Combining Shape from Shading and Stereo: A Variational Approach for the Joint Estimation of Depth, Illumination and Albedo

Daniel Maurer
maurer@vis.uni-stuttgart.de
Institute for Visualization and Interactive Systems
University of Stuttgart, Germany

Yong Chul Ju
ju@vis.uni-stuttgart.de
Institute for Visualization and Interactive Systems
University of Stuttgart, Germany

Michael Breuß
breuss@tu-cottbus.de
Institute for Applied Mathematics and Scientific Computing, BTU Cottbus-Senftenberg, Germany

Andrés Bruhn
bruhn@vis.uni-stuttgart.de
Institute for Visualization and Interactive Systems
University of Stuttgart, Germany

Abstract

Shape from shading (SfS) and stereo are two fundamentally different strategies for image-based 3-D reconstruction. While approaches for SfS infer the depth solely from pixel intensities, methods for stereo are based on a matching process that establishes correspondences across images. In this paper we propose a joint variational method that combines the advantages of both strategies. By integrating recent stereo and SfS models into a single minimisation framework, we obtain an approach that exploits shading information to improve upon the reconstruction quality of robust stereo methods. To this end, we fuse a Lambertian SfS approach with a robust stereo model and supplement the resulting energy functional with a detail-preserving anisotropic second-order smoothness term. Moreover, we extend the novel model in such a way that it jointly estimates depth, albedo and illumination. This in turn makes it applicable to objects with non-uniform albedo as well as to scenes with unknown illumination. Experiments for synthetic and real-world images show the advantages of our combined approach: While the stereo part overcomes the albedo-depth ambiguity inherent to all SfS methods, the SfS part improves the degree of details of the reconstruction compared to pure stereo methods.

1 Introduction

Reconstructing the 3-D surface of an object from one or more views of the same static scene is one of the key problems in computer vision. Two of the most frequently used techniques in this context are stereo and SfS. While stereo is based on finding corresponding pixels in multiple views and determining the depth via triangulation, SfS estimates the depth solely from a reflection model that relates the image brightness to the local surface normal – which requires information on both the illumination and the albedo of the scene.

In fact, the advantages and drawbacks of the two techniques are quite complementary. Stereo works well for textured objects, since this facilitates the matching process, while,

Pages 76.1-76.14
DOI: https://dx.doi.org/10.5244/C.30.76

at the same time, it requires spatial regularisation to compute estimates in homogeneous regions. In particular in the presence of fine surface details, however, this regularisation poses a problem as it tends to oversmooth the results. In contrast, SfS can produce detailed reconstructions, since it hardly needs regularisation, however, serious problems are caused by textured regions or equivalently by non-constant albedo, since the decomposition of the observed brightness into albedo/colour and depth is ambiguous. This in turn makes it difficult to apply SfS methods to real-world scenarios.

Given the advantages and drawbacks of both strategies, it seems natural to combine SfS and stereo to improve the quality of the reconstruction. Since the first ideas of Blake *et al*. [3] in 1985 a variety of methods for combining SfS and stereo has been proposed. The corresponding literature can essentially be divided into three types of approaches.

Fusion Approaches. Such approaches perform stereo and SfS independently and combine the results in terms of a post-processing step. Examples are, for instance, the method of Cryer *et al*. [9] that fuses depths maps from stereo and SfS in the frequency domain, or the approach of Haines *et al*. [16] that combines disparity information and surface normals within a probabilistic approach. Since the computations are performed separately in the case of fusion methods, the synergies between stereo and SfS are typically rather limited.

Subsequent Approaches. This type of approaches perform the stereo and SfS computation consecutively, where stereo provides an initialisation for SfS. Consequently, these techniques can be considered as shading-based refinement methods. First approaches such as the method of Leclerc *et al*. [22] and Hougen *et al*. [17] have been restricted to a simple orthographic camera model and a constant albedo, while assuming a global light direction and a polynomially parametrised reflectance map, respectively. Following [12], we denote these early techniques as view-centred, since they perform the refinement in the pixel domain. Recently, however, so-called object-centred approaches have become increasingly popular. Such methods operate directly on an initial closed 3-D mesh of an object. While they rely on a preceding stereo approach to obtain the initial mesh, stereoscopic cues are only implicitly exploited during the refinement by imposing shading cues on multiple views. As in the view-centred case, most of the object-centred approaches are based on the assumption that the scene consists of a single material [38, 39, 41, 42]. They either focus on generalising the reflectance model to non-Lambertian surfaces, e.g. by using the Phong model [41] or a view-independent reflectance map [42], or they aim towards estimating the illumination, e.g. by using spherical harmonics [38] or a general illumination vector field [39]. Among the few exceptions regarding the single material assumption are the approach from Yoon *et al*. [40] that estimates the reflectance of a dichromatic surface for a given illumination, and the approach of Valgaerts *et al*. [34] that exploits temporal constraints to cluster a spatially varying albedo. As expected, in the case of subsequent approaches, the shading based refinement can benefit significantly from the preceding stereo reconstruction. However, there is obviously no direct feedback in the sense that stereo cannot take advantage from any shading cues.

Joint Approaches. In contrast to subsequent methods, joint approaches exploit stereo and shading cues simultaneously when estimating the depth. For example, Fua *et al*. [12] propose to minimise an objective function that allows a slowly varying albedo, but assumes the illumination to be known. Moreover, in the context of face reconstruction Samaras *et al*. [29] developed a method that fits a face model to the stereo data and refines it while re-estimating illumination and albedo. Although the two previous approaches make use of both shading and stereo information during the refinement, they are still based on an initial stereo mesh and

hence cannot be considered to be fully joint. A method that does not require such an initial mesh is the level set approach of Jin *et al.* [19]. However, although the corresponding model considers ambient light as well as an explicit background, it is restricted to two regions with constant albedo as well as a global light direction. Moreover, also this approach cannot be considered fully joint, since it relies on multi-view SfS and hence only exploits stereo cues implicitly. Summarising, given the existing literature, it would be desirable to develop a *fully joint* approach that combines stereo and shading cues *explicitly* and *simultaneously* estimates *depth*, *illumination* and *albedo* from scratch (i.e. a general approach without initial mesh).

Contributions. In this paper, we present such a fully joint approach. We propose a novel view-centred variational method that combines data terms from multi-view stereo and SfS based on a separate parametrisation for depth, illumination and albedo. Moreover, we make use of an anisotropic second-order smoothness term that enables the detail-preserving reconstruction of non-fronto-parallel surfaces. As a result we obtain a rather general method that allows to estimate high-quality depth maps of Lambertian scenes with varying albedo under unknown illumination. Finally, we also propose a coarse-to-fine minimisation scheme based on a linearisation of the data terms. This in turn allows the joint estimation of all unknowns.

Organisation. In Section 2 we derive our novel variational model for combining SfS and stereo. The minimisation of the corresponding energy by means of an incremental coarse-to-fine approach is then discussed in Section 3. Experiments on both synthetic and real-world data are presented in Section 4. Finally, Section 5 concludes the paper with a summary.

2 Variational Model

In this section we introduce our variational framework which allows to jointly exploit shading and disparity cues. Our setting consists of $n+1$ perspective cameras C_i ($i \in \{0,\dots,n\}$, $n \geq 1$), where we choose C_0 to be the reference camera located at the origin of the coordinate system. Let $\pi_i : \mathbb{R}^3 \to \Omega_i$ denote the perspective projection of a 3-D point on the image plane $\Omega_i \subset \mathbb{R}^2$ corresponding to the camera C_i and let $\mathbf{I}_i : \Omega_i \to \mathbb{R}^3$ be the captured RGB colour image. Moreover, let the unknown Lambertian surface $\mathcal{S} : \Omega_0 \to \mathbb{R}^3$ be parametrised as $\mathcal{S}(\mathbf{x}_0) = (z \cdot x_0/\mathtt{f},\, z \cdot y_0/\mathtt{f},\, -z)^\top$ [20], where $\mathtt{f}$ denotes the focal length, $\mathbf{x}_0 = (x_0, y_0)^\top \in \Omega_0$ is the position on the reference image plane and $z(\mathbf{x}_0)$ stands for the depth orthogonal to the image plane. Finally, let the unknown illumination conditions be modelled in terms of an illumination vector field $\mathbf{l} : \Omega_0 \to \mathbb{R}^3$ similar to the overall parametrisation of Xu *et al.* [39] and let the vector-valued albedo (product of reflectivity and colour) be denoted by $\rho : \Omega_0 \to \mathbb{R}^3$. We then propose to compute the depth, the illumination vector field and the albedo as minimiser of the following energy functional:

$$E(z, \mathbf{l}, \rho) = D_{\mathsf{stereo}}(z) + \nu \cdot D_{\mathsf{sfs}}(z, \mathbf{l}, \rho) + \alpha_z \cdot R_{\mathsf{depth}}(z) + \alpha_{\mathbf{l}} \cdot R_{\mathsf{illum}}(\mathbf{l}) + \alpha_\rho \cdot R_{\mathsf{albedo}}(\rho)\,. \quad (1)$$

It is composed of two data terms and three regularisation terms. While the stereo data term D_{stereo} accounts for photo consistency between the reference image and the other views, the SfS data term D_{sfs} relates the reference image and the reprojected image based on depth, illumination, and albedo. In order to resolve ambiguities between the unknowns, the regularisers R_{depth}, R_{illum}, and R_{albedo} have been added. Finally, the terms are balanced by the weights ν, α_z, $\alpha_{\mathbf{l}}$, and α_ρ. Let us now discuss the data and the smoothness terms in detail.

Stereo Data Term. For the multi-view stereo data term we consider the depth-parametrised model of Robert and Deriche [28] based on the photo consistency (i.e. brightness constancy) of projected surface points. This model is frequently used by recent stereo approaches, see e.g. [1, 24, 31]. It is complemented by a gradient constancy assumption [5, 31], which may account for slight illumination changes between subsequently recorded views. This yields

$$D_{\mathsf{Stereo}}(z) = \int_{\Omega_0} \frac{1}{n} \sum_{i=1}^{n} \Psi_L\Big(\|\mathbf{I}_0(\mathbf{x}_0) - \mathbf{I}_i(\mathbf{x}_i)\|_2^2 \Big) + \gamma\, \Psi_L\Big(\|\nabla \mathbf{I}_0(\mathbf{x}_0) - \nabla \mathbf{I}_i(\mathbf{x}_i)\|_2^2 \Big)\, \mathrm{d}\mathbf{x}_0 \,. \quad (2)$$

Here, $\nabla = \left(\partial_{x_0}, \partial_{y_0}\right)^\top$ is the spatial gradient, $\mathbf{x}_i = \pi_i\left(\mathcal{S}\left(\mathbf{x}_0\right)\right)$ denotes the projection of the surface point $\mathcal{S}\left(\mathbf{x}_0\right)$ at the reference coordinates $\mathbf{x}_0$ onto the image $\mathbf{I}_i$, and γ is a weight to balance both terms. To improve the robustness of both assumptions with respect to outliers and occlusions, we finally apply the robust function $\Psi_L(s^2) = \sqrt{s^2 + \varepsilon^2}$ with $\varepsilon > 0$ [2, 6], that penalises violations of the constancy assumptions less severely than in a quadratic setting.

SfS Data Term. Before we define the SfS data term, let us first review the corresponding image formation process. This will lead us to the desired parametrisation in terms of depth, albedo and illumination. Considering the rendering equation [18, 21] we obtain the following Lambertian model for the RGB valued reference image

$$\mathbf{I}_0(\mathbf{x}_0) = \int_{\Omega(\mathbf{s})} \rho(\mathbf{x}_0) L(\mathbf{s}, \omega)(\omega^\top \mathbf{n})\, d\omega \,. \quad (3)$$

Here, $\mathbf{s} = \mathcal{S}\left(\mathbf{x}_0\right)$ denotes the surface point at position $\mathbf{x}_0$, $\mathbf{n}$ is the surface normal, $\Omega(\mathbf{s})$ represents the unit hemisphere centred around $\mathbf{n}$, ω is the negative incident light direction and $L(\mathbf{s}, \omega)$ stands for the incident radiance from direction ω at $\mathbf{s}$. Following Xu *et al.* [39] and integrating the contributions of the incident radiance over the angles of the hemisphere we obtain the following compact parametrisation in terms of an illumination vector field $\mathbf{l}$:

$$\mathbf{I}_0(\mathbf{x}_0) = \rho(\mathbf{x}_0) \underbrace{\left(\int_{\Omega(\mathbf{s})} L(\mathbf{s}, \omega)\, \omega\, d\omega \right)^\top}_{=\mathbf{l}(\mathbf{x}_0)} \mathbf{n}(\mathbf{x}_0) \quad (4)$$

In contrast to [39], however, we do not assume the albedo to be constant and thus explicitly separate it from the illumination vector field. This makes the resulting approach applicable to more realistic scenarios where the albedo is spatially varying. Finally, we are in the position to introduce the SfS data term. It is given by

$$D_{\mathsf{sfs}}(z, \mathbf{l}, \rho) = \int_{\Omega_0} \|\mathbf{I}_0(\mathbf{x}_0) - \mathbf{R}(\mathbf{x}_0)\|_2^2 \, \mathrm{d}\mathbf{x}_0 \,, \quad (5)$$

where $\mathbf{R}(\mathbf{x}_0) = \rho(\mathbf{x}_0)\left(\mathbf{l}(\mathbf{x}_0)^\top \mathbf{n}(\mathbf{x}_0)\right)$ and where the surface normal $\mathbf{n}$ can be computed as the cross product of the partial surface derivatives $\mathcal{S}_x$ and $\mathcal{S}_y$, which yields [20]:

$$\mathbf{n}(\mathbf{x}_0) = \frac{\mathbf{v}(\mathbf{x}_0)}{||\mathbf{v}(\mathbf{x}_0)||_2} \quad \text{with} \quad \mathbf{v}(\mathbf{x}_0) = \Big(\mathrm{f} \cdot z_x,\, \mathrm{f} \cdot z_y,\, \left(\nabla z^\top \mathbf{x}_0\right) + z \Big)^\top \,. \quad (6)$$

Depth Regularisation. In order to allow for slanted surfaces in the reconstruction, we refrain from using a first-order regulariser that inherently favours fronto-parallel surfaces. Instead, we resort to second-order smoothness terms that allow to model linear depth changes [20,

27, 30, 35]. To be more precise, we make use of the anisotropic second-order regulariser of Hafner *et al*. [15] that was originally proposed in the context of denoising and focus fusion. It combines the edge preservation properties of second-order coupling models with directional image information to guide the reconstruction. The corresponding smoothness term reads

$$R_{\text{depth}}(z) = \inf_{\mathbf{u}} \int_{\Omega_0} \left(C(z,\mathbf{u}) + \alpha_{\mathbf{u}} \cdot S(\mathbf{u}) \right) \mathrm{d}\mathbf{x}_0 \,, \tag{7}$$

with

$$C(z,\mathbf{u}) = \sum_{d=1}^{2} \Psi_C^d \left(\left(\mathbf{r}_d^\top \left(\nabla z - \mathbf{u} \right) \right)^2 \right) , \quad S(\mathbf{u}) = \sum_{d=1}^{2} \Psi_S^d \left(\sum_{m=1}^{2} \left(\mathbf{r}_m^\top \mathcal{J}(\mathbf{u}) \mathbf{r}_d \right)^2 \right) , \tag{8}$$

where $\mathbf{r}_1$ and $\mathbf{r}_2$ are orthogonal unit vectors that correspond to the dominant directions of the local structure of the reference image $\mathbf{I}_0 = \left(I_0^1, I_0^2, I_0^3\right)^\top$, respectively. These directions can be computed as eigenvectors of the colour structure tensor [10, 11]:

$$J := K_{\sigma_o} * \sum_{c=1}^{3} \left(\nabla (K_{\sigma_i} * I_0^c) \, \nabla (K_{\sigma_i} * I_0^c)^\top \right) , \tag{9}$$

where K_{σ_i} and K_{σ_o} are spatial Gaussians with standard deviation σ_i and σ_o for presmoothing and local integration, respectively, and $*$ is the convolution operator.

Let us now detail the two terms in (7) that are balanced by $\alpha_{\mathbf{u}}$: While the coupling term C connects the gradient ∇z of the depth field to a coupling variable $\mathbf{u}$, the smoothness term S ensures that the Jacobian $\mathcal{J}(\mathbf{u})$ of this coupling variable is small. Evidently, this strategy realises a second-order smoothness constraint on z. In fact, for the special case $\mathbf{u} = \nabla z$, the smoothness term S actually penalises the second-order directional derivatives $z_{\mathbf{r}_1\mathbf{r}_1}$, $z_{\mathbf{r}_1\mathbf{r}_2}$, $z_{\mathbf{r}_2\mathbf{r}_1}$ and $z_{\mathbf{r}_2\mathbf{r}_2}$. In this case, the resulting regulariser comes down to a second-order extension of the anisotropic complementary smoothness term of Zimmer *et al*. [44].

So far, we have discussed how second-order regularisation can be realised by using an auxiliary variable $\mathbf{u}$. Let us now explain how to achieve the desired anisotropic detail-preserving behaviour. Main concept in this context is the separate sub-quadratic penalisation of the two directions $\mathbf{r}_1$ and $\mathbf{r}_2$ in the coupling term (via Ψ_C^1, Ψ_C^2) and in the smoothness term (via Ψ_S^1, Ψ_S^2), respectively. This not only adapts the regularisation to the local image structure by considering the directions $\mathbf{r}_1$ and $\mathbf{r}_2$ from the structure tensor J, it also allows to preserve edges in *both* directions *independently*. This in turn allows to handle important scenarios such as corners (two edges), edges (one edge) and homogeneous areas (no edge).

Applying the separate penalisation to the smoothness and to the data term has different effects. While applying separate penalisation to the smoothness term yields an anisotropic regularisation of the auxiliary variable $\mathbf{u}$, applying it to the coupling term gives a second order anisotropic regularisation of the depth z. In the latter case only deviations from ∇z are penalised that cannot be explained by a piecewise constant (smooth) auxiliary variable $\mathbf{u}$. Such a piecewise constant (smooth) auxiliary variable $\mathbf{u}$ in turn corresponds in to a piecewise affine depth z which makes once again the second-order regularisation explicit. As penalising functions we chose the edge-enhancing Perona-Malik penaliser [25] $\Psi_C^1(s^2) = \Psi_S^1(s^2) = \lambda^2 \log\left(1 + s^2/\lambda^2\right)$ along the dominant direction (i.e. in $\mathbf{r}_1$-direction) and the edge-preserving Charbonnier penaliser [7] $\Psi_C^2(s^2) = \Psi_S^2(s^2) = 2\lambda^2(1 + s^2/\lambda^2)^{1/2}$ orthogonal to it (i.e. in $\mathbf{r}_2$-direction); see e.g. [36]. In both cases, λ plays the role of a contrast parameter.

Illumination Regularisation. As shown by Xu *et al.* [39] the illumination vector $\mathbf{l}$ is typically piecewise constant or only varies smoothly across the surface. Based on this finding we use the following isotropic first-order regulariser

$$R_{\text{illum}}(\mathbf{l}) = \int_{\Omega_0} \Psi_I\left(\|\mathcal{J}(\mathbf{l})\|_F^2\right) \mathrm{d}\mathbf{x}_0\,, \tag{10}$$

where $\mathcal{J}(\mathbf{l})$ denotes the Jacobian of $\mathbf{l}$, $\|\cdot\|_F$ is the Frobenius norm and Ψ_I represents the edge-preserving Charbonnier penaliser.

Albedo Regularisation. A common assumption in the field of intrinsic image decomposition is that pixels with similar chromaticity are likely to share a similar albedo [8]. Since we are interested in separating albedo from geometry and illumination, we follow this idea and make use of an isotropic first-order smoothness term which reduces smoothness at chromaticity edges. This is achieved using a positive, decreasing weighting function g applied to the Frobenius norm of the chromaticity Jacobian, which serves as fuzzy edge detector for chromaticity edges. Hence we obtain

$$R_{\text{albedo}}(\rho) = \int_{\Omega_0} g\left(\|\mathcal{J}(\mathbf{ch}(\mathbf{I_0}))\|_F^2\right) \cdot \|\mathcal{J}(\rho)\|_F^2 \,\mathrm{d}\mathbf{x}_0\,, \tag{11}$$

where $\mathbf{ch}(\mathbf{I_0}) = \mathbf{I_0}/(I_0^1+I_0^2+I_0^3)$ denotes the rg-chromaticity, $\mathcal{J}(\mathbf{ch}(\mathbf{I_0}))$ is the chromaticity Jacobian and $g(s^2) = 1/(1+s^2/\lambda^2)$ is the Perona-Malik diffusivity [25]. Since $g(s^2) \approx 0$ for large arguments $s^2 \gg \lambda^2$, jumps in the albedo are mainly aligned with chromaticity edges.

3 Minimisation

Differential Formulation. In order to minimise the non-convex energy functional in (1), we propose the use of an incremental coarse-to-fine fixed point strategy [5]. To this end, we approximate the original model by a series of differential energies. Given the known values z^k, $\mathbf{l}^k$, ρ^k and $\mathbf{u}^k$ from a coarser resolution level, we thereby compute the unknown increments dz^k, $\mathbf{dl}^k$, $\mathbf{d}\rho^k$ and $\mathbf{du}^k$ at each resolution level k. The differential formulation of the energy is derived as follows: In case of the stereo term we first linearise w.r.t. the depth increment and then introduce a constraint normalisation similar to [32, 44]. Regarding the SfS term we follow the recent work of Maurer *et al.* [23] and apply an upwind scheme approximation before performing the linearisation w.r.t. all unknowns. In contrast to the much simpler model of [23], however, we must additionally consider albedo and illumination when performing this linearisation. For the smoothness terms the differential formulation is straightforward. It essentially follows from the definition of the increments; see [5]. Finally, to ensure robustness w.r.t. large erroneous increments (e.g. if the linearisation is locally not valid), we extend our differential model by adding quadratic regularisation terms for the increments length. For more details regarding the differential formulation we refer to the supplementary material.

Numerical Solution. The resulting differential energy has to be minimised on each resolution level. To this end, the corresponding Euler-Lagrange equations are derived and discretised on a rectangular grid. Following [23], the derivatives of the linearised SfS term are computed numerically. Regarding the anisotropic depth regulariser we discretise the resulting divergence expressions as in [37]. For the remaining derivatives standard finite differences are used. Since the resulting system of equations is non-linear – due to the derivatives of the subquadratic penaliser functions Ψ_L, Ψ_C^d,Ψ_S^d and Ψ_I – we employ a second fixed-point iteration

in order to obtain a linear system of equations, where all the non-linear expressions are kept fixed; see [5]. Finally, the linear systems of equations are solved using the SOR method.

Implementation Details. Since we regularise the length of the increments, a single linearisation per resolution level is not sufficient. Hence, we perform several fixed point iterations per level. Moreover, at the coarsest level we initialise the depth z with a fronto-parallel plane, the illumination vector $\mathbf{l}$ with zero (not to prefer any particular direction), and the albedo ρ with the downsampled input image. Finally, to account for the fact that the zero initialisation of the illumination vector does not allow the SfS data term to provide any useful information at coarser levels, we introduce a weighting function that increases the SfS weight ν at finer resolutions. Also in this case we refer the reader to the supplementary material for more details.

4 Experiments

Experimental Setup. Experiments have been performed with a C++ implementation running on a single core with 3.40GHz on a standard desktop PC with Intel Core i7-2600 CPU. Using the parameters given in the supplementary material the runtime has been in the order of 1h 40m for input images of size 1536x1024.

Synthetic Data. In our first experiment, we compare the reconstruction quality of our combined approach with that of a pure stereo variant which is obtained by omitting the SfS term as well as the illumination and albedo regularisers. To this end, we created a synthetic dataset of *Blunderbuss Pete* with some artificial procedural Voronoi texturing using Blender. It consists of three views and the scene is illuminated by two distant light sources. The dataset comprises fine surface details and subtle geometry which pose a challenging task.

The reference image, the ground truth as well as the results for the pure stereo method and the combined approach are depicted in Figure 1. As one can see, the combined approach reconstructs fine surface details such as the eye and the beard much better than the pure stereo method that yields a somewhat coarser result. Moreover, the estimated albedo and the computed illumination direction in Figure 1 look quite reasonable. Thus it is not surprising that the clear visual improvement is also confirmed by a slight decrease of the *root mean square* (RMS) error of the computed depth maps: While the pure stereo method yields an error of $19.52 \cdot 10^{-5}$, the combined approach achieves $19.30 \cdot 10^{-5}$. In this context, one has to keep in mind that the improvement lies mainly in the reconstruction of small surface details.

Real-World Data. In order to investigate the performance of our approach when reconstructing real scenes and objects, we also conducted two experiments with real-world data. For our first experiment we used the *Angel* dataset from Wu *et al.* [38]. Once again, we computed results for our pure stereo method and for our combined approach, this time based on five views. Moreover, we added the results of the approach of Wu *et al.* [38] for comparison – a method that refines a pre-computed multi-view stereo mesh using shading information.

Once again, the corresponding reconstructions in Figure 2 show that the pure stereo variant is not able to capture all fine details such as the strands of hair, the disc area of the sunflower head or the toes. The method of Wu *et al.* does better, however, the overall reconstruction is far more smoothed. In particular, coarse structures such as the sunflower petals pointing towards the camera or the ringlet are over-smoothed. In contrast, our combined approach is able to recover both coarse and fine scale details accurately. This becomes particularly obvious when comparing our results to the reference input image.

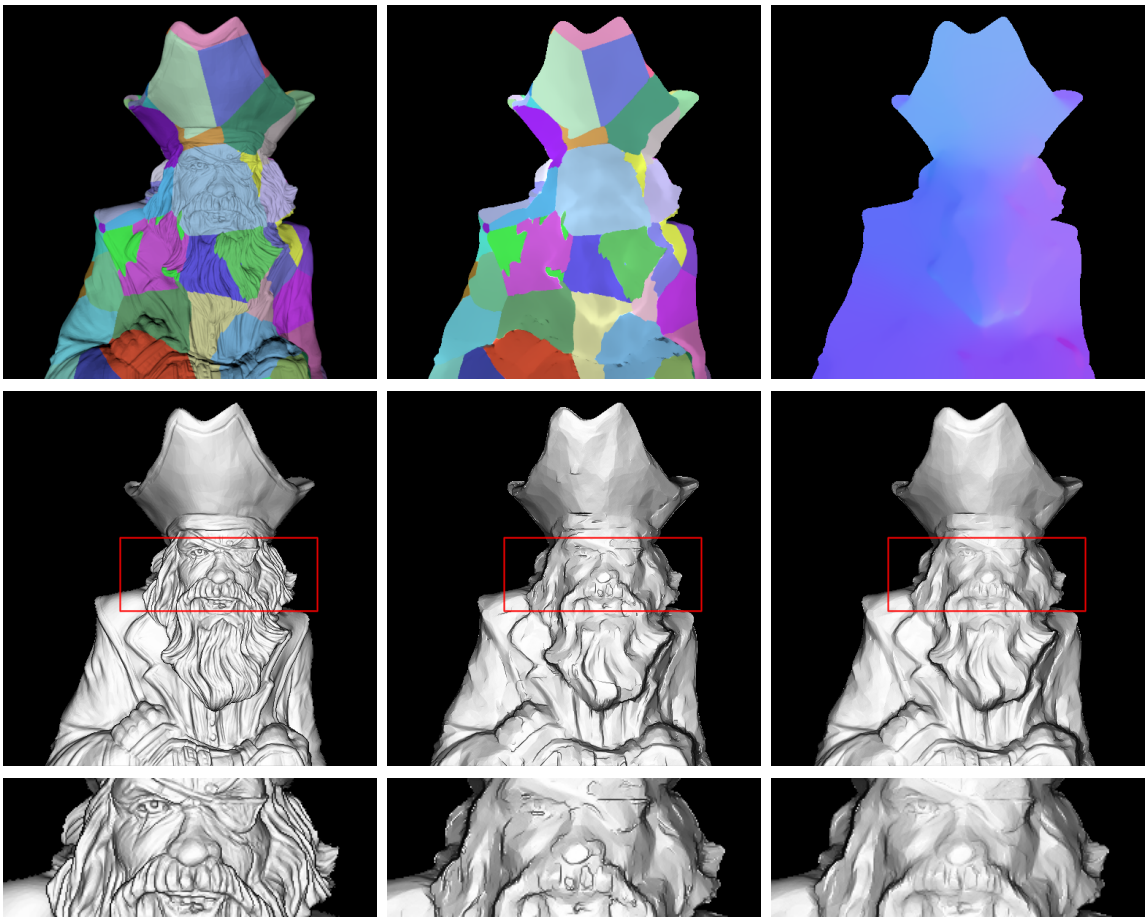

Figure 1: Synthetic *Blunderbuss Pete* dataset. **First row, from left to right**: Reference input image, computed albedo, computed illumination direction. **Second and third row, from left to right**: Shaded images showing the ground truth, pure stereo and our combined approach.

For our second real-world experiment we used the *Fountain-P11* and the *Herz-Jesu-P8* images of the Strecha dataset [33] for which the ground truth is available. Since the recovery of fine details strongly depends on the sharpness of the input data, we downsampled the slightly blurred images to half the resolution before reconstructing the scenes from only two views. This time, apart from the results of our combined method and its stereo variant, we also provide results for two recently proposed stereo approaches which are able to handle arbitrary camera settings and which provide source code publicly: On the one hand, we used the variational method of Graber *et al*. [14] that is based on minimal-surface regularisation. For the given data set this method has shown significant improvements compared to standard TV regularisation. On the other hand, we considered the basic approach of Galliani *et al*. [13] which is a multi-view variant of PatchMatch Stereo [4]. While Galliani *et al*. also proposed a final 3-D integration step in terms of fusing multiple reconstructions from different views, we had to omit this step here, since we focus on the computation of a single depth map.

Qualitative results for the *Fountain-P11* are depicted in Figure 3. While the multi-view PatchMatch method of Galliani *et al*. recovers major jumps very accurately, the corresponding reconstruction lacks fine details and contains significant outliers in occluded regions. The latter observation is a direct consequence of the lacking regularisation of the PatchMatch algorithm. In contrast, the approach of Graber *et al*. yields a more detailed reconstruction that is, however, very noisy. This in turn is a consequence of the minimal-surface regularisation that tends to round-off objects when suppressing local fluctuations and thus only preserves surface details if the amount of regularisation is chosen sufficiently low. In comparison, the

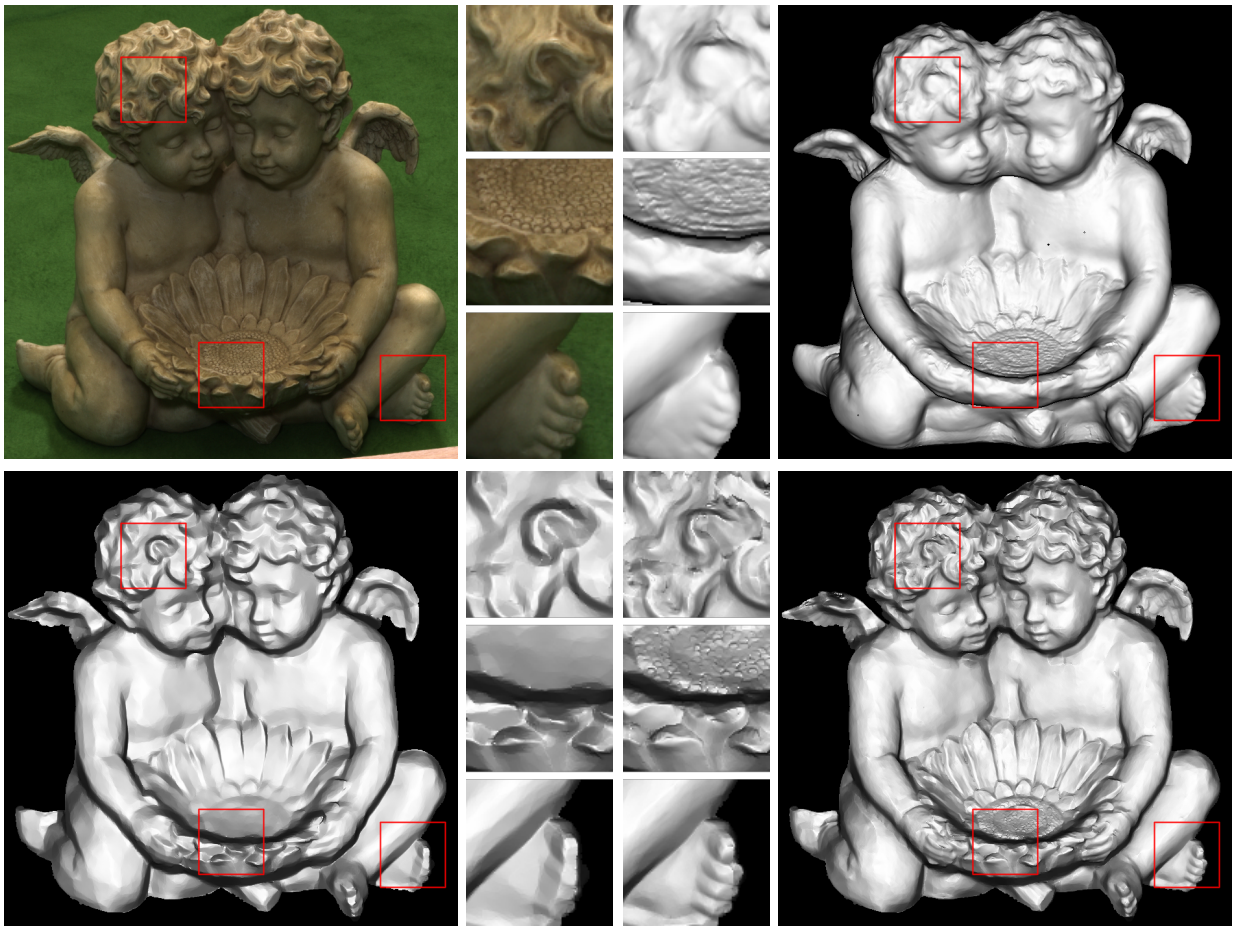

Figure 2: Real-world *Angel* dataset [38]. **Top left**: Reference image. **Top right**: Wu *et al.* [38]. **Bottom left**: Our pure stereo approach. **Bottom right**: Our combined approach.

reconstruction of our pure stereo method is already quite accurate. While flat surfaces are almost noise free, details of the fountain and the wall are more pronounced. This clearly shows the benefits of the edge-preserving anisotropic second-order regularisation. The visually most appealing reconstruction, however, is achieved by our combined approach. It recovers even fine scale details such as the mouth of the fish and the ornaments of the fountain. This in turn demonstrates the usefulness of shading information. Our findings are confirmed by the quantitative comparison of the results in Table 1. It shows that the RMS errors of our methods are significantly lower than those of the other two approaches both for the *Fountain-P11* as well as for the *Herz-Jesu-P8* data set. More results such as the *Herz-Jesu-P8*-reconstructions can be found in the supplementary material.

method	Fountain-P11		Herz-Jesu-P8	
	all	non-occluded	all	non-occluded
Graber *et al.* [14]	0.0688	0.0367	0.2217	0.0535
Graber *et al.* (CUDA)	0.0264[1]	-	-	-
Galliani *et al.* [13]	0.6124	0.0157	3.2813	0.9632
Ours (stereo)	0.0168	0.0023	0.0706	0.0328
Ours (stereo + SfS)	0.0134	0.0022	0.0695	0.0325

Table 1: RMS error for the *Fountain-P11* and *Herz-Jesu-P8* dataset.

[1] While the publicly available Python code does not achieve such low errors – even with optimised parameters – they have been reported for the non-publicly available CUDA code (for full resolution images) [14].

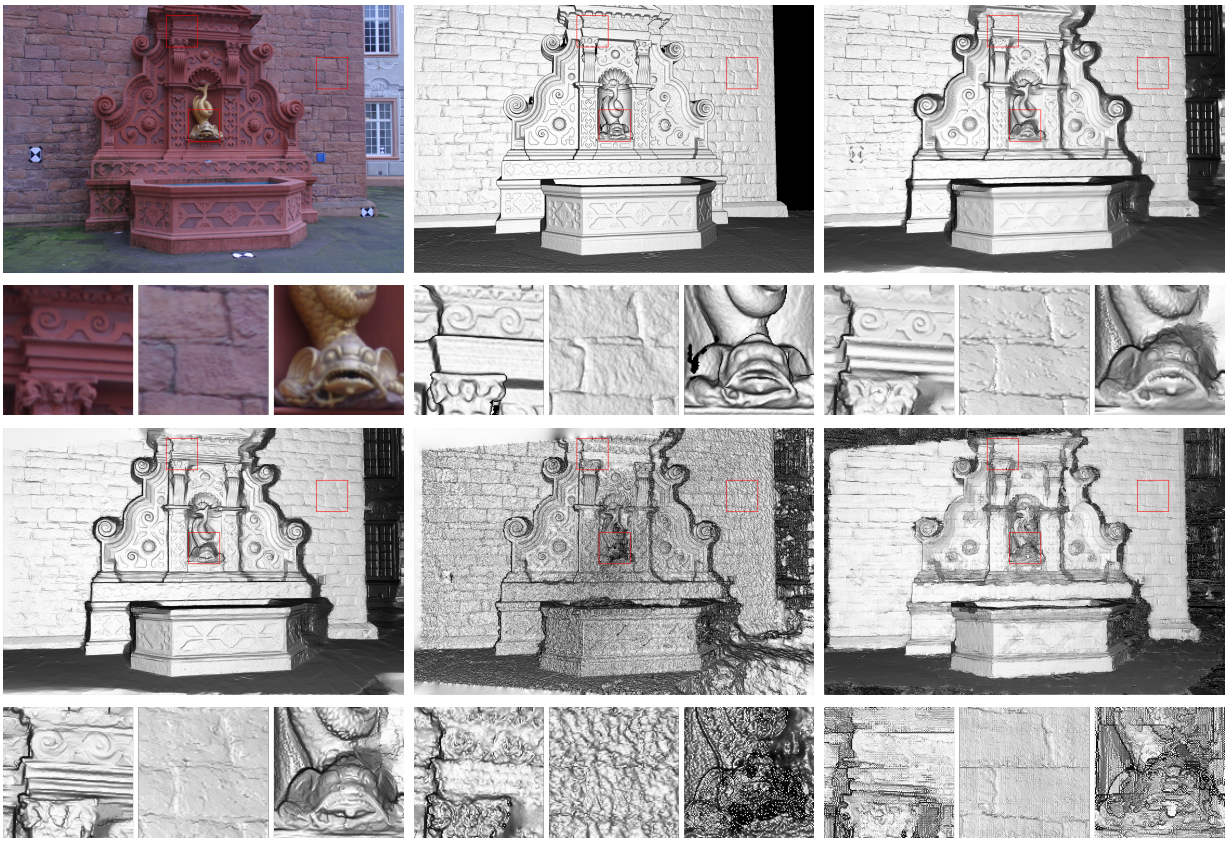

Figure 3: Real-world *Fountain-P11* dataset [33]. Two-view results. **Top left**: Reference image. **Top centre**: Ground truth. **Top right**: Our combined approach. **Bottom left**: Our pure stereo approach. **Bottom center**: Graber *et al.* [14]. **Bottom right**: Galliani *et al.* [13].

5 Conclusions and Outlook

In this paper we have proposed a novel variational method for combining shape from shading and stereo. In this context, our contribution was fourfold: (i) We showed how shading and disparity information can be integrated explicitly into a joint minimisation framework for estimating the depth. In contrast to most existing approaches we thereby refrained from using any form of stereo-based pre-estimation. (ii) We made use of an adaptive anisotropic second-order smoothness term. This term further encouraged the detail-preserving reconstruction of non-fronto-parallel surfaces. (iii) We extended this model in such a way that it additionally allows to estimate albedo and illumination. This made our approach applicable to more general scenarios including Lambertian objects with non-uniform albedo and scenes with unknown illumination. (iv) We derived a coarse-to-fine minimisation framework based on a linearisation of all data terms. This in turn enabled the application of standard optimisation techniques such as nested fixed point iterations. Experiments for synthetic and real-world images demonstrate that our combined approach allows for accurate and detailed reconstructions. Moreover, they show that shading information is indeed useful to improve upon pure stereo, in particular when it comes to the reconstruction of small-scale details.

Future work includes the use of more advanced reflection models, e.g. the Phong model for handling specular reflections [26], as well as the fusion of multiple depth maps to obtain a single high resolution reconstruction, see e.g. [13, 43].

Acknowledgements. This work has been partly funded by the German Research Foundation (DFG) within the joint project BR 2245/3-1 and BR 4372/1-1. Moreover, we thank the DFG for financial support within project B04 of SFB/Transregio 161.

References

[1] T. Basha, Y. Moses, and N. Kiryati. Multi-view scene flow estimation: a view centered variational approach. *International Journal of Computer Vision*, 101(1):6–21, 2012.

[2] M. J. Black and P. Anandan. Robust dynamic motion estimation over time. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 296–302, 1991.

[3] A. Blake, A. Zisserman, and G. Knowles. Surface descriptions from stereo and shading. *Image Vision Computation*, 3(4):183–191, 1985.

[4] M. Bleyer, C. Rhemann, and C. Rother. PatchMatch stereo-stereo matching with slanted support windows. In *Proc. British Machine Vision Conference*, pages 1–11, 2011.

[5] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *Proc. European Conference on Computer Vision*, pages 25–36, 2004.

[6] A. Bruhn and J. Weickert. Towards ultimate motion estimation: combining highest accuracy with real-time performance. In *Proc. IEEE International Conference on Computer Vision*, pages 749–755, 2005.

[7] P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud. Deterministic edge-preserving regularization in computed imaging. *IEEE Transactions on Image Processing*, 6(2):298–311, 1997.

[8] Q. Chen and V. Koltun. A simple model for intrinsic image decomposition with depth cues. In *Proc. IEEE International Conference on Computer Vision*, pages 241–248, 2013.

[9] J. E. Cryer, P.-S. Tsai, and M. Shah. Integration of shape from shading and stereo. *Pattern Recognition*, 28(7):1033–1043, 1995.

[10] S. Di Zenzo. A note on the gradient of a multi-image. *Computer Vision, Graphics and Image Processing*, 33:116–125, 1986.

[11] W. Förstner and E. Gülch. A fast operator for detection and precise location of distinct points, corners and centres of circular features. In *Proc. ISPRS Intercommission Conference on Fast Processing of Photogrammetric Data*, pages 281–305, 1987.

[12] P. Fua and Y. G. Leclerc. Object-centered surface reconstruction: Combining multi-image stereo and shading. *International Journal of Computer Vision*, 16(1):35–56, 1995.

[13] S. Galliani, K. Lasinger, and K. Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proc. IEEE International Conference on Computer Vision*, pages 873–881, 2015.

[14] G. Graber, J. Balzer, S. Soatto, and T. Pock. Efficient minimal-surface regularization of perspective depth maps in variational stereo. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 511–520, 2015.

[15] D. Hafner, C. Schroers, and J. Weickert. Introducing maximal anisotropy into second order coupling models. In *Proc. German Conference on Pattern Recognition*, pages 79–90, 2015.

[16] T. S. F. Haines and R. C. Wilson. Integrating stereo with shape-from-shading derived orientation information. In *Proc. British Machine Vision Conference*, pages 1–10, 2007.

[17] D. R. Hougen and N. Ahuja. Adaptive polynomial modelling of the reflectance map for shape estimation from stereo and shading. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 991–994, 1994.

[18] D. S. Immel, M. F. Cohen, and D. P. Greenberg. A radiosity method for non-diffuse environments. In *Proc. SIGGRAPH*, pages 133–142, 1986.

[19] H. Jin, D. Cremers, D. Wang, E. Prados, A. Yezzi, and S. Soatto. 3-D reconstruction of shaded objects from multiple images under unknown illumination. *International Journal of Computer Vision*, 76(3):245–256, 2008.

[20] Y. C. Ju, D. Maurer, M. Breuß, and A. Bruhn. Direct variational perspective shape from shading with Cartesian depth parametrisation. In *Perspectives in Shape Analysis*, Mathematics and Visualization. 2016.

[21] J. T. Kajiya. The rendering equation. In *Proc. SIGGRAPH*, pages 143–150, 1986.

[22] Y. G. Leclerc and A.F. Bobick. The direct computation of height from shading. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 552–558, 1991.

[23] D. Maurer, Y. C. Ju, M. Breuß, and A. Bruhn. An efficient linearisation approach for variational perspective shape from shading. In *Proc. German Conference on Pattern Recognition*, pages 249–261, 2015.

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

[25] P. Perona and J. Malik. Scale space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2:629–639, 1990.

[26] B. T. Phong. Illumination for computer generated pictures. *Communications of the ACM*, 18(6):311–317, 1975.

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

[28] L. Robert and R. Deriche. Dense depth map reconstruction: a minimization and regularization approach which preserves discontinuities. In *Proc. European Conference on Computer Vision*, pages 439–451, 1996.

[29] D. Samaras, D. Metaxas, P. Fua, and Y. G. Leclerc. Variable albedo surface reconstruction from stereo and shape from shading. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 480–487, 2000.

[30] C. Schroers, D. Hafner, and J. Weickert. Multiview depth parameterisation with second order regularisation. In *Proc. International Conference on Scale Space and Variational Methods in Computer Vision*, pages 551–562, 2015.

[31] B. Semerjian. A new variational framework for multiview surface reconstruction. In *Proc. European Conference on Computer Vision*, pages 719–734, 2014.

[32] E. P. Simoncelli, E. H. Adelson, and D. J. Heeger. Probability distributions of optical flow. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 310–315, 1991.

[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 *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[34] L. Valgaerts, C. Wu, A. Bruhn, H.-P. Seidel, and C. Theobalt. Lightweight binocular facial performance capture under uncontrolled lighting. *ACM Transactions on Graphics*, 31(6):1–11, 2012.

[35] O. Vogel, A. Bruhn, J. Weickert, and S. Didas. Direct shape-from-shading with adaptive higher order regularisation. In *Proc. International Conference on Scale Space and Variational Methods in Computer Vision*, pages 871–882, 2007.

[36] S. Volz, A. Bruhn, L. Valgaerts, and H. Zimmer. Modeling temporal coherence for optical flow. In *Proc. IEEE International Conference on Computer Vision*, pages 1116–1123, 2011.

[37] J. Weickert, M. Welk, and M. Wickert. L^2-stable nonstandard finite differences for anisotropic diffusion. In *Proc. International Conference on Scale Space and Variational Methods in Computer Vision*, pages 380–391, 2013.

[38] C. Wu, B. Wilburn, Y. Matsushita, and C. Theobalt. High-quality shape from multi-view stereo and shading under general illumination. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 969–976, 2011.

[39] D. Xu, Q. Duan, J. Zheng, J. Zhang, J. Cai, and T.-J. Cham. Recovering surface details under general unknown illumination using shading and coarse multi-view stereo. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1526–1533, 2014.

[40] K.-J. Yoon, E. Prados, and P. Sturm. Joint estimation of shape and reflectance using multiple images with known illumination conditions. *International Journal of Computer Vision*, 86(2-3):192–210, 2010.

[41] T. Yu, N. Xu, and N. Ahuja. Recovering shape and reflectance model of non-Lambertian objects from multiple views. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 226–233, 2004.

[42] T. Yu, N. Xu, and N. Ahuja. Shape and view independent reflectance map from multiple views. *International Journal of Computer Vision*, 73(2):123–138, 2007.

[43] C. Zach, T. Pock, and H. Bischof. A globally optimal algorithm for robust TV-L^1 range image integration. In *Proc. IEEE International Conference on Computer Vision*, pages 1–8, 2007.

[44] H. Zimmer, A. Bruhn, and J. Weickert. Optic flow in harmony. *International Journal of Computer Vision*, 93(3):368–388, 2011.