Refractive Light-Field Features for Curved Transparent Objects in Structure from Motion

Dorian Tsai[1], Peter Corke[1], Thierry Peynot[1], Donald G. Dansereau[2]

***Abstract*—Curved refractive objects are common in the human environment, and have a complex visual appearance that can cause robotic vision algorithms to fail. Light-field cameras allow us to address this challenge by capturing the view-dependent appearance of such objects in a single exposure. We propose a novel image feature for light fields that detects and describes the patterns of light refracted through curved transparent objects. We derive characteristic points based on these features allowing them to be used in place of conventional 2D features. Using our features, we demonstrate improved structure-from-motion performance in challenging scenes containing refractive objects, including quantitative evaluations that show improved camera pose estimates and 3D reconstructions. Additionally, our methods converge 15-35% more frequently than the state-of-the-art. Our method is a critical step towards allowing robots to operate around refractive objects, with applications in manufacturing, quality assurance, pick-and-place, and domestic robots working with acrylic, glass and other transparent materials.**

***Index Terms*—Computer Vision for Automation; Visual-Based Navigation; Computational Imaging; Light Fields**

I. Introduction

REFRACTIVE OBJECTS are often found in urban settings and industrial applications. However, many robotic vision algorithms find these objects particularly difficult to perceive. Assuming a Lambertian surface—that the appearance of a point on an object does not change with viewpoint—is common, but refractive objects violate this assumption. Their appearance from a particular camera pose is a distorted view of the scene behind them. Thus points on the object's surface can change dramatically in appearance with small changes in viewpoint. Consequently, robotic vision algorithms, including most approaches to structure-from-motion (SfM) and simultaneous localisation and mapping (SLAM), perform poorly around refractive objects. These algorithms yield incorrect camera trajectories and 3D shape estimates and sometimes fail to converge [1]–[3].

We propose a new feature detector for light field (LF) cameras that allows existing feature-based algorithms to operate in scenes dominated by refractive objects. *Image features* are distinct points of interest that can be repeatedly and reliably identified from different viewpoints. These form the basis for a range of robotic perception tasks including visual odometry and 3D reconstruction via SfM and SLAM [4], [5]. When image features are visible through a refractive object they exhibit apparent motion inconsistent with the scene geometry and camera trajectory, as seen in Fig. 1. We refer to these as *refracted image features*. Their view-dependent nature violates assumptions that underpin conventional vision algorithms, which can prevent them from operating as intended [1], [2].

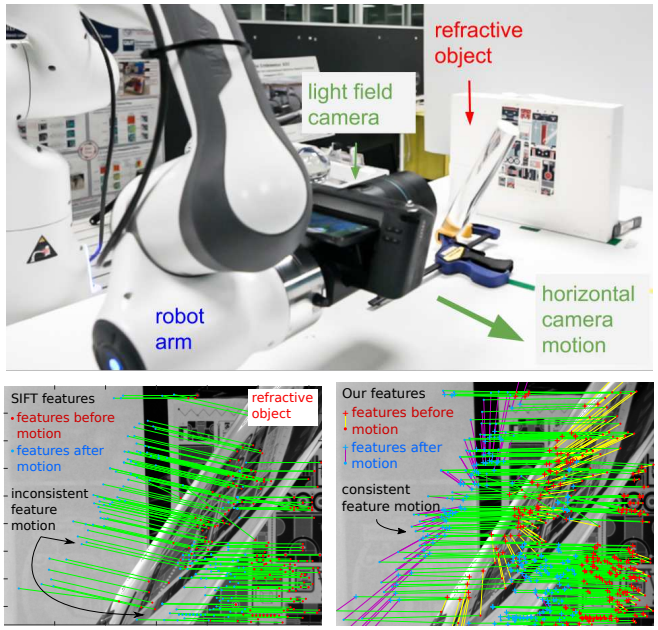


Fig. 1. Comparison of conventional and proposed image features: (top) a robot-mounted LF camera moves horizontally while observing an acrylic cylinder; (left) as seen by the camera's first central view, 2D SIFT features' apparent motion (green) between two frames (from red to blue) is inconsistent across the scene, due to distortion through the refractive object (causing the vertical shift in the refracted image features); (right) the proposed feature exhibits consistent apparent motion (green) between views (only exhibiting horizontal image feature motion), enabling structure-from-motion to operate correctly. The proposed approach yields two characteristic points for refracted features, simplifying to a single point for Lambertian features.

LF cameras capture dense and uniformly-sampled multiple views of the scene in one exposure [2], [6]. A single LF image describes view-dependent effects such as occlusion, specular reflection, and notably, refraction. We exploit this in the context of SfM, toward more reliable operation around refractive objects, such as an eye-in-hand robot grasping a transparent wine glass. We detect refracted image features based on the patterns of light passing through curved refractive

Manuscript received: Feb 24, 2021; Revised: May 19, 2021; Accepted: Jun, 28, 2021.

Editor Cesar Cadena Lerma upon evaluation of the Associate Editor and Reviewers' comments.

This research was partly supported by the Australian Research Council (ARC) Centre of Excellence for Robotic Vision (CE140100016).

[1]D. Tsai, T. Peynot and P. Corke are with the Australian Centre for Robotic Vision, Queensland University of Technology (QUT), Brisbane, Australia {dy.tsai, t.peynot, peter.corke}@qut.edu.au

[2]D. Dansereau is with the Sydney Institute for Robotics and Intelligent Systems, University of Sydney donald.dansereau@sydney.edu.au

Digital Object Identifier (DOI): see top of this page.

objects. From these we extract characteristic points with more consistent apparent motion. We show these can directly enable feature-based algorithms like SfM to operate around refractive objects.

Previous work has shown LF capture offers advantages in detecting reliable features [7] and ignoring refracted features [2]. None has to our knowledge described the detection and use of refracted features for robotic applications. We propose a novel feature detector for refractive objects: the refracted light field feature (RLFF). While conventional features identify patterns in the geometry or texture of objects, the RLFF considers the structure of light refracted by refractive objects, finding characteristic points in the free space between objects.

Our key contributions are:

- we describe a new kind of feature, the RLFF, that exists in the patterns of light refracted through objects;
- we propose efficient methods for detecting and extracting RLFF features from LF imagery, and for describing them in terms of characteristic points that can be employed in place of conventional features like SIFT; and
- we demonstrate that using RLFFs improves SfM performance in scenes dominated by refractive objects, yielding more accurate camera trajectory estimates, 3D reconstructions, and more robust convergence, even in complex scenes where state-of-the-art methods otherwise fail.

To evaluate the RLFFs we captured LF imagery using a Lytro Illum camera mounted on a robotic arm. We captured 218 LFs of 20 challenging scenes containing a variety of refractive and Lambertian objects. The dataset and code associated with this paper can be accessed at *https://tinyurl.com/rlff2021*.

Limitations: This work is inspired chiefly by applications dominated by smooth curved objects like drinking glasses and other manufactured transparent items. Evaluation with flat refractive objects is limited, and we expect adaptation of the method may be required. As with any feature-based method, the presence of texture is required for the algorithm to work. In particular, RLFFs only occur when texture is visible through a refractive object. The approach here will therefore not work with frosted or very complex surfaces through which scene content is not visible. We also assume a geometric ray-based optics approach, so strong defocus effects through highly distorting objects are not considered.

The rest of this paper is organised as follows. We review related work in Section II. In Section III, we discuss the optics of the lens elements that inform the behaviour of our RLFF. Next, the formulation and extraction of our RLFF are described in Section IV. Experimental results using our features in SfM and comparison to traditional 2D SIFT features are presented and discussed in Section V. Lastly, in Section VI, we conclude the paper and explore avenues for future work.

II. Related Work

A variety of approaches for detecting and reconstructing refractive objects using vision have been considered in previous work [1], [8]; however, many require known light sources with bulky configurations that make them impractical for mobile robotics. Other vision-based methods allow for robotic manipulation of refractive objects [9], [10]; however, they rely on having a 3D model of the object a priori. Complete and accurate 3D models and refractive indices of such objects are often difficult, time-consuming and expensive to obtain, or simply not available [1]. When such information is not available, localisation, manipulation and control of and around refractive objects become much harder.

Recently, using LF cameras for SfM has been explored [11], [12]. However, these methods employ conventional 2D image features that occur on 3D surfaces. In this paper, we propose a novel 4D feature that is defined by patterns of light that are not necessarily fixed to the surface of an object.

For LF-specific features, Tosic et al. developed a type of LF-edge feature [13]; however, our interest is in keypoint features, which tend to be more uniquely identifiable, and are more commonly applied to visual servoing and SfM tasks. Tsai et al. developed the first LF image-based visual servoing algorithm that uses a feature combining central-view image coordinates and depth-dependent LF slope [14]. Teixeira et al. used epipolar planar images (EPIs) to detect reliable Lambertian image features [15]. Similarly, Dansereau et al. proposed the Light-Field Feature (LiFF) [7], which focused on detecting and describing reliable Lambertian image features in a scale-invariant manner. However, all of these LF features were designed for Lambertian scenes. We show in the case of LiFF, its performance is affected by the same issues SIFT has with inconsistent apparent feature motion, making it unsuitable for describing refracted image features.

LFs have been considered for refractive object recognition. Maeno et al. proposed an LF distortion feature (LFD), which modelled an object's refraction pattern as image distortion [16]. However, the authors observed poor recognition performance due to specular reflections and changes in camera pose. Xu et al. used the LFD as a basis for refractive object image segmentation [3]. Corresponding image features from all views in the LF were fitted to the single normal of a 4D hyperplane using singular value decomposition (SVD). Tsai et al. extended this work to show that a 3D point manifests as a plane in 4D that has *two* orthogonal normal vectors, which yielded more accurate estimates of how closely an image feature follows the Lambertian model. These estimates helped distinguish more types of refractive objects with a higher rate of detection in order to reject refracted scene content [2].

In this paper, we propose a novel RLFF based on the appearance of background texture through a refractive object. We extend [2] to derive novel methods for detecting, extracting and estimating the 4D structure of an RLFF in the LF. We use the full LF to detect and extract each feature, making maximal use of available information. We employ the proposed RLFF to allow SfM to operate in scenes dominated by refractive objects. Notably, while prior work in [2] focused on detecting and rejecting the refracted scene content, our approach directly uses both the Lambertian and refracted scene content for more reliable camera pose and 3D shape estimation.

III. Describing Refractive Objects

We want to understand the visual appearance of background points imaged through refractive objects, so that we can locally approximate the surface of a refractive object as an astigmatic lens. We begin by investigating the behaviour of light as it travels from the background texture and enters the object. Where the light intersects with the object, the object's local surface curvature determines its path, just as in the case of the surface of a lens. We can describe the entire refractive object as a collection of surface patches, each distorting light based on its local curvature. Fig. 2 illustrates this concept for a toric refractive object, a specific case of an astigmatic refractive object. Here we highlight part of the torus' surface that has a local shape well described by two orthogonal axes of curvature, each with a corresponding radius of curvature. In the general case of an asymmetric surface, the axes of curvature need not be orthogonal, and the result is an astigmatic surface [17]. We can therefore locally approximate the surface of a refractive object as an astigmatic surface.

Other common optical surfaces can be described as special cases of the astigmatic surface. A spherical surface has two identical radii of curvature, and focuses a point source of light to a single point. A cylindrical surface has an infinite radius of curvature in one direction, and focuses a point to a line.

As light leaves the refractive object, it encounters a second surface and is again distorted. As with the first surface, the behaviour is determined by the local curvature of the object. In the general case in which both entrance and exit surfaces are astigmatic, their combined effect is to behave like an astigmatic lens [17].

Fig. 3 depicts the image of a point as seen through a general astigmatic lens. Note there are two focal lines at distinct depths C_1 and C_2. The shape of the bundle of rays passing through the astigmatic lens is known as an astigmatic pencil. Mathematician Jacques Sturm (1838) investigated the properties of the astigmatic pencil, and it is thus also known as Sturm's conoid [17]. The shortest line segment connecting the two focal lines is known as the interval of Sturm, ζ.

As a line segment, ζ can be described by two 3D points. These points have the desirable properties that they can be observed from different positions in the scene, they do not shift significantly as a function of viewpoint, and they can be estimated from a single LF image. They do not shift significantly with viewpoint because the locations of the focal lines are fixed, independent of viewing angle. These will therefore form the basis for our RLFF. Though our discussion is concerned with points, generalisation to more complex textural shapes, and in particular the corners or blobs that make up conventional image features, is straightforward.

IV. Refracted Light Field Features

Whereas a conventional 2D image feature is defined by a single location in space, e.g., the position of a textural corner or centroid of a blob, the RLFF is more complex, defined by ζ, depicted in Fig. 3. This section describes our method for detecting and estimating the RLFF from a single LF exposure.

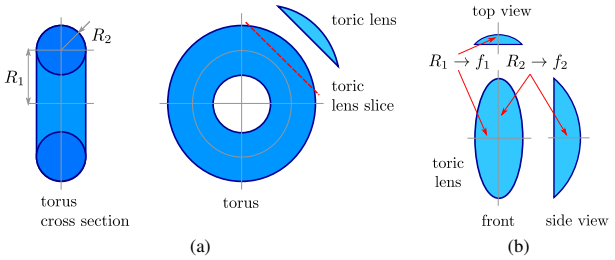


Fig. 2. A complex refractive object can be described in terms of the local curvature of small surface patches. (a) The surface of a torus with radii R_1 and R_2 is sliced (dashed red) to form a toric lens surface; (b) The lens surface is defined by two local radii of curvature, and will focus light to focal lines at two distinct focal depths. For the more general astigmatic surface, the axes of curvature need not be orthogonal.

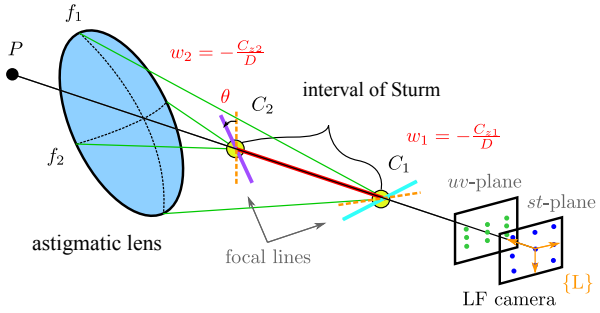


Fig. 3. A point P imaged through an astigmatic lens (blue) forms a distorted pencil of rays. Two lines of focus (purple and cyan) form at depths C_1 and C_2, and the shortest line connecting these is the interval of Sturm (red). The two 3D points C_1 and C_2 that describe the interval of Sturm that comprise our RLFF are shown (yellow/black circles). Observing the scene with an LF camera, we can estimate the orientations of the focal lines and the endpoints of the interval of Sturm. These phenomena are stable with respect to camera pose, and form the basis for our proposed refractive feature.

We parameterise the LF using the relative two-plane parameterisation [6]. A light ray ϕ has coordinates $\phi = [s, t, u, v]^T$, where s, t and u, v describe the points of intersection with two reference planes separated by an arbitrary distance D, s, t are chosen to be further from the scene as depicted in Figs. 3, 4, and ^{T} is the vector transpose. In the relative parameterisation, u, v are expressed relative to s, t. In the sampled LF, we employ the discrete variables i, j, k, l, where i, j select a sub-image, and k, l select a pixel from that image.

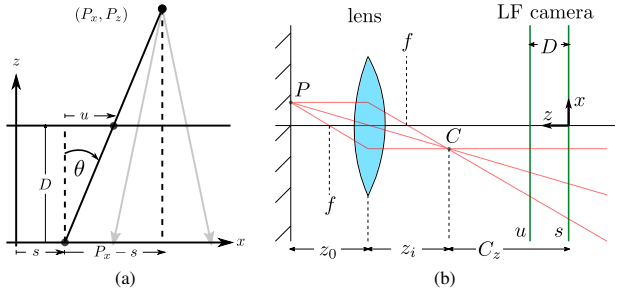


Fig. 4. (a) Geometry of the point-plane correspondence: in a slice of the scene, a Lambertian point P manifests as a line in s, u with slope inversely proportional to P_z [18]; the same holds in the t, v dimensions yielding a plane in the 4D LF (1). (b) Imaging the point through an axis-aligned astigmatic lens, in a single slice of the LF, P manifests as a focal line going into the page at C, yielding a line in s, u-space. In a similar slice (not shown), P appears as a second focal line at a different depth (see Fig. 3), yielding a line at a different slope in t, v-space; the net result is again a plane, but with unequal slopes for the two slices (2).

A. Feature Detector and Descriptor

Our goal is to exploit the local 4D structures of light refracted through objects to define a unique feature detector and descriptor, and in so doing identify and extract the refractive feature parameters. Thus, we leverage existing 2D tools to detect and describe RLFF features. In an approach similar to [11], we apply a SIFT detector to each sub-image of the LF, and match these between views. We use Root SIFT descriptors applied to the central sub-image of the 2D feature for more reliable matching [19], and match between LF sub-images based on the Euclidean distance between descriptors. Only allowing features that match across a minimum number of sub-images allows us to reject spurious detections. In future work, we envision extending a direct LF feature detector like LiFF, which detects features by simultaneously using the entire LF [7].

The detection process yields a set of discrete-space observations $\boldsymbol{n}$, each of the form $\boldsymbol{n_i} = [i, j, k, l]^T$ corresponding to the centroid of the detected SIFT feature k, l in each sub-image i, j. This approach works in the presence of refractive features because, although textural blobs have distorted apparent motion in the LF, their appearance is very similar across the LF sub-images, particularly for small-baseline cameras like the hand-held Lytro employed in this work.

Next, to effectively estimate ζ and extract the feature's parameters, we require the sub-images observing the feature to have sufficient diversity. Views must subtend a 2D space. To evaluate view diversity, we use the coefficient of determination R^2 of a line of best fit from the s, t coordinates of matching views. High R^2 corresponds to a mostly linear set of views, and we empirically determined $R^2 > 0.65$ to be a suitable criterion for rejection.

B. Feature Extraction

In this section we describe the process of estimating RLFF parameters. Feature detection yields a set of discrete space observations $\boldsymbol{n_i}$. For extraction, we first calibrate the camera to yield an LF intrinsic matrix [20] to convert $\boldsymbol{n_i}$ to a continuous-domain ray $\boldsymbol{\phi_i} = [s, t, u, v]^T$.

In the case of a Lambertian scene point, it is well established that the set of observations $\boldsymbol{\phi}$ will lie on a plane in the 4D light field [18]. The geometry for this point-plane correspondence is depicted in Fig. 4a, and is given by

$$\begin{bmatrix} u \\ v \end{bmatrix} = \left(-\frac{D}{P_z}\right) \begin{bmatrix} s - P_x \\ t - P_y \end{bmatrix}, \tag{1}$$

where $P = (P_x, P_y, P_z)$ is the position for a Lambertian point P with respect to the LF camera's nodal point. This can be interpreted as the intersection of two hyperplanes, where the slopes of the hyperplanes in the epipolar plane dimensions s, u and t, v are identical, given by $-D/P_z$.

We now generalise the point-plane correspondence by introducing an astigmatic lens between the camera and the Lambertian point, as depicted in Fig. 4b. We initially assume a thin toric lens with axes aligned with the s and t axes. This yields two orthogonal lines of focus at depths P_{z1}, P_{z2}. In s, u the behaviour is similar to that of a Lambertian point at depth P_{z1}, whereas in t, v the rays appear to emerge from a point at depth P_{z2}. We can write this more formally as

$$\begin{bmatrix} u \\ v \end{bmatrix} = S \begin{bmatrix} s - P_x \\ t - P_y \end{bmatrix}, \quad S = \begin{bmatrix} -D/P_{z1} & 0 \\ 0 & -D/P_{z2} \end{bmatrix}, \tag{2}$$

where P_{z1} and P_{z2} are apparent depths corresponding to the extremities of ζ, and S is a diagonal matrix of the two slopes for the RLFF.

For the more general case of non-camera-aligned lines of focus, i.e., the case of a general astigmatic lens, we apply a transformation H of the form

$$\begin{bmatrix} u \\ v \end{bmatrix} = H \begin{bmatrix} s - P_x \\ t - P_y \end{bmatrix}, \quad H = VSV^{-1}, \tag{3}$$

where V is a rotation matrix for the special case of a rotated toric lenses, and the concatenation of two potentially non-orthogonal axes $[V_1, V_2]$ for a general astigmatic lens. This transformation assumes ζ is close to parallel with the principal axis of the camera. For very wide-Field-of-View (FOV) LF cameras, features near the edges of images can violate this assumption, resulting in poor feature detection rates. However, for the Lytro cameras and the imaging scenarios considered here, this assumption holds well across the entire FOV.

Finally, separating the translation terms yields the generalised point-plane correspondence for points imaged through astigmatic lenses

$$\begin{bmatrix} u \\ v \end{bmatrix} = H \begin{bmatrix} s \\ t \end{bmatrix} + X, \quad X = -H \begin{bmatrix} P_x \\ P_y \end{bmatrix}. \tag{4}$$

Note that this general form also describes the Lambertian case, for which the depth P_{z1} equals P_{z2} and (4) reduces to (1).

C. Estimating Feature Parameters from Observations

From (4), observations of a scene point take the form

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} h_1 & h_2 & x_1 \\ h_3 & h_4 & x_2 \end{bmatrix} \begin{bmatrix} s \\ t \\ 1 \end{bmatrix}, \tag{5}$$

where $h_1 \dots h_4$ are the elements of H, and $X = [x_1, x_2]^T$. Given a set of $[s, t, u, v]^T$ observations for a single feature, we find the least squares solution to (5), directly yielding estimates $\hat{H}$ and $\hat{X}$.

Eigenvalue decomposition of $\hat{H}$ allows us to estimate V and S following (3). Equating terms from (2)–(4) allows us to solve for P_x, P_y, P_{z1}, P_{z2} as well as the directions of the axes V_1 and V_2, θ_1 and θ_2, respectively. These are the parameters of the focal lines caused by an astigmatic lens and ζ. These six parameters also compose our definition of the *Refracted Light Field Feature (RLFF)*:

$$RLFF = \left[P_x, P_y, P_{z1}, P_{z2}, \theta_1, \theta_2\right]^T. \tag{6}$$

Though physically realisable H matrices are symmetric, the estimate $\hat{H}$ may not be. Asymmetric matrices can result in imaginary eigenvalues. Prior to eigendecomposition we force $\hat{H}$ to be symmetric, $\hat{H}_S = (\hat{H} + \hat{H}^T)/2$. We note that using a constrained least squares estimator would likely yield improved noise performance. Eigenvalue decomposition on $\hat{H}_s$ yields V and S. When estimating the offsets, P_x and P_y,

we use the reconstructed $H_R = VSV^{-1}$ rather than $\hat{H}$, as this improves noise performance. The residual between $\hat{H}$ and H_R forms a convenient indicator for outlier features not well described by our assumptions.

An example of a feature extracted from captured LF imagery is shown in Fig. 5. The sub-image views are shown in blue, the set of s, t, u, v observations across the LF are shown in gray, and the estimated focal lines and ζ are shown as coloured line segments.

D. Driving SfM

The proposed feature definition comprises two infinite lines of focus and the interval between them. We anticipate constructing robotic vision algorithms that work directly off these features. However, existing systems like the popular SfM solution COLMAP [5] accept only 2D image features like SIFT. Thus, we propose two mechanisms for driving existing vision systems with the proposed feature for comparison with monocular- and stereo-based SfM approaches.

First, by projecting the 3D endpoints of ζ into the central view of the camera, they are reduced to 2D image features. This *RLFF mono* has the drawback of discarding all 3D information associated with the feature. Second, we therefore propose a variation of this approach in which we instead project the same endpoints into two separate LF views, separated by a baseline similar to that of the LF camera. This *RLFF stereo* preserves most of the 3D information of the feature, discarding only the orientations of the focal lines. This also preserves our knowledge of the baseline over which depth is being estimated, an important detail for reconstruction algorithms that consider uncertainty associated with short-baseline depth estimates. We evaluate both variations of this approach in Section V.

Note that our approach simultaneously detects both Lambertian and refracted features, and passes them all into SfM. In theory, Lambertian scene points yield a zero-length ζ, and so appear as a single point rather than as a pair. This is visible in our results, e.g., see the refracted and background Lambertian features in Fig. 1. Many applications will benefit by distinguishing Lambertian and refracted features, as Lambertian points generally correspond to an object's surface, while refracted features are images that exist in free space. Prior work has distinguished refracted features on the basis of slope differences in 3D subsets of the LF [2]. The RLFF effectively measures how Lambertian an image feature is in a more complete way, as the entire LF is employed. ζ has zero length for Lambertian scene points, and only takes on finite extent for refracted scene content. Otsu's method can be used to determine a suitable threshold when considering a histogram of ζ distances. However, some Lambertian features have a non-zero ζ in Fig. 1. We attribute this to the RLFF having more degrees of freedom than a simple Lambertian feature. The RLFF may be more susceptible to modelling errors, such as miscalibration. Thus, replacing SIFT for RLFF in Lambertian scenes, which is not proposed in this paper, requires deeper investigation into this phenomenon.

Note also that in the case of refracted features, we are passing two characteristic points to COLMAP as though they were separate 2D features. We require a descriptor for each of these to allow feature matching, but wish to disallow matching of front and back characteristic points C_1 and C_2 between frames. To this end we propose three approaches. First, the descriptor can be modified to reflect which focal line it belongs to, front or back, e.g., through addition of a bias term, scaling factor, of by raising to some power. Then matches could not occur between front and back features as their descriptors differ substantially. Second, one can perform matching externally to the SfM tool while keeping track of which characteristic point each feature corresponds to. Then potential matches between front and back points are simply not evaluated. Finally, the approach taken in this work is to employ identical descriptors for the two points, yielding extraneous putative matches, and relying on outlier detection to reject these on the basis of epipolar geometry.

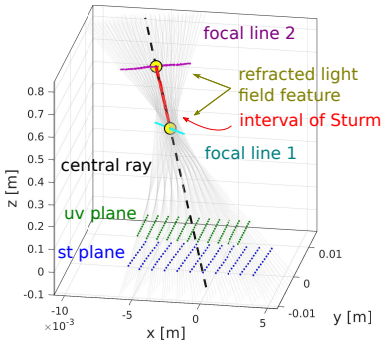


Fig. 5. An RLFF extracted from Illum imagery. The st sub-images (blue) and uv observations (green) are shown from distance $D = 0.1m$ from the st-plane. The rays projected by $stuv$ (grey) pass through both the first and second focal lines (cyan and magenta, respectively). The central ray of the feature is shown (dashed black). ζ (red), and the two 3D points (yellow) define our RLFF.

V. Evaluation

To evaluate the proposed feature we mounted an Illum LF camera on a Franka Emika Panda robot arm. The experimental setup is illustrated in Fig. 1. The LF camera was calibrated using the LF Toolbox [20]. To reduce the effect of extreme lens distortion, we cropped the 15×15 LF to a 13×13 array of sub-images. The robot arm was used for repeatability and ground truth trajectories with sub-millimeter positional accuracy. We used a variety of refractive objects, such as a water bottle, glass sphere, glass cups filled with water, combined with a variety of textured backgrounds.

A. Structure from Motion

To evaluate the RLFF, we used the popular COLMAP SfM implementation [5]. Following the procedure outlined in Section IV-D, we converted each detected RLFF to a pair of 2D image features, as seen from the central view of the LF camera (the middle view of a LF if we consider the LF as a grid of views). We also evaluated the alternative approach of projecting the feature into a stereo pair of LF views, preserving depth information. For comparison, we evaluated SIFT features as seen in the central LF view. To better understand the impact of projecting features into stereo pairs,

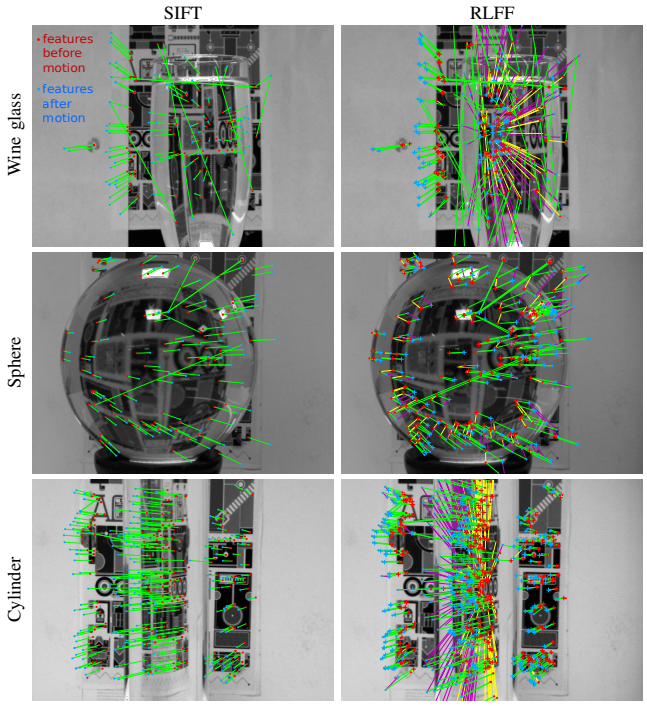


Fig. 6. Three example scenes for which COLMAP fails when using SIFT features (left) but succeeds when using the proposed RLFF. These examples also show different types of camera motion: the wine glass has forwards motion, the sphere has horizontal motion, and the cylinder has both. Only the first central view image is shown with red features with green feature motion to corresponding blue features of a subsequent image. RLFF show some inconsistent apparent motion near the edges of the refractive objects; however their motion is more consistent than SIFT, allowing COLMAP to converge.

we also compared with SIFT similarly projected into a pair of LF views, with the same baseline as for the RLFF test.

We also considered LiFF features. In an evaluation similar to that shown in Fig. 1, Fig. 7 shows the LiFF feature matches between two poses. Fewer refracted features were detected than SIFT, but LiFF demonstrated the same inconsistent feature motion exhibited by SIFT, making it unsuitable for use around refractive objects.

In all, we compared four methods using COLMAP, covering the proposed RLFF and SIFT for both monocular and stereo views: RLFF mono, RLFF stereo, SIFT mono and SIFT stereo.

We evaluated performance using COLMAP's sparse SfM. SfM was not able to reconstruct all scenes for all types of objects, as its solution did not always converge. This was especially problematic for scenes dominated by refracted content. Following [7] and [5], we evaluated performance in terms of the percentage of scenes for which COLMAP converged, the number of image features per image, putative image feature matches per image, inlier matches per image during SfM, putative match ratio, mean number of 3D points, track length, precision and matching score. The *putative match ratio* is the proportion of detected image features that yielded putative matches. The *mean number of 3D points* in the reconstructed models serves as an indicator of how many features were stable enough to be included into the model. The *track length* is the mean number of camera poses over which a feature was successfully tracked. The *precision* is the proportion of

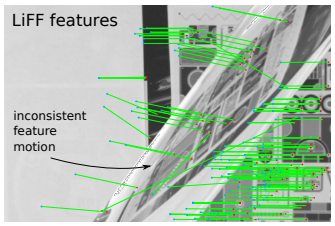


Fig. 7. Like the SIFT features shown in Fig. 1, LiFF shows inconsistent feature motion around refractive objects, resulting in poor reconstructions.

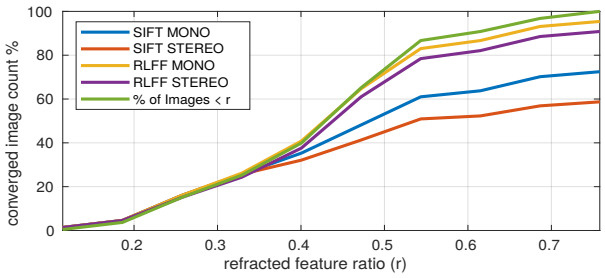


Fig. 8. Cumulative histogram of imagery successfully incorporated into the COLMAP SfM model (as a percent of 218 LFs) versus the refracted feature ratio r. For $r > 0.4$, imagery becomes increasingly challenging as more of image features become refracted and fewer images are being incorporated by each SfM solution, especially for SIFT. An ideal method would incorporate all images with a percent of images less than r (green), ending with a value of 100%, being able to incorporate all of the refracted scene imagery. Our RLFF-based SfM reconstructs more images in total, showing the strongest advantage with more refracted content.

putative image feature matches that yielded inlier matches. The *matching score* is the proportion of image features that yielded inlier matches. Note that for the RLFF features, we divided the number of image features, putative matches, inlier matches and 3D points by two because a single RLFF is represented by a pair of points, the extents of ζ. Similarly, we divided the track length by two for both stereo approaches, as twice the number of images were considered during the motion sequences.

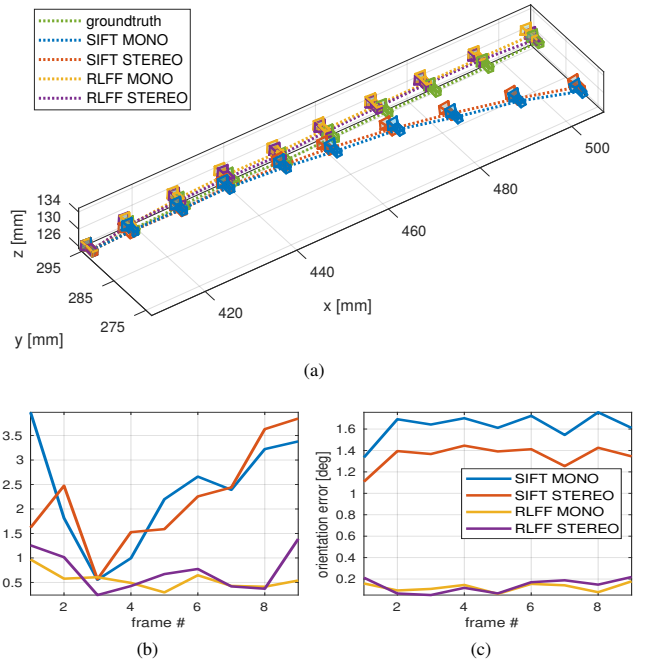
We collected 20 sequences consisting of 10 to 20 camera poses each, covering a variety of motion trajectories and a diverse set of scene content including spherical, cylindrical, and general astigmatic elements. The experimental setup shown in Fig. 1. The dataset contains 218 LFs in total. Example images highlighting the differences between SIFT and RLFF performance are shown in Fig. 6. In all of these scenes COLMAP was unable to converge while using either version of SIFT, but succeeded when using our RLFF.

To understand how refractive objects affect SfM performance, we evaluated the extent to which each LF is dominated by refractive features. We took the extent of ζ for each feature as an indication of how Lambertian it is. Lambertian scene points evaluated by our feature extractor yield a ζ of zero length, or equivalently, correspond to a plane with equal slopes in horizontal and vertical dimensions following (1).

A limitation of this approach is that it does not work for spherical lenses, which produce a well-formed image that behaves identically to Lambertian scene content. Human viewers are also susceptible to this, and it is the basis for some display technologies that produce images floating in air. Features refracted through spherical objects are thus not well

TABLE I
EVALUATING THE PROPOSED RLFF AND SIFT IN SFM, BOTH IN MONOCULAR AND STEREO MODES. MORE SCENES CONVERGE USING THE PROPOSED METHOD, AND IT OUTPERFORMS THE STATE-OF-THE-ART IN ALL MEASURES EXCEPT PRECISION.

Methods	# LFs converged	% Pass	# LFs common	Image Features / Img	Putative Matches / Img	Inlier Matches / Img	Putative Match Ratio	3D Points	Track Length	Precision	Matching Score
SIFT MONO	158	0.75	118	372	122	112	0.369	431	5.11	0.916	0.340
SIFT STEREO	128	0.6	118	370	131	122	0.369	524	4.16	**0.922**	0.342
RLFF MONO	208	**0.95**	118	321	155	134	0.448	437	**5.27**	0.842	0.385
RLFF STEREO	198	0.9	118	321	**168**	**147**	**0.487**	**684**	4.04	0.863	**0.426**



(a) (b) (c)

Fig. 9. Comparing pose accuracy for horizontal camera motion, as in Fig. 1: (a) Estimated trajectories show both proposed stereo and mono variants of RLFF outperforming both variants of SIFT; (b) Relative, instantaneous translational and (c) rotational error show the proposed methods outperforming SIFT in all cases.

detected by our method. They do, however, make excellent points for use in SfM, so while we do not detect them as being refracted, we do make use of them in the SfM solution.

We plotted the number of images correctly incorporated into an SfM solution by COLMAP. By sorting images on the horizontal axis according to the number of features identified as refractive over the total number of features per image, *the refracted feature ratio*, r, we obtain the cumulative histogram shown in Fig. 8. This shows both variants of the proposed method enabled SfM to succeed with almost all images, while the SIFT-based methods failed to converge for a significant number. Importantly, the difference in performance is due chiefly to scenes dominated by refracted content.

Finally, occluding edges remain an issue for RLFF, which exhibits some inconsistent apparent motion near the edges of the refractive objects in Fig. 6. This is likely because a support region at an occlusion boundary contains multiple depths, causing inconsistent depth estimates, though recent work focused on distinguishing these edges is promising [21].

B. 3D Reconstruction Performance

We ran COLMAP across all of the collected scenes, and summarise the results in Table I. The statistics show that RLFF methods have a higher proportion of putative and inlier matches, the number of 3D points in the reconstructions, the track length, and the overall matching score. Our methods converged 15-35% more frequently than their SIFT counterparts, and performed better in almost all measures except precision. The stereo version of RLFF generally showed higher performance, though the monocular version allowed more scenes to converge. We explain weaker precision performance of RLFF by noting more putative matches due to doubling the number of 2D image features, which subsequently quadrupled the number of potential 2D feature matches. In this preliminary work, descriptors were only duplicated for the RLFF projections to 2D image features. For refracted image features, the distance between feature projections was sufficiently large to discriminate. However, for Lambertian features where the projections are coincident, we increased the putative matches without increasing the inliers. This resulted in lower precision. To prevent this issue, we can adopt the strategies proposed earlier in Sec. IV-D for future work. These results showed that the proposed RLFF feature allowed COLMAP to operate in many scenes for which it could not previously operate.

C. Camera Trajectory Estimation

We also evaluated the accuracy of camera trajectories estimated by SfM. Ground truth was available by virtue of our use of a robotic arm to carry out the camera trajectories. A comparison of camera trajectory estimates and their corresponding translational and rotation error are shown in Fig. 9. For this scene, the proposed RLFF feature showed substantially higher performance in both rotational and translational error, yielding a much more accurate trajectory estimate.

Table II summarises SfM pose accuracy over all the test scenes, shown in terms of the instantaneous pose error $e_i = |(p_{s,i} - p_{s,i-1}) - (p_{g,i} - p_{g,i-1})|$ for image i in translation (e_{tr}[mm]) and rotation (e_{rot}[deg]), averaged over the entire sequence of images, where $p_{s,i}$ and $p_{g,i}$ are the estimated pose and groundtruth pose for image i, respectively. We compared the same four variants of SIFT and RLFF-based approaches as before. We separated scenes for which all methods converged, condensing 11 scenes into averages in the top half of the table, but noting that these correspond to the easiest scenes with a median r of 0.35. In contrast, the more challenging scenes where a refractive object dominated the image, had a median

TABLE II
COMPARISON OF SFM POSE ERROR. RLFF ALLOWED ALMOST ALL SEQUENCES TO CONVERGE, WHILE SIFT DID NOT. RLFF OUTPERFORMED SIFT BOTH ON EASIER SEQUENCES WHERE ALL METHODS CONVERGED (TOP), AND IN MORE CHALLENGING SEQUENCES (BOTTOM), DIFFERENTIATED VIA THEIR MEAN REFRACTED FEATURE RATIO, r.

			SIFT MONO		SIFT STEREO		RLFF MONO		RLFF STEREO	
Seq.	#LFs	r	e_{tr}	e_{rot}	e_{tr}	e_{rot}	e_{tr}	e_{rot}	e_{tr}	e_{rot}
All Converged										
Mean	118	0.34	5.69	2.08	5.27	1.49	**3.31**	**0.42**	4.00	0.92
Median	118	0.35	**2.21**	0.97	2.97	0.53	2.36	**0.30**	2.38	0.34
Not All Converged										
1	10	0.46	7.24	2.85	7.90	3.01	1.87	1.33	-	-
2	10	0.69	1.54	0.26	-	-	-	-	-	-
3	10	0.45	7.13	1.23	-	-	14.70	0.48	1.87	0.16
4	10	0.47	2.30	0.62	-	-	1.75	0.36	2.65	0.50
5	10	0.62	-	-	-	-	1.19	0.45	1.31	0.34
6	20	0.50	-	-	-	-	1.39	0.19	1.15	0.15
7	10	0.45	-	-	-	-	3.27	0.36	2.09	0.27
8	10	0.46	-	-	-	-	3.14	0.38	1.99	0.29
9	10	0.46	-	-	-	-	2.66	0.33	2.01	0.29
Mean	-	0.51	4.55	1.24	7.90	3.01	3.75	0.48	**1.87**	**0.29**
Median	-	0.46	4.72	0.95	7.90	3.01	2.26	0.37	**1.99**	**0.29**

r of 0.46. This corroborates Fig. 8, where SIFT's converged image count drops near $r = 0.4$. In all, the monocular variant of RLFF showed slightly stronger performance (lower error) for these easier scenes; whereas, the stereo version performed better for the more challenging scenes, in which at least one method failed to converge. The cases where RLFF failed to converge both involved low feature counts for scenes at the occluding boundaries of the refractive objects, which is a known limitation of the image-based features.

In sum, Table II shows our methods outperforming SIFT-based methods, and importantly, the proposed method converged in all scenes but one (mono) or two (stereo), while the SIFT-based methods failed to converge for five (mono) or eight (stereo). The proposed method allows SfM to operate in scenes where it previously could not, and it improves performance in camera pose estimation wherever these is refracted scene content, even for less challenging scenes.

VI. CONCLUSIONS

We proposed a novel 4D feature defined by the rays of light travelling through curved refractive objects, as opposed to the conventional 2D image features defined by a single 3D point in space. Advantageously, our feature captures both Lambertian points and features imaged through smooth refractive objects. We demonstrated methods for detecting and extracting the proposed RLFF from LF imagery captured by a hand-held LF camera, and for employing the resulting features in conventional vision algorithms including SfM. Finally, we evaluated RLFF's benefits in the context of SfM, comparing to conventional SIFT-based methods. We showed the proposed method allowed SfM to operate where it previously could not. We also show improved 3D reconstruction performance and 3D camera trajectory estimation. Our method is especially advantageous in scenes with over 46% refracted features, in which traditional approaches fail.

We intend to develop a feature that directly employs the local 4D structure of refracted features, as in the LiFF feature for Lambertian LFs [7]. Demonstrating RLFF in more scenarios is also of interest, including closed-loop control for visual servoing, and place recognition for localisation and SLAM. More complex and thicker refractive objects may be addressed by adopting the thick lens model or similar optics theory. Finally, it is well understood that reflection off smooth curved surfaces exhibits characteristics similar to refraction through transparent objects (multiple observable depths for a single feature). We expect generalisation of the RLFF to reflective scenes to be straightforward and to show similar performance advantages as in the refractive case.

REFERENCES

[1] I. Ihrke, K. Kutulakos, H. Lensch, M. Magnor, and W. Heidrich, "Transparent and specular object reconstruction," *Computer Graphics forum*, vol. 29, pp. 2400–2426, 2010.

[2] D. Tsai, D. G. Dansereau, T. Peynot, and P. Corke, "Distinguishing refracted features using light field cameras with application to structure from motion," *IEEE Robotics and Automation Letters (RA-L)*, vol. 4, no. 2, pp. 177–184, April 2019.

[3] Y. Xu, H. Nagahara, A. Shimada, and R. ichiro Taniguchi, "Transcut: Transparent object segmentation from a light field image," *Intl. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[4] P. Corke, *Robotics, Vision and Control*. Springer, 2013.

[5] J. Schoenberger and J.-M. Frahm, "Structure-from-motion revisited," *CVPR*, 2016.

[6] M. Levoy and P. Hanrahan, "Light field rendering," in *SIGGRAPH*. ACM, 1996, pp. 31–42.

[7] D. G. Dansereau, B. Girod, and G. Wetzstein, "LiFF: Light field features in scale and depth," in *CVPR*, 2019, pp. 8042–8051.

[8] D. Miyazaki and K. Ikeuchi, "Inverse polarisation ray-tracing: estimating surface shapes of transparent objects," *CVPR*, 2005.

[9] C. Choi and H. Christensen, "3D textureless object detection and tracking: An edge-based approach," in *Intl. Conference on Intelligent Robots and Systems (IROS)*, 2012.

[10] Z. Zhou, Z. Sui, and O. C. Jenkins, "Plenoptic Monte Carlo object localization for robot grasping under layered translucency," in *IROS*. IEEE, 2018, pp. 1–8.

[11] O. Johannsen, A. Sulc, and B. Goldluecke, "On linear structure from motion for light field cameras," in *Intl. Conference on Computer Vision (ICCV)*, 2015, pp. 720–728.

[12] S. Nousias, M. Lourakis, and C. Bergeles, "Large-scale, metric structure from motion for unordered light fields," in *CVPR*, 2019, pp. 3292–3301.

[13] I. Tosic and K. Berkner, "3D keypoint detection by light field scale-depth space analysis," in *Image Processing (ICIP)*. IEEE, 2014.

[14] D. Tsai, D. G. Dansereau, T. Peynot, and P. Corke, "Image-based visual servoing with light field cameras," *RA-L*, vol. 2, no. 2, pp. 912–919, 2017.

[15] J. A. Teixeira, C. Brites, F. Pereira, and J. Ascenso, "Epipolar based light field key-location detector," in *Multimedia Signal Processing*. IEEE, 2017.

[16] K. Maeno, H. Nagahara, A. Shimada, and R. Taniguchi, "Light field distortion feature for transparent object recognition," in *CVPR*. IEEE, Jun. 2013.

[17] E. Hecht, *Optics*, 4th ed. Addition-Wesley, 2002.

[18] E. H. Adelson and J. Y. A. Wang, "Single lens stereo with a plenoptic camera," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 2, pp. 99–106, 1992.

[19] R. Arandjelović and A. Zisserman, "Three things everyone should know to improve object retrieval," in *CVPR*. IEEE, 2012, pp. 2911–2918.

[20] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in *CVPR*. IEEE, Jun. 2013, pp. 1027–1034.

[21] C. Ham, S. Singh, and S. Lucey, "Occlusions are fleeting - texture is forever: Moving past brightness constancy," in *WACV*, 2017.