# A TV Prior for High-Quality Local Multi-View Stereo Reconstruction

Andeas Kuhn Helmut Mayer
Bundeswehr University Munich
{andreas.kuhn,helmut.mayer}@unibw.de

Heiko Hirschmüller
German Aerospace Centre
heiko.hirschmueller@dlr.de

Daniel Scharstein
Middlebury College
schar@middlebury.edu

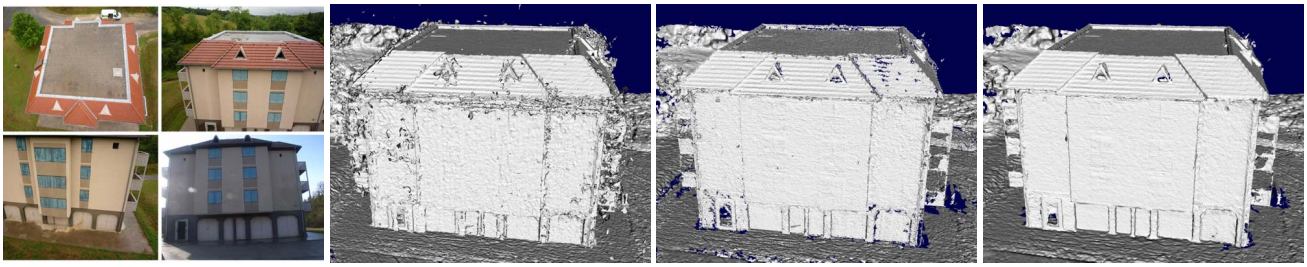

Figure 1. Challenging large-scale multi-view dataset only amenable by local methods. From left to right: four of 113 10-megapixel input images, and 3D reconstructions by Fuhrmann and Goesele [4], Kuhn et al. [11], and our method. Our method produces the cleanest surfaces and also fewer holes than [11]. It is able to adjust to varying disparity quality due to its flexible smoothing term employing a TV prior.

## Abstract

*Local fusion of disparity maps allows fast parallel 3D modeling of large scenes that do not fit into main memory. While existing methods assume a constant disparity uncertainty, disparity errors typically vary spatially from tenths of pixels to several pixels. In this paper we propose a method that employs a set of Gaussians for different disparity classes, instead of a single error model with only one variance. The set of Gaussians is learned from the difference between generated disparity maps and ground-truth disparities. Pixels are assigned particular disparity classes based on a Total Variation (TV) feature measuring the local oscillation behavior of the 2D disparity map. This feature captures uncertainty caused for instance by lack of texture or fronto-parallel bias of the stereo method. Experimental results on several datasets in varying configurations demonstrate that our method yields improved performance both qualitatively and quantitatively.*

## 1. Introduction

Constructing detailed geometric 3D models of the world is still a challenging and open problem in computer vision. Recent progress in Structure from Motion (SfM) and Multi-View Stereo (MVS) allows for fast reconstruction of surfaces from large image sets. 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 [28, 14] concerning completeness and accuracy, for example on the Middlebury multi-view benchmark [22]. Local methods, on the other hand, yield better scalability [4] and runtime performance [15, 24]. Even models of arbitrary size can by reconstructed in parallel without a complex fusion step [11].

Hu and Mordohai [9] show that local MVS methods have potential to reach a quality similar to global methods. In this paper we focus on local methods, arguing that the quality of their results can be further improved by modeling the uncertainties of 2D disparity maps.

## 2. Related Work

Surface reconstruction from depth maps has received considerable interest. 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) for MVS was introduced by Zach et al. [32]. 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 a publication of Rudin et al. [18] on the reconstruction of noisy 2D images. The idea was improved for MVS by further works [31, 10, 21, 16]. TV is important for our method, though we do not perform minimization via global convex optimization.

As mentioned, global methods are limited in practical applications due to poor scalability and runtime perfor-

978-1-4799-7000-1/14 $31.00 © 2014 IEEE
DOI 10.1109/3DV.2014.76

mance. The most promising local methods are volumetric ones employing range image integration, as proposed by Curless and Levoy [3], on stereo images [5]. The 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. Fuhrmann and Goesele [4] extend this method to handling varying surface qualities by introducing a dynamic voxel size. Kuhn et al. [11] propose an alternative probabilistic distance function for multi-resolution voxels, with an additional filtering step that delivers good results in challenging configurations that cause noisy spatial data. We employ this approach in our work.

A great challenge for local methods is the use of a regularization term. Existing methods consider a varying error in 3D [4, 14, 9, 11], 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 method. Our method analyzes the quality only based on the disparity map, independent of camera type and configuration. For integration in MVS we adapt a volumetric approach [3, 11]. Fig. 1 shows sample results by our method demonstrating improved performance over existing local methods.

Learning priors for stereo and MVS is not new. Scharstein and Pal [20] learn a CRF model for stereo from ground-truth disparities. Bao et al. [1] learn priors for semantic categories that consist of the object shape; their method also utilizes information from the SfM process. Häne et al. [6] demonstrate how learning semantic priors for classes such as buildings, ground, vegetation, and clutter can improve the surface quality. The method is based on joint segmentation, labeling, and classification. In contrast, our new TV prior is only based on the input disparities without a need for semantic modeling.

## 3. Volumetric Fusion

Scalable local volumetric surface reconstruction is based on the fusion of cumulative distance functions in efficient data structures like octrees. The depth of the octree, or respectively the size of the voxel, can be handled dynamically [4]. The choice of the voxel size is based on the expected uncertainty of the 3D point [11]. In this section we discuss our adaptation of the distance function and the corresponding fusion process, providing a brief introduction to both.

### 3.1. Signed Distance Function

Like Fuhrmann and Goesele [4] as well as Kuhn et al. [11], we evaluate distance functions depending on the quality of the depth value in 3D. In contrast to our approach, however, their approaches do not consider the quality of the disparities.

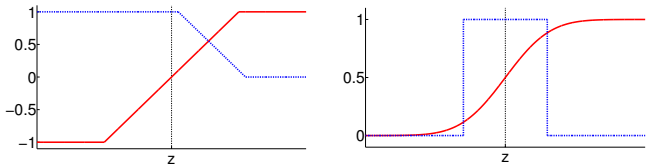


Figure 2. Two alternative cumulative distance functions for an estimated depth $z$. (a) Linear signed distance function $d$ (red) with a weighting function $w$ (blue) penalizing values behind the estimated surface [3, 4]. (b) Gaussian CDF (red) with indicator function (blue) bounding the area of influence [11].

A linear signed distance function $d$ [3, 4] assigns negative values to voxels in front of the estimated depth, and positive values to those behind (Fig. 2a). Following Kuhn et al. [11], we instead use a Gaussian cumulative distance function (CDF) (Fig. 2b), which is transformed into logarithmic ratio space as explained below. The voxels to be assigned a surface value are chosen by intersection of the octree with a part of the line of sight. Additionally, a second function defines the weight $w$ of this value.

The volumetric update process for voxels on the line of sight accumulates the contributions from individual pixels. For the linear function it follows two equations [3]:

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

$$D_{i+1}(v) = \frac{W_i(v)D_i(v) + w_{i+1}(v)d_{i+1}(v)}{W_{i+1}(v)}\,, \tag{2}$$

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 maintained using Equations (1) and (2). 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 [3] empirically adapt the linear weighting function depending on the the angle between line of sight and the normal vector, the slant, of the surface and on the distance to the next missing measurements. We instead propose a novel TV-based probabilistic function that also handles slant and missing measurements and is determined statistically from ground-truth data.

The probabilistic framework [11] is based on the idea of a point lying on the line of sight with Gaussian uncertainty. Hence, the CDF calculates the probability $p(v^1)$ of a point on the line of sight to lie behind the surface (Fig. 2b). This probability is allocated to each voxel intersected by the line of sight. To estimate the probability of a voxel seen from $n$

cameras, the probabilities have to be fused. The update process for the CDF is based on Binary Bayes Theory (BBT):

$$p(v^1|D=d) \propto p(v^1) \prod_j p(D_j = d_j|v^1) \; . \qquad (3)$$

In addition to MVS reconstruction [11], fusion of sensor data via BBT has also been applied to occupancy grid propagation [29, 17, 26].

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 a logarithmic ratio space, with values in the range $[-\infty, \infty]$, and fused via summation [11]:

$$l = log\frac{p(v^1)}{p(v^0)} = log\frac{p(v^1)}{1-p(v^1)} = \sum_j log\frac{p(v_j^1)}{1-p(v_j^1)} \; . \quad (4)$$

Instead of the weighting function $w$ from the linear case, in the probabilistic formulation we use a "boxcar" indicator function extending $\pm 2\sigma$ (Fig. 2b), which limits the influence of voxels to a narrow region around each estimated depth. This generally yields better results due to increased robustness to outliers. Furthermore, 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, and thus requires post-processing or filtering. However, multiple surfaces are possible even without bounding the influence with an indicator function, as disparity maps are generally incomplete.

In our method we adopt the probabilistic framework [11] since it integrates naturally with the probability functions learned from ground-truth data. It also allows extracting surface probabilities, which can be used for filtering.

### 3.2. Dynamic Integration

Multi-resolution methods allow for the fusion of data with strongly differing quality [4, 11]. The probabilistic signed distance functions from Section 3.1 are fused in a 3D volume on different levels. The level corresponds to the depth of the octree and the size of the voxel, and is chosen with respect to the uncertainty of the 3D point. From the point of view of a probabilistic method with an estimated standard deviation $\sigma_d^x$ of the disparity $d$ at pixel $x$, the 3D deviation can be seen as Gaussian [11]:

$$p(z_d^x) = \mathcal{N}(z_d^x, [\sigma_d^x \frac{(z_d^x)^2}{ft}\sqrt{2}]^2) \; , \qquad (5)$$

with depth $z_d^x$ of the 3D point, baseline $t$, and focal length $f$. The depth value $z_d^x$ is the depth estimated by stereo methods. The uncertainty in Eq. 5 can be derived by error propagation in stereo configurations, for instance as described by Molton and Bradey [13]. The estimation of the disparity uncertainty $\sigma_d^x$ is discussed in Section 4.3 below.

The voxel size is chosen using a linear relation of the error from Equation (5), resulting in a sidelength $v_s$ with $\sigma < av_s < 2\sigma$. We use a value of $a = 6$ in all experiments. On this octree level the voxels are updated by estimating the CDF for the intersection point with the line of sight. This probability is fused with probabilities obtained from other cameras as described in Section 3.1. After propagation from all disparity maps, each voxel obtains a probability of lying completely behind a surface.

### 3.3. Filtering and Meshing

Because of the possible occurrence of multiple surfaces, the probabilistic space cannot be transformed directly to polygons. Following Kuhn et al. [11], the probabilistic octree space is transformed into an optimized pointcloud by selecting neighboring voxels on the line of sight such that with maximum probability one is in front and the other behind the surface. The optimized point cloud is then again transformed to an octree which allows for efficient consistency checks by casting rays from all occupied voxels to those cameras the point has been seen from.

When conflicts are detected, we remove voxels with lower surface probability [11]. In addition, voxels on lower octree levels are removed by voxels on higher levels in order to preserve detail. Finally, the estimated clean point cloud is transformed to a triangle mesh using a local meshing method [2].

Fig. 1 illustrates the positive effect of this filtering step by comparing 3D models reconstructed from the same disparity maps in complex image configurations. The multi-resolution approach by Fuhrmann and Goesele [4], which is based on fusion of linear cumulative distance functions, has strongly noticeable visible artifacts due to multiple surfaces in the 3D model, which in turn are caused by holes in the disparity maps. This can be avoided using 3D probabilistic filtering.

## 4. Quality Features

In this section we discuss features strongly influencing the quality of disparities. Based on this discussion, we propose new TV-based feature classes for disparities covering a wide range of the influences. Additionally, we show how to learn the disparity uncertainty from ground-truth disparity maps in comparison with generated disparity maps for individual feature classes using an Expectation Maximization (EM) approach.

We employ Semi-Global Matching (SGM) with census matching costs for disparity estimation [7, 8]. SGM is especially suitable for large sets of images as it has low processing time, and still employs pixelwise matching, resulting in the reconstruction of small details.

Figure 3. Varying disparity quality. Left: Zoomed region of an input image of the Ettlingen30 sequence [25]. Right: Surface orientation of the disparity map computed by SGM visualized using a linear coding from $0^\circ$(light) to $90^\circ$(dark). The surface orientation gives a good impression of the reconstruction quality. Accuracy is lower in slanted and untextured regions (left and right red box), and higher in textured fronto-parallel regions (green box).

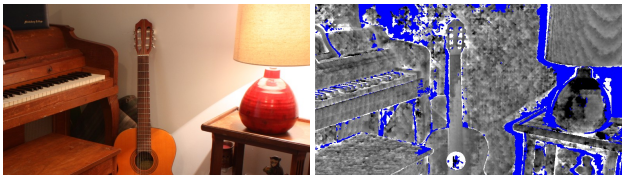

Figure 4. Disparity oscillations. Left: Half-resolution Middlebury Piano image [23]. Right: Signed disparity error of SGM w.r.t. ground truth, coded from -1 (white) to 1 (black). Missing values are in blue. The error exhibits oscillations with varying frequency and amplitude depending on the surface slant and the amount of texture present, as can be observed for instance on the lamp shade in the top right.

### 4.1. Uncertainties in Stereo Matching

Learning quality classes for disparities is not trivial. The quality of disparities is affected by many factors. It usually depends strongly on features such as texture strength and surface slant (see Fig. 3). Slanted surfaces are problematic for common priors employed by most stereo methods, including SGM, that 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 costs. SGM, for example, estimates subpixel disparities by fitting a parabola through the three costs values centered on the winning disparity. Depending on the geometry, this leads to varying uncertainties for subpixel precision.

Unfortunately, there are several additional features that can influence 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 also range from a fraction of a pixel to several pixels. Images from mobile phones are increasingly used in computer vision due to their availability in large numbers. 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 and are subject to motion blur. It is of prime importance to consider different qualities also for disparities.

Naively learning these qualities would result in a multivariate system where the learning space is defined by features covering all of the 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.

### 4.2. Local Total Variation

The key question is how disparity uncertainty can be classified. One possibility would be to estimate the pixelwise normal vector for slant, e.g., considering neighboring disparities in a window, and the image gradient for texture strength. This leads to two problems: First, the disparity maps show an oscillation with unknown frequency (see Fig. 4). The window could oversmooth them, but one could also obtain wrong measurements by undersampling. In Fig. 3 it can be seen that some normal vectors have large orientation errors in weakly-textured or slanted regions. Second, learning the distribution of a 2D function is a hard task 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. This feature represents 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 increased robustness against outliers. Such methods use TV-$L1$ for the estimation of a globally optimal surface from point clouds and aim to limit the 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 $y$:

$$TV(y) = \sum_{i,j \in \mathcal{N}_y} \sqrt{|d_{i+1,j} - d_{i,j}|^2 + |d_{i,j+1} - d_{i,j}|^2} \,. \tag{6}$$

$TV(y)$ represents the degree of the local oscillation in a certain neighborhood of pixel $y$. Unfortunately, oscillations in local neighborhoods have different frequencies. In par-

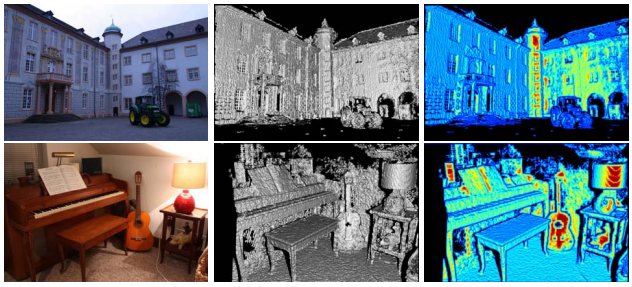

Figure 5. Color visualization of computed TV classes for each pixel. Left: Input images. Middle: Disparity surface orientation (as in Fig. 3) providing a good impression of local reconstruction quality. Right: TV classes measuring disparity smoothness, ranging from 1 (blue) to 20 (red). Fronto-parallel textured surfaces result in higher-numbered classes.

ticular, fronto-parallel planes cause low frequencies, while sloping planes lead to high frequencies (cf. Fig. 4). Hence, it is not feasible to set a constant window for TV estimation.

In addition, we need to discretize the TV term so we can learn the disparity variance from ground truth for each discrete 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 $\tau$. This can be written as:

$$\arg\max_{n}\left(\sum_{m=1}^{n}\frac{1}{8m}TV_{i,j\in x_m} < \tau\right), \tag{7}$$

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 the TV 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 $\infty$ is used. The number of pixels considered 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. 5 shows examples of the computed feature classes.

### 4.3. Learning TV priors

For the learning step 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 representing outliers. In the stereo case this mixture is considered a good approximation for the error distribution [27].

We learn our priors using the 2014 Middlebury stereo datasets with accurate ground truth [19]. We employ half-resolution versions of the seven images used in [23] 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 $n$.

The reason for using EM instead of Maximum Likelihood (ML) estimation is that 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. The calculation of expected value and variance with ML is:

$$\mu = \frac{1}{n}\sum_{i=1}^{n}(d_i - g_i)\,, \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(d_i - g_i - \mu)^2\,, \tag{8}$$

with disparity $d$ and ground truth $g$ for $n$ measurements.

These functions are used as initial state for the EM. For the estimation of an outlier probability we count measurements that lie in an area of five $\sigma$. The ratio of the number of outliers and the number of measurements defines an 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. Afterwards the outlier probability can be obtained as described above. We found that a single EM step yields sufficiently good results.

The resulting expected values and standard deviations for the EM estimation for the 20 classes are shown in Table 1. As expected, the standard deviations for the low-numbered classes (which represent large oscillations) are high, and then decrease quickly in the higher classes. Interestingly, it appears that there is also a positive disparity offset of up to one pixel in the low classes. The likely reason are foreshortening effects [30], since the TV also measures surface slant.

| TV class | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\mu_{EM}$ | 0.98 | 0.48 | 0.11 | 0.04 | 0.03 | 0.03 | 0 | -0.03 | -0.03 | -0.03 |
| $\sigma_{EM}$ | 4.44 | 3.11 | 1.65 | 1.07 | 0.67 | 0.50 | 0.40 | 0.33 | 0.34 | 0.34 |
| TV class | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| $\mu_{EM}$ | -0.03 | -0.03 | -0.02 | -0.02 | -0.02 | -0.01 | 0 | 0.01 | 0.01 | -0.01 |
| $\sigma_{EM}$ | 0.30 | 0.28 | 0.26 | 0.24 | 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.

## 5. Experiments

In this section we show qualitative and quantitative results on established test data as well as novel real-world datasets. We compare surfaces reconstructed using constant standard deviations $\sigma$ ranging from 0.5 to 4 pixels with surfaces reconstructed with variable TV-based standard deviation. (Kuhn et al. [11] use a constant $\sigma = 1$.) In all cases the TV prior yields results that match or exceed the best results obtainable with constant $\sigma$ 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.

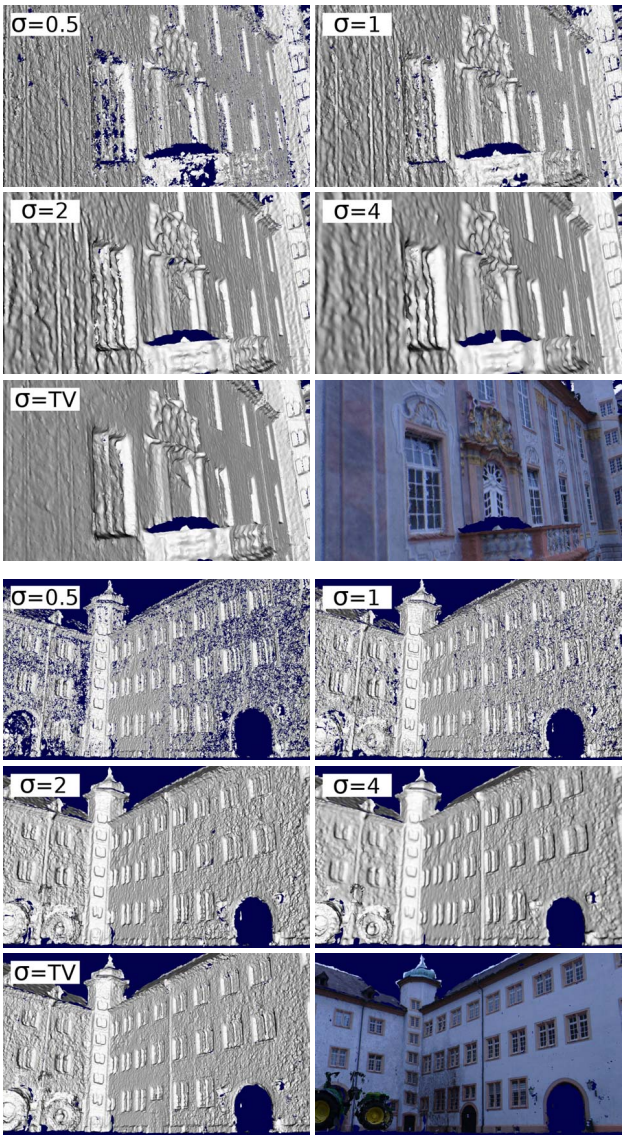


Figure 6. Constant vs. TV-based priors. Each group of 6 pictures shows models reconstructed from the Ettlingen30 dataset, comparing reconstructions with constant $\sigma$, our new TV-based prior, and the textured TV model. The TV-based reconstruction yields highest completeness and accuracy.

The Ettlingen30 dataset [25] (cf. Fig. 6) is well suited for demonstrating our method, as the images have varying perspectives and texture with varying strength. There are two types of difficulties affecting the disparity quality: a lack of texture because of the white walls, and the slanted surfaces, which produce uncertainties because of SGM's fronto-parallel bias. In both cases the TV prior highly weights the textured and fronto-parallel planes. The TV-derived standard deviation leads to the best quality, especially in the areas with many details, comparable to the one with standard deviation 0.5 or 1, but also best quality with respect to completeness, comparable to the results with standard deviation 4.

Additionally, a numerical evaluation on established datasets is of high importance. Strecha et al. [25] provided a numerical evaluation on real-world datasets, which unfortunately is no longer available. Fortunately, for two datasets the ground truth is made public. 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 shown.

Fig. 7 shows the 3D surface models by Fuhrmann and Goesele [4] and by our TV-based method. Our method does particularly well in recovering slanted surfaces.

For the numerical evaluation we employed the technique proposed by Strecha et al. [25]. In the original evaluation the errors are compared against the ground-truth uncertainty. As this information is not available, we employed the absolute error (cf. Fig. 8). To provide a comparison to a popular method, we also processed the disparity maps with the implementation by Fuhrmann and Goesele [4].

For the Herzjesu8 dataset the TV-based surface model is best in terms of accuracy and completeness. In com-

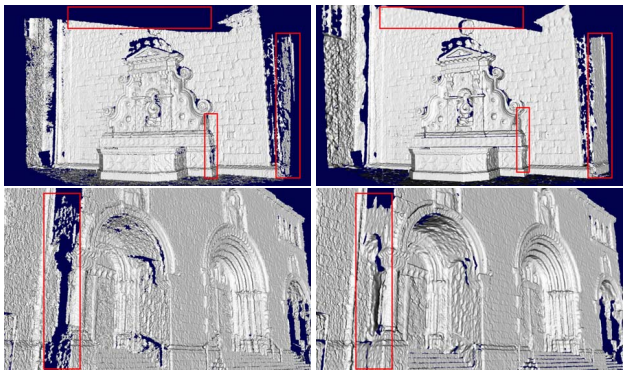

Figure 7. EttlingenFountain (top) and Herzjesu8 (bottom) 3D surface model by [4] (left) and our method (right). Our method yields particular improvement for slanted surfaces (red boxes).

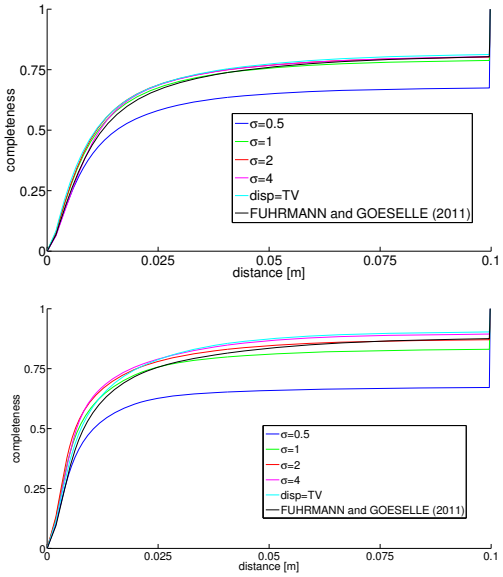


Figure 8. Unsigned cumulative distance functions of the absolute error from Herzjesu8 (top) and EttlingenFountain (bottom).

parison with the method by [4] the evaluation of the TV-based model shows a significant improvement. For the EttlingenFountain dataset the TV based surface quality shows a small loss in accuracy compared to a constant uncertainty. However, the evaluation results would differ in the high-resolution regions if the the ground-truth uncertainty were considered.

We also compared our method with constant-$\sigma$ versions using the Middlebury multi-view benchmark [22]. The datasets cannot show the strength of our method, as the objects do not have variable textures and difficult perspectives. Still, the numerical results (cf. Table 2) confirm the qualitative impression from the previous experiments that the TV results are best concerning accuracy and are mostly best in completeness. In comparison with $\sigma \in \{0.5,1,2,4\}$, the TV results are always close to the best individual accuracy and completeness scores. Thus, the model with TV prior combines the most details with the highest completeness.

To demonstrate the adaptability of the method, we present two additional datasets calibrated by a SfM method [12]. The first dataset comprises 26 36-megapixel images of a building from different perspectives. The second dataset with 31 images of 8 MP resolution depicts a painted junk

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

Table 2. Evaluation of Dino and Temple with $\sigma = 0.5/1/2/4/TV$ concerning accuracy and completeness. Best are marked bold.

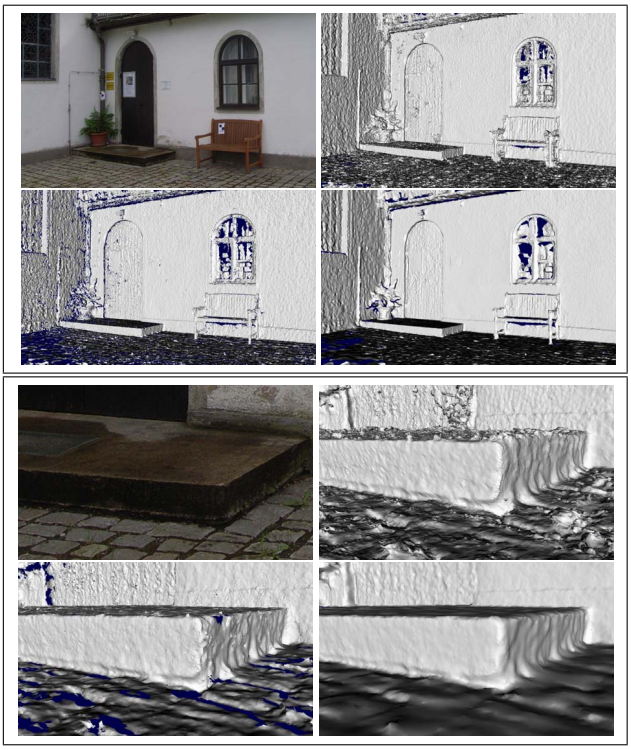

Figure 9. Reconstruction of a building. The two boxes each show: Upper left: Zoomed part of one of 26 36-megapixel images. Upper right: The resulting 3D surface by Fuhrmann and Goesele [4] shows noisy parts, e.g., around the seat bench, on the ground, and at the weakly textured door. Lower left: In the 3D model by Kuhn et al. [11] the surfaces contain holes in weakly textured and slanted areas. Lower right: Our method offers a good trade-off between completeness and clean surfaces and also excels at recovering highly slanted surfaces such as the ground in the bottom box.

car. Figs. 9 and 10 show both an example image and 3D models from competing methods as in Fig. 1. In both cases our method yields the best results and allows for the reconstruction of clean and dense surfaces also in weakly textured and slanted areas.

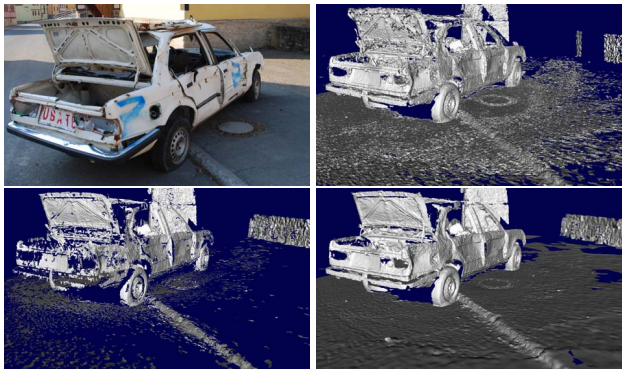

Figure 10. Reconstruction of a junk car. Upper left: Original image part. Upper right: 3D model by [4]. Lower left: 3D model by [11]. Lower right: our reconstruction.

## 6. Conclusion

In this paper, we presented a Total Variation (TV) based regularization term for Multi View Stereo (MVS). This regularization term defines disparity quality classes correlated with the disparity error. 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. Visual assessment on several datasets indicate a considerable improvement by this means of regularization. We consider this novel modeling of the disparity uncertainty as an important step towards a better quality concerning accuracy and completeness for local methods.

We believe that considering variable disparity quality offers great potential for local volumetric reconstruction. This is because the fusion area has to be known as small uncertainties tend not to intertwine, but larger uncertainties tend to lead to an oversmooth solution. In future work, we intend to verify additional features and to utilize additional information, particularly the registration uncertainty.

## References


[1] S. Bao, M. Chandraker, Y. Lin, and S. Savarese. Dense object reconstruction with semantic priors. In *CVPR*, 2013.

[2] T. Bodenmüller. *Streaming Surface Reconstruction from Real Time 3D Measurements*. PhD thesis, Technical University Munich, 2009.

[3] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In *SIGGRAPH*, 1996.

[4] S. Fuhrmann and M. Goesele. Fusion of depth maps with multiple scales. In *SIGGRAPH Asia*, 2011.

[5] M. Goesele, B. Curless, and S. Seitz. Multi-view stereo revisited. In *CVPR*, 2006.

[6] C. Häne, C. Zach, A. Cohen, R. Angst, and M. Pollefeys. Joint 3D scene reconstruction and class segmentation. In *CVPR*, 2013.

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

[8] H. Hirschmüller and D. Scharstein. Evaluation of stereo matching costs on images with radiometric differences. *PAMI*, 31:1582–1599, 2009.

[9] X. Hu and P. Mordohai. Least commitment, viewpoint-based, multi-view stereo. In *3DIMPVT*, 2012.

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

[11] A. Kuhn, H. Hirschmüller, and H. Mayer. Multi-resolution range data fusion for multi-view stereo reconstruction. In *GCPR*, 2013.

[12] H. Mayer, J. Bartelsen, H. Hirschmüller, and A. Kuhn. Dense 3D reconstruction from wide baseline image sets. In *15th International Workshop on Theoretical Foundations of Computer Vision*, 2011.

[13] N. Molton and M. Brady. Practical structure and motion from stereo when motion is unconstrained. *IJCV*, 39(1):5–23, 2000.

[14] P. Mücke, R. Klowsky, and M. Goesele. Surface reconstruction from multi-resolution sample points. In *VMV*, 2011.

[15] R. Newcombe et al. KinectFusion: Real-time dense surface mapping and tracking. In *ISMAR*, 2011.

[16] P. Ochs, A. Dosovitskiy, T. Brox, and T. Pock. An iterated L1 algorithm for non-smooth non-convex optimization in computer vision. In *CVPR*, 2013.

[17] K. Pathak, A. Birk, and S. Schwertfeger. 3D forward sensor modeling and application to occupancy grid based sensor fusion. In *IROS*, 2007.

[18] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. In *Physica D*, 1992.

[19] 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.

[20] D. Scharstein and C. Pal. Learning conditional random fields in stereo. In *CVPR*, 2007.

[21] C. Schroers, H. Zimmer, L. Valgaerts, A. Bruhn, O. Demetz, and J. Weickert. Anisotropic range image integration. In *DAGM*, 2012.

[22] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *CVPR*, 2006.

[23] S. Sinha, D. Scharstein, and R. Szeliski. Efficient high-resolution stereo matching using local plane sweeps. In *CVPR*, 2014.

[24] F. Steinbrücker, C. Kerl, J. Sturm, and D. Cremers. Large-scale multi-resolution surface reconstruction from RGB-D sequences. In *ICCV*, 2013.

[25] C. Strecha, W. von Hansen, L. V. Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *CVPR*, 2008.

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

[27] G. Vogiatzis and C. Hernández. Video-based, real-time multi-view stereo. *Image Vision Comput.*, 29:434–441, 2011.

[28] H.-H. Vu, P. Labatut, J.-P. Pons, and R. Keriven. High accuracy and visibility-consistent dense multiview stereo. *PAMI*, 34:889–901, 2012.

[29] O. Woodford and G. Vogiatzis. A generative model for online depth fusion. In *ECCV*, 2012.

[30] Y. Xiong and L. Matthies. Error analysis of a real-time stereo system. In *CVPR*, 1997.

[31] C. Zach. Fast and high quality fusion of depth maps. In *3DPVT*, 2008.

[32] C. Zach, T. Pock, and H. Bischof. A globally optimal algorithm for robust TV-L1 range image integration. In *ICCV*, 2007.