Durham University

Durham E-Theses

Depth Acquisition from Digital Images

DAVID NICHOLAS WILLIAMS

How to cite:

WILLIAMS, DAVID NICHOLAS (2011) Depth Acquisition from Digital Images. Masters thesis, Durham University.

Use policy

The full-text may be used and/or reproduced, and given to third parties in any format or medium, without prior permission or charge, for personal research or study, educational, or not-for-profit purposes provided that:

- a full bibliographic reference is made to the original source
- a https://etheses.durham.ac.uk/id/eprint/3334/ is made to the metadata record in Durham E-Theses
- the full-text is not changed in any way

The full-text must not be sold in any format or medium without the formal permission of the copyright holders.

Please consult the full Durham E-Theses policy for further details.

Academic Support Office, The Palatine Centre, Durham University, Stockton Road, Durham, DH1 3LE
e-mail: e-theses.admin@durham.ac.uk Tel: +44 0191 334 6107
https://etheses.durham.ac.uk

Depth Acquisition from Digital Images

A Per-Pixel Approach using Images Captured at Different Focus Settings with a Conventional Camera

David Nicholas Williams

Thesis submitted for the degree of

MSc by Research

School of Engineering and Computing Sciences

Durham University

2011

Abstract

Introduction

Depth acquisition from digital images captured with a conventional camera, by analysing focus/defocus cues which are related to depth via an optical model of the camera, is a popular approach to depth-mapping a 3D scene. The majority of methods analyse the neighbourhood of a point in an image to infer its depth, which has disadvantages. A more elegant, but more difficult, solution is to evaluate only the single pixel displaying a point in order to infer its depth. This thesis investigates if a per-pixel method can be implemented without compromising accuracy and generality compared to window-based methods, whilst minimising the number of input images.

Method

A geometric optical model of the camera was used to predict the relationship between focus/defocus and intensity at a pixel. Using input images with different focus settings, the relationship was used to identify the focal plane depth (*i.e.* focus setting) where a point is in best focus, from which the depth of the point can be resolved if camera parameters are known. Two metrics were implemented, one to identify the best focus setting for a point from the discrete input set, and one to fit a model to the input data to estimate the depth of perfect focus of the point on a continuous scale.

Results

The method gave generally accurate results for a simple synthetic test scene, with a relatively low number of input images compared to similar methods. When tested on a more complex scene, the method achieved its objectives of separating complex objects from the background by depth, and produced a similar resolution of a complex 3D surface as a similar method which used significantly more input data.

Conclusions

The method demonstrates that it is possible to resolve depth on a per-pixel basis without compromising accuracy and generality, and using a similar amount of input data, compared to more traditional window-based methods. In practice, the presented method offers a convenient new option for depth-based image processing applications, as the depth-map is per-pixel, but the process of capturing and preparing images for the method is not too practically cumbersome and could be easily automated unlike other per-pixel methods reviewed. However, the method still suffers from the general limitations of the depth acquisition approach using images from a conventional camera, which limits its use as a general depth acquisition solution beyond specifically depth-based image processing applications.

Contents

List of Figures, Tables and Equations

Acknowledgements

I would like to thank my supervisor, Ioannis Ivrissimtzis, for his supervision and guidance throughout this project. His enthusiasm, suggestions and ideas have been a source of inspiration for this work.

I would also like to thank Samuel Hasinoff for providing data from his work.

Declaration of Originality

I would like to state that all ideas and methods presented as my own in this thesis are completely original, to the best of my knowledge at the time of writing. Where this work builds upon theory and ideas from other work this is explicitly acknowledged, and all directly relevant references are listed in the bibliography. All written content, diagrams, methods and results presented in this thesis are entirely my own work, unless otherwise explicitly stated.

No part of this thesis has been submitted for any other degree at Durham University or any other institution.

Statement of Copyright

The copyright of this thesis rests with the author. No quotation from it should be published without the prior written consent and information derived from it should be acknowledged.

Glossary

Context-specific definitions of terms used commonly in this thesis.

Depth Acquisition

The algorithmic extraction of depth information about a scene from some input describing the scene. In this work, the input is digital images of the scene.

Depth Map

A two dimensional array describing the depths of scene points, with some pre-defined correspondence (projection) between elements and the location of the scene points they refer to. A depth-map is useful for describing the 3D geometry of a surface.

Conventional Camera

A camera system which can be readily purchased 'off-the-shelf', and functions as the vast majority of camera systems do: by focusing visible light rays through a lens onto a sensor. In this work it is assumed that the conventional camera has manual controls for (at least) focus setting, aperture setting, and exposure time, *i.e.* point-and-shoot compact cameras are not included in this definition.

Geometric Optics

A simple and classical model of the behaviour of visible light. Light is modelled as rays (vectors) which travel in absolutely straight lines, though they can be refracted (for example by a camera lens) to manipulate their direction. Therefore, using simple geometry combined with a mathematical model to describe the focusing behaviour of the camera lens, the propagation of light through the camera system can be predicted.

Depth from Focus (DFF)

The approach to depth acquisition where the level of focus of a point is quantified using some measure, obtained by modelling the visual features perfectly focused points should have. By evaluating the same point in a series of images with different focus levels at the point, the image which displays the point with the maximum level of focus is actively searched for. Assuming the camera settings used to capture the image are known, an optical model of the camera system can be used to relate this to the depth of the point.

Depth from Defocus (DFD)

The opposite approach to depth from focus, depth from defocus models the features of defocus blur in images, and then directly evaluates the blur (if any) around a point in an image in order to infer the magnitude of defocus. Magnitude of defocus is related to depth using an optical model. DFD is a more direct method of depth acquisition which can theoretically work with single input images.

Chapter 1

Introduction

1.1 **Background**

Photography using the conventional camera has been a highly popular form of image capture since its inception, and with the advent of digital photography the possibilities and accessibility of the technology have increased tremendously. A particularly important feature of digital photography is the ability to process digital images with a computer, which has opened up a field of research in Image Processing.

One of the main branches of Image Processing is concerned with the extraction of geometrical structure of the scene depicted in an image. Some of this structure may be directly visible in the image. Additionally, deeper geometrical properties which are not directly available in the image may be inferred by analysing the image in conjunction with some prior knowledge of, for example, the camera system, the scene, or properties of light.

One geometrical property of a scene which is not present in a raw digital image is depth. To the Human viewer an image may appear to depict depth as the relative depths of different objects in the scene seem to be differentiable, but this apparent depth is just an optical illusion based on cues such as perspective. In actual fact, an image is a 2D projection of the 3D scene it depicts (*Figure 1.1.1*).

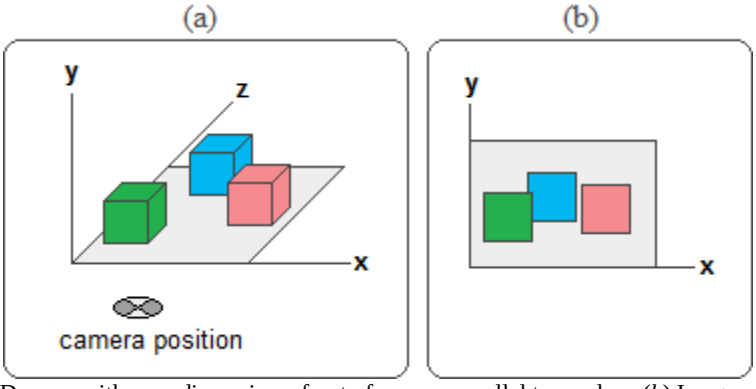


Figure 1.1.1 (*a*) 3D scene with *x,y,z* dimensions, front of camera parallel to *xy* plane (*b*) Image of the scene is a 2D projection, maintaining the *xy* dimensions and flattening the *z* dimension, or depth from the camera.

The 2D projection amounts to a preservation of geometric proportions in the plane parallel to the front of the camera lens, and a flattening along the dimension of depth or distance away from the camera. This loss of the depth dimension is an unavoidable consequence of the way a camera operates, and is therefore a fixed limitation of digital images.

However, if we can assume prior knowledge of the camera system used to capture the image, specifically the optics of the system, an interesting question arises: Is it possible, exploiting this prior knowledge, to reverse the effects of the projection and therefore infer depth information from the image? This question is the basis of an entire field of research, and is the basis of this work.

The justifications for depth acquisition from images are various, but they all stem from a common idea: because the physical world is 3D, the ability to image real-

world scenes in 3D is naturally advantageous. The acquisition of the depth dimension in an image will mean that every point in the image has coordinates in three dimensions. The resolution of 3D structure is a fundamental goal of Computer Vision in itself, but it should be emphasised that far more appropriate hardware than the camera exists to address this problem, for example laser scanners and radar.

The more specialised applications which are solved uniquely by combining digital images with depth information are of far more interest here. The general focus of such applications is on the imaging itself, and the goal is to enhance images beyond what it is possible to capture using the conventional camera, or indeed to achieve with traditional 2D Image Processing. A few examples include arbitrary image re-focusing, image segmentation by depth, object detection and recognition by depth, depth-based image filtering, and depth-based edge detection.

1.2 The Digital Camera

The functionality and operation of the digital camera are at the core of this work. The digital camera operates in much the same way as a traditional film camera, and indeed the basic concept of operation has remained constant since the invention of the device. Figure 1.2.1 shows the arrangement of basic components of the camera.

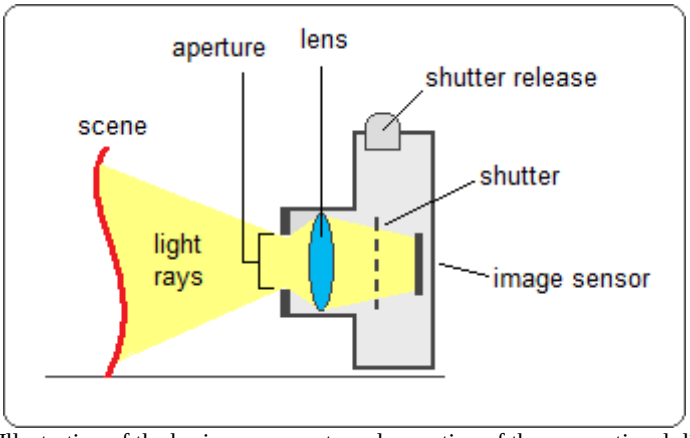


Figure 1.2.1 Illustration of the basic components and operation of the conventional digital camera.

1.2.1 Basic Camera Operation

A simple high-level overview of the operation of a digital camera is as follows: Light rays incident from the scene are allowed into the camera through an opening in front of the lens called the *aperture*. The light rays are focused by the *lens(es)* onto the *image sensor*. Normally the *shutter* blocks the light from reaching the sensor, but when the *shutter release* is pressed the shutter is removed for a pre-defined time interval (exposure time). The light collected by the image sensor during this time forms the *image*.

1.2.2 **The Digital Image**

The operation of the camera described above reveals precisely what a digital image is: A sampling of the visible light from the scene which is focused onto the image sensor. Specifically, the image sensor is made up of a uniform rectangular grid of sub-sensors, typically numbering in the millions. Each sub-sensor is capable of detecting the light particles (photons) which fall on it, from which it can determine the colour and intensity of the light. By collecting photons at each sub-sensor for the duration of the time interval that the shutter is up, a cumulative sample of light with a specific colour gained from the addition of the photons is built up by every sub-sensor.

The resulting digital image is therefore a rectangular array of these samples, where each sample is represented as a pixel. The familiar RGB colour format is assumed to be the norm in this work, where each pixel represents a colour as a combination of the three primary colour channels of Red, Green and Blue.

An important point to discuss in further detail here is the fact that the digital image is a 2D projection of the 3D scene. It is clear from the operation of the camera and specifically the way the digital image is captured by the sensor why this 2D projection occurs. The preservation of geometrical proportions in the plane parallel to the front of the camera is due to the image sensor also being parallel to this plane, and the flattening of the depth dimension due to the light being collected on this plane regardless of the depth of the point of origin.

1.2.3 **Camera Parameters**

It is important to emphasise that the components of the camera system shown in *Figure 1.2.1* are common to all conventional digital cameras. Moreover, all cameras which are manually controllable, regardless of specific functionality or model variations, share a set of three operational parameters which control these components. This is a crucial point as it means that we can discuss and make assumptions about the camera system in an abstract and general way, without worrying about the specific model of camera in use. The only assumption here is that we are referring to a manually controllable camera system as opposed to a fixed-parameter point-and-shoot camera, and from this point onwards that assumption is made.

To follow is a discussion of the three parameters of the camera system referred to above. Specifically, the parameters are focus setting, aperture setting, and exposure time. Generally, for a constant scene and camera position, these parameters can be thought of as entirely controlling the appearance of the final image, and any photographer with experience of manually controllable cameras will be aware of this. The way in which these parameters control and define the optics of the camera system and the effects they have on the image are crucial.

- *Focus Setting*

The focus setting of the camera is an intuitive concept. It determines the distance of the focal plane of the lens, or in other words, it determines the distance from the camera at which a point in the scene is sharply in-focus. Theoretically speaking, all points on the focal plane will be in perfect focus, and a point will become increasingly defocused as it moves away from the focal plane in either direction.

- *Aperture Size*

The aperture size and its effect on the image are less intuitive, but equally important. The aperture is an opening which allows light into the lens system. Conventionally, it is near-circular in shape and is made up of a series of interlocking blades which can slide in synchronisation to change the size or diameter of the opening (see *Figure 1.2.3.1*).

The effect of the aperture size is closely related to the focus setting; it controls the rate of defocus as a point moves away from the focal plane. This determines what is known as the depth-of-field, which is the distance either side of the focal plane at which scene points still appear to be in focus. Note that in reality all points away from the focal plane are defocused, but the defocus can be so slight as to be unnoticeable below a certain distance threshold, leading to the depth-of-field visual effect.

The extremes of aperture size best demonstrate the effect of this parameter. Some visual examples can be seen in *Figure 1.2.3.1*. With very small aperture size, the depth-of-field is very large and all the scene points appear to be in-focus. This effect is demonstrated by the classic pinhole camera which produces an all-focused image of a scene through a tiny opening. On the other hand, as the aperture size increases, defocus is more noticeable as a point moves away from the focal plane. With a very large aperture size, points in the image may appear to be almost instantly defocused if they are not exactly on the focal plane.

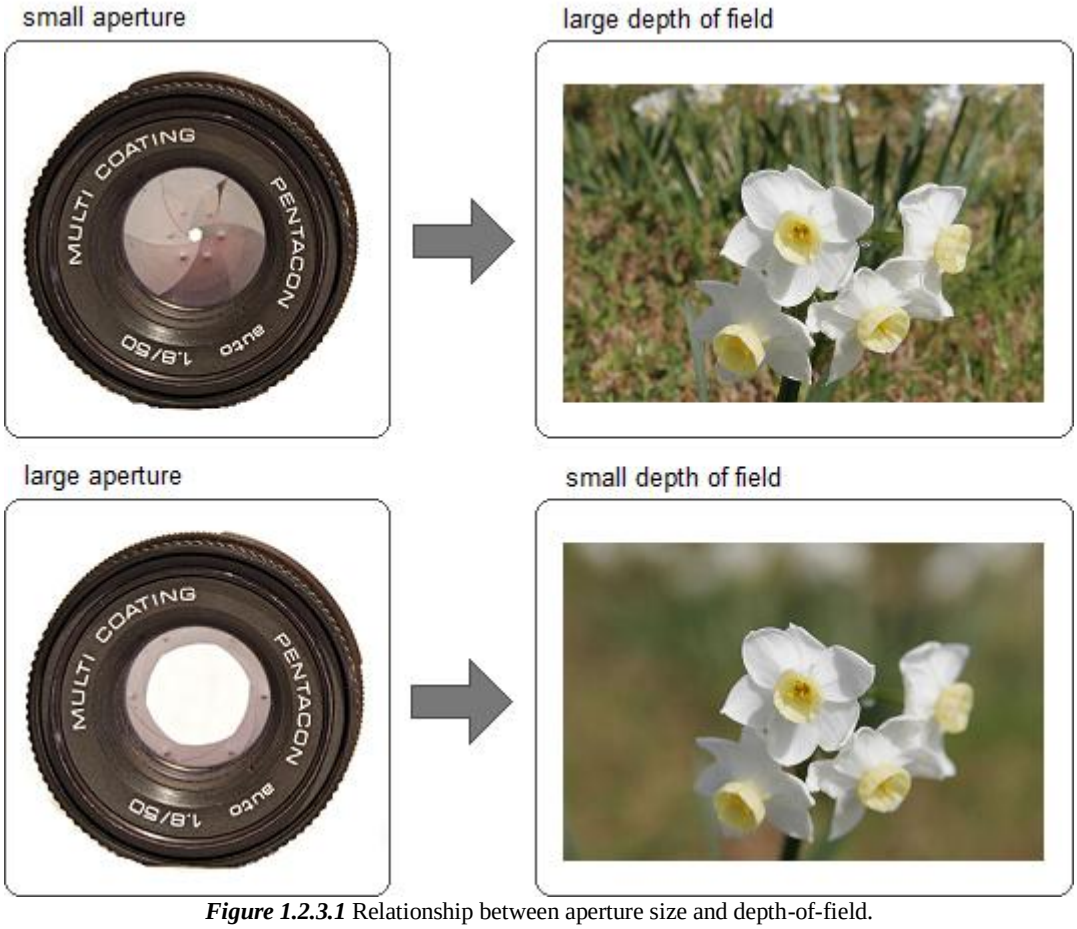


Figure 1.2.3.1 Relationship between aperture size and depth-of-field.
The visual effect of depth of field is shown in the example images. Images taken from www.wikipedia.com

- *Exposure Time*

The most straightforward parameter is exposure time, which is simply the time interval that the shutter is raised for when an image is captured, or the amount of time that light from the scene is allowed to reach the image sensor. Intuitively, if exposure time is too low it will create an under-exposed image which is too dark, and if it is too high it will create an over-exposed image which is too saturated. The goal when setting the exposure time is generally to get the average intensity to a level where colour balance and variance is optimal. Though this may appear trivial, it is a crucial consideration in an Image Processing context because the precise colour and intensity of individual pixels are the data being worked with.

Chapter 2

Optics

Having discussed the basics of digital photography, including the conventional digital camera and the digital images, this chapter will introduce and discuss the optics of the camera system. Using a digital image, the only data we have about a scene is the light received by the camera sensor to form the image. Therefore, in order to interpret the data in images it is essential to understand and develop a model of the optics of the camera system.

2.1 **Focus/Defocus and Depth**

When using a single camera, the most obvious physical cue to the depth of points in the scene is defocus blur. This fact is of paramount significance, and is the concept on which all the techniques and related research described in this work are based.

Defocus blur as a cue to depth is, on consideration, a rather intuitive notion. From common experience, the subject of an image being in sharp focus whilst other objects in front of or behind the subject are blurred due to defocus, should be familiar. Often, defocus blur is used purposefully to draw attention to particular objects in an image or to create visual effects.

As will be discussed in the sections to follow, defocus blur is much more than a visual cue to depth. It is a quantifiable and predicable optical phenomenon, and more importantly for the purposes of this work, it is a fixed function of depth from the camera, under certain assumptions.

In order to infer depth from defocus blur in an image, we must first analyse the optics of the camera system and, as already mentioned, develop a model which attempts to universally describe the function of defocus blur. After such a model has been developed, we have a basis for mathematically analysing the properties of the defocus blur and its functional relationship to depth, which can then be used to compute depth information directly from images.

2.2 **The Camera Lens**

The module of the camera system of most importance when developing a model of its optics is the camera lens. A typical camera lens is displayed in *Figure 2.2.1*.

Figure 2.2.1 A typical camera lens. Image taken from www.wikipedia.com

From the image of a typical camera lens in *Figure 2.2.1*, it is immediately obvious that the lens module is not optically trivial. The aperture and the focus settings are both included in the module, and typically the module contains a series of several lenses, although we will refer to it as a lens (singular) for convenience.

There is no way of predicting how different camera lenses will differ in their number and type of lenses, construction quality and range of aperture and focus settings, and there is indeed a huge variety of variation across these factors in different camera/lens models. However, though these variables exist, the actual function of the lens is known: to focus light from the scene onto the image sensor. The lens can therefore be abstracted as an optical 'black box' which, given a certain input (light from the scene) and a known set of parameters (focus and aperture settings) will always produce the same output (light focused onto the image sensor in a certain manner).

Since this functionality is universal across all lenses, and we know that the parameters will achieve certain pre-defined optical effects, we can assume that any lens will have a uniform set of optical properties. We can therefore include these properties in our optical model of the camera system without worrying about the specific lens being used.

2.3 **The Thin Lens Model**

The Thin Lens Model is a well-known classical optical model for lenses. It is a simple model based on first-order geometric optics. As will be discussed in detail in further sections, the Thin Lens Model will form the basis for both the optical model of the camera system and the model of defocus blur used in this work. The trade-offs involved in this choice of model will also be discussed in more depth, but as will become clear, a primary advantage of the Thin Lens Model is its inherent simplicity, whilst still being able to model all the optical effects required by this work.

It should be emphasised here that the main interest is in the functional predictions of the model, as opposed to the optics themselves. Though a clearly defined model is essential, a thorough examination of the Physics and Optics involved is outside the scope of this work, and therefore concepts are discussed with the assumption that the reader has a foundation in the basic terms and concepts of Optics.

2.3.1 **What is the Thin Lens Model?**

The Thin Lens Model is used to describe the first-order geometric optics of 'thin' lenses. Assuming that the lens in symmetrical around a centre *C*, has spherical edges, and has an optical axis which runs through *C* and is perpendicular to both edges, the definition of 'thin' is that the thickness of the lens (measured between the surfaces along the axis) is negligible in proportion to the focal length of the lens (the distance behind the lens along the axis at which rays from an infinitely distant point converge). This definition is clarified in *Figure 2.3.1.1*.

An important note here is that *Figure 2.3.1.1* depicts a converging lens, which projects a real image onto a screen. The opposite type of lens is a diverging lens, which produces a virtual image which cannot be projected onto a screen but can be viewed by an observer. In the camera system, where a real image is projected onto an image sensor, we are clearly interested in the properties of converging lenses. Therefore, the discussion from here on will focus on the converging thin lens, and unless otherwise stated this is what the term *lens* will refer to.

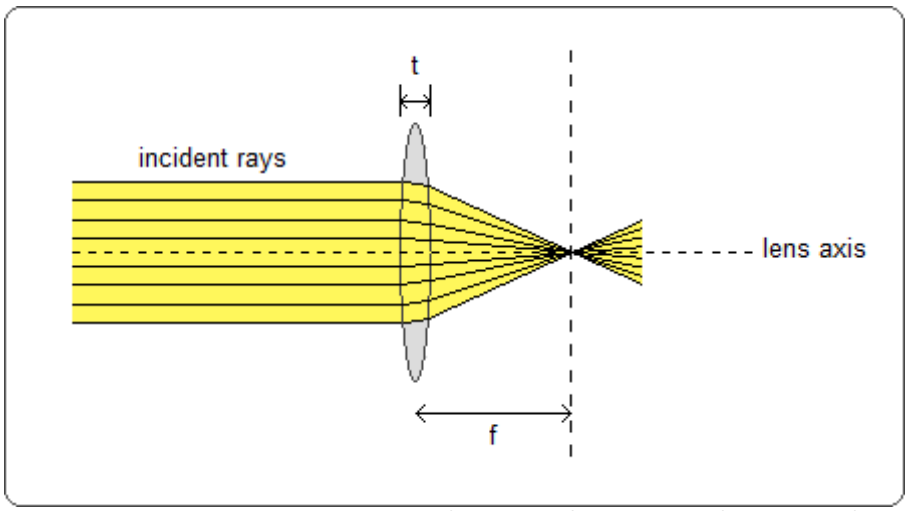


Figure 2.3.1.1 A thin lens with thickness *t* and focal length *f*. The lens is defined as thin if $t << f$. Focal length *f* is defined as the distance behind the lens that incident light rays from infinity converge.

In the Thin Lens Model, the first-order approximation of geometric optics is employed. This assumes that the angle between a ray of light and the optical axis is negligible, *i.e.* below around 10 degrees, an assumption referred to as the paraxial approximation. Under the paraxial approximation any optical effects due to the thickness of the lens, and the distance light rays travel through the lens perpendicular to the optical axis, can be overlooked. The most relevant consequence of this approximation is the widely-known thin-lens formula, which describes the focusing behaviour of the lens.

2.3.2 **The Thin Lens Formula**

$$\frac{1}{f} = \frac{1}{d_1} + \frac{1}{d_2} \qquad \textbf{(Eqn 2.3.2.1)}$$

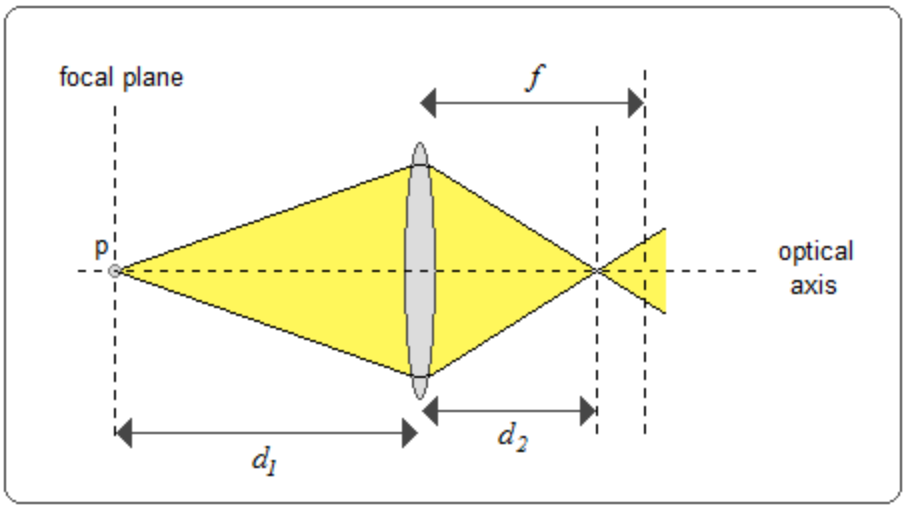


Figure 2.3.2.2 Diagrammatical representation of the Thin Lens Formula (ray diagram not to scale).

The Thin Lens Formula gives the relationship between the distance of a point p in front of the lens, d_1, and the distance behind the lens where light rays from p are brought into focus by the lens, d_2, where both distances are measured along the optical axis. In simpler terms, it describes how the focusing of a point varies with its depth in the scene. By examining the formula, some important properties of this relationship under the Thin Lens Model are revealed.

Firstly, as is implied by the Thin Lens Formula, and can be explained as a consequence of the paraxial approximation, the distances d_1 and d_2 (and f) are taken as parallel to the optical axis. This means that d_1 and d_2 actually specify planes which are perpendicular to the optical axis and parallel to the 'face' of the lens. This will be a fact central to the discussions to follow.

Secondly, we can make some statements about the variation of d_2 as d_1, the independent variable, varies. It is immediately clear that as d_1 increases (a point moves away from the lens) d_2 also increases. However, since f is a constant, the rate of change of d_2 as d_1 increases must slow exponentially, until the limit of $d_2 = f$ is reached when d_1 is infinite. This upper limit of d_2 is the situation shown in *Figure 2.3.1.1*, and indeed is the very definition of f. The lower limit of d_2 is set by the optical limits of the lens, however for simplicity we can assume here that a point will never be close enough to the lens (d_1 will never be low enough) for the lower limit of d_2 to be reached.

2.4 Optical Model of the Camera System

Having discussed the general properties of the Thin Lens Model, the concepts must be used to provide a model for the optics of the camera system, and specifically, to explain the behaviour of defocus blur and its relationship to depth. The optical model of the Camera System described in this section is one commonly used in the literature, and is the model used in this work.

2.4.1 **Assumptions about the Camera Lens**

Before applying concepts from the Thin Lens Model to the camera system, some important assumptions and simplifications have to be made regarding the camera lens. The lens module in a professional camera is typically a series of multiple individual lenses (see *Chapter 1*). This presents an apparent difficulty in applying the Thin Lens Model, which models the optics of a single lens. However, by making two strict assumptions about the lens, it is possible to sidestep this issue and treat the lens module as a single, converging thin lens.

Firstly, we must assume that despite being made up of (potentially) a mixture of converging and diverging lenses, the overall, functional behaviour of the lens module is converging, *i.e.* the lens module focuses incident light rays as a real image that can be projected onto a sensor.

Secondly, we must assume that the parameters of the lens module are fixed at the time of image capture, so that the functional optical properties of the lens module are constant.

If both of these assumptions hold then for a single image capture we can treat the lens module, functionally, as an abstract single converging thin lens with a fixed focal length.

2.4.2 **Application of the Thin Lens Model to the Camera System**

With the camera lens simplification assumed, modelling the camera system using the Thin Lens Model becomes straightforward. The basic configuration is shown in *Figure 2.4.2.1*.

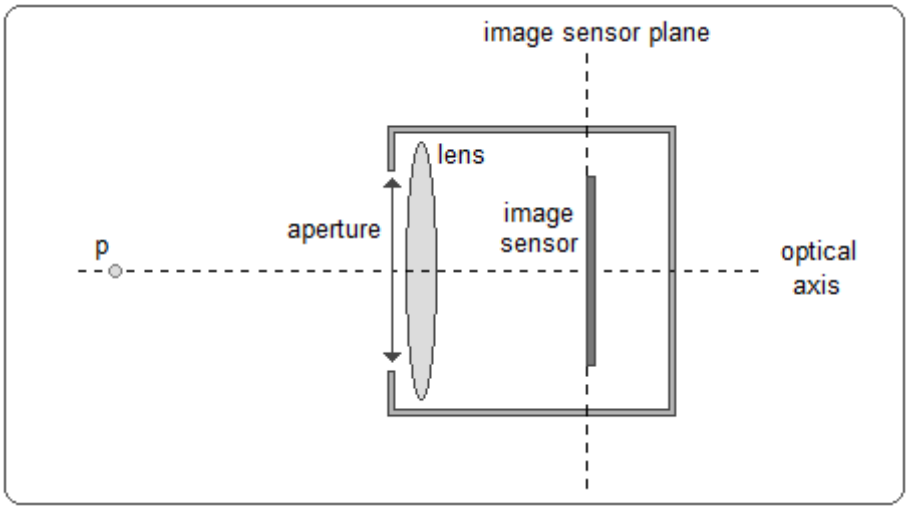


Figure 2.4.2.1 Components of the optical model of the camera system (not to scale).

Figure 2.4.2.1 displays a 2D representation of the camera system which will be used to diagrammatically illustrate the optical model of the camera system. Though only displayed in 2D, visualising the situation in 3D is achieved by simply assuming that the system is symmetrical (apart from the rectangular image sensor) around the optical axis, which of course runs through its centre.

For clarification, the individual components displayed in *Figure 2.4.2.1* are as follows:

- *Scene Point p*

The point *p* is a point in the scene in front of the camera. Scene point *p* represents a point in the scene of infinitesimal size, a concept useful for describing the optical properties of the system without worrying about resolution.

- *Aperture*

The aperture in this model is simply a circular opening in front of the lens, with a given diameter. The only optical effect of the aperture under this model is to block any incident light rays which do not travel freely through the opening, therefore we ignore in this model non-geometric optical effects such as light wave diffraction at the aperture edges.

- *Lens*

The camera lens, as explained in the previous paragraphs, is modelled as a single, converging, thin lens with a set focal length (at capture time). The lens has an optical axis which runs through its centre and indeed through the centre of the entire camera system.

- *Image Sensor*

The image sensor is modelled as being directly behind the lens, at a fixed distance less than or equal to the focal length of the lens (for any given parameters). Importantly, the image sensor lies on a plane which is perpendicular to the optical axis, and the sensor intersects with the optical axis at its centre.

2.5 **Defocus Blur**

Using the defined model of the camera system, both the optical cause and the visual effect of defocus blur can be explained in terms of the depth of points in the scene, and more importantly the relationship between these factors can be quantified. This provides the basis for inferring the depth of a point in an image, directly from the defocus blur (or lack of defocus blur) which can be seen visually in the image at that point.

2.5.1 **Visual Effect of Defocus Blur in Images**

Before the cause of defocus blur is examined, it will be useful to examine the visual effect of the phenomenon in images. The effect is commonplace in everyday photography, and amounts to some spatial area around the blurred point appearing softer, and of lower colour variance. The intuitive concept of a spatial area of some size around the point being related to the blur is crucial.

Perhaps the clearest way to examine the visual effect of defocus blur of a single point is by looking at a single point light source against a black background.

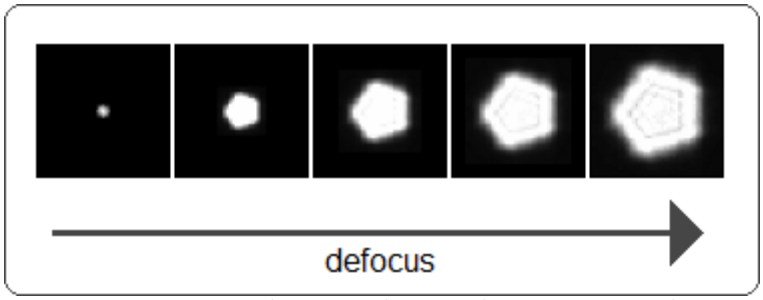


Figure 2.5.1.1 Increasing defocus blur of a point of light (real images from [28]).

Figure 2.5.1.1 shows actual images of a point of light at different levels of defocus. The leftmost image is the original point, and as the level of defocus increases, the magnitude of the blur around the point increases. From these images some important facts about the defocus blur of a point become clear.

- The actual point of light in the scene does not grow in size, but the region of blur around it does grow in size as the point becomes increasingly defocused. The spatial location (in the 2D image plane) of the original point in the scene is therefore at the centre of the blur region, and this location does not change if the camera position is constant.

- The defocus blur region appears to maintain a constant 'shape' which is scaled up as the magnitude of the blur increases. In fact, this shape or form of the blur is defined primarily by the aperture shape, and can be assumed practically constant as long as the aperture shape and size are constant. Here we see a pentagonal blur form indicating a conventional aperture was used to capture these images (see *Figure 1.2.3.1*). A crucial point to introduce here is that this blur form can be modelled mathematically as a point spread function.

2.5.2 **Cause of Defocus Blur in the Camera System**

With an understanding of the visual effect of defocus blur in images, the next logical step is to establish the cause of the effect using the optical model of the camera system. The ultimate goal in doing this is to reveal the relationship between the effect of defocus blur around a point in the image, the cause of this defocus blur in the camera system, and ultimately the depth of the point in the scene.

When examining the optics of defocus blur, the obvious starting point is to look at the special case of a point in perfect focus, *i.e.* when the point has no defocus blur in the image. This special case is shown in *Figure 2.5.2.1*.

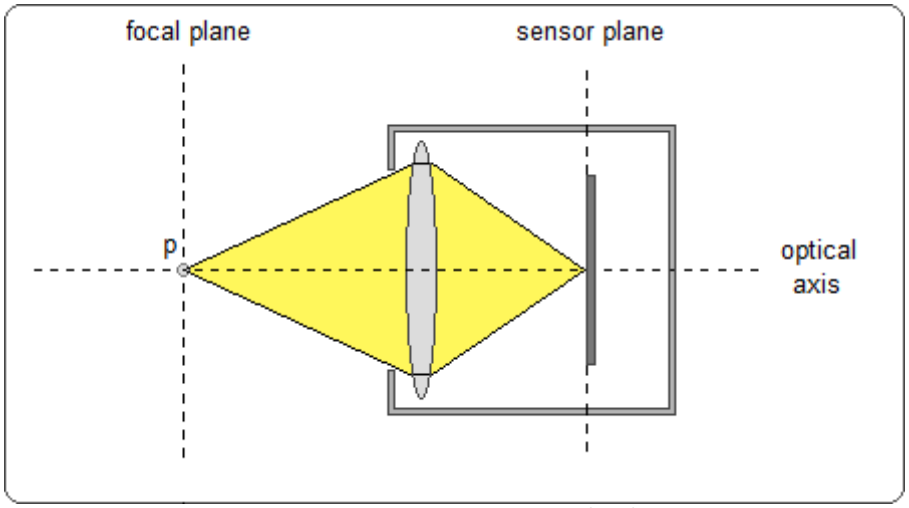


Figure 2.5.2.1 A scene point *p* in perfect focus.

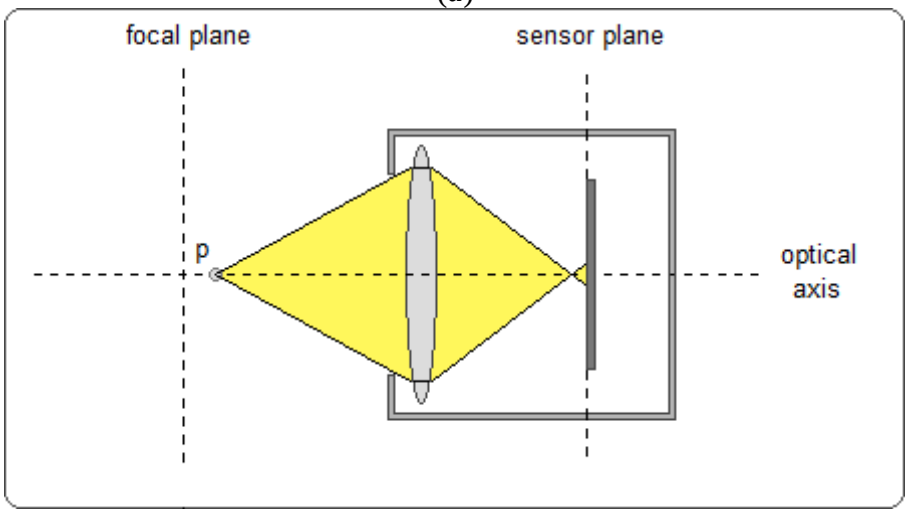
To give a definition, *p* is in perfect focus when the incident light rays from *p* are focused by the lens onto a single point on the image sensor. Hence, the point in the scene is mapped to a point in the image.

This special case of *p* in perfect focus allows us to specify a reference plane termed the focal plane. The focal plane can be defined as the equifocal plane of the sensor plane, *i.e.* it represents the distance along the optical axis at which points in front of the lens are focused perfectly onto the sensor plane, as given by the Thin Lens Formula. In other words, assuming all parameters in the system are constant, the focal plane can be taken as a constant, and any scene point lying exactly on the focal plane will be in perfect focus.

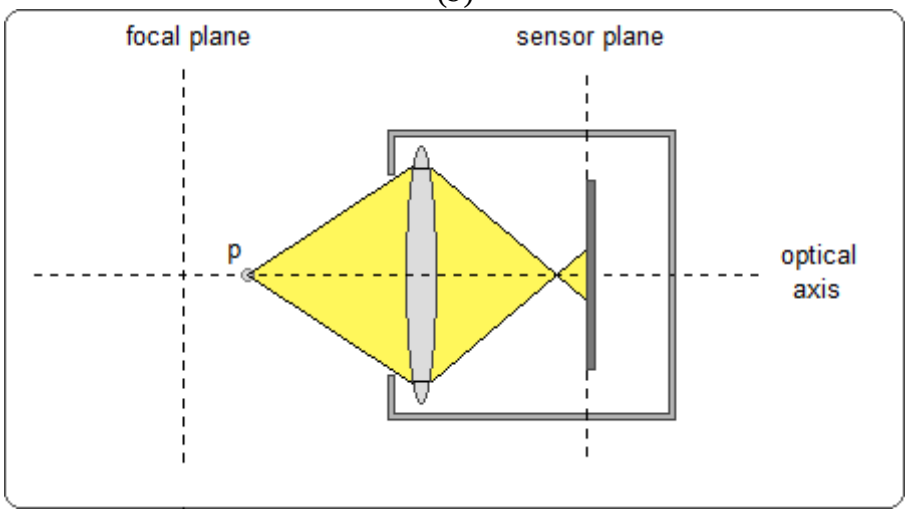
The focal plane is a somewhat abstract concept, yet has real practical value as a reference plane. This is because of a unique property which will be clarified in the discussions to follow: It represents the unique depth in the scene at which *p* has no defocus blur, and as *p* moves away from the focal plane in either direction, the magnitude of defocus blur of *p* increases. Therefore, the focal plane provides a convenient reference depth to which all other depths in the scene can be relative.

Figure 2.5.2.2 illustrates the optics of the system as *p* moves away from the focal plane, towards the camera. As expected, the consequence of this is *p* becoming increasingly blurred due to defocus in the image.

(*a*)



(*b*)



(*c*)

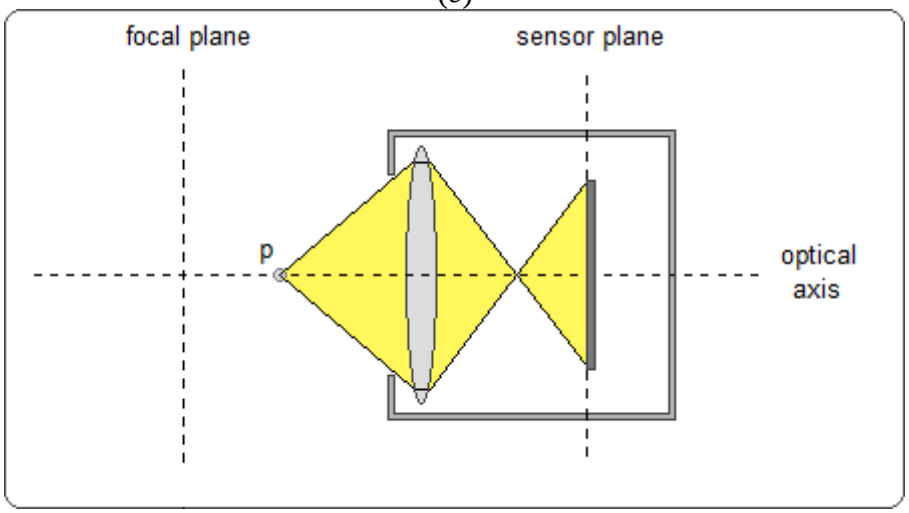


Figure 2.5.2.2 Ray diagrams (not to scale) illustrating the increase in the size of the defocus blur region as point *p* moves away from the focal plane towards the camera in (*a*), (*b*) and (*c*).

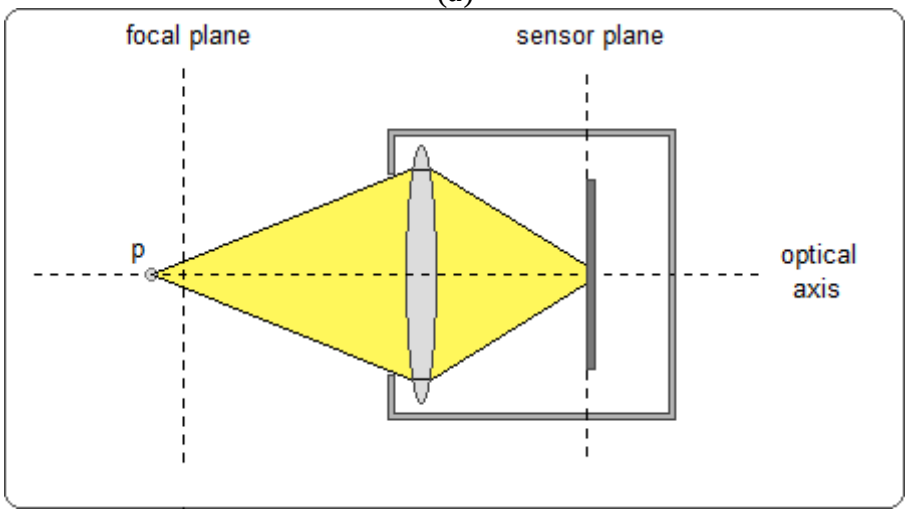
The cause of the defocus blur as *p* moves away from the focal plane and towards the lens can be seen directly in *Figure 2.5.2.2*. The optical explanation is simple, and comes from the basics of the Thin Lens Model. We know from the Thin Lens Formula that as a point moves towards the lens along the optical axis, light rays from the point are focused to a decreasing distance behind the lens.

This is precisely what happens here; the light rays from *p* are focused to a point further in front of the sensor plane as *p* moves further towards the camera. When the light rays converge before the sensor plane, we know from simple geometric optics that they will continue travelling on their respective vectors from the convergence point and hence re-spread. Therefore, the light rays arrive spread over a region of the image sensor, instead of at a point as in *Figure 2.5.2.1*.

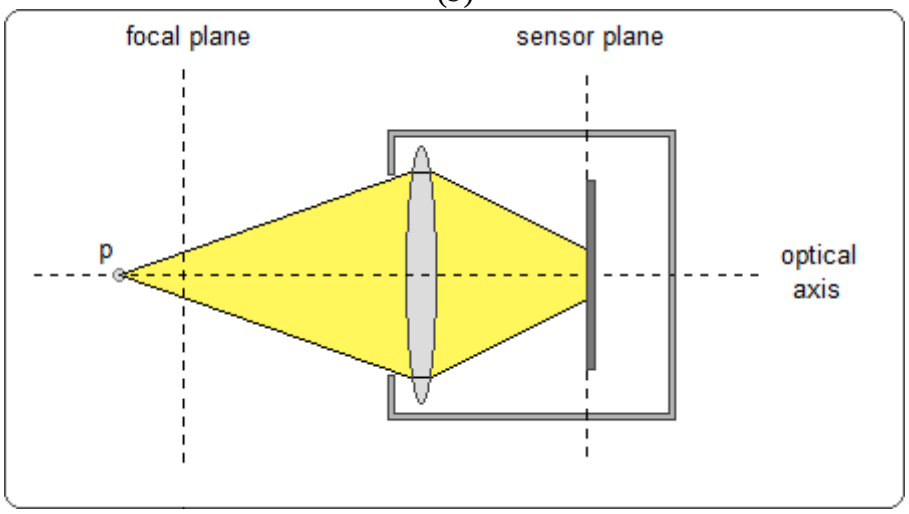
The increase in magnitude of defocus blur as *p* moves further away from the focal plane is simply due to the point of convergence moving further away from the sensor plane, hence the light rays having more distance over which to spread, hence arriving over a larger region of the image sensor, hence the visual effect of greater defocus blur. This is clear from *Figure 2.5.2.2*.

The opposing case of increasing defocus blur as *p* moves away from the focal plane and the camera towards infinity is similar in principle to the above, but with an important difference. In this case, the defocus blur is again caused by the light rays from *p* arriving over a region of the image sensor, but this happens because the light rays have not yet converged, *i.e.* the theoretical point of convergence is behind the image sensor (the light rays never reach this point as they are blocked by the image sensor). Again, this focusing behaviour is predicted by the thin lens formula. *Figure 2.5.2.3* illustrates the increase in defocus blur magnitude as *p* moves further towards infinity.

(*a*)



(*b*)



(*c*)

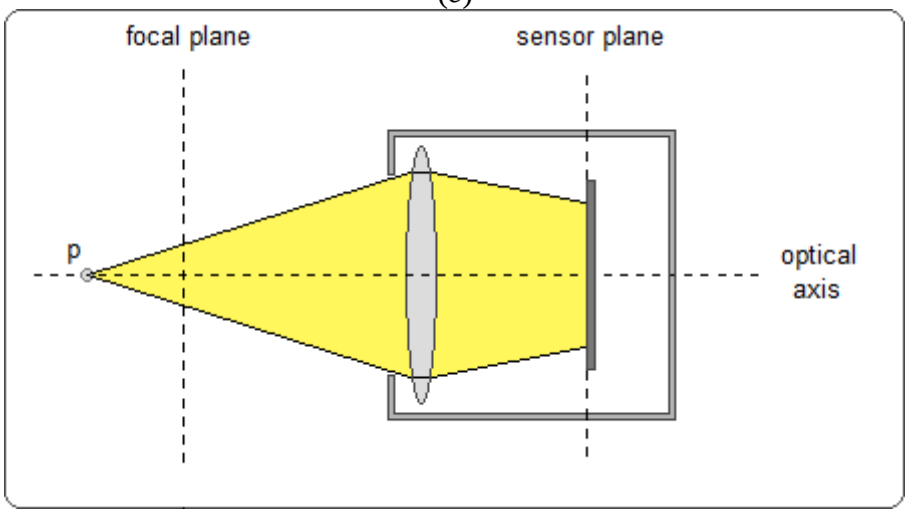


Figure 2.5.2.3 Ray diagrams (not to scale) illustrating the increase in the size of the defocus blur region as point *p* moves away from the focal plane towards infinity in (*a*), (*b*) and (*c*).

Chapter 3

Focus / Defocus and Depth

This chapter will discuss how, using the optical model of the camera system developed in *Chapter 2*, the depth of a 3D point in a scene can be related to, and therefore inferred from, the level of focus or magnitude of defocus blur of the point in an image captured using a conventional camera.

3.1 **Depth from Defocus**

The optical model of the Camera System can explain the cause of defocus blur, and gives some basic mathematical relationships between the depth of a point light source in a scene, and the magnitude of defocus blur around that point in the image.

It follows that the depth of a point could be directly inferred from the defocus blurring of the point in the image, by analysing the properties of the blur, and working in reverse through the optical model to infer the depth of the point in the scene.

This approach is referred to in the literature as Depth-From-Defocus (DFD). It is a very popular and mainstream approach to the problem of depth acquisition from digital images, and the most important aspects of this approach will be discussed in this section.

3.1.1 **Point Spread Functions**

Under the camera system model used here the defocus blur of a point, as previously mentioned, can be described using a point spread function (PSF). The concept of this is simple; the blur is caused by light from a point being spread over a region, hence a PSF can be used to model the blur. Note here that it is assumed that the in-focus point would appear at the centre of the region of blur in the image. This assumption can be justified using geometric optics, and can be seen clearly in *Figure 2.5.2.2* and *Figure 2.5.2.3* where the centre of the blur region, and the position of the in-focus point, is given by the optical axis.

A PSF describing blur can be said to have two components: form and scale. The form of the function describes the ‘shape’ of the blur region, and the scale describes the size (diameter) of the blur region. The concepts of shape and size of a blur region can be seen by referring again to *Figure 2.5.1.1*, where the blur of a single point keeps a pentagonal shape (PSF form) and increases in size (PSF scale) as the magnitude of blur increases.

To understand the meaning of the shape and scale of the blur region, it is best to visualise the situation in 3D. *Figure 2.5.2.1*, *Figure 2.5.2.2* and *Figure 2.5.2.3* illustrate the optics of defocus blur in 2D. By picturing the light ray triangles in these figures as 3D light cones in the real camera system, the 3D situation can be conceptualised easily. In 2D, the effect of the aperture on the shape of the light cone is not emphasised, but when visualising the light as a cone, it can be seen immediately that the cross-section of the cone must take the shape of the aperture opening, as the

aperture blocks completely any light rays which collide with its edges. This assumes a geometrical model of light rays, and ignores wave effects such as diffraction at the aperture edges.

Since the cross-section of the light cone takes the shape of the aperture, it follows that the intersection of the light cone with the image sensor must also take this shape (the paraxial approximation means here that any skewing of the shape due to angles of incidence can be assumed negligible).

Visualising the blur region as an intersection of the light cone with the image sensor, it is clear how the region can be said to have a constant shape related principally to the aperture, with a diameter which related to the depth of the point source of the light cone by the Thin Lens Formula and simple geometry.

Therefore, for a given camera system under constant parameters, the defocus blur can be modelled by establishing a functional form describing the shape of the blur. Then, the scale of the PSF gives the depth of the blurred point.

3.1.2 **PSF Form**

Establishing the functional form of the defocus blur PSF is non-trivial. From the optical model of the camera system, it might seem theoretically possible to attempt to directly calculate the shape of the blur region using ray-tracing, and hence estimate the form of the blur PSF under the optical assumptions of the model. However, in real images, complex optical effects which the model does not account for cause the real blur PSF to be more complex, both in overall shape and in the spread of light throughout the blur region.

See for example *Figure 2.5.1.1*. Though the outer edges of the blur region take the pentagonal shape of the aperture as predicted by first-order geometric optics, the spread of light within the region of blur is not constant, but forms a complex pattern. This is due to optical effects (both geometric and non-geometric) of the complex arrangement of individual lenses in a real camera which cannot be predicted by the model, and cannot be generalised across camera/lens module models.

The complexity of the 'real' blur PSF form means that in practice it must be approximated. There are two general approaches to approximating this PSF function form.

3.1.2.1 **Models Approach**

The first approach attempts to employ a simplified model, a 2D function with few parameters, to approximate the real PSF form. Such models are generally based on a combination of assumptions from optical models and observations of real camera systems. They are fully intended to provide a trade-off between accuracy and practicality. The practical advantage of these models lies not only in their simplicity, but their ability to be scaled continuously. This means that the blur PSF can be applied at any scale, where the loss in information is limited only by the resolution of the image being analysed. Some examples of commonly used defocus blur approximations in research are detailed below.

- *Pillbox Model*

The pillbox model is by far the simplest defocus blur model. In short, it is the approximation described in the first paragraph of this section which treats the blur region as aperture-shaped with a constant, uniform spread of light rays throughout the region. This is the model directly predicted by first-order geometric optics. As a further simplification, the shape of the aperture is taken as the ideal circle, rather than the more realistic pentagon. The function P_r, where r is the radius of the blur region, is shown in *Eqn 3.1.2.1.1*, and the plot of the function is illustrated in *Figure 3.1.2.1.2*.

$$P_r(x,y) = \begin{cases} \frac{1}{\pi r^2} & x^2 + y^2 \leq r^2 \\ 0 & x^2 + y^2 > r^2 \end{cases} \qquad \textbf{(Eqn 3.1.2.1.1)}$$

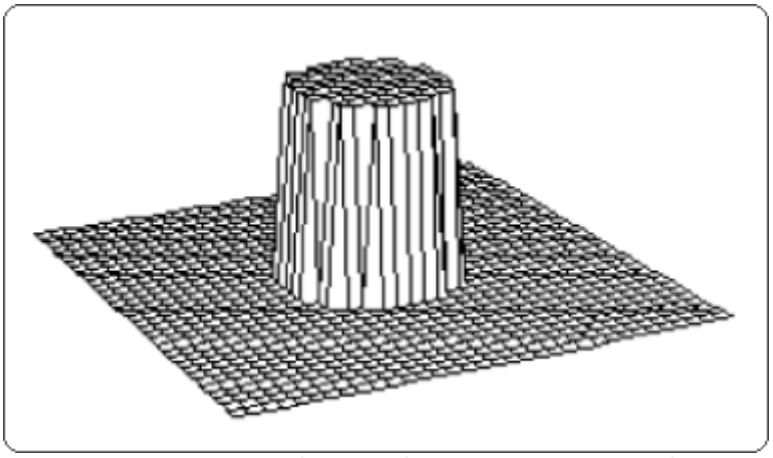


Figure 3.1.2.1.2 Illustration of the plot of the Pillbox Model. Taken from [23].

- *Gaussian Model*

As mentioned in the opening paragraph, the pillbox model of defocus blur is the prediction of simple first-order geometric optics. In practice, the optics of real camera lenses are very complex, and are affected by optical phenomena such as aberrations (both geometric and non-geometric) and diffraction, and also due to unpredictable lens configurations and imperfections. Because of this, the pillbox model with its uniform, constant spread of light is often an inappropriate approximation of the real blur PSF form.

A far more accurate model for blur in practice is the circular 2D Gaussian of blur in an image (blur magnitude is often seen 'falling off' away from the centre of the blur region). The equation for the circular Gaussian defocus blur model is given by *Eqn 3.1.2.1.3*, where r is the standard deviation of the circular Gaussian and is proportional to the radius of the blur region. The function plot is illustrated in *Figure 3.1.2.1.4*.

$$G_r(x,y) = \frac{1}{2\pi r^2} e^{-\frac{x^2+y^2}{2r^2}} \qquad \textit{(Eqn 3.1.2.1.3)}$$

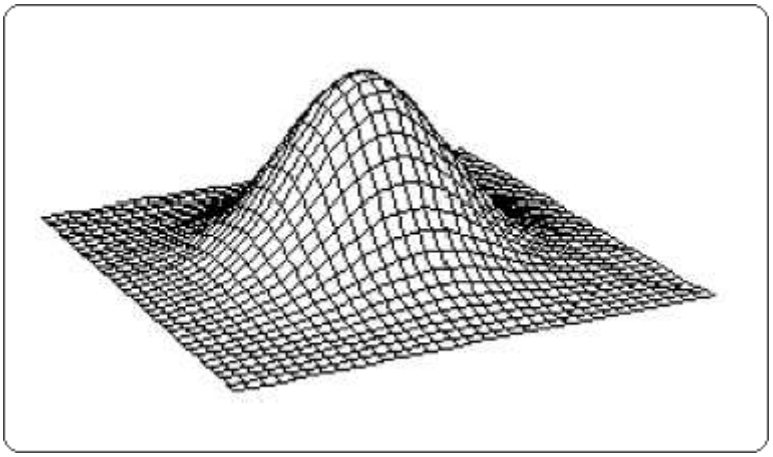


Figure 3.1.2.1.4 Illustration of the plot of the Circular Gaussian Model. Taken from [23].

3.1.2.2 **Empirical Approach**

The second approach to blur PSF form approximation is the empirical approach. Here the idea is to sidestep the modelling of complex lens optics, and acquire the blur PSF directly from images captured using the camera system.

A common method of doing this involves sampling real point-spreads from images captured in a calibration stage. For example, the images in *Figure 2.5.1.1* could be used (and probably were used) to provide samples of real point spreads for a particular camera system. In this way, a discrete approximation of the point-spread function of the camera system can be acquired from real image data. The advantage of this approach is its ability to capture the 'real' blur PSF for a particular camera system as a discrete function, completely sidestepping the issue of accurately modelling the complex optics of the system, which is a difficult problem and may result in inferior simplified models [49]. However, there are also some clear disadvantages compared to the first approach: Scales between those sampled can only be estimated by interpolation, the acquisition of samples could be prone to error, and of course the whole process is cumbersome considering it must be performed not only for each camera system used, but also for each set of lens parameters used.

3.1.3 **PSF Scale**

The approximation of the defocus blur PSF form is the essential first step in relating defocus blur in images to depth. However, assuming we have a model of the PSF form, how do we proceed to infer depth directly from images using this model?

The principle of relating defocus blur to depth, knowing the functional form of the blur PSF, is to identify the correct scale of the PSF which caused the blur. As discussed previously, by analysing the optics of the camera system we can conclude that there is a mathematical relationship between the radius of the defocus blur region and the depth of a blurred point. Hence, by identifying the correct scale of the PSF

which has caused the defocus blurring of a point in the image, we can directly infer the depth of the point in the scene.

This process can be broken down into two general stages. The most interesting stage here, the identification of the correct scale of blur PSF, leads to the second stage of relating this scale to depth by some conversion process.

By far the most popular approach to describing defocus blur in the Image Processing literature is as a linear filtering of a perfectly focused image with a defocus blur kernel (PSF). By following this convention, an image or a small window of an image which contains defocus blur can be described by the following convolution equation:

$$I = \beta_r * \hat{I} \qquad \textbf{(Eqn 3.1.3.1)}$$

In *Eqn 3.1.3.1*, I is the observed image window, $\hat{I}$ is the perfectly focused version of the image window, β_r is the defocus blur PSF with scale r, and * indicates the convolution operation.

The intuitive explanation of this convolution operation is that the observed image is a result of the filtering of the perfectly focused image signal with the blurring signal. The idea is therefore to remove the blurring signal from the observed image, to reveal the perfectly focused image. Only the correct scale of the defocus blur PSF, or the blurring signal, will reveal the perfectly focused image, and this is the basis on which we can identify the correct scale of blur. Mathematically, this process of removing the blur PSF from the observed image is known as deconvolution.

The importance of a prior model on the form of the blur PSF becomes apparent in the deconvolution stage, because it greatly simplifies the process. For example, one simple approach to the identification of the correct blur scale is to test a number of different scales of the known PSF, and then analysing the resulting image for each deconvolution. The image which is judged to be in best focus can then be used to identify the correct scale of blur PSF.

The specifics of blur scale identification through deconvolution are complex. There are many methods, and this process constitutes a field of research in Image Processing in its own right, therefore it is somewhat outside the scope of this work to give a full discussion of the various methods here. However, some of these methods which are specific to the context of this work will be examined in further detail in *Chapter 4*.

3.1.4 **Relationship between PSF and Depth**

Assuming that for a given blurred point, the correct scale of the defocus blur PSF which caused the blur has been identified, the remaining step in depth acquisition is to relate the blur PSF scale to the actual depth of the point in the scene.

This stage is considered in this work to be of lesser importance, as it is principally determined by the application, and can be considered as an additional independent stage. Nonetheless, it is important to discuss for completeness.

The actual type of depth information required for the application largely determines how blur scale is related to depth. For example, some applications require real, physical measurements of depth (in distance units) whereas others, particularly Computational Photography and Image Processing applications, require only a

relative measure of depth, or even as little detail as a grouping or simple ordering of points by depth.

From previous discussion of the Thin Lens Model and geometric optics, we know that the relationship between blur PSF scale and the depth of a point is mathematical, and predictable. Whilst it would in theory be possible to recover this relationship analytically using ray-tracing through the camera system, this method would likely be too inaccurate and/or cumbersome in practice. A far more common method would use an empirical calibration stage. For example, a scene with depth markers could be used to determine the actual scale of blur at different depths, directly from the images captured with the camera system under fixed settings.

In the case where only relative depth, or some form of depth index is required for the application, the relationship between the blur PSF scale and depth becomes far simpler. In this case, the depth value can be related directly to the blur scale.

3.2 **Depth from Focus**

Another mainstream approach to the problem of depth acquisition from digital images is termed Depth-from-Focus (DFF). As implied by the name, this approach takes the problem from a different perspective, and in many respects is the opposite of the Depth-from-Defocus (DFD) approach (*Section 3.1*).

With DFD the emphasis is on evaluating the properties of defocus blur of image points, and using prior models based on camera system optics to directly infer the depth of the points from the blur. DFF, on the other hand, is less concerned with directly modelling the optics of defocus blur (although the optics are always an important consideration), and more concerned with evaluating the actual effects of the blur in images to determine whether or not a point in the image is in-focus.

In general, DFF can be regarded as a search problem, where the goal is to identify, from a set of images of a scene captured using different camera parameters, the image where a point is most focused. This implies that a number of assumptions must be made about the set of images:

- *Single Viewpoint*

The analysis of variations over a single point implies that the set of images must be captured from a single, constant viewpoint by a single camera. There is also an implication here, of course, that the scene is static. Whilst initially seeming to be a limitation, the situation of a fixed camera position to 'frame' a scene is common practice in current photography, and in many application cases it is appropriate to assume that the scene is static.

- *Geometric Alignment*

In order to identify the same scene point in a set of images, the geometrical mapping between the 2D position of a scene point in any two images in the set must be known (often it is convenient to pre-process images so that this is a 1:1 mapping). This is one of the areas in DFF where a strict model of optics must be employed to determine the mapping.

- *Radiometric Alignment*

In addition to being geometrically aligned, it is often essential in DFF based methods that the set of images must be radiometrically aligned. The variation of camera parameters and stochastic variations in scene illumination between images can cause unwanted variations in the colour intensity of a scene point between images. Since it is sometimes this variation which is being analysed in a DFF method, it is obviously crucial to normalise the intensity of the set of images, taking these variations into account.

- *Global Depth Interval*

As the DFF approach is a search problem, it follows that the range of focus settings of the input images should be global over the scene. This means that every point in the scene should come into focus (so far as settings allow) in at least one image. This of course applies only to scene points within the depth range of interest to the application.

3.2.1 **The Optics of DFF**

In terms of optics, the DFF approach is somewhat the inverse of the DFD approach. In previous discussion of the optics of defocus blur, the emphasis has been on the points in the scene. The explanation of defocus blur used in DFD approaches explains blur in terms of light from each point source in the scene spreading over a region of the image sensor. The converse perspective of defocus blur, used in DFF approaches, is to consider the source(s) of the light which arrives at each point on the image sensor.

From this new perspective on defocus blur, a point in an image in defined as perfectly in-focus if all the light reaching that point on the image sensor came from a single point-source in the scene. An image point becomes defocused when it displays light from more than one point in the scene, with the magnitude of defocus increasing as the number of scene points where the light originated from increases.

The above definitions, as already mentioned, are just an alternative way of phrasing the same optics of defocus blur described in *Section 2.5*. *Figure 3.2.1.1 and Figure 3.2.1.2* illustrate this, and are similar to *Figure 2.5.2.1, Figure 2.5.2.2* and *Figure 2.5.2.3*, but with some important differences.

Firstly, it must be emphasised here that the camera settings are variable, and the scene is static. Therefore, a surface from the scene is pictured at a static position, while the focal plane of the lens varies. The varying focal plane signifies a varying focal length of the lens, which is a simplified abstraction of the varying focus setting of the camera.

Secondly, *Figure 3.2.1.1* and *Figure 3.2.1.2* show only the light which converges on a particular point on the image sensor. The implication is that these light rays are incident from any scene point which lies within the light cone. Of course, these are not the only light rays from these scene points which will end up reaching the image sensor, but they are all the light rays from the scene which are focused by the lens onto the specific point on the image sensor. This again highlights that we are looking at the optics of defocus blur from the perspective of a point on the image sensor rather than from the perspective of a point in the scene.

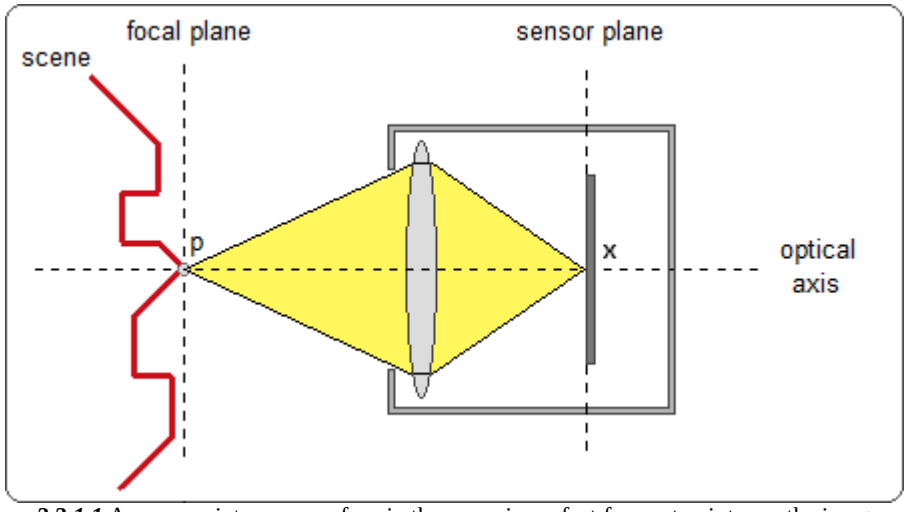
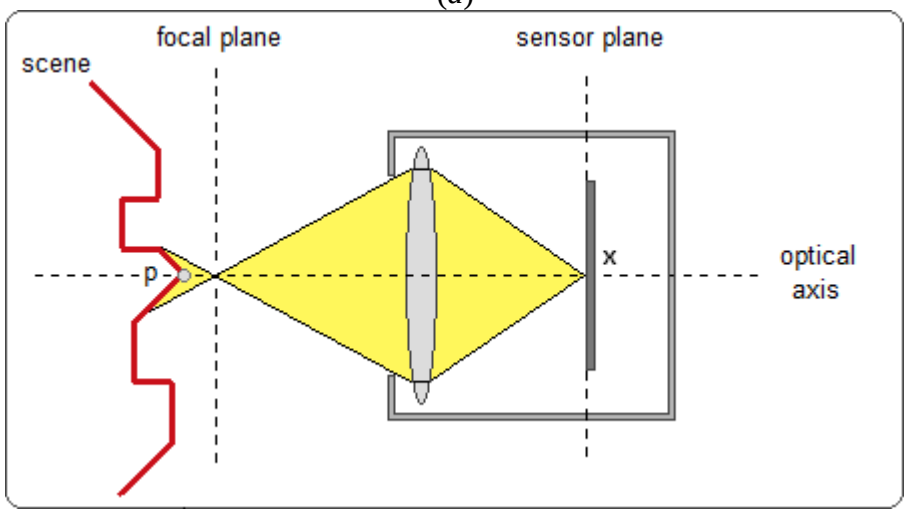


Figure 3.2.1.1 A scene point *p* on a surface in the scene in perfect focus at point *x* on the image sensor (not to scale).

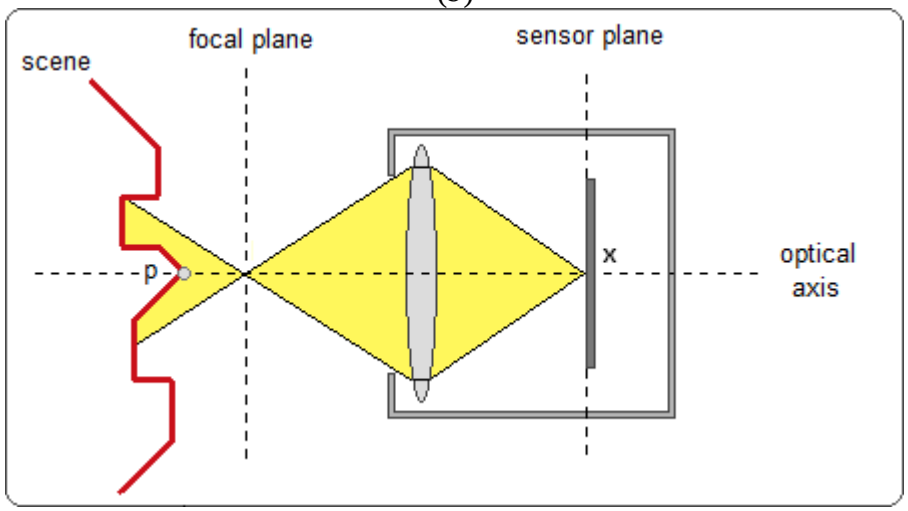
Figure 3.2.1.1 should be familiar from *Section 2.5*, as it is essentially identical to *Figure 2.5.2.1*. The reason for this is that *Figure 3.2.1.1* shows the special case of a scene point *p* being brought into perfect focus on the image sensor; a case which looks identical from any perspective. However here *p* is not the emphasis, but rather a point *x* on the image sensor where the light rays from *p* converge. The reason *x* will display a perfectly in-focus point in the image is that all the light rays reaching *x* on the image sensor originate from the single scene point *p*.

This special case of *x* displaying the perfectly focused point *p* occurs at precisely the focus setting which means that *p* lies on the focal plane. If the focus setting of the camera is changed, then the position of the focal plane will change and *x* will show a defocused point. As should be clear from *Section 2.5*, the further the focal plane is moved away from the perfect focus position, the more defocused the image will be at point *x* on then image sensor. *Figure 3.2.1.2* illustrates the optics of this increasing defocus at point *x* as the focal plane moves towards, and away from, the camera from the perfectly focused position.

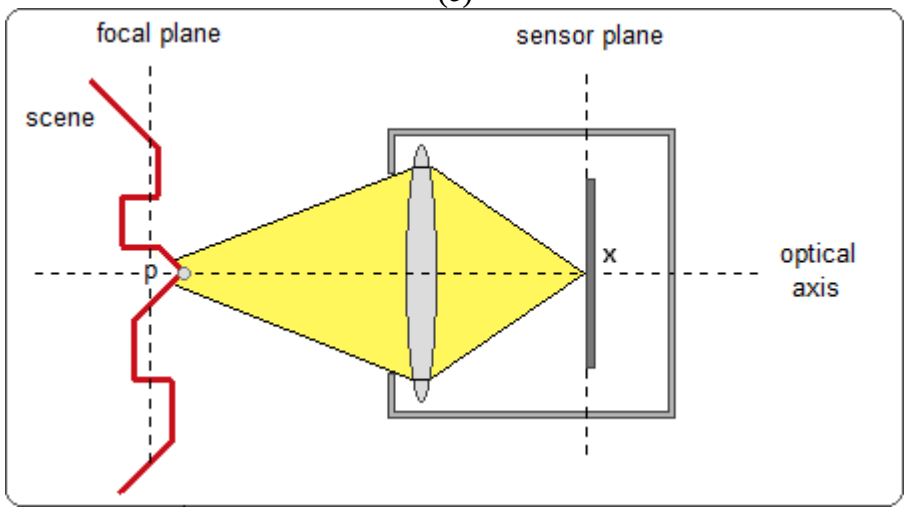
(*a*)



(*b*)



(*c*)



(*d*)

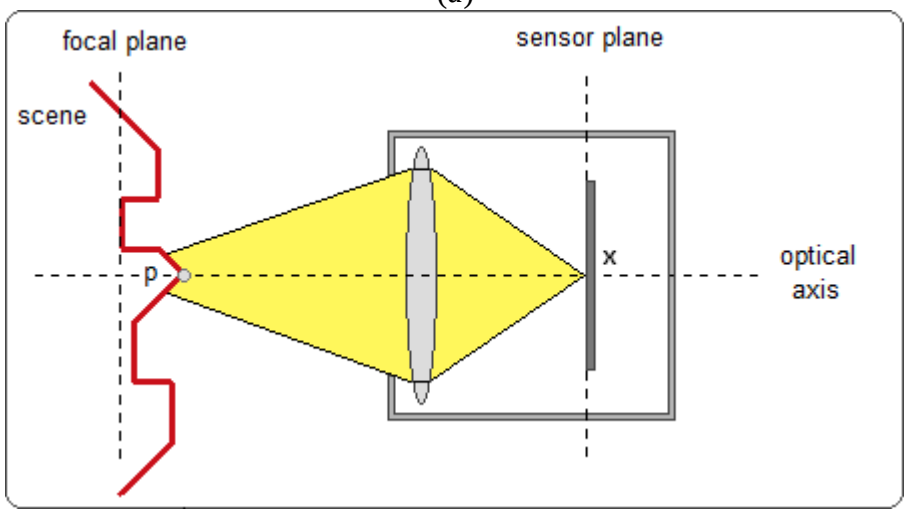


Figure 3.2.1.2 Ray diagrams (not to scale) illustrating the increasing size of the region of the surface in the scene around point *p* which is focused to point *x* on the image sensor as the focal plane moves towards the camera in (*a*), (*b*) and away from the camera in (*c*),(*d*).

Figure 3.2.1.2 is very similar to *Figure 2.5.2.2* and *Figure 2.5.2.3* (*Section 2.5*). Again, the defocus blur at point *x* on the image sensor is caused by a spread of the light cone, but from this perspective we think of the light cone as being spread over a region at the source, and converging to a precise point at the destination (the image sensor).

Here, again from simple geometric optics, we can see how as the focal plane moves away in either direction from the position where *p* is in perfect focus, the light cone integrates a larger region of the scene (where the region is parallel to the sensor plane) and hence point *x* in the image becomes more defocused.

The position of point *p* is included in all the diagrams to highlight the most important feature of the DFF approach: The goal is to find the image at which a given point in the image, *x*, shows a scene point *p* in perfect focus. We can see from *Figure 3.2.1.2* that some light from *p* will always reach point *x* on the image sensor. When *p* is in perfect focus, point *x* receives all the light rays coming from *p*, but as *p* becomes more defocused, fewer light rays from *p* end up reaching *x* as they are spread over a larger and larger region surrounding *x* on the sensor (see *Section 2.5*). Put another way, as *p* becomes more defocused, *x* receives a lower proportion of light rays from *p*, and an increasingly greater proportion of light rays from the region surrounding *p* in the scene. This fact is central to the DFF approach.

It should be noted here that the above interpretation of the optics of defocus only holds under certain assumptions about the scene: That surfaces in the scene are predominantly fronto-parallel, mostly continuous, fairly large relative to image resolution, and negligibly self-occluding.

3.2.2 **Depth Resolution**

Since DFF is a search based method, unlike DFD, the acquisition of depth using the DFF approach has an inherent limitation on the resolution of depth-mapping that can be achieved. The set of images used as input to a DFF search are equivalent to a depth-based sampling of the scene, where the resolution of this sampling depends on two key factors:

- *Grain of Focus Setting*

The focus setting of the camera system in use is a limitation on the depth resolution in the set of images. The focus setting on modern digital cameras is often on a discrete scale, but even if on a continuous scale the setting can only reasonably be set by the photographer to a certain precision. Since each successive focus setting will move the focal plane by some minimum distance in the depth dimension, this defines the maximum grain of depth resolution available using that particular camera system.

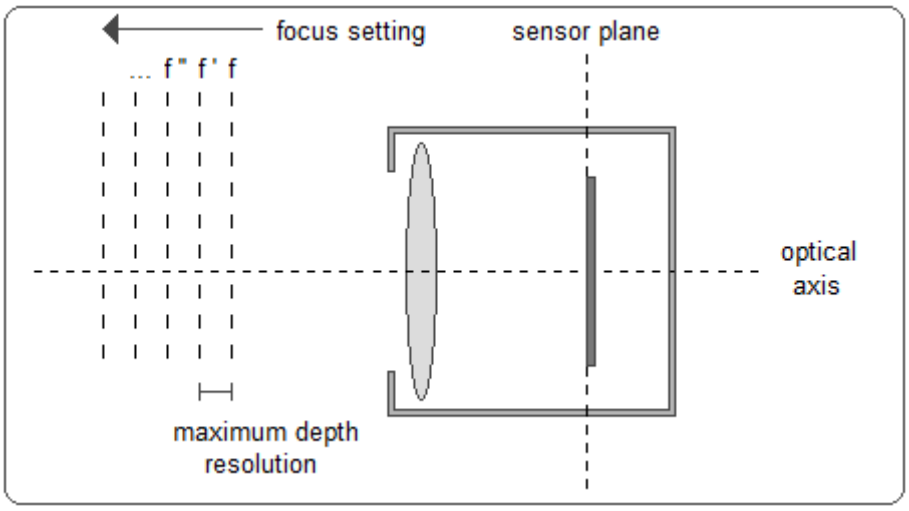


Figure 3.2.2.1 The minimum distance between the focal planes of two successive focus setting defines the maximum depth resolution of the camera system (diagram not to scale).

- *Depth-of-Field*

The maximum depth resolution of the camera system is not only dependent on the minimum distance between the focal planes at successive focus settings, but the ability to discriminate between the blur of a point between these focus settings.

For example, suppose a point *p* is in perfect focus at focus setting *f*, and that the next available focus setting is *f'*. The depth distance between the focal planes at *f* and *f'* defines the maximum depth resolution only if the defocus blur of *p* at *f'* is distinguishable from the defocus blur of *p* at *f*.

The minimum distance away from the focal plane at which defocus blur becomes distinguishable from perfect focus is termed the depth-of-field of the camera system. As previously explained, the depth-of-field is a visual phenomenon and not an optical one. The reality is that defocus blur increases continuously as a point moves away from the focal plane, but to the human viewer, and to some extent in Image Processing because of the limits of image resolution, there is a certain threshold beneath which the defocus blur is indistinguishable from perfect focus. This threshold defines the depth-of-field.

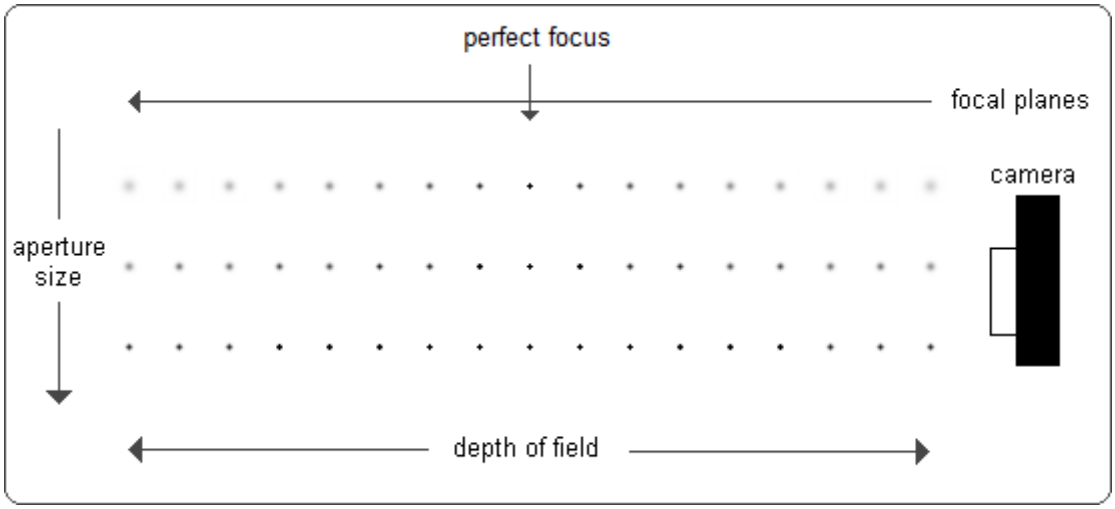


Figure 3.2.2.2 Illustration of the concept of depth-of-field. As the aperture size decreases, depth-of-field increases [41]. This means the point becomes less blurred due to defocus as it moves away from the focal plane (or the focal plane moves away from the point). Therefore, the point remains visually in-focus over a larger depth interval around the point of perfect focus.

The maximum depth resolution of a camera system is therefore set by a combination of focus setting grain, and depth-of-field. The focus setting grain sets a physical limit on resolution, whereas depth-of-field defines the level of discrimination between defocus blur at different depths, hence determining effective depth resolution.

3.2.3 **Identifying Focus**

Assuming that the input images conform to the required specification, *i.e.* they are aligned, cover a global range of focus settings for the scene, and are of appropriate depth resolution, the main concern of the DFF approach is the process of quantitatively evaluating focus (or conversely, defocus blur magnitude) so the image in input set can which displays a point in best focus be identified.

Importantly, it is not generally reasonable to assume that for any given scene point, *any* image from the set will display that point in *perfect* focus. This follows from the fact that the image set will have a limited depth resolution. However, in a pure DFF approach, it should be assumed from the global nature of the images that the focus of the point will be *maximised* by one distinct image from the set.

The specifics of some current DFF methods will be discussed in detail in *Chapter 4*. However, as a general overview, the following paragraphs briefly introduce some of the most mainstream approaches to evaluating focus in images.

- *Window-Based*

Window-based methods are similar in nature to the analysis of defocus blur seen in DFD, in that it involves analysing a spatial window in the image to evaluate the level of defocus blur. The important distinction however is that unlike DFD, the defocus blur is not directly modelled in terms of its functional effect on an image point. Rather, the effects of the blur over the entire region of the window are analysed. Typically, the blur is assumed to include every image point in the window.

An example approach to evaluating the blur effect quantitatively is to take the first or second differential of the window, and assume that the rate of variation of colour will decrease over the image as it becomes more blurred, due to the smoothing effect of blur.

The windowing approach is generally local with respect to the set of images, *i.e.* each image is evaluated individually without an explicit global relationship being drawn between images. It relies on the assumption that that the rate of colour variation in the scene is high when under perfect focus. A difficulty of the approach is choosing the window dimensions. If too small, different levels of blur are more difficult to distinguish, but as the window increases in size, the depth-map resolution decreases as all points within the window are assumed to be of equal depth.

- *Pixel-Based*

Pixel-based methods are unique to the DFF approach, and are based on the approach to the camera system Optics discussed previously in this chapter (*Section 3.2.1*). The idea of pixel-based analysis of defocus blur in a DFF approach is simple: If the image set is global, then any given pixel will show a certain scene point in its sharpest focus in *one* of the images. Therefore, the aim is to determine this in-focus image using the variation in colour of the pixel over the set of images (this is the only data available).

The basis on which we can infer the in-focus image from this colour variation data is from assumptions about the optics of the camera system, and to a lesser extent

assumptions about the scene. From *Section 3.2.1*, the model of the camera system used in this work can explain defocus from the perspective of individual image sensor points (image pixels), and this is the model on which per-pixel evaluation of defocus blur is based. Specific examples of this approach are too detailed to discuss here, but will be analysed in detail in *Chapter 4*. Indeed, the method proposed in this work follows a pixel-based approach to depth acquisition.

In the sense that the variation in pixel colour intensity is analysed over the entire image set, pixel-based methods can be considered to be global. They require very few assumptions about the defocus blur, typically fewer even than window-based DFF methods, and being per-pixel they maximise the resolution (in the image plane) of the depth-map, as the resolution matches the resolution of the input images.

3.3 **DFD Vs DFF**

To conclude this chapter, the relative advantages and disadvantages of the DFD and DFF approaches are evaluated in this short section. It is important to explicitly evaluate and highlight the differences and trade-offs involved in the two approaches, as these factors are central to discussions in the remaining chapters.

- *Depth-from-Defocus*

The overriding advantage of the DFD approach is the ability to use very few, or even single, input images. Because defocus blur is directly modelled, the behaviour and effects of the blur can be predicted, and so incomplete data is acceptable. The ability to work with fewer images means that DFD methods are generally less cumbersome, require less preparation of input images, and work better in application environments, particularly within general photography applications.

However, DFD approaches suffer from many practical difficulties. Firstly, the direct modelling of blur is difficult. Approximation models are inherently inaccurate, and empirical methods of sampling the defocus blur kernel are cumbersome to obtain. Moreover, identifying the correct scale of blur is in practice very difficult, particularly with conventional apertures, and the deconvolution of the image with a blur kernel is mathematically ill-posed, meaning that in practice strong scene priors are required to achieve accurate results. A further disadvantage of the DFD approach is the overheads and Image Processing difficulties of analysing the image using spatial windows around each image point being processed.

- *Depth-from-Focus*

The overriding advantage of the DFF approach (compared to DFD) is its simplicity. DFF sidesteps the practical difficulties of directly modelling defocus blur, taking the more empirical approach of analysing only the effects of the blur in the image data to estimate its magnitude. In addition, if the per-pixel analysis of images is employed (a per-pixel is by definition impossible with DFD) then Image Processing overheads and practical difficulties are reduced significantly, allowing for a greater depth-map resolution for the same processing time.

The clear disadvantage of the DFF approach is the fact that the image set must be global, *i.e.* each scene point in the depth interval of interest must come close to perfect focus in at least one image. This implies not only that single images cannot be used, but that the set of images must grow as the scene depth increases to achieve the

same depth resolution. The greater amount of input data required by DFF (compared to DFD) is unavoidable by the nature of the two approaches.

Chapter 4

Related Work

The previous three chapters have introduced the problem of depth acquisition from digital images, developed and discussed a general optical model of the camera system, and categorised the approaches to a solution most commonly found in the literature. In this chapter, this background information will be put into context by discussing in detail various previous work from the field on which the method proposed in this work is based, either directly or indirectly.

This chapter is loosely structured by the category which the work best fits into. First, there is a discussion of DFF techniques, and then there is a brief discussion of DFD techniques.

4.1 **DFF Techniques**

As will be discussed in the next chapter, the depth acquisition method proposed in this work is best categorised as a Depth-From-Focus (DFF) technique. Therefore, it shares significant similarities with, and draws inspiration from, many previous DFF techniques found in the literature.

Another important aspect of the proposed method is that it does not use image windowing but rather evaluates depth on a per-pixel basis using the variation in intensity of a single pixel across a variety of digital images captured with different camera settings. This multi-image per-pixel approach, whilst being less common than the windowing approach, has been an important feature of some previous work, most notably in [21, 22, 23]. Clearly, it is such work which is most directly relevant to the proposed method.

4.1.1 **Confocal Stereo**

A very recent method which is categorized as DFF and per-pixel is ‘Confocal Stereo’ developed in [21, 22, 23]. The method uses a hardware setup identical the one assumed in this work, *i.e.* a single unmodified digital camera at a fixed position, which is used to capture an array of images over a range of aperture and focus settings.

Assuming that the images are aligned such that a given scene point is projected to the same pixel in each image, each scene point is represented by a 2D array of intensities of that pixel (one from each aperture/focus pair) called an AFI (Aperture Focus Image). The approach is then to use the AFI data for a scene point to estimate the focus setting which shows the point in best focus, therefore this method can be categorised as DFF and per-pixel. The relationship between intensity (under known camera settings) and depth is derived from a general optical model of the camera system developed specifically for the method, which is very similar to the optical model employed in this work. This relationship is used to model the intensity variation pattern across the AFI when each focus setting is ‘correct’, therefore this

idealised model can be used as a search metric for the correct focus setting in real AFI data.

The modelling of intensity variation across the AFI is based on a property defined as 'confocal constancy'. The confocal constancy property is derived directly from the optical model of the camera system, which is based largely on the same geometric optical effects described in *Section 2.5*. In particular, this means that the following simplifications to the camera system are assumed:

- The camera lens does not absorb any energy from light rays which travel through it onto the image sensor.
- There is a perfect focus setting for every scene point.
- The aperture only blocks light rays from entering the camera lens, it does not have any effect on their direction (*i.e.* there is no diffraction).

As pointed out in [21, 22, 23] the above assumptions are well approximated by professional off-the-shelf lenses and DSLR camera systems. In addition to the simplifications assumed about the camera system, the method also relies on some strict assumptions (again, simplifications) about the scene:

- The scene must be complex in colour/texture (*i.e.* negligible regions of smooth colour gradient).
- The scene must have very low self-occlusion (*i.e.* all the light rays from a scene point reach the camera and are not blocked by other scene points closer to the camera).

As with the simplifications of the camera system, the above simplifications are acceptable, as it is clear that they can reasonably mostly hold in many natural scenes. An important point to make here is that the above scene assumptions must in general be made for a per-pixel approach to depth acquisition, as they follow from the limitations (defined by geometric optics) of the light intensity data available at a single point (pixel) in the image. Therefore, these scene assumptions must also be made for the method proposed in this work.

With the above assumptions in place, the confocal constancy property can be derived using simple geometric optics. The property can be introduced by looking at the special case of a scene point being perfectly in-focus. Taking $\mathrm{I}_{\alpha f}$ as the image at aperture setting α and focus setting f, and assuming a scene point p projects to the pixel (x,y) in every $\mathrm{I}_{\alpha f}$, then the light rays from p focused by the camera system to the pixel (x,y) are restricted to a cone, the solid angle of which is determined by α. With very low scene self-occlusion assumed, the intensity at the pixel (x,y) in $\mathrm{I}_{\alpha f}$ is proportional to the integral of the radiance of the light rays over this cone:

$$\mathrm{I}_{\alpha f}(x,y) = k \int_{\omega \in C_{xy}(\alpha,\mathrm{f})} L(p,\omega)\, d\omega$$

(***Eqn* 4.1.1.1**)

Where $C_{xy}(\alpha,f)$ is the light cone, ω is the solid angle, and $L(p,\omega)$ is the radiance of the light rays from p. The constant k is camera system specific and is determined by the response of the image sensor.

If we assume that the radiance of the light from p is constant over the light cone of the largest aperture (and therefore of all smaller apertures) then the integral is no longer necessary as $L(p) = L(p,\omega)$, and the intensity at $I_{\alpha f}(x,y)$ is proportional to the solid angle of the light cone $C_{xy}(\alpha,f)$.

$$I_{\alpha f}(x,y) = k \int_{\omega \in C_{xy}(\alpha,\mathrm{f})} d\omega\, L(p) = k \,||\, C_{xy}(\alpha,f) \,||\, L(p)$$

(*Eqn* 4.1.1.2)

Therefore, the radiance of the light cone defined by each aperture α, and the radiance of the light cone defined by a reference aperture α_1, should have a ratio which is constant and independent of the specific scene point in question or the scene in general:

$$\frac{I_{\alpha f}(x,y)}{I_{\alpha_1 f}(x,y)} = \frac{||\, C_{xy}(\alpha,f) \,||}{||\, C_{xy}(\alpha_1,f) \,||} = R_{xy}(\alpha,f)$$

(*Eqn* 4.1.1.3)

The constant $R_{xy}(\alpha,f)$ depends on the camera lens in use, and incorporates a variety of optical factors without the need to directly model them. In general, it varies across different α and f, as well as with the pixel position (x,y) in the image plane, hence the need to compute a separate constant $R_{xy}(\alpha,f)$ for the range of focus and aperture settings, and each pixel position.

It is important to state that, whilst the confocal constancy property is based on an optical model, it is argued in [21, 22, 23] that each $R_{xy}(\alpha,f)$ should be collected empirically in a calibration stage for a given camera, image resolution, and set of aperture and focus settings. Briefly, this is done by placing a planar surface in perfect focus for a given focus setting, and calculating the ratios of intensities at each aperture for each pixel in the knowledge that all points at that focus setting are in perfect focus.

With the constants of proportionality $R_{xy}(\alpha,f)$ known, the process of identifying the in-focus focus setting of a point from the AFI becomes fairly straightforward. By scaling each pixel in the AFI by the appropriate constant $R_{xy}(\alpha,f)$, the variation in intensity due to the aperture setting is counteracted for each focus setting in the AFI. Then, by the most direct application of the confocal constancy property, the in-focus focus setting should be the focus setting with the minimum intensity variance as the aperture setting changes. Another more advanced search metric applies the concept of confocal constancy more generally, by analysing not only the in-focus focus setting but also the other focus settings, to predict and search for constant-intensity 'regions' of the AFI rather than simply searching for the constant-intensity 'column' of the in-focus focus setting. The AFI model for each 'candidate' focus settings is fitted to the real AFI for a scene point, and the best fit is

deemed the correct focus setting and hence the depth for that scene point. The use of all the AFI data rather than just the intensities of the single candidate focus setting leads to a higher level of accuracy for this search metric, as reported in [21, 22, 23].

An important point to discuss is that the Confocal Stereo method relies on the assumption that a given scene point should project to the same pixel (x,y) in every image. In the above discussions of the basis of the method, this was assumed. However, in practice changing the camera settings often changes the geometric projection of light rays from the scene onto the image sensor. Moreover, changing the camera parameters can also vary radiometric factors which can affect pixel intensity locally and globally. Therefore, post-capture geometric and radiometric alignment of the set of images is essential to account for this. This is an important general point as it applies to all methods where one or more camera parameters are varied, and it is particularly emphasised as important for the Confocal Stereo method as the subtle variations in intensity which are analysed in the AFI rely on highly accurate and precise scene-point-to-pixel correspondence and the removal of other radiometric factors which vary intensity as camera settings change. See *Chapter 5* for an extended discussion of the alignment procedure in the context of the proposed method.

Briefly, the geometric alignment is performed using a model which incorporates both radial magnification caused by focus variation, and a stochastic shift factor parallel to the image plane caused by the mechanical movements of the camera as settings are changed (this is only relevant because of the very high image resolutions used in [21, 22, 23]. For lower image resolutions it is not an important factor). The radiometric alignment deals with global lighting change (between capturing successive images) by normalising all images to a reference image.

The Confocal Stereo method clearly takes a very similar approach to the method proposed in this work in terms of the hardware assumptions, scene assumptions, optical model and intensity-variation-based per-pixel approach to depth acquisition. The advantages of this approach are clear, and include a pixel-resolution depth-map, the ability to deal with very fine detail or a high depth differential, the avoidance of explicit blur modelling (one of the greatest practical difficulties discussed in the literature) and the simplicity of Image Processing due to the per-pixel approach.

Confocal Stereo does display general disadvantages of a per-pixel approach such as the need for accurate geometric and radiometric alignment, and strong assumptions about scene structure. In addition, whilst allowing for use of an off-the-shelf camera with no modifications, the method does require a cumbersome empirical calibration stage (on which the accuracy of the method depends entirely) for each camera system before it can be used. However, the most significant disadvantage of the Confocal Stereo method is the number of input images required. For a given depth resolution, *i.e.* number of focus settings, this method will have many more images than a DFF method which varies only focus setting (such as the method proposed in this work) because of the range of aperture settings required for each focus setting. The authors concede this limitation in [21, 22, 23], but argue that it is a trade-off of the increased accuracy of the method, which relies on use of the Confocal Constancy property, for which the data from both the change in focus setting and change in aperture setting is essential.

4.1.2 **Edge and Depth from Focus**

The method developed in [02] has many similarities of approach to the method proposed in this work. As in [21, 22, 23], a fixed-position conventional camera setup is used to capture images, and a simple geometric optical model is used to model the camera system. However, unlike [21, 22, 23], the method in [02] varies only the focus setting of the camera between input images. This, along with the general approach of the method, is what makes it most relevant to the method proposed in this work.
The method in [02] is interesting in terms of categorisation, as it can be regarded as a hybrid between both DFF/DFD and per-pixel/spatial windowing approaches. The method does not perform a classic search of the input images to locate the in-focus position of a point, but nor does it attempt to directly model or evaluate defocus blur to infer the depth of a point. Instead, it evaluates the intensity change of a point (pixel) as focus setting changes, using a one-dimensional row of neighbourhood pixels (not a 2D window) as part of this process. It is this aspect which places the method between a per-pixel and window-based technique.

In order to remain valid, the method assumes a very strict step-edge scene model, *i.e.* the scene is made up of overlapping planar surfaces which are parallel to the image place. An additional assumption is that at each edge (overlap of two surfaces), the intensity must be constant (and different) at each side of the edge. It is noted in [02] that these assumptions are a limitation of the method, as clearly they limit the practical application and validity of the method in natural scenes where these assumptions are unlikely to hold in general.

Again, in [02] a geometric optical model is used, and the spatio-focal image is defined using definitions based on geometric optics. The intensities on both sides of the edge are defined as *L1* and *L2*, respectively (*L1* > *L2*). Then, based on a blur kernel with circular symmetry, and geometric optics, the change in intensity at an edge point *p* can be reasoned about.

The basic concept is that where a point is near to an edge, as the blur kernel shrinks to the in-focus position and then grows again, the intensity at the point should reach a peak at *L1* (if on the *L1* side) or a trough at *L2* (if on the *L2* side). This happens because as the blur kernel radius increases, the intensity of the (blurred) point incorporates more of the other edge's intensity. However when the blur kernel shrinks to zero (*i.e.* the point is in focus) this will lead to a peak or trough at this intensity. This is similar in concept to the basis of the method proposed method in this work (see *Chapter 5*). It follows logically that edge points should be between 'peak' and 'trough' points in the image.

Once an edge points has been detected, its depth is calculated by looking at the pixels in a one-dimensional slice (perpendicular to the edge) of the spatial neighbourhood of the point, and how the intensities of this one-dimensional spatial neighbourhood change as focus setting changes. The slice of the neighbourhood is perpendicular to the edge so that, under a geometric optical model, the variation due to intensity on either side of the edge is maximised. A synthetic image termed a 'spatio-focal image' is built using the intensities of the spatial slice in one dimension, and the focal length in the other, to give an intensity distribution around the edge point as focus setting is changed. A model derived from geometric optics is then used to identify in this spatio-focal image the focal length at which the edge point is in focus. This model, which is based on the strict step-edge scene model, gives an ideal pattern of intensity distribution for an edge point based on the idea that the intensities *L1* and *L2* gradually merge across the spatial dimension as defocus increases, or

gradually separate across the spatial dimension as focus increases, to give a predictable pattern in which the in-focus focal length is a midpoint.

The main advantages of the method presented in [02] are in its general approach to the depth acquisition problem. This is an example of an approach which, using only a (discrete) variation in focus setting, can evaluate depth on a continuous scale by analysing only the change in intensity of (and around) a scene point as focus setting changes. This approach avoids the limitations and difficulties of using a model of focus as a search metric (classic DFF) and/or an explicit defocus blur model (classic DFD), instead analysing the effects of defocus blur in terms of intensity variation using simple and well-understood geometric optics (in addition to the strict scene model) to find the in-focus focal length. Importantly, this approach represents a third option between directly modelling focus (assumptions about the scene) in classic DFF techniques, and directly modelling blur (assumptions about the camera system) in classic DFD techniques. This is of great relevance here as it is very similar to the approach of the method proposed in this work.

Despite the importance and advantages of the general approach taken in [02], the method does have some major limitations which come as a result of their implementation of the approach. The most prominent of these limitations is the very strict, and practically unrealistic, assumptions about the scene, on which the modelling of the intensity variation around a point (in the spatio-focal image) is based. It is noted in [02] that although the method is stable where these assumptions hold, it would degrade in performance rapidly if the assumptions were not held. This represents a major limitation in the practical application and generality of the method. Another disadvantage of the method is that, although a new approach is taken, there is still a reliance on evaluating neighbourhood pixels around the point of interest. Though in [02] this only involves a one-dimensional slice rather than a full two-dimension window as in classic DFF, this still introduces difficulties such as selecting the correct neighbourhood size and general image processing issues seen in window-based techniques. A further disadvantage of the method is that, by definition, it can only be applied to edge points. Since a step-edge scene model is assumed, it follows that the method can produce a full depth-map using only the depth of edge points. However, in a 3D scene with complex geometry and textures the method would not work, *i.e.* it is not general enough to acquire the depth at arbitrary 3D points and cannot produce a full depth-map of a complex, highly textured 3D scene.

The method proposed in this work attempts to implement the same general approach taken in [02], and all the advantages of this approach, in a way which avoids the practical disadvantages of the method in [02]. For example, by evaluating depth on a per-pixel basis (no analysis of the neighbourhood of the point) and by allowing a far more flexible scene model which allows for much more generality (for example, the ability to estimate the depth of any scene point, and allowing for complex surface textures and more complex scene geometry). See *Chapter 5* for a full discussion of the method proposed in this work.

4.1.3 **Classic DFF Techniques**

The classic DFF approach as defined in the literature is a pure search problem, where a numerical measure of the level of focus of a scene point is used to identify, from a set of input images, the changing focus of the scene point.

Most commonly, the input images vary only in focus setting [23], as this provides a straightforward link between level of focus and depth, and the focus measure is estimated by evaluating the level of (defocus) blur in the local spatial neighbourhood of the scene point, which is described using an appropriately sized image window centred on the scene point. Examples of work which take this classical DFF approach can be seen in [27, 32, 51, 47, 42, 08, 30].

An essential point in classical DFF is that the focus measure used in the search should ideally be unimodal, monotonic about the mode, and maximum only when an image is perfectly focused [27]. An example of a measure which, according to optical theory, should satisfy these requirements is one based on gradient magnitude (*i.e.* an edge detector) around a scene point, for example see [27, 51, 47].

From these idealised requirements, it appears that a brute-force maximisation approach should be sufficient to solve the DFF search problem. However, due to noise introduced between input images by both the camera system and the scene, assuming that this is not perfectly corrected in the image alignment stage, such measures typically produce a focus setting/ focus measure profile with multiple local maxima and/or an incorrect global maximum.

For this reason, and because focus measure evaluation can be expensive, more sophisticated search techniques such as curve fitting (for example using a Gaussian-type function) are used in [27, 32, 51] to locate the correct focus measure from the data. An additional advantage of curve fitting is that, in principle, the true focal plane depth can be located between the sampled depths (input image focus settings), rather than simply selecting the sampled depth which shows the point in 'best' focus. This idea of fitting data to a functional relationship between depth and focus (derived from optical theory) is also adopted in the method proposed in this work (see *Chapter 5*).

The limitations of the classical DFF approach, particularly when compared to the proposed method, are discussed in detail at various points in this work; however a summary will be given here. The feature of the approach from which the greatest limitations arise is the necessity to analyse the spatial neighbourhood of a scene point in order to estimate its depth. Aside from the practical Image Processing difficulties, which in practice limit the depth-map resolution in the image plane [28], the reliance on the spatial neighbourhood of a scene point leads to some fundamental limitations.

Firstly, the neighbourhood must have non-uniform intensity for any focus measure to be discriminative [52]. Importantly, this excludes in practice not only areas of flat intensity but areas of constant intensity variation (assuming the focus measure is symmetrical), *i.e.* any region with a linear intensity gradient [23], both of which are abundant in natural scenes. Furthermore, in practice these restrictions are a theoretical minimum, with many classic DFF focus measures explicitly measuring high-frequency intensity content in a neighbourhood [01]. Examples of such focus measures are seen in [27, 51, 42, 31].

Secondly, the neighbourhood must be (generally) planar, and (generally) parallel to the image plane. If different regions of the neighbourhood are at different depth levels (for example at a step-edge or a complex surface with rapidly varying depth) then the optical assumptions on which the focus measure is based break down, and the focus measure may evaluate in an unpredictable manner. A related issue is that even if the window covering the local neighbourhood does not directly display multiple depth levels, other points near to the window in the image plane which are heavily blurred may introduce unrelated high-frequency information into the window, due to the spatial spread of the defocus blur region, which can cause the focus measure to evaluate unpredictably [52].

For these reasons, and because of general Image Processing issues such as computational efficiency increasing with increasing window size, the selection of an appropriate window size is an unavoidable problem of classical DFF. A smaller size increases depth-map resolution in the image plane and better avoids multiple levels of depth and contamination from nearby blurred points in the window, but a larger size is better able to distinguish between different levels of defocus blur. This trade-off is a problem with all window-based depth acquisition methods (including DFD methods, discussed in Section 4.2) which has no obvious general solution [28].

4.2 **DFD Techniques**

The Depth from Defocus (DFD) approach to depth acquisition is fundamentally different from DFF. In DFF, as discussed in the previous section, the idea is to search over a set of images captured with known camera parameters, to locate the parameters (and hence the depth) which yield the maximum focus of each scene point, as defined by some focus measure.

Conversely, DFD evaluates depth by directly evaluating the defocus of scene points, and linking the level of defocus (via an optical model of the camera system) to depth. It should be emphasised here that the method proposed in this work is not what is traditionally defined as a DFD method, as it does not directly model or evaluate defocus blur in order to infer depth. However, the reliance on the optical model of the camera system and the implicit modelling of defocus blur are key to the proposed method, and hence the method draws from various previous work on DFD techniques which explores these factors. Therefore in this section some relevant DFD techniques will be discussed.

4.2.1 **Image and Depth from a Conventional Camera with a Coded Aperture**

A recent example of a DFD technique which explicitly evaluates the form and scale of the blur PSF in order to infer depth is the technique proposed in [28]. In this work, defocus blur is modelled as a convolution of the focused image with a blur PSF, and the approach is to control the form of this PSF so that its scale can be linked to depth on an empirically derived scale. The PSF form is controlled by deliberately modifying the camera aperture with a pattern. The reason for doing this, and the key feature of the technique, is to maximise the level of discrimination between the blurring effect at different scales of the PSF, therefore maximising the accuracy and resolution of depth acquisition.

The basic principle of the technique in [28] is identical to the traditional DFD approach discussed in *Section 3.1, i.e.* linking the scale of the blur PSF to the depth of a point, where the form of the PSF is assumed known. However, the key observation here is that though the PSF produced by a conventional aperture can provide depth cues, it is very difficult to reliably differentiate between the blurring effect at similar scales of the conventional aperture PSF. Clearly, this limits the accuracy of linking PSF scale to depth, leading to low depth resolution and accuracy of depth-mapping. The solution in [28] is to place an occlusion pattern (2D piece of card) over a conventional aperture to modify or ‘code’ the blur PSF so that the form conforms to two key criteria:

1. Reliable discrimination between the blurring effects caused by different scales of the PSF is possible.
2. The PSF is easily invertible, so that the focused image can be recovered by deconvolution of the PSF with a blurred image.

The search for a coded aperture which conforms best to these criteria is guided by examining the Fourier transform of the PSF formed by the coded aperture. Taking a PSF *f*, the Fourier transform of *f* will have a number of zero frequencies. Looking at a particular one of these frequencies, ω, if an image *x* has been blurred (convolved) by *f*, then the Fourier transform of *x*, *X*, will also have a value of zero at frequency ω.

The basic principle of the search for a coded aperture pattern is that (the Fourier transforms of) different scales of the pattern will have different sets of zero frequencies. The goal is to find a pattern where these zero frequencies are distinct and distinguishable between different scales, particularly between similar scales. In doing this, the goal is to find a blur PSF which maximises the 'difference' of the effect on the blurred image as scale is varied, resulting in an easier identification of the correct scale by deconvolution with the blurred image.

In [28], the search for a pattern is implemented as a random search over a space of 'practical' patterns (symmetrical, cut from a single piece of card, no floating regions). For each pattern, the heuristic used is to take the minimum KL-divergence between the frequency distributions of the Fourier transforms of any two scales of the pattern. The goal of the search is to maximise this value.

Using the KL-divergence as a measure of distance between two scales of a pattern is key not only because it promotes differentiable zero-frequency values between any two scales, but because it promotes distributions where there is a differentiable amount of content at any given frequency. This is very important in practice, as noise introduced by the imaging system and other environmental variables will mean that no frequency is exactly zeroed. Therefore, it is important to be able to distinguish between 'noise' content and 'real' content at the zero frequencies of a particular scale, which is much easier if all other scales have a significant amount of 'real' content (*i.e.* above the noise level) at the zero frequency.

Having identified a coded aperture pattern, the depth acquisition process in [28] is fairly straightforward and conventional for a DFD technique. Importantly, it is possible to apply the process using only a single input image. This is an inherent theoretical advantage of the DFD approach and it is indeed employed in practice in [28].

First, a discrete set of PSF scales, along with their associated depths, is determined empirically. This is done by capturing the blur of a single point of white light against a black background over a set of known depths and using the actual blur images as the PSF for each of the set of depths. This works since the blurred image is of a point of light at intensity 1 isolated against a background of intensity 0, so if the PSF is a convolution of the original focused point with the blurred image, then the blurred image is identical to the PSF. Indeed, such an empirical approach can be preferable to a functional modelling of the PSF form as it takes into account non-geometric optical effects and any camera-system optical effects which are difficult to predict using an optical model.

The method then proceeds by dividing the image into small windows, assuming that the depth is constant over each window. Though it would be possible, within the boundary limits of the largest scale of the PSF, to evaluate each pixel-point separately, this is deemed in [28] to be too computationally expensive in practice.

This loss of depth-map resolution (in the plane orthogonal to the depth axis) is a key disadvantage of the window-based approach.

Each window is then deblurred by all scales of the PSF, and the correct scale is taken as the one producing the most plausible deblurred image, according to a reconstruction error based on a prior model of a focused natural image. The prior employed in [28] is a sparse derivative model, *i.e.* a natural image should be largely smooth with occasional sharp changes in intensity (for example at object edges).

The main strength of the technique is, clearly, that it is able to use only a single image as input. This means that the practical issues introduced by capturing and aligning multiple input images, which are necessary in all the DFF techniques discussed in the previous section as well as the technique proposed in this work, are conveniently avoided. However, it must be emphasised that the key strength of this work is the factor which allows a single image to be used in practice: the use of a coded aperture to reliably distinguish between similar blur scales, which means that reliance on additional information provided by the redundancy of multiple images is lowered to the point where using a single input image is feasible in practice.

Of course, the method does have limitations, which are significant even in the context of a single image input. As mentioned above, the method is expensive, meaning in practice only a course grain of resolution can be achieved in the plane orthogonal to depth as points must be grouped and evaluated together in small windows. This means that the method is inappropriate for scenes with highly complex 3D surfaces, which are not smooth. Moreover, the empirical sampling of a set of different scales of the PSF limits the depth resolution of the method. Ignoring the fact that depth resolution is limited in the first place by the arbitrarily chosen set of scales and associated depths, the spatial resolution of the samples is limited by the pixel resolution of both the samples themselves and the blurred input image (which must be identical). As noted in [28], this is a particular issue when attempting to differentiate between similar, very small, PSF scales. These issues result in a depth-map which is, in practice using a single image, quite low in resolution. Indeed, as discussed in [28] the raw depth-map produced by the technique may require user guidance in a post-processing stage to reach a useful level of accuracy.

4.2.2 **Relative Defocus**

Examples of DFD techniques which take a direct approach to modelling defocus blur are common in the literature, for example [28, 14, 19, 24, 49, 43]. However there is a fundamentally different DFD approach, for example in [36, 37], which looks at relative differences in blur to infer depth. In [36, 37] the method requires a minimum of two input images whose different defocus blurs are compared relatively (under a model of the blur PSF) as opposed to the approach of evaluating blur, and hence depth, directly from a single image.

The method in [36, 37], as is typical of DFD techniques, uses the basic geometric optical model of the camera system described in *Section 2.5*, with the minor difference that the 3D points focused to the image plane are assumed to originate from a curved (spherical) 3D ‘surface of focus’ in the scene, as opposed to a flat plane of focus. The method is based on capturing an all-focus image of a scene (*i.e.* using a pinhole aperture so every scene point is in-focus), and capturing a second image of the same scene using a larger aperture, such that depth of field is smaller, therefore

defocus blur is introduced with the familiar relationship between depth in the scene and blur magnitude.

The input therefore provides, for each scene point, a relative difference between the point perfectly in focus and the point blurred due to defocus, which depends exclusively on the depth of the point in the scene since the optical parameters of the camera system are known. The difference is described as relative because it is only meaningful in the context of other points in the scene, for example if a point *a* is more blurred than a point *b* in the second image, it indicates that *a* is at a greater distance from the surface of focus than *b*. This is the basis on which a depth-map of the scene can be constructed.

An important note about this technique is that the use of multiple images does not mean that alignment of the images must be performed prior to processing. This is an inherent advantage where, as is possible with some DFD techniques with strong scene priors, the method can be applied to a single image [23, 28]. However, this advantage is maintained in [36, 37] despite the multi-image input due to the convenient fact that varying aperture size does not alter the projection of the scene onto the image sensor at all, at least under a purely geometric optical model. Therefore, images captured at different aperture sizes, assuming a static camera and scene, generally require little or no geometric alignment to match up corresponding scene points. This advantage of using aperture rather than focus setting to control blur is commonly exploited in depth acquisition techniques, for example [44, 43]. Another method commonly employed to avoid the requirement for geometric alignment, even when varying the focus setting, is compensating for changing perspective projection using the camera zoom setting [50, 08]. In fact, as discussed in *Chapter 5*, modern camera lenses can by default provide a uniform image projection as focus setting varies, as is the case with the camera used to capture test images in this work.

Finally, because of the very low number of input images required in practice (a theoretical minimum of two), a hardware implementation which captures two (or more) images simultaneously is feasible in practice, which removes the requirement for the static scene assumption. Indeed, such a hardware implementation is presented in [37].

4.2.3 Alternative Approaches to DFD

- *Variational Bayesian Based Techniques*

So far in the discussion of DFD, the importance of the modelling of defocus blur to the approach has been referred to frequently. In particular, two distinct methods of modelling have been explored: the approach of modelling the PSF using a function with a set of parameters, and an approach of modelling by empirical sampling of the real blur PSF. The issues surrounding the former approach of functional modelling of defocus blur are explored in *Section 3.1*, with the important assumption that a functional model must be manually selected based on some compromise of real world optics basis and practical mathematical convenience.

However, there is an area of work in Machine Learning, Variational Bayesian-based methods of blind deconvolution, which provide another option for deriving a model for defocus blur. Such methods provide the advantages of both the approaches for deriving a blur PSF discussed above, in that PSF is derived from real data as in empirical sampling, but like functional modelling the PSF resolution is not limited to

a sampling resolution and the process is purely mathematical (*i.e.* there is no practically cumbersome, error-prone sampling stage).

Relevant examples of work in this area can be seen in [03, 04, 06, 46, 16]. In [03], for example, variational inference is used to estimate the blur PSF and the focused image directly from a single blurred image, using no assumptions about either. All that is required is that prior probability distributions of the blur and focused image (and noise) are supplied, along with prior distributions for the parameters (denoted hyperparameters) of these prior distributions. Then, variational inference is performed to simultaneously estimate PSF and focused image. Because variational inference is used, the posterior distributions (*i.e.* the probability distributions of the PSF and focused image which are converged to) are purely probabilistic so their uncertainty is measurable, and are derived only from the real data of the single blurred image (although the priors have some influence). It should be noted that the method in [03] is given generally, with no direct mention of a depth acquisition application. However, applying the variational Bayesian approach in DFD techniques is found commonly in the literature [48], such techniques include [39, 40, 26].

- *Active Illumination Techniques*

In the discussion of DFD so far, there has been an assumption of passive image capture using a conventional camera. This passive capture of visible light information imposes certain theoretical restrictions on depth acquisition using either DFF or DFD. As previously discussed, in the case of DFF the restriction is on the discrimination of a focus measure where there is low-frequency scene content, and in the case of DFD the restriction is on the ill-posed nature of the blur deconvolution problem which cannot be addressed without strong priors.

However, by projecting patterns of visible light directly onto the scene, and capturing images with a conventional camera as before, these difficulties can be reduced by inserting known depth cues into the images. This is known as active illumination, and though strictly an active depth acquisition technique, it is fundamentally very similar in terms of imaging hardware to the passive approach taken in this work, and so is relevant to the discussion here.

The trivial relative depth cues that scene illumination can provide in images are obvious, and are common to the Image Processing literature, for example in [45] where crude illumination from a camera flash is used to indicate objects in the 'foreground' of a scene. Far more precise depth cues can be provided by projecting specially designed light patterns onto a scene to control the frequency attributes of the neighbourhood of a point [13, 29, 33, 17, 18].

It is straightforward to see how modifying the frequency of the scene content can assist in DFF techniques by increasing the frequency of content generally, and by tailoring the focus measure specifically to the known dominant frequency characteristics where the projected pattern is in-focus (*i.e.* where the point is in focus) [33]. Where DFD is concerned, it can be shown that by controlling the illumination of a surface, it is theoretically possible to fully reconstruct a blurred image of that surface by deconvolution [23, 13]. This is similar to an idea seen in [28], where the frequency profile of a scene is estimated using a prior to reduce the difficulty of the blind deconvolution problem in DFD, except here the prior is not simply estimated but known.

The disadvantages of active illumination, however, are due to the very fact that light must be projected onto the scene. For example, this may be inappropriate or cumbersome in a practical setting, and in addition there is an issue of separating the

light pattern from the ‘true’ images of the scene in any application where the intent is to perform some Image Processing on the image(s) using the resolved depth information, for example post-exposure refocusing. In certain cases the projected light pattern may be too complex for separation to be possible [38].

- *Modifying camera optics to control defocus*

There are many examples in the literature of modifying the conventional camera optics in order to control the incident light rays in ways which result in a known modification of the defocus produced by the lens. This approach is similar in nature to the previously described active illumination approach, except that instead of modifying and controlling the light which is emitted from the scene, the light rays which have already been emitted from the scene are modified and controlled before they are captured by the camera sensor.

An example of this type of approach already discussed in detail in this chapter is [28], which uses a coded aperture to control the PSF of the defocus blur in a way which can be predicted. Similar work using a coded aperture to control defocus includes [15].

Another common method used to control defocus in order to improve depth discrimination in DFD techniques is the use of optical masks, or filters, which are placed in front of the camera lens to control the frequency characteristics of the light which passes through to the camera lens, providing control over properties of the defocus which can be exploited to infer depth [20, 12, 25]. Further examples of modification of light rays as they enter the camera system can be seen in [07, 10], in a technique known as Wavefront Coding, where light waves are coded in, for example, the aperture stop to produce a defocus blur PSF with known characteristics.

Chapter 5

Proposed Method

In this chapter, the method for depth acquisition proposed in this work is discussed. The method covers a process from the capture of raw images of a scene, to the pre-processing of the input images, to the computation of a raw depth-map of the scene from the input images. Broadly, the process can be thought of in terms of a *data collection* stage (image capture and alignment) and a *depth-mapping* stage (computing a depth-map of the scene from the input images). Though the latter stage is of primary interest in this work, the former stage is obviously an essential part of the process and must be clearly defined in order to understand and control the input.

5.1 **Overview**

The proposed method comes under the category of Depth-From-Focus (DFF). DFF is discussed in chapter 1, but briefly, this means that the depth of a point is acquired by locating the depth where the point is in best focus using a sample of images captured at focus settings across a global depth interval, *i.e.* a depth interval which contains all scene points.

The proposed method produces a per-pixel depth-map of a scene. This is an important feature of the proposed method, considering that DFF approaches typically use analysis of a spatial window around a pixel of interest to evaluate focus. The advantages of the per-pixel approach will be discussed in a section later in this chapter.

The idea on which the proposed method is based is predicted by the optical model of the camera system developed in *Chapter 2*, and specifically, the interpretation of the model discussed in *Section 3.2*. The specifics of this will be discussed in detail in the sections to follow, but to give some context to the remainder of the chapter it will be useful to briefly introduce the idea here.

The model predicts that when a scene point p lies on the focal plane, the lens brings p to exact focus at a pixel x on the image sensor, resulting in p being perfectly in-focus at pixel x in the image. As the focal plane moves away from p, either towards or away from the camera, light from an increasingly large region of the scene with p at its centre is focused to pixel x. Crucially, this can be interpreted as the centre of p still being projected to pixel x, even though all the light from p is not focused to pixel x, as it is when p is in-focus.

Assuming the above optical predictions, the basic idea of the proposed method is straightforward. By capturing N images at a sample of different focal plane depths over a global depth interval, ensuring that a given scene point p is projected to the same pixel x on the image sensor in every image, we have N intensities of p, taken from pixel x in each of the N input images. Assuming we know the depth of the focal plane for each image, on either a real or relative scale, we can link each intensity value to a known depth of focal plane. The goal is then to locate, from this sampling of intensity variation over depth, the intensity/depth which displays the point p in-focus. As shall be discussed, this can be done in a discrete manner (*i.e.* locate the

sample intensity/depth showing *p* in best focus) or a continuous manner (*i.e.* locate the intensity/depth of *p* in perfect focus along the continuous scales of intensity/depth).

5.2 The Proposed Method

In this section the process of the proposed method is introduced and discussed in detail. The theory behind the method is given, and to give the discussion some practical focus, the method is explained using a real test dataset, which is the set of images captured of the test scene used by the method to produce results in *Chapter 6*.

5.2.1 Data Collection Stage

5.2.1.1 Image Capture

In line with much of the related work in the literature, this technique assumes the use of a single digital camera at a fixed position. It may seem that these are arbitrary limitations, as using a multi-camera, multi-angle setup is likely to improve results by increasing the volume and redundancy of data. However it is important to emphasise that a general goal in any depth-acquisition method is to minimise the number of input images with respect to accuracy of results. The method can be extended to a multi-camera, multi-angle hardware implementation in an application by simple repetition.

The specifics of the hardware setup are as follows.

- *Camera Model*

The digital camera used for capture is assumed to be a conventional digital camera, *i.e.* it offers manual control over focus setting, aperture size and exposure time. It is beneficial for the camera to have a relatively large image resolution, as though in principle the method will apply to any resolution of image, in practice a large image resolution and therefore depth-map resolution can be advantageous.

To capture the test dataset, the camera model used was the Olympus Camedia E20-p, which fulfils the manual control requirements and has a 5 megapixel image resolution.

- *Initial Camera Settings*

In the proposed method, the focus setting is the only variable camera parameter. However, the initial settings chosen for aperture and exposure time must be considered in order to optimise the input data for the depth-mapping stage.

In order to minimise depth-of-field, as discussed in *Section 3.2.2*, the maximum available aperture size should be used. By minimising depth-of-field, we allow for the maximum depth resolution available with the camera in use. The aperture size used to capture the test dataset is *f*/2.0.

An equally important setting for capturing optimal data is the exposure time. In order to maximise the variance in intensity of a scene point over different focus settings, which is important to minimise the obscuring effects of noise on the input data, the exposure time must be correct. If it is too low then image noise will be very

obscuring since the variance of intensity will be low. If it is too high the pixels may be too saturated; another form of noise which will obscure the image data.

- *Scene*

The test scene used in this work is a synthetic scene, designed specifically for the purpose of a proof-of-concept of the method, and qualitatively and quantitatively comparing the accuracy of results. There is an intrinsic issue in evaluating the accuracy of depth-maps of any test scene, which is that a ground-truth image (*i.e.* a 'true' depth-map) of the scene is required. For complex scenes this is often impractical to obtain, as specialised hardware such as 3D laser scanners may be required. Therefore, the test scene in this work, the images of which make up the test dataset, is constructed in such a way as to avoid this issue whilst still providing a multi-depth scene of adequate intensity complexity to obtain meaningful results.

The scene consists of two books (with flat surfaces) covered with two real images. The faces of the books are parallel to the front of the camera, and the first is placed at 0.400m from the camera and the second at 0.600m from the camera. A millimetre scale was used to manually measure the distances, and so the estimated error on the distances is ±0.001mm. The books overlap, therefore the overall setup of the test scene is two complex-textured (but flat-surfaced) images on two different planes parallel to the front of the camera.

Figure 5.2.1.1.1 displays the arrangement of the two books and the position of the camera used to capture the test dataset from a top down perspective.

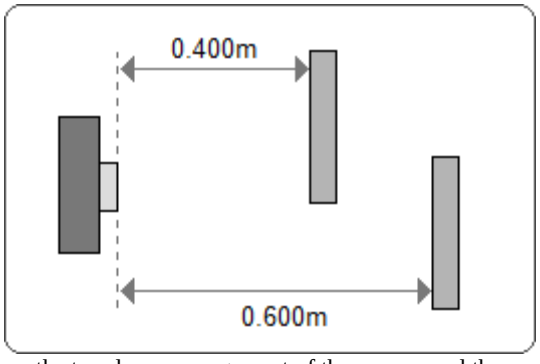


Figure 5.2.1.1.1 Shows the top-down arrangement of the camera and the scene in the test dataset.

- *Focus Settings*

The focus setting on the Olympus Camedia E20p is controlled by an electric motor, and can be changed only in discrete increments. The focus range of the camera is known (20cm to infinity) and a scale of 'focus setting notches' to distance of the focal plane has been calculated empirically to form a correspondence between focus setting and focal plane depth. With certain camera lenses, a highly accurate scale may be provided by the manufacturer, so this empirical estimation stage is unnecessary. The empirically estimated scale for the Olympus Camedia E20p is given in *Table 5.2.1.1.2.* In *Figure 5.2.1.1.3,* the plot of focus setting against depth is shown, illustrating how the increase in depth of the focal plane accelerates as focus setting increases. This relationship will be recognisable to photographers as the focus ranging from some initial depth to 'infinity' using a finite focus setting scale.

Image Index	Focus Setting (notches)	Depth (mm)
1	30	289
2	32	297
3	34	305
4	36	312
5	38	320
6	40	331
7	42	342
8	44	353
9	46	364
10	48	380
11	50	400
12	52	415
13	54	433
14	56	445
15	58	458
16	60	470
17	62	490
18	64	510
19	66	533
20	68	558
21	70	580
22	72	600
23	74	630
24	76	665
25	78	720
26	80	770
27	82	820
28	84	870
29	86	920
30	88	970
31	90	1020
32	92	1070

Table 5.2.1.1.2 Empirically estimated scale for the test dataset relating image index, corresponding focus setting value (in discrete notches) and depth from camera of focal plane.

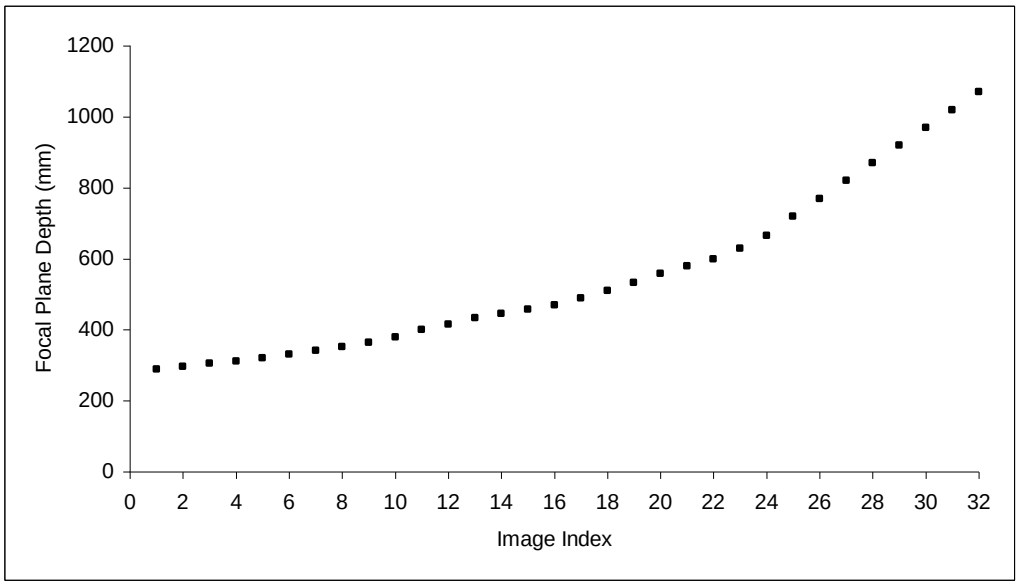


Figure 5.2.1.1.3 Plot showing the relationship given between image index and depth for the test dataset, as given by the scale in *Table 1*.

For the test dataset, it was decided to take 32 samples of focal plane depth; *i.e.* 32 different focus settings were used. The focal plane begins in front of the first book, and is moved forward by advancing the focus setting by 2 notches at a time. The final position of the focal plane is behind the second book, so that the test dataset is global. To re-emphasise the point, the focus setting is known for each image, and so by extension the depth of the focal plane for each image is known, by using the empirically estimated scale given in *Table 5.2.1.1.2*.

- *All images from the test dataset*

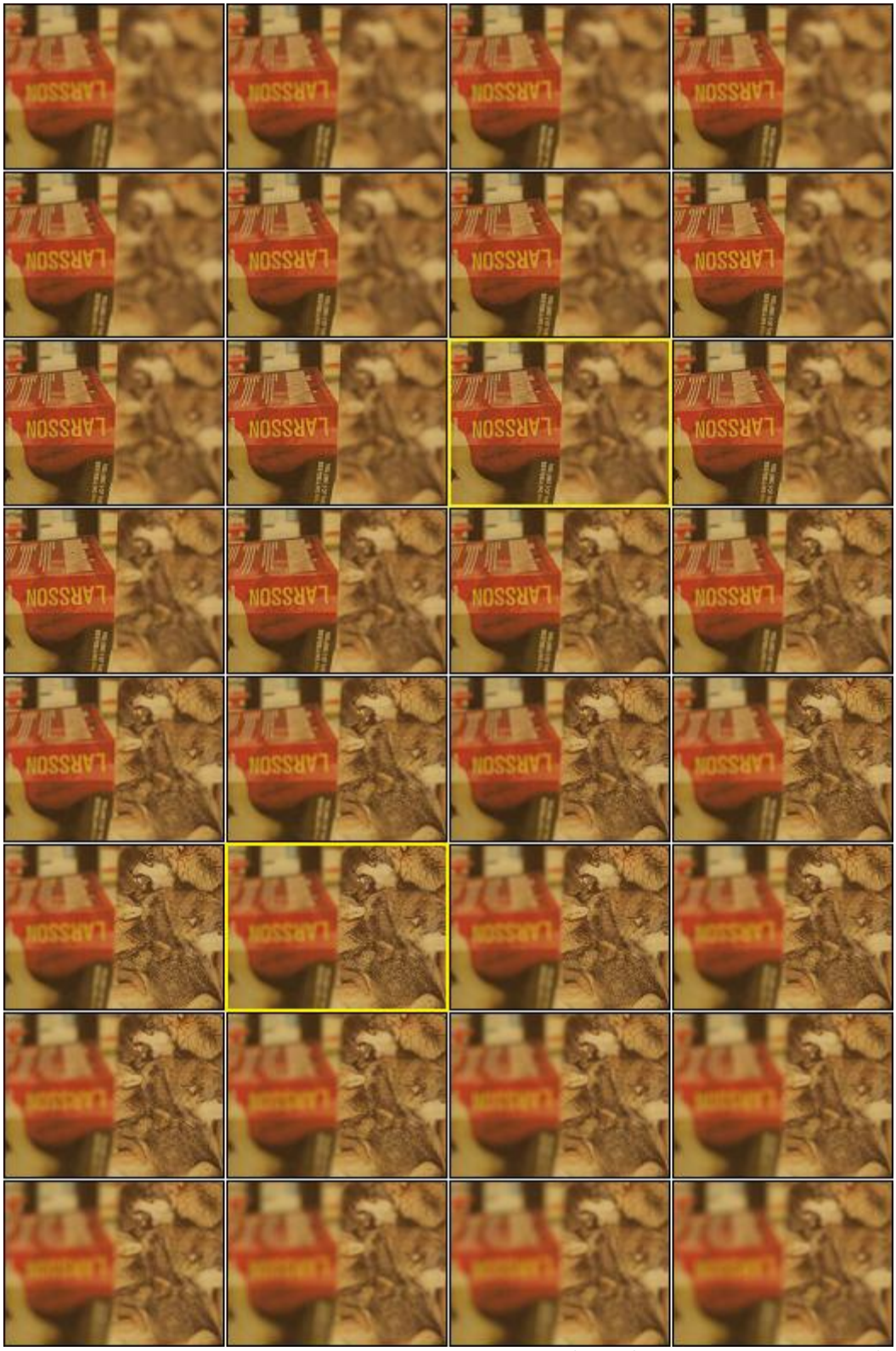


Figure 5.2.1.1.4 The 32 images in the test dataset, sequence is in rows from left to right, top to bottom. The first plane comes into best focus in image 11 (highlighted in yellow) and the second plane comes into best focus in image 22 (also highlighted in yellow).

5.2.1.2 **Image Alignment**

- *Geometric Alignment*

As previously mentioned, an essential feature of the input image set is that a given scene point must be projected to the same pixel in each image, *i.e.* that the images are geometrically aligned. In real camera systems, the projection of a scene point onto the image sensor can change in a non-linear fashion as camera parameters are changed, so the process of geometric alignment becomes a process of correcting these distortions.

The main geometric distortion comes from magnification caused by changes in either zoom or focus settings. As we assume zoom is fixed, only the magnification caused by change in focus is relevant here. The magnification caused by change in focus is radial, and is predicted by the simple geometric optics of the thin lens model. This radial magnification can be modelled directly, but in practice it may be more accurate to link the focus setting of a particular camera system to the level of radial magnification empirically, in a pre-calibration stage.

For the test dataset, using the Olympus Camedia E20p camera, the issue of radial magnification as focus setting is changed is conveniently sidestepped. The lens in this camera system is optically configured so that the magnification effect of focus change is countered and the projection of the scene remains constant as focus setting is changed. However, in other camera systems this may not be the case, and so for generality it is important to mention that an explicit magnification correction may be required.

Another geometric distortion to consider is shift parallel to the image plane. This shift can be caused by movement of the camera unit or movement in the scene. Furthermore, in very high-resolution images, a stochastic shift may be introduced by the mechanical movements of the camera lens as settings are changed [21, 22, 23].

Clearly, the effects of this shift are far easier to factor out than those of radial magnification. Using the test dataset, it is sufficient to ignore such a shift by assuming that both the camera and scene were static over the time interval where the input images are captured (a reasonable assumption in many practical photography situations), and the resolution of the images at 5 megapixels is low enough to avoid the tiny stochastic shift due to mechanical movements in the camera.

- *Radiometric Alignment*

Assuming the set of images are geometrically aligned, a scene point is represented by the variation of the intensity value of a certain pixel across the set of input images. As previously mentioned, this variation of intensity is the basis on which the in-focus intensity and therefore the in-focus image will be identified by the proposed method. However, this variation of intensity may become distorted by additional factors which 'weight' each successive intensity value between images. The process of radiometric alignment is the process of correcting for these additional factors so that the relationship between focal plane and intensity can be examined directly.

As with geometric distortions, changing camera parameters can cause radiometric distortions in a predictable way. The main parameters that cause radiometric distortion as they are changed are the aperture and exposure time, however in the proposed method these parameters are fixed, which is indeed an advantage. The one variable parameter, focus setting, can be said to have a radiometric effect because of the inverse-squared attenuation of light rays with the

distance travelled to the image sensor. However in general this effect is so subtle that it can be ignored in practice, and in the test dataset the distances involved are certainly low enough to disregard this effect.

A radiometric distortion which is far more applicable to the proposed method is global lighting shift. In a typical scene in practice, it is highly likely that ambient light levels are not constant over the time interval where the set of input images is captured. Indeed, even for the test dataset which was captured under controlled lighting conditions, this assumption cannot be made because of the unpredictable variation in intensity of electric lights.

Fortunately, normalising the global lighting level over a set of images is relatively straightforward. Under the assumption that the lighting is truly ambient and global, it can be modelled as a multiplicative factor for each input image. Then, all the input images can be normalised such that they have the same global lighting factor as some reference image, so that the effect of global lighting shift between the input images is corrected for, and the images are radiometrically aligned.

5.2.2 **Depth-Mapping Stage**

5.2.2.1 **Input Data**

The data collection stage deals with the capture and alignment of the set of input images. When this stage is complete, the set of input images Φ is suitable as input to the depth-mapping stage. Before detailing the specifics of the depth-mapping process using Φ as input, it will be useful to clarify some properties of Φ on which the entire process is based.

- Φ is the set of input images, of size N, where Φ_i ($1 \leq i \leq N$) refers to the i^{th} image.
- All Φ_i ($i = 1 \ldots N$) have the same resolution of X by Y pixels.
- Φ has been geometrically aligned, therefore $\Phi_i(x,y)$, the pixel at (x,y) in Φ_i, displays the centre of the same scene point for $i = 1 \ldots N$. The scene can therefore be described in terms of an 'image plane', where $\Phi_i(x,y)$ ($i = 1 \ldots N$) displays the scene point at (x,y) on this image plane, which we refer to as $P(x,y)$.
- The depth of the focal plane in any Φ_i is known in advance (from the image capture stage).
- The depth of the focal plane increases monotonically from $\Phi_i \ldots \Phi_N$, starting at depth d_1 in Φ_i and finishing at depth d_2 in Φ_N, where $(d_2 - d_1)$ is a global depth interval which contains the entire scene, *i.e.* the depth of any $P(x,y)$ is within this depth interval.
- Let the vector $I(x,y)$, where $I(x,y)_i = \Phi_i(x,y)$, $i = 1 \ldots N$ describe the intensity variation of the scene point $P(x,y)$ as the depth of the focal plane increases from image $\Phi_i \ldots \Phi_N$. Since we assume that for each i the depth of the focal plane is known, $I(x,y)$ actually gives a sample of intensity/depth pairs.

With the above definitions in place, it is straightforward to see how the input data is separated per-pixel. The vector $I(x,y)$ takes the intensity of the same single pixel (x,y) from each input image, and this provides input data for one unique scene point. This one-to-one relationship between the vector $I(x,y)$ and the corresponding scene point

$P(x,y)$ is an important feature of the proposed method. As should be expected, the method will process each vector $I(x,y)$ independently to find the depth of $P(x,y)$ to give a depth-map of resolution XY, the same as the image resolution.

5.2.2.2 **Intensity and Focal Plane Depth**

As indicated, the proposed method will infer the depth of each sampled scene point $P(x,y)$ in the image plane, using the variation in intensity of $P(x,y)$ as the depth of the focal plane moves from the front to the back of the scene. The intensity/depth samples are given by $I(x,y)$. The process is defined for one individual point, with the assumption that it can be repeated for all points (or indeed for as many points as are required) to output a depth-map of the scene.

As indicated in the overview of the method in the introduction to this chapter, the depth-map is computed by locating from the intensity/depth sampling given by $I(x,y)$, the intensity/depth where $P(x,y)$ is in-focus. In order to locate this point, it is necessary to model the relationship between the intensity of $P(x,y)$ and the focal plane depth. This modelling is based on the predictions of the optical model of the camera system outlined in *Chapter 2*.

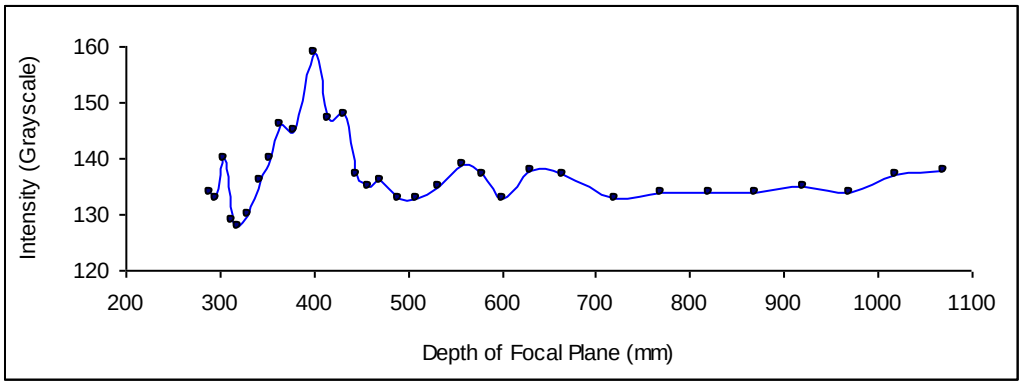
Let d indicate the depth of $P(x,y)$, and for simplicity let d be exactly equal to the focal plane depth in the image Φ_j ($1 \leq j \leq N$). This means that intensity $I(x,y)_j$ describes $P(x,y)$ in-focus. From the optical model of the camera system, $P(x,y)$ will be blurred in the images Φ_i ($i \neq j$), and the blur magnitude (diameter) will be greater the further away the focal plane depth of a given image Φ_i is from depth d. According to the optical model which is derived from the thin lens model, the PSF of the blur of $P(x,y)$ should be *identical* as the focal plane moves the same distance away from depth d in *either* direction, towards or away from the camera. Therefore, on a plot of $I(x,y)$, *i.e.* a plot of intensity of $P(x,y)$ against focal plane depth, there should be symmetry around the depth where $P(x,y)$ is in-focus.

Though the symmetry in the data is the key feature of the relationship between intensity and focal plane depth, the optical model of the camera system also predicts other features of the relationship. For example, it predicts that as we move away from the focal plane depth where a point $P(x,y)$ is in-focus, the intensity should represent an *average* intensity of the increasingly large region of blur with $P(x,y)$ at the centre. In real scenes, as we increase the size of a region around a certain point, we incorporate a series of increasing-sized local 'features'. Therefore as the region of blur expands it should incorporate a series of increasing-sized local features. The intensity should therefore keep converging to the average intensity of each local feature, resulting in an intensity/depth relationship which converges to a set of intermediary values, which themselves converge to some global value (where the extreme of 'global' is the average intensity of the whole scene).

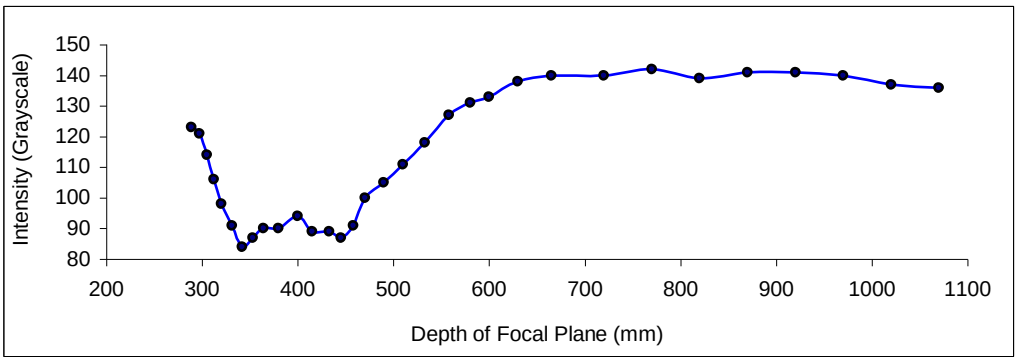
Of course, it is acknowledged that the optical model of camera system used is very simplistic and idealised, and perfect symmetry around the in-focus depth will not generally occur when using real lenses. However, the central hypothesis of the proposed method is that in practice, the optics of the camera system as predicted by the model will have a sufficiently greater effect on the intensity/depth variation than unaccounted-for optical effects, noise, and inaccuracies in image capture and alignment, that it will be possible to recognise the pattern of the relationship described above in real data and use it to locate the depth that a scene point is in-focus.

In practice, this pattern is indeed visible. What we typically see is a fast convergence to a particular intensity, which can be regarded as global. In some cases, the repeated re-convergence to more local average intensities is also recognisable, but this is often very subtle. Importantly, this typically results in a sharp local minimum or maximum point at the in-focus focal plane depth. *Figure 5.2.2.2.1* shows plots of the samples of intensity at different focal plane depths for four randomly selected scene points taken from the test dataset. As is clear from all four plots, the relationship predicted above is present in real data, and can therefore provide assistance in the search for the depth where a scene point is in-focus.

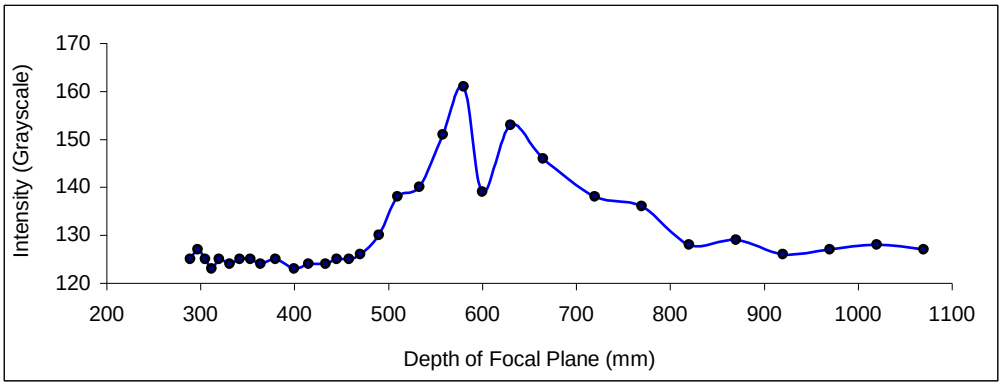
(*a*)



(*b*)



(*c*)



(*d*)

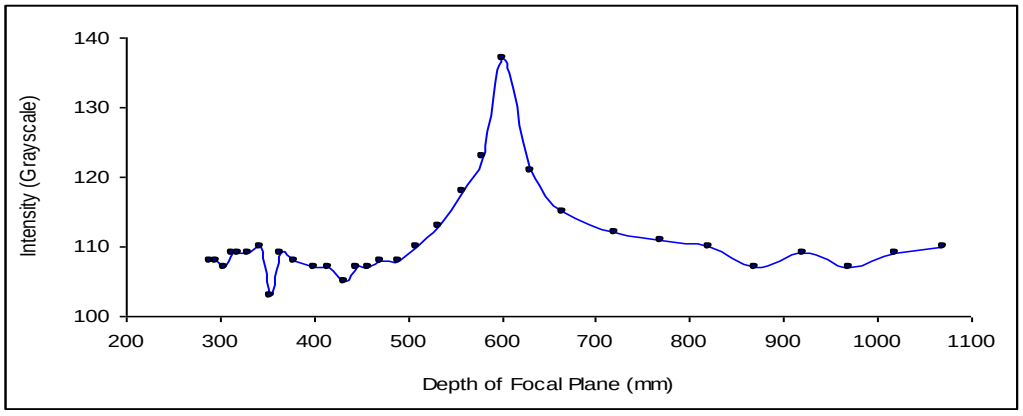


Figure 5.2.2.2.1 Four randomly selected scene points from the test dataset, plotting $I(x,y)$ against focal plane depth given in *Table 5.2.1.1.2*. (*a*) A straightforward sharp peak at in-focus depth (*b*) Trough trend with local peak at in-focus depth (*c*) peak trend with local trough at in-focus depth (*d*) Straightforward broad peak at in-focus depth.

When interpreting *Figure 5.2.2.2.1* it is important to remember that the test dataset contains only two depth levels: 400mm and 600mm. (*a*) and (*b*) show points from the first depth level at 400mm and (*c*) and (*d*) show points from the second depth level at 600mm.

Immediately obvious from all four plots in *Figure 5.2.2.2.1* is evidence supporting the symmetry hypothesis. There is some noise, and potentially some obscuring factors, for example in (*c*) and (*d*) where the intensity converged to close to the camera appears to be slightly lower than the intensity converged to away from the camera. However, the overriding effect is one of symmetry, as predicted by the optical model.

Another effect which can be observed in the plots is the repeated convergence to a local average intensity. In (*a*), the convergence to the global intensity value is rapid, and the effect is not clear. However, in (*b*) and (*c*), the effect is clearly displayed in the respective local deviations from the general trend of the intensity variation around the in-focus depth. In (*d*) the convergence seems relatively smooth, and this may be due to a lack of highly localised features near to that particular scene point, so that the blur averages the intensity smoothly towards the global average.

5.2.3 **Metrics for Finding the In-Focus Depth**

Having predicted the features of the intensity/depth relationship, the goal is to develop metrics which can be used to identify the in-focus depth by identifying those features in real data. The two metrics developed in this work are discussed in the following two sections.

5.2.3.1 ***Metric 1***

The first metric developed, *Metric 1*, is a classic Depth-From-Focus search. That is, from the N samples of intensity of a point $P(x,y)$ given in $I(x,y)$, find the intensity which represents $P(x,y)$ *most* in-focus. In a broader sense, this means that we are searching for the input image which shows $P(x,y)$ in the best focus. Because of this, the depth resolution of the depth-map will match the depth resolution of the input images. Therefore this metric is best suited for applications where this limit on depth resolution is acceptable, for example, Image Processing applications which only require information about the input images themselves.

Metric 1 is based on the hypothesised relationship between intensity and focal plane depth discussed in a previous section. Using the samples of intensity/depth for a point $P(x,y)$ given by $I(x,y)$, the aim is to locate the intensity/depth sample which shows $P(x,y)$ in best focus. Because only the samples in $I(x,y)$ are being searched over, the process is greatly simplified. In effect, we are searching for the point which is closest to the 'true' in-focus intensity/depth.

The task of searching for such a point can be performed by ranking each 'candidate' sample using a weighted sum of three factors, which quantify the features that the in-focus sample should have, as predicted by the optical model. The factors used in *Metric 1* can be summarised as symmetry around the candidate sample, sum of rates of change of intensity around the candidate sample, and distance of the candidate sample from the mean intensity. Importantly, all three factors are based on

the relationship of a candidate to the other samples from the set, as should be expected since the hypothesised relationship between intensity and depth of focal plane is being evaluated.

Where the term 'around the candidate' is used in the summary of the first and second factors, this refers to those samples adjacent to the candidate sample in the plot of the sample set, within a certain *interval* width either side of the candidate. This interval width, an integer, is a parameter used in the ranking function and literally refers to the number of samples to the left, and right, respectively, on the plot of the sample set to consider when evaluating the candidate on the first and second factors.

Though each factor has a basis in the optical model used in this work, the method of numerically evaluating the factor take some influence from empirical observation of data, to better emphasise the features seen in 'best-focus' samples in practice. Below the method of calculating each factor is explained, with the explanations taking the geometrical perspective of the sample set as a plot of intensity against depth of focal plane, for purposes of clarity. Finally the entire ranking function is presented.

- **Symmetry around the Candidate Sample**

The optical model predicts that the effect of the defocus blur on intensity should be identical when the plane of focus is at the same distance either side of a scene point, causing a global bilateral symmetry in the data with the line of symmetry at the in-focus focal plane depth (see *Chapter 2*). This symmetry is straightforward to evaluate by considering the differences in intensity between samples within the interval to the left of the candidate, and the estimated intensity (calculated by interpolation between the two nearest samples) at the same depth on the right side of the candidate. This is clarified for an example interval width of 3 in *Figure 5.2.3.1.1*.

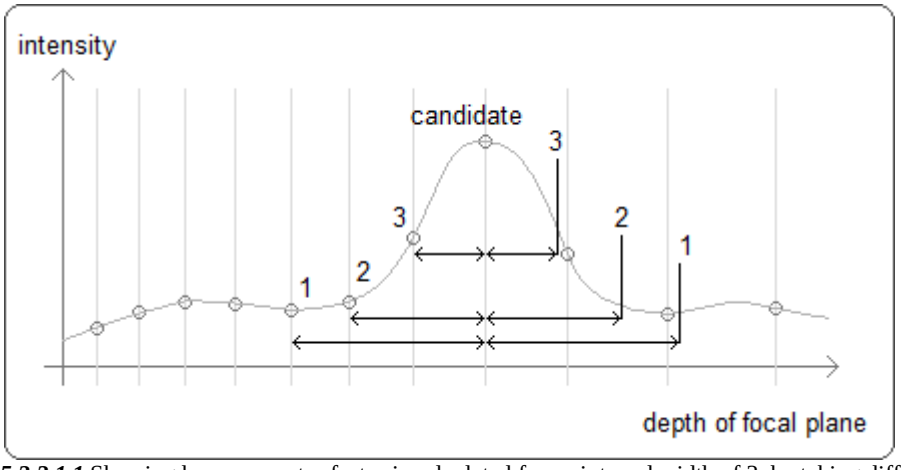


Figure 5.2.3.1.1 Showing how symmetry factor is calculated for an interval width of 3, by taking differences between the intensity of a sample to the left of the candidate and its corresponding intensity at the same depth to the right of the candidate, which is estimated by interpolation between the two samples either side.

After taking the difference for each of the i samples within the interval of width i, a total value for symmetry, S, is then calculated by summing each difference. Clearly as S increases there is less symmetry around the candidate, so in the ranking function the inverse of S is used so that the symmetry increases as this inverse value increases.

- **Sum of gradients around the Candidate Sample**

The optical model predicts that the rate of change of intensity with respect to depth of focal plane, *i.e.* the gradients between samples on the plot, is likely to be greater on average for samples closer to the in-focus focal plane depth. Briefly, this is because when the focal plane depth is closer to the in-focus focal plane depth *d*, the sample intensity is an average of a scene region with a smaller radius, which will generally be an increasingly local feature (with the most local feature being the scene point itself at focal plane depth *d*). This intensity is therefore likely to change rapidly near to d, whereas further from d it is likely to more smoothly converge to a more global average, resulting in greater rate of change of intensity with respect to focal plane depth closer to the in-focus focal plane depth.

In practice, the most straightforward way to calculate a numerical value representing this factor is to sum the absolute magnitude of the gradients between each successive sample within the interval, on both sides of the candidate, to produce a value *G* for this factor. This effectively takes a sampling of the rates of change of intensity around the candidate in the interval area, and emphasises those candidates which show large differences in intensity between successive samples within this region. As the value of *G* increases, it indicates that the candidate is more likely to be the best-focus sample.

The decision to use the sum of absolute gradient magnitude, with no specific weighting depending on the proximity of each sample to the candidate, is based on analysis of the gradients in real data. Although the optical model predicts that rate of change will generally be greater towards the in-focus focal plane depth, there is no obvious way to predict or even generalise the pattern of gradients around the candidate for individual samples. Therefore, evaluating the entire interval area only on gradient magnitude is judged, empirically, to be the best general way to evaluate the prediction of the optical model in real sample sets. Note there is no requirement to make the region evaluated equal on both sides of the candidate, as with the first factor, as rate of change (and not specific intensity values) is being evaluated, and this calculation is only intended as a general sampling of the region around the candidate.

Figure *5.2.3.1.2* clarifies the explanation of the calculation of this factor, again for a sample interval of 3.

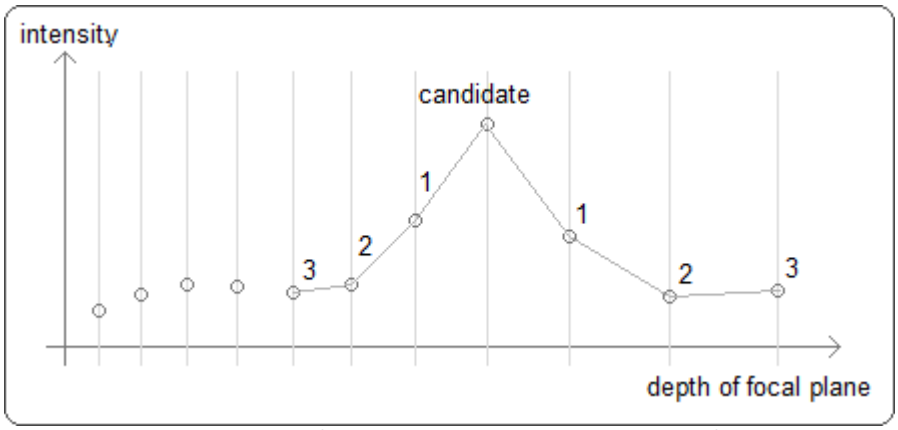


Figure 5.2.3.1.2 Showing how the sum of gradients around the candidate is taken for interval 3. Each gradient between the successive samples is taken, giving 2*i* gradients for interval width *i*, the absolute magnitude of these values is summed to give a value *G*.

- **Difference from Mean**

This factor is closely related to the previous factor, but isolates the evaluation of the absolute difference in intensity between a candidate sample point and the mean intensity of all sample points, to produce a value *D*. The optical model predicts that as the focal plane moves away from the in-focus depth, the intensity value of the samples will converge to some global average intensity *I*. Since the mean will therefore be biased towards *I*, it implies that the best-focus sample, which is more likely to have an intensity which differs significantly from *I* relative to the other samples, will generally be among the samples with intensity furthest from the mean.

Of course, the model does not explicitly predict that the best-focus sample will have the greatest distance from this mean. On the contrary, it is not forbidden by the model for the in-focus intensity to be equal to the mean intensity, even where the intensity varies arbitrarily at other focal plane depths. However, it can be seen empirically that in general, a significant difference from the mean intensity, relative to other samples, is a good indication that a sample is the best-focus sample, therefore in combination with the previous two factors it can be used effectively to rank a candidate sample.

- **Ranking Function**

Having introduced how the individual factors are calculated, the ranking function is a straightforward weighted sum of these, where S(*i*) is the symmetry value, G(*i*) is the sum of gradients value, and *D* is the difference from mean intensity for a candidate sample. The parameter *i* is the interval width that the calculations of S(*i*) and G(*i*) are based on.

$$Rank = \frac{a}{S(i)} + b\,G(i) + c\,D$$

(*Eqn* 5.2.3.1.3)

The ranking function will simply be maximised on every candidate sample from the sample set, to find the estimated best-focus sample from that set for each scene point.

To conclude the discussion of *Metric 1*, the weighting parameters {*a, b, c}* and the interval parameter *i* will be explained in more detail:

- The weighting parameters {*a, b, c}* will be empirically determined, and will be constant for any given implementation of *Metric 1*. Their purpose is to simultaneously normalise and weight the influence of each factor. Each factor has been quantified on a scale such that the likelihood of the candidate being the best-focus sample either increases or decreases monotonically as the value of the factor increases. However, with respect to each other the scales of the factors are somewhat arbitrary, and so require normalisation before being used together to give the rank. Even assuming the scales are normalised, however, the factors still need to be weighted according to their respective influence on the rank. Although each factor has basis in the optical model, there is no obvious way to predict from the model the influence each should have on the overall ranking in practice. The most straightforward way to achieve both normalisation and appropriate

weighting of the three factors is by incorporating both into the three weighting parameters {*a, b, c*} and calibrating these empirically for a set interval width.

- The interval width *i* is an important parameter. The integer value of *i* is essentially a trade-off between basing the calculation of factors S(*i*) and G(*i*) on more data as the value of *i* increases, thereby giving these values more data support and greater accuracy, but at the same time losing two potential candidate samples from either end of the sample set every time *i* is increased by 1. This happens because no candidate points can be considered within a distance *i* of the boundary (*i.e.* the first and last element) of the sample set, as if so there would not be an equal interval width on both sides of the candidate, which is essential to the calculation of both S(*i*) and G(*i*). The value of *i* producing the best results can be determined empirically, as will be the case in the results of *Metric 1* presented in *Chapter 6*.

5.2.3.2 ***Metric 2***

The second metric, *Metric 2*, takes a different approach to locating the intensity/depth where a point *P*(*x,y*) is in-focus. Here, the goal is to locate this intensity/depth on a continuous scale of intensity/depth, again using *I*(*x,y*) as input. In other words, this metric takes *I*(*x,y*) as a set of intensity/depth samples, assuming that these samples are taken from a function describing the relationship between intensity and focal plane depth. Therefore, unlike in *Metric 1*, the true in-focus intensity/depth point is searched for, which may be between samples in *I*(*x,y*).

Metric 2 therefore makes no assumptions about the input images. In theory, none of the input images could show *P*(*x,y*) in perfect or even good focus, but *Metric 2* could still identify the depth of focal plane where *P*(*x,y*) should be in-focus. Because of this, it is imprecise to describe *Metric 2* as a purely DFF approach; to some extent it is a hybrid of DFF and DFD as the in-focus depth is not estimated by a simple search over the input images to find the image of best focus, but rather the relationship between depth and blur is analysed to infer the depth of focus, even though blur is not directly modelled.

Metric 2 is more suitable for general depth-mapping applications, where we would like to infer the depth of each point on a continuous scale and not limit the depth resolution to that of the image set.

The basis of *Metric 2* is identical to that of *Metric 1*, in that the same relationship between intensity and focal plane depth, predicted by the optical model of the camera system and discussed in *Chapter 3*, is assumed. Clearly, we would like to model this relationship using a function relating intensity to depth of focal plane, where parameters of this function describe (at least) the depth of the in-focus focal plane. Then, by fitting the function to the data in *I*(*x,y*), we can take the best-fit parameter as the in-focus depth for the point *P*(*x,y*).

The function chosen to describe the relationship between intensity and focal plane depth was a Gaussian-style function, given by *Eqn 5.2.3.2.1*.

$$f(d) = ae^{\frac{-(d-d_f)^2}{b}} + c$$

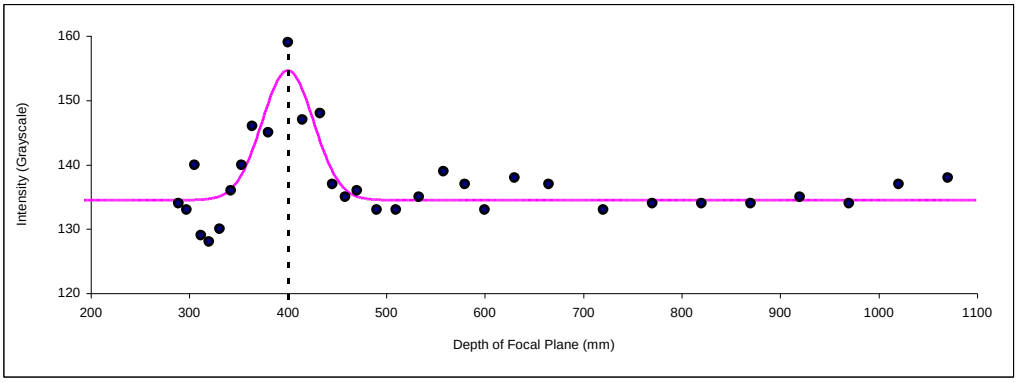
(*Eqn 5.2.3.2.1*)

The parameter we are most interested in recovering by fitting the function is the in-focus depth of focal plane, *i.e.* d_f in *Eqn 5.2.3.2.1*. As discussed in *Chapter 3*, the optical model predicts that this depth should be at the point of bilateral symmetry in the data. Furthermore, the model predicts that the intensity at depths away from this point should converge to some global value. *Eqn 5.2.3.2.1* is a very simple model which incorporates both of these features of the relationship between intensity and focal plane depth.

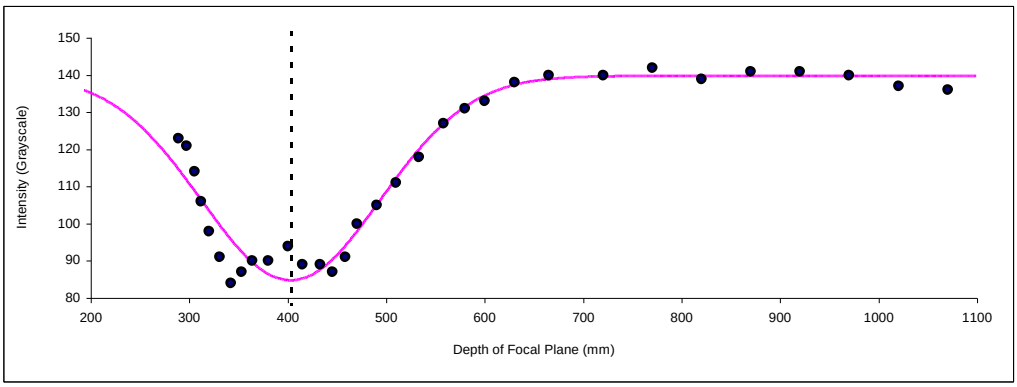
It is accepted that this is too simple a model to describe precisely the relationship between intensity and focal plane depth, as in many cases we are likely to see not a single convergence to a global value, but a repeated convergence to intermediate intensity values, as discussed earlier in this chapter. However, here our only concern is the recovery of the in-focus depth. If the precise intensity value at this depth were required, then a more comprehensive model would be advantageous. Such a model could be a linear combination of Gaussian-style functions with a common mean. Here though, it is important to emphasise that *Eqn 5.2.3.2.1* provides a model of the intensity-depth relationship which is sufficient for the purpose of recovering the in-focus depth, and a more comprehensive model would be more computationally expensive to fit to the data in practice.

In *Figure 5.2.3.2.2* below, the fitting of *Eqn 5.2.3.2.1*, and the inferred depth of focus, are shown for the four random scene points in the test dataset shown in *Figure 5.2.2.2.1*.

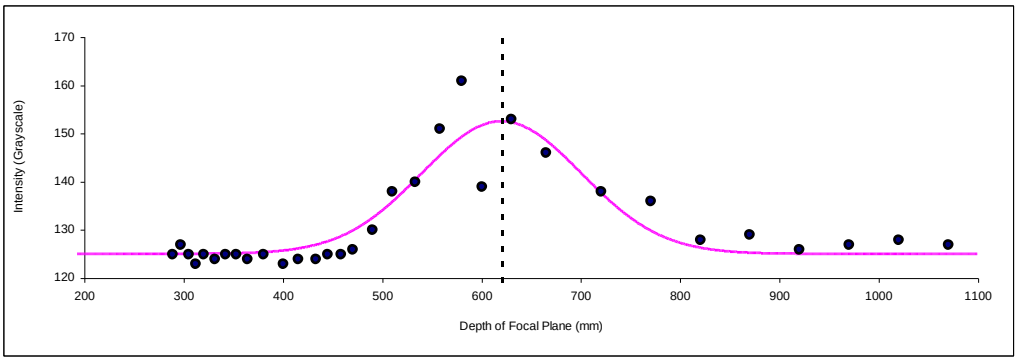
(*a*)



(*b*)



(*c*)



(*d*)

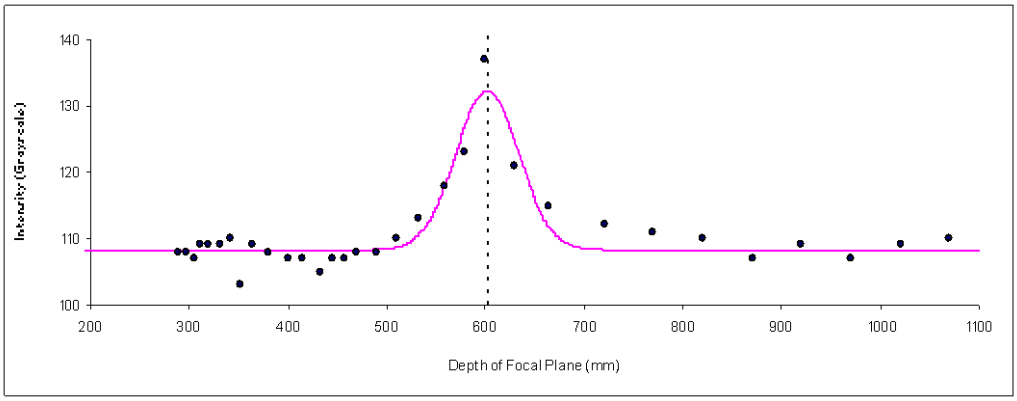


Figure 5.2.3.2.2 *Eqn 5.2.3.2.1* fitted to the data from the scene points $I(x,y)$ from four random scene points shown previously in *Figure 5.2.2.2.1*. The dotted line in each plot shows the point of symmetry, *i.e.* the inferred in-focus depth, which is the parameter d_f.

Figure 5.2.3.2.2 illustrates empirically the validity of the claim that fitting *Eqn 5.2.3.2.1* to the samples of intensity/depth from a scene point can find the point of symmetry in the data. The symmetry point is given by the best-fit value of the parameter d_f, shown as a dotted line.

It may seem obvious to point out that a much simpler iterative bilateral symmetry search could be employed to solve the problem of finding the symmetry point on the 'continuous' depth scale (continuous down to a small grain size resolution). However, such a simple approach would have difficulties which are dealt with by fitting *Eqn 5.2.3.2.1* to the data using an optimisation algorithm (such as trust region).

Firstly, a simple symmetry search could only consider the differences between points within a certain distance either side of a candidate symmetry point, meaning that boundary points would be less supported by data than points in the centre of the depth scale. There is no straightforward metric for comparing the confidence in the symmetry value based on very local data as opposed to more global data. Secondly, a simple symmetry search would not necessarily deal with slight skew and noise in the data in the same way that fitting *Eqn 5.2.3.2.1* does. It might be possible to control this problem using tactics such as smoothing the data, and weighting the symmetry of points closer to the candidate symmetry point as more relevant, but again there are no obvious general rules about precisely how this should be done.

Fitting *Eqn 5.2.3.2.1* to the data deals with both these problems automatically, to give a much more general solution to the problem. Because the actual function relating intensity to depth of focal plane is modelled (and not just the symmetry aspect of the relationship), all data can be taken into account for any candidate point, and any skew or noise in the data is dealt with automatically as the function is fitted to the global data, and the effect of local anomalies is therefore reduced.

5.3 Theoretical Strengths / Limitations of the Proposed Method

To conclude this chapter, the theoretical strengths and limitations of the proposed method will be discussed, in the context of the input hardware setup. In other words, the strengths and limitations of the method compared to other approaches with very similar input hardware setups. Note the strengths and limitations of the general depth acquisition approach, *i.e.* using multiple images captured with a conventional camera at a fixed position, are briefly discussed in *Chapter 1*.

A discussion of the practical strengths and limitations of the proposed method based on results using test data is given in *Chapter 6*.

5.3.1 Strengths

- *No Direct Blur Modelling*

The typical traditional approach to depth acquisition from images is to model defocus blur in an attempt to recover its magnitude, and hence the depth of the blurred point (see *Chapter 2*). However, directly modelling defocus blur caused by a camera system is a very difficult problem. General models of defocus blur are, by definition, loose estimations of the true blur model which varies from camera to camera, and specialised models for a particular camera system often require cumbersome modifications or pre-calibration to obtain an empirical model.

On the other hand, approaches which avoid directly modelling blur typically do so by searching only for characteristics of focus in images. These naïve pure DFF approaches, however, can only infer depth at the resolution which the input images cover. Clearly, this is not an adequate general solution to the problem of depth acquisition from images.

There is, however, a third class of approach which incorporates the relationship between defocus blur and depth, whilst avoiding the direct modelling of the blur. The proposed method takes this approach by analysing the change in intensity at the centre of the region of blur over a global range of focus settings. In other words, the effect of the blur in the image is utilised in depth acquisition, without having to directly infer the cause of the defocus blur effect (*i.e.* the magnitude of the blur) using a blur model.

- *Low Number of Input Images*

In the proposed method, the image set must span a global range of focus settings such that the focal plane moves from the front to the back of the scene. However, there is no specific requirement of the resolution or even the spacing of the focal plane of each image over this range. In addition, only the focus setting parameter is varied, *i.e.* aperture size and zoom are kept constant; therefore there is only one image per focal plane. Compared to other approaches which require an image set with a global focal plane range, the proposed method requires a low number of input images.

For example, in naïve DFF methods the depth resolution is determined by the resolution of the focal plane samples across the range, *i.e.* the number of images. Though this is true of the proposed method using *Metric 1*, *Metric 2* can theoretically be applied to image sets where no points are shown in-focus in any particular image. The problem of depth resolution being improved by the number of images also occurs in more advanced techniques [21, 22, 23]. The approach used in [21, 22, 23] also has the limitation of requiring several images per focal plane, as aperture size is also varied to provide the necessary input data.

- *Practicality of Image Capture*

The practical ease of image capture is a very important advantage of the proposed method. As mentioned above, only the depth of the focal plane is varied across the input image set and all other camera parameters such as zoom and aperture size are constant. In addition, there is no need to modify the camera system, and any pre-calibration of the camera in use is minimal. This means that the method can be readily applied in practice using existing hardware with very little difficulty.

- *Simple Image Processing*

Because the proposed method works in a per-pixel manner, it avoids many of the practical Image Processing difficulties associated with techniques which use windowing to evaluate focus/blur in the input images. For example, there are no difficulties with image boundaries, and the problem is readily parallelisable with no data redundancy. This is particularly applicable when working with high-resolution images/depth-maps.

5.3.2 **Limitations**

- *Scene/Camera System Assumptions*

One major limitation of the proposed method is that is relies on fairly strong assumptions about both the scene and the camera system in use.

As previously highlighted, the optical model of the camera system on with the theoretical aspects of the proposed method are based is very simplistic. However, it is claimed that in reasonable practical situations, the optics predicted by this model will be the overwhelmingly dominant factor in the true optics of the camera system, so in practice this should not present a problem.

The scene assumptions are also very specific. The optics on which the method is based require that the scene be largely fronto-parallel, and suitably complex in intensity variation (*i.e.* a flat-colour surface such as a blackboard would not be suitable). There are also very strong assumptions about the local environment of the scene. For example, the scene is assumed to be completely static with fixed non-ambient lighting (ambient lighting is allowed to vary) over the time interval of image capture. In practice these assumptions simply translate to limitations on the application of the proposed method. However, it is claimed here that there are many application areas, particularly in natural scenes where there is a fairly static local environment, where these assumptions are reasonable.

- *Susceptibility to Noise*

Since the proposed method relies on analysis of intensity variation as blur changes, the method is not very robust to noise. Noise would become a particular problem if ambient lighting conditions were not ideal (*i.e.* too low or too bright) or changed over the exposure interval (*i.e.* during image capture) or rapidly between the captures of successive images in the input set. Though as previously discussed the global lighting over the image set can be normalised, it must be recognised that in a typical practical situation this is likely to add noise to the input data.

The impact of noise, however, is mainly on the raw depth-map produced by the method. Since this method produces high-resolution depth-maps, it is likely that post-processing of the depth-map could reduce the impact of anomalies due to random noise, assuming that the majority of scene points are not significantly affected by noise.

Chapter 6

Results, Analysis and Evaluation

In this chapter, results of the proposed method using real data are presented. Firstly, the method is tested on a test scene. This provides both proof of concept for the method, and the basis for both quantitative and qualitative analysis of results. Secondly, results of the method when applied to a more natural test scene used in [21, 22, 23] are presented, and a comparison is given between these results and the results from [21, 22, 23]. Finally, in context of all the presented results, the strengths and limitations of the method are evaluated.

6.1 **Results of the Proposed Method using the Test Scene**

6.1.1 **Test Data**

As discussed in *Chapter 5*, the proposed method is tested using an artificial test scene with two depth levels. A full overview is given in *Chapter 5*, but as a brief re-introduction, this test scene is intentionally simple in structure so that the accuracy of the depth-maps produced by the method is straightforward to verify, but the scene simulates the variance of intensity in a 'natural' scene by using photographs of natural scenes at both levels, so in this sense the two depth levels cannot be inferred trivially.

Before presenting the results, two specific points about the use of the test data in producing these results must be specified:

- Firstly, the raw images of the test scene are of 5 megapixel resolution. Rather than use this entire image area, it was decided to use only a patch of 1024 by 1024 pixels at the centre of each image. This lower resolution allows for a lower running time and smaller input and output file sizes, allowing more a practical testing framework. Generally a greater volume of test data will produce more statistically meaningful results. However in the case of this test scene, this is not particularly true, since there are only two depth levels meeting at roughly the centre of the scene, therefore a 1 megapixel patch taken from the centre of the image area is a very good statistical sample of the entire 5 megapixel image area. Moreover, as the proposed method is per-pixel, a patch can be taken as a sample of the entire image area without introducing image processing obstacles.

- Secondly, the test data is converted from RGB to greyscale. The conversion from the three channels R, G, and B to a single greyscale intensity is done using the following formula:

$$I = 0.2989 * R + 0.5870 * G + 0.1140 * B$$

(*Eqn* 6.1.1.1)

Since the proposed method looks at intensity, conversion to greyscale is a convenient way to acquire a single intensity value from the separate intensities of the R, G and B channels. Again, another way of achieving this would be to evaluate the three channels separately and average the results, but for a more practical testing framework it was decided to do this averaging of the channels prior to running the method.

6.1.2 ***Metric 1***

To give a brief re-introduction, *Metric 1* is designed to produce a discrete depth-map where each scene point (pixel) is given a depth index referring to the image where that point is judged to be most in-focus. For a detailed discussion of *Metric 1* see *Chapter 5*.

Since *Metric 1* will produce a discrete, relative depth-map of the test scene, it is relatively straightforward to analyse the accuracy of the depth-map quantitatively, as the correctly focused image for each of the two levels in the test scene is obvious from observation. Specifically, in terms of index of the 32-image test set, the first level is in-focus at image 11 and the second level is in-focus at image 22. The in-focus images from the test data patch are shown in *Figure 6.1.2.1*.

Figure 6.1.2.2 shows the 'ground-truth' image that will be used to quantitatively evaluate the accuracy of *Metric 1*. This ground-truth image was produced manually based on the observation of the in-focus images of the two levels in the test scene being 11 and 22 respectively. It is estimated that this ground-truth image has an error of +-5 pixels per row, meaning +-5120 pixels overall, therefore an estimated percentage error of +-0.488% to 3 s.f.

Figure 6.1.2.1 Image 11 (left) and Image 22 (right) from the test data patch taken from the test scene. Image 11 shows the first level in-focus, Image 22 shows the second level in-focus.

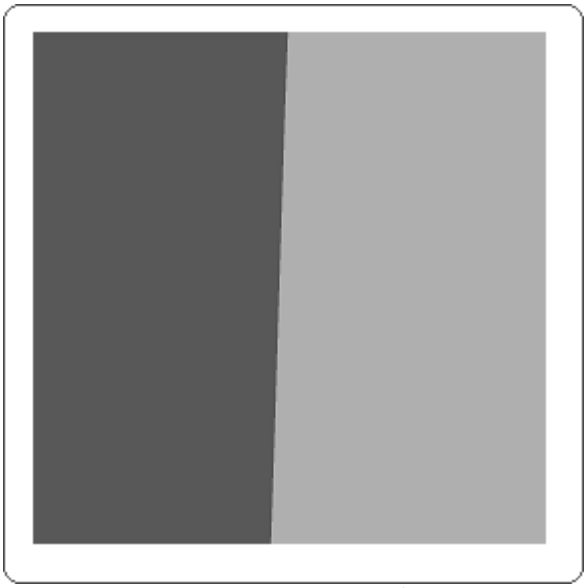


Figure 6.1.2.2 The manually produced ground-truth image for the test data patch of the test scene

6.1.2.1 **Parameters**

As discussed in *Chapter 5*, *Metric 1* has three parameters which weight the importance of the symmetry around the candidate sample, cumulative rate of change around the candidate sample, and deviance of the sample from the mean intensity. These parameters are referred to as *a*, *b*, *c* respectively.

Parameters *a*, *b* are reliant on another parameter, the interval width *i*, the number of samples around the candidate sample with which to evaluate *a* and *b* (see *Chapter 5*). The interval width will be varied as part of the test results, and the 'best values' of the other three parameters *a*, *b*, *c* were chosen empirically based on a trial-and-error calibration stage on the test scene. The values selected were $a = 0.2$, $b = 3.0$ and $c = 4.0$. Since these test results are intended as a proof-of-concept, it was deemed that choosing these values in this way is sufficient, rather than attempting to optimise the results over the entire range of parameters. It is proposed, however, that if these parameter values produce accurate results in the test scene, they are likely to have general validity for any natural scene.

On the other hand, it was decided to evaluate the effect of the change of interval width *i* on the results of *Metric 1* on this test scene, as this parameter is more closely related to the data itself. For different datasets which have, for example, a different number/frequency of focal plane depth sampling, it is likely that different interval widths will be appropriate to produce optimal accuracy in results because of the trade-off between the results being based on more data and the number of potential candidate samples being decreased.

6.1.2.2 **Results**

Figure 6.1.2.2.1 shows the depth-maps produced varying the interval width from 1 to 10, with 10 as the maximum possible value of the interval width without the 'true' depth values at sample 11 and 21 being unable to be evaluated. Of course, there is no assumption that this information is known *a priori*, but since this is a test and

evaluation of *Metric 1* it is not useful to evaluate with interval widths which are known to produce inaccurate results. Again, the parameters are set at $a = 0.2$ $b = 3.0$ $c = 4.0$. *Figure 6.1.2.2.1* also shows the ground-truth image for the test data patch, giving a visual indication of the accuracy of each depth-map.

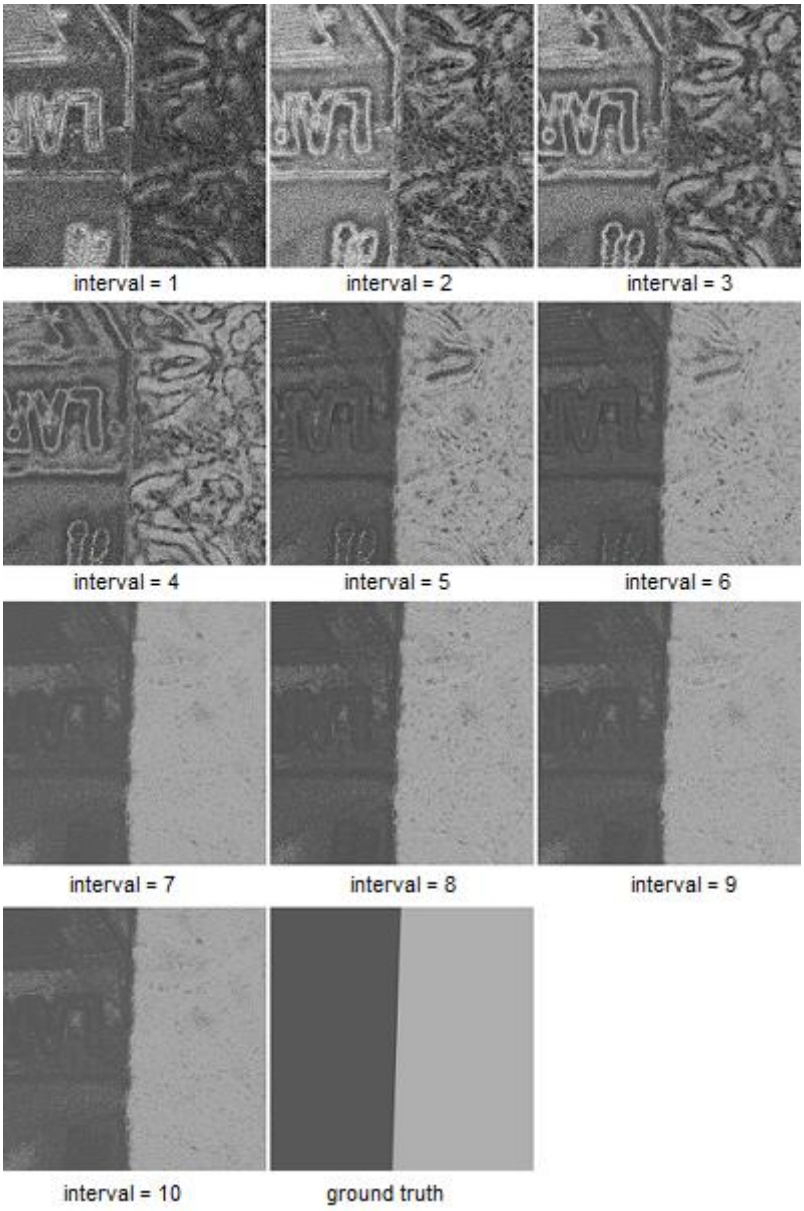
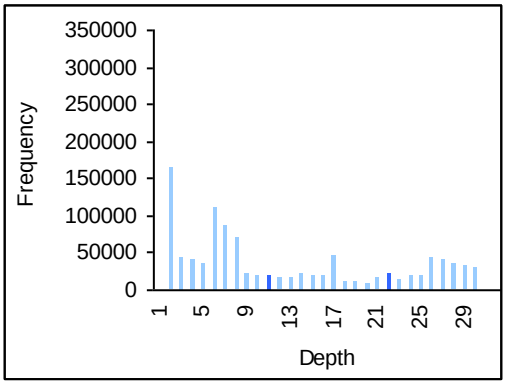
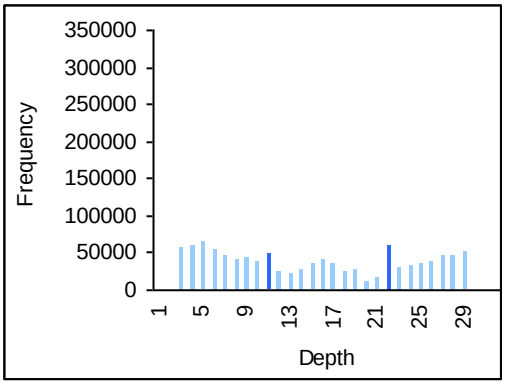


Figure 6.1.2.2.1 Depth maps produced by *Metric 1* with $a = 0.2$ $b = 3.0$ $c = 4.0$, and interval width 1…10. The ground-truth image is also shown for visual comparison.

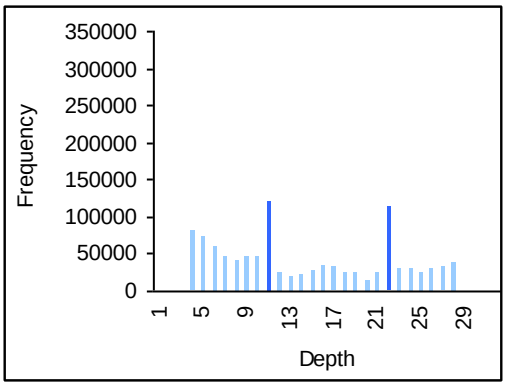
From visual comparison of the depth-maps with the ground-truth image in *Figure 6.1.2.2.1*, we can see that the accuracy of the depth-map appears to increase significantly as interval width increases. This trend can also be seen numerically by comparing histograms of the depth-maps (*Figure 6.1.2.2.2*), and comparing the distance and absolute error between each depth-map and the ground-truth image (*Figure 6.1.2.2.3*, *Table 6.1.2.2.4*, *Figure 6.1.2.2.5*).



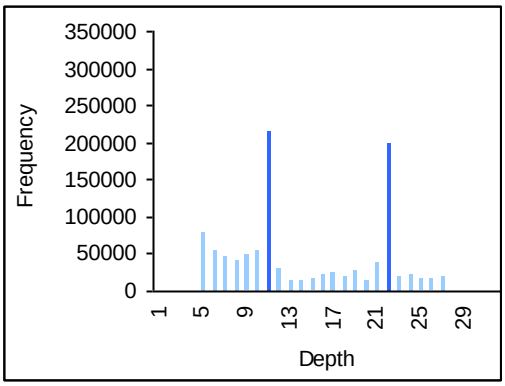
interval width = 1



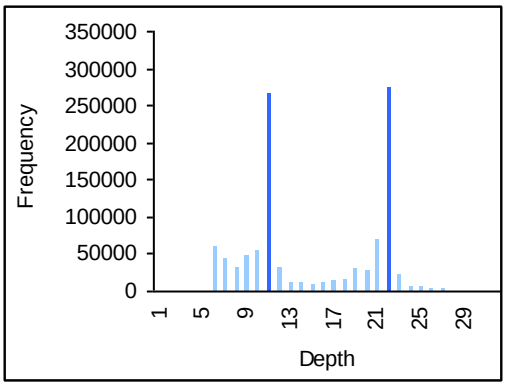
interval width = 2



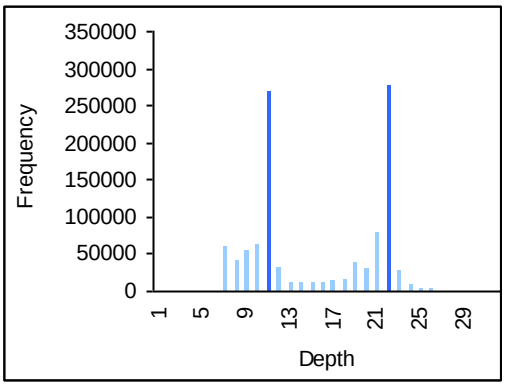
interval width = 3



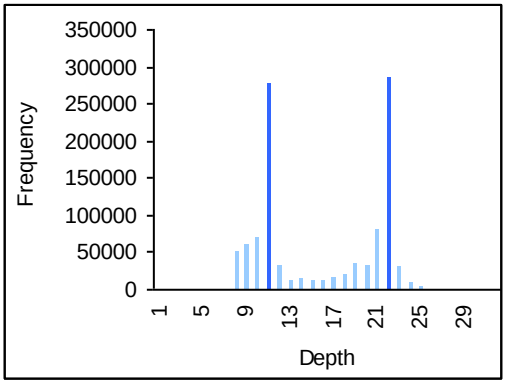
interval width = 4



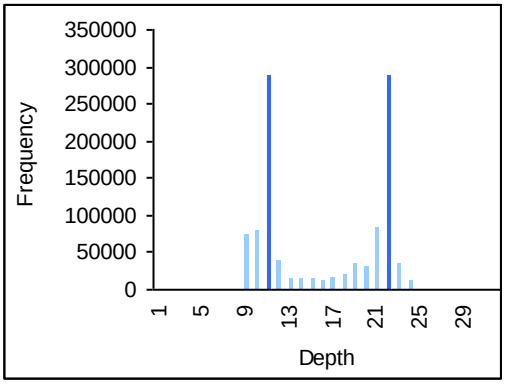
interval width = 5



interval width = 6



interval width = 7



interval width = 8

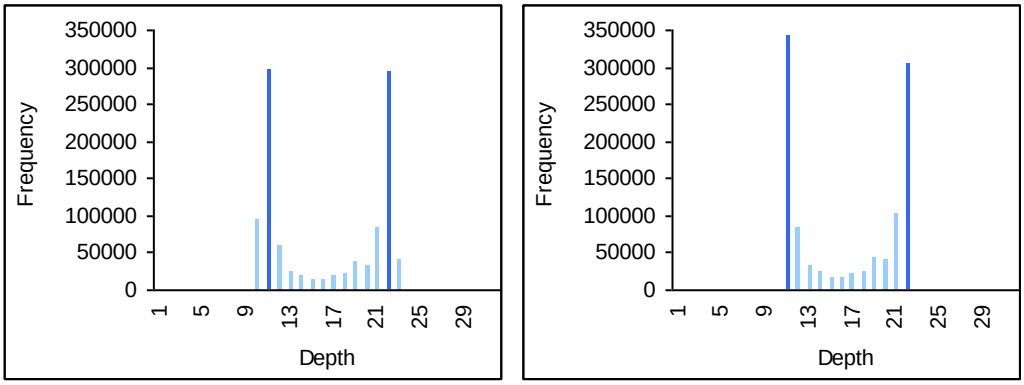


Figure 6.1.2.2.2 Histograms of the depth-maps in *Figure 6.1.2.2.1*, with the correct depth levels in darker blue. The frequency of depth values at or close to the correct depth values increases as the interval width increases (note this does not show whether or not the values are correct).

As seen from the ground-truth image, any point in the test data patch has a 'correct' depth value at either the value of 11 or 22. In the histograms in *Figure 6.1.2.2.2*, these 'correct' values are shown in a darker shade. As mentioned, we can see visually in *Figure 6.1.2.2.1* that as the interval width increases from 1 to 10, the accuracy of the depth-map (*i.e.* the similarity to the ground-truth image) increases. Although the histograms show nothing about the accuracy of individual points in the depth-maps, it is clear from a statistical point of view, supported by qualitative evidence in Figure 6.1.2.2.1, that they also show this increase in accuracy.

In addition to this conclusion, there is other interesting information available in *Figure 6.1.2.2.2* which is not shown directly in *Figure 6.1.2.2.1*. If we examine the distribution of the frequencies of depth values in each depth-map, as shown in the histograms, we see that as interval width increases the distribution of frequencies around the two 'correct' values converges to resemble two Gaussian distributions, with the 'correct' values at the mean, or the peaks, of each Gaussian. This implies that as interval width increases, not only is the absolute number of correct depth values increasing, but the incorrect results are becoming more generally accurate.

For example with interval width = 2, in addition to the correct depth values not having a relatively large frequency compared to the incorrect ones, the distribution is quite flat, meaning that there is little or no statistical evidence to support a hypothesis that a significant proportion of the incorrect values are 'almost' correct. Therefore we can conclude that the depth-map is largely meaningless.

However with interval width = 9, and in fact with widths 7, 8 and 10, we see two very clear Gaussian-like distributions with means at the correct values. Therefore statistically speaking, we can infer that the depth-maps have a high percentage of absolutely correct depth values, and also that any incorrect depth values are likely to be 'close' to being correct.

The above suggests that the hypothesis of a greater interval width increasing depth-map accuracy is correct. To provide further support for this conclusion, the differences between each depth-map and the ground-truth image (*Figure 6.1.2.2.3*), and the (non-normalised) absolute error of each depth-map (*Table 6.1.2.2.4*) are given.

Figure 6.1.2.2.5 shows the relationship between absolute error and interval width, and shows that the gain from increasing interval width in terms of accuracy decreases at an accelerating rate as interval width increases. This is an important

relationship to consider, as a greater interval width introduces the disadvantages of reducing the number of potential candidate samples, as well as (slightly) increasing processing time in practice. Although 'correct' results are possible in the test scene with interval width 10, this cannot be assumed in general. In practice, a lower interval may be chosen to incorporate a greater range of candidate samples, therefore the trade-off between interval width and accuracy is a very important consideration.

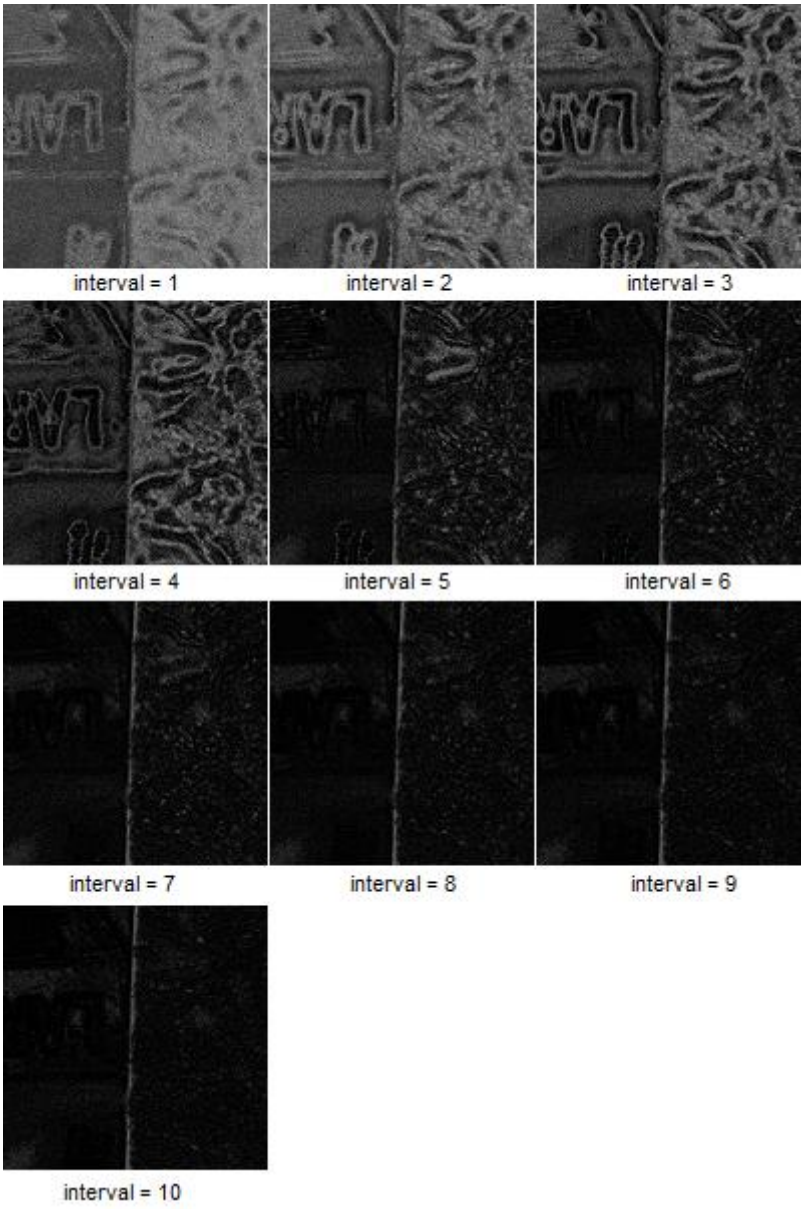


Figure 6.1.2.2.3 Images of absolute distance between depth-maps and ground-truth image (black = 0, white = 31)

Interval Width	Absolute Error From Ground Truth (non-normalised)
1	11508963
2	10392717
3	8430714
4	5572529
5	2672702
6	2147187
7	1749955
8	1527782
9	1380541
10	1362082

Table 6.1.2.2.4 Absolute error between depth-maps with interval widths 1…10 and ground-truth image (error ±0.488%). Note the error figures are not normalised or scaled; they are provided raw as a relative indication of change in error magnitude as interval width changes.

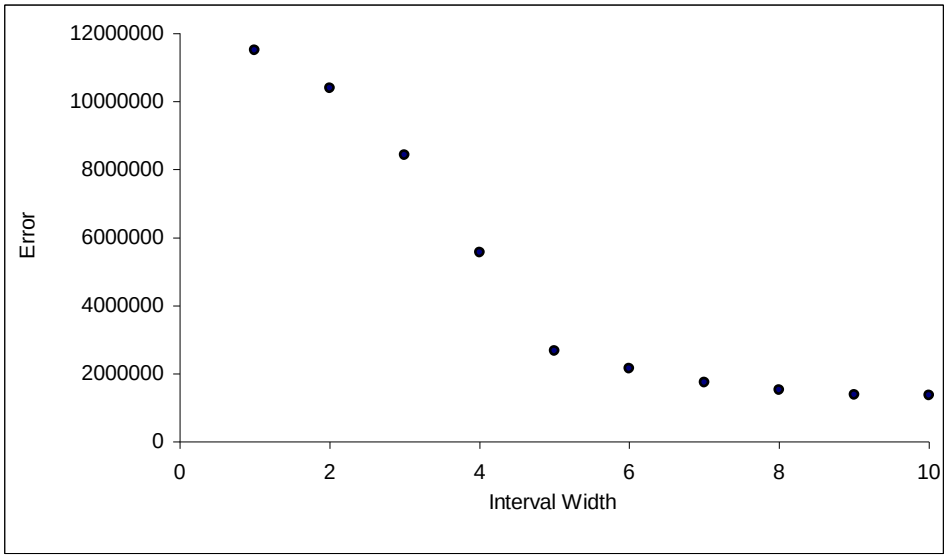


Figure 6.1.2.2.5 Plot of the error between depth-maps with interval widths 1…10 and ground-truth image (data from *Table 6.1.2.2.4*).

6.1.2.3 **Best Depth-Map Produced by *Metric 1***

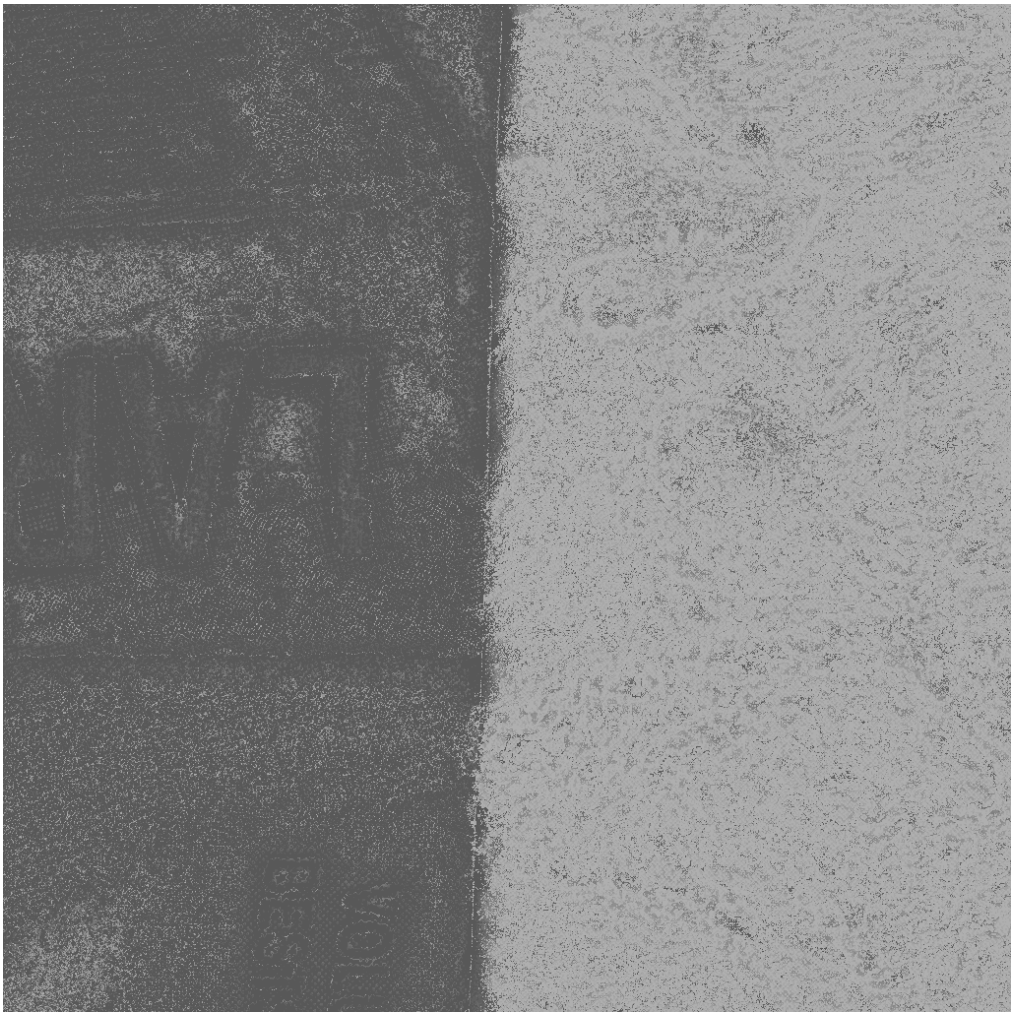


Figure 6.1.2.3.1 The best depth-map produced by *Metric 1* under the test parameters and interval width 10. Shown here in higher resolution so greater level of detail can be seen (actual depth-map resolution is 1024x1024 pixels).

6.1.3 ***Metric 2***

As a brief re-introduction, *Metric 2* is designed to produce a continuous, absolute depth-map. This means that the testing of *Metric 2* is different to the testing for *Metric 1* in a subtle but important way. With *Metric 1*, the depth-maps produced were discrete and the depth values were limited to the in-focus depths in each of the input images. In this way, the depth-maps produced by *Metric 1* could be described as relative, although absolute depth values can be assigned with prior knowledge of the in-focus depths in each image, which of course are available for the test scene.

By contrast, the depth-maps produced by *Metric 2* are continuous, *i.e.* the depth values inferred can lie between the set of in-focus depths from the input images. Therefore it is more natural to describe the depth values inferred by *Metric 2* as absolute rather than relative, assuming again that the in-focus depth of each input image is known. This distinction introduces an important consideration when testing *Metric 2* using the test scene. Rather than being able to assess the accuracy of results in a discrete, relative manner by comparing to the manually produced ground-truth as

for *Metric 1*, the accuracy of results must be assessed absolutely against the depth values of the ground-truth image.

Because the ground-truth for the test scene has been manually produced by selecting the depth levels which are visually the most 'in-focus', it would be inappropriate to assume that the depth values of the ground-truth are completely accurate in terms of absolute depth. Instead, a known error must be assigned to the 'estimated' ground-truth values to give the analysis of the error of the results validity. The error of the ground-truth has therefore been estimated simply as the sum of half the distance, in absolute terms, between consecutive depth levels around the two chosen 'in-focus' depth levels at image 11 and 22.

Assuming no human error in judging which depths are the most in-focus, which is reasonable considering the clear visual difference between the images, the logic behind this choice of error region is that the 'true' depth level could not lie closer to the less focused image than the more focused image, by simple geometric optics, therefore the 'halfway' point between the in-focus image and the less focused image, must be the boundary of the region where the true in-focus depth lies.

Therefore, when testing *Metric 2* we will keep the same ground-truth image used when testing *Metric 1*, but calculate the percentage errors of that ground-truth image using the empirically obtained depth scale seen in *Chapter 5*.

6.1.3.3 **Error of ground-truth**

As discussed, the images in the input set judged manually to show the first and second depth levels of the test scene in best focus are, respectively, image 11 and 22. The empirically determined focal plane depth of these images is 400mm and 600mm, respectively. In *Table 6.1.3.1.1* below, the empirically determined depths of focal plane of the images before and after the 'best-focus' images are given.

Image Index	Focus Setting (notches)	Depth (mm)
10	48	380
11	**50**	**400**
12	52	415
21	70	580
22	**72**	**600**
23	74	630

Table 6.1.3.1.1 The absolute depths of the focus settings surrounding the focus settings judged to be 'in focus' in *Metric 1.*

The estimates for the absolute depth of the two depth levels are therefore estimated as 400mm for the first depth level and 600mm for the second depth level. The percentage error of these estimates is calculated as half the distance between the in-focus image focal plane and the focal plane of the consecutive image in either direction. These calculations are given below:

First Depth Level (400mm)

$$\text{error towards camera} = \frac{\frac{400-380}{2}}{400} \times 100 = -2.50\%$$

$$\text{error away from camera} = \frac{\frac{415-400}{2}}{400} \times 100 = +1.88\%$$

Second depth level (600mm)

$$\text{error towards camera} = \frac{\frac{600-580}{2}}{600} \times 100 = -1.67\%$$

$$\text{error away from camera} = \frac{\frac{630-600}{2}}{600} \times 100 = +2.50\%$$

The percentage errors around both depth level estimates are of the estimates are similar and, in fact, quite low. This is strong evidence that, even though the ground-truth image is manually produced, it is viable for testing the accuracy of *Metric 2* over an absolute depth scale. As before, since the ground-truth was constructed manually, we will estimate that a maximum of 5 pixels per row are at the incorrect depth level at the boundary. This Figure is relevant when analysing the absolute error of the depth-maps, but can easily be avoided when qualitatively analysing the error images as high errors at the boundary between levels can be ignored.

6.1.3.2 **Implementations of *Metric 2***

The basic principle of *Metric 2*, as discussed in *Chapter 5*, is a model-fitting approach. In theory, the model fitting of *Metric 2* can be fully automated without any guiding heuristics; however in practice the use of heuristics is likely to improve results. To investigate this proposition, the results from two different implementations of *Metric 2* are evaluated.

The first uses no heuristic, selecting initial parameters for the model fitting randomly. However, the second follows the heuristic of keeping the initial parameters equal to the fitted parameters of the previously evaluated neighbouring pixel, where we begin evaluating the top-left pixel with random initial parameters and proceed through each row in turn, down through the scene area. This heuristic is based on an assumption that natural scenes are composed primarily of smooth surfaces (in terms of both depth and colour) so that, whilst not at the boundary of a surface, it is highly likely that neighbouring pixels will share a similar intensity/depth profile, and hence it is reasonable to assume that the fitted model parameters of a particular pixel will make good initial parameter estimates in the fitting of the model for a neighbouring pixel. It should be made clear here that this assumption is expected to hold well for this particular test scene, as it contains only two surfaces which are almost completely smooth in depth, and fairly smooth in colour, so we should expect greater than typical accuracy from the second implementation in the resulting depth-maps.

In *Figure 6.1.3.2.1* a depth-map produced by both implementations of *Metric 2* is given. Note that due to the pseudo-random nature of the implementation (even the second implementation has random initial parameter values for the first pixel evaluated), it can be said that these depth-maps were not arrived at deterministically and are therefore just 'samples' of results from *Metric 2*, unlike the depth-maps produced by *Metric 1* which are repeatable under the same input parameters. Also shown in *Figure 6.1.3.2.1* are the ground-truth image and a graphical representation of the absolute error of each depth-map (black indicates zero error, white indicates a maximum absolute error of 1050mm). These error images give a visual indication of the difference in accuracy between the two depth-maps.

A certain amount of qualitative analysis of the results can be performed by comparing the error images from *Figure 6.1.3.2.1*. The first implementation shows a significant proportion of low-intensity pixels at both depth levels, particularly the 'second' surface at 600mm from the camera. However, the consistency of results is poor, with accurate low-intensity pixels being interspersed with inaccurate high-intensity pixels across almost the entire depth-map. In fact, there is only one area, around the lettering on the first depth level, where the depth-map is consistently accurate. This can be seen by the patch of low-intensity pixels in that area.

The second implementation, on the other hand, clearly shows improvement in accuracy over the first implementation, since the average intensity of the error image is visibly lower, and there are clearly far larger 'patches' of very low intensity at both depth levels. As mentioned, this behaviour was expected, as neighbouring pixels are likely to have very similar intensity/depth profiles and therefore a similar fitting is likely to occur where the fitting of the model to each subsequent pixel uses the fitted model parameters of the previous pixel as initial parameters. However, whilst for large patches of the scene this has proved to be beneficial, it has also been counter-productive in other areas of the scene. In the depth-map, there is a very noticeable artefact of straight horizontal lines at very similar intensity. Where this intensity is low it indicates the advantage of using the heuristic; however where the intensity is high is indicates a major disadvantage of the heuristic. The hypothesis here is that the neighbouring pixels in a row which have produced very inaccurate depth results also have similar intensity-depth profiles. Because the first in the row has been poorly fitted, perhaps being caught in a local minima or maxima in the fitting algorithm, the remainder of the row which it is assumed share similar profiles have also been poorly fitted, due to poor initial parameters being used. In addition to this, there is also the possibility that these pixels have a profile 'shape' that is difficult to accurately fit the Gaussian model to, and so fairly meaningless results are produced for all these pixels.

An interesting general observation of both depth-maps, which is again visually clear from the error images, is that even where the accuracy across each respective depth level (particularly in the first implementation) is far from perfect and somewhat inconsistent, there remains a very clear distinction in average intensity between the two depth levels. This can be seen in both the depth-maps and the error images, and importantly, the difference is particularly noticeable at the boundary. This is an interesting observation because it suggests that even if accuracy of *Metric 2* is inconsistent and unpredictable at the per-pixel level, the average result over the entire scene may be more useful in practice. For example, it is clear just from visual inspection that either depth-map, with some additional processing, could be used to accurately extract the boundary between the two depth levels, therefore both implementations of *Metric 2* can in this case produce good input to depth-based edge detection.

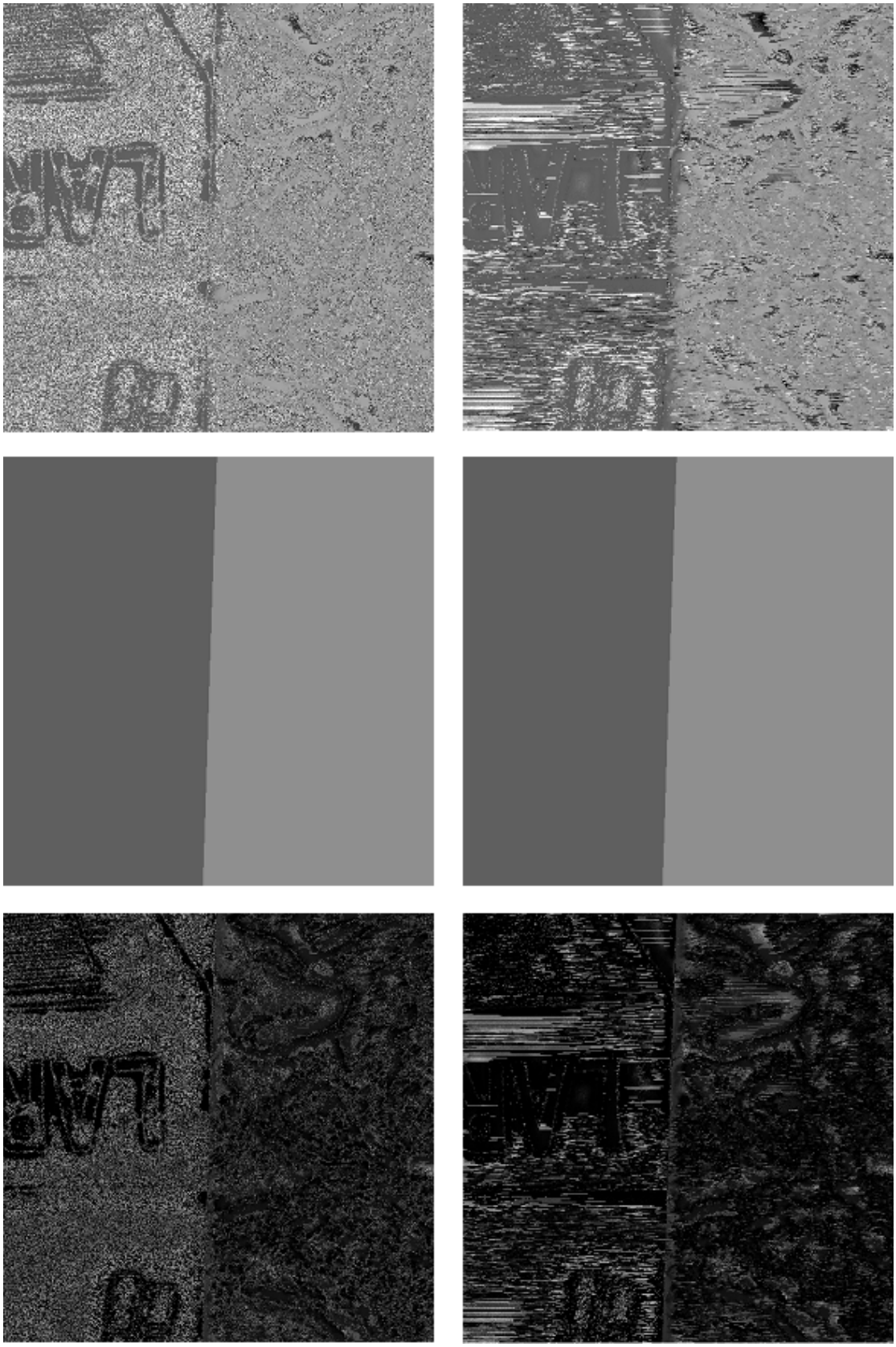


Figure 6.1.3.2.1 Raw depth-map image, ground-truth image, and absolute error image for *Metric 2*. The first implementation is the left column, the second implementation is the right column.

Again, as for *Metric 1*, we can analyse histograms of the depth-maps to perform some more quantitative analysis of the results. *Figures 6.1.3.2.2* and *6.1.3.2.3* display histograms of the depth-maps for the first and second implementation of *Metric 2*, respectively. Since these depth-maps are continuous, the depth levels have been separated into 32 depth ranges. The absolute depth range represented by each of the 32 depth ranges is detailed in *Table 6.1.3.2.4*.

The data in the histograms in *Figures 6.1.3.2.2* and *6.1.3.2.3* largely support the conclusions drawn from the qualitative analysis of the depth-maps. Both histograms are similar enough in 'shape' that the conclusions drawn from them can be drawn generally for *Metric 2*, regardless of the slightly different implementations. This in itself is good support for the validity of *Metric 2*. As with the histograms for the depth-map produced by *Metric 1*, there are clear peaks at the ranges containing the correct depth levels, that is levels 5 and 13. Here however the peaks are less Gaussian-like, most noticeably around level 13 where the peak falls off slowly over ranges 13-19, and in addition there appear to be a spike of incorrect results at range 17. It should be noted that there are also spikes of incorrect results at the end ranges. This can be explained by the fact that the fittings were limited to the range 289-1070mm for the mean parameter (representing the depth of the point). From the point of view of analysing the results, no significance should therefore be drawn from the spike of incorrect results at these ranges. These end ranges represent results which are so inaccurate that they are not worth consideration and should simply be ignored as failures of *Metric 2*.

The fact remains that both histograms show an abundance of results in the correct range for both implementations, which supports the conclusions which can be drawn visually from the depth-maps and error images. The qualitative comparison between the accuracy of results of the two implementations, that the second implementation shows generally more accurate results, is supported very well by the quantitative data in the histograms. This is because whilst the two histograms have very similar general shape, the second histogram clearly suggests that the second implementation produced a more generally accurate depth-map than the first implementation, as the peaks at the correct depth ranges show a greater frequency and each of the frequencies of the incorrect depth ranges is 'scaled down' *i.e.* each respective incorrect range has a lower frequency in the second histogram than in the first. Therefore this provides very good quantitative support of the conclusion that the second implementation provides a similar but generally more accurate depth-map than the first implementation of *Metric 2*.

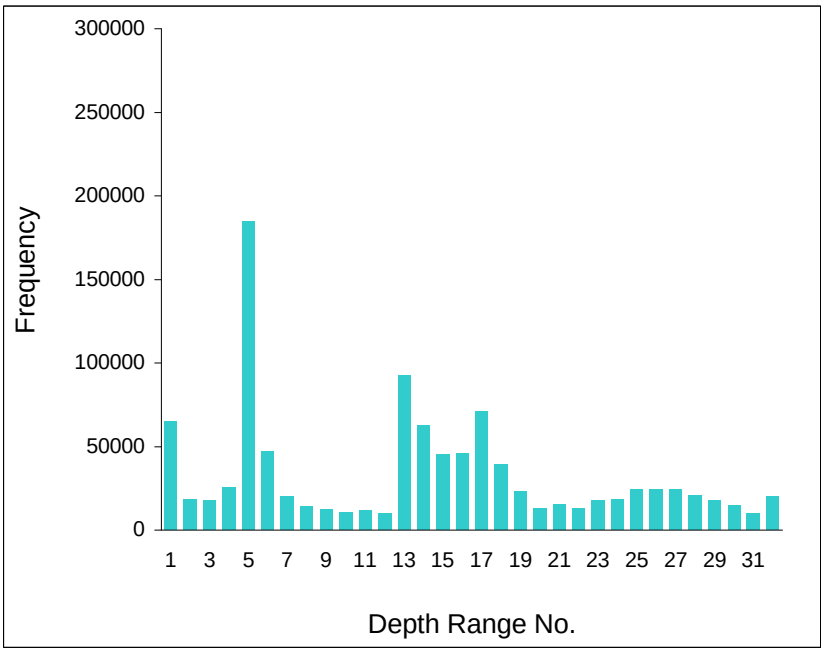


Figure 6.1.3.2.2 Histogram showing frequency of depth values (in 32 depth ranges) in the depth-map for the first implementation.

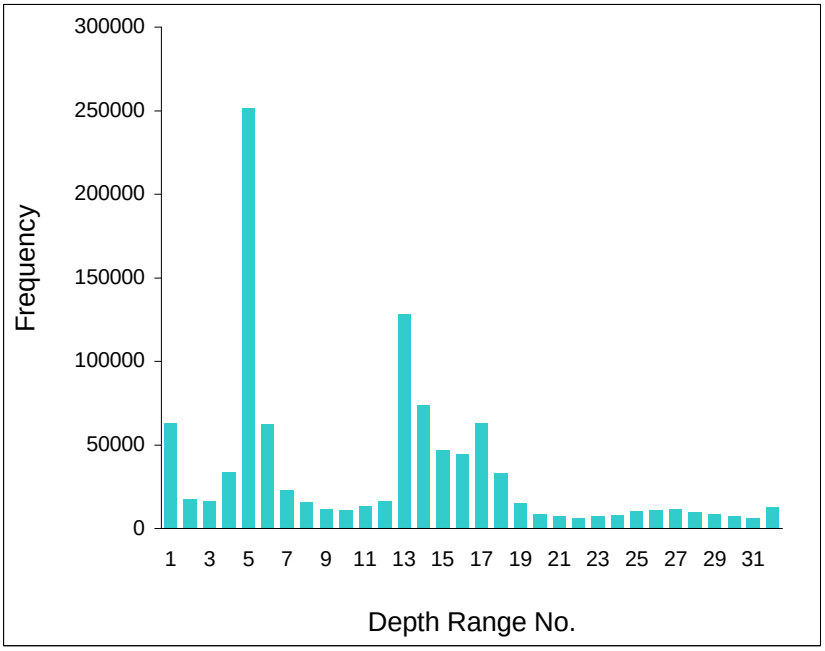


Figure 6.1.3.2.3 Histogram showing frequency of depth values (in 32 depth ranges) in the depth-map for the second implementation.

Depth Range No.	Lower Boundary (mm)	Upper Boundary (mm)
1	289.0	313.5
2	313.5	337.9
3	337.9	362.4
4	362.4	386.8
5	386.8	411.3
6	411.3	435.8
7	435.8	460.2
8	460.2	484.7
9	484.7	509.1
10	509.1	533.6
11	533.6	558.1
12	558.1	582.5
13	582.5	607.0
14	607.0	631.4
15	631.4	655.9
16	655.9	680.4
17	680.4	704.8
18	704.8	729.3
19	729.3	753.7
20	753.7	778.2
21	778.2	802.7
22	802.7	827.1
23	827.1	851.6
24	851.6	876.0
25	876.0	900.5
26	900.5	925.0
27	925.0	949.4
28	949.4	973.9
29	973.9	998.3
30	998.3	1022.8
31	1022.8	1047.3
32	1047.3	1070.0

Table 6.1.3.2.4 The absolute depth range covered by each depth range in the histograms shown in *Figures 6.1.3.2.2* and *6.1.3.2.3*. The 'correct' depth ranges are highlighted. Note the upper boundary of each range is non-inclusive.

6.2 **Results of the Proposed Method Using the Hair Dataset Scene**

The results from *Section 6.1*, using the simple test scene, provide a proof of concept for the method and a good basis for analysis of the method. In this section, the testing of the method will be extended by presenting results of the method using input data from a much more complex and 'natural' scene, which is used in [21, 22, 23] (data provided by Samuel Hasinoff, co-author of [21, 22] and author of [23]). This means that in addition to verifying that the proposed method produces meaningful results using more complex test scene data, there will be a basis for direct comparison of results with the method(s) in [21, 22, 23] on this test data.

6.2.1 **The Hair Dataset Scene**

The scene used is the scene referred to as the *Hair Dataset* in [23]. For full detail refer to [23] p66-67. The layout of the scene is displayed in *Figure 6.2.1.1*, and the key details of the scene and the full set of images of the scene used in [21, 22, 23] are summarised below.

Figure 6.2.1.1 The layout of the entire 'Hair Dataset' scene used in [21, 22, 23]. Image taken from [23].

The *Hair Dataset* scene, as seen in *Figure 6.2.1.1*, consists of a wig and some flowers in front of a mostly white-coloured board which is parallel to the front of the camera. This scene is much more complex and 'natural' than the test scene used in *Section 6.1*. It contains flowers which have complex, non-planar 3D geometry, has surfaces with different types of 3D surface texture, and has more complex-shaped objects, meaning that the object edges are not straight and uniform like the edge between the two depth levels in the test scene.

The full *Hair Dataset* consists of 61*13 images, captured using 61 different focus settings across 13 different aperture settings. See *Chapter 4* for details of the input data to the method in [21, 22, 23], which will be from here on referred to as *Confocal Stereo*.

To test the proposed method on the *Hair Dataset* scene only a subset of this full image set is used. Specifically, this subset consists of the 61 images captured at the largest aperture diameter available, to minimise depth of field and therefore maximise depth discrimination between images at different focus settings (see *Chapter 3*). Note that this significantly smaller subset of the input image data required for the proposed method in comparison to the full dataset required for Confocal Stereo is an important difference between the methods.

Another modification to the original images in the *Hair Dataset* for the purposes of testing the proposed method was that, as with the test dataset used in *Section 6.1*, the original colour images of the dataset were converted to greyscale, using the same weightings given in *Section 6.1.1*. Again, this was done to simplify testing by avoiding the practicality issues of repeating the method on three separate R, G and B colour channels, instead just running the method once on a single greyscale intensity channel. Again, the fact that the proposed method produced the presented results using only one intensity channel from the input images is a consideration to make when comparing results with those from Confocal Stereo.

As with the test dataset, only a small patch of the *Hair Dataset* scene images is used to produce results. This was done for reasons of practicality, but also so that the testing of the proposed method could be focused on specific areas of the scene. For example, it is more relevant to perform analysis of the performance of the proposed method at the boundaries between different objects, than in flat areas of similar texture like the hair of the wig (for example see *Figure 6.2.2.2*), since the proposed method is expected to produce better results in the former case and poorer results in the latter case.

The analysis of results will focus on one particular patch, which will be referred to as the sample patch, shown in Figure *6.2.1.2*. Additional examples of results from other patches of the scene are also presented in *Figure 6.2.2.2*.

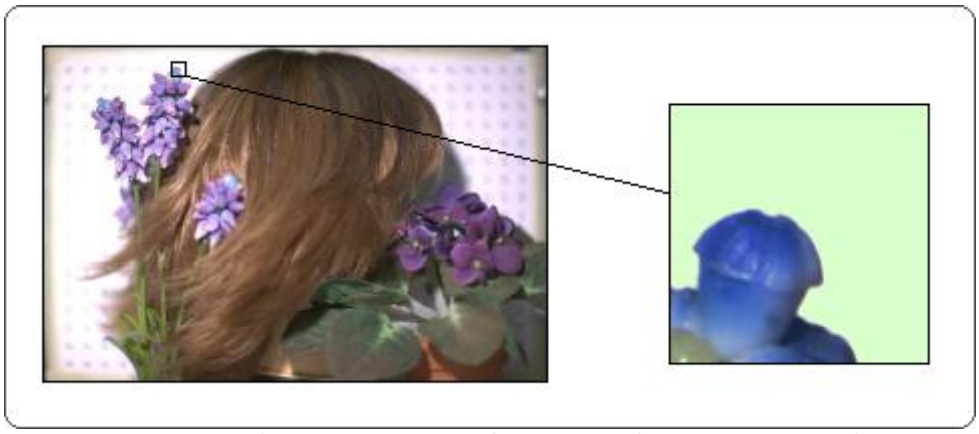


Figure 6.2.1.2 The location of the sample patch used for testing in a full resolution image of the scene. This illustrates how impractical it would be to test the method using the entire image area, which has a very large resolution.

The patch contains a flower set against the background board. The results from this sample were chosen for analysis to demonstrate that the proposed method can handle the complex geometry and texture of the flower, set against the planar background board which has fairly uniform colour and smooth geometry, simultaneously with the same set of input parameters (*Metric 1* will be used). Using this sample will test two main aims of the proposed method on the complex scene:

- Clearly and accurately separate the general depth level of the complex flower object from the plain background.
- Resolve the complex 3D surface of the flower object.

The extent to which the proposed method achieves these aims, which are quite general aims for natural complex scenes, will be a basis of analysis of the results produced from the sample of the *Hair Dataset* scene.

A final important note before results are presented is that there is no ground-truth for the *Hair Dataset* (a ground-truth was not captured for the *Hair Dataset* in [21, 22, 23]). However, where the results from the test dataset focused on the absolute quantitative accuracy of the proposed method, the analysis of results from the *Hair Dataset* sample will be more focused on the general performance of the proposed method on the complex geometry in the sample, and on relative comparison with the results produced for the sample by Confocal Stereo. For this type of analysis an absolute ground-truth is not necessary, although clearly if one were available it would provide useful additional quantitative analysis of the accuracy of the proposed method.

6.2.2 **Results**

Before presenting the results of the method using the *Hair Dataset*, a few points about the implementation of the method used in this testing must be clarified.

Firstly, *Metric 1* was used to produce the results. This is because no absolute depths of the focal planes at each focus setting for the *Hair Dataset* were available. To deal with this, it was assumed that the focal plane depths increased at a constant rate from focus setting 1 to focus setting 61, *i.e.* the depth between each consecutive focal plane is equal.

This assumption is acceptable for *Metric 1*, as it simply means that the symmetry factor in the ranking function (see *Chapter 5*) is done using matching 'pairs' of samples either side of the candidate sample. Since symmetry is only evaluated over the interval width, and assuming the interval width is suitably low relative to the number of samples (for example less than a third), then over the interval width it is reasonable to expect this assumption of uniformly distanced focal plane depths to be a good approximation of the actual focal plane depths.

However, for *Metric 2*, which attempts to fit a model precisely to the entire sample set, this assumption about the focal plane depths is inappropriate. Over the entire sample set, it is likely to result in a significant skew of the data which will have a significant effect of the fit of the model to the sample set.

Another reason why *Metric 1* is preferable to *Metric 2* as a test of the proposed method on the *Hair Dataset* sample is that *Metric 1* evaluates depth in a discrete, relative manner by selecting the 'best focus' sample from amongst the samples, whereas *Metric 2* evaluates depth absolutely and on a continuous scale. A relative depth-mapping is again the only appropriate option since the depths of the focal planes at each sample are unknown, and similarly a continuous depth-mapping would be meaningless if the absolute focal plane depths of the samples are unknown.

As should be expected, Confocal Stereo gives a discrete depth-mapping of the *Hair Dataset*, therefore yet another reason to use *Metric 1* is that the results of the proposed method using *Metric 1* can be directly compared to the results of Confocal Stereo on the *Hair Dataset*.

The images in *Figure 6.2.2.1* show an entirely focused greyscale image of the sample patch area of the scene showing all the detail in the patch, examples of different focus settings from the image set of the patch used as input to the proposed method, the best depth-map of the patch produced by Confocal Stereo, the depth-map produced by the proposed method (*Metric 1*) and the absolute error image between these depth-maps.

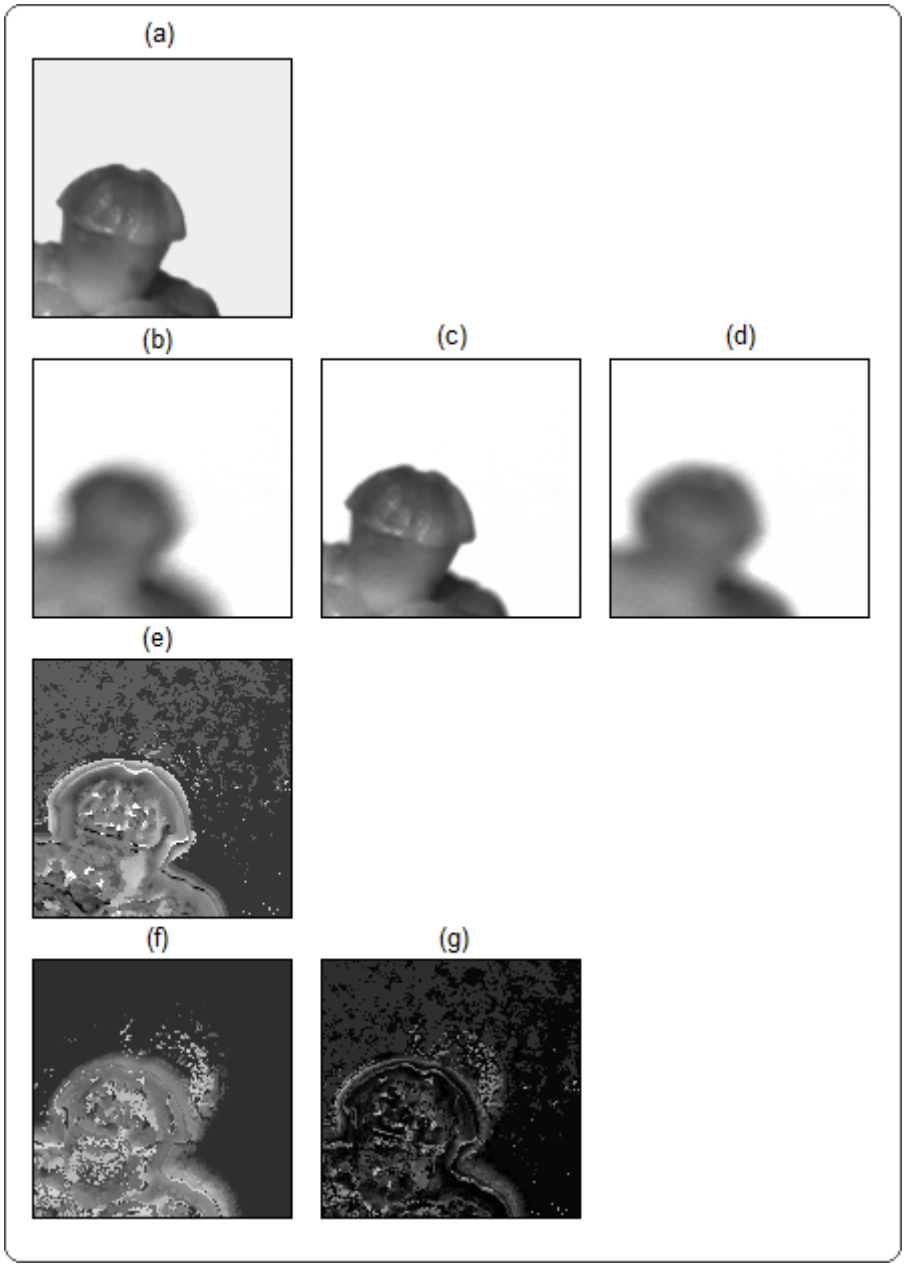


Figure 6.2.2.1 *(a)* All-focus greyscale image of the sample patch *(b)* Image from the input set at focus level 5 *(c)* focus level 30 *(d)* focus level 45 *(e)* Best depth-map produced by Confocal Stereo *(f)* depth-map produced by *Metric 1* with *interval width* = 10, a = 0.2, b = 3.0, c = 4.0 *(g)* Absolute error image between *(e)* and *(f)*, zero error indicated by black, max error of 60 focus settings indicated by white.

As mentioned, *Figure 6.2.2.2* provides two additional examples of patches from the scene, and compares the results with the results of Confocal Stereo, to give some additional informal support for the analysis and comparison of the methods to follow. For *Patch 1*, a flat area of hair, the *Metric 1* resolves the 'general' depth level quite well, but is not as successful (assuming the Confocal Stereo depth-map is more accurate) at resolving the finely detailed geometry of some of the hair strands (see error image). On the other hand, for *Patch 2* the 'general' level of depth is resolved by *Metric 1* quite differently (inaccurately) compared to Confocal Stereo, but *Metric 1* is at least successful in identifying the objects broadly, *i.e.* the stalk can be identified distinctly from the hair background in the depth-map.

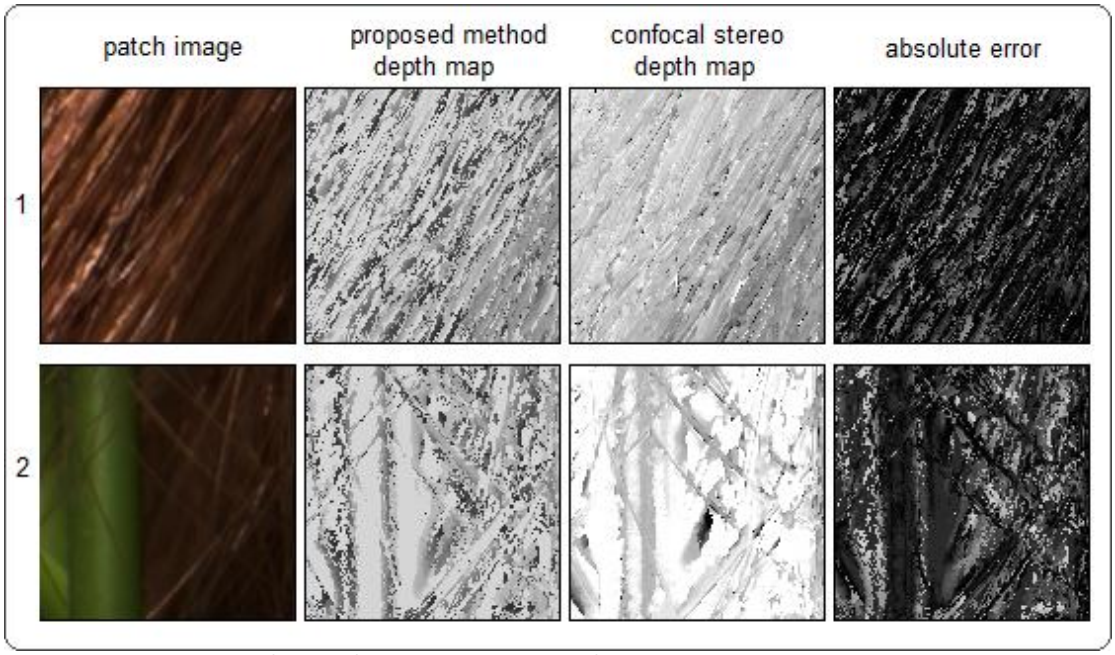


Figure 6.2.2.2 Examples of results from two other patches of the *Hair Dataset* scene, to give additional support to the analysis of results. *Metric 1* used with parameters *interval width* = 10, a = 0.2, b = 3.0, c = 4.0.

Bringing the focus of the analysis to the sample patch, the most interesting images in *Figure 6.2.2.1* are (*e*), the best depth-map produced by Confocal Stereo, (*f*), the depth-map produced by the proposed method, and (*g*) the error image between these two depth-maps.

The depth-map (*f*) was produced using the same parameters for *Metric 1* that were found to produce the best results for the test dataset in *Section 6.1.2*. That is, an interval width of 10, and weighting parameters a = 0.2, b = 3.0, c = 4.0. These input parameters were chosen in order to produce supporting evidence for the hypothesis in *Section 6.1.2* that the parameters should produce good results (for *Metric 1*) in general, and are not specific to a particular scene or type of scene.

In terms of the two main aims for testing the proposed method on the *Hair Dataset* sample patch, defined in *Section 6.2.1*, the depth-map (*f*) shows visually encouraging results. The first aim, that the depth-map should show a clear and accurate distinction in depth level between the flower object and the plain background, is clearly met by the depth-map produced. The edges of the flower object against the background are clearly defined in (*f*), and it can be seen from visual comparison with (*a*) that the flower object could easily be separated from the background using the depth information in (*f*), even with minimal or no post-processing of (*f*).

An important note to make here is that the depth level of the background in (*f*) is clearly inaccurate. Indeed, the depth-map is such that darker intensities indicate an object which is nearer to the camera (*i.e.* depth value increases as intensity increases). Therefore (*f*) appears to show the background being in front of the flower. However, this inaccuracy is not a concern, as the background is never in good focus in the input images using the widest aperture (it does not fall within the range of focal planes), and is very uniform in colour and texture. Such a surface is a known failure case for the proposed method, as the optical theory it is based on requires the defocus blurring of complex textures to reveal depth information (see *Chapter 5*). However, for the purposes of identifying that objects are broadly on different depth levels (*i.e.*

separating the flower from the background) all that is required is consistency in the failure case, as is seen in (*f*). The proposed method, particularly *Metric 1*, is tailored to provide such consistency since depth can be estimated in a deterministic way, meaning that for similar input (*i.e.* every scene point in the background) a consistent output is expected.

The second aim of the test was to verify that the complex 3D surface geometry of the flower can be meaningfully resolved. Since no ground-truth is available for the *Hair Dataset*, it is impossible to verify the accuracy of the depth-map (*f*) in an absolute quantitative manner. A qualitative analysis of (*f*) by visual comparison with (*a*) shows that the major geometrical features of the flower, particularly the edges, are identified in (*f*). However, the reconstruction of the surface of the flower is not smooth, and it is difficult to visually evaluate how accurately the changing depth of the flower surface has been resolved in (*f*). In order to perform a better analysis of (*f*) in terms of the fulfilment of the second aim, it is necessary to do a quantitative comparison of (*f*) and (*e*), which is the best depth-map of the sample produced by Confocal Stereo in [21, 22, 23].

The absolute error image (*g*) between the depth-maps produced by the proposed method using *Metric 1* and Confocal Stereo for the sample patch, shows absolute error as intensity. Black indicates an error of zero, whereas white indicates the maximum error of 60 focus settings. The error image (*g*) is a visual representation of the quantitative relative difference in results between the proposed method and Confocal Stereo. It can be seen from the generally low intensity of (*g*) that the depth-map *(f*) produced by *Metric 1* is similar to the depth-map (*e*) produced by Confocal Stereo. In particular, the error is very low at the edges of the flower, and there is also generally low error with several patches of almost no error across the surface of the flower. Assuming that depth-map (*e*) is approximately accurate, this provides some quantitative evidence that the proposed method has indeed resolved the surface of the flower generally accurately. This assumption that (*e*) is approximately accurate is also the basis of comparison between (*e*) and (*f*).

It is reasonable for comparison of the two methods to assume that depth-map (*e*) is more accurate than (*f*). Note that it would be inappropriate to explicitly assumed that (*e*) is the ground-truth of the patch, *i.e.* Confocal Stereo produced a perfect depth-map of the patch, as this is obviously not the case and would negate any sensible comparison of the results. However, it is reasonable to assume that (*e*) is sufficiently accurate that the proposed method should aim to produce a depth-map which is roughly as accurate, *i.e.* very similar to (*e*). This is because Confocal Stereo is expected to be more absolutely accurate than the proposed method, for several reasons.

Firstly, Confocal Stereo required significantly more input data to produce (*e*) than the proposed method required to produce (*f*), since the former requires images over 61 focus settings and 13 aperture settings, whereas the proposed method requires only a subset of these images across one aperture setting. In fact, by utilising the ability of the proposed method to estimate depth between focus settings using *Metric 2*, it is even possible to further reduce the number of input images by taking a subset of the focus settings, without necessarily losing depth resolution in the resulting depth-map. Related to this is the fact that (*e*) was produced using all 3 RGB channels from the input images whereas (*f*) was produced using only a single greyscale intensity channel. This alone effectively means dividing the amount of input data by 3 when using the proposed method (although it does not necessarily have to be done if the extra data redundancy of the three channels is wanted).

The fact that (*f*) is similar to (*e*), despite the significantly lower amount of input data used to produce (*f*), indicates a major advantage of the proposed method over Confocal Stereo. Indeed, it strongly suggests that much of the input data used for Confocal Stereo is redundant, since (*f*) was produced by the proposed method using a much smaller subset of the data used to produce (*e*). This is a particularly important point when it is considered that no explicit optimisation of *Metric 1* was performed to produce (*f*), and indeed, that *Metric 1* or *Metric 2* (if the data had been appropriate) could have been optimised to produce potentially even more accurate depth-maps than (*f*).

The focus of this section is on presenting and comparing the results from the proposed method, and Confocal Stereo, on the *Hair Dataset* sample patch. For further and more general discussions and comparisons of the strengths and limitations of the proposed method and Confocal Stereo, refer to the discussion of Confocal Stereo in *Chapter 4* and the general concluding discussion of the strengths and limitations of the proposed method in *Chapter 7*.

6.3 **Evaluation of Results**

Having presented and analysed various results of the proposed method and compared them with results of similar depth acquisition method [21, 22, 23], this chapter will be concluded by evaluating the strengths and limitations of the method in the context of the presented results and analysis from all previous sections.

6.3.1 **Comparison of *Metric 1* and *Metric 2***

The comparison of results of *Metric 1* and *Metric 2* must be placed in context. Thought the two metrics are based on the same underlying optical theory, they actually take quite different approaches to solving the depth acquisition problem. From the same set of samples of a scene point, *Metric 1* takes a discrete, search-based approach to finding the sample closest to the true in-focus depth of the point, and *Metric 2* fits a model to the sample set to estimate the true depth of the point absolutely on an effectively continuous scale.

The two approaches have intrinsic strengths and limitations which make them suited to different applications in practice. Therefore, an absolute and direct comparison of results from the two metrics without any application context has little value. However, it is worthwhile to evaluate the metrics against each other by identifying and comparing their respective strengths and limitations, and then comparing their practical effectiveness in real applications with respect to these strengths and limitations.

As a basic starting point, a direct comparison of the results presented in this chapter from *Metric 1* and *Metric 2* gives the obvious conclusion that, for this test scene, *Metric 1* produces results which are in the absolute sense more accurate. This can be seen quantitatively in the distribution of results in the best depth-maps produced by *Metric 1* and *Metric 2* (*Figures 6.1.2.3.1* and *6.1.3.2.1* respectively), and qualitatively from visual comparison of the depth-map and error images. This apparent difference in accuracy, however, cannot be assumed to be general, and must be analysed in context of the test scene.

Firstly, there is the issue of the accuracy of the test scene ground-truth image. The term 'accuracy' has a slightly different definition for *Metric 1* and *Metric 2*. For *Metric 1*, it is appropriate to assume that within the depth resolution of the sample set, the ground-truth is entirely accurate, with the exception of some accounted-for edge pixels since it is produced manually. This is because within the depth resolution of the input images of the test scene, it is clear visually which image (intensity/depth sample) shows both depth levels in best focus.

For *Metric 2* however, there is a different level of precision on the depth resolution (*i.e.* it is effectively continuous) therefore it is inappropriate to consider the accuracy of results without reference to the estimated error of the ground-truth, on the continuous scale. This is the fundamental reason why direct comparison of results from both metrics is flawed, however it is also the basis for analysing the strengths and limitations of both metrics, and the applications they are suited for.

Clearly, *Metric 1* produces results which are generally very accurate on the test scene. The best depth-map produced shows an overwhelming majority of results at the correct depth levels, from analysis of the histogram, error image and visual inspection of the depth-map itself. The test scene contains two very smooth, completely fronto-parallel surfaces, and this is precisely the type of scene that *Metric 1* is suited for. Since *Metric 1* operates with a discrete depth resolution, defined by the depth resolution of the set of input images, it is more appropriate for scenes with smooth surfaces which are largely parallel to these sample depth levels. It will not deal well with surfaces which are not (significantly) fronto-parallel, even where the depths of these surfaces transitions smoothly between the sampled depth levels, as these smooth transitions cannot be resolved very well with only a discrete set of possible depths, *i.e.* there will be one or more sharp edges between depth levels where in fact a smooth gradient should occur in the depth-map.

This is the most important strength/limitation trade-off with *Metric 1*. For the appropriate type of scene, *i.e.* smooth, mostly fronto-parallel surfaces without complex geometry in the depth dimension, it is likely to be more accurate than *Metric 2* whilst also being far less computationally complex (*i.e.* faster in practice). However, the limitation is in the strict restrictions on the scene. Not only must the conditions mentioned be satisfied, but the input images must contain a sampling of depth which is sufficient in resolution to accurately capture the depths of the scene surfaces.

The strengths and limitations of *Metric 2*, however, are biased in the opposite way to those of *Metric 1*. The strength of *Metric 2* lies in its ability to deal with complex geometry, within the general limitations of the optical model. Whilst very complex geometry which varies quickly, non-smoothly or without pattern is inherently difficult to deal with (though this is a limitation of passive depth acquisition methods in general [23]), *Metric 2* at least has no theoretical barrier to resolving smooth depth-maps of surfaces which are not fronto-parallel. This is particularly true as surfaces become smoother or have shallower depth gradients, for example surfaces which are angled less than 45 degrees from the image plane. In addition to this, the depth resolution of *Metric 2* is not defined by the depth resolution of the set of input images, although of course there is a natural correlation between a higher input resolution and greater accuracy of results (see *Chapter 5*).

Of course, the trade-off for the additional flexibility in the types of scene *Metric 2* can deal with is that it is liable to be far less accurate relative to running time than *Metric 1*. Because the process of fitting a model to the sample data is more computationally complex than a simple search over the samples using a ranking function, and the running time in practice is lengthened by the nature of the fitting

algorithm (where a longer running time generally correlates to greater accuracy), there is a significant increase in practical running time of *Metric 2* compared to *Metric 1*, which is a severe limitation of *Metric 2*. In addition to this the inherent difficulty of general model fitting to a dataset, as opposed to a more targeted search over the dataset relying on strict assumptions and heuristics, as in *Metric 1*, means that in general *Metric 2* is likely to be less accurate than *Metric 1*, assuming an appropriate depth resolution is present in the input images. This is in spite of, and not relative to, the significantly longer running time of *Metric 2*.

The complementary strengths and limitations of both metrics are compounded by general strengths and limitations of the method itself. These strengths and limitations were covered in detail in *Chapter 5*, but are worth mentioning explicitly here. A common strength of both metrics is that both are per-pixel, meaning that depth-map resolution matches input image resolution, and it is possible to capture fine-grain detail in geometry. As discussed in previous paragraphs, this is a particular strength of *Metric 2* which is better suited to dealing with complex geometry in the depth dimension than *Metric 1*. However, it is also relevant to *Metric 1*. For example the per-pixel approach allows the edges of surfaces to be more precisely resolved than, for example, a window-based DFF technique.

While *Metric 1* suffers from the depth-resolution issues common to search-based DFF techniques, *Metric 2* overcomes this limitation set by the input images by employing a method which is a hybrid between DFF and DFD. However, both metrics suffer from the limitations and restrictions placed on the scene which follow from the assumptions of the optical model used. This is particularly true since the methods are per-pixel, as strict assumptions about the composure of the scene must be made, such as largely fronto-parallel and smooth surfaces, complex intensity textures on surfaces, and a feature becoming increasingly global as the region of the scene examined becomes larger. Such assumptions are necessary because the method relies on all light rays from a region of the scene being summarised by one intensity value at a pixel, so strict assumptions must be made about the light sources in order to infer useful information from this value. At least in a window-based DFF technique, the evaluation of a region of the scene can be done on a case-by-case basis without being forced to make such generalised assumptions about the entire scene.

The strengths and limitations of the general method, and of the individual metrics provides a strong indication of which types of application the method, and each metric, is best suited for in practice. To give the discussion of practical application some context, *Chapter 1* provides detailed discussion of the types of applications that depth acquisition techniques can ultimately be applied to, and indeed the method presented in this work has the potential to be applied in practice to solve real computer vision problems.

However, in order for a reasonable discussion it is necessary to assume a practically viable input phase, that is the capture and calibration of image data for use by the method. Clearly, the manual capture and calibration of images for the test scene in this work is far too cumbersome and time-consuming as a process to be viable in practice. However, it is not difficult to see how hardware and software could be streamlined to automate this process and remove these difficulties; therefore for the remainder of this discussion it will be assumed that there is no barrier to practical use of the method because of difficulties in the input phase.

Generally speaking, the method presented in this work, using either metric, is appropriate for any application where a fine grain of resolution is required in the plane orthogonal to the depth dimension. This is because the method is per-pixel, offering a

depth-map at equal resolution to the input images. A good example of an application which can benefit from a high resolution in this plane is depth-based edge detection.

Depth-based edge detection is the extraction of edges in a scene where there is a difference in depth on either side of that edge only in depth. Many traditional edge detection techniques rely on intensity differentials to extract edges from scenes. Though intensity differentials are generally a good cue to a difference in depth at an edge, this is not necessarily always the case, and conversely if looking only for edges where there is a difference in depth, an intensity based edge detection algorithm can easily give false positives, for example in patterns on a planar surface.

Depth based edge detection benefits from a per-pixel depth acquisition approach as this allows for arbitrarily complex (down to a pixel resolution) edge shapes to be identified on a per-pixel basis, and allows pixels to be processed individually with no reference to their neighbourhood, so if desired only a subset of candidate 'edge pixels' can be processed without having to include the processing of all surrounding 'surface pixels'.

In a depth-based edge detection application, the choice of metric would be based on the limitations each metric imposes, in the context of the application goals and the scene.

Metric 1 would be appropriate for quickly identifying different depth levels in the scene, essentially segmenting the scene into a discrete set of depth levels which depend on the depth resolution of the set of input images. In practice, this could be very useful for extracting a particular object from a scene where it is known that the object is set a reasonable distance (relative to the depth sampling resolution) in front of or behind other objects in the scene. Here, we are less interested in the precise absolute depth of the object and more in its depth relative to the rest of the scene so that it can be extracted. *Metric 1* is therefore appropriate because it provides a much faster and probably more accurate way of solving this problem than *Metric 2*.

Metric 2, on the other hand, would be more appropriate where an object's borders must be extracted for edge detection, but the edges of the object are in focus at multiple different depth levels of the sample set, *i.e.* the object surface is not fronto-parallel. In this case, the simple depth level segmentation described above would not be an effective solution (assuming the depth resolution of the sample set is low relative to the difference in depth between different edges of the object), and a more general description of the changing depth of the object surface would be required. Here, the absolute and continuous depth-maps produced by *Metric 2* would be a far more appropriate basis for a solution.

Indeed, the example of depth-based edge detection reveals a general difference between the metrics when applied in practice. *Metric 1* is more appropriate for applications which require knowledge of the relative depths, or even just simple depth order, of various surfaces within a scene. *Metric 2* is more appropriate for any application where a general depth-mapping of a scene is required, so that complex surfaces spanning multiple levels of depth sampled by the input images can be meaningfully extracted and analysed.

Chapter 7

Conclusions and Future Work

To conclude this thesis, this chapter will make some general concluding remarks about the method and results presented in the context of previous work, and there will be a discussion of potential future work which was outside the scope of this thesis, but if investigated could extend and enhance the methods presented here.

7.1 **Conclusions**

A proof of concept of the methods presented in this work, with thorough discussion of results and analysis of strengths and limitations, is given in *Chapter 6*, so will not be repeated here. This section will instead make some general concluding remarks about the method in the context of the related work discussed in *Chapter 4*.

The main objectives of the method presented in this work can be summarised as follows:

- To use a set of input images captured with an off-the-shelf conventional camera system, with pre-defined settings, as input. Within this objective, there is an implied emphasis on minimising the number of input images used.
- To produce depth-map output with, at a maximum, the same resolution in the image plane as the input images.
- To perform all evaluation of the input images on a per-pixel basis, such that the depth of an arbitrary scene point can be evaluated using the single corresponding pixel in each input image, with no dependency on other pixels, and no specific dependency on pre-computed global scene parameters. In other words, the evaluation of a scene point can be completely isolated so that no Image Processing difficulties are introduced, and parallel processing of results is trivial.

In conclusion, it is clear that given the assumptions about the scene are held, and given the application is appropriate (see *Chapter 6*), the method presented in this work fulfils these objectives. Furthermore, through use of the two different metrics, the method is capable of producing a discrete, relative depth-map or a continuous, absolute depth-map.

The most appropriate recent method to compare the proposed method against is the Confocal Stereo method presented in [21, 22, 23]. A comparison of results from both of these methods on a sample patch of a scene is given in *Chapter 6*, and here a general comparison of the strengths and limitations of each method will be discussed.

Firstly, as the proposed method has been shown to work accurately using input images where only the focus setting is varied, as opposed to varying both focus and aperture settings in [21, 22, 23], the proposed method demonstrates far less data redundancy and represents a much more practical, and faster, way of acquiring image input. This relatively low amount of required input data in relation to similar methods is a major success of the proposed method.

Secondly, strongly related to the first point, the proposed method provides the possibility, through *Metric 2*, of extending beyond the limitations of traditional DFF so that depth can be inferred on an effectively continuous scale. Again, this represents an advantage over similar methods such as the method in [21, 22, 23], which despite relying on a significantly larger amount of input data, cannot offer more than a discrete depth evaluation, limited to the focal plane depths of the set of input images. This is strongly linked with the first point in the previous paragraph because again, it introduces a further opportunity to reduce the number of input images used, which was a major objective of the proposed method.

Although the initial objectives are fulfilled by the proposed method under certain conditions, the lack of generality of the method is its biggest limitation. For example, though the method is based on a very similar optical model to that used in other recent work such as [21, 22, 23, 28], the way in which the predictions of the model are generalised so that they can be applied on a per-pixel basis, rather than a more specific, case-by-case window-based approach as used by [28] or a more directly data-driven (albeit with much more input data) approach as used in [21, 22, 23], means that much of the analytical potential of the optical model is sacrificed in favour of minimising the input data and taking the per-pixel approach. As a result, it is likely that in practice, the results produced by the method presented in this work will have greater inherent limitations on accuracy compared to those of [21, 22, 23] and [28]. Note that this statement is intended to be very general, and made without consideration to the differences in input data.

A particular example of where the proposed method is inherently flawed is the evaluation of very fine grain complex scene detail. For example, it is claimed in [21] that the depth of a single strand of hair can be resolved on a per-pixel basis by the Confocal Stereo method, with a theoretical underpinning using a broadly similar optical model as is used in this work. However, this is because the Confocal Stereo method can work without any implicit evaluation of neighbourhood pixel intensities. The method in this work does employ such an implicit evaluation of the region around a scene point, and the model breaks down considerably if this neighbourhood is not (practically) on the same depth plane. A single hair set a significant distance in front of any other scene object, for example, is a good example of where this flaw in the model would affect the accuracy of results considerably. It is likely that the proposed method would not produce accurate results for such complex geometry (see *Figure 6.2.2.2*). Related to this point, the optical model employed in the proposed method will also struggle with other features of complex scenes such as shadowing, occlusion of objects, and reflections of light, although it must be emphasised that these are fundamental barriers to passive depth acquisition techniques using only light intensity data as input.

As a concluding remark, it is clear that under certain conditions, all of which are reasonable in a practical application where the scene domain is known and controlled, the proposed method fulfils its objectives. Though the trade-offs made to reduce the amount of input data and maximise the depth-map resolution place certain fundamental limitations on the types of scene the method can deal with, it is nonetheless true that the proposed method, given an appropriate scene, is capable of producing depth-maps of acceptable accuracy which are absolute and continuous, or if the application requires it, relative and discrete.

7.2 **Future Work**

Following naturally from the previous section is a discussion of future lines of investigation which could be undertaken to extend or enhance the method presented in this work. Although the research, development and results produced provided a satisfactory proof-of-concept for the method, and sufficient data on which to analyse and make reasonable conclusions about the method, there is no doubt that given a wider project scope there would be many opportunities to extend the work presented in this thesis to enhance both the absolute accuracy and range of application of the method. Some of the most prominent of these possible extensions are as follows:

- **Investigation of the effect of aperture setting**

Although from the optical model it was assumed that the maximum aperture setting of the camera system should be used for image capture, due to the maximum depth discrimination this facilitates, this assumption was not verified experimentally. It is hypothesised by the model that reducing the size of the aperture could have a 'smoothing' effect on the samples, and in practice is it entirely possible that this could make it easier to accurately fit a model to the set of samples in *Metric 2*. A future investigation could evaluate the trade-off between the positive effect of this smoothing and the negative effect of depth discrimination loss on the input data, and ultimately on the accuracy of results.

- **More explicit and rigorous modelling**

Though the results in this work provided a proof-of-concept for the method, there were a lot of assumptions placed on the modelling of the relationship between intensity and focal plane depth in the sample set for a scene point. For example, in *Metric 1* the weighting parameters used were manually selected rather than optimised. Such an optimisation could be driven by data or optical theory, and indeed this extends to optimising the ranking function that the search is based on in general.

For *Metric 2* there are several factors, again each representing a large scope of additional work, which could be investigated to optimise the accuracy of results. An obvious starting point is an extension of the model of the functional relationship between focal plane depth and pixel intensity. It is explicitly conceded in this work that using a single Gaussian-type function as a model makes many general assumptions about the data (which are likely to hold in practice). A more rigorous extension to the model, based on optical theory, would be to include a mixture of multiple Gaussian-type functions, with a common mean at the in-focus depth, to reflect the multiple peaks and troughs of intensity we expect to see as the focal plane depth approaches the depth of the scene point. Assuming we account for the extra difficulty of fitting a more complex model, such a model offers the general enhancement of a better fit to the data, *i.e.* more absolute accuracy of the mean and therefore the depth of the point, as well as extended capabilities of *Metric 2* such as the ability to better predict the intensity of the scene point at arbitrary distance out-of-focus. In practice this capability could be useful in applications such as computerised post-exposure re-focusing of the scene.

- **Consideration of non-geometric optical effects**

A very broad extension to the method would be to include the consideration of non-geometric optical effects, such as diffraction at the edges of the aperture, in the optical model. Such effects have been used successfully as defocus cues in [20] for example.

Although these effects have a minor impact on intensity compared to geometric optical effects, their impact increases in importance as input image resolution increases, since we are evaluating per-pixel and minor optical effects will be more emphasised in individual pixels as resolution increases. The optical model employed in this work simply disregards the non-geometric light wave effects.

Clearly, an investigation of the impact that these effects have on intensity could result in a more rigorous and ultimately accurate model, that as previously stated would be more appropriate for very high resolution (*i.e.* above ~16Megapixel) images.

- **Post-processing of depth-maps**

In *Chapter 6*, the results presented were 'raw' depth-maps, as produced directly by both metrics. It would be an interesting extension to this work to investigate how post-processing of these raw depth-maps, using *a priori* knowledge about for example the metric itself or the scene, could be done to improve the accuracy of the depth-maps by reducing the number of anomalous results. For example, for the best depth-map produced by *Metric 1*, a simple modal filter based on the knowledge that the scene consists of smooth planar surfaces, could almost eradicate erroneous results by setting the depth value of every pixel to be the same as the majority of pixels in its neighbourhood. Obviously, such post processing could be extended to model more complex scene expectations and assumptions, and would be implemented in an application-specific way.

This would be a relatively simple way to enhance the accuracy of the method in practice, by assuming that although the accuracy of individual pixels of the depth-map may be poor, the depth-map in general is accurate. Of course, such an enhancement would contradict the per-pixel nature of the method, in particular the ability to isolate results to individual pixels without needing to evaluate neighbouring pixels, but in a practical application this may be acceptable and appropriate.

Bibliography

[01] M. Aggarwal and N. Ahuja. *On generating seamless mosaics with large depth of field*. In International Conference on Pattern Recognition, Volume 1, pp588-591, 2000.

[02] N. Asada, H. Fujiwara, and T. Matsuyama. *Edge and depth from focus*. International Journal of Computer Vision, 26(2):153-163, 1998.

[03] S.D. Babacan, R. Molina, A.K. Katsaggelos. *Variational Bayesian Blind Deconvolution Using a Total Variation Prior*. IEEE Trans. on Image Processing. 18(1):12-26, 2009.

[04] S. D. Babacan, R. Molina, and A. K. Katsaggelos, *Total variation blind deconvolution using a variational approach to parameter, image, and blur estimation*, in EUSIPCO, Poznan, Poland, Sept. 2007.

[05] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[06] P. Campisi and K. Egiazarian, *Blind image deconvolution: theory and applications*, CRC press, 2007.

[07] W. Cathey and R. Dowski. *A new paradigm for imaging systems*. Applied Optics 41, pp1859-1866. 1995.

[08] T. Darrell and K.Wohn. *Pyramid based depth from focus*. In Proc. Computer Vision and Pattern Recognition, pp504-509, 1988.

[09] E.L. Dereniak, T.D. Dereniak. *Geometric and Trigonometric Optics*. Cambridge University Press, 2008.

[10] R. Dowski and W. Cathey. *Single-lens single-image incoherent passive-ranging systems*. Applied Optics 33, pp6762-6773. 1994.

[11] J. Ens and P. Lawrence. *An investigation of methods for determining depth from focus*. IEEE Trans. Pattern Anal. Mach. Intell., 15:97-108, 1993.

[12] H. Farid and E. P. Simoncelli. *Range estimation by optical differentiation*. Journal of the Optical Society of America 15, pp1777-1786. 1998.

[13] P. Favaro, A.Mennucci, and S. Soatto. *Observing shape from defocused images*. International Journal of Computer Vision, 52(1):25-43, 2003.

[14] P. Favaro and S. Soatto. *A geometric approach to shape from defocus*. IEEE Trans. on Pattern Analysis and Machine Intelligence, 27(3), pp406-417, 2005.

[15] E. Fenimore, and T. Cannon. *Coded aperture imaging with uniformly redundant rays*. Applied Optics 17, pp337-347. 1978.

[16] R. Fergus, B. Singh, A. Hertzmann, S.T. Roweis and W. Freeman. *Removing camera shake from a single photograph*. ACM Transactions on Graphics, Proc. SIGGRAPH 2006, 25, pp787-794. 2006.

[17] B. Girod and E. Adelson. *System for ascertaining direction of blur in a range-from-defocus camera.* In US Patent No. 4,965,422. 1990.

[18] B. Girod and S. Scherock. *Depth from focus of structured light*. In Proc. SPIE, vol. 1194, Optics, Illumination, and Image Sensing for Machine Vision. 1989

[19] P. Green,W. Sun,W.Matusik, and F. Durand. *Multi-aperture photography*. In Proc. ACM SIGGRAPH, 2007.

[20] A. Greengard, A Schechner and R. Piestun. *Depth from diffracted rotation*. Optics Letters 31, pp181-183. 2006.

[21] S. W. Hasinoff and K. N. Kutulakos, *Confocal Stereo*. International Journal of Computer Vision, 81(1), pp82-104, 2009

.

[22] S. W. Hasinoff and K. N. Kutulakos, *Confocal Stereo*. Proc. 9th European Conference on Computer Vision, ECCV 2006, pp620-634.

[23] S. W. Hasinoff, *Variable-Aperture Photography*. PhD Thesis, University of Toronto, Dept. of Computer Science, 2008.

[24] S. W. Hasinoff and K. N. Kutulakos. *A layer-based restoration framework for variable aperture photography*. In Proc. International Conference on Computer Vision, pp1-8, 2007

[25] S. Hiura and T. Matsuyama. *Depth measurement by the multi-focus camera*. In CVPR, IEEE Computer Society, pp953-961. 1998.

[26] H. Jin and P. Favaro. *A variational approach to shape from defocus*. In Proc. European Conference on Computer Vision, volume 2, pp18-30, 2002

[27] E. P. Krotkov. *Focusing*. International Journal of Computer Vision, pp223-237, 1987.

[28] A. Levin et al. Image and *Depth from a Conventional Camera with a Coded Aperture*. In Proc. ACM SIGGRAPH, 2007.

[29] F. Moreno-Noguer, P. N. Belhumeur, and S. K. Nayar. *Active refocusing of images and videos*. In Proc. ACM SIGGRAPH, 2007.

[30] H. Nair and C. Stewart. *Robust focus ranging*. In Proc. Computer Vision and Pattern Recognition, pp309-314, 1992

[31] S. K. Nayar. *Shape from focus system*. In Conference on Computer Vision and Pattern Recognition, pp302-308, 1992.

[32] S. K. Nayar and Y. Nakagawa. *Shape from focus: An effective approach for rough surfaces*. In International Conference on Robotics and Automation, pp218-225, 1990.

[33] S. Nayar, M. Watanabe, and M. Noguchi. *Real-time focus range sensor*. IEEE Trans. on Pattern Analysis and Machine Intelligence, 18(12):1186-1198, 1996.

[34] Olympus Camedia E20p Specifications - http://www.olympus.co.uk

[35] F. L. Pedrotti, L. S. Pedrotti, L. M. Pedrotti. *Introduction to Optics*. Pearson Prentice Hall, 2007.

[36] A. P. Pentland. *A new sense for depth of field*. IEEE Trans. on Pattern Analysis and Machine Intelligence, 9(4):523-531, 1987.

[37] A. Pentland, T. Darrell, M. Turk, and W. Huang. *A simple, real-time range camera*. In Proc. Computer Vision and Pattern Recognition, pp256-261, 1989.

[38] M. Proesmans and L Van Gool. *One-shot active 3d image capture*. In Proceedings SPIE, vol. 3023, Three-Dimensional Image Capture, pp50-61. 1997.

[39] A.N. Rajagopalan and S. Chaudhuri, *A Variational Approach to Depth from Defocus*, Proc. Int'l Conf. Intelligent Robotic Systems, pp45-48. 1995.

[40] A. N. Rajagopalan and S. Chaudhuri. *A variational approach to recovering depth from defocused images*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(10):1158–64, 1997.

[41] Y. Y. Schechner and N. Kiryati. *Depth from defocus vs. stereo: How different really are they?* International Journal of Computer Vision, 39(2):141-162, 2000.

[42] M. Subbarao, T. Choi, and A. Nikzad. *Focussing Techniques*, Optical Engineering, 32(11):2824-2836, 1993.

[43] M. Subbarao and G. Surya. *Depth from defocus: A spatial domain approach.* International Journal of Computer Vision, 13(3):271-294, 1994.

[44] M. Subbarao. *Parallel depth recovery by changing camera parameters*. In Proc. International Conference on Computer Vision, pp149-155, 1988.

[45] J. Sun, S. Kang, Z. Xu, X. Tang, H.Y. Shum. *Flash cut: Foreground extraction with flash/no-flash image pairs*. In CVPR, 2007.

[46] Y-W. Tai, H. Tang, M. Brown, S. Lin, *Detail Recovery for Single-image Defocus Blur*. IPSJ Transactions on Computer Vision and Applications, vol. 1, pp1-10, 2009.

[47] J. M. Tenenbaum, *Accomodation in Computer Vision*, PhD thesis, Stanford University, 1970.

[48] Various Authors. *Special issue on blind system identification and estimation*. Proceedings of the IEEE, 86(10), 1998.

[49] M. Watanabe and S. K. Nayar. *Rational filters for passive depth from defocus*. International Journal of Computer Vision, 27(3):203-225, 1998.

[50] R.Willson and S. Shafer. *Dynamic lens compensation for active color imaging and constant magnification focusing*. Technical Report CMU-RI-TR-91-26, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA,1991.

[51] Y. Xiong and S. Shafer. *Depth from Focusing and Defocusing*. In Proc. Computer Vision and Pattern Recognition, pp68–73, 1993.

[52] N. Xu, K. Tan, H. Arora, and N. Ahuja. *Generating omnifocus images using graph cuts and a new focus measure*. In Proc. International Conference on Pattern Recognition, volume 4, pp697-700, 2004.