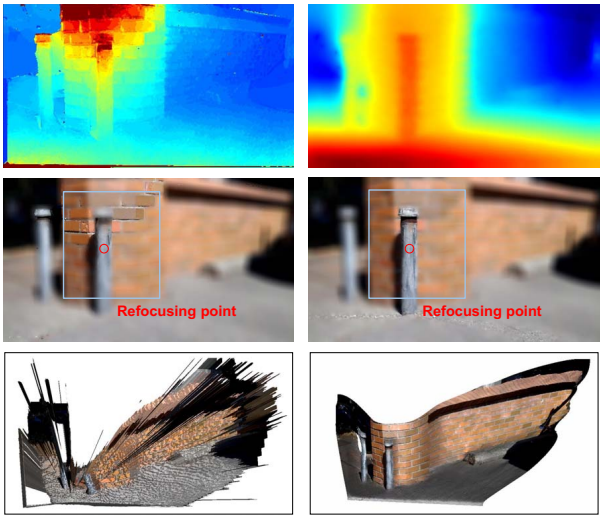
High Quality Structure from Small Motion for Rolling Shutter Cameras

Sunghoon Im Hyowon Ha Gyeongmin Choe Hae-Gon Jeon Kyungdon Joo In So Kweon
Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea
{shim, hwha, gmchoe, hgjeon, kdjoo}@rcv.kaist.ac.kr, iskweon77@kaist.ac.kr

Abstract

We present a practical 3D reconstruction method to obtain a high-quality dense depth map from narrow-baseline image sequences captured by commercial digital cameras, such as DSLRs or mobile phones. Depth estimation from small motion has gained interest as a means of various photographic editing, but important limitations present themselves in the form of depth uncertainty due to a narrow baseline and rolling shutter. To address these problems, we introduce a novel 3D reconstruction method from narrow-baseline image sequences that effectively handles the effects of a rolling shutter that occur from most of commercial digital cameras. Additionally, we present a depth propagation method to fill in the holes associated with the unknown pixels based on our novel geometric guidance model. Both qualitative and quantitative experimental results show that our new algorithm consistently generates better 3D depth maps than those by the state-of-the-art method.



(a) Yu and Gallup [33] (b) Our results

Figure 1. Comparison of the proposed method with the state-of-the-art. Top : Depth maps. Middle : Synthetic defocused images based on the depth maps. Bottom : 3D meshes.

1. Introduction

With the widespread use of commercial cameras and the continuous growth observed in computing power, consumers are starting to expect more variety with the photographic applications on their mobile devices. Various features, such as refocusing, 3D parallax and extended depth of field, are a few examples of sought-after functions in such devices [1, 2]. To meet these needs, estimating 3D information is becoming an increasingly important technique, and numerous research efforts have focused on computing accurate 3D information at a low cost.

Light-field imaging and stereo imaging have been explored as possible solutions. Light-field imaging products utilize a micro-lens array in front of its CCD sensor to capture aligned multi-view images in a single shot. The captured multi-view images are used to compute depth maps and to produce refocused images. The problem with this approach is that it requires highly specialized hardware and it also suffers from a resolution trade-off, which significantly reduces the resulting 3D spatial resolution *e.g.* Lytro [1] and Pelican [30]. Stereo imaging is an alternative method that works by finding correspondences of the same feature points between two rectified images of the same scene [2, 34]. Although this method shows reliable depth results, both cameras are required to be calibrated beforehand and must maintain their calibrated state, which makes it cumbersome and costly for many applications.

One research direction that has led to renewed interest is the depth estimation of narrow-baseline image sequences captured by off-the-shelf cameras, such as DSLRs or mobile phone cameras [15, 33, 13]. The main advantage of these approaches is that 3D information can be estimated by an off-the-shelf camera without the need for additional devices or camera modifications. However, these methods use images with a narrow-baseline, a few *mm*, often failing to generate reasonable depth maps if existing multi-view stereo such as [8] were to be applied directly. Additionally, we observe that the rolling shutter (RS) used in most digital cameras causes severe geometric artifacts and results in severe errors in 3D reconstruction. These artifacts commonly occur when the motion is at a higher frequency than

1550-5499/15 $31.00 © 2015 IEEE
DOI 10.1109/ICCV.2015.102

the frame rate of the camera, like when the user's hands are shaking [7, 12].

In this paper, we propose an accurate 3D reconstruction method from narrow-baseline image sequences taken by a digital camera. We call this approach *Structure from Small Motion (SfSM)*. Our major contributions are threefold. We first present a model for a RS which effectively removes the geometrical distortions even under narrow-baseline in Sec. 3.2. Secondly, we extract supportive features and accurate initial camera poses to use as our bundle adjustment inputs in Sec. 3.3. Finally, we propose a new dense reconstruction method from the obtained sparse 3D point cloud in Sec. 4. To demonstrate the effectiveness of our algorithm, we evaluate our results on both qualitative and quantitative experiments in Sec. 5.2. To measure the competitiveness, we draw comparisons with the results from the state-of-the-art method [33] and depth from the Microsoft Kinect2 [14] in Sec. 5.3. In terms of its usefulness, we show the user-friendliness of our work, providing a realistic digital refocusing application in Sec. 5.4.

2. Related Work

Our algorithm is composed of two modules: the first module estimates accurate 3D points from narrow-baseline image sequences, and the second module computes a dense 3D depth map via linear propagation based on both color and geometric cues. We refer the reader to [26, 11] for a comprehensive review of 3D reconstruction with image sequences.

Depth from narrow baseline As is widely known, 3D reconstruction from a narrow baseline is a very challenging task. The magnitude of the disparities are reduced to sub-pixel levels, and the depth error grows quadratically with respect to the decreasing baseline width [9]. In this context, there are other ways to estimate 3D information from the narrow-baseline instead of the conventional correspondence matching in computer vision.

Kim *et al*. [16] capture a massive number of images from a DSLR camera with intentional linear movement and compute high-resolution depth maps by processing individual light rays instead of image patches. Morgan *et al*. [20] present sub-pixel disparity estimation using phase-correlation based stereo matching and demonstrate good depth results using satellite image pairs. However, these approaches work well under a controlled environment but cannot handle moving objects in the scene.

A more general approach is to use video sequences as presented in [33, 15]. Yu and Gallup [33] utilize random depth points relative to a reference view and identical camera poses for the initialization of the bundle adjustment. The bundle adjustment produces the camera poses and sparse 3D points. Based on the output camera poses, a plane sweeping algorithm is performed to reconstruct a dense depth map. Joshi and Zitnick [15] compute per-pixel optical flow to estimate camera projection matrices of image sequences. Then, the computed projection matrices are used to align the images, and a dense disparity map is computed by rank-1 factorization.

While the studies in [33, 15] have a purpose similar to our work in terms of depth from narrow-baseline image sequences, we observe that the performance depends on the presence of the RS effect.

Rolling shutter Most off-the-shelf cameras are equipped with a RS due to the manufacturing cost. However, the RS causes distortions in the image when the camera is moving. This distortion limits the performances of 3D reconstruction algorithms, such as Structure from Motion (SfM). Many works in [7, 12, 17, 22] have recently studied how to handle the RS effect. Forssen *et al*. [7] rectify the RS video through a linear interpolation scheme for camera translations and a spherical linear interpolation (SLERP) [27] for camera rotations. Hedborg *et al*. [12] formulate the RS bundle adjustment for general SfM using the SLERP schemes. While the RS bundle adjustment is effective in refining the camera poses and 3D points in a wide-baseline condition, it is inadequate for being applied to the bundle adjustment for small motion due to the high order of the SLERP model. Therefore, we formulate a new RS bundle adjustment with a simple but effective interpolation scheme for small motion.

Depth propagation Depth propagation is an important task that produces a dense depth map. Conventional depth propagation assumes that pixels with similar color are of similar depth to that of neighboring pixels [6]. Wang *et al*. [31] propose a closed-form linear least square approximation to propagate ground control points in stereo matching. Park *et al*. [23] propose a combinatory model of different weighting terms that represent segmentation, gradients and non-local means for depth up-sampling. However, the assumption is too strong because the geometric information is barely correlated with color intensity, as mentioned in [3]. In our framework, we propose a linear least square approximation with a new geometric guidance term. The geometric guidance term is computed using the normal information from a set of initial 3D points and ultimately helps to obtain a geometrically consistent depth map.

3. Structure from Small Motion (SfSM)

The main objective of the proposed method is to reconstruct a dense 3D structure of the scene captured in image sequences with small motion. To achieve this goal, it is extremely important to recover the initial skeleton of the 3D structure as accurately as possible. In this section, we explain the proposed SfSM method for accurate 3D reconstruction of sparse features.

3.1. Geometric Model for Small Motion

The geometric model of the proposed method is based on the conventional perspective projection model [11] which describes the relationship between a 3D point in the world and its projection onto the image plane for a perspective camera. According to this projection model, a 3D coordinate of a world point $\mathbf{X} = [X, Y, Z, 1]^\intercal$ and its corresponding 2D coordinate in the image $\mathbf{x} = [u, v, 1]^\intercal$ are described as follows:

$$s\mathbf{x} = \mathbf{KPX}, \ \textit{where}\ \mathbf{K} = \begin{bmatrix} f_x & \alpha & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}, \quad (1)$$

where s is a scale factor, $\mathbf{K}$ is the intrinsic matrix of a camera that contains focal lengths f_x and f_y, principal points c_x and c_y, and skew factor α.

In SfSM, the camera pose is modified to adopt the small angle approximation in rotation matrix representation [4]. Yu and Gallup [33] point out that this small angle approximation is the key to estimating the camera poses and the 3D points without any prior pose or depth information. Under small angular deviations, the camera extrinsic matrix for SfSM can be simplified as

$$\mathbf{P} = [R(\mathbf{r})|\ \mathbf{t}], \ \textit{where}\ R(\mathbf{r}) = \begin{bmatrix} 1 & -r^z & r^y \\ r^z & 1 & -r^x \\ -r^y & r^x & 1 \end{bmatrix}, \quad (2)$$

where $\mathbf{r} = [r^x, r^y, r^z]^\intercal$ is the rotation vector and $\mathbf{t} = [t^x, t^y, t^z]^\intercal$ is the translation vector of the camera. The function R transforms the rotation vector $\mathbf{r}$ into the approximated rotation matrix.

Since the geometric model is designed for small motion, it needs highly accurate camera poses and feature correspondences. However, a RS camera captures each row at different time instances, and each row belongs to different camera poses when the camera is moving. This causes significant error in 3D reconstruction with small motion. Therefore, we propose a new camera model covering RS cameras with small motion in Sec. 3.2, as well as a method to accurately extract features and correspondences in Sec. 3.3.

3.2. Rolling Shutter Camera Model

To overcome the RS effect, several works [12, 22] have focused on modeling the RS effect in the case of conventional SfM. In their approaches, the rotation and translation of each feature are assigned differently according to their vertical position in the image by interpolating the rotation and translation between two successive frames. To interpolate the changes of rotation and translation, usually the SLERP [7, 27] method is used for rotation and a linear interpolation is used for translation. The SLERP method is designed to cover the discontinuous change of the rotation vector caused by the periodic structure of the rotation matrix. Accordingly, it contains a complex equation for being applied in the bundle adjustment for small motion, which can hardly be achieved with a high-order model.

To include the RS effect in our camera model without increasing its order, we simplify the rotation interpolation by reformulating its expression under a linear form. Though the linear interpolation of the rotation vector is simple, it is effective in modeling the continuously changing rotation for small motion, where the rotation matrix is composed not of periodic functions, but only of linear elements. The rotation and translation vector for each feature between two consecutive frames are modeled as

$$\begin{aligned} \mathbf{r}_{ij} &= \mathbf{r}_i + \frac{ak_{ij}}{h}(\mathbf{r}_{i+1} - \mathbf{r}_i) \\ \mathbf{t}_{ij} &= \mathbf{t}_i + \frac{ak_{ij}}{h}(\mathbf{t}_{i+1} - \mathbf{t}_i). \end{aligned} \quad (3)$$

where $\mathbf{r}_{ij}$ and $\mathbf{t}_{ij}$ are the rotation and translation vectors for the j-th feature on the i-th image respectively, and a is the ratio of the readout time of the camera for one frame. h denotes the total number of the rows in the image, and k_{ij} stands for the row number of each feature. The readout time of the camera can be calculated by using the method developed by Meignast *et al.* [18]. For the global shutter camera, a is set to zero. The camera poses $\mathbf{P}_{ij}$ for RS projection model are formulated by Eq. (2) using the new $\mathbf{r}_{ij}$ and $\mathbf{t}_{ij}$. We use this camera model to build our bundle adjustment function described in Sec. 3.4.

3.3. Feature Extraction

Since the baseline for SfSM is narrow, a small error in feature correspondence results in significant artifacts on the whole reconstruction. Thus, the accurate extraction of features and correspondences is a crucial step in the proposed method. For initial feature extraction, we utilize well-known Harris corner [10] and Kanade-Lucas-Tomasi (KLT) tracker [28] to extract sub-pixel corner features in the reference frame and track them through the sequence. This scheme is feasible when the pixel changes in the subsequent frames are small.

As the next step, we filter out the outliers since the features can suffer from slipping on lines or blurry regions, and even be shifted by moving objects or the RS effect. For this process, we compute the essential matrix $\mathbf{E}$ using a 5-point algorithm based on the RANSAC [21], and then we calculate the fundamental matrix as follows: $\mathbf{F} = \mathbf{K}^{-\intercal}\mathbf{E}\mathbf{K}^{-1}$ [11]. The fundamental matrix $\mathbf{F}$ describes the relationship between two images which defined as

$$\mathbf{l}_2 = \mathbf{F}\mathbf{x}_1, \ \ \mathbf{l}_1 = \mathbf{F}^\intercal\mathbf{x}_2, \ \ \mathbf{l}^\intercal{}_1\mathbf{x}_1 = 0, \ \ \mathbf{l}^\intercal{}_2\mathbf{x}_2 = 0 \quad (4)$$

where $\mathbf{x}_1$ and $\mathbf{x}_2$ are the corresponding points in consecutive frames, and $\mathbf{l}_1$ and $\mathbf{l}_2$ are their corresponding epipolar

lines. In practice, the points are not exactly on the lines so that the line-to-point distance is used to check the inliers. For each pair of the reference frame and another, an essential matrix is estimated to contain the maximum number of inlier features with the line-to-point distance under 1 pixel. The final inlier set κ is only composed of points visible on 90 percent of the frames. Additionally, the extrinsic parameters, $\mathbf{r}_i$ and $\mathbf{t}_i$ are estimated by the decomposition of essential matrices for all frames [11].

3.4. Bundle Adjustment

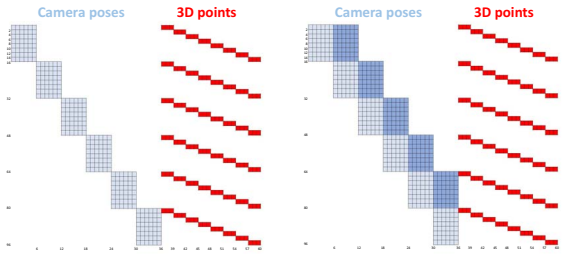
Bundle adjustment [29, 33] is a well-studied nonlinear optimization method which iteratively refines 3D points and camera parameters by minimizing the reprojection error. We formulate a new bundle adjustment for our geometric model with the proposed camera model from Sec. 3.2 and the features from Sec. 3.3. The cost function C is defined as the squared sum of all reprojection errors as follows:

$$C(\mathbf{r}, \mathbf{t}, \mathbf{X}) = \sum_{i=1}^{N_I} \sum_{j=1}^{N_J} ||\mathbf{x}_{ij} - \varphi(\mathbf{K}\mathbf{P}_{ij}\mathbf{X}_j)||^2, \quad (5)$$

where $\mathbf{x}$, $\mathbf{K}$, $\mathbf{P}$, and $\mathbf{X}$ follow the previously introduced geometric model in Eq. (1, 2), and $\mathbf{r}$, $\mathbf{t}$ follow the proposed camera model in Eq. (3). N_I and N_J are the number of images and features, and φ is a normalization function to project a 3D point into the normalized coordinate of camera as follows $\varphi([X, Y, Z]^\intercal) = [X/Z, Y/Z, 1]^\intercal$.

The bundle adjustment refines the camera parameters $\mathbf{r}$, $\mathbf{t}$ and the world coordinates $\mathbf{X}$ with a reliable initialization. We set the initial camera parameters as the decomposition of essential matrices from Sec. 3.3. We set the initial 3D coordinates for all pixels as the multiplication of their normalized image coordinates $\hat{\mathbf{X}}_j = [\hat{x}_j, \hat{y}_j, 1]^\intercal$ and a random depth value $\hat{z}$. To estimate camera poses and the 3D points that minimize the cost function in Eq. (5), the Levenberg-Marquardt (LM) method [19] is used. For computational efficiency, we compute the analytic Jacobian matrix for the proposed SfSM bundle adjustment, which is different from the Jacobian matrix for the conventional SfM. Since our rotations and translations are linearly interpolated for two consecutive frames, each residual is related to the extrinsic parameters of two viewpoints. Thus, the Jacobian matrix for the proposed method is computed as depicted in Fig. 2.

By the proposed bundle adjustment, accurate 3D reconstruction of the feature points in the images can be successfully achieved for a RS camera performing small motion. The 3D points obtained from this step are used in the subsequent stage for dense reconstruction as the robust initialization of the scene.



(a) General Jacobian matrix (b) Proposed Jacobian matrix

Figure 2. Example Jacobian matrices with 8 points (16 parameters) and 6 cameras.

4. Dense Reconstruction

The initial points obtained from Sec. 3 are geometrically well reconstructed, but the points are not dense enough for 3D scene understanding because it highly depends upon the scene characteristics and feature extraction. To overcome this sparsity, we propose a depth propagation method for dense reconstruction.

4.1. Objective Function

Our propagation can be formulated as minimizing an energy function for a depth $\mathbf{D}$ on every single pixel point. Our energy function consists of three terms: a data term $E_d(\mathbf{D})$, a color smoothness term $E_c(\mathbf{D})$ and a geometric guidance term $E_g(\mathbf{D})$ expressed as follows:

$$E(\mathbf{D}) = E_d(\mathbf{D}) + \lambda_c E_c(\mathbf{D}) + \lambda_g E_g(\mathbf{D}), \quad (6)$$

where λ_c and λ_g are the relative weights to balance the three terms. Since we formulate the three terms in quadratic forms, the depth that minimizes $E(\mathbf{D})$ is calculated from

$$\nabla E(\mathbf{D}) = 0. \quad (7)$$

The solution of Eq. (7) is efficiently obtained by solving a linear problem in the form of $Ax = b$. The explanations of the three terms follow with details.

Data term In Eq. (6), the data term indicates the initial sparse points obtained from Sec. 3.4, which is designed as

$$E_d(\mathbf{D}) = \sum_j \Big(D_j - Z_j \Big)^2, \quad (8)$$

where D_j is the targeted depth of the pixel j where the initial sparse depth Z_j is computed from Sec. 3.

Color smoothness term The color smoothness term is defined as

$$E_c(\mathbf{D}) = \sum_p \sum_{q \in W_p} \left(D_p - \frac{w^c_{pq}}{\sum_q w^c_{pq}} D_q \right)^2, \quad (9)$$

where p is a pixel on the reference image and q is the pixel in the 3×3 window W_p centered at p. The weight term w^c_{pq}

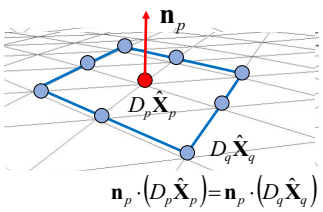


Figure 3. Geometric guidance term.

is the color affinity which is defined as follows:

$$w_{pq}^{c} = \exp\left(\sum_{\mathbf{I}\in lab} -\frac{|\mathbf{I}_p - \mathbf{I}_q|}{2\max(\sigma_p^2, \epsilon))}\right), \quad (10)$$

$$where\ \sigma_p^2 = \sum_{q\in W_p} (\mathbf{I}_p^2 - \mathbf{I}_q^2), \quad (11)$$

where $\mathbf{I}$ is the color intensity vector of the reference image in *lab* color space and ϵ is a maximum bound. This color similarity constraint was presented in [31] and is based on the assumption that each object consists of consistent color variation in the scene. Although it demonstrates reliable propagation results, it could not cover the continuous depth changes on the slanted surface with sparse control points while many real-world scenes have slanted objects with complex color variations.

Geometric guidance term To overcome the limitations of using only the color smoothness term, we include a geometric guidance term, which provides a geometrical constraint between adjacent pixels to have similar surface normals. Assuming that the depth of the scene is piecewise smooth, we define the geometric constraint $E_g(\mathbf{D})$ using the pre-calculated normal map described in Sec. 4.2 :

$$E_g(\mathbf{D}) = \sum_{p}\sum_{q\in W_p} w_p^g \left(D_p - \frac{\mathbf{n}_p \cdot \hat{\mathbf{X}}_q}{\mathbf{n}_p \cdot \hat{\mathbf{X}}_p} D_q\right)^2, \quad (12)$$

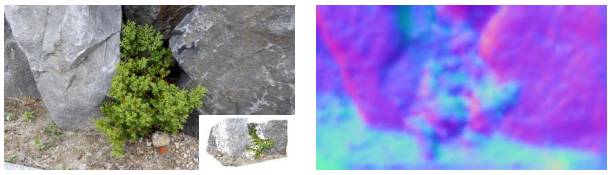
where $\mathbf{n}_p = [n_p^x, n_p^y, n_p^z]^\intercal$ is the normal vector of p and $\hat{\mathbf{X}}_p$ is the normalized image coordinate of p. w_p^g is a weight of the consistency of the normal directions between neighboring pixels.

$$w_p^g = \frac{1}{N_g}\sum_{q\in W_p} \exp\left(\frac{-(1-\mathbf{n}_p\cdot\mathbf{n}_q)}{\gamma_g}\right), \quad (13)$$

where γ_g is a parameter which determines the steepness of the exponential function, and N_g is the number of neighboring pixels in the window W_p. If the normal vectors of neighboring pixels are barely correlated with the normal vector of the center pixel, then the optimized depth $\mathbf{D}$ is less affected by the geometric guidance term.

4.2. Normal map Estimation

To incorporate the geometric guidance term in the objective function Eq. (6), a pixel-wise normal map should be previously estimated as shown in Fig. 4. First, we determine the normal vector for each sparse 3D point using local plane fitting. The sparse normal vectors are used for the data term of the normal propagation, and each normal component in xyz is propagated by the color smoothness term in Eq. (9). Since we observe that the normal vectors of adjacent pixels with high color affinity tend to be similar [32], the color-based propagation produces reliable dense normal map.



(a) Reference image & 3D points (b) Normal map

Figure 4. Normal map estimation - *Plant*.

5. Experimental Results

Our method is evaluated under three different perspectives. First of all, we demonstrate the effectiveness of each module of our framework by quantitative and qualitative evaluation in Sec. 5.2. Second, we compare our 3D reconstruction results with those obtained from the conventional state-of-the-art method [33] in Sec. 5.3. For a fair comparison, we use author-provided datasets[1] taken with a Google Nexus. Finally, our results are compared with the depth maps from the Microsoft Kinect2 which is valid for being used as ground truth [24].

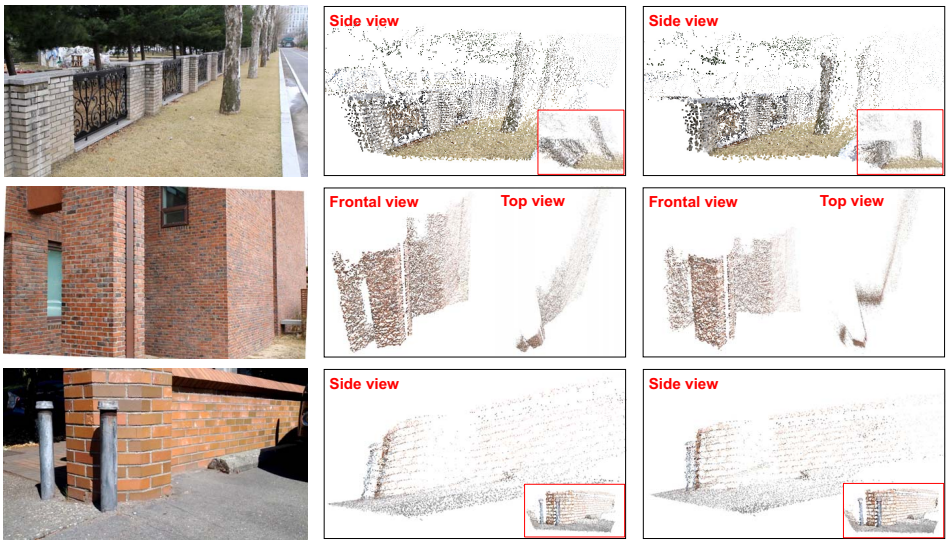
5.1. Experiment Environment

We capture various indoor and outdoor scenes with a Canon EOS 60D camera using the video capturing mode. We obtain 100 frames for 3 seconds. While capturing each image sequence, the camera is barely moved with only inadvertent motion by the photographer.

The proposed algorithm required ten minutes for 10000 points over 100 images in MATLAB™. Among all computation steps, the feature extraction is the most time-consuming. However, we expect that parallelized computing using GPU makes the overall process more efficient. A machine equipped with an Intel i7 3.40GHz CPU and 16GB RAM was used for computation.

We set the parameters as follows: the steepness of geometric guidance weight γ_g is fixed as 0.001 and the maximum bound ϵ as 0.001. The pre-calculated ratio of the readout time a is set as 0.5, 0.7 and 0.3, respectively, for the Canon EOS 60D, Google Nexus and Kinect2 RGB. The resolution of all the images among the dataset is 1920 $\times$ 1080.

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



(a) Reference images (b) SfSM without RS handling (c) SfSM with RS handling

Figure 5. SfSM result with/without RS handling - *Grass* (Top) & *Building1* (Middle) & *Wall* (Bottom).

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

Figure 6. Dense reconstruction result with/without geometric guidance - *Building2*. (a) Reference image. (b) Estimated normal map. (c) 3D mesh and depth without geometric guidance. (d) 3D mesh and depth map with geometric guidance.

5.2. Evaluation of the proposed method

In this subsection, we show the effectiveness of our RS handling method. We set $a = 0.5$ for the RS-handled case and $a = 0$ for the RS-unhandled case and compare the results. The qualitative and quantitative results are shown in Fig. 5 and Fig. 7, respectively. In Figure 5, we observe that the RS effect is removed, so that perpendicular planes are not distorted and are geometrically correct. Fig. 7 reports the average reprojection errors between the two cases. For all datasets, our RS handling method significantly reduces reprojection errors.

To verify the usefulness of our geometric guidance term in Sec. 4, we compare the results with and without the term as shown in Fig. 6. The result using only the color smoothness term causes severe artifacts on the slanted plane with multiple colors due to the lack of geometric information for an unknown depth. On the other hand, the geometric guidance term assists in preserving the slanted structures.

5.3. Comparison to state-of-the-art

To qualitatively evaluate our method, we first compare it with the state-of-the-art method [33]. In Figure 8, results

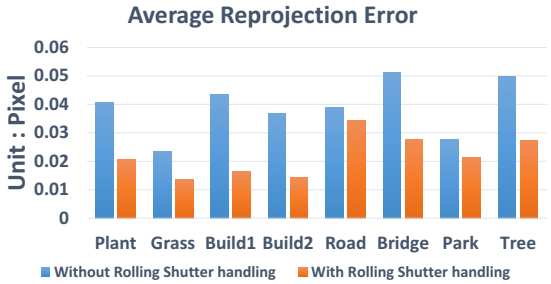
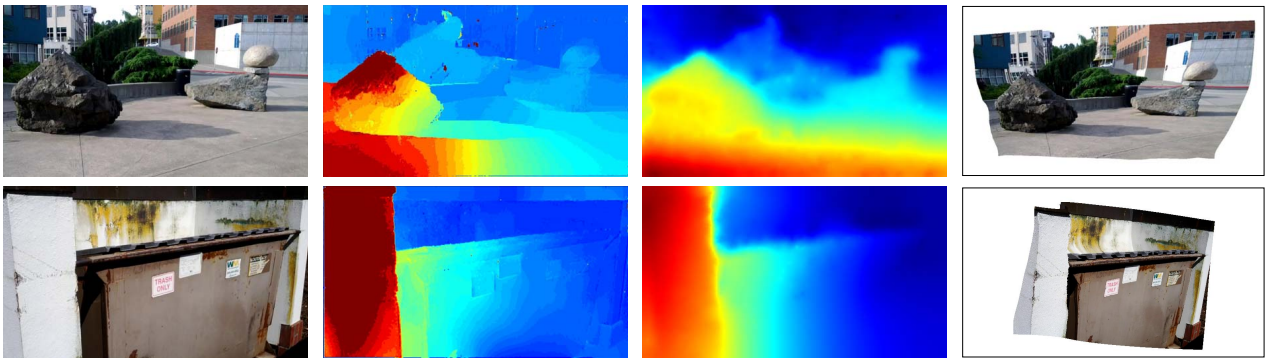


Figure 7. Average reprojection error for 8 datasets without RS handling, and with RS handling (Unit : pixel).

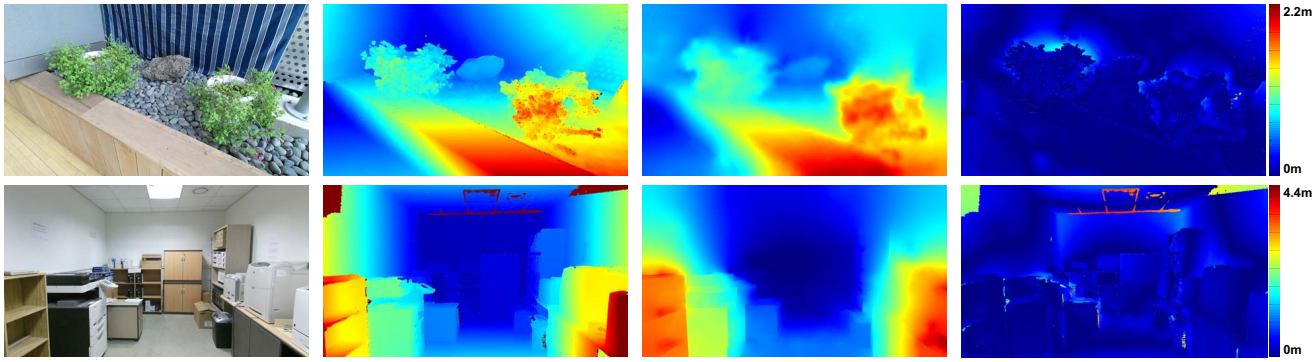
from [33] show distorted variations on depth maps. This is due to their disregard of the RS effect and plane-sweeping algorithm [5] for dense reconstruction whose data term is too noisy for narrow-baseline images. On the other hand, our bundle adjustment and propagation produce accurate depth maps which are continuously varying and geometrically correct.

For quantitative evaluation, we also compare our method with Kinect fusion [14] in a metric scale. The Kinect depth is aligned from the mesh using ray tracing with the known extrinsic matrix of the Kinect RGB to the depth sensor.



(a) Reference images [33] (b) Depth maps [33] (c) Our depth maps (d) Our 3D meshes

Figure 8. Result comparison with [33] - *Stone* (Top), *Trash* (Bottom).



(a) Reference images (b) Kinect depth maps (c) Our depth maps (d) Depth difference maps

Figure 9. Result comparison with the depth from Kinect fusion - *Pot* (Top), *Room* (Bottom)

Table 1. The percentages of depth error from Kinect2.

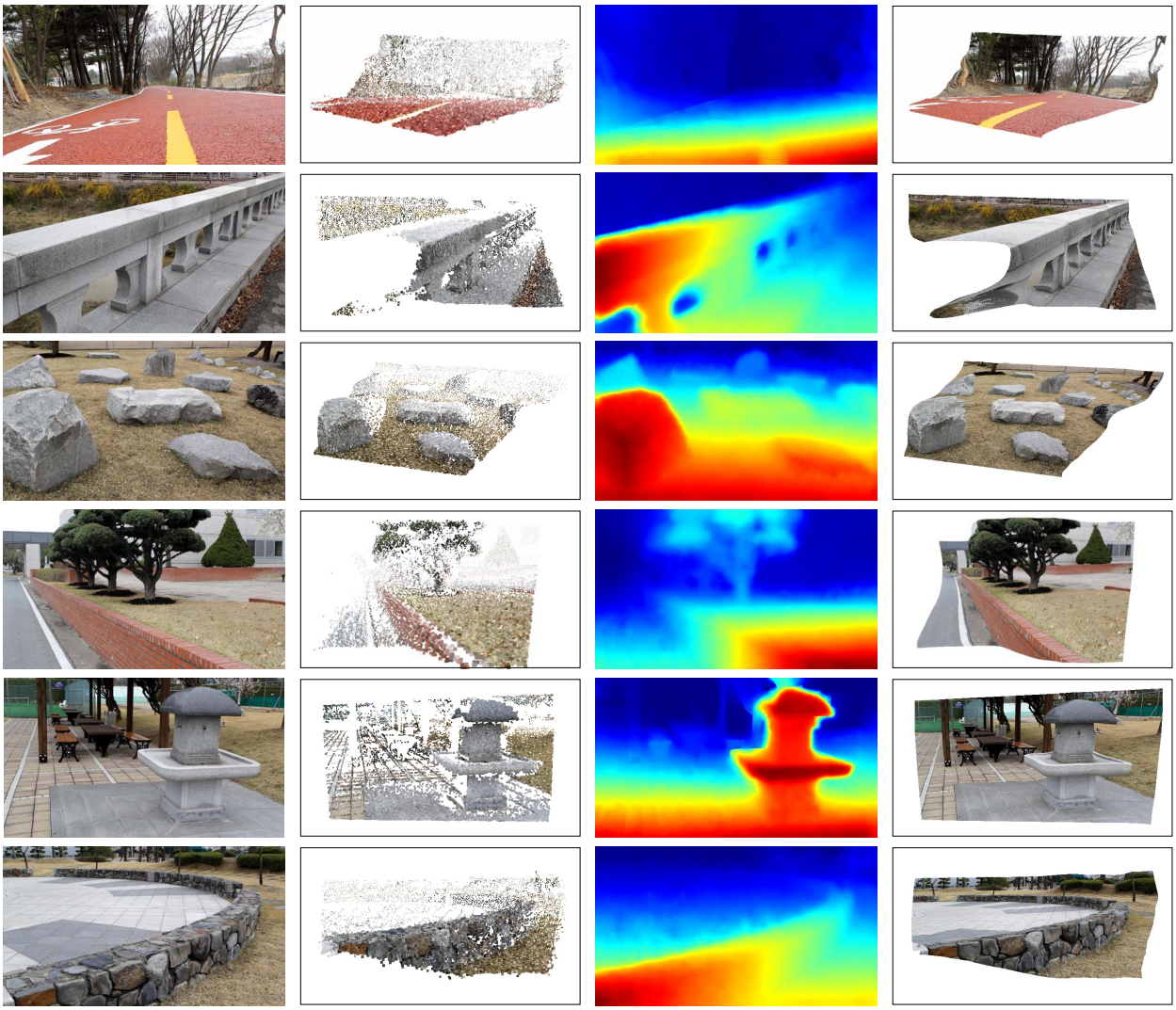
Dataset	*Max. depth*	R10	R20
Pot	2.2*m*	94.14%	99.07%
Room	4.4*m*	85.50%	96.31%

Due to the scale ambiguity of our results, the scale of each depth map is adjusted to the scale of the depth map from the Kinect using the average depth value. As shown in Fig. 9, the scale-adjusted depth maps from our method are similar enough to the depth maps from the Kinect fusion. For more detailed analysis, we utilize a robustness measure frequently used in a Middlebury stereo evaluation system [25]. Specifically, R10 and R20 respectively denote the percentage of pixels that have a distance error of less than 10% and 20% of the maximum depth value in the scene. Except for the occluded regions, only 5.86% pixels of the *Pot* dataset have over 22*cm* error compared to the depth from Kinect2, which is a reasonable error.

5.4. Applications

One of the emerging applications in the computer vision field is digital refocusing, changing a point or level of focus after taking a photo [1, 30, 15, 33, 13, 2]. With a depth map, we can add a synthetic blur by applying different amounts of blur depending on the pixels' depth as shown in Fig. 10. For realistic digital refocusing, an accurate depth map is necessary. To show a noticeable improvement of the application, we synthetically render defocused images based on depth maps from the proposed method and [33]. As shown in Fig. 1, since the depth map from [33] is geometrically inaccurate, the defocused image rendered using the depth map does not look natural. On the other hand, the result from our algorithm shows a distinctively more realistic refocusing effect.

Figure 10. Refocusing based on our depth maps in Fig. 11



(a) Reference images (b) SfSM results (c) Depth maps (d) Our 3D meshes

Figure 11. Our final results - *Road* (First) & *Bridge* (Second) & *Rocks* (Third) & *Tree* (Fourth) & *Faucet* (Fifth) & *Park* (Sixth).

6. Discussion

Conclusion This paper has presented an accurate and dense 3D reconstruction method using only a narrow-baseline image sequence captured from small motion. Three major contributions have been introduced: efficient outlier feature removal, a novel SfSM method with RS bundle adjustment and a dense reconstruction algorithm with geometric guidance constraint. By virtue of our RS handling procedure, the proposed method is very practical and generic since it can be applied to both global shutter and rolling shutter cameras. Furthermore, to overcome the limitation of point-based sparse reconstruction, an accurate depth propagation method has been designed. Finally, a large variety set of experiments have been conducted, highlighting the high-quality depth maps and 3D meshes obtained with our method. These results have been compared against the existing methods with different criterion, bringing to light the strong improvements offered by our approach.

Limitation & Future work We have proposed a practical system that has the potential to benefit a number of vision applications. Because we have focused on the high-quality depth only from small motion, the performance of the proposed method is not guaranteed for datasets with large rotations. As an important part of future work, we plan to accurately reconstruct 3D regardless of the amount of rotations. In addition, while our depth propagation shows reliable dense reconstruction results, there is still room for mitigating the over-smoothing effect. Moreover, our depth result is not represented in the metric scale since the estimated camera poses are up to scale. For metric scale estimation, usage of a camera equipped with a gyro-sensor, such as smartphone may be a good solution.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) grant funded by Korea government (MSIP) (No.2010- 0028680), and partially supported by the Study on Imaging Systems for the next generation cameras funded by the Samsung Electronics Co., Ltd (DMC RD center) (IO130806-00717-02). Hae-Gon Jeon was partially supported by Global PH.D Fellowship Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (NRF-2015H1A2A1034617).

References

[1] The lytro camera. http://www.lytro.com/. 1, 7

[2] J. T. Barron, A. Adams, Y. Shih, and C. Hernández. Fast bilateral-space stereo for synthetic defocus. In *Proc. of Comp. Vis. and Pattern Rec. (CVPR)*, 2015. 1, 7

[3] M. Bleyer, C. Rhemann, and C. Rother. Patchmatch stereo-stereo matching with slanted support windows. In *Proc. of British Machine Vision Conference (BMVC)*, 2011. 2

[4] M. L. Boas. *Mathematical Methods in the Physical*. John Wiley & Sons., Inc, 2006. 3

[5] R. T. Collins. A space-sweep approach to true multi-image matching. In *Proc. of Comp. Vis. and Pattern Rec. (CVPR)*, 1996. 6

[6] J. Diebel and S. Thrun. An application of markov random fields to range sensing. In *Advances in Neural Information Processing Systems (NIPS)*, 2005. 2

[7] P.-E. Forssén and E. Ringaby. Rectifying rolling shutter video from hand-held devices. In *Proc. of Comp. Vis. and Pattern Rec. (CVPR)*, 2010. 2, 3

[8] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 2010. 1

[9] D. Gallup, J.-M. Frahm, P. Mordohai, and M. Pollefeys. Variable baseline/resolution stereo. In *Proc. of Comp. Vis. and Pattern Rec. (CVPR)*, 2008. 2

[10] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey vision conference*, 1988. 3

[11] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 2, 3, 4

[12] J. Hedborg, P.-E. Forssén, M. Felsberg, and E. Ringaby. Rolling shutter bundle adjustment. In *Proc. of Comp. Vis. and Pattern Rec. (CVPR)*, 2012. 2, 3

[13] C. Hernández. Lens blur in the new google camera app. http://googleresearch.blogspot.kr/2014/04/lens-blur-in-new-google-camera-app.html. 1, 7

[14] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, and A. Fitzgibbon. Kinectfusion: Real-time 3d reconstruction and interaction using a moving depth camera. In *Proc. of the 24th Annual ACM Symposium on User Interface Software and Technology*, 2011. 2, 6

[15] N. Joshi and C. L. Zitnick. Micro-baseline stereo. *Microsoft Research Technical Report MSR-TR-2014-73*, 2014. 1, 2, 7

[16] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross. Scene reconstruction from high spatio-angular resolution light fields. In *Proc. of SIGGRAPH*, 2013. 2

[17] L. Magerand, A. Bartoli, O. Ait-Aider, and D. Pizarro. Global optimization of object pose and motion from a single rolling shutter image with automatic 2d-3d matching. In *Proc. of European Conf. on Comp. Vis. (ECCV)*, 2012. 2

[18] M. Meingast, C. Geyer, and S. Sastry. Geometric models of rolling-shutter cameras. *arXiv preprint cs/0503076*, 2005. 3

[19] J. J. Moré. The levenberg-marquardt algorithm: implementation and theory. In *Numerical analysis*. 1978. 4

[20] G. L. K. Morgan, J. G. Liu, and H. Yan. Precise subpixel disparity measurement from very narrow baseline stereo. *IEEE Transactions on Geoscience and Remote Sensing*, 2010. 2

[21] D. Nistér. An efficient solution to the five-point relative pose problem. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 2004. 3

[22] L. Oth, P. Furgale, L. Kneip, and R. Siegwart. Rolling shutter camera calibration. In *Proc. of Comp. Vis. and Pattern Rec. (CVPR)*, 2013. 2, 3

[23] J. Park, H. Kim, Y.-W. Tai, M. S. Brown, and I. Kweon. High quality depth map upsampling for 3d-tof cameras. In *Proc. of Int'l Conf. on Comp. Vis. (ICCV)*, 2011. 2

[24] M. Reynolds, J. Doboš, L. Peel, T. Weyrich, and G. J. Brostow. Capturing time-of-flight data with confidence. In *Proc. of Comp. Vis. and Pattern Rec. (CVPR)*, 2011. 5

[25] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int'l Journal of Computer Vision (IJCV)*, 2002. 7

[26] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proc. of Comp. Vis. and Pattern Rec. (CVPR)*, 2006. 2

[27] K. Shoemake. Animating rotation with quaternion curves. In *Proc. of SIGGRAPH*, 1985. 2, 3

[28] C. Tomasi and T. Kanade. *Detection and tracking of point features*. School of Computer Science, Carnegie Mellon Univ. Pittsburgh, 1991. 3

[29] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon. Bundle adjustmenta modern synthesis. In *Vision algorithms: theory and practice*. Springer, 2000. 4

[30] K. Venkataraman, D. Lelescu, J. Duparré, A. McMahon, G. Molina, P. Chatterjee, R. Mullis, and S. Nayar. Picam: An ultra-thin high performance monolithic camera array. *ACM Trans. on Graph.*, 2013. 1, 7

[31] L. Wang and R. Yang. Global stereo matching leveraged by sparse ground control points. In *Proc. of Comp. Vis. and Pattern Rec. (CVPR)*, 2011. 2, 5

[32] R. J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical engineering*, 19(1):191139–191139, 1980. 5

[33] F. Yu and D. Gallup. 3d reconstruction from accidental motion. In *Proc. of Comp. Vis. and Pattern Rec. (CVPR)*, 2014. 1, 2, 3, 4, 5, 6, 7

[34] C. Zhou, A. Troccoli, and K. Pulli. Robust stereo with flash and no-flash image pairs. In *Proc. of Comp. Vis. and Pattern Rec. (CVPR)*, 2012. 1