Point Cloud Fusion Algorithm for High-Quality Digital Surface Model Generation from Multi-Date Stereo Images

D. Díaz-Blanco , Paula Gonzalez , Izar Azpiroz , Giovanni Paolini , Mikel Maiza
Vicomtech Foundation
Basque Research and Technology Alliance (BRTA)
Donostia-San Sebastián 20009, Spain
e-mail: {dadiaz, | pgonzalez, | iazpiroz, | gpaolini, | mmaiza}@vicomtech.org

Abstract—**This paper presents a novel methodology for generating high-quality Digital Surface Models (DSMs) through the fusion of point clouds obtained from multi-date stereo images. By applying a custom fusion algorithm to the point clouds generated by the Context-Aware Reconstruction of Scenes (CARS) software, the proposed approach enhances DSM quality in terms of completeness and error metrics compared to the original DSM. The fusion process effectively integrates multiple DSMs, resulting in a more comprehensive and accurate terrain representation. This method addresses challenges such as shadow occlusions and temporal variations, demonstrating significant improvements. The technique shows potential for applications in precision agriculture and other fields requiring detailed terrain models. Validation using the Intelligence Advanced Research Projects Activity (IARPA) challenge dataset highlights the method's robustness in mixed terrains, offering a notable increase in completeness and solving issues related to data gaps in shadowed areas.**

Keywords-Point Cloud Fusion; Digital Surface Model (DSM); Multi-Date Stereo Images; Terrain Representation.

I. Introduction

The world population has exceeded the 8 billion barrier according to a United Nations press release of November 2022. Moreover, the world population is expected to reach 8.5 billion by 2030, and 10.4 billion by 2100. This rapid growth is expected to place increasing pressure on land and other natural resources, presenting significant challenges to food security [1]. The growing need to produce more and higher-quality food with unsustainable agricultural practices, as well as climate change and urban growth, are accelerating the loss of available arable land, threatening sustainability in terms of productivity and environmental impact [2]. It is important to note that climate change will lead to extreme environmental events that will require a rapid and efficient response from the agricultural sector. The agricultural sector must adapt effectively to mitigate the adverse effects of such events and ensure global food security [3].

Several studies have already indicated that modernisation processes are crucial to overcome the difficulties caused by agricultural land change [4]. Among these processes is where Precision Agriculture (PA) can be mentioned and highlighted as one of the solutions to ensure food security for the whole world. The PA, also known as Smart Farming or Agriculture 4.0, is an agricultural management strategy focused on improving the efficiency in the use of resources, productivity, quality, profitability, and sustainability of agricultural production [5]. This discipline implements technologies and resources of all kinds, including, among others, Digital Surface Models (DSM). A DSM is a type of elevation model that not only represents the height of the terrain in areas devoid of objects but also considers all features present on the terrain, including buildings, tree canopies, and other elements on the earth's surface [6]. DSM can have a wide range of applications in the field of PA, notably in evaluating the suitability of terrain for agricultural use, crop yield monitoring, and biomass estimation [7]. DSMs are a fundamental starting point for the development of other models, among which the Digital Elevation Model (DEM) stands out. The latter represents the earth's surface once the elements that are not part of it have been removed, providing crucial information in various disciplines within the environmental field [6]; these models stand as pivotal spatial information tools in geomorphological applications, enabling the extraction of essential attributes like slope, aspect, profile curvature, and flow direction [8].

The extraction of elevation models can be derived from a variety of techniques; however, historically, aerial photogrammetry and LiDAR have been the most widely used methods for their generation. Nowadays, techniques derived from optical satellite imagery are also used. Among these, interferometric techniques based on radar images have been extensively investigated. Nevertheless, their application requires more complex processing involving the use of specialised algorithms and software, compared to techniques based on optical satellite imagery. In addition, optical imagery offers better interpretability and is more widely accessible and available [9].

One of the most commonly used techniques for DSM generation from optical satellite imagery is the stereo method. DSMs are generated using dense point clouds acquired from stereoscopic satellite imagery [10]. Point clouds are detailed sets of three-dimensional points that capture terrain features (buildings, vegetation, etc) using advanced image-matching algorithms [11]. Some research already mentions the importance that point cloud fusion brings to the quality and accuracy of DSM [12]. By integrating information from multiple viewpoints, point cloud fusion overcomes the occlusions and inaccuracies inherent in individual stereo pairs, resulting in more complete and detailed terrain representations. This approach not only improves spatial resolution and accuracy, but also facilitates the extraction of finer details [10].

This paper presents a methodology to generate high-quality

ISBN: 978-1-68558-323-1

DSMs by fusing point clouds obtained from stereo images captured at different dates. This approach uses point clouds generated by the Context-Aware Reconstruction of Scenes (CARS) software [13]. The main objective is to study the improvement produced by the fusion of DSMs and compare the results with other similar works like [10], where different software (S2P: Satellite Stereo Pipeline [14]) were used to generate the point clouds. As CARS appear to generate better results in the used stereo images dataset than S2P [13], it is significant to analyse the improvement made by the fusion of the results from CARS.

The rest of the paper is structured as follows. The Methodology section is divided into three parts. First, we present the Align and Fusion methodology. Then, the considered real context for validation purposes is detailed. The Methodology section is closed with a description of the obtained results. Finally, the Conclusion section summarizes the main contributions of the presented work and future perspectives.

II. Methodology

A. Align and Fusion methodology

The present subsection describes the iterative procedure to fusion P point clouds generated by CARS from different multi-date image pairs in order to create a unique fused DSM for a given Region of Interest (ROI). Each point cloud $p \in 1, ..., P$ is aligned with a reference point cloud selected by the user, and the P aligned point clouds are fused along with the reference one. The final fused point cloud is rasterised into an image for display. The proposed procedure includes the following steps:

1) Preparation: Pre-alignment processing of the point clouds composed from each image pair.

- CARS generates multiple point clouds for each DSM, as it separates the processing into tiles. Subsequently, the point cloud files generated for each tile are merged to generate a single point cloud for each DSM to be fused. The common set of pixels in both point clouds is projected for each DSM onto a grid with a specified node spacing, ideally equal to or greater than the image resolution. This forms a two-dimensional matrix in which the value of each cell represents the estimated height z in the coordinates (x, y).
- Applying a grayscale-closing interpolation, single pixel holes are filled averaging the values of elements in a surrounding 3x3 area, while the larger holes are kept as non-data. These larger holes are usually consequence of being located in shadow areas. This creates the new DSMs, that will be the reference DSM and the DSM p to be fused: E_{ref} and E_p respectively, which will be the inputs to the fusion step (Section II-A3).
- Another pair of DSMs D_{ref}, D_p is generated from the previous E_{ref}, E_p by completely filling all holes using the lowest hole edge values (using the 5th percentile), so that the occluded parts where no data has been generated are assumed to be at ground level. These DSMs will be used for alignment purposes (Section II-A2).

2) Alignment: Due to the pointing errors of RPCs (*Rational Polynomial Coefficients*) models [15], 3D point clouds obtained from different images are usually not aligned. The usual method for adjusting the parameters of all cameras uses correspondences between images (e.g., by Scale-invariant Feature Transform algorithm [16] matching). However, this method is sensitive to noise and radiometric changes, which are common in a multidate analysis [10]. The error induced in the point clouds by the pointing error is mainly a 3D translation, so following the strategy proposed by Facciolo et al. [10] the translation of D_p that maximizes the Normalized Cross-Correlation (NCC) over D_{ref} is calculated as follows:

$$NCC(\mathbf{u}, \mathbf{v}) \equiv \frac{1}{|\hat{\Omega}|} \sum_{j \in \hat{\Omega}} \frac{(\mathbf{u_j} - \mu_{\mathbf{u}}(\hat{\Omega}))(\mathbf{v}_j - \mu_{\mathbf{v}}(\hat{\Omega}))}{\sigma_{\mathbf{u}}(\hat{\Omega})\sigma_{\mathbf{v}}(\hat{\Omega})} \tag{1}$$

where $\mathbf{u}, \mathbf{v}$ are each one of the DSMs to align (in this case D_{ref}, D_p respectively), $\hat{\Omega} \equiv \Omega_{\mathbf{u}} \cap \Omega_{\mathbf{v}}$ is the intersection of the sets of known pixels in both DSMs, j represents an index that iterates over the pixels within the intersection set $\hat{\Omega}$, $\mathbf{u}_j$ refer to the pixel values of DSM $\mathbf{u}$ at position j. $\mu_{\mathbf{u}}$ and $\sigma_{\mathbf{u}}$ represents the simple mean and the standard deviation of $\mathbf{u}$, respectively. The same notation applies to $\mathbf{v}$.

We then look for the pair (dx^*, dy^*) under which the offset dx, dy maximizes NCC:

$$(dx^*, dy^*) = \arg\max_{dx,dy} NCC(\mathbf{u}, \mathbf{v}_{dx,dy}) \tag{2}$$

where $\mathbf{v}_{dx,dy}$ represents the DSM $\mathbf{v}$ shifted dx and dy. A search for (dx^*, dy^*) is applied following a coarse-to-fine method:

- 1) Shift v in coarse steps (e.g:25 cells) and calculate the NCC at each shift.
- 2) The offset that gives the largest NCC value in the initial search is selected.
- 3) New consecutively smaller steps (e.g: 5 and 1 cells) are added to the coarse shift (shifting in total always less than the value of the previous coarse shift) until (dx^*, dy^*) that maximizes NCC is found.

Shift in z (dz^*) is calculated as the difference between the height means of D_{ref} and D_p.

Finally, the translation (dx^*, dy^*, dz^*) is applied to E_p to obtain $E_{p,aligned}$, which is saved as a point cloud file.

3) Fusion: In this step all point cloud files aligned in the previous step are combined into a single matrix:

$$M(x, y, k) = \begin{cases} E_{ref}(x, y) & \text{if } k = 0 \\ E_{p,aligned}(x, y) & \text{for } k = 1, 2, \ldots, P \end{cases} \tag{3}$$

A three-dimensional matrix is generated, where x, y represent the pixel location and every value of k is a layer which represents one of the point cloud in the fusion. The dimension k has a maximum value equal to the number of fused point clouds. The value of each cell in the matrix, $M(x, y, k)$, represents a height z. For each pixel x, y we perform a k-medians clustering analysis of the values of the heights along k with a similar approach than [10], increasing the number of clustering from a single one to a maximum number n_{max}, with the difference that in [10] n_{max} is always 8, and in our approach $n_{max} = min(8, length(k) - 1)$, such that n_{max} is equal to the number of existing heights minus 1 if there are equal or less than 8 heights or 8 if there are more. By this way, we are able to perform better with small numbers of fused DSMs, where using 8 clusters with less than 8 heights values has no sense. The number of clusters is increased until every cluster has a span between the minor and maximum height of each cluster less than a predefined value (arbitrary value used in this analysis: grid resolution + 1m). If one or two clusters are detected, the lowest level is saved, and if more are detected, it is saved as non-data, since the results are not considered coherent. Once the cluster is saved, the value of its median is taken as the height in that pixel, forming the DSM merged in a two-dimensional matrix.

The objective of this method is to obtain the best estimation of the height, by saving a height level which is similar between some of the DSM to be fused. The intention is to prevent objects above ground level, such as variable vegetation, from distorting the result by preserving the value that should represent the ground level, using the lowest height cluster. The heights corresponding to the object that perturbs the height value at that point would be stored in another cluster, obtaining the one cluster corresponding to the ground and that of the upper object. If more than two clusters are obtained, the value of the height at that point is considered invalid because it does not fit any of the cases, assumed to be a spurious result and stored as non-data.

4) Rasterization: Once the fused DSM has been obtained as a matrix, the objective is to generate a georeferenced raster image. This goal is achieved creating a 2D matrix for the DSM with dimensions defined by the resolution and the boundaries of the region of interest:

$$\text{width} = \left[\frac{x_{\text{max}} - x_{\text{min}}}{r}\right] + 1 \tag{4}$$

$$\text{height} = \left[\frac{y_{\text{max}} - y_{\text{min}}}{r}\right] + 1 \tag{5}$$

Where x_{min}, x_{max} is the maximum and minimum co-ordinate respectively in the chosen Coordinate Reference System (CRS), and the same for y_{min}, y_{max}. r represents the resolution of the DSM grid.

The matrix is initialized with NaN (Not a Number) values to indicate the absence of data. Each point in the point cloud is inserted into the DSM matrix. For each point (x, y), the corresponding position in the matrix is calculated and the value of z is assigned to that position.

To smooth the DSM and reduce noise, a weighted Gaussian filter is applied. This filter considers the proximity of the points and their height values to generate a more accurate DSM. The motivation for this method, not applied in [10], is to use one equivalent to the one used in the CARS rasterization [13], in order to make a fairer comparison between an original CARS DSM and the DSM resulting from the fusion.

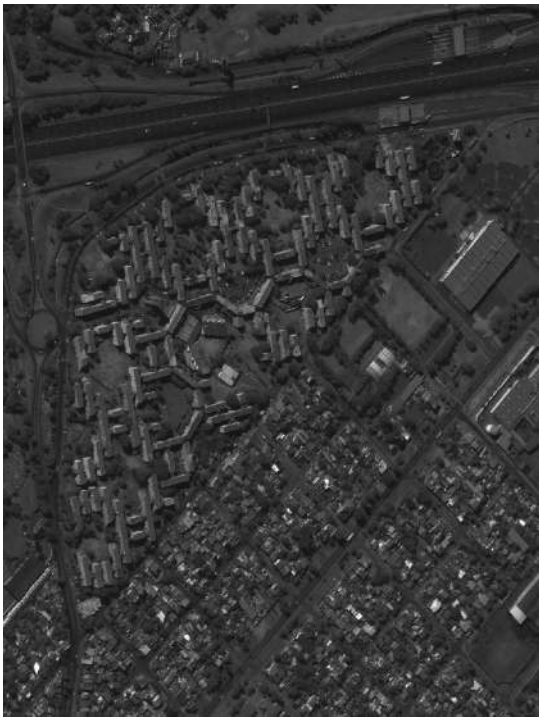


Figure 1: Panchromatic band image (PAN) from the IARPA database

Finally, a raster image is created. The image is georeferenced using the geographic coordinates of one of the corners (typically the upper left corner) and the defined resolution. The geographic projection is established using a CRS corresponding to the worked portion of the surface. The final raster image, representing the DSM, is saved in GeoTIFF format. Each pixel of the image corresponds to a cell in the DSM matrix and its brightness value represents the height.

B. Real context validation

1) Dataset: The algorithm is used for fusing DSMs generated from the IARPA challenge dataset [17], which covers the city of Buenos Aires, Argentina. This dataset contains, among other files, 30 cm resolution NITF images from World-View 3 satellite, which can be converted specifying ROI to TIF images as in Figure 1 and GEOM files with the RPCs corresponding to each image. The specific site analyzed contains high- and low-density urban areas corresponding to city areas. They do not represent agricultural fields but contain some tree zones and a flat highway area, thus allowing the study of the algorithm's behavior in different types of terrain.

Based on the fusion method presented in Section II-A, different DSMs obtained from pairs of manually selected images have been fused under two of the criteria selected by [10]:

- The angle between the views of the image pair must be within 5º and 45º.
- Temporal proximity

For the generation of the DSMs and the visualization of the subtended angle between the views, a graphical interface of our own creation was used, which uses CARS for the generation of the point clouds for each pair of images.

2) Metrics: We compared the fused DSM with the original DSM generated by CARS using a very high quality LiDAR-generated DSM as ground truth. The following DSM quality metrics are used:

- Completeness: Percentage of pixels with valid values (not NaN).
- Root Mean Square Error (RMSE).
- Standard Deviation (STD).

C. Results

Following the procedure described in Section II, we obtain point clouds and their corresponding raster images as shown in Figures 2 and 3, respectively. Most of the occluded and vegetated areas visible in Figure 2 indicate that ground data has been obtained after the fusion (examples in red circles, where tree crowns, represented as groups of points higher than the ground with reddish colours, have been removed and ground points have been obtained).

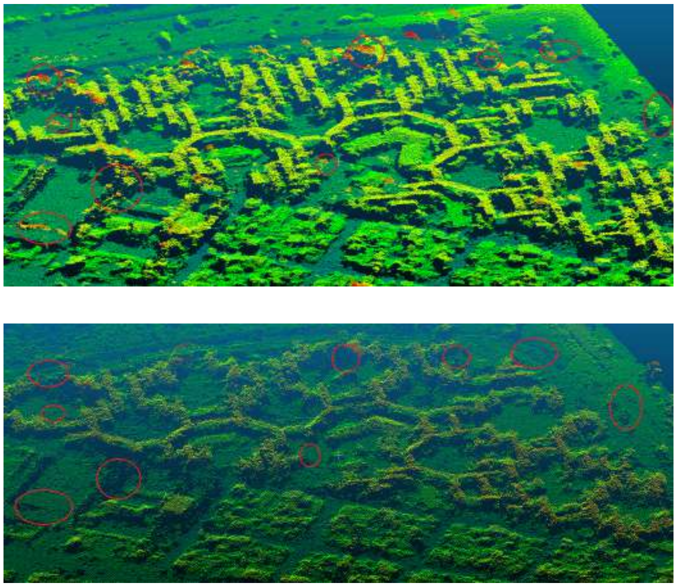
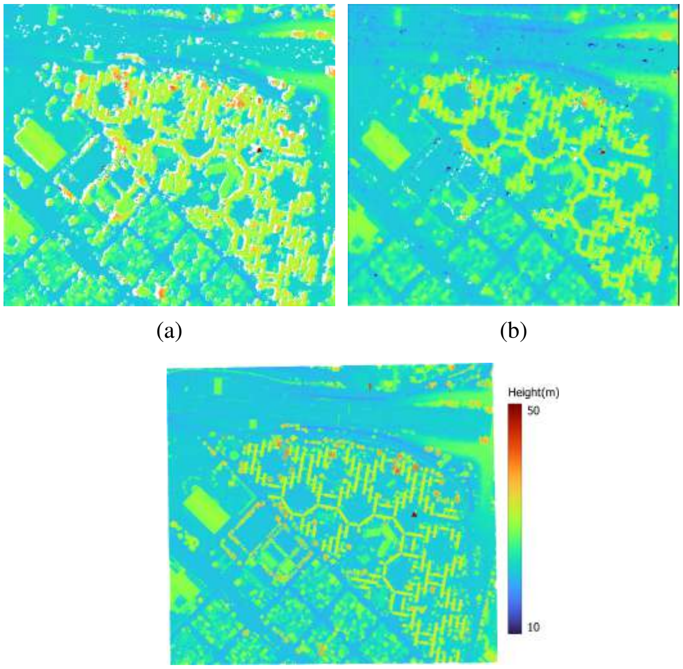


Figure 2: Original and fused point clouds. Top: Original DSM generated by CARS. Bottom: Fusion of 8 DSMs by applying the procedure described in section II-A.

This increase in completeness for number of DSMs within 3-12 is shown in Figure 4, where sections of the DSMs are shown: in Figures 4(a-I) and 4(a-II) we observe shadowed areas with no data (white color), whereas in the fused DSM of Figures 4(b-I) and 4(b-II) those areas are complete. It must be mentioned that in Figure 4 it is easy to see how some part of the trees have been removed in the fusion, and the more percentage of ground is shown, thanks to obtaining data on their height from the different views and dates of the DSMs.

The quality metrics of the fused DSM are plotted in Figure 5. In this case, we observe that there is a general trend of RMSE and STD reduction in Figure 5(a), and a quick increase in completeness, followed by a reduction from 12 fused DSMs.



(a)

(b)

(c)

Figure 3: DSM comparison: a) Best individual DSM (from one pair only) among the ones used in the fusion. b) Fused DSM obtained from 8 DSMs from individual pairs. c) DSM obtained by LiDAR, used as ground truth.

Figure 6 shows the difference between the fused DSM and the ground truth taken by the LiDAR.

It must be mentioned that the improvement in results occurs with a lesser number of fused DSMs compared to [10], where the best results were obtained at around 50 fused DSMs. On the other hand, by adding a significant amount of DSM to fuse the completeness drop, as more pixel heights are considered as non-data. It is not clear whether this different behavior from [10] is due to differences in the algorithm used in the present work, or differences in the characteristics of the point clouds generated by CARS and S2P.

One of the advantages provided by this method is the possibility of removing a large part of the trees from the fused DSM by simply adding DSMs generated from images taken in the leafless trees season or by fusing DSMs generated from different views, so that data can be obtained for the occluded area. In Figure 6, we observe that the error of the merged DSM is significantly concentrated in the tree areas, as in the merged DSM the latter were eliminated, while being present in the image taken by the LiDAR. The k-clustering algorithm takes the cluster with the lowest value, which should correspond to the ground, and stores it as the height at that point. We can observe this phenomenon in Figure 4, where many trees have been removed. This has a negative impact on the error metrics, as this removal of trees, although not detrimental to the terrain representation, increases the error with respect to a LiDAR image with trees, so the overall STD and RMSE values do not accurately represent the improvement of the fused DSM with respect to the surface and are not reduced as much as possible due to the increasing of the error in the

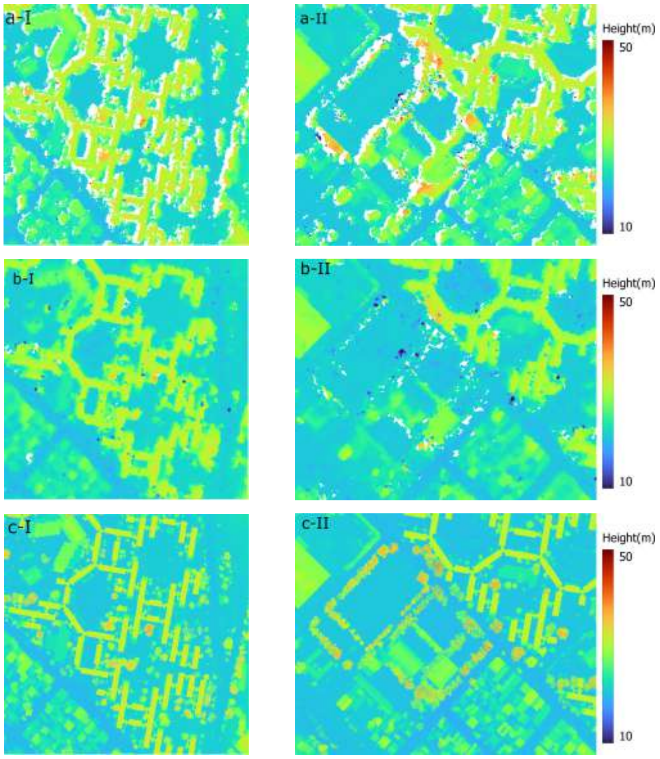


Figure 4: Sections of the DSM where different types of surfaces are shown. Horizontally: I) area with buildings and trees, II) area with trees on a flat sports field. Vertically: a) Best individual DSM (lowest RMSE value) b) Fused DSM. c) DSM obtained by LiDAR, used as ground truth.

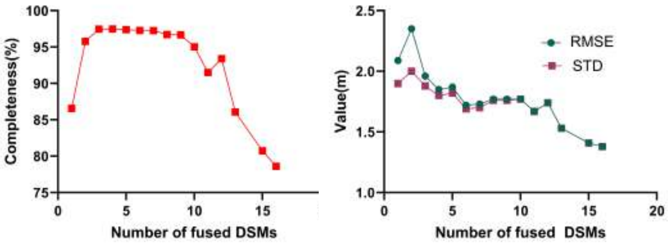


Figure 5: Metrics of the results obtained. Left: Completeness of the resulting DSM. Right: STD and RMSE values.

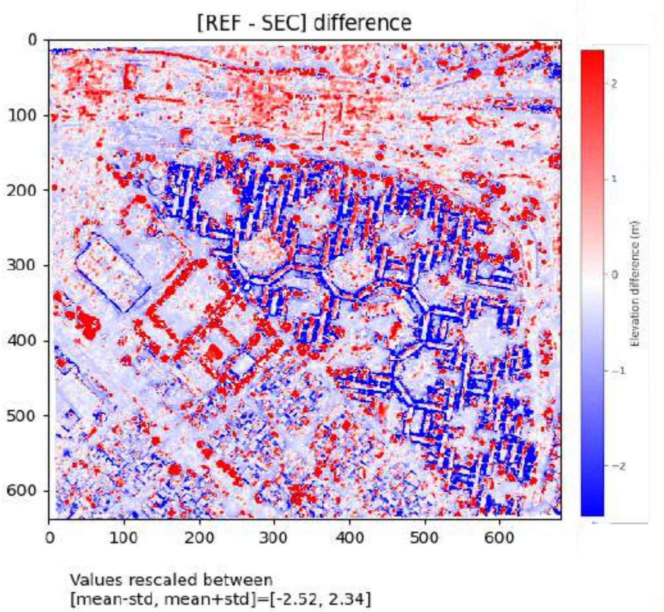


Figure 6: Difference between the fused DSM from 8 point clouds and the actual LiDAR value.

areas with trees. It should also be mentioned that while the improvement in RMSE and STD is around 15%-20% and 7%-12% respectively for the fused DSM with best results (with a completeness higher than the original DSM, using between 6-12 DSMs), the improvement in completeness is remarkable, offering a fused DSM with values greater than 97%, so that the problem of shadow areas without data in the original DSM is practically solved by this method.

III. Conclusions

A methodology to generate high-quality DSMs through the fusion of point clouds obtained from stereo images taken at different dates has been presented. This approach leverages the CARS software to generate point clouds, offering an improvement over previous software such as S2P. The study demonstrates that the fusion process significantly enhances the quality of the DSMs, particularly in terms of completeness and error metrics. The results validate the effectiveness of the proposed method in a mixed terrain, showcasing its potential for applications in precision agriculture and other fields requiring detailed terrain models. The successful integration of multiple DSMs results in a more comprehensive and accurate representation of the terrain, addressing challenges like shadow occlusions and temporal variations in the data.

These results confirm that the DSM fusion procedure improves the quality of the results, having improved them using a similar procedure from point clouds generated with different software.

Considering future continuation of this work, the quality metrics of our DSMs could potentially be improved by employing a more sophisticated procedure. This would involve generating all possible DSMs from pairs of images, organizing these DSMs based on their parameters, and selecting the most suitable ones. Additionally, incorporating enough different dates for covering the maximum surface area while considering changes in vegetation and luminosity would ensure a more comprehensive analysis. This approach, aimed at enhancing the accuracy and completeness of the DSMs, remains a subject for future work.

Finally, it should be mentioned that sustainable farming practices can be improved through the use of static DSMs, as they provide valuable insights for efficient irrigation, soil erosion prevention, optimized fertilizer application, and other key activities.

References

[1] R. Bongiovanni and J. Lowenberg-DeBoer, “Precision agriculture and sustainability,” *Precision agriculture*, vol. 5, pp. 359–387, 2004.

[2] A. Sharma, A. Jain, P. Gupta, and V. Chowdary, “Machine learning applications for precision agriculture: A comprehensive review,” *IEEE Access*, vol. 9, pp. 4843–4873, 2020.

[3] D. B. Lobell, W. Schlenker, and J. Costa-Roberts, “Climate trends and global crop production since 1980,” *Science*, vol. 333, no. 6042, pp. 616–620, 2011.

[4] M. J. Grundy *et al.*, “Scenarios for australian agricultural production and land use to 2050,” *Agricultural systems*, vol. 142, pp. 70–83, 2016.

[5] C. Cantini, P. E. Nepi, G. Avola, and E. Riggi, “Direct and indirect ground estimation of leaf area index to support interpretation of ndvi data from satellite images in hedgerow olive orchards,” *Smart Agricultural Technology*, vol. 5, p. 100 267, 2023.

[6] Q. Zhou, “Digital elevation model and digital surface model,” *International Encyclopedia of Geography: People, the Earth, Environment and Technology*, pp. 1–17, 2017.

[7] J. Torres-Sánchez, F. López-Granados, I. Borra-Serrano, and J. M. Peña, “Assessing uav-collected image overlap influence on computation time and digital surface model accuracy in olive orchards,” *Precision Agriculture*, vol. 19, pp. 115–133, 2018.

[8] M. M. Ouédraogo, A. Degré, C. Debouche, and J. Lisein, “The evaluation of unmanned aerial system-based photogrammetry and terrestrial laser scanning to generate dems of agricultural watersheds,” *Geomorphology*, vol. 214, pp. 339–355, 2014.

[9] D. B. Gesch, “Global digital elevation model development from satellite remote-sensing data,” *Advances in Mapping from Remote Sensor Imagery: Techniques and Applications; Yang, X., Li, J., Eds*, pp. 92–109, 2012.

[10] G. Facciolo, C. de Franchis, and E. Meinhardt, “Automatic 3d reconstruction from multi-date satellite images,” *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1542–1551, 2017.

[11] H. Hirschmuller, “Stereo processing by semiglobal matching and mutual information,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 328–341, 2007.

[12] K. Gong and D. Fritsch, “Point cloud and digital surface model generation from high resolution multiple view stereo satellite imagery,” *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 42, pp. 363–370, 2018.

[13] J. Michel *et al.*, “A new satellite imagery stereo pipeline designed for scalability, robustness and performance,” *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. V-2-2020, pp. 171–178, 2020. DOI: 10.5194/isprs-annals-V-2-2020-171-2020.

[14] C. de Franchis, E. Meinhardt-Llopis, J. Michel, J.-M. Morel, and G. Facciolo, “An automatic and modular stereo pipeline for pushbroom images,” *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. II-3, pp. 49–56, 2014. DOI: 10.5194/isprsannals-II-3-49-2014.

[15] *Rpc (rational polynomial camera model), the nitty gritty*, https://edgybees.com/rpc-rational-polynomial-camera-model-the-nitty-gritty/, Accesed: 03-07-2024.

[16] D. G. Lowe, “Object recognition from local scale-invariant features,” *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, 1150–1157 vol.2, 1999.

[17] M. Bosch, Z. Kurtz, S. Hagstrom, and M. Z. Brown, “A multiple view stereo benchmark for satellite imagery,” *2016 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pp. 1–9, 2016.