A Review of Advancements in Multi-view Stereo

Jackie Lin
jackiel4@illinois.edu

Christopher Conway
cconway4@illinois.edu

Abstract

3D scene reconstruction is a prevalent problem in applications such as robotics, Virtual Reality (VR), autonomous driving, and more. Multi-view stereo (MVS) techniques generate a comprehensive 3D representation of a scene from a series of images. This review presents a taxonomy of various MVS methods, starting with sensor types and then depth information merging techniques. Key evaluation metrics for accuracy or completeness and the related datasets with ground truth information used to benchmark MVS algorithms are presented. Trade-offs between reconstruction accuracy, efficiency, and generalizability are discussed in the state-of-the-art. Finally, ongoing challenges and possible future research directions for large-scale or challenging scenes are explored.

1. Introduction

3D scene reconstruction from image series has been an enduring challenge in computer vision. Multi-view stereo (MVS) techniques aim to create 3D models of scenes from series of images. MVS has been used for view synthesis and image-based rendering.

MVS algorithms can be broadly categorized as conventional or learning-based. Conventional MVS can be split into two stages: first is computing per-pixel depth from either stereo matching passive camera frames or obtaining the depth maps directly from active sensors, and second is to merge the depth information over multiple frames using consistency measures to form a dense 3D representation. Learning-based approaches are a more recent development that focus on using data to learn global semantic information that helps guide the depth optimization in noisy and difficult MVS scenarios, such as specular reflections and weak texture patterns.

This survey paper is organized as follows: Section 2 presents the taxonomy of key design decisions and techniques of MVS. Section 3 reviews key error metrics for accuracy or completeness. It also contains a summary of popular MVS datasets used to benchmark methods. Section 4 assesses current capabilities and gaps in MVS. Lastly, Section 5 identifies future areas of MVS research including recent advancements of the task with learning-based approaches.

2. Taxonomy

Multi-View Stereo is a fundamental technique in computer vision for reconstructing 3D models from multiple 2D images taken from different viewpoints. By leveraging overlapping depth information from multiple images, MVS can produce highly detailed 3D reconstructions. This section provides a taxonomy of MVS methods, focusing on the depth information acquisition and depth merging techniques. Refer to [3, 19] for the development of MVS over the past two decades.

2.1. Sensors to Obtain Depth Information

Two types of sensor methods exist for capturing and precomputing depth images: passive and active methods.

Passive sensors use two or more RGB cameras that capture the visible light of the scene. In passive sensing, when the relative camera poses on the sensor device are known, traditional stereo matching methods can be applied to estimate the depth map. Stereo matching methods [15] exploit both geometric constraints (epipolar geometry) and photometric consistency (color similarity across images) to compute pixel-wise depth. The advantage of passive sensors is that they are relatively cheap and easy to implement with RGB cameras. The disadvantage of passive methods is that textureless, featureless regions and repeating patterns on surfaces result in multiple disparity solutions with the same cost, which leads to depth ambiguity.

Depth information can also be obtained through *active* sensors, such as structured light cameras [13], time-of-flight cameras [25], Microsoft Kinect [8] which directly measure depth by emitting and analyzing light reflections. The disadvantages is that active sensors perform poorly in outdoor and sun illuminated environments because the emitted light from the sensor is overpowered by outdoor light.

The objective of MVS, however, is to capture depth information from multiple sensor poses around the scene which requires moving the sensor device. Often the world space position of the sensor and the corresponding captured

depth information may not be known. Thus, MVS is commonly employed with Structure from Motion (SfM) [17,21] which estimates both the extrinsic sensor poses and a sparse 3D point cloud. SfM computes the camera positions and a preliminary 3D reconstruction based on matched feature points across images. Additionally, bundle adjustment [20] is performed to refine 3D points and camera parameters found in SfM by minimizing the reprojection error. This is necessary because MVS is highly sensitive to reprojection error due to the use of epipolar geometry. Once the initial camera poses and sparse points are recovered, depth merging techniques are applied to generate a dense reconstruction.

The challenge in MVS lies in how to efficiently and accurately merge information from multiple depth frames into a unified, dense 3D representation. Three primary approaches for merging depth information in MVS are (1) voxel-based methods, (2) feature point-growing MVS, and (3) depth map fusion techniques.

2.2. Merging Depth Information

2.2.1 Voxel-Based Methods

In voxel-based methods, the 3D space is discretized into a grid of small volumetric elements, or voxels. These methods assume the scene volume to be initially filled with voxels and use the input images to determine which voxels are consistent with the scene surface.

One of the foundational voxel-based algorithms is **space carving** [9]. In this approach, voxels that do not conform to color consistency constraints across multiple views are incrementally "carved" away, leaving only those voxels that belong to the scene surface. This process relies on a visibility check, where each voxel must be visible and consistent across multiple views to remain part of the reconstruction. While voxel-based methods are simple and effective for coarse reconstructions, they can be computationally expensive due to the 3D discretization and large memory requirements, particularly when dealing with high-resolution scenes.

Improvements in voxel-based MVS include **volumetric graph-cut** [22] methods, which model the scene as a graph of voxels and seek an optimal surface boundary using energy minimization techniques. These methods can better handle noisy input data and occlusions but still face limitations in resolution and computational efficiency.

2.2.2 Feature-Based Methods

Feature point-growing methods build upon sparse feature points (e.g., from SfM) and iteratively expand the set of points by matching and triangulating additional features across images. A prominent example is **Patch-Based Multi-View Stereo** (PMVS) [4], which begins with a sparse set of 3D points and grows patches around these points to create a dense 3D point cloud. This method focuses on ensuring photometric consistency, as each new point is added only if it can be matched in multiple images with minimal color variance.

The feature-based approach is particularly useful for reconstructing scenes with textured surfaces, where distinctive features can be reliably matched across views. However, it may struggle with textureless or homogeneous regions, where feature matching is ambiguous or unreliable. Additionally, feature-based methods are generally less dense than voxel-based approaches, but they offer higher accuracy and efficiency when feature detection is robust.

2.2.3 Depth Map Fusion Methods

Depth map fusion methods work by first estimating a depth map for each image using stereo matching techniques, and then fusing these depth maps into a single 3D representation. This category encompasses several prominent techniques, including plane-sweeping stereo, PatchMatch stereo, and deep learning-based approaches.

Plane-Sweeping Stereo [5] This technique involves "sweeping" a plane through the scene at different depths and projecting the images onto this plane. By evaluating the color consistency across views for each plane position, the algorithm can estimate the depth for each pixel. Plane-sweeping stereo is effective for large-scale scenes and is particularly well-suited for handling textureless regions. However, it can be computationally expensive due to the need to evaluate many planes across different depths.

PatchMatch Stereo [1]: This approach introduces a fast, randomized search technique for depth estimation. It begins by randomly assigning depth hypotheses to each pixel and then iteratively refines these estimates by propagating good depth values to neighboring pixels. PatchMatch stereo is highly efficient and capable of producing dense and accurate depth maps. It is particularly advantageous in complex scenes with irregular surfaces, as it can quickly converge to the correct depth solution.

Truncated Signed Distance Function Fusion [11]: This approach represents a depth image as a truncated signed distance function which is a voxel grid where the voxel containing the surface is represented with zero, and voxels in front and behind are 1 and -1 respectively. The TSDF of the scene is recursively updated through volume integration of new depth observations. This method allows for the smooth fusion of multiple depth maps and can be implemented with parallelism since each voxel can be integrated independently.

Deep Learning-Based MVS: Recent advances in deep learning have introduced data-driven approaches to MVS, with MVSNet [26] being one of the earliest and prominent

examples. MVSNet uses a convolutional neural network (CNN) to predict per-pixel depth maps for each input image and constructs a cost volume to evaluate depth hypotheses. By learning geometric and photometric consistency from large datasets, MVSNet can produce highly accurate reconstructions, even in challenging scenarios with occlusions and textureless surfaces.

2.3. Summary

The taxonomy of MVS techniques reveals a wide range of methods for both depth acquisition and fusion. Voxel-based methods are suited for coarse reconstructions, while feature-based approaches excel in capturing fine details in textured regions. Depth map fusion techniques, including deep learning-based methods, strike a balance between accuracy and efficiency, offering a promising direction for future research. The review of MVS fusion techniques shows that a variety of 3D representations is produced, such as voxels, point clouds, and surfaces.

3. Evaluation

Evaluation of MVS techniques considers multiple performance metrics such as accuracy and completeness, as well as efficiency of reconstruction. These metrics can be calculated using various methods, such as comparison to ground truth data or evaluation of synthetic images obtained by warping reference images [15]. Many datasets have been published to benchmark MVS models, such as the Middlebury datasets [16] and Tanks and Temples [16]. This section will discuss quantitative evaluation metrics in greater detail, and highlight the some of the more popular benchmark datasets.

3.1. Metrics

Evaluation metrics are necessary to quantitatively compare the performance of MVS algorithms. These are often error measurements that compare estimated results to ground truth, examining overall depth maps or at a finer pixel level. Evaluators may also choose to focus on a subsection of the image, such as occluded regions, regions near disparity jumps, or regions with lower texture. Alternatively, quality can be judged using prediction error, where estimated results for novel views can be compared to data for these additional views.

Root-mean-squared error (RMS) is a common disparity error calculation to quantify accuracy. The formula considers the root of the sum of the squared errors, such as the error between the computed and the ground truth depth maps over the total number of pixels. Similarly, Mean absolute error (MAE) can be computed to consider accuracy. MAE considers the sum of the absolute value of the errors over the total, where the errors are again between the computed and the ground truth depth maps.

Completeness may be more difficult to quantify, as the regions to consider must be selected based on the availability of ground truth data, or with problematic areas such as low texture and occluded regions considered. One measure for completeness is the proportion of pixels that attain a depth within a tolerance region of the ground truth value. Consideration of the regions with higher disparity may yield insights as to benefits and weaknesses of particular algorithms.

MVS algorithms have been evaluated using these and similar metrics for various applications such as satellite imagery [6], using multi-camera videos [18], for cultural heritage preservation [12], and others. Demonstrations highlight that many state-of-the-art methods can create high quality reconstructions using various data sources ranging from hand-held devices to high-resolution satellite captures.

3.2. Datasets

Adequate datasets are critical to compare the performance of various MVS algorithms. In particular, ground truth data often in the form of point clouds is necessary to quantify accuracy. Rich datasets with many views of each scene further enable evaluation.

Datasets of varying sources and kinds have been published over the decades of research on MVS. Middlebury datasets have continuously enabled further study of MVS algorithms. Many earlier datasets were lower resolution captures of planar scenes or simple objects such as cones, with ground truth data from structured light [15]. More recent datasets have placed emphasis on high-resolution and accurate ground truth disparities, sometimes with sub-pixel accuracy [14]. Some of the most recent datasets explore challenging scenarios such as images taken from mobile devices on richer scenes.

Deep learning based MVS approaches have motivated the generation of very large datasets for training purposes. Due to difficulties related to capturing many images of real scenes, researchers have considered the development of synthetic datasets. Synthetic datasets such as MVS Synth [7] have been shown to be photorealistic and successful in improving reconstruction predictions. Synthetic datasets can be generated from the recreation of varying lighting conditions and the use of colored image rendering techniques in conjunction with real image inputs. Such synthetic datasets have been created with over 15000 images of various scenes and objects [27].

4. Capabilities and Gaps in the State of the Art

Many state-of-the-art MVS methods have demonstrated high quality 3D scene reconstruction on a range of tasks, but work is ongoing to improve performance for scenarios with inadequate lighting or texture, scalability to large

scenes, and generalizability for diverse environments and object shapes.

Though conventional methods lead the MVS field for a long time learning based methods have seen wider adoption in recent years ([24]), due to their ability to learn robust feature representations and information from multiple views. However, one issue with learning based approaches is generalizability outside of the training set. To address this, larger and larger training sets such as BlendedMVS [27] have been made available to provide more scenes and objects to the algorithm. The possible tradeoff to richer scene variety and density however is the algorithm efficiency.

The potential for real-time MVS, such as from video data, has driven exploration of high efficiency MVS algorithms. Some MVS methods consider unlimited computation time, and may have demonstrated results on smaller or lower resolution datasets. Others have demonstrated intermediate or high resolution reconstructions in real-time using both conventional [23] and learning approaches [2]. These approaches can be enabled through updating previous estimations with each new frame, and light-weighting calculations such as viewpoint approximation between time steps.

To address challenges related to occlusion and poor lighting conditions, researchers have considered estimation of additional parameters such as visibility, identification of local regions, and the use multi-sensor inputs. Visibility estimations consider uncertainty ranges for matching pixel positions, and quantify probabilities to identify pixels that are visible or occluded between various support images. Challenging areas such as areas with identified occlusions or specularities can be isolated, and specific local windows, often called patches, with the desired scene information can be matched rather than just globally matching full images [28]. This consideration of global and local features can also be considered through the use of sensors with different resolutions. Some of the most recent approaches such as Hybrid-MVS [10] have utilized both patch matching and the combination of high resolution digital camera images with low resolution RGB-D frames to aid in robustness.

5. The Future of MVS

Progress in MVS is ongoing, aiming to achieve the highest accuracy scene reconstruction from the smallest amount of or most challenging views in the fastest time span. A variety of methodologies and their benefits have been briefly introduced in this review, ranging from conventional approaches to more recent advances that leverage deep learning. The future for MVS likely combines the most compelling advancements in various methods to create hybrid MVS pipelines [10].

One of the most promising advancements in MVS is the use of coarse to fine methods. The trade off between superior performance with high resolution images against the increased computational cost is difficult to balance. Hybrid methods that consider at first low resolution inputs to generate initial estimations, then scale to use high resolution inputs optimizes the computational trade off. Similarly, using patches of high resolution images in local matching in tandem with global matching achieves a balance between performance and efficiency. Future work will likely involve novel algorithm structures that cascade these varying levels of information, or utilize parallel computing for tasks such as calculating many patches at once to further improve efficiency.

Another promising area for advancement is in more intelligent sampling strategies. In a given multiview dataset, a variety of occlusions or issues with specularity may occur, but particular subsets of images may not have this issue. Probabalistic methods that can identify good subsets of images or subsets of image patches can be used to tune reconstruction to achieve higher completeness with minimal additions in computational complexity. The balance between increasing completeness through such calculations and maintaining a lightweight method that can compute in real time must be further explored.

Further work in MVS to combine the benefits of these approaches is made possible through advancements in hardware, such as larger memory capacity and accelerated calculation time. In particular, progress with GPU memory and acceleration presents the opportunity to achieve real time performance even for large and complex scenes.

References

[1] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009. 2

[2] Arda Duzceker, Silvano Galliani, Christoph Vogel, Pablo Speciale, Mihai Dusmanu, and Marc Pollefeys. Deepvideomvs: Multi-view stereo on video with recurrent spatio-temporal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15324–15333, June 2021. 4

[3] Yasutaka Furukawa, Carlos Hernández, et al. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 1

[4] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009. 2

[5] David Gallup, Jan-Michael Frahm, Philippos Mordohai, Qingxiong Yang, and Marc Pollefeys. Real-time plane-sweeping stereo with multiple sweeping directions. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 2

[6] Alvaro Gómez, Gregory Randall, Gabriele Facciolo, and Rafael Grompone von Gioi. An experimental comparison of multi-view stereo approaches on satellite images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 844–853, January 2022. 3

[7] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[8] Kourosh Khoshelham and Sander Oude Elberink. Accuracy and resolution of kinect depth data for indoor mapping applications. *sensors*, 12(2):1437–1454, 2012. 1

[9] Kiriakos N Kutulakos and Steven M Seitz. A theory of shape by space carving. *International journal of computer vision*, 38:199–218, 2000. 2

[10] Chenchen Li, Liyang Zhou, Hanqing Jiang, Zhuang Zhang, Xiaojun Xiang, Han Sun, Qing Luan, Hujun Bao, and Guofeng Zhang. Hybrid-mvs: Robust multi-view reconstruction with hybrid optimization of visual and depth cues. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12):7630–7644, 2023. 4

[11] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE international symposium on mixed and augmented reality*, pages 127–136. Ieee, 2011. 2

[12] Massimiliano Pepe, Vincenzo Saverio Alfio, and Domenica Costantino. Uav platforms and the sfm-mvs approach in the 3d surveys and modelling: A review in the cultural heritage field. *Applied Sciences*, 12(24), 2022. 3

[13] CMPPC Rocchini, Paulo Cignoni, Claudio Montani, Paolo Pingi, and Roberto Scopigno. A low cost 3d scanner based on structured light. In *computer graphics forum*, volume 20, pages 299–308. Wiley Online Library, 2001. 1

[14] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. volume 8753, pages 31–42, 09 2014. 3

[15] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47:7–42, 2002. 1, 3

[16] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 1, pages I–I, 2003. 3

[17] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 2

[18] Thomas Schops, Johannes L. Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3

[19] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 1, pages 519–528. IEEE, 2006. 1

[20] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms Corfu, Greece, September 21–22, 1999 Proceedings*, pages 298–372. Springer, 2000. 2

[21] Shimon Ullman. The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 203(1153):405–426, 1979. 2

[22] George Vogiatzis, Carlos Hernández Esteban, Philip HS Torr, and Roberto Cipolla. Multiview stereo via volumetric graph-cuts and occlusion robust photo-consistency. *IEEE transactions on pattern analysis and machine intelligence*, 29(12):2241–2246, 2007. 2

[23] George Vogiatzis and Carlos Hernández. Video-based, real-time multi-view stereo. *Image and Vision Computing*, 29(7):434–441, 2011. 4

[24] Fangjinhua Wang, Qingtian Zhu, Di Chang, Quankai Gao, Junlin Han, Tong Zhang, Richard Hartley, and Marc Pollefeys. Learning-based multi-view stereo: A survey, 2024. 4

[25] Aloysius Wehr and Uwe Lohr. Airborne laser scanning—an introduction and overview. *ISPRS Journal of photogrammetry and remote sensing*, 54(2-3):68–82, 1999. 1

[26] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. 2

[27] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3, 4

[28] Zhaokun Zhu, Christos Stamatopoulos, and Clive S. Fraser. Accurate and occlusion-robust multi-view stereo. *ISPRS Journal of Photogrammetry and Remote Sensing*, 109:47–61, 2015. 4