Approaching Image Processing and Computer Vision Problems via Light Field Imaging

Garcia Moll, Clara

Curs 2018-2019

Director: COLOMA BALLESTER I PATRICIA VITORIA

GRAU EN ENGINYERIA DE SISTEMES AUDIOVISUALS

upf. Universitat Pompeu Fabra Barcelona
Escola Superior Politècnica

Treball de Fi de Grau

Approaching Image Processing and Computer Vision Problems via Light Field Imaging

Clara Garcia Moll

BACHELOR THESIS UPF / YEAR 2019

SUPERVISOR
Coloma Ballester and Patricia Vitoria
Image Processing Group

Universitat Pompeu Fabra
Barcelona

Als meus pares,

Acknowledgements

First of all, I would like to express my deepest gratitude to my supervisors Dr. Coloma Ballester and Patricia Vitoria for their continuous support and guidance throughout the elaboration of this project and for introducing me to the field of image processing.

I am also grateful to everyone in the Image Processing and Computer Vision Group. I feel privileged to have learned from this group of professionals and to make me enjoyed their lessons.

I would also like to thank W.Williem to introduce me in light field imaging and for those amusing lessons.

Many thanks to my friends for the great times we get, for their company and their support. And also, the rest of the fantastic people I have met in the past four years at Pompeu Fabra University that make this stage very pleasant.

Finally, a special thanks to my parents, Lluis Garcia and Maria Moll who are, for sure, the best teachers I ever have had. Thanks for giving me the opportunity to make it possible and their constant support.

Abstract

Light field imaging is an emerging technology which allows capturing an image with richer information than just a two-dimensional image. Traditional cameras capture only light intensity. In contrast, light field cameras, also known as plenoptic cameras, are able not only to capture light intensity but also its reaching direction. This additional information allows for new strategies permitting to come up with a solution of several classical problems in image processing and computer vision such us image refocusing, occlusion detection, depth estimation, and 3D reconstruction. However, this new technology, known as light field imaging, has even more challenges than the traditional one. Mostly, these challenges are focused on the difficulty of dealing with the huge quantity of data when a light field image is captured, the way to compress this data and also how to display these kinds of images. On this project, develope some methods to show the results obtained using light field images and discusse the advantages and improvements.

Resum

El camp del *light field imaging* és molt nou i té la possibilitat de capturar molta més informació del món real que no la que pot capturar una imatge bidimensional. Les càmeres tradicionals capturen únicament la intensitat de la llum incident. D'altra banda, les *light field cameras*, també conegudes com a càmeres plenòptiques, no tan sols poden capturar la intensitat de la llum sinó que també la direcció de la qual provenen els rajos. Aquesta informació addicional permet l'estudi de noves estratègies per solucionar problemes del camp del processament d'imatge o visió de computador tals com a *image refocusing*, *occlusion detection*, *depth estimation*, i *3D reconstruction*. No obstant, aquesta tecnologia, coneguda com *light field imaging*, presenta nous problemes respecte a la fotografia tradicional. La gran majoria d'ells centrats en la dificultat de tractar amb la gran quantitat d'informació que ens proporciones les *light field images*, així com els problemes de compressió o de *display*. En aquest projecte, desenvolupar alguns metodes i mostrar els resultats obtinguts utilitzant *light field images* i comentar els avantatges i les possibles millores.

Contents

List of Figures **xii**

1 INTRODUCTION **1**
1.1 Motivation 3
1.2 Project Goals 3
1.3 Project Structure 4

2 PREVIOUS CONCEPTS **5**

3 RELATED WORK **11**

4 DATASET **13**

5 LIGHT FIELD IMAGE APPLICATIONS **15**
5.1 Contribution 16
5.2 Image Refocus 17
5.2.1 Related Work 17
5.2.2 Method 18
5.2.2.1 First step: Finding correspondences between images 18
5.2.2.2 Second Step: Create the Refocus Image 19
5.2.3 Results and Discussion 20
5.3 Depth Estimation 23
5.3.1 Related Work 23
5.3.2 Method 25
5.3.2.1 Angular Entropy Response 26
5.3.2.2 Adaptive Defocus Response 26
5.3.3 Results and Discussion 27
5.4 Occlusion Detection 31
5.4.1 Related Work 32
5.4.2 Method 33

5.4.2.1 Fixed Threshold 35
5.4.2.2 Adaptive Threshold 35
5.4.3 Results and Discussion 36
5.5 3D Reconstruction 39
5.5.1 Related Work 39
5.5.2 Method 40
5.5.2.1 Zhang's Method 40
5.5.2.2 Approach based on Light Field Imaging 41
5.5.3 Results and Discussion 46

6 CONCLUSIONS AND FUTURE WORK 51
6.1 Achievements 51
6.2 Applications Conclusions 51
6.3 Future Work and Improvements 52

List of Figures

1.1 The mechanism of the plenoptic camera. First, it has a sensor, which is the one that captures the light. Secondly, the main lens that from one incoming ray creates a ray for each miro-lens of the array, which is the part that makes plenoptic cameras able to capture the extra information. 2

1.2 An example of some images obtained with the plenoptic camera where there is a small variation on the point of view. 2

2.1 This image represents how the micro-lens array works. The array of micro-lens is placed between the lens and the sensor. From one incoming ray the lens create a ray for each micro-lens. 5

2.2 Two subaperture images of a scene captured by a light field camera. 6

2.3 Each of these images are angular patches. 7

2.4 3D point projected on 2 images plane under different perspective. 8

2.5 A graphical example of the triangulation method. 9

4.1 Subaperture images from the 4D light field benchmark dataset, which is the one used in this project for all the different tasks and applications. 14

5.1 3 images which the depth on focus is a different on each image. Image courtesy [Bando and Nishita, 2007]. 17

5.2 It is seen the different angular patches by zooming on the sheared light field image. 19

5.3 Zoom on one part of the image that is not on focus 20

5.4 Zoom on one part of the image that is on focus 20

5.5 Two examples of the refocus image solution. On the first row the chosen disparity is the lower one, on the second row the chosen disparity is the higher one, finally, on the third row the chosen disparity is one in the middle. 22

5.6 First row shows the 3D reconstruction and on the second row a depth map. Images from [Im et al., 2019] 23

5.7 Two examples of depth maps. On the first and the second rows using the Angular Entropy Response, on the third row using the Adaptive Defocus Response and on the last row using both data costs. 29
5.8 *Buddha* image 30
5.9 An example of a tracking method robust to occlusions. Image courtesy [Tang et al., 2014]. 31
5.10 An example of a 3D reconstruction robust to occlusions. Image courtesy [Tuấn Trần et al., 2018] 32
5.11 An example of a 3D reconstruction not robust to occlusions. Image courtesy [Tuấn Trần et al., 2018] 32
5.12 An example of an accurated depth map. Image courtesy [Huguet, 2009]. 33
5.13 Example of the histograms of different angular patches. Image source: [Williem and Kyu Park, 2016]. In this Figure, each row corresponds to the same pixel but using different depth. In the left column a non-occluded pixel is shown and the right column shows an occluded pixel. 34
5.14 Using 3 different values. From left to right thr = 3.1, thr = 2.1 and thr = 1. 36
5.15 Using 3 different values. From left to right thr = 0.2, thr = 0.25 and thr = 0.9. 37
5.16 Using 3 different values. From left to right thrfix = 2.1 thradap = 1.5 , thrfix = 2.3 thradap = 0.5, thrfix = 1.5 thradap = 0.5. 37
5.17 Example of 3D face Reconstruction using light field images. Image courtesy of [Feng et al., 2018]. 39
5.18 The 4 corners chose to compute the correspondences 43
5.19 The different image planes in order to obtain the angular information of 1 light field image 44
5.20 The 121 points which are the correspondences computed in the first step 44
5.21 Some matchings between two subaperture images. As we can see, there are some mismatchings in the flowers from the background due to the repetition of the flowers. 45
5.22 The 3D reconstruction compared with the 2D image. 47
5.23 Examples from different points of view of the 3D reconstruction using only two cameras. 49
5.24 Examples from different points of view of the 3D reconstruction using four cameras. 50

Chapter 1

INTRODUCTION

Emerging developments in the camera industry have provided additional information compared to traditional photography able to facilitate classical computer vision tasks. One of these improvements has been made in the photography field. Until now, most of the people only knew about traditional photography, which is only able to capture two-dimensional images containing spatial information. However, opposed to conventional photography, it appeared plenoptic cameras which are able to capture additional information of the scene, obtaining not only spatial information but also angular information referred to the angle light rays reach a given point. With the emergence of this development, a new field of study in the image processing and computer vision field called light field imaging has emerged.

Light field images are taken by plenoptic cameras. Plenoptic cameras, also known as light field cameras, have an internal mechanism able to capture complete information about the scene. This mechanism is composed of a traditional camera lens (the so-called main lens), a sensor, and an array of micro-lens, which is the novelty. This structure could be seen in Figure 1.1.

Furthermore, the resulting image taken by a plenoptic camera is an array of images each one having a different perspective of the scene.

Intuitively, it could be like, using the traditional camera, but by taking the image from different but closer points view. Additionally, plenoptic cameras provide us the exact position of each view, hard to control by conventional cameras.

An example of some of the images resulting from the output array of images from a light field camera is shown in Figure 1.2. On this image it is shown some variations such as the distance between the bottom margin of the photo and the wood, the distance between the top margin of the photo and the buddha, or also the part where the dice occludes the wood.

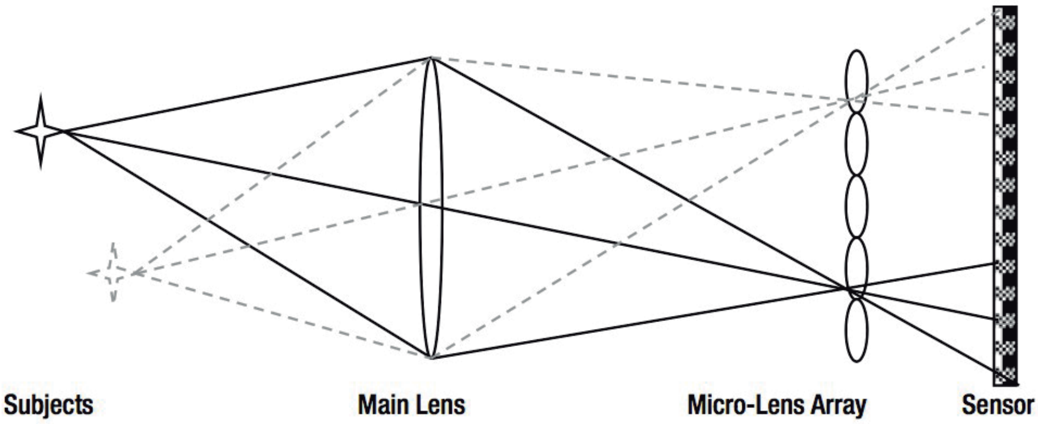


Figure 1.1: The mechanism of the plenoptic camera. First, it has a sensor, which is the one that captures the light. Secondly, the main lens that from one incoming ray creates a ray for each miro-lens of the array, which is the part that makes plenoptic cameras able to capture the extra information.

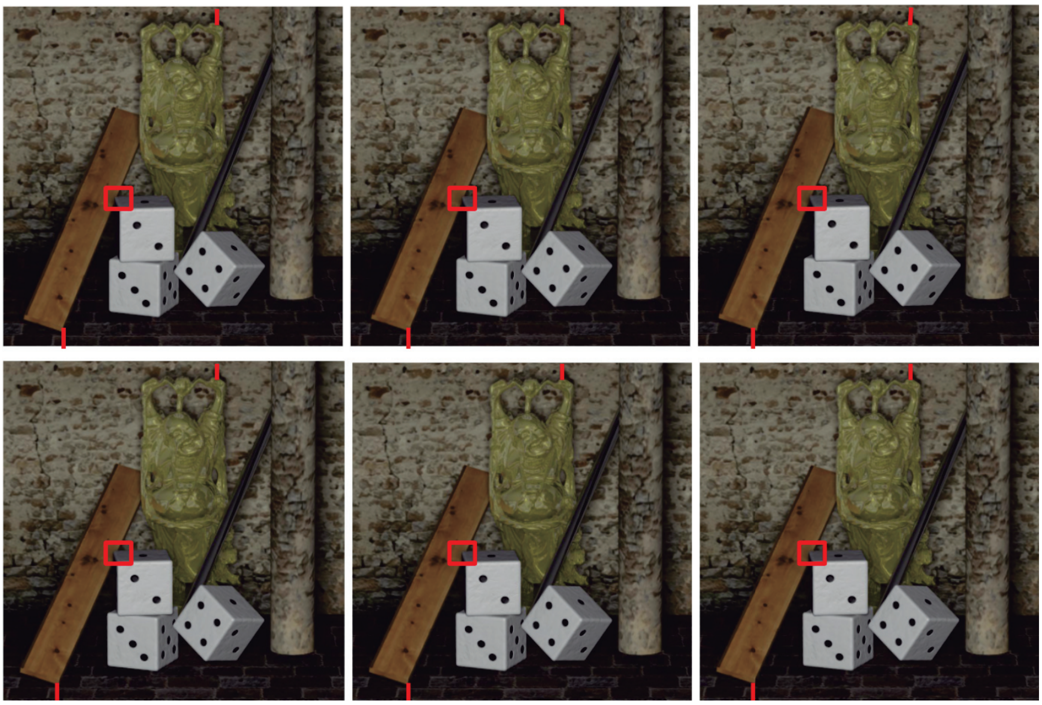


Figure 1.2: An example of some images obtained with the plenoptic camera where there is a small variation on the point of view.

Moreover, the additional information captured makes it possible to solve some challenging tasks in the computer vision and image processing fields such as 3D reconstruction [Feng et al., 2018], image matting [Cho et al., 2014], segmentation [Zhu et al., 2017], image refocusing [Fu et al., 2015], depth estimation [Williem and Kyu Park, 2016], among others.
However, there are still many challenges such as the improvement of the methods proposed in the literature for the mentioned previous tasks, or other challenges which remain mainly unsolved such as how to display this kind of images, how to compress them efficiently, among others. All of these challenges appeared due to the acquisition of high-dimensional data and the difficulty of managing this huge quantity of information.

1.1 Motivation

There are many applications of image processing and computer vision very useful in biomedical fields such as depth estimation or 3D reconstruction. However, in some cases these application have not accurated results due to the lack of information. Hence, the novel field of light field imaging, which captures more information from a single scene and allows dealing with problems of image processing and computer vision from a different point of view and strategies.

Considering all these aspects, I got attracted by the idea of contributing to the research of this novel field by studying the improvements on different applications of image processing and computer vision.

1.2 Project Goals

The main goal of this project is to widen the knowledge about light field imaging and to go deeper into analyzing some classical problems of image processing and computer vision field. The objectives are listed below.

1. Understand the information provided by a light field camera and how to use this information in order to solve some problems of image processing and computer vision fields.

2. Understand how could be done image refocusing using light field images.

3. Learn how it is possible to do an algorithm of depth estimation robust to occlusions using light field images.

4. Obtain an occlusion map using the idea proposed by W.Williem [Williem and Kyu Park, 2016] focused on the computation of the entropy. On this occlusion map will be marked those pixels which are occluded in any sub-apperture image.

5. Obtain a 3D Reconstruction using only one light field image. And to take some conclusions of how could be improved the 3D Reconstruction if it will be used more than one light field image.

1.3 Project Structure

In this first Chapter, an introduction about plenoptic cameras has been done, together with the motivation to carry out this research and its goals. Chapter 2 explains some previous concepts which are important to fully understand the algorithms presented in this work. In Chapter 3, a brief description of related work in the field is described. Later on, in Chapter 4, an introduction of the dataset used for all the tasks approached is given. Chapter 5 focuses on some of the applications that light field imaging has, the methods implemented to tackle with each of the tasks, as well as, some results obtained for each of the different applications. Finally, Chapter 6 includes a discussion and the conclusions extracted after all the developed project.

Chapter 2

PREVIOUS CONCEPTS

In this chapter, some fundamental concepts that will be needed to understand the remaining parts of this project will be detailed.

- **Plenoptic or Light-field Camera:** This camera uses a different mechanism formed by the traditional camera lens, the sensor, and a micro-lens array as it is seen in Figure 2.1. The function of the micro-lens array is to allow the sensor to record additional information about the incoming rays, including the light coming from different distances.

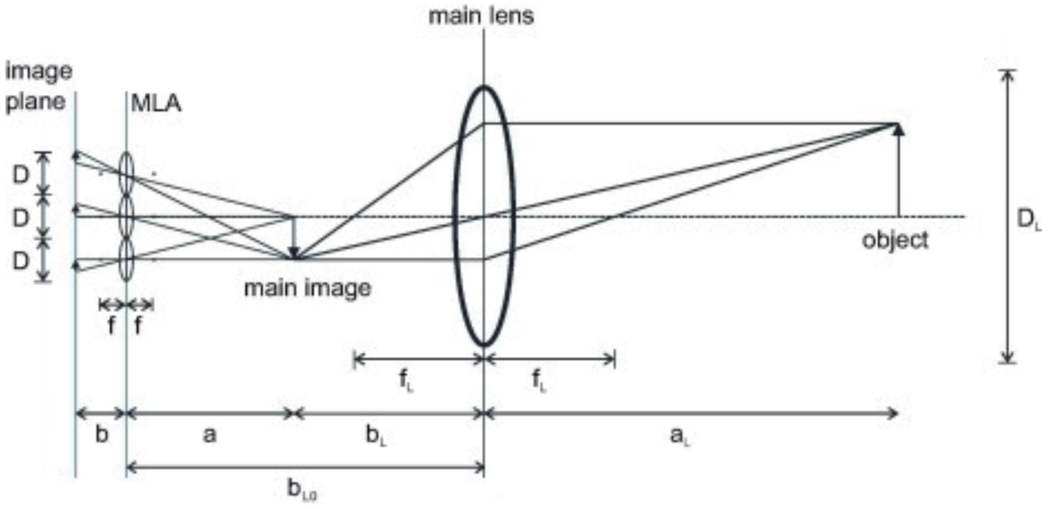


Figure 2.1: This image represents how the micro-lens array works. The array of micro-lens is placed between the lens and the sensor. From one incoming ray the lens create a ray for each micro-lens.

- **Light Field Images:** Images obtained by a plenoptic camera. These images contain spatial and angular information. A light field image is mathematically defined as $L(x, y, u, v)$ where (x, y) refer to the spatial coordinates and (u, v) refer to the angular ones.

  - **Subaperture Images:** The subaperture images are images created by keeping the angular coordinates constant and varying the spatial coordinates. An example of some subaperture images of the same captured scene is shown in Figure 2.2. In both images there are some parts where it is noticeable the difference of the points of view.

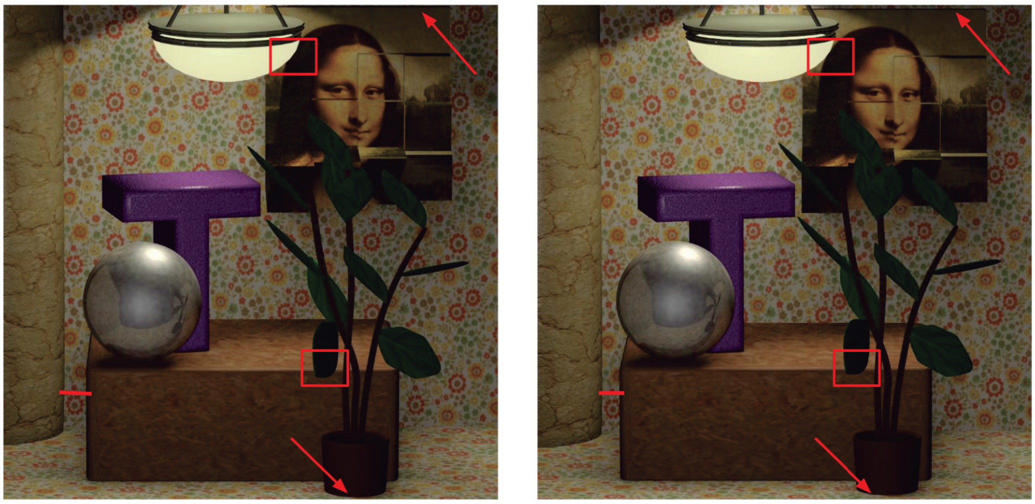


Figure 2.2: Two subaperture images of a scene captured by a light field camera.

  - **Angular Patch:** Represent an image which has been created in an opposite way than subaperture images. More precisely, for each angular patch the spatial coordinates are constant while the angular ones are varying. An example of some angular patches obtained from the subaperture images is shown in Figure 2.3.

- **Entropy:** The entropy is a measure of the randomness of a certain distribution P, and it is defined mathematically as

$$H = -\sum_i P_i log(P_i) \tag{2.1}$$

  where P_i denotes the probability that the event i happens.

- **Photo-Consistency:** Photo-Consistency (frequently referred as **color constancy** or **brightness constancy** in the case of graylevel images) is the property of constancy of color of a same scene point but when seen onto images

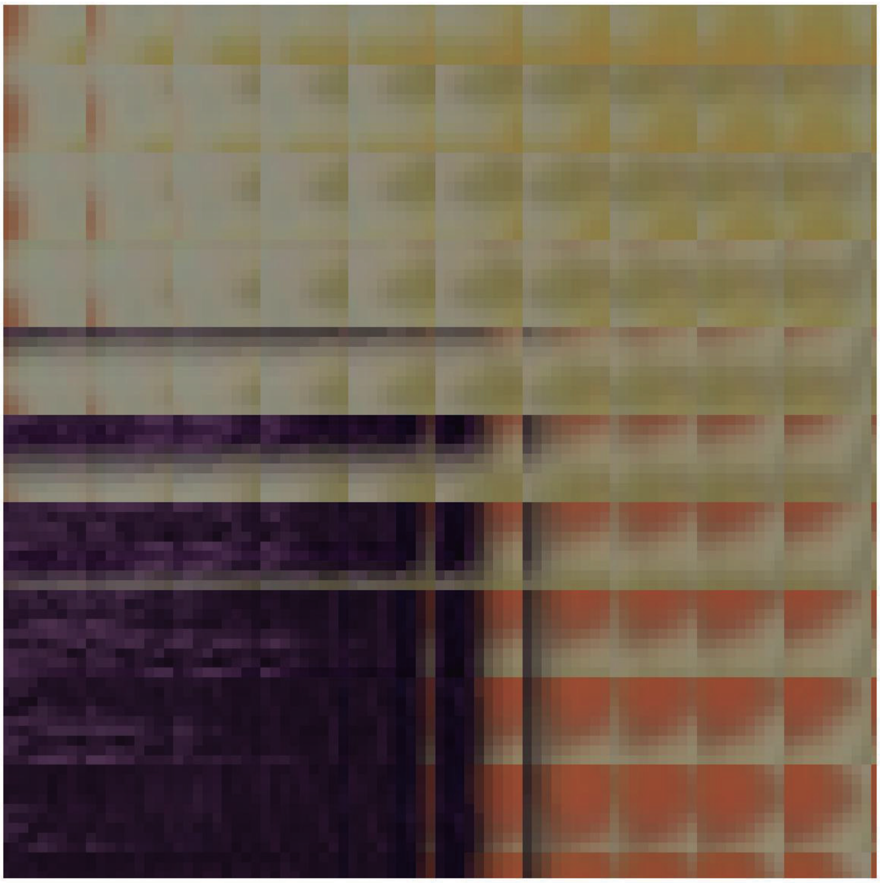


Figure 2.3: Each of these images are angular patches.

representing the captured scene from different points of view or cameras. It is a term used in the computer vision and image processing fields to determine if there is similarity between pixels or patches (a patch is a neighborhood or connected group of pixels) from different photos under different views. In light field imaging, the photo-consistency frequently refers to the similarity of the pixels of the angular patch.

There are some cases where the photo-consistency might be broken; for instance, when there is an occlusion. In occlusion regions, the color-constancy does not hold.

- **Disparity:** Disparity is the distance between points of a conjugate pair when the two images are superimposed. In other words, the displacement between the locations of the two features in the image plane is called the disparity [Jain et al., 1995]. It could be defined mathematically as:

$$x_2 = x_1 + d(x_1) \tag{2.2}$$

where x_1 and x_2 are the points projected on each image plane. In the case of light field imaging x_1 and x_2 are points on different subaperture images and $d(x_1)$ is the disparity for each point in the central subaperture image. This situation is shown in Figure 2.4.

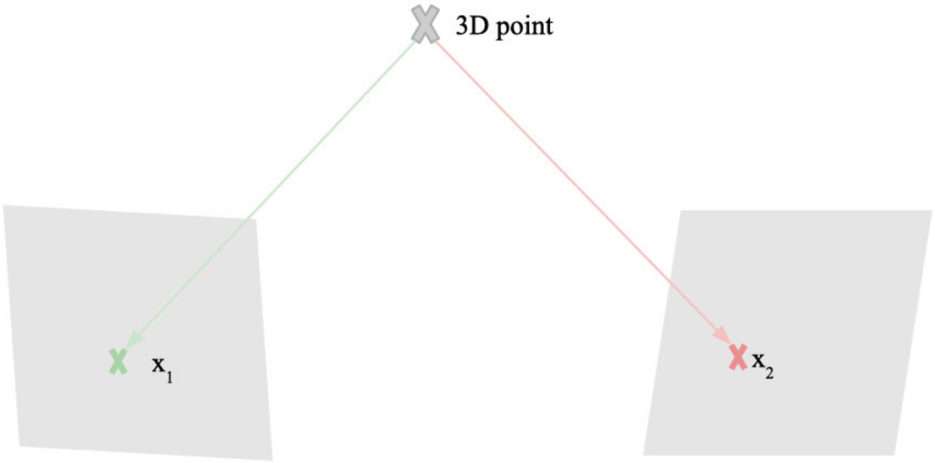


Figure 2.4: 3D point projected on 2 images plane under different perspective.

Moreover, the disparity is inversely proportional to the depth.

- **SIFT:** SIFT (Scale Invariant Feature Transform) is a method of feature extraction used in image processing and computer vision fields. This algorithm aims to extract invariant and highly distinctive features. In order to do it, the descriptors of these features are invariants to image scale and rotation, easing the process of matching it correctly with other images under such transformations, which are related to the change of point of view from which an image has been captured. Moreover, this algorithm is invariant to blur, contrast and illumination changes, noise, occlusions, and invariant to affine transformation.

- **Triangulation:** The main goal of triangulation methods is to estimate the position of a set of points in the 3D space given their projection (2D plane) on two or more images under different points of view. The set of points are correspondences across the various images used. In order to apply the triangulation method, the correspondences and the *Projection matrices* of each camera are needed.

A graphical example of how triangulation method works is shown in Figure 2.5. In this example, it can be seen how the 3D points are generated. It is created by the intersection of the two rays created by the 2D points from different cameras (these 2D points need to be correspondences).

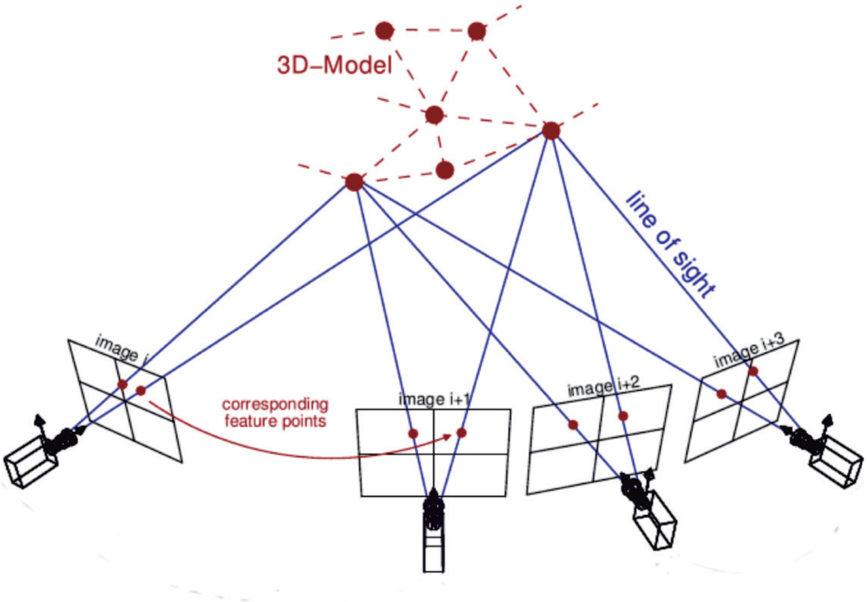


Figure 2.5: A graphical example of the triangulation method.

Chapter 3

RELATED WORK

Nowadays, many studies are focused on light field imaging since it is a novel field with a large number of applications such as depth estimation [Williem and Kyu Park, 2016, Jeon et al., 2015], image matting [Cho et al., 2014], image segmentation [Zhu et al., 2017], and among others. Indeed, the fact that a light field image contains the same scene seen from different but close points of view and, moreover, without dismissing cameras position, involves several advantages such as managing to deal with some problems of image processing and computer vision, being able to capture images under different perspective without the need of several cameras, to mention just a few. Furthermore, this field gives the opportunity to understand better the scene through the information obtained by a camera. Many years ago, the studies were focused on how the function of the plenoptic camera could be described mathematically [Adelson et al., 1991]. Now, the most remarkable studies are focused on different applications such as image refocusing [Fu et al., 2015, Ng, 2005], or occlusion detection, for instance, in order to better understand the image formation process and artificial vision related problems, and this understanding be applied in other methods, and also how the different techniques could be improved.

On the other hand, 3D reconstruction is also a field of computer vision intensely studied in order to be able to recover a 3D scene using several images captured from different points of view. Some methods to do the 3D reconstruction rely on triangulation using the algebraic method [Bardsley and Li,], or Voxel-based methods [Seitz and Dyer, 1999, Kutulakos and Seitz, 2000], which are calibrated methods.

Despite applications based on 3D reconstruction using light field images are not a hot topic, 3D face reconstruction is attracting significant attention [Feng et al., 2018].

Chapter 4

DATASET

As mentioned above, light field imaging is very novel. Possibly as consequence of that, the number of publicly available datasets is limited. In addition, due to limited resources, having a plenoptic camera to take our own images was unfeasible. Finally, we choose to work with the same dataset used in [Williem and Kyu Park, 2016], which is called *4D light field benchmark* [Wanner et al., 2013]. The main reason for that is that the *4D light field benchmark* is a widely used dataset in research and this fact make it easier in order to provide a comparison of our work with the one of others.

4D light field benchmark dataset contains seven light field synthetic images in h5 format. This format is used to store and organize large amounts of data such as the ones obtained with a plenoptic camera. Moreover, a Matlab function called *hdf5info* contained in the *Light Field Toolbox v0.4* allows us to extract the information present in the h5 file, such as the focal length, size of the image, angular and spatial resolution, among others. In order to obtain the subaperture images, we use a function already provided by Matlab called *hdf5read*.

Below, in Figure 4.1 there is an example of some of the different light field images and different subaperture images of the chosen dataset.

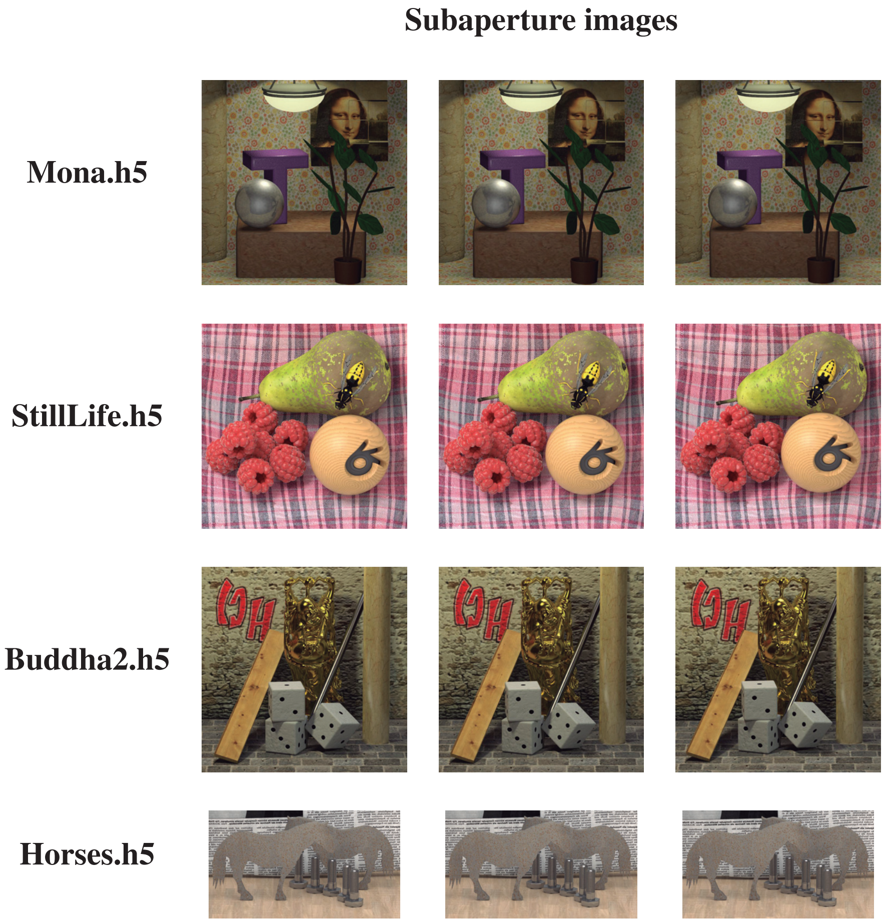


Figure 4.1: Subaperture images from the 4D light field benchmark dataset, which is the one used in this project for all the different tasks and applications.

Chapter 5

LIGHT FIELD IMAGE APPLICATIONS

In this section, first, the contributions and then, the applications on image refocusing, depth estimation, occlusion detection, and 3D reconstruction that have been developed in this project are explained. Each application is structured in four sections. First, a brief introduction and overview of the concrete application; after that, an overview of existing related work on the topic; then the proposed method is explained; and, finally, some results will be shown and a brief discussion on the topic is provided.

The applications of image refocus (detailed in Section 5.2) and depth estimation (explained in Section 5.3) ground on the algorithm proposed by W.Williem [Williem and Kyu Park, 2016]. An approach to occlusion detection is presented in Section 5.4. We propose a method that exploits the concept of entropy proposed by W. Williem in his work on depth estimation [Williem and Kyu Park, 2016]. In particular, as explained in Section 5.4, the method proposed in Williem's work includes a management of occlusions that we propose to leverage to explicitly detect regions of occlusions. Finally, Section 5.5 is devoted to obtain an approximate 3D reconstruction. The proposal is based on the calibration method proposed on [Wang et al., 2018] and also SIFT and triangulation method in order to approximately recover the 3D scene from the light field images of it.

An important point to mention is that all the applications are done using the images from *4D light field benchmark* dataset mentioned in previous Section 4.

5.1 Contribution

In order to learn how it is applied different methods such as image refocusing and depth estimation using light field image we used the method proposed by W. Williem [Williem and Kyu Park, 2016] and also his code.
Moreover, in order to obtain a method of occlusion detection, we use the idea of using the entropy proposed by W. Williem [Williem and Kyu Park, 2016] and we proposed the different thresholds in order to determine if there is an occlusion.
Finally, in order to obtain the 3D reconstruction, the first step is to compute the extrinsic and intrinsic parameters and then do the triangulation. To obtain the parameters, we use the code proposed by [Wang et al., 2018]. Furthermore, to do the triangulation, the code is proposed by us using the knowledge obtained on the subject of 3D Vision.

The contribution of this paper is summarized as follows:

- Keen observation on the method proposed by W. Williem [Williem and Kyu Park, 2016] in order to obtain the first and the second application and detect if any improvement could be done.
- Propose different thresholds in order to obtain an occlusion map.
- Obtain a 3D reconstruction using one light field image by triangulation method and detect if any improvement could be done in order to compute the intrinsic and extrinsic parameters.

5.2 Image Refocus

The term refocusing refers to the fact of focusing a specific part of the image when the photography is already taken. The refocus could be applied only to all objects which are at a certain depth, to all the objects at different depths, or to all the objects at all the depths. This last case would output the so-called *all focus image* and is a very challenging problem in traditional photography with conventional cameras. In fact, this procedure relates to changing the depth of field (DOF).

Although, image refocusing is a problem hard to manage by using only the information provided by a conventional camera in traditional photography, refocusing a light field image is an easier task due to the redundant information that plenoptic cameras provide. This additional data is the angular information, which allows understanding better the scene and as a result, be able to refocus the image. An example of some image refocusing results are shown in Figure 5.1 where in the left hand side image is seen that the depth which is on focus is the nearest one, in the central image the depth on focus is one in the middle and on the right hand side photo the depth on focus is the farthest one.

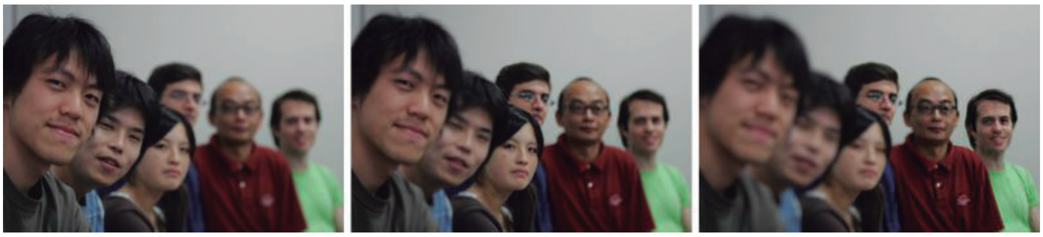


Figure 5.1: 3 images which the depth on focus is a different on each image. Image courtesy [Bando and Nishita, 2007].

This application is widely applied in light field imaging due to the ease to do refocusing using light field images.

5.2.1 Related Work

In 2008, the first study, [Xiao et al., 2008], of light field image refocusing appeared, where the information of the frequency domain is used. Since that moment, some studies related to the obtention of accurated refocus algorithms began. Most of these algorithms are based on the obtention of the distance measurement [Hahne et al., 2016], having an evaluation method using the frequency or the spatial domain.
In 2014 appeared another algorithm of image rectification and refocusing. However, this method was computationally less effective than previous methods.

For this reason, Wenxing Fu et al. [Fu et al., 2015] presented an algorithm using the information of the space domain and the frequency domain, computed using the fast Fourier transform.

5.2.2 Method

The method for light field image refocusing proposed in this project grounds on the method by W. Williem [Williem and Kyu Park, 2016]. This method exploits the information of the spatial domain and the angular information.

A light field image is used as an input from which the different subaperture and the angular images are obtained. The algorithm is done following the two steps presented below in Sections 5.2.2.1 and 5.2.2.2, respectively.

5.2.2.1 First step: Finding correspondences between images

The main goal of this step is to find the correspondences between the central image and the subaperture images. First, each pixel of the light field image $L(x, y, u, v)$ is remapped to the so-called *sheared light field image*, which is an image created in such a way that, for each angular coordinates (u, v) fixed, we extract the disparity-compensated information from the subaperture image associated to (u, v), but having into account a disparity label candidate α. The disparity label candidate is related with the depth resolution and as similar as the disparity of a determined object, more focued will be the object. The sheared light field image is done using the following equation.

$$L_\alpha(x, y, u, v) = L\left(x + \nabla_x(u, \alpha), y + \nabla_y(v, \alpha), u, v\right) \tag{5.1}$$

where (x, y) are the spatial coordinates and (u, v) the angular ones, $\nabla_x(u, \alpha) = (u - u_c)\,\alpha k$ and $\nabla_y(v, \alpha) = (v - v_c)\,\alpha k$. In fact, $\nabla_x(u, \alpha)$ and $\nabla_y(v, \alpha)$ play the role of disparity factors, that is, spatial displacements. Doing this step, it is obtained as many sheared light field image as depth label candidates we have chosen.

The key of the disparity label candidates is that those objects which are on the depth correspondent to disparity, α, will appear on focus. The objects at other depths will not appear on focus and as farther as the α, more blurred are the objects.

Once the sheared light field images are computed, it could be generated the angular patches by extracting the pixels in the angular images. In other words, the sheared light field image is composed of angular images as it is seen in Figure 5.2. On this image, it is shown the different angular patches, which size is $9x9$.

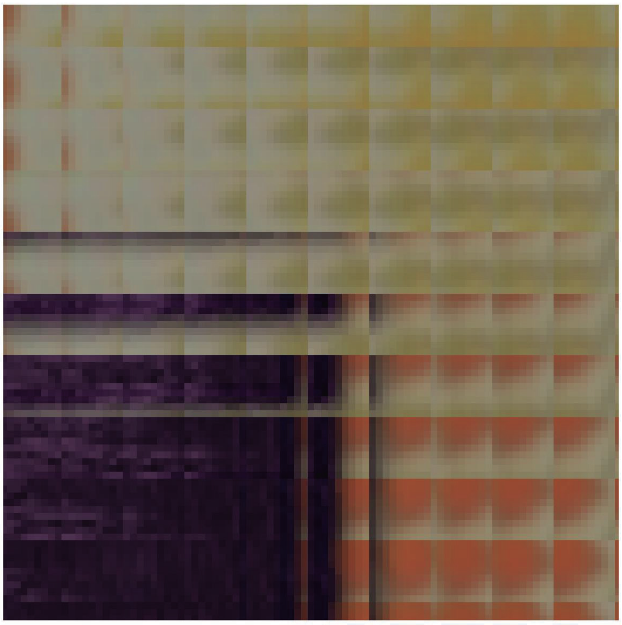


Figure 5.2: It is seen the different angular patches by zooming on the sheared light field image.

5.2.2.2 Second Step: Create the Refocus Image

The goal of this second step is to obtain the refocus image from the angular images. This is done by averaging the values in the angular domain of the sheared light field image, according to the following formula:

$$\overline{L}_\alpha(x, y) = \frac{1}{N} \sum_{u,v} L_\alpha(x, y, u, v) \tag{5.2}$$

If the angular patch belongs to an object which is not in focus (not in the same depth we are looking at using the sheared light field image), the resulting mean of the angular patch will not be in focus. However, if the angular patch belongs to an object on focus, the pixels resultant after doing the mean will be in focus. This difference is seen in Figures 5.3 and 5.4.

As it is seen in Figure 5.3 those objects which are not on focus are not clear on the sheared light field image. Having a look on the zoom which is done on the dice, an object which is not on focus, the edge of the black point is not sharp.

However, Figure 5.4 is a good example which shows that if the object is on focus, the shape is sharper as it could be seen on the dice. Specifically, the edge of the dice, as it could be seen in the zoom, is very sharp. It means that on this Figure the

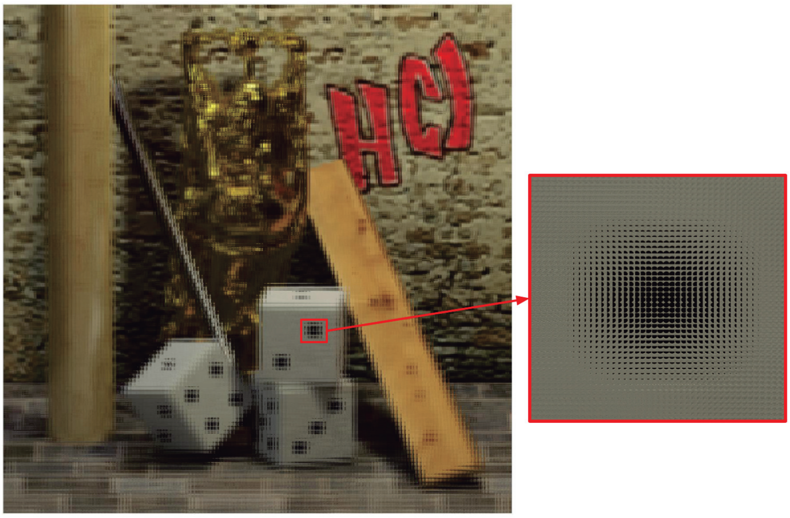


Figure 5.3: Zoom on one part of the image that is not on focus

chosen depth label candidate corresponds to the depth where is placed the dice.

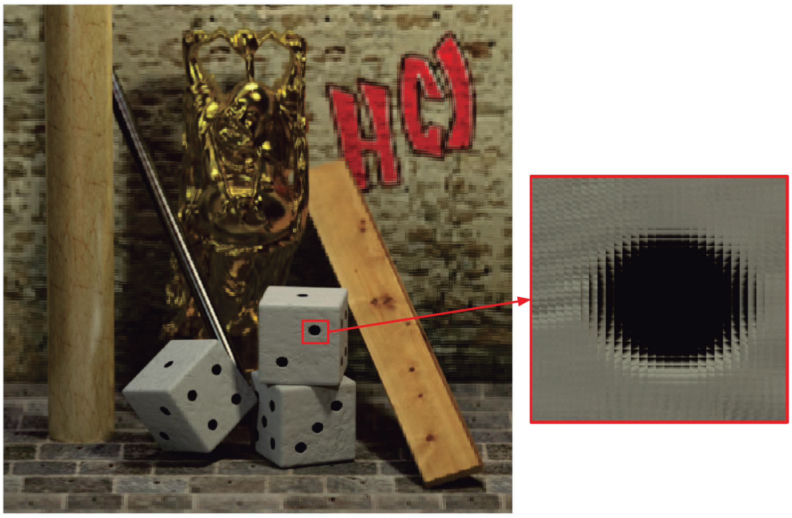


Figure 5.4: Zoom on one part of the image that is on focus

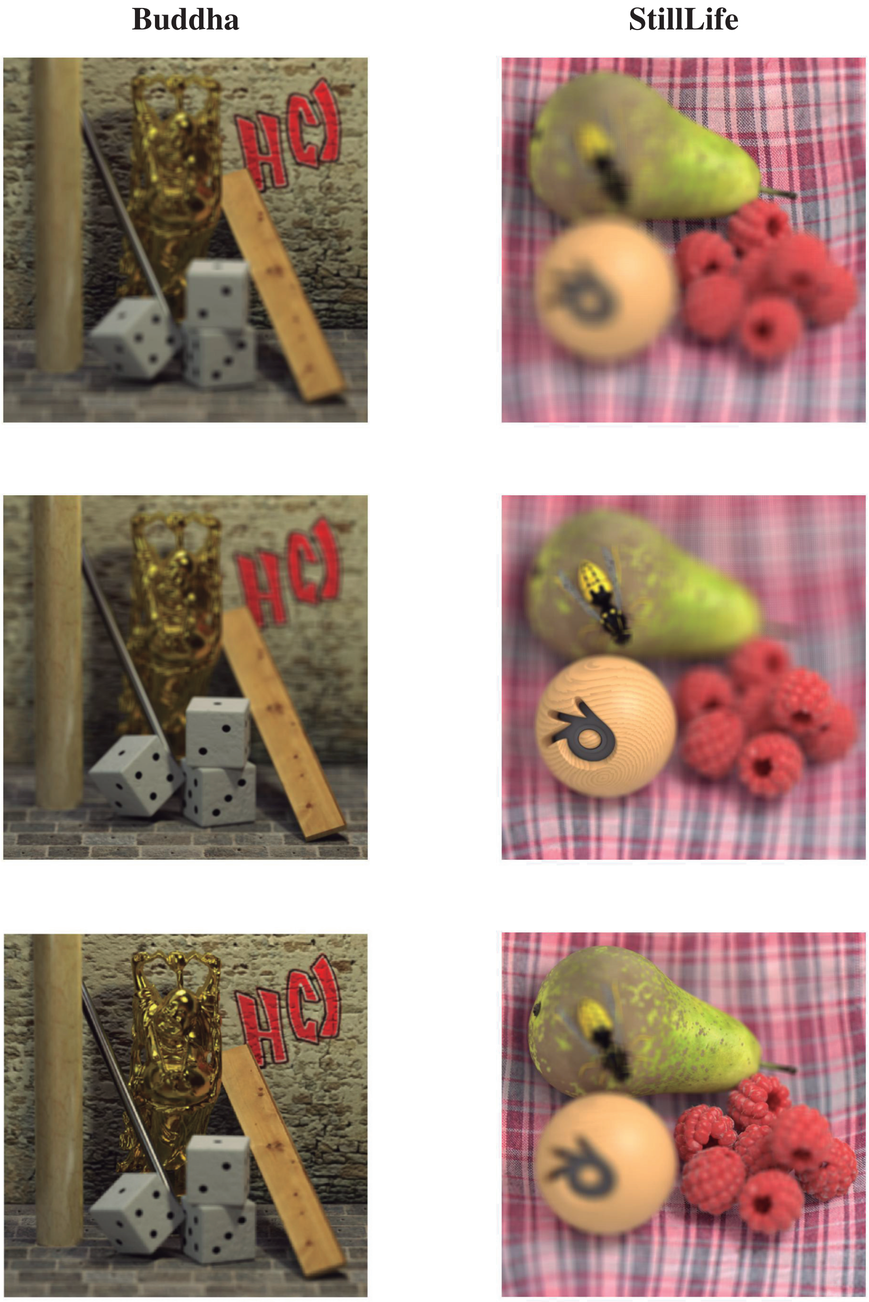
5.2.3 Results and Discussion

When the subsequent previous steps are completed, we obtain an image which is refocused. This image has on focus all the objects which are on the chosen depth label. For instance, two possibles results are shown on the first row in Figure 5.5

where two images appear. In both of them, it appears a part which is on focus such as the bottom part of the wall or the tablecloth. It means that the chosen disparity label candidate α corresponds to the lower disparity (the farthest depth). In addition, in the left image, the rear part of the floor is also on focus because it is at the same depth than the wall. Moreover, other possibles results are the ones which are seen on the second row in Figure 5.5 where the part which is on focus is the nearest part of the floor, also the closest part of the wood, the dices, the wasp, and finally the ball because the chosen disparity label candidate corresponds to the lower depth one (closer to the camera).

In all the images, it could be observed that as farther from the focused depth more blurred is the image. A good example to observe it is on the case of the floor where, gradually, is becoming more blurred. It is clearly seen in the right image of all rows in Figure 5.5, where in the first one the focused part is the tablecloth and in the second row, the part which is focused is the opposite part which is the ball and the wasp.

Finally, in order to show the difference and the progress between the lowest disparity and the highest disparity, on the third row in Figure 5.5 it is shown a result obtained choosing a middle disparity. On these images the parts on focus are the column, the buddha, the metal and also some berries.
Moreover, on the left image of this third row, what is being observed is that the wall is more focused than the case of the left image of the first row. This is strange because is the farthest part of the image and it should be related to the lowest disparity. This problem could be produced due to an illumination change that makes the algorithm fail. It is seen in the bottom part of the wall where in the left image on the first row is less blurred than the left image on the third row. However, the statue is also on focus, which is correct as it is placed in front of the wall. Also, the gradual blur we have mentioned above can be observed, for instance, on the floor on the left image of the third row.



Buddha

StillLife

Figure 5.5: Two examples of the refocus image solution. On the first row the chosen disparity is the lower one, on the second row the chosen disparity is the higher one, finally, on the third row the chosen disparity is one in the middle.

5.3 Depth Estimation

Depth estimation is one of the many problems of computer vision that have captured a lot of attention over the years. A few years ago, when the plenoptic cameras were not invented nor the stereo cameras, depth estimation was done using different images (usually a pair) which were taken from different points of view. However, by using plenoptic cameras is not necessary to take different images from different point of views because a light field image is an array of images which show different but closer perspectives of the scene. As previously mentioned, the images on this array are called subaperture images.
Depth estimation is an important tool for 3D reconstruction because it provides the position where the object is placed. An example of a 3D reconstruction algorithm where depth estimation is used is shown in Figure 5.6.

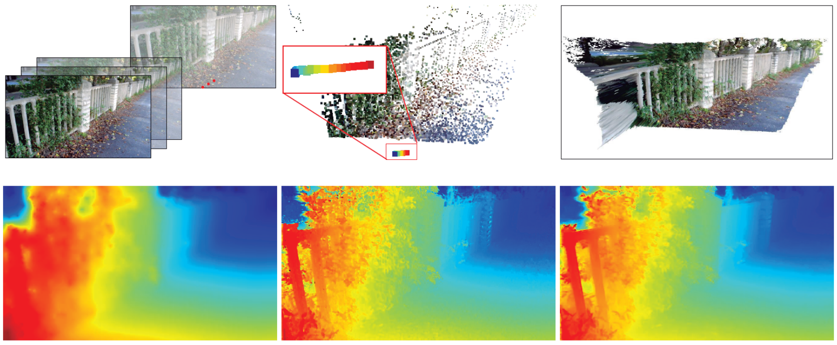


Figure 5.6: First row shows the 3D reconstruction and on the second row a depth map. Images from [Im et al., 2019]

Light field depth estimation has become one of the most important applications in the field. The basic idea behind it is that using the subaperture images, the algorithm is able to estimate the depth where each object is located. Depth based on light field images is able to be estimated since, the subaperture images capture the scene from different but close points of views. In other words, the subaperture images play the role of the images from different points of view in traditional photography.

5.3.1 Related Work

Due to the importance of light field depth estimation, there is a large number of algorithms in this field that let us estimate the depth using various light field characteristics in order to compute the data costs.

The first algorithm did by Wanner and Goldluecke [Wanner and Goldluecke, 2012] took into account knowledge of trigonometry and projective geometry. They measure the local line orientation in the epipolar plane image. In this case, the method was not robust due to the dependency on the angular line.

After this method, Tao *et al.* [Tao et al., 2013] used the information of correspondences and defocus cues to improve the method that had been proposed before. In particular, this method uses the variance in the angular patch as correspondence data cost and sharpness as defocus data cost. Moreover, this method was extended by adding a regularization term and also modifying the correspondence and defocus data measure, which in turn is computed by the sum of absolute differences.

In addition, Jeon *et al.* proposed a method [Jeon et al., 2015] to estimate the depth based on the phase shift theorem. In this method, the data costs are computed using the sum of absolute differences and gradient differences. Despite of the accuracy obtained using this method, it fails when an occlusion is produced.

Another method was proposed by Chen *et al.* [Chen et al., 2014]. It is focused on the bilateral consistency metric on the angular patch in order to obtain the data cost. Although its robustness to occlusions, this method is sensitive to noise.

In 2015, a method was proposed by Wang *et al.* [Wang et al., 2015] who make an assumption related to the invariance of the edge orientation in angular and spatial patches. In this method, the goal is to find the minimum cost in each patch for both data costs. They introduced a new regularization term to make the method robust to occlusions. However, this only lets the appearance of a single occluder in each angular patch.

Another contribution was made by Lin *et al.* [Lin et al., 2015] who introduced the infocus and consistency measure. But there were no many comparisons for each data costs without global optimization. After that, Kolmogorov and Zabih [Kolmogorov and Zabih, 2002] used the visibility constraint to model the occlusion and afterwards it optimize it by using graph cut. However, this method fails when an occlusion is produced. Wei and Quan [Wei and Quan, 2005] tried to solve this problem adding a smoothness term.

Bleyer *et al.*[Bleyer et al., 2016] proposed another method which is based on the visibility of a pixel in corresponding images to design the occlusion cost. The main problem was the difficulty when there is a huge number of views. Vaish *et al.* proposed a method [Vaish et al., 2006] where the entropy has already used in order to obtain the data cost.

In 2016, [Williem and Kyu Park, 2016] introduced a method based on using the information of the entropy they made the algorithm robust to occlusion without using any constraint. And at the same time, this method achieves to be less sensitive to noise. Our proposed method grounds on this paper.

5.3.2 Method

As we have seen in the previous section, there are many algorithms focusing on estimating the depth. However, these algorithms fail when an occlusion is produced or there is presence of noise on the image. The method used in this section is robust to occlusions and less sensitive to noise. First, in order to estimate the depth, two data cost are computed, the *correspondence data costs* and *adaptive defocus response*. In the *correspondence data cost* is computed for each angular patch the pixel color randomness and in the *adaptive defocus response* is computed to achieve robustness in presence of occlusions.
To compute the data cost for each disparity candidate α an angular patch for each pixel and a refocus image must be estimated. In order to create the angular patch, first the correspondences between every pixel on the center image and all the subaperture images are created. To create every angular patch, each pixel in the angular images is extracted from the sheared light field $L_\alpha(x, y, u, v)$ based on the disparity label candidate.

As many problems in computer vision, the depth is estimated by minimizing an energy. This energy is computed by computing two data costs that measure how proper the label α of each pixel is. This is defined by the following formula:

$$E = \sum_p E_{unary}(p, \alpha(p)) + \lambda \sum_p \sum_{q \in N(p)} E_{binary}(p, q, \alpha(p), \alpha(q)) \tag{5.3}$$

where p is a pixel of the image, $\alpha(p)$ denotes the disparity label at p and $N(p)$ is the neighborhood around the pixel p. Moreover, $E_{binary}(p, q, \alpha(p), \alpha(q))$ is the smoothness restriction that forces the consistency between neighbor pixels, which is multiplied by $\lambda > 0$ which represents a weighting factor [Williem and Kyu Park, 2016]. $E_{binary}(p, q, \alpha(p), \alpha(q))$ is defined by the following formula:

$$E_{binary}(p, q, \alpha(p), \alpha(q)) = \nabla I(p, q) \min(|\alpha(p) - \alpha(q)|, \tau) \tag{5.4}$$

where $\nabla I(p, q)$ is the intensiti difference between the pixel p and the pixel q and τ is the threshold.
Finally, $E_{unary}(p, \alpha(p))$ is the term accounting for how proper the label α is. This data term is formed by the sum of the correspondence response and the adaptive refocus response as in Equation 5.5.

$$E_{unary}(p, \alpha(p)) = C(p, \alpha(p)) + D(p, \alpha(p)) \tag{5.5}$$

where $C(p, \alpha(p))$ corresponds to the *correspondence response* that is computet using the Angular Entropy metric and $D(p, \alpha(p))$ is the *adaptive defocus response*.

5.3.2.1 Angular Entropy Response

Mostly, the correspondence data cost is focused on the similarity between pixels in the angular patch. However, if there is an occlusion, the photo-consistency is totally broken. This is the reason why many algorithms tried to be robust to occlusions in order to estimate the depth.
In the paper [Williem and Kyu Park, 2016] this work is grounded on, it is proposed a novel occlusion-aware correspondence data cost by computing the entropy in the angular patch, which is referred to as *angular entropy metric*. The entropy is computed for each channel (R,G,B) independently and then the average between the three channels is calculated to obtain the total entropy, using the following equation:

$$C_avg(p,\alpha) = \frac{H_R(p,\alpha) + H_G(p,\alpha) + H_B(p,\alpha)}{3} \tag{5.6}$$

The idea behind this metric is the following: if an occlusion is produced, the entropy will be higher due to the fact that randomness increases. The advantage is that the entropy is computed for each angular patch, which is formed by pixels from the different subaperture images and each subaperture image contains different views. Having it into account, it may well be possible that, in one subaperture image a certain pixel is not occluded and, in another subaperture image the correspondent pixel is occluded. This will be reflected on the angular patch.
Moreover, if in the angular patch there are no occlusions, the entropy on the incorrect depths is even higher than the entropy on the correct depth. However, if in a certain pixel there is an occlusion the entropy between the wrong depth and the proper depth is still existing but, is not as different as in the other case. It is in that way, how it is possible to estimate the depth being robust to occlusions.

5.3.2.2 Adaptive Defocus Response

The idea of the adaptive defocus response is to find the minimum response among the neighborhood regions. In other words, this data cost plays the role of the similarity constraint, because it is trying to achieve a minimum difference between one pixel and its neighbor. This data cost is the one that make the algorithm less robust to noise. The novel part is to make the region, where the response is computed, smaller in order to avoid the effect of the blurry artifact.
In order to obtain this cost, the refocus image $\bar{L}_\alpha$, given by equation 5.2 in Section 5.2, must be computed. This part is done, as it is explained in the Refocus section (Section 5.2), by computing the average of the angular patch. The Adaptive Defocus Response is computed by following the equation:

$$D_c(p,\alpha) = \frac{1}{\mid N_c(p) \mid} \sum_{q \in N_c(p)} \mid \overline{L}_\alpha(q) - P(q) \mid \tag{5.7}$$

where $N_c(p)$ is the subpatch of the pixel p and P is the center image.

Finally, in order to obtain the depth map, it is needed to compute the correct depth. To do so, the energy is minimized, in other words, choosing the minimum energy for each pixel.
In addition, in the method of W. Williem he uses the *graph cut* method in order to optimize efficiently the energy of the algorithm [Boykov et al., 1999].

5.3.3 Results and Discussion

Once all the process is done, if it is only used the method of the angular entropy response, the results obtained are the ones shown on the first and the second row in Figure 5.7. It can be clearly seen some noise on the dices or on the T. This happens because the adaptive defocus response is not computed, which is the one that makes the algorithm less sensitive to noise. However, when an occlusion is produced the algorithm detects it in an accurated way.

However, if the method is done using only the adaptive defocus response, the results obtained are worse than the result shown on the first and the second row in Figure 5.7. As it is shown on the third row in Figure 5.7 the result is less accurated because there are estimated different depths on the same object, when this object is on the same depth, for instance, on the case of the dice.

Finally, if we use both data costs the result obtained is shown on the last row in Figure 5.7. As it is seen the result obtained has not a significant difference to the result obtained using only the angular entropy response. So in terms of efficiency, it is better to use only the first data cost. In fact, it happens because the entropy not only is higher when there is an occlusion but also when the angular patch does not have the correct depth. So, by using only the information of the entropy it could be estimated the depth being robuts to occlusions. In fact, it is seen on the image of the first row where the cube occludes the wall and it is well estimated, or also on the case of the plants.

Notice, in all cases of *Buddha* image, there is one part of the image where the algorithm fails. This part is on the column, which is always on the same depth but the algorithm detects the upper part as it would be in a higher depth. Now if we take a look at the image in Figure 5.8, it could be seen that in the column there is a soft illumination change which could be the responsible of this error. Another point to comment is in the case of the *Mona* image where the depth of the ball is estimated correctly even though it is a spherical volume. And also, the gradual change of depth on the case of the cube and the letter T, which is also well estimated.

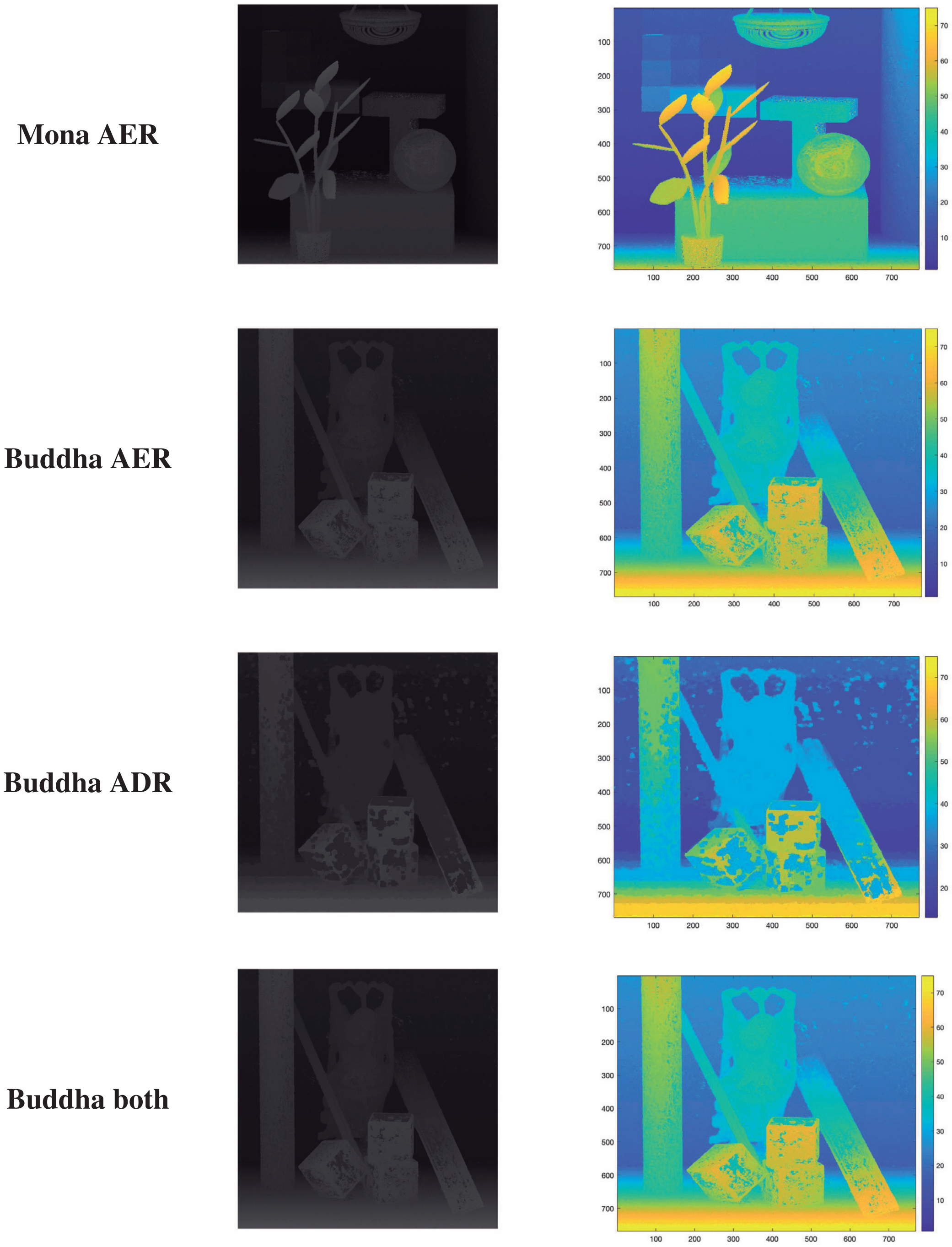


Figure 5.7: Two examples of depth maps. On the first and the second rows using the Angular Entropy Response, on the third row using the Adaptive Defocus Response and on the last row using both data costs.

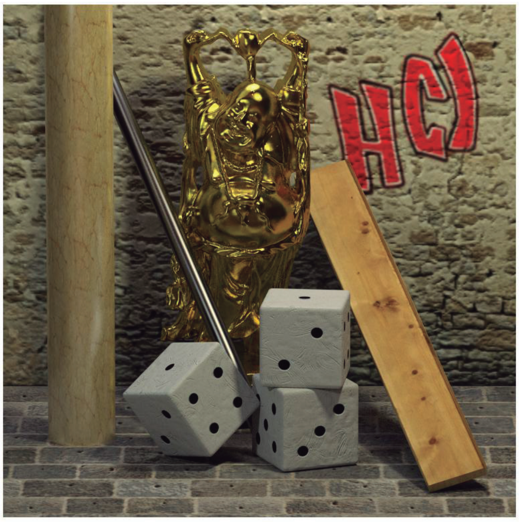


Figure 5.8: *Buddha* image

5.4 Occlusion Detection

Occlusion detection is a method used in many problems of image processing and computer vision field such as video tracking, 3D reconstruction, depth estimation and among others. For instance, in the case of video tracking, it is important to detect the occlusions to construct an algorithm able to track objects even in cases where they are partially occluded. If an occlusion appears during the process of video tracking, the tracked object is lost, so making the algorithm robust to occlusions make it easier to find the object when the occluder disappear. An example of the result obtained after applying a method robust to occlusion could be seen in Figure 5.9. In this example, the algorithm can detect a person even being partially occluded.

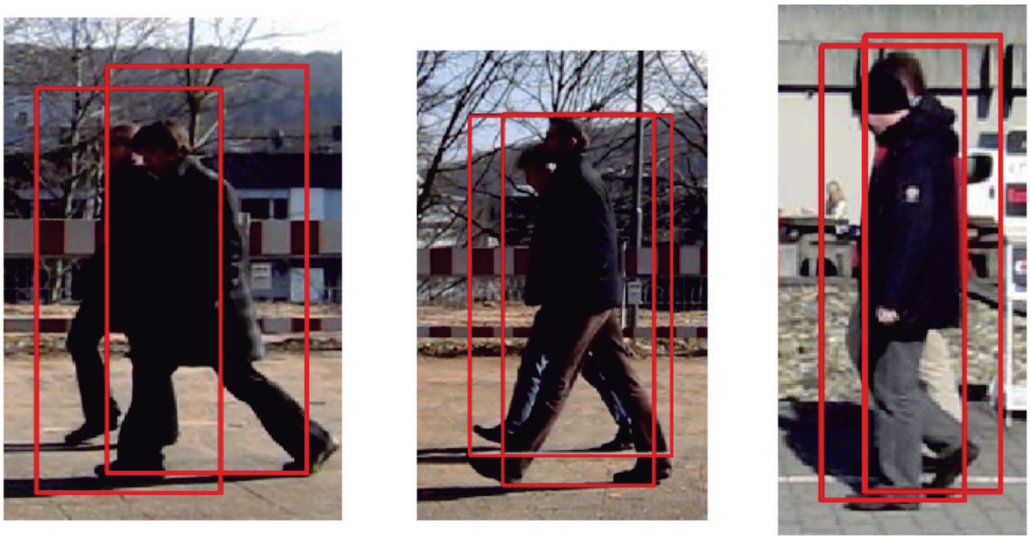


Figure 5.9: An example of a tracking method robust to occlusions. Image courtesy [Tang et al., 2014].

Moreover, while performing 3D reconstruction, an occlusion region prevents to establish the correspondences between those objects which are occluded in one or more views. An example of the result obtained after applying a 3D reconstruction algorithm robust to occlusions is shown in Figure 5.10. In this example, it is easily seen that the mouth and the eyes which are occluded, are reconstructed quite good. However, there are other cases where performing the 3D reconstruction is hard to achieve because the occluder is too large or there are many changes and it is impossible no know how to reconstruct the object. However, in the case shown in Figure 5.11 it is seen a 3D reconstruction which is not robust to occlusions. It is seen clearly that the eye is not well reconstructed.

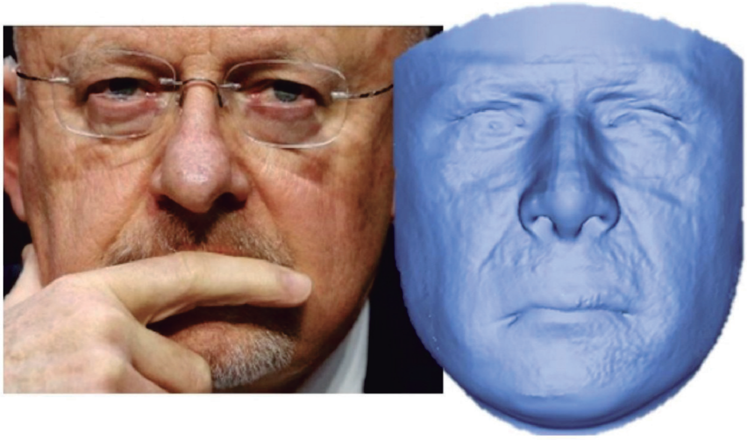


Figure 5.10: An example of a 3D reconstruction robust to occlusions. Image courtesy [Tuấn Trần et al., 2018]

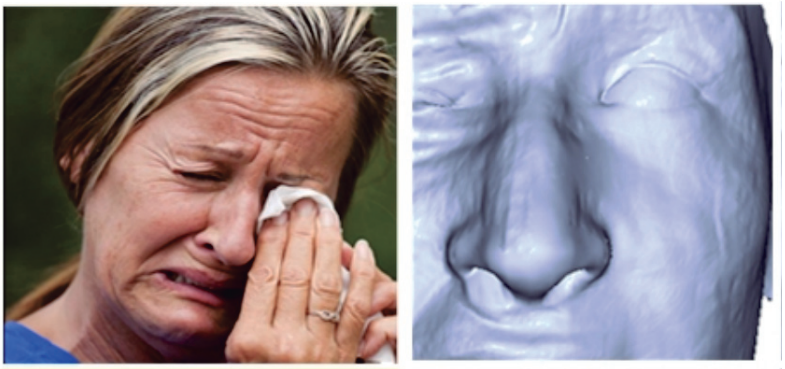


Figure 5.11: An example of a 3D reconstruction not robust to occlusions. Image courtesy [Tuấn Trần et al., 2018]

In the case of light field imaging, as far as we know, there are no previous works focused on occlusion detection. Nonetheless, some works have been done in depth estimation being robust to occlusions being the goal to estimate the depth correctly instead of estimating the occluders. An example of an accurated depth map estimation by [Huguet, 2009] is shown in Figure 5.12. On this depth map, it is seen that even if the presence of occlusions, the depth is estimated correctly.

5.4.1 Related Work

In the field of light field imaging, there is a particular interest in finding an algorithm able to detect occlusions in the given image. Since many years ago, many

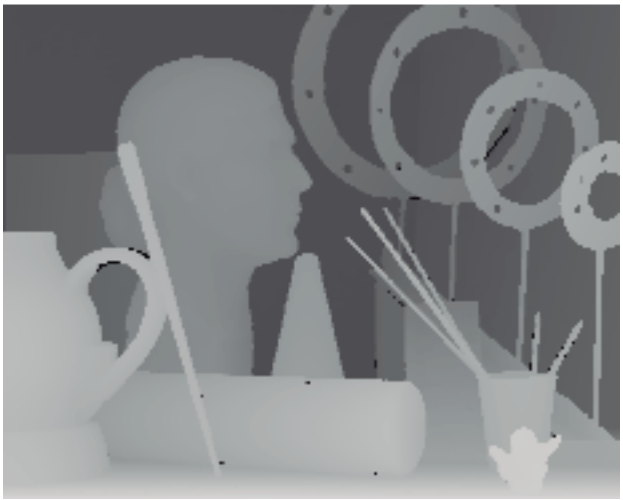


Figure 5.12: An example of an accurated depth map. Image courtesy [Huguet, 2009].

studies are focused on achieving this goal to be able to create a method of depth estimation robust to occlusions, most of them fail on specific cases, for instance, when appears more than one occlusion regions.
W. Williem and In Kyu Park [Williem and Kyu Park, 2016] have achieved the goal of creating an algorithm for depth estimation robust to occlusions, three years ago by using the angular entropy metric.

5.4.2 Method

The method for occlusion detection presented on this project uses the entropy information to check if there is an occlusion. Our method follows some steps:

1. First, the angular patch for each pixel is computed.
2. Second, we compute the histogram associated to each angular patch.
3. Third, the entropy related to each histogram is computed.
4. Finally, we threshold the entropy values to determine if there is an occlusion.

As explained in the *refocus section* (Section 5.2), a pixel which is not in the correct depth will not be on focus. As a result, this pixel is blurred in every subaperture image, so the entropy of this angular patch will be higher that the entropy computed on the angular patch which is on focus.

Moreover, when there is a pixel occluded in some subaperture image, the entropy of the angular patch that contains that pixel will be higher. Both cases are shown in Figure 5.13, where the entropy is high not only when the angular patch is not on focus but also when there is any pixel of the angular patch which is occluded.

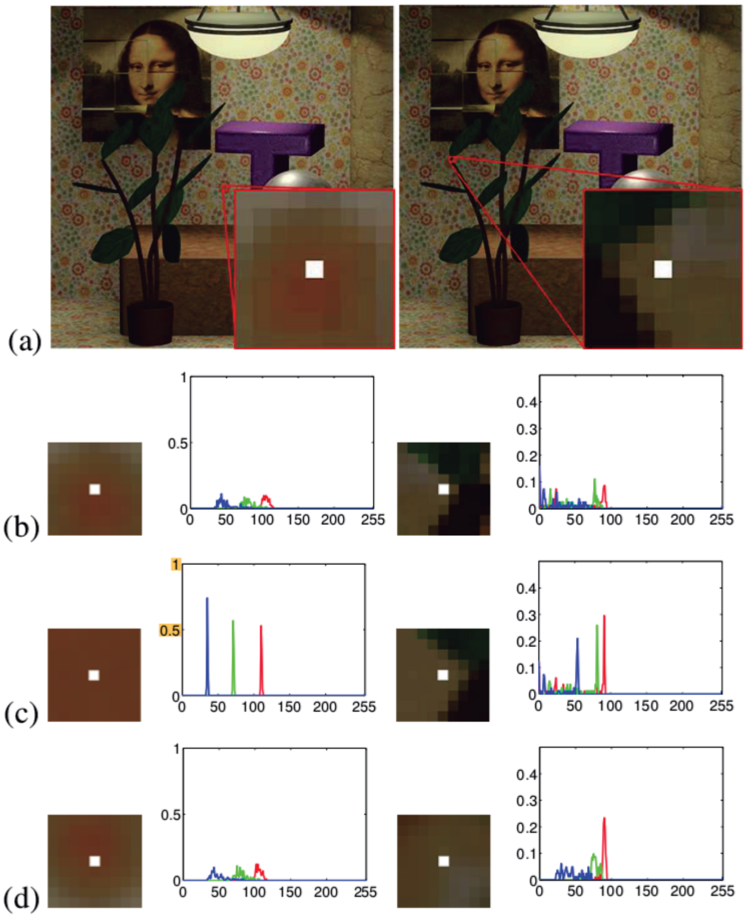


Figure 5.13: Example of the histograms of different angular patches. Image source: [Williem and Kyu Park, 2016]. In this Figure, each row corresponds to the same pixel but using different depth. In the left column a non-occluded pixel is shown and the right column shows an occluded pixel.

Once the entropy is computed, a threshold should be determined in order to decide if the pixel is occluded or not. This threshold could be fixed or adaptive.

5.4.2.1 Fixed Threshold

In the case where we have a fixed threshold, the basic idea is to fix a certain value for the threshold which would determine if the pixel is occluded or not. In order to determine if there is an occlusion, the condition that the value of the entropy should accomplish is to be higher than the value of the threshold. If it is not higher, the pixel will not be occluded.
The idea behind this condition is, as we mentioned above when a pixel is occluded, the entropy of the angular patch is high due to the change of intensity between the pixel in one subaperture image and the other subaperture image.

This number is chosen by computing the mean of all the entropies computed. Specifically, the value of the threshold would be a little bit higher than the value of the mean of all the entropies. As a result, it will be detected the outliers of the entropy, which intuitively, should correspond to those pixels occluded.

5.4.2.2 Adaptive Threshold

The adaptive threshold requires some more computations since the idea behind this kind of threshold is to compute the entropy of the same angular patch, but for different depths labels candidates, which are not only the estimated one. Then, the mean of these entropies is computed with Equation 5.8. After that, we compute the difference between the entropy of the correct depth label and the mean of entropies as it is shown in Equation 5.9. Finally, the condition accomplished in order to be an occluded pixel is that the value of the difference between the entropies should be lower than the value of the threshold. If this condition is not achieved, the pixel will not be occluded.

$$H_{mean} = \frac{H_{\alpha 1} + H_{\alpha 2} + H_{\alpha 3} + H_{\alpha 4}}{4} \tag{5.8}$$

$$Diff_H = \mid H_{mean} - H_{\alpha_{good}} \mid \tag{5.9}$$

In this case, we use, if exists the information of the depth estimated, the depth estimated $+10$, the depth estimated $+20$, the depth estimated -10 and the depth estimated -20. It means that it is used the information of 5 different depths. However, in some cases, the object is placed in a low depth and the depth estimated -20 does not exist. So, depth estimated -20 would be equal to the lower depth and, using this logic, the same happens if the estimated depth is a high depth.

The logic behind this condition is that, as we mentioned above, when a pixel is occluded the entropies of the angular patch for different depth labels are not really different. However, if the pixel is not occluded, the entropies are much more different.

Finally, the pixels that accomplish the condition will be indicated by changing the value of those pixels to 255 and the other ones will be changed to 0.

5.4.3 Results and Discussion

Using the method explained in the previous section, in Figure 5.14 it is shown the result obtained after using the Fixed Threshold strategy, where the white pixels are those pixels that in, at least, one subaperture image are occluded. On the left hand side image, the threshold is too high, so there are few pixels with this entropy and not all the pixels occluded are marked as white. On the contrary, on the right hand side image, the threshold used is too low, so there is a lot of pixels with this entropy and most of them are not occluded in any subaperture image. Finally, on the image in the center, it can be seen that the threshold is the best suited for this application because most of the pixels in white are edges and it will be occluded in some subaperture image. However, it could be seen that the method of detecting occlusions using the entropy fails when there is an illumination change.

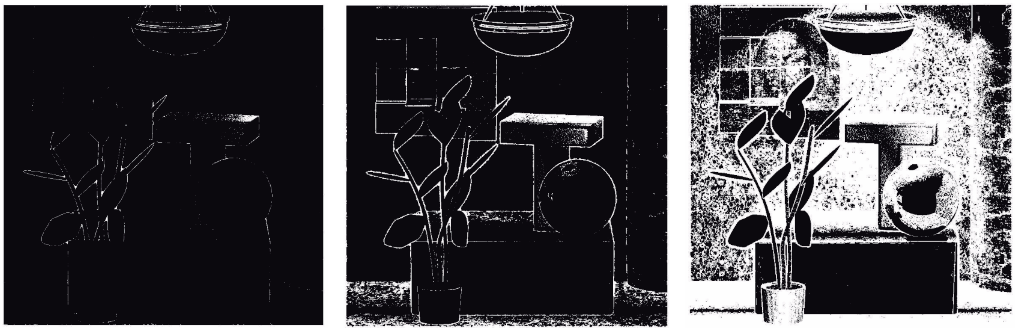


Figure 5.14: Using 3 different values. From left to right thr = 3.1, thr = 2.1 and thr = 1.

Moreover, in Figure 5.15 it can be seen the result after applying the method to obtain the occlusions using the Adaptive Threshold strategy. If this case is compared with the one using a Fixed Threshold, the Adaptive Threshold works worse because this threshold selects pixels as occluded when they are not.
It could be seen that most of the pixels chosen as pixels occluded are placed in points where there are illumination changes. Furthermore, as it is seen on the right

hand side image, there are selected pixels placed on objects with textures, for example, the wall. Both problems are caused because when there are abrupt changes from one pixel to another, for instance, in the case of illumination changes, the entropy will be high. So, when is computed the difference between the entropy on the correct depth and the entropy on the other depths, the difference would be small.

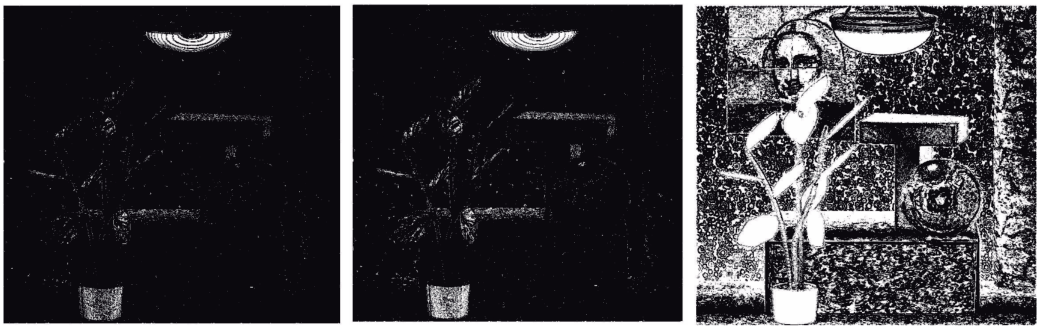


Figure 5.15: Using 3 different values. From left to right thr = 0.2, thr = 0.25 and thr = 0.9.

Finally, if we use both thresholds at the same time the result obtained is the one reflected in Figure 5.16. Using both thresholds we achieve better results than the ones obtained using the Fixed Threshold or the Adaptive Threshold. Using both thresholds the method is more robust to illumination changes, it is seen in the central image which is the best result obtained because it is the most accurated one. In the image on the left, there are few pixels chosen and in the image on the right, there are many pixels selected and most of them could not be occluded.

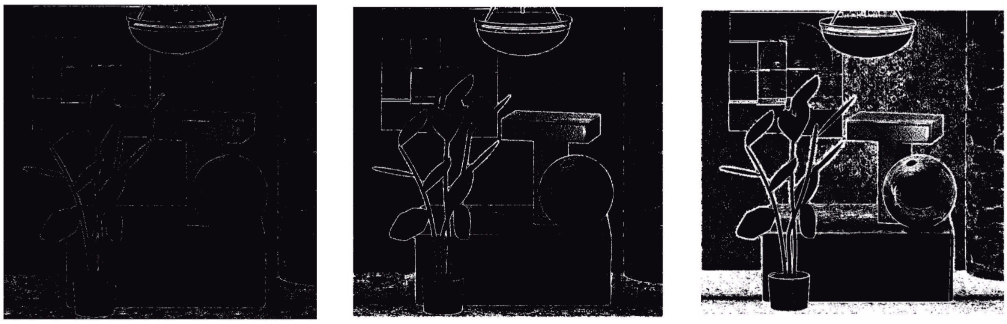


Figure 5.16: Using 3 different values. From left to right thrfix = 2.1 thradap = 1.5 , thrfix = 2.3 thradap = 0.5, thrfix = 1.5 thradap = 0.5.

Mostly, the pixels selected as occluded pixels are on the edges of the objects, because the subaperture images have different but close points of view, so the occluded points between the different subaperture images are those placed in the edges.

5.5 3D Reconstruction

3D reconstruction refers to the creation of three-dimensional shapes from a set of images under different points of view. This is a field very explored from a few years ago until now, especially in computer vision and image processing fields, and it is mostly solved using projective geometry. The redundant information provided by light field cameras made research on 3D reconstruction a top subject on that field. This fact is produced due to the contribution that light field imaging has on the field of 3D reconstruction, which is considered an advantage because light field images have more angular information. The angular information provides more details in order to do the 3D reconstruction so that the results could be more accurated.

5.5.1 Related Work

The studies realized until now in 3D reconstruction using light field images are focused on 3D face reconstruction [Feng et al., 2018]. An example of this application is shown in Figure 5.17. Also, other studies are focused on the extraction of the calibration parameters as a pre-processing step before performing the reconstruction [Wang et al., 2018].

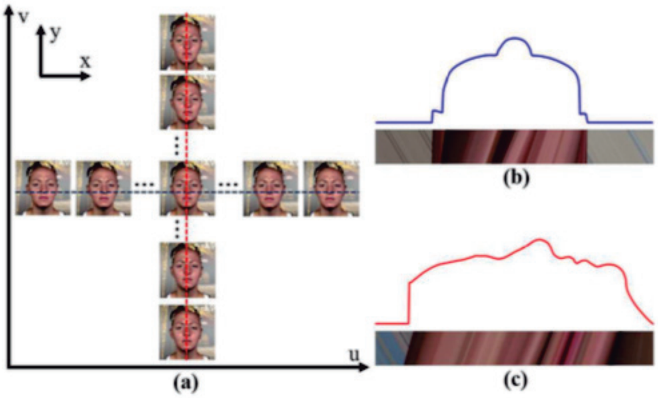


Figure 5.17: Example of 3D face Reconstruction using light field images. Image courtesy of [Feng et al., 2018].

The use of light field images for 3D face reconstruction is really useful due to the high resolution of the plenoptic cameras, especially the angular resolution, and also because a 4D light field image captures the intensity at each pixel for each channel and the direction of light rays. Most of the methods used to reconstruct 3D faces use common techniques such as *Shape from Shading* [Atick et al., 1996], *3D Morphable Models* [Blanz et al., 1999] or *CNN*s [Jackson et al., 2017]. In *Shape from Shading*, the variation of shading is used to reconstruct the face and in the case of *3D Morphable Models*, it is projected the 3D face in a low-dimensional subspace.

5.5.2 Method

The method used in order to perform 3D reconstruction is divided into two parts. The first one is the computation of the calibration parameters. In this part, the calibration parameters are computed using a recent adaptation of Zhang Method [Wang et al., 2018] to the case of light field images, which is going to be explained later. On the other hand, the second part is based on making the 3D reconstruction by using the triangulation method in order to find the point in the 3D space.
In the following subsections, first of all, we are going to explain the theory of Zhang's method using conventional images, then it is going to be described the adaptation in order to be able to working using light field images [Wang et al., 2018]. And finally, how to achieve the 3D reconstruction using the triangulation method.

5.5.2.1 Zhang's Method

Zhang's Method is a method for camera calibration. The goal is to obtain the intrinsic and extrinsic parameters of the camera that will allow us to convert a 2D point into a 3D point. The intrinsic parameters are contained into a matrix K, with the structure shown in following equation:

$$K = \begin{pmatrix} \alpha_x & \delta & u_0 \\ 0 & \alpha_y & v_0 \\ 0 & 0 & 1 \end{pmatrix} \tag{5.10}$$

where $\alpha_x = f \cdot m_x$ and $\alpha_y = f \cdot m_y$ are the components of the focal length in terms of pixels, δ is the skew coefficient and u_0 and v_0 are the coordinates of the principal point in terms of pixels. If the focal length has no scale factor, the principal point would be in the center of the image and the matrix would follow the following structure

$$K = \begin{pmatrix} f_x & \delta & x_0 \\ 0 & f_y & y_0 \\ 0 & 0 & 1 \end{pmatrix} \tag{5.11}$$

In order to compute the previously mentioned parameters we will follow the following steps:

1. First, to compute the correspondences between the planar pattern and the images under different perspectives.
2. Secondly, to build matrix V. This matrix is composed by the homographies that relates the images.
3. Then, to find ω by doing the SVD of V. This matrix is known as Image of the Absolute Conic, which is independent of the camera position.
4. Finally, to extract the calibration parameters, K, by Cholesky factorization of ω.

On the other hand, the extrinsic parameters denote the coordinates system transformation from the 3D world coordinates to the camera coordinates. These extrinsic parameters are computed having into account that camera calibration is done by using a planar pattern. Having into account this fact, the equation used in order to find R and t is

$$[Kr_1 Kr_2 Kt] \sim [h_1 h_2 h_3] \tag{5.12}$$

From this equation, by isolating we can obtain the equation

$$r_1 = \frac{K^{-1}h_1}{|| K^{-1}h_1 ||}; r_2 = \frac{K^{-1}h_2}{|| K^{-1}h_2 ||}; r_3 = r_1 \times r_2; t = \frac{K^{-1}h_3}{|| K^{-1}h_1 ||} \tag{5.13}$$

The resulting matrix is composed by a rotation and a translation and has the following structure

$$M = \begin{pmatrix} R & T \\ 0 & 1 \end{pmatrix} \tag{5.14}$$

where R is the rotation, which is a matrix $3x3$ and T is the translation, which is a vector $3x1$.

5.5.2.2 Approach based on Light Field Imaging

Focusing on the 3D reconstruction from light field images, the process is really similar than the previous one, but using a modified algorithm for light field images, which contain a significant amount of information compared with traditional images. An important point to take into account, it is the fact that due to the difficulty of having a plenoptic camera, this application will be done using only one

light field image formed by 81 subaperture images.
Then, the steps to perform in order to obtain the calibration parameters are the following ones:

1. First, to compute the correspondences between the planar object and every subaperture image for each light field image (in our case only one light field image).

2. Second, to solve the H, which is the homography matrix that relates two images. On the algorithm this project grounds on is used the Levenberg-Marquardt algorithm.

3. Third, to calculate ω in order to compute K, which is the matrix of intern parameters.

In order to compute the camera calibration matrix's parameters, it is important to use a light field image which contains a planar surface, for example, a wall, a painting, or some other planar object. This object or surface could not be occluded. These conditions are needed to be accomplished because to obtain the correspondences, is required to mark four points that must be the same on each subaperture image. As it is indicated in Figure 5.18. These four points would be four of the correspondences which would be used to obtain the transformations between the different subaperture images. However, this method of looking for the correspondences has the problem of being not as accurated as other methods are because it is very challenging to mark the exact point not only on the 81 subaperture images but also in the different light field images.

Once the camera calibration matrix's parameters are obtained, the extrinsic parameters of the camera, that is the rotation and the translation with respect to world coordinates, can be computed by using the matrix of intern parameters K and the homography H.
Finally, using the intrinsic and extrinsic parameters, the camera projection matrix, usually denoted by P, is constructed. It is used to project 3D points of the real world into the image plane which is in 2D. This projection is done by applying the formula

$$x = PX \tag{5.15}$$

where X are the points in the 3D real world coordinates and x are the points into the image plane. Notice that both of them (X and x) are represented in homogeneous coordinates. The Matrix P is composed by the product of two matrices: the camera calibration matrix K and the matrix of extrinsic parameters $[R|t]$ as seen in the following equation:

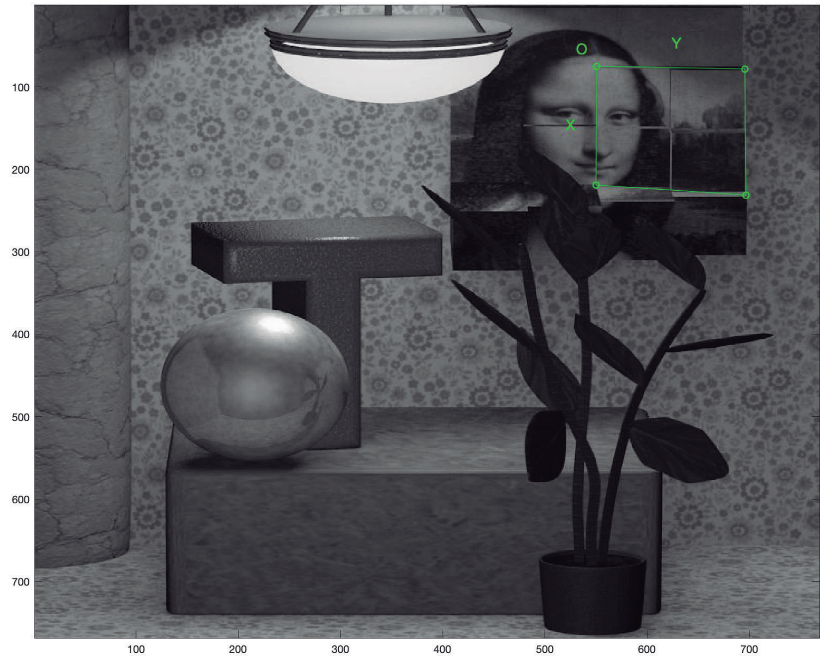


Figure 5.18: The 4 corners chose to compute the correspondences

$$P = K_{3x3}[R_{3x3}|t_{3x1}] = \begin{pmatrix} f & \delta & x_0 \\ 0 & f & y_0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_x \\ r_{31} & r_{32} & r_{33} & t_x \end{pmatrix} \tag{5.16}$$

After applying the previous concepts and formulas, the result obtained is the intrinsic and the extrinsic parameters. The extrinsic parameters are different for each image, so it is obtained a matrix P, which will be different for each subaperture images. By using both kind of parameters, we are able to know the position of the cameras respect to the image plane and vice versa as it can be seen in Figure 5.19. If others light field images were used, there will be groups of planes capturing different points of view, where each group corresponds to a light field image.

The next step is to recompute the correspondences using a well-known method for feature extraction called ***SIFT***. It is necessary to recompute the correspondences because the goal is to recover all the scene and, until now, we only have the correspondences of one part of the image shown in Figure 5.20. As a result, it will

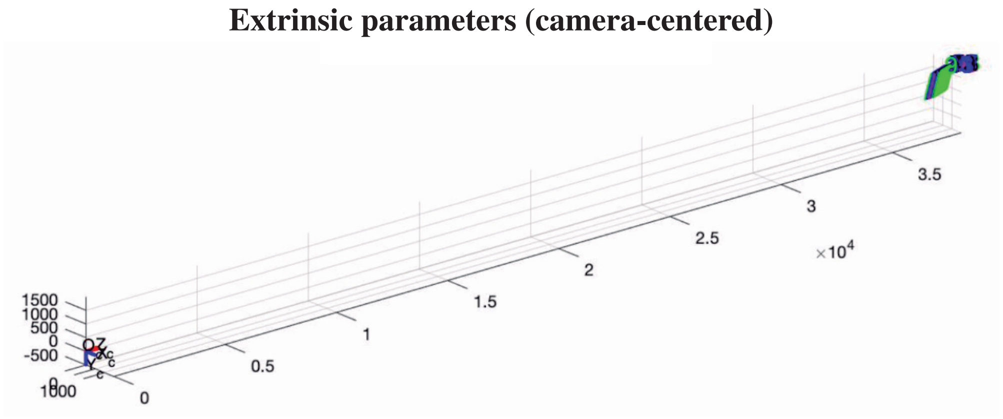


Figure 5.19: The different image planes in order to obtain the angular information of 1 light field image

be reconstructed only this part. However, if we recompute the correspondences of the whole image it will be possible to reconstruct all the scene. That is the reason why it is recomputed the correspondences using SIFT.

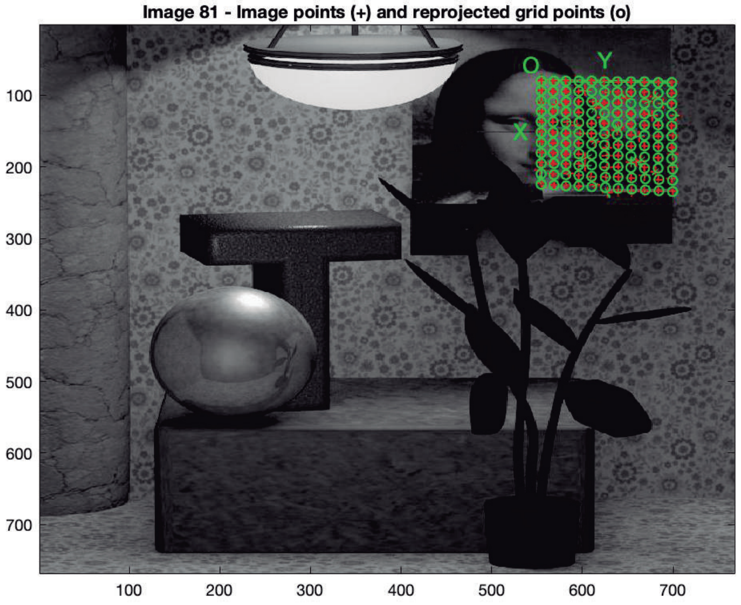


Figure 5.20: The 121 points which are the correspondences computed in the first step

After applying SIFT and finding the correspondences between two subaperture images is obtained. An example can be seen in Figure 5.21. On this image, it is seen that the correspondences are homogeneously dispersed, therefore, there are parts of the image where there are many correspondences and others where are few correspondences. As a result, on the 3D reconstruction will be less information to reconstruct the light than the wall or the cube.

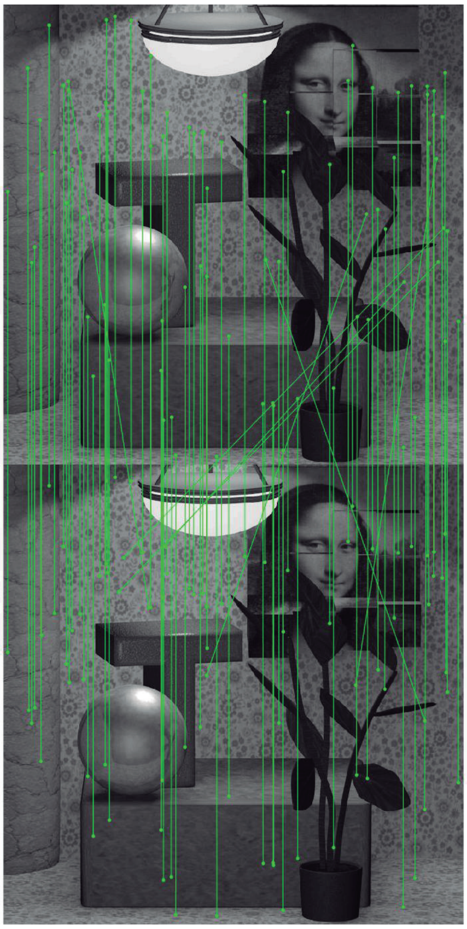


Figure 5.21: Some matchings between two subaperture images. As we can see, there are some mismatchings in the flowers from the background due to the repetition of the flowers.

Once the correspondences are recomputed, it is going to be used the triangulation method in order to find the 3D point. To do this process, it is needed, at least, two subaperture images and the correspondents matrices P. Afterwards, we want to find the point in 3D space which corresponds to the point in 2D space. Mathematically the goal is to find $x = PX$ and $x' = P'X$, where the 3D point is the same, x and x' are the correspondences in both subaperture images and P and P' are the **projection matrices**. Both formulas give us the system of equations shown in Equation 5.17, which is going to be solved.

$$\begin{cases} x_1(p_3X) = p_1X \\ x_2(p_3X) = p_2X \\ x'_1(p'_3X) = p'_1X \\ x'_2(p'_3X) = p'_1X \end{cases} \tag{5.17}$$

After that, the 3D point correspoding to every correspondence are obtained. And it is going to be reconstructed the 3D scene.

5.5.3 Results and Discussion

The result obtained by using the proposed method of marking on the different subaperture images the four points and obtaining the correspondences, is the matrix K, which contains the internal parameters of the camera. In this case, the K has the structure shown in the following equation:

$$K \simeq \begin{pmatrix} 9 \times 10^3 & 0 & 383.5 \\ 0 & 8 \times 10^3 & 383.5 \\ 0 & 0 & 1 \end{pmatrix} \tag{5.18}$$

All the value obtained in matrix K are possible. First of all, the central points are the correct ones because the size of the subaperture images is 767×767. Also if it is taken into account the fact that the variation of perspective between the different subaperture images is very small, the focal length should be very high. The reason is because the method detects each image as it would be taken from different cameras located very far away from the scene. In brief, if the variation is small, the focal length should be high, which coincides with our result.

Once we did the triangulation and we plot the 3D reconstruction using four subaperture images, the result obtained is shown in Figure 5.22. In the 3D plot, which is a profile view, we are able to distinguish some of the objects of the scene, such as the wood cube or the ball. Moreover, some purple points can be seen, which are points ont the T object and some green points which come from the plant. An important point to mention is that the back and profile sides of the wood cube do not exist. This is because as we have discussed above, subaperture images have different but close points of view, so there is not any image taken from the perpendicular or more slanted point of view. As a result, the corresponding profile could not be reconstructed. Also, it can be seen that some points of the wall which are behind the cube are not reconstructed and it is because the cube is an occluder.

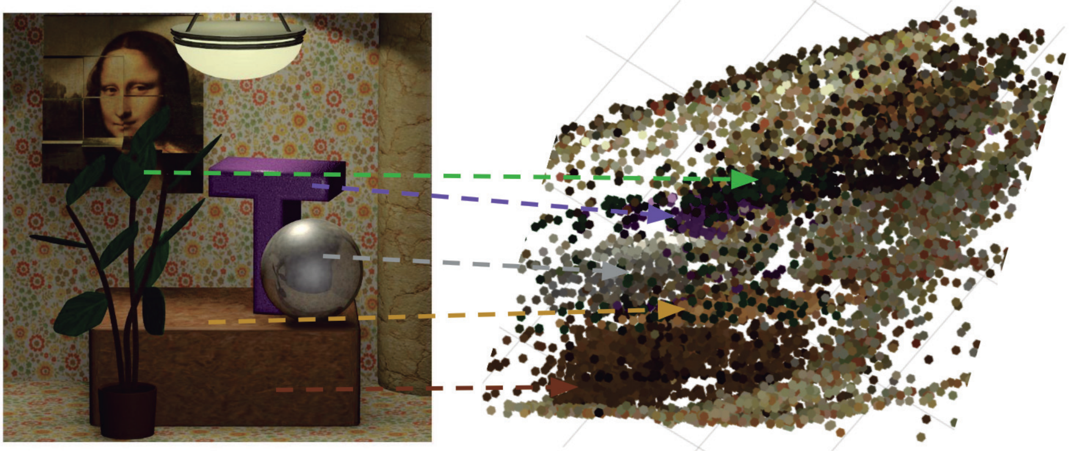


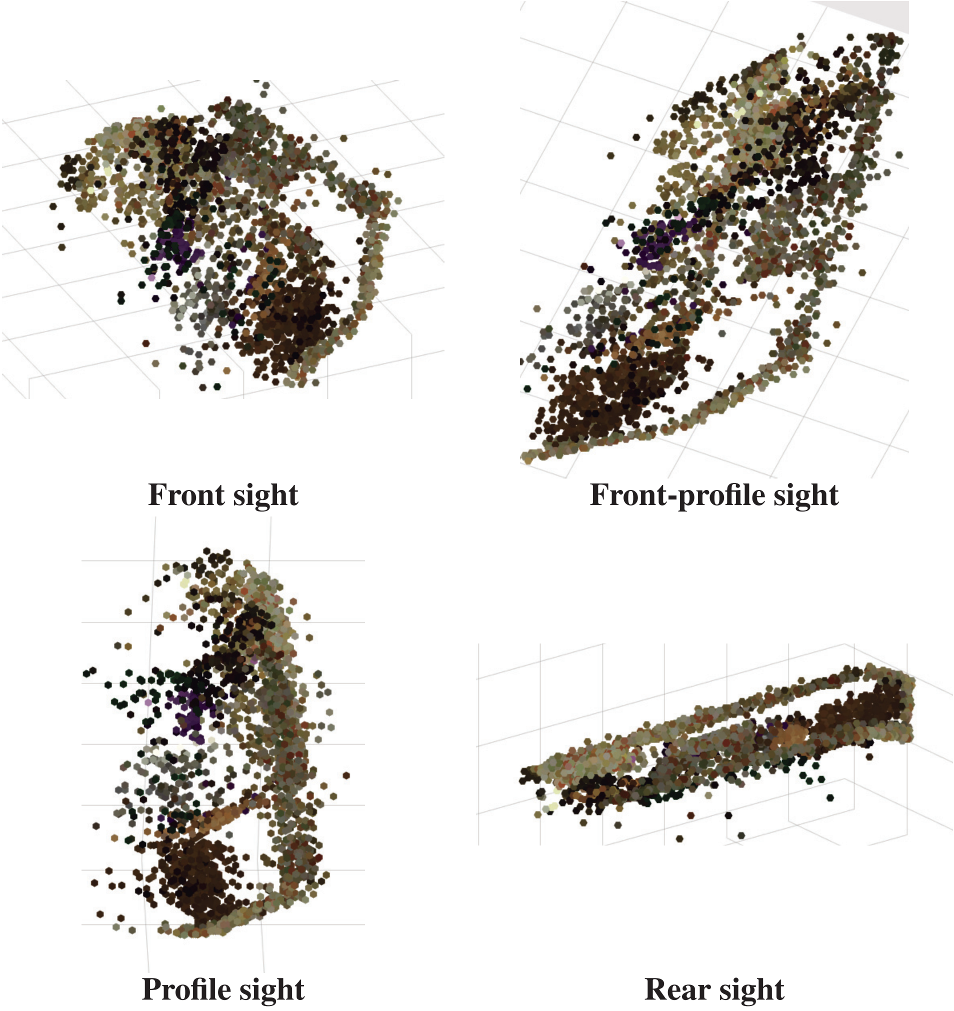
Figure 5.22: The 3D reconstruction compared with the 2D image.

Furthermore, in Figure 5.23 it is shown some views from the 3D reconstruction using only two cameras. The image of the first row left is seen from the front sight and it could be seen the lack of information behind the cube as it is also seen in the image of the first row right. On the image of the second row left, it is clearly seen the wall (which is actually a plane in the 3D world) and also it is seen the cube which is empty due to the lack of perspective. Finally, the volume is seen from the rear sight, where is seen the wall with those parts where are not information because in both of the images there is an occluder.

In addition, in Figure 5.24 it is shown some results obtained using four cameras from different points of view. The first point of view is from the Front-profile in which it is able to see the lack of information behind the cube. On the second view, it is seen from Rear-profile and it could be seen the depth where is placed the cube, the ball (which corresponds to the gray points) and the wall. In addition, in the 3rd view seen from the rear part, it could be seen the cube, the wall, and part of the picture. Finally, on the last view(second row right), it is seen the top part of the cube, which has a different color than the front part.

Once, we have seen different results from different points of view by using two and four cameras, it is asserted that using only two subaperture images, there are few points, but depending on the point of view it is easier to see the object. While using the information from four subaperture images, some volumes are more visible because there is more information about this object.

As far as we are concerned, the result obtained could be better if instead of using the method of marking four points manually, the correspondences would be found using other methods such as SIFT.
Finally, it is an important point to comment, if we would be able to capture different light field images from the same scene, the 3D reconstruction would have been even better due to the huge amount of data we have had.



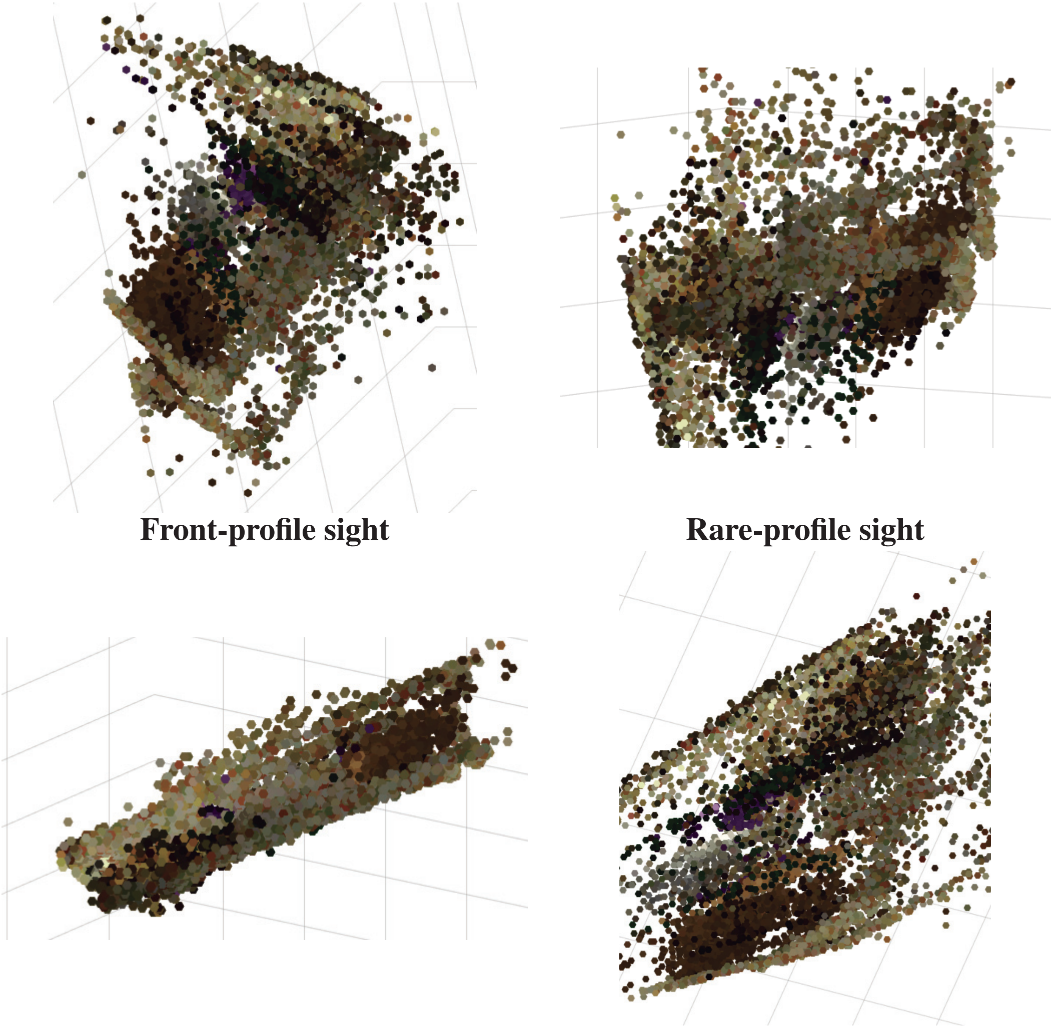
Front sight

Front-profile sight

Profile sight

Rear sight

Figure 5.23: Examples from different points of view of the 3D reconstruction using only two cameras.



Front-profile sight

Rare-profile sight

Rare sight

Profile sight

Figure 5.24: Examples from different points of view of the 3D reconstruction using four cameras.

Chapter 6

CONCLUSIONS AND FUTURE WORK

In this last section, an evaluation of the overall project is done. Firstly, the compliance of the objectives defined at the beginning are commented. Secondly, some conclusions extracted during the development of this project are expressed and finally, some future work and improvements suggested are disclosed.

6.1 Achievements

The main goal of this project is to widen the knowledge significantly on light field imaging because is a novel field introduced in one subject of the Bachelor degree. In order to learn how to deal with the data obtained using a light field camera a rigorous documentation process was done, in which [Wu et al., 2017, Williem and Kyu Park, 2016] were essential documents in order to introduce the key concepts. Moreover, several nice results for different applications have been shown, permitting to come up with solutions to solve some classical problems in image processing and computer vision.
Reviewing the objectives of this project, most of them are successfully achieved. However, if the database would not have been a restriction, some of these results would have been even better.

6.2 Applications Conclusions

For each application, some conclusions are extracted.

- Refocusing is easy to apply by using light field images. Also, the method is really powerful and different applications using refocusing could be done such as a mobile APP which applies some filters on the part which is not on focus.

- Use the information of entropy in order to estimate the depth is a smart option due to the accuracy of results obtained and the efficiency of the process.
- The method proposed to detect the occlusions could be used in many other applications such as 3D reconstruction. For example, when an occlusion is detected, it could be defined to fill the unknown part with the information of the neighborhood.
- 3D reconstruction using light field images is really powerful due to the result that could be obtained, since using only one light field image, the result seems a 3D scene if it is used many light field images the result would be very realistic.

Only reviewing these conclusions above, it could be assumed that light field imaging in the near future is going to be able to come up with even better solutions in order solve some classical problems in the field of image processing and computer vision.

6.3 Future Work and Improvements

Focusing on the application of refocusing, a possible future work would be the one proposed in the previous section. In other words, to program an APP able to apply different filters to those part which is not refocused.
In the field of the 3D reconstruction, a possible improvement would be to use SIFT in order to estimate the parameters of the camera instead of manually marking the needed four points.
Finally, mixing 3D reconstruction and occlusion detection could be applied for reconstructing a 3D depth map that in real time would be able to show the occluded parts.

Bibliography

[Adelson et al., 1991] Adelson, E. H., Bergen, J. R., et al. (1991). The plenoptic function and the elements of early vision.

[Atick et al., 1996] Atick, J. J., Griffin, P. A., and Redlich, A. N. (1996). Statistical approach to shape from shading: Reconstruction of three-dimensional face surfaces from single two-dimensional images. *Neural computation*, 8(6):1321–1340.

[Bando and Nishita, 2007] Bando, Y. and Nishita, T. (2007). Towards digital refocusing from a single photograph. In *15th Pacific Conference on Computer Graphics and Applications (PG'07)*, pages 363–372. IEEE.

[Bardsley and Li,] Bardsley, D. and Li, B. 3d reconstruction using the direct linear transform with a gabor wavelet based correspondence measure. Technical report, Technical Report [online]. 2004,[cit. 2011-05-03]. Available from:¡ http â.

[Blanz et al., 1999] Blanz, V., Vetter, T., et al. (1999). A morphable model for the synthesis of 3d faces. In *Siggraph*, volume 99, pages 187–194.

[Bleyer et al., 2016] Bleyer, M., Rother, C., and Kohli, P. (2016). Surface stereo with soft segmentation.

[Boykov et al., 1999] Boykov, Y., Veksler, O., and Zabih, R. (1999). Fast approximate energy minimization via graph cuts. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 1, pages 377–384. IEEE.

[Chen et al., 2014] Chen, C., Lin, H., Yu, Z., Bing Kang, S., and Yu, J. (2014). Light field stereo matching using bilateral statistics of surface cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1518–1525.

[Cho et al., 2014] Cho, D., Kim, S., and Tai, Y.-W. (2014). Consistent matting for light field images. In *European Conference on Computer Vision*, pages 90–104. Springer.

[Feng et al., 2018] Feng, M., Zulqarnain Gilani, S., Wang, Y., and Mian, A. (2018). 3d face reconstruction from light field images: A model-free approach. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 501–518.

[Fu et al., 2015] Fu, W., Tong, X., Shan, C., Zhu, S., and Chen, B. (2015). Implementing light field image refocusing algorithm. In *2015 2nd International Conference on Opto-Electronics and Applied Optics (IEM OPTRONIX)*, pages 1–8. IEEE.

[Hahne et al., 2016] Hahne, C., Aggoun, A., Velisavljevic, V., Fiebig, S., and Pesch, M. (2016). Refocusing distance of a standard plenoptic camera. *Opt. Express*, 24(19):21521–21540.

[Huguet, 2009] Huguet, F. (2009). *Modélisation et calcul du flot de scène stéréoscopique par une méthode variationnelle*. PhD thesis, Université Joseph-Fourier-Grenoble I.

[Im et al., 2019] Im, S., Ha, H., Choe, G., Jeon, H.-G., Joo, K., and Kweon, I. S. (2019). Accurate 3d reconstruction from small motion clip for rolling shutter cameras. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):775–787.

[Jackson et al., 2017] Jackson, A. S., Bulat, A., Argyriou, V., and Tzimiropoulos, G. (2017). Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1031–1039.

[Jain et al., 1995] Jain, R., Kasturi, R., and Schunck, B. G. (1995). *Machine vision*, volume 5. McGraw-Hill New York.

[Jeon et al., 2015] Jeon, H.-G., Park, J., Choe, G., Park, J., Bok, Y., Tai, Y.-W., and So Kweon, I. (2015). Accurate depth map estimation from a lenslet light field camera. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1547–1555.

[Kolmogorov and Zabih, 2002] Kolmogorov, V. and Zabih, R. (2002). Multi-camera scene reconstruction via graph cuts. In *European conference on computer vision*, pages 82–96. Springer.

[Kutulakos and Seitz, 2000] Kutulakos, K. N. and Seitz, S. M. (2000). A theory of shape by space carving. *International journal of computer vision*, 38(3):199–218.

[Lin et al., 2015] Lin, H., Chen, C., Bing Kang, S., and Yu, J. (2015). Depth recovery from light field using focal stack symmetry. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3451–3459.

[Ng, 2005] Ng, R. (2005). Fourier slice photography. In *ACM transactions on graphics (TOG)*, volume 24, pages 735–744. ACM.

[Seitz and Dyer, 1999] Seitz, S. M. and Dyer, C. R. (1999). Photorealistic scene reconstruction by voxel coloring. *International Journal of Computer Vision*, 35(2):151–173.

[Tang et al., 2014] Tang, S., Andriluka, M., and Schiele, B. (2014). Detection and tracking of occluded people. *International Journal of Computer Vision*, 110(1):58–69.

[Tao et al., 2013] Tao, M. W., Hadap, S., Malik, J., and Ramamoorthi, R. (2013). Depth from combining defocus and correspondence using light-field cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 673–680.

[Tuấn Trần et al., 2018] Tuấn Trần, A., Hassner, T., Masi, I., Paz, E., Nirkin, Y., and Medioni, G. (2018). Extreme 3d face reconstruction: Seeing through occlusions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3935–3944.

[Vaish et al., 2006] Vaish, V., Levoy, M., Szeliski, R., Zitnick, C. L., and Kang, S. B. (2006). Reconstructing occluded surfaces using synthetic apertures: Stereo, focus and robust measures. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2331–2338. IEEE.

[Wang et al., 2015] Wang, T.-C., Efros, A. A., and Ramamoorthi, R. (2015). Occlusion-aware depth estimation using light-field cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3487–3495.

[Wang et al., 2018] Wang, Y., Qiu, J., Liu, C., He, D., Kang, X., Li, J., and Shi, L. (2018). Virtual image points based geometrical parametersâ calibration for focused light field camera. *IEEE Access*, 6:71317–71326.

[Wanner and Goldluecke, 2012] Wanner, S. and Goldluecke, B. (2012). Globally consistent depth labeling of 4d light fields. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 41–48. IEEE.

[Wanner et al., 2013] Wanner, S., Meister, S., and Goldluecke, B. (2013). Datasets and benchmarks for densely sampled 4d light fields. In *VMV*, volume 13, pages 225–226. Citeseer.

[Wei and Quan, 2005] Wei, Y. and Quan, L. (2005). Asymmetrical occlusion handling using graph cut for multi-view stereo. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 902–909. IEEE.

[Williem and Kyu Park, 2016] Williem, W. and Kyu Park, I. (2016). Robust light field depth estimation for noisy scene with occlusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4396–4404.

[Wu et al., 2017] Wu, G., Masia, B., Jarabo, A., Zhang, Y., Wang, L., Dai, Q., Chai, T., and Liu, Y. (2017). Light field image processing: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 11(7):926–954.

[Xiao et al., 2008] Xiao, X., WANG, Z., SUN, C., and BAI, J. (2008). A range focusing measurement technology based on light field photography. *Acta Photonica Sin*, 37(12):2539–2543.

[Zhu et al., 2017] Zhu, H., Zhang, Q., and Wang, Q. (2017). 4d light field superpixel and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6384–6392.