Enhancement of Depth Map by Fusion using Adaptive and Semantic-Guided Spatiotemporal Filtering

Hessah Albanwan [1], Rongjun Qin [1,2] *

[1] Geospatial Data Analytics Laboratory, Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, OH, USA - albanwan.1@osu.edu

[2] Department of Electrical and Computer Engineering, The Ohio State University - qin.324@osu.edu

Commission VI, WG VI/4

KEY WORDS: Multi-depth Fusion, Digital Surface Model (DSM), Adaptive Spatiotemporal Fusion, Multi-view Stereo (MVS)

ABSTRACT:

Extracting detailed geometric information about a scene relies on the quality of the depth maps (e.g. Digital Elevation Surfaces, DSM) to enhance the performance of 3D model reconstruction. Elevation information from LiDAR is often expensive and hard to obtain. The most common approach to generate depth maps is through multi-view stereo (MVS) methods (e.g. dense stereo image matching). The quality of single depth maps, however, is often prone to noise, outliers, and missing data points due to the quality of the acquired image pairs. A reference multi-view image pair must be noise-free and clear to ensure high-quality depth maps. To avoid such a problem, current researches are headed toward fusing multiple depth maps to recover the shortcomings of single-depth maps resulted from a single pair of multi-view images. Several approaches tackled this problem by merging and fusing depth maps, using probabilistic and deterministic methods, but few discussed how these fused depth maps can be refined through adaptive spatiotemporal analysis algorithms (e.g. spatiotemporal filters). The motivation is to push towards preserving the high precision and detail level of depth maps while optimizing the performance, robustness, and efficiency of the algorithm.

1. INTRODUCTION

1.1 Background

Over the last few decades, a large number of Very High Resolution (VHR) Satellites are established to provide sub-meter resolution imageries, with frequent re-visiting times during the year to allow extracting comprehensive 3D geometrical information about the scene. Algorithms such as Multi-view stereo (MVS) highly depend on the spatial and temporal resolutions of sensors to facilitate generating reliable 3D reconstructed models. However, dense stereo image matching algorithms often rely on the quality of the image pair and nature of the captured surface. Images captured by satellite sensors are prone to spectral inconsistencies and distortions, which may affect the dense stereo matching algorithm and produce incorrect or missing height information, and therefore, degrade the quality of the depth map. In particular, MVS algorithms are very sensitive to temporal inconsistencies between the images, thus, they cannot be used directly to obtain 3D models and generate height information like the Digital surface model (DSM). The acquisition conditions and measurement errors such as distance to sensor, the lighting conditions, occlusions in the scene (e.g. Tree obstructing a building), and not enough overlap between the images can also complicate unique feature matching (Qin, 2019). Object properties and their pattern in the scene can also increase the uncertainty of the generated height, such as thin structures, texture-less surfaces, featureless areas, and repeated patterns or structures directly affect stereo matching. All these errors can cause the height data to be temporally inconsistent and lead to holes, noise, missing data, blurry artifacts, and fuzzy edges and boundaries, hence, incomplete and unreliable representation of the 3D information.

To enhance low-quality depth maps (i.e. height map) generated from MVS algorithms, researchers suggest fusing several depth maps to utilize the redundant information in the temporal data. A common approach to fusing depth maps is the simple median filtering (Kuschk, 2013; Matyunin et al., 2011; Ozcanli et al., 2015; Qin, 2017). The median filter is preferred in many works related to depth refinement and processing due to its simplicity, and ability to eliminate outliers while preserving the details. Other methods to process fusion include global approaches such as Markov Random Field (MRF) and total variation (TV) which optimizes the solutions (Zhu et al., 2010; Liu et al., 2015 ; Kuschk & d'Angelo, 2013; Lasang et al., 2016; Kuschk, et al., 2017). However, they are mostly used to fused depth maps resulted from RGB-D images from Kinect or video scenes, and despite the effectiveness of these algorithms, there are still some limitations that strict their usages. One limitation is the necessity to acquire noise-free and clear set of images to improve stereo matching and the corresponding depth map, and unlike Kinect and video scenes where many images are captured indoor within a few milliseconds with consistent acquisition and lighting conditions, satellite images are more exposed to noise and outliers due to atmospheric conditions, seasonal variations, sun and satellite angles, image time, occlusions, etc. (Qin, 2019). Another issue is that current fusion algorithms are not adaptive to the scene objects. Urban features made of concrete and asphalt like buildings and roads are time-invariant, which means they experience a low rate of change in the temporal depth map. Vegetation, on the other hand, is time-variant, where they tend to change frequently during the year due to atmospheric conditions and seasonal changes. The time variance is not the only factor that should be considered while designing the fusion algorithm; the characteristics of the object should also be considered. For example, narrow objects like road and ground tend to be blocked by shadows, or trees, which increase their height uncertainty. The discrepancies in the height of objects in the depth maps produces a nonlinear type of change, which cannot be resolved directly using simple filters or fusion techniques that process all objects in the scene similarly. Therefore, in our work, we emphasize on the importance of analysing the type of class to develop an adaptive spatiotemporal fusion that processes each pixel based on its class stability.

* Corresponding author 2036 Neil Avenue, Columbus, Ohio, USA. qin.324@osu.edu

1.2 Related Works and Rationale

Generating high-quality depth maps using very high-resolution multi-view satellite images to improve 3D reconstruction model has been an ongoing research topic in the past few years. Multi-view Stereo (MVS) methods are widely used to extract elevations from multiple satellite views due to its efficiency and lower cost in comparison to direct methods like LiDAR. However, the depth map generated from MVS may be of bad quality and come with noises, outliers, and missing data due to the temporal and spectral inconsistencies between the images pair used to generate the depth map. In dense image matching the number of matched points between the image pair can play a major role in depth quality, if no or few points can be matched between the image pair due to other factors such object surface properties (e.g. smooth or texture-less surface, repeated patterns, etc.), it can produce inaccurate depth map.

A solution to recover the depth map generated from MVS algorithms is multi-view depth fusion, which has been explored by researches in two contexts either global or local approaches. The global approaches mainly involve optimization and energy function to minimize the losses and sometimes include smoothing and regularization terms as additional constraints. Markov Random Field (MRF) is one example that has been widely used in depth fusion. For instance, Zhu et al., (2010) and Liu et al.,(2015) both used spatiotemporal MRF to fuse depth maps by achieving temporal coherence. Weighted total variation (TV) and total generalized variation (TGV) methods are also popular approaches in global fusion of depths (Kuschk & d'Angelo, 2013; Lasang et al., 2016); Kuschk, et al., 2017). Neural networks are also effective fusion techniques that have been mostly used to fuse depth extracted from video scenes. Although global methods are useful and can provide accurate depth results, they are mostly applied to Kinect RGB-D sensors or video scenes, which in comparison to satellite images have an optimal indoor environment to capture numerous numbers of frames in a few milliseconds, therefore, provides lower distortions and better dense image matching. Additionally, background and foreground objects in the frames can be separated easily because of the fixed acquisition and lighting conditions, unlike satellite images where all objects are interrelated and difficult to extract directly. For depths generated from satellite images, local approaches are more popular, where fusion is performed mostly using local filtering. The most common fusion technique is the median filter because of its stability and robustness to outliers (Kuschk, 2013; Matyunin et al., 2011; Ozcanli et al., 2015; Qin, 2017). Other algorithms such as (Reinartz, Müller, Hoja, Lehner, & Schroeder, 2005) used average filtering to fuse DSMs obtained from stereo techniques using SPOT-5 and radar data obtained from SRTM, but because average filtering techniques smooth high-frequency data it tends to discard high level of details and generate outliers. Recent methods include median clustering which has been proposed in (Facciolo et al., 2017; Rumpler et al., 2013), where clustering is considered as an effective method to assess the temporal homogeneity of height data by measuring the inter- and intra-class similarity and dissimilarity. Local strategies are efficient in terms of time and robust to small outliers but are not able to solve the problems of extensive nonlinear noise and object boundaries. Moreover, the current filtering methods used in fusion often ignores objects class and process the image using fixed parameters. Since objects in the scene and nature have different responses and reflection properties in the captured satellite image, it is important to assess the elevation uncertainty for each class and incorporate it into the fusion algorithm.

1.3 Proposed Method and Rationale

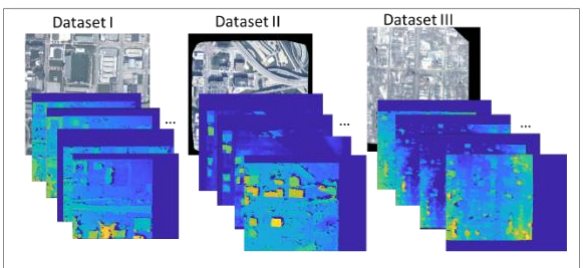
The spatiotemporal analysis is an effective way to solve problems related to data consistency, noise, missing data, outliers, etc. Using redundant data, we can fuse the depth maps to result in a reliable and accurate single depth map. Fusing depth maps using an adaptive spatiotemporal algorithm is an ongoing research topic where very few researches have investigated this area of research (Qin, 2017). Therefore, in our work, we aim to investigate and analyze the role of class to improve the multi-depth fusion algorithm that is adaptive to scene elements, efficient, robust, and able to recover the gaps mentioned in the literature.

This paper is organized in the following order, in Section 2, we will mention the methodology and analysis, where we will discuss the dataset used, the pre-processing methods, and the analysis that supports our proposed work. Section 3 includes the experimental results, with reasonable explanations and validations. Finally, the conclusion and future works will be addressed in Section 4.

2. METHODOLOGY AND ANALYSIS

2.1 Data Description and Pre-Processing

In our experiment, we chose 3 different datasets with varying complexities to examine and fuse; dataset I, is a commercial buildings area, dataset II is a more open area with natural objects such as vegetation and a water surface, and dataset III is a condensed housing area (for more details see figure 1).
We follow the same pre-processing method for all datasets, wherein each dataset we use VHR image pairs from Worldview 3 satellite to generate the corresponding multispectral orthophoto and the temporal DSMs. We use RSP (RPC Stereo Processor) software developed by (Qin, 2016) that performs hierarchical semi-global matching (SGM) algorithm (Hirschmuller, 2008) to generate and register the Orthophoto and the DSMs. We then generate the mask for each class using the 8-band multispectral orthophoto. We categories the classes into tree, grass, buildings, and a combined category for ground/road for all datasets. We use indices such as the Normalized difference vegetation index (NDVI) along with normalized DSM (nDSM) generated using top har reconstruction (Qin and Fang, 2014) to extract the masks. For instance, the NDVI helps to determine the locations of trees and grass, and with the appropriate nDSM we can find the position of trees based on their heights, and determine the location of buildings, ground, and road accordingly. For more details on the pre-processing steps, see the diagram in figure 2.



Dataset	I	II	III
Number of image pairs	400	51	51
DSM generated from good image pairs	219	37	26
Image and DSM size	1089x1113	1080x1109	1043x949

Figure 1. Details on the dataset.

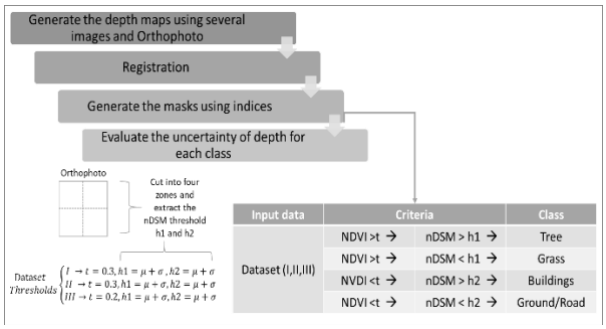


Figure 2. Pre-processing steps. Note h1, h2, and t are empirically determined based on the dataset.

2.2 Data Analysis

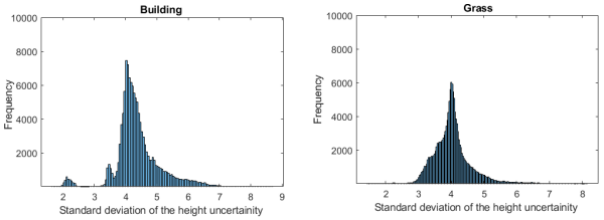
In our data analysis, we aim to investigate the stability and uncertainty of the height of objects and the way it varies according to the class. We categories the objects in the scene to either:

1) Time-invariant (e.g. buildings, ground/roads.) or time-variant (e.g. vegetation) objects.
2) Based on its surface properties (e.g. smooth texture-less, flat, etc.).

To assess the height of classes, we perform per-pixel processing by taking the standard deviation in the temporal DSMs of the DSMs for each class. We then summarize the analysis results by viewing the histogram of standard deviations to be able to see the distribution of every class and its average standard deviation (see Figure 3). From figure 3, we can see that each class has different standard deviation distribution. Most classes follow normal distribution, which can give us a clue about which type of probability to use in the weight measurement of our adaptive spatiotemporal fusion. The average standard deviation in Table 1 tells us the average uncertainty of height for each class. We can notice from Table 1 that for all three datasets vegetation (including trees and grass) has higher uncertainty than other classes, which means that height in the temporal DSM varies more than other objects. The buildings class comes in second to have higher height uncertainty. Among all classes, ground/ road appear to have better and higher elevation stability, where in all datasets it had the lowest values of uncertainty.

Class\ Dataset	I	II	III
Building	4.0968	8.8098	7.4692
Ground/ road	3.8197	8.5997	7.3950
Tree	**4.1097**	**8.9717**	**8.8059**
Grass	**4.6147**	**9.0144**	**9.6021**

Table 1. The standard deviation of the uncertainty per class (meter)



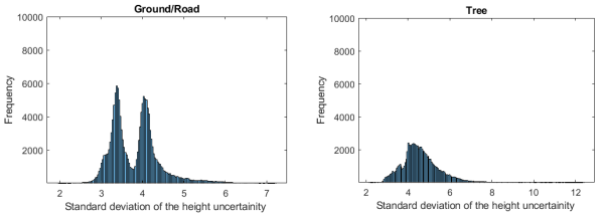


Figure 3. The distribution of the standard deviations of temporal heights in dataset I. The x-axis shows the standard deviation.

2.3 The Proposed Spatiotemporal Fusion Algorithm

Our fusion method is based on spatiotemporal filter; the generic formula for the fusion process is as follows

$$DSM_f(i,j) = \frac{1}{W_T} * \sum_{i=1}^{Width} \sum_{j=1}^{Height} W_r * W_s * W_h * h(i,j,t)_{med} \quad (1)$$

Where DSM_f= final fused pixel
i, j = location of pixels along the width and height
h_{med} = median height in the temporal DSMs
W_r= spectral weight
W_s= spatial weight
W_h= temporal height weight
W_T= total weight

Inspired from the basic median filter, where its purpose is to suppress noise and maintain the sharpness of edges, we extract the median height (h_{med}) from the temporal for each pixel. We choose the median as the base to get the final fused height value and to calculate the height weight as in the following equation

$$W_h(i,j) = exp^{\frac{-||hmed - h(i,j,t)||^2}{2\sigma_h^2}} \quad (2)$$

Where σ_h= the height bandwidth

The height bandwidth determines the extent of filtering in the temporal direction of the DSMs. σ_h is assigned for every pixel based on its class, and its value is decided empirically. With the prior knowledge of the extracted labeled image (mentioned in the pre-processing section 2.1.), we can determine which σ_h to use for each pixel.

$$\sigma_h = \begin{cases} \sigma_{Building} \rightarrow if\ pixel\ (i,j)\ is\ building \\ \sigma_{Ground/road} \rightarrow if\ pixel\ (i,j)\ is\ ground/road \\ \sigma_{tree} \rightarrow if\ pixel\ (i,j)\ is\ tree \\ \sigma_{grass} \rightarrow if\ pixel\ (i,j)\ is\ grass \\ \sigma_{water} \rightarrow if\ pixel\ (i,j)\ is\ water \end{cases} \quad (3)$$

We also analyze each DSM spatially, where spatial information is known to best reflect similarities in the neighborhood. To further improve the results, we use the spectral information in the orthophoto (to get the RGB image) and smooth the final DSM. Thus, we calculate the weights (W_r, W_s) in the formula as follows

$$W_r(i,j) = exp^{\frac{-||(I(i,j) - I(k,l)||^2}{2\sigma_r^2}} \quad (4)$$

$$W_s(i,j) = exp^{\frac{-((i-l)^2 + (i-k)^2)}{2\sigma_s^2}} \quad (5)$$

Where I = RGB image
i, j, l, k = the location of pixels in the window
σ_r = range bandwidth
σ_s = spatial bandwidth

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-3-2020-227-2020 | © Authors 2020. CC BY 4.0 License.

3. EXPERIMENTAL RESULTS

The results of our proposed spatiotemporal fusion are shown and discussed in this section. We will specify the parameters used and the outcomes for the three datasets. We validate our results per pixel, and against other existing methods.

3.1 Parameters

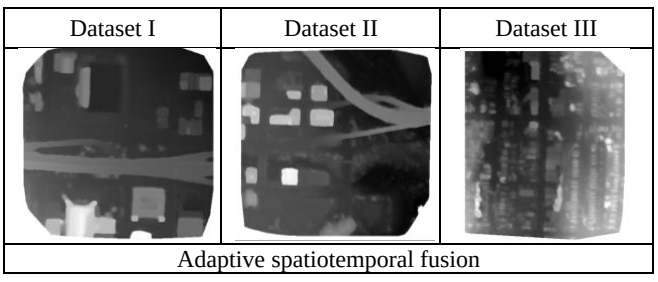
The two main parameters that we require in this work are the window size and bandwidths (spectral, spatial, and height), the choice of these parameters is made empirically. The window size is set and fixed to a moderate value of 7. Similarly, the spatial and spectral bandwidths (σ_r, σ_s) are set to (11, 50) (Tomasi and Manduchi, 1998). Validating the adaptive spatiotemporal fusion concept requires examining the bandwidth under different scenarios. For example, for objects with high uncertainty such as the time-variant objects like vegetation, we might need high σ_h. For flat, narrow, and featureless objects like ground and road, we might also need high σ_h . We also want to compare our adaptive approach to the fixed value of σ_h. We also chose the height bandwidth (σ_h) empirically, but in a manner that allows validating the assumption and the analysis made in section 2.2. Therefore, we chose σ_h based on several criteria indicated in Table 2.

σ_h	Notes
High for time-invariant objects	Give more emphasis to urban structure likes buildings and ground
High for time-variant and flat featureless objects	Give more emphasis to vegetation and ground/road
High for flat featureless objects	Give more emphasis to urban ground/road
Fixed for all classes	

Table 2. The choice categories for σ_h.

3.2 Results and Validations

The fused DSM for all datasets is shown in figure 4, where we can see that most of the distortions such as noise, missing datasets, and holes are filled and taken into consideration. We can notice that the adaptive spatiotemporal fusion algorithm produces good results visually, in comparison to other methods. For instance, if we compared it to the simple median filter we can see that median filter fusion results can produce overly smoothed outcomes, in addition to blurring some details. The adaptive median, on the other hand, is better in terms of capturing the details of the buildings since they use an adaptive window for buildings, but smaller details in ground are also overly smoothed. The C-median clustering can generate fuzzy and partially noisy results as in dataset II and III.



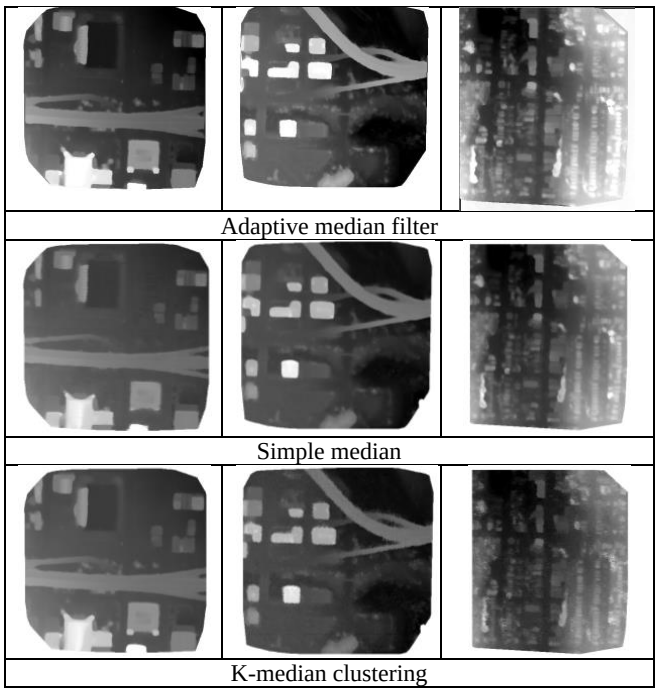


Figure 4. The fusion results of the three datasets using adaptive spatiotemporal fusion, the adaptive median filter, simple median filter, and C-median clustering.

We evaluate the results of the adaptive spatiotemporal fusion statistically by showing a comparison between different values of σ_h for all classes, and a comparison between other existing approaches. We compared the results of the proposed adaptive spatiotemporal fusion to the ground truth height data (from LiDAR) and measured the accuracy of each dataset to a 6-meter level of difference. The results are shown in tables 3, 4, and 5.

	sigma h				OA (%)	Accuracy per class (%)			
	T	GRS	B	GRD		T	GRS	B	GRD
Dataset I	11	11	17	13	95.4002	96.0384	94.3252	92.7269	95.6825
	13	13	15	12	95.4071	96.0834	94.3663	92.7449	95.6744
	15	18	15	14	95.4815	96.1677	94.3689	92.8167	95.7605
	15	15	13	14	95.4926	96.1677	94.3689	92.7808	95.7825
	23	27	20	23	95.5975	96.1396	94.3689	92.9065	95.9292
	30	33	23	27	**95.6154**	96.1115	94.3689	**92.9335**	**95.9556**
	18	18	17	21	95.5750	96.1283	**95.9045**	92.8078	94.3689
	23	28	33	30	95.5887	96.1396	94.3689	92.8975	95.9255
	15	15	15	15	95.5036	**96.1733**	94.3689	92.8167	95.7954

	sigma h				(OA) (%)	Accuracy per class (%)			
	T	GRS	B	GRD		T	GRS	B	GRD
Dataset II	13	13	17	12	99.0345	99.9246	99.5249	96.2123	99.2214
	23	28	33	30	99.0021	99.9113	99.6858	96.4428	99.4177
	13	13	17	17	98.97118	99.92902	99.29387	95.87279	99.42895
	15	15	11	13	**99.0925**	**99.9645**	99.6691	95.5644	99.4282
	18	18	15	17	99.0536	99.9601	99.6747	95.7015	99.4640
	37	37	35	35	98.9925	99.9379	99.7264	96.2497	99.4401
	18	18	17	23	98.9574	99.9556	99.5120	95.8292	**99.4700**
	17	33	35	27	98.9895	99.8314	**99.7412**	**96.5020**	99.3379
	15	15	15	15	99.0669	99.9601	99.5878	95.7856	99.4379

	sigma h				(OA) (%)	Accuracy per class (%)			
	T	GRS	B	GRD		T	GRS	B	GRD
Dataset III	11	11	17	13	99.9912	99.9558	99.8954	99.9873	99.9975
	13	13	15	11	99.9931	99.9853	99.9564	99.9873	99.9965
	23	28	33	30	99.9935	99.9890	99.9739	99.9873	99.9949
	15	15	13	14	99.9939	99.9890	99.9564	99.9848	99.9965
	18	18	15	17	99.9954	99.9890	99.9651	99.9873	99.9970
	21	27	20	23	99.9958	99.9890	99.9739	**99.9899**	99.9975
	15	18	15	21	99.9912	99.9484	99.8780	99.9873	99.9975
	30	30	25	35	**99.9962**	**99.9926**	**99.9826**	**99.9899**	**99.9975**
	15	15	15	15	99.9943	99.9890	99.9564	99.9848	99.9965

Table 3. Accuracy assessment of dataset II. Accuracy assessment of dataset I. Note: T=Tree, GRS=Grass, B=Building, GRD= Ground/Road, and OA = Overall Accuracy.

From the table 3, we note that that the highest overall accuracies (95.6254%, 99.0925%, and 99.9962%) for all datasets are located in the second category (explained in Table 2.), where greater values of weight W_h are given to objects with high elevation uncertainty. We also note that the adaptive approach provides slightly better results (0.02% higher) than the fixed bandwidth parameter (see last rows in table 3). The second part of the table explains how different classes correspond to different sigma values. We also show this in Figures 5, 6, and 7. For instance, we can notice from Figure 7 that trees and grass achieve the highest accuracies at large values of σ_h (30 and 35), while buildings require lower σ_h to achieve high accuracy at σ_h = 20 to 25. In all three datasets, grass always require larger σ_h, whereas urban objects such as buildings and ground can achieve high accuracies at moderate values of $\sigma_h \approx$ = 20 to 25. This can lead us to conclude that the height of time-variant objects are less certain and require larger compensation using higher σ_h. We can also conclude that fixed bandwidth parameters do not achieve optimal results; this can be seen from Figures 5, 6, and 7, and confirmed from the overall accuracy in table 3 at a fixed value of σ_h=15. We can use the information in these figures, to determine the optimal sigma values for each class and use it to get the optimal fused depth map. However, we can see that the optimal height bandwidth patterns between the classes differ with the dataset depending on the complexity and objects in the scene. For instance, for areas with few trees and many large commercial buildings as in dataset I, trees and grass required least σ_h of values 13 and 18 respectively, while, dominant objects like buildings and ground required larger σ_h of values 23 and 27 respectively. On the contrary, dataset II and III had trees and grass as dominant objects, thus, they require larger σ_h than the other classes.

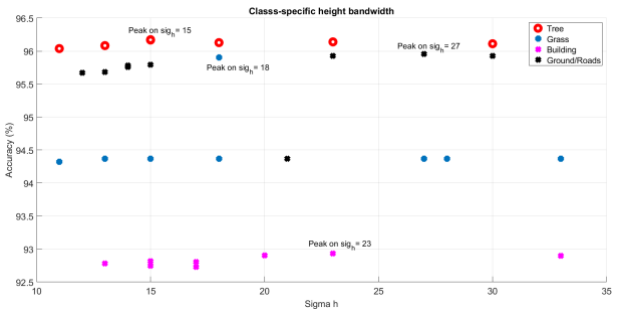


Figure 5. The Accuracy of each class in dataset I according to the height bandwidth.

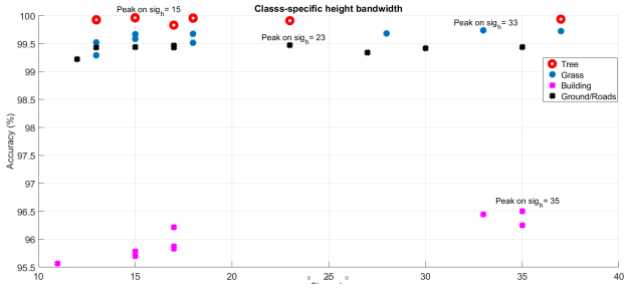


Figure 6. The Accuracy of each class in dataset II according to the height bandwidth.

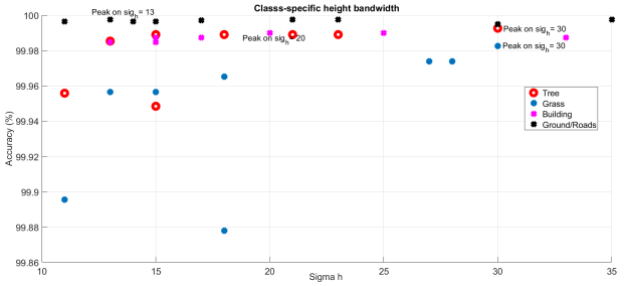


Figure 7. The Accuracy of each class in dataset III according to the height bandwidth.

We also use existing methods such as simple median filter, adaptive median filter by (Qin et al., 2017), and C-median clustering by (Facciolo et al., 2017) to evaluate our method (see table 4). We find that our adaptive method provides marginally higher accuracy with an increased range between almost 0.01-2%. Similarly, the accuracy of the majority of classes has higher accuracy in the adaptive case than the other methods with fixed bandwidths.

Dataset I: Accuracy per class (%)					
Method	(OA) (%)	B	GRD	GRS	T
ADPT_MED	95.4689	92.9335	95.7626	**94.3714**	95.9373
SIMP_MED	95.4468	**92.9604**	95.7088	94.3689	95.9485
C-MEDCLUST	94.6945	92.8347	95.0283	94.3047	92.5995
ADPT_STF	**95.6154**	92.9335	**95.9556**	94.3689	**96.1115**

Dataset II: Accuracy per class (%)					
Method	OA	B	GRD	GRS	T
ADPT_MED	98.7492	**95.6341**	99.1757	98.9671	99.9156
SIMP_MED	98.6427	89.0975	98.6440	94.8945	99.2275
C-MEDCLUST	95.9599	94.4836	97.0258	95.6869	96.8017
ADPT_STF	**99.0925**	95.5644	**99.6691**	**99.4282**	**99.9645**

Dataset III: Accuracy per class (%)					
Method	OA	B	GRD	GRS	T
ADPT_MED	99.6796	99.1686	98.6753	99.8710	99.3224
SIMP_MED	99.9874	98.7149	84.2601	99.5619	93.2096
KMEDCLUST	98.1956	99.0900	87.2494	98.6332	93.1286
ADPT_STF	**99.9962**	**99.9899**	**99.9826**	**99.9975**	**99.9926**

Table 4. Accuracy assessment for all datasets. Note: ADPT_MED= adaptive median filter, SIMP_MED= simple median filter, C-MEDCLUST= C-median clustering, and ADPT_STF= adaptive spatiotemporal fusion.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-3-2020-227-2020 | © Authors 2020. CC BY 4.0 License.

4. CONCLUSION AND FUTURE WORKS

In our work, we show that the adaptive spatiotemporal fusion technique can provide a better solution for objects with a high level of elevation uncertainty. The overall accuracy in all three datasets showed that optimal results could be achieved using a class-adaptive approach rather than the fixed parameter. Our analysis also shows that for classes with a high level of uncertainty like vegetation, more emphasis should be given by adjusting their height bandwidths to larger values. We also compare our results to existing work and found that it achieved slightly better overall accuracy ranging from 0.01 to 2%. In the next step, we will extend this work to determine the value of the height bandwidth automatically based on the scene information-using machine learning (ML) methods. We also would like to obtain the label image more efficiently using better nonparametric classification methods with indices such as NDVI, Morphology index, etc. to extract varying objects, their class and the corresponding classification map.

ACKNOWLEDGMENT

The authors would like to express their gratitude for the Johns Hopkins University Applied Physics Laboratory and IARPA and the IEEE GRSS Image Analysis and Data Fusion Technical Committee for the ONR data used in this work.

REFERENCES

Facciolo, G., Franchis, C.D., & Meinhardt, E. (2017). Automatic 3D Reconstruction from Multi-date Satellite Images. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1542-1551.

Hirschmuller, H. (2005). Accurate and Efficient Stereo Processing by Semi-Global Matching and Mutual Information. *Proceedings, 2*, 807-814.

Kuschk, G. (2013). Large Scale Urban Reconstruction from Remote Sensing Imagery. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. XL-5/W1. 139-146. 10.5194/isprsarchives-XL-5-W1-139-2013.

Kuschk, G., d'Angelo, P. (2013). Fusion of Multi-Resolution Digital Surface Models. *ISPRS - International Archives of the Photogrammetry*, Remote Sensing and Spatial Information Sciences. XL-1/W3. 247-251. 10.5194/isprsarchives-XL-1-W3-247-2013.

Kuschk, G., d'Angelo, P., Gaudrie, D., Reinartz, P., & Cremers, D. (2017). Spatially Regularized Fusion of Multiresolution Digital Surface Models. *IEEE Transactions on Geoscience and Remote Sensing, 55*, 1477-1488.

Liu, J., Li, C., Fan, X., & Wang, Z. (2015). Reliable Fusion of Stereo Matching and Depth Sensor for High Quality Dense Depth Maps. *Sensors*.

Lasang, P., Kumwilaisak, W., Liu, Y., & Shen, S. (2016). Optimal depth recovery using image guided TGV with depth confidence for high-quality view synthesis. *J. Visual Communication and Image Representation, 39*, 24-39.

Özcanli, Ö.C., Dong, Y., Mundy, J.L., Webb, H.F., Hammoud, R.I., & Tom, V. (2015). A comparison of stereo and multiview 3-D reconstruction using cross-sensor satellite imagery. *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 17-25.

Qin, R., (2016). RPC Stereo Processor (RSP) – A software package for digital surface model and orthophoto generation from satellite stereo imagery. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, III-1*, 77–82.

Qin, R., & Fang, W. (2014). A Hierarchical Building Detection Method for Very High Resolution Remotely Sensed Images Combined with DSM Using Graph Cut Optimization. *Photogrammetric Engineering & Remote Sensing, 80*(9), 873–883.

Qin, R., (2017), Automated 3d recovery from very high resolution multi-view satellite images, *ASPRS (IGTF) annual Conference*, Baltimore, Maryland, USA.

Qin, R., (2019) A Critical Analysis of Satellite Stereo Pairs for Digital Surface Model Generation and A Matching Quality Prediction Model. *ISPRS Journal of Photogrammetry and Remote Sensing*. 154 (2019).

Rumpler, M., Wendel, A., & Bischof, H. (2013). Probabilistic Range Image Integration for DSM and True-Orthophoto Generation. *SCIA*.

Tomasi, C., & Manduchi, R. (1998). Bilateral filtering for gray and color images. *In IEEE: Sixth international conference on computer vision (ICCV'98)*, pp. 839-846.

Zhu, J., Wang, L., Gao, J., & Yang, R. (2010). Spatial-Temporal Fusion for High Accuracy Depth Maps Using Dynamic MRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32, 899-909.