Noname manuscript No.
(will be inserted by the editor)

A TV Prior for High-Quality Scalable Multi-View Stereo Reconstruction

Andreas Kuhn[1,2] · **Heiko Hirschmüller**[2,3] · **Daniel Scharstein**[4] · **Helmut Mayer**[1]

Received: date / Accepted: date

Abstract We present a scalable multi-view stereo method able to reconstruct accurate 3D models from hundreds of high-resolution input images. Local fusion of disparity maps obtained with semi-global matching (SGM) enables the reconstruction of large scenes that do not fit into main memory. Since disparity maps may vary widely in quality and resolution, careful modeling of the 3D errors is crucial. We derive a sound stereo error model based on disparity uncertainty, which can vary spatially from tenths to several pixels. We introduce a feature based on Total Variation (TV) that allows pixel-wise classification of disparities into different error classes. For each class, we learn a disparity error distribution from ground-truth data using Expectation Maximization (EM). We present a novel method for stochastic fusion of data with varying quality by adapting a multi-resolution volumetric fusion process that uses our error classes as a prior and models surface probabilities via an octree of voxels. Conflicts during surface extraction are resolved using visibility constraints and preference for voxels at higher resolutions. Experimental results on several challenging large-scale datasets demonstrate that our method yields improved performance both qualitatively and quantitatively.

Keywords Multi-View Stereo · 3D Modeling · Scalable 3D surface reconstruction

Andreas Kuhn
E-mail: andreas.kuhn@unibw.de

Heiko Hirschmüller
E-mail: heiko.hirschmueller@roboception.de

Daniel Scharstein
E-mail: schar@middlebury.edu

Helmut Mayer
E-mail: helmut.mayer@unibw.de

[1] Bundeswehr University Munich, Neubiberg Germany

[2] German Aerospace Centre, Munich, Germany

[2] Roboception GmbH, Munich, Germany

[2] Middlebury College, Middlebury, USA

1 Introduction

Constructing detailed geometric 3D models of the world is still a challenging and open problem for many applications of computer vision. Recent progress in structure from motion (SfM) and multi-view stereo (MVS) already allows for a fast reconstruction of surfaces from large image sets. Methods focusing on community photo collections of prominent tourist sites filter important subsets from a large amount of input images (Frahm et al., 2010).

Many existing algorithms can handle real-world datasets with tens of input images with minor variability, e.g., the datasets introduced by Strecha et al. (2008). Unfortunately, there are very few methods that are scalable in a way that allows reconstruction of full 3D models from hundreds or even thousands of high-resolution images with tens of megapixels without significant loss in detail. Furthermore, almost no existing methods can deal well with image configurations with a wide range of object distances, which for instance occur when combining images from unmanned aerial vehicles (UAVs) and from the ground.

3D modeling methods can be categorized into global and local methods based on the underlying optimization method.

Global methods tend to produce the best surface quality (Vu et al., 2012; Mücke et al., 2011) concerning completeness and accuracy as demonstrated for example on the Middlebury multi-view benchmark (Seitz et al., 2006). Local methods, on the other hand, yield better scalability (Fuhrmann and Goesele, 2011) and runtime performance (Newcombe et al., 2011; Steinbrücker et al., 2013). Even models of arbitrary size can be reconstructed in parallel without complex fusion (Kuhn et al., 2013).

In this paper we focus on local fusion of semi-globally optimized disparity maps, because they present a very good trade-off between runtime and the quality of the results. We show that 3D reconstruction can be improved by modeling the uncertainties of disparity maps. This paper describes the scalable 3D reconstruction method (Kuhn et al., 2013) with improvements by considering a variable disparity error (Kuhn et al., 2014).

The paper is organized as follows: Section 2 gives an overview of related work. Section 3 presents our reconstruction pipeline from registered image sets to 3D surface models. The potential of parallel processing for scalable 3D surface reconstruction is discussed in Section 4. Section 5 describes the derivation of disparity classes from a Total Variation (TV) based feature and its correlation with disparity uncertainty (Kuhn et al., 2014). The improvement obtained by using this TV prior for scalable volumetric MVS (Kuhn et al., 2013) is shown in Section 6. Finally, Section 7 presents experiments on a variety of popular and novel datasets and a comparison to state-of-the-art methods, and Section 8 concludes the paper.

2 Related Work

Surface reconstruction from depth maps has received considerable interest in recent years. Though we focus on methods based on local optimization, some of the global ones have to be mentioned as they give the best results. The idea of using Total Variation (TV) was introduced for MVS by Zach et al. (2007). They estimate the surface by minimizing a global energy function containing a TV-$L1$ regularization term for increased robustness to outliers, while still allowing efficient convex minimization. The use of TV regularization dates back to Rudin et al. (1992) for the reconstruction of noisy 2D images. It was improved for MVS by further work (Zach, 2008; Kolev et al., 2009; Schroers et al., 2012; Ochs et al., 2013). TV is important for our method, although we do not perform minimization via global convex optimization.

The drawback of global methods is that they are limited in practical applications due to poor scalability and runtime performance. Furthermore, many methods require an initial solution near the global optimum. The most promising local methods are volumetric and employ range image integration, as proposed in the seminal paper by Curless and Levoy (1996), to stereo images (Goesele et al., 2006). The basic idea is to extract an iso-surface from numerically occupied voxels whose values are estimated by the fusion of signed distance functions derived from the depth values. The resulting volumetric zero crossing defines the surface. Sagawa et al. (2005) adapt the approach by Curless and Levoy (Curless and Levoy, 1996) to dynamic voxel sizes depending on a consensus of multiple measurements (Wheeler et al., 1998). Fuhrmann and Goesele (2011) extend (Goesele et al., 2006) to varying surface qualities from MVS images.

We propose an alternative probabilistic distance function for multi-resolution voxels, with an additional filtering step that delivers good results in challenging configurations that lead to noisy spatial data (Kuhn et al., 2013). The filtering is based on free-space constraints, which was proposed by Merrell et al. (2007) and improved by further work (Bailer et al., 2012; Hu and Mordohai, 2012; Wei et al., 2014). In contrast to image-based consistency filtering, we propose probabilistic filtering in volumetric 3D space.

Probabilistic fusion on 3D volumes is also used by methods for occupancy grid propagation (Woodford and Vogiatzis, 2012; Pathak et al., 2007; Thrun, 2003). These methods propagate individual 3D points into single larger volumes and are therefore not suitable for high-quality 3D reconstruction. Nonetheless, the probabilistic framework has also been shown to be important for volumetric fusion (Hernández et al., 2007; Furukawa et al., 2007).

A great challenge for local methods is the use of a regularization term. Existing methods consider a varying error in 3D (Fuhrmann and Goesele, 2011; Mücke et al., 2011; Hu and Mordohai, 2012; Kuhn et al., 2013), but a constant error for 2D disparity. However, disparity quality can vary widely due to many reasons, including texture variability, motion blur, defocus, low-quality cameras, bad lighting conditions, compression artifacts, and priors employed by the stereo matching method. Our method analyzes the quality based solely on the disparity map, independent of camera type and configuration. For integration in MVS, we adapt the volumetric method (Curless and Levoy, 1996).

Since the above methods extract surfaces in 3D voxel space at different resolutions, they result in an ill-defined 4D regularization problem. To avoid this, we regularize according to the 2.5D disparity map. Our regularization term is derived from features describing the local oscillation of the disparities which influences the quality of the 3D points and hence the choice of the voxel size.

The correlation between this feature and the quality of the surface is learned from ground-truth data, in particular the 2014 Middlebury stereo datasets (Scharstein et al., 2014).

Learning priors for stereo and MVS is not new. Scharstein and Pal (2007) learn a conditional random field model for stereo from ground-truth disparities. Bao et al. (2013) learn priors for semantic categories in the form of the object shape; their method also utilizes information from the SfM process. Häne et al. (2013) demonstrate how learning semantic priors

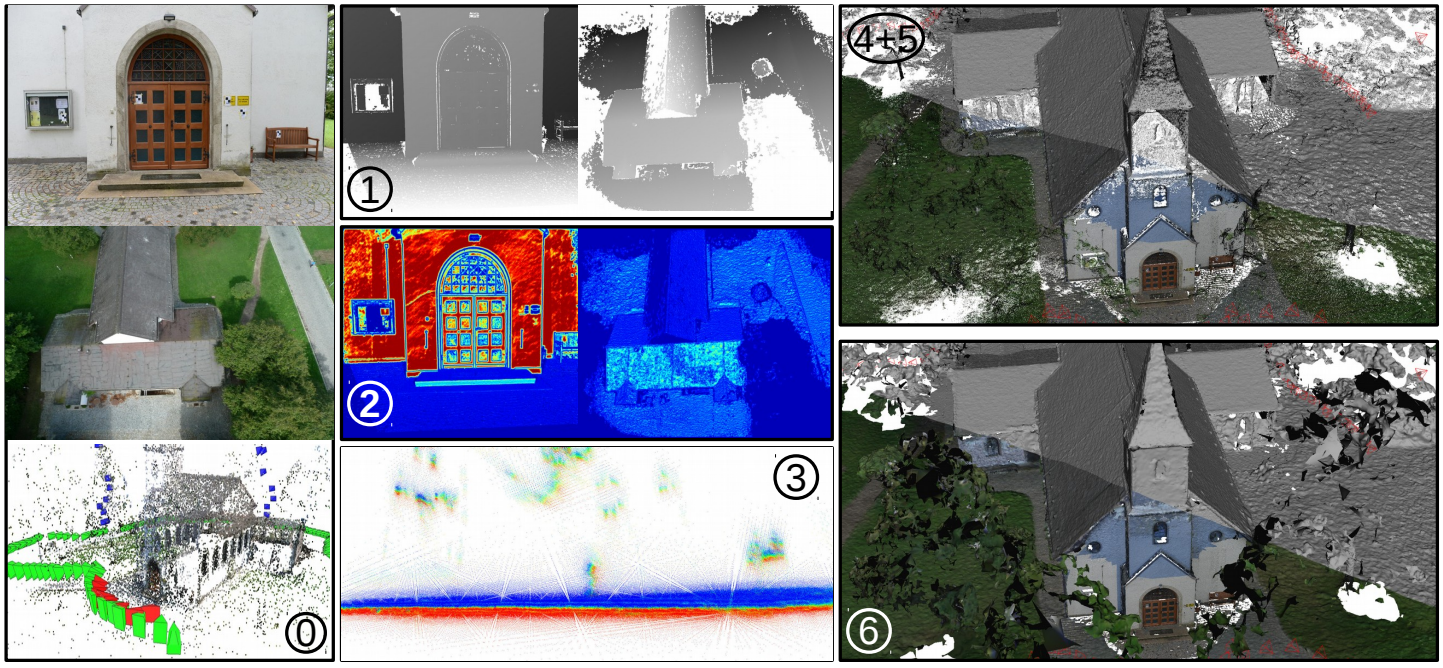

Fig. 1: Processing chain of the 3D reconstruction method described in this paper: (0) Image registration, (1) Stereo Matching using SGM, (2) Quality estimation for disparities, (3) Generation of a volumetric probabilistic space, (4) Point optimization considering the probabilistic space, (5) Filtering of outliers in the point cloud, (6) Triangulation of the point cloud. Fast parallel and hence scalable 3D modeling techniques in steps 2–5 are the main focus of this paper.

for classes such as buildings, ground, vegetation, and clutter can improve the surface quality. Their method is based on the global optimization approach of Zach (2008) and employs joint segmentation, labeling, and classification. In contrast, our new TV prior is only based on the input disparities without a need for semantic modeling.

Next, we give an overview of our 3D reconstruction pipeline followed by a detailed description of high quality improvements for scalable MVS.

3 Reconstruction Pipeline

After all input images are registered, the pipeline of our 3D reconstruction algorithm consists of the following steps (Fig. 1):

1. Estimation of disparity maps from stereo pairs by Semi Global Matching (SGM) (Hirschmüller, 2008; Hirschmüller and Scharstein, 2009).
2. Quality estimation for individual disparities.
3. Propagation of discrete 1D probability functions on the lines of sight into 3D space.
4. Optimization of points on the surface based on the probability function.
5. Filtering by visibility checks considering probabilistic information.
6. Local triangulation of the optimized point cloud.

For the estimation of disparity maps we use SGM, as it maintains small details due to pixelwise matching and has a low processing time for large images. Estimating the disparity quality in step 2 and expressing the disparity uncertainties as probabilistic functions is discussed with our geometric error model in Section 5.2. The propagation of the probability functions in step 3 as well as the optimization and probabilistic filtering in steps 4 and 5 are complex and described in detail in Sections 6.2 to 6.4. For step 6 we use a local triangulation building the final triangle mesh incrementally (Bodenmüller, 2009). Our pipeline allows fast parallel processing; steps 2–5 are the main focus of this paper.

4 Scalability and Parallel Processing

For fast reconstruction of large 3D models from high-resolution images, parallel computation is essential. To guarantee scalability for very large numbers of images it is even not possible to process all data at once. Steps 1 and 2 (Fig. 1) can be performed for all suitable image pairs separately. Hence, parallel processing is straightforward. The parallelization of the next steps is more challenging as the volumetric space has to be divided and the results need to be merged at the end. This allows processing on systems with limited main memory and offers scalability for very large scenes. It also makes the computation much faster if clusters with hundreds of cores are available.

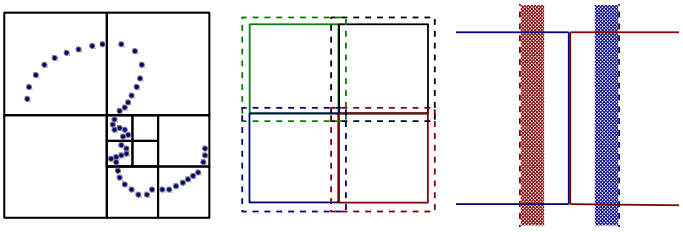

Fig. 2: The left image shows a line of points representing a 3D point cloud. Depending on the number of points, the reconstruction space is divided incrementally into four subspaces (eight in 3D). Neighboring subspaces are overlapping as shown in the center. The right image illustrates that by setting an overlap, the border of neighboring voxels consists of similar 3D information. The parts in the shaded red and blue areas are discarded after meshing.

For space division, the algorithm runs in a preprocessing step through all depth maps and divides the overall volume into subvolumes whose sizes depend on model resolution, memory size and number of cameras a point was captured from. To handle point cloud areas with varying density which usually appear in complex stereo image configurations, a simple incremental algorithm divides the specific volume in eight subvolumes if the number is beyond a threshold (Fig. 2).

In divide-and-conquer strategies, especially the merging can be highly complex. For merging the subvolumes, the overlap of neighboring subvolumes has to be large enough so that meshes are equivalent in the merged volumes. More precisely, the overlap has to be at least twice the local neighborhood used for meshing. This constraint allows for a very easy fusion as the meshes are equivalent in the inner half of the overlapping area. 3D points or respectively triangles in the outer half are simply not considered (Fig. 2). For a more detailed description of fast space division for scalable 3D reconstruction see Kuhn and Mayer (2015).

5 Disparity Quality Modeling

In this section we first review a popular stereo error model for MVS, and then discuss how disparity quality classes can be learned from ground-truth data. In general, the 3D error strongly depends on the disparity quality which is not constant and varies spatially from tenths of pixels to several pixels. To handle this variation, we discuss strong influences on the quality of disparities, and propose novel TV-based feature classes for disparities covering a wide range of these influences. Additionally, we show how to learn the disparity uncertainty from ground-truth disparity maps in comparison with disparity maps generated by SGM for individual feature classes using an Expectation Maximization (EM) approach.

5.1 Stereo Error Model

Uncertainties of the disparities result in 3D reconstruction errors. Here we employ the compact error model proposed by Molton and Brady (2000). This ellipsoidal error model propagates the disparity error Δp in image space into the error in three space dimensions ΔP_x, ΔP_y and ΔP_z (Fig. 3).

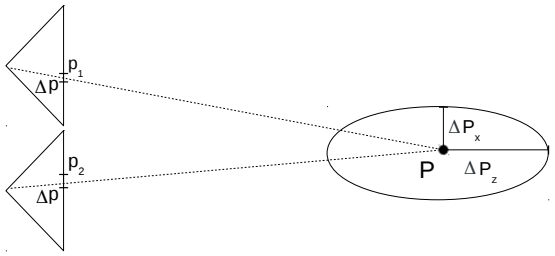

Fig. 3: A pair of parallel cameras. The disparity uncertainty Δp in the image leads to an ellipsoidal uncertainty ΔP of the 3D point.

The error in x and y direction of the camera coordinate system rises linearly with P_z:

$$\Delta P_x = \Delta p \frac{P_z}{ft} \sqrt{(t - P_x)^2 + P_x^2} \ , \tag{1}$$

$$\Delta P_y = \Delta p \frac{P_z}{ft} \sqrt{2P_y^2 + \frac{t^2}{2}} \ , \tag{2}$$

while the error in z direction rises quadratically with P_z:

$$\Delta P_z = \Delta p \frac{P_z^2}{ft} \sqrt{2} \ . \tag{3}$$

Focal length f and baseline t are known for registered image sets. Coordinates P_x, P_y, P_z follow from the estimated disparities. Yet, the disparity error Δp is not known and not constant. It follows an unknown function ranging from subpixel to several pixels. For local MVS, which cannot regularize such uncertainties globally, it is important to consider this uncertainty function.

5.2 Uncertainties in Stereo Matching

Learning quality for disparities is difficult since the quality of disparities is affected by many factors. It usually depends strongly on influences such as texture strength and surface slant (Fig. 4). Slanted surfaces are problematic because common priors employed by most stereo methods, including SGM, favor constant disparities and thus introduce a fronto-parallel bias. Both local and global stereo methods tend to propagate disparities from textured into textureless regions, which can lead to errors on slanted and curved surfaces. This needs to be considered during the fusion of depth maps.

Efficient stereo methods generally obtain subpixel accuracy indirectly by interpolation of neighboring cost values. For instance, SGM estimates subpixel disparities by fitting a parabola through the three costs values centered at the

Fig. 4: Varying disparity quality. Left: Zoomed region of an input image of the Ettlingen30 sequence (Strecha et al., 2008). Right: Surface orientation derived from the disparity map computed by SGM visualized using a linear coding from 0°(light) to 90°(dark). The surface orientation gives a good impression of the reconstruction quality. Accuracy is lower in slanted and untextured regions (left and center red box), and higher in textured fronto-parallel regions (right green box).

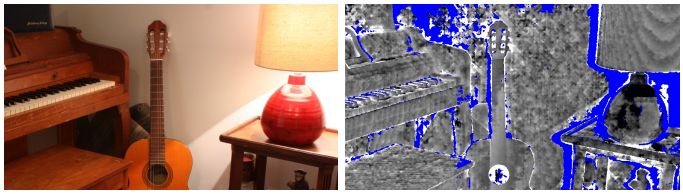

Fig. 5: Disparity oscillations. Left: Half-resolution Middlebury Piano image (Scharstein et al., 2014). Right: Signed disparity error of SGM w.r.t. ground truth, coded from -1 (white) to 1 (black). Missing values are in blue. The error exhibits oscillations with varying frequency and amplitude depending on the surface slant and the amount of texture present, as can be observed particularly well on the lamp shade in the top right.

winning disparity. Depending on the geometry, this leads to varying uncertainties in subpixel precision.

Unfortunately, there are several additional influences on the accuracy. MVS is often used for complex scenes based on registration information from SfM methods. Depending on scene geometry and texture, the bundle adjustment error can range from a fraction of a pixel to several pixels. Images from mobile phones are increasingly used in computer vision due to their widespread availability. It is well known that the quality of images from small chips and lenses is limited. Even high-quality cameras have a limited depth of field and are subject to motion blur. It is, therefore, of prime importance to consider different qualities also for disparities.

Naively learning all these influences would result in a multivariate system where the learning space is defined by features covering all uncertainties. Two of these features would be texture strength and surface slant. The corresponding multivariate uncertainty would have to be learned for all camera types and perhaps even all types of scenes. As this is not possible in a generic way, and it is very expensive to generate ground truth, we focus on estimating the uncertainty from the disparity map directly. For this, we propose a feature covering important aspects of the uncertainty, particularly those caused by slant and texture, by analyzing the local oscillation behavior of the disparity map.

5.3 Classification of Error Levels Based on TV Features

The key question is how disparity uncertainty can be classified. A particular problem is that our disparity maps show oscillations with unknown frequency (Fig. 5). A window that is too large could oversmooth them, but one could also obtain wrong measurements by undersampling. In Fig. 4 it can be seen that some normal vectors have large orientation errors in weakly-textured or slanted regions. In addition, learning the distribution of a 2D function is difficult since it can lead to the estimation of wrong correlations.

To avoid these problems, we introduce feature classes based on Total Variation (TV) for estimating the local oscillation behavior. These classes represent the disparity quality in a stable way and can be learned from ground truth directly.

MVS methods typically use TV in combination with the $L1$ norm for the estimation of a globally optimal surface from point clouds with a limited influence of 3D outliers. In contrast, we use the $L2$ norm since we are interested in measuring the quality of the disparities, which includes both noisy measurements and outliers. We focus on local optimization and employ TV on 2D disparity maps instead of spatial surfaces. We can thus use the original formulation for 2D signals to express the TV of disparities d over a neighborhood $\mathcal{N}_y$ for a pixel y:

$$TV(y) = \sum_{i,j \in \mathcal{N}_y} \sqrt{|d_{i+1,j} - d_{i,j}|^2 + |d_{i,j+1} - d_{i,j}|^2} \; . \tag{4}$$

$TV(y)$ represents the degree of the local oscillation in a certain neighborhood of pixel y. Unfortunately, oscillations in local neighborhoods have different frequencies. In particular, fronto-parallel planes cause low frequencies, while sloping planes lead to high frequencies (Fig. 5). Hence, it is not feasible to set a constant window for TV estimation.

In addition, we need to discretize the TV term so that we can learn the disparity variance from ground truth for each level. A reasonable way to limit the discretization levels is to compute the TV over square windows with increasing radius m while requiring the TV to stay below a threshold τ. This can be written as:

$$\arg\max_n \Big(\sum_{m=1}^{n} \frac{1}{8m} TV_{i,j \in x_m} < \tau \Big) \; , \tag{5}$$

where x_m describes a series of concentric square "ring-shaped" neighborhoods with radius m and $|x_m| = 8m$. That is, in the first step the TV term is calculated for the eight neighboring pixels. If the value exceeds the threshold, the discretized value $n = 1$ defines the TV class. If it does not exceed the threshold, TV is calculated considering the next 16 $(8m, m = 2)$ pixels, until a maximum of $n = 20$. This can be done in linear time. For pixels with missing disparities, a value of ∞ is used. The number of pixels considered

Fig. 6: Visualization of computed TV classes for each pixel. Top: Input images. Middle: Disparity surface orientation (as in Fig. 4) providing an impression of the local reconstruction quality. Bottom: TV classes measuring disparity smoothness, ranging from 1 (blue) to 20 (red). Fronto-parallel textured surfaces lead to higher class numbers.

for level m rises with $8m$. Hence, the sum of TV increases with the size of the level. This can be accounted for by a division of the sum by $8m$. In our experiments this regularization leads to better results. We use a threshold of $\tau = 1$ for all experiments. This limits the average oscillation of the pixels in the neighborhood of pixels considered in step m to a maximum of one disparity. Fig. 6 shows examples of the computed feature classes.

5.4 Learning Error Distributions

For the learning we relate the estimation of the uncertainty of the disparity to the TV classes $n = [1, 20]$ introduced in the previous section.

We assume that the error for each class follows a combination of a Gaussian $\mathcal{N}_n(\mu_n, \sigma_n)$ with parameters $\theta_n = \mu_n, \sigma_n$ modeling the disparities, and a uniform distribution $\mathcal{U}$ representing outliers. In the stereo case this mixture is considered a good approximation for the error distribution (Vogiatzis and Hernández, 2011).

We learn our priors using the 2014 Middlebury stereo datasets with accurate ground truth (Scharstein et al., 2014). We employ half-resolution versions of the seven images used in Sinha et al. (2014) for which public floating-point ground-truth disparities are available.

After generating the disparity maps, we calculate the TV class for all pixels of the SGM result with a valid disparity. The Gaussian is estimated for all classes $0 < n \leq 20$ by an Expectation Maximization (EM) method $\arg\max_{\theta_n} p(\theta_n|\mathcal{D}_n)$. The data $\mathcal{D}_n$ describes the set of measured differences between the ground truth and the value based on the SGM results, assigned to class n.

We use EM instead of Maximum Likelihood (ML) estimation, because we consider mixture functions. It is well known that for EM learning a good initial estimation is required. We found that by a ML estimation suitable initial functions can be obtained. Expected value and variance are obtained by ML as:

$$\mu = \frac{1}{n}\sum_{i=1}^{n}(d_i - g_i) \ , \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(d_i - g_i - \mu)^2 \ , \tag{6}$$

with disparity d and ground truth g for n measurements.

These functions are used as initial state for the EM. For the estimation of the outlier probability we count measurements that lie outside the area of five σ. The ratio of the number of outliers and the number of measurements defines the outlier probability and is used for the uniform function of the mixture. In the E step the measurements are assigned to the Gaussian or the uniform function depending on their probability. In the M step ML estimation is used again for estimation of the Gaussian parameters, e.g., for μ:

$$\mu_{EM} = \frac{1}{||N||}\sum_{k \in N}(d_k - g_k) \ , \ N = \{k|\mathcal{N}(d_k) > \mathcal{U}(d_k)\} \ . \tag{7}$$

Afterwards the outlier probability can be obtained as described above. We found that a single EM step yields good results while multiple EM steps lead to an unreasonable shrinking of the variance. This lack of convergence is caused by the fact that disparity noise and error cannot be clearly separated as multiple disparities are implicitly derived by interpolation.

The resulting expected values and standard deviations for the 20 classes are given in Table 1. As expected, the standard deviations for the low-numbered classes (which represent large oscillations) are high, but decrease quickly for the higher classes. Interestingly, it appears that there is also a positive disparity offset of up to one pixel in the low-numbered classes. The likely reason are foreshortening effects, since the TV also measures surface slant. For an analysis of systematic errors in stereo estimation, which is beyond the scope of this paper, we refer to Xiong and Matthies (1997).

6 Volumetric Modeling

As described in Sec. 5 we represent depth as a random variable following a set of Gaussians. In this section we describe

TV class	1	2	3	4	5	6	7
μ_{EM}	0.98	0.48	0.11	0.04	0.03	0.03	0
σ_{EM}	4.44	3.11	1.65	1.07	0.67	0.50	0.40
TV class	8	9	10	11	12	13	14
μ_{EM}	-0.03	-0.03	-0.03	-0.03	-0.03	-0.02	-0.02
σ_{EM}	0.33	0.34	0.34	0.30	0.28	0.26	0.24
TV class	15	16	17	18	19	20	
μ_{EM}	-0.02	-0.01	0	0.01	0.01	-0.01	
σ_{EM}	0.22	0.22	0.21	0.20	0.19	0.18	

Table 1: Learned expected value and standard deviations in pixels for the 20 TV classes.

how this can be used for volumetric MVS. The set of uniform functions is disregarded at the volumetric fusion step, because we want to perform fusion of noisy measurements. The filtering of outliers will be considered afterwards. Because it comprises the most important part, at the moment we only consider the error in the direction of the line of sight. To this end, we use the dominant error in z-direction from the error model derived via error propagation, which depends on four geometric parameters (Eq. (3))

While focal length f, baseline t and distance P_z are constant for registered images with corresponding disparity maps, the error Δp follows a Gaussian derived in Sec. 5.4. Hence, we consider a Gaussian $\mathcal{N}(\mu, \sigma)$ with expected value $\mu = d_i + \mu_{EM}$, where d_i is the depth at coordinate i derived by SGM and μ_{EM} the offset from Table 1, and standard deviation $\sigma = \Delta P_z = \Delta p \frac{P_z^2}{ft}\sqrt{2}$ with $\Delta p = \sigma_{EM}$ from Table 1.

In summary, the function for the expected noise of a depth value can be expressed by

$$p(d_i^x) = \mathcal{N}(P_z, (\sigma_{EM} \frac{P_z^2}{ft}\sqrt{2})^2) \ , \tag{8}$$

with $P_z = ft/(d_i + \mu_{EM})$.

We extend multi-scale volumetric fusion based on octree data structures, as formerly proposed by Sagawa et al. (2005); Fuhrmann and Goesele (2011). To this end, in addition to the choice of the voxel size (Sec. 6.1), we show below how to directly propagate probability functions into volumetric space (Sec. 6.2), and how the probabilities can be used for optimization and filtering in 3D space (Secs. 6.3 and 6.4).

6.1 Choice of Voxel Size

An important step to efficiently handle disparity maps of varying density is the choice of the voxel size v_s in the octree. In our case the octree cubes correspond to the voxels. Because the fusion is only reasonable for related data, the algorithm chooses the voxel size for all disparities individually. The idea is that data are fused (Section 6.2) with others having at minimum half and at maximum double the quality.

This accounts for the fact that the voxel size in an octree increases by a factor of 2. Hence, the voxel size v_s is chosen such that $\sigma < a v_s < 2\sigma$, where a is a smoothness term.

For practical applications we found $a \in [4, 8]$ to be suitable to maintain the details, but to avoid "pitted" surfaces. In our experiments we use $a = 8$ in order to maintain the smallest details, while $a = 4$ would speed up the processing by about a factor of four.

6.2 Propagation into Probabilistic Space

After choosing an octree depth, data fusion can be regarded as volumetric fusion on a regular grid as proposed by Curless and Levoy (1996). In their formulation a linear signed distance function assigns negative values to voxels in front of the estimated depth, and positive values to those behind (Fig. 7a). The voxels to be assigned a surface value are chosen by intersection of the octree with a part of the line of sight. Instead of using a linear signed distance function, we directly propagate a Gaussian cumulative distribution function (CDF) (Fig. 7b), which directly transforms the values from Equation (8) into logarithmic ratio space as explained below. Additionally, a second function defines the weight w of this value (Fig. 7).

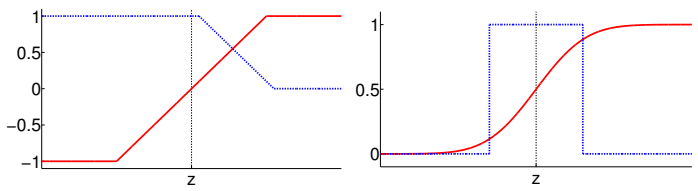

Fig. 7: Two alternative cumulative distance functions for an estimated depth z. (a) Linear signed distance function d (red) with a weighting function w (blue) penalizing values behind the estimated surface (Curless and Levoy, 1996; Fuhrmann and Goesele, 2011). (b) Gaussian CDF (red) with indicator function (blue) bounding the area of influence (Kuhn et al., 2013).

The volumetric update process for voxels on the line of sight accumulates the contributions from individual pixels from individual disparity maps. For the linear function it follows the two equations (Curless and Levoy, 1996):

$$W_{i+1}(v) = W_i(v) + w_{i+1}(v) \ , \tag{9}$$

$$D_{i+1}(v) = \frac{W_i(v)D_i(v) + w_{i+1}(v)d_{i+1}(v)}{W_{i+1}(v)} \ , \tag{10}$$

where $d_i(v)$ is the discretized value of the cumulative signed distance function and $D_i(v)$ characterizes the current discrete representation of voxel v at iteration i. This value has a range of $[-1, 1]$ and is propagated using Equations (9) and (10). The individual discretized weight functions $w_i(v)$ are accumulated in $W_i(v)$. The weight function reduces the weight of depth values behind the measured distance. This is reasonable because the voting is only meaningful on the line of sight in front of the point.

Curless and Levoy (1996) empirically adapt the linear weighting function depending on the angle between line of sight and the normal vector (i.e., slant) of the surface, as well as on the distance to the next missing measurements. We instead use the novel TV-based probabilistic function that implicitly handles slant and missing measurements and is determined statistically from ground-truth data.

Instead of a linear signed distance function, we use the probabilistic functions (8). It has been shown that the original linear formulation optimizes measurements with Gaussian noise in a least-squares sense. We argue that a least-squares optimization is sensitive to outliers. We, therefore, propose the direct propagation of probability function which can be used for probabilistic filtering. A detailed derivation of the functions is given in Kuhn (2014).

Furthermore, instead of the weighting function w from the linear case, we use in the probabilistic formulation a "boxcar" indicator function with a width of $\pm 2\sigma$ (Fig. 7b). This limits the influence of voxels to a narrow region around each estimated depth, and therefore generally yields better results due to an increased robustness to outliers. Additionally, it significantly decreases the number of voxels to process which allows for limited memory resources. On the other hand, it can lead to multiple estimated surfaces for one real surface, and thus requires post-processing in the form of filtering. However, multiple surfaces are possible even without bounding the influence with an indicator function, as disparity maps are generally incomplete.

When estimating the probability that voxel v_i lies behind the detected surface along the line of sight (Fig. 8), we use $p(v_i^0)$ and $p(v_i^1)$ for the probability that a voxel lies in front or behind the surface, respectively.

As the probability $p(v_i^1)$ is the integral of the Gaussian from $-\infty$ to the distance a_i of the camera center to the intercept point of the line of sight and the voxel v_i, one can take the Gaussian CDF instead of the Probability Density Function (PDF) to estimate it immediately:

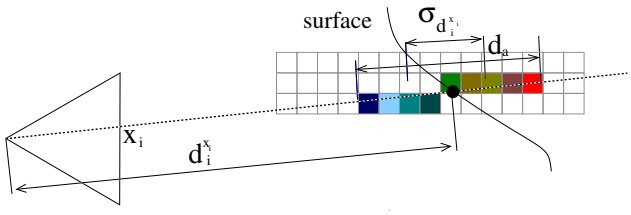

Fig. 8: Discrete probability of a surface – Point with pixel coordinate x_i and expected distance $d_i^{x_i}$. $\sigma_{d_i^{x_i}}$ is the standard deviation of the 3D point position along the line of sight. The colored voxels represent the probability that a voxel lies behind the surface from blue (low) to red (high).

$$p(v_i^1) = \int_{-\infty}^{a_i} \mathcal{N}_{pdf}(x)dx = \mathcal{N}_{cdf}(a_i) \tag{11}$$

The Gaussian CDF is numerically estimated with the Gauss error function, available for instance in the C++ standard library.

Probabilities on different rays from the same image that fall in the same voxel are averaged. Similar to other occupancy grid approaches, we employ a Bayesian statistical framework for the fusion of data from different images.

Since the probabilistic state of the surface is binary, i.e., a voxel is either occupied or not, we use the Binary Bayes Filter for probabilistic fusion. In addition to MVS reconstruction (Kuhn et al., 2013), fusion of sensor data via Binary Bayes Theory has also been applied for occupancy grid propagation (Woodford and Vogiatzis, 2012; Pathak et al., 2007; Thrun, 2003).

Each occupied voxel has a probability $p(v^1)$ of lying completely behind the surface, and conversely a probability $p(v^0) = 1 - p(v^1)$ of lying at least partially in front of the surface. These probabilities in the range $[0, 1]$ are transformed into logarithmic ratio space, with values in the range $[-\infty, \infty]$, and fused via summation (Kuhn et al., 2013):

$$l_i = log\frac{p(v_i^1)}{p(v_i^0)} = log\frac{p(v_i^1)}{1 - p(v_i^1)} = \sum_j log\frac{p(v_{ij}^1)}{1 - p(v_{ij}^1)}. \tag{12}$$

Like the linear formulation (10), this function can be reformulated as an incremental update process:

$$l_{i+1} = l_i + log\frac{p(v_{i+1}^1)}{1 - p(v_{i+1}^1)}, \ l_0 = 0, \tag{13}$$

which allows for sequential processing of the images. Hence, at no point in the processing more than one disparity map has to be held in memory. The initial probability l_0 derives from the assumption that the prior probability of a voxel to be occupied is 0.5.

6.3 Optimization of Point Positions on the Surface

The surface is characterized by neighboring voxels for which the probability that one is in front of the surface and the other is behind the surface is maximum. To achieve a better accuracy than the estimated distances d_x^i, we make use of the probabilistic voxel grid described above.

In the original formulation from Curless and Levoy (1996) surfaces are directly derived from volumetric space using the marching cubes (MC) algorithm. While standard MC is only applicable for regular grids (Fuhrmann and Goesele, 2011), there are extensions for a generation of watertight surfaces for irregular grids (Kazhdan et al., 2007). To this end, the probabilistic space has to be transformed into a volumetric point cloud space (Fuhrmann and Goesele, 2011). Unfortunately, the watertight constraint does not comply with the space division approach presented in Sec. 4. Furthermore, watertight surfaces are not guaranteed to represent the real world reconstructed from a limited number of camera positions. We therefore transform the probabilistic voxel grid into a 3D point cloud and use an incremental local meshing method (Bodenmüller, 2009) whose processing area depends on the voxel size. In our case the points are connected with other points in an area which is five times the voxel size.

Initially we define a search interval by shifting the estimated point along the line of sight by twice the standard deviation in both directions. This is advantageous since multiple surfaces may exist in this interval and the use of the viewing direction allows an extraction with subvoxel accuracy. We consider all voxels I in the interval and take the adjacent pair of voxels maximizing the probability that one is in front and the other is behind the surface:

$$i^* = \arg\max_{i \in I} (p(v_i^0) p(v_{i+1}^1)) \ . \tag{14}$$

To obtain the position d^* with subvoxel accuracy, we fit a Gaussian to the neighboring voxels of v_{i^*} and v_{i^*+1}. For this we use a probabilistic weighting of the distances:

$$d^* = \frac{1}{4} \sum_{j=i^*-1}^{i^*+2} d_j \ |0.5 - p_j| \ , \ p_j = \frac{e^{l_j}}{1 + e^{l_j}} \ . \tag{15}$$

The Gaussian CDF at μ is 0.5, which is not its maximum value. Hence, we weigh each term by the absolute difference $|0.5 - p|$. A regression of the Gaussian CDF would also be possible by solving a system of equations of the linear parts. We found that the results differ only marginally and the computation by Equation (15) is faster.

6.4 Filtering Voxels using Visibility Checks

Octrees allow for efficient local consistency checking based on ray tracing. Even though ray tracing is a (semi-)global method, we use it locally because outliers have a detrimental influence on the surface quality particularly if they occur nearby. We do not deal with solitary outliers, because they are ignored in the later step of mesh generation.

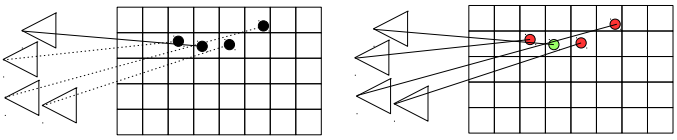

Fig. 9: Geometric consistency checks that consider surface probabilities. Left: The second point from the left is cast to the upper camera, resulting in a conflict with the leftmost point. It is not obvious which point has to be filtered. Right: After considering multiple points with surface probabilities, which are all in conflict, only the point with the highest surface probability is maintained.

For filtering, we cast a ray along the line of sight for ten times the voxel size. Local ray tracing filters conflicting points having a relatively lower quality occluding others with a better quality (Fig. 9). In our case the quality is described by the maximum probability $p(v_i^0)p(v_{i+1}^1)$ of the voxel surface estimated by Equation (14). In case of a conflict, only the point with the higher surface probability is maintained.

For all voxels v classified as occupied above, consistency is checked. Rays are cast to those cameras the voxel has been seen from. We found that even one camera consistency check is usually enough. If there is another occupied voxel on the ray, a conflict is detected. For all voxels the maximum quality of conflicting voxels i on all rays is saved. Those voxels are filtered whose qualities are worse than the maximum quality for all conflicting voxels.

If a conflict occurs for voxels on different octree levels, the voxels on the lower resolutions are filtered out immediately. This allows the processing of configurations with a wide range of distances. The substitution of lower-resolution voxels by higher-resolution voxels is reasonable in our processing pipeline because our TV prior leads to more stable points at higher resolution levels. Hence, the proposed TV prior acts as a regularization term.

Of course it is not always necessarily true that a lower-resolution point has lower surface probability than a higher-resolution point. This results in an ill-defined 4D regularization problem. Fuhrmann and Goesele (2011) proposed heuristic averaging on neighboring levels. In our case this did not lead to an improvement, and the implicit regularization via our TV prior allows a good estimate of the chosen octree level.

7 Experiments and Evaluation

In this section we show qualitative and quantitative results on established test data as well as novel real-world datasets. In addition to demonstrating scalability we also evaluate surface quality, with particular focus on the improvement due to the TV prior.

Fig. 10: Left: four sample images of a registered image set of 822 high-resolution images. Right: The complete textured reconstructed model, as well as zoomed views rendered without texture to visualize the preservation of small details. Despite the space division strategy, the 3D model has seamless and consistent surfaces.

We first demonstrate the potential for reconstructing 3D surfaces from large sets of high-resolution images as this is one important benefit of our method. The space division strategy described in Sec. 4 allows for fast parallel reconstruction of potentially arbitrarily large scenes. This is demonstrated in Fig. 10 with a large dataset consisting of 822 images of an entire village captured both from a UAV and from the ground. Because of this complex configuration of viewpoints, the distances, and thus also the quality of the reconstructed surface parts, varies greatly. This is handled by means of the octree representation. The division of 3D space allows for parallel processing of thousands of subspaces in a couple of hours on a cluster system. The local processing constraint leads to consistent surfaces without requiring a complex fusion strategy.

To demonstrate the improvement due to our novel TV prior, we compare surfaces reconstructed using constant standard deviations $\sigma \in \{0.5, 1, 2, 4\}$ and surfaces reconstructed using our variable TV-based standard deviation $\sigma = \sigma_{EM}$ (Eq. (8)). Fig. 11 shows the results on the Ettlingen30 dataset (Strecha et al., 2008), which is a good test case since the images have varying perspectives and varying amounts of texture. There are two types of difficulties affecting the disparity quality: Lack of texture because of the white walls, and slanted surfaces, which produce uncertainties because of SGM's fronto-parallel bias.

In all cases the TV prior yields results that match or exceed the best results obtainable with constant σ in terms of accuracy and completeness. In general, the surfaces produced with $\sigma = 4$ tend to be best in terms of completeness (fewest holes) but are overly smooth, while the surfaces for $\sigma = 0.5$ appear best in terms of accuracy (most detail captured), but have many holes. Using a variable TV-based prior ($\sigma = TV$) combines the best of both worlds.

For an objective assessment of an approach, a numerical evaluation on established datasets is highly important. Strecha et al. (2008) provided a numerical evaluation on real-world datasets, which unfortunately is no longer available. Fortunately, for two datasets the ground truth is publicly available. However, for the special configurations of these datasets with mostly fronto-parallel surfaces and highly textured areas only a small increase in quality can be expected.

Fig. 12 shows the 3D surface models derived from SGM disparity maps by the method of Fuhrmann and Goesele (2011) and by our TV-based method. Our method does particularly well in recovering slanted surfaces. Fuhrmann and Goesele use optimized disparity maps from a community photo collection approach (Goesele et al., 2007). For the comparison we employed their fusion method on SGM disparity maps.

For the numerical evaluation we used the technique proposed by Strecha et al. (2008). In the original evaluation the errors are compared against the ground-truth uncertainty. As this information is not available, we employed the absolute error (Fig. 13).

For the Herzjesu8 dataset the TV-based surface model is best in terms of accuracy and completeness. In comparison with the method by Fuhrmann and Goesele (2011) the evaluation of the TV-based model presents a significant improvement. For the EttlingenFountain dataset the TV based surface is slightly less accurate than the surface based on a constant uncertainty. However, it is expected that the evaluation results would differ in the high-resolution regions if the ground-truth uncertainty could have been considered.

We also compared our TV-based method with constant-σ versions using the Middlebury multi-view benchmark (Seitz et al., 2006). The datasets are not suitable to demonstrate the strength of our method, as the objects do not have variable texture and difficult perspectives. Still, the numerical results (Table 2) confirm the qualitative impression from the previous experiments that the TV results are best concerning accuracy and are mostly best concerning completeness. In comparison with $\sigma \in \{0.5,1,2,4\}$, the TV results are always close to the best individual accuracy and completeness

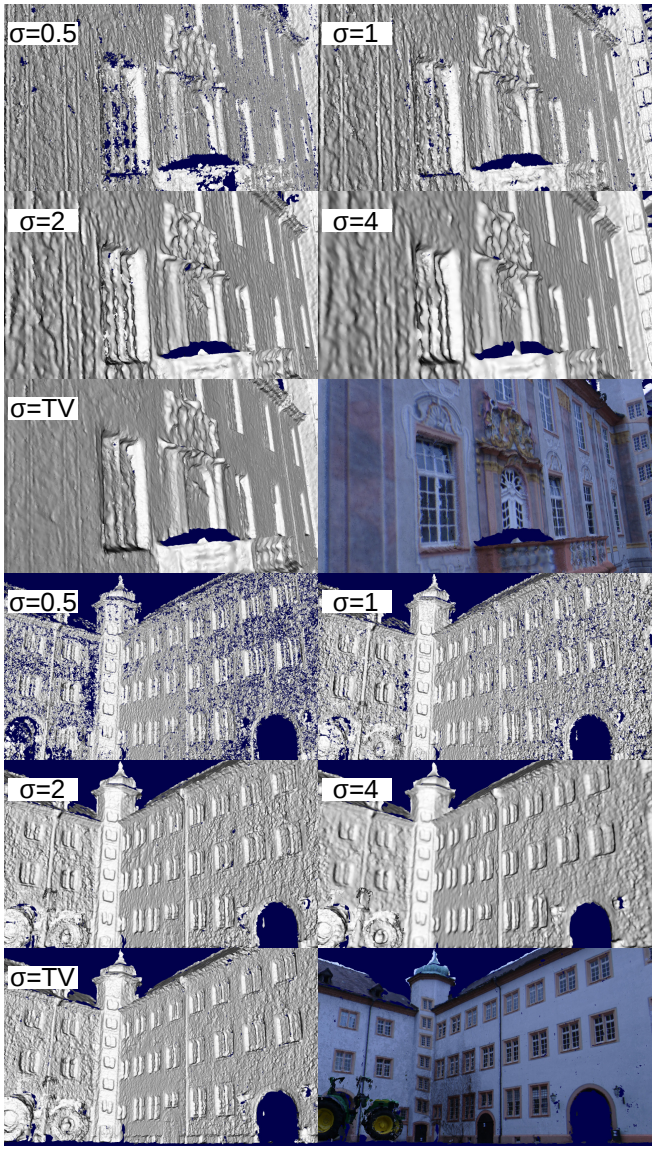

Fig. 11: Constant vs. TV-based priors. Each group of six pictures shows models reconstructed from the Ettlingen30 dataset, comparing reconstructions with constant σ, our new TV-based prior, and the textured TV model. The TV-based reconstruction yields the highest completeness and accuracy.

scores. Thus, the model with TV prior combines the finest details with the highest completeness.

To demonstrate the adaptability of the method, we present results for additional real-world datasets calibrated using a highly precise SfM method (Mayer et al., 2011).

The first dataset consists of 31 images with a resolution of 8 megapixels (MP), and depicts a painted junk car (Fig. 14). The dataset is challenging since the ground is only captured slanted and the object is weakly textured.

The second dataset, "Haus51," consists of 112 10MP images of a standalone building captured from a UAV and from the ground. Fig. 15 shows example images and a comparison of surfaces from different views and at varying level of

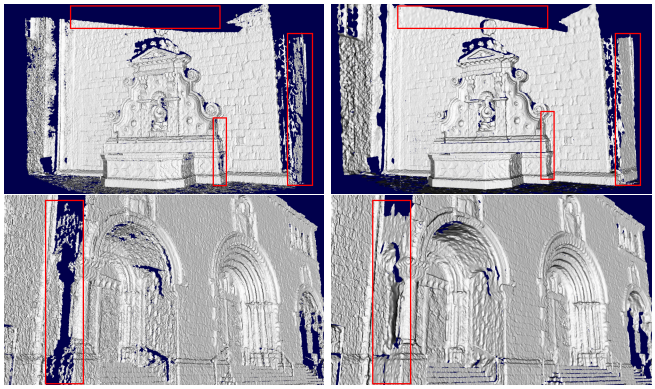

Fig. 12: 3D surface models by Fuhrmann and Goesele (2011) (left) and our method (right) on EttlingenFountain (top) and Herzjesu8 (bottom). Our method yields particularly an improvement for slanted surfaces (red boxes).

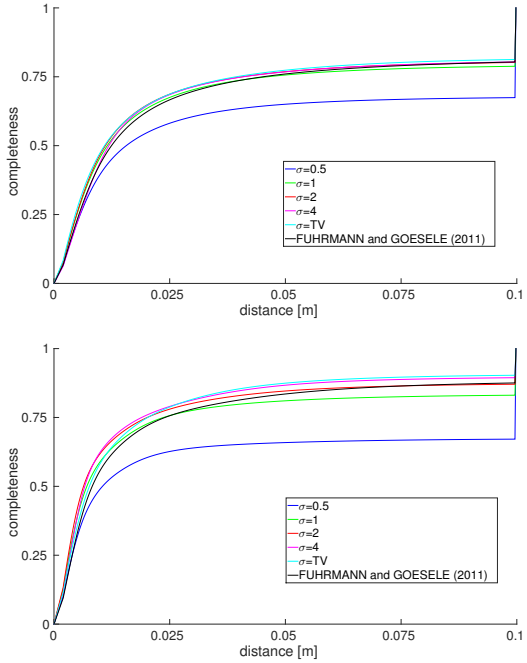

Fig. 13: Unsigned cumulative distance functions of the absolute error from Herzjesu8 (top) and EttlingenFountain (bottom).

	Temple	Dino
Acc.	0.48/0.45/0.49/0.80/**0.43**	0.50/0.44/0.46/0.79/**0.39**
Compl.	59.5/92.8/**97.7**/95.1/96.9	71.8/97.1/**98.3**/95.4/96.3
	TempleRing	DinoRing
Acc.	**0.47**/0.49/0.59/1.09/0.48	0.49/0.47/0.50/1.12/**0.43**
Compl.	77.9/88.3/95.0/91.7/**95.7**	81.2/94.1/**96.7**/89.7/95.3
	TempleSparseRing	DinoSparseRing
Acc.	**0.44**/0.46/0.52/0.71/0.48	0.71/**0.49**/0.51/1.05/**0.49**
Compl.	60.0/77.0/81.8/83.9/**84.9**	72.1/87.0/**92.4**/88.3/89.6

Table 2: Evaluation of Dino and Temple (Seitz et al., 2006) with $\sigma = 0.5/1/2/4/TV$ concerning accuracy and completeness. Best results are marked bold.

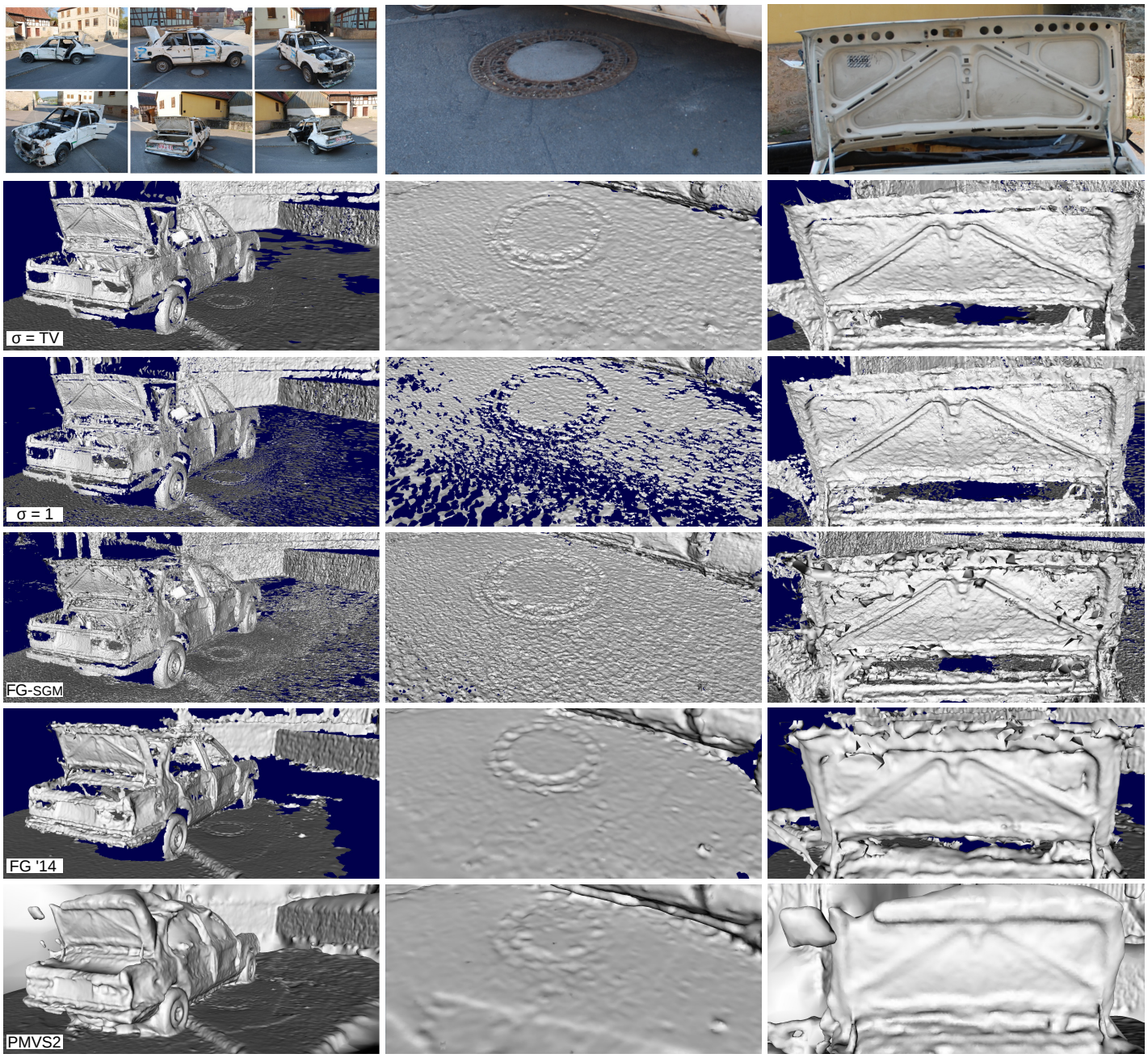

Fig. 14: Junk car dataset with 31 8MP images. The top row shows sample images and zoomed views illustrating the main challenges: the slanted ground plane, and weakly textured surfaces on the car. The remaining five rows show the results by our method, with constant σ, and by three competing methods, FG-SGM (Fuhrmann and Goesele, 2011), FG '14 (Fuhrmann and Goesele, 2014), and PMVS2 (Furukawa and Ponce, 2010) (see text for detail). Our method can compensate for uncertainties by means of the TV prior, recovering small details while minimizing outliers.

detail. The camera configuration is characterized by varying perspectives and degree of texturedness of the building. Furthermore, the building contains challenging 3D structures such as the balconies.

The third dataset, "Unikirche," combines 10MP images of a building captured from the ground and from a UAV with 36MP images of the building door (Fig. 16). This dataset is very challenging since in addition to multiple resolutions, it also contains varying perspectives, distances, degrees of texturedness, motion blur, and radiometric differences.

For an extensive qualitative comparison with state-of-the-art MVS methods, we compare five algorithms on these three datasets. We compare our method with and without TV prior ($\sigma = \text{TV}$ and $\sigma = 1$) against three popular publicly available SfM and MVS methods, which we denote PMVS2, FG-SGM, and FG '14.

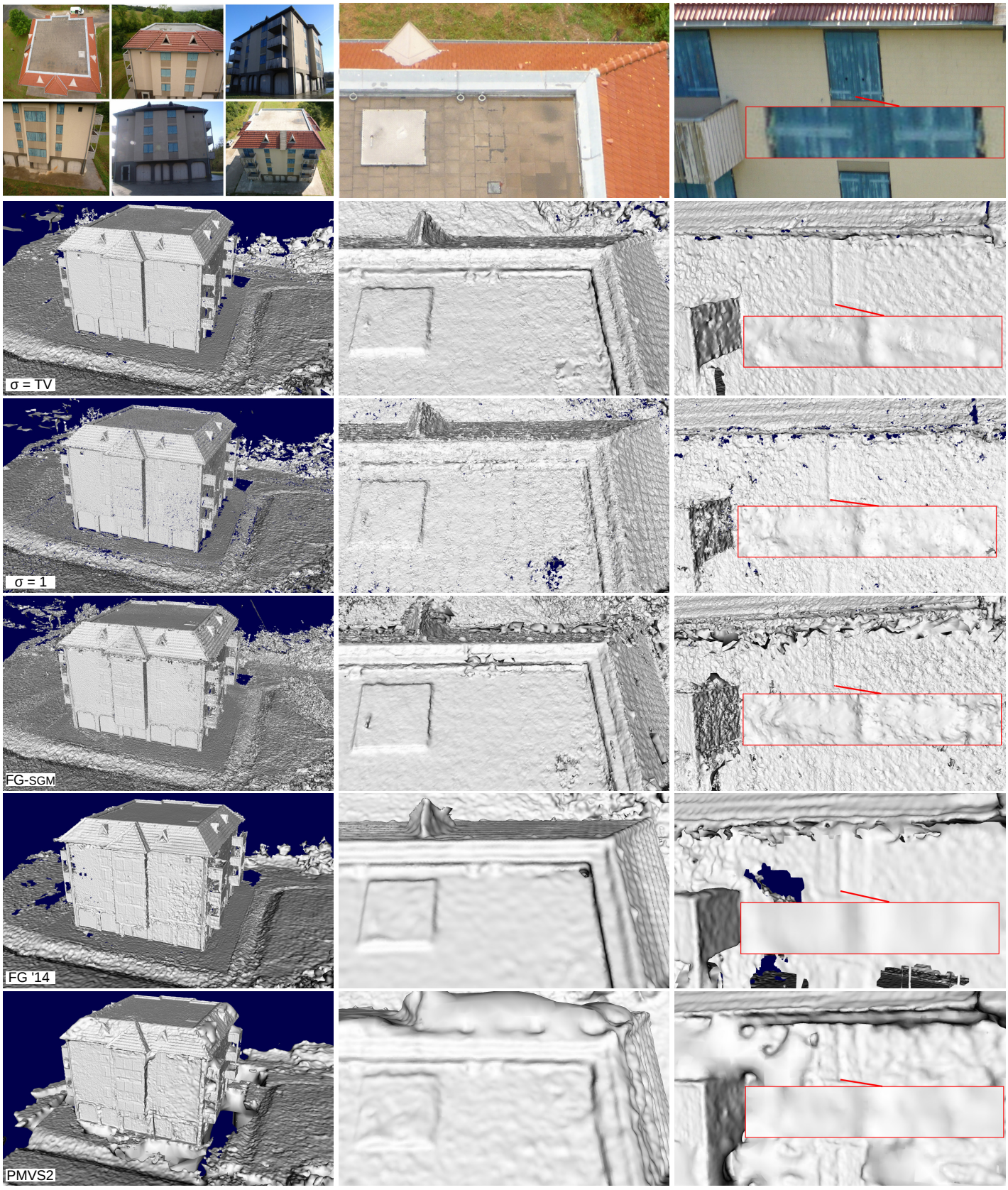

Fig. 15: Haus51 dataset with 112 10MP input images captured from a UAV and from the ground. The top row shows sample images and zoomed views. The varying perspectives, degree of texturedness, and changing lighting conditions are particularly challenging for MVS. The remaining rows show comparative results. Generally our TV prior yields best results combining smoothness and completeness while preserving small details and minimizing outliers.

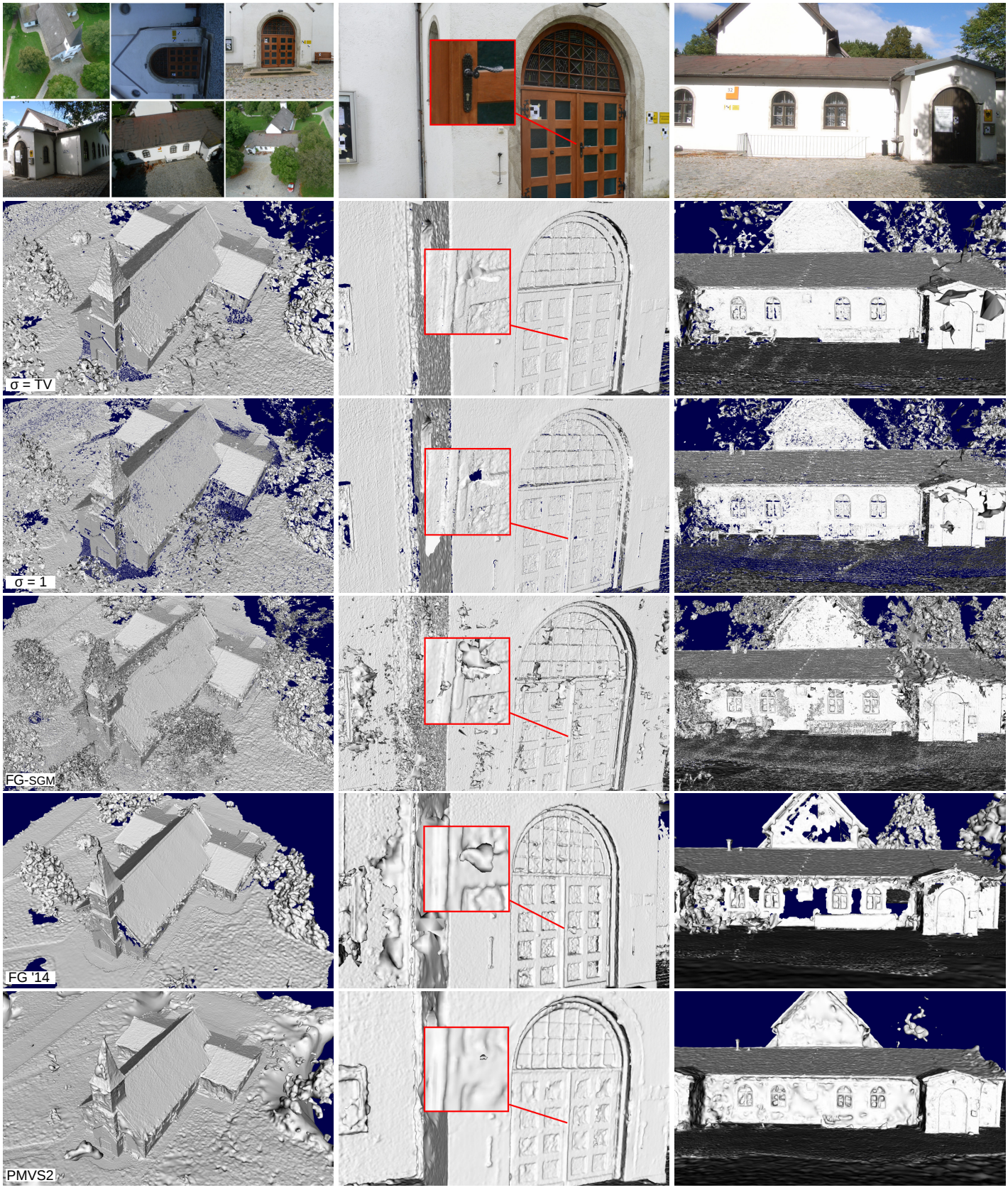

Fig. 16: Unikirche dataset, consisting of 234 input images with mixed 10MP and 36MP resolution, captured from a UAV and from the ground. Challenges include poor lighting, motion blur, weakly textured areas, and large differences in resolution, which result in noisy depth maps, outliers, and incomplete surfaces. Again, our method produces the best results combining smoothness and completeness with high accuracy and minimal outliers.

PMVS2 is the popular reconstruction pipeline consisting of VisualSfM (Wu et al., 2011; Wu, 2013) for SfM, PMVS2 (Furukawa and Ponce, 2010) for MVS, and Poisson reconstruction (Kazhdan et al., 2006) for surface generation.

FG-SGM is the disparity map fusion method by Fuhrmann and Goesele (2011) for SGM disparities, with which we already compared earlier.

Finally, FG '14 represents Fuhrmann and Goesele (2014), who present an improvement of volumetric fusion using 3D basis and weighting functions. As a comparison of the complete pipeline is interesting, we consider their method in their original pipeline using VisualSFM, region-growing-derived disparity maps (Goesele et al., 2007) and mesh cleaning.

For all dense approaches, the results are generated from disparity maps at half resolution as this has been empirically found to give the best trade-off between accuracy and completeness.

As demonstrated in Figs. 14–16, our reconstruction pipeline with TV prior performs best overall, combining smoothness and completeness with high accuracy.

Examining Fig. 14, it is obvious that filtering without the TV prior (row 3) generates incomplete surfaces. The 3 competing methods, which do not perform probabilistic filtering, all yield multiple "ghost" surfaces (right column in rows 4–6). Only our probabilistic fusion with TV prior (row 2) is able to reconstruct complete and detailed surfaces with minimal outliers.

In Fig. 15, our method (row 2) is able to reconstruct fine detail such as the window hinges (see zoomed region on the right), unlike PMVS2 (Furukawa and Ponce, 2010) and FG '14 (Fuhrmann and Goesele, 2014) (rows 5–6). Without the TV prior, closed surfaces have holes in the 3D model (row 3). Holes and ghost surfaces also appear in the competing methods in rows 5–6. The FG-SGM method (Fuhrmann and Goesele, 2011) based on dense SGM disparity maps (row 4) has fewer holes but many more outliers, e.g., at the border of the building roof in the right column.

Similar observations can be made in Fig. 16, which demonstrates that our method (row 2) is able to recover significantly more detail than the competing methods such as PMVS2 (Furukawa and Ponce, 2010) (row 6), for instance the door handle shown in the zoomed region in the middle column. Our results are again more complete than FG '14 (Fuhrmann and Goesele, 2014) (row 5). In general, when non-dense depth maps are used for 3D surface reconstruction, FG '14 tends to yield incomplete surfaces with many holes (row 5), while PMVS2 often results in interpolated ghost surfaces (row 6). The door shown in the middle column, which was imaged from very different distances and with varying resolutions, demonstrates the power of our multi-resolution approach. The FG-SGM method (Fuhrmann and Goesele, 2011) (row 4) generates complete surfaces, but has a much higher outlier rate around complex 3D structures imaged at multiple resolutions. Our method successfully filters such outliers via ray tracing (Sec. 6.4) while taking into account probability information (Sec. 6.3). Finally, comparing the zoomed views in rows 2 and 3, one can observe clear improvements due to our TV prior in terms of smoothness and completeness.

8 Conclusion

In this paper, we have proposed a Total Variation (TV) based regularization term for Multi-View Stereo (MVS). This regularization term is derived from disparity quality classes correlated with the disparity uncertainty. The uncertainty is learned from the difference between generated disparity maps and ground-truth disparities with an Expectation Maximization (EM) method, considering noise and outliers. The knowledge about the quality classes is employed for local volumetric surface reconstruction, which allows for parallel processing of very large models. The uncertainties of the classes are considered when fusing disparity maps into a multi-scale octree structure. To this end, we extended the well-known volumetric fusion of signed distance functions to probabilistic fusion with filtering considering surface quality. Quantitative evaluation, but particularly visual assessment on several datasets indicate a considerable improvement by the proposed means of regularization.

We believe that considering variable disparity quality offers great potential to improve the accuracy and completeness of local volumetric reconstruction, because the fusion area has to be known: Small uncertainties do not tend to interfere, but larger uncertainties often lead to an oversmooth solution. In future work we plan to utilize additional uncertainty information, in particular the registration uncertainty.

References

Christian Bailer, Manuel Finckh, and Hendrik Lensch. Scale robust multi view stereo. In *ECCV*, 2012.

Sid Bao, Manmohan Chandraker, Yuanqing Lin, and Silvio Savarese. Dense object reconstruction with semantic priors. In *CVPR*, 2013.

Tim Bodenmüller. *Streaming Surface Reconstruction from Real Time 3D Measurements*. PhD thesis, Technical University Munich, 2009.

Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *SIGGRAPH*, 1996.

Jan-Michael Frahm, Pierre Georgel, David Gallup, Tim Johnson, Rahul Raguram, Changchang Wu, Yi-Hung Jen, Enrique Dunn, Brian Clipp, Svetlana Lazebnik, and Marc Pollefeys. Building Rome on a cloudless day. In *ECCV*, 2010.

Simon Fuhrmann and Michael Goesele. Fusion of depth maps with multiple scales. In *SIGGRAPH Asia*, 2011.

Simon Fuhrmann and Michael Goesele. Floating scale surface reconstruction. In *SIGGRAPH*, 2014.

Ryo Furukawa, Tomoya Itano, Akihiko Morisaka, and Hiroshi Kawasaki. Improved space carving method for merging and interpolating multiple range images using information of light sources of active stereo. In *ACCV*, 2007.

Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *PAMI*, 32:1362–1376, 2010.

Michael Goesele, Brian Curless, and Steven Seitz. Multi-view stereo revisited. In *CVPR*, 2006.

Michael Goesele, Noah Snavely, Brian Curless, Hugues Hoppe, and Steven Seitz. Multi-view stereo for community photo collections. In *ICCV*, 2007.

Christian Häne, Christopher Zach, Andrea Cohen, Roland Angst, and Marc Pollefeys. Joint 3D scene reconstruction and class segmentation. In *CVPR*, 2013.

Carlos Hernández, George Vogiatzis, and Roberto Cipolla. Probabilistic visibility for multi-view stereo. In *CVPR*, 2007.

Heiko Hirschmüller. Stereo processing by semi-global matching and mutual information. *PAMI*, 30:328–341, 2008.

Heiko Hirschmüller and Daniel Scharstein. Evaluation of stereo matching costs on images with radiometric differences. *PAMI*, 31:1582–1599, 2009.

Xiaoyan Hu and Philippos Mordohai. Least commitment, viewpoint-based, multi-view stereo. In *3DIMPVT*, 2012.

Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Eurographics*, 2006.

Michael Kazhdan, Allison Klein, Ketan Dalal, and Hugues Hoppe. Unconstrained isosurface extraction on arbitrary octrees. In *Eurographics*, 2007.

K. Kolev, M. Klodt, T. Brox, and D. Cremers. Continuous global optimization in multiview 3D reconstruction. *IJCV*, 84:80–96, 2009.

Andreas Kuhn. *Scalable 3D Surface Reconstruction by Local Stochastic Fusion of Disparity Maps*. PhD thesis, University of the Bundeswehr, 2014.

Andreas Kuhn and Helmut Mayer. Incremental division of very large point clouds for scalable 3D surface reconstruction. In *ICCV Workshop (ICCVW)*, 2015.

Andreas Kuhn, Heiko Hirschmüller, and Helmut Mayer. Multi-resolution range data fusion for multi-view stereo reconstruction. In *GCPR*, 2013.

Andreas Kuhn, Helmut Mayer, Heiko Hirschmüller, and Daniel Scharstein. A TV prior for high-quality local multi-view stereo reconstruction. In *3DV*, 2014.

Helmut Mayer, Jan Bartelsen, Heiko Hirschmüller, and Andreas Kuhn. Dense 3D reconstruction from wide baseline image sets. In *15th International Workshop on Theoretical Foundations of Computer Vision*, 2011.

Paul Merrell, Amir Akbarzadeh, Liang Wang, Philippos Mordohai, Jan-Michael Frahm, Ruigang Yang, David Nistér, and Marc Pollefeys. Real-time visibility-based fusion of depth maps. In *CVPR*, 2007.

Nicholas Molton and Michael Brady. Practical structure and motion from stereo when motion is unconstrained. *IJCV*, 39(1):5–23, 2000.

Patrick Mücke, Ronny Klowsky, and Michael Goesele. Surface reconstruction from multi-resolution sample points. In *VMV*, 2011.

Richard Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. KinectFusion: Real-time dense surface mapping and tracking. In *ISMAR*, 2011.

Peter Ochs, Alexey Dosovitskiy, Thomas Brox, and Thomas Pock. An iterated L1 algorithm for non-smooth non-convex optimization in computer vision. In *CVPR*, 2013.

Kaustubh Pathak, Andreas Birk, and S. Schwertfeger. 3D forward sensor modeling and application to occupancy grid based sensor fusion. In *IROS*, 2007.

Leonid Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. In *Physica D*, 1992.

Ryusuke Sagawa, Ko Nishino, and Katsushi Ikeuchi. Adaptively merging large-scale range data with reflectance properties. *PAMI*, 27(3):392–405, 2005.

D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nesic, X. Wang, and P. Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *GCPR*, 2014.

Daniel Scharstein and Chris Pal. Learning conditional random fields in stereo. In *CVPR*, 2007.

Christopher Schroers, Henning Zimmer, Levi Valgaerts, Andrés Bruhn, Oliver Demetz, and Joachim Weickert. Anisotropic range image integration. In *DAGM*, 2012.

Steven Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *CVPR*, 2006.

Sudipta Sinha, Daniel Scharstein, and Richard Szeliski. Efficient high-resolution stereo matching using local plane sweeps. In *CVPR*, 2014.

Frank Steinbrücker, Christian Kerl, Jürgen Sturm, and Daniel Cremers. Large-scale multi-resolution surface reconstruction from RGB-D sequences. In *ICCV*, 2013.

Christoph Strecha, Wolfgang von Hansen, Luc Van Gool, Pascal Fua, and Ulrich Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *CVPR*, 2008.

Sebastian Thrun. Learning occupancy grid maps with forward sensor models. *Auton. Robots*, 15:111–127, 2003.

George Vogiatzis and Carlos Hernández. Video-based, real-time multi-view stereo. *Image Vision Comput.*, 29:434–441, 2011.

Hoang-Hiep Vu, Patrick Labatut, Jean-Philippe Pons, and Renaud Keriven. High accuracy and visibility-consistent dense multiview stereo. *PAMI*, 34:889–901, 2012.

Jian Wei, Benjamin Resch, and Hendrik Lensch. Multi-view depth map estimation with cross-view consistency. In *BMVC*, 2014.

Mark Wheeler, Yoichi Sato, and Katsushi Ikeuchi. Consensus surfaces for modeling 3D objects from multiple range images. In *ICCV*, 1998.

Oliver Woodford and George Vogiatzis. A generative model for online depth fusion. In *ECCV*, 2012.

Changchang Wu. Towards linear-time incremental structure from motion. In *3DV*, 2013.

Changchang Wu, Sameer Agarwal, Brian Curless, , and Steven Seitz. Multicore bundle adjustment. In *CVPR*, 2011.

Yalin Xiong and Larry Matthies. Error analysis of a real-time stereo system. In *CVPR*, 1997.

Christopher Zach. Fast and high quality fusion of depth maps. In *3DPVT*, 2008.

Christopher Zach, Thomas Pock, and Horst Bischof. A globally optimal algorithm for robust TV-L1 range image integration. In *ICCV*, 2007.