Plenoptic Cameras

Bastian Goldlücke, Oliver Klehm,
Sven Wanner, and Elmar Eisemann

5.1 Introduction

The light field, as defined by Gershun in 1936 [Gershun 36] describes the radiance traveling in every direction through every point in space. Mathematically, it can be described by a 5D function which is called the *plenoptic function*, in more generality sometimes given with the two additional dimensions time and wavelength. Outside a scene, in the absence of occluders, however, light intensity does not change while traveling along a ray. Thus, the light field of a scene can be parameterized over a surrounding surface; light intensity is attributed to every ray passing through the surface into any direction. This yields the common definition of the light field as a 4D function. In contrast, a single pinhole view of the scene only captures the rays passing through the center of projection, corresponding to a single 2D cut through the light field.

Fortunately, camera sensors have made tremendous progress and nowadays offer extremely high resolutions. For many visual-computing applications, however, spatial resolution is already more than sufficient, while robustness of the results is what really matters. Computational photography explores methods to use the extra resolution in different ways. In particular, it is possible to capture several views of a scene from slightly different directions on a single sensor and thus offer single-shot 4D light field capture. Technically, this capture can be realized by a so-called plenoptic camera, which uses an array of microlenses mounted in front of the sensor [Ng 06]. This type of camera offers interesting opportunities for the design of visual computing algorithms, and it has been predicted that it will play an important role in the consumer market of the future [Levoy 06].

The dense sampling of the light field with view points lying closely together may also offer new insights and opportunities to perform 3D reconstruction. Light fields have thus attracted quite a lot of interest in the computer vision community. In particular, there are indications that *small changes* in view point, are important for visual understanding. For example, it has been shown that even minuscule changes at occlusion boundaries from view point shifts give a powerful perceptional cue for depth [Rucci 08].

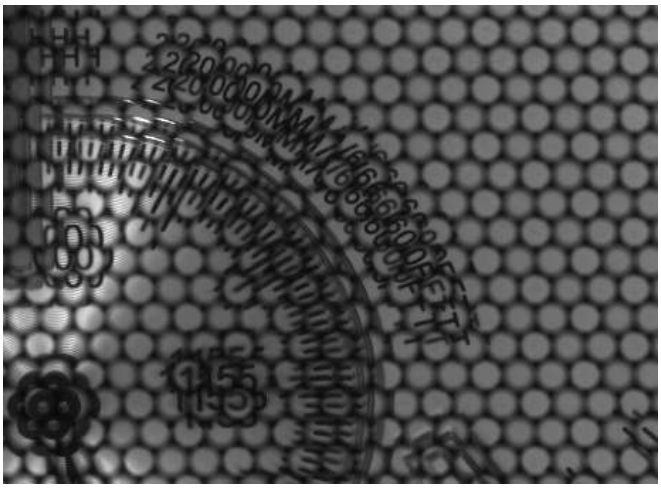


Figure 5.1: Detail of a raw image captured by a plenoptic 2.0 camera by Raytrix. Objects closer to the camera are visible in more microlens images. The camera makes use of different types of microlenses to increase depth of field, which can be distinguished in this image by comparing the sharpness of the projections.

5.2 4D Light Field Acquisition

Considering the special case that the light field is recorded on a planar surface, the 4D light field in this sense can be viewed as an intensity function that not only depends on the 2D position on the imaging plane, but also on the 2D incident direction. Many ways to record light fields have been proposed and can be classified into three main categories [Wetzstein et al. 11]. *Multi-sensor capture* solves the problem essentially on the hardware level. One can assemble multiple (video) cameras into a single array, with the cameras lying on a common 2D plane [Wilburn et al. 05]. This solution is quite expensive and requires careful geometric and photometric calibration of the sensors [Vaish et al. 04], as well as considerable effort to process and store the huge amount of data streamed by the array in real time. However, with temporal synchronization of the camera triggers, one can also apply camera arrays to the recording of dynamic scenes [Wilburn et al. 05]. Furthermore, they allow some interesting applications due to their wide baseline.

In contrast, with *time-sequential imaging* one is limited to static scenes, but only a single sensor is required. Different view points of the scenes are captured consecutively by moving the camera [Levoy and Hanra-

han 96, Gortler et al. 96], rotating a planar mirror [Ihrke et al. 08], or programmable aperture, where only parts of the aperture are opened for each shot, allowing to re-assemble the light field from sevel such images by computational means [Liang et al. 08]. Besides cost considerations, an advantage of the sensor being the same for all views is that calibration is simplified, Chapter 1.

Finally, a technology which recently has become available in commercial cameras is *single-shot multiplexing* where a 4D light field is captured with a single sensor in a single shot, which also makes it possible to record videos. In all cases, one faces a trade-off between resolution in the image ("spatial") and view point ("angular") domain. In plenoptic cameras [Ng 06, Bishop and Favaro 12, Georgiev et al. 11, Perwass and Wietzke 12], spatial multiplexing is realized in a straightforward manner by placing a lenslet array in front of the sensor, which allows to capture several views at the same time. Other techniques include coded aperture imaging [Lanman et al. 08] or, more exotically, a single image of an array of mirrors can be used to create many virtual view points [Manakov et al. 13].

Of the light field acquisition techniques above, plenoptic cameras are gaining increasing interest in the vision community since they are now commercially available as affordable consumer hardware.

5.3 Plenoptic Cameras

While normal 2D cameras only record irradiance from different directions at a single view point in space, plenoptic cameras capture the complete 4D light field on the sensor plane. The idea originates in the early 20th century. First described using a grid of pinholes inside a camera by Ives in 1903 [Ives 03], Lippmann proposed the use of microlenses in front of the image plane in 1908 [Lippmann 08]. Several improvements to the design have been proposed. For example, cameras manufactured by the company Raytrix employ multiple types of microlenses to accomplish a larger depth of field, Fig. 5.1.

At the time of writing, plenoptic cameras are commercially available from two manufacturers. The Lytro camera is based on the "plenoptic 1.0" design and targeted at the consumer market, while the Raytrix camera is based on the "plenoptic 2.0" design and targeted at industrial applications. This is reflected in both price as well as the bundled software.

The plenoptic camera 1.0 (Lytro camera) design is based on a usual camera with a digital sensor, main optics, and an aperture. In addition, a microlens array is placed in the focal plane of the main lens exactly at the focal length f_{MLA} from the sensor, Fig. 5.2. This way, instead of integrating the focused light of the main lens on a single sensor element, the microlenses

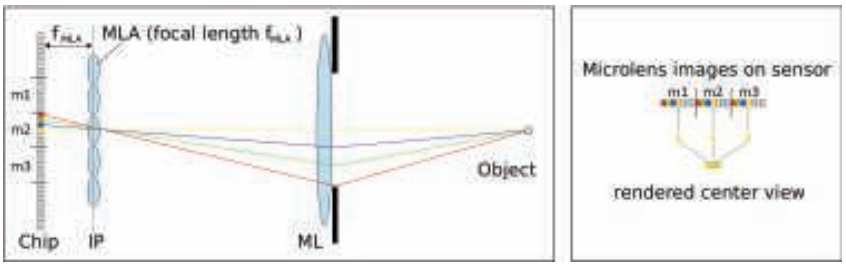


Figure 5.2: *Left:* one-dimensional sketch of a Plenoptic Camera 1.0 setup. Light rays emitted by the object are focused by the main lens (ML). The microlens array (MLA) is placed at the image plane (IP) of the main lens and thus separates the rays according to their direction. *Right:* a single view point, indexed by (s, t), here the center view, is extracted by collecting the corresponding pixels of each micro image m_i.

split the incoming light cone according to the direction of the incoming rays and map them onto the sensor area behind the corresponding microlens. In particular, one has direct access to the radiance $L(u, v, s, t)$ of each ray of the light field by choosing the micro-image of the microlens corresponding to spatial position (s, t) and pixel corresponding to direction (u, v) of the underlying micro-image. The size of each microlens is determined by the aperture or f-number of the main optics. If the microlenses are too small compared to the main aperture, the images of adjacent microlenses overlap. Conversely, sensor area is wasted if the microlenses are too large. Since light passing the main aperture also has to pass a microlens before being focused on a pixel, what actually happens is that the camera integrates over a small 4D volume in light field space. The calibration of unfocused lenslet-based plenoptic cameras like the ones commercially available from Lytro is discussed in [Dansereau et al. 13].

The main disadvantage of the 1.0 design is the poor spatial resolution of the rendered views, which is equal to the number of microlenses. By slightly changing the optical setup, one can increase the spatial resolution dramatically. As another way to compactly record 4D light fields, the focused plenoptic camera has been developed, often called the plenoptic camera 2.0 (Raytrix camera) [Lumsdaine and Georgiev 09, Perwass and Wietzke 12].

The main difference in the optical setup between the cameras is the relative position of the microlens array. The microlenses are no longer placed at the principal plane of the main lens, but are now focused onto the image plane of the main lens. In effect, each microlens now acts as a single pinhole camera, observing a small part of the virtual image inside

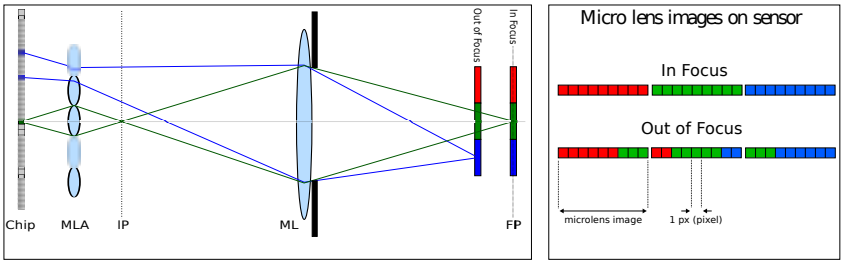


Figure 5.3: *Left:* one-dimensional sketch of a Plenoptic Camera 2.0 setup. Light rays emitted by the object are focused by the main lens (ML) onto the image plane (IP). The micro-lens array (MLA) is placed so that the micro-lenses are focused onto the image plane of the main lens, mapping fractions of the virtual image onto the sensor. Green rays originate from an object in focus of the main lens (FP), blue rays from an object away from the principal plane of the main lens. *Right:* resulting micro-images of an object in and out of focus.

the camera. This small part is then mapped with high spatial resolution onto the sensor. The scene points have to lie in a valid region between the principal plane of the main lens and the image sensor. Scene features behind the principal plane cannot be resolved.

Scene points that are not in focus of the main lens but within this valid region are imaged multiple times over several neighboring microlenses, thus encoding the angular information over several micro-images, Fig. 5.3 [Lumsdaine and Georgiev 09, Perwass and Wietzke 12]. Angular information is encoded while at the same time preserving high resolution. Due to multiple imaging of scene features, however, rendered images from this camera have a lower resolution than the inherent sensor resolution promises. The light field is encoded in a complicated way, and it is necessary to perform an initial depth estimate at least for each microlens in order to decode the sensor information into the standard 4D light field data structure [Wanner et al. 11]. External and internal calibration of plenoptic 2.0 cameras has been investigated in [Johannsen et al. 13].

5.4 4D Light Field Structure and Depth Reconstruction

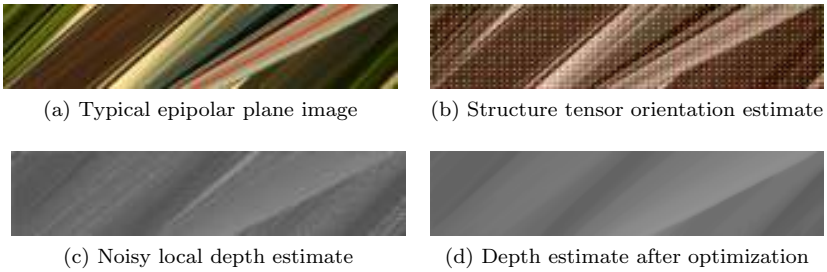
Since a 4D light field can be understood as a dense collection of multiple views, off-the-shelf correspondence search techniques can be applied to infer

Figure 5.4: One way to visualize a 4D light field is to think of it as a collection of images of a scene, where the focal points of the cameras lie in a 2D plane. The rich structure becomes visible when one stacks all images along a line of viewpoints on top of each other and considers a cut through this stack (denoted by the green border). The 2D image one obtains in the plane of the cut is called an *epipolar plane image (EPI)*.

3D structure (Chapter 8). Due to the rich information content in the light field data, however, also specialized methods can be developed, which work more efficiently and robustly.

One line of research follows the philosophy of the earliest works on the analysis of epipolar volumes [Bolles et al. 87], and rely on the fact that 3D scene points project to lines in the epipolar-plane images. The reason is that a linear camera motion leads to a linear change in projected coordinates, Fig. 5.4. These lines can be more robustly detected than point correspondences which has been exploited in several previous works [Bolles et al. 87,Berent and Dragotti 06,Criminisi et al. 05]. A recent advanced method aims at accurate detection of object boundaries and is embedded in a fine-to-coarse approach, delivering excellent results on very high-resolution light fields [Kim et al. 13].

In the same spirit, an efficient and accurate approach, which is however limited to only small disparity values and thus has a limited depth range, computes a direct estimate of the local orientation of the pattern [Wanner and Goldluecke 14], Fig. 5.5. Here, orientation estimation is performed using an Eigenvector analysis of the first-order structure tensor of the EPI. This approach can be extended to detect multiple overlaid patterns to efficiently reconstruct reflections or transparent objects [Wanner and Goldluecke 13]. Since local depth estimates from any source (including, e.g., stereo matching) are usually noisy, global smoothing schemes can be employed to improve the result. By careful construction of the regularizers and constraints, one can obtain consistent estimates over the complete light field which respect occlusion ordering across all views [Goldluecke



(a) Typical epipolar plane image (b) Structure tensor orientation estimate

(c) Noisy local depth estimate (d) Depth estimate after optimization

Figure 5.5: Depth estimation on an epipolar plane image (a). Standard 2D Pattern analysis using the structure tensor yields a robust orientation estimate (b), whose slope encodes the (still noisy) depth map for the EPI (c). Global optimization techniques result in a consistent estimate across all views (d).

and Wanner 13, Wanner and Goldluecke 14].

Other 3D reconstruction method specific to light fields exist, including focus stacks in combination with depth-from-focus methods [Nayar and Nakagawa 94, Perez and Luke 09]. Multiple methods that make use of depth maps to warp individual light field views to densify the light field from a sparse set of views have been proposed, Section 17.2.

5.5 Spatial and Angular Super-Resolution

Since plenoptic cameras trade off sensor resolution for the acquisition of multiple viewpoints.it is not surprising that super-resolution techniques are one focus of research in light-field analysis. Such methods have been investigated using priors regarding statistics of natural images [Bishop and Favaro 12] as well as modified imaging hardware [Lumsdaine and Georgiev 09].

In the classical Bayesian approach, an image formation model is set up to obtain the known input images from the desired super-resolved target image. In particular, when one transforms the target image into the image domain of an input image and performs a downsampling operation (usually modeled via a blur kernel), one should obtain an exact copy of the input image. In practice, however, this property will not be satisfied exactly due to sensor noise or sampling errors. Thus, the set of equations is enforced as a soft constraint in a minimization framework, where the desired super-resolved image appears as the minimizer of some energy functional [Bishop

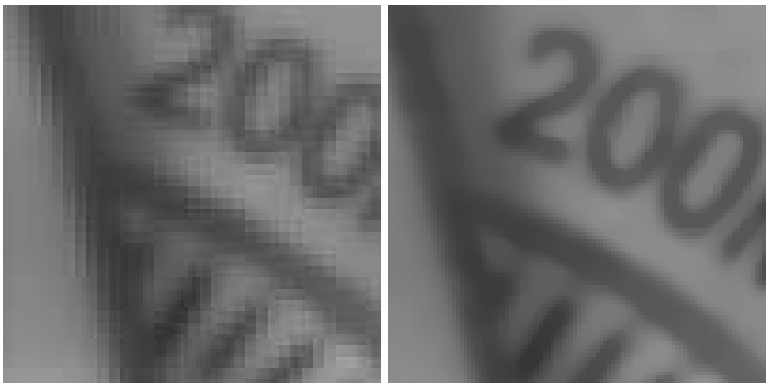


Figure 5.6: By solving a single inverse problem, one can create super-resolved novel views from a 4D light field captured with a Raytrix plenoptic camera [Wanner and Goldluecke 12]. Above are close-ups of one of the 7×7 input views (left) and the result from the super-resolution algorithm (right).

and Favaro 12, Wanner and Goldluecke 12].

In particular, some frameworks also allow to generate views in new locations, thus, solving an image-based rendering task in the same step [Wanner and Goldluecke 12]. In some recent work, a Bayesian framework was explored which also models uncertainties in the depth estimates and which is able to mathematically derive many of the heuristics commonly used in image-based rendering [Pujades et al. 14]. The topic of image-based rendering is explored in detail in Chapter 17.

5.6 Refocusing and other Applications

In this section, methods are presented that allow to simulate the intrinsics of a usual camera by relying on a light field as input. The two additional dimensions of a 4D light field compared to a conventional 2D image (quantities: radiance $[\mathrm{W\,m^{-2}\,sr^{-1}}]$ vs. irradiance $[\mathrm{W\,m^{-2}}]$) make it possible to produce effects such as changing the aperture or refocusing (adjusting the focal plane) even *after* a photo has been taken.

For many of these effects, a depth image has to be computed first, which can be directly derived from the light field, Section 5.4. In particular, the plenoptic camera 2.0 requires a reasonable depth estimate to reconstruct any meaningful image from the captured light field. This depth reconstruction is possible because the light field stores partially redundant

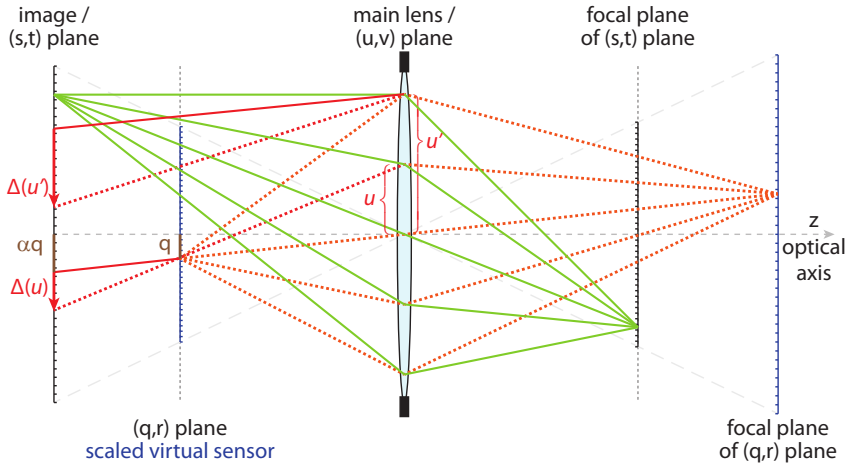


Figure 5.7: Refocusing example: a new virtual sensor at plane (q, r) is introduced, causing the a different focal plane. The mapping from coordinates (q, r) in the local space of the scaled sensor to coordinates (s, t) in the space of the original image plane reduces to a constant translation $\Delta(u, v)$.

information. More precisely, objects in a scene tend to have similar appearance under slightly different viewing angles. While the redundancy can be directly used for compression of light field data [Levoy and Hanrahan 96], it has recently exploited for the reconstruction of a light field from very sparse data [Marwah et al. 13].

The basis of the following examples is to sample or integrate the 4D light field to synthesize a new 2D image. The classical (u, v, s, t) parameterization [Levoy and Hanrahan 96] of a light field uses two distinctive planes that are aligned with the optical axis z: $[u, v]^T$ denotes the coordinates on the plane at the main lens (ML) z_{UV} and $[s, t]^T$ the coordinates on the focal plane of the main lens. As points on the focal plane uniquely map to points on the image plane (IP), $[s, t]^T$ also denotes the coordinates on the image plane at z_{ST}.

The light field can be used to fetch radiance for a new plane (q, r) at distance z_{QR}, parallel to the (s, t) plane. The mapping to the original coordinates is simple as it only requires to find the intersection of the ray, originating at $[q, r, z_{\mathrm{QR}}]^T$ with direction $[u, v, z_{\mathrm{UV}}]^T - [q, r, z_{\mathrm{QR}}]^T$ with the (s, t) plane at z_{ST}. The $[s, t]^T$ coordinates of the intersection point can be determined in two steps: first, a scaling α of $[q, r]^T$ depending on the positions of the (s, t), (q, r), and (u, v) planes is computed: $\alpha = \frac{z_{\mathrm{UV}} - z_{\mathrm{ST}}}{z_{\mathrm{UV}} - z_{\mathrm{QR}}}$.

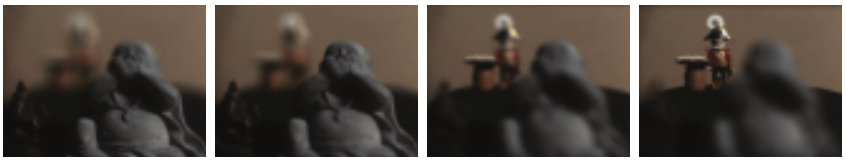


Figure 5.8: Example of a refocusing sequence. Left to right: the focal plane is moved from front to back, shifting the focus from the buddha statue to the pirate. The scene was captured with the kaleidoscope camera add-on [Manakov et al. 13] with an 50mm f/1.4 main lens. While this light-field camera add-on only captures nine directions, these views are sufficient to estimate depth and perform view interpolation, allowing for smooth out-of-focus blur.

Second, a translation by $\Delta(u,v) = -\beta \cdot [u,v]^T$ with $\beta = \frac{z_{\mathrm{QR}} - z_{\mathrm{ST}}}{z_{\mathrm{UV}} - z_{\mathrm{QR}}}$ yields the final coordinates in the (s,t) plane: $[s,t]^T = \alpha \cdot [q,r]^T + \Delta(u,v)$, Fig. 5.7.

While a pinhole camera could, in theory, have an infinitesimal aperture, such a camera would not produce any image, because no light would be detected. Hence, cameras rely on a larger aperture and use a lens to refocus the rays. All points on a so-called focal plane project to exactly one location on the sensor; outside the focal plane, points can project to several locations. Adjusting the focal plane right is a major challenge in photography. Imagining light rays leaving a camera, all rays from a given pixel will meet on the focal plane. Traversing these light rays in the opposite direction, all rays will be integrated at the given sensor pixel. With 4D light fields, it is possible to perform this integration in a post-capture process, Fig. 5.8.

A usual camera with the sensor at the image plane (IP) is simulated by integrating over all directions, hence, the (u,v) plane. Roughly, for a plenoptic camera 1.0, all pixels under a microlens are summed up as: $L(s,t) := \sum_u \sum_v L(u,v,s,t)$. The focal plane depends on the distance of the IP to the ML. Assuming the thin lens model, the original focal plane is at a distance $d_{\mathrm{org}} = (1/f - 1/(z_{\mathrm{UV}} - z_{\mathrm{ST}}))^{-1}$ from the main lens, where f is the focal length of the main lens. A virtual move of the image plane to a (q,r)-plane at z_{QR} causes the focal plane to change. Precisely, the new focal plane will be located at a distance $d_{\mathrm{refocus}} = (1/f - 1/(z_{\mathrm{UV}} - z_{\mathrm{QR}}))^{-1}$. To evaluate the result with the new focus plane, from a point on (q,r) all rays towards (u,v) are integrated, whereby (u,v,q,r) is mapped to (u,v,s,t) coordinates by the ray/plane intersection method described above.

This approach can be rendered more efficiently by splatting individual views (each indexed by their (u,v) coordinates, Fig. 5.2 right). Entire scaled views indexed by (u,v) can be accumulated on the sensor: $L(q,r) =$

$\sum_u \sum_v L(\alpha q + \Delta(u), \alpha r + \Delta(v), u, v)$ with $\Delta(u), \Delta(v)$ denoting the first resp. second component of Δ.

The main challenge of refocusing is that it requires a very high number of different view points in order to achieve a smooth out-of-focus blur. For a large blur kernel, banding or ghosting artifacts can remain visible. As none of the existing plenoptic cameras provide a sufficiently high number of view points, it is often essential to perform view interpolation (Sections 5.5 and 17.2).

In photography, *Bokeh* defines the rendering of out-of-focus areas by a camera lens. For small and very bright out-of-focus lights, this effect can be strong and is used as a stylization method. The shape of the Bokeh is indirectly defined by the shape of the lens aperture. As the lens aperture in a standard camera cannot be changed, photographers often attach an additional aperture with reduced size in front of the lens. The attachment simply blocks incoming light from certain directions. With the 4D light field, it is very simple to simulate such a behavior: $L'(u, v, s, t) = b(u, v)L(u, v, s, t)$ with b being a function mimicking the aperture shape. In order to control the aperture, incident light rays are thus scaled by a weighing factor (usually a binary mask). Additionally, it is possible to make this influence depend on the wavelength.

As refocusing practically requires interpolation in the (u, v) domain to generate additional views (Sections 5.4 and 5.5), the same pipeline can also perform extrapolation. Extending the available directional domain corresponds to photography with a larger aperture, which allows for very narrow depth-of-fields. Manakov et al. [Manakov et al. 13] report the simulation of a lens with an aperture of up to $f/0.7$ from a single snapshot light-field.

In practice, pixels of a light-field camera do not correspond to exact rays. Instead, each pixel records the incident irradiance within a small cone of directions. Each view point that relates to the microlens corresponds to an image taken with a lens of small aperture, Fig. 5.2 right. Consequently, these views also exhibit depth-of-field, and any refocusing operation is limited to the depth-of-field range imposed by these optics. Similarly, the captured light field might not be sufficient to deal with large apertures as some light rays necessary for the border pixels might be missing.

In a 4D light field, when keeping (s, t) constant, varying the (u, v) parameters results in a view of the scene, Fig. 5.2 right, which roughly corresponds to a capture of the scene with a pinhole camera centered at (u, v). Changing the (u, v) parameters causes a 'lens-walk' and offsets the corresponding image. This 2D effect relates to the Ken Burns-effect using a 2D pan and zoom, but with a light field this walk can also be extended to 3D by varying the (s, t) coordinates. Hereby, a parallax effect is induced due to the different viewpoints. A trivial extension is the generation of stereo images by sampling two (u, v) views. More details on multi-view-stereo are

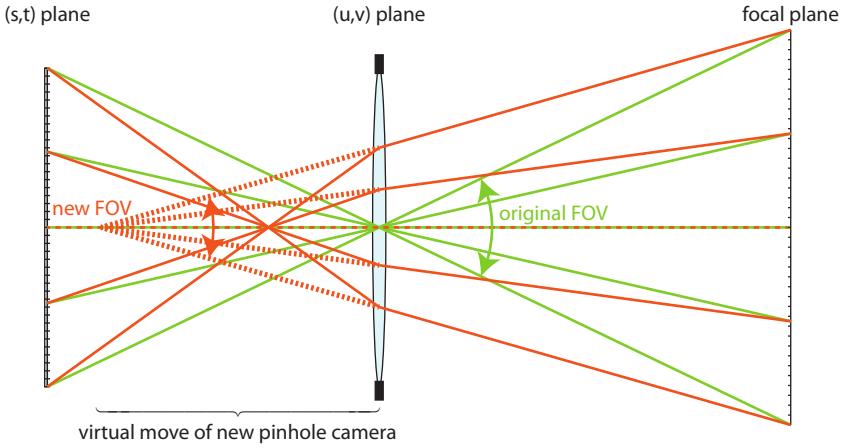


Figure 5.9: Dolly zoom: the light field allows to pick a (u, v) per (s, t). While the center view (green) corresponds to a pinhole camera with center of projection at the origin of the (u, v) plane, it is possible to simulate a moving pinhole camera with different field-of-view. Here, objects in front of the focal plane shrink while objects behind the focal plane grow in projected size (and are potentially cut off).

described in Section 8.3.

A computationally more involved effect is the 'dolly zoom' or 'Hitchcock zoom', where a change of the field-of-view (FOV) and camera motion along the viewing direction are synchronized, while focusing on an object in the scene. It causes out-of-focus regions to warp due to the changing FOV while the focal plane position and its imaged extent remain the same. Typically, this effect is used to shrink/grow the background, while keeping the object in focus at the same scale for a dramatic effect. To achieve this result, the image is rendered by: $L(s, t) = L(s, t, \gamma s, \gamma t)$ with γ defining the strength and direction of the effect. Here, a single ray sample is taken from each view, Fig. 5.9.

While refocusing processes thousands of views for high-quality rendering, the dolly zoom requires a single view per pixel. In turn, the computational complexity stems from the fact that it requires dense directional information. In practice, angular interpolation is strictly needed. While splatting of entire views, as in the refocusing application, is not possible, some computational simplifications can be made. The effect is most efficiently implemented by coupling the ray selection $(s, t, \gamma s, \gamma t)$ and directional interpolation in a GPU shader.

5.7 Summary

With the advent of consumer-grade plenoptic cameras, light field imaging has become comparatively cheap. Acquisition of a 4D light field is now as simple as taking a picture with a standard digital camera. Consequently, in addition to the traditional light field applications in computational photography and image-based rendering, a lot of research interest has been geared recently to leverage light fields for computer vision challenges like non-Lambertian 3D reconstruction.

Related to this chaper, Chapter 8 deals with 3D reconstruction from light field correspondence estimation, while Chapter 17 covers image-based rendering in more detail.