STATE OF THE ART IN DENSE IMAGE MATCHING COST COMPUTATION FOR HIGH-RESOLUTION SATELLITE STEREO

Yue Wang [1], Danchao Gong [2], Hualong Hu [3], Shugen Wang [3], Yilong Han [3, *],Yuan Wang [4], Xiaoliang Ma [5]

[1] Wuhan Geomatics Institute, Wuhan, China - ywangwhu@163.com
[2] State Key Laboratory of Geo-information Engineering, Xi'an, China - sx_gdch@sina.com
[3] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China - (huhualong, wangsg, hanyl)@whu.edu.cn
[4] School of Geographic and Oceanographic Science, Nanjing University, Nanjing, China - wangyuan1204@smail.nju.edu.cn
[5] Xi'an Geovis Spatial Data Technology Co., Ltd, Xi'an, China - maxl@geovis.com.cn

Commission II, WG II /2

KEYWORDS: High-resolution Satellite Images, Dense Image Matching, Cost Computation, Disparity, Semi-global Matching, Census Transform, Convolutional Neural Network

ABSTRACT:

Large-scale Digital Surface Model (DSM) generated with high-resolution satellite images (HRSI) are comparable, cheaper, and more accessible when comparing to Light Detection and Ranging (LiDAR) data and aerial remotely sensed images. Several photogrammetric commercial/open-source software packages are being developed for satellite image-based 3D reconstruction, in which, most of them adopt a modified version of Semi-Global Matching (SGM) algorithm for dense image matching. With the continuous development of matching cost computation methods, the existing methods can be divided into classical (low-level) and learning-based algorithms (non-end-to-end learning and end-to-end learning methods). On Middlebury and KITTI datasets, learning-based algorithms has shown their superiority compared to SGM derived methods. In this context, we assume that matching cost is the key factor of DIM. This paper reviews and evaluates Census Transform, and MC-CNN on a WorldView-3 typical city scene satellite stereo images on the premise that the overall SGM framework remains unchanged, providing a preliminary comparison for academic and industrial. We first compute the cost valume of these two methods, obtains the final DSM after semi-global optimization, and compares their gemetric accuracy with the corresponding LiDAR derived ground truth. We presented our comparison and findings in the experimental section.

1. INTRODUCTION

Everything moves. However, with the development of photogrammetry and computer vision, we can achieve roaming in the four-dimensional world (three-dimensional physical world with the time dimension). Benefit from the revolution of satellite sensors and orbit revisit technology, more and more high-resolution Earth Observation Satellites have been launched, such as the WorldView series from the United States, the Pléiades series from France, SuperView-1, and the Gaofen series from China. Dense image matching (DIM) leverage the growing numbers of high-resolution satellite images to generate large-scale usable 2.5/3D/4D (3D with the time dimension) the Earth's surface models, which fueling remote sensing applications in a variety of domains, such as ecological monitoring, city-scale 3D/4D modeling, urban planning and monitoring, and navigation (Gruen, 2012; Haala & Kada, 2010).

Dense image matching (DIM) algorithms aim to find the pixel-wise correspondences between two or more epipolar images. Matching costs evaluate the similarity of two pixels in the stereo images, which is the very first and important step for dense image matching. The result after matching cost computation is cost volumes, where the z-axis contains the cost value over the whole disparity range, while the x and y-axis depict the image coordinates. Given the fact that the geometric and radiation condition changes exist in the considered stereo, which makes the images no longer meet the luminosity consistency condition, and the complexity of the features of the Earth's surface brings great challenges to the matching cost computation. Shadows, moving objects, "soft" vegetation canopies, and texture-less regions also challenge the results of dense image matching methods. Other challenges include effects caused by view difference and parameter configuration for various scenarios. These will lead to poor matchablility for low-level matching cost, and generally reduce the accuracy of the matched points cloud and DSM products(Han, Liu, et al., 2020; Han, Wang, et al., 2020). An algorithm computing Matching Cost based on Convolutional Neural Networks (MC-CNN) is proposed recently (Zbontar & Lecun, 2015, 2016), based on the appearance similarity of stereo image patches and outperforms many previous methods on KITTI2012 (Geiger et al., 2013), KITTI 2015 (Menze and Geiger, 2015), and Middlebury (Scharstein and Szeliski, 2002, 2003; Scharstein and Pal, 2007; Hirschmueller and Scharstein, 2009; Scharstein et al., 2014) stereo datasets. On the other hand, existing works mainly focus on the close-range or aerial frame images in the computer vision community, little research on the satellite images, given the vast number of pixels for the satellite images created a considerable gap between the photogrammetry and computer vision community. The orientation parameters are provided as Rational Polynomial Coefficients (RPC) files by different satellite image vendors, which helps both in hiding the physical sensor model and allowing high accurate mapping and surveying (Gong, 2003; Gong et al., 2020).

* Corresponding author

This paper reviewed both classical and learning based matching cost computation methods for dense image matching with high-resolution satellite images. Related works are analyzed in Section 2. Section 3 gives the details on the experimental dataset and the methodology. We did the experiments and gave the analysis in Section 4. Last but not least, Section 5 concludes this paper.

2. RELATED WORKS

2.1 Dense Image Matching (DIM) or Light Detection and Ranging (LiDAR)?

Gone are the days when laser scanning sensors, both airborne and terrestrial, provided the fundamental point clouds in the past two decades, photogrammetric dense image matching can deliver comparable dense and detailed 3D point clouds these days, thanks to the significant investigations in both photogrammetry and computer vision, which have contributed to the current automatic dense image matching methods; examples are feature detection/ matching algorithms (Bay et al., 2006; Förstner & Gülch, 1987; Harris & Stephens, 1988; Lowe, 2004); bundle adjustment (Gruen & Beyer, 2001; Hartley & Zisserman, 2003); and dense image matching (DIM) (Heiko Hirschmüller & Scharstein, 2009; Scharstein et al., 2014; Scharstein & Szeliski, 2002).

LiDAR sensors, although lighter and cheaper than decades ago, are still relatively expensive for large-scale 2.5D/3D/4D modeling, mapping compared to photogrammetric sensors. High-resolution Earth Observation Satellites made it possible for surveying and mapping in the case of without a flight permit, or inaccessible places.

2.2 Dense Image Matching Cost Computation

Almost all the present photogrammetric pipeline implemented these components aforementioned in Section 2.1, and DIM might be the most challenging part. Dense image matching aims to find pixel-wise correspondences between stereo images to recover the scene depth information. In the past decades, DIM methods have been facing unprecedented development. Many algorithms were developed and adopted both in the photogrammetry and computer vision community, examples include, multi-image matching (Zhang, 2005), dynamic programming (Van Meerbergen et al., 2002), semi-global matching (H. Hirschmüller, 2005; Heiko Hirschmüller, 2008), patch-based matching (Furukawa & Ponce, 2010, 2012), and deep learning-based stereo matching (Zbontar & Lecun, 2015, 2016).

The Semi-global Matching algorithm was designed one decade ago, implemented in several commercial/open-source software packages and outperformed most of the existing DIM methods when considering both accuracy and efficiency (Han, Qin, et al., 2020). However, with different matching cost computation method adopted, the result of SGM vary dramatically. Different matching costs evaluate the similarity of two pixels in the stereo images, which is the very first and important step for dense image matching. Typical examples are Census transform and MC-CNN for traditional and learning-based matching cost computation methods. Hence, in this paper, we evaluate the performance of the classical Census transform and exploring the benefits of MC-CNN, which adopting CNN for large-scale DSM generation.

2.2.1 Census

Census uses a non-parametric transformation to convert the local grayscale difference between the neighborhood pixel and the central pixel in the neighbor pixels window Ω (the size of the window Ω is $n \times m$, and both n and m are odd numbers) into a string to express the local texture information of the image. This string is the Census transform feature descriptor C_S of the central pixel $p(u, v)$, as shown in formula (1):

$$C_S(u,v) := \otimes_{i=-n'}^{n'} \otimes_{j=-m'}^{m'} \xi\big(I(u,v), I(u+i, v+j)\big) \quad (1)$$

where, $n' \le \frac{n}{2}$, $m' \le \frac{m}{2}$, and $n' \in N^+, m' \in N^+$. $\otimes$ is the bitwise concatenation operation of the string, $\xi(\cdot)$ is defined by formula (2):

$$\xi(x,y) = \begin{cases} 0 & if\ x \le y \\ 1 & if\ x > y \end{cases} \quad (2)$$

Given Census transform uses the relative relationship of the local texture information of the image, it can avoid the influence of abnormal values such as noise to a certain extent; at the same time, it can also obtain relatively ideal results in the disparity jump area such as the edge of buildings. The matching cost of the Census transform is to calculate the Hamming distance of the Census transform values of the center pixels of the two windows corresponding to the left and right epipolar images, as shown in formula (3):

$$C(u,v,d) := Hamming\big(C_{Sl}(u,v), C_{Sr}(u-d,v)\big) \quad (3)$$

We summarised the matching cost computation method based on Census transform as Figure 1.

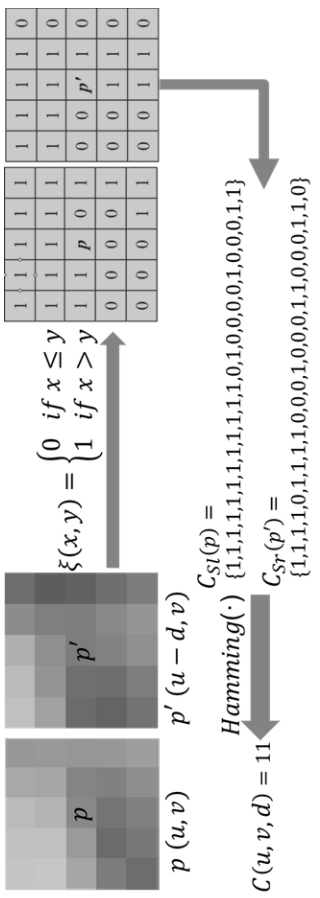


Figure 1. The matching cost computation based on Census.

The Hamming distance is the number of different characters on the corresponding bits of the two Census transform feature descriptors. The calculation method is to perform an XOR

operation on two Census transformed values, and the number of the result string that is not 1 is the Hamming distance of the two Census transformed values, that is, the $L1$ norm of the string XOR operation result, as shown in the formula (4).

$$\rho_{Census}(f,g) = |Census(f) - Census(g)|_1 \quad (4)$$

The range of Hamming distance is $[0, n \times m]$, and the size of the window Ω is $n \times m$.

2.2.2 Matching Cost-CNN

Dense matching based on deep learning is generally divided into two strategies: learning only part of the four steps of classic dense matching, namely non-end-to-end learning and end-to-end learning. The former includes MC-CNN (only used to learn the matching cost, as shown in Figure 2, the cost aggregation, left and right consistency check, median filtering and bilateral filtering and other post-processing steps refer to SGM) and SGM-Net (in SGM, introduce CNN learning penalty items to solve the problem of difficulty in adjusting penalty parameters). It is still necessary to introduce artificially designed post-processing steps to optimize the matching results for these methods. The end-to-end learning methods predict the disparity map directly with the stereo image, including DispNet (a fully convolutional network for disparity map prediction), GC-Net (Geometry and Context Network, which uses geometric and semantic information between pixels to construct a 3D tensor, Learning disparity maps from 3D features) and PSMNet (Pyramid Stereo Matching Network, a network composed of spatial pyramid pooling and three-dimensional convolutional layers, incorporating global background information into stereo matching to achieve occlusion areas, untextured areas or repetitions Reliable estimation of disparity in texture area). The end-to-end methods produce disparity directly, thus we will discuss and compare these methods in the future. All learning strategies need to prepare tens of millons training data containing epipolar images and true disparity values obtained via LiDAR or structured light methods for retrain, which also requires more human and material support. Thus, in this paper, we adopt the pretrained weights from KITTI stereo dataset for MC-CNN in the comparison.

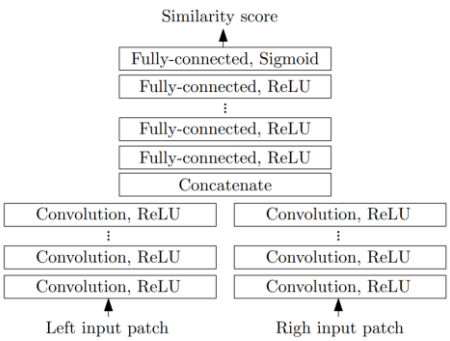


Figure 2. The accurate architecture of the MC-CNN. (Zbontar & Lecun, 2015, 2016)

3. DATASET AND METHODOLOGIES

In our experiments, a 1×1 km^2 area of interest with complicated ground object classes is focused on, charactered with the elevated road/bridge, low-rise residential buildings, high-rise buildings are selected. The satellite images and corresponding ground truth LiDAR data are covering this region created for the Creation of Operationally Realistic 3D Environment, CORE3D (Marc Bosch et al., 2016; M Bosch et al., 2017). WorldView-3 satellite imagery is provided courtesy of DigitalGlobe, and HSIP provides ground truth LiDAR. As shown in Figure 3, there exist 26 WorldView-3 satellite images collected between 2014 and 2016, and 1 WorldView-2 satellite images collected in 2010 covering 100 km^2 area in Jacksonville, Florida. The ground sampling distance (GSD) of the panchromatic images is approximate 0.3 m. The reference DSM at 0.3 m GSD is generated from the LiDAR data as the ground truth. The ground truth DSM in this paper is shown in Figure 1(middle).

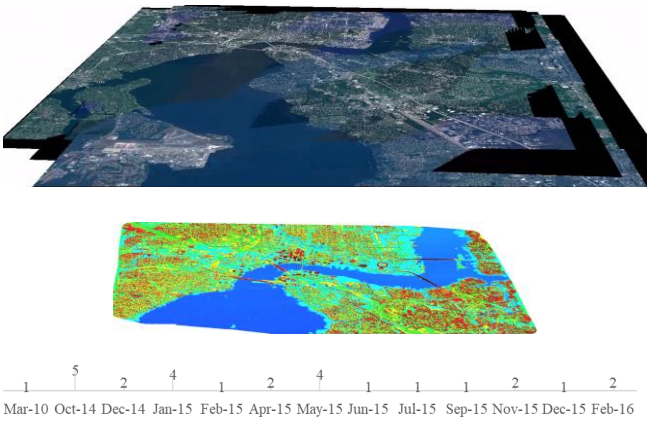
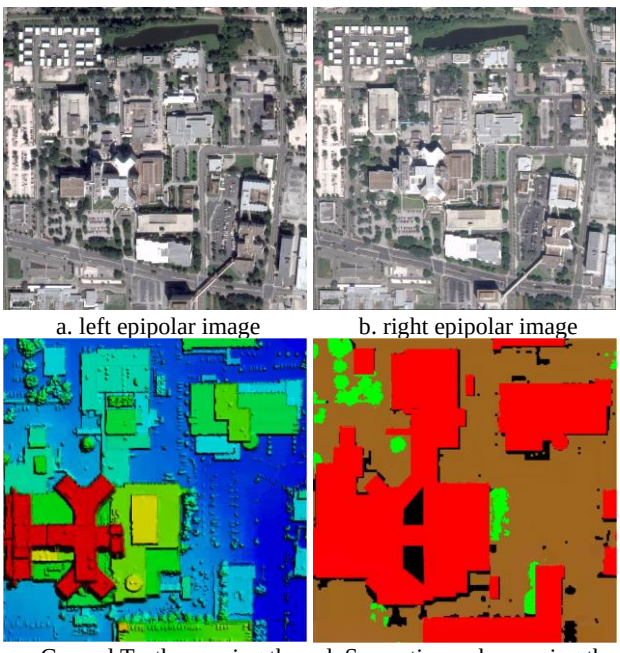


Figure 3. Overview of the 1 WorldView-2 and 26 WorldView-3 images in two years (above), and the acquired LiDAR-DSM (middle). The imaging time is summarized (bottom).

The Johns Hopkins University Applied Physics Laboratory and the Intelligence Advanced Research Projects Agency (IARPA) based on the CORE3D dataset, provided satellite multi-view images, aerial LiDAR, and semantic masks for the 2019 IEEE Data Fusion Competition The competition data set is called grss_dfc_2019 (Contest, 2019). The experimental data in subsequent sections comes from this data set.

An area from one WorldView-3 satellite stereo images is selected in this paper, as shown in Figure 4.



a. left epipolar image b. right epipolar image

c. Ground Truth covering the middle area of the left epipolar d. Semantic mask covering the middle area of the left epipolar

Figure 4. Epipolar images generated from the selected WorldView-3 Stereo Image and corresponding ground truth DSM and the Semantic mask provided in the CORE3D dataset

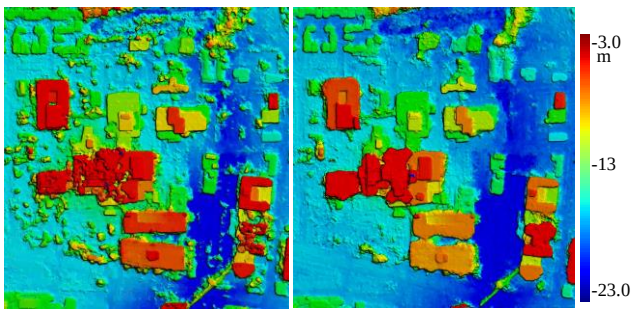
The experiment is designed to compare the performance of Census with 9x9 window size and MC-CNN with pre-trained weights from KITTI stereo dataset for matching cost comparison. We take the epipolar images and corresponding RPC file as the input, the matching costs are computed by Census and MC-CNN, respectively. Then, semi-global optimization is adopted on these cost valumes to compute the disparity image. At last, the computed disparity maps are further triangulated to produce DSMs for the fact that DSM is the usable product in practice.

4. EXPERIMENTS AND RESULTS

The computed DSMs with both Census and MC-CNN based matching cost are shown in Figure 5. The first column in Figure 5 shows the DSM computed using Census based matching cost and SGM, while the second column shows DSM MC-CNN based matching cost and SGM.

Figure 5 shows that the two considered matching costs can achieve good digital surface models result in the area of interest, including elevated road/bridge, low-to-high buildings. However, some mismatches occurred in untextured or weakly textured or shadow regions, e.g., roads, building roofs, and building boundaries, due to matching uncertainties on these objects. For building boundaries, there exists an extension for all the matched DSMs when comparing with LiDAR-DSM.

Moreover, some areas covering with trees (especially the trees in the shadow caused by the tall building boundaries) were not well reconstructed, due to there existing a time gap between imaging and LiDAR acquisition, besides seasonal changes on the vegetation areas among temporal images, which caused the high matching uncertainty reason.



a. Census based DSM b. MC-CNN based DSM

Figure 5. Results visualization from considered matching cost computation methods.

For better analysis on the performance of different solutions quantitatively, firstly, the ground truth DSM is generated from the LiDAR points cloud. Given a matched DSM from one of the three latest solutions and the LiDAR-derived gound truth DSM, the root means square error (RMSE) of all the pixels on the DSM is computed.

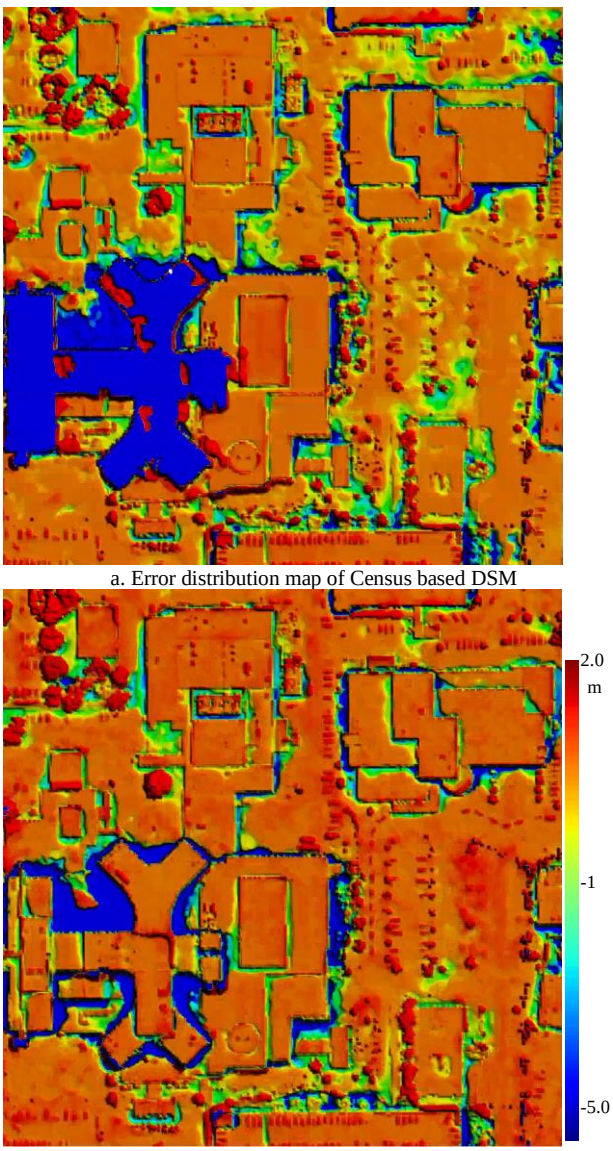
Table 1 shows the RMSEs between the matched DSM and the LiDAR ground truth. By comparing each column, it can be seen that the DSM generated using MC-CNN based matching cost and SGM got a smaller RMSE.

Solution	DSM_{Census}	DSM_{MC-CNN}
RMSE	1.939	1.324

Table 1. Quantitative analysis between the DSM computed from the selected matching cost computation methods and the LiDAR-DSM (meter).

To give a better visual comparison and determine the digital surface modeling performance of an individual matching cost computation method on specific ground feature, we also computed the spatial error distribution maps, as shown in Figure 6.

The spatial error distribution indicates the distance between the matched DSM with the LiDAR-DSM, where red and blue indicate the most considerable distance. It can be seen that the DSM computed using MC-CNN fits the LiDAR-DSM better, while the Census performs slightly worse. However, blunders can be found on the boundary of buildings and vegetation areas. It can be seen from the Census-based DSM, that more blunders exist on the complicated building area. Errors on the boundary areas are caused by the occlusion of objects, which are texture-less and are challenging for matching. The natural seasonal change often causes errors in the vegetation areas. Even with recent CNN methods, it is difficult to reconstruct occlusion objects.



a. Error distribution map of Census based DSM

b. Error distribution map of MC-CNN based DSM

Figure 6. Error distribution map of the computed DSMs from considered matching cost computation methods (red and blue indicate the largest distance).

5. CONCLUSION

Matching cost computation in dense image matching is challenging due to the complexity of imaging conditions, thus it is necessary to compute matching cost as accurately as possible as the basis for the final DSM product. Given that learning-based algorithms has shown their superiority compared to SGM derived methods on Middlebury and KITTI datasets, we assume that matching cost is the key factor of DIM. This paper reviews and evaluates Census Transform, and MC-CNN on a WorldView-3 typical city scene satellite stereo images on the premise that the overall SGM framework remains unchanged. We first compute the cost volume of these two methods, obtains the final DSM after semi-global optimization, and compares their gemetric accuracy with the corresponding LiDAR derived ground truth. The result computed from MC-CNN is based on a pretrained weights from KITTI stereo dataset, achieved a better geometric accuracy on the selected area, which demonstrates its potential power for dense image matching. However, compared with Census, the MC-CNN has higher requirements for computer hardware and a longer running time.

We will consider more aspects, such as more ground covers, more satellite images with different metadata, such as intersection angle and solar angles for a better understanding of the performance for classical and learning based matching cost computation methods in future research.

ACKNOWLEDGEMENTS

The authors would like to thank the Johns Hopkins University Applied Physics Laboratory and IARPA for providing the data used in this study, and the IEEE GRSS Image Analysis and Data Fusion Technical Committee for organizing the Data Fusion Contest. The authors would also like to thank the anonymous reviewers and editor for their thoughtful and constructive comments and suggestions to improve this study.

REFERENCES

Bay, H., Tuytelaars, T., & Van Gool, L. (2006). SURF: Speeded Up Robust Features. In (pp. 404-417): Springer Berlin Heidelberg.

Bosch, M., Kurtz, Z., Hagstrom, S., & Brown, M. (2016). *A multiple view stereo benchmark for satellite imagery*. Paper presented at the 2016 IEEE Applied Imagery Pattern Recognition Workshop (AIPR).

Bosch, M., Leichtman, A., Chilcott, D., Goldberg, H., & Brown, M. (2017). Metric Evaluation Pipeline for 3D Modeling of Urban Scenes. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences, 42*.

Contest, I. G. D. F. (2019). Large-Scale Semantic 3D Reconstruction. Retrieved from http://www.grss-ieee.org/community/technical-committees/data-fusion/2019-ieee-grss-data-fusion-contest-data/

Förstner, W., & Gülch, E. (1987). *A fast operator for detection and precise location of distinct points, corners and centres of circular features*. Paper presented at the Proc. ISPRS intercommission conference on fast processing of photogrammetric data.

Furukawa, Y., & Ponce, J. (2010). Accurate, Dense, and Robust Multiview Stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*(8), 1362-1376. doi:10.1109/tpami.2009.161

Furukawa, Y., & Ponce, J. (2012). Patch-based multi-view stereo software (pmvs-version 2). *PMVS2, University of Washington, Department of Computer Science and Engineering. Web. Downloaded from on May*.

Gong, D. (2003). *Models and algorithms on processing of high-resolution satellite remote sensing stereo images.* Information Engineering University, Zhengzhou.

Gong, D., Han, Y., & Zhang, L. (2020). Quantitative Assessment of the projection trajectory-based epipolarity model and epipolar image resampling for linear-array satellite images. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 1*, 89-94. doi:10.5194/isprs-annals-V-1-2020-89-2020

Gruen, A. (2012). Development and Status of Image Matching in Photogrammetry. *The Photogrammetric Record, 27*(137), 36-57. doi:10.1111/j.1477-9730.2011.00671.x

Gruen, A., & Beyer, H. A. (2001). System calibration through self-calibration. In *Calibration and orientation of cameras in computer vision* (pp. 163-193): Springer.

Haala, N., & Kada, M. (2010). An update on automatic 3D building reconstruction. *ISPRS journal of photogrammetry and remote sensing, 65*(6), 570-580. doi:10.1016/j.isprsjprs.2010.09.006

Han, Y., Liu, W., Huang, X., Wang, S., & Qin, R. (2020). Stereo Dense Image Matching by Adaptive Fusion of Multiple-Window Matching Results. *Remote Sensing, 12*(19), 3138.

Han, Y., Qin, R., & Huang, X. (2020). Assessment of dense image matchers for digital surface model generation using airborne and spaceborne images – an update. *The Photogrammetric Record, 35*(169), 58-80. doi:10.1111/phor.12310

Han, Y., Wang, S., Gong, D., Wang, Y., Wang, Y., & Ma, X. (2020). State of the art in digital surface modelling from multi-view high-resolution satellite images. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2*, 351-356. doi:10.5194/isprs-annals-V-2-2020-351-2020

Harris, C. G., & Stephens, M. (1988). *A combined corner and edge detector*. Paper presented at the Alvey vision conference.
Hartley, R., & Zisserman, A. (2003). *Multiple view geometry in computer vision*: Cambridge university press.

Hirschmüller, H. (2005). *Accurate and Efficient Stereo Processing by Semi-Global Matching and Mutual Information.* Paper presented at the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05).

Hirschmüller, H. (2008). Stereo Processing by Semiglobal Matching and Mutual Information. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30*(2), 328-341. doi:10.1109/tpami.2007.1166

Hirschmüller, H., & Scharstein, D. (2009). Evaluation of Stereo Matching Costs on Images with Radiometric Differences. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(9), 1582-1599. doi:10.1109/tpami.2008.221

Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision, 60*(2), 91-110. doi:10.1023/b:visi.0000029664.99615.94

Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., & Westling, P. (2014). *High-resolution stereo datasets with subpixel-accurate ground truth*. Paper presented at the German conference on pattern recognition.

Scharstein, D., & Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision, 47*(1/3), 7-42. doi:10.1023/a:1014573219977

Van Meerbergen, G., Vergauwen, M., Pollefeys, M., & Van Gool, L. (2002). A hierarchical symmetric stereo algorithm using dynamic programming. *International Journal of Computer Vision, 47*(1/3), 275-285. doi:10.1023/a:1014562312225

Zbontar, J., & Lecun, Y. (2015). *Computing the stereo matching cost with a convolutional neural network.* Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Zbontar, J., & LeCun, Y. (2016). Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches. *Journal of Machine Learning Research, 17*(1-32), 2.

Zhang, L. (2005). *Automatic digital surface model (DSM) generation from linear array images.* ETH Zurich,