International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XL-1/W3, 2013
SMPR 2013, 5 – 8 October 2013, Tehran, Iran

FUSION OF MULTI-RESOLUTION DIGITAL SURFACE MODELS

Georg Kuschk, Pablo d'Angelo

Remote Sensing Technology Institute, German Aerospace Center (DLR), D-82234 Wessling, Germany
(georg.kuschk , pablo.angelo)@dlr.de

Commission WG I/4

KEY WORDS: DSM, 3D Reconstruction, Data Fusion

ABSTRACT:

This paper proposes an algorithm for fusing digital surface models (DSM) obtained by heterogenous sensors. Based upon prior confidence knowledge, each DSM can be weighted locally adaptively and therefore strengthen or lessen its influence on the fused result. The proposed algorithm is based on variational methods of first and second order, minimizing a global energy functional comprising of a data term forcing the resulting DSM being similar to all of the input height information and incorporating additional local smoothness constraints. By applying these additional constraints in the form of favoring low gradients in the spatial direction, the surface model is forced to be locally smooth and in contrast to simple mean or median based fusion of the height information, this global formulation of context-awareness reduced the noise level of the result significantly. Minimization of the global energy functional is done with respect to the L_1 norm and therefore is robust to large height differences in the data, which preserves sharp edges and fine details in the fused surface model, which again simple mean- and median-based methods are not able to do in comparable quality. Due to the convexity of the framed energy functional, the solution furthermore is guaranteed to converge towards the global energy minimum. The accuracy of the algorithms and the quality of the resulting fused surface models is evaluated using synthetic datasets and real world spaceborne datasets from different optical satellite sensors.

1 INTRODUCTION

Digital surface models (DSM) are a basic component for many applications, such as orthophoto creation, mapping, visualisation and 3D planing in many application fields. Today, many technologies for DSM generation exist, such as airborne LiDAR, SAR interferometry and automatic image matching, each resulting in a different quality and characteristics. As a result, multiple datasets of DSMs are available for most parts of the earths landmass and it is therefore interesting to fuse these into a single, higher accuracy DSM. Depending on the underlying satellite characteristics like ground sampling distance (GSD), the DSMs capture different parts of the scene in different quality, which even can be mutually exlusive to some extent. For example, high resolution sensors like World-View 2 with a GSD of 0.5m perform very well in urban areas, whereas the results in forest areas are somewhat moderate. In contrast, Cartosat-1 with a GSD of 2.5m performs quite opposite in these areas. Even with the same sensor, a different exposure time can drastically alter the results in shadow areas or in highly reflective areas like glaciers. Apart from the obvious aspect of fusing multiple different DSMs, some techniques produce multiple DSMs for the same area, which need to be fused as well. For example, many multi-view image matching techniques are based on matching individual stereo pairs and later fusing these stereo pairs into a common height model, see e.g. (Hirschmueller, 2008), (Kuschk, 2013), (Rumpler et al., n.d.).

Our work focuses on the fusion of 2.5D DSM grids, with a resolution from several decimeters to a few meters. DSM fusion has been considered by various authors previously. The simplest method is based on weighted averaging of two or more height maps (Schultz et al., 1999), (Reinartz et al., 2005). As weighted averaging cannot deal with outliers or blunders in the DSMs, a median fusion is often used for multi-DSM fusion, sometimes followed by weighted averaging of the inliers (Hirschmueller, 2008). Both median and weighted averaging does process each pixel independently, and thus cannot take into account the local surface shape, which is usually very regular. Applying additional mean or median based filtering spatially reduces the amount of noise to some extent, at the cost of blurring potentially sharp edges. An example for context aware fusion algorithms is the use of sparse representations (Papasaika et al., 2011), where a DSM patch is computed as a sparse linear combination of dictionary DSM patches. Except for median fusion, pixelwise error maps are required by weighted averaging and sparse representations. A comparison between weighted averaging and sparse representations (Schindler et al., 2011) found that the quality of the fused DSMs is mostly determined by the quality of these pixel based error maps.

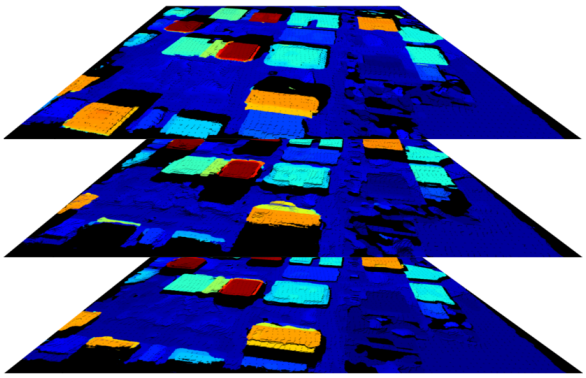


Figure 1: Fusion of multiple DSMs covering the same area

Another direction of work (Pock et al., 2011) aims to formulate a global energy function, minimizing the distance of the fused result to all input DSMs simultaneously and additionally incorporates the assumption of the world being locally planar. Due to its simple structure and theoretically well founded minimization procedure, we build upon this work and extend it to a weighted, multi-resolution, fusion framework.

2 METHOD

As most computer vision problems are generally ill-posed (e.g. image segmentation, stereo reconstruction, image fusion), additional regularizers (constraints) are needed for physical meaningful solutions. In our case of 2.5D image fusion, these regularizers are the assumption of the world being locally planar, meaning that height value of each pixel of the DSM depends on its local context and e.g. is highly unlikely to have a signifanctly different height value than its surrounding pixels. This smoothness constraint typically is implemented by minimizing the gradient of the resulting DSM. Together with the data term this results in large systems of partial differential equations (pde) which are time-consuming to solve and special care has to be taken to still allow for strong discontinuities along building edges and to not smooth them over. The design of the energy functional has to be chosen such that it is convex in the variable to solve for. Otherwise it would be very hard for non-linear minimization method to not getting stuck in local minima.
In recent years, Total Variation based methods (TV) for minimizing energy functionals have seen a lot of attention in the research community. One reason is that these algorithms are very well-suited for parallelization and, together with the recent advances of GPU-based computational power, lead to efficient algorithms, solving these optimization problems efficiently. And as the energy functional of our image fusion problem can be written in a convex formulation, the solution is globally optimal.

2.1 TV-L_1 Fusion

Based upon the ROF-model for image denoising (Rudin et al., 1992), the extension for multiple image fusion, together with replacing the quadratic data term by the more robust L_1 norm as in (Pock et al., 2011) is written by

$$\min_{\mathbf{u}} \left\{ \|\nabla \mathbf{u}\|_1 + \lambda_d \sum_{k=1}^{K} \|\mathbf{u} - \mathbf{g_k}\|_1 \right\} \tag{1}$$

where $\mathbf{u} \in \mathbb{R}^{M \cdot N}$ is our fused DSM to solve, the K input DSMs are given as $\mathbf{g_k}$ and the scalar factor λ_d balances the impact of the smoothness term and the data term. While this model already provides good results by smoothing flat areas and preserving sharp discontinuities, it suffers from the so-called staircasing effect. This effect is a direct result of the regularizer, whose assumption is a locally planar world - where planar unfortunately refers to locally fronto-parallel. This staircasing effect of the TV-L_1 algorithm is visible in Figure 2.

2.2 TGV-L_1 Fusion

To overcome the fronto-parallel assumption of TV-L_1 minimization, (Bredies et al., 2010) introduced the mathematical model of Total Generalized Variation (TGV) as a higher-order extension of Total Variation which favors the solution to consist of piecewise polynomial functions (e.g. fronto-parallel, affine, quadratic). Especially the 2nd order is of high interest, as it forces the solution to consist of piecewise planar functions. In contrast to the TV-L_1 model, now also including slanted planes. (Pock et al., 2011) applied this model to DSM fusion, resulting in the following optimization problem

$$\min_{\mathbf{u},\mathbf{v}} \left\{ \lambda_s \|\nabla \mathbf{u} - \mathbf{v}\|_1 + \lambda_a \|\nabla \mathbf{v}\|_1 + \lambda_d \sum_{k=1}^{K} \|\mathbf{u} - \mathbf{g_k}\|_1 \right\} \tag{2}$$

Now, before the variation of the image $\mathbf{u}$ is measured, a 2D vector field $\mathbf{v}$ is subtracted from the gradient of $\mathbf{u}$. An affine surface in the image $\mathbf{u}$ has a constant gradient $\nabla \mathbf{u}$, so by coupling and minimizing $|\nabla \mathbf{u} - \mathbf{v}|$, the vector field $\mathbf{v}$ will also be constant and it's gradient $\nabla \mathbf{v}$ therefore zero. Regarding our overall optimization problem, this means that the energy term will be lower, if affine functions can be found in the image, whereas non-affine function get additional penalties by $|\nabla \mathbf{v}|$. The values $\lambda_s, \lambda_a, \lambda_d$ are scalar weights and balance the impact of the smoothness term, the affine term and the data term.

2.3 Weighted TGV-L_1 Fusion

When fusing DSMs it is desirable to weight the input DSMs on a per pixel base, to be able to incorporate additional prior knowledge into the fusion process. This prior knowledge for example can be based on the different sensor characteristics used to generate the DSM, confidence measures during the 3D reconstruction process itself, information about occluded and therefore unknown areas in each DSM, etc. We therefore extend Equation 2 with a weighting matrix W_k for each input DSM

$$\min_{\mathbf{u},\mathbf{v}} \left\{ \lambda_s \|\nabla \mathbf{u} - \mathbf{v}\|_1 + \lambda_a \|\nabla \mathbf{v}\|_1 + \lambda_d \sum_{k=1}^{K} W_k \|\mathbf{u} - \mathbf{g_k}\|_1 \right\} \tag{3}$$

This optimization problem (and the ones in Equation 1 and 2) is very parameter dependent, as we need to adapt the influence of the data term λ_d manually for datasets with different ranges of $g_k^{(i,j)} \in \mathbf{g_k}$ as well as for a different number K of input images. To achieve independence of the data range of the input DSMs, we scale all input data to the interval $[0..1]$

$$g_k^{(i,j)} = \frac{g_k^{(i,j)} - g_{min}}{g_{max} - g_{min}} \tag{4}$$

with $g_{min} = \min_{i,j,k} g_k^{(i,j)}$ and $g_{max} = \max_{i,j,k} g_k^{(i,j)}$. The independence from K is achieved by normalizing the influence of the data term w.r.t. the two-image case and using the adaptive

$$\lambda_d^K = \frac{2}{K} \lambda_d \tag{5}$$

Note that all these extensions and modifications apply to the TV-L_1 method similarly. In the next section we will go into detail about how to solve these optimization problems numerically.

3 ALGORITHM

To solve for the fused DSM $\mathbf{u} \in \mathbb{R}^{M \times N}$ (in the following written as stacked vector $\mathbb{R}^{MN \times 1}$) in Equation 3, we need to overcome the non-differentiable L_1-norm, which complicates any gradient descent based minimization scheme. An efficient algorithm which elegantly circumvents the differentiability problem of the gradient operator is the primal-dual algorithm of (Chambolle and Pock, 2011).
By applying the Legendre-Fenchel transform we obtain the dual formulation / conjugate of the separate terms as

$$\begin{aligned} \lambda_s \|\nabla \mathbf{u} - \mathbf{v}\|_1 &= \max_{\mathbf{p} \in P} \{\langle \nabla \mathbf{u} - \mathbf{v}, \mathbf{p} \rangle\} \\ \lambda_a \|\nabla \mathbf{v}\|_1 &= \max_{\mathbf{q} \in Q} \{\langle \nabla \mathbf{v}, \mathbf{q} \rangle\} \\ \lambda_d \sum_{k=1}^{K} W_k \|\mathbf{u} - \mathbf{g_k}\|_1 &= \max_{\mathbf{r_k} \in R} \left\{ \sum_{k=1}^{K} \langle \mathbf{u} - \mathbf{g_k}, W_k \mathbf{r_k} \rangle \right\} \end{aligned} \tag{6}$$

such that the saddle-point problem in the primal variables $\mathbf{u}, \mathbf{v}$ and the dual variables $\mathbf{p}, \mathbf{q}, \mathbf{r_k}$ with constraints

$$\begin{aligned} P &= \{\mathbf{p} \in \mathbb{R}^{2MN} : \|\mathbf{p}\|_\infty \leq \lambda_s\} \\ Q &= \{\mathbf{q} \in \mathbb{R}^{4MN} : \|\mathbf{q}\|_\infty \leq \lambda_a\} \\ R &= \{\mathbf{r_k} \in \mathbb{R}^{MN} : \|\mathbf{r_k}\|_\infty \leq \lambda_d\} \end{aligned} \tag{7}$$

is

$$\min_{\mathbf{u},\mathbf{v}} \max_{\mathbf{p},\mathbf{q},\mathbf{r_k}} \left\{ \langle \nabla \mathbf{u} - \mathbf{v}, \mathbf{p} \rangle + \langle \nabla \mathbf{v}, \mathbf{q} \rangle + \sum_{k=1}^{K} \langle \mathbf{u} - \mathbf{g_k}, W_k \mathbf{r_k} \rangle \right\} \tag{8}$$

Please note that due to the stacked vector notation, the input weights are denoted as diagonal matrices W_k and the corresponding multiplication $W_k \mathbf{r_k}$ is actually a pixelwise multiplication.
The saddle-point problem above can be solved by iteratively performing gradient descents on the primal variables and gradient ascents on the dual variables. Applying this primal-dual algorithm leads to the following optimization scheme:

do

$$\begin{aligned} \mathbf{p}^{n+1} &= \Pi_P \left(\mathbf{p}^n + \tau_p (\nabla \overline{\mathbf{u}}^n - \overline{\mathbf{v}}^n) \right) \\ \mathbf{q}^{n+1} &= \Pi_Q \left(\mathbf{q}^n + \tau_q \nabla \overline{\mathbf{v}}^n \right) \\ \mathbf{r_k}^{n+1} &= \Pi_R \left(\mathbf{r_k}^n + \tau_r (\overline{\mathbf{u}}^n - \mathbf{g_k}) \right) \\ \mathbf{u}^{n+1} &= \mathbf{u}^n + \tau_u \mathrm{div} \mathbf{p}^{n+1} - \tau_u \sum_{k=1}^{K} (W_k \mathbf{r_k}^{n+1}) \\ \mathbf{v}^{n+1} &= \mathbf{v}^n + \tau_v \mathbf{p}^{n+1} + \tau_v \mathrm{div} \mathbf{q}^{n+1} \\ \overline{\mathbf{u}}^{n+1} &= 2\mathbf{u}^{n+1} - \mathbf{u}^n \\ \overline{\mathbf{v}}^{n+1} &= 2\mathbf{v}^{n+1} - \mathbf{v}^n \end{aligned} \tag{9}$$

while $(n < nIterations) \;||\; (\Delta Energy < \Delta EnergyThres)$

To ensure the constraints of Equation 7, the corresponding proximal mappings above are given as

$$\begin{aligned} \Pi_P(\mathbf{p}) &= \frac{\mathbf{p}}{\max\{1, \|\mathbf{p}\|/\lambda_s\}} \\ \Pi_Q(\mathbf{q}) &= \frac{\mathbf{q}}{\max\{1, \|\mathbf{q}\|/\lambda_a\}} \\ \Pi_R(\mathbf{r_k}) &= \frac{\mathbf{r_k}}{\max\{1, \|\mathbf{r_k}\|/\lambda_d\}} \end{aligned} \tag{10}$$

In the analytical derivation of the primal-dual scheme, we require the gradient and divergence operators to be negative adjoint, such that $\langle \nabla \mathbf{u}, \mathbf{p} \rangle = -\langle \mathbf{u}, \mathrm{div} \mathbf{p} \rangle$ and $\langle \nabla \mathbf{v}, \mathbf{q} \rangle = -\langle \mathbf{v}, \mathrm{div} \mathbf{q} \rangle$. Therefore we use finite forward differences with Neumann boundary conditions for the gradient operators and for the divergence operators finite backward difference with Dirichlet boundary conditions. The step sizes of the gradient ascents/descents are bound to the norm of the gradient/divergence operators, see (Chambolle and Pock, 2011), and are set to $\tau_u = \tau_p = \tau_r = 1/\sqrt{K_1}$ and $\tau_v = \tau_q = 1/\sqrt{K_2}$, with $K_1 = 2(6 + K)$ and $K_2 = 2(4 + K)$.

The whole algorithm stops, if either the maximum number of iterations has been reached ($nIterations = 1000$) or the energy change between successive iterations drops below a relative threshold $\Delta EnergyThres = 0.1\%$.
Again, the algorithm for the TV-L_1 fusion is derived similarly.

4 EVALUATION

4.1 Artificial Tests

The first evaluation is done on synthetic data. A given ground truth DSM $\mathbf{g}$ is perturbed with noise to simulate different noisy observations of the scene. Five of these noisy DSMs are then given as input to the fusion algorithms and the accuracy of the output DSM $\mathbf{u}$ is measured by the logarithmic signal-to-noise ratio:

$$SNR = 10 \log_{10} \left(\frac{I^2_{\text{signal}}}{I^2_{\text{noise}}} \right) = 10 \log_{10} \left(\frac{||\mathbf{g}||^2}{||\mathbf{u} - \mathbf{g}||^2} \right) \tag{11}$$

In Figure 2, visual and numerical results are given, showing a significantly higher accuracy of the global optimization methods for DSM fusion over simple mean and median based fusion. Please note that we applied mean and median filtering for all height values of a single pixel as well as its neighboring pixels to also include some spatial regularization.

4.2 Unimodal DSM fusion

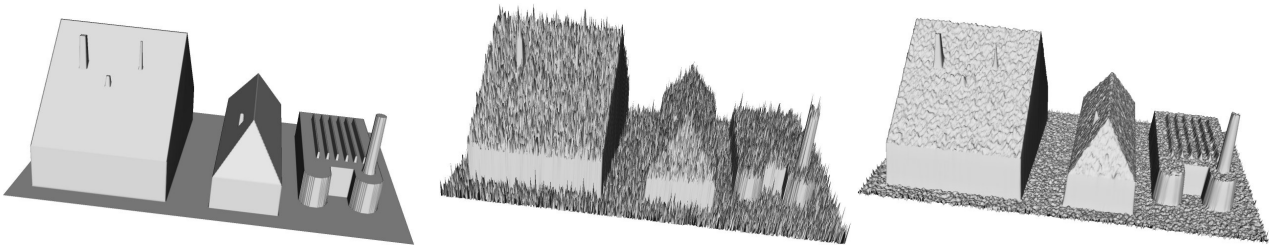
In our second evaluation, we created 10 different DSMs of the same 1km × 1km area of the inner city of London using a stereo reconstruction framework as proposed in (Kuschk, 2013). For this we have a collection of 25 WorldView-2 images, taken from different positions during one pass of the satellite. The ground sampling distance (GSD) of these images are 0.5m and for evaluation purposes, we obtained a LiDAR measurement of the same area by aerial laser scanning, unfortunately only having a GSD of 1.0m. Two of the satellite images together with the computed heightmaps are shown in Figure 3. The 10 selected heightmaps are projected in the same orthogonal UTM coordinate system and the resulting DSMs are again fed into the different fusion algorithms. Figure 4 shows the result of median based fusion versus TGV-L_1 fusion, again with a lower noise ratio visible for the TGV-L_1 algorithm. The accuracy of the fused DSMs w.r.t. to the LiDAR ground truth is given in Table in the common error metrics Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Normalized Median Absolute Deviation (NMAD). Here the improvements are hardly detectable at all, with all algorithms exhibiting similar numerical results. As of yet we do not have further explanation for these results, but strongly suspect the coarse resolution / GSD of our ground truth data mentioned above (in average we only have 1 LiDAR point covering the area of 4 DSM pixels).

	MAE [m]	RMSE [m]	NMAD [m]
median	1.74	3.89	1.62
TV-L_1	1.75	3.65	1.55
TGV-L_1	1.69	3.64	1.51

Table 1: London dataset: Accuracy of the fused DSM w.r.t. ground truth obtained by aerial laserscanning (LiDAR)

4.3 Multimodal DSM fusion

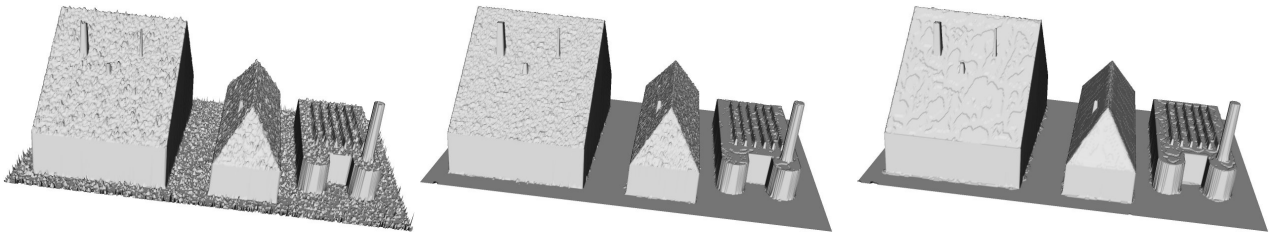
Our third evaluation is investigating the results of fusing DSMs derived from different sensors and different spatial resolutions. The test data is taken from the ISPRS benchmark (Reinartz et al., 2010) and consists of 3 different scenes (hilly forest = *La Mola*, mountains = *Vacarisses*, city = *Terassa*) near Barcelona, Spain. For each scene, we compute a DSM from the two given CartoSat-1 images (GSD=2.5m) and a DSM from the two given



(a) Groundtruth data

(b) 1 of 5 noisy input images

(c) Fusion using pixel based plus spatial mean filtering, $SNR = 22.76$

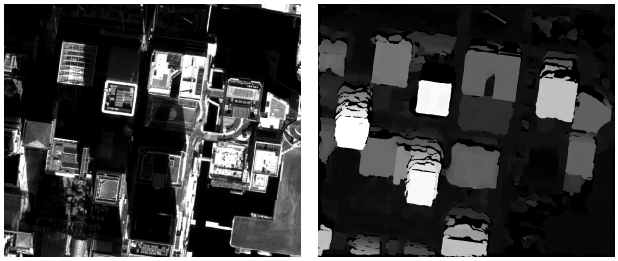


(d) Fusion using pixel based plus spatial median filtering, $SNR = 32.61$

(e) TV-L_1 fusion, $SNR = 39.60$

(f) TGV-L_1 fusion, $SNR = 40.32$

Figure 2: Comparison of local fusion method versus global optimization methods. Both numerical results and visual appearance show the benefit of the latter ones.



(a) Input image A

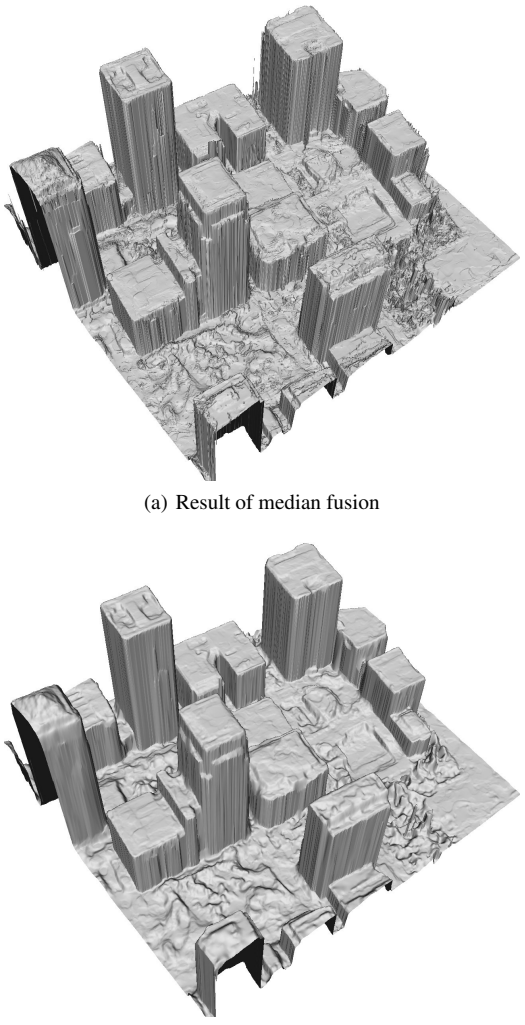
(b) Heightmap of image A

(c) Input image B

(d) Heightmap of image B

Figure 3: London dataset: 2 out of 10 input images and their corresponding heightmaps

WorldView-2 images (GSD=0.5m). Fusion of the different resolution DSMs is done by projecting the DSMs into the same coordinate frame and setting the weights w_k for the invalid pixels of the coarser resolution to zero. As reference we have LiDAR data with a GSD of about 1.0m. The numerical results of local median fusion and global TV-L_1 and TGV-L_1 fusion are given in Table 2, indicating an improvement of both local and global fusion algorithms over the coarse CartoSat-1 DSM, but only the global fusion algorithm improves the result of the high resolution WorldView-2 DSM.



(a) Result of median fusion

(b) Result of TGV fusion

Figure 4: London dataset: Fusion results

(a) WorldView-2, *La Mola* (b) WorldView-2, *Terassa* (c) WorldView-2, *Vacarisses*

Figure 5: ISPRS dataset: Exemplary WorldView-2 images of the three sub datasets

Algorithm	La Mola			Terassa			Vacarisses		
	MAE[m]	RMSE[m]	NMAD[m]	MAE[m]	RMSE[m]	NMAD[m]	MAE[m]	RMSE[m]	NMAD[m]
CartoSat-1	4.82	12.77	2.18	2.80	5.90	1.81	3.76	8.80	2.11
WV-2	4.43	8.31	4.00	2.31	3.66	2.06	3.05	5.00	3.07
median	4.51	10.45	3.03	2.44	4.71	1.79	3.07	6.17	2.49
TGV-L_1	4.31	8.11	3.89	2.27	3.61	2.00	2.99	4.89	3.00

Table 2: Results of local median fusion and global TGV-L_1 fusion for heterogenous sensor data (CartoSat-1 and WorldView-2 satellite images). The first two rows show the accuracy of the unfused DSM of each satellite separately, whereas the two bottom rows show the fusion results of the two DSMs per scene.

5 CONCLUSION

In this paper we proposed global optimization algorithms for fusing multi-resolution DSM obtained by heterogenous sensors. These global optimization algorithms are based on adaptively weighted TV-L_1 and TGV-L_1 optimization problems, allowing for a context-aware fusion of multiple DSMs. In three different evaluations, both synthetic and real world data sets, a significant improvement of the accuracy was shown with respect to mean and median based filtering methods.

ACKNOWLEDGMENT

The authors would like to thank European Space Imaging (EUSI) for providing the WorldView-2 data of London.

REFERENCES

Bredies, K., Kunisch, K. and Pock, T., 2010. Total generalized variation. SIAM Journal on Imaging Sciences 3(3), pp. 492–526. 2

Chambolle, A. and Pock, T., 2011. A first-order primal-dual algorithm for convex problems with applications to imaging. Journal of Mathematical Imaging and Vision pp. 1–26. 2, 3

Hirschmueller, H., 2008. Stereo processing by semiglobal matching and mutual information. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 328–341. semi global matching - das gute und alles zusammengefasste paper. 1

Kuschk, G., 2013. Large scale urban reconstruction from remote sensing imagery. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 5, pp. W1. 1, 3

Papasaika, H., Kokiopoulou, E., Baltsavias, E., Schindler, K. and Kressner, D., 2011. Fusion of digital elevation models using sparse representations. Photogrammetric Image Analysis pp. 171–184. 1

Pock, T., Zebedin, L. and Bischof, H., 2011. Tgv-fusion. Rainbow of computer science pp. 245–258. 1, 2

Reinartz, P., dAngelo, P., Krauß, T., Poli, D., Jacobsen, K. and Buyuksalih, G., 2010. Benchmarking and quality analysis of dem generated from high and very high resolution optical stereo satellite data. In: ISPRS Symposium Commission I. 3

Reinartz, P., Müller, R., Hoja, D., Lehner, M. and Schroeder, M., 2005. Comparison and fusion of dem derived from spot-5 hrs and srtm data and estimation of forest heights. In: Proc. EARSeL Workshop on 3D-Remote Sensing, Porto. 1

Rudin, L., Osher, S. and Fatemi, E., 1992. Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena 60(1-4), pp. 259–268. 2

Rumpler, M., Irschara, A., Wendel, A. and Bischof, H., n.d. Rapid 3d city model approximation from publicly available geographic data sources and georeferenced aerial images. 1

Schindler, K., Papasaika-Hanusch, H., Schuetz, S. and Baltsavias, E., 2011. Improving wide-area dems through data fusion–chances and limits. 1

Schultz, H., Riseman, E. M., Stolle, F. R. and Woo, D.-M., 1999. Error detection and dem fusion using self-consistency. In: Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on, Vol. 2, IEEE, pp. 1174–1181. 1