A splitting-based algorithm for multi-view stereopsis of textureless objects

Jean Mélou[1,3(✉)], Yvain Quéau[2], Fabien Castan[3], and Jean-Denis Durou[1]

[1] IRIT, UMR CNRS 5505, Université de Toulouse, France
[2] GREYC, UMR CNRS 6072, Caen, France
[3] Mikros Image, Paris, France
✉ jeme@mikrosimage.eu

Abstract. We put forward a simple, yet effective splitting strategy for multi-view stereopsis. It recasts the minimization of the classic photo-consistency + gradient regularization functional as a sequence of simple problems which can be solved efficiently. This framework is able to handle various photo-consistency measures and regularization terms, and can be used for instance to estimate either a minimal-surface or a shading-aware solution. The latter makes the proposed approach very effective for dealing with the well-known problem of textureless objects 3D-reconstruction.

Keywords: Multi-view stereo · 3D-reconstruction · Shape-from-shading

1 Introduction

Multi-view stereopsis consists in reconstructing dense 3D-geometry from multi-view images. A common approach to this problem is to estimate a mapping (depth) between pixels in a reference view and 3D-geometry, by maximizing the photo-consistency of the reference image with the others. To measure photo-consistency, the reference image is warped to the other views using the estimated depth map and the (known) relative poses, and compared against the target images. However, photo-consistency is not significant in textureless areas (see Figure 1): the optimization problem needs to be regularized by constraining variations in the 3D-geometry. If $z : \Omega \subset \mathbb{R}^2 \to \mathbb{R}^+\backslash\{0\}$ denotes the unknown depth map, with Ω the image domain, surface variations under perspective projection can be measured as a function of $\frac{\nabla z}{z} = \nabla \log z : \Omega \to \mathbb{R}^2$. Multi-view stereo can then be formulated in a classic and generic manner [13] as the minimization of the sum of a fidelity term f inversely proportional to photo-consistency, and of a regularization term g. We thus consider in this work the following variational problem:

$$\min_z f(z) + g(\nabla \log z) \tag{1}$$

(the choice of applying regularization in log-space is discussed in Section 2).

Possible choices for f and g are discussed in Section 2. Section 3 introduces our main contribution: a generic multi-view stereo algorithm, which recasts (1) as a sequence of nonlinear-yet-local and global-yet-linear problems, both of which can be solved efficiently. We present in Section 4 appropriate regularizers for the 3D-reconstruction of textureless objects, before empirically evaluating the potential of our algorithm in Section 5. Eventually, our conclusions are drawn and future research directions are suggested in Section 6.

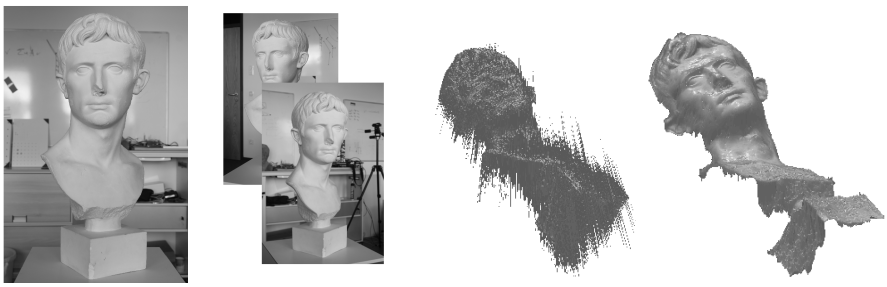

Fig. 1. Given a reference view of a textureless object (first column), and a set of $t \geq 1$ target views (second column), multi-view stereopsis based solely on photo-consistency optimization fails to estimate a reasonable mapping (depth) between the reference view and 3D-geometry (third column). Much more satisfactory results are obtained when introducing shading-aware and/or minimal-surface regularizations (fourth column).

2 Preliminaries

In this work we focus on solving the discrete counterpart of (1). In the following, z is thus a vector in $\mathbb{R}^p$ containing the p unknown depth values, $\log z$ is to be understood in a element-wise manner and $\nabla \in \mathbb{R}^{2p \times p}$ is a first-order, forward finite differences matrix such that $\nabla z \in \mathbb{R}^{2p}$ approximates the depth gradient.

Fidelity term. Let $\pi_z^{-1}(i)$ be the back-projection from the i-th pixel, $i \in \{1, \ldots, p\}$, in the reference view to its conjugate 3D-point, given a depth map z and the (known) intrinsics of the camera. Let $\left\{\pi^j\right\}_{j \in \{1,\ldots,t\}}$ be the projections from 3D-points to pixels in the $t \geq 1$ other cameras (hereafter "target cameras"), using the (known) camera poses and their (known) intrinsics. Let $v_i \in \mathbb{R}^m$ be a m-dimensional feature vector at pixel i in the reference view. Such a vector can be the brightness value in pixel i ($m = 1$), the RGB values in that pixel ($m = 3$), the concatenation of brightness values in a 3×3 neighborhood centered in i ($m = 9$), etc. For a given target camera $j \in \{1, \ldots, t\}$, photo-consistency measures the adequation between v_i and the feature vector $v^j_{\pi^j \circ \pi_z^{-1}(i)} \in \mathbb{R}^m$ at the matched pixel $\pi^j \circ \pi_z^{-1}(i)$ in the target view, in the sense of some loss function ρ.

The fidelity term can then be constructed by averaging the photo-consistency contributions from all target cameras and summing over all pixels:

$$f(z) = \frac{1}{t} \sum_{i=1}^{p} \sum_{j=1}^{t} \rho \left(v_i, v^j_{\pi^j \circ \pi_z^{-1}(i)} \right). \tag{2}$$

One could consider as loss function ρ the normalized sum of squared deviations $\rho_{\text{SSD}}(x,y) = \frac{1}{m} \sum_{c=1}^{m} (x_c - y_c)^2$ or a robust variant of it, and then linearize (2) using first-order Taylor expansion as in [7,12]. However, linearization requires small depth increments. Robustness can also be reached by replacing SSD with the normalized sum of absolute deviations $\rho_{\text{SAD}}(x,y) = \frac{1}{m} \sum_{c=1}^{m} |x_c - y_c|$ or a loss function based on the zero-mean normalized cross-correlation such as $\rho_{\text{ZNCC}}(x,y) = \frac{1}{2} \left[1 - \frac{(x-\overline{x})^\top (y-\overline{y})}{\|x-\overline{x}\| \|y-\overline{y}\|} \right]$ (see [5, Chapter 2]). Photo-consistency measures are then further normalized within $(0,1)$ using a nonlinear operator, e.g., the exponential transform $\rho(x,y) := 1 - \exp\left\{ -\frac{\rho(x,y)^2}{\sigma^2} \right\}$ with user-defined parameter σ. With all these choices, the fidelity term may become nonlinear, non-smooth and non-convex, and the optimization tedious. Therefore, minimization of f is usually carried out using bruteforce grid-search over the sampled depth space. This "winner-takes-all" strategy was first advocated in [8]. Despite its simplicity, it is remarkably efficient, and impressive depth map reconstructions of highly textured scenes have long been demonstrated [6].

Regularization. Nevertheless, in textureless areas, the fidelity term degenerates (in each pixel, there are multiple depth values for which it is globally minimized). Obviously, increasing the number of views will have no effect on this issue, and one should rather rely on regularization. For instance, one may introduce a total variation prior [16]. However, under perspective projection total variation does not enforce physically sound constraints on the geometry: smoothing should rather be carried out using the minimal surface prior [7]. Yet, this would turn the regularizer into a bilinear form involving both the depth z and its gradient ∇z, making the optimization challenging. Re-parameterization avoids this issue: introducing the change or variable $\tilde{z} = \sqrt{z}$, the total surface area can be rewritten as a function $g(\nabla \tilde{z})$ [7]. A logarithmic change of variable $\tilde{z} = \log z$ can also be considered for the same purpose, as well as to derive a physics-based regularization term based on shape-from-shading under natural illumination [14]. This might be particularly interesting for us, because shape-from-shading algorithms typically assume constant scene reflectance, which is essentially another wording for textureless scenes. Because we believe it is interesting to compare smoothness-based and physics-based priors for multi-view stereopsis of poorly textured objects within the same numerical framework, we opt for the logarithmic change of variable in the regularization term, which explains the form of the variational model (1). Shading-aware multi-view stereo has long been identified as a promising track [3], and theoretical guarantees on uniqueness exist [4]. Still, there is a lack of practical numerical solutions. Jin *et al.* presented in [9] a variational one, yet it assumes a single, infinitely distant light source, while we

are rather interested in natural illumination. Other methods combining stereo and shading information under natural illumination have also recently been developed [10,11,17], but they only consider photometry as a way to refine an existing multi-view 3D-reconstruction, which remains the baseline of the process. We would rather like to follow an end-to-end joint approach, as for instance in the very recent work [12]. In comparison with [12], the approach presented in the next section avoids linearization of the fidelity term and is therefore slightly more generic, since any robust photo-consistency measure (including those based on non-differentiable or non-convex loss functions) can be considered.

3 A generic splitting strategy for multi-view stereo

In this section, we show how to turn the discrete counterpart of the variational problem (1) into the simpler problem (6), and we introduce Algorithm 1 for solving the latter.

Proposed variational model. As discussed in the previous section, the fidelity term f in (1) is often chosen as a robust non-smooth cost function, hence the non-regularized problem is already challenging. The coupling induced by the gradient operator in the regularization term g makes things even worse. We separate those difficulties by splitting the optimization over f and g. Introducing an auxiliary variable $u = z \in \mathbb{R}^p$, (1) is equivalently rewritten as follows:

$$\begin{array}{ll} \min\limits_{u,z} & f(u) + g(\nabla \log z) \\ \text{s.t.} & u = z \end{array} . \tag{3}$$

In (3), the u-subproblem is still non-smooth and possibly non-convex, but at least it is now small-scale (and hence, parallelizable) since f is separable (each term in the outer sum in (2) only involves the depth in a single pixel). Minimization of f can be carried out using bruteforce grid-search over a set of sampled depth values. Moreover, assuming g is smooth, its minimization can be achieved using gradient-based optimization. However, the hard constraint $u = z$ would prevent z from capturing thin surface variations. Thus, we relax the hard constraint $u = z$ in (3) into a quadratic penalization term:

$$\min_{u,z} \; f(u) + g(\nabla \log z) + \beta \left\| \log u - \log z \right\|^2 , \tag{4}$$

with $\beta > 0$ a tunable hyper-parameter. Let us remark that the penalization is applied in log-space: in this way, z appears in (4) only through its logarithm. We can thus equivalently optimize over $\tilde{z} = \log z$, and recover $z = \exp \tilde{z}$ at the end of the process. The new optimization problem becomes:

$$\min_{u,\tilde{z}} \; f(u) + g(\nabla \tilde{z}) + \beta \left\| \log u - \tilde{z} \right\|^2 . \tag{5}$$

As mentioned in the previous section, recent studies have advocated in favor of nonlinear regularization terms g, and thus the $\tilde{z}$-subproblem in (5) remains challenging. We simplify it through a second splitting: introducing an auxiliary variable $\theta = \nabla\tilde{z} \in \mathbb{R}^{2p}$, Problem (5) is turned into the following, equivalent one:

$$\begin{array}{l} \min\limits_{u,\theta,\tilde{z}} \; f(u) + g(\theta) + \beta \left\| \log u - \tilde{z} \right\|^2 \\ \text{s.t.} \;\; \theta = \nabla\tilde{z} \end{array} . \tag{6}$$

Numerical solving of (6). The linear constraint in (6) could be handled, e.g., by resorting to an augmented Lagrangian approach, but in this preliminary work we rather follow a simpler strategy consisting in approximating the solution of (6) by iteratively solving quadratically-penalized problems of the form

$$\min_{u,\theta,\tilde{z}} f(u) + g(\theta) + \alpha^{(k)} \left\| \theta - \nabla\tilde{z} \right\|^2 + \beta \left\| \log u - \tilde{z} \right\|^2 , \tag{7}$$

with values of $\alpha^{(k)} > 0$ increasing to infinity with iterations k. We want the hard constraint in (6) to be satisfied at convergence i.e., when $k \to +\infty$, in contrast with the one in (3) which we purposely replaced by a quadratic penalization with fixed parameter β. For each value $\alpha^{(k)}$, we approximately solve (7) by one sweep of alternating optimization. As discussed above, the u-subproblem can be solved by grid-search. We focus on smooth and separable regularizers g (cf. Section 4), so the θ-subproblem can be solved using parallelized gradient-based iterations. Eventually, the $\tilde{z}$-subproblem is a sparse linear least-squares problem which can be solved using conjugate gradient. We repeat this process until the relative residual between two estimates of $z = \exp\tilde{z}$ falls below a threshold set to 10^{-4}. This algorithm is sketched in Algorithm 1. Intuitively, it iteratively estimates a rough depth map by optimizing photo-consistency (Equation (8)), then regularizes the depth variations (Equation (9)), and integrates the refined gradient into the log-depth map (Equation (10)). The values $\alpha^{(0)} = 1$ and $\beta = 0.1$ were empirically found to yield reasonable results and were used in all experiments. As initial depth map $z^{(0)}$, a fronto-parallel plane was always considered, with depth values taken as the mean of the ground-truth ones.

4 Regularizers for textureless multi-view stereopsis

Given that f and g might be non-convex, it seems difficult to draw a theoretical convergence analysis of Algorithm 1. We thus leave this analysis for the future, and rather focus in this exploratory work on evaluating the efficiency of the algorithm on real-world multi-view stereo problems. In particular, we now turn our attention to the challenging problem of reconstructing poorly textured objects, and discuss, in view of this, the choice of a suitable regularizer. This requires clarifying the notion of "textureless" objects, hence let us first recall some photometric notions.

input : Initial depth map $z^{(0)}$, $\alpha^{(0)} > 0$, $\beta > 0$
output: Refined depth map z
$\tilde{z}^{(0)} = \log z^{(0)}$, $k = 0$, $r^{(0)} = +\infty$;
while $r^{(k)} > 10^{-4}$ **do**

`// Photo-consistency optimization`

$$u^{(k+1)} = \underset{u}{\operatorname{argmin}}\ f(u) + \beta \left\| \log u - \tilde{z}^{(k)} \right\|^2 ; \tag{8}$$

`// Regularization of depth variations`

$$\theta^{(k+1)} = \underset{\theta}{\operatorname{argmin}}\ g(\theta) + \alpha^{(k)} \left\| \theta - \nabla \tilde{z}^{(k)} \right\|^2 ; \tag{9}$$

`// Integration`

$$\tilde{z}^{(k+1)} = \underset{\tilde{z}}{\operatorname{argmin}}\ \alpha^{(k)} \left\| \nabla \tilde{z} - \theta^{(k+1)} \right\|^2 + \beta \left\| \tilde{z} - \log u^{(k+1)} \right\|^2 ; \tag{10}$$

`// Auxiliary updates`

$\alpha^{(k+1)} = 1.5\,\alpha^{(k)}$; $z^{(k+1)} = \exp \tilde{z}^{(k+1)}$; $r^{(k)} = \frac{\left\| z^{(k+1)} - z^{(k)} \right\|}{\left\| z^{(k)} \right\|}$; $k = k + 1$;

end

Algorithm 1: Generic splitting strategy for multi-view stereo.

Lambertian image formation model. The fidelity term in (2) is derived from the common assumption that the brightness of a surface patch is invariant to changes in the viewing angle. In other terms, the surface is assumed to be Lambertian, and its reflectance is characterized by the albedo. Assuming a single point light source at infinity, the brightness I_i in the reference view at pixel i is then the product of albedo and shading:

$$I_i = a_i \, \max\left\{0, n_i^\top l\right\}, \tag{11}$$

with $a_i > 0$ the albedo at the 3D-point $\pi_z^{-1}(i)$ conjugate to pixel i, $n_i \in \mathbb{S}^2 \subset \mathbb{R}^3$ the unit-length surface normal at this 3D-point, and $l \in \mathbb{R}^3$ the lighting vector (in intensity and direction). The surface normal depends on the gradient of the log-depth map i.e., on θ, according to (see, for instance, [14]):

$$n_i := n(\theta_i) = \frac{1}{d(\theta_i)} \begin{bmatrix} \mathrm{f}\,\theta_i \\ -1 - [x, y]^\top \cdot \theta_i \end{bmatrix}, \tag{12}$$

where $\theta_i = \begin{bmatrix} \theta_i^1 \\ \theta_i^2 \end{bmatrix} \in \mathbb{R}^2$ denotes the depth gradient in pixel i (vector $\theta \in \mathbb{R}^{2p}$ introduced in (6) is thus the concatenation of all θ_i, $i \in \{1, \ldots, p\}$), $\mathrm{f} > 0$ is the focal length of the perspective camera, and $(x, y) \in \mathbb{R}^2$ are the centered coordinates of pixel i. The unit-length constraint on n_i is ensured thanks to the normalization by

$$d(\theta_i) = \sqrt{\mathrm{f}^2 \|\theta_i\|^2 + \left(1 + [x, y]^\top \cdot \theta_i\right)^2}. \tag{13}$$

Model (11) is valid for a single light source at infinity, which is rather unrealistic in practical scenarios. However, natural illumination can be represented as a

collection of infinitely-distant light sources, and the brightness at pixel i is then obtained by integrating the right-hand side of Equation (11) over the upper hemisphere. Approximating this integral using second-order spherical harmonics, one obtains (see [2] for details):

$$I_i = a_i \, \tilde{n}_i^\top \tilde{l}, \tag{14}$$

with $\tilde{l} \in \mathbb{R}^9$ a low-order lighting representation which can be calibrated beforehand using a reference object with known geometry and reflectance, and $\tilde{n}_i \in \mathbb{R}^9$ a "pseudo-normal" vector depending solely on the three components of $n_i = \left[n_i^1, n_i^2, n_i^3\right]^\top$ and thus, again, on θ_i:

$$\tilde{n}_i := \tilde{n}(\theta_i) = \begin{bmatrix} n_i \\ 1 \\ n_i^1 \, n_i^2 \\ n_i^1 \, n_i^3 \\ n_i^2 \, n_i^3 \\ \left(n_i^1\right)^2 - \left(n_i^2\right)^2 \\ 3\left(n_i^3\right)^2 - 1 \end{bmatrix} \overset{(12)}{=} \begin{bmatrix} \frac{\mathrm{f}\,\theta_i}{d(\theta_i)} \\ \frac{-1-[x,y]^\top \cdot \theta_i}{d(\theta_i)} \\ 1 \\ \frac{\mathrm{f}^2 \theta_i^1 \theta_i^2}{d(\theta_i)^2} \\ \frac{\mathrm{f}\theta_i^1\left(-1-[x,y]^\top \cdot \theta_i\right)}{d(\theta_i)^2} \\ \frac{\mathrm{f}\theta_i^2\left(-1-[x,y]^\top \cdot \theta_i\right)}{d(\theta_i)^2} \\ \frac{\mathrm{f}^2\left(\left(\theta_i^1\right)^2 - \left(\theta_i^2\right)^2\right)}{d(\theta_i)^2} \\ \frac{3\left(-1-[x,y]^\top \cdot \theta_i\right)^2}{d(\theta_i)^2} - 1 \end{bmatrix}. \tag{15}$$

In highly-textured scenes, the albedo values a_i in (14) strongly differ from one pixel i to another. As a consequence, so do the brightness values I_i and the feature vectors v_i, which makes the optimization of the fidelity term $f(z)$ in (2) meaningful, even in the absence of regularization. However, in textureless scenes the albedo is uniform, say equal to one:

$$a_i = 1 \quad \forall i \in \{1, \ldots, p\}, \tag{16}$$

and thus, according to (14), brightness variations are purely geometric i.e., due to variations in n_i. Such variations may be extremely subtle and thus unsuitable for use in a fidelity term such as (2), and regularization is required. Next we discuss two possible choices of regularizers.

Shading-aware regularization. Equation (14) can serve as a guide in multi-view stereo, in order to let the Lambertian image formation model disambiguate the matching problem in textureless areas. For instance, if we assume that the Lambertian model is satisfied up to a homoskedastic, zero-mean Gaussian noise, we can minimize the difference between both sides of Equation (14) in the sense of the quadratic loss, in the spirit of the variational approach to shape-from-shading under natural illumination introduced in [14]. This yields the following regularization function (recall that $a_i = 1$):

$$g_{\mathrm{Shading}}(\theta) = \lambda \sum_{i=1}^{p} \left(\tilde{n}(\theta_i)^\top \tilde{l} - I_i\right)^2, \tag{17}$$

with $\lambda > 0$ a tunable hyper-parameter. Using the regularization (17) in Algorithm 1 yields a solution to shading-aware multi-view stereo. Let us emphasize that $g_{\text{Shading}}(\theta)$ is smooth and separable (each term in the sum involves only $\theta_i = [\theta_i^1, \theta_i^2]^\top \in \mathbb{R}^2$), so (9) can be recast as a series of p two-dimensional non-linear problems which can be solved in parallel using, e.g., BFGS iterations.

Minimal-surface regularization. The latter regularizer requires knowledge of the lighting vector $\tilde{l}$. In some situations, calibrating lighting might be tedious or impossible, and one may prefer not to use an explicit image formation model. In such cases, it is possible to simply limit the surface variations, for instance by penalizing the total surface area. Following [14], this can be achieved by penalizing the ℓ^1-norm of the map d defined in Equation (13), which yields the following minimal-surface regularizer:

$$g_{\text{MS}}(\theta) = \mu \sum_{i=1}^{p} d(\theta_i). \tag{18}$$

Again, $g_{\text{MS}}(\theta)$ is smooth and separable, so (9) can be solved using parallelized BFGS iterations.

Combined regularization. Obviously, the minimal-surface regularizer (18) will tend to favor smooth surfaces and may miss thin structures. Conversely, the shading-aware regularizer (17) will tend to explain all thin brightness variations in terms of surface variations, which may be a source of noise misinterpretation. Therefore, it might be interesting to combine both shading-aware and minimal-surface regularizations, and the experiments in the next section are carried out using the following regularizer:

$$g(\theta) = \underbrace{\lambda \sum_{i=1}^{p} \left(\tilde{n}(\theta_i)^\top \tilde{l} - I_i \right)^2}_{g_{\text{Shading}}(\theta)} + \underbrace{\mu \sum_{i=1}^{p} d(\theta_i)}_{g_{\text{MS}}(\theta)}, \tag{19}$$

which remains smooth and separable, and yields the shading-aware solution if $\lambda > 0$ and $\mu = 0$, and the minimal-surface one if $\lambda = 0$ and $\mu > 0$.

5 Experimental results

In all our experiments, the feature vectors v_i are the concatenation of brightness values in a 3×3 neighborhood. Unless stated otherwise, the loss function ρ in (2) is the exponential-transformed SAD (with $\sigma = 0.2$). We first test our model on a synthetic dataset, using a renderer to generate images of size 540×540 of the well-known "Stanford's Bunny" with uniform albedo, knowing the lighting $\tilde{l}$ and the camera parameters. Gaussian noise with standard deviation equal to 1% of the maximum intensity is added, in order to get closer to real images.

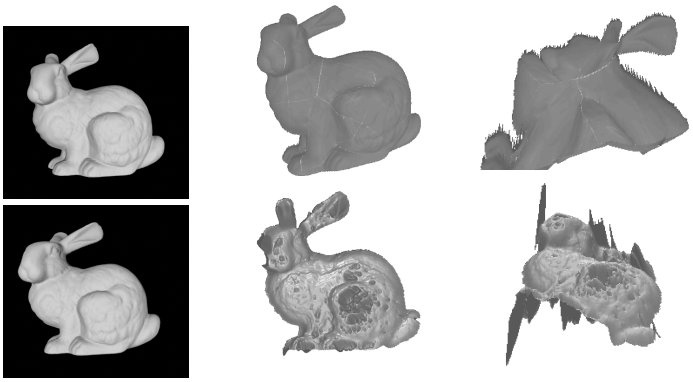

Fig. 2. Top row: in shape-from-shading, thin details in the input synthetic image (left) are finely recovered (center, the estimated depth is rendered frontally), yet the overall shape is biased due to the concave/convex ambiguity (right, rendering from another viewing angle). Bottom row: non-regularized multi-view stereo ($t = 1$), where the target synthetic image (left) is generated by translating the perspective camera. The overall shape is reasonable, yet thin details are missing and artifacts appear because photo-consistency degenerates in textureless areas (same viewing angles as above).

To preface, we highlight in Figure 2 the main issue of single-view shape-from-shading (i.e., $f(z) = 0$, $\lambda = 1$ and $\mu = 0$): though the reference image is well explained, the resulting depth map is obviously prone to the concave/convex ambiguity. Non-regularized multi-view stereo (i.e., $f(z) = (2)$, and $\lambda = \mu = 0$) is not satisfactory either: adding a second view (here, a simple translation of the perspective camera) and optimizing photo-consistency results in a noisy surface due to ambiguities in matching textureless patches.

As shown on the top row of Figure 3, results improve with the introduction of regularization. When they are not set to zero, the hyper-parameters are set to $\lambda = 5.10^{-4}$ and $\mu = 5.10^{-5}$ (those values were determined empirically). As expected, minimal-surface regularization allows to estimate a noise-free depth map which is globally reasonable, yet fine-scale details are missing. Using shading-aware regularization, fine-scale details are recovered, but a single target view ($t = 1$) is not enough to remove all concave/convex ambiguities. On the other hand, a joint approach gives satisfactory results, since the advantages of both regularization terms are combined.

Let us also remark that, since we did not explicitly take into account visibility issues, the estimated depth is not valid around parts which are not visible in the target image: see, for instance, the right edge of the bunny on the top row of Figure 3. To deal with visibility issues, we can simply increase the number t of target images ($t = 6$ in the example on the bottom row of Figure 3) so that each part in the reference view is covered in a few target pictures, the remaining occlusions being treated as outliers in the robust fidelity term $f(z)$.

This largely improves the results, as confirmed when evaluating the root mean squared error (RMSE, expressed in millimeters, knowing that the ground truth values stand within a 800-millimeter interval) with respect to ground truth. Let us remark that shading-aware regularization alone seems sufficient and minimal-surface regularization tends to smooth out fine-scale details.

This is confirmed by Figure 4: increasing the number t of target views removes all the concave/convex ambiguities of shape-from-shading, so minimal-surface regularization should be decreased, since it is not physics-based and tends to systematically flatten the surface.

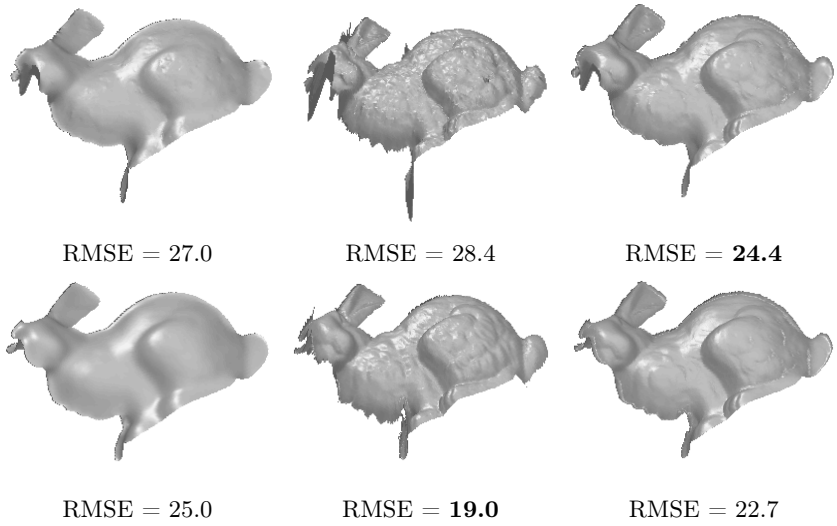

Fig. 3. Top row: Multi-view stereo with $t = 1$ target view (the two images are those of Figure 2). Bottom row: Multi-view stereo with $t = 6$ target views. From left to right: minimal-surface regularization (λ set to zero), shading-aware regularization (μ set to zero), and combined regularization ($\lambda > 0$ and $\mu > 0$). The errors due to occlusions on the first row have largely been reduced on the second row.

Finally, we put this work in real context, using the Augustus dataset from [17]. We used an existing photogrammetric pipeline [1] to estimate the camera parameters, as well as a rough depth map from which we could estimate lighting and the position of the initial plane (let us emphasize that this rough depth map was not used any further, e.g., as initial estimate). To demonstrate the ability of our framework to handle various photo-consistency measures, we show results obtained with the exponential-transformed SAD or ZNCC loss functions. From the results in Figure 5, the last reconstruction (bottom right), which uses exponential-transformed ZNCC and combined regularization, is the most satisfactory, at least from a qualitative point of view.

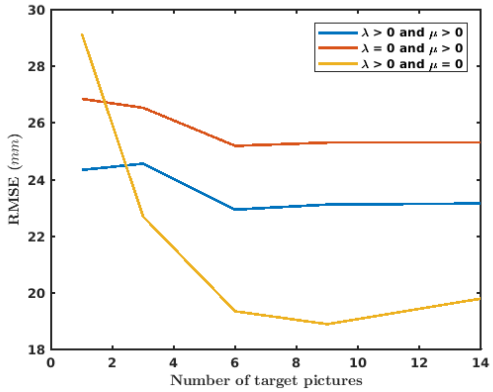

Fig. 4. RMSE for different regularization terms and for different numbers t of target images. If the combined approach gives better results with a low number of images, shading-aware regularization alone works better as soon as $t > 3$.

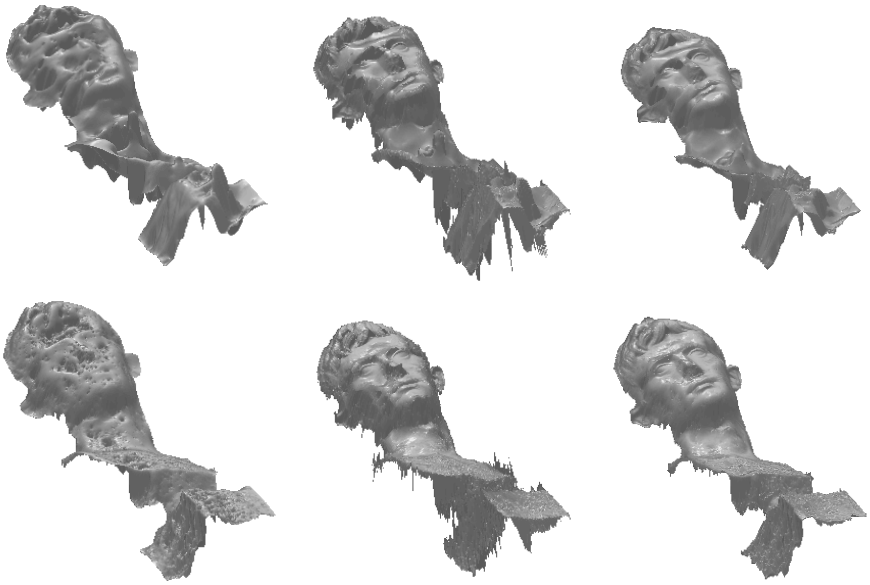

Fig. 5. Real-world multi-view stereo with $t = 6$ target views (two of them are shown in Figure 1), using SAD (top row) or ZNCC (bottom row) loss functions. From left to right: minimal-surface regularization ($\lambda = 0$, $\mu = 1.10^{-5}$), shading-aware regularization ($\lambda = 5.10^{-3}$, $\mu = 0$), and combined regularization ($\lambda = 5.10^{-3}$, $\mu = 1.10^{-5}$).

6 Conclusion and perspectives

We have introduced a generic splitting algorithm for multi-view stereo. It handles a broad class of photo-consistency measures and regularization terms, and is a suitable approach to 3D-reconstruction of textureless objects with few parameters to tune. Now, the proposed numerical scheme could be extended to the use of higher order regularization terms [15], to a joint estimation of depth, reflectance and lighting as in [12], and the whole approach could be turned into a volumetric one in order to recover a full 3D-model, as for instance in [11].

References

1. AliceVision. `https://github.com/alicevision/AliceVision`
2. Basri, R., Jacobs, D.P.: Lambertian reflectances and linear subspaces. PAMI 25(2), 218–233 (2003)
3. Blake, A., Zisserman, A., Knowles, G.: Surface descriptions from stereo and shading. IVC 3(4), 183–191 (1985)
4. Chambolle, A.: A uniqueness result in the theory of stereo vision: coupling shape from shading and binocular information allows unambiguous depth reconstruction. Annales de l'IHP - Analyse non linéaire 11(1), 1–16 (1994)
5. Furukawa, Y., Hernández, C.: Multi-view stereo: A tutorial. Foundations and Trends in Computer Graphics and Vision 9(1-2), 1–148 (2015)
6. Goesele, M., Curless, B., Seitz, S.M.: Multi-view stereo revisited. In: Proc. CVPR (2006)
7. Graber, G., Balzer, J., Soatto, S., Pock, T.: Efficient minimal-surface regularization of perspective depth maps in variational stereo. In: Proc. CVPR (2015)
8. Hernández, C., Schmitt, F.: Silhouette and stereo fusion for 3D object modeling. CVIU 96(3), 367–392 (2004)
9. Jin, H., Cremers, D., Wang, D., Yezzi, A., Prados, E., Soatto, S.: 3-D Reconstruction of Shaded Objects from Multiple Images Under Unknown Illumination. IJCV 76(3), 245–256 (2008)
10. Langguth, F., Sunkavalli, K., Hadap, S., Goesele, M.: Shading-aware Multi-view Stereo. In: Proc. ECCV (2016)
11. Maier, R., Kim, K., Cremers, D., Kautz, J., Nießner, M.: Intrinsic3D: High-Quality 3D Reconstruction by Joint Appearance and Geometry Optimization with Spatially-Varying Lighting. In: Proc. ICCV (2017)
12. Maurer, D., Ju, Y.C., Breuß, M., Bruhn, A.: Combining shape from shading and stereo: A joint variational method for estimating depth, illumination and albedo. IJCV 126(12), 1342–1366 (2018)
13. Newcombe, R.A., Lovegrove, S.J., Davison, A.J.: DTAM: Dense tracking and mapping in real-time. In: Proc. ICCV. pp. 2320–2327 (2011)
14. Quéau, Y., Mélou, J., Castan, F., Cremers, D., Durou, J.D.: A variational approach to shape-from-shading under natural illumination. In: Proc. EMMCVPR (2017)
15. Schroers, C., Hafner, D., Weickert, J.: Multiview Depth Parameterisation with Second Order Regularisation. In: Proc. SSVM (2015)
16. Wendel, A., Maurer, M., Graber, G., Pock, T., Bischof, H.: Dense reconstruction on-the-fly. In: Proc. CVPR (2012)
17. Zollhöfer, M., Dai, A., Innman, M., Wu, C., Stamminger, M., Theobalt, C., Nießner, M.: Shading-based refinement on volumetric signed distance functions. ACM Transactions on Graphics 34(4), 96:1–96:14 (2015)