange line) whose orientations we want to encourage, and their respective regions of influence (shaded). (a) 2D orientation priors extend across all disparities. (b) Rasterized version; the disparity jumps ±1 (red and blue circles) do not depend on d and can be stored in an *offset image*. (c) 3D priors allow multiple surface hypotheses per pixel (here, two on the left and one on the right). Each point is influenced by its closest surface in the d direction, so surfaces define Voronoi cells. (d) When multiple surfaces are present, disparity jumps vary with d and define an *offset volume*. The offsets are computed for each surface segment within its respective Voronoi cell.

a 2D Markov Random Field (MRF),

$$E(D) = \sum_{\mathbf{p}} C_{\mathbf{p}}(d_{\mathbf{p}}) + \sum_{\mathbf{p},\mathbf{q}\in\mathcal{N}} V(d_{\mathbf{p}}, d_{\mathbf{q}}), \quad (1)$$

where $C_{\mathbf{p}}(d)$ is a unary data term that represents the cost of matching pixel $\mathbf{p}$ at disparity $d \in \mathcal{D} = \{d_{\min}, \ldots, d_{\max}\}$, and $V(d, d')$ is a pairwise smoothness term that penalizes disparity differences between neighboring pixels. Specifically, V implements a first-order smoothness assumption,

$$V(d, d') = \begin{cases} 0 & \text{if } d = d' \\ P_1 & \text{if } |d - d'| = 1 \\ P_2 & \text{if } |d - d'| \geq 2. \end{cases} \quad (2)$$

Instead of minimizing the 2D MRF, which is NP-hard, SGM efficiently minimizes a 1D version of Eqn. 1 along 8 cardinal directions $\mathbf{r}$ via dynamic programming [26]. For each direction $\mathbf{r}$, SGM computes an aggregated matching cost $L_{\mathbf{r}}(\mathbf{p}, d)$ recursively defined from the image boundary:

$$L_{\mathbf{r}}(\mathbf{p}, d) = C_{\mathbf{p}}(d) + \min_{d'\in\mathcal{D}} (L_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, d') + V(d, d')). \quad (3)$$

The 8 aggregated costs are summed at each pixel, yielding an aggregated cost volume

$$S(\mathbf{p}, d) = \sum_{\mathbf{r}} L_{\mathbf{r}}(\mathbf{p}, d) \quad (4)$$

whose per-pixel minima are chosen as the winning disparities

$$d_{\mathbf{p}} = \arg\min_{d} S(\mathbf{p}, d). \quad (5)$$

Drory et al. [12] observe that the sum of the 8 individual minima of $L_{\mathbf{r}}(\mathbf{p}, d)$ is a lower bound on the minimum of the aggregated cost $S(\mathbf{p}, d)$ at each pixel $\mathbf{p}$, and define an uncertainty measure $U_{\mathbf{p}}$ as the difference between the two:

$$U_{\mathbf{p}} = \min_{d} \sum_{\mathbf{r}} L_{\mathbf{r}}(\mathbf{p}, d) - \sum_{\mathbf{r}} \min_{d} L_{\mathbf{r}}(\mathbf{p}, d). \quad (6)$$

The intuition is that $U_{\mathbf{p}}$ will be zero at locations where the 8 minimum-cost paths agree, e.g., in textured regions where incorrect disparities have high unary costs $C_{\mathbf{p}}$. In untextured regions, however, multiple disparities will have similar unary costs, and the 8 individual minima of $L_{\mathbf{r}}$ will likely occur at different disparities, in particular on slanted surfaces. We use $U_{\mathbf{p}}$ in our experiments to plot disparity error as a function of uncertainty, and also in Fig. 1 to select high-confidence matches.

3.2. SGM-P

In order to utilize surface priors, the basic idea is to modify SGM's smoothness penalties to favor surfaces with the expected surface slant. The problem is that we cannot represent fractional surface slants in algorithms that use discrete disparities, such as SGM. Thus, the key idea is to *rasterize* the disparity surface, i.e., render it in the 3D pixel grid, and record the locations of the steps in the discretized disparity values. At these locations we then shift SGM's smoothness penalties V so that the zero-cost transitions coincide with these steps. See Fig. 2 for illustration. We first discuss the case where we have only one orientation prior per pixel.

3.3. 2D orientation priors

Assume we are given a real-valued disparity surface prior S whose orientation at any given pixel $\mathbf{p}$ we would like to encourage across all possible disparities (Fig. 2a).

Let $\mathbf{r}$ be the current "sweeping direction" of SGM. Given a pixel $\mathbf{p}$ and its successor $\mathbf{p}' = \mathbf{p} + \mathbf{r}$, we rasterize the surface S to integer disparities

$$\hat{S}(\mathbf{p}) = \text{round}(S(\mathbf{p})) \quad (7)$$

and compute the discrete disparity steps (or *jumps*)

$$j_{\mathbf{p}} = \hat{S}(\mathbf{p}') - \hat{S}(\mathbf{p}). \quad (8)$$

We replace the original smoothness penalty V with a new function V_S that incorporates the disparity jumps:

$$V_S(d_{\mathbf{p}}, d'_{\mathbf{p}}) = V(d_{\mathbf{p}} + j_{\mathbf{p}}, d'_{\mathbf{p}}). \tag{9}$$

At pixels where the value of j is nonzero, V_S favors taking that disparity step and encourages the disparity surface to stay parallel to S. We can efficiently compute V_S by storing the jumps j in an *offset image* (Fig. 2b). Since the jumps depend on the direction $\mathbf{r}$, four different offset images are needed, one for each pair of opposing directions.

However, we do not need to precompute and store all four offset images simultaneously. Instead, we only need to store $\hat{S}$. Before running scanline optimization on each pair of opposing directions, we generate the associated offset image using Eqns. 7 and 8. When reversing the direction, we flip the signs of the offsets.

3.4. 3D orientation priors

We now relax the requirement of a single orientation prior per pixel and allow multiple overlapping surface hypotheses at different depths. As before, each surface should only act as an *orientation* prior, but should influence all nearby points (Fig. 2c). Assume that at pixel $\mathbf{p}$ we have K disparity surfaces $\{S_k^{\mathbf{p}}\}, k = 1 \ldots K$. For a given disparity d, we find the closest surface in terms of disparity

$$\tilde{S}_d^{\mathbf{p}} = \arg\min_k |S_k^{\mathbf{p}}(\mathbf{p}) - d|. \tag{10}$$

Then, we rasterize it to integer disparities

$$\hat{S}_d(\mathbf{p}) = \text{round}(\tilde{S}_d^{\mathbf{p}}(\mathbf{p})), \tag{11}$$

and again compute the discrete disparity jumps between adjacent pixels. This time, however, they depend on d:

$$j_{\mathbf{p}}(d) = \hat{S}_d(\mathbf{p}') - \hat{S}_d(\mathbf{p}). \tag{12}$$

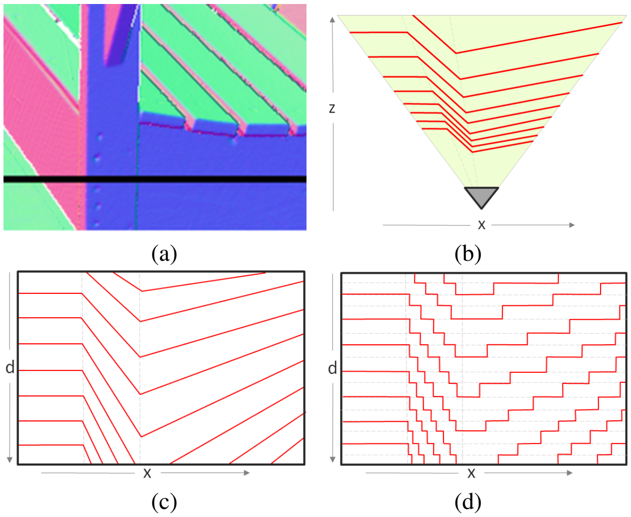
The new smoothness penalty term now becomes

$$V_S(d_{\mathbf{p}}, d'_{\mathbf{p}}) = V(d_{\mathbf{p}} + j_{\mathbf{p}}(d_{\mathbf{p}}), d'_{\mathbf{p}}). \tag{13}$$

Fig. 2d illustrates this disparity-dependent orientation prior. In order to compute V_S efficiently, we store the precomputed values of $\tilde{S}_d^{\mathbf{p}}$ for all $\mathbf{p}$ and d in an auxiliary volume. This can be seen as a 1D discrete Voronoi diagram for each pixel along the disparity axis in the volume. The values in each column can be computed efficiently with a forward scan followed by a backward scan after rendering all disparity surfaces into the column at this pixel.

3.5. Surface normal priors

So far we have considered using disparity surfaces as orientation priors. We now focus on the case where we are given a surface normal map as prior, obtained for instance



(a) (b)

(c) (d)

Figure 3. Converting surface normals to disparity orientation priors. (a) Color rendering of a surface normal map with a scanline spanning three piecewise-planar surfaces. (b) Integrating normals under perspective projection gives rise to parallel surfaces with varying scale and depth. (c) Converting to disparity space, viewing rays become parallel, and surface slant now varies with disparity. (d) The offset volume encodes all possible surface orientations; note that disparity steps are no longer aligned vertically.

via photometric stereo [13] or from Manhattan-world priors [32]. Such normal maps can be used in our SGM-P algorithm, though the situation is more complex than one might expect. First, we cannot use orientations directly, since SGM-P requires an actual surface that can be rasterized. Thus, the normal map must be integrated. Second, we need to distinguish between scene space (in Euclidean world coordinates) and disparity space. While a surface normal map can be considered a 2D prior since it encodes a single surface orientation per pixel, this orientation is given in scene space. As we show below, when converting to disparity space under a perspective projection model, the surface orientation becomes depth-dependent and results in a 3D disparity orientation prior, requiring an offset volume representation. See Figure 3 for illustration.

We now derive the relationship between surface normals in scene coordinates and the orientation of disparity surfaces. Given the surface normal vector $(n_x^p, n_y^p, n_z^p)^T$ at pixel $\mathbf{p}$, the equation of the tangent plane of the surface at $\mathbf{p}$ in scene space coordinates (x, y, z) is $n_x^p x + n_y^p y + n_z^p z = h^p$. Here, h^p encodes the plane's unknown depth. Under perspective projection we have $x = uz/f$ and $y = vz/f$ for image coordinates (u, v) and a camera at the origin with focal length f. Substituting these values we obtain

$$z = h^p f / \left(n_x^p u + n_y^p v + f n_z^p\right). \tag{14}$$

For stereo pairs we have $z = bf/d$, where b and d are the baseline and disparity respectively. Substituting z into

Eqn. 14 we obtain the disparity plane equation

$$d(u,v) = \frac{b}{h^p}\left(n_x^p u + n_y^p v + f n_z^p\right). \tag{15}$$

Note that the disparity plane orientation depends on h^p, which encodes the depth of the tangent plane in scene space. Therefore, a scene plane with fixed orientation but unknown depth yields a family of disparity planes whose orientation depends on the associated disparity.

In order to use surface normal priors in SGM-P, they first need to be integrated into a surface. We do this integration in scene space under a perspective projection model using a least-squares approach [45] (see the supplementary materials for more details). The result is a z-surface in scene space coordinates, initially at an arbitrary depth. We scale this surface by an appropriate sequence of scale factors (Fig. 3b) and convert to d to arrive at roughly equally-spaced d-surfaces covering the full disparity range (Fig. 3c). Finally, we construct an offset volume from this family of disparity surfaces as described in the previous section, resulting in a 3D orientation prior with varying disparity surface slants. (Fig. 3d).

4. Experiments

We now demonstrate the utility of our new algorithm by comparing a baseline SGM implementation with various version of SGM-P employing different types of priors. For a fair comparison we use the same matching cost and smoothness weights across all versions of the algorithms.

To compensate for global and local rectification errors, we first robustly fit a global model $y' = ay + b$ to matched feature points and warp the right image accordingly before computing the matching costs. During matching, for each horizontal disparity, we evaluate matching costs corresponding to vertical disparities of {-1, 0, +1} pixels and select the smallest of the three costs.

For SGM's unary data term we use negated and truncated normalized cross correlation (NCC):

$$C_{\mathbf{p}}(d) = 1 - \max(0, \text{NCC}(\mathbf{p}, d)), \tag{16}$$

where $\text{NCC}(\mathbf{p}, d)$ compares 5×5 grayscale image patches centered at $\mathbf{p}$ and $\mathbf{p} - (d, 0)^T$ in the left and right image, respectively. Image intensities are in the range $[0, 255]$. We add a small value $\epsilon = 1.0$ to the NCC denominator to suppress the effect of noise in untextured regions. We scale $C_{\mathbf{p}}(d)$ by 255 and round it to the nearest integer. We can use unsigned shorts for SGM's aggregated costs, which reduces the memory overhead.

We use NCC as matching cost since it is commonly employed in real-world systems. As mentioned, our method is orthogonal to the choice of matching cost. Below, and in the supplementary materials, we also evaluate MC-CNN, the state-of-the-art matching cost by Zbontar and LeCun [49], and show that it yields similar performance.

No Prior	
SGM	– Baseline method
2D Prior (offset image representation)	
SGM-EPi	– Estimated segmented planes
SGM-GS	– GT surface
SGM-GP	– GT surface, planar approximation
SGM-GNi	– GT normals (fixed-z "strawman")
3D Prior (offset volume representation)	
SGM-EPv	– Estimated overlapping planes
SGM-GNv	– GT normals (accurate version)
SGM-MW	– Manhattan-world prior

Table 1. The algorithm variants compared in our experiments. See Section 4.1 for details.

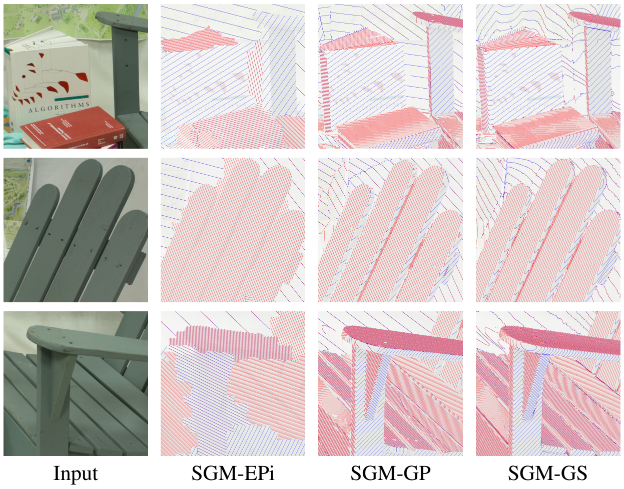


Figure 4. Visualization of different 2D orientation priors ("offset images") on zoomed regions of the Adirondack image pair.

For SGM's smoothness penalty (Eqn. 2) we use the following settings: $P_1 = 100$, $P_2 = P_1(1 + \alpha e^{-|\Delta I|/\beta})$, where $\alpha = 8$, $\beta = 10$, and $|\Delta I|$ is the absolute intensity difference at neighboring pixels. Our choice of P_2 favors large disparity jumps at high-contrast image edges.

4.1. Algorithm variants

In order to evaluate the full potential of SGM-P, we evaluate a number of different priors. We will substitute P with a combination of letters to distinguish the algorithm variants (see Table 1 for a summary and Fig. 4 for visualizations). The first letter distinguishes priors G derived from ground-truth disparities with priors E estimated from the input images. The former versions can be considered *oracles* that provide an upper bound on the potential benefit of our idea, while the latter versions give an indication of the actual realizable benefit. The second letter denotes the type of surfaces acting as priors: S for arbitrary (e.g., curved) surfaces, P for planar surfaces, and N for the case when only surface

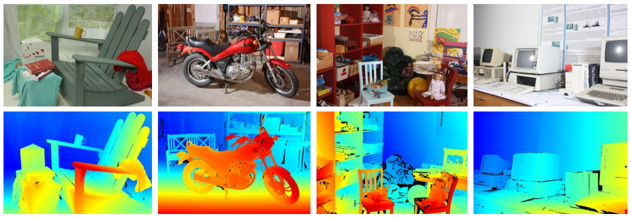

Adirondack Motorcycle Playroom Vintage

Figure 5. Challenging high-resolution Middlebury datasets with untextured slanted surfaces.

normals are available. Finally, we use i and v to distinguish between 2D priors only requiring an offset image, and 3D priors requiring an offset volume.

For the quantitative analysis we focus on only one E variant that estimates priors from the input images. We do this by running SGM at a coarser resolution and clustering the resulting disparities into disparity plane hypotheses [30]. We use these hypotheses and the associated pixel-to-plane label map to generate 2D and 3D orientation priors (EPi and EPv). For the 2D variant, SGM-EPi, we segment the image into superpixels [1] and then select for each superpixel the plane most often assigned to its constituent pixels. Superpixels with low support for any of the plane hypotheses are set to an arbitrary fronto-parallel plane.

The more powerful 3D variant, SGM-EPv, allows modeling of multiple disparity surface hypotheses at the same pixel. We use the same pixel-to-plane label map as for SGM-EPi but obtain potentially overlapping disparity hypotheses by bounding each 3D disparity plane by the convex hull of its constituent pixels in the label map.

For the oracle priors G based on ground-truth disparities we compare all three surface variants (S, P, and N) in order to explore the benefits and limitations of the different types of priors. Of these, SGM-GS uses the ground-truth disparity surface as 2D prior directly, which is the best possible prior available. Next, SGM-GP uses a piecewise-planar approximation of the ground-truth surface, again as 2D prior, constructed in the same manner as SGM-EPi. We omit the suffix "i" in both cases since we do not have corresponding 3D priors. (While a 3D variant of SGM-GP with multiple overlapping planar hypotheses is possible, we found that it yielded no benefit over the 2D version.) Next, SGM-GNv discards the original ground-truth surface and uses only its normal map, which results in a 3D prior as explained in Section 3.5. We also include a "strawman" 2D version, SGM-GNi, which we obtain by integrating a single z-surface at an arbitrary depth. Finally, we investigate a 3D normal prior estimated from the images using a Manhattan-world assumption [32]; deviating from our naming scheme we simply call it SGM-MW (more details on this below).

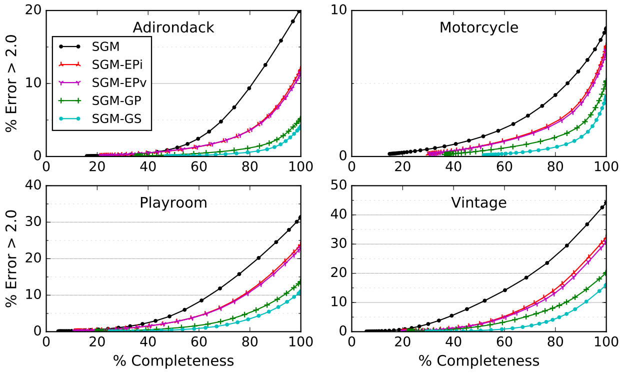


Figure 6. ROC curves plotting error rate vs. completeness for baseline SGM, estimated planar priors SGM-EPi and SGM-EPv, and ground-truth plane and surface priors SGM-GP and SGM-GS.

4.2. Quantitative analysis

We start by evaluating the promise of SGM-P on a subset of the high-resolution stereo pairs from the training set of the Middlebury stereo evaluation v3 [38], for which high-quality ground-truth disparities are available. We select 4 challenging image pairs with untextured slanted surfaces, depicted in Fig. 5. In the experiments below, we use the full-resolution (5–6 MP) versions of these datasets; see the supplementary materials for additional results, including other resolutions.

In our first experiment, we compare the baseline SGM method with our SGM-P algorithm using both estimated planar priors (SGM-EPi and SGM-EPv) and ground-truth priors (SGM-GP and SGM-GS). Fig. 6 shows disparity error rates (percentage of pixels whose disparity error is greater than t=2.0) as a function of completeness (inverse sparsity). We obtain disparity maps with increasing completeness by raising the allowable uncertainty $U_\mathbf{p}$ (Eqn. 6) from 0 to U_max. Our plots are similar to ROC curves and allow the comparison of sparse (or semi-dense) stereo methods that leave uncertain regions unmatched [29]. The error rate for the dense result (100% completeness) is the rightmost point on each curve.

The plots in Fig. 6 show that the four variants of our SGM-P algorithm all significantly outperform the baseline SGM algorithm on these four image pairs. As expected, the best performance is obtained with perfect orientation priors derived from the ground-truth surface (SGM-GS), which yields a dramatical improvement over SGM, with error rates ranging from one half to one fifth of the original errors. As mentioned, this provides an upper bound on the potential benefit of our idea. A more realistic upper bound is given by SGM-GP, which utilizes piecewise planar priors derived from the ground-truth disparities. This results in a slight decrease in performance compared to SGM-GS, but still a dramatical increase over the baseline.

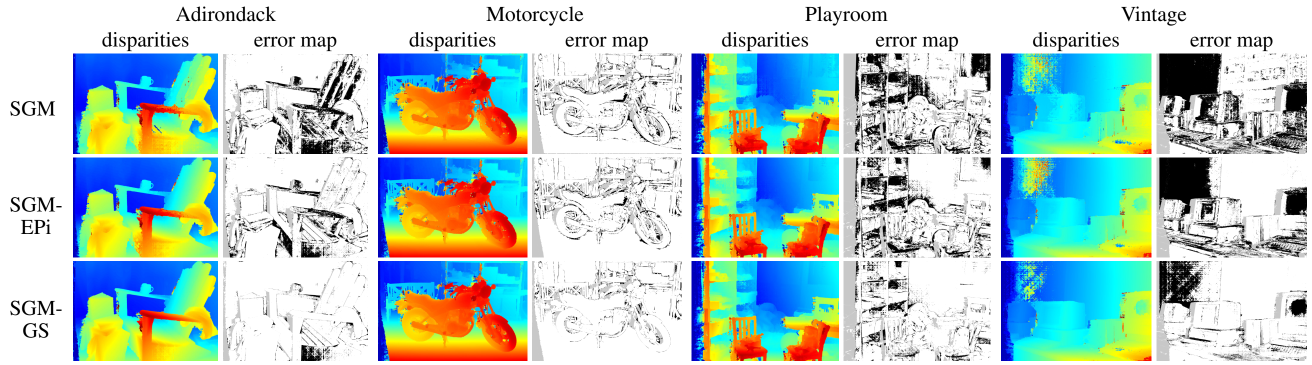


Figure 7. Disparity maps and error maps corresponding to the plots in Fig. 6 at 100% completeness. Black regions in the error maps indicate disparity errors > 2.0 in non-occluded regions. SGM-EPi and SGM-GS yield a noticeable reduction of errors on slanted surfaces.

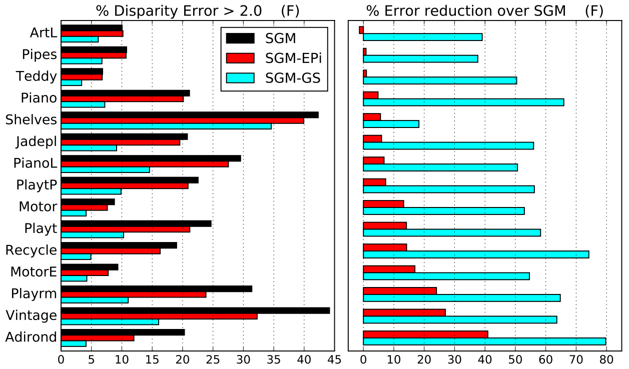


Figure 8. Performance on all 15 Middlebury training sets at full resolution, sorted by increasing performance gain (error reduction) of SGM-EPi over SGM.

Most importantly, even without utilizing ground-truth information, we still get a significant improvement from planar priors estimated from the input images (SGM-EPi and SGM-EPv). The two versions, which are almost indistinguishable in terms of performance, achieve between 25% and 50% of the upper bounds, resulting in an improvement over the original SGM errors by 13–41%. For now we will focus on the simpler SGM-EPi method; we will discuss the potential of 3D priors (SGM-EPv) below. Fig. 7 shows the disparity maps and error maps for the dense results (100% completeness) for SGM, SGM-EPi, and SGM-GS.

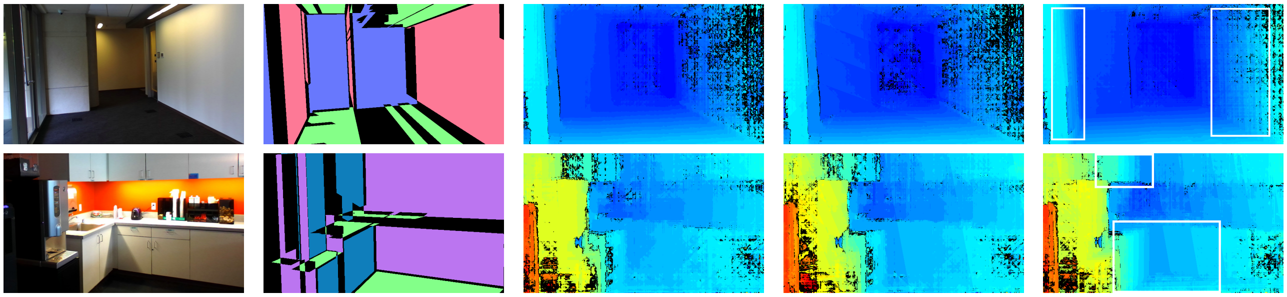
It should be noted that the benefit of SGM-P strongly depends on the scene structure. In scenes with mostly fronto-parallel surfaces, SGM-P yields little improvement. An important question is whether estimated priors can hurt the performance. Fig. 8 shows the performance of SGM-EPi and SGM-GS on all 15 Middlebury training pairs, sorted by rate of error reduction of SGM-EPi over SGM. It can be seen that the performance gains range from around -1% to 41%, with an average gain of 12%. Importantly, the performance never significantly decreases. We see the same trend for other matching costs; see the supplementary materials.

Cost	SGM		SGM-EPi		SGM-EPv		SGM-GS	
	avg err	rank	avg err	rank	avg err	rank	avg err	rank
NCC	18.9	28	16.3	24	16.2	24	8.27	4
gain over SGM	-		14%		14%		56%	
MC-CNN	15.6	23	13.7	18	13.5	18	7.31	2
gain over SGM	-		12%		13%		53%	

Table 2. Performance of different SGM-P variants and matching costs on the Middlebury online evaluation for the training sets.

We also submitted the results of SGM-EPi to the Middlebury stereo evaluation [39]. Since only a single submission per paper is allowed, we do not have a baseline for the 15 Middlebury test pairs. We thus cannot show the improvement ratios for the test pairs, as we do in Fig. 8 for the training pairs. The public evaluation results, however, show that our method outperforms all existing SGM entries in the public table (especially the full-resolution SGM entry). The largest performance gains are on scenes with slanted textureless surfaces, including Classroom, Crusade, and Stairs. SGM-EPi ranks 20th and 24th overall on test and training sets, respectively; among the full-resolution submissions it ranks 3rd on both sets. Table 2 shows the official (weighted) average training error rates, as well as the table ranks, for the different SGM-P variants for both NCC and MC-CNN [49] matching costs. While MC-CNN yields slightly lower errors, both costs result in similar performance gains. Recall that our goal is *not* to create a top-ranked stereo method, but rather to improve upon SGM, one of the most widely-used stereo methods. The rankings clearly show the potential of SGM-P for high-resolution stereo matching.

Finally, the fact that SGM-EPi and SGM-EPv produce very similar numerical results is not too surprising since most regions in the Middlebury images can be well explained with single surfaces. In the supplementary materials we show qualitative evidence that SGM-EPv is better at recovering surface creases by utilizing multiple overlapping hypotheses. Harnessing the full power of SGM-EPv, however, would require more powerful methods for generating priors that extend over larger regions of the image.



(a) Input image (b) Normal map (c) SGM (d) SGM-EPi (e) SGM-MW

Figure 9. (a, b) Input images and Manhattan-world normal estimates [32]. (c, d) Both SGM and SGM-EPi have trouble reconstructing some of the untextured slanted surfaces. (e) SGM-MW utilizes the normal map (b) and does a better job, in particular in the highlighted regions. Note that incorrect or missing normal information does not hurt performance if sufficient texture is present.

4.3. Manhattan-world priors

We exploit Manhattan-world layouts as a type of surface normal priors in our SGM-MW method. Using a method for scene layout recovery based on vanishing points [32], we obtain a semi-dense pixel labeling of the Manhattan world's principal surface normals in the left input image (Fig. 9b). We turn these normal priors into 3D disparity orientation priors by rendering planar segments in scene (z) space. Rather than covering the depth range uniformly, we select depths with local evidence for a surface, found by running SGM at a coarser resolution. Instead of fitting planes to these disparities (as we do for SGM-EPi) we convert the disparity map to z space, fit constrained planes with known orientation to the 3D points and convert those planes back to d space (see Section 3.5). Finally, these disparity planes are rasterized and used as 3D priors.

Fig. 9 shows qualitative results of SGM, SGM-EPi, and SGM-MW on two challenging indoor image pairs. We see that both SGM and SGM-EPi struggle in regions with slanted untextured surfaces, where SGM-EPi does not find a supporting plane. However, SGM-MW recovers smoother slanted surfaces in these regions by utilizing the Manhattan-world normal estimates.

In the supplementary materials we also test our surface normal prior idea with oracle priors derived from ground-truth data. Recall that a planar surface patch with known normal n but unknown depth z yields a family of disparity planes whose orientation depends on z, which requires a 3D prior (SGM-GNv). We show that modeling this depth dependance is crucial by demonstrating that SGM-GNv is significantly more accurate than the “strawman” 2D version SGM-GNi, which integrates the surface normals in scene space at one arbitrary depth and uses the resulting disparity surface as a 2D prior.

4.4. Runtimes

Fig. 10 compares the runtime of various SGM-P versions with the baseline SGM method on the 15 full-resolution Middlebury training pairs. All timings were measured on a computer with a 3.4 GHz Xeon E5-2643 v4 processor and 32 GB RAM. Our C++ implementation is not yet optimized for speed. However, the timings show that SGM-GS has almost no overhead over SGM. SGM-EPi and SGM-EPv have similar runtimes, with average runtime overheads over SGM of about 7%. This overhead is mainly due to the cost of extracting the orientation priors.

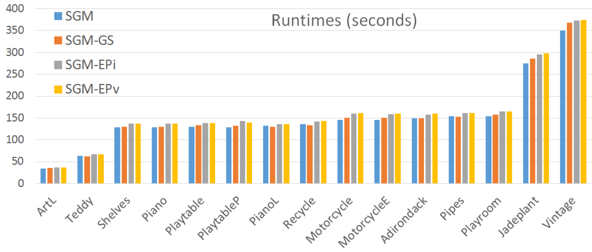


Figure 10. Comparison of runtimes.

5. Conclusion

We have presented a simple extension to semi-global matching (SGM) that allows surface orientation priors to be incorporated as soft constraints. Using priors derived from stereo matching at coarser resolution, our SGM-P method consistently yields improved accuracy for challenging indoor scenes that contain slanted weakly-textured surfaces. We also demonstrate the potential of orientation priors derived from single images. Our analysis involving oracle priors demonstrates the potential for large performance gains.

Avenues for future work includes recovering more accurate orientation priors, possibly via semantic analysis [2] or revisiting binocular photometric stereo [13]. Combining orientation priors with depth priors, obtained either from coarse resolution or via commodity depth sensors, is also worth exploring. Finally, it might be possible to extend our method to other MRF optimization frameworks with first-order smoothness terms.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE TPAMI*, 34(11):2274–2282, 2012. 6

[2] A. Bansal, B. Russell, and A. Gupta. Marr revisited: 2D-3D model alignment via surface normal prediction. In *CVPR*, 2016. 2, 8

[3] C. Banz, H. Blume, and P. Pirsch. Real-time semi-global matching disparity estimation on the GPU. In *Mobile Vision Workshop, ICCV*, pages 514–521, 2011. 1

[4] A. Barry, H. Oleynikova, D. Honegger, M. Pollefeys, and R. Tedrake. FPGA vs. pushbroom stereo vision for MAVs. In *IROS Workshop on Vision-based Control and Navigation of Small Lightweight UAVs*, 2015. 1

[5] F. Besse, C. Rother, A. Fitzgibbon, and J. Kautz. PMBP: PatchMatch belief propagation for correspondence field estimation. *IJCV*, 110(1):2–13, 2014. 2

[6] M. Bleyer, C. Rhemann, and C. Rother. PatchMatch stereo – stereo matching with slanted support windows. In *BMVC*, volume 11, 2011. 2

[7] M. Bleyer, C. Rother, and P. Kohli. Surface stereo with soft segmentation. In *CVPR*, pages 1570–1577, 2010. 2

[8] M. Bleyer, C. Rother, P. Kohli, D. Scharstein, and S. Sinha. Object stereo—joint stereo matching and object segmentation. In *CVPR*, pages 3081–3088, 2011. 2

[9] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE TPAMI*, 23(11):1222–1239, 2001. 2

[10] M. Brown, D. Burschka, and G. Hager. Advances in computational stereo. *IEEE TPAMI*, 25(8):993–1008, August 2003. 2

[11] R. Collins. A space-sweep approach to true multi-image matching. In *CVPR*, pages 358–363, 1996. 2

[12] A. Drory, C. Haubold, S. Avidan, and F. Hamprecht. Semi-global matching: a principled derivation in terms of message passing. In *GCPR*, pages 43–53, 2014. 1, 3

[13] H. Du, D. Goldman, and S. Seitz. Binocular photometric stereo. In *BMVC*, 2011. 4, 8

[14] G. Facciolo, C. De Franchis, and E. Meinhardt. MGM: A significantly more global matching for stereovision. In *BMVC*, 2015. 2

[15] D. Fouhey, A. Gupta, and M. Hebert. Data-driven 3D primitives for single image understanding. In *ICCV*, pages 3392–3399, 2013. 2

[16] D. Fouhey, A. Gupta, and M. Hebert. Unfolding an indoor origami world. In *ECCV*, pages 687–702, 2014. 2

[17] Y. Furukawa, B. Curless, S. Seitz, and R. Szeliski. Manhattan-world stereo. In *CVPR*, pages 1422–1429, 2009. 2

[18] D. Gallup, J.-M. Frahm, P. Mordohai, Q. Yang, and M. Pollefeys. Real-time plane-sweeping stereo with multiple sweeping directions. In *CVPR*, 2007. 2

[19] S. Gehrig, F. Eberli, and T. Meyer. A real-time low-power stereo vision engine using semi-global matching. In *Computer Vision Systems*, pages 134–143, 2009. 1

[20] A. Geiger, M. Roser, and R. Urtasun. Efficient large-scale stereo matching. In *ACCV*, 2010. 2

[21] S. Hadfield and R. Bowden. Exploiting high level scene cues in stereo reconstruction. In *ICCV*, pages 783–791, 2015. 2

[22] C. Hane, L. Ladicky, and M. Pollefeys. Direction matters: Depth estimation with a surface normal classifier. In *CVPR*, pages 381–389, 2015. 2

[23] S. Hermann, R. Klette, and E. Destefanis. Inclusion of a second-order prior into semi-global matching. *Advances in Image and Video Technology*, pages 633–644, 2009. 2

[24] H. Hirschmüller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *CVPR*, volume 2, pages 807–814, 2005. 1

[25] H. Hirschmüller. Stereo vision in structured environments by consistent semi-global matching. In *CVPR*, volume 2, pages 2386–2393, 2006. 2

[26] H. Hirschmüller. Stereo processing by semiglobal matching and mutual information. *IEEE TPAMI*, 30(2):328–341, 2008. 1, 2, 3

[27] H. Hirschmüller. Semi-global matching — motivation, developments and applications. In *Photogrammetric Week*, pages 173–184, 2011. 1

[28] K. Konolige, M. Agrawal, R. Bolles, C. Cowan, M. Fischler, and B. Gerkey. Outdoor mapping and navigation using stereo vision. In *International Symposium on Experimental Robotics*, 2006. 2

[29] J. Kostliva, J. Cech, and R. Sara. Feasibility boundary in dense and semi-dense stereo matching. In *CVPR BenCOS workshop*, 2007. 6

[30] A. Kowdle, S. Sinha, and R. Szeliski. Multiple view object cosegmentation using appearance and stereo cues. In *ECCV*, pages 789–803, 2012. 2, 6

[31] L. Ladicky, B. Zeisl, and M. Pollefeys. Discriminatively trained dense surface normal estimation. In *ECCV*, pages 468–484, 2014. 2

[32] D. Lee, M. Hebert, and T. Kanade. Geometric reasoning for single image structure recovery. In *CVPR*, pages 2136–2143, 2009. 2, 4, 6, 8

[33] G. Li and S. Zucker. Differential geometric inference in surface stereo. *IEEE TPAMI*, 32(1):72–86, 2010. 2

[34] C. Olsson, J. Ulén, and Y. Boykov. In defense of 3D-label stereo. In *CVPR*, pages 1730–1737, 2013. 2

[35] C. Rabe, T. Müller, A. Wedel, and U. Franke. Dense, robust, and accurate motion field estimation from stereo image sequences in real-time. In *ECCV*, pages 582–595, 2010. 1

[36] S. Ramalingam, J. Pillai, A. Jain, and Y. Taguchi. Manhattan junction catalogue for spatial reasoning of indoor scenes. In *CVPR*, pages 3065–3072, 2013. 2

[37] M. Rothermel, K. Wenzel, D. Fritsch, and N. Haala. SURE: Photogrammetric surface reconstruction from imagery. In *Proceedings LC3D Workshop, Berlin*, volume 8, 2012. 2

[38] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nesić, X. Wang, and P. Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *GCPR*, pages 31–42, 2014. 6

[39] D. Scharstein and R. Szeliski. Middlebury stereo vision page. http://vision.middlebury.edu/stereo/. 7

[40] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1-3):7–42, 2002. 2

[41] G. Schindler, P. Krishnamurthy, and F. Dellaert. Line-based structure from motion for urban environments. In *3DPVT*, pages 846–853, 2006. 2

[42] S. Sinha, D. Scharstein, and R. Szeliski. Efficient high-resolution stereo matching using local plane sweeps. In *CVPR*, 2014. 2

[43] S. Sinha, D. Steedly, and R. Szeliski. Piecewise planar stereo for image-based rendering. In *ICCV*, pages 1881–1888, 2009. 2

[44] T. Taniai, Y. Matsushita, and T. Naemura. Graph cut based continuous stereo matching using locally shared labels. In *CVPR*, pages 1613–1620, 2014. 2

[45] A. Tankus and N. Kiryati. Photometric stereo under perspective projection. In *ICCV*, volume 1, pages 611–616, 2005. 5

[46] K. Wenzel, M. Rothermel, N. Haala, and D. Fritsch. SURE–the ifp software for dense image matching. In *Photogrammetric Week*, pages 59–70, 2013. 2

[47] O. Woodford, P. Torr, I. Reid, and A. Fitzgibbon. Global stereo reconstruction under second-order smoothness priors. *IEEE TPAMI*, 31(12):2115–2128, 2009. 2

[48] K. Yamaguchi, T. Hazan, D. McAllester, and R. Urtasun. Continuous Markov random fields for robust stereo estimation. In *ECCV*, pages 45–58, 2012. 2

[49] J. Žbontar and Y. LeCun. Stereo matching by training a convolutional neural network to compare image patches. *JMLR*, 17(65):1–32, 2016. 2, 5, 7

[50] S. Zhang, W. Xie, G. Zhang, H. Bao, and M. Kaess. Robust stereo matching with surface normal prediction. In *ICRA*, 2017. 2

SUPPLEMENTARY MATERIALS

Semi-Global Stereo Matching with Surface Orientation Priors

Daniel Scharstein
Middlebury College

Tatsunori Taniai
RIKEN AIP

Sudipta N. Sinha
Microsoft Research

1. Other matching costs and resolutions

In addition to NCC we also evaluate MC-CNN [49] as a matching cost, on both full-resolution (6MP) and half-resolution (1.5MP) versions of the Middlebury training images. Fig. 1 shows barplots for baseline SGM, estimated priors SGM-EPi, and ground-truth priors SGM-GS, on all four combinations of matching cost and image resolution. Table 1 summarizes the average performance gains.

For SGM-EPi, which estimates plane hypotheses from coarser matching results, we run SGM with NCC matching costs at quarter resolution for all of these combinations. We found that MC-CNN is tuned for half resolution and does not work well at quarter resolution.

Fig. 1 and Table 1 show that SGM-EPi yields comparable average performance gains of 10-12% on all combinations except for MC-CNN at half resolution, which overall yields the lowest errors and thus smaller average gains. However, individual gains on difficult scenes such as Adirondack and Vintage remain high across all combinations. The oracle prior SGM-GS performs well across all combinations, with average error reduction of at least 50%. Overall, these results demonstrate that our SGM-P method has great potential independent of matching cost.

		% error reduction	
Cost	Resolution	SGM-EPi	SGM-GS
NCC	Full	12.1	54.8
	Half	10.1	52.8
MC-CNN	Full	10.7	51.1
	Half	3.8	50.0

Table 1. Average performance gains for different matching costs and image resolutions.

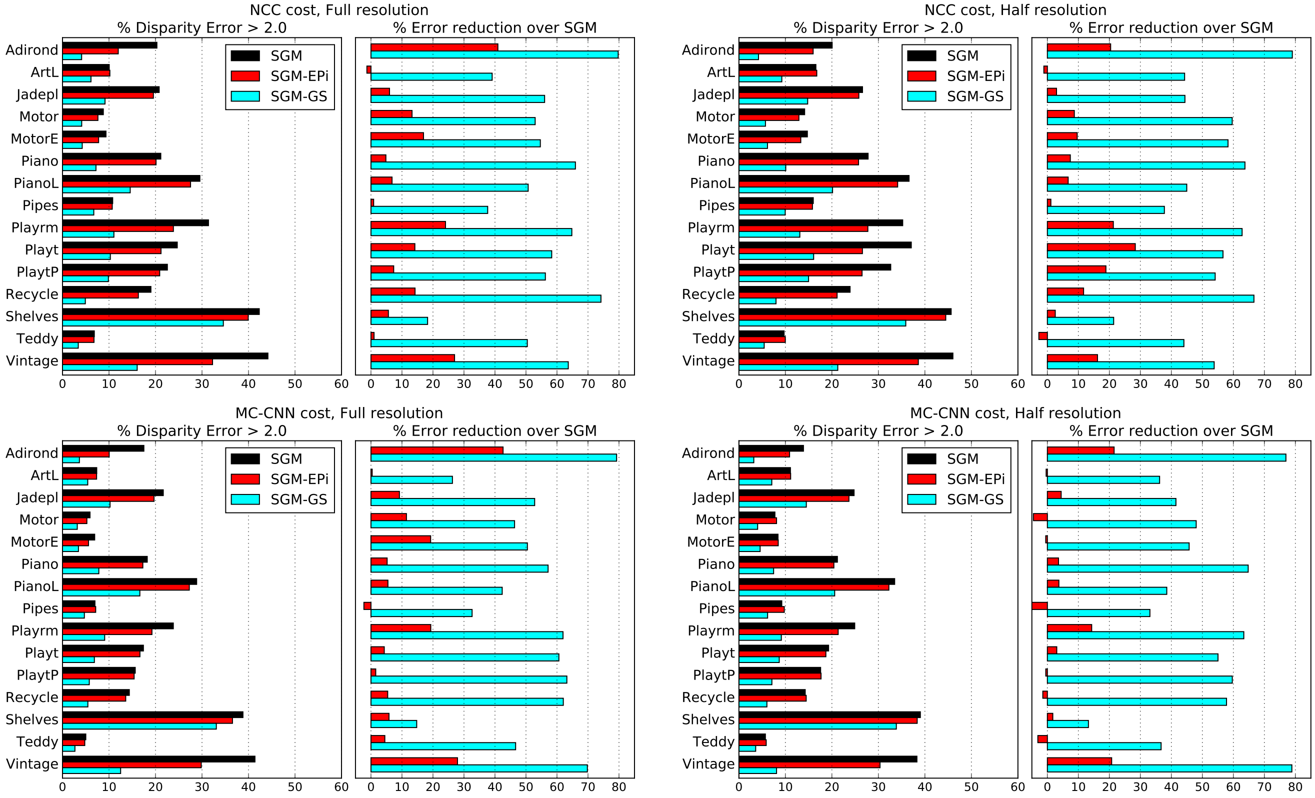


Figure 1. Comparison of matching costs NCC (top) and MC-CNN (bottom), at full (left) and half (right) resolution.

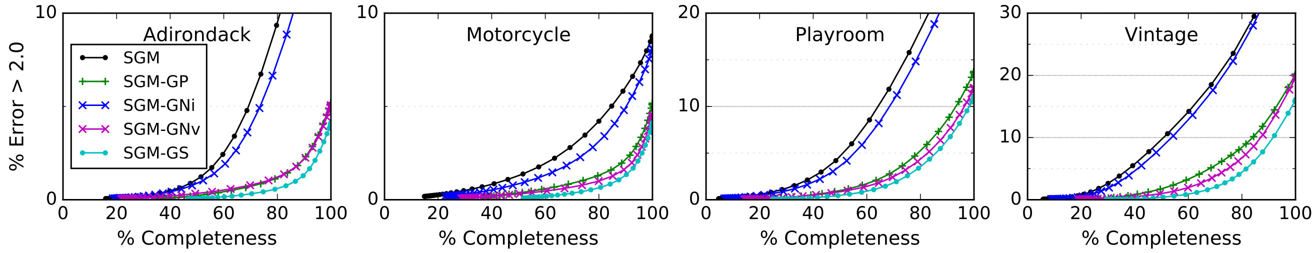
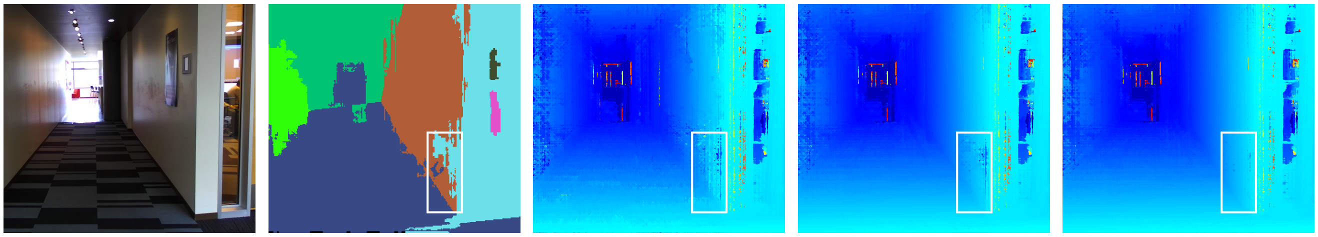


Figure 2. Comparison of oracle orientation priors SGM-GS (true surface), SGM-GP (planar approximation), SGM-GNv (3D prior derived from true surface normals), and SGM-GNi (2D “strawman” prior derived from true surface normals).



(a) Input image (b) Plane labels (c) SGM (d) SGM-EPi (e) SGM-EPv

Figure 3. Qualitative results using estimated priors. (a) Challenging image pair.(b) Planar priors extracted from noisy matching results at lower resolution. (c-e) Comparison of baseline (SGM) with estimated 2D priors (SGM-EPi) and 3D priors (SGM-EPv). Despite the low quality of the estimated planes, the priors result in significantly cleaner disparity maps compared to the baseline. In addition, SGM-EPv yields cleaner surface transitions (“creases”) in the presence of noisy labels and/or competing priors, for instance near the front edge of the right corridor wall (highlighted).

2. Oracle surface normal priors

To allow an accurate comparison of the achievable benefit of priors derived from surface normals (as opposed to the surfaces directly) we include additional experiments using oracle priors derived from ground-truth data.

In order to compute oracle priors from ground-truth surface normals (SGM-GNv), we integrate these normals in scene space under a perspective projection model using a least-squares approach [45]. The result is a z-surface in world coordinates, initially at an arbitrary depth. We implement the 2D integration step using a sparse solver employing conjugate gradients. We divide the image into a coarse grid and independently integrate a surface in each grid cell, arbitrarily fixing one depth value in each cell. In order for the integration to succeed, it is crucial that the discontinuities in the normal map are known. Otherwise any integration method, including our least-squares approach, will not produce a locally accurate z-surface.

Recall from Section 3.5 in the paper that a planar surface patch with known normal but unknown depth yields a family of planar disparity surfaces whose orientation depends on z, resulting in a 3D prior SGM-GNv. To investigate the importance of modeling this depth dependance, we compare the accurate 3D version SGM-GNv with an (inaccurate) 2D version SGM-GNi which we obtain by integrating the surface normals in scene space using an arbitrary starting depth, and using the resulting disparity surface as a 2D prior. We compare both the accurate 3D version SGM-GNv and the 2D “strawman” SGM-GNi with the ground-truth surface prior SGM-GS and its piecewise-planar approximation SGM-GP. Fig. 2 shows the performance of these variants. It can be observed that the accurate normal prior SMG-GNv is close to the upper bound SGM-GS, often outperforming the planar approximation SGM-GP, while the strawman SGM-GNi performs much worse. The difference between SGM-GNv and SGM-GNi is less pronounced on other image pairs where fewer slanted surfaces are present, or where the single integration result coincides with a large actual surface in the scene.

3. Estimated 2D vs. 3D priors

Recall from Table 2 in the paper that estimated 3D priors (SGM-EPv) perform quantitatively slightly better than 2D priors (SGM-EPi). Fig. 3 shows a qualitative example illustrating why 3D priors are advantageous. It can be seen that both SGM-EP methods clearly produce much cleaner surfaces than the baseline SGM algorithm. In addition, SGM-EPv produces smoother results especially near plane transitions at discontinuities and orientation changes (“creases”) between planes. Since SGM-EPi only has a single orientation prior per pixel, its performance degrades when the pixel-to-plane labeling is noisy or incomplete. SGM-EPv allows multiple overlapping disparity hypotheses and is thus more robust in the presence of noisy labels.