# Critical Examination of 3D Building Modelling through UAV Frame and Video Imaging

Hassan Rezvan [2], Elham Rezagholi [2], Masood Varshosaz [1,2]

[1] Institute of Artificial Intelligence, USX, Zhejiang, China – varshosazm@usx.edu.cn, varshosazm@kntu.ac.ir
[2] Geomatics Engineering Faculty, K. N. Toosi University of Technology, Tehran, Iran –
(h.rezvan, eli.rezagholi)@email.kntu.ac.ir,

**Keywords:** Photogrammetry, Videogrammetry, Unmanned Aerial Vehicle, 3D Modelling.

**Abstract**

Data capture in UAV photogrammetry is carried out using two main methodologies: frame-based and video-based. Frame-based data gathering is the preferred method among UAV projects because to its inherent reliability in calibration. Nonetheless, circumstances involving moving objects or occlusions inside the measured region may produce unsatisfactory results utilizing this method. In response to these challenges, video-based data collecting appears as a potential option, capable of creating a series of successive images that together alleviate the constraints outlined above. In this study we aim to compare the usefulness of frame and video images in building 3D models, using both oblique and vertical image orientations. Rigorous evaluations produced many outputs, including dense point clouds, digital surface models (DSM), meshes, and orthophotographs. The evaluation criteria included data acquisition velocity, processing efficiency, calibration precision, distortion analysis, residual plots, scale correctness, and reprojection error. The empirical results demonstrated the advantages of video-frame captures in improving the quality of the resulting 3D models. Notably, the use of video frames resulted in a significant reduction in reprojection error by 16%, calibration residuals by 36%, distortions by up to 51%, and processing time by 27%. Thus, it seems that integrating video frames improves data gathering accuracy and speeds up processing, replacing standard frame images with video counterparts for increased efficacy.

## 1. Introduction

The field of photogrammetry is witnessing significant advancements in 3D modelling (Yalcin and Selcuk, 2015), a process that involves creating a three-dimensional representation of an object from two-dimensional images (Karami et al., 2023). This technology has found applications across various domains, including surveying engineering, road pavement monitoring, digital terrain modelling, as-built surveying, quality control, building information models (BIM), and computer-aided design (CAD) (Alsadik and Khalaf, 2022). Additionally, it plays a crucial role in medical fields such as prosthesis creation, disease diagnosis, and dental reconstruction (Lerma et al., 2018). Beyond healthcare, 3D modelling is instrumental in cultural heritage restoration, historical documentation, digital preservation, conservation, virtual reality motion graphics (Remondino, 2011, Marshall et al., 2019, Herraez et al., 2021), and urban design and management (Yalcin and Selcuk, 2015, Alsadik and Khalaf, 2022). These applications underscore the versatility and importance of 3D modelling in today's technology landscape.

3D modelling primarily involves capturing still images with a certain degree of overlap to facilitate the creation of three-dimensional representations of objects. Studies suggest that achieving an overlap of 80% in both side and end views is adequate for generating 3D models (Alsadik and Khalaf, 2022). However, this method is not without its challenges. Issues such as occlusion, where parts of the object are not visible due to being blocked by other parts (Nunes, 2010), and the distortion caused by moving objects, which can lead to poor matching and stretching in images, can compromise the quality of the 3D models. These problems can result in a reduction of information content within the dataset images, leading to weaker 3D modelling outcomes.

The advancements in video resolution, utilizing both consumer and professional cameras, have expanded the potential applications of 3D modelling, including its use in sports event broadcasting and cinematography (Alsadik and Khalaf, 2022). It is anticipated that fields such as geoinformation science, photogrammetry, and image-based 3D city modelling will also benefit from these technological advancements. A notable development in this area is videogrammetry, a technique that involves capturing video of an object to create 3D models (Singh et al., 2014). Videogrammetry, which extracts 3D coordinates over time, facilitates the acquisition of multitemporal data for dynamic objects (Lerma et al., 2018). This method is becoming increasingly popular for extracting 3D models due to its flexibility and efficiency compared to traditional still photography. Video capture allows for continuous recording without the need to adjust the camera shutter speed or determine optimal waypoints along flight trajectories (Shilov et al., 2021, Alsadik and Khalaf, 2022), addressing the complexity of determining the optimal number of photos and angles for a satisfactory 3D model (Shilov et al., 2021).

Recently, great attention has been paid to using drones for mapping purposes. To this end, in this paper, our focus is on evaluating the effectiveness of using frame and video images captured from Unmanned Aerial Vehicles (UAVs) in constructing 3D models, utilizing both oblique and vertical image orientations. The process generates various outputs, including dense point clouds, Digital Surface Models (DSM), meshes, and orthophotos. These outputs are then assessed through a combination of visual comparison, data acquisition and processing time, and statistical evaluations.

The rest of the paper is structured as follows. In Section 2 we give an overview of some of the related works in this area before outlining our materials and proposed method in Section 3. Section 4 shows the results of our method when applied to a large dataset. Finally, we in Section 5 conclusions are reached and outlined.

## 2. Related works

The use of video frames in 3D modelling of objects has already been investigated in several studies. Also, comparing the results of 3D modelling using both image and video frames has been demonstrated previously.

One of the main applications of 3D modelling is DEM generation. Kwasnitschka et al. (2013) exploited the tremendous potential of Remotely Operated Vehicles (ROVs) in video and photographic data mining by observing and sampling seafloor. They presented a new workflow to create synthetic model visualizations of the sea floor with the aim of deriving fundamental field geology information such as quantitative stratigraphy and tectonic structures from ROV-based photo and video materials. Bhushan et al. (2021) developed an automated, open-source workflow to refine the SkySat camera models and improve absolute geolocation accuracy using external reference DEMs. These refined cameras are then used to produce accurate DEM and orthoimage composites from both the SkySat triplet stereo and video products.

As stated before, another 3D modelling usage is in medical science. Lerma et al. (2018) investigated the suitability of smartphone video cameras to create 3D models for cranial deformation analysis compared to digital single-lens reflex (SLR) cameras traditionally used in close-range photogrammetry. Two models were obtained, the first one from slow-motion video recorded from a smartphone, and the second one from SLR camera imagery. They evaluated the results of two models with themselves and with the best-fitting ellipsoid that allows the determination of the cranial deformation. The average distance between models were 0.5 mm and below 1 mm for 86% of the model points. Also, Shilov et al. (2021) emphasized the use of modern computer vision algorithms, photogrammetry and machine learning methods to create 3D foot models based on video streams obtained from a smartphone. Their proposed method has a minimum linear deviation of 0.95 mm in foot length and width.

In the field of cultural heritage restoration, the study of Herraez et al. (2021) showed the restoration by video projection of the vault paintings of a church in Valencia that were destroyed by fire. For this, they used two black and white analogue frames taken before the fire. Considering the ceiling of the case study was an irregular hemispherical vault, they projected images on its surface with video cannons in original positions and without metric deformation. To define geometry with the greatest accuracy that each partial image is projected on the real surface, they first calculated a complete 3D virtual image of paintings on the mathematical modelling of this irregular surface and then calculated 3D partial virtual images.

Alsadik and Khalaf (2022) evaluated the potential use of blur-free drone ultra-high-definition videos for 3D city modelling. In their research, the impact of using UHD video cameras onboard drones was investigated on the 3D reconstructed city models. In their experiment, it was shown that increasing the video resolution not only improved the density but also the internal and external accuracies of the created 3D models. According to the results, the point density and reconstruction accuracy improved up to 90% when using 8K compared with the HD videos taken from the same drone. Additionally, GSD was improved four times when the 8K image resolution was used compared with the HD resolution while maintaining the same flying height.

In the extension of the above approaches, we aim to generate a 3D model of the same building by both frame-based and video-based approaches. Both datasets are captured by UAV. Then two sets of outputs are compared using evaluation criteria.

## 3. Materials and method

### 3.1 Study area

The study object to model is the building of the dormitory of technical school of Loshan city located at 49°31’35” E and 36°37’14” N in Rudbar city, Gilan province (Figure 1).

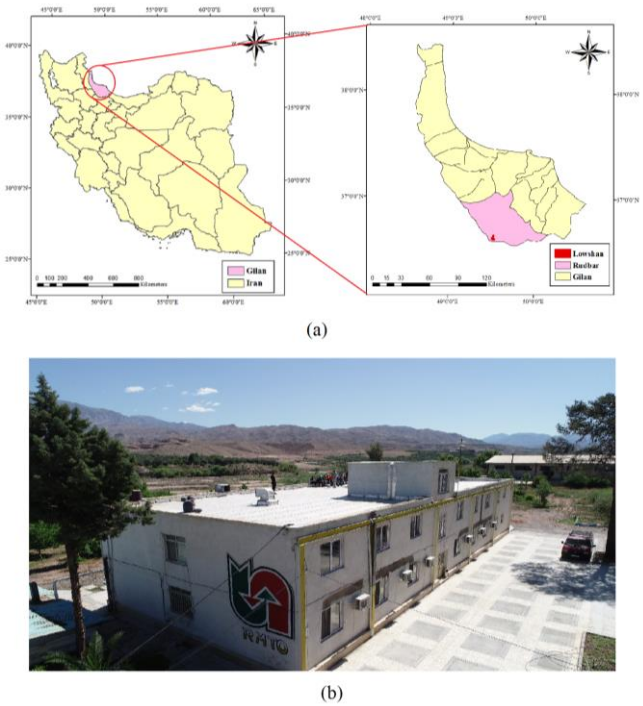



Figure 1. Study area; a) Location b) Study building

### 3.2 Data

In this study we used a Phantom 4 Pro drone equipped with a 8.8 mm focal length and a 2 cm pixel ground sampling distance (GSD) to gather the data, including both vertical and oblique images. The drone's capability to record 4K quality video is also leveraged for capturing video footage of the target object. For the generation of 3D models, it's essential to have control and check points, which were established using four points measured by a total station. This setup ensures the accuracy and reliability of the 3D modelling process. More details of Phantom 4 Pro are shown in Table 1.

| Item | Value |
|---|---|
| Resolution | 20 MP |
| Image size | 5472*3648 |
| Pixel size | 2 cm |
| Flight height | 500 m |
| Flight range | 7 Km |
| Focal length | 8.8 mm |
| Max flight time | ≈ 30 min |
| Video quality | 4K |

Table 1. Technical details of Phantom 4 Pro

### 3.3 Methodology

Our workflow is illustrated in Figure 1. Two sets of datasets were acquired by DJI Phantom4Pro UAV.

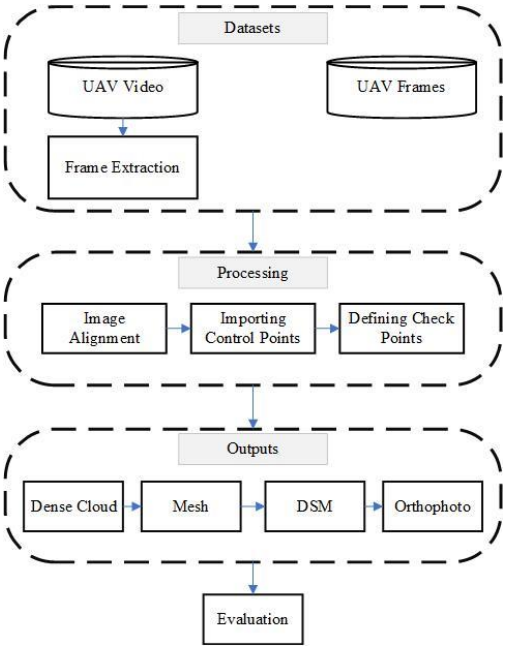



Figure 2. Proposed workflow

Flight planning is crucial for ensuring the success of UAV missions and meeting the objectives of mapping projects. Various flight planning strategies can be employed based on the specific requirements of the flight and the area being surveyed (Alsadik and Khalaf, 2022). Considering the desired object, which is a building in our case, circular flight was selected as flight planning for taking several 2D images and video recording, because it allows the capturing of stereoscopic photographs, supporting stereo restitution (Gómez-López et al., 2020).

Both datasets were captured by a pilot, manually with the frame ones taken at an overlap approximately equal to 80%. Also, the video of the building's surroundings was taken in both vertical and oblique modes. Since the frames were mostly captured at a high rate of 25 frames/sec, a frame sampling process was implemented at fixed intervals to ensure the overlap was sufficient for 3D modelling (Alsadik and Khalaf, 2022). Therefore, video was imported into MATLAB environment and its properties were extracted, revealing a frame rate of 23.9 frames/sec and a total of 4666 frames. By selecting every 40th frame, the dataset was reduced, which helped in avoiding data redundancy, reducing processing time, and improving the quality of the 3D model by filtering out blurry images to reach a better 3D model and have a geometrically stronger configuration (Alsadik and Khalaf, 2022). Due to the variable speed of the UAV during video recording, some frames were very similar, especially in areas where the drone moved slowly. Consequently, frames with more than 80% coverage or duplicates were manually removed to prevent the creation of redundant data. Ultimately, 140 images were extracted in both oblique and vertical modes, ready for further processing.

After creating the two datasets, they were imported into Agisoft Metashape software as separate files. The initial stage involved aligning the photos, which was accomplished through camera calibration and the formation of a sparse point cloud using detected key points and image matching. As previously mentioned, four control points at the corners of the building were measured using a total station. The subsequent stage involved importing these control points into the software, considering five scenarios: four cases where one point was used as a check point and the others as control points, and one case where all points were considered control points.

To evaluate the results, it is essential to generate various products and compare them. Therefore, after creating a sparse point cloud, we generated the depth maps from the images in both datasets. These depth maps were then used to create a dense point cloud, offering a detailed representation of the object and its surroundings. Subsequently, a mesh, Digital Surface Model (DSM), and orthophoto were also produced.

In the end, to evaluate the two approaches, the evaluation criteria were:

- Data acquisition velocity
- Processing efficiency
- Scale correctness
- Calibration precision
- Distortion analysis
- Residual plots
- Reprojection errors

## 4. Results

Sample outputs are illustrated in Figures 3 to 6.

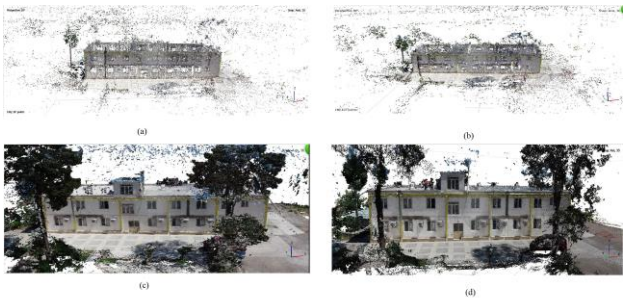
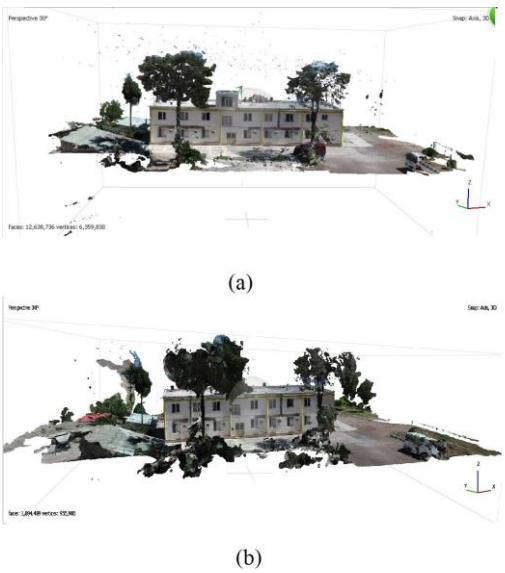


Figure 3. Sparse and dense point clouds: The first and second columns are related to frame and video images respectively.



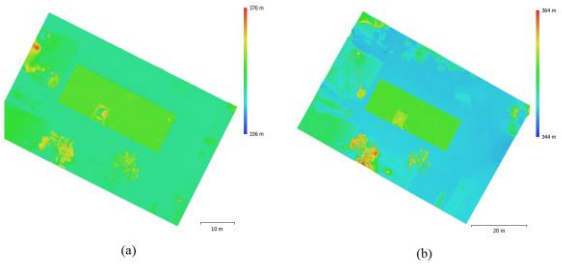
(a)

(b)

Figure 4. Mesh: a: frame image; b: video frame

(a) (b)

Figure 5. Mesh: a: frame image; b: video frame



(a) (b)

Figure 6. DSM: a: frame image; b: video frame

(a) (b)

Figure 6. Orthophoto: a: frame image; b: video frame

The number of tie points identified in frame mode model was 148107 and in video mode was 71613.

By visually comparing the products obtained from frame images and video, it can be seen that frame products have a higher quality representing fewer gaps and visual blunders than the products obtained from video images.

In addition to the visual evaluation, other criteria have been considered to evaluate the error and accuracy. One of the evaluation criteria considered to compare the models obtained from two datasets was control scale bars, the results of which are shown in Table 2.

| Scale item | Distance (m) | Error (mm) | |
|---|---|---|---|
| | | Frame | Video |
| 1-2 | 33.2 | -2.5 | 2.4 |
| 2-3 | 12.47 | -6.4 | 1.9 |
| 3-4 | 33.18 | 3.4 | 5.4 |
| 1-3 (Check) | 12.47 | 3.2 | 6.0 |
| Control scale bars | ---------- | 4.2 | 3.6 |
| Check scale bars | ---------- | 3.2 | 6.0 |

Table 2. Scale bar accuracy evaluation.

According to Table 2, three scale distances are considered as the control scale, and one scale distance between control points 1 and 3 is considered as a check, it can be seen that the amount of error is 2.7 mm in the video image more than that of the frame mode. The difference does not suggest any difference in terms of scale change in both models, as 2.7mm is not notable for the map scale of these evaluations. Therefore, to better evaluate the accuracy of the models, we decided to investigate the accuracy of some control and check points. Thus, the next criterion that was considered for evaluation was the accuracy assessment of control and check points. The results of this evaluation for different modes are given in Table 3.

| #GCP-#CP-No.CP | Points type | Frame | | Video | |
|---|---|---|---|---|---|
| | | Error (mm) | Error (pix) | Error (mm) | Error (pix) |
| 4-0-0 | Control | 38 | 0.588 | 26 | 0.537 |
| 3-1-1 | Control | 36 | 0.583 | 21 | 0.54 |
| | Check | 82 | 0.542 | 110 | 0.215 |
| 3-1-2 | Control | 26 | 0.43 | 12 | 0.517 |
| | Check | 11 | 0.798 | 143 | 0.456 |
| 3-1-3 | Control | 19 | 0.603 | 12 | 0.526 |
| | Check | 111 | 0.55 | 140 | 0.364 |
| 3-1-4 | Control | 22 | 0.747 | 475 | 0.0217 |
| | Check | 116 | 0.252 | 628 | 0.0939 |

Table 3. Control and check points errors.

As the number of ground points in our case was limited, we took three points as control and one point as check. The first digit in the first column of the Table 3, shows the number control points, while the second digit represents the number of check point which is in all rows equal to 1. The third digit shows the number of the check point used to evaluate the accuracy of the resulting model. As can be seen, for almost all rows, the accuracy of the control points is a relatively low value and not much different in both frame and video-based models. This was expected as the photogrammetric model always fits itself to the base control coordinate system, almost regardless of the quality of the model. However, it is observed that that in all check points 1 to 4, the error in the video-based mode is notably larger than the corresponding frame-based model. This suggests the frame-based model is established more accurately. This could stem from the fact that the videos are less stable compared to frames and, thus, leading to more accurate models.

To address the variability in error levels due to differences in the number of control and check points, an alternative metric, reprojection error, was employed to assess accuracy. This metric showed values of 1.7 pixels for frame images and 0.897 pixels for video images, with frame images exhibiting a higher error

rate. Another criterion considered for evaluation is the distortions resulting from two models. Distortion plots are illustrated in Figure 3.

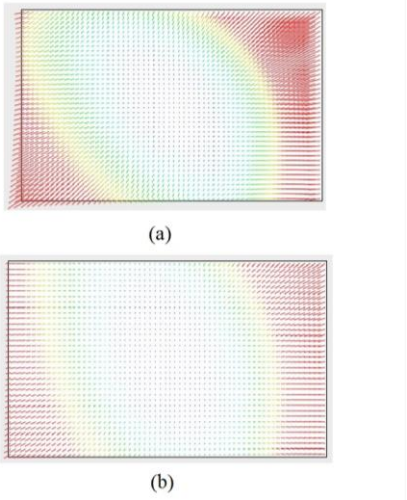



Figure 7. Distortion plots: a) frame image; b) video frame

Figure 7 illustrates that distortions are more on the sides of the image, reflecting the typical distortion behaviour in images. However, frame images exhibited greater distortions compared to video images. The highest distortion was recorded as 11.2 pixels in frame images and 5.25 pixels in video images. Furthermore, the presence of large residuals, particularly in the centre of the image, was more significant in frame images compared to video images.

Also, the residual plot is considered as another evaluation criterion. Residual plots are illustrated in Figure 8.

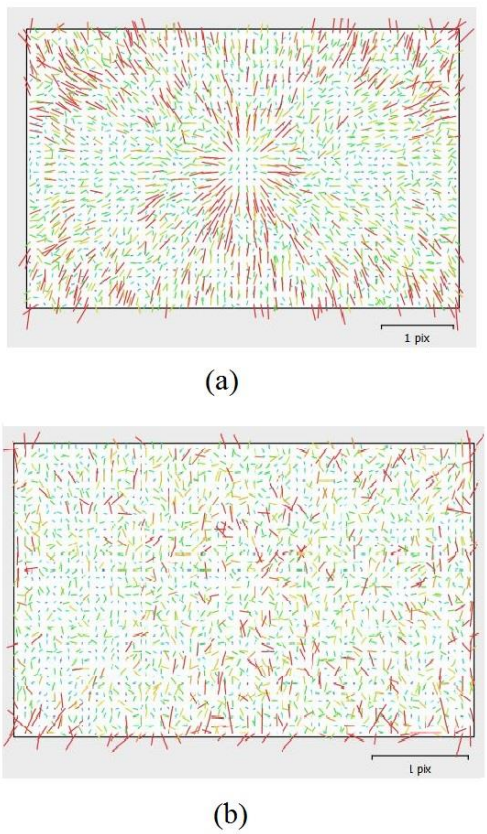



Figure 8. Residual plots: a: frame image; b: video frame

According to Figure 8, it can be seen that the frame mode has shown a larger residual amount compared to the video mode. The maximum remaining value was 0.111 pixels in frame images and 0.0707 pixels in video images. In addition, the density of large residuals in parts of the image, including in the center of the image, was higher in the frame images mode compared to the video images mode. In video mode, larger residual vectors are observed on the sides of the image than in the centre.

As mentioned earlier, calibration accuracy has also been one of the evaluation criteria. The information related to this criterion is given in Table 4.

| Parameters | Error | |
|---|---|---|
| | Frame | Video |
| F | 0.3423 | 0.2427 |
| Cx | 0.1657 | 0.0881 |
| Cy | 0.2477 | 0.1101 |
| B1 | 0.3409 | 0.2395 |
| B2 | 0.1864 | 0.0827 |
| K1 | 0.0003 | $5.6*10^{-5}$ |
| K2 | 0.0010 | 0.0001 |
| K3 | 0.0011 | 0.0001 |
| P1 | $3.7*10^{-5}$ | $7.9*10^{-6}$ |
| P2 | $1.9*10^{-5}$ | $3.8*10^{-6}$ |

Table 4. Calibration parameters errors

According to Table 4, the error value for all calibration parameters is higher in frame images concerning its value in video ones.

Another criterion for comparing frames and video modes is the total processing time allocated to each approach. This time is 17 hours and 16 minutes in frame image mode and 14 hours and 7 minutes in video image mode. The longer processing time in frame image mode can be attributed to the number of tie points detected in frame mode compared to video and their processing.

| Stage | Frame (hour, min, sec) | Video (hour, min, sec) |
|---|---|---|
| Align | 0,8, 6 | 0,9, 34 |
| Dense point cloud | 16, 3, 33 | 12, 29, 23 |
| Texture | 0,14, 31 | 1, 9, 36 |
| DSM | 0,1, 40 | 0, 0,17 |
| Orthophoto | 50,0,14 | 0,30, 22 |
| Total | 17, 16,0 | 14, 7 |

Table 4. Calibration parameters errors

In addition, one of the challenges of working with video frames is the lack of video support in Agisoft Metashape software and the extraction of desired frames with the help of other software such as MATLAB. Also, the method of sampling and extracting frames depends on the speed of the UAV, which sometimes requires a human operator to be involved into this process. If the UAV movement speed is uniform, frame extraction can be done automatically, otherwise, like our case, the initial frames are extracted automatically, and then duplicate frames without new information are removed manually. Furthermore, the quality of the extracted frames in the in some images in video mode was lower than the frame images, which made it difficult to mark the control points.

## 5. Conclusion

The method of obtaining data in photogrammetry can be divided into two modes: frame-based and video-based. The purpose of this research was to compare the results of the 3D modelling of the building in these two cases based on different

criteria, including visual interpretation, statistical evaluation, and hardware performance. For this, criteria such as data acquisition velocity, processing efficiency, scale correctness, calibration precision, distortion analysis, residual plots and reprojection errors have been taken into account. In the case of modelling by video frames, we observed a reduction of reprojection error by 16%, calibration residuals by 36%, distortions by up to 51%, and processing time by 27% compared to frame images. On the other hand, in terms of visual evaluation, the model obtained from frame images is a more integrated model than the model of video images due to more tie points extraction. Finally, according to the cases examined in this research, mentioned advantages and drawbacks of employing video frames in Section 4, replacing frame images with video images can lead to an improvement in modelling accuracy and processing speed.

## Acknowledgements

We would like to express our sincere gratitude to Aseman Negar Tin Company, for their invaluable assistance in acquiring UAV data for our research. Their expertise and support significantly contributed to the successful execution of our project. Their commitment to advancing the field of UAV technology and their willingness to share their knowledge and resources have been instrumental in our progress. We are deeply appreciative of their collaboration and look forward to future opportunities to work together.

## References

Alsadik, Bashar, and Yousif Hussein Khalaf. 2022. “Potential Use of Drone Ultra-High-Definition Videos for Detailed 3D City Modelling.” ISPRS International Journal of Geo-Information 11 (1): 34. https://doi.org/10.3390/ijgi11010034.

Bhushan, Shashank, David Shean, Oleg Alexandrov, and Scott Henderson. 2021. “Automated Digital Elevation Model (DEM) Generation from Very-High-Resolution Planet SkySat Triplet Stereo and Video Imagery.” ISPRS Journal of Photogrammetry and Remote Sensing: Official Publication of the International Society for Photogrammetry and Remote Sensing (ISPRS) 173: 151–65. https://doi.org/10.1016/j.isprsjprs.2020.12.012.

Herraez, Jose, Jose L. Denia, Enrique Priego, Pablo Navarro, Maria T. Martin, and Jaime Rodriguez. 2021. “Cultural Heritage Restoration of a Hemispherical Vault by 3D Modelling and Projection of Video Images with Unknown Parameters and from Unknown Locations.” Applied Sciences (Basel, Switzerland) 11 (12): 5323. https://doi.org/10.3390/app11125323.

Gómez-López, José Miguel, José Luis Pérez-García, Antonio Tomás Mozas-Calvache, and Jorge Delgado-García. 2020. “Mission Flight Planning of RPAS for Photogrammetric Studies in Complex Scenes.” ISPRS International Journal of Geo-Information 9 (6): 392. https://doi.org/10.3390/ijgi9060392.

Karami, A., M. Varshosaz, F. Menna, F. Remondino, and T. Luhmann. 2023. “Fft-Based Filtering Approach to Fuse Photogrammetry and Photometric Stereo 3d Data.” ISPRS Annals of Photogrammetry Remote Sensing and Spatial Information Sciences X-4/W1-2022: 363–70. https://doi.org/10.5194/isprs-annals-x-4-w1-2022-363-2023.

Kwasnitschka, Tom, Thor H. Hansteen, Colin W. Devey, and Steffen Kutterolf. 2013. “Doing Fieldwork on the Seafloor: Photogrammetric Techniques to Yield 3D Visual Models from ROV Video.” Computers & Geosciences 52: 218–26. https://doi.org/10.1016/j.cageo.2012.10.008.

Lerma, José Luis, Inés Barbero-García, Ángel Marqués-Mateu, and Pablo Miranda. 2018. “Smartphone-Based Video for 3D Modelling: Application to Infant’s Cranial Deformation Analysis.” Measurement: Journal of the International Measurement Confederation 116: 299–306. https://doi.org/10.1016/j.measurement.2017.11.019.

Marshall, M. E., A. A. Johnson, S. J. Summerskill, Q. Baird, and E. Esteban. 2019. “Automating Photogrammetry for the 3d Digitisation of Small Artefact Collections.” ISPRS - International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences XLII-2/W15: 751–57. https://doi.org/10.5194/isprs-archives-xlii-2-w15-751-2019.

Nunes, Jorge Luís. “OCCLUSION DETECTION IN DIGITAL IMAGES THROUGH BAYESIAN NETWORKS.” (2010).

Remondino, Fabio. 2011. “Heritage Recording and 3D Modelling with Photogrammetry and 3D Scanning.” Remote Sensing 3 (6): 1104–38. https://doi.org/10.3390/rs3061104.

Shilov, Lev, Semen Shanshin, Aleksandr Romanov, Anastasia Fedotova, Anna Kurtukova, Evgeny Kostyuchenko, and Ivan Sidorov. 2021. “Reconstruction of a 3D Human Foot Shape Model Based on a Video Stream Using Photogrammetry and Deep Neural Networks.” Future Internet 13 (12): 315. https://doi.org/10.3390/fi13120315.

Singh, Surendra, Kamal Jain and Venkata Ravibabu Mandla. “3D Scene Reconstruction from Video Camera for Virtual 3D City Modelling.” (2014).

Yalcin, Guler, and Osman Selcuk. 2015. “3D City Modelling with Oblique Photogrammetry Method.” Procedia Technology 19: 424–31. https://doi.org/10.1016/j.protcy.2015.02.060.