Multi-View Stereo Using Discontinuity-Aware Planar Parameters

HanShin Lim
Realistic Media Research Group
Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea
hslim@etri.re.kr

Hyon-Gon Choo
Realistic Media Research Group
Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea
hyongonchoo@etri.re.kr

***Abstract*—One of the successful approaches in the traditional field of Multi-View Stereo is PatchMatch-based methods. However, because the approaches rely heavily on photometric consistencies, the results are unreliable in low-textured areas. This paper proposes a Multi-View Stereo method that enforces the reliability of similarity costs by utilizing planar parameters refined through spatial consistencies in the PatchMatch process. In the proposed method, initially, depth maps and normals are generated by applying a conventional PatchMatch method. Afterward, a triangular mesh is generated from the selected nodes, and planar parameters are computed from the resulting triangular mesh model. The discontinuity-aware PatchMatch is then processed by leveraging the planar parameters refined through adjacent triangular meshes for similarity cost computation. The results of the experiments demonstrate that the proposed method achieves superior performance compared to existing Multi-View Stereo methods for publicly available High-Resolution images.**

***Index Terms*—Multi-View Stereo, Depth Map Generation, 3D Reconstruction**

I. Introduction

In the field of Multi-View Stereo, two critical elements are intensively addressed: the precise reconstruction of detailed structures and the representation of surfaces that lack photometric consistencies, such as areas with low-textured or occluded regions. In terms of accuracy, the main challenge lies in the effective and precise identification of matching points across neighboring images. Within a range of traditional methods, the PatchMatch [1] algorithm has been effectively utilized, leveraging its efficient reduction of solution space and its capability for parallel processing. However, in spite of their effectiveness, PatchMatch-based methods encounter drawbacks when dealing with cases containing numerous or large low-textured or occluded areas, as these regions often lead to unstable values. To improve the completeness within these regions, several approaches have been introduced, all based on the idea of effectively utilizing spatial information within each image's domain.

This paper proposes a Multi-View Stereo method that is enforcing the reliability of similarity costs by utilizing planar parameters refined through spatial consistencies in the PatchMatch process. Figure 1 shows a color image from the Office image set in the ETH3D training set, and the generated depth map, normal map, and the point cloud by the proposed method. Main contributions of the paper are as follows:

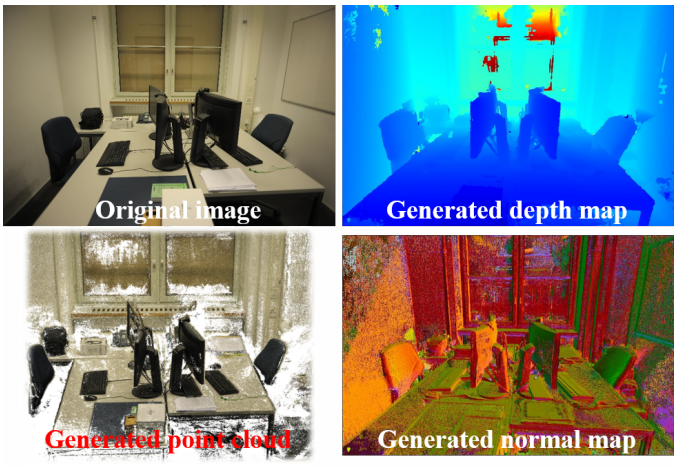

Fig. 1. A color image from the Office image set in the ETH3D training set, and the generated depth map, normal map, and the point cloud by the proposed method.

1) We propose a similarity cost model that incorporates photometric consistencies along with planar parameters refined through spatial consistencies, aiming to achieve more reliable estimates of depth values.
2) We demonstrate that the proposed approach achieves superior results compared to existing methods on a high-resolution public dataset.

In section 2, we briefly review the Multi-View Stereo approaches. Fundamental concept and overall procedure of the proposed Multi-View Stereo using discontinuity-aware planar prior is explained in section 3. Section 4 presents the experimental results and we conclude the paper in section 5.

This work was supported by ETRI grant funded by the Korean government [23ZH1200, The research of the fundamental media·contents technologies for hyper-realistic media space].

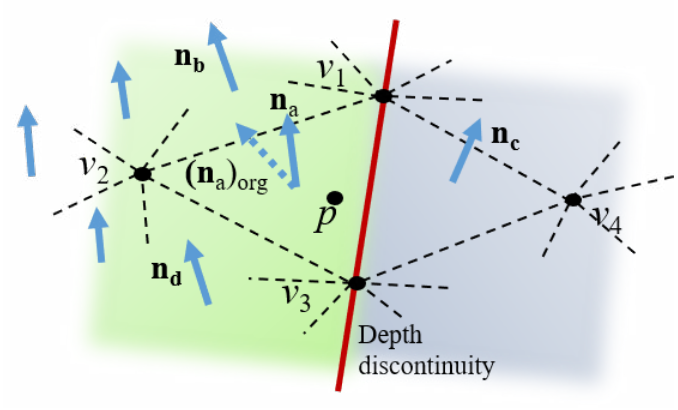

Fig. 2. Fundamental concept of the proposed method. Since the original planar parameters have limitations in representing depth discontinuities, the original planar parameter $(\mathbf{n_a})_{\mathrm{org}}$ is refined to be $\mathbf{n_a}$ by utilizing the planar parameters $\mathbf{n_b}$, $\mathbf{n_c}$, and $\mathbf{n_d}$ and color values of the adjacent triangular meshes. The refined planar parameter is then incorporated into the similarity cost model to estimate the depth of the pixel p.

II. Previous Works

A. Traditional Approaches

One of the main goals of the Multi-View Stereo is to increase the accuracy of reconstructing structures. Many approaches have emerged to tackle this issue, focused on accurately estimating pixel depths, primarily by evaluating the photometric consistencies of patches from both the reference image and nearby images. PatchMatch [1] emerges as one of the most prominent techniques for generating depth maps in the traditional Multi-View Stereo domain. The primary advantages of PatchMatch-based approaches lie in their ability to efficiently reduce the solution space, while also enabling parallel processing. [2] adopted a variational inference approach in the depth propagation step. [3] used additional geometric priors for accurate view selection. [4] proposed an efficient propagation step for parallel processing.

The main drawback of the PatchMatch-based approach is that the approach cannot handle low-textured regions effectively due to the reliance on photometric consistencies during the best cost value selection process in the propagation step. Moreover, because the matching consistency is often ambiguous and unstable in several regions, the outliers cannot be effectively reduced. In order to reduce the ambiguity of the color consistencies, [5] and [6] adopted a pyramidal structure and [7] applied mesh-based planar priors.

B. Learning-based Approaches

MVSNet [9] is one of the most successful learning-based Multi-View Stereo approaches. The method make use of a deep learning network to optimize cost volumes to generate depth maps. However, since the size of the cost volume increases rapidly as the resolution of the image increases, the approach is not suitable for high-resolution applications. To address to problem, various approaches [10] [11] [12] have been proposed. The adoption of the self-attention mechanism [13] within the learning-based Multi-View Stereo field has yielded promising results in recent studies [14]. In addition, various approaches to incorporating implicit neural representation techniques [15] into volumetric 3D reconstruction are gaining much attention.

As mentioned earlier, learning-based approaches become inefficient at higher resolutions due to the rapid increase in memory size. In contrast, the proposed method, which is PatchMatch-based, shows reliable results for high-resolution images, as demonstrated in the Experimental Results section.

III. Proposed Method

A. Fundamental Concept of the Proposed Method

The basic concept of the proposed method is to incorporate refined planar parameters into the similarity cost model to improve the reliability of the estimated depth values. Figure 2 shows the Fundamental concept of the proposed method. When applying a conventional PatchMatch approach with a photometric cost model, the estimated depth value of a pixel p is unreliable since the region is a homogeneous area. In order to address the problem, planar prior term was added to the similarity cost model [7]. However, planar parameters computed from the depths and positions of the initial nodes have a limitation in representing depth discontinuities because the structure of the generated triangular model is continuous. To deal with the depth-discontinuity representation problem, the proposed method refine the planar parameters by using the information of the adjacent triangular meshes. In Figure 2, the original planar parameter $(\mathbf{n_a})_{\mathrm{org}}$ is refined to be $\mathbf{n_a}$ by utilizing the planar parameters $\mathbf{n_b}$, $\mathbf{n_c}$, and $\mathbf{n_d}$ and color values of the adjacent triangular meshes.

B. Overall Procedure of the Proposed Method

The proposed method is largely composed of three steps: initial depth map and normal generation, triangular model generation, and discontinuity-aware PatchMatch process. Figure 3 shows the overall procedure of the proposed method. The following is more detailed explanation of the proposed method.

1) Initial depth map and normal map generation: In the proposed method, first, initial depth maps and normals are generated by applying an existing PatchMatch method [5] to the original color images with their camera parameters. In the PatchMatch process, conventional photometric costs are used for similarity measure.

2) Triangular model generation: The second step involves selecting pixel positions with locally high photometric costs as initial nodes from the image domain. A triangular mesh is then generated by applying the 2D Delaunay triangulation from the selected nodes, and initial planar parameters are computed from the generated triangular mesh model.

3) Discontinuity-aware PatchMatch: In the third step, Discontinuity-aware PatchMatch is executed by leveraging the planar parameters of both corresponding and adjacent triangular meshes for similarity cost computation.

Fig. 3. Overall procedure of the proposed method, which is largely composed of three steps of initial depth map and normal generation, triangular model generation, and discontinuity-aware PatchMatch process. In the proposed method, first, initial depth maps and normals are generated by applying a conventional PatchMatch method. After that, triangular mesh is generated from the selected nodes, and planar parameters are computed from the generated triangular mesh model. Discontinuity-aware PatchMatch is then processed by leveraging the planar parameters refined through adjacent triangular meshes for similarity cost computation.

In the proposed method, the original planar parameter $(\mathbf{n_a})_{\text{org}}$ is refined to be $\mathbf{n_a}$ by weighted averaging the original and the adjacent planar parameters as follows:

$$\mathbf{n_a} = w_a(\mathbf{n_a})_{\text{org}} + w_b\mathbf{n_b} + w_c\mathbf{n_c} + w_d\mathbf{n_d} \quad (1)$$

where the weights w_a, w_b, w_c, and w_d are computed by averaging the color values in each planar mesh with the ratio of distance of the edges the two planar meshes are faced.

The depth values of each plane is recomputed from the initial nodes and the refined parameter by applying a simple method such as least squares optimization.

The similarity cost $C(p, \{d, \mathbf{n}\})$ of the pixel p for hypothesis depth d and normal $\mathbf{n}$ is then computed as follows:

$$C(p, \{d, \mathbf{n}\}) = Photo(p, \{d, \mathbf{n}\}) - Plane(p, \{d, \mathbf{n}\}) \quad (2)$$

where $Photo(p, \{d, \mathbf{n}\})$ is the photometric consistency term and $Plane(p, \{d, \mathbf{n}\})$ is the planar prior term leveraged by the refined planar parameters as follows:

$$Plane(p, \{d, \mathbf{n}\}) = \log(\gamma + \exp -(\frac{(d - d_{\text{r}})^2}{\sigma_d} + \frac{\arccos^2 \mathbf{n}^{\text{T}}\mathbf{n}_{\text{r}}}{\sigma_{\mathbf{n}}})) \quad (3)$$

where σ_d and $\sigma_{\mathbf{n}}$ are respectively bandwidth of depth difference and bandwidth of normal difference [7], and d_{r} and $\mathbf{n}_{\text{r}}$ are respectively depth and normal values obtained by the refined planar parameters.

IV. Experimental Results

The experiment was conducted under the AMD CPU with RTX-series GPU environment, and the proposed method was implemented with C++. The comparative evaluation of the performance of the proposed method was conducted on the Office and Pipes, Meadow, and Terrace sets in the ETH3D training set [16], where the resolutions are 6048x4032. We calculated F-scores, the harmonic mean of precision and completeness, of the reconstructed point clouds to quantitatively evaluate the performance of the existing approaches and the proposed method.

Figure 4 shows the original color images, triangulation results, generated depth maps and normal maps, and generated point clouds by the proposed method for Office, Pipes, Meadow, and Terrace sets in the ETH3D training set. Table I shows the quantitative evaluation (F-score) results

TABLE I
Quantative evaluation (F-score) of the reconstructed dense 3D point clouds for Office, Pipes, Meadow, and Terrace image sets in the ETH3D training set.

	Office	Pipes	Mea.	Terr.	Mean
COLMAP [3]	47.32	50.72	49.96	84.94	58.24
EPPMVSNet [12]	68.68	72.09	49.77	89.12	69.92
MVSTER [14]	68.86	63.76	61.39	87.56	70.39
ACMM [5]	63.01	69.26	71.49	89.76	73.38
MARMVS [8]	64.7	77.04	68.25	88.22	74.55
ACMP [7]	74.45	69.16	71.00	89.92	76.13
Proposed	76.36	71.92	71.21	90.77	77.57

of the reconstructed dense 3D point clouds for the Office, Pipes, Meadow, and Terrace sets in the ETH3D training set. As shown in the quantitative evaluation results, the proposed method outperforms both traditional (COLMAP, ACMM, MARMVS, ACMP) and learning-based (EPPMVS-Net, MVSTER) approaches with respect to the F-scores for high-resolution images.

Figure 5 shows qualitative comparison of the depth maps generated by COLMAP, ACMP, and the proposed method for Office and Pipes image sets from ETH3D training set. As shown in the generated depth maps, the proposed method produces qualitatively more reliable and less noisy depth values especially in the low-textured regions such as the walls and the panels on the desks.

Although the proposed method is based on the traditional approaches, if some learning-based techniques can be elaborately combined with the proposed method, we expect much improvement can be possible.

V. Conclusion

This paper proposed a Multi-View Stereo method that is enforcing the reliability of similarity costs by using the planar parameters refined through spatial consistencies in the PatchMatch process. The experimental results show that the proposed method successfully improves the quality of depth maps and point clouds compared to the existing approaches.

Future work will be aimed at applying the proposed method to various challenging applications such as the 3D recon-

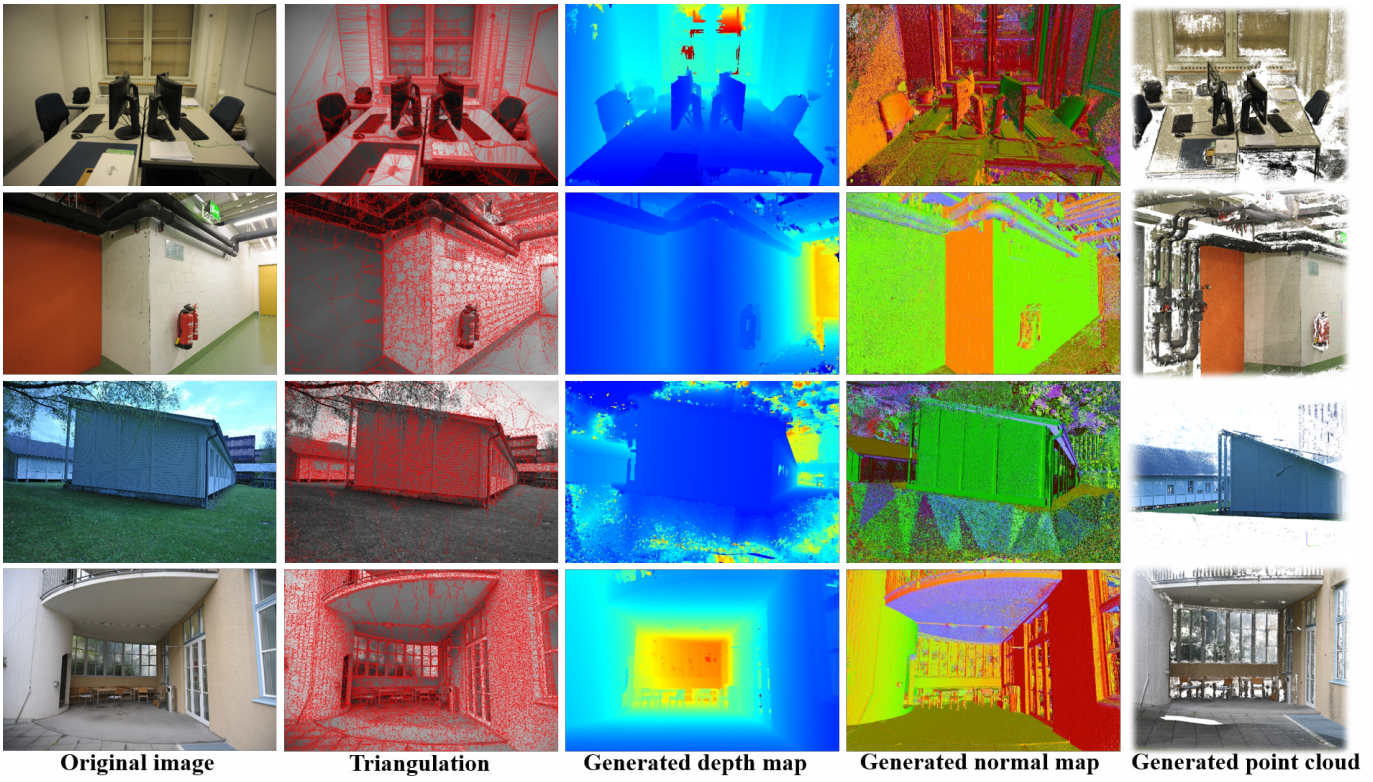

Fig. 4. Original images, triangulation, generated depth maps and normal maps, and generated point clouds by the proposed method for ETH3D training set (Office, Pipes, Meadow, and Terrace).

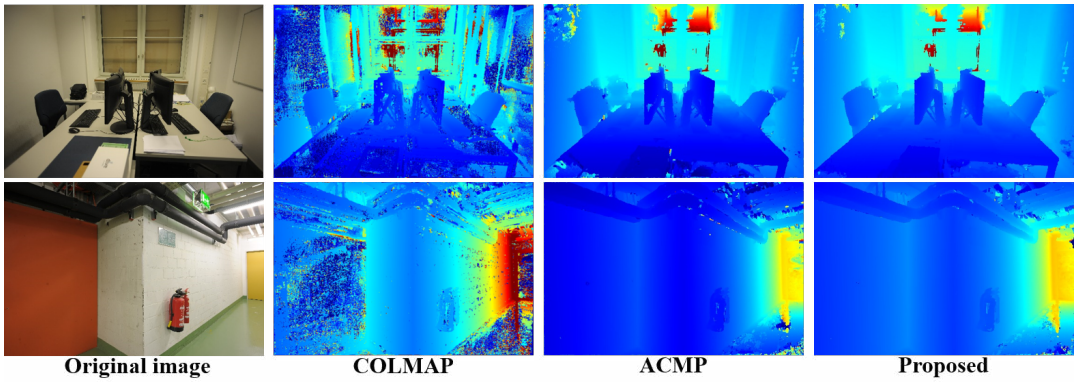

Fig. 5. Qualitative comparison of the depth maps generated by COLMAP, ACMP, and the proposed method for Office and Pipes image sets from ETH3D training set.

struction and modeling of indoor scenes, in combination with various machine-learning approaches.

REFERENCES

[1] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "PatchMatch: A Randomized Correspondence Algorithm for Structural Image Editing," in *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 2009.

[2] E. Zheng, E. Dunn, V. Jojic, and J.M. Frahm, "PatchMatch Based Joint View Selection and Depthmap Estimation," *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[3] J. L. Schönberger, E. Zhang, J. M. Frahm, and M. Pollefeys, "Pixelwise View Selection for Unstructured Multi-View Stereo," *European Conference on Computer Vision*, 2016.

[4] S. Galliani, K. Lasinger, and K. Schindler, "Massively parallel multiview stereopsis by surface normal diffusion," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[5] Q. Xu and W. Tao, "Multi-Scale Geometric Consistency Guided Multi-View Stereo," *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[6] J. Liao, Y. Fu, Q. Yan, and C. Xiao, "Pyramid multi-view stereo with local consistency," *Computer Graphics Forum*, 2019.

[7] Q. Xu and W. Tao, "Planar Prior Assisted PatchMatch Multi-View Stereo," *AAAI Conference on Artificial Intelligence*, 2020.

[8] Z. Xu, Y. Liu, X. Shi, Y. Wang, and Y. Zheng, "MARMVS: Matching Ambiguity Reduced Multiple View Stereo for Efficient Large Scale Scene Reconstruction," *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[9] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "MVSNet: Depth Inference for Unstructured Multi-view Stereo," *European Conference on Computer Vision*, 2018.

[10] Y. Yao, Z. Luo, T. Shen, S. Li, T. Fang, and L. Quan, "Recurrent MVSNet for High-Resolution Multi-View Stereo Depth Inference," *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[11] X. Gu, Z. Fan, Z. Dai, S. Zhu, F. Tan, and P. Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[12] X. Ma, Y. Gong, Q. Wang, J. Huang, L. Chen, and F. Yu, "Epp-mvsnet: Epipolar-assembling based depth prediction for multi-view stereo," *International Conference on Computer Vision*, 2021.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[14] F. Qin, Y. Ye, G. Huang, X. Chi, Y. He, X. Wang, Z. Zhu, and X. Wang, "MVSTER: Epipolar Transformer for Efficient Multi-View Stereo," *Proceedings of the European conference on computer vision*, 2022.

[15] B. Mildenhall, P. P. Srinivasan, Ma. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," *Proceedings of the European conference on computer vision*, 2020.

[16] T. Schöps, J. L. Schönberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger, "A Multi-View Stereo Benchmark with High-Resolution Images and Multi-Camera Videos," *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.