The road toward smart city infrastructures: a review on 3d facade reconstruction using images

Youssef Arhrib*, *Omar* El Kharki, *Meriam* Wahbi, *Otmane* Yazidi Alaoui, *Mustapha* Maatouk, and *Hakim* Boulaassal

Geomatics, remote sensing and cartography Unit, FSTT, Abdelmalek Essaadi University, Tetouan, Morocco

Abstract. Numerous types of solid structures protect coastal activities and cities from the damaging effects of tides and waves all around the globe. Thus, having a three-dimensional digital representation of the physical environment would help decision-makers in understanding the dynamic nature of coastal environments and implementing effective mitigation strategies. Generally speaking, Feature matching, Structure from Motion (SFM) and Multi-View Stereo (MVS) algorithms are used in this order to achieve realistic results. The Literature shows that there is a constant evolution of new techniques and technologies either with learning based or hand-crafted approach, which gives a possibility to integrate different method to optimize each step of the three-dimensional reconstruction process. The aim of this paper is to present the progress of three-dimensional modelling methods that use ground-level images by providing an overview of the latest applications and a comparison of their results. Overall, the state-of-the-art in three-dimensional building modelling using ground-level imagery is rapidly evolving, and new ways are being developed to improve the efficiency, accuracy and scalability of the process
Keywords: Smart cities, 3D building model, SFM, MVS, 3D reconstruction.

1. Introduction

For several years, the scientific community has been working on the subject of three-dimensional modelling of objects from various types of data, whether aerial or terrestrial optical images, LIDAR data or RADAR images.

Recovering three-dimensional scenes using images as the main input and the principles of photogrammetry and computer vision techniques has become more preferred to generate a detailed three-dimensional model than laser scanning approach which is considered expensive and require user-knowledge for better results.

While photogrammetry helps to obtain accurate measurements of an object, computer vision would infer geometric and other properties from the three-dimensional space using either one or multiple images [1].

We live in a three-dimensional world, which is different from two-dimensional (2D) maps and drawings, for this reason, integrating three-dimensional models is a must in decision-making process and could have multiple applications in urban management and urban planning [2]. The fact that urban areas are mostly composed of buildings, leads numerous scientists last decades to investigate their time in enhancing efficiency of pipelines designated to three-dimensional building reconstruction.

Targeting building façade require images to be at a ground level, in recent years, terrestrial mobile mapping systems (MMSs) have been widely used to collect street-view images. Different scientists worked on improving methods used in the photogrammetric pipeline, by using computer vision techniques, starting from pre-processing images [3] to extracting key points [4], and applying structure from motion [5] as well as generating dense points clouds [6].

All the work that has been done achieved a high level of detail which is beneficial for multiple applications [7,8], but others could require simple models like level of detail (LOD) models [9].

Storing and exchanging three-dimensional models of man-made scenes is widely used by different communities, for that reason, CityGML standard is used to share three-dimensional scene models like in [10,11]. Generally, three-dimensional models are classified under CityGML to five classes, from the simple version LOD0 to a detailed model LOD4 (including indoor Features).

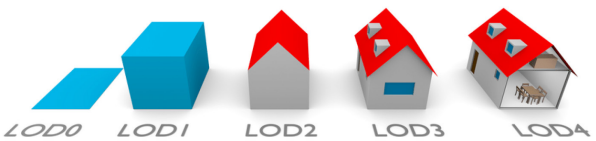

Fig. 1. LOD definition under CityGML 2.0 standard [11]

Simultaneously, the extensive use of machine learning and deep learning techniques in photogrammetry due to the increasing of computational power and the availability of image data helps in optimizing the pipeline components like segmenting three-dimensional heritage data [12] or removing outliers from point clouds [13].

* Corresponding author: youssef.arhrib@etu.uae.ac.ma

Compared with the three-dimensional reconstruction of building that depends on laser scanning, the image-based three-dimensional reconstruction of building has its specifics characteristics and deals with particular challenges. Regardless of, the several ways to get the images either by using monocular cameras, binocular cameras, or multi-cameras, challenges still similar. Shadows and intensity of light decrease the image quality, also dynamic and self-occlusion preclude researchers from observing the building [14].

The extraction of point cloud from images comes with so much noise and obstacle, deleting redundant points would help simplifying data processing efficiency.

Point cloud are known by having only three-dimensional coordinate information, in order to know from the three-dimensional model, the type of the building, labeling point cloud from the color and texture in RGB images is beneficial for achieving a better model. In order to bypass the above challenges many research have been carried out, although, most of the review articles tend to focus on one aspect of the process [15].

Therefore, this paper is aiming at identifying the relationship between methods by highlighting the advantages and limitations of each one; creating a research map as a reference for researchers with a discussion about the current challenges of image-based three-dimensional reconstruction of building and explore possible solutions.

2. Feature-based methods and dense methods

2.1. From multiple images

Recovering three-dimensional geometry from images requires camera poses to be estimated by calculating at least three feature points from an image under the condition that they are evenly distributed and their 3D coordinates are known. The relative orientation of two images can be estimated using five pairs of homologue image points [16].

For a larger number of images, methods have been implemented, that take a sequence of images from uncalibrated cameras and generate three-dimensional models [17] by tracking feature points across multiple views then through bundle adjustment methods relation between these views are calculated [18]. Those steps, would require a very short baselines between the images.

Agarwal and other scientists successed to recover automatically from large, unorganized images sparse scene geometry [19] by the use of structure from motion algorithm, key points at each image are extracted using SIFT [20] then for each pair of image fundamental matrix is estimated relying on the eight-point [21] and Ransac algorithm [22].

Subsequently, camera parameters and three-dimensional coordinates of the matched points are estimated using Levenberg Marquardt. Based on this research, Microsoft created Bundler an open-source software that rely on SfM using large unordered set of internet photos [23].

The same idea has been used by Furukawa ad his team where they introduce an approach for enabling existing multi-view stereo methods to generate a dense reconstruction of the target scene [24].

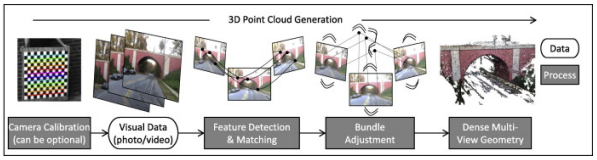

Fig. 2. dense reconstruction process [15]

Under Manhattan-world assumption on building regularity, generating automatically three-dimensional models from images of building facades captured along the streets gives good results [25]. At the same year, other scientists used a sequence of panoramic street view images as input under the piecewise planar structure constraint. [26]

Other researches focused on getting good results on texture-less areas by applying a quadtree structure to generate depth and normal hypothesis for every pixel [27].

2.2. From single image

Three-dimensional building reconstruction from single images captured substantial notice from the photogrammetry community; The literature shows that existing methods that use single images can be classified to either metric or nonmetric methods. Nonmetric reconstruction methods are used to generate three-dimensional model not geometrically accurate but visually close to the truth [28] [29].Saxena and *al.* in this article [28] worked on producing depth maps from single images by applying multiscale Markov random field to model the depths and the relation between depths at different image locations using dataset that consist of several images and their ground truth depth maps generated from a laser scanner. [23]

Metric reconstruction works on recovering a precise three-dimensional geometry of the target scene, but without having any assumptions about the scene structure, the use of a single image will generate a three-dimensional model that has several interpretations [23].These assumptions can be either model-based or constraint-based; In urban scenes, building have generally rectilinear or piecewise shapes which are used in model-based methods [30], the objective is that through a collection of primitive shapes we find the best fit to image scene structure. While constraint-based approach takes advantage of geometric properties of the image scene structure like planarity, parallelism, and orthogonality to recover three-dimensional scene geometry.

The main drawback of these methods is that model-based approach is limited to recovering scene structure that are stored in the collection of primitive

shapes, while constraint-based approach uses a limited number of geometric properties.

3. Challenges and Limitations

In order to make image-based three-dimensional modeling techniques a standard practice, method limitations need to be highlighted and addressed [31].

3.1. Full Automation

Literature proved that recovering a well-defined building structure from images is a complex task and many data processing techniques require setting up manually processing parameters. Thus, giving a low level of automation. Enhancing automation of three-dimensional building reconstruction is done through the minimization of human intervention which could be done by creating more flexible workflows.

3.2. Data collection

Obtaining a complete dataset of urban environment has its own difficulties, disparities in buildings dimensions lead to an inconsistent dataset, point cloud generated for tall building will be less dense in higher parts. Moreover, there is occlusion that happen when capturing building facades, camera vision is usually blocked by different objects that create gaps in data thus generating a mediocre reconstruction of the building facades. The occlusion could be either static when it is caused by vegetation or dynamic when pedestrians and vehicles obscure the captured view. Dynamic occlusions could be limited when collecting data at different periods of the day [32] while static occlusions is handled by introducing data from oblique cameras that show obscured parts of the building façade Camera is considered as an optical sensor, for this reason weather conditions is constantly changing image quality, when there is a bad lighting conditions the collected images are blurry and the generated point cloud is noisy.

3.3. Quality of the model

Due to the current level of automation in three-dimensional building reconstruction, there is a necessity in human intervention either in data collection, data processing techniques or evaluating final result. Every user has its own experience when he processes data, which could impact 3D

3.4. model quality

Consequently, creating a fully automatic data processing technique helps to minimize the biased intervention of users. The specific use case of 3D modelling building and the method of acquiring and processing data implies a unique level of accuracy and a unique level of detail model.

3.5. Point Cloud Processing Dilemma

In order to generate a dense 3D model, collected images must have sufficient overlaps to cover the target scene [33] and also there is a need to process tremendous amount of point cloud which will excess the use of computer resources and it will take too much time to remove noisy points. [34]

3.6. Prior Information Contribution

Occlusions make the process of obtaining geometric properties only through images a hard task, literature proved that using prior information as physical relationship between building components [35,36] or geometric primitives is beneficial in inferring building façade geometry. However, there is a need to visually check the final 3D model in order to be consistent with the captured scene.

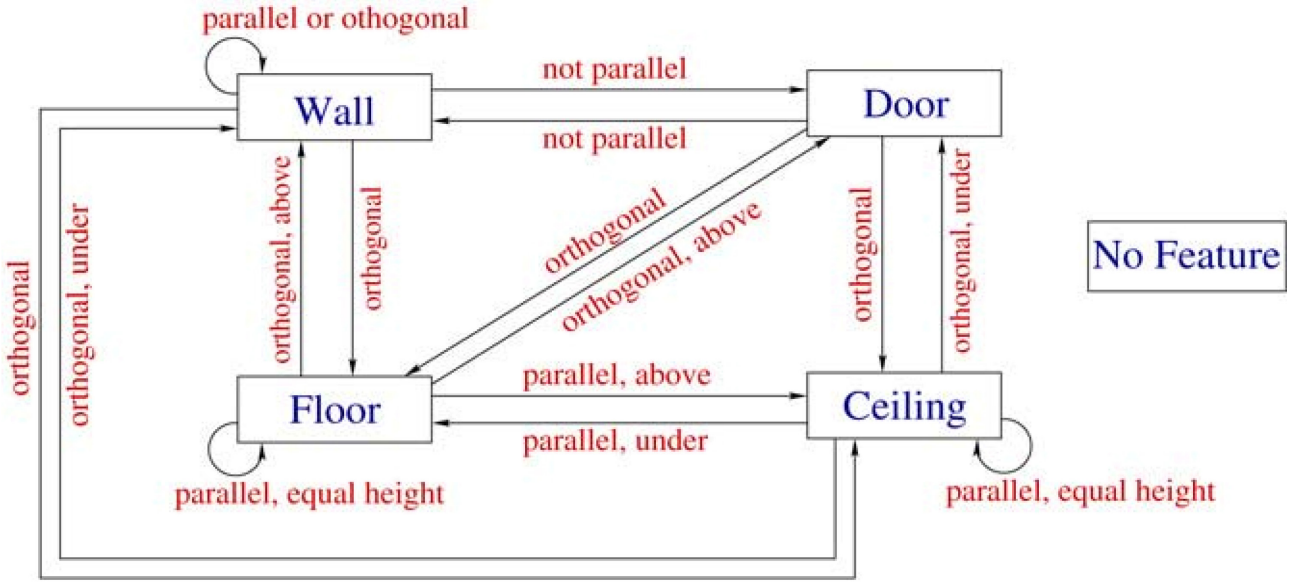

Fig. 3. prior information of building components[37]

4. Knowledge Gaps and Feature directions

4.1. Completeness

When reconstructing three-dimensional building from images, completeness could be used to validate the generated 3D building model. This metric is based on commission and omission errors, commission is related to the classification of elements as building components while they are not, on the other hand, omission is based on the building components that were neglected during reconstruction. In the case of generating 3D building model from terrestrial imagery, the evaluation of 3D model completeness is a complex task because of different challenges such as occlusions and lighting conditions. Accordingly, it is necessary to implement robust methods that assess 3D building model completeness.

4.2. Deep-Learning usefulness

Over the last years, deep-learning methods have been implemented to process images and point cloud, nevertheless, the use of these methods as a standalone pipeline that generate 3D building models still in progress. This set back can be explained by the focus on implementing techniques that models indoor scenes which is totally different from outdoor scenes which is known by more noise and outliers. Developing a deep-learning workflow require processing huge amount of data to train the model for the purpose of inferring building components, this issue can be avoided with working on quality of the dataset than quantity which could improve the 3D building models.

4.3. Benchmark Datasets

Having the ability to test 3D building modelling techniques on benchmark datasets opens the possibility to compare results with existing workflows [38], but literature shows a lack of two-dimensional image-based datasets. Moreover, the generation of a dense 3D model requires building façade openings to be also modelled. Which in this case, image-based datasets need to have internal and external camera parameters. Heavy borrowing of datasets was first introduced by machine leaning community [39] which is basically mean that a dataset was created to solve a specific problem but it is used to solve another one. From this idea, multiple benchmark datasets mainly created for building facade segmentation could be used for three-dimensional building reconstruction [40,42].
However, using this approach could yield problems for learning-based 3D models, training a model on a dataset collected in a specific location would not give good result on images for another place.

4.4. Computer resources and big data issues

The fact that 3D building reconstruction becomes a great interest for many applications created the needs for data collection at different scales, which in turn, expanded the time needed to produce good results. All the big data processing issues can be tackled with high performance computing technologies.it is used to avoid the time-consuming data processing steps due to CPU limitations [43].

5. Conclusions

In three-dimensional building modelling, the use of photogrammetry and computer vision techniques has gained attention from research and industrial communities, either with recovering target scene from multiple images or a single one. When it's required to generate a dense 3D model, façade openings need to be detected using segmentation. The major challenges include incomplete data and data inconsistency due to weather conditions and occlusions. Furthermore, the low-level of automation, created the need for a user intervention which is impacting 3D model quality.
Minimizing obscured parts of building components can be done through data fusion were aerial oblique imagery shows the hiding parts. Moreover, the creation of multiple benchmark datasets representing urban scenes for a specific geographic location (Morocco) could definitely help creating robust workflow that generate a 3D building model. While not losing sight of the fact, that establishing quality metrics is important to evaluate pipelines efficiency.

References

1. T. Alessandro Verri, "Introductory techniques for 3D computer vision", Prentice_Hall (1998)
2. R. Billen et al., "3D City Models and urban information: Current issues and perspectives", European COST Action TU0801, Liège, Belgium, p. **I**–118, (2014)
   G. Verhoeven, W. Karel, S. Štuhec, M. Doneus, I. Trinks, and N. Pfeifer, "MIND YOUR GREY TONES – EXAMINING THE INFLUENCE OF DECOLOURIZATION METHODS ON INTEREST POINT EXTRACTION AND MATCHING FOR ARCHITECTURAL IMAGE-BASED MODELLING," Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci., vol. **XL-5/W4**, pp. 307–314, (2015)
3. W. Hartmann, M. Havlena, and K. Schindler, "Recent developments in large-scale tie-point matching," ISPRS J. Photogramm. Remote Sens., vol. **115**, pp. 47–62, (2016)
4. J. L. Schonberger and J.-M. Frahm, "Structure-from-Motion Revisited," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 4104–4113, (2016)

5. F. Remondino, M. G. Spera, E. Nocerino, F. Menna, and F. Nex, "State of the art in high density image matching," Photogramm. Rec., vol. **29**, no. 146, pp. 144–166, (2014)
6. N. Kassotakis and V. Sarhosis, "Employing non-contact sensing techniques for improving efficiency and automation in numerical modelling of existing masonry structures: A critical literature review," Structures, vol. **32**, pp. 1777–1797, (2021)
7. R. De Marco and S. Parrinello, "Digital surveying and 3D modelling structural shape pipelines for instability monitoring in historical buildings: a strategy of versatile mesh models for ruined and endangered heritage," ACTA IMEKO, vol. **10**, no. 1, p. 84, (2021)
8. Y. Verdie, F. Lafarge, and P. Alliez, "LOD Generation for Urban Scenes," ACM Trans. Graph., vol. **34**, no. 3, pp. 1–14, (2015)
9. H. Arefi, J. Engels, M. Hahn, and H. Mayer, "LEVELS OF DETAIL IN 3D BUILDING RECONSTRUCTION FROM LIDAR DATA," (2008)
10. F. Biljecki, H. Ledoux, and J. Stoter, "An improved LOD specification for 3D building models," Comput. Environ. Urban Syst., vol. **59**, pp. 25–37, (2016)
11. E. Grilli, D. Dininno, G. Petrucci, and F. Remondino, "FROM 2D TO 3D SUPERVISED SEGMENTATION AND CLASSIFICATION FOR CULTURAL HERITAGE APPLICATIONS," Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci., vol. **XLII–2**, pp.399–406, (2018)
12. C. Stucker, A. Richard, J. D. Wegner, and K. Schindler, "SUPERVISED OUTLIER DETECTION IN LARGE-SCALE MVS POINT CLOUDS FOR 3D CITY MODELING APPLICATIONS," ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci., vol. **IV–2**, pp. 263–270, (2018)
13. J. Xue, X. Hou, and Y. Zeng, "Review of Image-Based 3D Reconstruction of Building for Automated Construction Progress Monitoring," Appl. Sci., vol. **11**, no. 17, p. 7840, Aug. (2021)
14. H. Fathi, F. Dai, and M. Lourakis, "Automated as-built 3D reconstruction of civil infrastructure using computer vision: Achievements, opportunities, and challenges," Adv.Eng. Inform., vol. **29**, no. 2, pp. 149–161, (2015)
15. Hongdong Li and R. Hartley, "Five-Point Motion Estimation Made Easy," in 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, China, pp. 630–633, (2006)
16. F. Remondino and S. El-Hakim, "Image-based 3D Modelling: A Review: Image-based 3D modelling: a review," Photogramm. Rec., vol. **21**, no. 115, pp. 269–291, (2006)
17. B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle Adjustment — A Modern Synthesis," in Vision Algorithms: Theory and Practice, vol. **1883**, B. Triggs, A.Zisserman, and R. Szeliski, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, (2000)
18. S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski, "Building Rome in a day," in 2009 IEEE 12th International Conference on Computer Vision, Kyoto, pp. 72–79, (2009)
19. D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," Int. J. Comput.Vis., vol. **60**, no. 2, pp. 91–110, (2004)
20. R. Hartley and A. Zisserman, Multiple view geometry in computer vision, 2nd ed. Cambridge, UK: Cambridge University Press, (2004)
21. M. A. Fischler and R. C. Bolles, "Random sample consensus," vol. **24**, no. 6, (1981)
22. R. Wang, "3D building modeling using images and LiDAR: a review," Int. J. Image Data Fusion, vol. **4**, no. 4, pp. 273–292, (2013)
23. Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski, "Towards Internet-scale multi-view stereo," in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, pp. 1434–1441, Jun. 2010
24. J. Xiao, T. Fang, P. Tan, P. Zhao, E. Ofek, and L. Quan, "Image-based Fac¸ade Modeling", Association for Computing Machinery, pp. 1–10, (2008)
25. B. Micusık and J. Kosecka, "Piecewise Planar City 3D Modeling from Street View Panoramic Sequences", *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, pp. 2906-2912, (2009)
26. E. K. Stathopoulou, R. Battisti, D. Cernea, A. Georgopoulos, and F. Remondino, "Multiple View Stereo with quadtree-guided priors," ISPRS J. Photogramm. Remote Sens., vol. **196**, pp.197–209, (2023)
27. A. Saxena, Min Sun, and A. Y. Ng, "Make3D: Learning 3D Scene Structure from a Single Still Image," IEEE Trans. Pattern Anal. Mach. Intell., vol. **31**, no. 5, pp. 824–840, (2009)
28. D. Hoiem, A. A. Efros, and M. Hebert,"Geometric context from a single image," in Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume **1**, Beijing, China, pp. 654-661,(2005)
29. B. G. Pantoja-Rosero, R. Achanta, M. Kozinski, P. Fua, F. Perez-Cruz, and K. Beyer, "Generating LOD3 building models from structure-from motion and semantic segmentation," Autom. Constr., vol. **141**, p. 104430, (2022)

30. P. Musialski, P. Wonka, D. G. Aliaga, M. Wimmer, L. van Gool, and W. Purgathofer, "A Survey of Urban Reconstruction" Comput. Graph.Forum, vol. **32**, no. 6, pp. 146–177, (2013)
31. H. Omar, L. Mahdjoubi, and G. Kheder, "Towards an automated photogrammetry-based approach for monitoring and controlling construction site activities," Comput. Ind., vol. **98**, pp. 172–182, (2018)
32. R. A. Galantucci and F. Fatiguso, "Advanced damage detection techniques in historical buildings using digital photogrammetry and 3D surface anlysis," J. Cult. Herit., vol. **36**, pp.51–62, (2019)
33. B. Bortoluzzi, I. Efremov, C. Medina, D. Sobieraj, and J. J. McArthur, "Automating the creation of building information models for existing buildings," Autom. Constr., vol. **105**, p. 102838, (2019)
34. P. Tang, D. Huber, B. Akinci, R. Lipman, and A. Lytle, "Automatic reconstruction of as-built building information models from laser-scanned point clouds: A review of related techniques," Autom. Constr., vol. **19**, no. 7, pp. 829–843, (2010)
35. A. Braun, S.Tuttas, A.Borrmann,U.Stilla. "A Concept for Automated Construction Progress Monitoring Using BIM-Based Geometric Constraints and Photogrammetric Point Clouds." *Journal of Information Technology in Construction* 20, pp.68–79 (January 1, 2015).
36. A. Nüchter and J. Hertzberg, "Towards semantic maps for mobile robots," Robot. Auton. Syst., vol. **56**, no. 11, pp. 915–926, (2008)
37. Y. Xie, J. Tian, and X. X. Zhu, "Linking Points With Labels in 3D: A Review of Point Cloud Semantic Segmentation," IEEE Geosci. Remote Sens. Mag., vol. **8**, no. 4, pp. 38–59, (2020)
38. B. Koch, E. Denton, A. Hanna, and J. G. Foster, "Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research." arXiv, (2021)
39. R. Tyleček and R. Šára, "Spatial Pattern Templates for Recognition of Objects with Regular Structure," in Pattern Recognition, vol. **8142**, J. Weickert, M. Hein, and B. Schiele, Eds.Berlin, Heidelberg: Springer Berlin Heidelberg, (2013)
40. H. Riemenschneider, A. Bódis-Szomorú, J. Weissenberg, and L. Van Gool, "Learning Where to Classify in Multi-view Semantic Segmentation," in Computer Vision – ECCV 2014, vol. **8693**, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, pp. 516–532, (2014)
41. F.Korˇc, W.Forstner. "eTRIMS Image Database for Interpreting Images of Man-Made Scenes," Dept. of Photogrammetry, University of Bonn, (2009)
42. A. Nowogrodzki, "Eleven tips for working with large data sets," Nature, vol. **577**, no. 7790, pp. 439–440, (2020)