Structure From Plenoptic Imaging
Simão Pedro da Graça Oliveira Marto
smarto@isr.tecnico.ulisboa.pt
Instituto Superior Técnico, Universidade de Lisboa, Portugal
November 2017

Abstract
Plenoptic cameras are becoming more available. They give extra information on a captured scene. In this report, methods for scene reconstruction based on single snapshots are studied and compared. Precision and accuracy of depth reconstructions are assessed both on synthetic and real datasets. Reconstruction results are promising for a variety of focused depth settings of the camera.

I. Introduction
Conventional cameras are functionally similar to human eyes. They are composed of an image sensor and a single main lens. Plenoptic cameras, on the other hand, are more closely related to compound eyes, like those of insects such as the mantis or the housefly.

These devices display what is called a lightfield, that is, a sampling of the light rays in a given scene, discriminated not only by the position of the rays, as in a conventional photograph, but also by their direction.

In plenoptic cameras, the many micro-lenses are equivalent to the many ommatidia in compound eyes. In the case of Lytro’s first generation camera, shown in Figure 1, there are about a hundred thousand micro-lenses, each being sampled by about a hundred pixels.

A detail of an image captured by a Lytro first generation camera is shown in Figure 1, where several small hexagons can be seen. These correspond to the projected image of each microlens.

A. The Lightfield and the Plenoptic Function
To better understand plenoptic cameras, it is important to understand the concept of lightfield, and the plenoptic function, introduced by Adelson and Bergen [1]. The plenoptic function, represented as \( L(u,v,\lambda,t,x,y,z) \), describes the intensity of a light ray with wavelength \( \lambda \), direction defined by the direction coordinates \( (u,v) \), that passes through position \( (x,y,z) \), and at an instant \( t \).

A simplified version, the lightfield, is used throughout this article. The lightfield assumes a static monochromatic light ray whose intensity does not change throughout the ray’s path. These simplifications reduce the plenoptic function to a four dimensional space, \( L(x,y,u,v) \), the lightfield [12]. The parametrization of a lightfield will be described in more detail in Section II-A.

Plenoptic cameras and arrays of conventional cameras are capable of sampling this 4D space. Any lightfield image has in it enough information to produce a 2D array of 2D images, called viewpoints, which implies multi-view stereo is possible and, therefore, so is depth reconstruction.

B. Related Work
The idea of capturing the lightfield with a single camera can be traced back to the work of G. Lippmann in 1908 [11]. He proposed a camera that contained a transparent sheet with several spherical bumps on its anterior face, which would function like the microlenses in today’s plenoptic cameras.

Later, in 1992, Adelson and Wang [2], proposed a plenoptic camera with main lens and a microlens array located at the focal plane. With this setup depth information could be extracted from a scene in a simpler form than with binocular imaging.

In 2006, Ren Ng founded the Lytro company which produced the first consumer plenoptic cameras, the Lytro first generation cameras. In [16], [17], Ren Ng et al. introduce the concept of digitally refocusing lightfield images after they have been captured.

1) Depth Estimation: Along with the development of lightfield cameras, methods for extracting depth from their images were also being produced.
Dansereau et al. [7], Monteiro et al. [15] and Wanner et al. [20] have showed that depth can be estimated by gradient operations on epipolar plane images (EPIs) extracted from the acquired lightfield. The latter two use the structure tensor on EPIs to estimate disparity. The concept of EPIs and what they represent will be discussed in more detail in Section III-B. The gradient based method in Section III-B is based in these approaches.

There are also other works using depth from focus, such as [10] by Diebold et al. and [19] by Tao et al. In this work, a similar approach, that focuses on correspondence information, is evaluated in Section III-C.

2) Plenoptic Camera Calibration: The work by Dansereau et. al in [8] was instrumental in this paper, as it delineates a method for calibrating a plenoptic camera. This allows the pixel information to be turned into metric information, and thus metric reconstruction can be made of a scene.

C. Problem Formulation

There is a large amount of work on methods for reconstructing scenes with plenoptic cameras, but not much assessing reconstruction accuracy on real scenes. The purpose of this work is to evaluate the effectiveness of these methods for those purposes. Thus, various reconstruction methods will be analyzed with respect to their depth range and accuracy, on both virtual images, and images captured with a first generation Lytro lightfield camera, Figure 1.

1) Motivation: In this dissertation, methods to obtain depth reconstruction from lightfield are developed and assessed, so as to study the depth estimation capabilities of a plenoptic camera.

Previous work on depth reconstruction has so far focused on obtaining valid disparities, which showed objects appearing in correct order. Obtaining actual depth values, and evaluating their accuracy, was outside the scope most of these works.

D. Thesis Structure

Chapter 1 introduces the problem to approach in the thesis, depth estimation accuracy. In particular, it presents a short discussion on the state of the art on plenoptic imaging and the processing of lightfield images to obtain depth. Section 2 presents a study on the properties of a plenoptic camera. Section 3 describes methods for reconstructing depth from lightfield images. Section 4 provides an overview of the different experiments executed as well as the results attained. Section 5 summarizes the work performed and highlights the main achievements in this work. Moreover, this Section proposes further work to extend the activities described in this document.

The work developed in this thesis has been partially published in [14]. Part of section II has been submitted to the journal Computer Vision and Image Understanding (CVIU).

II. MODELLING A PLENOPTIC CAMERA

Plenoptic imaging can discriminate the contribution of pixels with varying positions and directions by sampling the lightfield.

In order to properly interpret the lightfield images, it is necessary to have a precise model of the camera. For this purpose, the model described by Dansereau et al. [8] is used. In this chapter, the model is briefly described, and adapted for our needs in analyzing a lightfield camera’s depth accuracy. Namely, back-projection and projection models are developed based on the model proposed in [8].

Before proceeding to a description of the camera model, we describe the lightfield parametrization that is used throughout this document.

A. Lightfield Parametrization

For the analysis of a lightfield, it is necessary to have it parametrized. Two lightfields are defined, based on [8]: for the object space and for the image space. The lightfield in the object space describes the rays using metric coordinates, the lightfield in the image space describes the ray based on how it was captured by the camera. A ray in the image space is often referred to as a “raxel” [13] in analogy to the term pixel.

The lightfield in the object space can be described using a two-plane parametrization (2PP). The rays have coordinates (s, t, u, v) where (s, t) describe the position the ray intersects a certain plane at, and (u, v) describe the derivative of the ray’s x and y positions with z, where axis Z is perpendicular to the first plane.

The lightfield images recovered from the camera, and decoded by Dansereau’s toolbox [6], are considered to be in the image space. A raxel is defined by coordinates (i, j, k, l), where (k, l) are the indexes of the microlens it passes through, and (i, j) are the indexes of the pixel it hits on the image sensor, measured relative to the image formed by the microlens on the image sensor. In Figure 2, both parametrizations for rays are shown.

![Fig. 2. Diagram showing two rays, Ψi and Ψj, and respective horizontal coordinates in the object space (s and u) and in the image space (k and l).](image-url)
B. Camera Intrinsics

In this work, the back-projection model by Dansereau et al. [8] is used to convert a ray from its image space representation to the object space. The parameters for the model are obtained through calibration, using a dataset of checkerboard images.

This model allows mapping the lightfield in the image space \( \Phi = [i, j, k, l, t]^T \) to the lightfield in the object space \( \Psi = [s, t, u, v, l]^T \), like so:

\[
\begin{bmatrix}
  s \\
  t \\
  u \\
  v \\
  1
\end{bmatrix} =
\begin{bmatrix}
  h_{si} & 0 & 0 & h_{si} \\
  0 & h_{ti} & 0 & h_{ti} \\
  h_{ui} & 0 & h_{uk} & 0 \\
  0 & h_{uj} & 0 & h_{uj} \\
  0 & 0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
i \\
j \\
k \\
l
\end{bmatrix},
\]

(1)

where \( \Phi \) is a matrix containing intrinsic parameters. The model described in [8] also includes distortion correction, but this is ignored in order to simplify the model.

1) Simplifications: The matrix \( \Phi \) defined by Dansereau has a total of 12 parameters. Of these, 4 are constant terms \( h_s, h_t, h_u, \) and \( h_v \), while the remaining 8 are coefficients that give the derivative of coordinates in the object space relative to coordinates in the image space. Of the former, \( h_s \) and \( h_t \) are redundant with translation parameters of the extrinsics. For this reason they are chosen so as to make the central \((i, j)\) values correspond to \((s, t) = (0, 0)\).

It is possible to reduce the number of parameters to consider by making some assumptions that do not compromise the ability of the model to represent an actual camera. More specifically, by assuming that the density of pixels and microlenses in both directions is the same, and translating the origin of the coordinate system. These simplifications are similar to the ones presented by Birkbauer et al. [3]. The intrinsic matrix that results of these simplifications is

\[
\hat{\Phi} =
\begin{bmatrix}
h_{si} & 0 & 0 & h_s \\
0 & h_{ti} & 0 & h_t \\
h_{ui} & 0 & h_{uk} & 0 \\
0 & h_{uj} & 0 & h_u \\
0 & 0 & 0 & 1
\end{bmatrix},
\]

(2)

where \( \hat{h}_s = h_s - h_u h_{sk}/h_{uk} \), \( \hat{h}_t = h_t - h_v h_{tk}/h_{uk} \), and \( \hat{h}_l = h_l - h_v h_{lk}/h_{uk} \). This intrinsics matrix only depends on three coefficient parameters, \( h_s, h_u \) and \( h_{uk} \).

The fact that a ray’s position \((s, t)\) depends only on coordinates \(i\) and \(j\) makes it clear that all rays share the same \((i, j)\) coordinates, which form a viewpoint, originate from the same position. This implies that viewpoints are pinhole cameras forming a square grid with centers of projection with coordinates \( \hat{h}_s [i, k, 0]^T \).

2) Camera Array Intrinsics: Lightfields can also be captured using camera arrays [21], comprised of several identical cameras arranged in a rectangular array, each pointing in the same direction. For these setups, an intrinsic matrix \( \Phi \) can be built which relates with the properties of each of the cameras used.

If the cameras are equally spaced on a rectangular grid, with horizontal spacing \( d_s \) and vertical spacing \( d_t \), while having focal length \( f \), principal point \((k_0, l_0)\) and no skew, the resulting intrinsic matrix is

\[
\hat{H} =
\begin{bmatrix}
d_s & 0 & 0 & s_0 \\
0 & d_t & 0 & t_0 \\
0 & 0 & 1/f & -k_0/f \\
0 & 0 & 1/f & -l_0/f \\
0 & 0 & 0 & 1
\end{bmatrix}.
\]

(3)

C. Back-projection model

The back-projection model relates a certain ray in the lightfield with the possible 3D positions that its light could have originated from. It will be useful in defining a projection model, as well as the reconstruction methods. For an arbitrary point \( \mathbf{m} = [x, y, z]^T \), the rays \( \Psi = [s, t, u, v, l]^T \) that pass through it obey the following relation, also defined in the work of Grossberg and Nayar [13]

\[
\begin{bmatrix}
x \\
y \\
z
\end{bmatrix} =
\begin{bmatrix}
s \\
t
\end{bmatrix} + \lambda \begin{bmatrix}
0 \\
1
\end{bmatrix}, \lambda \in \mathbb{R}.
\]

(4)

D. Projection Model

The projection model maps an arbitrary point in the object space, \( \mathbf{m} = [x, y, z]^T \), to the lightfield on the sensor plane, \( \Phi \), knowing the intrinsic matrix \( \Phi \). This model is developed to produce sets of synthetic rays that can be used to assess the accuracy of plenoptic cameras at performing depth estimation.

Combining Equation (4) with Equation (1), one obtains

\[
\begin{bmatrix}
i \\
j
\end{bmatrix} = f(k; \mathbf{m}, \Phi), \quad g(l; \mathbf{m}, \Phi)
\]

(5)

where \( f(k; \mathbf{m}, \Phi) \) and \( g(l; \mathbf{m}, \Phi) \) are affine on the variables \( k \) and \( l \). The coordinates of the lightfield in the sensor plane \((i, j, k, l)\), in general cannot be all integers.

Let the slope of the projection lines \( m_{i,j} \) corresponds to the derivative of the pixel relatively to the micro-lens coordinate. The results of Equation (5) show that \( m_{i,j} \) is constant for points at the same depth. \( b_{i,j} \) is the \( i \)- or \( j \)-intercept. The reader is reminded of the simplifications delineated in subsection II-B1, which are used here, namely that \( m_i \) is the same as \( m_j \).

Equation (5) shows that a point in the object space defines lines on the spaces defined by each pair of coordinates \((i, k)\) and \((j, l)\), which implies that it also defines a plane in the full lightfield.

1) Set of Imaged Points: Unlike common projection problems, as in the pinhole camera model, in a standard plenoptic camera a point \( \mathbf{m} \) in the object space normally has multiple projections. In order to produce a set of imaged rays, there needs to be a rasterisation procedure applied. By rasterisation it is meant that out of the infinite points defined by the line expressed by, e.g. \( i = f(k; \mathbf{m}, \Phi) \), a finite set of \((i, k)\) pairs is chosen to represent the projection of \( \mathbf{m} \) onto the lumigraph.

One way to achieve this is to pre-select the values for \( i \) and \( j \), and compute the values for \( k \) and \( l \) afterwards using Equation (5), or vice-versa.
In rigorous terms, if the viewpoint indexes \((i, j)\) are the pre-selected ones, the output of this method is given by
\[
\begin{align*}
\{ [i, j, k, l, 1]^T & : \]
\begin{align*}
k & = \lfloor f^{-1}(i; m, H) \rfloor, \quad i \in \chi, \quad k \in \kappa, \\
l & = \lfloor g^{-1}(j; m, H) \rfloor, \quad j \in \zeta, \quad l \in \iota
\end{align*}
\end{align*}
\]
where \(\lfloor \cdot \rfloor\) indicates rounding to the nearest integer, and \(\chi = \{1, \ldots, N_i\}\), \(\iota = \{1, \ldots, N_l\}\), \(\kappa = \{1, \ldots, N_k\}\), \(\zeta = \{1, \ldots, N_j\}\), and \(N(\cdot)\) corresponds to the number of samplings made of variable \((\cdot)\). If the micro-lens indexes are pre-selected, one gets
\[
\begin{align*}
\{ [i, j, k, l, 1]^T & : \]
\begin{align*}
i & = \lfloor f(k; m, H) \rfloor, \quad i \in \chi, \quad k \in \kappa, \\
j & = \lfloor g(l; m, H) \rfloor, \quad j \in \zeta, \quad l \in \iota
\end{align*}
\end{align*}
\]
where the red pixels are omitted if variable \(k\) is pre-selected, but not if \(i\) is instead. Since \(m_k = m_i^{-1}\), whenever \(m_i\) is greater than one, \(m_k\) is lower.

Choosing for pre-selection the variable for which \(|m_{(\cdot)}|\) is larger than one will lead to less rays being sampled than if the other variable had been chosen. This can be seen in Figure 3, where the red pixels correspond to the projections skipped by assuming integer values for the micro-lens \(k\).

Therefore, the projection model used will output Equation (6) if \(|m_{(\cdot)}| < 1\) and (7) otherwise, in this way avoiding skipping rays that are valid.

III. 3D-RECONSTRUCTION METHODS FOR PLENOPTIC IMAGES

In this chapter, various methods that can be used to reconstruct a scene from a lightfield image are presented and analyzed. Namely, a point based reconstruction method where a feature’s position is estimated from a collection of the rays that sample it; a gradient based method where depth is estimated from gradient calculations performed on epipolar plane images (EPIs); and a photo-similarity method, where the similarity between translated viewpoint images is used to measure disparity.

A. Point Based Reconstruction

If the coordinates of a set of rays that intersect a 3D point are known, it is possible to find the coordinates of that point from their intersection. Using a least squares method, the position of the point can be recovered. This forms the basis for the reconstruction method described in this section. The ray coordinates corresponding to a 3D point can be found, for example, using SIFT feature matching algorithms.

1) Multi-View Stereo Method: Let us consider that we have a set of \(Z\) rays on the sensor plane that correspond to a given point \(m\) in the object space and that the intrinsic matrix \(H\) is known. This allows to convert the set of rays \(\{\Phi_1, \ldots, \Phi_Z\}\) in the image space to a set of rays \(\{\Psi_1, \ldots, \Psi_Z\}\) in the object space.

Using the defined in Equation (4), and combining with Equation (1), we obtain for the set of rays

\[
\begin{bmatrix}
1 & 0 & -h_1 \Phi_1 \\
0 & 1 & -h_2 \Phi_1 \\
\vdots & \vdots & \vdots \\
1 & 0 & -h_3 \Phi_Z \\
0 & 1 & -h_4 \Phi_Z
\end{bmatrix}
\]

where \(h_i\) corresponds to the \(i\)-th row of the intrinsic matrix \(H\). This is a problem that can be readily solved using a least-squares method.

2) Imposing Projection Priors: Due to the discretization that occurs at the sensor level, the projection rays do not define a line in the ray-space defined by the pair of coordinates \((i, k)\) and \((j, l)\), but a staircase. Therefore, the reconstruction is improved if we ensure that the projection rays in these spaces define a line, by fitting a line to these coordinates, and correcting the less discriminative coordinate in each pair. Let us call these, the projection priors. This allows to have more precision on the coordinates used to define the projection rays and is likely to improve the reconstruction when the new set of projection rays is then given to the reconstruction methodology defined by Equation (8).

B. Gradient Based Reconstruction

The method defined in the previous Section III-A does not take advantage of the fact that a 3D feature forms a 2D plane in the image space, and its \((k, l)\) positions do not vary more than a few pixels. This can be used to obtain a computationally simpler, dense reconstruction method, that performs matching and depth estimation simultaneously, as we will show in this section.

For a lightfield in the image space, \(L(i, j, k, l)\) slices with \((i, j)\) constant are a viewpoint image, and slices made by fixing \((i, k)\) or \((j, l)\) are vertical or horizontal epipolar plane images (e.g. Figure 4).

Fig. 4. An example of an epipolar image, taken from a synthetic lightfield, made using Matlab’s VR tools.

Since each viewpoint image is similar to an image captured by a pinhole camera at a slightly different position (see Section II-B1), due to parallax, the position of objects in those viewpoints \((k, l)\) will vary according to their depths, causing...
slopes. This variation is referred to as disparity (see Definition III.1), like in stereoscopic vision.

**Definition III.1** (Lightfield Disparity). The disparity \( d \) of a feature on the lightfield is given as \( \frac{\partial k}{\partial i} \) or \( \frac{\partial l}{\partial j} \), which are the same when one keeps the assumptions made in subsection II-B1. Both these derivatives are taken while maintaining \((x, y, z)\) constant.

In the epipolar plane images, a feature is rendered as a line, a feature line, whose slope is related to its parallax. The relation between these slopes and the actual depth can be obtained by manipulating Equation (5), obtaining

\[
\frac{\partial k}{\partial i} = \frac{h_{ij} + h_{ik} \frac{\partial k}{\partial i}}{h_{ij} + h_{ik} \frac{\partial k}{\partial i}} \quad \text{or} \quad \frac{\partial l}{\partial j} = \frac{h_{ij} + h_{il} \frac{\partial l}{\partial j}}{h_{ij} + h_{il} \frac{\partial l}{\partial j}}.
\]

This equation relates depth to disparity \( \frac{\partial k}{\partial i} \) or \( \frac{\partial l}{\partial j} \). Afterwards, \( x \) and \( y \) positions can be determined from Equation (5).

The method described in this section extracts disparities by measuring the slopes of the feature lines in EPIs using structure tensor on epipolar plane image. The complete reconstruction algorithm encompasses three main steps, namely disparity estimation, disparity regularization and conversion to metric structure, which will be described in the next sections.

1) **Disparity From Gradient**: As previously stated, the disparity of the objects in the scene will be calculated from the slopes of the feature lines in the epipolar plane images. These slopes are calculated using the structure tensor. The structure tensor is used to obtain the predominant direction of the gradient in a certain region. The disparity corresponds to the slope of the feature lines, which are perpendicular to the gradient direction (see Figure 5). Thus, the disparity is related to the gradient by

\[
\frac{\partial k}{\partial i} = -\nabla_{L} L, \quad \frac{\partial l}{\partial j} = -\nabla_{L} L.
\]

![Fig. 5. Gradient vector \( \nabla L \) and disparity \( \frac{\partial k}{\partial i} \) on the epipolar image of a green object over a blue background.](image)

Keeping with the assumption made in Subsection II-B1 both of the derivatives in Equation (10) are the same.

The predominant gradient direction on the epipolar image is calculated using a structure tensor. This is preferred to using the gradients directly, because averaging the gradients could lead to gradients in opposing directions canceling each other. e.g. if a small black object is in front of a white background, the gradients on the EPI at the left of the object have opposing direction to those at the right. To obtain a depth map \( d(k, l) \), we will need to obtain a structure tensor matrix \( S(k, l) \). This is done by computing the structure tensor on each epipolar plane image of the lightfield, and averaging the results in the \( i \) and \( j \) directions.

2) **Sparse to Dense Disparities**: In areas where the confidence measure does not meet a certain threshold, the disparity values extracted were ignored. If the method were to proceed with only the remaining values, only a small amount of depth values would be present in the reconstruction. To improve the quality of the results, the areas with missing values are filled with an inpainting algorithm published by John D’Errico in [9]. A total variation regularization procedure by Gabriel Peyre [18] is then applied. Further study of the inpainting and regularization methods is outside of the scope of this thesis.

**C. Photo-Similarity Based Reconstruction**

The method described in the previous Section III-B works by evaluating the variation in position \((k, l)\) of a feature across the viewpoints, i.e., disparity. However, it does not attempt to match the data found in different EPIs that might correspond to the same features.

The method presented in this section does this by shearing the lightfield as described in Subsection III-C1 for various disparity values, creating a lightfield where features with a disparity equal to the one being tested will appear on the same \((k, l)\) positions on all viewpoints. In Subsection III-C2, a cost is then attributed to that pixel for all tested disparities that measures how similar the intensities on the same \((k, l)\) positions on all viewpoints are, producing a lightfield generalization of the concept of disparity space image (DSI) described in [4]. Finally, in Subsection III-C4, a method is proposed that can improve the accuracy of the estimations without increasing the running time of the algorithm.

1) **Lightfield Shearing**: In this section, the process of shearing the lightfield is explained. It was shown by Ren Ng in [17] that shearing the lightfield is equivalent to changing the world plane which is in focus in the microlens plane. This makes objects at these depths have zero disparity. In Tao et al. [19], this process is referred to as refocusing, although Ren Ng uses refocusing to refer to the method that produces a refocused conventional image. The process of shearing the lightfield consists of translating the position of each viewpoint by an amount proportional to the viewpoint distance from the center

\[
L_{\alpha}(i, j, k_{\alpha}, l_{\alpha}) = L(i - \alpha(i - i_{centre}), j - \alpha(j - j_{centre})).
\]

If a certain disparity \( \frac{\partial k}{\partial i} \) is to be tested, we want to shear \( L \) into \( L_{\alpha} \) such that features at that disparity become in focus, which happens for \( \alpha = \frac{\partial k}{\partial i} \).

2) **Lightfield Disparity Space Image**: The lightfield DSI defined in Definition III.2 is the basis of this photo-similarity method. It produces an array \( V \) indexed by microlens pixel \((k, l)\) and disparity \( \alpha \). A low value for \( V(k, l, \alpha) \) implies the pixels \((k, l)\) of all viewpoints in sheared lightfield \( L_{\alpha} \) are similar, which might mean that the feature is in focus, but not always, as will be discussed in Subsection III-C3.

**Definition III.2** (Lightfield DSI). The lightfield Disparity Space Image (DSI) is defined as the photo-similarity measure
computed for various sheared versions of the original lightfield (Equations 12 and 13).

\[ V(k, l, \alpha) = \frac{1}{N_i - 1} \frac{1}{N_j - 1} \sum_{i} \sum_{j} (L_\alpha(i, j, k, l) - \mu(k, l, \alpha))^2 \]
where

\[ \mu(k, l, \alpha) = \frac{1}{N_i N_j} \sum_{i} \sum_{j} L_\alpha(i, j, k, l) \]  

Fig. 6. Plot of variance as a function of disparity value considered to shear. The red line corresponds to a pixel in a low gradient region, while the blue line corresponds to a pixel in a high gradient region.

3) Confidence Metric: A problem of the method delineated so far is that a low value of \( V(k, l, \alpha) \) is not enough to determine that the right disparity has been found. Some regions of a lightfield \( L(i, j, k, l) \) are very homogeneous, which means that regardless of the disparity \( \alpha \) being tested, a low value is produced for \( V(k, l, \alpha) \). To improve the quality of the disparity data, it is necessary to find a way to remove these values by thresholding a confidence measure.

The plots in Figure 6 suggest that a confidence measure for the disparity estimations made can be obtained by using the mean curvature at an interval around the minimum. This way, areas that have a large variance outside of the minimum, due to being areas with strong gradients, are given higher confidence. Furthermore, this measure is directly related to the intensity of the gradients in the lightfield images.

Property III.1 (DSI Parabola). For a linearized lightfield, the lightfield DSI, \( V(k, l, \alpha) \), is a parabola on the variable \( \alpha \) with minimum at local disparity and curvature that grows with viewpoint image gradient:

\[ V(k, l, \alpha) = C(k, l) (\alpha - d(k, l))^2 \]

where \( C_{kl} \), the curvature, is directly related to the gradient in the viewpoint images.

The curvature grows with the intensity of the gradient of the feature on the viewpoint images, i.e., \( \nabla_k L, \nabla_l L \). This suggests that the curvature is a reasonable measure of confidence for a disparity measured by this algorithm. An important consequence of this fact, is that there needs to be a strong enough gradient in the viewpoint images for disparity to be measurable this way.

A way to estimate the curvature \( C(k, l) \) is by fitting a parabola using a least-squares minimization,

\[ \arg \min_{C_{kl}, b_{kl}, m_{kl}} \left\{ \sum_{\alpha} \left( C_{kl} \alpha^2 + b_{kl} \alpha + m_{kl} - V(k, l, \alpha) \right)^2 \right\} \]

where the values for \( \alpha \) are chosen in a contiguous segment around the minimum found.

4) Optimizing Disparity Estimation: The method presented in this section assumes pre-determined values for disparity. This leads to a discretization, i.e. a "staircase" effect\(^1\). It is possible to solve this by setting the disparity of a point \((k, l)\) to be the minimum of the parabola defined by parameters in Equation (15), resulting in

\[ d(k, l) = -\frac{2b(k, l)}{C(k, l)} \]

IV. EXPERIMENTS

In this chapter the methods previously described are tested on both synthetic and real lightfield datasets. The description of the datasets is presented before the reconstruction experiments. The reconstruction experiments are organized as one section for each method. Within these sections, each subsection describes the results of using that method for each applicable dataset.

A. Datasets Used in the Experiments

In this section the datasets used in the experiments are detailed. The datasets used can be divided in two types - synthetic and real. The synthetic datasets allow for a more systematic approach to testing the method, while the real datasets demonstrate their effectiveness in practice. Some of these datasets consist of sets of rays, grouped by which feature they correspond to; while others consist of 4D lightfield images.

1) Synthetic Point-Projections: To test the theoretical accuracy of the lightfield camera, synthetic sets of rays were produced, corresponding to specified 3D points generated using the projection algorithm described in Section II-D. The only reconstruction method described in this paper that takes sets of rays as input is the point-based method, and as such that is the one that was used for this dataset. For the purposes of defining the virtual camera, the intrinsic matrix \( H \) was obtained using the parameters provided by Dansereau et al. [8].

The accuracy at each depth was evaluated by randomly selecting \( P = 500 \) points from the field of view of the plenoptic camera and computing the reconstruction error after projection and reconstruction using the two variants defined in Section III-A. The results of testing on this dataset can be found in Section IV-B1

\(^1\)This effect is shown experimentally in Figure 14.
2) Synthetic Lightfield: For further testing of the reconstruction algorithms, a synthetic scenario was created in order to create synthetic 4D lightfield images. In our case, the scene was built in the Virtual Reality Markup Language (VRML) and the images were captured using the Matlab Virtual Reality (VR) toolbox. Given the VRML scenario, an array of images is captured, and then arranged into a lightfield. The intrinsics matrix is obtained using J. Y. Bouguet’s calibration toolbox [5], since Matlab VR tools do not clarify what the intrinsics are when an image is captured.

The VRML scene is composed of a horizontal plane (i.e. normal to the y direction, considered as up) with a grassy texture upon which rests a box with a wooden texture. The center of the coordinate system corresponds to the center of the box, which has dimensions $0.6 \times 0.04 \times 0.1$ in the $x$, $y$ and $z$ directions, respectively. Floating above this box are two spheres. One sphere is textured to look like the Earth, and has its center at coordinates $(-0.05, 0.15, 0)$ and radius 0.1. The other is textured with the moon’s surface, centered on $(0.15, 0.15, 0)$ and radius 0.02. The $H$ matrix can be obtained by applying Equation (3).

![Central viewpoint and different view of the scene](image)

(a) Central viewpoint  (b) Different view of the scene

Fig. 7. Captured Scene. Central viewpoint of size $378 \times 378$ pixels, of a $11 \times 11$ set of viewpoint images forming the captured lightfield (a). Conventional image from a different point of view (b).

3) Calibration Lightfields: The reconstruction and projection algorithms defined in this work require the intrinsic properties of a lightfield camera. In order to obtain these parameters, the camera must be calibrated using a calibration dataset along with Dansereau’s toolbox [6]. For this purpose, two calibration datasets were used, some publicly available datasets provided by Dansereau et al. [8], and some that were captured by us.

The public calibration datasets provided by Dansereau et al. [8], were collected using a 1st generation Lytro camera identical to the one used in this work. These calibration datasets encompass different calibration grid sizes and different poses at different depths to calibrate the plenoptic camera (Table I).

![Graph showing results of performing calibrations on the datasets focused at various distances](image)

Fig. 8. Results of performing calibrations on the datasets focused at various distances. In the graphs are the coefficient parameters, averaged between the horizontal and vertical counterparts. Error bars show the difference between them.

4) Test-Scenario Lightfield: To study the practical applicability of these methods, real images of scenes with some known physical properties were used. The methods are applied onto the images, to test if the reconstructed scene is similar to the real object. Example scenes would be those composed of objects at well defined distances from each other, or cubes with known side length.

The lightfield image used, captured by a Lytro model F01 (first generation plenoptic camera), is represented in Figure 9.a. The scene is composed of two checkerboard textured cubes, stacked on top of each other. Under the cubes, to the left, is a page with text, and to the right, a calendar. These objects were chosen because of the high amount of image texture, that lends well to depth reconstruction.

In order to avoid darker viewpoints, due e.g. to vignetting, a “ring” composed of the outermost viewpoints was rejected. Furthermore, to deal with stuck pixels the viewpoint images are processed by a 2D median filter.

5) Depth Accuracy Estimation Dataset: Several images of a set of cubes were captured at varying distances from the camera. They were taken when the calibration dataset was

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of Cells</th>
<th>Cell Size (mm x mm)</th>
<th>Number of Poses</th>
<th>Maximum Distance from Camera (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>19 x 19</td>
<td>3.61 x 3.61</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>B</td>
<td>19 x 19</td>
<td>3.61 x 3.61</td>
<td>18</td>
<td>20</td>
</tr>
<tr>
<td>C</td>
<td>19 x 19</td>
<td>7.22 x 7.22</td>
<td>12</td>
<td>30</td>
</tr>
</tbody>
</table>
being captured, and the camera was focused at a distance of 30cm. The distance between the cubes and the camera was varied between 0.05m and 1.3m. The purpose of these images is to perform an analysis of the camera’s depth reconstruction abilities, the results being shown in Section IV-E.

B. Point Based Reconstruction

1) Synthetic sets of rays: In Section III we described a projection model and a reconstruction methodology for a given point in the scene. In this section, we evaluate the depth estimation accuracy using these models to define the depth range for these cameras.

The reconstruction error is defined as the distance between the reconstructed point and the generated point in the object space. The mean reconstruction error $r_c$ is defined as

$$ r_c = \frac{1}{P} \sum_{i=1}^{P} \| m_i - \hat{m}_i \| . \quad (17) $$

In Figure 10.a we compare the estimated depth with the ground truth depth in the camera coordinate system while in Figure 10.b, the reconstruction error defined in Equation (17) is depicted. The ground truth depth for the generated points correspond to their depth values before projection and reconstruction. Assuming that the deviation starts to occur when the reconstruction error $r_c$ normalized by the ground truth depth is greater than 10%, the estimates start to deviate at 0.7 m, 1.2 m and 1.0 m for datasets A, B and C, respectively. From 10.a, one can see that the mean value for the estimated depth is in accordance with the ground truth depth. Nonetheless, the uncertainty associated with the reconstruction error increases significantly from depth values 1.5 m for dataset C, 1.4 m for dataset B, and 1.0 m for datasets A. This suggests that the standard plenoptic camera cannot reconstruct accurately points that are at a depth greater than 1.5 m, and that the depth range depends on the focused depth of the camera.

2) Synthetic Lightfield Points-based Reconstruction: The point based algorithm was also tested on lightfield images, namely on the same VRML lightfield image that is to be used to test the other methods. As previously mentioned, doing this requires feature detection and matching. For these purposes, a SIFT feature detector and matcher was employed. Because the feature matcher was designed for sequences of images, the viewpoints were processed in a sequence forming a spiral pattern. In Figure 11, the $(k, l)$ positions of some of the detected features are over the image of the central viewpoint, connected in the same order as they were processed, leading to the square spiral shapes.

Only features that were detected in 4 or more viewpoints were kept, resulting in 424 features. Of those, 223 were detected on all viewpoints. The resulting point cloud is displayed in Figure 11. There are too few points to discern the features of the scene, but it is possible to see that the grassy plane underneath is further away at the top due to perspective, and that an object (the Earth globe) is on top of it.

To get a numerical measurement of the error, the root mean square of the error in depth was calculated. The formula for the error is Equation (18), the root mean squared error (RMSE) of depth. The result of this experiment was 0.0554, and the depth of the elements of the scene vary between 0.63 and 1.15. Of the total features detected, only 14 were had a relative error above 10%. This corresponds to 3.3% of all detected features.

$$ RMSE = \sqrt{\frac{1}{P} \sum_{i=1}^{P} (z_i - \hat{z}_i)^2} . \quad (18) $$

C. Gradient Based Reconstruction

1) Synthetic Lightfield: The gradient based algorithm was also tested with the synthetic lightfield. The accuracy of the algorithm was evaluated by comparing with the ground truth obtained from the information used to define the scene in the
VRML file. The results are presented in Figure 12. The relative
error is below 10% for 88% of the image, and the RMS of the
error is 0.0574, where the globe has a diameter of 0.2. If the
grass background plane is ignored, the RMS becomes 0.0841.
The resulting point cloud approximates well the scene, espe-
cially large planar features such as the grassy plane. However,
more detailed areas are significantly distorted and in areas
where occlusion occurs, "ramps" seem to form, connecting the
sphere to the grassy background. This is a result of averaging
in the i or j direction, which averages structure tensors of two
distinct features, significantly separated in space.

In order to quantify the reconstruction error, points have
been extracted from the point cloud corresponding to the three
visible faces of the cube at the top (black squares). The results
from each face were then fitted into a plane using a least
squares method. The RMS of the distances between each
point and the fitted plane was then obtained. The results are
4.1mm, 5.6mm and 3.6mm for the left, right and top faces,
respectively. Furthermore, the edge lengths for the faces are
close to 6cm, the real dimensions for the object.

To test the performance of the photo-similarity based
algorithm, it was applied to the same real image the gradient
based method was also applied to. The resulting reconstructed
scene is clearly organized into layers, each at one of the depths corresponding
to one of the sampled disparity values by the algorithm.

1) Synthetic Lightfield: The algorithm was applied to the
VRML image using the default disparity step of 0.1. The
optimization described in Section III-C4 was not applied. The
resulting 3D point cloud can be seen in Figure 14.a. Because
the step was too large, the reconstructed scene is significantly
organized into layers, each at one of the depths corresponding
to the step values. However, the quality of the reconstruction improves
significantly nonetheless. The resulting reconstructed scene
can be seen in Figure 14.b, where the layering effect has also
disappeared.

2) Test-Scenario Lightfield: The photo-similarity based
method was also applied to the same real image the gradient
based reconstruction method was tested on. The results are
similar, with the only clear difference being that there is less
distortion at the boundaries of objects due to occlusion. This
happens because in this method there is not an averaging of
data that might correspond to different features at different
depths, like in the gradient based method.

Once again, points from each of the three visible faces of
the cube at the top (black squares) were extracted and then
fitted into a plane using a least squares method. The RMS of the
distances between each point and the fitted plane was then obtained. The results are
3.1mm, 4.5mm and 3.3mm for the left, right and top faces,
respectively. Furthermore, the edge lengths for the faces are
close to 6cm, the real dimensions for the object.

E. Comparison of Reconstruction Methods

In order to compare both methods, various tests were
performed. Firstly, in Table II, are the results obtained using
synthetic lightfields. Based on these metrics alone, the photo-
similarity metric appears to be the superior method, producing
results much closer to the ground truth than both other
methods. On the other hand, it is much slower than the other

D. Photo Similarity Based Reconstruction

The photo similarity based algorithm was applied to the
same datasets. The results are analyzed in this section.

2) Test-Scenario Lightfield: The lightfield image of the
cubes was passed as input to the gradient based algorithm.
The resulting depth map is presented in Figure 13. It correctly
displays lower depth values in regions of the scene that are
closer, such as the cube, and higher in regions that are further,
such as the monitor. Moreover, basic geometry is preserved.

In order to compare both methods, various tests were
performed. Firstly, in Table II, are the results obtained using
synthetic lightfields. Based on these metrics alone, the photo-
similarity metric appears to be the superior method, producing
results much closer to the ground truth than both other
methods. On the other hand, it is much slower than the other

2) Test-Scenario Lightfield: The lightfield image of the
cubes was passed as input to the gradient based algorithm.
The resulting depth map is presented in Figure 13. It correctly
displays lower depth values in regions of the scene that are
closer, such as the cube, and higher in regions that are further,
such as the monitor. Moreover, basic geometry is preserved.

In order to compare both methods, various tests were
performed. Firstly, in Table II, are the results obtained using
synthetic lightfields. Based on these metrics alone, the photo-
similarity metric appears to be the superior method, producing
results much closer to the ground truth than both other
methods. On the other hand, it is much slower than the other

2) Test-Scenario Lightfield: The lightfield image of the
cubes was passed as input to the gradient based algorithm.
The resulting depth map is presented in Figure 13. It correctly
displays lower depth values in regions of the scene that are
closer, such as the cube, and higher in regions that are further,
such as the monitor. Moreover, basic geometry is preserved.

In order to compare both methods, various tests were
performed. Firstly, in Table II, are the results obtained using
synthetic lightfields. Based on these metrics alone, the photo-
similarity metric appears to be the superior method, producing
results much closer to the ground truth than both other
methods. On the other hand, it is much slower than the other

2) Test-Scenario Lightfield: The lightfield image of the
cubes was passed as input to the gradient based algorithm.
The resulting depth map is presented in Figure 13. It correctly
displays lower depth values in regions of the scene that are
closer, such as the cube, and higher in regions that are further,
such as the monitor. Moreover, basic geometry is preserved.

In order to compare both methods, various tests were
performed. Firstly, in Table II, are the results obtained using
synthetic lightfields. Based on these metrics alone, the photo-
similarity metric appears to be the superior method, producing
results much closer to the ground truth than both other
methods. On the other hand, it is much slower than the other

2) Test-Scenario Lightfield: The lightfield image of the
cubes was passed as input to the gradient based algorithm.
The resulting depth map is presented in Figure 13. It correctly
displays lower depth values in regions of the scene that are
closer, such as the cube, and higher in regions that are further,
such as the monitor. Moreover, basic geometry is preserved.

In order to compare both methods, various tests were
performed. Firstly, in Table II, are the results obtained using
synthetic lightfields. Based on these metrics alone, the photo-
similarity metric appears to be the superior method, producing
results much closer to the ground truth than both other
methods. On the other hand, it is much slower than the other

methods. However, it is possible that CPU implementations make comparable the computation times of the methods. In the columns of Table II we have the root mean squared error calculated using Equation 18, metric BadDepth, i.e., the number of depth values with a relative error above 10%, and the number of estimated features.

<table>
<thead>
<tr>
<th>Method</th>
<th>RMSE</th>
<th>BadDepth</th>
<th>N estimated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Point Based</td>
<td>0.0554</td>
<td>5.3%</td>
<td>424</td>
</tr>
<tr>
<td>Gradient Based</td>
<td>0.0674</td>
<td>12.43%</td>
<td>141 376</td>
</tr>
<tr>
<td>Photo Similarity Based</td>
<td>0.0169</td>
<td>0.6%</td>
<td>133 876</td>
</tr>
</tbody>
</table>

The methods were also applied on real images, from the dataset from Subsection IV-A5, namely the lightfield images captured when the camera was focused at a depth of 30cm.

![Fig. 15. Depth reconstruction for cubes at various distances from the camera.](image)

The depth values corresponding to the cubes were extracted and the results are in Figure 15. It can be observed that both methods have similar results, with the gradient based method being somewhat better. The difference in the results compared to the previous experiment may be due to, among other things, the lack of sharp variations in depth, which produce occlusion problems that affect the photo similarity method less than the gradient based one.

V. CONCLUSION AND FUTURE WORK

The work described in this thesis consisted in the study of the properties of plenoptic cameras, and in the use of these cameras to perform metric reconstructions of scenes from lightfield images. The results were convincing. The proposed methodologies are quite successful at reconstructing 3D data, structure, despite some distortion being visible. Geometry is mostly kept in a recognizable form after reconstruction, and the measured dimensions of the objects were similar to their real values.

As further work, the option of combining lightfields into lightfield panoramas as a way to improve depth estimation accuracy. Furthermore, it would be interesting to investigate how well these cameras are suited for SLAM applications compared to traditional depth cameras.