High-quality Depth from Uncalibrated Small Motion Clip

Hyowon Ha[†] Sunghoon Im[†] Jaesik Park[‡] Hae-Gon Jeon[†] In So Kweon[†]
[†]Korea Advanced Institute of Science and Technology [‡]Intel Labs

Abstract

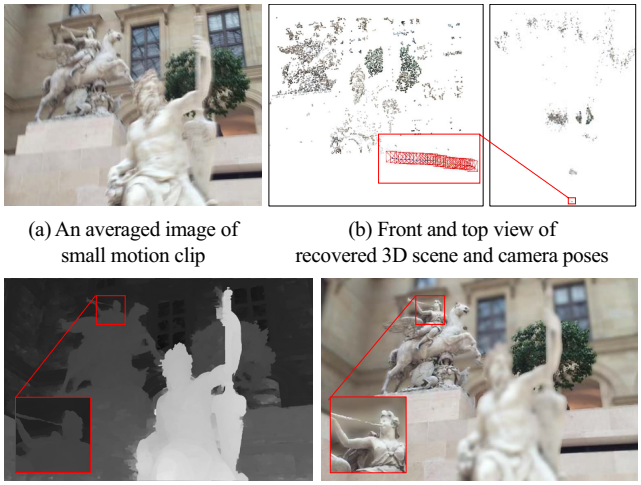
We propose a novel approach that generates a high-quality depth map from a set of images captured with a small viewpoint variation, namely small motion clip. As opposed to prior methods that recover scene geometry and camera motions using pre-calibrated cameras, we introduce a self-calibrating bundle adjustment tailored for small motion. This allows our dense stereo algorithm to produce a high-quality depth map for the user without the need for camera calibration. In the dense matching, the distributions of intensity profiles are analyzed to leverage the benefit of having negligible intensity changes within the scene due to the minuscule variation in viewpoint. The depth maps obtained by the proposed framework show accurate and extremely fine structures that are unmatched by previous literature under the same small motion configuration.

1. Introduction

Small motion in a hand-held camera commonly happens when a user moves the device slightly to find a better photographic composition, or even when the user tries to hold the camera steady before pressing the shutter. If we were able to restore the geometry of the scene using the small motion clip captured at that moment, it could be useful for a variety of applications, such as synthetic refocusing or view synthesis. Figure 1 shows an example of the small motion clip. The averaged image of the entire sequence gives a sense of how small the camera motion is.

In this paper, we propose an effective pipeline for depth acquisition from a small motion clip. At the core of our approach is the novel bundle adjustment scheme that is specially devised to be applied to the small motion case. Unlike to prior approaches, our algorithm can jointly estimate the intrinsic parameters and poses of the camera from a small motion footage, which imbues the proposed method with practicality and severs the need for camera calibration.

By virtue of reliably estimating the intrinsic and extrinsic camera parameters, a plane sweeping based dense stereo matching algorithm can be directly applied to produce a dense depth map in a unified framework. A notable benefit



(a) An averaged image of small motion clip

(b) Front and top view of recovered 3D scene and camera poses

(c) Depth map

(d) Synthetic refocusing

Figure 1. A small motion clip is our sole input. (a) An averaged image showing the overall camera motion. (b) Visualization of recovered 3D scene points and camera poses. (c) Our depth map result. (d) A synthetic refocusing result as an application example. Please note that our method does not require camera calibration.

to the small motion clip is that the observed image intensities for a given point in the scene are almost identical along the sequence due to the low variation in viewpoint. The dozens of intensity observations also give a better chance in finding reliable matchings. In order to leverage this benefit, our dense stereo matching algorithm utilizes the variance of the intensity profile as the cost measure for plane sweeping, which is less likely to be affected by the noise of the reference image compared to other pair-wise intensity difference based methods.

As opposed to previous approaches [9, 10, 27] that target the same goal for small motion, the distinctive points of our approach can be summarized as follows:

- A unified framework for depth from an uncalibrated small motion clip is proposed, which can allow the user to acquire a high-quality depth map from a single instance of capture.
- Our bundle adjustment can even jointly estimate the camera intrinsic parameters (*i.e.* focal length and radial

1063-6919/16 $31.00 © 2016 IEEE
DOI 10.1109/CVPR.2016.584

IEEE computer society

distortion) as well as the camera poses and the scene geometry from a single small motion clip.

- Our dense stereo matching that analyzes the intensity and gradient profiles in plane sweeping can generate depth maps exhibiting extremely fine structures, which have not been demonstrated in previous literature under the same small motion conditions.

2. Related Work

3D reconstruction from a hand-held camera is a widely studied topic. SfM successfully recovers the sparse 3D geometry and camera poses for wide baseline images [7, 22]. The bundle adjustment [25, 4] minimizes the reprojection errors using an optimization framework. Other approaches [11, 21, 12] use the L_∞ norm instead of the L_2 norm to make the cost function convex, but they are more susceptible to outliers. As opposed to SfM, multi-view stereo (MVS) can provide a depth for each pixel via dense matching of the images [16]. Gallup *et al*. [5] present an effective image matching method that selects a proper baseline and image resolution adapted for the scene depth.

Conventional SfM and MVS approaches can reconstruct accurate 3D geometry using wide-baseline images, but users often cannot capture such images. Yu and Gallup [27] propose an inspirational method that can estimate camera trajectory even from a clip with hand-shaking motion. They recover a dense depth map from a random depth initialization and perform a plane sweeping [3] based image matching that incorporates a Markov Random Field [13]. Although it is a well-known fact that narrow baselines affects the accuracy of the estimated 3D geometry [17, 18, 19], the inverse depth representation [27] is successfully demonstrated in challenging small motion scenarios. Im *et al*. [9] extends [27] with the consideration of rolling the shutter effect. Instead of performing dense image matching, they propagates the tracked 3D points into the canonical image domain. As the propagation is regularized by smooth surface normal map obtained from sparse depth points, the resulting depth map is also smooth. Joshi and Zitnick [10] adopts a homography based image warping for dense image matching with the micro baseline assumption. Their algorithm even targets the tremble of a camera mounted on a tripod.

The proposed approach also targets small motion clips obtained by monocular cameras. To the best of our knowledge, we are the first to demonstrate that even the camera intrinsic and lens parameters can be reasonably estimated from a small motion clip. This allows us to introduce a fully automatic pipeline that performs a self-calibration of the camera and estimation of a high-quality depth map. As opposed to [27] that computes a pair-wise consistency between the images, our dense stereo measures the consistency of the observed intensities by looking at the intensity distributions. Our high fidelity depth map is more reliable that that of [9, 10, 27].

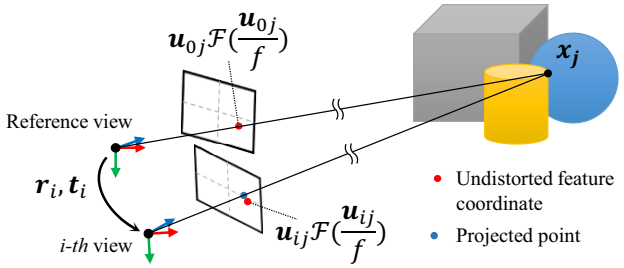


Figure 2. Small motion geometry used in our bundle adjustment for an uncalibrated camera (Sec. 3.1). We adopt the distorted-to-undistorted mapping function $\mathcal{F}$ to utilize the inverse depth representation in an analytic form.

3. Our Approach

We introduce the two consecutive stages of the proposed framework. The small motion image sequence is first processed in the bundle adjustment to estimate the camera parameters, then the undistorted images and the acquired parameters are utilized in the dense stereo matching.

3.1. Bundle Adjustment

The key aspect to the image sequences in question is that the baseline between the small motion images is significantly smaller than that of the conventional SfM problem. This makes the feature matching easier, but also results in a much higher depth uncertainty that causes conventional SfM approaches to fail.

In order to handle this challenging problem, Yu and Gallup [27] introduce two practical clues: (1) the small angle approximation of the camera rotation matrix and (2) the inverse depth based 3D scene point parameterization. It is shown that the former reduces the complexity of the cost function, and the latter helps to regularize the scales of the variables in the bundle adjustment. This idea is validated well in challenging real-world datasets, but they assume that the calibrated focal length is known a priori and do not account for the effects of the lens distortion.

Based on the two aforementioned insights, we propose a novel bundle adjustment framework that is carefully designed to estimate the focal length and radial distortion parameters in addition to the 3D scene points and camera poses. Compared to a prior work targeting the same objective for a wide baseline 3D reconstruction [23], our approach is tailored for the inverse depth representation and is more effective on small motion clips.

Conventional SfM algorithms map the projected 3D points from the Undistorted image domain coordinates to the Distorted image domain coordinates when measuring

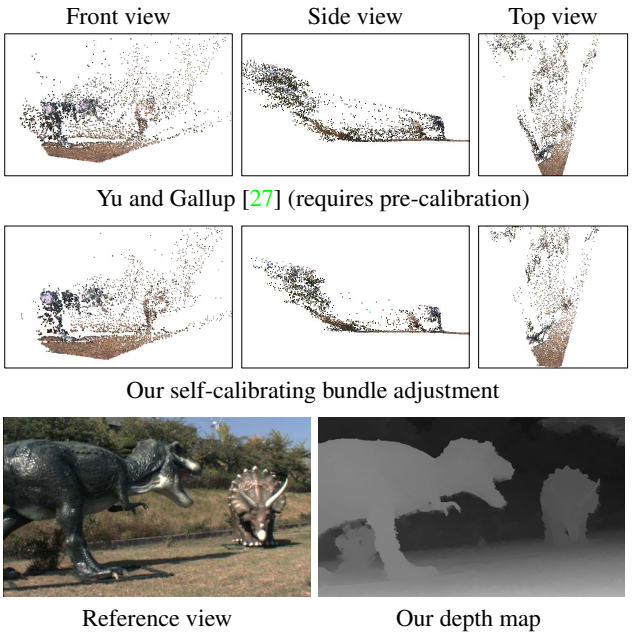


Figure 3. Comparing reconstructed 3D point clouds. Our approach recovers reliable 3D point clouds like the result of Yu and Gallup [27], but our approach is capable of camera self-calibration. We use calibrated camera parameters to apply [27].

the reprojection error (we refer this approach as to the U-D model). When it comes to the inverse depth representation, computing the reprojection error becomes rather complex: a point must be back-projected into the 3D space before it is projected onto the other image. However, describing the back-projected point in an analytic form is not straightforward when using the U-D model because it is difficult to get an exact analytic inverse function for the radial distortion model [15].

Our bundle adjustment scheme is built in a slightly different way to avoid losing its analytic form that is essential for non-linear optimization. We accomplish this by instead adopting the D-U radial distortion model that maps the point in the Distorted image domain into the Undistorted image domain [24]. While the U-D model measures the reprojection error in the distorted image domain, our D-U model *measures the error in the undistorted image domain*. This idea is fitting for the inverse depth representation because any feature point can be directly mapped onto the undistorted image domain for the back-projection or comparison with the reprojected point.

We follow a reasonable approximation of the camera model using one focal length f and two radial distortion parameters k_1, k_2, where the principal point and radial distortion center are assume to be equal to the image center, as done in [23]. The small motion geometry used in our bundle adjustment is depicted in Figure 2. If $\mathbf{u}_{ij}$ is the distorted coordinates for the j-th feature in the i-th image relative to the image center, its undistorted coordinates can be calculated as $\mathbf{u}_{ij}\mathcal{F}\left(\frac{\mathbf{u}_{ij}}{f}\right)$, where $\mathcal{F}$ is the D-U radial distortion function that is defined by:

$$\mathcal{F}\left(\cdot\right) = 1 + k_1 \left\|\cdot\right\|^2 + k_2 \left\|\cdot\right\|^4 . \tag{1}$$

If $i = 0$ for the reference image, the back-projection of the feature $\mathbf{u}_{0j}$ to its 3D coordinates $\mathbf{x}_j$ is parameterized using its inverse depth w_j by:

$$\mathbf{x}_j = \begin{bmatrix} \frac{\mathbf{u}_{0j}}{fw_j}\mathcal{F}\left(\frac{\mathbf{u}_{0j}}{f}\right) \\ \frac{1}{w_j} \end{bmatrix} . \tag{2}$$

We now introduce a projection function π to describe the projection of $\mathbf{x}_j$ onto the i-th image plane as

$$\pi\left(\mathbf{x}_j, \mathbf{r}_i, \mathbf{t}_i\right) = \left\langle \mathcal{R}\left(\mathbf{r}_i\right)\mathbf{x}_j + \mathbf{t}_i \right\rangle, \tag{3}$$

$$\mathcal{R}\left(\mathbf{r}_i\right) = \begin{bmatrix} 1 & -r_{i,3} & r_{i,2} \\ r_{i,3} & 1 & -r_{i,1} \\ -r_{i,2} & r_{i,1} & 1 \end{bmatrix}, \tag{4}$$

$$\left\langle [x, y, z]^\intercal \right\rangle = [x/z, y/z]^\intercal , \tag{5}$$

where $\mathbf{r}_i \in \mathbb{R}^3$ and $\mathbf{t}_i \in \mathbb{R}^3$ indicate the relative rotation and translation from the reference image to the i-th image, $\{r_{i,1}, r_{i,2}, r_{i,3}\}$ are the elements of $\mathbf{r}_i$, and $\mathcal{R}$ is the vector-to-matrix function that transforms the rotation vector $\mathbf{r}_i$ into the small-angle-approximated rotation matrix.

The undistorted image domain coordinates of the projected point is then calculated as $f\pi\left(\mathbf{x}_j, \mathbf{r}_i, \mathbf{t}_i\right)$. We use the distance between these coordinates and the undistorted coordinates as the reprojection error of $\mathbf{u}_{ij}$. Finally, our bundle adjustment is formulated to minimize the reprojection errors of all the features in the non-reference images by:

$$\underset{K,R,T,W}{\operatorname{argmin}} \sum_{i=1}^{n-1}\sum_{j=0}^{m-1} \rho\left(\mathbf{u}_{ij}\mathcal{F}\left(\frac{\mathbf{u}_{ij}}{f}\right) - f\pi\left(\mathbf{x}_j, \mathbf{r}_i, \mathbf{t}_i\right)\right), \tag{6}$$

where n is the number of images, m the number of features, $\rho\left(\cdot\right)$ the element-wise Huber loss function [8], K the set of the intrinsic camera parameters $\{f, k_1, k_2\}$, R and T the sets of the rotation and translation vectors for the non-reference images, and W the set of inverse depth values.

To obtain the feature correspondences, we first extract the local features using the Harris corner detector [6] in the reference image and find the corresponding feature locations in the other images by using the Kanade-Lukas-Tomashi (KLT) algorithm [14]. Each tracking is performed forwards and backwards to reject outlier features with bidirectional error greater than 0.1 pixel.

For the initial parameters of the bundle adjustment, we set the rotation and translation vectors to zero, which is mentioned to be reasonable for the small motion case [27]. The focal length is set to the larger value between the image width and height. The two radial distortion parameters

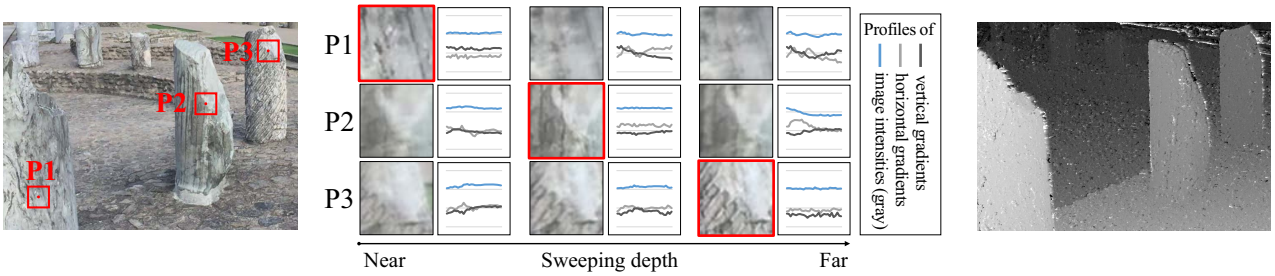


Figure 4. An example of plane sweeping stereo. The image sequences correspond to the small motion clip are warped according to the sweeping depth. (a) One of input images. Three local regions of different scene depth – $P1$, $P2$, and $P3$ are marked for the illustration. (b) The mean image of warped input images and its corresponding intensity/gradient profile are displayed. If the sweeping depth is correct for the local region, its profiles become flat. (c) Recovered depth map after applying winner-takes-all scheme on the computed cost volume.

are also set to zero. For the inverse depths, a random value between 0.01 and 1.0 is given to each feature.

The proposed bundle adjustment has several benefits: 1) It can successfully handle images captured by conventional cameras having mild lens distortion without any pre-calibration. 2) The use of the robust Huber loss function helps disregard the effects of outliers. Therefore, filtering through the use of random sample consensus (RANSAC) based two view relation estimation is not necessary. In practice, the two view approach can be unstable for small baselines [27].

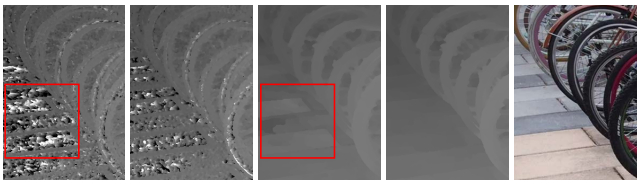
Although our formulation requires a higher order compared to that of [27] due to the addition of the intrinsic camera parameters, we find that our bundle adjustment successfully converges with a reasonable approximation for the parameters in most of our experiments (see Figure 3).

3.2. Dense Stereo Matching

Once we have obtained the intrinsic and extrinsic camera parameters from the previous stage, we can utilize these parameters in our plane sweeping based dense stereo matching algorithm to recover a dense depth map. The distortion in the input images are rectified using the estimated intrinsic parameters. The rectified images are then used in this step.

The original idea of the plane sweeping algorithm [3] is to back-project the tracked features onto an arbitrary virtual plane perpendicular to the z-axis of the canonical view point. If the back-projected points from all viewpoints are gathered in a small region of the virtual plane, we can conclude that the depth of the tracked feature is equivalent to that of the plane. Otherwise, this step repeats using other virtual planes. This simple but powerful idea is extended to dense stereo matching by warping the images onto the sweeping plane and measuring the photo consistency of each pixel of the warped images. The representative approach [2] computes the absolute intensity differences between the reference image and the other images for the consistency measure.

Inspired by this reliable framework, our approach takes



(a) WTA-SAD (b) WTA-VAR (c) Refined-SAD (d) Refined-VAR (e) Reference view

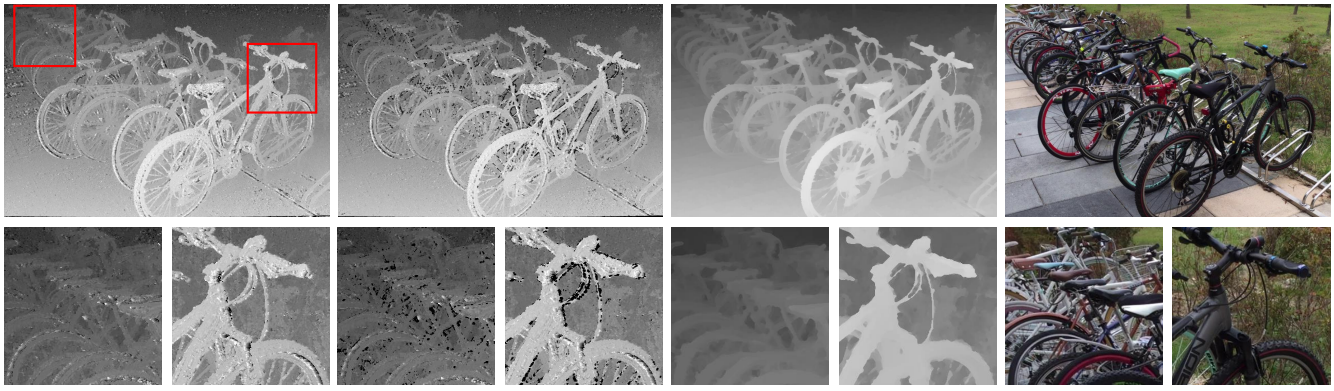
Figure 5. Comparison with SAD based cost and VAR based cost. After applying winner-takes-all (WTA) strategy, depth from SAD based cost (WTA-SAD) shows higher depth noise than that of VAR based cost (WTA-VAR). As a result, WTA-SAD gives incorrect depth when depth refinement algorithm [26] is applied.

into account the distribution of the intensities acquired from the pixels in the warped images that correspond to the same point on the virtual plane, which we collectively call the *intensity profile*. Consider the observed intensities in the profile acquired from the correct sweeping depth; it is reasonable to assume that the captured intensities are almost identical because the camera response function, white-balance, scene illumination, and observed scene radiance are unchanged with the kind of small viewpoint variation we are dealing with. Therefore, the profile will be uniform if the sweeping plane is at the correct depth for that pixel. Figure 4 shows an example supporting this idea.

Now we will introduce our dense stereo algorithm devised for small motion clips, step-by-step.

Building intensity profile. For the k-th depth in n_k sweeping depths, all the images are warped by back-projecting them onto a virtual plane at a given inverse-depth[1] w_k from the reference viewpoint, and then projected onto the reference image domain. The plane-induced homography $\mathbf{H}_{ik} \in \mathbb{R}^{3\times3}$ that describes the transformation from the reference image domain coordinates to the i-th image domain coordinates when passing through the virtual plane at the

[1]To prevent the abuse of the notation, we view the inverse depth $w_k = \frac{1}{z_k}$ as a practical analog to the depth candidate.



(a) Winner-takes-all (b) Outlier removal (c) After refinement (d) Reference View

Figure 6. The sequential procedures to get the dense depth map. (a) Rough depth map acquired by applying winner-takes-all strategy on the cost volume, (b) after the removal of unreliable depth pixels through our approach, and (c) after the depth refinement algorithm is applied. (d) The reference image. Note the fine structures observable in the bicycles.

k-th sweeping depth can be formulated by:

$$\mathbf{H}_{ik} = \mathbf{K} \begin{bmatrix} 1 & -r_{i,3} & r_{i,2} + w_k t_{i,1} \\ r_{i,3} & 1 & -r_{i,1} + w_k t_{i,2} \\ -r_{i,2} & r_{i,1} & 1 + w_k t_{i,3} \end{bmatrix} \mathbf{K}^{-1}, \quad (7)$$

where $\{t_{i,1}, t_{i,2}, t_{i,3}\}$ are the elements of $\mathbf{t}_i$ and $\mathbf{K}$ is defined as $\left[\begin{smallmatrix} f & 0 & p_u \\ 0 & f & p_v \\ 0 & 0 & 1 \end{smallmatrix}\right]$. (p_u, p_v) in $\mathbf{K}$ is the principal point coordinates equated to the image center. Using this homography, the i-th undistorted image I_i^u can be warped into the reference image domain through the operation described by the following formulation:

$$I_{ik}(\mathbf{u}) = I_i^u(\langle \mathbf{H}_{ik}\mathbf{u} \rangle), \quad (8)$$

where $\langle \cdot \rangle$ is the same function defined in Eq. (5) and I_{ik} is the warped i-th image according to the k-th sweeping depth. After warping n images, every pixel $\mathbf{u}$ in the reference image domain has an intensity profile $\mathcal{P}(\mathbf{u}, w_k) = [I_{0k}(\mathbf{u}), \cdots, I_{(n-1)k}(\mathbf{u})]$ for the inverse depth candidate w_k.

Compute matching cost volume. Based on earlier discussions, we measure the consistency of the intensity profile to evaluate the alignment of the warped images. Our matching cost $\mathcal{C}_I$ for pixel $\mathbf{u}$ and depth candidate w_k is defined as follows:

$$\mathcal{C}_I(\mathbf{u}, w_k) = \mathrm{VAR}\Big(\big[I_{0k}(\mathbf{u}), \cdots, I_{(n-1)k}(\mathbf{u}) \big] \Big), \quad (9)$$

where $\mathrm{VAR}(\mathbf{p})$ is the variance of vector $\mathbf{p}$. In order to enforce the matching fidelity on the edge regions of the image, we introduce two additional costs $\mathcal{C}_{\delta u}$ and $\mathcal{C}_{\delta v}$ defined as the horizontal and vertical gradients of the images, respectively. $\mathcal{C}_{\delta u}$ is defined as follows:

$$\mathcal{C}_{\delta u}(\mathbf{u}, w_k) = \mathrm{VAR}\left(\left[\frac{\delta I_{0k}}{\delta u}(\mathbf{u}), \cdots, \frac{\delta I_{(n-1)k}}{\delta u}(\mathbf{u}) \right] \right), \quad (10)$$

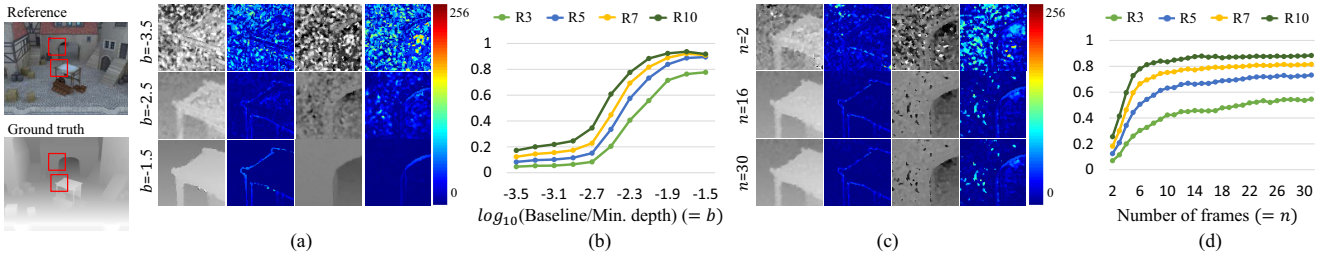
where $\frac{\delta I}{\delta u}$ indicates the image gradient in the horizontal direction. We use the first-order gradient filter $\mathbf{F} = \left[-1 \; 0 \; 1 \right]$ to approximate $\frac{\delta I}{\delta u}$ in the discrete and finite image space. $\mathcal{C}_{\delta v}$ is similarly calculated using the vertical gradients obtained by using $\mathbf{F}^\intercal$. The comprehensive matching cost $\mathcal{C}$ is defined as

$$\mathcal{C} = \mathcal{C}_I + \lambda(\mathcal{C}_{\delta u} + \mathcal{C}_{\delta v}). \quad (11)$$

Although the numerous intensity observations in the profile give a reliable matching cost, the variance operator used in Eq. (9) and (10) may not robustly compute the deviation of the profile in the presence of outliers. The application of a 3×3 box filter on $\mathcal{C}$ can suppress some of the noise in the costs. The proposed matching scheme is found to recover the fine structures in the depth map results, which will be shown later.

The proposed cost volume $\mathcal{C}$ is analogous to the conventional plane sweeping stereo [2], which computes the sum of absolute intensity difference (SAD) between the reference image and the other images. Here, the notable difference to our method is that the pairwise matching costs depend on the reference image; if the reference image includes a large amount of noise, the matching cost will be less reliable. By contrast, our variance operator handles every element equally. Figure 5 shows the depth map comparison between the SAD based cost and the VAR based cost after applying the winner-takes-all strategy and the successive refinement.

Depth refinement. After applying the winner-takes-all strategy on the cost volume $\mathcal{C}$, we get the depth map D_{win}. Although D_{win} gives a reasonable depth estimate, some values of D_{win} can be noisy when homogeneous textures are present in that region as may not give an obvious cost minimum. To handle noisy depth, we define a confidence measure described by the formulation $\mathcal{M}(\mathbf{u}) = 1 - \mathcal{C}_I(\mathbf{u}, D_{win}(\mathbf{u}))/\bar{\mathcal{P}}(\mathbf{u}, D_{win}(\mathbf{u}))$, where $\bar{\mathcal{P}}$ is the mean



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

Figure 7. Synthetic experiments with respect to the magnitude of the baseline ($n = 31$) and the number of images ($b = -2.1$). (a), (c) Depth maps obtained by the proposed WTA and their difference maps. (b), (d) Robustness measure results denoting the percentage of pixels within 3, 5, 7, and 10 label differences from the ground truth.

Algorithm 1: Depth from Small Motion Clip

Input: Image sequence $\{I_i\}_{i=0}^{n-1}$
Output: Depth map D_{out}

- Compute relative pose $\{\mathbf{r}_i, \mathbf{t}_i\}_{i=0}^{n-1}$, 3D points $\{\mathbf{x}\}$ of the scene, and lens parameters f, k_1, k_2 (Sec. 3.1)
- Undistort $\{I_i\}_{i=0}^{n-1}$ to get $\{I_i^u\}_{i=0}^{n-1}$
- n_k = number of label, z_{min} = nearest depth of $\mathbf{x}$
for $k = 1 : n_k$ **do**
  - Set $w_k = \frac{k}{n_k z_{min}}$
  for $i = 0 : (n-1)$ **do**
    - Warp I_i^u using w_k, $\mathbf{K}$, $\mathbf{R}_i$, and $\mathbf{t}_i$ (Eq. (7))
  - Build intensity/gradient profiles and compute cost volume $\mathcal{C}$ (Eq. (9, 10, and 11))
- Winner-takes-all on $\mathcal{C}$ and get D_{win}
- Refine D_{win} and get D_{out} (Sec. 3.2)

of the intensity profile $\mathcal{P}$ used for normalizing the confidence map scale. If $\mathcal{M}(\mathbf{u})$ is smaller than a constant threshold, we declare them as outlier pixels. D_{win} and $\mathcal{M}(\mathbf{u})$ are used for the depth refinement algorithm [26] that enforces smoothness on the depth image using the guidance of the reference color image. [26] is based on the minimum spanning tree structure, and it requires only 2 addition/subtraction operations and 3 multiplication operations in total for each pixel. Figure 6 shows an example of the outlier detection and the refinement. The overall pipeline of our approach is summarized in Algorithm 1.

4. Experimental Results

4.1. Synthetic Dataset

We devise a synthetic experiment to analyze the algorithm's dependence on the magnitude of the baseline motion and on the number of images. We render 11 clips each containing one reference image and 30 non-reference images all captured at a fixed distance around the z-axis of the reference image. The baselines are determined relative to the minimum depth of the scene from the reference image. All the images are rendered using the Blender™ software at a set resolution of 640×480. Since our algorithm determines the camera pose and scene depth up to a scaled factor, the obtained depth map and the ground truth has to be normalized into the same range, from 1 to 256 in our case, for an accurate assessment. Fig. 7 shows the depth map results obtained by the proposed WTA scheme and their corresponding difference maps by varying magnitudes of the baseline, shown in Fig. 7.(a), and the number of images used, Fig. 7.(c), respectively. Additionally, it shows the results of the quantitative evaluations, Fig. 7.(b) and Fig. 7.(d), using the robustness measure employed by the Middlebury stereo evaluation system [20]; $R5$ denotes the percentage of pixels that come within an absolute distance of 5 labels from the ground truth label. These two evaluations demonstrate that our method can produce a depth with an $R5$ score of at around 80% as long as the magnitude of the baseline is greater than 1% of the nearest scene depth and the number of frames captured exceeds 30 frames. This roughly equates to a 1 second small motion clip.

	R3	R5	R7	R10	MAD
SAD	43.331	64.940	78.337	86.419	8.650
VAR	**44.349**	**67.728**	**81.646**	**90.201**	**5.763**

Table 1. Quantitative comparison between the SAD based cost and VAR based cost using one of our synthetic clips ($b = -2.1$, $n = 31$). MAD denotes the mean of absolute depth label difference.

We also carried out a quantitative comparison by using either the SAD based cost or the proposed VAR based cost in the plane sweeping stage. In this experiment, the ground-truth camera parameters of our synthetic clip ($b = -2.1$, $n = 31$) are given to the depth map acquisition pipeline. As shown in Table 1, the WTA depth map using the SAD based cost gives worse results than that of the VAR based cost. The reason behind this result is that the SAD based cost is easily affected by quality of the reference image, which may contain noise, whereas the VAR based cost has no bias and considers the importance of all input images equally.

4.2. Real-world Dataset

Camera setup. We have tested our method using a machine vision camera, Flea3 from Point Grey, Inc. (1280×960 resolution at 30fps), and an iPhone 6 with two video modes

Figure 8. Depth maps acquired by our algorithm. For each dataset, the left image shows the averaged image of the footage to indicate the amount of camera movement, and the right image shows the estimated depth map using our algorithm. The clips are captured by a hand-held iPhone 6 device using the default video capturing mode. Note that the capture times of the displayed clips are only *one second*.

			Point Grey Flea3		iPhone6	
			#1 (11)	#2 (6)	#1 (7)	#2 (7)
Focal length (pixel/scale)	GT		1685.95	1685.02	1283.05	1293.73
	Initial		1280.00	1280.00	1000.00	1000.00
	Refined	Min	1656.44	1634.14	1253.45	1279.52
		Mean	1707.74	1684.74	1300.06	1301.56
		Max	1771.43	1748.79	1351.44	1332.44
Distortion error(pixel)	Initial		5.76	5.89	3.65	3.25
	Refined	Min	0.09	0.02	0.89	0.97
		Mean	0.52	0.53	1.48	1.14
		Max	1.26	1.20	2.17	1.49

Table 2. Evaluation on the estimated intrinsic camera parameters (*i.e.* focal length and radial distortion). We have tested 31 clips (identified by the number in the parenthesis) grabbed by two cameras with different lens settings. The ground truth camera parameters (GT) are acquired using the camera calibration toolbox [28].

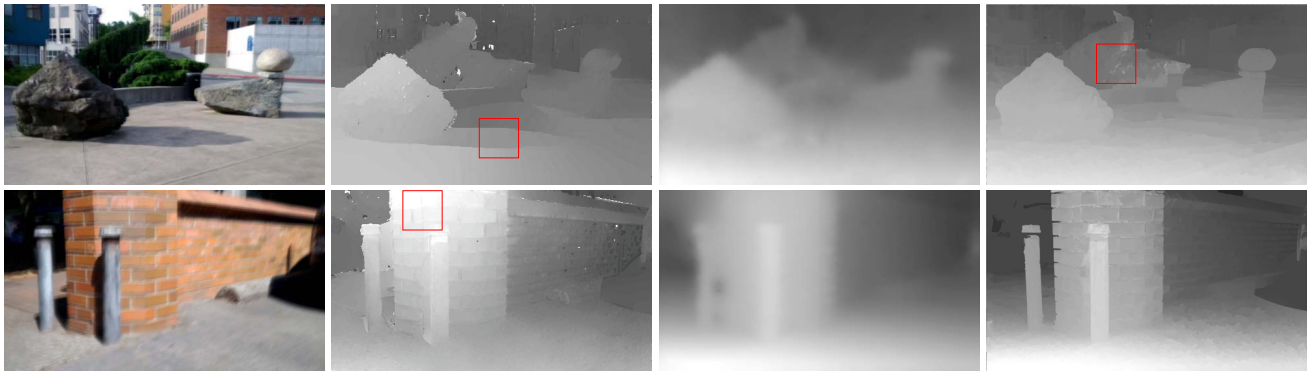
(1920 × 1280 at 30fps and 1280 × 720 at 240fps). For the 240 fps videos, 30 frames are uniformly sampled from the first 240 frames. As the proposed method can utilize the full resolution of the images, the estimated depth maps have the same resolution as the inputs. The datasets are captured by multiple users independent to this research. The small motion clips contain various types of motions, such as the one-directional or waving motion. The maximum distances for the captured scenes range from a wide variety of distances. Figure 8 shows high-quality depth maps from various types of small motions.

Computational time. In order to process each small motion footage (30 frames of 1280 × 720 res. images), our unoptimized implementation takes about one minute to perform feature extraction, tracking, and bundle adjustment. We use the Ceres solver for the sparse non-linear optimization [1]. The dense stereo matching stage takes about 10 minutes. The reported time is measured without CPU parallelization. We use a desktop computer equipped with an Intel i7-4970K 4.0Ghz CPU and 16GB RAM. The data and program are released on our project website.

Evaluation on the camera self-calibration. As the proposed bundle adjustment is designed to self-calibrate the intrinsic camera parameters, we devise a quantitative evaluation method for the camera parameters obtained by our approach. For this experiment, we use the calibrated Flea3 and iPhone6 cameras, each while on two different lens settings[2] . Table 2 compares the estimated focal length and radial distortion against the ground truth. Here, we intentionally set the initial focal length to be significantly different from the ground truth. For measuring the distortion error, we generate a pixel grid and transform their coordinates using the estimated D-U function $\mathcal{F}$. The transformed coordinates are again applied with the ground-truth U-D model found in the camera pre-calibration. If the estimated $\mathcal{F}$ is reliable, these sequential transformation should be identity. The distortion error is measured in pixels using the mean of absolute distances. The results show that the estimated parameters are close to the ground truth. Notably, the mean distortion error from the Flea3 datasets is around 0.5 pixels, while the initial parameter ($k_{1,2} = 0$) had an error of 5.8 pixels.

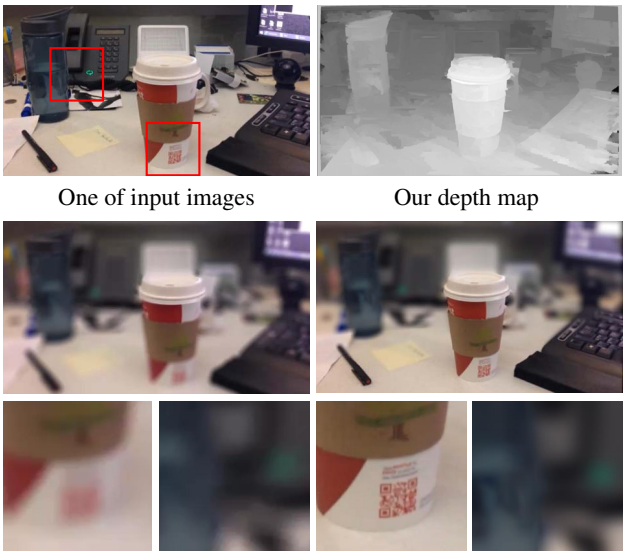
Comparison with [9, 27]. Figure 9 shows the depth maps acquired by Yu and Gallup [27], Im *et al.* [9], and our approach. We use the dataset and results provided by their

[2]iPhone can hold a fixed focal length if the user touches the region of interest in the preview screen for a long time.



Averaged image of the clips | Yu and Gallup [27] | Im *et al*. [9] | Our results

Figure 9. Comparison with state-of-the-art approaches. The amount of camera movement is quite small as indicated by the averaged images on the far left. The results of Yu and Gallup [27] show an inaccurate depth discontinuity on the shadow and indicate that the brick wall is nearer than the ground. Our result shows a similar depth tendency to [9] while exhibiting much sharper depth discontinuities. The erroneous region in our approach is also marked in red.



One of input images | Our depth map

Joshi and Zitnick [10] | Our synthetic refocusing

Figure 10. Comparison with Joshi and Zitnick [10] on the application of synthetic refocusing. Our high-quality depth map generates an admirable foreground focused image from the small motion clip. Note that our result shows realistic defocus blurs on the background whilst preserving sharp edges in the foreground (bottom region of the cup).

respective websites for the pair comparison. The results by [27] have inaccurate depth values as marked with the red rectangles. The results of [9] show better depth values, indicating that the ground plane is nearest to the camera. This is due to their explicit consideration of the rolling shutter effect. However, the depth discontinuity is too smooth, resulting in blurry object boundaries unsuitable for the use of synthetic refocusing. Our results have similar depth values to [9] but show much sharper edge boundaries. Note that our approach performs a self-calibration of the camera parameters for the displayed results, whereas [27, 9] utilize the factory settings for the camera focal length and do not account for the lens distortion.

Comparison with [10]. The synthetic refocusing image obtained by using our depth map is compared with that of [10]. We also use the dataset and result provided by [10] for the pair comparison. As shown in Fig. 10, our foreground focused image consistently gives realistic defocus blurs.

Limitations. Although our algorithm is specially designed for small baselines, the estimated camera poses become unreliable if the motion is unreasonably small. However, according to Sec. 4.1, the required minimum baseline to apply our approach is reasonable, and such failure cases rarely happen in real hand-held scenarios. We observe that the lack of features near the image border results in the erroneous estimation of the radial distortion parameters. Our approach does not explicitly take into account the effects of occlusion/dis-occlusion because the detected outliers in our depth post-processing stage reasonably correspond to such regions. However, any sophisticated occlusion inference algorithm could be applied here.

5. Conclusion

We have introduced a practical algorithm that recovers a high-quality depth map from a small motion clip recorded by commercial cameras. Our self-calibrating bundle adjustment estimates the camera parameters that are shown to be close to the ground truth, even if only one second of the clip is used. Our dense stereo matching step analyzes the statistics of intensity profile and shows a superior depth map than that of the previous approaches. In the future, we plan to study the convergence properties of the proposed bundle adjustment scheme.

Acknowledgement This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No.2010-0028680).

References

[1] S. Agarwal, K. Mierle, and Others. Ceres solver. http://ceres-solver.org.

[2] A. Akbarzadeh, J. m. Frahm, P. Mordohai, C. Engels, D. Gallup, P. Merrell, M. Phelps, S. Sinha, B. Talton, L. Wang, Q. Yang, H. Stewenius, R. Yang, G. Welch, H. Towles, D. Nistr, and M. Pollefeys. Towards urban 3d reconstruction from video. In *in 3DPVT*, pages 1–8, 2006.

[3] R. T. Collins. A space-sweep approach to true multi-image matching. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 358–363. IEEE, 1996.

[4] D. Crandall, A. Owens, N. Snavely, and D. Huttenlocher. Discrete-continuous optimization for large-scale structure from motion. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3001–3008. IEEE, 2011.

[5] D. Gallup, J.-M. Frahm, P. Mordohai, and M. Pollefeys. Variable baseline/resolution stereo. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008.

[6] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey vision conference*, volume 15, page 50, 1988.

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

[8] P. J. Huber. Robust estimation of a location parameter. *Annals of Statistics*, 53(1):73–101, 1964.

[9] S. Im, H. Ha, G. Choe, H.-G. Jeon, K. Joo, and I. S. Kweon. High quality structure from small motion for rolling shutter cameras. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[10] N. Joshi and C. L. Zitnick. Micro-baseline stereo. Technical report, Technical Report MSR-TR-2014-73, Microsoft Research, 2014.

[11] F. Kahl. Multiple view geometry and the l-norm. In *Proceedings of International Conference on Computer Vision (ICCV)*, volume 2, pages 1002–1009. IEEE, 2005.

[12] F. Kahl, S. Agarwal, M. K. Chandraker, D. Kriegman, and S. Belongie. Practical global optimization for multiview geometry. *International Journal on Computer Vision (IJCV)*, 79(3):271–284, 2008.

[13] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *arXiv preprint arXiv:1210.5644*, 2012.

[14] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. *IJCAI*, 81:674–679, 1981.

[15] L. Ma, Y. Chen, and K. L. Moore. Rational radial distortion models of camera lenses with analytical solution for distortion correction. *International Journal of Information Acquisition*, 1(12):135–147, 2004.

[16] M. Okutomi and T. Kanade. A multiple-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 15(4):353–363, 1993.

[17] J. Oliensis. Computing the camera heading from multiple frames. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 203. IEEE, 1998.

[18] J. Oliensis. A multi-frame structure-from-motion algorithm under perspective projection. *International Journal on Computer Vision (IJCV)*, 34(2-3):163–192, 1999.

[19] J. Oliensis. The least-squares error for structure from infinitesimal motion. *International Journal on Computer Vision (IJCV)*, 61(3):259–299, 2005.

[20] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal on Computer Vision (IJCV)*, 47(1-3):7–42, 2002.

[21] K. Sim and R. Hartley. Recovering camera motion using linfty minimization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 1230–1237. IEEE, 2006.

[22] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: Exploring photo collections in 3d. In *Proceedings of ACM SIGGRAPH*, pages 835–846, New York, NY, USA, 2006. ACM Press.

[23] N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the world from Internet photo collections. *International Journal of Computer Vision*, 80(2):189–210, November 2008.

[24] T. Tamaki, T. Yamamura, and N. Ohnishi. Unified approach to image distortion. In *Pattern Recognition (ICPR), 2002. Proceedings. 16th International Conference on*, pages 584–587. IEEE, 2002.

[25] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon. Bundle adjustmenta modern synthesis. In *Vision algorithms: theory and practice*, pages 298–372. Springer, 2000.

[26] Q. Yang. A non-local cost aggregation method for stereo matching. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1402–1409. IEEE, 2012.

[27] F. Yu and D. Gallup. 3d reconstruction from accidental motion. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3986–3993. IEEE, 2014.

[28] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 22(11):1330–1334, 2000.