

Efficient Large-Scale Stereo Matching

Andreas Geiger¹, Martin Roser¹, and Raquel Urtasun²

¹ Dep. of Measurement and Control, Karlsruhe Institute of Technology

² Toyota Technological Institute at Chicago
geiger@kit.edu, martin.roser@kit.edu, rurtasun@ttic.edu

Abstract. In this paper we propose a novel approach to binocular stereo for fast matching of high-resolution images. Our approach builds a prior on the disparities by forming a triangulation on a set of support points which can be robustly matched, reducing the matching ambiguities of the remaining points. This allows for efficient exploitation of the disparity search space, yielding accurate dense reconstruction without the need for global optimization. Moreover, our method automatically determines the disparity range and can be easily parallelized. We demonstrate the effectiveness of our approach on the large-scale Middlebury benchmark, and show that state-of-the-art performance can be achieved with significant speedups. Computing the left and right disparity maps for a one Megapixel image pair takes about one second on a single CPU core.

1 Introduction

Estimating depth from binocular imagery is a core subject in low-level vision as it is an important building block in many domains such as multi-view reconstruction. In order to be of practical use for applications such as autonomous driving, disparity estimation methods should run at speeds similar to other low-level visual processing techniques, e.g. edge extraction or interest point detection. Since depth errors increase quadratically with the distance [1], high-resolution images are needed to obtain accurate 3D representations. While the benefits of high resolution imagery are already exploited exhaustively in structure-from-motion, object recognition and scene classification, only few binocular stereo methods deal efficiently with large images.

Stereo algorithms based on local correspondences [2, 3] are typically fast, but require an adequate choice of window size. As illustrated in Fig. 1 this leads to a trade-off between low matching ratios for small window sizes and border bleeding artifacts for larger ones. As a consequence, poorly-textured and ambiguous surfaces cannot be matched consistently.

Algorithms based on global correspondences [4–9] overcome some of the aforementioned problems by imposing smoothness constraints on the disparities in the form of regularized energy functions. Since optimizing such MRF-based energy functions is in general NP-hard, a variety of approximation algorithms have been proposed, e.g., graph cuts [4, 5] or belief propagation [6]. However, even on low-resolution imagery, they generally require large computational efforts and high

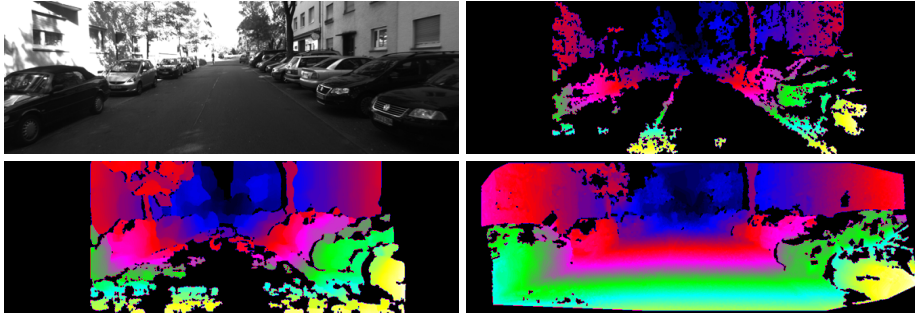


Fig. 1. Low-textured areas often pose problems to stereo algorithms. Using local methods one faces the trade-off between low matching ratios (top-right, window size 5×5) and border bleeding effects (bottom-left, window size 25×25). Our method is able to combine small window sizes with high matching ratios (bottom-right).

memory capacities. For example, storing all messages of a one Megapixel image pair requires more than 3 GB of RAM [10]. In these approaches, the disparity range usually has to be known in advance, and a good choice of the regularization parameters is crucial. Furthermore, when increasing image resolution, the widely used priors based on binary potentials fail to reconstruct poorly-textured and slanted surfaces, as they favor fronto-parallel planes. Recently developed methods based on higher-order cliques [7] overcome these problems, but are even more computationally demanding.

In this paper we propose a generative probabilistic model for stereo matching, called *ELAS* (*Efficient LArge-scale Stereo*)³, which allows for dense matching with small aggregation windows by reducing ambiguities on the correspondences. Our approach builds a prior over the disparity space by forming a triangulation on a set of robustly matched correspondences, named ‘support points’. Since our prior is piecewise linear, we do not suffer in the presence of poorly-textured and slanted surfaces. This results in an efficient algorithm that reduces the search space and can be easily parallelized. As demonstrated in our experiments, our method is able to achieve state-of-the-art performance with significant speedups of up to three orders of magnitude when compared to prevalent approaches; we obtain 300 MDE/s (million disparity evaluations per second) on a single CPU core.

2 Related work

In the past few years much progress has been made towards solving the stereo problem, as evidenced by the excellent overview of Scharstein et al. [2]. Local methods typically aggregate image statistics in a small window, thus imposing

³ C++ source code, Matlab wrappers and videos online at <http://www.cvlibs.net>

smoothness implicitly. Optimization is usually performed using a winner-takes-all strategy, which selects for each pixel the disparity with the smallest value under some distance metric [2]. Weber et al. [3] achieved real-time performance using the Census transform and a GPU implementation. However, as illustrated by Fig. 1, traditional local methods [11] often suffer from border bleeding effects or struggle with correspondence ambiguities. Approaches based on adaptive support windows [12, 13] adjust the window size or adapt the pixel weighting within a fixed-size window to improve performance, especially close to border discontinuities. Unfortunately, since for each pixel many weight factors have to be computed, these methods are much slower than fixed-window ones [13].

Dense and accurate matching can be obtained by global methods, which enforce smoothness explicitly by minimizing an MRF-based energy function which can be decomposed as the sum of a data fitting term and a regularization term. Since for most energies of practical use such an optimization is NP-hard, approximate algorithms have been proposed, e.g. graph-cuts [4, 5], belief propagation [6]. Klaus et al. [14] extend global methods to use mean-shift color segmentation, followed by belief propagation on super-pixels. In [15], a parallel VLSI hardware design for belief propagation that achieves real time performance on VGA imagery was proposed. The application of global methods to high-resolution images is, however, limited by their high computational and memory requirements, especially in the presence of large disparity ranges. Furthermore, models based on binary potentials between pixels favor fronto-parallel surfaces which leads to errors in low-textured slanted surfaces. Higher order cliques can overcome these problems [7], but they are even more computationally demanding.

Hirschmüller proposed semi-global matching [16], an approach which extends polynomial time 1D scan-line methods to propagate information along 16 orientations. While reducing streaking artifacts and improving accuracy compared to traditional methods based on dynamic programming, computational complexity increases with the number of computed paths. ‘ground control points’ are used in [17] to improve the occlusion cost sensitivity of dynamic programming algorithms. In [18, 19] disparities are ‘grown’ from a small set of initial correspondence seeds. Though these methods produce accurate results and can be faster than global approaches, they do not provide dense matching and struggle with textureless and distorted image areas. Approaches to reduce the search space have been investigated for global stereo methods [10, 20]. However, they mainly focus on memory requirements and start with a full search using local methods first. Furthermore, the use of graph-cuts imposes high computational costs particularly for large-scale imagery.

In contrast, in this paper we propose a Bayesian approach to stereo matching that is able to compute accurate disparity maps of high resolution images at frame rates close to real time without the need for global optimization. The remainder of this paper is structured as follows: In Section 3 we describe our approach to efficient large-scale stereo matching. Experimental results on real-world datasets and comparisons to a variety of other methods on large-scale

versions of the Middlebury benchmark images are reported in Section 4. Finally, Section 5 gives our conclusions and future work.

3 Efficient Large-Scale Stereo Matching

In this section we describe our approach to efficient stereo matching of high-resolution images. Our method is inspired from the observation that despite the fact that many stereo correspondences are highly ambiguous, some of them can be robustly matched. Assuming piecewise smooth disparities, such reliable ‘support points’ contain valuable prior information for the estimation of the remaining ambiguous disparities. Our approach proceeds as follows: First, the disparities of a sparse set of support points are computed using a full disparity range. The image coordinates of the support points are then used to create a 2D mesh via Delaunay triangulation. A prior is computed to disambiguate the matching problem, making the process efficient by restricting the search to plausible regions. In particular, this prior is formed by computing a piecewise linear function induced by the support point disparities and the triangulated mesh. For simplicity of the presentation, we will assume rectified input images, such that correspondences are restricted to the same line in both images.

3.1 Support Points

As support points, we denote pixels which can be robustly matched due to their texture and uniqueness. While a variety of methods for obtaining stable correspondences are available [17, 21, 22], we find that matching support points on a regular grid using the ℓ_1 distance between vectors formed by concatenating the horizontal and vertical Sobel filter responses of 9×9 pixel windows to be both efficient and effective. In all of our experiments we used Sobel masks of size 3×3 and a grid with fixed step-size of 5 pixels. A large disparity search range of half the input image width was employed to impose no restrictions on the disparities. We also experimented with sparse interest point descriptors such as SURF [23], but found that they did not improve matching accuracy while being slower to compute.

For robustness we impose consistency, i.e., correspondences are retained only if they can be matched from left-to-right and right-to-left. To get rid of ambiguous matches, we eliminate all points whose ratio between the best and the second best match exceeds a fixed threshold, $\tau = 0.9$. Spurious mismatches are removed by deleting all points which exhibit disparity values dissimilar from all surrounding support points. To cover the full image, we add additional support points at the image corners whose disparities are taken to be the ones of their nearest neighbors.

3.2 Generative Model for Stereo Matching

We now describe our probabilistic generative model which, given a reference image and the support points, can be used to draw samples from the other im-

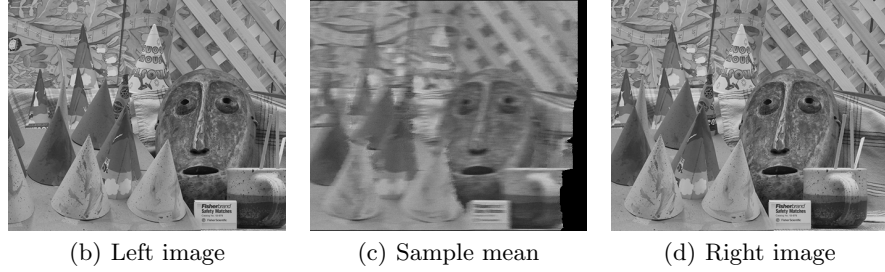
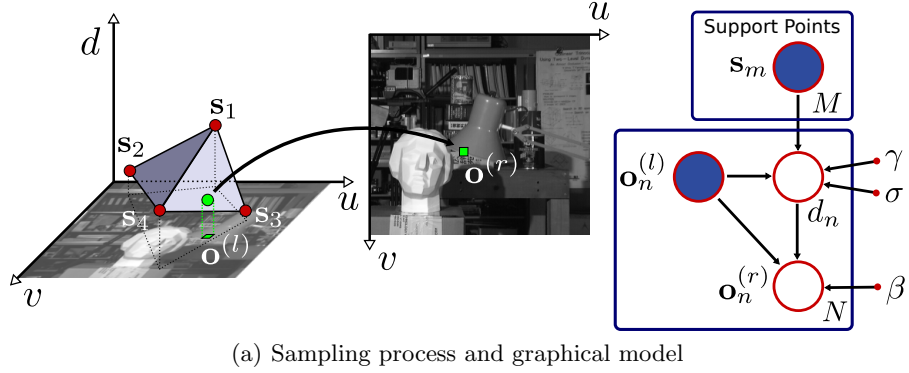


Fig. 2. Illustration of the sampling process. (a) Graphical model and sampling process: Given support points $\{\mathbf{s}_1, \dots, \mathbf{s}_M\}$ and an observation in the left image $\mathbf{o}_n^{(l)}$, a disparity d is drawn. Given the observation on the left image and the disparity, we can draw an observation in the right image $\mathbf{o}_n^{(r)}$. (c) Repeating this process 100 times for each pixel and (d) computing the mean results in a blurred version of the right image.

age. More formally, let $\mathbf{S} = \{\mathbf{s}_1, \dots, \mathbf{s}_M\}$ be a set of robustly matched support points. Each support point, $\mathbf{s}_m = (u_m, v_m, d_m)^T$, is defined as the concatenation of its image coordinates, $(u_m, v_m) \in \mathbb{N}^2$, and its disparity, $d_m \in \mathbb{N}$. Let $\mathbf{O} = \{\mathbf{o}_1, \dots, \mathbf{o}_N\}$ be a set of image observations, with each observation $\mathbf{o}_n = (u_n, v_n, \mathbf{f}_n)^T$ formed as the concatenation of its image coordinates, $(u_n, v_n) \in \mathbb{N}^2$, and a feature vector, $\mathbf{f}_n \in \mathbb{R}^Q$, e.g., the pixel's intensity or a low-dimensional descriptor computed from a small neighborhood. We denote $\mathbf{o}_n^{(l)}$ and $\mathbf{o}_n^{(r)}$ as the observations in the left and right image respectively. Without loss of generality, in the following we consider the left image as the reference image.

Assuming that the observations $\{\mathbf{o}_n^{(l)}, \mathbf{o}_n^{(r)}\}$ and support points \mathbf{S} are conditionally independent given their disparities d_n , the joint distribution factorizes

$$p(d_n, \mathbf{o}_n^{(l)}, \mathbf{o}_n^{(r)}, \mathbf{S}) \propto p(d_n | \mathbf{S}, \mathbf{o}_n^{(l)}) p(\mathbf{o}_n^{(r)} | \mathbf{o}_n^{(l)}, d_n) \quad (1)$$

with $p(d_n | \mathbf{S}, \mathbf{o}_n^{(l)})$ the prior and $p(\mathbf{o}_n^{(r)} | \mathbf{o}_n^{(l)}, d_n)$ the image likelihood. The graphical model of our approach is depicted in Fig. 2(a). In particular, we take the prior to be proportional to a combination of a uniform distribution and a sam-

pled Gaussian

$$p(d_n | \mathbf{S}, \mathbf{o}_n^{(l)}) \propto \begin{cases} \gamma + \exp\left(-\frac{(d_n - \mu(\mathbf{S}, \mathbf{o}_n^{(l)}))^2}{2\sigma^2}\right) & \text{if } |d_n - \mu| < 3\sigma \vee d_n \in N_{\mathbf{S}} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

with $\mu(\mathbf{S}, \mathbf{o}_n^{(l)})$ a mean function linking the support points and the observations, and $N_{\mathbf{S}}$ the set of all support point disparities in a small 20×20 pixel neighborhood around $(u_n^{(l)}, v_n^{(l)})$. We gain efficiency by excluding all disparities farther than 3σ from the mean. The condition $d_n \in N_{\mathbf{S}}$ enables the prior to locally extend its range to better handle disparity discontinuities in places where the linearity assumption might be violated.

We express $\mu(\mathbf{S}, \mathbf{o}_n^{(l)})$ as a piecewise linear function, which interpolates the disparities using the Delaunay triangulation computed on the support points. For each triangle, we thus obtain a plane defined by

$$\mu_i(\mathbf{o}_n^{(l)}) = a_i u_n + b_i v_n + c_i \quad (3)$$

where i is the index of the triangle the pixel (u_n, v_n) belongs to, and $\mathbf{o}_n = (u_n, v_n, \mathbf{f}_n)^T$ is an observation. For each triangle, the plane parameters (a_i, b_i, c_i) are easily obtained by solving a linear system. Hence, the mode of the proposed prior, μ , is a linear interpolation between support point disparities, serving as a coarse representation.

We express the image likelihood as a constrained Laplace distribution

$$p(\mathbf{o}_n^{(r)} | \mathbf{o}_n^{(l)}, d_n) \propto \begin{cases} \exp(-\beta \|\mathbf{f}_n^{(l)} - \mathbf{f}_n^{(r)}\|_1) & \text{if } \begin{pmatrix} u_n^{(l)} \\ v_n^{(l)} \end{pmatrix} = \begin{pmatrix} u_n^{(r)} + d_n \\ v_n^{(r)} \end{pmatrix} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where $\mathbf{f}_n^{(l)}, \mathbf{f}_n^{(r)}$ are feature vectors in the left and right image respectively, and β is a constant. The if-condition ensures that correspondences are located on the same epipolar line and matched via the disparity d_n . In our experiments, the features \mathbf{f}_n are taken as the concatenation of image derivatives in a 5×5 pixel neighborhood around (u_n, v_n) , computed from Sobel filter responses, leading to $2 \times 5 \times 5 = 50$ -dimensional feature vectors. Note that in [14] similar features were shown to be robust to illumination changes (i.e., additive bias). Empirically, we found features based on Sobel responses to work noticeably better than features based on Laplacian-of-Gaussian (LoG) filters. We do not make use of any color information, although this could be incorporated with little additional effort. We refer the reader to [24, 2] for a comprehensive study of dissimilarity metrics for stereo matching.

An advantage of having a generative model is that we can use it to draw samples, see Fig. 2(a) for an illustration. Given the support points and an observation in the left image, samples from the corresponding observation in the right image can be obtained as follows:

1. Given \mathbf{S} and $\mathbf{o}_n^{(l)}$ draw a disparity d_n from $p(d_n|\mathbf{S}, \mathbf{o}_n^{(l)})$
2. Given $\mathbf{o}_n^{(l)}$ and d_n draw an observation $\mathbf{o}_n^{(r)}$ from $p(\mathbf{o}_n^{(r)}|\mathbf{o}_n^{(l)}, d_n)$

Fig. 2(b-d) depicts the left input image, as well as the mean of the samples drawn from the right image given the left image and the support points. In order to obtain a comprehensive visualization in Fig. 2, here we use pixel intensities as features and draw 100 samples for each pixel. As expected, the sample mean corresponds to a blurred version of the right image.

3.3 Disparity Estimation

In the previous section we have proposed a prior and an image likelihood for stereo matching. We have also shown how to draw samples of the right image given support points and observations in the left image. At inference, however, we are interested in estimating the disparity map given the left and right images. We rely on maximum a-posteriori (MAP) estimation to compute the disparities

$$d_n^* = \operatorname{argmax} p(d_n|\mathbf{o}_n^{(l)}, \mathbf{o}_1^{(r)}, \dots, \mathbf{o}_N^{(r)}, \mathbf{S}), \quad (5)$$

where $\mathbf{o}_1^{(r)}, \dots, \mathbf{o}_N^{(r)}$ denotes all observations in the right image which are located on the epipolar line of $\mathbf{o}_n^{(l)}$. The posterior can be factorized as

$$p(d_n|\mathbf{o}_n^{(l)}, \mathbf{o}_1^{(r)}, \dots, \mathbf{o}_N^{(r)}, \mathbf{S}) \propto p(d_n|\mathbf{S}, \mathbf{o}_n^{(l)})p(\mathbf{o}_1^{(r)}, \dots, \mathbf{o}_N^{(r)}|\mathbf{o}_n^{(l)}, d_n). \quad (6)$$

The observations along the epipolar line on the right image are structured, i.e., given a disparity associated with $\mathbf{o}_n^{(l)}$, there is a deterministic mapping to which observations have non-zero probability on the line. We capture this property by modeling the distribution over all the observations along the epipolar line as

$$p(\mathbf{o}_1^{(r)}, \dots, \mathbf{o}_N^{(r)}|\mathbf{o}_n^{(l)}, d_n) \propto \sum_{i=1}^N p(\mathbf{o}_i^{(r)}|\mathbf{o}_n^{(l)}, d_n). \quad (7)$$

Note that from Eq. (4), there is only one observation with non-zero probability for each d_n . Plugging Eq. (2) and (4) into Eq. (6) and taking the negative logarithm yields an energy function that can be easily minimized

$$E(d) = \beta \|\mathbf{f}^{(l)} - \mathbf{f}^{(r)}(d)\|_1 - \log \left[\gamma + \exp \left(-\frac{[d - \mu(\mathbf{S}, \mathbf{o}^{(l)})]^2}{2\sigma^2} \right) \right] \quad (8)$$

with $\mathbf{f}^{(r)}(d)$ the feature vector located at pixel $(u^{(l)} - d, v^{(l)})$. Note that from the definition of the image likelihood, the energy $E(d)$ is required to be evaluated only if $|d - \mu| < 3\sigma$, or d is an element of the neighboring support point disparities. A dense disparity map can be obtained by minimizing Eq. (8). Importantly, this can be done in parallel for each pixel as the support points decouple the different observations.

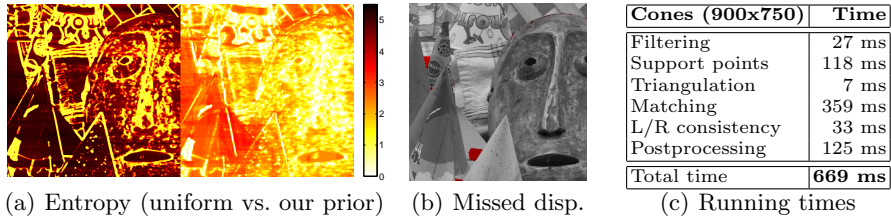


Fig. 3. (a) Entropy of the posterior in Eq. (5) in nats using a uniform prior and our method for a subset of the *Cones* image pair. (b) Pixels for which our proposal distribution does not contain the ground truth disparity are shown in red. (c) Running time of the individual parts of our algorithm for an image of size 900×750 pixels on a single CPU core.

Although in this section we have focused on obtaining the disparity map of the right image, similarly one can obtain the left disparity map. In practice, we apply our approach to both images, and perform a left/right consistency check to eliminate spurious mismatches and disparities in occluded regions. Following [16] we also remove small segments with an area smaller than 50 pixels.

4 Experimental Evaluation

In this section we compare our approach to state-of-the-art methods in terms of accuracy and running time. Throughout all experiments we set $\beta = 0.03$, $\sigma = 3$, $\gamma = 15$ and $\tau = 0.9$ which were found to empirically perform well. We employed the library ‘Triangle’ [25] to compute the triangulations. All experiments were conducted on a single i7 CPU core running at 2.66 GHz. Since our goal is to achieve near real time rates on high resolution imagery, we compare all approaches on large images in the Middlebury dataset. This is in contrast to the 450×375 resolution typically used in the literature.

4.1 Entropy Reduction

We first evaluate the quality of our prior by evaluating the entropy of the posterior in Eq. (6). We expect a good prior to reduce the matching entropy since it gets rid of ambiguous matches. Fig. 3(a) shows the posterior entropy in nats on the *Cones* image set from the Middlebury dataset. Our approach disambiguates the problem as the matching entropy is significantly lower compared to using a uniform prior. As shown in Fig. 3(b), the pixels for which our prior is erroneous mainly occur in occluded regions. Notably, our algorithm is able to compute the disparity maps for the left and right image in 0.6 s. An overview of the running times of the different parts of our algorithm is shown in Fig. 3(c). Post-processing refers to the aforementioned removal of small segments and constant interpolation of missing disparities in order to obtain dense disparity maps.

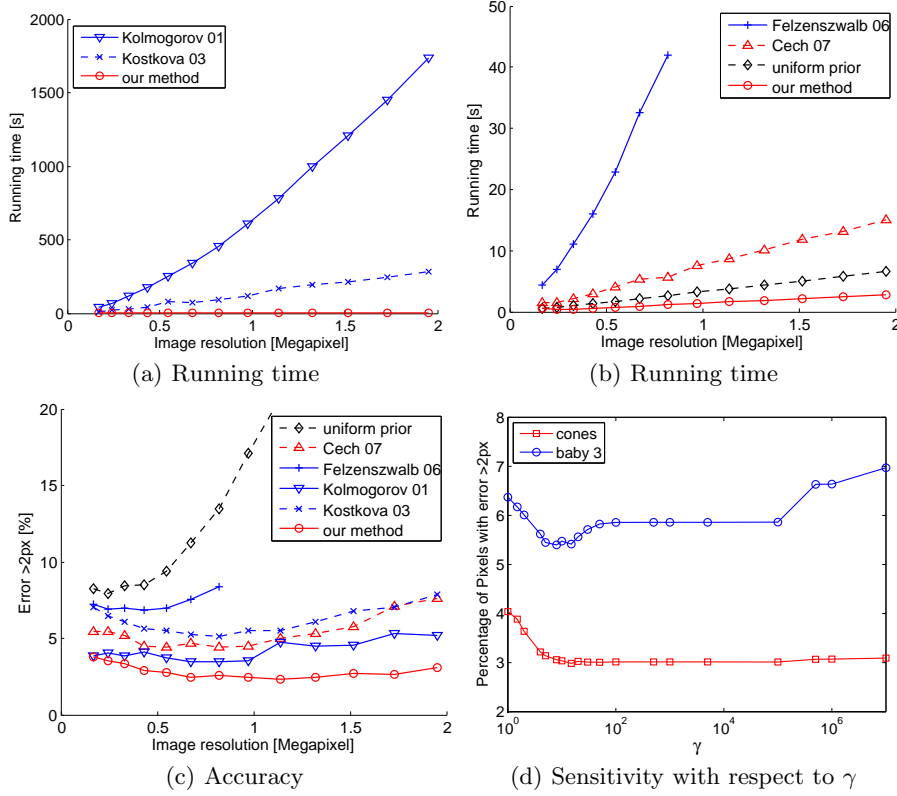


Fig. 4. Comparison to state-of-the-art and sensitivity on the *Cones* image pair. (a,b) Running time and (c) accuracy as a function of the image resolution. (d) Sensitivity of our algorithm wrt. γ .

4.2 Accuracy and Running Time for the Middlebury Dataset

We compare our approach to a wide range of baselines on medium to high resolution images from the Middlebury benchmark. In particular, we compare against two global methods [5, 6] and two seed-and-grow algorithms [19, 18], using their publicly available implementations⁴. We also compare our prior to a *uniform prior* that uses the same image likelihood as our method, and selects disparities using a winner-takes-all strategy over the whole disparity range.

For graph-cuts with occlusion handling [5] we use $K = 60$, $\lambda_1 = 40$, $\lambda_2 = 20$, the l_2 norm with an intensity threshold of 8 in combination with the Birchfield-Tomasi dissimilarity measure. We iterate until convergence to obtain best results. These parameters are set based on suggestions in [5] and adaptations to the large-scale imagery we use. We run efficient hierarchical belief propagation [6] using 4

⁴ <http://people.cs.uchicago.edu/~pff/>, <http://www.cs.ucl.ac.uk/staff/V.Kolmogorov/>, <http://cmp.felk.cvut.cz/~stereo/>, <http://cmp.felk.cvut.cz/cechj/GCS/>


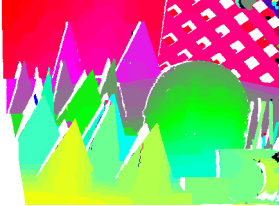
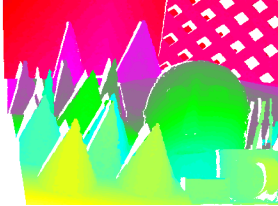


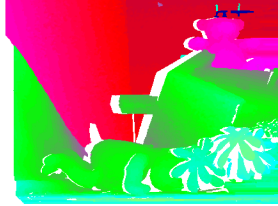
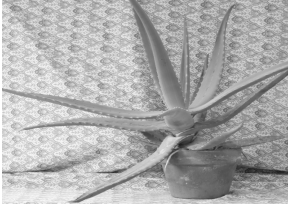
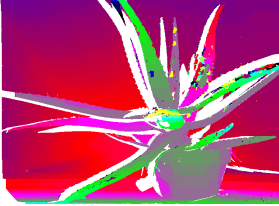
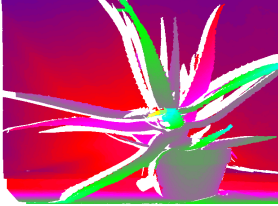
	left image		Kolmogorov 01		our method					
										
										
										
	Cones	Teddy	Art	Aloe	Dolls	Baby3	Cloth3	Lamp2	Rock2	
Image width	900	900	1390	1282	1390	1312	1252	1300	1276	
Image height	750	750	1110	1110	1110	1110	1110	1110	1110	
Support points	3236	3095	6164	6268	8241	5901	6805	4424	6670	
Correct points	99.2%	99.1%	99.1%	99.8%	99.4%	98.7%	100.0%	98.9%	100.0%	
Triangles	6376	6128	12237	12417	16353	11689	13473	8769	13215	
Missed pixels	0.7%	4.0%	5.3%	1.6%	1.1%	1.0%	0.4%	9.7%	0.6%	
Non-occluded pixels: Error > 1										
uniform prior	18.0%	37.5%	43.0%	12.8%	33.1%	49.4%	7.6%	74.9%	7.8%	
Felzenszwalb 06	15.2%	18.7%	23.3%	12.8%	20.9%	13.0%	6.1%	32.0%	7.6%	
Kolmogorov 01	8.2%	16.5%	30.3%	13.5%	28.7%	26.2%	4.3%	65.7%	10.4%	
Cech 07	7.2%	15.8%	18.8%	9.2%	19.8%	17.4%	2.8%	36.7%	3.6%	
Kostkova 03	7.2%	13.5%	17.9%	7.2%	14.4%	14.2%	2.7%	31.5%	3.0%	
our method	5.0%	11.5%	13.3%	5.0%	11.0%	10.8%	1.4%	17.5%	1.9%	
Non-occluded pixels: Error > 2										
uniform prior	16.4%	35.0%	41.1%	11.3%	29.6%	46.9%	7.3%	74.2%	7.3%	
Felzenszwalb 06	7.8%	11.4%	16.5%	7.8%	10.5%	7.0%	3.5%	26.0%	3.1%	
Kolmogorov 01	4.1%	8.1%	21.0%	8.1%	17.0%	19.0%	1.8%	60.7%	6.0%	
Cech 07	4.4%	10.2%	11.2%	4.8%	10.6%	9.7%	1.8%	27.1%	2.1%	
Kostkova 03	5.3%	10.1%	13.0%	4.8%	8.2%	8.2%	2.2%	26.7%	2.2%	
our method	2.7%	7.3%	8.7%	3.0%	5.3%	4.5%	0.9%	10.4%	1.0%	

Fig. 5. Results on Middlebury data set. White regions highlight occluded areas.

scales, 10 iterations per scale, $\lambda = 0.1$ and $\sigma = 1$. For the seed-and-grow methods [19, 18] we set $\alpha = \beta = 0$, which empirically gave best performance. For all baselines which are not able to estimate the disparity search range automatically, we set it manually to the largest disparity in the ground truth. We use the aforementioned large search space for our method and [19, 18]. To obtain dense results for all methods, missing disparities are interpolated using a piecewise constant function on the smallest valid neighbor in the same image line.

We first evaluate running time and accuracy for varying resolutions ranging from 0.16 Megapixel (Middlebury benchmark size) to 2 Megapixels on the *Cones* image pair. For best performance, we adjust the smoothing parameter of the global methods to adapt linearly with the image scale. Since hallucinating occluded areas is not the focus of this paper, we evaluate the error in all non-occluded regions, i.e. all non-white pixels in Fig. 5. As shown in Fig. 4(a,b), our method, which takes about three seconds to process a 2 Megapixel image, runs up to three orders of magnitude faster than global methods. Note that while when using a uniform prior the error rate increases quickly with image size due to an increase in the number of ambiguities, as depicted by Fig. 4(c), our method’s performance remains constant. Also note that due to memory limitations, we were only able to process images up to 0.8 Megapixels for [6]. Fig. 4(d) shows the sensitivity of our method with respect to the choice of γ on the *Cones* and *Baby3* image pairs. While our approach is insensitive to the precise value of γ , we observed best performance for $\gamma = 15$, especially for poorly-textured images.

We also compare the accuracy of our approach to the baselines in a variety of stereo images from the Middlebury data set [2], i.e., *Cones*, *Teddy*, *Art*, *Aloe*, *Dolls*, *Baby3*, *Cloth3*, *Lampshade2* and *Rock2*. As before, we use the non-occluded pixels to compute error rates. The top row of Fig. 5 depicts the left camera image, the disparity maps created using Kolmogorov’s graph-cuts [5] and our method. The upper rows of the table depict statistics of the input images as well as of our prior. We refer as ‘correct points’ the ratio of correctly matched support points and ‘missed pixels’ as the amount of ground truth disparities not contained in the prior; this is a lower bound on the error of our method. Note that, for most of the images, more than 98% of the correct disparities are included in our prior. A comparison to the baselines is depicted in the bottom two rows of the table. Due to memory limitations, we downsampled the images bicubically by $\frac{2}{3}$ for [6]. The final disparity maps were up-sampled again, using the nearest neighbor disparities. Note that even though our method mainly aims for efficient matching, for all images we perform competitively to global methods. As expected, smallest errors are achieved for highly textured objects: *Cloth3* and *Rock2*. Worst performance is obtained for the *Lamp2* image pair, as it is poorly textured. The bad performance of the uniform prior is mostly due to its inability to match textureless regions. In contrast, our approach handles them via the proposed prior. This is particularly evidenced in the *Dolls*, *Baby3* and *Lamp2* image pairs. Note that even though the graph-cuts baseline is able to capture disparity discontinuities more accurately than our method, it is not able

to perform very well overall. This is mainly due to the inability of the Potts model to truthfully recover poorly-textured slanted surfaces, as shown in Fig. 5.

4.3 Urban Sequences and Face Image Set

We now demonstrate the effectiveness of our approach on challenging real-world matching problems using high resolution imagery (1382×512 pixels) recorded from a moving vehicle. Typical challenges are poorly-textured regions and sensor saturation. While some regions are unmatched due to half occlusions (i.e., disparities depicted in black), most of the scene is accurately estimated by our approach. For this dataset, we obtain frame rates of ≥ 2 fps on a single core. Real time 3D scene reconstruction is made possible by computing one disparity map every 0.5 seconds, or by exploiting the parallel nature of our algorithm.

The bottom row of Fig. 6 shows the disparity map of a 1 Megapixel face image [26], generated by our algorithm. Importantly, note that all computations took only 1 second. Fine facial structures are clearly visible, while only few ambiguous regions could not be matched.

5 Conclusion and Future Work

In this paper we have proposed a Bayesian approach to stereo matching that is able to compute accurate disparity maps of high resolution images at frame rates close to real time. We have shown that a prior distribution estimated from robust support points can decrease stereo matching ambiguities. Our experiments on the Middlebury benchmark and real-world imagery show that our approach performs comparably to state-of-the-art approaches, while being orders of magnitude faster. Importantly, exploiting the parallel nature of our algorithm e.g., in a GPU implementation, will enable real time stereo matching at resolutions above 1 Megapixel. To increase the robustness of our method and account for small surfaces, we intend to study the use of adaptive support windows to estimate support points. We further plan to incorporate scene segmentation to better model disparity discontinuities.

Acknowledgement. We thank the reviewers for their feedback, Thabo Beeler for providing the face test image and the *Karlsruhe School of Optics and Photonics* and the *Deutsche Forschungsgemeinschaft* for supporting this work.

References

1. Gallup, D., Frahm, J.M., Mordohai, P., Pollefeys, M.: Variable baseline/resolution stereo. In: CVPR. (2008) 1–8
2. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. Journal of Computer Vision* **47** (2002) 7–42
3. Weber, M., Humenberger, M., Kubinger, W.: A very fast census-based stereo matching implementation on a graphics processing unit. In: IEEE Workshop on Embedded Computer Vision. (2009)

4. Boykov, Y., Veksler, O., Zabih, R.: Markov random fields with efficient approximations. In: CVPR. (1998) 648–655
5. Kolmogorov, V., Zabih, R.: Computing visual correspondence with occlusions using graph cuts. In: In International Conference on Computer Vision. (2001) 508–515
6. Felzenszwalb, P., Huttenlocher, D.: Efficient belief propagation for early vision. *International Journal of Computer Vision* **70** (2006) 41–54
7. Woodford, O., Torr, P., Reid, I., Fitzgibbon, A.: Global stereo reconstruction under second-order smoothness priors. *PAMI* **31** (2009) 2115–2128
8. Cheng, L., Caelli, T.: Bayesian stereo matching. *Computer Vision and Image Understanding* **106** (2007) 85–96
9. Kong, D., Tao, H.: Stereo matching via learning multiple experts behaviors. In: BMVC. (2006) –97
10. Wang, L., Jin, H., Yang, R.: Search space reduction for mrf stereo. In: European Conference on Computer Vision. (2008)
11. Konolige, K.: Small vision system. hardware and implementation. In: International Symposium on Robotics Research. (1997) 111–116
12. Kanade, T., Okutomi, M.: A stereo matching algorithm with an adaptive window: Theory and experiment. *ICRA* (1994)
13. Yoon, K.J., Member, S., Kweon, I.S.: Adaptive support-weight approach for correspondence search. *PAMI* **28** (2006) 650–656
14. Klaus, A., Sormann, M., Karner, K.: Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In: ICPR. (2006)
15. Liang, C.K., Cheng, C.C., Lai, Y.C., Chen, L.G., Chen, H.H.: Hardware-efficient belief propagation. In: Computer Vision and Pattern Recognition. (2009)
16. Hirschmüller, H.: Stereo processing by semiglobal matching and mutual information. *PAMI* **30** (2008) 328–341
17. Bobick, A.F., Intille, S.S.: Large occlusion stereo. *International Journal of Computer Vision* **33** (1999) 181–200
18. Cech, J., Sára, R.: Efficient sampling of disparity space for fast and accurate matching. In: Computer Vision and Pattern Recognition. (2007)
19. Kostkova, J., Sára, R.: Stratified dense matching for stereopsis in complex scenes. In: BMVC. (2003)
20. Veksler, O.: Reducing search space for stereo correspondence with graph cuts. In: British Machine Vision Conference. (2006)
21. Xiaoyan Hu, P.M.: Evaluation of stereo confidence indoors and outdoors. In: CVPR. (2010)
22. Sára, R.: Finding the largest unambiguous component of stereo matching. In: ECCV, London, UK, Springer-Verlag (2002) 900–914
23. Bay, H., Tuytelaars, T., Gool, L.V.: Surf: Speeded up robust features. In: ECCV, Graz Austria (2006)
24. Hirschmüller, H., Scharstein, D.: Evaluation of cost functions for stereo matching. In: CVPR. (2007) 1–8
25. Shewchuk, J.R. In: *Applied Computational Geometry: Towards Geometric Engineering*. Volume 1148. Springer, Berlin (1996) 203–222
26. Beeler, T., Bickel, B., Beardsley, P., Sumner, B., Gross, M.: High-quality single-shot capture of facial geometry. *SIGGRAPH* **29** (2010)

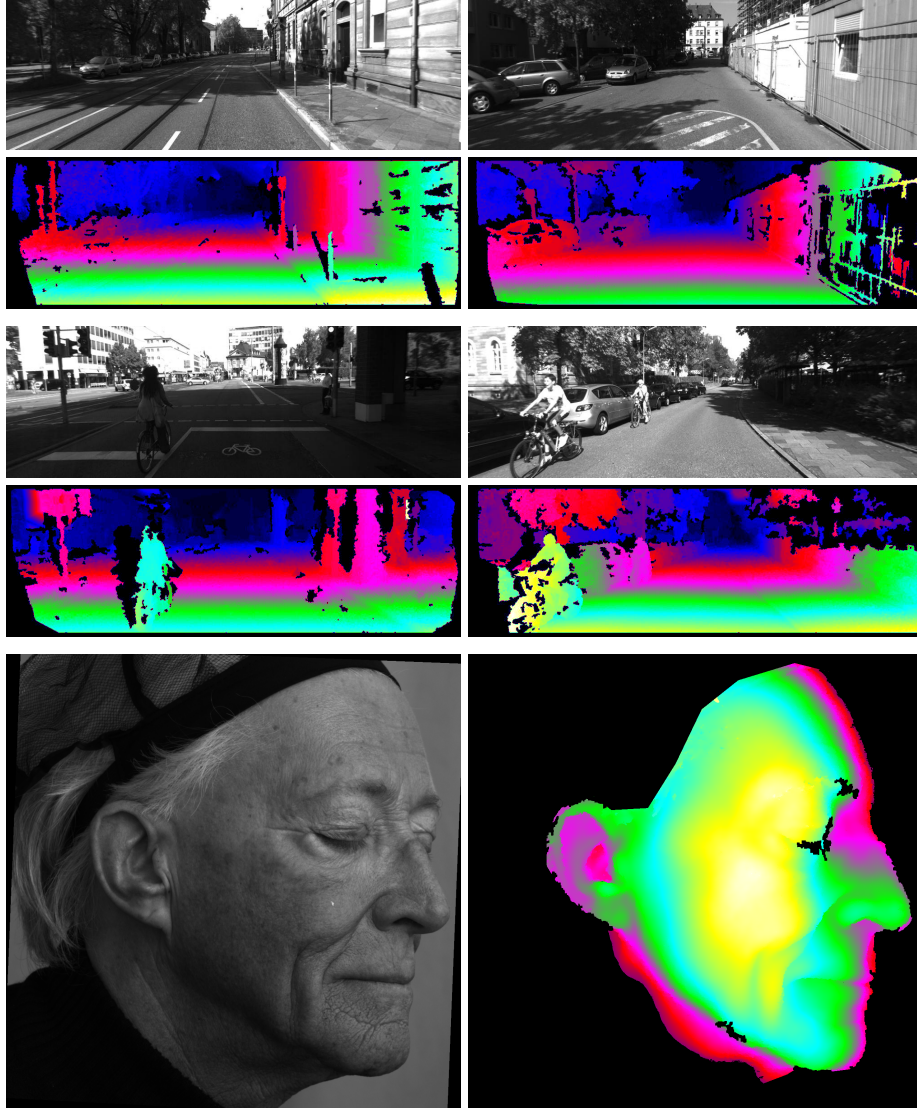


Fig. 6. Challenging urban scene and face dataset. We evaluate our algorithm on an urban video sequence at 1382×512 pixels resolution. By computing disparity maps at 2 fps and using visual odometry with sparse features, we were also able to obtain a 3D scene reconstruction in real time (see <http://www.cvlibs.net>). The bottom row shows a one Megapixel face image from the dataset of [26] and the corresponding disparity map obtained using our algorithm. Best viewed in color.