ISSN 2464-4617 (print)
ISSN 2464-4625 (online)
Computer Science Research Notes - CSRN 3301
http://www.wscg.eu
WSCG 2023 Proceedings

SAIL: Semantic Analysis of Information in Light Fields: Results from Synthetic and Real-World Data

Robin Kremer
Saarland University
Saarland Informatics Campus
Campus Building C6 3
Germany 66123, Saarbrücken, Saarland
kremer@cs.uni-saarland.de

Thorsten Herfet
Saarland University
Saarland Informatics Campus
Campus Building C6 3
Germany 66123, Saarbrücken, Saarland
herfet@cs.uni-saarland.de

ABSTRACT

Computational photography has revolutionized the way we capture and interpret images. Light fields, in particular, offer a rich representation of a scene's geometry and appearance by encoding both spatial and angular information. In this paper, we present a novel approach to light field analysis that focuses on semantics. In contrast to the uniform distribution of samples in two-dimensional images, the distribution of samples in light fields varies for different scene regions. Some points are sampled from multiple directions, while others may only be captured by a small portion of the light field array. Our approach provides insights into this non-uniform distribution and helps guide further processing steps to fully leverage the available information content.

Keywords

Light fields, Computational photography, semantic analysis, information content, MPEG-I

1 INTRODUCTION

The objective of enhancing the immersive experience for visual content has been a long-standing pursuit, dating back several decades, starting at analog photography and progressing over digital cameras to 3D video [Sch09]. In recent years computational photography has become one of the most important parts of the complete visual pipeline and is used to optimally use all available data [Lam03; Lib19; Sam21]. The latest frontier in immersive content is the creation of interactive experiences that enable users to freely adjust their viewpoint in real-time. This type of content can encompass a range of immersive experiences, from 3 Degrees of Freedom (DoF), where users can change the direction of their viewpoint, to 6 DoF content, which allows users to also move their position in space [MPE18]. To enable such interactive applications, it is essential to capture a scene from multiple viewpoints, which is typically achieved using light field cameras or light field arrays. Although the resulting data is also visual in nature, there are several significant differences between light fields and traditional 2D imaging. One of the main being that, properties such as view-dependent appearances and occlusions caused by scene geometry are lost in traditional 2D imaging, whereas these effects are captured in light fields. Although the raw data rate of light field capture systems can be upwards of ten gigabits per second [Che20], which presents significant challenges for current processing, network, and storage devices, this additional data provides unique post-processing options. In contrast to recent work that focuses on these applications and makes certain assumptions about the available data (e.g. everything is Lambertian), this work presents a method for analyzing the distribution of information in light fields.

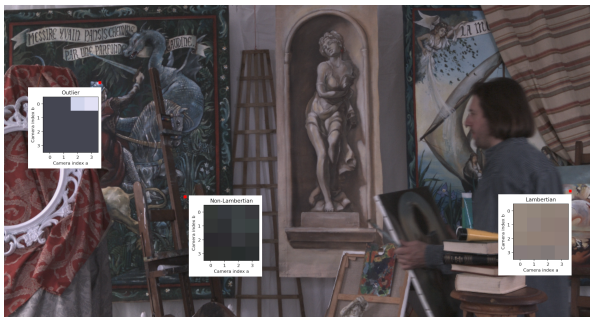


Figure 1: Painter scene from InterDigital [Sab17] captured on a 4 by 4 light field array. Highlighted are three scene regions (froxels). The displayed diagrams show the rays assigned to each froxel. As each ray is captured by a different camera, the color distributions give insights about the view dependency of the regions.

2 RELATED WORK

The underlying theory behind light fields has been established for several decades [Ber91; Lev96], similar to the field of machine learning. However, many applications only recently became computationally feasible. Today, a diverse range of light field filters and processing techniques is available.

With MPEG Immersive video (MIV), which is part of ISO/IEC 23090 MPEG-I, the Moving Picture Experts Group introduced a standard to store and distribute highly immersive 3D content. A content database has been published with scenes captured on a variety of different setups. The general idea is to remove redundancy by merging all views into one and creating a patch atlas for occluded regions. In practice, diverse options to create the atlases are available.

Methods like "Linear Volumetric Focus" [Dan15], "Calibration and Auto-Refinement for Light Field Cameras" [Ani21] and "Fourier Disparity layers" [Le 19] are based on traditional algorithms and filters. These techniques enable a range of effects, such as adjusting the focal distance and depth of field or synthesizing new views in real-time. To achieve these results, certain assumptions are often made, such as treating the entire scene as being Lambertian or assuming limited occlusions. If a scene does not adhere to these assumptions, visual artifacts may occur.

Similar to many other fields, such as image segmentation or gigapixel compression, machine learning methods also became a popular choice for light field processing. Techniques like "Deepview" [Fly19] and "Immersive Video" [Bro20] use gradient decent optimisation to represent a scene as a Multi-Plane Images (MPI) or Multi-Sphere Images (MSI) enabling real-time view interpolation even for light field videos. However, it should be noted that the training of these techniques is computationally intensive and can take multiple tens of hours per frame. In the work of "Local Light Field Fusion" (LLFF) [Mil19], MPIs are also utilized, but they employ a single trained network to promote each input view into an MPI (local light field). This significantly reduces the total time from capture to view synthesis (roughly 10 minutes), while still enabling real-time view interpolation.

One of the most disruptive innovations for the field of light field processing in recent years has been the emergence of Neural Radiance Fields (NeRFs)[Mil20]. These encode the information of a light field in multilayer perceptron (MLP) neural network. Specifically, the MLP takes both the 3D spatial location and viewing direction as input and outputs color and volume density information. In the context of light field theory this amounts to a continuous representation of the underlying plenoptic function and thereby enables synthesis of arbitrary viewpoints. Training NeRFs is computationally expensive, requiring multiple GPU hours. Additionally, synthesizing novel views from the original NeRF implementation is not feasible in real-time, typically requiring multiple tens of seconds. Although network inference is relatively fast, volumetric rendering necessitates the use of multiple samples per pixel, resulting in millions of inferences required to produce a high-definition view. Despite these limitations, the visual quality of NeRF-generated images is exceptionally high and capable of gracefully handling non-Lambertian and opaque objects. In addition, scenes represented as NeRFs require significantly less memory compared to LLFF, and often are even smaller in size than the original input views.

Numerous techniques have been developed that build upon the fundamental idea of NeRF, enabling the handling of specific types of scenes and addressing limitations of the original implementation. Mip-NeRF [Bar21] increases the visual fidelity by sampling conical frustums instead of rays. This was further developed in Mip-NeRF 360 [Bar22] to better handle unbounded 360 degree scenes. While methods like "NeRF in the dark" [Mil21] and "HDR-NeRF" [Hua22] focus on dynamic range and noise handling. Other methods like "Fastnerf" [Gar21] speed up the inference time significantly rendering 100+ frames per second on modern GPUs. "PixelNeRF" [Yu21], on the other hand, drastically reduces the number of input images compared to traditional NeRF. Works like "Instant Neural Graphics Primitives" [Mül22] can produce high quality results after just 5 minutes of training. Recent techniques such as "Space-time NeRF" [Xia21] and others [Par21; Pum21] have extended the capabilities of NeRFs to be able to handle videos.

In summary, since their introduction, NeRFs have emerged as the most prominent method for processing light fields and have spawned a new research field focused on extending their capabilities. While neural techniques are likely to dominate light field processing, it is crucial for applications such as codecs, post-processing and novel view generation to be capable of real-time operation and to possess an understanding of the underlying scene properties. As a result, this paper does not aim to compete with the processing capabilities of NeRF and its derivatives. Rather, it seeks to address a more fundamental aspect of the complete visual pipeline, namely, what information is captured by a light field array and how it is distributed.

3 THEORY OF LIGHT FIELDS

The theory behind light fields is based around the plenoptic function, which contains all information about the propagation of light in a certain space-time region [Ber91]. A light field is created by sampling this continuous function at certain positions using

cameras. As such, only a small portion of the overall information contained in the plenoptic function is sampled. Nonetheless, the amount of information contained within a light field is substantial and allows for a vast variety of post processing techniques as described earlier.

Although static scenes can be captured by moving a camera on a gantry or handheld, the majority of light fields are captured using multiple cameras that are rigidly fixed together in a camera array (light field array) like the "Stanford Multi-Camera Array" [Wil05] or the light field array from InterDigital [Sab17]. Light fields captured by these rigs are called "forward-facing" since all cameras are oriented into the same direction. The two-plane parameterization (compare Figure 2) is particularly intuitive, as it closely aligns with the physical arrangement of the cameras [Cam98]. The first plane corresponds to the plane on which the cameras are mounted (a,b-plane / camera), while the second plane represents the plane on the camera sensors (u,v-plane / pixel). Due to this close correspondence to the physical arrangement of cameras, the two-plane parameterization is a common starting point for a wide range of light field processing techniques.

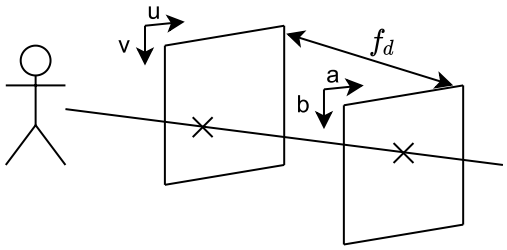


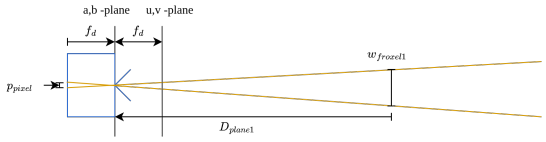
Figure 2: In the two-plane parameterization a light ray is described by it's intersection with two parallel planes.

One characteristic of the two-plane parameterization, as well as similar formats like the Direction and Point Parameterization (DPP) and Two-Sphere Parameterization (2SP) [Cam99], is that they present the light field information in a camera-centric format. Although this is intuitive, it hides how the captured information is arranged in a light field. This information distribution is a crucial difference between light fields and traditional 2D imaging. Because unlike single-camera imaging, where pixels evenly sample rays from a scene and capture each visible point exactly once, light fields can have a non-uniform sample distribution. Some scene points are visible in all cameras and are therefore sampled multiple times, while others may be occluded for most cameras and are only sampled by a small subset. The shape of this distribution plays a pivotal role in determining which post-processing techniques can be effectively applied to the captured light field data. For instance, achieving high-quality results with the "Light Field Superresolution" technique [Bis09] requires a sufficient number of samples per scene region, making the distribution of captured information a critical factor for the effectiveness of this and many other methods. Nevertheless, many light field processing techniques rely on certain assumptions about the distribution of captured information and deviations from these assumptions can lead to artifacts as in [Le 19; Dan15]. In the subsequent chapters, a scene-centric light field parameterization will be explored that enables straightforward analysis of captured information distribution.

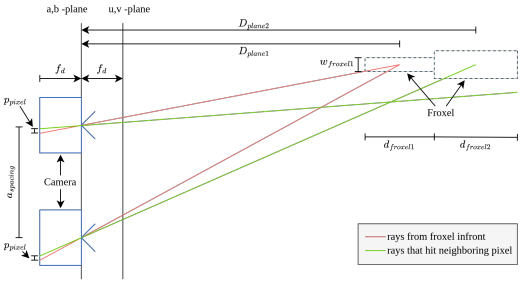
4 THE IDEA OF FROXELS

In order to analyze the distribution of information that is contained in a light field more easily, a scene centric parameterization is needed. In order to achieve this, we make use of the "froxel" concept, which involves discretizing the view frustum of the light field array into frustum-shaped voxels [Eva15]. This is accomplished by populating the view frustum with froxels of specific sizes, which are designed to match the resolution of the light field array. By choosing the size of the froxels appropriately, we can achieve a discretization raster that perfectly matches the array resolution. As a result, if an object in the scene is moved by one froxel, its image will shift exactly one pixel on a camera sensor. Unlike a single camera, which does not capture any information about scene depth and therefore does not require discretization along the depth axis, light field arrays do capture this information [Ber91]. As a result, the view frustum of a light field array has to be discretized in all three dimensions. For the two axes parallel to the camera plane, this discretization scheme is straightforward, since the region covered by a single pixel increases linearly with the distance from the camera (compare figure 3a). However, the resolution along the depth axis of the light field array, and thus the size of a froxel, is dependent on the specific geometry of the array. In light fields, depth information is captured as the disparity experienced by objects within the scene. As a result, the disparity is responsible for the depth resolution and, ultimately, the size of the froxels along the depth axis. Because disparity is inversely proportional to depth, froxels that are closer to the camera plane have smaller depth and become larger as they move further away from the camera. As the largest disparity is experienced between the furthest cameras in an array, the size of the froxels is chosen such that moving an object one froxel closer or further away from the camera plane results in a one-pixel change in its position between these two cameras (compare figure 3b). The exact dimension of the froxels can be calculated with (1) and (2). Where w_{froxel}, h_{froxel} and d_{froxel} are the width, height and depth of a froxel at a certain distance D_{plane} from the camera plane.

$$w_{froxel} = h_{froxel} = \frac{p_{pixel} D_{plane}}{f_d} \tag{1}$$



(a) The froxel width w_{froxel} and height h_{froxel} scale linearly with the distance D_{plane} from the camera plane (a,b-plane)



(b) The froxel depth is based on the maximum disparity that the array can capture

Figure 3: The froxel width, height and depth are chosen to perfectly match the resolution of the light field array

$$d_{froxel} = D_{plane}^2 / \left(\frac{f_d \cdot s_{max}}{p_{pixel}} - D_{plane} \right) \qquad (2)$$

It is assumed that all cameras have the same intrinsic parameters such as f_d, which is the focus distance. The maximum distance between two cameras in the light field array is denoted as s_{max} and governs the largest disparity that can be observed at a given depth.

Once the discretization raster is created, each ray captured in the light field is assigned to the froxel that contains the object from which it originated in the scene. This origin is calculated by combining the two-plane parameterization with a depth map. Due to the way the raster is designed, two rays captured by the same camera can not be assigned to the same froxel. However, when two rays are captured by different cameras and originate from the same scene point, they will be assigned to the same froxel. This means one froxel can at most have as many rays assigned to it as there are cameras in the light field array. Froxels that have rays assigned to them will be refereed to as non-empty froxels. Once all rays have been assigned to their respective origin froxel, the resulting set of non-empty froxels contains all of the information captured by the light field. The resulting froxel parameterization represents the information in a scene-centric manner, in contrast to the two-plane parameterization, which is camera-centric. The term "scene-centric" refers to the fact that this parameterization allows for easy analysis of the light field captured from a specific scene region, as it facilitates a straightforward examination of how rays and information are distributed throughout the scene.

5 OPTIMIZING THE FROXEL REPRESENTATION

As discussed in the previous section, the transformation from two-plane parameterization to the froxel parameterization relies on depth maps to determine the origin of a captured ray. For the froxel parameterization to be most effective, it is essential that rays originating from the same scene point are assigned to the same froxel. As a result, the accuracy of the depth maps has a significant impact on the achievable quality. This is particularly true, since the froxel sizes are designed to precisely match the resolution of the light field array.

Thus, the most crucial characteristic of the depth maps used in the froxel parameterization is good multi-view consistency, which means that the depth maps of each camera must assign the same depth to a given scene point. This consistency ensures that rays captured by different cameras and originating from the same scene point are assigned to the same froxel, resulting in an accurate representation of the underlying scene.

Upon analyzing multiple datasets that contained depth maps generated using various techniques, it was determined that while the resulting froxel parameterizations were acceptable, there remained potential to improve the meaningfulness. In theory, a wall that is parallel to the camera plane should result in a plane of non-empty froxels located exactly at the depth of the wall. However, in practice, the froxel representations are often narrowly distributed around the true position of the wall. To improve the meaningfulness of the parameterization and reduce the total number of non-empty froxels, a consolidation step is employed. This is based on the idea of reassigning rays from non-empty froxels with few rays to other froxels that already have more rays assigned to them. When reassigning a ray to a different froxel, only the non-empty froxels along the ray's original path are considered to avoid altering the representation too much. By searching for a new froxel within a few neighboring layers, the consolidation step can already reduce the total number of non-empty froxels significantly.

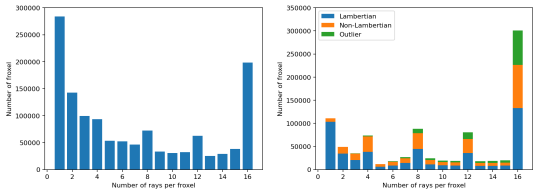
6 SEMANTIC ANALYSIS

Once a light field has been transformed into the froxel representation, it becomes significantly easier to analyze how the captured information is distributed. The number of rays assigned to each froxel following the conversion is a good starting point to analyze the information distribution. Firstly, since the majority of scenes typically contain a significant amount of free space, many froxels will remain unoccupied following the conversion process. Consequently, the froxels that do have rays assigned to them approximate the hull of the scene. However, within these non-empty froxels, there can be substantial differences in the amount of

information present. For instance, a froxel that corresponds to a scene point, which is captured by all of the cameras in a light field array, should have the same number of rays assigned to it as there are cameras in the array. Other froxels that are occluded for part of the array contain fewer rays. Consequently, the number of rays per froxel directly indicate how densely the underlying scene region is sampled. One approach for visualizing this distribution is to use fristograms (froxel + histogram) [Her21]. They are created by grouping froxels according to the number of rays assigned to them and then generating a histogram based on these groupings (compare 4a). They provide an initial indication of the level of uniformity with which a scene is sampled.

Another approach for visualizing the distribution of samples is to count the number of non-empty froxels along a ray. If there is more than one, it indicates that the corresponding scene point is likely occluded in the current viewpoint and was captured from a different perspective by a different camera. This allows to easily locate occluded scene regions that are only visible by a subset of all cameras in the light field array (compare figure 5d). This information may be used in various applications, such as virtual viewpoint rendering or aid in the generation of atlases.

Analyzing the distribution of rays within a froxel reveals additional semantic information. All rays that are assigned to a single froxel originate from the same scene point, but were captured from different directions. Consequently, by analyzing the color distribution of these rays, it is possible to infer the visual properties of the underlying object. If the rays within a froxel exhibit similar colors, this is an indication that the corresponding object behaves as a Lambertian radiator. On the other hand, if there is a significant amount of color variation among the rays, this suggests non-Lambertian behavior [Kop14]. This information is crucial for post-processing, as different techniques may only be effective for certain types of surfaces. , The froxel representation also provides a means for quantifying the information captured by a light field, allowing for the comparison of different capture setups. By analyzing the distribution of froxels and their associated rays, it is possible to evaluate the level of scene sampling and coverage achieved by a particular light field capture setup. To quantify the information content I_{total} of a light field, each ray is assigned a specific value that reflects its contribution to the overall scene information. For example, rays originating from a Lambertian surface point may be assigned a lower value compared to those from non-Lambertian or occluded regions, as the former contribute less unique information. In practice, this is often done by grouping rays of a froxel together if their color differs by less than a just-noticeable-difference (JND) [Sha17]. The probability of a ray p_i is then calculated with (4), where n_{rays}



(a) Fristogram (b) Consolidated and classified fristogram

Figure 4: Fristograms of the painter scene from InterDigital

is the total number of rays in the light field and $n_{cluster_i}$ is the size of the cluster to which the ray belongs. From this the total information content of the light field can be calculated with (3) [Sha48].

$$I_{total} = \sum_{i=1}^{n_{rays}} -\mathrm{ld}(p_i) \tag{3}$$

$$p_i = \frac{n_{cluster_i}}{n_{rays}} \tag{4}$$

This information can be used to optimize the design of future light field acquisition systems for specific applications.

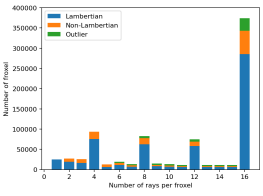
To demonstrate the effectiveness of the froxel representation, the surface properties present in a scene, where analysed by a simple froxel classification. The proposed technique works by analyzing the color distribution of rays assigned to individual froxels. Froxels are classified as non-Lambertian when the standard deviation of their associated rays surpasses a predetermined threshold, while those whose standard deviation is below the threshold are considered Lambertian. Additionally, a third category of "Outliers" is established by identifying non-Lambertian froxels that have at least one ray with a z-score that exceeds a certain value. This indicates that while the majority of the rays associated with a froxel have a uniform distribution of colors, there are a few outliers that do not conform to this pattern. This can be caused by specular highlights or due to wrongly assigned rays (compare figure 1).

The semantics acquired through this method can be utilized to direct post-processing procedures in a manner that minimizes visual artifacts while maximizing the use of all available information.

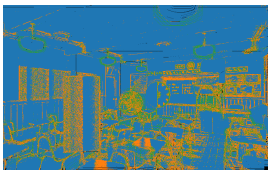
7 RESULTS

The developed pipeline was tested on synthetic data generated in blender, an open source 3D animation software, and real-world data sourced from the MPEG-I content database. The depth maps utilized during development were either generated in Blender, exported

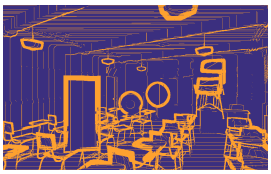
(a) Top left camera view



(b) Classified Fristogram



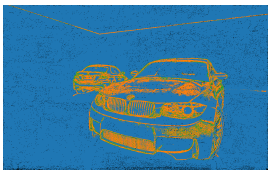
(c) Classified regions. Blue: Lambertian; Orange: non-Lambertian; Green: outlier



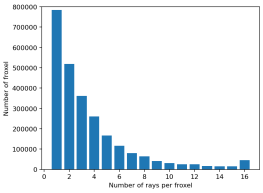
(d) Occluded regions. Violet: no occlusions captured; Orange: occlusions captured

Figure 5: Custom Blender Classroom scene, captured on a 4-by-4 light field array with blender depths

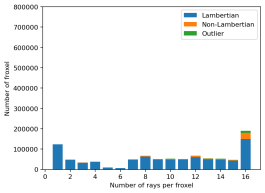
(a) Top left camera view



(b) Classified regions. Blue: Lambertian; Orange: non-Lambertian; Green: outlier



(c) Fristogram before consolidation: 2,569,795 non-empty froxels



(d) Classified fristogram after consolidation: 942,232 non-empty froxels

Figure 6: Blender BMW scene with depth maps extracted from NeRF

from a NeRF, or provided together with the content. Although our methods are capable of accommodating arbitrary forward-facing light field arrays, the results presented in this paper are all based on light fields captured by uniform 4-by-4 arrays to increase conciseness.

Figure 5 displays results generated with high quality depth maps that were generated in blender. Upon examining the corresponding fristogram, it is clear that there is a prominent peak at 16 rays per froxel. This indicates that the majority of the scene was captured by all the cameras in the 4-by-4 array. Additionally, distinct bumps can be observed at 4, 8, and 12 rays per froxel, which correspond to edges in the scene that align with the arrangement of the cameras in the array, such as the floating cork board in the foreground. These edges create occlusions for multiple cameras simultaneously, resulting in noticeable patterns in the fristogram. Moreover, by including the froxel classes, the fristogram reveals that most of the scene behaves in a Lambertian manner. Examining Figure 5c, it is apparent that the wall and ceiling are classified as Lambertian, whereas the table desks with a glossy finish and the reflective metal chair legs are classified as non-Lambertian. Occlusions that occur in the light field are displayed in 5d. The presence of orange stripes on the edges of objects signifies the existence of samples that lie behind the foreground object within the light field. This information can be leveraged to reveal occluded areas within the scene, or even to identify objects that could potentially be completely eliminated during view reconstruction.

In the shown example of the Classroom scene, the depth maps were generated within Blender, which allowed for access to the scene geometry and, as a result, yielded depth maps of exceptionally high quality. Since such high quality depth maps are not always available especially for real world scenes[Luo20; Jan20; Kop21], the developed methods were also tested on depth maps acquired by other means. Specifically, we showcased the compatibility of our approach with NeRF by training a NeRF model, extracting the corresponding depth maps, and using them into our pipeline.

Upon examining the fristogram of the BMW scene (see figure 6c) generated from the NeRF depth maps, it becomes evident that the shape is markedly different from that of the Classroom scene. Despite the fact that much of the scene is visible to all 16 cameras, the majority of the froxels are assigned fewer than four rays. An inspection of the scene reveals that a substantial portion of it consists of featureless, monotonous background, which presents inherent challenges in generating depth maps accurately from visual data [Sch16]. This leads to bad multi-view consistency, which artificially inflates the number of non-empty froxels. To address this issue, we leverage the techniques outlined in Chapter 5 to consolidate froxels. This drastically reduced the number of non-empty froxels and created a clear peak at 16 rays per froxel. Although, not as pronounced as previously small peaks at 8 and 12 rays per froxel are also visible. Looking at the resulting classification (see figure 6b) the background is correctly marked as Lambertian, while reflective features on the car are identified as non-Lambertian. This demonstrates that our method is capable of generating dense froxel representations that hold significant meaning, even in situations where access to the scene geometry is not available. Nevertheless, the information value that can be extracted increases with the quality of the depth maps.

Table 1: Information Content

Scene	Captured	Minimum
Classroom	182,664,675 bits	176,354,697 bits
BMW	178,224,039 bits	176,354,697 bits
Painter	182,190,389 bits	170,124,571 bits

Table 2: Impact of ray reduction on visual quality

	Classroom			BMW		
Method	all rays	one sample	Ours	all rays	one sample	Ours
PSNR↑	30.460	29.400	30.180	36.360	32.890	35.620
SSIM↑	0.9153	0.8882	0.9076	0.9801	0.9677	0.9778
LPIPS↓	0.0596	0.0922	0.0689	0.0276	0.0489	0.0357
Ray Count	9.21 M	1.1 M	3.22 M	9.21 M	1.03 M	1.81 M

Furthermore, we showcase the potential of the froxel representation for real-world scenes. As an example, Figure 7 depicts the Painter scene sourced from InterDigital [Sab17], which was captured using a 4-by-4 light field array. This scene is listed in the MPEG-I content database and comes supplied with depth maps. An examination of the fristogram (refer to Figure 7b) reveals distinct peaks at 16 rays per froxel indicating that most of the scene is sampled by all 16 cameras. Peaks at 12, 8, and 4 rays per froxel suggest occlusions that are roughly aligned to the camera pattern of the light field array. These regions can be seen in figure 7d. Looking at the distribution of Lambertian, non-Lambertian and outlier froxels it becomes evident, that these scene contains many more than the previous two. This is a consequence caused by the limitations of color matching between cameras and the fact that real objects always exhibit at least some level of Lambertian reflectance [Geo07]. Therefore, the classification thresholds could be adjusted for real world scenes, but for the sake of comparison, they were kept the same.

Calculating the information content for each scene, based on the method described in chapter 6, reveals the additional information captured due to occlusions and non-Lambertian surfaces. Table 1 displays the result for the three discussed scenes. The listed minimum information content would be achieved if all cameras captured exactly the same information (e.g. a Lambertian wall a depth infinity) and therefore is only depends on the total number of rays and cameras. The Classroom and BMW scenes were captured using the same virtual light field camera, enabling direct comparison of their results. Notably, the Classroom scene exhibits a significantly higher information content due to a larger number of occlusions compared to the BMW scene.

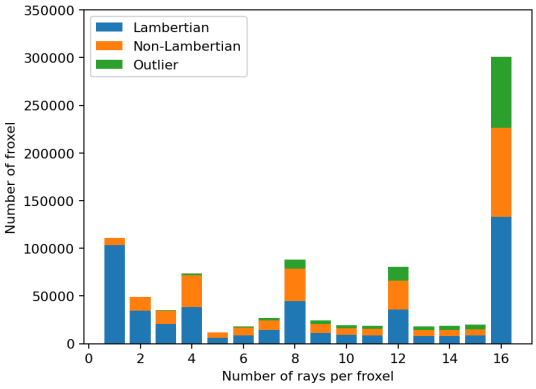
This semantic analysis can be used to guide further post processing steps. As an example a surface aware ray reduction was implemented. This is based on the idea that a Lambertian surface can be accurately described with a single view-independent sample, while non-Lambertian surfaces require multiple samples. In practice, the rays assigned to a Lambertian froxel were filtered using a mean filter, whereas those assigned to non-Lambertian froxels remained unaltered. The original views of the light field were generated using this reduced set of froxels and compared against views generated using all rays, as well as ones generated using only one sample per froxel. The results of the proposed ray reduction technique are presented in Table 2. It can be observed that the visual quality achieved with the reduced set of rays is comparable to that of the unfiltered representation, while containing significantly fewer rays. Although the representation that utilizes only one sample per froxel contains even fewer rays, it results in notably lower quality.

We evaluated the entire processing pipeline on supplementary Blender scenes, such as "The Wanderer" by Daniel Bystedt and "Mr. Elephant" by Glenn Melenhorst, yielding consistent results. Obtaining further real-world data posed challenges due to the limited availability of suitable datasets.

8 CONCLUSION

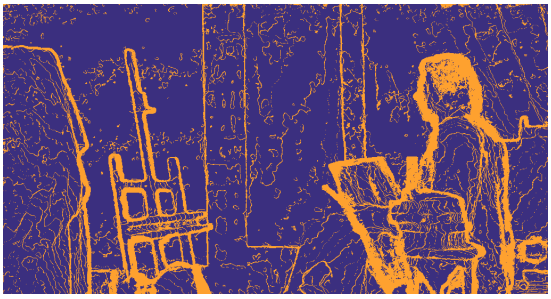
In this paper we demonstrated how the froxel representation can be leveraged to perform semantic analysis of the information contained within a light field. Specifically, we illustrated methods for quantifying the sampling density of a captured scene and classifying surface properties. Rather than challenging methods like NeRF, that prioritize novel view reconstruction, our proposed approach instead enables visualization and quantization of the information distribution. This can be leveraged to effectively adapt post-processing steps to the available data. This enables creative professionals to understand the types of processing feasible with the acquired data, while also facilitating efficient light field encoding. One such application was demonstrated with a surface property aware ray reduction. Furthermore, we showed that our pipeline is robust against imperfect depth maps and can be applied to real-world scenes. A limitation, that the current pipeline shares with MPEG Immersive Video (MIV) is the assumption that the region between the cameras and the scene hull is free space. While in theory the froxel parameterization is capable of handling a more nuanced representation, this limitation is due to the fact, that the used depth maps only assign one specific depth to each ray. This limitation could be overcome by utilizing more complex depth formats and would permit better analysis of complex visual phenomenons such as fog. Moreover, the presented method of semantic analysis is compatible with the notion of time, enabling the analysis of light fields video (e.g. quantify the difference in information content captured by sub-framing).

(a) Top left camera view



(b) Classified Fristogram after consolidation

(c) Visualization of the classified regions. Blue: Lambertian; Orange: non-Lambertian; Green: outlier



(d) Visualization of occluded regions. Violet: no occlusions captured; Orange: occlusions captured

Figure 7: Example visualization of the painter scene from InterDigital [Sab17]

9 ACKNOWLEDGMENT

This work is supported by the German Research Foundation (DFG) under grant number 429078454.

REFERENCES

[Ani21] Yuriy Anisimov, Gerd Reis, and Didier Stricker. “Calibration and Auto-Refinement for Light Field Cameras”. In: *arXiv preprint arXiv:2106.06181* (2021).

[Bar21] Jonathan T Barron et al. “Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 5855–5864.

[Bar22] Jonathan T Barron et al. “Mip-nerf 360: Unbounded anti-aliased neural radiance fields”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 5470–5479.

[Ber91] James R Bergen and Edward H Adelson. “The plenoptic function and the elements of early vision”. In: *Computational models of visual processing* 1 (1991), p. 8.

[Bis09] Tom E Bishop, Sara Zanetti, and Paolo Favaro. “Light field superresolution”. In: *2009 IEEE International Conference on Computational Photography (ICCP)*. IEEE. 2009, pp. 1–9.

[Bro20] Michael Broxton et al. “Immersive light field video with a layered mesh representation”. In: *ACM Transactions on Graphics (TOG)* 39.4 (2020), pp. 86–1.

[Cam98] Emilio Camahort, Apostolos Lerios, and Donald S Fussell. “Uniformly sampled light fields.” In: *Rendering Techniques* 98 (1998), pp. 117–130.

[Cam99] Emilio Camahort and Don Fussell. “A geometric study of light field representations”. In: *Technical Report TR99-35* (1999).

[Che20] Kelvin Chelli et al. “A Versatile 5D Light Field Capturing Array”. In: *NEM Summit 2020*. 2020.

[Dan15] Donald G. Dansereau, Oscar Pizarro, and Stefan B. Williams. “Linear Volumetric Focus for Light Field Cameras”. In: *ACM Transactions on Graphics (TOG)* 34.2 (2015).

[Eva15] Alex Evans. *"Learning from failure: A Survey of Promising, Unconventional and Mostly Abandoned Renderers for 'Dreams PS4', a Geometrically Dense, Painterly UGC Game*. SIGGRAPH, 2015.

[Fly19] John Flynn et al. "Deepview: View synthesis with learned gradient descent". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 2367–2376.

[Gar21] Stephan J Garbin et al. "Fastnerf: High-fidelity neural rendering at 200fps". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 14346–14355.

[Geo07] Georgi T Georgiev and James J Butler. "Long-term calibration monitoring of Spectralon diffusers BRDF in the air-ultraviolet". In: *Applied Optics* 46.32 (2007), pp. 7892–7899.

[Her21] Thorsten Herfet et al. "Fristograms: Revealing and Exploiting Light Field Internals". In: *arXiv preprint arXiv:2107.10563* (2021).

[Hua22] Xin Huang et al. "Hdr-nerf: High dynamic range neural radiance fields". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 18398–18408.

[Jan20] Joel Janai et al. "Computer vision for autonomous vehicles: Problems, datasets and state of the art". In: *Foundations and Trends® in Computer Graphics and Vision* 12.1–3 (2020), pp. 1–308.

[Kop14] Sanjeev J. Koppal. "Lambertian Reflectance". In: *Computer Vision: A Reference Guide*. Ed. by Katsushi Ikeuchi. Boston, MA: Springer US, 2014, pp. 441–443. ISBN: 978-0-387-31439-6.

[Kop21] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. "Robust consistent video depth estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 1611–1621.

[Lam03] Edmund Y Lam. "Image restoration in digital photography". In: *IEEE Transactions on Consumer Electronics* 49.2 (2003), pp. 269–274.

[Le 19] Mikael Le Pendu, Christine Guillemot, and Aljosa Smolic. "A fourier disparity layer representation for light fields". In: *IEEE Transactions on Image Processing* 28.11 (2019), pp. 5740–5753.

[Lev96] Marc Levoy and Pat Hanrahan. "Light field rendering". In: *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. 1996, pp. 31–42.

[Lib19] Orly Liba et al. "Handheld mobile photography in very low light." In: *ACM Trans. Graph.* 38.6 (2019), pp. 164–1.

[Luo20] Xuan Luo et al. "Consistent video depth estimation". In: *ACM Transactions on Graphics (ToG)* 39.4 (2020), pp. 71–1.

[Mil19] Ben Mildenhall et al. "Local light field fusion: Practical view synthesis with prescriptive sampling guidelines". In: *ACM Transactions on Graphics (TOG)* 38.4 (2019), pp. 1–14.

[Mil20] Ben Mildenhall et al. "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis". In: *ECCV*. 2020.

[Mil21] Ben Mildenhall et al. "NeRF in the Dark: High Dynamic Range View Synthesis from Noisy Raw Images". In: *arXiv* (2021).

[MPE18] WG 11 MPEG-I. *MPEG-I Phase 1 Use Cases (v1.5)*. Standard. International Organization for Standardization, 2018.

[Mül22] Thomas Müller et al. "Instant neural graphics primitives with a multiresolution hash encoding". In: *ACM Transactions on Graphics (ToG)* 41.4 (2022), pp. 1–15.

[Par21] Keunhong Park et al. "Nerfies: Deformable neural radiance fields". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 5865–5874.

[Pum21] Albert Pumarola et al. "D-nerf: Neural radiance fields for dynamic scenes". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 10318–10327.

[Sab17] Neus Sabater et al. "Dataset and Pipeline for Multi-View Light-Field Video". In: *CVPR Workshops*. 2017.

[Sam21] Jaroslaw Samelak et al. "Efficient Immersive Video Compression using Screen Content Coding". In: (2021).

[Sch09] Heidrun Schaaf et al. "evolution of photography in maxillofacial surgery: from analog to 3D photography–an overview". In: *Clinical, cosmetic and investigational dentistry* (2009), pp. 39–45.

[Sch16] Johannes Lutz Schönberger and Jan-Michael Frahm. "Structure-from-Motion Revisited". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[Sha17] Gaurav Sharma and Raja Bala. *Digital color imaging handbook*. CRC press, 2017.

[Sha48] Claude E Shannon. "A mathematical theory of communication". In: *The Bell system technical journal* 27.3 (1948), pp. 379–423.

[Wil05] Bennett Wilburn et al. "High performance imaging using large camera arrays". In: *ACM SIGGRAPH 2005 Papers*. 2005, pp. 765–776.

[Xia21] Wenqi Xian et al. "Space-time neural irradiance fields for free-viewpoint video". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 9421–9431.

[Yu21] Alex Yu et al. "pixelnerf: Neural radiance fields from one or few images". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 4578–4587.