This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Large-Scale, Metric Structure from Motion for Unordered Light Fields

Sotiris Nousias
University College London
sotiris.nousias.15@ucl.ac.uk

Manolis Lourakis
FORTH
lourakis@ics.forth.gr

Christos Bergeles
King's College London
christos.bergeles@kcl.ac.uk

Abstract

This paper presents a large scale, metric Structure from Motion (SfM) pipeline for generalised cameras with overlapping fields-of-view, and demonstrates it using Light Field (LF) images. We build on recent developments in algorithms for absolute and relative pose recovery for generalised cameras and couple them with multi-view triangulation in a robust framework that advances the state-of-the-art on 3D reconstruction from LFs in several ways. First, our framework can recover the scale of a scene. Second, it is concerned with unordered sets of LF images, meticulously determining the order in which images should be considered. Third, it can scale to datasets with hundreds of LF images. Finally, it recovers 3D scene structure while abstaining from triangulating using very small baselines. Our approach outperforms the state-of-the-art, as demonstrated by real-world experiments with variable size datasets.

1. Introduction

Structure from Motion (SfM) employs a set of 2D images acquired by a moving camera to estimate the 3D geometry of a scene and the camera motion. The vast majority of relevant research assumes that images have been acquired with ordinary pinhole cameras, which collect converging light rays. Hence, most existing SfM frameworks cannot be directly applied to generalised cameras, *i.e.* cameras which do not share a single centre of projection [9, 37].

In this work, we focus on SfM techniques suitable for generalised, non-central projection cameras with overlapping fields-of-view, such as multi-ocular stereo rigs and multi-camera arrays. Multi-camera arrays have long been used in plenoptic, or Light Field (LF), imaging [25, 52]. Contrary to pinhole cameras that integrate the light rays that intersect each pixel from every direction, a LF image measures the light along each ray reaching the imaging sensor, avoiding angular integration. Thus, a LF captures a 4D slice of the plenoptic function [1] and can be post-processed to support a wide variety of applications [52].

Another approach to plenoptic imaging multiplexes different 2D slices to capture a LF within a single portable camera body [52]. Portable plenoptic cameras typically involve a microlens array placed between the sensor and main lens [32, 36]. The spatial arrangement of the microlenses permits the scene to be captured from multiple viewpoints during a single exposure, essentially trading off spatial resolution on the image sensor with angular resolution. Each independent viewpoint corresponds to a sub-aperture image [52], therefore a single-body plenoptic camera can be considered as a system of multiple pinhole cameras with overlapping fields-of-view. The sub-aperture image whose coordinate frame coincides with that of the LF image will be referred to as the *central* sub-aperture image.

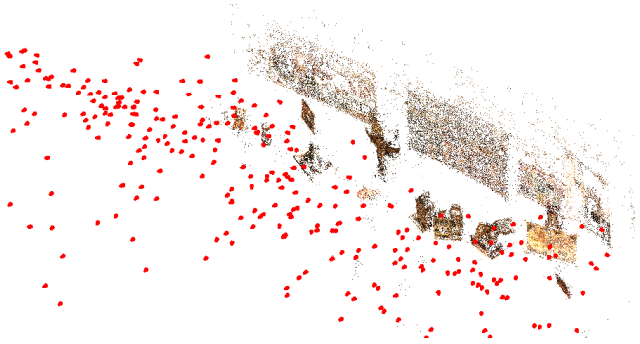

Figure 1. Point cloud and camera poses (indicated with red pyramids) reconstructed with *uLF-SfM* from 303 views of an indoor scene captured by Lytro Illum.

It is also worth noting an emerging trend to include plenoptic imaging capabilities in smartphones, exemplified by Huawei P20 Pro that features three rear cameras, or the Samsung Galaxy A9 that features four. Thus, it is timely to devise bespoke SfM pipelines for plenoptic systems.

We discuss related prior work in Sec. 2. The operation of our pipeline, called *uLF-SfM* and depicted schematically in Fig. 2, is described afterwards. Specifically, we first explain how correspondences are established among LF images (Sec. 3) and how an initial reconstruction is obtained (Sec. 4). Subsequently, we discuss how additional LF images are introduced to the reconstruction and extend it (Sec. 5). Throughout the pipeline, care is taken to cope

with outliers. Sparse bundle adjustment for reconstruction refinement is discussed in Sec. 6. Section 7 compares *uLF-SfM* with the state-of-the-art. We conclude in Sec. 8.

2. Related Work

SfM for pinhole cameras. This research strand has undergone an impressive evolution in recent years and is nowadays capable of reconstructing accurate camera positions and realistic scene models from large, unordered image collections [12, 44], while it can operate in real-time on ordered image sequences [8, 42]. High quality software implementations are also publicly available [41]. Traditional SfM customarily alternates between pose estimation and triangulation (*i.e.*, resection-intersection) steps, an approach we also adopt in our pipeline.

A seemingly straightforward choice for dealing with a set of LF images, is to consider each constituent sub-aperture image as an ordinary image and process it with traditional SfM techniques. Treating sub-aperture images independently, however, creates large image sets. Furthermore, it neglects that their optical centres are regularly arranged on a planar grid and that this arrangement remains constant within LFs. Sub-aperture images also present challenging peculiarities such as tiny baselines and low resolution. Therefore, it is essential to design efficient and robust SfM pipelines specifically for LFs. We note that a medium-sized set of 100 LF images acquired with Lytro Illum contains 2.5K sub-aperture images, 4.4M point features in total, and 20K feature tracks, each giving rise to a 3D point. Sub-aperture image baselines can be as small as 0.5 mm. In the following, the term *LF frame* will imply an LF image that has been acquired by a calibrated single body plenoptic camera and can be decomposed in a collection of sub-aperture images.

State-of-the-art LF-SfM pipelines. The first LF-SfM pipeline was developed by Johannsen *et al.* [15], who assume ordered LF frame sequences. They derive a 2D linear subspace constraint on ray bundles passing through a certain 3D point, which they call the ray manifold constraint. This constraint leads to a linear system on the camera motion parameters. The linear subspace is first recovered within a single LF via a process resembling 3D reconstruction (as also pointed out in [53]). Then, the ray manifold constraints are combined to recover the LF pose with a numerical scheme borrowed from [26]. However, the ultra-small baseline of LF sub-aperture images renders the triangulation of a 3D point from them an ill-conditioned problem. Bundle adjustment (BA) was omitted from [15] but introduced in more recent work [16].

Zhang *et al.* recently presented *P-SfM*, a sequential pipeline for plenoptic SfM that, in addition to points, uses lines and planes as geometric features [53]. They study how ray manifolds associated with such geometric features

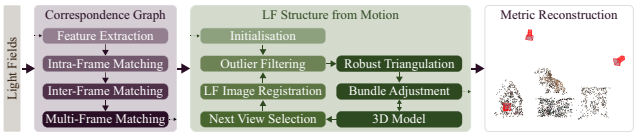

Figure 2. Schematic illustration of our *uLF-SfM* pipeline.

transform under pose changes and exploit these transforms to recover LF camera poses. The point-ray manifold is identical to that in [15]. The complete pipeline, however, is not easily reproducible as its description in the paper lacks important details. For instance, there is no explanation on how the line-ray manifold (*i.e.*, set of rays through a 3D line) is derived from noisy line correspondences, nor how the optimal motion is determined with a non-geometric cost function arising from corresponding manifold constraints. Further, apart from a final refinement, *P-SfM* treats the estimation of motion and structure as two separate problems.

We complete this discussion by noting that although both [15] and [53] perform joint refinement of motion and structure resembling BA, they do not exploit the sparseness of the problem. In other words, they minimise the cumulative reprojection error with standard, dense non-linear optimisation techniques which do not scale well as they incur very large execution times even for a few tens of LF frames [28]. This, in turn, leads to inability to process large datasets: the maximum number of LFs employed in [15], and [53], are limited to 21, and 5, respectively.

Two-view relative motion estimation. To bootstrap SfM for generalised cameras when no information about the motion or scene structure is available, the relative motion between two views must be estimated. Such work was first reported by Pless [37] who substituted image rays for pixels and presented the generalised epipolar constraint (GEC) for a pair of cameras, rigidly attached to a body frame. The GEC decouples motion and scene structure estimation and, contrary to conventional epipolar geometry, enables metric 3D reconstruction as scale can be recovered from prior calibration which determines the true distance between the origins of two projection rays, and, hence, scale [27, 38]. Stewénius *et al.* [45] combined the GEC with Gröbner basis techniques to develop the 6-pt algorithm, which computes the relative pose between two generalised cameras from 6 corresponding rays. The 6-pt algorithm, however, is far from practical as it is computationally expensive and needs to disambiguate up to 64 solutions.

The GEC was also used by Li *et al.* [26] to develop the 17-pt algorithm, a linear, non-minimal technique requiring 17 ray correspondences for estimating the essential matrix. Kneip and Li [19, 21] proposed an iterative solution for generalised relative pose from at least 7 correspondences. This algorithm is based on eigenvalue minimisation and is sensitive to initialisation, being prone to getting trapped in local

minima. Larsson *et al.* [22] study automatically generated polynomial solvers for a wide variety of geometric problems, among which that of generalised 6-pt relative pose.

Absolute motion estimation. Starting with a reconstruction obtained from a set of frames, new ones can be added by estimating their pose with respect to the already reconstructed 3D points. Estimating the pose of a calibrated generalised camera from n given 2D-3D correspondences is known as the generalised Perspective-n-Point (gPnP) problem. Minimal solvers for gPnP require triplets of correspondences between 3D points and the viewing rays of their corresponding image projections [18, 23, 34]. These solvers involve octic polynomials that are solved iteratively. Efficient solvers are proposed in [4, 20].

Contributions. Relying on sparse point features, our work builds on ideas from classical pinhole SfM and develops a pipeline for structure and motion recovery from large numbers of LF frames. Specifically:

- We propose a novel method for large-scale, metric and unordered SfM with generalised cameras having overlapping fields-of-view, and demonstrate it with LF cameras. In particular, we develop techniques for feature matching, outlier filtering and robust triangulation adapted to LFs.
- We demonstrate that our method significantly outperforms current state-of-the-art approaches, both in terms of accuracy and attainable input size. We also show that it is more effective for SfM with LFs than mature, pinhole pipelines such as COLMAP [41].
- We describe an extension of standard sparse bundle adjustment to accommodate LFs.

3. Building the Correspondence Graph

The first stage of our pipeline concerns the construction of a correspondence graph and is composed of four steps. It starts with feature detection and descriptor extraction, followed by intra-frame matching, which identifies sets of features linked to the same 3D point within a certain LF frame. These sets of features are then matched between different LF frame pairs, yielding inter-frame pairwise matches. Finally, pairwise matches are converted to multi-image feature tracks in the multi-frame matching step. In the rest of this section, we assume that individual sub-aperture images have been extracted by a calibration process, *e.g.* [3, 6, 35].

3.1. Feature extraction and intra-frame matching

For every LF frame, this step identifies sets of features that putatively are projections of the same 3D points in the frame's sub-aperture images. To this end, for each sub-aperture image, sparse point features are detected using the difference of Gaussians (DoG) cornerness measure [29], and their RootSIFT descriptors are computed [2]. Using the standard distance ratio test [29], RootSIFT descriptors from the central sub-aperture image are subsequently matched with the RootSIFT descriptors of every other sub-aperture image. Comparing RootSIFT descriptors with the Euclidean distance is equivalent to comparing the original SIFT descriptors with the Hellinger distance, which is more effective for comparing histograms [39]. Sets of matching features that appear in less than a minimum number of sub-aperture images (4 in our implementation) are discarded. Furthermore, to prune mismatched features between the sub-aperture images, we perform filtering which relies on the observation that for matching features, their pixel disparity in both image coordinates is small. Thus, using the median and the median absolute deviation (MAD) as robust estimators of location and scale, we compute the modified Z-score [13] for each image coordinate of the putative intra-frame matches. Then, matches with an absolute Z-score in either coordinate greater than a cutoff (set to 3.0), are discarded as erroneous.

3.2. Inter-frame matching

Inter-frame matching provides sets of features matched between pairs of LF frames. For efficiency, this step employs only the descriptors obtained from the central sub-aperture images. The ratio test [29] combined with left-right consistency checking is used to determine pairwise matches using the central sub-aperture descriptors for all LF frames. Since these descriptors are expected to be less similar across different LF viewpoints, inter-frame matching applies the ratio test with a laxer threshold compared to that for intra-frame matching. To account for mismatches, we fit an essential matrix using the 5-pt algorithm [33] within a RANSAC [7] framework and discard the outliers.

Acknowledging that robust 3D triangulation requires sufficient parallax induced by the translational component of the relative motion between the viewpoints involved, we wish to avoid triangulating with LF frames that are related with a homography. Hence, we also fit a homography to the matches of each central sub-aperture image pair and use Torr's geometric robust information criterion (GRIC) [50] to determine the most likely model (*i.e.* essential matrix or homography). GRIC has been employed in sequential SfM to detect and avoid homographies when selecting keyframes [48]. In our case, frame pairs that are best described with an essential matrix are called *geometrically verified* and are used to perform triangulation in later stages of our pipeline. For example, we consider only geometrically verified pairs for SfM initialisation (cf. Sec. 4.1) and avoid triangulating newly established matches between pairs that have not passed the verification (cf. Sec. 5).

Features from the central sub-aperture images participate in intra-frame matches extracted as described in Sec. 3.1. Thus, matching central descriptors permits the association of intra- and inter-frame features. The extension of pairwise matches to multi-frame ones is addressed next.

3.3. Establishing multi-frame matches

This step identifies lists of matches for the same 3D points across multiple LF frames. To obtain multi-frame feature matches from the pairwise matches determined in Sec. 3.2, we construct an undirected graph that has a vertex for every matched feature of the central sub-aperture images and an edge between the vertices of each pairwise match. Owing to the very large number of features, this graph has a large adjacency matrix that cannot directly fit in memory. Nevertheless, it is very sparse and can hence be economically represented using the compressed sparse row (CSR) storage format that supports efficient random access. Feature tracks, *i.e.* matches across multiple LF frames, correspond to connected components of the graph. These can be determined in time linear to the number of graph vertices and edges by repeatedly performing breadth-first searches (BFS) until all graph vertices have been visited [5]. BFS were preferred over transitive closure computation with the Floyd-Warshall algorithm since the latter has cubic complexity in the number of vertices [5].

Multi-frame matches could in principle be determined with the recent work of Tron *et al*. [51], who use density-based clustering to determine multi-image matches from the modes of a non-parametric density function estimated in feature space. However, [51] does not scale well to the thousands of matches arising from a set of even 100 frames. Specifically, we used the author's implementation[1] to perform the transitive loop closure among the pairwise central sub-aperture feature matches. Despite that the algorithm converged fast for small sets, *e.g.* up to 20 frames, the amount of memory required for 100 frames exceeded that available, causing the algorithm to abort prematurely. Indeed, the authors argue in [51] that their algorithm can handle approximately up to 20K features, but this is only a fraction of the number of features obtained in a set of even 50 frames. An alternative approach is [30], which makes use of the spectral properties of the pairwise matches' permutation matrices. Although [30] scales to hundreds of images, it requires prior knowledge of the number of expected features. Furthermore, its runtime depends on the feature universe size, requiring a couple of minutes for around 50 images [30]. In comparison, our approach is more practical and faster, completing, *e.g.*, in just 40 seconds for 100 images with 250K features.

4. Structure and Motion Initialisation

Given the correspondence graph from the previous section, reconstruction starts by selecting a geometrically verified frame pair, and proceeds to relative pose estimation and robust triangulation using matched features.

[1] https://bitbucket.org/tronroberto/quickshiftmatching

4.1. Choosing the initial pair

As also noted in [40, 41, 49], initialisation is critical in unordered SfM since it may never recover from a poor initial-pair choice. We empirically observed that the scene scale is not estimated accurately using only co-planar correspondences. Therefore, candidates for the initial pair are the geometrically verified pairs obtained using GRIC [50] (see Sec. 3.2). We select the pair with the maximum number of pairwise matches, and fit them with a generalised essential matrix using the 17-pt algorithm with RANSAC. Pairs for which either the inlier ratio is less than a specified threshold or RANSAC exceeds a maximum number of 200 iterations are discarded and the next best initial pair candidate is evaluated. Since this step is critical for scale recovery, we employ a high inlier ratio threshold of 0.7. Having chosen a candidate initial LF frame pair, following subsections describe how to accurately estimate its relative pose using ray-to-ray correspondences and remove outliers.

4.2. From light field features to spatial rays

Each of the pairwise inter-frame correspondences obtained in Section 3.2 consists of sets of features between two LFs. Using the two-plane parameterisation [25], each feature is defined by a quadruple of coordinates (u, v, s, t), where $(u, v) \in \mathbb{R}^2$ encode the pixel location on the sub-aperture image centred on $(s, t) \in \mathbb{Z}^2$ in the LF camera grid. State-of-the-art LF camera calibration techniques [3, 35] provide the calibration matrix $\mathbf{K}$ for each sub-aperture image, which is the same for all sub-aperture images of a micro-lens based LF camera, and map the $s - t$ coordinates to sub-aperture image centres in metric coordinates. Thus, a pixel in a sub-aperture image can be directly mapped to a spatial ray with direction $\mathbf{d} = \mathbf{K}^{-1} \ [u \ v \ 1]^T$.

Each inter-frame feature match can be transformed to a ray correspondence, which gives rise to a constraint based on the GEC [37]. Assuming that there are N inter-frame correspondences, each of which contains l_i intra-frame features from the first LF and m_i from the second, $i = 1...N$, we obtain a total of $\sum_{i=1}^{N} l_i m_i$ ray correspondences. These are the input to the 17-pt algorithm discussed next.

4.3. Relative pose algorithm selection

After selecting the initial pair of frames, we compute their relative pose. To determine the most suitable approach, we compared in simulation several algorithms for estimating the relative pose of generalised cameras: The 17-pt algorithm [26], 17-pt with RANSAC, 17-pt with RANSAC followed by non-linear refinement, 6-pt with RANSAC [45] and the algorithm of Johannsen *et al*. [15] (which also employs RANSAC). We simulated a realistic LF camera, similar to Lytro Illum, with 5×5 sub-aperture views, a baseline of 0.5mm between neighbouring cameras on the grid, and focal length of 600 pixels. For

each noise level, we carried out 200 tests, each of which consists in randomly selecting 30 3D points having a distance between 0.5m and 8m to the world origin, resulting in $25 \times 25 \times 30 = 18750$ ray correspondences. The first LF camera is at the origin of the world frame and aligned with the axes whereas the second is chosen with a random translation in the cube $[-2,\ 2]^3$ and a random rotation from $[-0.5,\ 0.5]$ rad for each axis. To evaluate the algorithms in a challenging scenario, half of the 3D points were chosen so that their disparity in the neighbouring sub-aperture images was less than 0.1 pixels. These points lie at a distance larger than 3m from the world origin. The percentage of outliers was 20%. Regarding implementation, we used the code provided from the author's website[2] for [15], whereas for the rest of the algorithms we relied on OpenGV [17].

Figure 3 summarises the simulation results using the median of the translation and rotation relative pose errors for all algorithms; note that the translation error in the left graph is absolute. On one hand, it is evident that the 17-pt algorithm with RANSAC is the most accurate. Furthermore, we observe that the estimation of translation is accurate for up to two pixels noise. On the other hand, the 17-pt algorithm of Johannsen *et al*. [15] results in larger errors when applied to 3D points with small disparities. Thus, we select the 17-pt algorithm for our pipeline, followed by a refinement step minimising the ray reprojection error of the 17-pt inliers.

4.4. Outlier filtering

The effect of outliers on SfM is detrimental, thus successfully removing them is of utmost importance. Contrary to central cameras where each pair of correspondences contains unique features in each image, in LF frames a certain feature might be included in multiple pairs of corresponding rays, due to the construction of ray correspondences described in Sec. 4.2. An outlier in a set of intra-frame feature matches in one LF frame will result in a set of outliers in the ray correspondences. Thus, simply removing features from the outlier set may result in discarding correct feature matches in addition to mismatched ones, resulting in a sparser point cloud. To avoid this, we remove a feature only if it is labelled as a 17-pt RANSAC outlier more than a certain number of times (4 in our implementation). This procedure is repeated for the features of the other LF frame. If all intra-frame features of either LF are removed, the inter-frame correspondence is eliminated altogether.

4.5. Robust triangulation

Triangulation is performed using all the matched intra-frame features between a pair of LF frames. For a certain 3D point observed in two LF frames by M and N sub-aperture images, the input is a set of $M+N$ projection matrices each of which corresponds to one of the sub-aperture images, and a set of $M+N$ sub-aperture image projections. Although we have removed most mismatched features using either the thresholding on the image coordinates or through RANSAC for pose estimation (Secs. 3.1, 4.4), some may still be present. To safe-guard against erroneous matches but also unstable sub-aperture viewpoint configurations, we perform robust triangulation as follows.

If $M + N \leq 16$, we examine all possible camera pairs, otherwise we select $7(M + N)$ pairs at random. For each camera pair examined, we perform triangulation with the midpoint method [11]. This determines a triangulated 3D point as the midpoint of the shortest line segment (*i.e.*, common perpendicular) between two, possibly skew, back-projected rays. We only consider ray pairs whose common perpendicular (*i.e.*, distance) is shorter than a fraction of the pair's baseline (set to 5%), and the angle of triangulation (defined with the aid of the midpoint) is also above a threshold (set to 5°). These checks avoid triangulation with tiny baselines (*e.g.*, sub-aperture images of the same LF frame) or nearly parallel triangulating rays (*e.g.*, forward motion).

Midpoint triangulation was preferred over DLT [11] due to being more computationally efficient and lending itself to an intuitive, geometrically meaningful check for assessing whether a pairwise triangulation is well-conditioned. We retain the 3D point corresponding to the minimum length common perpendicular and use it to identify outliers by projecting it on every sub-aperture image, calculating the reprojection error and removing projections whose error exceeds a threshold determined with the X84 rejection rule [49]. Finally, the triangulated 3D point is refined using Levenberg-Marquardt to non-linearly minimise its cumulative reprojection error over the inlying sub-aperture images. As an extra precaution, we keep only the points whose average reprojection error is less than 1 pixel.

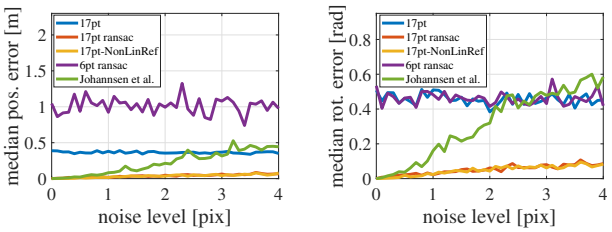

Figure 3. Comparison of relative pose estimation algorithms for generalised cameras with overlapping fields-of-view.

5. Extending the Reconstruction

The initial reconstruction is extended to include more LF frames and 3D points by first selecting the next LF frame to be registered. Then, the new frame is registered, points are estimated via triangulation and trajectories are verified. This process repeats until all frames have been registered.

[2] https://www.cvia.uni-konstanz.de/code-and-datasets

5.1. Next best view selection

Choosing which frame to register next is crucial, as it affects the accuracy of both the pose estimates and the triangulation. Inaccurate pose estimates may lead to spurious 3D points causing the reconstruction to fail. A popular strategy is to simply choose the image which captures most of the scene [43]. Usually, this is also the convergence point of covariance propagation algorithms for view planning [10]. Lepetit *et al*. [24] experimentally showed that the accuracy of absolute solvers is affected by both the number of points and their spatial distribution in the image.

Schönberger [41] proposed a multi-scale approach where an image is discretised into bins for each scale. The next-view candidate set consists of images that already see at least a predefined number of points. For each scale, the number of bins that a point is visible in contributes to the image score. The image with the highest score is selected as the new frame to be registered. In that way, images with better spatially distributed 3D-2D correspondences will result in higher scores and so will be registered first. This approach is practical, easy to implement, and less computationally expensive than covariance propagation.

Considering that the sub-aperture images in a LF frame have large field-of-view overlap, it is computationally more efficient to use only the central sub-aperture image for the next view selection. Therefore, *uLF-SfM* uses the view selection of [41] applied to the central sub-aperture images.

5.2. Light Field frame registration

Provided with an initial reconstruction, new LF frames can be registered to it by solving the generalised PnP problem and determining their absolute pose. The input to the generalised PnP problem is a set of 3D points along with their corresponding LF features. Instead of one-to-one correspondences between reconstructed 3D points and spatial rays, we obtain N point-ray correspondences, where N is the total number of intra-frame feature matches of the particular 3D point. Using a simulation scenario similar to that in Sec. 4.3, we employed synthetic data to compare the performance of several absolute pose estimation algorithms for generalised cameras. Specifically, we compared the following solvers embedded in RANSAC: the minimal solver *gP3P* [18], *gPnP* [18] which is an n-point solver extending *EPnP* [24] to the non-central case, *g1P2R* [4] which employs one point-point and two point-ray correspondences, and the *UPnP* algorithm of [20]. We employed the author's implementation[3] for *g1P2R* and OpenGV [17] for the rest.

Figure 4 illustrates the performance of the algorithms with regard to the median translation and rotation absolute pose errors. Note that *g1P2R* does not perform well since it needs to locally triangulate a point, which in the case of a LF frame is inaccurate as it is performed with a small baseline. *gP3P* outperforms all other algorithms, therefore it is employed with RANSAC in *uLF-SfM* to estimate the pose of the LF to be registered. If the fraction of inliers is less than 0.3, the frame is kept unregistered, to be reconsidered in the future. Otherwise, the pose obtained through RANSAC is further refined using non-linear minimisation of the ray reprojection error pertaining to the inliers [17]. Intra-frame features corresponding to RANSAC outliers are removed from the correspondence graph, as explained in Sec. 4.4.

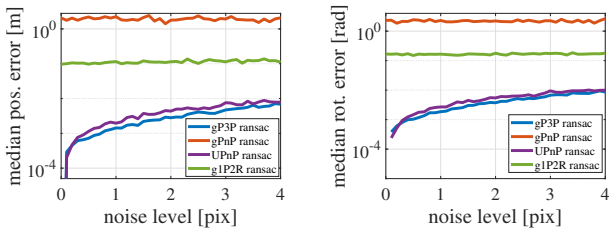

Figure 4. Comparison of absolute pose estimation algorithms. Note the logarithmic scale in the vertical axes.

5.3. Incremental mapping

Once a new frame has been registered, we perform robust triangulation as in Sec. 4.5. Triangulation involves geometrically verified pairs only. We use inter-frame correspondences between the new frame and the already registered ones, as available from the correspondence graph. Reconstructed points whose average reprojection error is above 1 pixel are discarded. Also removed are outlying intra-frame LF features identified by the triangulation step.

Occasionally, the correspondence graph might contain a few outliers, *e.g.* due to erroneous matching. Thus, for each new 3D point, we compute the reprojection error for the intra-frame correspondences of already registered frames and remove features whose error is higher than 1 pixel.

6. Bundle Adjustment for Light Fields

Bundle adjustment (BA) bounds drift by refining the 3D structure and camera poses to minimise the average reprojection error across frames. Although the standard practice in conventional SfM is to frequently perform local BA [31, 49] and resort less frequently to the more expensive global BA, we deviate from this since our robust triangulation already includes a non-linear refinement of the reprojection error. Thus, we periodically employ only global BA, which is invoked either when the 3D point cloud has grown by a certain percentage (15% in our implementation) or when the number of newly registered frames exceeds 10.

In the case of a LF frame consisting of S sub-aperture images, a particular 3D point projects to up to S image projections. Each of these sub-aperture images has a constant

[3] http://people.inf.ethz.ch/fcampose/publications

pose relative to the pose of the LF frame, thus its absolute pose can be directly related to the latter. Therefore, we perform BA by keeping the relative poses of sub-aperture images fixed and minimising the reprojection error with respect to only the LF frame pose and the 3D structure. To achieve this, we have extended the SBA [28] generic BA engine to handle projections in as many as S sub-aperture images per LF frame. Specifically, SBA supports arbitrary camera projection functions where the details of projection as well as the pose, structure and image projection parameters are at its user's discretion. Thus, we use a 2D point for each sub-aperture image a 3D point appears in (up to $2S$ concatenated image coordinates in total) and a 6D rigid transformation to represent LF frame poses. Camera rotations are parameterised with modified Rodrigues parameters [47] and the projection Jacobian is derived analytically.

Consider N 3D points viewed by M LF frames and let $\mathbf{x}_{ijk}$ denote the projection of the i-th point on the k-th sub-aperture image of LF frame j, $\mathbf{a}_j$ the pose of LF frame j, $\mathbf{b}_i$ the coordinates of point i, and $\mathbf{c}_k$ the relative pose of the k-th sub-aperture image. BA for LFs amounts to the following non-linear least squares problem:

$$\min_{\mathbf{a}_j,\mathbf{b}_i} \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{S} v_{ijk}\, d\left(Q(\mathbf{a}_j, \mathbf{b}_i, \mathbf{c}_k)\,,\ \mathbf{x}_{ijk}\right)^2, \qquad (1)$$

where $Q(\mathbf{a}_j, \mathbf{b}_i, \mathbf{c}_k)$ is the predicted projection of point i on the k-th sub-aperture image of LF frame j, $d(.,.)$ is the Euclidean distance, and v_{ijk} is the visibility mask. Notice that the $\mathbf{c}_k$ are not modified during minimisation since they are identical for all LF frames and have been estimated during camera calibration. For efficiency, (1) is minimised exploiting the sparseness of the underlying normal equations [28].

7. Experimental Evaluation

This section presents experiments demonstrating the accuracy of reconstructions recovered by *uLF-SfM* and the correctness of their scale. It also compares the performance of *uLF-SfM* with the state-of-the-art represented by *LF-SfM* [15] and *COLMAP* [41], treating the latter as the gold standard. More results are in the suppl. material. A comparison with *P-SfM* [53] was not possible as its implementation was not made available to us and, as explained in Sec. 2, is not easily reproducible based solely on the publication.

Using a Lytro Illum LF camera, we captured 4 increasingly larger datasets, namely "Octopus", "House", "Toycar", and "Chameleon", from different scenes shown in the first row of Fig. 5. The number of LF frames is given in the second column of Table 1. The LF camera was geometrically calibrated with [3] for use with *uLF-SfM* and *COLMAP*, attaining a reprojection error of 0.20 mm. On the other hand, *LF-SfM* required calibration with [6], which led to an error of 0.22 mm. Sub-aperture images are 552×383 pixels, arranged on 5×5 grids within LF frames. For all pipelines, identical initialisation, feature detection and matching parameters were used.

	# LFs	# Registered		# 3D points		Avg. reprojection error [pix]		
		COLMAP^C	*Ours*	*COLMAP*^{C}	*Ours*	*COLMAP*^{C}	*Ours*	*Ours*^{C}
Octopus	7	7	7	1045	1169	0.20	0.29	0.16
House	16	16	16	1654	1756	0.26	0.31	0.24
Toycar	103	103	103	9917	13512	0.17	0.72	0.23
Chameleon	303	303	303	28281	35597	0.19	0.81	0.21

Table 1. Comparison of *uLF-SfM* with state-of-the-art, classical SfM [41] applied to the central sub-aperture images.

	Transl. Difference		Rot. Difference (°)	
	Ours	*LF-SfM* [15]	*Ours*	*LF-SfM* [15]
Octopus	0.16 ± 0.13	4.39 ± 2.36	0.34 ± 0.01	1.41 ± 0.93
House	0.13 ± 0.08	3.34 ± 1.51	0.18 ± 0.02	3.52 ± 1.87
Toycar	0.11 ± 0.04	(32.41 ± 26.36)	0.67 ± 0.01	(7.03 ± 2.64)
Chameleon	0.12 ± 0.08	–	0.89 ± 0.43	–

Table 2. Differences in poses obtained with *uLF-SfM* and *LF-SfM* [15] using COLMAP [41] as reference. *LF-SfM* registered only 63 frames on "Toycar" and failed on "Chameleon".

Run time. On a PC with an Intel Core i7 CPU at 2.2 GHz and 16 Gb of RAM, *uLF-SfM* required 11 sec, 80 sec, 38 min, and 131 min for the "Octopus", "House", "Toycar", and "Chameleon" datasets, respectively[4]. *LF-SfM* required 25 min and 115 min for "Octopus" and "House", of which 23 min and 112 min were spent for its final non-linear refinement step. *LF-SfM* aborted due to insufficient memory after the first iteration of the final refinement on "Toycar" and completely failed on "Chameleon". On the two smaller datasets ("Octopus" and "House"), *COLMAP* succeeded in registering all sub-aperture images in 114 and 325 sec (note that these are longer than *uLF-SfM*). On "Toycar", *COLMAP* aborted after 5 non-convergent BA steps, having registered 1.3K images in 10 hrs. This is because *COLMAP* treats each sub-aperture image independently, not accounting for the special structure of LF frames. Thus, it is confronted with a fairly large problem involving 2.6K images. Consequently, we ran *COLMAP* on "Toycar" and "Chameleon" using only the central sub-aperture images.

Structure estimation quality. The reconstructions obtained with our *uLF-SfM* are shown in Fig. 5, illustrating that object shapes and boundaries are faithfully recovered. Reconstructed points numbers and mean reprojection errors compared to those of *COLMAP* applied to the central sub-aperture images are shown in Table 1. Fig. 5 also includes reconstructions recovered by *LF-SfM* after providing it with LF frames in the order they were registered by *uLF-SfM* (cf. Sec. 5.1). Since the final refinement step of *LF-SfM* fails when processing more than 20 LF frames, we visualise the output of *LF-SfM* only for the two smaller datasets. Fig. 6 shows that *uLF-SfM* accurately recovers scale by comparing metric reconstructions with measured object dimensions.

Since *uLF-SfM* uses only central images for inter-frame matching, it is fair to compare the density of its recon-

[4] Apart from feature detection and BA, our pipeline is coded in non-optimised Matlab, as is *LF-SfM* [15]; *COLMAP* [41] is written in C++.

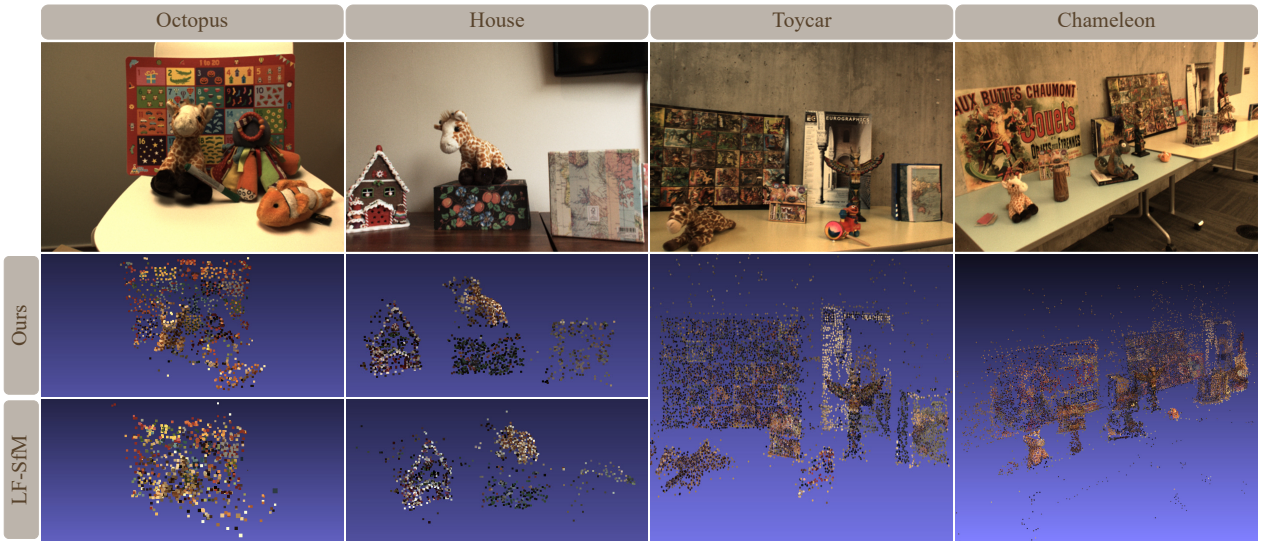

Figure 5. Frames from "Octopus", "House", "Toycar" and "Chameleon" (top); structure obtained with our pipeline (middle left & bottom right); structure obtained with *LF-SfM* [15] for "Octopus" and "House" (bottom left). *LF-SfM* failed for "Toycar" and "Chameleon".

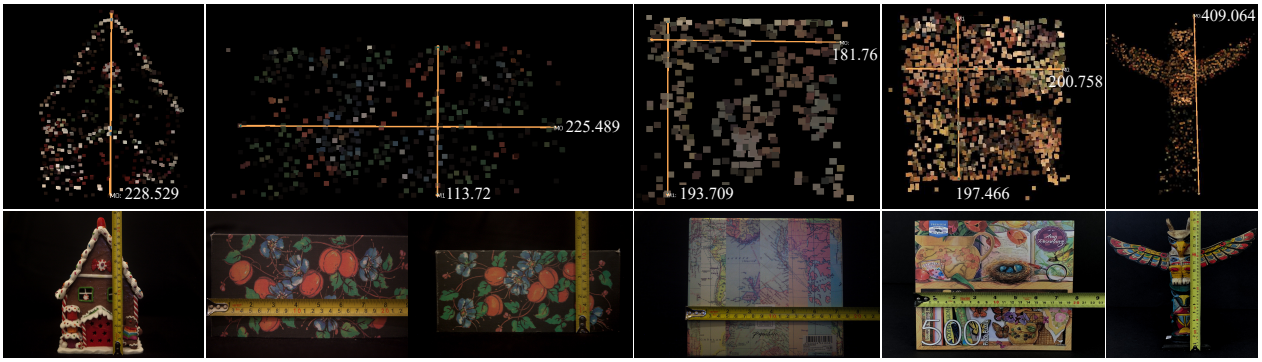

Figure 6. Fidelity of metric reconstruction. The measurements of the reconstructed objects are in mm.

structions with *COLMAP*. In all datasets, *uLF-SfM* provides denser reconstructions with accuracy comparable to *COLMAP* (see Table 1). Our mean reprojection error is slightly higher compared to *COLMAP* when calculated for all sub-aperture images. However, as shown in the last column of Table 1 (superscript C), the reprojection error is similar to *COLMAP*'s, when computed for the central sub-aperture images only. This discrepancy has been also observed in [3, 6] and relates to astigmatism and field curvature [46] that affect microlens-based LF cameras [14].

Pose estimation fidelity. To quantify the performance of camera pose estimation, we compared the output of *uLF-SfM* and *LF-SfM* with *COLMAP*. Since the latter cannot recover scale, we scaled the camera translation vectors for both *uLF-SfM* and *LF-SfM* so that the translation between the two first frames has unit norm. Table 2 presents the differences between the translations and rotations, measured with the L_2-norm and the single axis residual rotation angle, respectively. Clearly, the poses of *uLF-SfM* and *COLMAP* are very similar. *LF-SfM* managed to register only 63 frames on "Toycar" without refinement and failed entirely on "Chameleon".

8. Conclusion

We presented an SfM algorithm capable of dealing with several hundred unordered LF frames. Our pipeline outperforms the state-of-the-art by orders of magnitude in computation time and input size, with accuracy very similar to that attained by [41] applied to central sub-aperture images. Our code and data are available at https://github.com/sotnousias/uLF-SfM.git.

Acknowledgements: This work was supported by an EPSRC Centre for Doctoral Training in Medical Imaging [EP/L016478/1], an AMS Springboard Award [SBF001/1002], an ERC Starting Grant [714562], and EU's H2020 Programme [sustAGE No 826506].

References

[1] Edward H. Adelson and James R. Bergen. The plenoptic function and the elements of early vision. In *Computational Models of Visual Processing*, pages 3–20, 1991.

[2] Relja Arandjelovic and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2911–2918, 2012.

[3] Yunsu Bok, Hae-Gon Jeon, and In So Kweon. Geometric calibration of micro-lens-based light field cameras using line features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(2):287–300, Feb 2017.

[4] Federico Camposeco, Torsten Sattler, and Marc Pollefeys. Minimal solvers for generalized pose and scale estimation from two rays and one point. In *European Conference on Computer Vision (ECCV)*, pages 202–218, 2016.

[5] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. The MIT Press, 3rd edition, 2009.

[6] Donald G. Dansereau, Oscar Pizarro, and Stefan B. Williams. Decoding, calibration and rectification for lenselet-based plenoptic cameras. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1027–1034, Jun 2013.

[7] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981.

[8] Steffen Gauglitz, Chris Sweeney, Jonathan Ventura, Matthew Turk, and Tobias Hollerer. Model estimation and selection towards unconstrained real-time tracking and mapping. *IEEE Transactions on Visualization and Computer Graphics*, 20(6):825–838, June 2014.

[9] Michael D. Grossberg and Shree K. Nayar. The raxel imaging model and ray-based calibration. *Int'l Journal of Computer Vision*, 61(2):119–137, Feb 2005.

[10] Sebastian Haner and Anders Heyden. Covariance propagation and next best view planning for 3d reconstruction. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *European Conference on Computer Vision (ECCV)*, pages 545–556, 2012.

[11] Richard I. Hartley and Peter Sturm. Triangulation. *Computer Vision and Image Understanding*, 68(2):146–157, Nov. 1997.

[12] Jared Heinly, Johannes L. Schönberger, Enrique Dunn, and Jan-Michael Frahm. Reconstructing the world in six days. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3287–3295, June 2015.

[13] Boris Iglewicz and David Hoaglin. Volume 16: How to detect and handle outliers. In *The ASQC Basic References in Quality Control: Statistical Techniques*. 1993.

[14] Hae-Gon Jeon, Jaesik Park, Gyeongmin Choe, Jinsun Park, Yunsu Bok, Yu-Wing Tai, and In So Kweon. Accurate depth map estimation from a lenslet light field camera. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1547–1555, 2015.

[15] Ole Johannsen, Antonin Sulc, and Bastian Goldluecke. On linear structure from motion for light field cameras. In *IEEE International Conference on Computer Vision (ICCV)*, pages 720–728, 2015.

[16] Ole Johannsen, Antonin Sulc, Nico Marniok, and Bastian Goldluecke. Layered scene reconstruction from multiple light field camera views. In *Asian Conference on Computer Vision (ACCV)*, pages 3–18, 2017.

[17] Laurent Kneip and Paul Furgale. OpenGV: A unified and generalized approach to real-time calibrated geometric vision. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–8, May 2014.

[18] Laurent Kneip, Paul Furgale, and Roland Siegwart. Using multi-camera systems in robotics: Efficient solutions to the NPnP problem. In *IEEE International Conference on Robotics and Automation*, pages 3770–3776, May 2013.

[19] Laurent Kneip and Hongdong Li. Efficient computation of relative pose for multi-camera systems. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 446–453, 2014.

[20] Laurent Kneip, Hongdong Li, and Yongduek Seo. UPnP: An optimal O(n) solution to the absolute pose problem with universal applicability. In *European Conference on Computer Vision (ECCV)*, pages 127–142, 2014.

[21] Laurent Kneip, Chris Sweeney, and Richard Hartley. The generalized relative pose and scale problem: View-graph fusion via 2D-2D registration. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, 2016.

[22] Viktor Larsson, Kalle Åström, and Magnus Oskarsson. Efficient solvers for minimal problems by syzygy-based reduction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2383–2392, 2017.

[23] Gim Hee Lee, Bo Li, Marc Pollefeys, and Friedrich Fraundorfer. Minimal solutions for the multi-camera pose estimation problem. *The International Journal of Robotics Research*, 34(7):837–848, 2015.

[24] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. EPnP: An accurate O(n) solution to the PnP problem. *International Journal of Computer Vision*, 81(2):155–166, Jul 2008.

[25] Marc Levoy. Light fields and computational imaging. *Computer*, 39(8):46–55, Aug. 2006.

[26] Hongdong Li, Richard Hartley, and Jae-Hak Kim. A linear approach to motion estimation using generalized camera models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

[27] Manolis Lourakis and Xenophon Zabulis. Accurate scale factor estimation in 3D reconstruction. In *Computer Analysis of Images and Patterns (CAIP)*, pages 498–506, 2013.

[28] Manolis I. A. Lourakis and Antonis A. Argyros. SBA: A software package for generic sparse bundle adjustment. *ACM Trans. Math. Software*, 36(1):1–30, March 2009.

[29] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, Nov 2004.

[30] Eleonora Maset, Federica Arrigoni, and Andrea Fusiello. Practical and efficient multi-view matching. In *IEEE In-*

ternational Conference on Computer Vision (ICCV), pages 4578–4586, Oct 2017.

[31] Etienne Mouragnon, Maxime Lhuillier, Michel Dhome, Fabien Dekeyser, and Patrick Sayd. Real time localization and 3D reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 363–370, 2006.

[32] Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, and Pat Hanrahan. Light field photography with a hand-held plenoptic camera. Technical Report CSTR 2005-02, Stanford University, Apr. 2005.

[33] David Nistér. An efficient solution to the five-point relative pose problem. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(6):756–777, Jun 2004.

[34] David Nistér and Henrik Stewénius. A minimal solution to the generalised 3-point pose problem. *Journal of Mathematical Imaging and Vision*, 27:67–79, Jan 2007.

[35] Sotiris Nousias, Francois Chadebecq, Jonas Pichat, Pearse Keane, Sebastien Ourselin, and Christos Bergeles. Corner-based geometric calibration of multi-focus plenoptic cameras. In *IEEE International Conference on Computer Vision (ICCV)*, pages 957–965, 2017.

[36] Christian Perwaß and Lennart Wietzke. Single lens 3D-camera with extended depth-of-field. In *Proceedings of SPIE*, volume 8291, 2012.

[37] Robert Pless. Using many cameras as one. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 587–593, 2003.

[38] Srikumar Ramalingam, Suresh K. Lodha, and Peter Sturm. A generic structure-from-motion framework. *Comput. Vis. Image Underst.*, 103(3):218–228, Sept. 2006.

[39] Yossi Rubner, Jan Puzicha, Carlo Tomasi, and Joachim M Buhmann. Empirical evaluation of dissimilarity measures for color and texture. *Comput. Vis. Image Underst.*, 84(1):25–43, Oct. 2001.

[40] Frederik Schaffalitzky and Andrew Zisserman. Multi-view matching for unordered image sets, or "How do I organize my holiday snaps?". In *European Conference on Computer Vision (ECCV)*, pages 414–431, 2002.

[41] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, June 2016.

[42] Thomas Schöps, Torsten Sattler, Christian Häne, and Marc Pollefeys. Large-scale outdoor 3D reconstruction on a mobile device. *Computer Vision and Image Understanding*, 157:151–166, 2017.

[43] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Modeling the world from internet photo collections. *International Journal of Computer Vision*, 80(2):189–210, Nov 2008.

[44] Noah Snavely, Ian Simon, Michael Goesele, Richard Szeliski, and Steven M. Seitz. Scene reconstruction and visualization from community photo collections. *Proceedings of the IEEE*, 98(8):1370–1390, Aug 2010.

[45] Henrik Stewénius, David Nistér, Magnus Oskarsson, and Kalle Åström. Solutions to minimal generalized relative pose problems. In *Workshop On Omnidirectional Vision*, 2005.

[46] Huixuan Tang and Kiriakos N. Kutulakos. What does an aberrated photo tell us about the lens and the scene? In *IEEE International Conference on Computational Photography (ICCP)*, pages 1–10, April 2013.

[47] George Terzakis, Manolis Lourakis, and Djamel Ait-Boudaoud. Modified Rodrigues Parameters: An Efficient Representation of Orientation in 3D Vision and Graphics. *J. Math. Imaging Vis.*, 60(3):422–442, Oct 2017.

[48] Thorsten Thormählen, Hellward Broszio, and Axel Weissenfeld. Keyframe selection for camera motion and structure estimation from multiple views. In *European Conference on Computer Vision (ECCV)*, pages 523–535, 2004.

[49] Roberto Toldo, Riccardo Gherardi, Michela Farenzena, and Andrea Fusiello. Hierarchical structure-and-motion recovery from uncalibrated images. *Computer Vision and Image Understanding*, 140:127–143, Nov 2015.

[50] Philip H. S. Torr. An assessment of information criteria for motion model selection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 47–53, 1997.

[51] Roberto Tron, Xiaowei Zhou, Carlos Esteves, and Kostas Daniilidis. Fast multi-image matching via density-based clustering. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4077–4086, Oct 2017.

[52] Gaochang Wu, Belen Masia, Adrian Jarabo, Yuchen Zhang, Liangyong Wang, Qionghai Dai, Tianyou Chai, and Yebin Liu. Light field image processing: An overview. *IEEE J. of Selected Topics in Signal Processing*, 11(7):926–954, Oct 2017.

[53] Yingliang Zhang, Peihong Yu, Wei Yang, Yuanxi Ma, and Jingyi Yu. Ray space features for plenoptic structure-from-motion. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4641–4649, Oct 2017.