3D Reconstruction from Accidental Motion

Fisher Yu
Princeton University

David Gallup
Google Inc.

(a) Input image sequence (b) Foreground points of SfM (c) Dense reconstruction (d) Synthetic aperture

Figure 1: We investigate reconstruction from image sequence with very small motion as shown in (a), where the motion between images are hard to observe. (b) and (c) shows our reconstruction results and our method leads to interesting applications such as synthetic aperture effect as shown in (d), where the foreground object is in focus and the background is blurred.

Abstract

We have discovered that 3D reconstruction can be achieved from a single still photographic capture due to accidental motions of the photographer, even while attempting to hold the camera still. Although these motions result in little baseline and therefore high depth uncertainty, in theory, we can combine many such measurements over the duration of the capture process (a few seconds) to achieve usable depth estimates. We present a novel 3D reconstruction system tailored for this problem that produces depth maps from short video sequences from standard cameras without the need for multi-lens optics, active sensors, or intentional motions by the photographer. This result leads to the possibility that depth maps of sufficient quality for RGB-D photography applications like perspective change, simulated aperture, and object segmentation, can come "for free" for a significant fraction of still photographs under reasonable conditions.

1. Introduction

When a person captures a still photo by hand, it usually takes several seconds between pointing the camera to the scene and pressing the shutter button. During this time, while one intends to hold the camera still, there is inevitable motion due to hand shaking or heart beating, especially when a lightweight camera like a smartphone, is used. We call this type of motion *accidental motion*. If a camera were to capture a short video before and/or after the capture of a still, would it be possible to use the baseline (translation) from accidental motion for 3D reconstruction? We demonstrate in this paper that indeed 3D reconstruction can be achieved, and that the resulting reconstruction can be used for a variety of applications.

In this paper we investigate the properties of accidental motion and find a method to reconstruct 3D information of the image sequences. There are two main challenges to this problem. First, the commonly used Structure from Motion (SfM) approaches assume that a good two-view reconstruction can be obtained with algebraic methods, which in turn depend on adequate baseline between overlapping views. In accidental motion, the maximum viewing angle for a 3D point is usually less than 0.2 degrees, where algebraic methods are very unstable. Second, the depth uncertainty is very large due to small baseline and, therefore, the previous multiview stereo methods can produce serious artifacts.

To address these issues, we find that we can use multiple images together to do SfM directly. Due to accidental motion, we use inverse depth relative to a reference view to parameterize the 3D points, which helps regularize bundle adjustment. We find that random depth and identical camera poses are good initialization for bundle adjustment with all the images. We also find that many images can help reduce uncertainty.

Given camera poses, the depth estimation of most of the pixels is noisy and has high uncertainty. Because the depth signal is weak and noisy, we find that the popular first-order CRF is not very effective in regularizing depth, and can often result in an oversmoothed depth map, as shown in Figure 2. We propose to use long range connections, and we show that direct connections between a pixel and its bigger

neighborhood can improve the dense reconstruction in our case.

We have conducted a user study that yields empirical evidence that there is several-millimeter translation throughout the capture of a still photo. Under reasonable conditions, such as 3-meter depth, focal length of 2000 pixels, and localization standard deviation of 1 pixel, a baseline of 3 mm over 100 frames (a few seconds at 30 fps) is enough to estimate depth with a standard deviation of 150 mm, which is low enough uncertainty for many applications.

We test our algorithm on a variety of scenes captured by a variety of users. The proposed method can indeed produce high quality depth maps, and these depth maps are good enough for RGB-D photography applications, such as synthetic aperture (focus change) and parallax effects.

2. Previous Work

We follow the common pipeline to build dense 3D models from a collection of images. We first do SfM to estimate the viewing parameters of each image and then use them to do multiview stereo to get dense reconstruction. A wealth of previous work has studied this two problems, and we mention some of them here to show the difference of our system.

Structure from motion has been actively studied for a long time and we have got a good understanding of the geometric properties of estimating sparse structure and camera poses [10]. Bundle adjustment is commonly used to obtain the optimal estimates [31]. Nonlinear least squares is used to measure the projection errors because of its nice error modeling properties. But it is usually difficult to optimize the nonlinear cost function and a good initialization is critical. [26] presents a successful way to do incremental bundle adjustment, which relies on two-view reconstruction. However, when the motion is very small as in our case, the two-view reconstruction is ill conditioned and therefore it can't provide reliable initialization. Discrete optimization [7] is also proposed to initialize structure and camera parameters. But the optimization itself is a hard problem and there is a tradeoff between accuracy and complexity. To work around the nonlinearity of the cost functions, some other error measures are also proposed. [12, 25, 13] propose to use L_∞ norm instead of L_2 to measure the reprojection error because the resulting cost function is convex. But L_∞ is not robust to outliers, which are unavoidable in most of the applications. We will show that even in our case, where the feature matching is supposed to be easier than the general case due to little view point and illumination change, we still need to deal with outliers in feature matching. Robustifing the cost function can help improve the reconstruction result.

Instead of doing bundle adjustment with multiple images, some works [29, 5, 27, 30, 20] propose factorization methods to do multiview SfM directly. Potentially, those methods should be used as initialization for bundle adjustment. However, in our experiments, we find that in presence of feature localization noise and outliers, these methods are unstable and our proposed initialization is the most effective.

Several works [19, 21, 4] study the ambiguity properties of structure from small motion and propose some algorithms. But the analysis of the bundle adjustment is mainly for two-view case. In this paper, we will present analysis for the multiview case and show that with the assumption of small motion, tasks such as estimating point depth can be easier to solve. Although several researchers [20, 4] have proposed methods to reconstruct sparse structure, to our knowledge, our method is the first to deal successfully with outliers and to work in practice. A recent work [18] proposes to use a similar initialization approach to ours to initialize a tracking system. But their goal is not to find a 3D structure and we find that random depth initialization works better than their proposed constant depth initialization.

Multiview stereo When the camera motion is very small, the view change is very small. We aim to estimate depth for each pixel in the reference view instead of a complete 3D model. Therefore stereo methods are more relevant here. Even if SfM can provide perfect camera parameters, the photo-consistency measurement at each pixel can still be noisy due to various reasons such as image noise and the aperture problem. Various methods have been proposed to solve this problem by smoothing or regularizing the depth estimation. The Conditional Random Field (CRF) framework is one of the most successful methods [3, 11]. A probabilistic model is used to associate adjacent pixels to encourage them to have similar depth values. Second order Markov Random Field (MRF) is also proposed [14, 32, 15] to avoid the fronto-parallel bias. However, those methods can only connect adjacent pixels, although they are global methods. In our experiments, we find that the low order connection can't regularize our depth effectively. Therefore, we propose to connect pixels over even longer ranges. The inference is made possible by the recent development of high dimensional Gaussian Filtering [1] and the mean field method [16]. We will show that this method can effectively regularize noisy depth maps estimated from weak data terms. Some local methods [22] based on cost-volume filtering have also been proposed to solve the stereo problem. Our method bears a similarity to the filtering methods, but our method is based on a global formulation, which usually performs better than local methods, as evaluated on Middlebury benchmark [23].

3. Structure from Motion

Given feature correspondences between images, we use bundle adjustment to get the 3D structure and camera poses of these images. It is well known that the cost function of bundle adjustment is nonlinear and it is easy to get stuck in a local minimum that is far away from the global minimum. It is hard to even solve part of the problem [9]. Incremental bundle adjustment based on two view reconstruction is often used to get a good initialization. Surprisingly, we find experimentally that in the small motion case, identical camera poses and random point depth are good initialization for the cost function. What's more, because the view change is small, we can parameterize the 3D point position as depth in the reference view, which also contributes to the success of bundle adjustment. The small motion assumption also makes the analysis of the cost function in bundle adjustment easier. In this section, we first analyze the cost function of bundle adjustment with the assumption of small motion (both rotation and translation). Although the bundle adjustment is still a complicated optimization problem under this assumption, we can show that it has some nice properties. When the camera poses are fixed, it is convex to get the depth of a feature relative to a reference view. Also, it is convex to optimize the rotation for the points at infinity when an approximation is used. We will present our method after proofs of the properties. In Section 5, we demonstrate that our method is effective in reasonably restricted environments.

3.1. Definitions

Assume we have an image sequence with N_c images and N_p points in 3D, where every point is visible to all the images. Let the camera of the first image be the reference view, and the i-th camera is related to it by a relative rotation matrix R_i followed by relative translation $\mathbf{T}_i = [T_i^x, T_i^y, T_i^z]^T$. Assume P_j is the position of the j-th point in the coordinate system of the reference camera. Its position in the coordinate system of the i-th camera is $\mathbf{R}_i\mathbf{P}_j + \mathbf{T}_i$.

Let $\mathbf{\Theta} = [\theta_i^x, \theta_i^y, \theta_i^y]$ be the rotation angles of the i-th camera. With the assumption of small angles, $\mathbf{R}_i$ can be approximated by

$$\mathbf{R}_i = \begin{bmatrix} 1 & -\theta_i^z & \theta_i^y \\ \theta_i^z & 1 & -\theta_i^x \\ -\theta_i^y & \theta_i^x & 1 \end{bmatrix}. \tag{1}$$

To make the resulting optimization easier, we parameterize each 3D point by its inverse depth. so we have $\mathbf{P}_j = \frac{1}{w_j}[x_j, y_j, 1]^T$, where (x_j, y_j) is the projection of $\mathbf{P}_j$ in the reference image. The projection of $\mathbf{P}_j$ on the i-th image is $\mathbf{p}_{ij} = [p_{ij}^x, p_{ij}^y]^T$. Let $\pi : \mathcal{R}^3 \to \mathcal{R}^2$ be the projection function, that is, $\pi([x, y, z]^T) = [x/z, y/z]^T$.

3.2. Analysis

We use the L_2 norm to measure the reprojection error because it has nice statistical interpretation and can be robustified [31].

Based on the above definitions, we can define the cost function of bundle adjustment in the retina plane as

$$\begin{aligned} F &= \sum_{i=1}^{N_c}\sum_{j=1}^{N_p} ||p_{ij} - \pi(R_iP_j + T_i)||^2, \\ &= \sum_{i=1}^{N_c}\sum_{j=1}^{N_p} (\frac{e_{ij}^x + f_{ij}^x w_j}{c_{ij} + d_{ij}w_j})^2 + (\frac{e_{ij}^y + f_{ij}^y w_j}{c_{ij} + d_{ij}w_j})^2, \end{aligned} \tag{2}$$

where

$$\begin{aligned} a_{ij}^x &= x_j - \theta_i^z y_j + \theta_i^y, \\ b_{ij}^x &= T_i^x, \\ a_{ij}^y &= y_j - \theta_i^x + \theta_i^z x_j, \\ b_{ij}^y &= T_i^y, \\ c_{ij} &= -\theta_i^y x_j + \theta_i^x y_j + 1, \\ d_{ij} &= T_i^z, \\ e_{ij}^x &= p_{ij}^x c_{ij} - a_{ij}^x, \\ f_{ij}^x &= p_{ij}^x d_{ij} - b_{ij}^x, \\ e_{ij}^y &= p_{ij}^y c_{ij} - a_{ij}^y, \\ f_{ij}^y &= p_{ij}^y d_{ij} - b_{ij}^y. \end{aligned} \tag{3}$$

Depth Estimation Assume that the correct camera poses are given and fixed. The depth estimation is to find the depth of a point minimizing

$$F_i(w_j) = \sum_{i=1}^{N_c} f_j^x(w_j) + f_j^y(w_j), \tag{4}$$

where $f_j^x(w_j) = (\frac{e_{ij}^x + f_{ij}^x w_j}{c_{ij} + d_{ij}w_j})^2$ and $f_j^y(w_j) = (\frac{e_{ij}^y + f_{ij}^y w_j}{c_{ij} + d_{ij}w_j})^2$. We will prove estimating the depth is easier in the context of small motion.

First, consider the general form of f_j^x and f_j^y, $f(x) = (\frac{x-a}{x-b})^2$, where a and b are the zero and pole of the function, respectively. When $a > b$, the function is convex in $(b, \frac{3a}{2} - \frac{b}{2})$. When $a < b$, the function is convex in $(\frac{3a}{2} - \frac{b}{2}, b)$.

Assume that $f_j^x(\bar{w}_j^x) = 0$, that is, $\bar{w}_j^x = -\frac{e_{ij}}{f_{ij}}$. Because $c_{ij} \approx 1$ and $|d_{ij}| \ll \frac{1}{w_j}$, $w_j \ll |\frac{c_{ij}}{d_{ij}}|$. So $f_j^x(w_j)$ is convex in $(0, |\frac{c_{ij}}{2d_{ij}}|)$, so is $f_j^y(w_j)$. Hence, $F(w_j)$ is convex in $(0, \min_i |\frac{c_{ij}}{2d_{ij}}|)$. Since $|\frac{c_{ij}}{2d_{ij}}|$ is supposed to be far greater than reasonable values of w_j, we can easily optimize w_j for the reprojection error in Equation 4. Also, note that if there is noise in the detection $\mathbf{p}_{ij}$, it doesn't change c_{ij} and d_{ij}, and hence the convex interval $(0, \min_i |\frac{c_{ij}}{2d_{ij}}|)$ of $F_i(w_j)$. What's more, the convexity analysis of the cost function doesn't depend on the approximation of the rotation matrix. It is an exact property of depth estimation with small motion.

Points at Infinity If the points are approximately at infinity, the cost function in Equation 2 can be approximated by

$$F \approx \sum_{i=1}^{N_c} \sum_{j=1}^{N_p} (e_{ij}^x)^2 + (e_{ij}^y)^2. \tag{5}$$

It is a convex function of the camera rotation angles on the domain around 0.

Depth Uncertainty Consider a rectified stereo pair separated by a baseline b, observing a point at inverse depth w. The relationship between disparity and depth is given by $w = \frac{d}{fb}$, where d is the disparity and f is the focal length. Ignoring quantization errors and mismatches, we can obtain the inverse depth estimation at any single pixel, namely

$$\mathrm{Var}[\hat{w}] = \mathbf{E}[(\frac{d+\epsilon}{fb} - \frac{d}{fb})^2] = \frac{\mathrm{Var}[\epsilon]}{f^2 b^2}, \tag{6}$$

where ϵ is the feature localization error. Unlike analyzing the variance of depth, we don't have to take first-order approximation here. Similarly, assuming that we have n observations of the point and they have the same variance, we can get the variance of the combined estimation $\hat{w} = \frac{1}{n}\sum_{i=1}^{n} \hat{w}_i$:

$$\begin{aligned} \mathrm{Var}[\hat{w}] &= \frac{1}{n^2 f^2 b^2} \mathbf{E}[(\sum_{i=1}^{n} \epsilon_i)^2] \\ &= \frac{1}{f^2 b^2}(\frac{1}{n} + \rho(1 - \frac{1}{n}))\mathrm{Var}[\epsilon], \end{aligned} \tag{7}$$

where $\mathrm{Cov}[\epsilon_i, \epsilon_j] = \rho \mathrm{Var}[\epsilon]$ for all i, j between 1 and n and $i \neq j$, and $\mathrm{Var}[\epsilon_i] = \mathrm{Var}[\epsilon]$. This indicates that if the feature detection errors are independent, the standard deviation of the inverse depth estimation decrease linearly with $\sqrt{n}$. However, if the feature detection errors are fully correlated, multiple observations don't help reduce uncertainty. Similar conclusion can be drawn for depth [8].

3.3. Initialization

A good initialization is crucial to finding good minima of reprojection errors. Because of the results in Section, 3.2 we conjecture that a random initialization for structure may give good results. Given a sequence of images, we select a reference view and initialize all the camera poses with zero rotation and translation. As mentioned above, the points are parameterized by inverse depth. The projections of the 3D points are proposed by feature tracking across the images. First, corner features [24] are detected in the reference image. Then, instead of tracking the corners in the image sequence order, we track all the corners from the reference image to each of the other images with Kanade-Lucas-Tomasi (KLT) [17, 28] feature tracker. This can effectively reduce the accumulative localization error of feature tracking. To remove the tracking outliers, we require that all the features can be tracked to all the non-reference images and the maximum color gradient difference per pixel between the two patches should be under a threshold. KLT method can provide subpixel accuracy, and this is critical when the camera motion and therefore the feature movement are small.

3.4. Optimization

We optimize the cost function of bundle adjustment in Equation 2 with Ceres Solver [2]. Robustifiers are optionally used in the cost function. The camera of the reference view is fixed at the coordinate origin. Usually, the outliers can be neglected after the feature tracking and selection in initialization. However, we find cases where robustifiers can improve the reconstruction results. On the other hand, after each optimization, we remove the points with negative depth and optimize again with the remaining points.

4. Dense Reconstruction

After getting structure from motion, we want to densely reconstruct the 3D scene by estimating the depth of the images. Because all the input images capture the scene from a similar viewpoint, we can only get a 3D structure seen from the common viewpoint. Therefore, we aim to get a depth map of a reference view as the 3D reconstruction output. Because the depth signal at each pixel tends to be noisy in our case, we adopt plane sweeping together with the CRF framework [11] to solve a smooth depth map.

One distinct attribute of multiview stereo from small baseline images is that the confidence of depth minima is low in general instead of just in textureless areas. Therefore, the details can be easily smoothed out, as shown in Figure 2. To preserve the details while smoothing the depth map, we propose to use long range connection between pixels in the CRF energy function, which can pass information to a pixel effectively.

4.1. Formulation

The input is a set of images. Let $\mathcal{I}$ be the index set of the pixels in a reference view I, and $\mathrm{I}(i), i \in \mathcal{I}$, is the color of the i-th pixel. The goal is to determine a dense depth map, D, of the reference view. Let L map each pixel index $i \in \mathcal{I}$ to a 2D location in the image. Let P be the photo-consistency function such that $\mathrm{P}(i, d)$ is the photo-consistency score of the i-th pixel at distance d.

The energy we intend to minimize is

$$E(\mathrm{D}) = E_p(\mathrm{D}) + \alpha E_s(\mathrm{D}). \tag{8}$$

E_p is the standard photo-consistency term of the form

$$E_p(\mathrm{D}) = \sum_{i \in \mathcal{I}} \mathrm{P}(i, \mathrm{D}(i)), \tag{9}$$

which can be obtained by plane sweeping algorithm [6].

E_s is the smoothness term to regularize the depth estimation. It often represents first-order or second-order

(a) Reference View (b) WTA (c) Long Range Connection

(d) Less first-order Smoothness (e) First-order Smoothness (f) More first-order Smoothness

Figure 2: Comparison of first-order and long range connection. (b) shows the data term. (c) is the result optimized based on the long range connected model. (d) to (f) shows the graph cut solution of the first-order smoothness with increasing regularization. Because the data term is very noisy, first-order regularization always oversmooth the estimated depth to reduce noise.

CRF model to connect and pass information between adjacent pixels. However, we find that those adjacent connected model can't effectively regularize the noisy data term. Hence, we propose to connect pixels with longer range so that the photo-consistency measurement can be effectively aggregated from an area to a pixel in it.

To build a connection between pixels that are not adjacent, we introduce the function $\mathrm{C}(i, j, \mathrm{I}, \mathrm{L}, \mathrm{D})$, which gives a score for the depth assignment of the i-th and the j-th pixels based on the color intensities and their locations in the reference images. So E_s is the long range connection term of the form

$$E_s(\mathrm{D}) = \sum_{i \in \mathcal{I}, j \in \mathcal{I}, i \neq j} \mathrm{C}(i, j, \mathrm{I}, \mathrm{L}, \mathrm{D}), \qquad (10)$$

and

$$\mathrm{C}(i, j, \mathrm{I}, \mathrm{L}, \mathrm{D}) = \rho_c(\mathrm{D}(i), \mathrm{D}(j)) \times \exp\left(-\frac{||\mathrm{I}(i) - \mathrm{I}(j)||^2}{\theta_c} - \frac{||L(i) - L(j)||^2}{\theta_p}\right), \qquad (11)$$

where ρ_c is robust measurement of depth difference, and θ_c and θ_p are parameters to control the connection strength and range. We choose ρ_c to be the truncated linear function, i.e., $\rho_c = \min(t, |\mathrm{D}(i) - \mathrm{D}(j)|)$, where t is a threshold. The purpose of E_c is to connect pixels within an area with similar colors such that they can have consistent depth, since they are more likely to belong to the same object.

4.2. Optimization

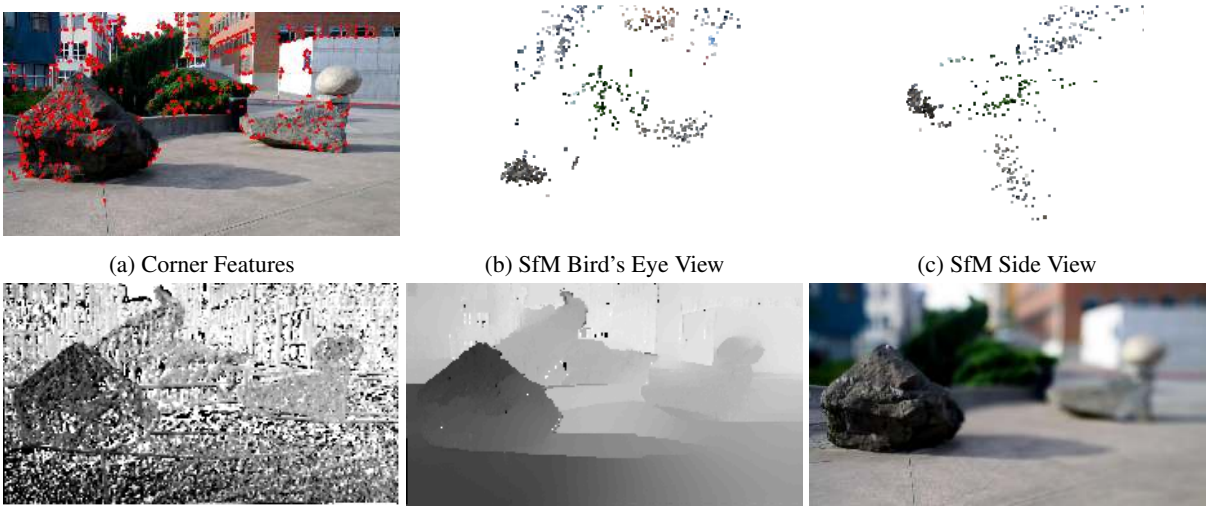
We use the mean field method with an efficient implementation proposed in [16] to optimize Equation 8. It can solve the dense CRF model and give a smooth depth map efficiently.

5. Experiments

We evaluate our methods on both synthetic and real data. The real data is collected by a smartphone camera, and it is captured in the video mode at 24 frames per second. To make our system practical to real world applications, we limit the number of images to 100, which is about a 4-second video. The camera intrinsic parameters are calculated from the factory specification of the phone and the image distortion is not accounted for. Better results are expected when a better camera is used. More results are shown in the supplemental material.

Google Nexus 4 (smartphone)				
		Translation		
All	Translation	stdev. (mm) after		
users	speed (mm/s)	1s	2s	3s
Mean	18.07	2.18	3.35	3.81
Stdev.	6.67	1.11	1.99	2.31
Canon PowerShot S95 (point-and-shoot)				
		Translation		
All	Translation	stdev. (mm) after		
users	speed (mm/s)	1s	2s	3s
Mean	9.23	1.71	3.02	3.99
Stdev.	2.10	0.65	1.23	1.88

Figure 4: Camera translation statistics obtained from a user study of 9 participants. Users were asked to record video of a calibration pattern and hold the camera steady, as if they were capturing a photograph. Although the smartphone moves faster than the point-and-shoot (perhaps due to the weight and form), both cameras exhibit similar standard deviation of translation (camera centers).



(a) Corner Features (b) SfM Bird's Eye View (c) SfM Side View

(d) WTA depth estimation (e) Smoothed depth estimation (f) Synthetic aperture

Figure 3: Pipeline of our system. (a) We select the first image in a sequence as the reference view. Corner features (Red dots) are extracted in the reference view and tracked to the other images. (b) and (c) show the SfM result with initialization of random structure and identical camera poses. (d) WTA of the photo-consistency at each pixel (e) The smoothed depth estimation based on our proposed energy function with long range connections. (f) Given the depth map, we can refocus on part of the image.

5.1. User Behavior

We have conducted a user study to determine the magnitude of accidental translational motion during still photography. To measure camera motion, we asked users to capture videos of a calibration pattern at a distance of roughly 0.5 meters. Users were instructed to hold the camera steady, as if they were capturing a photograph, for a duration of 5 seconds. We evaluated 9 participants and two cameras: a Google Nexus 4 smartphone, and a Canon PowerShot S95 point-and-shoot. The results are shown in Figure 4. From this study, we observe that after 3 seconds, the camera centers exhibit a standard deviation of 3.9 mm, which yields sufficient baseline to obtain a good reconstruction under reasonable conditions. For example, for a scene depth of 3 meters, 100 frames of video, feature localization (or disparity) standard deviation of 1 pixel, and a focal length of 2000 pixels, we would expect a depth standard deviation of 0.115 meters, assuming measurements are uncorrelated. We have asked several users to capture a 4-second video of a natural scene using a Galaxy Nexus smartphone, and our algorithm generates similar results shown in this paper.

5.2. Structure from Motion

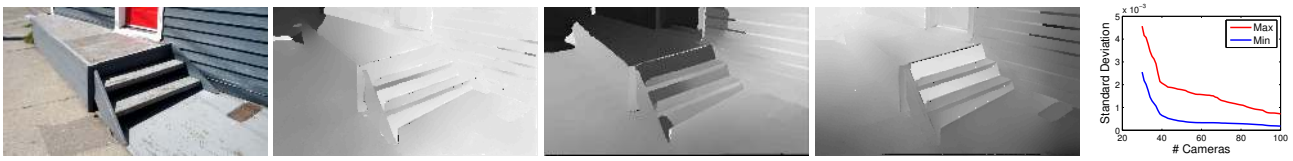
We follow the method described in Section 3 in our experiments. An image sequence of a video is taken as the input. The first image of a sequence is selected as the reference view. When we remove the feature tracking outliers by average pixel difference in a patch, we usually use 6 as the threshold for a 8-bit encoded gray image. All the 3D positions of the feature points are parametrized by their inverse depth relative to the reference view. Before the bundle adjustment, all the cameras have zero rotation and translation, and all the points have uniformly random depth between 2 and 4 meters.

SfM Results The bundle adjustment results are shown in Figure 3b and 3c. It demonstrates that our simple initialization method is effective in the small motion scenario. Since we don't have to do two-view reconstruction for each pair of images or solve hard optimization problem [7], SfM is very fast. With about 1000 points and 100 cameras, it usually takes several seconds on a modern desktop.

Feature tracking outliers are inevitable, but in most of the cases, they don't affect the result. However, when there are too many outliers, a robustifier can be used as in the general structure from motion problem. We observe that when the robustifier is not necessary, the SfM results look better without it.

Multiple Images To understand how the multiple images help the reconstruction, we can first look at the depth estimation uncertainty. As shown in Figure 5e, the depth uncertainty of the 3D points decreases with more input images. It shows that in the case of KLT tracked features, more images can help reduce the tracking noise.

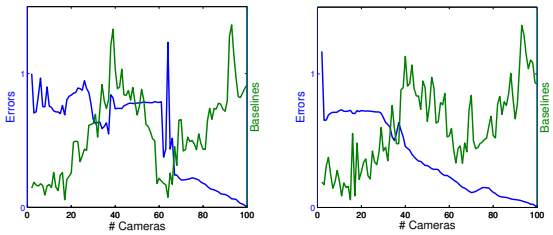
To understand how different numbers of images change final structure, we did bundle adjustment with different number of images while fixing the detected features and their matching. Since the camera intrinsic paramters are known and the 3D points are reconstructed up to scale, we first normalize the inverse depth values to have the same mean and variance. One of the results is shown in Figure 6. The blue curve shows the structure error measured by sum of squared difference between the models reconstructed by



(a) Reference Image (b) 10 Images (c) 50 Images (d) 100 Images (e)

Figure 5: Change of smoothed depth maps with different number of images. Darker color indicates closer depth. More images decrease the uncertainty of the reconstruction and also reduce the influence of outliers. (b) to (d) show the change the structure with their smoothed depth map. (e) shows the change of depth estimation uncertainty with number of cameras in this example. The Y axis shows the standard deviation of the inverse depth. The maximum and minimum of the uncertainty continue decreasing with addition of images. The depth uncertainty is measured with camera poses fixed. Please note darker means closer.

certain number of images and all the images. The green curve shows the baseline between a camera and the reference camera in the model reconstructed by all the images. only the points with middle 90% depth ranking are considered in normalization to reduce the effects of outliers. As we can observe in Figure 6a, there is a big error jump between 60 and 70 cameras. Since the baseline doesn't significantly increase, the error change may be because of the matching outliers. If robustifier is added to the cost function, there is no sudden error change, as shown in Figure 6b.



(a) (b)

Figure 6: How multiple images help the reconstructed result. (a) shows that there is a sudden change in the reconstruction error. (b) shows that this sudden change is due to matching outliers and in general, multiple images can help reduce the effect of outliers.

Figure 5 shows the evolution of the structure in 2D with depth maps. When more images are used, the structure gradually becomes better. In some cases, we observe that the structure is already good enough when 50 images are used, although more images can decrease the point position uncertainty.

Therefore, more images can help reduce the reconstructed depth uncertainty and the effects of possible outliers.

Points at Infinity In Section 3, we mentioned that optimizing with points at infinity is equivalent to convex optimization of rotation. In practice, as the distant points are approximately at infinity, they play an important role in resolving the ambiguities between camera rotation and translation. If we remove the distant points, the bundle adjustment can easily get stuck in a local minimum. Even if we initialize the bundle adjustment with a good structure, the bundle adjustment can still distort the structure due to feature noise.

5.3. Dense Reconstruction

After getting camera poses from SfM, we can do a dense reconstruction using the method in Section 4. We will show the role of each term in the energy function in Equation 8 and argue that the terms are necessary to get plausible depth map.

Data Term If we only optimize E_a of the energy function in Equation 8, we will get the noisy depth map that optimize the photo consistency at each pixel, which is winner-take-all (WTA), as shown in Figure 3d. We observe that the planes, such as the ground and the wall, present consistent depth values in general, though the values are noisy.

First-Order Smoothness Figure 2d-f show regularized depth maps with first-order smoothness. We observe that although some areas of the depth map are still noisy, part of it is already oversmoothed. When the regularization is weak, the estimated depth is still noisy. When the noise is reduced to a good level, the estimated depth is oversmoothed and an object is reconstructed to several layers. This motivates us to seek long connection between pixels to pass the information more effectively.

Long Range Connection Instead of only connecting adjacent pixels for smoothness, we connect pixels with longer range. This can effectively accumulate the information from a selected neighborhood. Inspired by the recent works of joint segmentation and stereo estimation, we first smooth the reference image with mean shift before using its color to compute the pixel connection weight in Equation 11. For an image of size 480 by 270, we normally choose θ_c from 20 to 30 and θ_p from 5 to 9. Greater θ_p should be used for higher resolution image. Because of the efficient implementation of mean field inference, the running time doesn't change with the values of θ_c and θ_p. The connection threshold t used in Equation 11 is chosen to be a fixed percentage of the total label number, which is 15% in our system. Because the truncated linear function can be implemented as two convolutions of 1D box filtering, the running time is linear to the number of depth labels. The results are shown in Figure 3e.

5.4. Application

The reconstructed depth map can facilitate a lot of applications that are nearly impossible with a single color image. For example, we can use the 3D information to simulate different aperture effects or synthesize new views. To test our depth map is good enough for such applications, we can do refocus of the reference image. As shown in Figure 3f, the generated depth map can clearly show the depth change of the objects in the scene.

6. Conclusion

We propose the first practical system to reconstruct 3D structure from small motion image sequences. We discover that in the case of small motion, random point depth relative to a reference view and identical camera poses are good initialization for the bundle adjustment cost function, even in presence of outliers. Although the reconstructed 3D points at the background have very high uncertainty, the foreground points clearly show the 3D structure. We provide some analysis of the cost function and find some of its nice properties with the assumption of small motion. Further, based on the noisy nature of the photo consistency measurement, we propose to use long range connection between pixels to regularize the depth map, and the resulted depth map looks much better than only using connections between adjacent pixels. We also demonstrate that the resulting depth map has enough quality to make perceptualy plausible refocused images.

7. Future Work

The proposed method works well in practice but more theoretical work is necessary to analyze the ambiguities of multiview reconstruction with small motion and find exact conditions when the method works. Also, we want to know more about how people capture the scene and the success rate of our methods under different conditions.

Acknowledgments We thank the Google Seattle Lightfield team for helpful discussion and support during this work, especially Steve Seitz. We also thank Carlos Hernández and Simon Fuhrmann for rendering the synthetic aperture results.

References

[1] A. Adams, J. Baek, and M. A. Davis. Fast high-dimensional filtering using the permutohedral lattice. *CGF*, 2010. 2

[2] S. Agarwal and K. Mierle. *Ceres Solver: Tutorial & Reference*. Google Inc. 4

[3] N. D. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *ECCV*, 2008. 2

[4] A. Chiuso, R. Brockett, and S. Soatto. Optimal structure from motion: Local ambiguities and global estimates. *IJCV*, 2000. 2

[5] S. Christy and R. Horaud. Euclidean shape and motion from multiple perspective views by affine iterations. *TPAMI*, 1996. 2

[6] R. Collins. A space-sweep approach to true multi-image matching. In *CVPR*, 1996. 4

[7] D. Crandall, A. Owens, N. Snavely, and D. Huttenlocher. Discrete-continuous optimization for large-scale structure from motion. In *CVPR*, 2011. 2, 6

[8] C. Geyer, T. Templeton, M. Meingast, and S. Sastry. The recursive multi-frame planar parallax algorithm. In *3DPVT*, 2006. 4

[9] R. Hartley, F. Kahl, C. Olsson, and Y. Seo. Verifying global minima for L_2 minimization problems in multiple view geometry. *IJCV*, 2013. 3

[10] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Second edition, 2004. 2

[11] C. Hernández and G. Vogiatzis. Shape from photographs: A multi-view stereo pipeline. In *Computer Vision*, Studies in Computational Intelligence. 2010. 2, 4

[12] F. Kahl. Multiple view geometry and the L_8-norm. In *ICCV*, 2005. 2

[13] F. Kahl, S. Agarwal, M. Chandraker, D. Kriegman, and S. Belongie. Practical global optimization for multiview geometry. *IJCV*, 2008. 2

[14] P. Kohli, L. Ladicky, and P. H. S. Torr. Robust higher order potentials for enforcing label consistency. In *CVPR*, 2008. 2

[15] N. Komodakis and N. Paragios. Beyond pairwise energies: Efficient optimization for higher-order mrfs. In *CVPR*, 2009. 2

[16] P. Krähenbühl and V. Koltun. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. In *NIPS*, 2011. 2, 5

[17] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, 1981. 4

[18] A. Mulloni, G. Reitmayr, D. Wagner, R. Grasset, and S. Diaz. User Friendly SLAM Initialization. *ISMAR*, 2013. 2

[19] J. Oliensis. Computing the camera heading from multiple frames. In *CVPR*, 1998. 2

[20] J. Oliensis. A multi-frame structure-from-motion algorithm under perspective projection. *IJCV*, 1999. 2

[21] J. Oliensis. The least-squares error for structure from infinitesimal motion. *IJCV*, 2005. 2

[22] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz. Fast cost-volume filtering for visual correspondence and beyond. In *CVPR*, 2011. 2

[23] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 2002. 2

[24] J. Shi and C. Tomasi. Good features to track. In *CVPR*, 1994. 4

[25] K. Sim and R. Hartley. Recovering camera motion using linfty minimization. In *CVPR*, 2006. 2

[26] N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the world from internet photo collections. *IJCV*, 2008. 2

[27] P. Sturm and B. Triggs. A factorization based algorithm for multi-image projective structure and motion. In *ECCV*. 1996. 2

[28] C. Tomasi and T. Kanade. Detection and tracking of point features. 1991. 4

[29] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *IJCV*, 1992. 2

[30] B. Triggs. Factorization methods for projective structure and motion. CVPR, 1996. 2

[31] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon. Bundle adjustmenta modern synthesis. In *Vision algorithms: theory and practice*. 2000. 2, 3

[32] O. Woodford, P. H. S. Torr, I. Reid, and A. Fitzgibbon. Global stereo reconstruction under second order smoothness priors. In *CVPR*, 2008. 2