Efficient Minimal-Surface Regularization of Perspective Depth Maps in Variational Stereo

Gottfried Graber[1] Jonathan Balzer[2,4] Stefano Soatto[2] Thomas Pock[1,3]
[1]Graz University of Technology [2]UCLA [3]AIT Austrian Institute of Technology [4]Vathos GmbH
{graber,pock}@icg.tugraz.at, {jbalzer,soatto}@ucla.edu

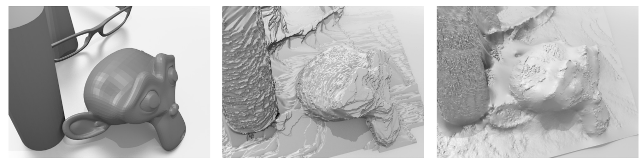
Abstract

We propose a method for dense three-dimensional surface reconstruction that leverages the strengths of shape-based approaches, by imposing regularization that respects the geometry of the surface, and the strength of depth-map-based stereo, by avoiding costly computation of surface topology. The result is a near real-time variational reconstruction algorithm free of the staircasing artifacts that affect depth-map and plane-sweeping approaches. This is made possible by exploiting the gauge ambiguity to design a novel representation of the regularizer that is linear in the parameters and hence amenable to be optimized with state-of-the-art primal-dual numerical schemes.

1. Introduction

Reconstructing three-dimensional (3-d) scenes from multiple images, including stereo or video, is intrinsically ill-posed as there are infinitely many surfaces that are compatible with the images. The problem is usually cast as an optimization problem within the calculus of variation, whereby the choice of which solution to pick among the infinitely many possible is determined by a *regularizer*. This is a functional that penalizes solutions that do not respect prior assumptions, and plays a key role both in the quality of the reconstruction, as well as in the efficiency of the numerical optimization scheme.

The most principled approaches to 3-d reconstruction aim to infer a collection of (multiply-connected, piecewise smooth) surfaces directly, represented intrinsically without regards to the images [2, 10, 18, 28, 38, 21, 42], as evident by the large body of literature on shape space and shape optimization. In these methods, both the geometry and the topology is then inferred to fit the available images. This is desirable as one can enforce priors on the surfaces based on physically meaningful regularizers. The disadvantage is that inferring topology is difficult and requires computation of visibility at each iteration of the algorithm, with obvious



(a) Ground truth (b) TV regularization (c) Our method

Figure 1. TV regularization tends to favor piecewise-constant functions, which is detrimental in the case of depth maps that represent a 3-d surface. Our regularizer, while being defined on the image, just as TV, respects the inner geometry of the 3-d surface.

repercussions on computational efficiency.

On the other hand, one could use the image plane to parametrize *charts* on the scene, corresponding to *depth maps* that associate a positive number (distance) to each pixel [7, 11, 13, 14, 19, 32, 36, 37]. Such depth maps will have to be combined in an additional fusion step to yield the global surface. The advantage of using depth maps is that they conveniently confine the data (images), the optimization variable (surfaces) and hence the objective function to the same domain, the image plane of a reference view. This makes computation efficient, and the method of choice for real-time applications. Unfortunately, the image plane is not the natural place to enforce regularization of the surface. In the vicinity of depth discontinuities, caused by occlusions, neighboring pixels do not necessarily correspond to points which are close in 3-d space. Thus typical image-based regularizers, such as total variation (TV), favor piecewise fronto-parallel depth maps, resulting in staircasing artifacts (Fig. 1(b)).

In this paper, we seek to combine the advantages of shape space methods with those of range maps. The advantage of the former is the availability of surface-based, physically plausible regularizers; the use of range maps allows us to avoid the inference of scene topology. To the best of our knowledge, this has not been done in the literature on variational stereo and is made possible by a number of technical contributions summarized in Sect. 1.1.

In addition to variational reconstruction algorithms described thus far, the research on 3-d reconstruction spawned a variety of methods that seek to bypass the complexities of computing topology or visibility by localizing the surface representation to subsets of the image plane. These image patches are small enough that correspondence with a topologically-connected surface patch can be maintained [1, 3, 5, 4, 6, 12, 22, 31, 39]. Often, the optimization is restricted to a collection of small planar facets rotating and moving along the viewing rays of the reference camera, a process referred to as *plane sweeping*. In the variational setting that we adopt here, the object of inference, including the depth map, is a *function*. This distinguishes our approach from patch-based methods which we will – except for a brief review in Sect. 1.2 – consider no further. The majority of variational methods resort to *implicit* handling of depth discontinuities by TV regularization [15, 26, 25, 34, 40].

1.1. Contribution and overview

While TV effectively handles depth discontinuities in images, it does not impose geometrically meaningful constraints on the depth map: In Sect. 2.2, we show that TV is a proxy of the minimal-area functional *provided the depth map is orthographic*, a rather unrealistic assumption. Straightforward coupling of TV with a perspective re-projection error ceases to be physically plausible and yields undesirable staircasing artifacts (Fig. 1). Therefore, designing image-based regularizers that impose a geometrically meaningful prior on the surface is the first goal of this paper. In Sect. 2.3, we derive the correct area form for the perspective case and embed it in a novel regularization term for variational stereo. While this makes the regularizer plausible, it makes the resulting optimization challenging. Thus, our second goal is to devise an efficient optimization method tailored to this regularizer. On this topic, our core contribution is to leverage on *feedback linearization* [20], a technique from differential geometric control theory, to re-parametrize the regularizer into a form that is linear in the optimization variable and thus amenable to highly efficient primal-dual solvers. This is made possible by the gauge freedom in the parametrization: we exploit the fact that there are infinitely many equivalent parameterizations to our advantage. The implementation details of our method are provided in Sects. 3.2 and 3.3. In order to facilitate reproducible research, we will make our source code publicly available at `www.gpu4vision.org`. A series of experiments on synthetic and real data confirm our theoretical findings and demonstrate a gain in reconstruction quality (Sect. 4).

1.2. Other related work

In two companion papers [23, 24], Li and Zucker recognize the need for richer geometric representations in stereo vision. Their work has initiated a series of enhancements of patch-based methods [3, 4, 35, 41, 43] that all include some crude approximation of surface curvature in the proposed energy functional. Recently, Heise et al. [17] proposed to augment the PatchMatch algorithm with a term reminiscent of a Huber norm applied to normal changes across different patches. All of the aforementioned approaches depart from a discrete, label-based formulation of the problem, whose solution is accomplished by combinatorial optimization. Combinatorial optimization is however contrary to the calculus of variations, which we have chosen as our paragon here. Weighing the advantages and disadvantages of both paradigms against each other is beyond the scope of this paper, but we believe that the latter offers more flexibility in accurately modelling the inner geometry of regular surfaces. In the variational setting, generalized total variation (TGV) has helped to diminish staircasing by enriching the piecewise constant basis that spans the space of functions of minimal TV with polynomials of higher order [16, 30]. Still, TGV is a generic regularizer not specifically designed for surface parametrization, unlike the regularizer introduced here. Re-parametrizations of range maps to the benefit of optimization have appeared previously, e.g., in the realm of shape from shading [29] or self-localization and mapping [9].

2. Variational model

2.1. Data term

The following derivations consider the binocular stereo problem. An extension to more than two views is conceptually straightforward. We model the relative position and orientation of the sensors by an element $g \in \mathsf{SE}(3)$ in the special Euclidean group. By $\pi : \mathbb{R}^3 \to \mathbb{R}^2$, we denote the canonical pinhole projection with parameters $f_1, f_2 \in \mathbb{R}$, the focal lengths, and $c_1, c_2 \in \mathbb{R}$, the location of the principal point. Two radiance images $\mathcal{I}_r, \mathcal{I} : \Omega \to \mathbb{R}_+$, the former being the *reference* image and $\Omega \subset \mathbb{R}^2$ denoting the image domain, give rise to the *re-projection error*

$$r = \mathcal{I} \circ w(\boldsymbol{x}, z(\boldsymbol{x})) - \mathcal{I}_r(x). \tag{1}$$

The domain warping $w = \pi \circ g \circ \pi^{-1}$ is obtained by projecting pixels $\boldsymbol{x} \in \Omega$ back onto the surface $S \subset \mathbb{R}^3$ and then into the projection center of the camera displaced by g. Finally, the data term accumulates some robust Huber function $|\cdot|_\epsilon$ of the re-projection error over S:

$$E(S) = \int_\Omega |r|_\epsilon \,\mathrm{d}\boldsymbol{x}. \tag{2}$$

Note that this integral is computed over Ω which rules out trivial solutions such as $S = \emptyset$ that may occur in shape-space-based methods discussed in Sect. 1.2. Also note that the back-projection π^{-1} needed to compute the warping w depends on the – initially unknown – surface S, and thus also on its parameterization which we will turn to now.

2.2. Orthographic minimal-area regularizer

In the simplest case where π is orthographic, we can model S by the graph of a scalar function $z \in C^2(\Omega)$. Each point $\boldsymbol{X} \in S$ can then be written as

$$\boldsymbol{X} = p(\boldsymbol{x}) = \begin{pmatrix} x \\ y \\ z(x,y) \end{pmatrix}. \quad (3)$$

Note that conceptionally, except for the degree of smoothness, the surface has not much in common with the depth map, yet z appears in the parametrization and thus will influence the inner geometry of S. For example, the *metric tensor* at a point $\boldsymbol{X} = p(\boldsymbol{x})$ reads

$$\mathbf{I}_p = \begin{pmatrix} \langle \boldsymbol{X}_x, \boldsymbol{X}_x \rangle & \langle \boldsymbol{X}_x, \boldsymbol{X}_y \rangle \\ \langle \boldsymbol{X}_x, \boldsymbol{X}_y \rangle & \langle \boldsymbol{X}_y, \boldsymbol{X}_y \rangle \end{pmatrix}, \quad (4)$$

where the tangent vectors $\mathbf{X}_x = p_x(\boldsymbol{x})|_{\boldsymbol{x}}$ and $\mathbf{X}_y = p_y(\boldsymbol{x})|_{\boldsymbol{x}}$ are obtained by partial differentiation w.r.t. x and y. Typically, $\mathbf{I}_p$ is used to measure infinitesimal lengths and angles on the surface. In particular, $\sqrt{\det \mathbf{I}_p}$ determines the distortion of the two infinitesimal area elements $\mathrm{d}x$ and $\mathrm{d}S$. A scalar function $f : S \to \mathbb{R}$ defined on the surface can be pulled back by p to the parametric domain Ω. The pullback also relates the domains of integration S and Ω with each other:

$$\int_S f(\boldsymbol{X}) \mathrm{d}S = \int_\Omega f \circ p(\boldsymbol{x}) \sqrt{\det \mathbf{I}_p}\, \mathrm{d}\boldsymbol{x}. \quad (5)$$

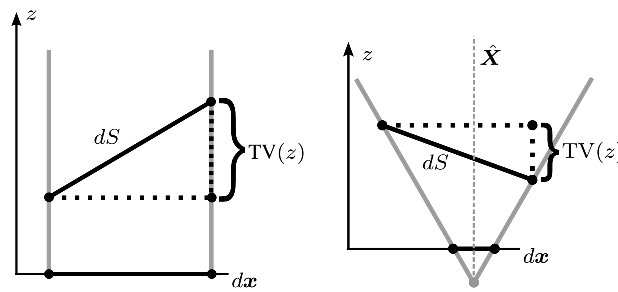
Setting $f = 1$ and substituting (3) into (4) and then (5) yields the total area

$$A(z) = \int_\Omega \sqrt{\det \mathbf{I}_p}\, \mathrm{d}\boldsymbol{x} = \int_\Omega \sqrt{z_x^2 + z_y^2 + 1}\, \mathrm{d}\boldsymbol{x} \quad (6)$$

of the graph of z.

Eq. (6) looks much like the TV of z if it were not for the additional value 1 under the square root. As can be seen in Fig. 2(a), it is this difference that allows measuring the area of a surface element. Still, both the area form (6) and $\mathrm{TV}(z)$ admit precisely the same set of global minimizers, namely the set of piecewise fronto-parallel surfaces. This underlines the well-known fact that the TV-regularizer favors piece-wise constant functions.

This property may be desirable in image processing applications, where z corresponds to some image intensity distribution over Ω, but certainly not when z is a depth map that parameterizes a geometric surface. Meanwhile, the assumption of an orthographic camera model in reconstruction is quite unrealistic for practical purposes, and so is the



(a) Orthographic (b) Perspective

Figure 2. TV and surface area under orthographic projection (a) differ in that the TV measures only the jumps along z, whereas a surface element $\mathrm{d}S$ also takes into account the component of $\mathrm{d}\boldsymbol{x}$ parallel to the image plane. To reduce the area of a non-minimal surface element, the only option in both cases is to rotate until fronto-parallelity is achieved, i.e. $\mathrm{TV}(z) = 0$. The area form induced by perspective projection (b) on the other hand has an additional degree of freedom: the area of a non-minimal surface element can either be decreased by rotating it until perpendicular to the pixel viewing ray, or by moving it closer to the center of projection.

use of $\mathrm{TV}(z)$ as a regularizer, although that appears to be common practice in previous works [15, 26, 25, 34, 40]. So in the following section, let us clarify how the area form of a perspective depth map parametrization looks like, and highlight its interplay with the TV.

2.3. Perspective minimal-area regularizer

In the perspective parameterization, for which – to reduce the notational burden – we maintain the symbol p, the depth z influences all three spatial coordinates of a surface point. More precisely, we have $\boldsymbol{X} = p(\boldsymbol{x}) = z\hat{\boldsymbol{X}}$ where

$$\hat{\boldsymbol{X}} = \begin{pmatrix} \hat{\boldsymbol{x}} \\ 1 \end{pmatrix} = \begin{pmatrix} \hat{x} \\ \hat{y} \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{x-c_1}{f_1} \\ \frac{y-c_2}{f_2} \\ 1 \end{pmatrix}$$

is the direction of the viewing ray associated with a pixel. It is obtained by multiplying $\boldsymbol{x}$ with the inverse intrinsic parameter matrix. From the tangent vectors

$$\boldsymbol{X}_x = \begin{pmatrix} \hat{x} z_x + \frac{z}{f_1} \\ \hat{y} z_x \\ z_x \end{pmatrix}, \quad \boldsymbol{X}_y = \begin{pmatrix} \hat{x} z_y \\ \hat{y} z_y + \frac{z}{f_2} \\ z_y \end{pmatrix}, \quad (7)$$

both, the fundamental form and the square root of its determinant, immediately follow:

$$\sqrt{\det \mathbf{I}_p} = \frac{z}{f_1 f_2} \sqrt{\|\nabla_{\boldsymbol{f}} z\|^2 + (\langle \nabla_{\boldsymbol{f}} z, \hat{\boldsymbol{x}} \rangle + z)^2}. \quad (8)$$

Here, we have introduced an abbreviation $\nabla_{\boldsymbol{f}}$ for the nabla operator whose components are weighted by the focal lengths $\boldsymbol{f} = (f_1, f_2)$, i.e., $\nabla_{\boldsymbol{f}} z = (f_1 z_x, f_2 z_y)^\top$.

Eq. (8) presents the central ingredient in our regularization term, so a few remarks are in order: First of all, the factor z in front of the square root makes the surface area form $\sqrt{\det \mathbf{I}_p}\mathrm{d}S$ distance-dependent. Consequently, a mean curvature flow is furnished with an additional degree of freedom. As shown in Fig. 2(b), one can reduce surface area (locally) by moving points towards the center of projection. Vice-versa, minimal surfaces, unless they constitute the global optimum $z = 1$ at which $\mathrm{d}S = \mathrm{d}\boldsymbol{x}$, are not necessarily piecewise constant in depth. As we will later verify empirically, this helps reduce staircasing artifacts generated by methods with "naive" TV regularization. On the downside, Eq. (8) is neither equal to the TV nor to the norm of some linear operator applied to z. Hence, combining (2) and (6) with (8) as area form does not yield a functional of ROF-type, for which a plethora of solvers is available. We will address this issue in the next section by transforming our variational problem such that (8) becomes tractable for a powerful off-the-shelf primal-dual algorithm.

3. Optimization

3.1. Algorithm

Let us first summarize the continuous variational problem we wish to solve:

$$\min_z \int_\Omega \frac{z}{f_1 f_2}\sqrt{|\nabla_{\boldsymbol{f}} z|^2 + (\langle \nabla_{\boldsymbol{f}} z, \hat{\boldsymbol{x}}\rangle + z)^2}\,\mathrm{d}\boldsymbol{x} + \lambda \int_\Omega |\mathcal{I} \circ w(\boldsymbol{x}, z(\boldsymbol{x})) - \mathcal{I}_r(\boldsymbol{x})|_\epsilon \,\mathrm{d}\boldsymbol{x}. \quad (9)$$

As customary, $\lambda \in \mathbb{R}_+$ is a scalar parameter controlling the trade-off between data fidelity and smoothness. We call the model (9) *flow-based stereo* for its connection to optical flow methods. In the latter, the data term is usually some form of photoconsistency measure between the reference image and the warped second image, where the warp is parameterized by a flow vector field. In our model, the warp is parameterized by the depths of pixels. A similar approach was used recently by the authors of [34] to compute dense depthmaps in real-time, albeit employing the raw TV for regularization. Flow-based stereo carries out a continuous search for correspondences along the epipolar lines. It can thus be seen as the variational counterpart of the planesweep algorithm, however, with the advantages that it requires no resource-hungry spatial data structure. Also, the extension to multiple views is easy to achieve by summing up the reprojection error over a number of image pairs.

Looking at the first term of (9), we see that in general it is non-convex because of the bilinear form involving z and its derivative. If we use the fact that $\sqrt{\det \mathbf{I}_p}$ equals the length of the surface normal $\|\boldsymbol{n}\| = \|\boldsymbol{X}_x \times \boldsymbol{X}_y\|$, the situation improves slightly. From (7), it becomes clear, though, that

$$\boldsymbol{n} = \begin{pmatrix} -\frac{zz_x}{f_2} \\ -\frac{zz_y}{f_1} \\ \frac{1}{f_1 f_2}(z^2 + \hat{x} f_1 z z_x + \hat{y} f_2 z z_y) \end{pmatrix} \quad (10)$$

in general is still non-convex in z. Remarkably, this can be fixed by re-parametrizing S as stated in the central

Proposition 3.1. *Substituting z in the perspective depth map parametrization by $z = \phi(\zeta)$ with $\phi(\zeta) = \sqrt{2\zeta}$, the Gauss map becomes a linear function of ζ. In particular, it holds*

$$\boldsymbol{n}(\zeta) = \begin{pmatrix} -\frac{\zeta_x}{f_2} \\ -\frac{\zeta_y}{f_1} \\ \frac{\hat{x}\zeta_x}{f_2} + \frac{\hat{y}\zeta_y}{f_1} + \frac{2\zeta}{f_1 f_2} \end{pmatrix}. \quad (11)$$

Proof. We start by applying the chain rule to the non-convex term zz_x (similarly to zz_y), which yields for a re-parameterization $z = \phi(\zeta)$

$$zz_x = \phi \frac{\mathrm{d}\phi}{\mathrm{d}\zeta}\frac{\partial \zeta}{\partial x}. \quad (12)$$

If we now require that

$$\phi \frac{\mathrm{d}\phi}{\mathrm{d}\zeta} = 1, \quad (13)$$

we are left with the (linear, thus convex) term $\frac{\partial \zeta}{\partial x}$. Eq. (13) constitutes a first-order ordinary differential equation, which can be solved for ϕ by *separation of the variables*:

$$\begin{aligned} \phi \mathrm{d}\phi &= \mathrm{d}\zeta, \\ \int \phi \mathrm{d}\phi &= \int \mathrm{d}\zeta, \\ \frac{\phi^2}{2} &= \zeta. \end{aligned} \quad (14)$$

From (14), we get $\phi = \sqrt{2\zeta}$. Inserting this into (10) and using $\frac{\mathrm{d}\phi}{\mathrm{d}\zeta} = \frac{1}{\phi}$, it follows immediately that

$$\begin{aligned} zz_x &= \phi \frac{\mathrm{d}\phi}{\mathrm{d}\zeta}\frac{\partial d\zeta}{\partial x} = \sqrt{2\zeta}\frac{\mathrm{d}\sqrt{2\zeta}}{\mathrm{d}\zeta}\zeta_x = \zeta_x, \\ zz_y &= \phi \frac{\mathrm{d}\phi}{\mathrm{d}\zeta}\frac{\partial \zeta}{\partial y} = \zeta_y, \\ z^2 &= \phi^2 = 2\zeta, \end{aligned} \quad (15)$$

and hence the claim. □

Let us remark that the transformation ϕ is bijective and differentiable over $(0, \infty)$, which is sufficient since we may assume all surface points to be located in front of the camera.

We are now left with the non-convexity of the data term. Since the optimization variable z appears as an argument to the warping w in (9), it is clear the only way to get around the non-convexity is to linearize the data term. This calls for an iterative optimization strategy, in which at each step, say $k \in \mathbb{N}$, a local convex approximation of the data term is minimized. The convergence rate will be highly influenced by how faithful the approximation is to the original functional. If the approximation is linear, the resulting iteration will correspond to a gradient descent. Higher rates can be achieved by leveraging on the special structure of the problem, where the energy consists of a (Huber) norm $|r|_\epsilon$ applied to some residual function r, here the re-projection error (1). The idea is to first compute a Taylor expansion of the *residual*

$$r(z) \approx r(z_k) + \frac{\mathrm{d}r}{\mathrm{d}z}\Big|_{z_k}(z - z_k) \tag{16}$$

around the current iterate z_k, and only then apply the norm to it, yielding a local approximation of (2):

$$\hat{E}(z) := \int\limits_\Omega \left| r(z_k) + \frac{\mathrm{d}r}{\mathrm{d}z}\Big|_{z_k}(z - z_k) \right|_\epsilon \mathrm{d}\boldsymbol{x}. \tag{17}$$

The overall strategy is reminiscent of the classical Gauss-Newton method for nonlinear least-squares problems [2]. The last piece we need in order to finalize treatment of the data term is the derivative of the residual w.r.t. the function z, which is given by

$$\frac{\mathrm{d}r}{\mathrm{d}z} = \nabla \mathcal{I}|_{\pi \circ g(z\hat{\boldsymbol{X}})} D\pi|_{g(z\hat{\boldsymbol{X}})} Dg\hat{\boldsymbol{X}}. \tag{18}$$

To conclude this section, let us state the full variational problem in the variable ζ following re-parametrization:

$$\min_\zeta \int_\Omega |\boldsymbol{n}(\zeta)|\,\mathrm{d}\boldsymbol{x} + \lambda \int_\Omega |\mathcal{I} \circ w(\boldsymbol{x}, \zeta(\boldsymbol{x})) - \mathcal{I}_r(\boldsymbol{x})|_\epsilon \,\mathrm{d}\boldsymbol{x}. \tag{19}$$

Note that since the original warp w depends on z, it also has to be reformulated in terms of ζ.

3.2. Discretization

We discretize the image domain Ω on the regular Cartesian grid of size $M \times N$ with indices (i, j). An approximation of (19) is then given by

$$\min_\zeta \|L\zeta\|_{2,1} + \lambda \tilde{E}(\zeta), \tag{20}$$

where $\tilde{E}(\zeta) := \|\mathcal{I}^w - \mathcal{I}_r\|_\epsilon$ is the discrete data term and $\mathcal{I}^w$ denotes the warped image. We apply the $2,1$-matrix norm to the result of the multiplication $L\zeta$ (a vector of dimension $3MN$), which, by slight abuse of notation, is implicitly rearranged to a $MN \times 3$ matrix. Given a matrix $A \in \mathbb{R}^{M \times N}$, the $2,1$-norm takes the ℓ_2-norm across rows and the ℓ_1-norm across columns and is defined as follows: $\|A\|_{2,1} = \sum_{m=1}^M \left| \sqrt{\sum_{n=1}^N A_{m,n}^2} \right|$. The operator $L \in \mathbb{R}^{3MN \times MN}$ is linear and for every element of ζ, it computes its normal vector according to Eq. (11):

$$L = \begin{pmatrix} -\frac{1}{f_2} D_x \\ -\frac{1}{f_1} D_y \\ \frac{\hat{x}}{f_2} D_x + \frac{\hat{y}}{f_1} D_y + \frac{2}{f_1 f_2} \end{pmatrix}. \tag{21}$$

It contains first-order finite-difference approximations $D_x, D_y \in \mathbb{R}^{MN \times MN}$ of the partial derivatives of a function which are defined as follows:

$$D_x \zeta^{i,j} = \begin{cases} \zeta^{i,j+1} - \zeta^{i,j} & \text{if } j < N, \\ 0 & \text{else,} \end{cases} \tag{22}$$

$$D_y \zeta^{i,j} = \begin{cases} \zeta^{i+1,j} - \zeta^{i,j} & \text{if } i < M, \\ 0 & \text{else.} \end{cases} \tag{23}$$

3.3. Implementation

Recalling that our strategy to solve the originally non-convex problem (9) is to linearize the data term around some estimate ζ_k and subsequently solve a series of locally-convex approximations, each of the sub-problems looks like

$$\min_\zeta \|L\zeta\|_{2,1} + \lambda \|r + a(\zeta - \zeta_k)\|_\epsilon, \tag{24}$$

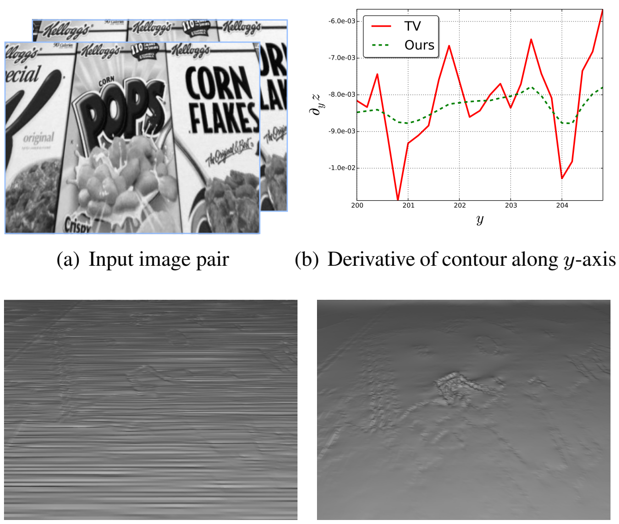
where r is the residual at ζ_k and $a := \frac{\mathrm{d}r}{\mathrm{d}\zeta}\big|_{\zeta_k}$ is the derivative of r w.r.t. ζ. We point out that while (24) and (9) are equivalent, they differ in the important aspect that the regularizer in (9) is non-convex in the optimization variable, whereas the regularizer of (24) is linear. For this reason, we can apply the highly efficient first order primal-dual algorithm due to Chambolle and Pock [8] to solve (24). Dualizing the first term, the primal-dual formulation reads

$$\min_\zeta \max_{\|q\|_\infty \le 1} \langle L\zeta, q \rangle + \lambda \|r + a(\zeta - \zeta_k)\|_\epsilon, \tag{25}$$

where $q \in \mathbb{R}^{3MN}$ is the dual variable. The inner iteration – denoted by l to distinguish it from the outer iterations k – delivers a minimizer of (25) with superlinear rate of convergence:

$$\begin{aligned} q^{l+1} &= \mathrm{proj}_{\|q\|_\infty \le 1}(q^l + \Sigma L \bar{\zeta}^l), \\ \zeta^{l+1} &= \mathrm{prox}(\zeta^l - T L^* q^{l+1}), \\ \bar{\zeta}^{l+1} &= 2\zeta^{l+1} - \zeta^l. \end{aligned} \tag{26}$$

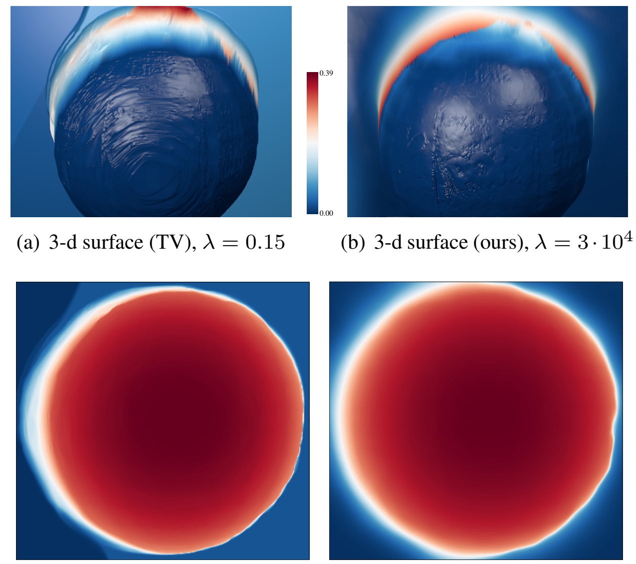
Here, $T = \mathrm{diag}(\tau_n)$, $n = 1, \ldots, MN$ and $\Sigma = \mathrm{diag}(\sigma_m)$, $m = 1, \ldots, 3MN$ are diagonal preconditioning matrices which, according to [27], account for the bad



(a) Input image pair (b) Derivative of contour along y-axis

(c) 3-d surface (TV) (d) 3-d surface (ours)

Figure 3. Results for the synthetic experiment consisting of a tilted plane to show the behavior of the regularizer in case of a slanted surface. The TV regularizer (c) produces the well-known staircasing artifacts while the surface area regularizer (d) shows no bias towards fronto-parallelity. (b) The zig-zagging of TV is clearly visible in the derivative of a cut through the depth map along the y-axis.



(a) 3-d surface (TV), $\lambda = 0.15$ (b) 3-d surface (ours), $\lambda = 3 \cdot 10^4$

(c) Depth map TV (d) Depth map Ours

Figure 4. Results for a synthetic curved surface. Whereas the depth maps (c)-(d) of the sphere surface look similar, a 3-d visualization (a)-(d) allows a more thorough examination of reconstruction quality: the plateau structure is visible in case of TV regularization (especially at the top), area regularization on the other hand is more faithful to the true curved surface.

scaling of the operator L, and $\text{proj}_{\|q\|_\infty \le 1}(\cdot)$ is a simple pointwise projection onto the unit ball. The proximal operator $\text{prox}(\cdot)$ is defined as the solution of the following minimization problem:

$$\text{prox}(\tilde{\zeta}) = \min_{\zeta} \frac{\|\zeta - \tilde{\zeta}\|^2}{2\tau} + \lambda \|r + a(\zeta - \zeta_k)\|_\epsilon. \quad (27)$$

Defining $b := r - a\zeta_k$, a solution can be obtained by the following explicit formula applied to each of the elements of ζ separately:

$$\zeta = \begin{cases} \tilde{\zeta} - \tau\lambda a & \text{if } a\tilde{\zeta} + b > \tau\lambda a^2 + \epsilon \\ \tilde{\zeta} + \tau\lambda a & \text{if } a\tilde{\zeta} + b < -\tau\lambda a^2 - \epsilon \\ \frac{\tilde{\zeta} - \tau\lambda ab/\epsilon}{1 + \tau\lambda a^2/\epsilon} & \text{else} \end{cases} \quad (28)$$

As it is common in many optical flow algorithms, we embed the whole procedure into a coarse-to-fine warping framework to account for large discontinuities in depth. We created a highly parallel implementation using the CUDA toolkit, which makes the method attractive for (near) real-time applications.

4. Experimental studies

All results were computed on a desktop PC equipped with a 3.2GHz i7 QuadCore CPU and a Geforce 780Ti GPU. We used a pyramid scale factor of 0.75 throughout, and computed 30 warps per pyramid level and 60 iterations per warp.

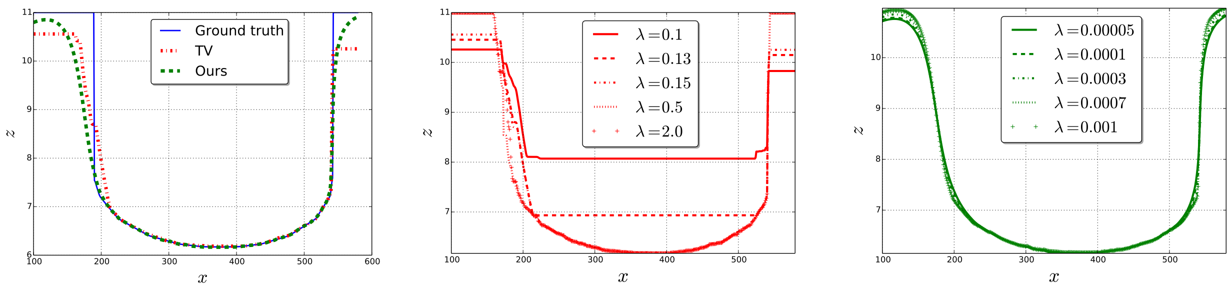
For a real-world configuration with a scale factor of 0.5, 20 warps and 30 iterations (note that this is rarely needed for convergence and can be trimmed further), the runtime is 0.14 s for a resolution of 640×480 and 1.9 s for 3072×2048 respectively.

4.1. Synthetic data

To empirically verify the theoretical properties of our regularizer (see Sec. 2.3), we conducted a number of experiments with synthetic (i.e., perfect) input data and known ground truth. Despite the fact that our model is conceptually capable of using multiple views, we restricted ourselves to classical binocular stereo for all experiments.

The first example consists of a plane rotated 30° around the x-axis. As depicted in Fig. 3, TV regularization clearly shows staircaising artifacts, whereas our surface area regularizer produces a smooth slanted plane. We emphasize the fact that the regularization strength for TV has been hand-tuned to be as smooth as possible before breaking down (e.g., approximating the slanted surface by a series of very large fronto-parallel steps).

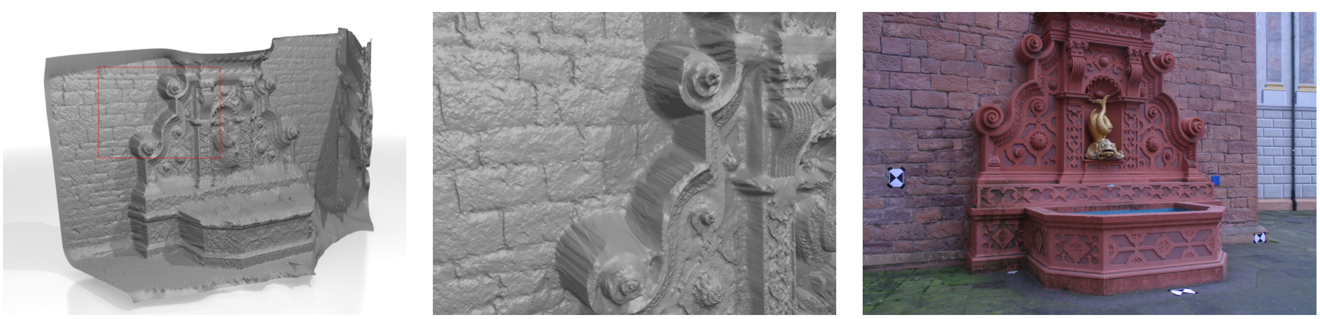
The next experiment (Fig. 4) involves a hemisphere set against a fronto-parallel background. We use the hemisphere to assess the ability of the regularizer to reconstruct curved surfaces. Figs. 4(a) and (b) show a 3-d visualization



(a) Reconstruction vs. groundtruth (b) TV at different levels of regularization (c) Ours at different levels of regularization

Figure 5. (a) Horizontal cross-section through the depth maps for the sphere experiment. Note that the depth discontinuity at the left side is occluded in the input images; the right side is co-visible. We find that in occluded regions, where the task of the regularizer is hallucination, TV produces the familiar steps. Area regularization on the other hand smoothly bridges the occluded area.

(a) TV (b) TV (closeup) (c) Input image 1



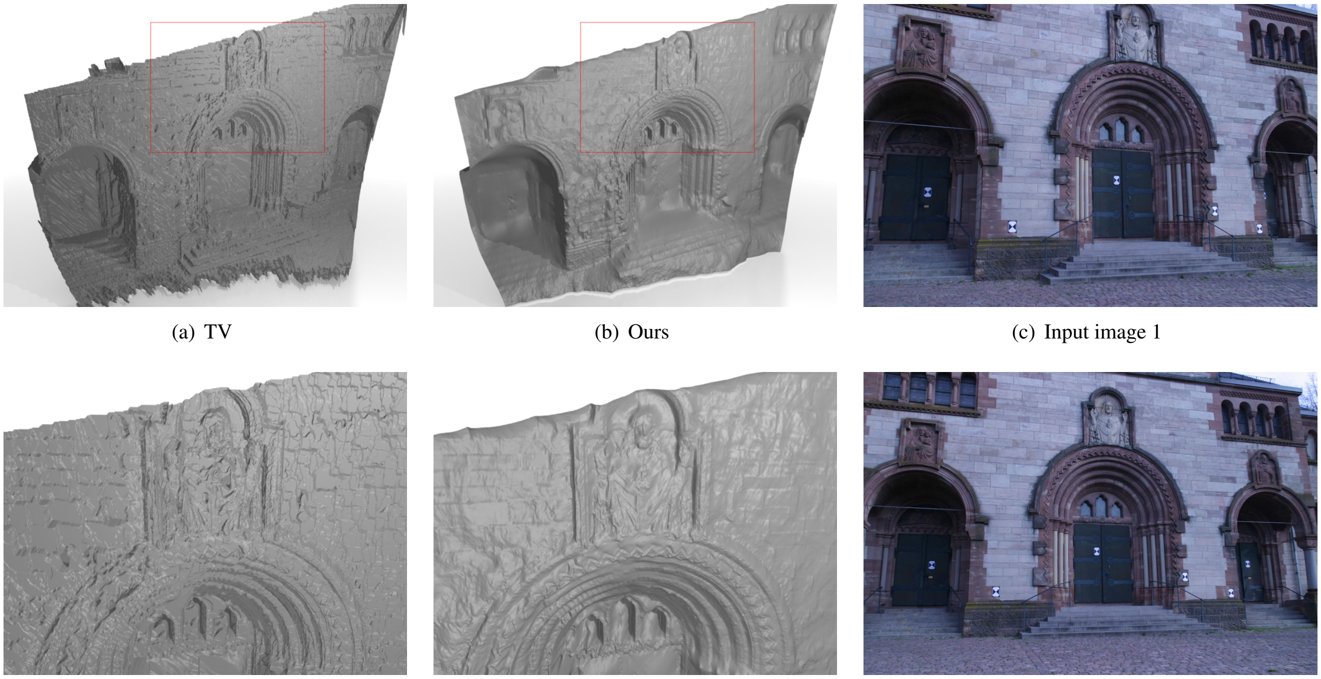
(d) Ours (e) Ours (closeup) (f) Input image 2

Figure 6. Results for the *Fountain-P11* scene using the views 6 and 7, with 6 as reference view.

of the result for TV and surface area regularization respectively, where color encodes error relative to ground truth. Despite of the depth maps 4(c) and (d) of the sphere looking similarly smooth, one can see qualitative differences in the 3-d surface: Where TV tends to approximate the half-sphere by fronto-parallel plateaus, area regularization produces a more pleasing result. This underlines the importance of using 3-d visualizations when assessing the quality of stereo algorithms. Fig. 5 illustrates the behaviour of the regularizer under varying values of the regularization parameter λ. TV (Fig. 5(b)) suffers from sudden breakdown whereas area regularization (Fig. 5(c)) remains stable over orders of magnitude. Stable parameters are of great interest to the practitioner because most of these are still hand-tuned. Note that for the visualizations of the sphere (Figs. 4(a) and (b)) we again chose favorable regularization strength for TV (i.e., close to breakdown) and medium regularization for surface area.

4.2. Real data

We tested our regularizer on real-world examples taken from the Strecha dataset [33]. It consists of a number of high-resolution (3072×2048) images for dense multiview stereo algorithms. All results were obtained using only two images. Fig. 6 depicts 3-d renderings of results from the *Fountain-P11* scene, renderings for *Herz-Jesu-P25* are



(a) TV (b) Ours (c) Input image 1

(d) TV closeup (e) Ours closeup (f) Input image 2

Figure 7. Results for the *Herz-Jesu-P25* scene using views 5 and 6, with 5 as reference view.

	RMS error		Reduction
	TV	Ours	$\frac{\mathrm{TV}-\mathrm{Ours}}{\mathrm{TV}}$
TiltedPlane10	5.5527e-4	1.0819e-4	80.5%
TiltedSine10	0.0167188	0.0114407	31.6%
Fountain-P11	0.05342	0.02644	50.5%
Herz-Jesu-P25	0.264935	0.222491	16.0%

Table 1. RMS error against ground-truth depth-maps for different datasets. The last column is the error reduction, i.e., the percentaged gain in reconstruction quality achieved by our method over TV regularization.

found in Fig. 7. We also provide the input images for reference. Despite our best efforts and a very strong data term (as can be seen in Fig. 7(a) by the artifacts at the bottom and the top), we did not succeed in getting a smooth reconstruction of the church facade by means of TV regularization. The reason for this is that the facade is slightly slanted w.r.t. the reference camera image plane. The gaps between the individual weakly-textured bricks provide a gradient for the TV regularizer to hold on to. It therefore tends to approximate every individual brick by its own fronto-parallel facet, as can be seen in the closeup Fig. 7(d). The surface area regularizer 7(e) is able to recover small depth discontinuities (see for instance the little arches above the front door), while maintaining a smooth facade.

Tab. 1 shows a quantitative comparison of the root mean square (RMS) error against ground-truth depth-maps over different datasets.

5. Conclusion

We have introduced a new regularizer for variational stereo, which is defined on the image but regularizes a geometrically meaningful quantity on the surface. Exploiting gauge freedom, a re-parameterization makes the regularizer compatible with highly efficient primal-dual solvers for large scale problems. We evaluated important properties of the regularizer such as the ability to reconstruct smooth surfaces using both synthetic and real world data. In particular, a comparison to the widely used TV regularizer showed that the minimal surface regularizer does not suffer from the staircaising effect. Because the computational cost remains basically the same when going from TV to surface area regularization, our method is suited for real-time applications.

6. Acknowledgments

This work was supported by the research initiative "Mobile Vision" with funding from the AIT and the Austrian Federal Ministry of Science, Research and Economy HRSM programme (BGBl. II Nr. 292/2012), by the START project BIVISION No. Y729, and by NGA HM02101310004 and AFRL FA8650-11-1-7154.

References

[1] C. Bailer, M. Finckh, and H. P. A. Lensch. Scale Robust Multi View Stereo. *European Conference on Computer Vision*, 1:398–411, 2012.

[2] J. Balzer and S. Soatto. Second-order Shape Optimization for Geometric Inverse Problems in Vision. *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[3] F. Besse, C. Rother, A. Fitzgibbon, and J. Kautz. PMBP: PatchMatch Belief Propagation for Correspondence Field Estimation. In *British Machine Vision Conference*, 2012.

[4] M. Bleyer, C. Rhemann, and C. Rother. PatchMatch Stereo - Stereo Matching with Slanted Support Windows. In *British Machine Vision Conference*, pages 1–11, 2011.

[5] A. Bodis-Szomoru, H. Riemenschneider, and L. V. Gool. Fast, Approximate Piecewise-Planar Modeling Based on Sparse Structure-from-Motion and Superpixels. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 469–476, 2014.

[6] D. Bradley, T. Boubekeur, and W. Heidrich. Accurate multi-view reconstruction using robust binocular stereo and surface meshing. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[7] N. D. F. Campbell, G. Vogiatzis, C. Hern, and R. Cipolla. Using Multiple Hypotheses to Improve Depth-Maps for Multi-View Stereo. *European Conference on Computer Vision*, pages 766–779, 2008.

[8] A. Chambolle and T. Pock. A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[9] J. Civera, A. J. Davison, and J. Montiel. Inverse depth parametrization for monocular SLAM. *IEEE Transactions on Robotics*, 24(5):932–945, 2008.

[10] A. Delaunoy and M. Pollefeys. Photometric Bundle Adjustment for Dense Multi-View 3D Modeling. *IEEE Conference on Computer Vision and Pattern Recognition*, 1(1):1486–1492, 2014.

[11] S. Fan and F. P. Ferrie. Photo Hull regularized stereo. *Image and Vision Computing*, 28(4):724–730, 2010.

[12] Y. Furukawa and J. Ponce. Accurate, dense, and robust multi-view stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1362–1376, 2010.

[13] D. Gallup. 3D Reconstruction from Accidental Motion. *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[14] M. Goesele, B. Curless, and S. Seitz. Multi-View Stereo Revisited. *IEEE Conference on Computer Vision and Pattern Recognition*, 2:2402–2409, 2006.

[15] G. Graber, T. Pock, and H. Bischof. Online 3D reconstruction using convex optimization. *IEEE International Conference on Computer Vision*, pages 708–711, Nov. 2011.

[16] S. Heber and T. Pock. Shape from Light Field meets Robust PCA. In *European Conference on Computer Vision*, 2014.

[17] P. Heise, S. Klose, B. Jensen, and A. Knoll. PM-Huber: PatchMatch with Huber Regularization for Stereo Matching. *IEEE International Conference on Computer Vision*, pages 2360–2367, Dec. 2013.

[18] C. Hernández and F. Schmitt. Silhouette and stereo fusion for 3D object modeling. *Computer Vision and Image Understanding*, 96(3):367–392, 2004.

[19] X. Hu and P. Mordohai. Least Commitment, Viewpoint-Based, Multi-view Stereo. In *International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission*, pages 531–538, 2012.

[20] A. Isidori. *Nonlinear control systems*, volume 1. Springer, 1995.

[21] H. Jin, R. Tsai, L. Chen, A. Yezzi, and S. Soatto. Estimation of 3d surface shape and smooth radiance from 2d images: A level set approach. *Journal of Scientific Computing*, 19(1-3):267–292, 2003.

[22] A. Klaus, M. Sormann, and K. Karner. Segment-Based Stereo Matching Using Belief Propagation and a Self-Adapting Dissimilarity Measure. In *International Conference on Pattern Recognition*, pages 15–18, 2006.

[23] G. Li and S. W. Zucker. Differential Geometric Consistency Extends Stereo to Curved Surfaces. In *European Conference on Computer Vision*, volume 3953, pages 44–57, 2006.

[24] G. Li and S. W. Zucker. Surface Geometric Constraints for Stereo in Belief Propagation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2355–2362, 2006.

[25] Y. Liu, X. Cao, Q. Dai, and W. Xu. Continuous depth estimation for multi-view stereo. pages 2121–2128, 2009.

[26] R. Newcombe, S. J. Lovegrove, and A. J. Davison. DTAM: Dense tracking and mapping in real-time. *IEEE International Conference on Computer Vision*, pages 2320–2327, Nov. 2011.

[27] T. Pock and A. Chambolle. Diagonal Preconditioning for First Order Primal-dual Algorithms in Convex Optimization. In *IEEE International Conference on Computer Vision*, pages 1762–1769, 2011.

[28] J.-P. Pons, R. Keriven, and O. Faugeras. Multi-View Stereo Reconstruction and Scene Flow Estimation with a Global Image-Based Matching Score. *International Journal of Computer Vision*, 72(2):179–193, 2006.

[29] E. Prados and O. Faugeras. A generic and provably convergent shape-from-shading method for orthographic and pinhole cameras. *International Journal of Computer Vision*, 65(1-2):97–125, 2005.

[30] R. Ranftl, S. Gehrig, T. Pock, and H. Bischof. Pushing the limits of stereo using variational stereo estimation. In *Intelligent Vehicles Symposium*, pages 401–407, 2012.

[31] C. Raposo, M. Antunes, and J. P. Barreto. Piecewise-Planar StereoScan: Structure and Motion from Plane Primitives. *European Conference on Computer Vision*, pages 48–63, 2014.

[32] M. Sormann, C. Zach, J. Bauer, K. Karner, and H. Bishof. Watertight Multi-View Reconstruction Based On Volumetric Graph-Cuts. In *Scandinavian Conference on Image Analysis*, pages 1–10, 2007.

[33] C. Strecha, W. von Hansen, L. Van Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[34] J. Stühmer, S. Gumhold, and D. Cremers. Real-Time Dense Geometry from a Handheld Camera. In *Pattern Recognition (Proc. DAGM)*, pages 11–20, Darmstadt, Germany, Sept. 2010.

[35] T. Taniai, Y. Matsushita, and T. Naemura. Graph Cut Based Continuous Stereo Matching Using Locally Shared Labels. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1613–1620, 2014.

[36] P. Tanskanen, K. Kolev, L. Meier, F. Camposeco, O. Sauer, and M. Pollefeys. Live Metric 3D Reconstruction on Mobile Phones. *IEEE International Conference on Computer Vision*, 1:65–72, 2013.

[37] E. Tola, C. Strecha, and P. Fua. Efficient large-scale multi-view stereo for ultra high-resolution image sets. *Machine Vision and Applications*, 23(5):903–920, 2011.

[38] R. Tyleček and R. Šará. Refinement of Surface Mesh for Accurate Multi-View Reconstruction. *International Journal of Virtual Reality*, 9(1):45–54, 2010.

[39] Z.-F. Wang and Z.-G. Zheng. A region based stereo matching algorithm using cooperative optimization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[40] A. Wendel, M. Maurer, G. Graber, T. Pock, and H. Bischof. Dense reconstruction on-the-fly. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1450–1457, 2012.

[41] O. Woodford, P. Torr, I. Reid, and A. Fitzgibbon. Global Stereo Reconstruction Under Second-Order Smoothness Priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(12):2115–2128, 2009.

[42] A. Zaharescu, E. Boyer, and R. Horaud. Topology-adaptive mesh deformation for surface evolution, morphing, and multiview reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(4):823–37, Apr. 2011.

[43] C. Zhang, Z. Li, R. Cai, H. Chao, and Y. Rui. As-Rigid-As-Possible Stereo under Second Order Smoothness Priors. *European Conference on Computer Vision*, pages 112–126, 2014.