Generating Digital Elevation Models using Drone Footage through Structure-from-Motion

Pedro Afonso
PARIS LODRON UNIVERSITY OF SALZBURG

Contents

Abstract

The generation of Digital Elevation Models (DEMs) does not have to rely on high-cost methods such as LIDAR or InSAR. The current technology (both on the software and hardware levels) makes it possible to create highly accurate DEMs with low-end recreational drones using Structure-from-Motion (SfM) photogrammetry

This paper focuses on the new City Hall of the Municipality of Valongo, in Portugal, to showcase how this can be achieved with a 10-minute flight of a DKI 4K Mini Drone. The processing step was done using both Aerial Model's free plan and Agisoft Metashape free trial period.

The final result was a 5-cm DEM of the area of interest, which had a 90 meter error overall when compared to the ground truth, but that could be easily fixed once this was realized.

Keywords: Digital Elevation Models, Structure-from-Motion, Drone Imagery

1 Introduction

Digital Elevation Models (DEMs) are one of the staples of Remote Sensing and Geoinformatics. They can be generated using a variety of techniques, most famously LIDAR (whether it be Airborne, Mobile, or Terrestrial) and InSAR. However, these two methods are often associated with high costs, either because of the transportation of the material or the complexity of said material, which, in addition to the expected measurements, must also register the 3D location and orientation of the camera in order to virtually reconstruct the 3D landscape later on.

Most of these issues can be overcome by using low-end recreational drones for image acquisition and processing the images with Structure-from-Motion (SfM) photogrammetry, a technique that requires no location or orientation data, relying solely on finding correspondence between images and triangulation.

This technique traces its origins to a landmark paper by Wallach and O'connell (1953) and its introduction of the "Kinetic depth effect", which shows that it is possible to perceive the shape of 3D objects by observing the rotation of their silhouette. The perception of the tri-dimensional world through bi-dimensional representations, in the field of computer vision, continued to be explored in the following decades, culminating in a paper by Ullman (1979), where the Structure-from-Motion Theorem was first stated:

Given three distinct orthographic projections of four non-coplanar points in a rigid configuration, the structure and motion compatible with the three views are uniquely determined up to a reflection about the image plane.

The motion must, of course, be rotational, since the recovery of a translation in depth is impossible, and the recovery of other translations would be a trivial solution. Moreover, this is a mathematical conclusion. In practice, it's important to have more than 3 views of each set of 4 points, in order to mitigate the position errors and improve the quality of the results. Nonetheless, the idea was there: by collecting several (overlapping and converging images) of an object, as shown in Figure 1, and identifying tie points between them, it should be possible to recreate that object's 3D structure.

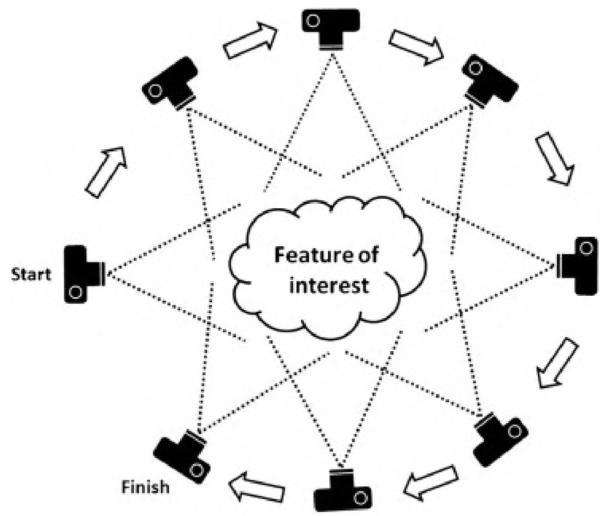

Figure 1: Diagram explaining the process of Image Collection (Westoby et al., 2012)

Despite the elegance of the theorem, however, there was yet a long (metaphorical) road ahead before any of these could be used for recreating 3D landscapes with simple drone images. The next big step in that direction would come with the creation of the SIFT (Scale Invariant Feature Transform) algorithm, which is capable of detecting the same objects/features in different images in spite of luminosity changes, rotations, and scale differences (Lowe, 1999).

This made it possible to automatize the generation of the tie point cloud, which otherwise would have to be created manually, by identifying and selecting every chosen tie point, on every image.

Finally, with the advent of publicly-accessible satellite positioning, the feature detection algorithms became even more precise! Furthermore, it automatically provided the whole system with a reference frame, which, once again, would have to be defined manually otherwise.

In 2012, the first applications of this technique to geosciences emerged (Westoby et al., 2012; James and Robson, 2012), setting the field for everyone that would come after.

2 Materials and methods

2.1 Area of focus

The chosen focus area was the new City Hall of the Municipality of Valongo, in North-western Portugal (41.191° N, 8.497° W), more specifically the entire construction site around it, which measured approximately 14 thousand m^2, and, as illustrated in Figure 2, was designed to resemble a Trilobite, an extinct animal whose fossils are abundant in the region. This site was chosen for two reasons: the first one was to simply showcase the new City Hall of the author's hometown to the world, and the second (and more important) was the opportunity to have up-to-date data of an area that suffered big changes since Portugal's latest 50 cm DEMs were generated, in late 2024.

Figure 2: New Citty Hall of Valongo, enviosioned in the shape of a Trilobite.

2.2 Data collection

The data was acquired using a DJI 4K Mini drone (`www.dji.com`), seen in Figure 3 on the following page. The drone is equipped with a GPS-assisted hover function and a 12 megapixel on-board camera (1/2.3-inch CMOS sensor) with an f/2.8 aperture, an 83° field of view, and a 24 mm full-frame equivalent focal length, mounted on a three-axis mechanical gimbal.

Since DGT, "Direção Geral do Território" (the portuguese General Directorate for Territory) make 50 cm DEMs of the country available on their website (as mentioned in Section 2.1), it was possible to consult the DEM of the area and assess that there was a 50 meter-tall building in the vicinity of the site. It was then decided the drone would have to be flying at least 60 meters above ground (for its own protection); on the other hand with the camera's inclination set at 45°, that would prevent the drone from going more than 10 meters away from the building, so a height of 70 meters above ground was finally chosen, even though this varied a little due to wind.

Figure 3: Drone used for the data collection, alongside the (unused) spared battery and the controller plugged to a smart phone.

The drone was manually controlled from the ground using the DJI remote controller, since the model didn't have "planned flight" capabilities. A total of 49 pictures were taken, during a single flight that lasted 10 minutes. All 49 pictures have been represented in Figure 4.

There was a conscious attempt to keep the focus on the building to avoid the so-called dome effect, which results from taking parallel images. Inadvertently, this resulted in the surrounding areas often lacking some details.

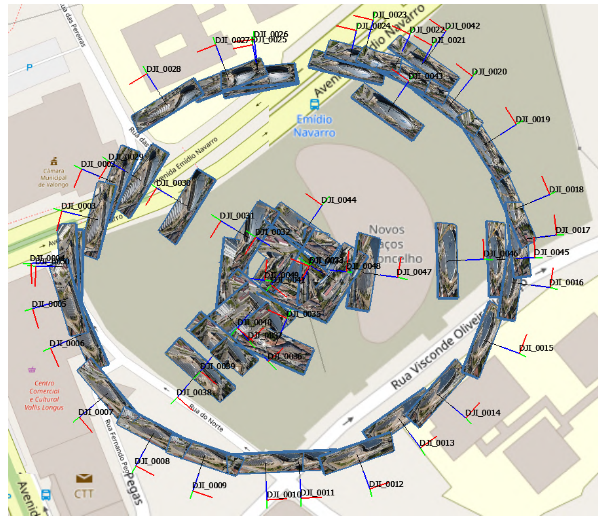

Figure 4: Position and orientation of the photos overlayed on OpenStreetMap (aptly edited by the author).

2.3 Processing

2.3.1 Software Specifications

The first software used was Aerial Model (accessible online at www.aerialmodel.com), since it has a free plan and a simple interface. However, Aerial Model's free plan only allows the usage of 25 Images, and its "simple interface" means the whole process is automatic (i.e. non-customizable). Still, the results were a good indication that this entire endeavor was possible, despite some misalignment of images, as shown in Figures 5, 6 and 7 on the next page.

The second software used was Agisoft Metashape 2.3.0 Professional Edition, which includes a free trial period. Unlike Aerial Model, Agisoft Metashape does not include the option to generate the model with one click, requiring more familiarity with the theory behind the algorithms, in order to achieve better results. A step-by-step description of the processing choices follows in the upcoming sections.

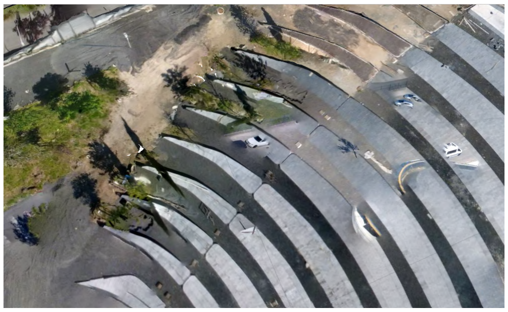

Figure 5: Instances of cars in the City Hall Square as an example of texture failure after poor image alignment in Aerial Model.

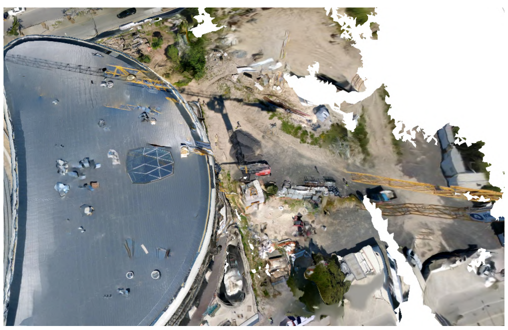

Figure 6: Projection of the construction crane on the building's roof and the floor as an example of texture failure after poor image alignment in Aerial Model.

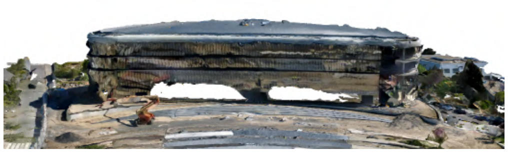

Figure 7: Large holes in the building as an example of texture failure after poor image alignment in Aerial Model.

2.3.2 Generating Tie Points

After loading all 49 images into a new project, the first step was running "Estimate Image Quality", which analyzes contrast between pixels. With all images having quality values between 0.8 and 1.0 (out of 1.0), no image was discarded on this basis.

The second step was to "Align Photos", accessible in the "Workflow" tab. As explained in Section 1, the no location data is not mathematically necessary[1], but it mitigates errors, which is why it was used. The Highest Accuracy was selected, and the "Adaptive camera model fitting" option was also enabled. As for the "Key Point Limit" and the "Tie Point Limit" (which defines how many key points are used across multiple photos), both were kept at their default values of 40,000 and 4,000, respectively.

The result was a Point Cloud with 59,146 points.

2.3.3 Cleaning Tie Points

High Quality of Images does not translate to flawless Tie Points. Therefore, it is good practice to clean them up a little, by following "Tools -> Tie Points -> Clean Tie Points" (called "Gradual Selection" in older versions of the program). Table 1 on the following page shows the applied filters and the respective number of points left afterwards.

[1]In fact, Agisoft Metashape Standard Edition doesn't even have the option to use it, but the Professional Edition does

Filter	Points
Reconstruction Uncertainty <30	50,860
Reprojection Error <0.45 pixels	48,400
Projection accuracy <6	43,905

Table 1: Applied filters and number of points remaining after their successive implementation.

After having removed about a quarter of the points, the next step was to "Optimize Camera Alignment" (on the Reference Tab), with "Adaptive camera model fitting" enabled once more. With this alone, it is already possible to create a tolerable Textured Model in a few seconds, but the goal is to reach something that can be compared to the 50 cm DEM, and that is not possible yet.

2.3.4 Building Dense Cloud

Following the cleaning of Tie Points, it was time to "Build [a] Dense Cloud" (also accessible in the "Workflow" tab, and called "Build Point Cloud" in newer versions of the program). If no model has been created yet, the only possible data source are Depth Maps (which will be generated during the process), and, thus, that was the chosen option. Once more, the highest quality was chosen. In the advanced options, the depth filtering was left as "Mild", and the options to "Calculate point colors" and "Calculate point confidence" were enabled.

The result was a Dense Point Cloud with 53,697,412 points, which was subjected to one final cleaning step, by doing "Tools -> Point Cloud -> Filter by Confidence", selecting [0,1] and deleting those points.

2.3.5 Building Model and DEM

A high-resolution DEM can be built directly from the Dense Point Cloud (by clicking "Build DEM", in the "Workflow" tab), but after all these steps it would be unreasonable not to build a textured model to visualize the results.

The newly created model's 13,567,370 faces contrast with the 251,146 faces generated by Aerial Model; and Figures 8, 9 and 10 can be compared with Figures 5, 6 and 7 on the previous page for quality control.

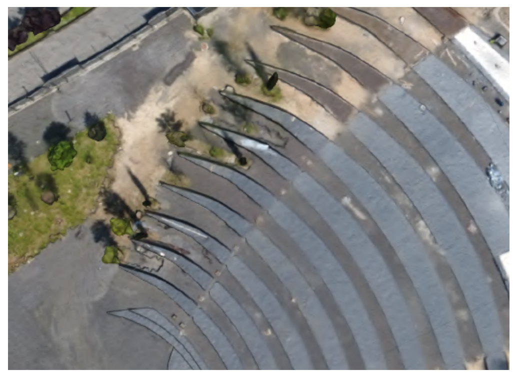

Figure 8: View of the City Hall Square after (more) successful image alignment.

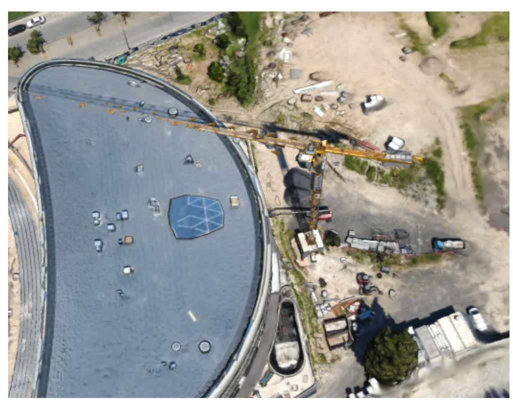

Figure 9: View of the construction crane after (more) successful image alignment.

Figure 10: View of the building's main facade after (more) successful image alignment.

The DEM can then be built from the Dense Point Cloud, from the Model, or directly from the Depth Maps, which provide the best resolution (theoretically lower than 5cm).

3 Results and Discussion

The Digital Elevation Model, represented in Figure 11, was generated with a resolution of 5cm from the Depth Maps. However, this doesn't mean it has the same accuracy everywhere. The areas around the city hall building have a very high accuracy, while the areas further away from it (especially those facing the opposite direction) have large errors; on the other hand, simple shapes like the building roof and the road are easier to map and thus also have higher accuracies than complex objects, like the construction crane and tree canopies.

After making an isolated qualitative analysis and verifying that everything looks as expected, it's important to conduct a comparative analysis, namely comparing the results to the DGT 50 cm DEM, illustrated in Figure 12.

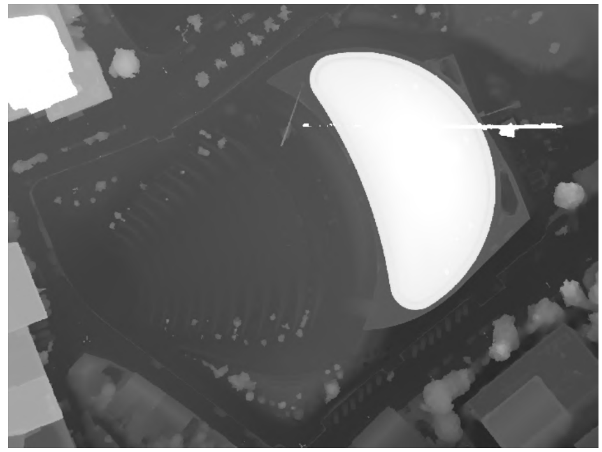

Figure 11: 5-cm Digital Elevation Model generated using Drone Footage through SfM. The whitest the color, the most elevated it is.

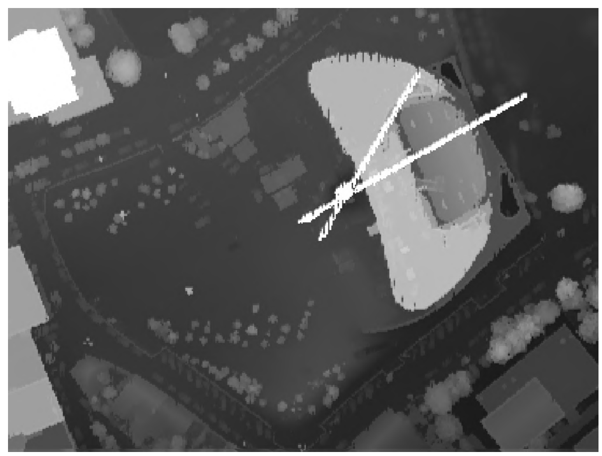

Figure 12: 50-cm Digital Elevation Model generated through airborne LIDAR. The whitest the color, the most elevated it is.

At a first glimpse, everything appears to be according to the expected:

- the DGT DEM is coarser in resolution
- all the buildings and trees appear to be in the right place and with similar relative height;
- the tallest structures are the "skyscrapper" on the top-left corner and the crane (which has moved position from on image to the other)
- the street on the lower side of the image appears to be slightly less elevated than the one on the top side
- the New City Hall building grew in height in the 2 years that passed between images

There's more to the results than meets the eye, however: when looking at the absolute values of each image, the DEM generated through SfM has heights varying between 41 and 97 meters above sea level; while the DGT DEM (correctly) has heights varying between 133 and 197 meters above sea level.

This is not a processing mistake, as the elevation values come from the images' metadata, which states the camera was 122 meters high. As mentioned in Section 2.2, the drone was flown at 70 meters above ground, which would give the point of lift-off an elevation of 52 meters, exactly what is shown in the result.

The cause of this error is not entirely clear, but it is most likely due to a miscalibration of the heigh reference system in the drone systems. That being said, that doesn't mean it can't be fixed; it should suffice to increase all the values a certain threshold. Yet, identifying said threshold is not so straight-forward.

With the DGT DEM as a ground truth, the most tempting way of find the threshold is comparing extrema (either minima or maxima). The problem with this is that the image acquisition did not anticipate this problem and, thus, the skyscraper in the top-left corner was not thoroughly photographed, meaning the values might not be the most accurate. On the other hand, the minimum values appear to be in the holes behind the new City Hall Building, which either got filled up between images or were also not thoroughly photographed, resulting in different relative height values. With anything inside the construction area inadmissible due to severe changes between the images, there was one good candidate left: the streets.

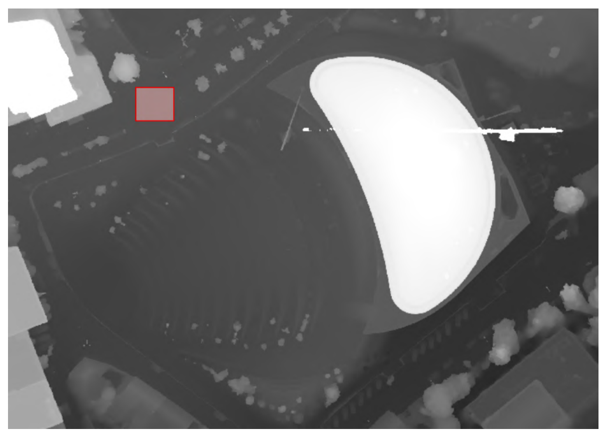

Figure 13: Area where the mean value was calculated (in red), overlayed on the 5-cm DEM generated using Drone Footage through SfM.

By analyzing a section of the street (illustrated in Figure 13), it was ascertained that it had a mean elevation of 143 meters, while the DEM generated through Structure-from-Motion recorded a mean elevation of 53 meters in the same area. The final DEM, seen in great detail in Figure 14 on the following page, was generated by adding 90 meters to all the values.

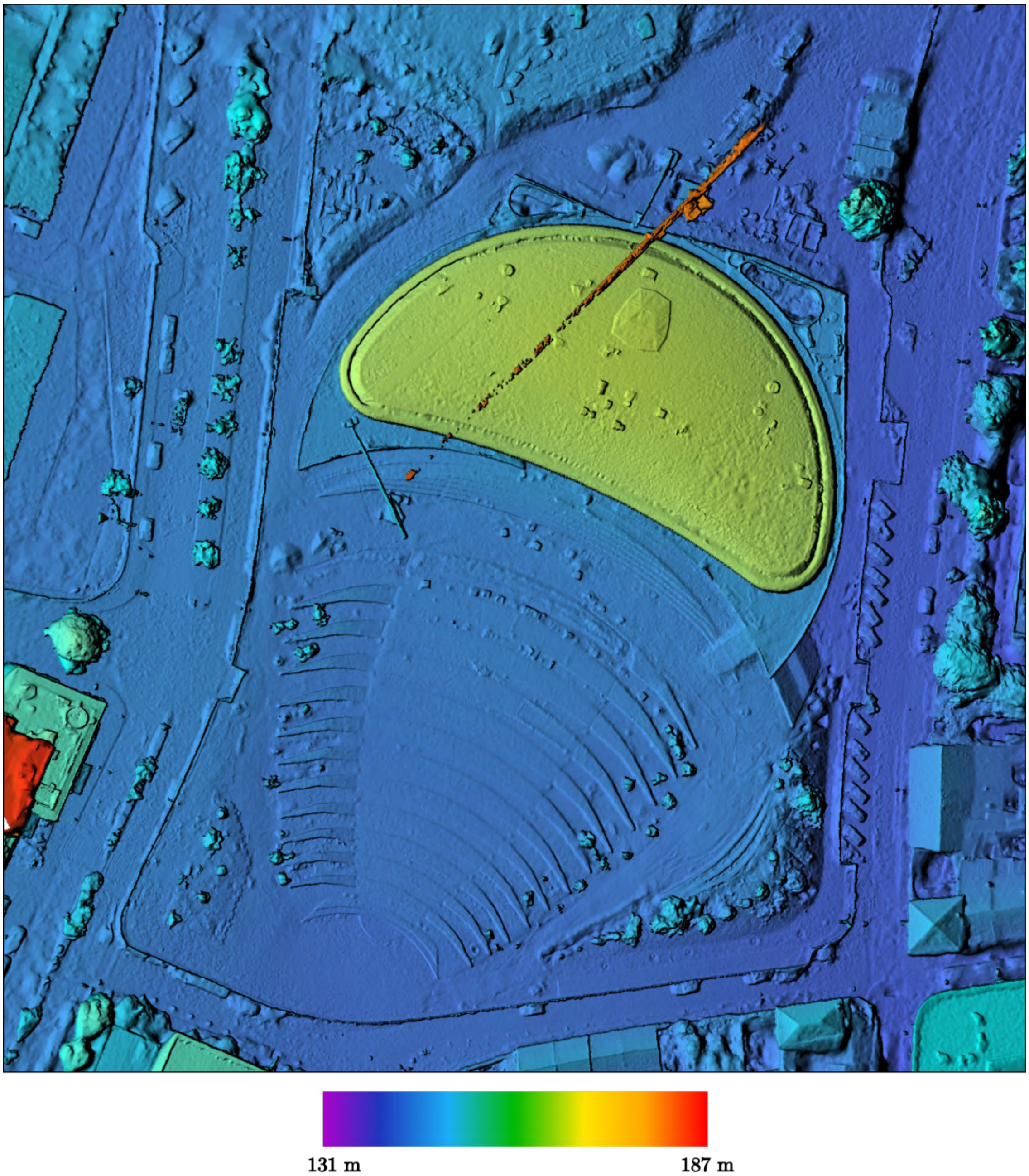

Figure 14: Final Digital Elevation Model of the Area of Interest represented with shaded relief with colors closer to blue representing lower values than red hues.

4 Conclusion

The Structure-from-Motion technique is a powerful tool at the disposal of anyone with a low-end recreational drone. The freely available software, such as Aerial Model's free plan, has its limitations, but, perhaps, with more carefully planned photo sessions, it can yield adequate results too.

Nevertheless, the true potential of the technique only shines in paid proprietary software such as Agisoft Metashape, where the ability to visualize every step of the process and filter through both the tie point cloud and the dense point cloud yield almost unbelievable results.

After correcting an (unfortunate) height error, the initial goal to have up-to-date data of the area was accomplished, and it was possible to ascertain that, in the time between the two images, the new City Hall building grew by a total of 8 meters, with its (nearly complete) roof now standing at an elevation of 164 meters above sea level, and a height of 21 meters above the ground.

References

James, M. R. and Robson, S. (2012). Straightforward reconstruction of 3d surfaces and topography with a camera: Accuracy and geoscience application. *Journal of Geophysical Research: Earth Surface*, 117(F3).

Lowe, D. (1999). Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157 vol.2.

Ullman, S. (1979). The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 203(1153):405–426.

Wallach, H. and O'connell, D. N. (1953). The kinetic depth effect. *J. Exp. Psychol.*, 45(4):205–217.

Westoby, M., Brasington, J., Glasser, N., Hambrey, M., and Reynolds, J. (2012). 'structure-from-motion' photogrammetry: A low-cost, effective tool for geoscience applications. *Geomorphology*, 179:300–314.