DEPTH FROM ACCIDENTAL MOTION USING GEOMETRY PRIOR

Sung-Hoon Im, Gyeongmin Choe, Hae-Gon Jeon, In So Kweon

Robotics and Computer Vision Lab, KAIST, Korea

ABSTRACT

We present a method to reconstruct dense 3D points from small camera motion. We begin with estimating sparse 3D points and camera poses by Structure from Motion (SfM) method with homography decomposition. Although the estimated points are optimized via bundle adjustment and gives reliable accuracy, the reconstructed points are sparse because it heavily depends on the extracted features of a scene. To handle this, we propose a depth propagation method using both a color prior from the images and a geometry prior from the initial points. The major benefit of our method is that we can easily handle the regions with similar colors but different depths by using the surface normal estimated from the initial points. We design our depth propagation framework into the cost minimization process. The cost function is linearly designed, which makes our optimization tractable. We demonstrate the effectiveness of our approach by comparing with a conventional method using various real-world examples.

Index Terms— Structure from motion, Small baseline, Depth propagation

1. INTRODUCTION

Estimating a depth map from multiview images is a major problem in the computer vision field since the depth map plays an important role in many applications such as scene understanding and photographic editing. The most representative approach is SfM [1] which estimates 3D points and camera poses at the same time. Moreover, bundle adjustment [2] optimizes both the 3D points and camera poses accurately. Although the solution is theoretically optimal, when the camera baseline is too small, the metric depth error increases quadratically even with small errors in disparity [3].

Various attempts in computational photography have been made to measure depth information of a scene without camera motion. In [4, 5, 6], modifications of camera apertures for robust depth from defocusing were proposed as an alternative way to compute depth maps. Another approach is by light-field photography which uses an angular and spatial information of incoming light ray in an image domain. This allows to obtain multi-view images of a scene in a single shot [7, 8], and the multi-view images are used for depth map estimation [9, 10, 11]. Although these recent progresses of computational photography show alternative ways to compute depth map, they are not available in practice because they require either camera modification or loss of image resolution.

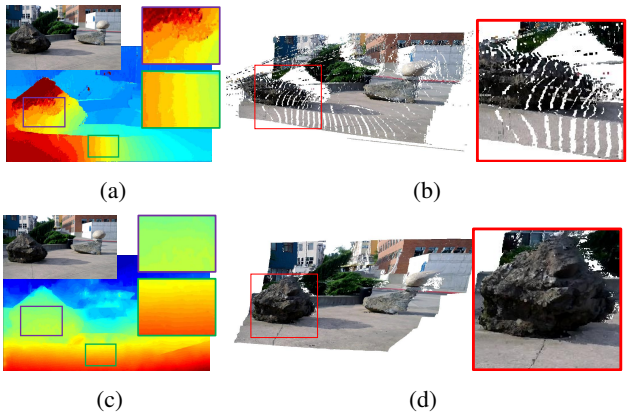


Fig. 1: Result comparison. (a) Depth map by [12]. (b) 3D point cloud from (a). (c) Our depth map. (d) 3D point cloud from (c).

Recently, Yu and Gallup [12] show that 3D points can be reconstructed from multiple views via an accidental motion caused by hand shaking or heart beating. It has the potential to overcome the limitations of coded aperture and light-field by capturing a short video without large movement. Since the baseline between consecutive image sequences is narrow, the camera poses at each view can be initialized as identical and can be used for the bundle adjustment. Although the underlying assumption is reasonable, the depth map and 3D point cloud in Fig. 1-(a), (b) show inaccuracy in the results because of unreliable sparse 3D points, which is unsatisfactory for the high-quality 3D modeling.

In this paper, we target to obtain an accurate 3D point cloud as well as a depth map from small motion. First, we use SfM to estimate initial sparse 3D points and camera poses from the small motion. In contrast to [12], our method provides a good initial solution of the camera poses by homography decomposition for accurate sparse 3d reconstruction. For dense 3D reconstruction, we propose a depth propagation method using both a color prior from the images and a geometry prior from the initial points while the conventional propagation method [13] uses only a color prior. Our depth propagation is designed into the linear cost minimization framework, which effectively improves the depth quality as shown in Fig. 1-(c), (d).

(a) (b)

Fig. 2: Feature handling. (a) Features on moving objects. (b) Feature removal - Stone [12] : 1920 × 1080

2. PROPOSED METHOD

In this section, we describe our overall dense 3D reconstruction framework for tiny baseline images. Our system is similar to the 3D reconstruction from accidental motion proposed in [12]. The overall procedure consists of feature extraction, initial sparse 3D reconstruction, and dense 3D reconstruction. We improve the performance in every step on [12], and depth propagation in Section 2.2 which is the most significant improvement in this paper.

2.1. Structure from small motion

This section describes the way to extract reliable features and reconstruct 3D points for small motion precisely. The key observation is homography between reference view and the other views.

Feature extraction It is important to extract and match features precisely for narrow baseline multiview. If there are some unreliable feature matchings and features on moving objects, it causes significant error and should be removed. [12] tracked corner features by Kanade-Lucas-Tomasi (KLT) tracker [14] and removed feature outlier by maximum color gradient difference. It can only handle features with high localization error, but not features on moving objects. We filter out the features on moving objects by RANSAC [15] for the homography as shown in Fig. 2. Homography H for each image is computed by matched feature points and homography outliers can be detected by RANSAC. If the features are classified as outliers more than m times, we regard them as features on moving objects.

Sparse 3D reconstruction The key idea of 3D reconstruction from small motion is to directly apply bundle adjustment with approximated initial depth and camera poses. [12] assumed that the identity matrix for rotation matrix R, zero translation T, and randomly assigned depth for initial depth d are good initializations for bundle adjustment input. However, the initial parameters are rough approximations that they are not suitable for our purpose which is computing reliable 3D points. We propose to set initial camera poses as the decomposition of homography [16]. Homography H is composed as:

$$H = K(R^T + \frac{R^T T \mathbf{n}^T}{d})K^{-1}, \quad (1)$$

Dataset	Stone (Fig. 1)	Wall (Fig. 4)	Shop (Fig. 5)
[12]	0.038	0.308	0.749
Proposed	**0.003**	**0.094**	**0.072**

Table 1: Initial average reprojection errors comparison between the conventional [12] and proposed initial camera poses (Unit : pixel).

where K and $\mathbf{n}$ are camera intrinsic parameter and surface normal. Even though the decomposed rotation matrix R and translation matrix T with unknown normal vector $\mathbf{n}$ and depth d are not precise camera poses, they reduce initial reprojection error leading to well refined camera poses and 3D points. Thus decomposed camera poses can be reliable initial camera poses for bundle adjustment. Table 1. represents the initial projection error comparison between rough initial camera poses and decomposed initial camera poses. Initial reprojection errors of our initial camera poses are less than that of the conventional method, and more accurate sparse points can be obtained.

With reliable initial parameters, bundle adjustment successfully refine depth and camera poses. The cost function of bundle adjustment is the L_2 norm of the reprojection error defined as:

$$F = \sum_{i=1}^{N_I} \sum_{j=1}^{N_F} ||p_{ij} - \pi(K(R_i P_j + T_i))||^2, \quad (2)$$

where N_I and N_F are the number of images and features. Matrix R_i and T_i are camera rotation and translation for each view point i. World coordinates $P_j = d_j[x_{1j}, y_{1j}, 1]$ for each feature point j are the depth d_j multiplication with normalized image coordinates $p_{ij} = [x_{ij}, y_{ij}]$ of reference view. The function $\pi : \mathbf{R}^3 \rightarrow \mathbf{R}^2$ is the projection function from 3D to 2D coordinates. The cost function in Eq. (2) is optimized with Levenberg marquardt (LM) algorithm [17].

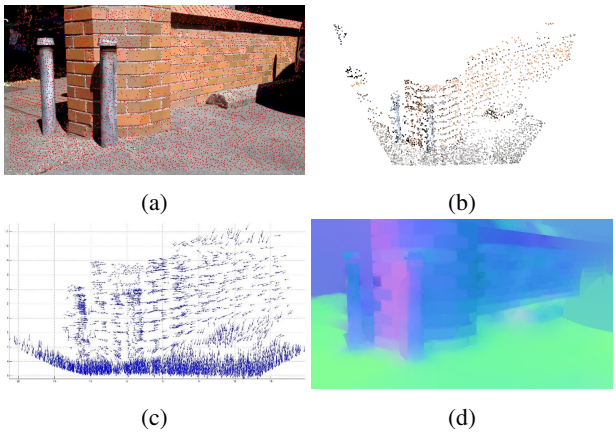
2.2. Depth propagation

This subsection describes a method to propagate the sparse depth points initially estimated in Sec. 2.1 into dense depth. Generally, depth propagation using a color cue [18, 19, 13] is frequently used, but there are many cases where crucial artifacts can occur, especially, in the region where neighboring pixels with similar color have different depth. To handle this unreliability, we propose a novel optimization method. We design our propagation method minimizing cost function defined as:

$$E(D) = E_c(D) + \lambda E_g(D), \quad (3)$$

where D is the optimal depth. E_c and E_g are color and geometry terms with regularization parameter λ.

Color consistency A color consistency term is designed based on [13] . We assume that the scene depth field is always piecewise smooth. The main idea is that a similar depth



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

Fig. 3: Normal vector estimation. (a) Reference image with features. (b) Sparse 3D reconstruction. (c) Sparse normal vectors. (d) Normal map.

value tends to be assigned to an adjacent pixel with high color affinity. The cost function for color is defined as:

$$E_c(D) = \sum_{p \in I_{ref}} \left(D_p - \sum_{q \in N_8(p)} w^c_{pq} D_q \right)^2, \quad (4)$$

where q is the 8 neighbors of p which belong to pixels in the reference image I_{ref}. The color similarity weight w^c_{pq} between pixel p and q in lab color space is defined as

$$w^c_{pq} = \frac{1}{N_c} \exp\left(\frac{-|I^{lab}_p - I^{lab}_q|}{\gamma_c} \right), \quad (5)$$

where N_c is the normalization factor which makes the sum of 8-neighboring w^c_{pq} equal to one and γ_c is the strength parameter for similarity measure which is manually tuned.

Geometric consistency We assume that a normal vector at point p is perpendicular to vectors from point p to adjacent sparse control points q on the same plane. The geometry term E_g is defined as the sum of the inner products between normal vector at point p and vectors pointing to adjacent sparse control points q.

$$E_g(D) = \sum_{p \in I_{ref}} \sum_{q \in G_w(p)} w^g_{pq} |\mathbf{n}_p \cdot (D_p X_p - D_q X_q))|, \quad (6)$$

where q is the set of sparse control points $G_w(p)$ at the center p within the 2D local window l_w. The normalized image coordinate X_p and the normal vector $\mathbf{n}_p$ are represented by $[x_p, y_p, 1]$ and $[a_p, b_p, c_p]$ at the pixel p, respectively. Normal vector $\mathbf{n}_p$ is computed by plane fitting in advance. A plane can be fit using adjacent sparse control points in 3D space and the fitted plane estimates the adjacent normal vectors. Algorithm 1. shows the normal vector estimation step. To ensure sparse points q are on the same plane with point p, the normal similarity measure w^g_{pq} between points p and q is defined as

$$w^g_{pq} = \frac{1}{N_g} \exp\left(\frac{-(1 - \mathbf{n}_p \cdot \mathbf{n}_q)}{\gamma_g} \right), \quad (7)$$

Algorithm 1 Normal vector estimation

$N_d \leftarrow$ The number of adjacent sparse control points within a 3D sphere with the radius r_d
if $N_d < 2$ **then**
  Skip to calculate normal vectors
else
  Compute normal vectors $\{\mathbf{n}\}$ by plane fitting
  $[U, D, V] = svd(\mathbf{n}'\mathbf{n})$
  Refined normal vectors $\leftarrow$ First column of U
end if
Normal vector propagation by color energy function

where N_g is the number of sparse control points within local window l_w and γ_g controls the strength of similarity measure. It measures how much the normal vectors of p and q are correlated.

Linear equation Our cost minimization is efficiently solved by the linear equations $Ax = b$. To find the optimal solution of the color energy function, we solve $\nabla E_c(D) = 0$ which is defined as:

$$\nabla E_c(D) = (I - W^c)D, \quad (8)$$

where I is the $M \times M$ matrix (M is the number of pixels) and W^c is the pairwise color similarity (w^c_{pq}) matrix. D is a one dimensional vector that we optimize.
For the geometry energy function, we derive from (6),

$$\sum_{p \in I_{ref}} \sum_{q \in G_w(p)} w^g_{pq} D_p - s_{pq} = 0, \quad (9)$$

$$\text{where } s_{pq} = w^g_{pq} D_q \frac{a_p x_q + b_p y_q + c_p}{a_p x_p + b_p y_p + c_p}. \quad (10)$$

and we obtain matrix form as:

$$W^g D - S = 0, \quad (11)$$

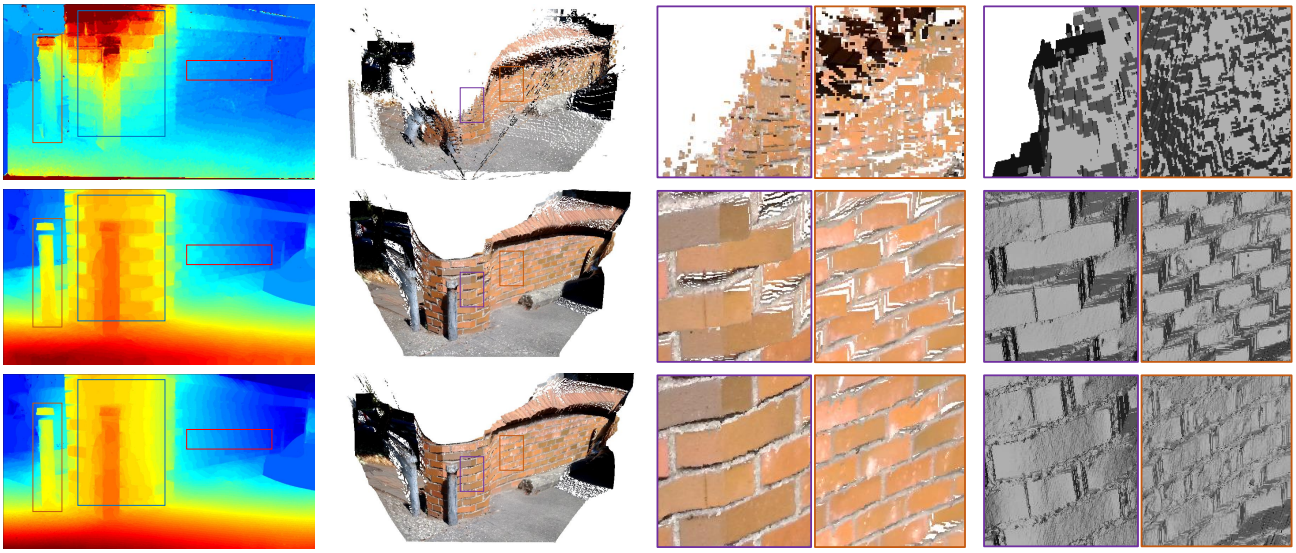
where W^g is the $M \times M$ the pairwise normal similarity (w^g_{pq}) matrix and S is the pairwise element of s_{pq}.
Both terms are combined with a regularization term λ :

$$A_t(p,q) = \begin{cases} I(p,q) & p \in G; \\ ((I - W^c) + \lambda W^g)(p,q) & p \notin G. \end{cases} \quad (12)$$

$$b_t(p) = \begin{cases} G_p & p \in G; \\ \lambda S_p & p \notin G. \end{cases} \quad (13)$$

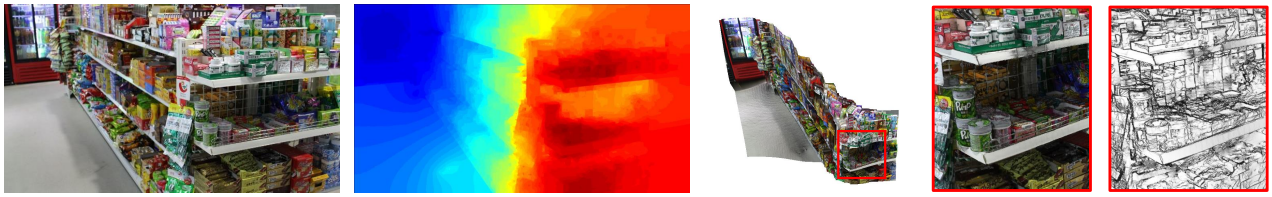
3. EXPERIMENTAL RESULTS

The proposed algorithm is carried out with the parameters as follows. We set both strength of similarity measure γ_c and γ_g as 0.01 and use 49 $\times$ 49 support window l_w. The radius r_d is calculated by dividing the initial depth over 20. To evaluate the performance of the proposed algorithm, we conduct



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

Fig. 4: Experiment Result - Wall [12] : 1920 × 1080. (a) Depth map. (b) Overall 3D point cloud. (c) Enlarged box of (b). (d) Mesh without texture mapping; First row : Yu and Gallup [12], Second row : conventional propagation, Third row : proposed propagation.



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

Fig. 5: Result of own dataset - Shop : 1296 × 728. (a) Reference image. (b) Depth map. (c) Mesh with texture mapping. (d) Enlarged box of (c). (d) Mesh without texture mapping.

experiments using datasets provided by the author[1] (Fig. 1-4) and our own dataset (Fig. 5). To verify the effectiveness of our initial camera pose estimation, we compare our result with the results obtained from [12] and they are shown in the $1st$ and $3rd$ rows of Fig. 4. As expected, our dense reconstruction result has less artifacts and more sense of reality. Furthermore, we compare our novel depth propagation method with the conventional method shown in the $2nd$ and $3rd$ rows of Fig. 4, respectively. We show detailed 3D point cloud and mesh in Fig. 4(c)-(d). While the conventional depth propagation method makes a lot of discontinuity across the bricks, the proposed method continuously propagates the depth values, which reconstructs the overall 3D points more reliably.

Fig. 5 is the result from our own dataset taken by Cannon EOS60D. We take 90 frames for 3 seconds with less than 10mm translation. The depth range of our dataset is from 1m to 10m. As shown in Fig. 5(b)-(c), the overall depth range is accurately reconstructed with our method. The details of the 3D model are represented in Fig. 5(d)-(e). Supplementary video and high-resolution results of various outdoor and indoor scenes are available online.

[1] http://yf.io/p/tiny/

4. CONCLUSION AND DISCUSSION

We have presented a novel method to reconstruct initial sparse 3D points and propagate them into dense 3D structure under narrow-baseline, multi-view imaging setup. By using features with less localization errors and more reliable initial camera poses, more accurate sparse 3D points were obtained. Furthermore, we were able to achieve better dense 3D points by solving simple linear equations. In the future, we will try to improve our method robust to motion and depth range by using additional sensors in a cell-phone or DSLR, such as gyro-sensor.

5. ACKNOWLEDGEMENT

This research is supported by the Study on Imaging Systems for the next generation cameras funded by the Samsung Electronics Co., Ltd (DMC R&D center) (IO130806-00717-02).

6. REFERENCES

[1] Richard Hartley and Andrew Zisserman, *Multiple view geometry in computer vision*, Cambridge university press, 2003.

[2] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon, "Bundle adjustmenta modern synthesis," in *Vision algorithms: theory and practice*, pp. 298–372. Springer, 2000.

[3] David Gallup, J-M Frahm, Philippos Mordohai, and Marc Pollefeys, "Variable baseline/resolution stereo," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2008.

[4] Changyin Zhou, Stephen Lin, and Shree K Nayar, "Coded aperture pairs for depth from defocus and defocus deblurring," *Int'l Journal of Computer Vision*, vol. 93, no. 1, pp. 53–72, 2011.

[5] Paul Green, Wenyang Sun, Wojciech Matusik, and Frédo Durand, "Multi-aperture photography," *ACM Trans. on Graph.*, vol. 26, no. 3, pp. 68, 2007.

[6] Ayan Chakrabarti and Todd Zickler, "Depth and deblurring from a spectrally-varying depth-of-field," in *Proc. of European Conf. on Computer Vision (ECCV)*. 2012, Springer.

[7] Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, and Pat Hanrahan, "Light field photography with a hand-held plenoptic camera," *Stanford University Computer Science Technical Report CSTR*, vol. 2, no. 11, 2005.

[8] Andrew Lumsdaine and Todor Georgiev, "The focused plenoptic camera," IEEE, 2009.

[9] Sven Wanner and Bastian Goldluecke, "Globally consistent depth labeling of 4d light fields," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012.

[10] Michael W Tao, Sunil Hadap, Jitendra Malik, and Ravi Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in *Proc. of Int'l Conf. on Computer Vision (ICCV)*, 2013, pp. 673–680.

[11] Hae-Gon Jeon, Jaesik Park, Gyeongmin Choe, Jinsun Park, Yunsu Bok, Yu-Wing Tai, and In So Kweon, "Accurate depth map estimation from a lenslet light field camera," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2015.

[12] Fisher Yu and David Gallup, "3d reconstruction from accidental motion," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014.

[13] Liang Wang and Ruigang Yang, "Global stereo matching leveraged by sparse ground control points," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2011.

[14] Carlo Tomasi and Takeo Kanade, *Detection and tracking of point features*, School of Computer Science, Carnegie Mellon Univ. Pittsburgh, 1991.

[15] Martin A Fischler and Robert C Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[16] Yi Ma, *An invitation to 3-d vision: from images to geometric models*, vol. 26, springer, 2004.

[17] Jorge J Moré, *The Levenberg-Marquardt algorithm: implementation and theory*, Springer, 1978.

[18] J. Park, H. Kim, Y.-W. Tai, M. S. Brown, and I. S. Kweon, "High quality depth map upsampling for 3d-tof cameras," in *Proc. of Int'l Conf. on Computer Vision (ICCV)*, 2011.

[19] G. Choe, J. Park, Y.-W. Tai, and I.S. Kweon, "Exploiting shading cues in kinect ir images for geometry refinement," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2014.