Efficient Joint Stereo Estimation and Land Usage Classification for Multiview Satellite Data

Ke Wang[1], Craig Stutts[2], Enrique Dunn[1], Jan-Michael Frahm[1]
[1]Department of Computer Science, The University of North Carolina at Chapel Hill
[2]Applied Research Associates, Inc.
kewang@cs.unc.edu, cstutts@ara.com, dunn@cs.unc.edu, jmf@cs.unc.edu

Abstract

*We propose an efficient algorithm to jointly estimate geometry and semantics for a given geographical region observed by multiple satellite images. Our joint estimation leverages an efficient PatchMatch inference framework defined over lattice discretization of the environment. Our cost function relies on the local planarity assumption to model scene geometry and neural network classification to determine semantic (*e.g. *land use) labels for geometric structures. By utilizing the commonly available direct (i.e. space to image) rational polynomial coefficients (RPC) satellite camera models, our approach effectively circumvents the need for estimating or refining inverse RPC models. Experiments illustrate both the computational efficiency and high quality scene geometry estimates attained by our approach for satellite imagery. To further illustrate the generality of our representation and inference framework, experiments on standard benchmarks for ground-level imagery are also included.*

1. Introduction

Recent advances in remote sensing techniques have powered multiple commercial vendors to provide global coverage of satellite images. Captured at multiple spectral bands and different resolutions, satellite images enable various applications, including navigation in online maps, geolocalization of photos [20], changes detection and planning in urban environment [19], ocean fluid motion estimation [16], and land usage analysis [23].

Establishing accurate relationships between the underlying 3D geometry and the 2D image observation is critical for extracting valuable information from such geo-spatial datasets. Since satellite images from different commercial vendors are usually captured by very different sensor platforms, formulating precise physical camera models to describe such geometric relationship is complex. Thus, the study of efficient and accurate methods to establish such correspondences, has drawn attention from both the computer vision community and the remote sensing community.

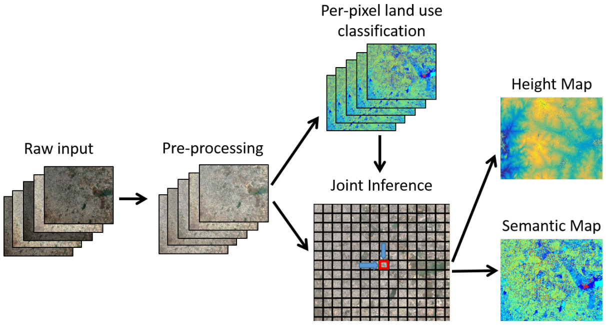

Figure 1. Schematic description of our pipeline. Pixel values of the given set of satellite images are first converted from native radiance to top-of-atmosphere reflectance. Land usage labels and converted images are used together in the MRF inference process to get refined geometric and semantic estimates.

The rational polynomial coefficients (RPC) camera model, first proposed by Hartley and Saxena [12], is capable of maintaining accuracy and sensor independence, and has become the standard abstraction for geometric modeling of satellite images. However, in practice, satellite image vendors may only provide a one-way space-to-image RPC mapping. Estimating the missing image-to-space mapping, is either inaccurate [27] or time-consuming [31]. The absence of accurate bidirectional mappings severely undermines the feasibility of many satellite images applications. Moreover, modern satellite platforms are usually capable of imaging at extremely high resolutions, while capturing multiple bands of the light spectrum. Panchromatic imagery can have ground sampling distances down to 0.2 meters, which depending on the sampled area can readily achieve pixel resolutions an order of magnitude greater than current consumer cameras.

In this paper, we propose an efficient dense multiple view stereo algorithm for satellite images, which solves the geometry and semantic estimation problem directly in the 3D space (see Figure 1 for an overview). Utilizing only the provided space-to-image mapping, our method effectively bypasses the limitation of a missing or inaccurate image-to-space mapping. By using an efficient inference framework, our proposed method can achieve state-of-the-art results in significantly less time. To summarize, our contributions are:

1. We propose to solve the dense stereo problem directly in a 2.5D volumetric representation of the 3D space, effectively by-passing the limitation of missing/inaccurate image-to-space model.
2. We propose an efficient PatchMatch based inference framework to address the high-resolution dense satellite stereo problem.
3. We demonstrate that utilizing semantic land usage information can help to improve the accuracy of dense stereo in satellite problems.

2. Related Work

Accurate and fast dense stereo algorithms have long been a focus in computer vision community. While much progress has been made to improve the accuracy [29], various efforts are also made to boost the efficiency of dense stereo methods. For example, Wang *et al*. [26] uses sequential sampling to reduce the disparity search space, thus saving computation overhead for high resolution satellite images. Barron *et al*. [4] solves stereo correspondence problems in the “bilateral space” to produce “defocus” images.

PatchMatch, initially proposed [2, 3] to solve for nearest neighbor field (NNF) in photo edit operations, shows great capabilities in minimizing complex unary energy formulations. Thus PatchMatch has been introduced to solve over-parameterized dense stereo correspondence [22] and optical flow problems [21]. Zheng *et al*. [30] extended PatchMatch to solve depth estimation problems in multiple view settings. Besse *et al*. [5] combined PatchMatch with Belief Propagation to explicitly control the smoothness in the output correspondence field. Unlike the Markov Chain model adopted in Zheng *et al*. [30], we follow Besse *et al*. [5] to formulate the multiple view stereo problem as a Markov Random Field (MRF) inference process.

A camera model must be known beforehand in order to solve dense stereo correspondence problems on satellite images. RPC camera models, calibrated either from physical sensor models, or using ground control points [15], provides a well-accepted form of approximation for satellite imaging process. Numeric instability and inaccuracy of RPC fitting process can severely undermine the latter stereo matching accuracy, requiring the use of bundle adjustment [8] or additional sensor characteristics information to be applied to further improve accuracy. With known camera models, different vision methods can be applied to satellite images. For example, Tao *et al*. [24] reconstruct the approximate 3D points iteratively given matched image features. Semi-global matching [13] is adopted by [7, 9] to extract height map or digital surface model.

Higher order information, such as image semantic labels, can also help solving low level vision problems. For example, by representing a 3D scene as a collections of objects, Bleyer *et al*. [6] solves stereo matching together with object segmentation. Wang *et al*. [25] refines coarse single image depth estimation within same semantic regions. Hane *et al*. [11] and Lubor *et al*. [17] both argue that augmenting stereo matching with semantic labeling can boost performance for both problems.

3. Methodology

Given no less than two satellite image observations for the region of interest, our goal is to establish dense correspondences across the reference view image and all the rest matching images, so that geometry and semantic information can be extracted on top of such correspondences.

3.1. Data Pre-processing

Solar illuminance geometry can greatly impact satellite return photon values at different wavelengths, allowing for satellite images captured at different times and illuminance conditions to have very different native radiance values. Thus, converting multiple-spectral images from native radiance values to top-of-atmosphere reflectance values can help increase stereo matching and land use classification accuracy [1]. Such conversion is described by Equation 1:

$$\rho_{Pixel,Band} = \frac{L_{Pixel,Band} \cdot d_{ES}^2 \cdot \pi}{E_{Band} \cdot \cos(\theta_s)} \tag{1}$$

where $\rho_{Pixel,Band}$ is the top-of-atmosphere reflectance per pixel per band, $L_{Pixel,Band}$ is the top-of-atmosphere band-averaged radiance, d_{ES} is the earth-sun distance, E_{Band} is the band-averaged solar spectral irradiance, and θ_s is the solar zenith angle.

3.2. Multiple View Stereo Formulation

We propose to jointly estimate the geometry structure and land usage information for a geographical region observed by multiple satellite images. Such a geographical region of interest is represented by a uniform vector grid $\{\mathbf{u}\}_{s=1}^{n}$. Each node corresponds to a geographical point in 3D space, with known latitude lat and longitude lon, but unknown altitude alt. Following the slanted surface formulation proposed in [22], we associate an unknown local 3D plane $\mathbf{f} = (a_f, b_f, c_f)$ to each geographical point (lat, lon, alt). The altitude of the 3D point satisfies:

$$alt = a_f * lat + b_f * lon + c_f \tag{2}$$

Once the 3D planes are estimated, 3D points lying on the plane can be projected onto a specific satellite image I by

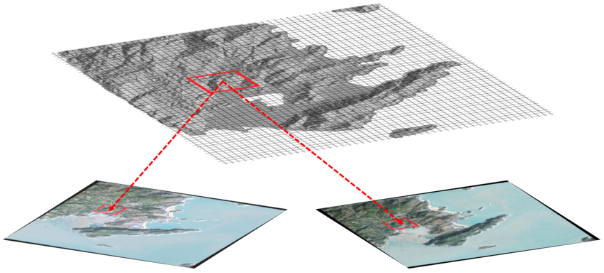

Figure 2. Visualization of the grid representation of our proposed method. Each grid point is parameterized by a local 3D plane and land usage information. Photo consistencies are accumulated across multiple images.

using the image specific RPC camera models (see Figure 2 for example). In addition to the geometry 3D plane, each node $\mathbf{u}_s$ also encodes land usage information for the given geographical 3D point.

Under our formulation, our goal of extracting geometry structure can be achieved by finding the optimal parameterization for each of the vector grid points that minimizes photometric and semantic observation conflicts between the designated reference view r and matching views, while maintaining local smoothness in the underlying geometry structures and semantic representations. Pixel-level view selection has been proved successful in increasing stereo accuracy and robustness in [30]. In order to account for the drastic view and capture condition changes across multiple satellite images, we include a similar pixel-level view selection scheme in our formulation.

The state of each grid point $\mathbf{u}_s = (\mathbf{f}_s, v_s, l_s)$ encodes information for an unknown local 3D plane $\mathbf{f}_s \in \mathbb{R}^3$, a unknown matching view index v_s, and an unknown land usage categories label l_s. Thus our goal is to optimize the following energy function:

$$E(\mathbf{u}_1, \ldots, \mathbf{u}_n) = \sum_{s=1}^{n} \psi_s(\mathbf{u}_s) + \beta_1 \sum_{s=1}^{n} \left[\sum_{t \in N(s)} \psi_{st}(\mathbf{u}_s, \mathbf{u}_t) \right] \tag{3}$$

where $N(s)$ being the pairwise neighborhood of node s. The unary energy $\psi_s(\mathbf{u}_s)$ measures consistency of image appearance and semantic labeling of the associated 3D plane between different image views. To compute the unary energy for a node $\mathbf{u}_s$, we first collect a set of nearby 3D points P_s lying on its local 3D plane. Each 3D point $p_s \in P_s$, is projected onto the reference view r and the selected match view v_s. The obtained image pixel color/intensity (I_r, I_{v_s}) and image pixel labeling (L_r, L_{v_s}) is compared to build the following cost:

$$C(p_s) = \min(\tau_c, \|I_r - I_{v_s}\|) + \beta_2 \delta(L_r, L_{v_s}) \tag{4}$$

where τ_c is truncation threshold for stereo cost, $\|I_r - I_{v_s}\|$ is the L_1 color difference, and δ is a Kronecker delta function defined as:

$$\delta(L_r, L_{v_s}) = \begin{cases} 0 & \text{if } L_r = L_{v_s} \\ 1 & \text{if } L_r \neq L_{v_s} \end{cases} \tag{5}$$

The unary data term is then defined over the 3D point set P_s:

$$\psi_s(\mathbf{u}_s) = \sum_{p_s \in P_s} \omega(p, p_s) \left[C(p_s) + \gamma \mathcal{L}(l_s) \right] \tag{6}$$

where $\omega(p, p_s)$ is the adaptive weight between center point p and neighboring 3D point p_s [28], and $\mathcal{L}(l_s)$ is the cost of classifying node $\mathbf{u}_s$ as label l_s, which can be obtained through classifier output (see Section 3.3).

On the other hand, the pairwise term $\psi_{st}(\mathbf{u}_s, \mathbf{u}_t)$ explicitly considers smoothness between adjacent 3D planes, matching view selections, and semantic label assignment:

$$\begin{aligned} \psi_{st}(\mathbf{u}_s, \mathbf{u}_t) =& (|\mathbf{n}_s \cdot (\mathbf{x}_s - \mathbf{x}_t)| + |\mathbf{n}_t \cdot (\mathbf{x}_t - \mathbf{x}_s)|) \\ &+ \omega_1 \delta(v_s, v_t) + \omega_2 \delta(l_s, l_t) \end{aligned} \tag{7}$$

where $\mathbf{n}_s$ represents the unit normal vector of plane $\mathbf{f}_s$, and $\mathbf{x}_s$ is a point on the plane $\mathbf{f}_s$. The smoothness term $\psi_{st}(\mathbf{u}_s, \mathbf{u}_t)$ will have a zero cost value if and only if two nodes lie on the same plane, have the same semantic label and match view selection. In our experiments, only selecting best view can better handle occlusion. In case multiple equally good views exists for a given MRF node, the matching view that's more consistent with adjacent nodes would be selected via MRF inference.

3.3. Land Use Classification

Captured at multiple spectral bands, pixel values from satellite images naturally carry physical response information from different land types, thus providing strong clues on land usage information. Thus, we propose to use a simple three-layer fully connected neural network to perform pixel-wise land usage classification based on pixel characteristics.

The land usage classification neural network consists of an input layer with 8 neurons to take input from the 8 channel pixel values, 40 neurons for the hidden layer, and 13 output neurons in the output layer. The neural network output is inverted and serves as the labeling cost term $\mathcal{L}(l_s)$ in the unary energy function (see Equation 6).

3.4. Inference

Our proposed MRF formulation in Equation 3 can be efficiently solved using the PatchMatch Belief Propagation method (PBMP) [5]. We initialize each node $\mathbf{u}_s$ with a random plane $\mathbf{f}_s$, a random semantic label l_s and a random match view choice v_s. MRF is then inferred through spatial-propagation and local resampling. We empirically set the

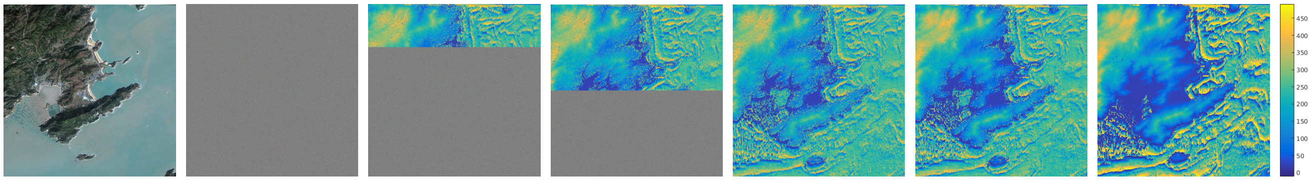

Image Initialization 1/4 iteration 1/2 iteration 1 iteration 3 iteration 10 iteration

Figure 3. Qualitative illustration of convergence. We show the height map obtained at different stages of the optimization process. The first iteration already extracted coarse geometrical structures from the random initializations. Results obtained after 3 iterations are visually very close to the final results with 10 iterations.

number of particles at each node $\mathbf{u}_s$ to 3, which achieved good balance between computation overhead and converging speed.

We show a quantitatively illustration of inference process in Figure 4, and a qualitative visualization in Figure 3.

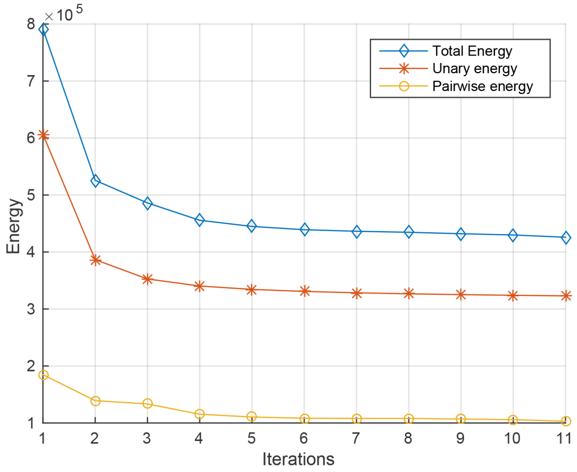

Figure 4. An example of PMBP inference on the Xiapu dataset (See Table 5 for dataset details). PMBP solver shows fast convergence speed in practice.

4. Experiments

4.1. Implementation

We evaluate our proposed method in a parallel C++ implementation. All the benchmark numbers are collected on a 32 core Intel Xeon CPU running at 2GHz. We use 5×5 patches for stereo matching. PMBP solver is run for 3 iterations with 3 particles at each node. Parameters are set as $\beta_1 = 0.4, \beta_2 = 15.0, \tau_c = 30.0, \omega_1 = 0.2, \omega_2 = 0.2, \lambda = 20.0$.

4.2. Ground Level Stereo Experiments

To justify the incorporation of semantic information in dense stereo problems, we quantitatively evaluate our proposed formulation of joint dense stereo and semantic object classification on the Leuven dataset [17] and the Kitti dataset [10]. Both datasets provide rectified stereo image pairs, simplifying the correspondence search space to horizontal lines. Accordingly, we define a corresponding grid node over each pixel in the reference image.

Table 1. Stereo accuracy on Leuven dataset. Out-All: percentage of erroneous pixels in total; Avg-All: average disparity in total.

	ALE [17]	PMBP Stereo	Joint PMBP
Out-All	0.3766%	0.2930%	**0.2714**%
Avg-All	5.1849px	3.8885px	**3.5608**px

Table 2. Stereo experimentation results computed at 3 pixel error threshold. Outperforming comparison baselines, our proposed joint estimation method shows close stereo accuracy to the state-of-art results on Kitti benchmark list. (As of August 2015). Out-Noc: percentage of erroneous pixels in non-occluded areas; Out-All: percentage of erroneous pixels in total; Avg-Noc: average disparity in non-occluded areas; Avg-All: average disparity in total.

Method	Out-Noc	Out-All	Avg-Noc	Avg-All
ALE [17]	5.27%	8.48%	1.4607px	1.9871px
PMBP Stereo	4.64%	5.96%	1.1820px	1.6693px
Joint PMBP	**3.82%**	**5.69%**	**0.9932**px	**1.2650**px

On both datasets, we compare our method against one of the state-of-art joint estimation framework, Automatic Labelling Environment (ALE) [17]. Specially, we used the ground-truth semantic labels from [18] to train the ALE classifier on the Kitti dataset.[1] We directly use the semantic classification output as the semantic label initialization of our MRF formulation. Quantitative evaluations on semantic classification accuracy on two datasets can be found in Table 3 and Table 4 respectively.

Though quantitative improvements upon semantic classifications being modest, Table 1 and Table 2 show that stereo matching can benefit from semantic regularization, as our proposed method outperforms ALE and pure PatchMatch stereo method. We attribute the insignificant semantic improvements to the highly accurate initialization with over 99% accuracy.

[1] Notice we perform quantitatively evaluation on this subset of labeled stereo image pairs provided in [18], instead of the original Kitti stereo dataset.

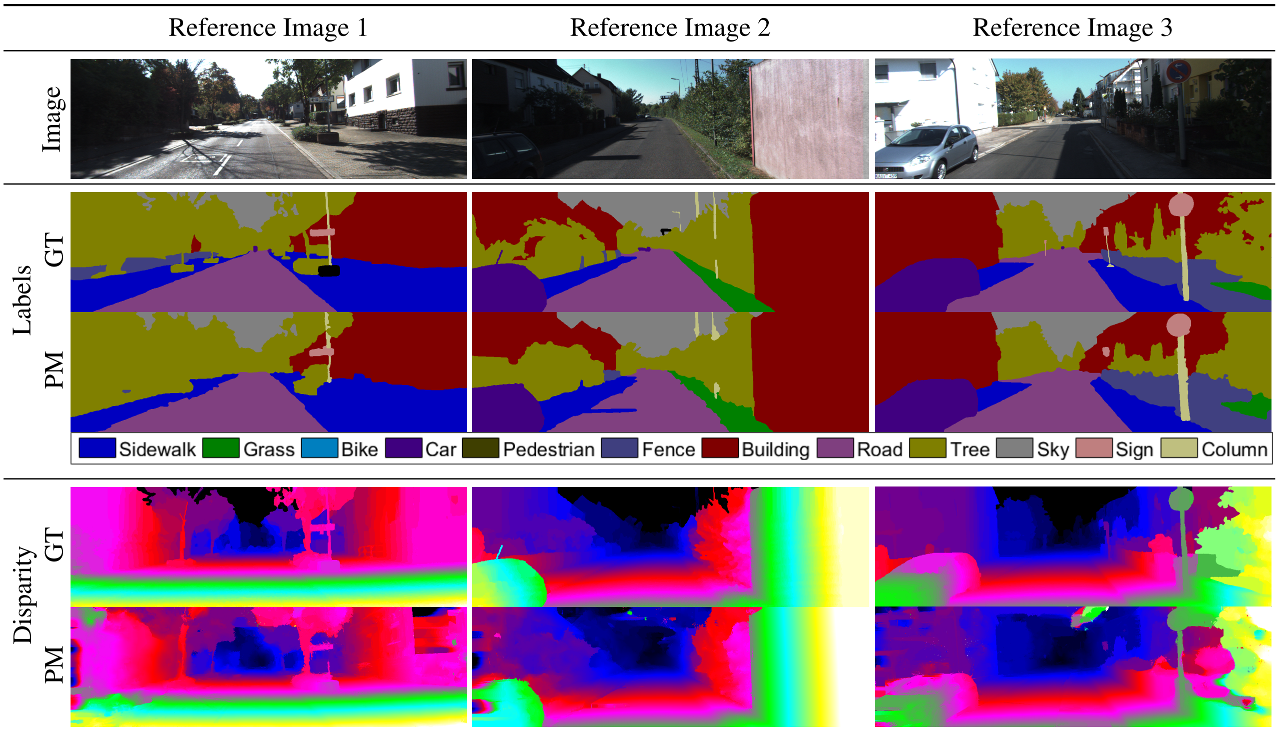

Figure 5. Visualization of semantic and stereo results of our proposed method on the Kitti dataset [10]. Through joint-optimization, clear boundaries are preserved in semantic segmentations, while stereo ambiguity can be resolved through high-order semantic information.

Table 3. Semantic classification accuracy on the Leuven test dataset. Seven semantic categories are defined for typical outdoor scenes. Despite being simple, our method can slightly improve upon state-of-the-art by simply smoothing the classification results together with stereo disparity maps.

Method	All	Pavement	Person	Bike	Car	Building	Road	Sky
ALE[17]	0.9948	**0.6118**	0	0.6765	0.9042	0.9729	0.9885	0.9967
Joint PMBP	**0.9949**	0.6110	0	**0.6767**	**0.9046**	**0.9733**	**0.9887**	**0.9968**

4.3. Satellite Images Experiments

4.3.1 Datasets

The possible drastic change in solar illuminance, weather, and atmosphere conditions for a designated region, poses great challenges for the dense stereo problem on satellite images. To counter such difficulties, commercial satellite image vendors provide multiple image captures within the same orbit pass over a given region of interest. Taken within very short time-intervals, such one-pass captures simplify stereo matching by providing similar illuminance conditions and image appearances.

In order to show the robustness and effectiveness of our proposed satellite stereo method, we collect multi-pass satellite image datasets, with images taken at different time and on different orbits (see Table 5 for details). Compared to one-pass captures, such datasets have higher availability, but contain greater variety in viewing angles, solar illuminance, and scene contents, thus posing greater challenges for the stereo matching solver. For example in Figure 6, the Zarqa dataset is captured within a time span of 3 years.

Without loss of generality, we used multi-spectral images from WorldView-2 satellite sensors. Such images have ground sampling distance as low as 0.2 meters. Eight spectral bands provide spectrum coverage from optical wavelength to infrared wavelength. The 11 bits dynamic range also enriches the discrimination between different image pixels, which can potentially improve stereo results. Example results of our joint estimation can be found in Figure 6.

4.3.2 Land Usage Classification

In order to train the land usage classification neural network, 12,000 samples are manually collected for each of the 13 semantic classes. A random split of 70% is used for training, 15% for cross-validation, and 15% for testing. The three layer neural network is trained by scaled gradient back-propagation algorithm.

A simple per-pixel classification can be obtained by se-

Table 4. Kitti semantic classification accuracy evaluation.

Class	Overall	Sidewalk	Grass	Bike	Car	Pedestrian	Fence	Building	Road	Tree	Sky	Sign	Column
ALE [17]	0.9370	0.9683	0.9520	0.9645	0.9081	0.9364	0.9427	0.9569	0.9280	0.9709	0.9036	**0.4770**	**0.0630**
Joint PMBP	**0.9378**	**0.9694**	0.9520	**0.9653**	**0.9132**	**0.9377**	**0.9433**	**0.9576**	**0.9289**	**0.9721**	**0.9050**	0.4649	0.0622

Table 5. Quantitative evaluation on multiple satellite datasets. Empirically, our proposed method expands linearly in terms of memory and computation time.

	Dataset	Resolution	Height range	Pass	View	SGM	BidPMBP	GridPMBP
Time	Xiapu	26.2 Mpx	500 m	1	2	0.48 Hours	0.57 Hours	**0.18** Hours
	Bengaluru	66.1 Mpx	1000 m	5	5	9.73 Hours	5.75 Hours	**1.78** Hours
	Zarqa	58.7 Mpx	1000 m	7	7	12.96 Hours	7.66 Hours	**2.37** Hours
Memory	Xiapu	26.2 Mpx	500 m	1	2	1.8 GB	1.4 GB	**1.4** GB
	Bengaluru	66.1 Mpx	1000 m	5	5	11.3 GB	8.8 GB	**8.8** GB
	Zarqa	58.7 Mpx	1000 m	7	7	12.0 GB	9.4 GB	**9.4** GB

lecting the most likely semantic label for each pixel. Such maximum likelihood classification can lead to noisy semantic maps, as shown in Figure 7. By combining semantic classification with stereo correspondence estimation, we can effectively smooth the semantic map, leading to less noisy predictions.

4.3.3 Dense Stereo Benchmark

To evaluate our performance on dense stereo estimation tasks, we apply traditional image-space multiple-view stereo methods on our testing satellite image dataset as baselines.

As discussed in Section 2, commercial satellite images usually either lack image-to-3D RPC camera model, or suffer from inaccuracies issues. We incorporated the minimal solver from [31] to establish accurate bi-directional correspondences between 3D points and pixels. We embedded this minimal solver into traditional image space dense stereo solvers to ensure their functionality on satellite images.

Semi-Global matching (SGM) approximates a globally optimal solution for dense stereo problems by aggregating pixel-wise matching cost from multiple directions. SGM has been proved robust and successful for reconstructions in both standard pin-hole camera images, as well as aerial/satellite images [7, 9].

The original SGM method proposed in [13] needs to store the entire matching cost volume for multi-direction cost aggregation. For a reference image with width D, height H, and disparity range D, such $O(WHD)$ memory requirement can hardly be fulfilled for high-resolution satellite images. So we adopted an memory efficient variant of the original SGM [14], which decreases the memory complexity to $O(kWH)$ where k is the number of aggregation directions.

We also compare our proposed method to standard image-space multiple view PatchMatch stereo method. In this case, we define each node $\mathbf{u}_s$ over each pixel instead of a 3D space grid point. The unary and pairwise energy naturally follows the formulation in Section 3.2. We run bidirectional PatchMatch (BidPMBP) stereo on satellite image datasets with same patch radius and optimization iterations.

A detailed comparison on memory usage and run time can be found in Table 5. Our proposed method (GridPMBP) achieved on average 3.2X speedup against bidirectional PatchMatch, and at least 5 times faster than SGM. Since the PatchMatch based method smartly traverse the disparity/altitude search space without iterating through the entire cost volume, our proposed method inherits the lower memory requirement. Both BidPMBP and GridPMBP shows less memory usage than SGM. Qualitative zoom-in comparisons for three methods can be found in Figure 8.

As Figure 8 shows, our datasets covers different terrain types, including mountainous and urban areas. Compared with SGM, GridPMBP accurately captured finer details of the terrain shape on different terrain types, demonstrating great robustness and application ability. GridPMBP also achieved good stereo results on both one-pass stereo captures, and multi-pass captures. Thus our proposed method enables the use of satellite images captured from multiple-passes for accurate dense stereo problem.

4.4. Discussion

Overall, our proposed method achieved good results on satellite image datasets, as well as on standard stereo benchmark datasets. Especially, our proposed method has lower memory footprint and higher computation efficiency, which is more suitable for multi-view high resolution satellite image applications.

In our MRF formulation, the unary term $\psi_s(\mathbf{u}_s)$ is only evaluated between the reference view r and the per-node

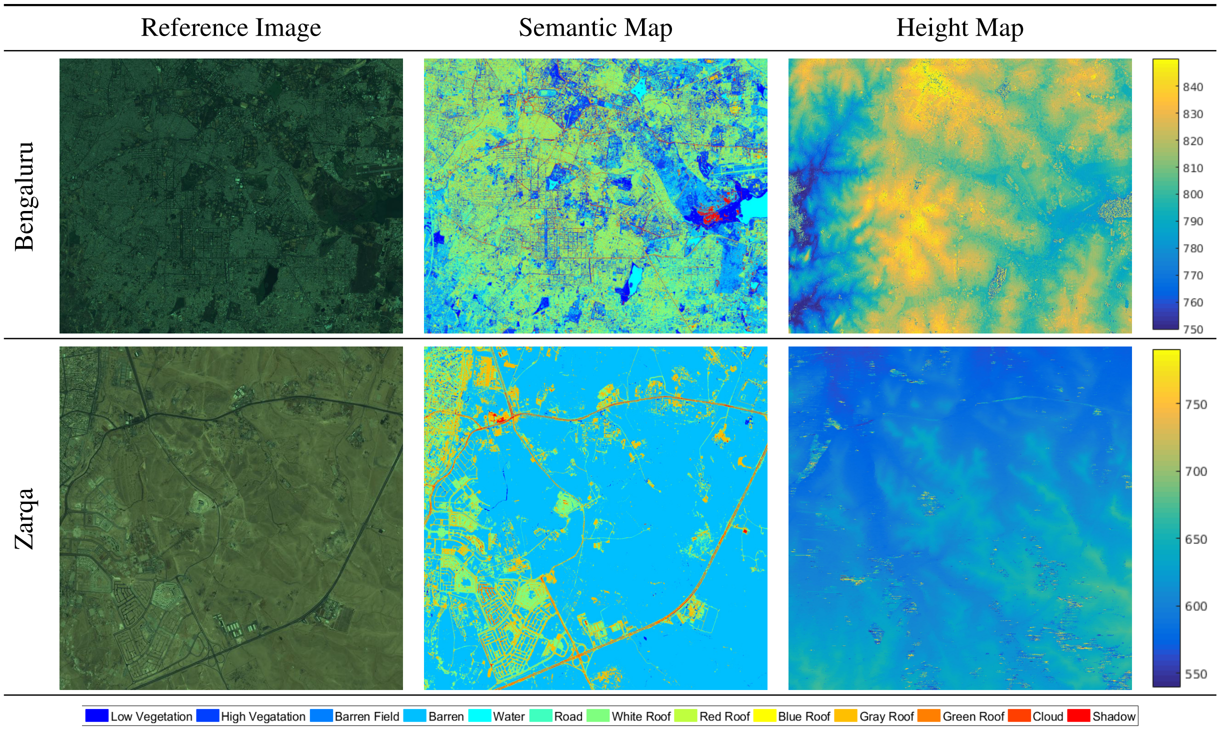

Figure 6. Example results obtained via GridPMBP solver on selected datasets. (Best view in color.)

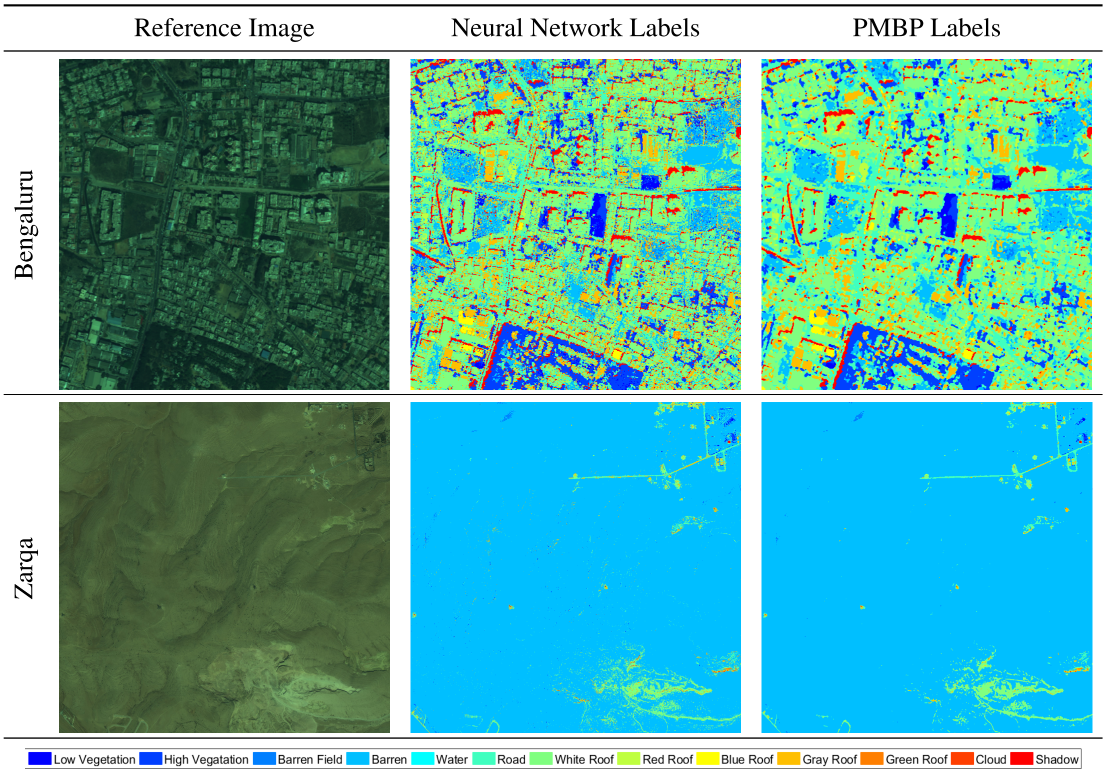

Figure 7. Comparison of semantic labels. Per-pixel based classification can lead to noisy semantic labels, see column 2. By joining dense stereo estimation together with per-pixel classifications, noisy semantic maps can be smoothed, see column 3. (Best view in color.)

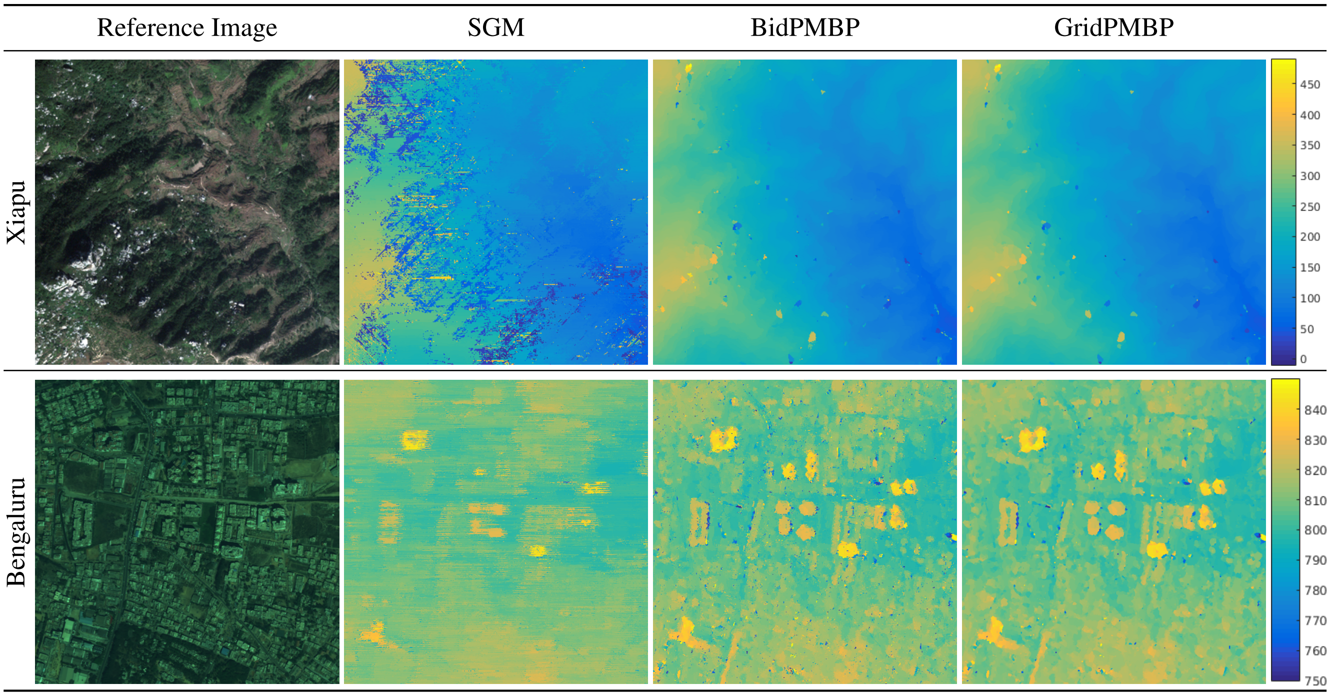

Figure 8. Zoom in comparisons of satellite height map estimation results. GridPMBP provides cleaner results with much smaller computation and memory overhead. (Best view in color.)

selected match view v_s, thus our proposed method has a low complexity with matching view numbers. Increasing dataset cardinality won't lead to catastrophic runtime increase. Also, the piecewise smoothness in a typical image domain, and spatial propagation scheme employed by PatchMatch based methods, greatly amortizes the optimization burden amongst neighboring pixels. This also make the optimization of a complex state space and unary cost function feasible. Last but not least, by using a grid space representation, we effectively by-pass the limitation of missing image-to-space RPC models, and thus save the overhead of computing such inverse mapping on-the-fly.

As can be seen in Figure 7, pixel-based local classification can be error-prone. However, our PMBP solver is independent from the methods used to attain these labels (though we do require a confidence value for each label). Hence, improved classification results can be easily integrated for better results. We consider this as part of the future work.

By adopting the space grid representation, our method has advantages that are not seen in the baseline methods:

1. Compared with traditional multiple-view-stereo pipeline, which first establishes dense correspondences to extract depth/height maps, then does fusion to extract dense geometry, our space grid representation directly optimizes for 3D geometry structure. Huge computation overhead can be thus saved in later fusion stage.
2. Depth maps can be obtained by projecting the estimated geometry structure to the desired view. Our space-based representation also create possibilities for virtual view synthesis.
3. The spatial resolution of the MRF grids can be changed to achieve a good balance between geometry fine-details and computation resource budget.

5. Conclusion

To summarize, we propose an efficient method to estimate geometric and semantic information for a given geographical region. By directly using a grid representation in 3D space, we by-pass the limitation of missing RPC camera models, and entirely avoid the inaccuracy and computation overhead caused by fitting the camera model.

To further speed up the estimation, we plan to port the implementation onto GPUs. Also, in order to increase the semantic parsing accuracy, we can try to use convolutional neural network (CNN) features for land usage classification [23].

Acknowledgement: Supported by the Intelligence Advance Research Projects Activity (IARPA) via Air Force Research Laboratory (AFRL), contract FA8650-12-C-7214. The U.S Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government.

References

[1] Radiometric use of worldview-2 imagery. http://www.digitalglobe.com/sites/default/files/Radiometric_Use_of_WorldView-2_Imagery%20(1).pdf. 2

[2] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 2009. 2

[3] C. Barnes, E. Shechtman, D. B. Goldman, and A. Finkelstein. The generalized patchmatch correspondence algorithm. In *ECCV*. 2010. 2

[4] J. T. Barron, A. Adams, Y. Shih, and C. Hernandez. Fast bilateral-space stereo for synthetic defocus. In *CVPR*, 2015. 2

[5] F. Besse, C. Rother, A. Fitzgibbon, and J. Kautz. Pmbp: Patchmatch belief propagation for correspondence field estimation. *IJCV*, 2013. 2, 3

[6] M. Bleyer, C. Rother, P. Kohli, D. Scharstein, and S. Sinha. Object stereo - joint stereo matching and object segmentation. In *CVPR*, 2011. 2

[7] P. dÁngelo and G. Kuschk. Dense multi-view stereo from satellite imagery. In *IGARSS*, 2012. 2, 6

[8] G. Dial and J. Grodecki. Block adjustment with rational polynomial camera models. In *ISPRS*, 2002. 2

[9] S. Gehrke, K. Morin, M. Downey, N. Boehrer, and T. Fuchs. Semi-global matching: An alternative to lidar for dsm generation. In *Proceedings of the 2010 Canadian Geomatics Conference and Symposium of Commission I*, 2010. 2, 6

[10] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 4, 5

[11] C. Hane, C. Zach, A. Cohen, R. Angst, and M. Pollefeys. Joint 3d scene reconstruction and class segmentation. In *CVPR*, 2013. 2

[12] R. Hartley and T. Saxena. The cubic rational polynomial camera model. In *Image Understanding Workshop*, 1997. 1

[13] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *TPAMI*, 2008. 2, 6

[14] H. Hirschmüller, M. Buder, and I. Ernst. Memory efficient semi-global matching. *ISPRS*, 2012. 6

[15] Y. Hu and C. Tao. Updating solutions of the rational function model using additional control information. *Photogrammetric engineering and remote sensing*, 2002. 2

[16] T. Isambert, J.-P. Berroir, and I. Herlin. A multi-scale vector spline method for estimating the fluids motion on satellite images. In *ECCV*. 2008. 1

[17] L. Ladick, P. Sturgess, C. Russell, S. Sengupta, Y. Bastanlar, W. Clocksin, and P. H. S. Torr. Joint Optimization for Object Class Segmentation and Dense Stereo Reconstruction. *IJCV*, 2011. 2, 4, 5, 6

[18] L. Ladicky, J. Shi, and M. Pollefeys. Pulling things out of perspective. In *CVPR*, 2014. 4

[19] W. Li, X. Li, Y. Wu, and Z. Hu. A novel framework for urban change detection using vhr satellite images. In *ICPR*, 2006. 1

[20] T.-Y. Lin, S. Belongie, and J. Hays. Cross-view image geolocalization. In *CVPR*, 2013. 1

[21] J. Lu, H. Yang, D. Min, and M. Do. Patch match filter: Efficient edge aware filtering meets randomized search for fast correspondence field estimation. In *CVPR*, 2013. 2

[22] C. R. Michael Bleyer and C. Rother. Patchmatch stereo - stereo matching with slanted support windows. In *BMVC*, 2011. 2

[23] O. A. B. Penatti, K. Nogueira, and J. A. dos Santos. Do Deep Features Generalize From Everyday Objects to Remote Sensing and Aerial Scenes Domains? In *CVPR Workshop*, 2015. 1, 8

[24] C. V. Tao and H. Yong. 3d reconstruction methods based on the rational function model. *Photogrammetric engineering and remote sensing*, 2002. 2

[25] K. Wang, E. Dunn, J. Tighe, and J.-M. Frahm. Combining semantic scene priors and haze removal for single image depth estimation. In *WACV*, 2014. 2

[26] Y. Wang, K. Wang, E. Dunn, and J.-M. Frahm. Stereo under sequential optimal sampling: A statistical analysis framework for search space reduction. In *CVPR*, 2014. 2

[27] X. Yang. Accuracy of rational function approximation in photogrammetry. In *ISPRS*, 2000. 1

[28] K.-J. Yoon and I. S. Kweon. Adaptive support-weight approach for correspondence search. *TPAMI*, 2006. 3

[29] J. Zbontar and Y. LeCun. Computing the Stereo Matching Cost With a Convolutional Neural Network. In *CVPR*, 2015. 2

[30] E. Zheng, E. Dunn, V. Jojic, and J.-M. Frahm. Patchmatch based joint view selection and depthmap estimation. In *CVPR*, 2014. 2, 3

[31] E. Zheng, K. Wang, E. Dunn, and J.-M. Frahm. Minimal solvers for 3d geometry from satellite imagery. In *ICCV*, 2015. 1, 6