Special Paper
방송공학회논문지 제28권 제7호, 2023년 12월 (JBE Vol. 28, No. 7, December 2023)
https://doi.org/10.5909/JBE.2023.28.7.925
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

A Simple and Efficient Method to Combine Depth Map with Planar Prior for Multi-View Stereo

Hanshin Lim[a)‡] and Hyon-Gon Choo[a)]

Abstract

This paper proposes a simple and efficient method to combine planar prior obtained from a triangular model with the depth values computed by PatchMatch. Through this approach, the accuracy of depth estimation was improved in low-textured areas where existing PatchMatch-based methods showed low depth estimation accuracy, and in detailed areas where the triangular model showed limitations in surface modeling. In the proposed method, first, initial depth and normal maps are obtained via an existing PatchMatch approach. Subsequently, a triangular model is generated from the selected node, and initial planar priors are derived from this model. After refining the initial planar priors via a bilateral technique, the initial depth and normal maps are combined with the refined planar priors by applying a simple and efficient combining process. Experimental results show that the proposed method achieves superior results compared to existing approaches.

Keyword : Multi-View Stereo, Depth Map Generation, Point Cloud Generation

Ⅰ. Introduction

Technologies for reconstructing and modeling the accurate 3D structure of a scene from multi-view images taken from various viewpoints generally include extracting and matching feature points, estimating camera parameters, generating a dense point cloud, and reconstructing a 3D surface. Among these, the generation of a dense point cloud aims to accurately represent the structure of scenes and objects in multi-view images as a 3D point cloud, utilizing both the images and camera parameters. The most critical step in this process is obtaining high-quality dense depth maps from these images.

This paper proposes a method to obtain high-quality dense depth maps by simply combining the depth maps computed by the traditional PatchMatch[1] technique and the triangular model. Through this approach, the accuracy

a) Electronics and Telecommunications Research Institute
‡ Corresponding Author : Hanshin Lim
E-mail: hslim@etri.re.kr
Tel: +82-42-860-1155
ORCID: https://orcid.org/0000-0003-4829-2893
※ 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].
· Manuscript November 29, 2023; Revised December 11, 2023; Accepted December 11, 2023.

of depth estimation is improved in low-textured areas, where existing PatchMatch-based methods showed low depth estimation accuracy, and in detailed areas, where the triangular model showed limitations in surface modeling. Fig. 1 shows a color image from the Auditorium image set in the Tank and Temples Advanced test set, and the depth map, normal map, and the point cloud generated by applying the proposed method.

Main contributions of the paper are as follows:

1) We propose a simple and efficient method to combine planar prior obtained from a triangular model with the depth values computed by PatchMatch to take advantage of both approaches.
2) We demonstrate that the proposed approach achieves superior results compared to the existing methods on a public dataset.

II. Previous Works

Recent multi-view stereo technologies can be broadly classified into learning-based[2][3][4] and traditional approaches[5][6][7]. Among the traditional approaches for generating high-quality depth maps from multi-view images, the PatchMatch method is one of the most widely used. In multi-view stereo, depth values are estimated based on the color similarity between a reference image and several source images, making it crucial to accurately and efficiently reduce the number of depth hypotheses. Since the PatchMatch technique efficiently reduces these hypotheses, it has been successfully applied in the field of multi-view stereo for depth map generation. The PatchMatch technique consists of random initialization, propagation, and random search steps. Based on the strong spatial coherence between current and neighboring pixels, the values of the neighboring pixels become hypotheses for the current pixel in the propagation step, and random search values become hypotheses to avoid local minima in the random search step. However, PatchMatch-based methods are known to perform poorly in estimating depth values in low-textured areas, where assessing depth based on photometric similarities is difficult.

Compared to the PatchMatch method, segment-based approaches such as the triangular model can obtain more stable results in areas where photometric consistency is unreliable, such as low-textured regions. However, these approaches have a limitation in representing detailed areas.

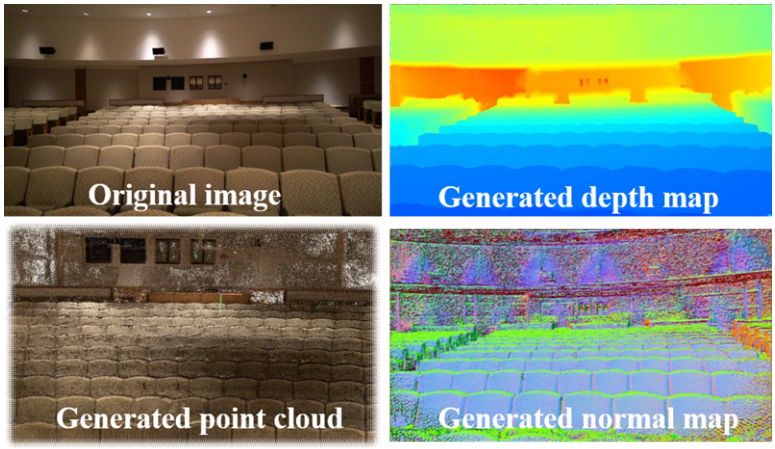


Fig. 1. A color image from the Auditorium image set in the Tank and Temples Advanced test set, and the depth map, normal map, and the point cloud generated by applying the proposed method.

ACMP[7] incorporated a planar prior term obtained by the triangular model into the matching cost in the PatchMatch process to take advantage of both the PatchMatch and triangular model methods. ACMP repeats the conventional PatchMatch process by applying the depth and normal maps obtained via the planar-prior-assisted matching cost as initial values.

Ⅲ. Proposed Method

1. Fundamental Concept

As mentioned earlier, while PatchMatch approaches can obtain relatively accurate depth values in textured or detailed areas, triangular models represent surfaces of low-textured areas more reliably. ACMP added a planar prior term to the matching cost in the PatchMatch process to leverage the advantages of both approaches. However, since the solution space, a set of the hypotheses, is not altered during the matching process, the planar prior is not adequately integrated into the depth map generated by PatchMatch.

The proposed method reduces the solution space for each pixel (x, y) as follows:

$$\{(d_{cur}, n_{cur}), (d_{planar}, n_{planar})\} \quad (1)$$

where (d_{cur}, n_{cur}) is the initial depth and normal values and (d_{planar}, n_{planar}) is the planar prior computed in the triangular model for each pixel (x, y).

It is observed that reducing the solution space is equivalent to simply combining the planar prior with the initial depth values computed by PatchMatch. Since a sufficient planar prior is added to the depth map, the result also becomes a better initialization for additional PatchMatch processes.

In addition, the proposed method refines the planar priors prior to the combining process, enhancing the reliability of the planar parameters. Fig. 2 shows the fundamental concept of the proposed method.

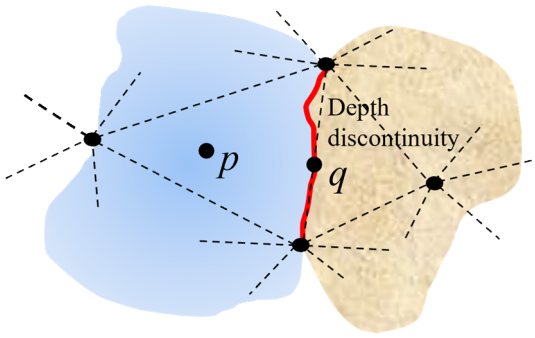


Fig. 2. Fundamental concept of the proposed method: While the PatchMatch approaches can obtain relatively accurate depth values in the textured or detailed areas, such as pixel q, triangular models represent surfaces in the low-textured areas, such as pixel p, more reliably. The proposed method simply selects the more reliable one from these two depth values computed by the two approaches for each pixel.

2. Overall Procedure

Fig. 3 shows the overall procedure of the proposed method. More detailed explanation of the proposed depth

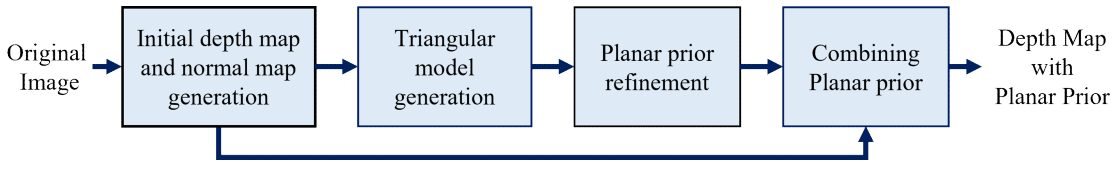


Fig. 3. Overall procedure of the proposed depth map and planar prior combining strategy. First, initial depth and normal maps are obtained from the original images. After that, a triangular model is generated from the selected nodes, and an initial planar prior is computed from the triangular model. The initial planar prior is then refined by applying a bilateral approach. The planar prior is combined with the initial depth and normal maps using a simple selection method.

map and planar prior combining process is as follows:

2.1 Initial depth map and normal map generation

First, initial depth and normal maps are obtained via one of the existing PatchMatch approaches. ACMH[6] was applied in this paper.

2.2 Triangular model generation

For each nxn block in the image domain, a pixel position is selected as a node if 1) it has the largest matching cost in the block, and 2) more than 25% of the pixels in the block have matching costs that exceed a certain threshold. From the selected nodes, a triangular model is generated, and initial planar priors are computed based on the locations and depth values of the nodes.

2.3 Planar prior refinement

Although the nodes are elaborately selected at the prior step, the initial planar priors are inaccurate in some areas because of erroneous and noisy depth values. In order to improve the reliability of the planar priors, the proposed method applies a bilateral technique to refine the planar priors in the triangular model.

2.4 Combining Planar prior

The refined planar priors are combined with the initial depth and normal maps by applying the planar-prior-assisted matching cost[7] with the reduced solution space.

$$C(d,n) = Photo(d,n) + \log\left[\gamma + e^{-\left(\frac{(d-d_{planar})^2}{2\lambda_d} + \frac{\arccos^2 n^\top n_{planar}}{2\lambda_n}\right)}\right] \quad (2)$$

where (d_{planar}, n_{planar}) is the refined planar prior, $\gamma = 0.5$, and λ_d and λ_n are respectively depth and normal bandwidths.

In the combining process, for each pixel, among the two hypotheses (d_{cur}, n_{cur}) and (d_{planar}, n_{planar}), the one with the higher matching cost becomes the depth and normal value in the combined depth and normal maps.

After the combining process, conventional PatchMatch processes are repeated, using the combined depth and normal maps as initial values.

Ⅳ. Experimental Results

1. Test Condition

The experiment was performed using an 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 Tanks and Temples Advanced test set[8], where the resolutions are 1920x1080. The camera parameters were obtained using the Structure-from-Motion module in COLMAP.

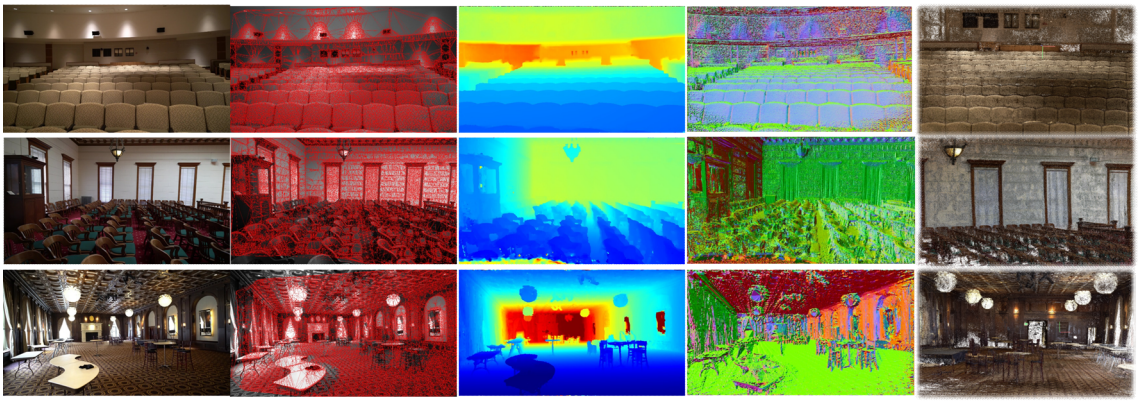
F-scores, which are the harmonic mean of precision and completeness, were calculated for the generated 3D point clouds from the depth and normal maps to quantitatively evaluate the performance of both the previous and the proposed approaches.

2. Test Results on Tanks and Temples Advanced Test Set

Table 1 shows the quantitative evaluation results for the Tanks and Temples Advanced test set. As shown in the results, the proposed method outperforms the previous learning-based and traditional PatchMatch-based approaches in terms of the mean F-scores. Fig.4 shows the original images, triangular models, depth maps, normal maps, and point clouds generated by applying the proposed method to the Tanks and Temples Advanced test set. Considering the indoor scenes (Auditorium, Ballroom, Courtroom, and

Table 1. Quantitative evaluation results of the 3D point clouds for the Tanks and Temples Advanced test set

	Auditorium	Ballroom	Courtroom	Museum	Palace	Temple	Mean
COLMAP[5]	16.02	25.23	34.70	41.51	18.05	27.94	25.83
ACMM[6]	23.41	32.91	41.17	48.13	23.87	34.60	34.02
EPP-MVSNet[3]	21.28	39.74	35.34	49.21	30.0	38.75	35.72
MVSNet[2]	24.04	44.52	36.64	49.51	30.23	39.09	37.34
ACMP[7]	30.12	34.68	44.58	50.64	27.20	37.43	37.44
MVSTER[4]	26.68	42.14	35.65	49.37	32.16	39.19	37.53
Proposed	29.88	36.15	44.25	51.74	27.39	38.73	38.02



(a) (b) (c) (d) (e)

Fig. 4. (a) Original images, (b) triangular models, (c) depth maps, (e) normal maps and (e) point clouds generated by applying the proposed method for Tank and Temples Advanced test set (Auditorium, Ballroom, and Courtroom)

Museum), which have larger low-textured regions, the mean F-score also increased to some degree. This suggests that the proposed method combines planar priors more efficiently compared to ACMP.

Ⅴ. Conclusion

This paper proposes a simple and efficient method to combine planar prior obtained from a triangular model with the depth values computed by PatchMatch. Through this approach, the accuracy of depth estimation was improved in low-textured areas where existing PatchMatch-based methods showed low depth estimation accuracy, and in detailed areas where the triangular model showed limitations in surface modeling. The experimental results showed that the proposed method produces high-quality depth maps and point clouds compared to other existing approaches by successfully combining the two approaches.

References

[1] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "Patch-Match: A Randomized Correspondence Algorithm for Structural Image Editing," *ACM Transactions on Graphics (Proc. SIGGRAPH)*, Vol.28, Issue 3, No.24, pp.1-11, 2009.
doi: https://doi.org/10.1145/1531326.1531330

[2] Y. Yao, Z. Luo, T. Shen, S. Li, T. Fang, and L. Quan, "Recurrent MVSNet for High-Resolution Multi-View Stereo Depth Inference," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.5525-5534, 2019.
doi: https://doi.ieeecomputersociety.org/10.1109/CVPR.2019.00567

[3] X. Ma, Y. Gong, Q. Wang, J. Huang, L. Chen, and F. Yu, "Eppmvsnet: Epipolar-assembling based depth prediction for multi-view stereo," *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp.5732-5740, 2021.
doi: https://doi.org/10.1109/ICCV48922.2021.00568

[4] 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,* pp.573-591, 2022.
doi: https://doi.org/10.1007/978-3-031-19821-2_33

[5] J. L. Schonberger, E. Zhang, J. M. Frahm, and M. Pollefeys, "Pixelwise View Selection for Unstructured Multi-View Stereo," *Proceedings of the European conference on computer vision*, pp.501-518, 2016.
doi: https://doi.org/10.1007/978-3-319-46487-9_31

[6] Q. Xu and W. Tao, "Multi-Scale Geometric Consistency Guided Multi-View Stereo," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.5478-5487, 2019.
doi: https://doi.ieeecomputersociety.org/10.1109/CVPR.2019.00563

[7] Q. Xu and W. Tao, "Planar Prior Assisted PatchMatch Multi-View Stereo," *Proceedings of the AAAI Conference on Artificial Intelligence*, pp.12516-12523, 2020.
doi: https://doi.org/10.1609/aaai.v34i07.6940

[8] A. Knapitsch, J. Park, Q. Y. Zhou, and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM Transactions on Graphics(Proc. SIGGRAPH)*, Vol.36, Issue 4, No.78, pp.1-13, 2017.
doi: https://doi.org/10.1145/3072959.3073599

Introduction Authors

Hanshin Lim

- 2004.02 : Dept. of Electronic and Electrocal Engineering (Minor: Mathematics), Yonsei University (B.S)
- 2006.02 : Dept. of Electronic and Electrocal Engineering, KAIST (M.S)
- 2007.09 ~ 2007.12 : Visiting Researcher, TU Berlin
- 2014.02 : Dept. of Electronic and Electrocal Engineering, KAIST (Ph.D.)
- 2014.03 ~ Current : Senior Researcher, ETRI
- ORCID : https://orcid.org/0000-0003-4829-2893
- Research interests : 2D/3D Image Processing, Computer Vision, 3D Reconstruction and Modeling, Video Coding

Hyon-Gon Choo

- Principal Researcher, ETRI, Daejeon, Korea(2005~)
- Director of the Digital Holography Section(2015~2017)
- Visiting Researcher, Warsaw University of Technology, Poland(2017-2018)
- ORCID : https://orcid.org/0000-0002-0742-5429
- Research interests : Multimedia Signal Processing with emphasis on Digital Video Processing Techniques and Applications, including 3D imaging and holography/3D depth imaging/3D broadcasting system/Computer vision