The Light Field 3D Scanner

Yingliang Zhang[1] Zhong Li[2,3] Wei Yang[2,3] Peihong Yu[1] Haiting Lin[3] Jingyi Yu[1,2]
[1]ShanghaiTech University, Shanghai, China. {zhangyl,yuph,yujy1}@shanghaitech.edu.cn
[2]University of Delaware, Newark, DE, USA. {lizhong,wyangcs}@udel.edu
[3]Plex-VR Inc. {haiting.lin}@plex-vr.com

Abstract

We present a novel light field structure-from-motion (SfM) framework for reliable 3D object reconstruction. Specifically, we use the light field (LF) camera such as Lytro and Raytrix as a virtual 3D scanner. We move an LF camera around the object and register between multiple LF shots. We show that applying conventional SfM on sub-aperture images is not only expensive but also unreliable due to ultra-small baseline and low image resolution. Instead, our LF-SfM scheme maps ray manifolds across LFs. Specifically, we show how rays passing through a common 3D point transform between two LFs and we develop reliable technique for extracting extrinsic parameters from this ray transform. Next, we apply a new edge-preserving stereo matching technique on individual LFs and conduct LF bundle adjustment to jointly optimize pose and geometry. Comprehensive experiments show our solution outperforms many state-of-the-art passive and even active techniques especially on topologically complex objects.

1. Introduction

Recovering high-fidelity 3D models from imagery data is a long standing problem in computer vision. Significant progress has been made in the past decade, on both passive and active fronts [33, 18, 31]. New active solutions based on structured light and time-of-flight [16][12] can now acquire 3D models in real-time for applications such as interactive gaming. They are, however, mainly limited to indoor applications to avoid interference from environment lighting. Further, most affordable active sensors are of a low spatial and depth resolution. In contrast, passive techniques such as Structure-from-Motion (SfM) have been perfected in the few two decades and can handle both indoor and outdoor scenes as large as an urban environment [8, 1, 2, 32]. The final reconstruction, however, is generally sparse and requires elaborate triangulation and model fitting [7].

The renewed interest on virtual reality (VR) and augmented reality (AR) has also brought new demand

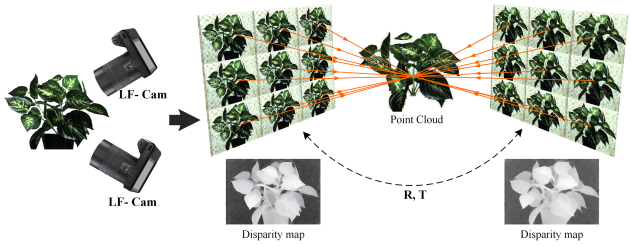

Figure 1. The light field 3D scanner framework for producing ultra high quality 3D reconstruction.

for 3D reconstruction - 3D real objects need to be "scanned" efficiently, accurately, and most importantly omni-directionally, so that VR headset users can freely lean towards or walk around the captured subject. The best known solution is the USC light stage where structure light and multi-camera acquisition are coupled for shape, texture, reflection and motion capture. A miniature version developed by 8i, a leading VR startup company, uses a similar dome setup to acquire photorealistic human characters. Although effectively, such solutions are bulky, expensive, and unsuitable for onsite 3D acquisition. Other solutions such as Microsoft KinectFusion [18] yields coarse 3D models due to low range resolution as shown in Fig. 7.

In this paper, we present a 3D object scanning solution that uses the LF camera as a virtual 3D scanner (Fig. 1). Light fields are image-based representations that were originally designed for producing special photographic effects such as refocusing and view morphing. With the availability of commodity LF cameras such as Lytro and Raytrix, a light field can now be directly acquired in a single shot. Compared with classical multi-view geometry, such as Light Stage, LFs exhibit several unique features amenable for 3D reconstruction, including regular sampling pattern [26, 36], dense angularly sampling density [6, 15, 22], and subpixel disparity [19, 39]. By far light fields have shown success in obtaining disparity maps through stereo matching. Yet, for full 3D scanning using multiple light fields, several key issues such as light field pose estimation and depth fusion

need to be addressed.

We present a novel light field structure-from-motion (SfM) framework for reliable 3D object reconstruction. Specifically, we use the light field (LF) camera such as Lytro and Raytrix as a virtual 3D scanner. We move an LF camera around the object and register between multiple LF shots. We show that applying conventional SfM on subaperture images is not only expensive but also unreliable due to ultra-small camera baseline and low image resolution. Instead, our LF-SfM scheme maps ray manifolds across the LFs. Specifically, we show how rays passing through a common 3D point transform between two LFs and we develop reliable technique for extracting the extrinsic parameters from this ray transform.

Next, we apply a new edge-preserving stereo matching technique on individual LFs. Our new LF stereo matching scheme preserves the sharpness of the occlusion boundary which is essential for high-fidelity 3D reconstruction. Finally, we conduct light field bundle adjustment to jointly optimize pose and geometry estimation. To validate our approach, we compare our scheme with two classical approaches: SfM with all views in all LFs cameras and Microsoft KinectFusion [18]. We show that our scheme provides a much more reliable solution especially when reconstructing topologically complex objects.

1.1. Related Work

As a passive sensing scheme, our work is most related to structure-from-motion and light field stereo vision.

Structure-from-Motion (SfM) aim to simultaneously recover camera motion and scene structure from multiple images [9, 35], presumably captured by a perspective camera. With immense computational powers, SfM can now be used to recover very large scale 3D models [37, 34, 17, 10, 25], e.g., from community photo collections shared on the internet [33]. The key component in SfM is to establish reliable feature correspondences. For cultural heritage architecture composed mainly of piecewise linear facades, SfM has shown great success in obtaining extremely realistic models [21, 5, 27, 28]. However, for geometric models that exhibit similar texture and complex topology (e.g., trees or plants), the results are often less satisfactory since only a small number of reliable feature correspondences can be established across views and thereby be reconstructed, as shown in Fig. 7. We refer the reviewers to a comprehensive survey [13] that characterizes the cons and pros of SfM.

We adopt the LF camera as the capture apparatus. Light fields are image-based representations that use densely sampled rays as a scene description. They were traditionally acquired using an array of cameras positioned on a 2D regular grid and can now be acquired through a single shot using a light field camera. For example, using an 11MP sensor and 0.1 million microlenses, the Lytro camera can acquire 100 views of the scene. Conceptually, LFs can be viewed as a special form of multi-view geometry. Real light fields, especially the ones captured by the plenoptic cameras, exhibit two unique features: 1) it preserves a regular sampling pattern and 2) the angular sampling is generally much denser. A number of new approaches have been recently developed that effectively exploit these two features [26, 36, 6, 15, 22, 19, 39].

The dense angular sampling, for example, enables tensor analysis that can be further used to optimize the direction/depth field in 2D Epipolar slices [39]. It also enables angular ray statistics analysis where the color distribution of rays can be used separate specular vs. diffuse, occlusion vs. non-occlusion, transparent vs. opaque 3D points. Chen et al. further proposed a bilateral consistency metric based on ray statistics for reliable stereo matching under heavy occlusions [6]. The regular sampling property, on the other hand, provides a useful prior. Heber et al. model depth from LF as a rank minimization problem [15]. Lin et al. proved that the aliasing artifacts in LF rendering preserves symmetry with respect to disparity due to regular sampling [26]. They have further tailored a data term that effectively uses such symmetry to tackle noise and undersampling.

Although each LF image can be used to generate a disparity map, in order to "weave" multiple disparity maps into a 3D model, it is essential to conduct online extrinsic calibration between LF images, a process analogous to SfM but applied to LFs. Brute-force approaches that directly treat each LF view as a camera yield suboptimal results (Fig. 7). Johannsen et al. [20] recently derived the relationship between scene geometry and light field structure under the Plucker ray coordinates for image registration. Their derivation assumes that the two LFs uses identical parameterization whereas we derive a more general case between two arbitrary LFs. Further, we demonstrate how to directly conduct 3D modeling by registering multiple LFs.

2. Pose Registration of Light Fields

In this section, we describe our light field pose estimation method in ray space. To represent a light field, we adopt the two-plane-parametrization (2PP) for its simplicity [24, 11]. In 2PP, each ray is parameterized by its intersections with two parallel planes Π_{uv} and Π_{st}. In this paper, we set Π_{uv} as the camera plane at $z = 0$ and Π_{st} as the image plane at $z = 1$. Hence each ray can be represented as the combination of $x - y$ component of its intersections with Π_{st} and Π_{uv}. To further simplify our derivation, we use the $\sigma = s - u$ and $\tau = t - v$ to parameterize the ray direction as $[\sigma, \tau, 1]$. All rays can be parameterized as a 4-tuple $[\sigma, \tau, u, v]$. The light field coordinate system is set as follows: the origin corresponds to the CoP of the central camera and the x, y axis are concordant with the subaperture image axis.

Given a reference LF $\mathbf{F}$ and a target LF $\mathbf{F}'$, we assume $\mathbf{F}$ corresponds to the world coordinate system and set out to register $\mathbf{F}'$ via rotation R and translation T.

To register the LFs, we first derive ray space mapping from the first LF to the second. In the reference LF, a ray $r[\sigma, \tau, u, v]$ that pass through a 3D point $P = [p^x, p^y, p^z]$ should satisfy:

$$\begin{cases} u + \lambda\sigma & = p^x \\ v + \lambda\tau & = p^y \\ \lambda & = p^z \end{cases} \tag{1}$$

This yields to two linear constraints:

$$\underbrace{\begin{bmatrix} p^z & 0 & 1 & 0 & -p^x \\ 0 & p^z & 0 & 1 & -p^y \end{bmatrix}}_{M} \begin{bmatrix} \sigma \\ \tau \\ u \\ v \\ 1 \end{bmatrix} = 0 \tag{2}$$

Now that consider a ray $r'[\sigma', \tau', u', v']$ in the target LF that also pass the same 3D point P. Since the two LFs are parameterized in their own coordinate systems, we first map r' to r's coordinate system as:

$$w \begin{bmatrix} \sigma^* \\ \tau^* \\ 1 \end{bmatrix} = R \begin{bmatrix} \sigma' \\ \tau' \\ 1 \end{bmatrix}, \quad \begin{bmatrix} \hat{u} \\ \hat{v} \\ \hat{q} \end{bmatrix} = R \begin{bmatrix} u' \\ v' \\ 0 \end{bmatrix} + T \tag{3}$$

Specifically, we can trace the ray from point $[\hat{u}, \hat{v}, \hat{q}]$ along direction $[\sigma^*, \tau^*, 1]$ towards the parametrization plane of r as:

$$\begin{bmatrix} u^* \\ v^* \\ 0 \end{bmatrix} = \begin{bmatrix} \hat{u} \\ \hat{v} \\ \hat{q} \end{bmatrix} - \hat{q} \begin{bmatrix} \sigma^* \\ \tau^* \\ 1 \end{bmatrix} \tag{4}$$

r' is then represented as $[\sigma^*, \tau^*, u^*, v^*]$ in the reference LF.

Transform r' using R, T to the reference LF as $\hat{r}$ where $\hat{r}$ should still satisfy Eqn. 2 as the ray also passes through P:

$$M \cdot \hat{r}(R, T) = 0 \tag{5}$$

We call this constraint ray manifold constraint. To recover this mapping, the brute-force approach is to recover the 3D point P first, i.e., the parameters of M, as shown in Johannsen et al. [20]. In reality, due to low image resolution of the subaperture image and ultra-small baseline, reliably computing the depth of the feature points is very difficult. Instead, we adopt the ray-ray intersection constrains to find the optimal R, T.

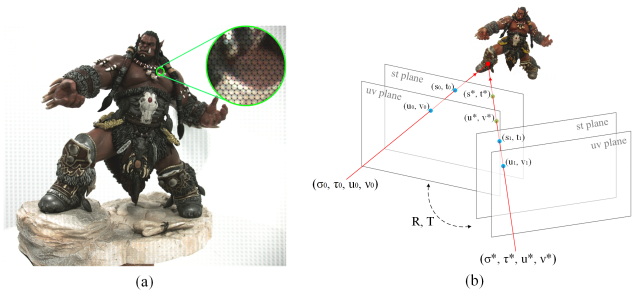

Figure 2. (a) Subaperture images from Lytro Illum Camera. (b) To model LF registration, we map the ray r_1 to r^* in the reference LF and r^*, r satisfy the side operator.

Consider any two rays $r_0[\sigma_0, \tau_0, u_0, v_0]$ and $r_1[\sigma_1, \tau_1, u_1, v_1]$ in different subaperture images pass through the same 3D point should satisfy the side operator $(\sigma_1 - \sigma_0)(v_1 - v_0) = (\tau_1 - \tau_0)(u_1 - u_0)$ [41].

The constraint applies to r and r' as

$$\frac{(\sigma^* - \sigma)(v^* - v)}{(\tau^* - \tau)(u^* - u)} = 1 \tag{6}$$

Notice that $\sigma^*, \tau^*, u^*, v^*$ is a function in R and T. Therefore, if we obtain ray-ray correspondences, we can apply Eg. 6 as the pose constraint.

It is important to note that the ray-ray intersection constraint applies to any 3D point. Therefore, we can combine all pairs of matched feature rays r_i and r_j^* between $\mathbf{F}$ and $\mathbf{F}'$ even if they correspond to different 3D points. Finally, we can form an objective function:

$$E = \sum_{i,j} ||(\sigma_i^* - \sigma_j)(v_i^* - v_j) - (\tau_i^* - \tau_j)(u_i^* - u_j)||_2^2 \tag{7}$$

To find the optimal R, T between $\mathbf{F}$ and $\mathbf{F}'$, we can minimize the energy function via the gradient based optimization method [30]. To robustly establish ray-ray correspondences across LFs, we first detect the SIFT features of each subaperture image within the same LF and establish correspondences between subaperture images within each LF to form groups of features, each potentially corresponding to a 3D point. We prune the outliers using RANSAC and eliminate the groups that do not contain sufficient number of rays. Next, we match groups of ray features across the LFs and again apply RANSAC to remove the outliers. Finally, we use the remaining matching groups for minimizing the energy function 7.

3. Light field Stereo Matching

Almost all depth/disparity estimation methods adopt some kind of smoothness prior for the result. The prior helps to correct false matches and propagate depths to occluded regions where no match could be found. Accurate

prior is crucial to robust matching and occlusion handling. Traditional cost aggregation based depth estimation methods usually assume that within a supporting window the depth is locally constant. The aggregated matching cost is computed as patch matching cost according to aggregation weights based on color similarity. However, this locally constant assumption is only valid for fronto-parallel surfaces. It is inaccurate for curved or slanted surfaces, or around depth boundaries. In practice, the surfaces are always not perfectly planar and do not face right to the camera. Most disparity errors occur around depth boundaries where occlusion happens.

We propose a more flexible prior which is capable of modeling curved and slanted surfaces, and respects depth boundaries. Similar to the guided image filtering [14], we assume that locally, the disparity can be represented as a linear combination of three color channels of the image. Mathematically, the disparity q_i for pixel i in color image I is represented as:

$$q_i = a_k I_i + b_k, \forall i \in \omega_k, \tag{8}$$

where (a_k, b_k) are some linear coefficients assumed to be constant in the supporting local window ω_k.

We verify this assumption by plotting the errors between the ground truth disparity and the represented disparity from three color channels, using the patches from the data set of image disparity pairs [29]. For locally constant assumption, the optimal disparity for a local patch is the mean disparity of the whole patch. From Fig. 3 we can see that, with a small enough supporting window, the representation error is negligible and is much less than that of the locally constant assumption.

Eq. 8 suggests a result that minimize the energy function for the whole disparity map:

$$J(q, a, b) = \sum_{k \in I} \Big(\sum_{i \in \omega_k} (q_i - \sum_c a_k^c I_i^c - b_k)^2 + \epsilon \sum_c (a_k^c)^2 \Big), \tag{9}$$

where the superscript c indicates the color channel, and the second term on a_k is for numerical stability and slightly favors constant disparity with a small weight ϵ. Following the derivation in [23], eliminating (a_k, b_k) by minimizing the cost $J(q, a, b)$, Eq. 9 yields a pure regularization on disparity map q:

$$J(q) = q^T L q, \tag{10}$$

where the Laplacian L is an $N \times N$ matrix, whose (i, j)-th element is

$$\sum_{k|(i,j)\in\omega_k} \Big(\delta_{ij} - \frac{1}{|\omega_k|} \big(1 + (I_i - \mu_k)(\Sigma_k + \frac{\epsilon}{|\omega_k|} I_3)^{-1} (I_j - \mu_k)\big), \tag{11}$$

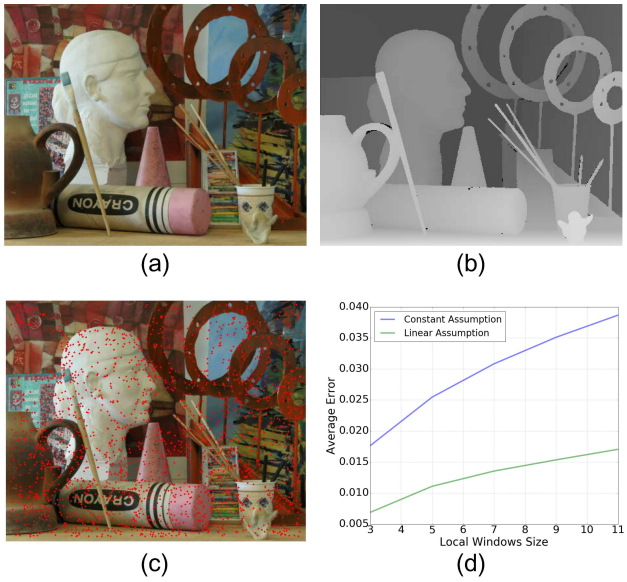

Figure 3. Verification of the linear combination assumption. (a) and (b) are the image and its groundtruth disparity. Red dots in (c) shows the patch locations used for error computation. (d) plots the error curves (blue: locally constant assumption; green: our assumption) vs. patch sizes ($3 \times 3, 5 \times 5, \ldots, 11 \times 11$).

where Σ_k is a 3×3 covariance matrix, μ_k is a 3×1 mean vector of the colors in window ω_k, and I_3 is the 3×3 identity matrix.

The regularization term or prior J Eq. 10 has several benefits that facilitate a high quality disparity estimation: first, surface curvature and orientation usually produce their corresponding shading effects and will eventually be incorporated into the prior by combining color information; second, disparity discontinuities will align with the edges in the referent color image. In other words, the structure information within the color image can be incorporated into the prior.

We integrate this prior into a global formulation for depth estimation from LF data. Let I_r be the center reference light field view (or the center subaperture image when captured using a plenoptic camera), and I_o be the second subaperture image at the 2D position $(o - r)$. We set out to compute the disparity map by minimizing the following energy function:

$$E(q) = \sum_o \sum_i (I_r(i) - I_o(i + q_i * (o - r)))^2 + \lambda q^T L q, \tag{12}$$

where the first term corresponds to data fidelity and λ is a balancing weight. Since the baseline between views in the LF is usually very small, I_o can be expanded as $I_o(i + q_i * (o - r)) \approx I_o(i) + \nabla_{(o-r)} I_o(i) q_i$, where $\nabla_{(o-r)} I_o(i)$ is the gradient along direction $(o - r)$. Then Eq. 12 can be reduced as $E(q) = \sum_o \sum_i (I_r(i) - I_o(i) -$

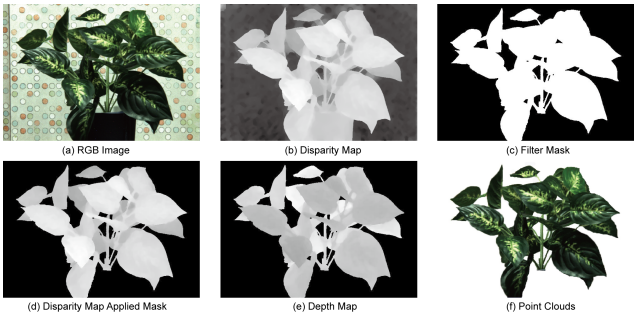

Figure 4. Intermediate Result of Point Cloud Generation. (a) is corresponding RGB image, (b) is disparity map, (c) is filter mask after threshold and Graph Cut, (d) is disparity map after apply mask, (e) is depth map and (f) is point cloud result.

$\nabla_{(o-r)} I_o(i) q_i)^2 + \lambda q^T L q$. It only involves quadratic costs and can be efficiently solved. To further improve its efficiency and robustness, we employ a multi-scale approach that starts from coarse downsampled inputs and recovers the corresponding disparity map of low resolution. Then we linearly scale and upsample the low resolution disparity map to higher resolution, and treat it as an initialization for the disparity estimation of the higher resolution level.[1] This process continues until the original resolution is reached.

4. Bundle Adjustment

After we obtain both the LF poses and their corresponding depth maps acquired at different viewpoints, we aim to fuse the results. The simplest approach would be to directly combine the point clouds. In reality, since both the pose estimation and depth maps can contain errors due to small baselines and low image resolution, such direct fusion can produce noisy results.

We therefore add an additional bundle adjustment stage as commonly used in SfM to simultaneously refine scene geometry and camera pose estimation. Different from traditional SfM method that formulates this optimization as least square minimization, we combine 3D geometry and ray geometry consistency terms.

Recall two rays r in $\mathbf{F}$ and ray r^* in $\mathbf{F}'$ satisfy the side operator:

$$\frac{z-1}{z} = \frac{(s^*-s)}{(u^*-u)} = \frac{(v^*-v)}{(t^*-t)} \tag{13}$$

Since we have computed the depth z, e.g., through stereo matching at each LF, we can map depth z' of P in $\mathbf{F}'$ to $\mathbf{F}$ as z^* and use Eq. 13 as constraint. In addition, the depth estimation should also satisfy the size operator constraints, therefore, putting all these constraints together, we set out to minimize the following energy function for each pair of r and r^*:

$$\begin{aligned}\hat{R}, \hat{T}, \hat{z} \leftarrow \arg\min_{R,T,z} \sum_{r,r^*} & ||(s^*-s) - \frac{\hat{z}-1}{\hat{z}}(u^*-u)||_2 \\ & + ||(t^*-t) - \frac{\hat{z}-1}{\hat{z}}(v^*-v)||_2 \\ & + ||\hat{z}-z|| + ||\hat{z}-z^*||\end{aligned} \tag{14}$$

Notice that this optimization problem is non-linear. We there apply the gradient based optimization method [30] to iteratively refine $\hat{R}, \hat{T}, \hat{z}$. To initialize the optimization process, we first use the estimated pose in Sec. 2 and the fused depth maps as inputs. Table 1 compares our technique vs. the Iterative Closest Point and Johannsen's method [20] based on synthetic data. Our method outperforms both solutions in robustness and reliability.

5. Experiments

5.1. Experiment setup

We evaluate our approach on different types of scenes and additional results (including videos) can be found in the supplementary materials. We choose the Lytro Illum camera as the scanning device. We also use the geometric calibration [4] for intrinsic calibration and subaperture image generation. After calibration, we can map each pixel (k, l) in a sub-view (microlens) image (i, j) onto the light field coordinate $[\sigma, \tau, u, v]$ in the camera coordinate system. Each Lytro LF image is decoded to 5×5 subaperture images (light field views), each at a resolution of 552×383.

To conduct 3D scanning, the simplest approach is to move the LF camera around and capture the object at different poses. Since our approach is largely based on stereo matching, the results would be sensitive to the background. To avoid this problem, we place the object on a rotation table and position our camera on a tripod. The object is position on a patterned background so that background disparity can be robustly computed and then eliminated using the disparity map. The configuration of our setup is shown in Fig. 6. It is important to note that rotating the object maps to a combination of translation and rotation between LFs rather than pure rotation. Further, we do not assume known rotation speed or enforce the object be positioned perfected at the center.

We first compare our light field stereo matching method with the state of arts on the dataset created by Wanner et al [40]. Fig. 5 shows that the outputs of our method contain

[1] With initial disparity q^0, $E(q)$ can be rewritten as $E(\Delta q) = \sum_o \sum_i (I_r(i) - I_o(i + (q_i^0 + \Delta q_i) * (o - r)))^2 + \lambda \Delta q^T L \Delta q$ with $\Delta q = q - q^0$. It improves the result since the accuracy in the expansion of $I_o(i + q_i * (o - r)) \approx I_o(i + q_i^0) + \nabla_{(o-r)} I_o(i + q_i^0) \Delta q_i$ is more accurate.

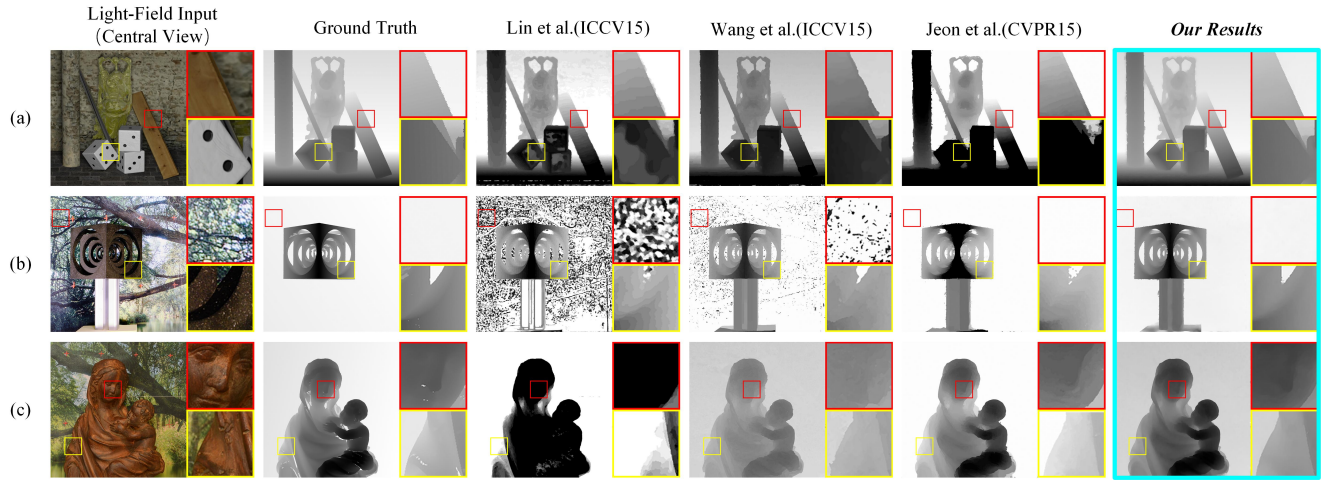

Figure 5. Our stereo matching results compared with state of art methods.

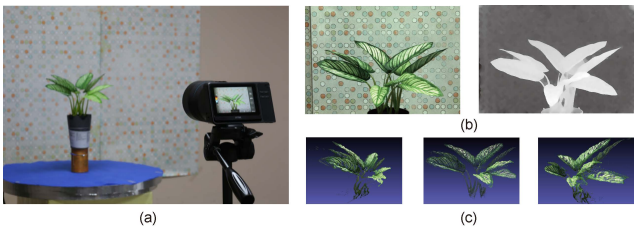

Figure 6. (a) is Our 3D scanning setup, (b) is RGB image and corresponding disparity map, (c) is the point cloud generated by our method viewed from different perspectives.

much less error, especially for regions with less textures and complex static background [26, 38, 19]. The depth maps produced by our method tend to be smooth and visually pleasing, as we use linear assumption Eq. 8.

Fig. 7 shows our results on 5 different models: three plants, a wooden toy house, and a wooden hen. We show our reconstruction results using 1, 2, and 5 light field views, respectively. Using 1 view directly corresponds to the disparity/depth map and we observe that although the results are able to reveal the overall shape of the object, they contain many holes due to heavy occlusions, especially for the 3 plant models. For the house model, our results reveal very fine geometric details as the LF can be used to detect subpixel disparity. Using 2 light fields, the reconstructions are significantly improved and we observe many missing regions have been filled up. With 5 light fields, we are able to obtain high quality 3D models in all 5 cases with nearly all holes filled up.

Next, we compare our technique vs. SfM and Kinect. SfM has shown great success in recent years on both pose estimation and 3D reconstruction. However, SfM generally generates only a sparse set of 3D points which are generally insufficient to produce highly detailed models as the ones shown in the paper. For fairness, we use all light field views (subaperture images) as input to the Agisoft PhotoScan [3]. Fig. 7 and 8 compares the SfM results vs. ours. With 405 images, SfM still generates rather sparse 3D point cloud and brute-force triangulation produces wrong geometry and topology.

Next, we compare our technique with the popular active depth sensor Kinect, along with software KinectFusion for fusing depth maps acquired from different captures. In our experiments, we position the Kinect sensor next to the LF camera so that they capture from approximately the same angle and position. A first look at the Kinect depth map reveals that it contains a large number of holes, mainly due to heavy occlusions. Further, the depth map quality is rather poor due to low spatial and range resolution. In fact, the plant leaves are nearly grouped together into a region of uniform depth and the final fused results exhibit even larger errors since ICP fails due to incorrect depth maps.

Potentially, one can use a higher quality depth sensor (e.g., a ToF with a much higher depth resolution than Kinect). In reality, such solutions require elaborate designs of the ToF unit and cannot be easily adjusted to handle dynamic depth ranges once designed. Further, as a common problem of active depth sensing, they may fail under natural lighting due to interference. These limitations contrast the advantages of passive sensing techniques as ours.

All our experiments were conducted using Matlab implementation on a PC with Intel Core i7-5820K CPU @ 3.30 GHz with 32.0 GB memory. Table 2 shows the running time of our approach on the datasets shown in the paper. For clarity, we separate the running time into three components: LF registration, LF stereo matching and depth map fusion. The running time shown in the Table 2 is the averaged time

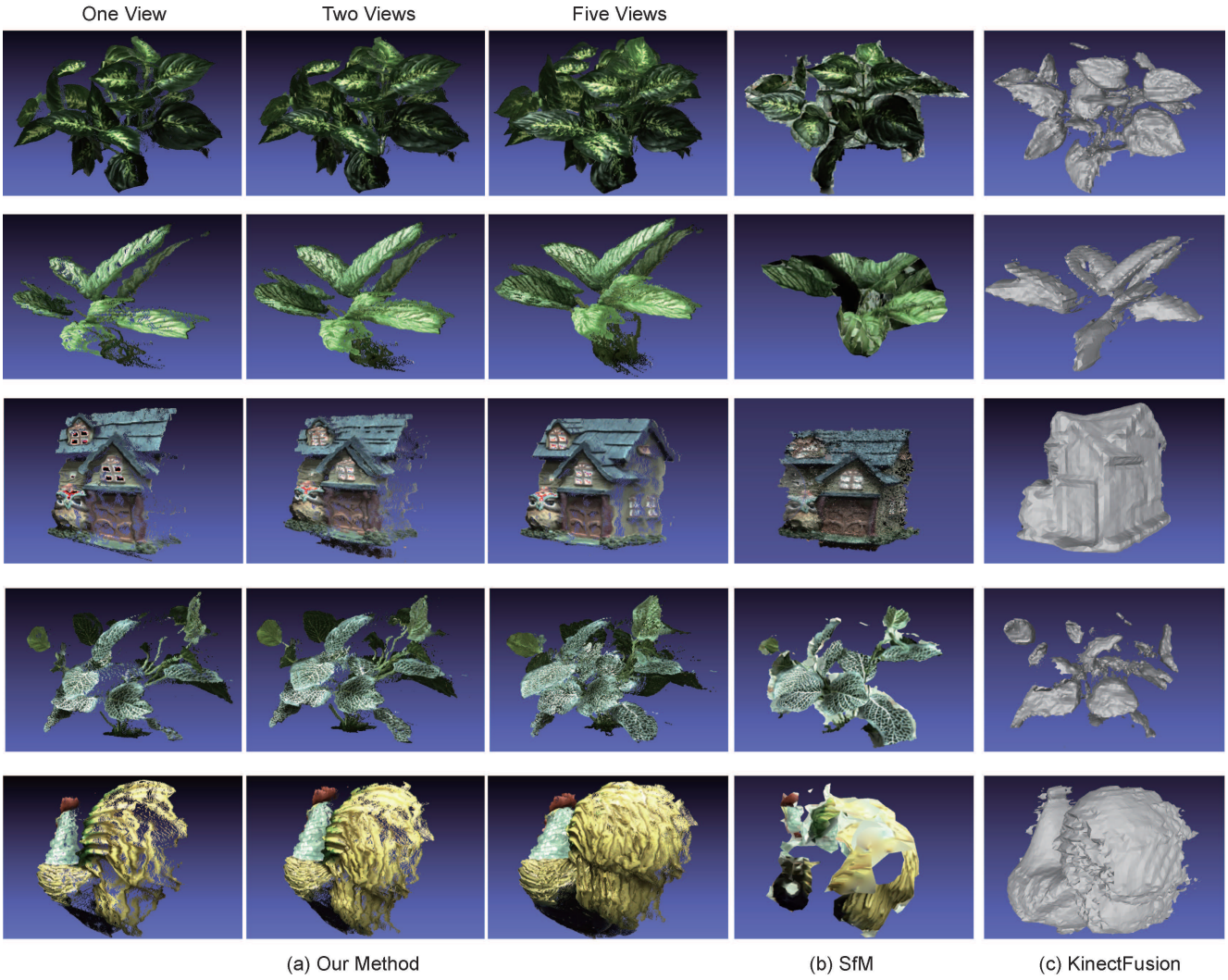

Figure 7. Our results vs. SfM and Kinect. For each model, we illustrate our reconstruction result using 1 view (i.e., the disparity map), 2 views, and 5 views. We also compare with the results generated using SfM and Kinect.

	GT	R: 16^o, T: 62mm				R: 23^o, T: 78mm				R: 30^o, T: 121mm			
		1	2	3	average	1	2	3	average	1	2	3	average
Rot. Err	ICP	0.36	0.16	0.64	0.39	1.15	0.99	1.38	1.17	0.21	0.21	0.22	0.21
	Johannsen [20]	0.07	0.39	0.88	0.45	0.86	0.89	1.23	0.99	0.6	0.29	0.36	0.42
	Ours	0.06	0.03	0.35	0.15	0.8	0.67	0.84	0.77	0.08	0.14	0.12	0.11
Tran.Err	ICP	1.51	0.68	1.66	1.28	2.08	1.16	1.36	1.53	0.35	0.32	0.33	0.33
	Johannsen [20]	0.11	0.6	0.09	0.27	1.2	0.72	1.67	1.2	0.65	0.25	0.32	0.41
	Ours	0.05	0.11	0.19	0.12	1.27	1.07	1.29	1.21	0.13	0.17	0.16	0.15

Table 1. Accuracy of the different methods: we set up virtual scenes and capture LF images from different view points to generate synthetic data to compare our method with others, the number $1, 2, 3$ means different noise. The result shows that our method is more reliable and accurate.

for conducting registration of a pair of LF, stereo matching on one LF, fusing two depth maps. Not surprisingly, the most time consuming component is stereo matching which requires sophisticated optimization.

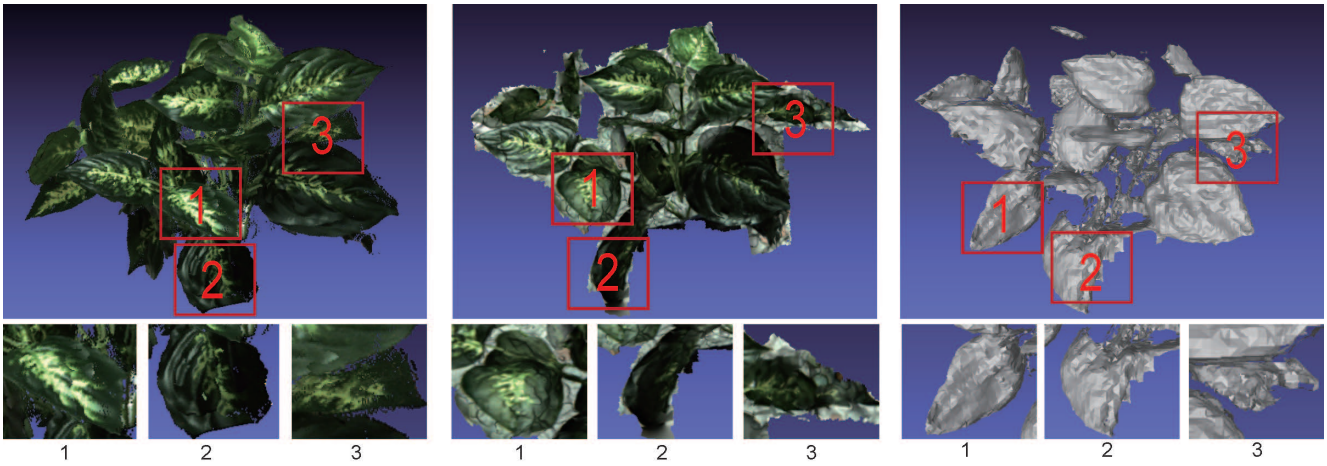

Figure 8. Closeup views of the results using our technique vs. SfM and Kinect fusion. Notice that SfM produces very sparse results whereas Kinect generates many holes to due to occlusion.

	Plant1	Plant2	Plant3	House	Chicken
LF registration	241.578	267.031	298.344	335.563	288.984
LF stereo matching	2039.06	2061.73	2052.63	1988.72	2018.44
Depth map fusion	13.462	3.449	2.381	5.244	3.962
Total run time	2294.1	2332.21	2353.355	2329.527	2311.386

Table 2. Run time statistics (in seconds)

Although our results outperform both classical SfM and active sensing (Kinect), they still exhibit certain artifacts. For example, the ghosting artifacts in the final results (Fig. 7) are caused by inconsistent disparity maps. Even though the feature points are accurately registered thanks to the LF-SfM registration framework, 3D points on textureless regions are slightly misaligned, causing ghosting. This is because stereo matching relies on the smoothness prior to fill in these textureless regions and the results are thus overly smooth and even flat. Although each disparity map appears plausible, it still contains large errors respect to the ground truth on these regions.

6. Conclusions

In this paper, we have developed a 3D object scanning solution that uses the LF camera as a virtual 3D scanner. By exploring the unique light field geometric structures, we have developed a novel LF pose estimation method that exploits ray-ray correspondences to gain accurate extrinsic calibration. We have further developed a new LF stereo matching algorithm that incorporate image gradients into regularization to preserve occlusion boundaries in the depth map. The results are refined via a ray-space bundle adjustment.

Our current approach separates pose calibration and stereo matching estimations. Conceptually, the two processes can be integrated into a unified framework, i.e., finding the optimal pose parameters that yield to the most consistent disparity maps. The problem, however, would require more sophisticated modeling and optimization and is our immediate future work. There are other quite interesting directions. For example, how to couple a light field camera with a ToF camera. Ideally, the ToF can provide a reliable but low quality depth map that can be refined via LF stereo matching. One can further add a high resolution still camera to potentially generate a high resolution light field. As different types of sensors are getting more readily available, hybrid sensing problems as discussed above will surely attract attentions from both the vision and imaging communities.

References

[1] S. Agarwal, Y. Furukawa, N. Snavely, B. Curless, S. M. Seitz, and R. Szeliski. Reconstructing rome. *Computer*, 43(6):0040–47, 2010.

[2] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. Building rome in a day. In *ICCV*, 2009.

[3] Agisoft. Agisoft photoscan. `http://www.agisoft.com/`.

[4] Y. Bok, H.-G. Jeon, and I. S. Kweon. Geometric calibration of micro-lens-based light-field cameras using line features. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2014.

[5] E. Bylow, J. Sturm, C. Kerl, F. Kahl, and D. Cremers. Real-time camera tracking and 3d reconstruction using signed distance functions. In *Robotic Science and System, RSS*, 2013.

[6] C. Chen, H. Lin, Z. Yu, S. Bing Kang, and J. Yu. Light field stereo matching using bilateral statistics of surface cameras. In *CVPR*, 2014.

[7] A. Cohen, T. Sattler, and M. Pollefeys. Merging the unmatchable: Stitching visually disconnected sfm models. In *ICCV*, 2015.

[8] D. Crandall, A. Owens, N. Snavely, and D. Huttenlocher. SfM with MRFs: Discrete-continuous optimization for large-scale structure from motion. *IEEE TPAMI*, 2013.

[9] O. Enqvist, F. Kahl, and C. Olsson. Non-sequential structure from motion. In *Workshop on Omnidirectional Vision, Camera Networks and Non-Classical Cameras*, 2011.

[10] J.-M. Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y.-H. Jen, E. Dunn, B. Clipp, S. Lazebnik, and M. Pollefeys. Building rome on a cloudless day. In *ECCV*, 2010.

[11] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen. The lumigraph. SIGGRAPH '96. ACM, 1996.

[12] M. Gupta, S. K. Nayar, M. Hullin, and J. Martin. Phasor Imaging: A Generalization Of Correlation-Based Time-of-Flight Imaging. *ACM Transactions on Graphics (TOG)*, 2015.

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

[14] K. He, J. Sun, and X. Tang. Guided image filtering. In *ECCV*, 2010.

[15] S. Heber and T. Pock. Shape from light field meets robust PCA, booktitle = ECCV, year = 2014.

[16] F. Heide, L. Xiao, W. Heidrich, and M. B. Hullin. Diffuse mirrors: 3D reconstruction from diffuse indirect illumination using inexpensive time-of-flight sensors. In *CVPR*, page to appear, 2014.

[17] J. Heinly, J. L. Schönberger, E. Dunn, and J.-M. Frahm. Reconstructing the World* in Six Days *(As Captured by the Yahoo 100 Million Image Dataset). In *CVPR*, 2015.

[18] 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. ACM Symposium on User Interface Software and Technology, October 2011.

[19] H.-G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y.-W. Tai, and I. S. Kweon. Accurate depth map estimation from a lenslet light field camera. In *CVPR*, 2015.

[20] O. Johannsen, A. Sulc, and B. Goldluecke. On linear structure from motion for light field cameras. In *ICCV*, 2015.

[21] C. Kerl, M. Souiai, J. Sturm, and D. Cremers. Towards illumination-invariant 3d reconstruction using tof rgb-d cameras. In *International Conference on 3D Vision (3DV)*, 2014.

[22] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. H. Gross. Scene reconstruction from high spatio-angular resolution light fields. In *CVPR*, 2012.

[23] A. Levin, D. Lischinski, and Y. Weiss. A closed form solution to natural image matting. In *CVPR*, 2006.

[24] M. Levoy and P. Hanrahan. Light field rendering. SIGGRAPH '96. ACM, 1996.

[25] X. Li, C. Wu, C. Zach, S. Lazebnik, and J.-M. Frahm. Modeling and recognition of landmark image collections using iconic scene graphs. In *ECCV*, 2008.

[26] H. Lin, C. Chen, S. Bing Kang, and J. Yu. Depth recovery from light field using focal stack symmetry. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[27] D. Martinec and T. Pajdla. Structure from many perspective images with occlusions. In *ECCV*, 2002.

[28] D. Martinec and T. Pajdla. 3d reconstruction by fitting low-rank matrices with missing data. In *CVPR*, 2005.

[29] Middlebury. Middlebury stereo benchmark. `http://vision.middlebury.edu/stereo/data/`.

[30] J. J. Moré. The levenberg-marquardt algorithm: implementation and theory. In *Numerical analysis*, pages 105–116. Springer, 1978.

[31] J. L. Schönberger, F. Radenović, O. Chum, and J.-M. Frahm. From single image query to detailed 3d reconstruction. In *CVPR*, 2015.

[32] Q. Shan, C. Wu, B. Curless, Y. Furukawa, C. Hernandez, and S. Seitz. Accurate geo-registration by ground-to-aerial image matching. In *3DV14*, pages 525–532, 2014.

[33] N. Snavely, I. Simon, M. Goesele, R. Szeliski, and S. M.Seitz. Scene reconstruction and visualization from community photo collections. *Proceedings of the IEEE*, 2010.

[34] F. Steinbruecker, C. Kerl, J. Sturm, and D. Cremers. Large-scale multi-resolution surface reconstruction from rgb-d sequences. In *ICCV*, 2013.

[35] L. Svärm, O. Enqvist, M. Oskarsson, and F. Kahl. Accurate localization and pose estimation for large 3d models. In *CVPR*, 2014.

[36] M. Tao, P. Srinivasa, J. Malik, S. Rusinkiewicz, and R. Ramamoorthi. Depth from shading, defocus, and correspondence using light-field angular coherence. In *CVPR*, 2015.

[37] V. Usenko, J. Engel, J. Stueckler, and D. Cremers. Reconstructing street-scenes in real-time from a driving car. In *Proc. of the Int. Conference on 3D Vision (3DV)*, 2015.

[38] T.-C. Wang, A. Efros, and R. Ramamoorthi. Occlusion-aware depth estimation using light-field cameras. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[39] S. Wanner and B. Goldluecke. Globally consistent depth labeling of 4d light fields. In *CVPR*, 2012.

[40] S. Wanner, S. Meister, and B. Goldlücke. Datasets and benchmarks for densely sampled 4d light fields. In *Vision, Modeling & Visualization*, pages 225–226, 2013.

[41] J. Yu and L. McMillan. Modelling reflections via multiperspective imaging. In *CVPR*, 2005.