Brunel University London

Depth Estimation from a Single Holoscopic 3D Image and Image Up-sampling with Deep-learning

By

Akuha Solomon Aondoakaa

A thesis submitted for the degree of

DOCTOR OF PHILOSOPHY

In

Department of Electronic & Computer Engineering

Collage of Engineering, Design and Physical Sciences

BRUNEL UNIVERSITY LONDON

January 2020

ABSTRACT

3D depth information is widely utilized in industries such as security, autonomous vehicles, robotics, 3D printing, AR/VR entertainment, cinematography and medical science. However, state-of-the-art imaging and 3D depth-sensing technologies are rather complicated or expensive and still lack scalability and interoperability. The research identified, entails the development of an innovative technique for reliable and efficient 3D depth estimation that deliver better accuracy.

The proposed (1) multilayer Holoscopic 3D encoding technique reduces the computational cost of extracting viewpoint images from complex structured Holoscopic 3D data by 95%, by using labelled multilayer elemental images. It also addresses misplacement of elemental image pixels due to lens distortion error. The multilayer Holoscopic 3D encoding computing efficiency leads to the implementation of real-time 3D depth-dependent applications. Also, (2) an innovative approach of a deep learning-based single image super-resolution framework is developed and evaluated. It identified that learning-based image up-sampling techniques could be used regardless of inadequate 3D training data, as 2D training data can yield the same results.

(3) The research is extended further by implementation of an H3D depth disparity -based framework, where a Holoscopic content adaptation technique for extracting semi-segmented stereo viewpoint image is introduced, and the design of a smart 3D depth mapping technique is proposed. Particularly, it provides a somewhat accurate 3D depth estimation from H3D images in near real-time. Holoscopic 3D image has thousands of perspective elemental images from omnidirectional viewpoint images and (4) a novel 3D depth estimation technique is developed to estimates 3D depth information directly from a single Holoscopic 3D image without the loss of any angular information and the introduction of unwanted artefacts.

The proposed 3D depth measurement techniques are computationally efficient and robust with high accuracy; these can be incorporated in real-time applications of autonomous vehicles, security and AR/VR for real-time interaction.

ACKNOWLEDGEMENT

I will like to thank my supervisor, Dr Rafiq Mohammed Swash, for the great support he has given me during this research.

I will like to thank my family, specifically my father, Aondoakaa Michael Kaase, and mother, Aondoakaa Nguvan Susanna, for supporting throughout this research.

I will like to thank my fellow PhD colleagues for making my research experience an unforgettable one.

STATEMENT OF ORIGINALITY

The work contained with this thesis is purely that of the author unless otherwise stated. Except what is acknowledged, none of the work presented here has been published or distributed by anyone other than the author.

Akuha Solomon Aondoakaa

March 2020, London

Table of Contents

List of Figures

List of Tables

List of Equations

List of Abbreviations

1D	One-dimension
2D	Two-dimension
3D	Three-dimension
5D	Five-dimension
H3DI	Holoscopic 3D Image
OH3DI	Omni-directional Holoscopic 3D Image
UH3DI	Uni-directional Holoscopic 3D Image
SH3DI	Synthetic Holoscopic 3D Image
ML	Micro Lens
MLA	Micro Lens Array
LED	Light-emitting Diode
LCD	Liquid-crystal Display
CCD	Charged-coupled Device
EI	Elemental Image
VI	Viewpoint Image
VIP	Viewpoint Image Pixel
RVI	Reference Viewpoint Image
TVI	Target Viewpoint Image
RVIP	Reference Viewpoint Image Pixel
TVIP	Target Viewpoint Image Pixel
MV	Multiview
HD	High Definition
SD	Standard Definition
SISR	Single Image Super Resolution
LR	Low Resolution
HR	High Resolution
MSE	Mean Square Error
RMS	Root Mean Square
H3DDD	Holoscopic 3D Depth from Disparity
DDH	Direct Depth from Holoscopic
DFD	Depth from Disparity
SFM	Structure from Motion

HCA	Holoscopic Content Adaptation
PSNR	Peak Signal to Noise Ratio
SSIM	Structural SIMilarity
SAD	Sum of Absolute Difference
SSD	Sum of Square Difference
NCC	Normalised Cross Correlation
BT	Birchfield Tomasi
CT	Census Transform
MRF	Markov Random Field
BP	Belief Propagation
DD	Depth from Defocus
PDV	Pixel Declining Value
POV-RAY	Persistence of Vision Raytracer
SURF	Speeded-Up Robust Features
SIFT	Scale-Invariant Feature Transform
LoG	Laplacian of Gaussian
HoG	Histogram of Oriented Gradients
AR	Augmented Reality
VR	Virtual Reality
CMOS	Complementary Metal-oxide-semi Conductor
RDN	Residual Dense Network
SFE	Shallow Feature Extraction
RDB	Residual Dense Blocks
DFF	Dense Feature Fusion
CM	Contiguous Memory

CHAPTER 1: INTRODUCTION

The research overview is comprehensively documented in this chapter, outlining the research aim and objectives. The chapter layout is as follows: 1.1 PhD Research Overview, 1.2 The Research Aim and Objectives, 1.3 Research Motivations, 1.4 The Research Contributions, 1.5 Thesis Chapter Outline and 1.6 Author’s Publications.

1.1. PhD Research Overview

In research fields such as robotics, artificial intelligence, AR/VR digitisation, and medical sciences, where 3D depth information is often used either for autonomous navigation, object and gesture recognition, and analyses of bio cells. Has sparked a new wave of research that centres around effective ways of digitally recording and extracting 3D depth information [1] [2][3][4]. An example of how 3D depth information is crucial to the operational success in some of the mentioned research areas are as follows:

i. Robotics – assuming a task where a robot has to navigate itself from point A to B in a moderately busy environment. The robot will have to use 3D depth information to be able to identify the distance between itself and objects of the scene, helping it to accurately make decisions that will improve navigation and obstacle avoidance while making its way to point B (the desired destination).

ii. Content digitisation – 3D depth information is essential in the sense that, for a user to successfully digitise an asset, the volumetric content captured and reconstructed in virtual space is only made possible when 3D depth information is present.

iii. Cinematography – 3D depth, in this case, helps the cinematographer to easily segment a shot or refocus at different objects within the shot. 3D depth information can also be used in the creation of a realistic AR/VR realities.

As the world becomes more connected due to the constant growth of internet-of-things (IoT) and ultra-fast networks, there is a need for smart digital devices capable of recording high-quality contents on small digital spaces. The Holoscopic 3D (H3D) imaging system is one of the few digital devices that fall into this category, as the imaging system can record the full parallax of any given scene in a single snapshot. This leading to the research topic "*Robust 3D depth Estimation form an H3D Image*", where the implementation of a scalable viewpoint extraction technique is presented, the implementation of a learning-based image up-sampling technique and the development of an innovative H3D depth estimation technique are all presented in this thesis.

The mentioned research goal resulted in the comprehensive analysis of existing state-of-the-art 3D depth estimation techniques. This also resulted in the analysis of advanced 3D sensing systems and their recording principles. The research and analysis were done on the 3D sensing system used for this research, (Holoscopic 3D imaging system) were present limitations are addressed by specially developed H3D pre-processing techniques, documented in chapter three. However, due to high cost needed to optimise the H3D imaging system to record 3D image data of varying micro lens array specifications, the use of a cost-effective light field software known as POV-Ray is employed to render Synthetic Holoscopic 3D (SH3D) data. This facilitated and aided the deduction of the ideal micro lens array (MLA) specification for recording depth ready H3D data. The SH3D data is also used for formulating the fundamental principle behind the innovative 3D depth estimation framework and pre-processing techniques presented in chapters three to five, as the SH3D data has no distortion errors.

3D sensing systems can be used to record 3D information of any given scene successfully. These 3D sensing systems can be categorised either as Active, Stereo/Multiview, or H3D imaging systems (Integral light field imaging systems). Active imaging system employs the use of a controlled light source as part of its principle for estimating and recording 3D depth information. The Stereo and Multiview 3D imaging systems are similar as they are both modelled against the human binocular viewing system. The difference between the two being the stereo 3D imaging system uses two 2D imaging systems to record the 3D depth information of a scene, while the Multiview imaging system is an extension of the Stereo imaging system that uses more than two 2D imaging systems to record 3D depth information. H3D imaging system, on the other hand, uses a micro lens array in place of an expensive multisensory rig to record the 3D depth information of a scene. The following being one of the main reasons the H3D imaging system is used to record all 3D data presented in this thesis.

The H3D imaging "Integral Imaging" system, as mention earlier, is the most cost-effective means of recording 3D data. First proposed by Ives [5] and Lippmann[6][7], the H3D imaging principle is based on Holographic imaging, where slightly different views of a scene are projected to the viewers' eyes for the brain to reconstruct the 3D scene [8]. However, the H3D imaging system has some drawbacks that make the task of 3D depth estimation more complex to accomplish. First, the relatively small baseline between viewpoint images compared to the traditional stereo systems, primarily caused by the MLA used for recording H3D data. Secondly, the relatively low spatial resolution of viewpoint images (VI) and elemental images (EI), making the task of feature matching and cost calculation more tedious than usual. The lack of efficient pre-processing platforms for H3D data has continued to allow the following limitations to hinder the use of H3D imaging system as a primary device for recording 3D depth information in industries. To overcome the H3D image limitations, the detailed analysis of the H3D imaging system is conducted and presented in chapter two and three, identifying H3D image properties that are used to aid efficient 3D depth estimation.

3D depth estimation techniques require only image data as its primary data source (In this case, a single H3D image). Although other Active 3D depth estimation techniques that employ the use of special control light to estimate 3D depth information have higher accuracy results, the lack of flexibility options has allowed 3D techniques to grow in popularity. The following, coupled with the advancement in super microprocessors, has made the use of 3D depth estimation technique a reliable form of extracting 3D depth information from any image data. 3D depth estimation techniques can be further subcategorised into two groups, known as local and global 3D depth estimation techniques. Local 3D depth estimation techniques are computationally fast but error-prone, they employ the use of a finite neighbouring window to estimate 3D depth. Global 3D depth estimation techniques, on the other hand, generate high-quality 3D depth maps but are computationally expensive due to the complex nature, rendering it infeasible for real-time applications. It has been indicated that the 3D depth estimation techniques generally include the following four steps: feature matching or cost estimation, cost accumulation, disparity estimation and disparity optimisation [3]. However, this does not mean the 3D depth estimation framework has been standardised, therefore, choosing the appropriate framework for estimating 3D depth depends on the list of circumstances presented below:

i. Camera lens type – this could range from ultra-wide-angle (fisheye) lenses to super-telephoto lenses.
ii. Inter-axial distance – this is the distance between the two cameras of a stereo system, also known as a baseline.
iii. Data type – this ranges from, images, video, well-textured or lightly textured image data sets.
iv. Application area – this could range from application areas that require real-time estimation to application areas that do not require detailed 3D depth maps.

To link as to how current state-of-the-art 3D depth estimation techniques, designed for a stereo 2D data set, is used for estimating 3D depth information from H3D data. They heavily depend on inefficient H3D data adaptation means of converting H3D images into stereo 2D images. The current adaptation technique cannot handle complex H3D data in real-time and do not take into account other suitable viewpoint up-sampling techniques. Having identified these limitations, the design and development of two robust 3D depth estimation frameworks are presented in chapter four and five. The H3D depth from disparity framework is designed based on the proposed Holoscopic 3D content adaptation technique that is capable of handling complex H3D data with 95% less computational effort in comparison to current adaptation techniques. This H3D depth from disparity framework also takes advantage of the H3D data ability to refocus at different focal planes to extract semi segmented viewpoint images for stereo 3D depth estimation. Design of a 3D depth mapping framework is proposed to reduce the computation cost accumulated during subpixel matching and pyramid optimisation. The following framework is also designed to aid estimation of depth information from H3D images with standard stereo algorithms. However, this framework still shares the same limitations as current state-of-the-art stereo frameworks regardless of the significant improvements. This is the introduction of wanted artefacts in the extracted viewpoint images and the loss of angular information. The second 3D depth framework presented in chapter five is designed to by-pass the Holoscopic 3D content adaptation process, estimating 3D depth information directly from the H3D image. This is achieved by the implementation of a unique similarity measure that estimates 3D depth information by calculating the disparity between viewpoint pixels (VIPs) EI positions.

The resulting 3D depth information or 3D depth maps are presented either as colour, 2D grayscale image or in Pseudo form.

1.2. The Research Aim and Objectives

The main aim of this research program is to develop a robust 3D depth estimation framework that is capable of estimating 3D depth information from a single H3D image. The mentioned aim is further subdivided into objectives that can be categorised into the following groups: literature exploration in 3D depth-related fields, general data review and analysis, experimentations and implementation of efficient and practical techniques. Below is the highlighted list of aim and objectives of this research.

List of objectives

i. To carry out comprehensive research on state-of-the-art 3D imaging systems, including related sensors and applications and evaluate the principles to scope the research.

ii. To investigate 3D depth estimation techniques and evaluate them to identify suitable methods to estimate 3D depth from a single H3D image.

iii. To carry out experiments on different micro lens array (MLA) specification to identify correct MLA parameters for 3D depth estimation as well as to find out the best trade-off between MLA aperture and disparity depending on the nature of the 3D depth estimation approach.

iv. To investigate Holoscopic 3D (H3D) image processing for effective 3D depth estimation to improve computational performance with scalable H3D pre-processing techniques.

v. To design, implement and evaluate a scalable 3D depth estimation technique that is capable of estimating 3D depth directly from a low-resolution H3D image without the need of any pre-processing.

1.3. Research Motivations

3D depth has a wide application area as mention earlier, therefore researching on this topic will provide flexible and broader career options.

Vision is one of the most important senses for humans, and this contributes vastly to the ability to learn and improve on existing talents. Human visual systems are designed to slightly see a different view of the same scene in each eye, allowing the brain to perceive the world in three dimensions (3D). The use of a cost-effective Holoscopic 3D imaging system as a primary source of recoding 3D depth data is an improvement from the traditional 2D stereo systems. The fact a single Holoscopic 3D image can be used to produce stereo views, and with its ability to refocus on any object in the scene after capture, proves that Holoscopic 3D image is on the way and will eventually be the next generation imaging system. This being one of the motivations to go ahead and be amongst the first to introduce this technology to the world by proving one of the various ways a single Holoscopic 3D image can be used in multiple applications, as the imaging system undoubtedly records more information of a scene than any conventional 2D imaging system.

Computer vision plays an essential role in the improvement of computer intelligence, since humans use all five senses to help accurately understand their surroundings, it is only reasonable to develop more straightforward ways to estimate 3D depth information, taking advantage of how 3D data is stored in a coded 2D format called an "elemental image" (EI), which can be used for different purposes in 3D depth-related fields. The computer vision society common goal serves as a motivation and will very much want to be part of, and academically contribute to this society.

1.4. The Research Contributions

The original contributions of this research can be grouped into the three categories presented in Figure 1.1 below.

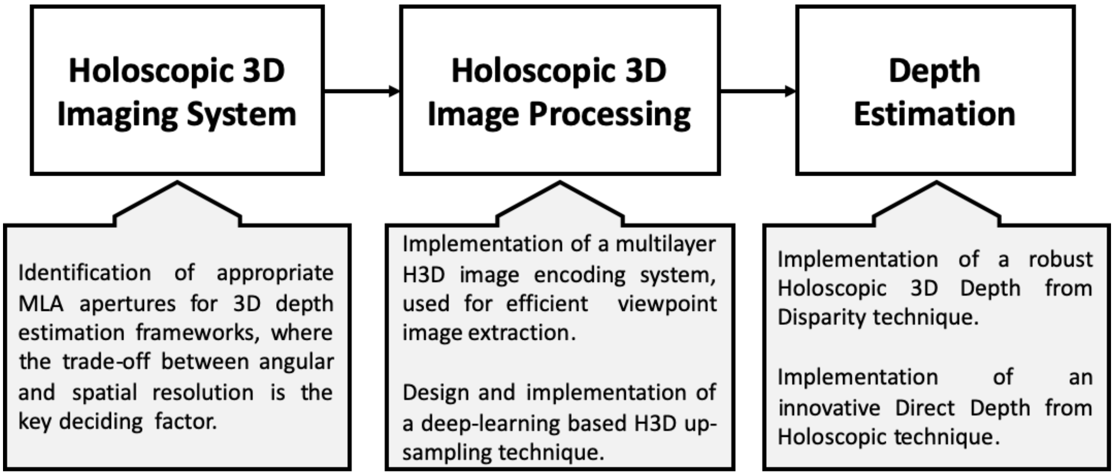

Holoscopic 3D Imaging System
Holoscopic 3D Image Processing
Depth Estimation
Identification of appropriate MLA apertures for 3D depth estimation frameworks, where the trade-off between angular and spatial resolution is the key deciding factor.
Implementation of a multilayer H3D image encoding system, used for efficient viewpoint image extraction.
Design and implementation of a deep-learning based H3D up-sampling technique.
Implementation of a robust Holoscopic 3D Depth from Disparity technique.
Implementation of an innovative Direct Depth from Holoscopic technique.

Figure 1.1- Summary of the research contributions and area of application.

The research contributions in Figure 1.1 is presented in more detail below, ranging from contributions made for optimisation of the H3D imaging system to the development of 3D depth estimation frameworks and implementation of H3D image pre-processing techniques.

I. The implementation of a multilayer Holoscopic 3D encoding technique.
II. The proposal of the use of a deep neural network for Holoscopic 3D image up sampling.
III. The implementation of a Holoscopic content adaptation technique for depth estimation through disparity.
IV. The implementation of a novel direct depth from Holoscopic technique.

1. Holoscopic 3D Imaging System

The detailed analysis of the H3D imaging system resulted in the identification of the best trade-off between spatial and angular information. Depending on the nature of the 3D depth estimation approach, the use of a smaller or larger MLA is employed, resulting in high-quality depth maps.

The proposed multilayer H3D encoding system reduces the computation cost of viewpoint extraction by 95%, as content adaptation is a crucial step for current state-of-the-art H3D depth estimation techniques. The multi-layered H3D is a stack of labelled elemental images that aids more efficient identification of viewpoint images and feature extraction during depth estimation. The following contribution should be used in the development of the second generation H3D imaging system, where the installation of an inbuilt H3D image encoding and calibration system will make the H3D system mature enough for industrial use.

The underlying contribution is the development of a scalable Holoscopic content adaptation and 3D depth mapping technique for extracting reliable semi-segmented image data for standard H3D depth frameworks.

2. Holoscopic 3D Image Pre-processing

The H3D imaging system is known for its inability to record high-resolution viewpoint images. This limitation affects the feature matching process, a crucial stage of any depth estimation framework. The comprehensive investigation and evaluation of current single image up-sampling techniques currently used in handing the H3D low-resolution problem are executed and documented. The findings of the subsequent research led to the design and development of a learning-based single image up-sampling technique. This H3D learning-based network uses down-sampled multilayer EI stacks and 2D images as training data, using their default images as reference data. The network then learns the residual difference between the images to reproduce high-quality EI stacks that are reconstructed into a super-resolution H3D image. The proposed deep-learning single image resolution technique outperforms current state-of-the-art image up-sampling techniques like Bicubic and Bilinear. The results are evaluated using the industry-standard image quality assessment matrix, specifically Peak Signal to Noise Ratio (PSNR) and Structure Similarity Index (SSIM).

The underlying contributions are the design and development of a deep-learning-based single image super-resolution (SISR) framework for up-sampling Holoscopic 3D images.

3. Depth Estimation

The design and development of two robust H3D depth estimation technique make up the final part of this research contributions. The first depth technique, referred to as H3D depth from disparity (H3DDD), is designed to estimate 3D depth information from stereo viewpoint images extracted from a single H3D image. The H3D adaptation process in this depth estimation framework makes use of the H3D imaging system's ability to refocus at any given point after capture, extracting semi-segmented viewpoint images that improve feature matching results. The framework also contains a proposed smart depth mapping technique that is designed to improve computational efficiency. However, the above framework still has limitations that are associated with H3D adaptation, specifically, loss of angular information and the introduction of unwanted artefacts that affect feature matching. These limitations were the motivation behind the design and development of the innovative Direct Depth from Holoscopic technique (DDH). The DDH technique estimates 3D information directly from an H3D image due to its unique ability to calculate disparity information at the EI level.

The underlying contribution is the design and development of an innovative 3D depth estimation technique that estimates 3D depth information directly from a single Holoscopic 3D image.

1.5. Thesis Chapters Outline

This thesis is divided into five main chapters. The subsections below contain a brief outline of the chapters presented in this thesis.

Chapter 2 – Literature Review on 3D Systems and 3D depth Estimation Techniques

This chapter presents literature on current 3D imaging systems and the imaging principles, depth estimation techniques, learning based image up-sampling techniques and image evaluation metrics. The chapter starts by briefly introducing the principles behind Active imaging and Stereo/Multiview imaging systems before detailing light field/3D imaging principles. Elaborating on the critical components in 3D imaging systems that make them the future in optics, and simultaneously justifying why it is used as the primary device for collecting 3D image data in this research.

This chapter also presents various techniques used in estimating 3D depth information from a 2D image(s), presenting ways in which 3D depth information or 3D depth maps are usually represented. The chapter also presents 3D depth correspondence constraints that are used later on in chapter four and five to determine the complexity of a 3D depth technique. The chapter concludes by presenting the classes of 3D depth estimation techniques and listing notable methods currently been used in the computer and stereo vision society.

The chapter concludes by presenting the literature on learning-based image up-sampling techniques and state-of-the-art image evaluation metrics used in the evaluation of Holoscopic 3D image up-sampling and depth map results used throughout this thesis.

Chapter 3 – Holoscopic 3D Image Pre-processing

This chapter elaborates on the image registration process of H3D imaging systems, highlighting image registration errors associated with H3D images that can affect 3D depth estimation. The chapter presents all techniques implemented and designed to aid H3D depth estimation based on the following registration errors. The following techniques include a lens correction technique and an optimised viewpoint extraction technique based on the proposed H3D encoding technique. The analysis of current interpolation techniques such as the Nearest neighbour, Bilinear and Bicubic, currently been used to resolve the H3D low-resolution problem is also presented in this chapter. The chapter concludes with the presentation of the proposed learning-based single image up-sampling technique for Holoscopic 3D images.

Chapter 4 – Holoscopic 3D Depth Estimation from Disparity

This chapter presents the first of the two H3D depth estimation frameworks. This framework estimate 3D depth information from stereo viewpoint images extracted from a single Holoscopic 3D image. The process of extracting the stereo viewpoint images is referred to as Holoscopic content adaptation (HCA). The chapter starts by presenting the overall framework of the 3D depth from disparity (H3DDD) formulation, followed by a step by step process of how 3D depth information is estimated. This chapter concludes with the presentation of all the 3D depth maps estimated from various Holoscopic 3D image dataset, evaluating the results against state-of-the-art 3D depth estimation techniques.

Chapter 5 – Innovative Direct 3D Depth Estimation from Holoscopic 3D Image
This chapter presents the second 3D depth estimation framework implemented uniquely for H3D images. This framework is implemented to by-pass the Holoscopic 3D content adaptation stage, totally eliminating the introduction of unwanted artefacts and loss of angular information that might affect the quality of estimated 3D depth maps. The technique also has the ability to estimated disparity information in both directions of an omnidirectional H3D image. The unique feature that enables this technique to estimate 3D depth information directly is the similarity measure, where the estimation of disparity is done at an EI level. This chapter concludes with the presentation of all the 3D depth maps estimated from varying Holoscopic 3D image dataset, evaluating the results against state-of-the-art 3D depth estimation techniques.

Chapter 6 – Conclusions and Future Works
The overall conclusion of this research is presented in this chapter, suggesting future works based on these research contributions.

1.6. Author’s Publications

A. S. Aondoakaa, M. R. Swash, and A. Sadka, “3D depth estimation from a holoscopic 3D image,” in *2017 4th International Conference on Signal Processing and Integrated Networks (SPIN)*, 2017, pp. 320–324

Belhi A., Bouras A., Alfaqheri T., Akuha A., Sadka A., and Foufou S., “Machine Learning and Digital Heritage: The CEPROQHA Project Perspective” *Fourth*

*International Congress on Information and Communication Technology. Advances in Intelligent Systems and Computing, vol 1027. Springer Singap*ore, 03 January 2020.

A. Belhi, A. Bouras, T. Alfaqheri, A. Solomon, and A. Hamid, "Signal Processing : Image Communication Investigating 3D holoscopic visual content upsampling using super-resolution for cultural heritage digitization ☆," *Signal Process. Image Commun.*, vol. 75, no. March, pp. 188–198, 2019.

CHAPTER 2: LITERATURE REVIEW ON 3D SYSTEMS AND 3D DEPTH ESTIMATION TECHNIQUES

This chapter presents qualitative research on current 3D imaging systems, state-of-the-art 3D depth estimation techniques, deep learning-based image up-sampling techniques and image up-sampling techniques. This chapter also introduces the evaluation constraints used for estimating the computational complexity of the 3D depth techniques presented in this thesis. The chapter layout is as follows: 2.1 3D Systems, 2.2 3D Depth Estimation Techniques, 2.3. Investigation of Learning Based Image Up-sampling Techniques, 2.4. H3D Evaluation Matric: PSNR and SSIM and 2.5 Summary.

2.1. 3D Systems

2.1.1 Introduction

Beyond standard two-dimensional (2D) imaging and photography, three-dimensional (3D) imaging systems records more information of the world and has attracted a lot of interest in research societies such as computer vision [10] and life science [11]. The additional information recorded by 3D imaging systems is related to the structural composition of the scene, attracting an increasing amount of attention in recent years, as shown in Figure 2.1 below.

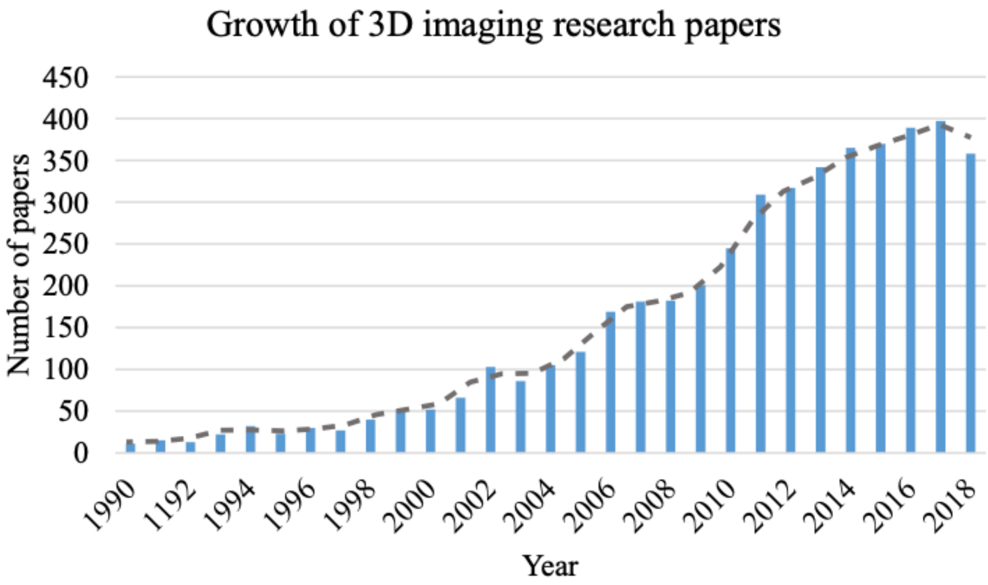


Figure 2.1 - Indication of the growing interest in 3D imaging found in Google Scholar during the past decade.[1]

The 3D depth or structural composition of objects recorded in addition to what 2D imaging systems record is sometimes referred to as amplitude and phase imaging or wavefront imaging. Depending on the properties and density of an object, the invisible electromagnetic wave is inevitably changed as it passes through an object, inducing intensity changes. However, for 3D depth information to be estimated successfully from digital image data, specific 3D depth cues apart from the intensity variations are considered.

[1] The recent growth of 3D imaging can be found at google scholar online at https://scholar.google.co.uk [Accessed: 30th December 2018].

2.1.2 Active Imagining Principles

Active imaging principles employs the use of a specially controlled light source as part of its technique to estimate 3D depth information of any given scene. The active lighting incorporates some form of temporal or spatial modulation of the illumination. This technique was proposed before the 3D techniques (3D depth-through-disparity and 3D depth-through-defocus) because of the fact that the microprocessing was not yet invented [1] [12]. From a computational point of view, Active imaging systems tend to be less demanding, as special illumination is used to simplify some of the steps in the 3D recording process [13]. However, their applicability is restricted to environments where the special illumination techniques can be applied. A second distinction is between the number of vantage points from where the scene is observed, as a single vantage point is used to record and estimate 3D depth information of a scene. In the case where there are multiple viewing or illumination components positioned very close to each other, ideally, the illumination components could coincide. The latter (illumination components) can sometimes be realised virtually through optical means like semi-transparent mirrors. For multi-vantage systems to work well, the different components often have to be positioned far enough from each other. One could say the 'baseline' between the components has to be wide enough. Technologies like 3D scanners, Vicon motion-capture and Xbox Kinect use this imaging principle to estimate 3D depth or produce 3D data directly. These technologies are both fitted with lasers and 3D scanners that project a laser beam onto the object and record the shape the beam makes or the time it takes to reflex the light source back to the sensor. Figure 2.9 below presents a visual representation of how this imaging principle works.

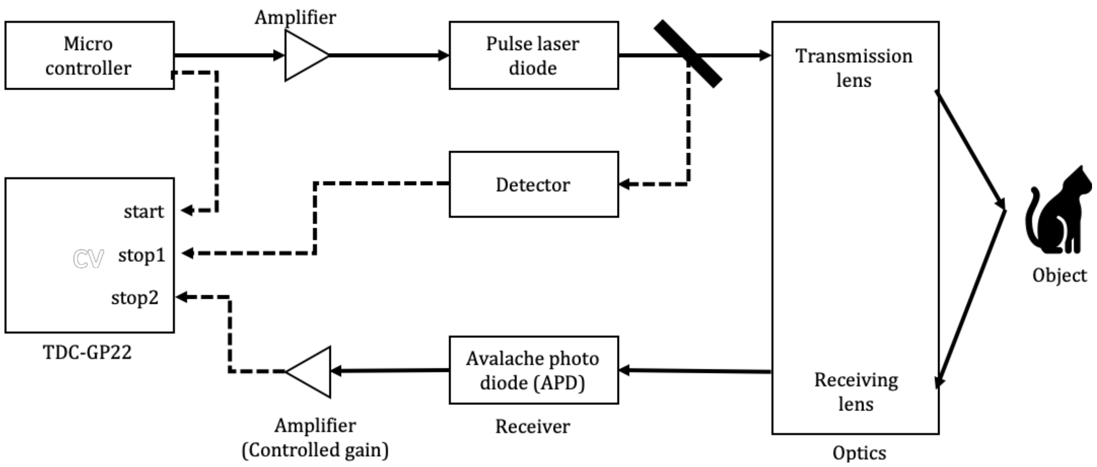


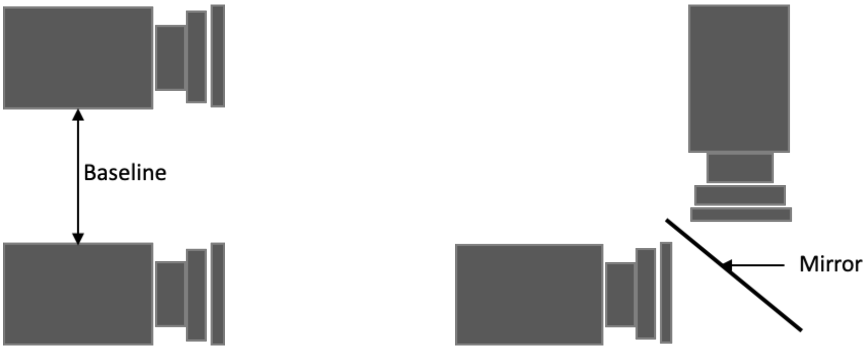
Figure 2.2 - Active and Time of flight (ToF) image recording principle.

Although the Active imaging system has some disadvantages over the 3D systems, one of its great advantage is its accuracy in obtaining the ground truth depth information in comparison to 3D depth estimation techniques.

2.1.3 Stereo / Multiview Imagining Principles

Stereo/Multiview imaging principles are the fundamental procedure taken by all 2D imaging systems to capture and display 3D data. In the early '50s, due to the decline in movie theatre attendance caused by television, stereoscopic cinema was seen as a method to regain this audience [14]. "This explains the first wave of commercial stereoscopic application and relevance in the '50s.

Sir Charles Wheatstone first proposed this imaging technique in 1838 [15], where the use of stereo imaging systems is employed to record slightly different stereo views of a scene. The concept behind stereography is derived from the human visual system. The human optical system is approximately two-and-a-half inches apart, resulting in the visualization of the same scene from slightly different angles and perspective. The left image is shown only to the left eye and the right image to the right eye, the brain then combines the images to give a perception of 3D depth; this is called binocular vision. Stereoscopic relates to seeing space three-dimensionally as a result of binocular disparity. The stereo imaging systems are usually set up to mimic the above system, either by a side-by-side camera rig or mirror camera rig, shown in Figure 2.12 below.



(a) side by side stereo rig (b) Mirror stereo rig

Figure 2.3-(a) A stereo side-by-side camera rig, (b) A stereo mirror camera rig.

The stereo imaging system has an interaxial separation between the stereo 3D imaging systems, similar to that of the human visual system. This distance is referred to as a baseline of the stereo imaging system, allowing the user to capture slightly different angles of a scene, and the difference between the stereo views is known as disparity. When disparity is processed and displayed with the right conditions, this information gives the viewer the perception of 3D depth like in a real-world scenario [16].

The Multiview imaging system is an extension of the Stereo 3D imaging system. This 3D imaging system has attracted increasing attention thanks to the rapid drop in the cost of digital imaging systems. 3DTV and free point TV are the most popular application areas where Multiview 3D imaging systems are applied to expand user experience beyond what has traditionally been offered by 2D media. The multiview system is developed by the convergence of new technologies from computer vision, multimedia, computer graphics and related fields. The Multiview 3D imaging systems use more than two camera array systems to capture different viewing angles of a scene. More than one user can visualise 3D depth at a time, this setup shown in Figure 2.13 below.

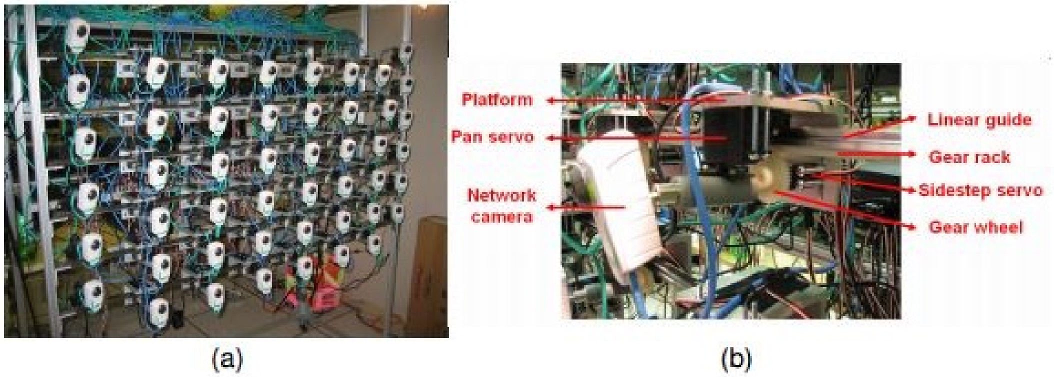
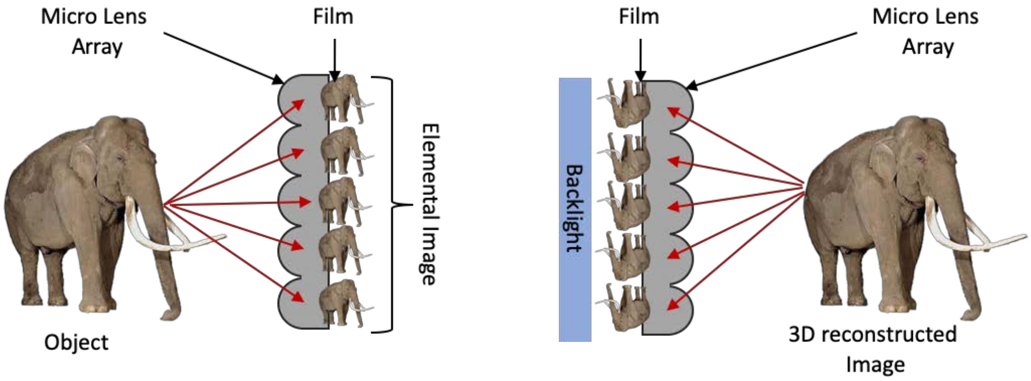


Figure 2.4 - (a) A 48 Multiview array camera system. (b) A mobile camera unit.[17]

The captured data from the above set up in Figure 2.13 usually overlap each other, giving the user the ability to match distinctive features between them. This occurrence has been studied intensively to deduce a technique of 3D depth estimation in the field of stereo and computer vision. This imaging principle is not used in recording data for this research as the cost acquired when recording multi-viewpoint images prove too costly compared to the 3D imaging system.

2.1.4 Light Field Imaging Principles

Light field "Holoscopic 3D" imaging can be seen as the next generation of imaging system in the optics industry due to its ability to record the full parallax of any given scene and its ability to refocus in post-processing. This imaging system has a wide application area such as robotics, autonomous navigation, AR/VR and many other imaging areas [1]. Lightfield is a true auto-stereo system that can provide all four-eye mechanism: binocular disparity, motion parallax, accommodation and convergence [6]. When using this system to capture accurate 3D images, a considerable amount of tightly packed distinct micro-images is obtained, referred to as elemental images (EI). The use of micro-lens or series of lenses was first proposed by the physicist Prof M. Lippmann (1845-1921) rather than the use of opaque barrier lines [2]. The micro-lens used enabled him to record the full parallax a scene; the microlens array also referred to as fly's eye lens array is used to record and playback the image as shown in Figure 2.5 below.



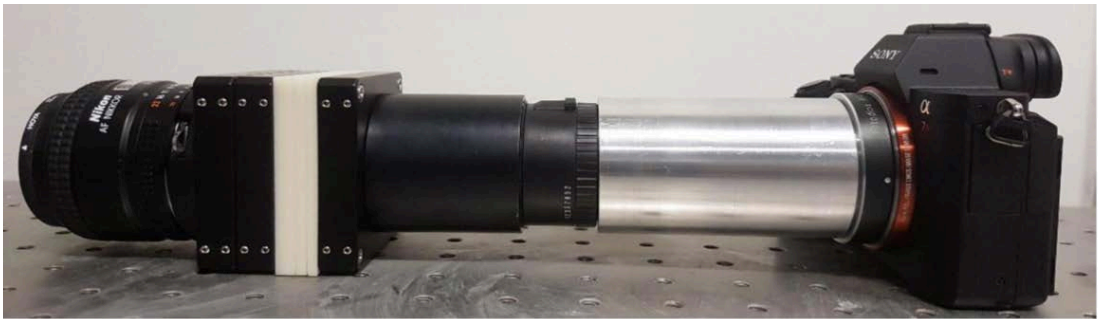
(a) Recording Process (b) Replay Process

Figure 2.5 - Principle of light field imaging system, (a) capture (recording) object process and (b) display reconstruction process.

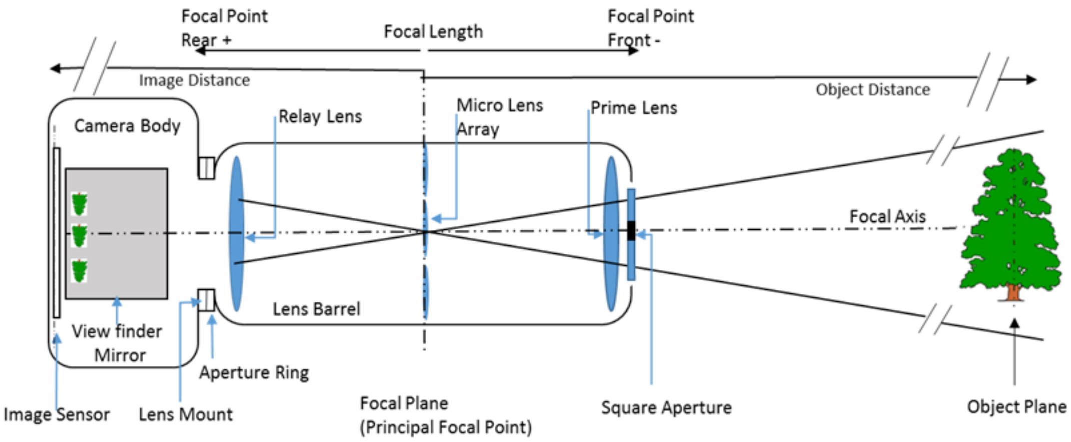
The next section presents the detail specifications of the Holoscopic 3D image system and its image properties.

2.1.4.1 Brunel Holoscopic 3D Camera

The Holoscopic 3D (H3D) imaging system is an adaptation of a standard commercial camera, where the image lens system is redesigned to enable the recording of the full parallax of any given scene. The lens architecture consists of a prime lens, microlens array and a relay lens as presented below in Figure 2.6.



(a)



(b)

Figure 2.6 - (a) Square Aperture H3D lens integration with Sonny MKII sensor (b)H3D system schematics.

Figure 2.6 description: Image sensor size= 35.9x24mm, Lens Barrel = Sonny 35mm F2 wide-angle lens, Lens Mount = Sonny Fmount, Relay lens = Rodagon 50mm F2.8 ×1.89, Image Sensor/Camera Body = Sonny Alpha MKII.

The main difference between any H3D imaging system and a 2D imaging system is the introduction of the microlens array placed before the camera sensor, as shown clearly in Figure 2.6b. The microlens enables H3D imaging systems to record the spatial and angular information of any given scene in a single snapshot, serving as an advantage over 2D imaging systems. The MLA mention is usually group into two categories, (i) Omni-directional and (ii) unidirectional microlens array.

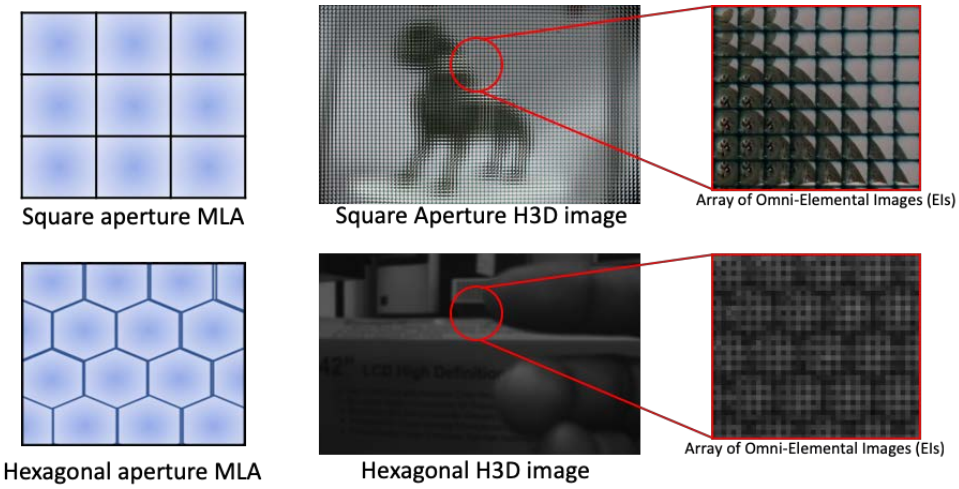

Square aperture MLA
Square Aperture H3D image
Array of Omni-Elemental Images (EIs)
Hexagonal aperture MLA
Hexagonal H3D image
Array of Omni-Elemental Images (EIs)

Figure 2.7 - Omnidirectional MLAs and their corresponding H3D images.

Figure 2.7 above presents the Omni-directional MLAs capable of recording the full parallax of any given scene in both vertical and horizontal directions. This group of MLAs usually come in a square or hexagonal shaped aperture, however, when extracting viewpoint images from an H3D image recorded with a square-shaped MLA aperture. The process of viewpoint extraction is relatively straight forward compared to an H3D image recorded with a hexagonal shaped aperture. This is as a result of the consistency of the viewpoint image displacement in square apertures compared to the irregular displacements of viewpoint images in hexagonal shaped apertures.

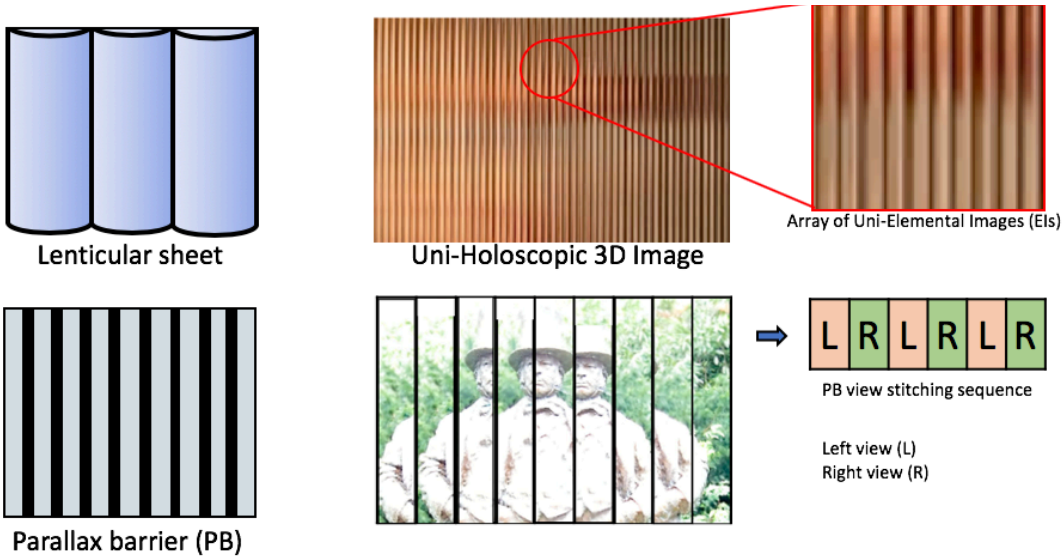


Figure 2.8 - Unidirectional MLAs and their respective H3D images.

Figure 2.8 above presents the unidirectional MLAs, depending on how it is positioned, it records the parallax of any given scene in either vertical or horizontal direction (one direction only). Unidirectional MLA type includes lenticular sheet and parallax barriers, and each has its own characteristics and drawbacks. The lenticular sheet main advantage over the parallax barrier is its ability to let more light be pushed through, resulting in the perception of sharper viewpoint images. However, the number of VI that can be displayed is fixed. Parallax barrier, on the other hand, is more flexible when it comes to the alternation of VI displayed. However, due to its poor transparency properties, light pushed through this aperture is significantly deemed down, resulting in dull VI. The schematics of the above MLA types are presented in Figure 2.9 below.

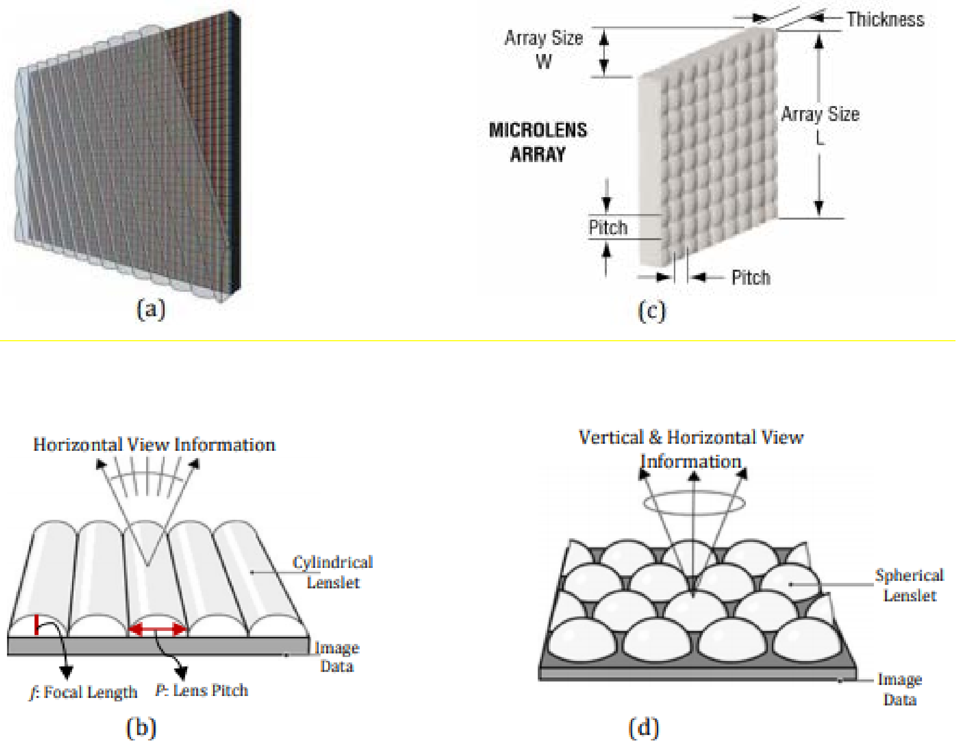
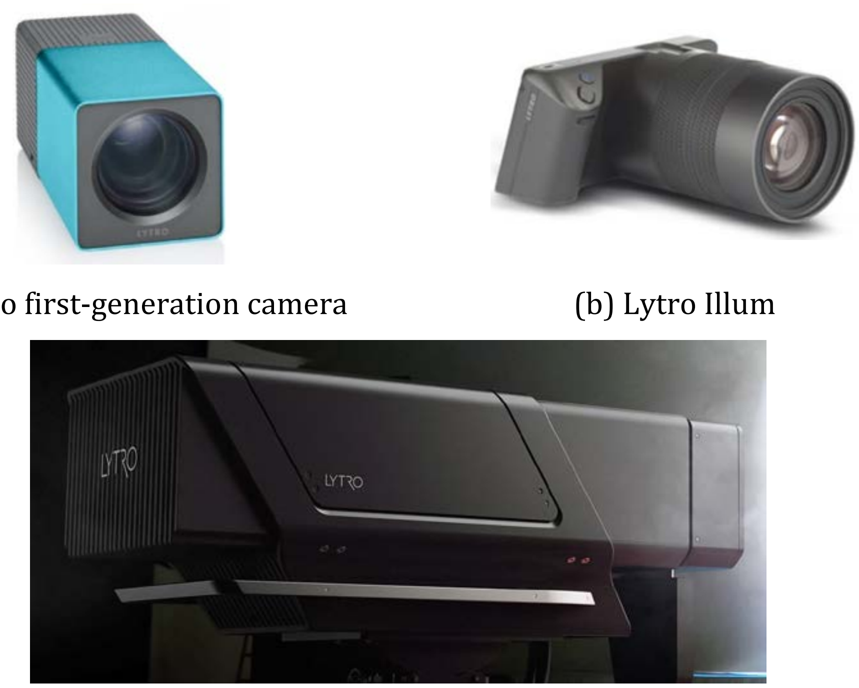


Figure 2.9 - Micro lens array schematics [88][89]. (a,b) Unidirectional lenticular sheet and (c,d) spherical micro lens array.

Given a lens pitch of a microlens, the number of pixels in an individual microlens can be defined as, EI Size = Pitch size /dot pixel pitch, Where EI Size is the maximum number of pixels an EI can capture, while the dot pixel pitch is the size of a single-pixel a specific imaging sensor can record. The total number pixels of an EI image is equal to the number of viewpoint images recorded by the H3D image system.

2.1.4.2 Lytro Light Field Systems

The Lytro imaging technology is of three types, the first generation, second generation and Lytro cinema. These cameras have the same properties of an H3D camera except for the fixed hexagonal-shaped MLA used for capture. The microscopic lens array (often in the range of 100,000) with tiny focal lengths as low as 0.15mm and split up what would become 2D pixel into individual light rays just before reaching the sensor. The resulting raw image is a composition of multiple hexagonal-shaped Elemental Images (See Figure 2.7). This camera is released with its image processing software for extracting viewpoint images and refocusing at different planes within the 3D image. Figure 2.10 below are images of the Lytro light field technologies.



(a) Lytro first-generation camera

(b) Lytro Illum

(c)Lytro cinema

Figure 2.10 - Lytro imaging technologies. a/b) been the first and second-generation cameras, while c) is the latest camera.[2]

2.1.4.3 Direct Comparison of Brunel H3D Systems and Lytro Systems

Some of the positives that the Brunel H3D imaging system have over the Lytro imaging systems are as follows,

i. Compared to the first and second generation Lytro cameras, the H3D imaging system has a much bigger imaging sensor for recording over the Lytro imaging systems.

ii. The Microlens array placed right before the imaging sensor can be easily changed in the H3D imaging technology than that of the Lytro imaging systems.

iii. Raw H3D image is available and easy to access, where the raw image of the Lytro imaging systems is more difficult to access and adapt, for application use.

[2] The images of the Lytro imaging systems on this page where all downloaded from Lytro official website www.lytro.com [Accessed: 29th January 2016].

iv. Viewpoint images are restricted to a maximum of two for Lytro imaging systems while the H3D imaging system provides as much as the size of the Elemental image.

v. H3D images give the user more refocusing options in comparison to the Lytro image data that is restricted to three planes.

vi. The H3D imaging system is financially less expensive than Lytro imaging systems.

Some of the positives the Lytro imaging system have over Brunel H3D imaging system are as follows,

i. The Lytro imaging system is more mobile as the H3D imaging system still is at its prototype stage.

ii. The fact that the calibration process of the H3D imaging system is done manually exposes the H3D data to human error, while the Lytro MLA is factor fitted and always ready for use.

iii. The Lytro imaging system has a standard software for Lytro image processing while the Holoscopic 3D imaging system is still at its development stage of standard software.

Based on the direct comparison of the H3D imaging system and the Lytro imaging system, the H3D imaging system is used due to its low cost, high-resolution sensor and easy access to raw image file for 3D depth estimation.

The next section presents the 3D depth estimation techniques that are currently used to estimate 3D depth information from any image data.

2.2. 3D Depth Estimation Techniques

2.2.1. Introduction

3D depth estimation also referred to as stereo correspondence problem, require only image data as its primary input to calculate the depth information of any given scene. Resulting to lower equipment cost that constitutes one of the advantages 3D depth estimation technique have over sensor-based or active 3D depth estimation techniques [18][19]. The most common 3D depth estimation techniques are documented in this section and classified either as a global or local technique. The complexity range of 3D depth estimation problem is defined and presented in this chapter as stereo constraints in section 2.2.3. Depending on the number of restrictions a stereo problem requires to produce reliable results, the complexity of the stereo problem can be estimated as well. The stereo constraints presented in section 2.2.3 is used for evaluation of all H3D depth frameworks presented in this thesis.

The stereo correspondence problem is usually tackled within four steps: matching cost calculation, cost accumulation, disparity calculation and lastly disparity refinement. However, a more detail framework for disparity estimation is presented in Figure 2.11 below.

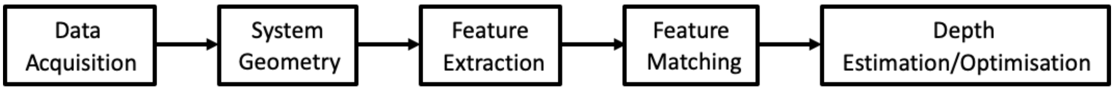


Figure 2.11 - Current state of the art 3D depth estimation framework.

I) Data Acquisition

Data acquisition in 3D depth estimation requires only the use of image recording systems. However, the data type can consist of a single image frame or multiple image frame. The process of estimating 3D depth information from such data is also referred to as structure from motion (SFM). The imaging systems used for recording the 3D depth estimation content can be classified into three major types, namely, (i) a single or monocular camera system, (ii) stereo camera or Multiview camera system and (iii) lastly, H3D camera imaging system. Single-camera system requires 3D depth to be estimated from just a single 2D image. Single camera is not a popular technique of recording 3D data as the absence of disparity makes 3D depth estimation extremely complicated. In the case where this single image system records multiple image frames for estimating 3D depth, misalignment problems and camera calibration problems also make the estimation of 3D depth information a very complex task [20][21]. The principle behind Multiview imaging system is the use of two or more cameras to record a scene, when using this imaging setup to record 3D depth information for 3D depth estimation, the ability to control the baseline between viewpoint images coupled with the high resolution gotten from this stereo imaging system, makes it the most popular technique of acquiring image data for 3D depth estimation. However, if the number of viewpoint images of a scene increases, the cost of acquiring those system increases as well. The fragile nature of the Multiview rig renders the recording setup immobile. Finally, the 3D imaging systems, these imaging systems can record the full parallax of any given scene and refocus after capture. However, due to its inability to produce high-resolution viewpoint images and small baseline range, makes it challenging to extract reliable 3D depth information. The implantation of a robust 3D depth estimation technique presented in this thesis makes the following H3D related issues irrelevant, making the H3D the most cost-effective way of recording image data for 3D depth estimation purposes.

II) System Geometry

System geometry, also known as camera calibration, is a crucial but time-consuming step in depth estimation frameworks. The information extracted from an imaging system indicates the actual 3D coordinates (intrinsic parameters) and the position of the imaging system in relation to the scene during capture (extrinsic parameters) as presented in Figure 2.12 below. These intrinsic and extrinsic parameters imaging do not only help in enforcing the Epipolar constraint but also help in correcting lens distortions errors.

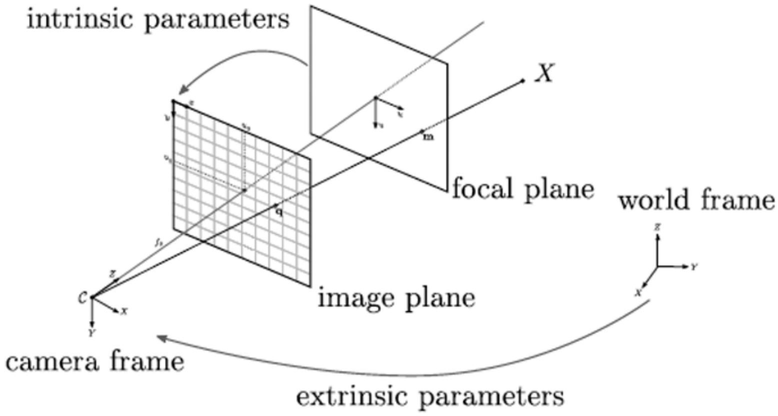


Figure 2.12 – An oriented central projective camera from a pinhole model.

Several dynamic camera calibration techniques are used in extracting the imaging system position during data acquisition, but one of the most common is the checkerboard technique. H3D data captured from the H3D imaging system does not require any form of camera calibration for 3D depth estimation as all viewpoint images are recorded orthographically.

III) Disparity Estimation (Feature Extraction/ Matching-cost/ Depth Computation)

The feature extraction, matching, and initial 3D depth estimation stages often work hand in hand. Depending on the technique used, they can be classified either as a local or global 3D depth estimation technique, presented in more detail in sections 2.2.4 and 2.2.5. As mentioned earlier, current 3D depth estimation techniques use stereo data recorded from a stereo imaging system. The stereo systems usually have a baseline that is targeted to be the same as the distance between the human visual systems. Resulting in the registration of scene features at different locations onto the 2D stereo imaging sensors. The difference in this feature registration is referred to as disparity.

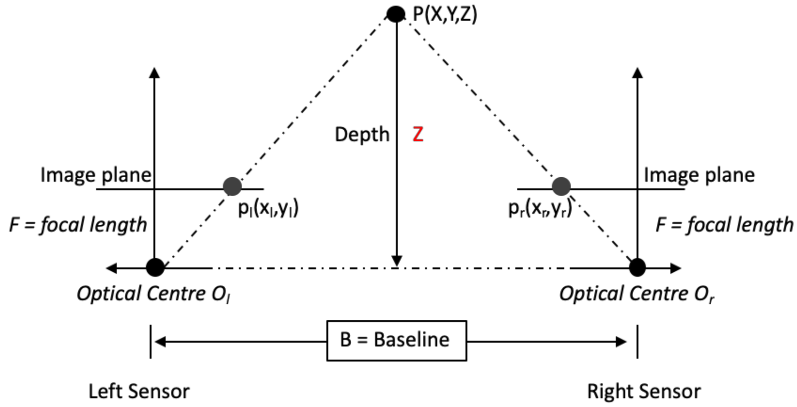


Figure 2.13 - Disparity representation of a rectified pinhole camera model.[1]

The depth, Z, of point, P(X, Y, Z), presented in Figure 2.13 above is estimated through triangulation as the point in space is projected onto two image views at different positions $p_l(x_l,y_l)$ and $p_r(x_r,y_r)$, with a fixed baseline 'B'. The baseline distance between the optical systems is the main factor that results in the projected points been recorded at different positions. Projected points of objects closer to the optical system have more significant disparity while projected points or features of objects further away have lesser disparity between them. This basic principle is the main idea on which various 3D techniques are derived. The following defined below, in Equation (1) as:

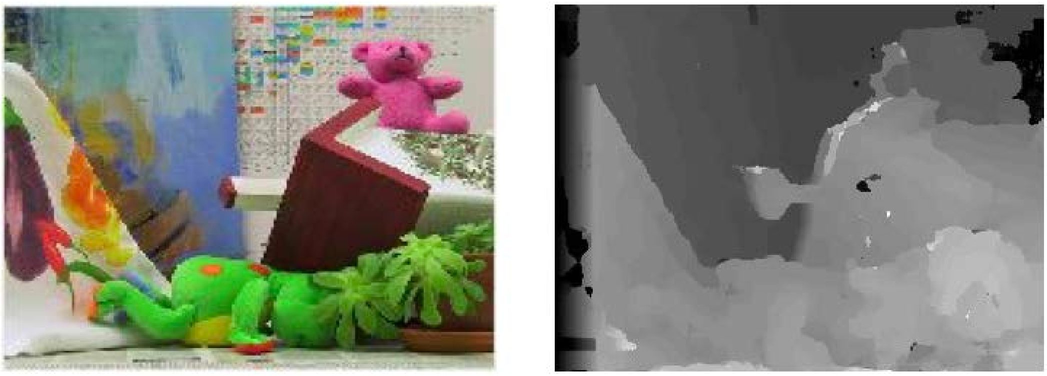
$$Z = \frac{fT_x}{d} \qquad (1)$$

Where, f is the focal length, T_x is the baseline and d is the disparity. The depth information mentioned can be represented or stored in one of the following 3D depth registration formats listed below.

2.2.2. 3D Depth Map Representations

I. Greyscale 2.5D Depth Map Representation

Greyscale images are made up of pixels ranging from black and white intensity values. This mode of 3D depth representation registers 3D depth information by registering pixels closer to the recording systems with higher pixel intensity values than pixels further away, or vice versa, producing a grayscale image usually known as a greyscale 3D depth map, shown in Figure 2.14 below.

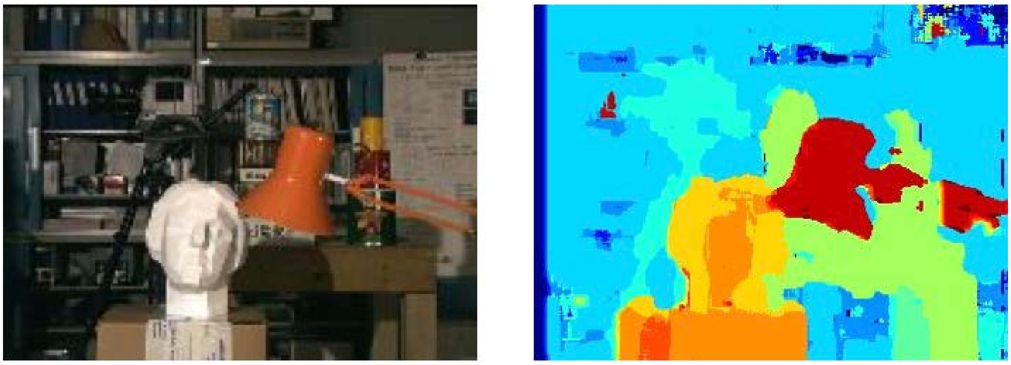


(a)Reference viewpoint image (b) Depth map result

Figure 2.14 - Greyscale 3D depth map representation.[22] [23]

II. Colour 2.5D Depth Map Representation

Similar to the greyscale 3D depth representation, this mode of representation uses all RGB values instead of a range of pixel values between a colour value to represent 3D depth information. Usually using darker RGB values to register objects further away from the recording system and lighter values for objects closer, shown in Figure 2.15 below.

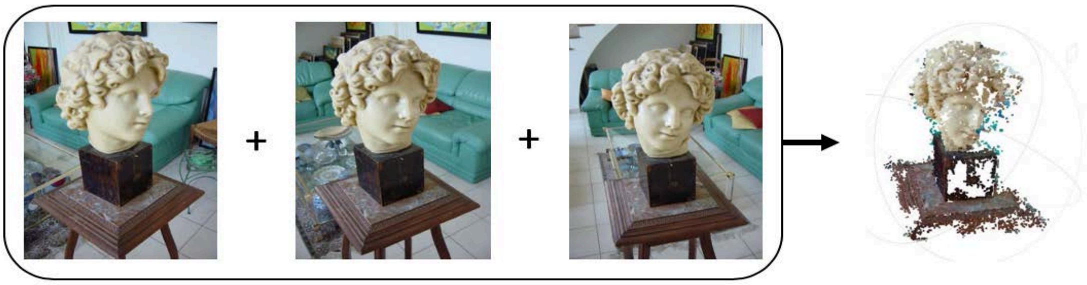


(a)Reference viewpoint image (b) Depth map result

Figure 2.15 - Colour 3D depth map representation. [22] [23]

III. Pseudo-3D Depth Map Representation

This mode of representation provides different viewpoints of the scene reconstructed with a point cloud, and in some cases, the point cloud is converted into a 3D mesh creating a virtual environment. This 3D depth representation is often used when multiple viewpoints are used to the estimate 3D depth information of a scene, shown in Figure 2.16 below.



(a)Reference viewpoint images (b) Depth map result

Figure 2.16 - Pseudo-3d 3D depth map representation.[53]

The following 3D depth representation techniques are used throughout this thesis. Although the use of colour and grayscale 3D depth representation methods makes subjective evaluation between state-of-the-art 3D depth estimation algorithms easier, the limited angular information between viewpoint images recorded by the H3D system makes it tedious to represent depth data in a pseudo form.

2.2.3. Stereo Correspondence Problem

The stereo correspondence problem is the process where two or more images of a scene are used to estimate the 3D depth information of a scene. The following achieved by determining the disparity of matching pixels between the stereo viewpoint images. This problem can be defined or represented in different ways, however, in this thesis, the stereo matching problem is defined as a series of constraints and presented in subsections 2.2.3.1 to 2.2.3.6.

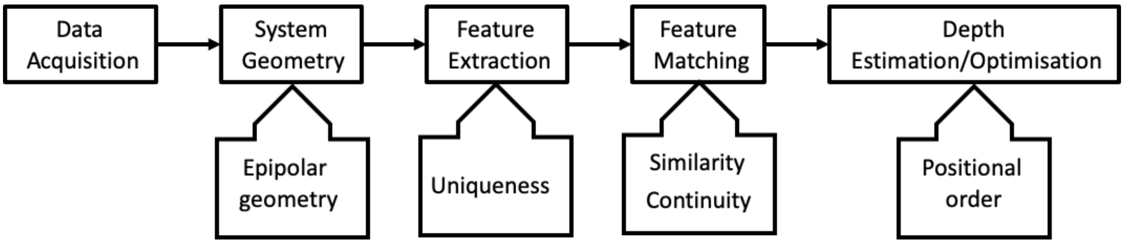


Figure 2.17 - 3D depth estimation framework along with various stereo correspondence problems placed at the stage of rectification.

2.2.3.1. Similarity Constraint

The similarity constraint is the foundation of every 3D depth estimation technique, where the correlation between pixels registered on the reference and target viewpoints images are estimated. This is executed at the Feature extraction stage, enforcing the point that both projections of the same three-dimensional entity should have similar properties or attributes; like shapes, colours, sizes, vertex and number.

2.2.3.2. Epipolar Geometry Constraint

Estimation of 3D depth from stereo images require a pixel from one view to match against to a pixel on the other view. Due to imaging errors like image misalignment and lens distortion, pixels projected onto a reference view is not usually projected onto the same horizontal plan onto the target view. In order to reduce the number of potential correspondences, the exploitation of additional information of the camera position is used, resulting in less computational cost and more reliable feature matching. This is known as the Epipolar geometry, and when applied to stereo viewpoint images, it is known as image rectification.

2.2.3.3. Uniqueness Constraint

This restriction applies the condition that a feature from the reference viewpoint image has one, and only one, feature related to it on the target viewpoint image. This is done to accommodate scenes with the object of similar texture, which leads to the possibility of a single feature having more than one match. Viewpoint images with occluded areas can further complicate the implementation of this restriction as well, resulting in most stereo algorithms working better on well-textured scenes than scenes with less texture.

2.2.3.4. Positional Order Constraint

The Positional order restriction implies that on the target viewpoint image, features have to appear in the same order as the reference image. Given the situation where there is a cross-eyes projection, the disparity between those features can be more tedious to estimate. However, most stereo techniques do not have this problem, and specifically in the case of the H3D data, where the microlens array tightly packed together, prevent cross-eyes projection or positional disorder.

2.2.3.5. Disparity Continuity Constraint

Disparity continuity assumes that changes in the image disparity are smooth, i.e. if a disparity map is to be considered reliable, it is presented continuously except for expected discontinuities. This principle also appears in different forms and sometimes with some small variations, as the case of minimum differential disparity [24][25].

2.2.3.6. Structural Relations Constraint

Structural relations impose that object made of edges, verticals or surfaces with a specific structure and geometry arrangement between the elements.

The stereo matching problem/constraints described above can be applied in different orders depending on the application they are used. However, not all the restrictions are applied when it comes to the extracting 3D depth information from an H3D image. For example, the positional order restriction would not be needed due to the slight pixel shift in perspective between reference and target viewpoint images extracted from an H3D image. This is caused by the microlens closely placed together. In a typical scenario, the most used H3D restrictions are similarity, uniqueness and continuity.

The change in the order of the following constraints listed above, produces two typical alternatives, local and global techniques. A detail explanation of the local and global depth estimation techniques is presented in sections 2.2.4 and 2.2.5 below.

2.2.4. Local Depth Estimation Techniques

Local depth estimation techniques apply constraints on a small number of pixels around the pixel been inspected (reference pixel). They are usually efficient but sensitive to local ambiguities of the regions (i.e. regions of occlusion or region with uniform texture) which are one its drawbacks. However, local depth estimation techniques are computationally less demanding and can be grouped as an area-based technique or feature-based technique [26][27].

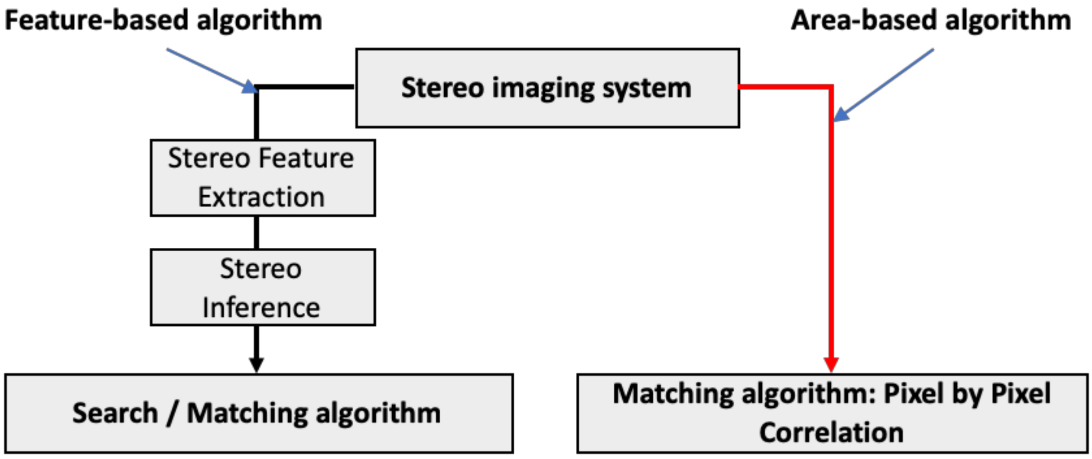

Feature-based algorithm
Area-based algorithm
Stereo imaging system
Stereo Feature Extraction
Stereo Inference
Search / Matching algorithm
Matching algorithm: Pixel by Pixel Correlation

Figure 2.18 - A local area and feature-based matching workflow, presenting the difference in their matching process.

3D depth estimation techniques that are usually associated with the following Feature-based and Area-based local techniques, presented in Figure 2.18 above are further explained in sections 2.2.4.1 and 2.2.4.2 below.

2.2.4.1. Area-based Matching Technique

Area-based techniques solve matching problems by using the intensity patterns of the neighbourhood of a reference pixel to determine its correlation. It estimates the correlation between the distribution of disparity for each pixel in an image using a window centred at the reference pixel. A window of the same size centred on the target pixel along the epipolar scanline, presented in Figure 2.19 below.

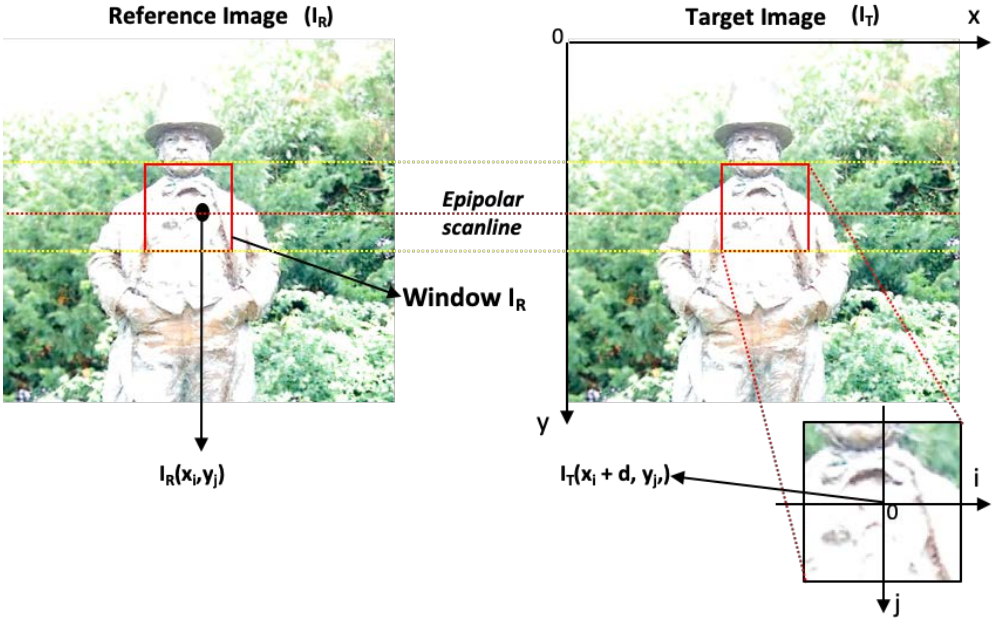


Figure 2.19 - Local similarity measure, where $I_R(x_i,y_j)$ is reference pixel, and $I_T(x_i+d,y_j)$ is the target pixel, matched along the epipolar scanline.

The effectiveness of this technique depends largely on the window size taken. It can be assumed that the larger the window, the better the outcome. However, the larger the window size becomes, the higher its computational cost as well. Therefore, the biggest problem of this method is to find the appropriate window size that can ensure finding a correspondence between two viewpoint images, having in mind that if the window size is too large, it could cause a huge latency to the system. Also, if the window size is close to the total image size, it would be deriving to the global methods, which were not considered because of their computational inefficiency.

A list of commonly used area-based depth estimation techniques is listed below.

I) Sum of Absolute Difference (SAD)

The sum of absolute difference measures the similarity of a pixel from a reference image in the target image. By summing the intensity values of window including the discrete pixel from the target image and finding the same best match of that block in the reference image. Depending on the point in which the best match is found, the absolute difference is referred to that discrete pixel's disparity value. The following is defined in Equation (2) below as:

$$C_{SAD}(x, y, d) = \sum_{(x,y)\in WINDOW} (I_R(x, y) - I_T(x - d, y)) \tag{2}$$

Where, $I_R(x,y)$ is the reference image and reference pixel in which its disparity is estimated in the target image I_T, and d is the disparity value. WINDOW is the pixel block containing the reference pixel.

II) Sum of Squared Differences (SSD)

Similar to the Sum of Absolute Difference (SAD) technique, the Sum of Squared Differences is a pixel by pixel similarity measure technique that estimates the sum of squared of a window between the target and reference viewpoint images. This technique estimates disparity of a pixel by finding it minimal squared value, as defined in Equation (3) below,

$$C_{SSD}(x, y, d) = \sum_{x=0}^{M}\sum_{y=0}^{N}(I_R(x, y) - I_T(x - d, y))^2 \tag{3}$$

M and N is the maximum size of the array, d is the disparity, while x and y are values for pixel coordinates for the reference and target viewpoint image, I_R and I_T. The SSD directly uses the formulation of the sum of square error [28].

III) Birchfield-Tomasi Measure (BT)

By linearly estimating the interpolated values of a window match and its nearest linear neighbouring pixels, the Birchfield-Tomasi dissimilarity measure is insensitive to image sampling [29]. However, for this technique to work well, the intensity functions must vary predictably between pixels as long as aliasing does not occur. This technique is Equation (4) below,

$$C_{BT}(x_R, y, d) = min\left\{ \min_{x_R-\frac{1}{2}\leq x\leq x_R+\frac{1}{2}} \left|I_R(x_R, y) - \tilde{I}_T(x + d, y)\right|, \min_{x_R-\frac{1}{2}\leq x\leq x_R+\frac{1}{2}} \left|I_T(x_R + d, y) - \tilde{I}_R(x, y)\right| \right\} \tag{4}$$

Where, *I_R (x, y)* and *I_T (x, y)* are the reference and target image viewpoint image and respective pixels, and *d* is the disparity value of the reference pixel. This equation is derived with the assumption that the viewpoint images are rectified.

IV) Normalized Cross-Correlation (NCC)

Normalized cross-correlation (NCC) derives the matching cost between stereo viewpoint images using Cauchy Schwarz inequality [30][31]. Technically, this is achieved when a feature from a reference image is matched in the target image by locating its maximum value in the image matrices. NCC is computationally more expensive compared to the SAD and SSD techniques as it is more robust due to its involvement of numerous multiplications, division and square root operations. This technique is defined in Equation (5) below,

$$C_{NCC}(x, y) = \frac{\sum_{i,j\in A} \sum [I_R(i, j) * I_T(x + i, y + i)]}{\left[\sum_{i,j\in A} \sum {I_T}^2(x + i, y + i)\right]^{\frac{1}{2}}} \tag{5}$$

Where the ***x*** and ***y*** variables are shift component that travels along the axis of the target viewpoint image, the numerator term defined in Equation (5) is the cross-correlation between the reference and target viewpoint image. However, the cross-correlation alone cannot be used as a similarity measure because it will produce false results. Therefore, the denominator term is used as a normalized cross-correlation to acquire a correct match.

V) Census Transform (CT)

Census transform reduces the image intensity composition of an image data into binary intensity values depending on the value of the centre pixel [32]. This technique is often used in the stereo vision community as it is not sensitive to global radiometric differences, such as global illumination differences [33].

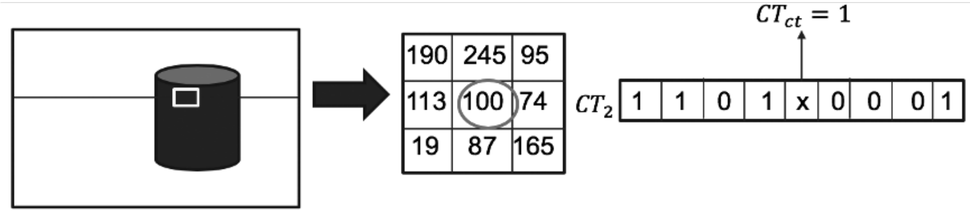


Figure 2.20 - Census transform technique.[63]

The Census transform technique shown in Figure 2.20 above is highly dependent on the centre pixels and depending on the window size used could result in high computationally cost.

2.2.4.2. Feature-based Matching Techniques

This group uses some specific features like edges, shapes and curves [34][35][36] resulting to the use of a differential operator (typically Laplacian or Laplacian of Gaussian, as in [37][38]). Feature-based matching also requires a convolution of a minimum of 3x3 to a maximum window size equivalent to the reference and target viewpoint image; resulting to an increase in computational load as the size of the operator grows. However, these algorithms do not allow real-time implementations as oppose to area-based techniques. The main difference being feature-based matching techniques require an extra step where images are pre-processed to extract suitable and reliable features for the matching step (see Figure 2.18 above). This pre-processing stage usually extract features from both images, resulting in the identification of features of each image. This step is closely linked to the matching stage of the respective matching techniques in which it is used because, without this step, the technique would not be able to have enough information to make an inference and obtain the image correlation.

For feature-based depth estimation, the most widely used features within the stereo images are breakpoints, isolated chains of edge points or regions defined by borders. Once the important aspects of edges are extracted as shown in Figure 2.21, the techniques then use arrays of edge points to represent straight segment, not straight segments and closed geometric structures, defined or unknown.

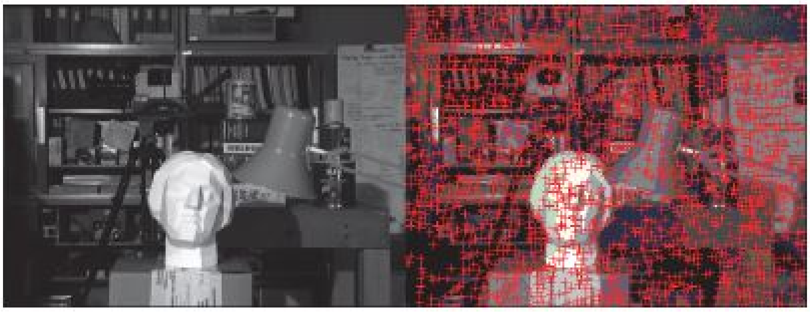


Figure 2.21 - Edge detections in a feature-based algorithm. [1]

Other primitive regions that can be used in feature-based techniques apart from the edges are regions within an image were an area that is typically associated with a given surface in the 3D scene is bounded by borders, basically, objects of the 3D scene that are made up of well know geometric shapes.

Depending on the matching technique used and the number of object features, an additional segmentation step may be necessary. In the segmentation step, additional information would be extracted from the known features which are calculated based on inferences from known characteristics. Therefore, the matching technique that receives the inferred data possesses much more information than the technique that works directly on the pixel intensity. A list of the commonly used feature extraction techniques in Stereo/Computer vision is presented below.

I. Scale Invariant Feature Transform (SIFT)

SIFT feature extraction technique is insensitive to image rotation, translation, scaling and partly robust to illumination changes. This feature is classified as a part-based approach [39] and consists of three main steps. First, the Laplacian of Gaussian (LoG) filter with different sigma value or kernel size is applied on multiple copies of the same local area in which a feature is examined within the stereo viewpoint images. However, finding the right sigma value when using the Laplacian of Gaussian (LoG) can be challenging as there is no true way of deriving the best value suitable for the local feature scales. Secondly, the calculation of Difference of Gaussians to locate the extrema within the localised area. Finally, the maxima along edges within the local area are suppressed to reduce the possibility of ambiguous matches [40]. Once this is completed, reliable features suitable for disparity estimation is matched.

II. Speeded-Up Robust Gradients (SURF)

SURF feature descriptors share similar properties as SIFT descriptors, where mixing of crudely localised information and distribution of gradient related features are used for feature extraction. Specifically detecting feature points by the use of integer approximation of the determinant of Hessian blob detector [41]. The main difference between the two is that the SURF descriptors computationally cost less compared to SIFT descriptors due to its less complex nature.

III. Histogram of Oriented Gradients (HoG)

Histogram of Oriented Gradients (HoG) is, by default, a global feature detector. However, the same technique can be optimised to work with local 3D depth estimation algorithms. The HoG technique consists of the following steps, i) computing the centred horizontal and vertical gradients with no smoothing, followed by the computation of gradient orientation and magnitudes. In cases of colour images, only one colour channel gradient and magnitude are computed. HoG descriptors are sensitive to occlusion and are commonly used for human detection [42].

2.2.5. Global Depth Estimation Techniques

Global depth techniques estimate 3D depth information from images by applying restrictions on the entire image, they are usually less sensitive to local peculiarities, and they add support to regions that are difficult to study in a local way, for instance, occluded regions. To overcome such ambiguities, restriction terms or functions are implemented to minimises the global cost or energy for more reliable estimation of 3D depth maps. This technique is also referred to as an energy minimisation problem [43][33][44][45][46]. The following is defined in Equation (6) as:

$$E(f) = E_{data}(f) + E_{occ}(f) + E_{smooth}(f) \tag{6}$$

Where the data energy function $E_{data}(f)$, measures the agreement or disagreement between pixels, based on their assumed disparities. $E_{data}(f)$ is defined below, imposing a penalty based on the intensity differences of matching pixels ***p*** and ***q.***

$$E_{data} = \sum_{l(p,q)=1} D_{(p,q)} \tag{7}$$

The smoothness energy function $E_{smooth}(f)$ measures the disparity smoothness between pixel pairs. This is enforced when pixels of the same disparity segment have relatively close disparities or minor disparity difference between them, keeping a consistent change between them. This is defined below as:

$$E_{smooth} = \sum_{\{(p,q),(p',q')\} \in N} K_{\{(p,q),(p',q')\}} \bullet T_{(l_{(p,q)} \neq l_{(p',q')})} \tag{8}$$

The presence of occlusions makes it more complex to accurately enforce the smoothing function above and relying on the data energy function alone will not improve 3D depth results. In order to rectify this issue, the occlusion energy function, $E_{occ}(f)$ is enforced, where pixels are modified to present pairs of pixels which potentially correspond. This is defined below as:

$$E_{occ} = \sum_{p \in I_1 \cup I_2} C_p \cdot T(p\ is\ occluded) \tag{ 9}$$

Where a penalty term is imposed on a specific pixel *p* in the stereo image pair *I₁* or *I₂* is occluded, and *T(.)* is the indication function. The occlusion term above could lead to an excess of regularisation, and the penalty function or cost of assigning different disparity to neighbouring pixels is not always convex, resulting in the estimation of relative disparities.

Although global depth estimation techniques often produce dense 3D depth maps, they tend to be computationally expensive and not suitable for real-time 3D depth applications. Techniques such as the Sum of Absolute Differences (SAD), Sum of Squared Differences (SSD) and Census Transform (CT) can be optimised to perform as a semi-global 3D depth estimation technique by the introduction of a

3D depth optimisation stage or smoothness function to refine initially estimated 3D depth map.

Formulations and graph models associated with Global depth estimation techniques are presented in section 2.2.5.1 below.

2.2.5.1. Graph Cut Formulations and Models

Graph Cuts formulations take into consideration the dimensional nature of the stereo correspondence problem. The dynamic use of graph theory in solving this problem was initially proposed by Roy and Cox 1998 [47], then reformulated by Veksler 1999 [48] in which the stereo correspondence problem is considered as an energy minimisation problem. This was followed by the introduction of an iterative graph-cut algorithm by Kolmogorov and Zabih in 2001, 2002a, and 2002b [49][50][51][52]. Boykov et al. 1998 & 1999 also proposed another graph cut technique that is formulated by a Markov Random Field (MRF)[53], accepting nonlinear penalties for discontinuities making its estimate more precise disparity maps near object edges or occlusion edges. The use of Markov random field (MRF) or different variants of MRFs such as loopy-BP, iterative BP, Markov Belief Propagation (BP) can be associated with the global 3D depth estimation structure presented in Equation (6) above[54][55]. The workflow of the mentioned graph cut formation is presented in Figure 2.22 below, where information is communicated between nodes, and the cost is represented as the weight of their edge. This group of techniques splits and smooths the image to match disparity regions, estimating 3D depth information in the process.

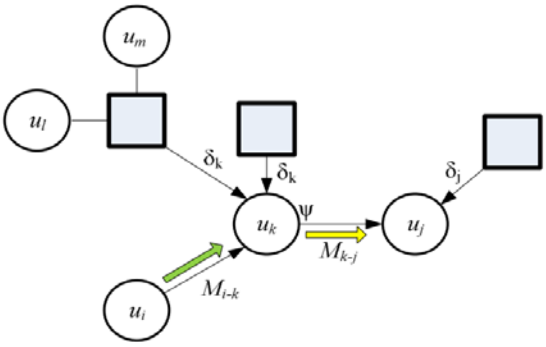


Figure 2.22 - Belief propagation (BP) model, where computing of new incoming information to neighbouring nodes is represented by the yellow arrow from information gotten from the node with the green arrow.[44]

2.2.5.2. Dynamic Programming Techniques

3D depth estimation or matching techniques classified under dynamic programming do not have a straightforward technique. As techniques under this group use specialised techniques that changes drastically depending on the 3D depth problem. 3D depth from defocus and 3D depth from a single 2D image often fall into this group. 3D depth from defocus takes into account the focal properties of its image data to estimate 3D depth, while 3D depth from single 2D images takes into account the monocular cues in the VI to estimate 3D depth information.

I) 3D Depth from Defocus

3D depth from defocus (DD) was initially used to estimate 3D depth in the early 1980s [56][57], but due to change in imaging technologies and their setting, various DD techniques have been suggested [58][21][59][60][61]. This approach to 3D depth estimation has received renewed attention in recent years. The method of estimation requires scene points that lie on a focal plane located at a certain distance from the lens be correctly focused onto the sensor, while points at greater distances or further away from the imaging sensor will appear increasingly blurred due to defocus. The traditional method of achieving this is by the capture of two images at camera settings with different focusing characteristics. One can then infer the 3D depth of each point in the scene from their comparative focus. 3D depth from defocus is more robust to occlusion problems compared to the shape reconstruction techniques.

II) 3D depth from a Singe Monocular View

3D depth from a single monocular view depends heavily on image segmentation and the monocular 3D depth cues found within the image view. Segmentation is based on edges, hue and saturation [20]. This technique often relies on known object structures that are made up of shaped like squares, circles and cylinders. This is often the case when estimating the 3D depth information of buildings with a single image. Other single image techniques rely on supervised learning [62][63], where training image data of buildings, trees, sidewalks, and various structured and unstructured environments are collected. Supervised learning is then used to predict the value of the 3D depth map as the function of the image.

2.3. Investigation of Learning Based Image Up-sampling Techniques

This section presents a brief overview of current learning-based single image up-sampling networks that inspired the framework of the proposed learning-based solution presented in chapter three.

Single image super resolution approaches are mainly classified either as interpolation based, reconstruction based and lastly learning based techniques. Interpolation based techniques up-samples images by weighting the value of neighbouring pixels. The bicubic technique being the most accurate out of the interpolation-based technique. Reconstruction based techniques solve only specific super-resolution problem, therefore lacking generalisation accommodation to other domains [64]. Finally, learning-based techniques which are mostly populated with machine learning techniques are currently the most robust image up-sampling technique. Learning-based techniques analyse the visual cues and learn the nonlinear mapping between low-resolution images and super-resolution images [65]. As a result, this class technique is investigated and applied to Holoscopic 3D image data.

Deep learning is an aspect of machine learning that is based on principles of deep neural networks [66]. Originally used as classifiers and trained using pairs of labels [66], this class of technique has been used in across image processing domains such object recognition, natural language processing, gesture recognition and computer vision [67][68]. The main aim of deep-learning techniques is to efficiently train classifiers to generalise unnoticed data samples, making it one of the major advantages of deep-learning techniques over other machine learning techniques, due to its handling and generalisation capability of unstructured data. Due to its high success, deep learning has been implemented for other applications such as time series prediction [69].

Single Image Super-Resolution (SISR) is a class of image up-sampling technique that aims to increase the resolution of low-resolution images (LR) to higher resolution images (HR) without losing the image's structural information and its high-frequency details. Other SR techniques such as legacy have several limitations concerning the irregularities of their LR-HR mapping and the inefficiency in handling a large amount of data. However, deep-learning-based techniques are proven in the mentioned area and continue to maintain their high standard of modelling due to their high-level ability to learn abstractions from training data [70].

Before presenting the framework for SR Holoscopic 3D data, a brief investigation of other deep-learning techniques follows.

2.3.1. State-of-the-Art Super Resolution Deep-learning Networks

I) SRCNN (Image Super-Resolution Using Deep Convolutional Networks)

By learning an end-to-end mapping between low/high-resolution images, a deep convolutional neural network (CNN) proposed by Chao Dong et al. [65] takes the low-resolution image as input data and outputs a high-resolution one. Let the high-resolution reference image be donated as X and the subject low-resolution test image donated as *Y,* is up-sampled using an interpolation-based technique such as bicubic interpolation to match the reference image size X. The main goal is then recovering an image for *F(Y)* to match the same spatial and feature quality of the reference X. This is achieved by the learned mapping F between Y and X. This mapping consists of the following three operations:

i. Patch extraction and representation: this operation extracts overlapping matches from the low-resolution image denoted as Y and represents each patch as a high dimensional vector that is made up of a set of feature maps, which equal to the dimensionality of the vectors.

ii. Nonlinear mapping: this is a nonlinear operation that maps each high-dimensional vector onto another high-dimensional vector, as these vectors comprise another set of the feature map, where each mapped vector is conceptually the representation of a high-resolution patch.

iii. Reconstruction: this operation aggregates the high-resolution patch representation to reconstruct the final high-resolution image. The resulting image is expected to have similar qualities to the ground truth **X.**

The following operations form the convolutional neural network presented in Figure 2.23, where the overview of the network is depicted.

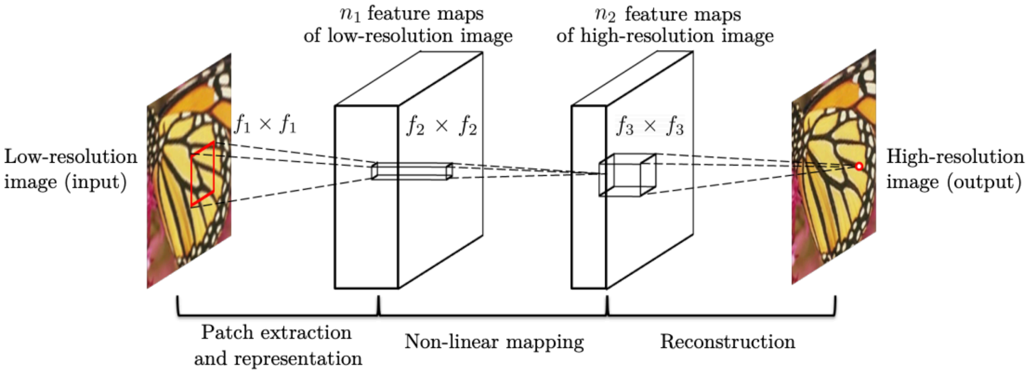

n_1 feature maps of low-resolution image
n_2 feature maps of high-resolution image
$f_1 \times f_1$
$f_2 \times f_2$
$f_3 \times f_3$
Low-resolution image (input)
High-resolution image (output)
Patch extraction and representation
Non-linear mapping
Reconstruction

Figure 2.23 – SRCNN work flow, where the first convolutional layer of the SRCNN extracts a set of feature maps from low-resolution image **Y.** The second layer maps extract features nonlinearly to high-resolution patches. Finally, the last layer combines the predictions within a spatial neighbourhood to produce the high-resolution image F(**Y**).[65]

II) VDSR (Accurate Image Super-Resolution Using Very Deep Convolutional Networks)

VDSR single super-resolution(SR) image using very deep convolution networks is inspired by Simonyan and Zisserman [71] and proposed by Kim et al. [72] where the technique uses a very deep convolutional network to reproduce single super-resolution images (SSRI). This technique is regarded as deep due to the involvement of up to 20 weigh layers. Apart from the first and last layers, the used layers have the same 64 filters of size 3 x 3 x 64, where a single filter operates on a 3x3 spatial area across 64 channels or feature maps. The first layer operates on the input image data while the last layer used for reconstruction of SR image, taking in an interpolated low-resolution image as input data, then predicting image details. However, with very deep layers, convergence speed becomes a critical problem at the data training stage. The following issue is resolved by feeding only residual data to the network to learn, outperforming networks like SRCNN. In residual-learning, instead of using the exact copy of the input that goes through all layers until it reaches the output layer resulting in an end-to-end relation requiring very long-term memory often associated with SRCNN networks. The use of residual-learning to solve the vanishing gradient problem [73] is essential, where residual image r = y - x, where most values are likely to be zero or small as the input and output image are mostly similar. The VDSR network configuration is presented in Figure 2.24 below.

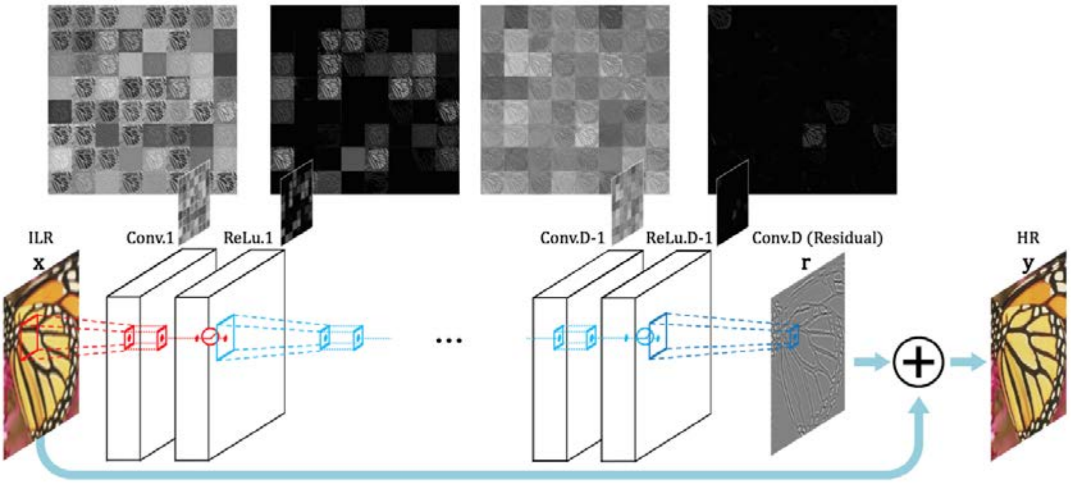


Figure 2.24 – VDSR network structure. Where an interpolated low-resolution image (ILR) goes through hidden layers to transform into a high-resolution (HR) image. The following network uses 64 filers for each convolutional layer, and some sample feature maps are drawn for visualization. Most features, after applying rectified linear units (ReLu) are zero. [72]

III) ESPCN (Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network)

Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network (ESPCN) are proposed by Wenzhe Shi et al. [74]. This network is the first to compute super-resolution videos of 1080p in real-time using a single K2 GPU. This is achieved by the extraction of feature maps in LR space and the introduction of an efficient sub-pixel convolution layer which learns an array of upscaling filters to upscale the final LR feature maps into the HR output. The ESPCN configuration is presented in Figure 2.25 below.

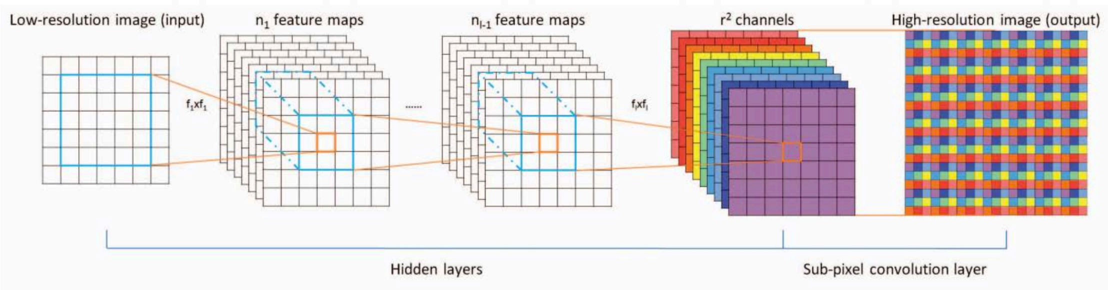


Figure 2.25 – Architecture of the efficient sub-pixel convolutional neural network (ESPCN), with two layers for feature maps extraction, and a subpixel layer that aggregates the feature maps from LR space and builds the SR image in a single step. [74]

On like the SRCNN and VDSR convolutional neural networks, the SISR estimates an HR image given an LR image downscaled from the corresponding HR image, instead of the usual upscaling to match the HR. The subsampling operation reproduces LR image from HR image by first convolving the HR image using a Gaussian filter (simulating the camera's point spread function) then subsampling the image by a factor ***r***. The following factor can also be referred to as the upscaling ratio. The first layer of the convolutional neural network is applied directly to the LR image. A sub-pixel convolution layer then upscales the LR feature maps to produce HR image.

IV) RDN (Residual Dense Network for Image Super-Resolution)

As most very deep convolutional neural network based on SR models do not make full use of hierarchical features from the original low-resolution (LR) images, thereby achieving to relatively low performance. A novel residual dense network (RDN) proposed by Yulun Zhang et al. [75] is used to address the following problem in image SR. The configuration of this RDN network is presented in Figure 2.26 below.

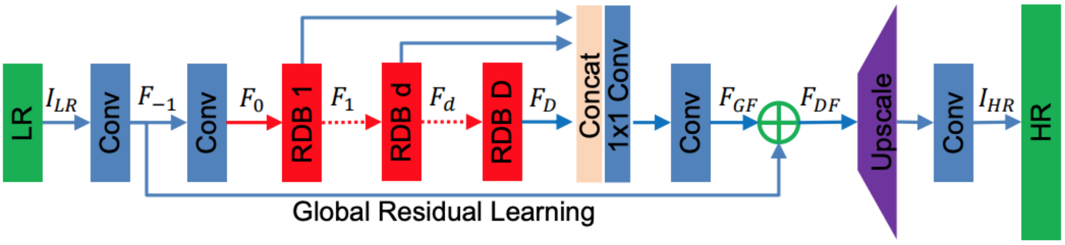


Figure 2.26 – Architecture of the proposed residual dense network (RDN) by Yulun et al. [75]

The proposed network fully exploits the hierarchical feature from all the convolutional layers. As presented in Figure 3.20 above, the RDN network mainly consists of four stages: shallow feature extraction net (SFENet), residual dense blocks (RDB), dense feature fusion (DFF), and lastly the up-sampling net (UPNet). The RDN network uses two convolutional layers to extract shallow features, where the first layer extract features from the LR input used for further shallow feature extraction and global residual learning. The second convolution layer extracts shallow features used as input for residual dense blocks (RDB). Specifically, the RDB is used to extract abundant local features through dense connected convolutional layers. The RDB further accommodates direct connections from the state of preceding RDB to all the layers of current RDB, resulting in contiguous memory (CM) mechanism. The fused local features in RDB is used to adaptively learn more effective features from previous and current local features and stabilizes the training of a wider network. After fully obtaining dense local features, the use of a global feature fusion to jointly and adaptively learn global hierarchical features is executed in a holistic way.

V) WDSR (Wide Activation for Efficient and Accurate Image Super-Resolution)

The expansion of features before ReLU activation layer without computational overhead results to better performance for single image super-resolution (SISR), when the same parameters and computational budgets are the same. The resulting WDSR network proposed by Jiahui Yu et al. [76] widens the slim identity mapping pathway to (2x to 4x) channels before activation in each residual block. This can be further widened to (6x to 9x) without computational overhead by the introduction of linear low-rank convolution into SR networks and achieving better accuracy-efficiency trade-offs. The configuration of the WDSR network is presented in Figure 2.27 below.

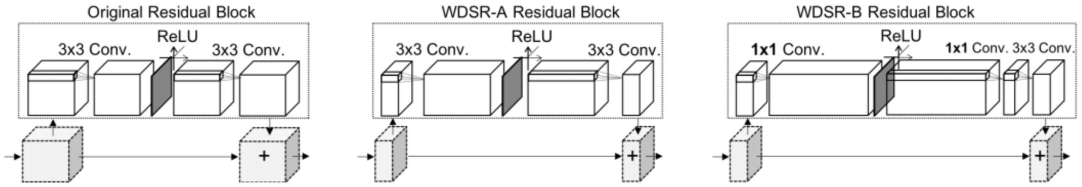


Figure 2.27 – Configuration of a residual block (left) widened before activation (middle) and further extended by the introduction of linear low-rank convolution into the SR network (right).[76]

2.4. H3D Evaluation Matric: PSNR and SSIM

This section presents the two image quality assessment techniques, namely Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [77], used in evaluation of image up-sampling techniques that where investigated to in order to improve the Holoscopic content adaptation technique. This is because any processing applied on an image may result to important loss of information or quality. In the case of depth estimation, this important information could be an important feature that might result to better feature matching.

Image evaluation techniques are classified either as objective or subjective based techniques [78][79]. Subject based image evaluation are based on human judgement without any explicit numerical criteria [80]. Object based image evaluation techniques require explicit numerical criteria and image reference or ground truths to evaluate an image[81][82]. The PSNR and SSIM techniques presented in this section is classified under this category. Object image evaluation techniques are mainly based on human judgement and require not explicit reference point.

The use of the objective based image evaluation techniques mentioned below, to analyse current interpolation techniques currently used for H3D up-sampling is key in the determination of the trade-off between angular and spatial information, details of the resulting evaluation presented in chapter three.

Given a reference image *r* and a target image *t,* both of size *MxN,* the PSNR between the two images *r* and *t* can be defined as:

$$PSNR(r,t) = 10log_{10}(\frac{255^2}{MSE(r,t)}) \qquad (10)$$

where

$$MSE(r,t) = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(r_{ij} - t_{ij})^2 \tag{11}$$

As the PSNR value approaches infinity the MSE approaches zero, the means that the higher PSNR value provides a higher image quality. Inversely, small PSNR values implies high numerical differences between images.

The SSIM developed by Wang et al. [83] is a popular image quality metric used to measure the similarity between two images and considered to be correlated with the quality perception of the human visual system (HVS). On like the PSNR, where traditional error summation techniques are employed, the SSIM is designed as a combination of the three factors that are loss of correlation, luminance distortion and contrast distortion. The SSIM is defined as:

$$SSIM(r,t) = l(r,t)c(r,t)s(r,t) \tag{12}$$

where

$$\begin{cases} l(r,t) = \dfrac{2\mu_r\mu_t + C_1}{\mu_r^2 + \mu_t^2 + C_1} \\ c(r,t) = \dfrac{2\sigma_r\sigma_t + C_2}{\sigma_r^2 + \sigma_t^2 + C_2} \\ s(r,t) = \dfrac{\sigma_{rt} + C_3}{\sigma_r\sigma_t + C_3} \end{cases} \tag{13}$$

The first term $l(r,t)$, in Equation (13) is the luminance similarity measure function, estimating the closeness of given image mean luminance (μ_r *and* μ_t). The following factor is maximal and equal to 1 only if $\mu_r = \mu_t$. The second term $c(r,t)$, is the contrast similarity measure, estimating the closeness of the contrast or the standard deviation (σ_r *and* σ_t) of the reference (r) and target image (t). The following is maximal and equal to 1 only if $\sigma_r = \sigma_t$. The last term $s(r,t)$, is the structure similarity measure between given images r and t, where σ_{rt} is the covariance between the given images. A result value of 0 means no correlation between images, and 1 means $r = t$, where [0,1] are the positive values of the

SSIM index. The position constants, $C_1, C_2 \, and \, C_3$, in equation (13) are used to avoid a null denominator.

As there are no exact rules for selection of the SSIM or PSNR when evaluation of images quality, both methods are used in this project as they together evaluate different aspects and feature of an image.

2.5. Summary

The following chapter is made up of four sections, where the principles of existing 3D system, 3D depth estimation techniques, learning-based up-sampling techniques and image evaluation metric are presented.

The 3D systems presented range from Active, Stereo/Multiview and Light field imaging system. The principles of the Active, Stereo and Multiview techniques are briefly discussed, the light-field imaging systems and its imaging principle is discussed in greater detail. The first section then concludes with the comprehensive examination of the H3D imaging system used in this project, where the direct comparison of the Lytro imaging system and Brunel H3D imaging system is presented. The Brunel Holoscopic 3D imaging system is used as the primary research 3D imagining system for recoding 3D image data due to cost-effective reasons, bigger recording sensor, easy accessibility its raw 3D image data, and easy ability to interchange micro lens arrays amongst a host of other reasons. However, as H3D imaging has not yet reached full maturity, there is a lack of a pre-processing platform for H3D data. Therefore, a detail investigation and evaluation of current processing H3D techniques is conducted, followed by the design and development of efficient pre-processing H3D techniques, all presented in chapter three.

Documentation of state-of-the art depth estimation technique is documented where the core principle and reasoning behind each technique is documented. This is done to identify the most suitable depth estimation techniques for the Holoscopic 3D image. Due to the one of the Holoscopic 3D image main drawbacks, which is its inability to extrapolate high resolution viewpoint images, a survey of deep learning-based super resolution techniques is explored. Concluding the chapter with state-of-the-art image evaluation metrics used in this project.

CHAPTER 3: HOLOSCOPIC 3D IMAGE PRE-PROCESSING

This chapter presents used and newly developed Holoscopic 3D imaging pre-processing techniques needed for H3D depth estimation. The motivation has been to implement scalable pre-processing techniques to address drawbacks associated with H3D data that affect H3D depth estimation. The chapter layout is as follows: 3.1 Introduction, 3.2 Distortion Error Correction, 3.3 Optimised Viewpoint Extraction Technique, 3.4 Evaluation of State-of-the-art Interpolation Techniques, 3.5 Proposed Deep-learning Framework for H3D Up-sampling and 3.6 Summary.

3.1. Introduction

3D depth estimation from a Holoscopic 3D image is a new research area. As a result, it has no standard 3D depth estimation framework. Specifically, there is no technique to extract 3D depth information directly from a Holoscopic 3D image apart from the innovative Direct 3D depth from Holoscopic (DDH) presented in chapter five. Current 3D depth estimation frameworks rely heavily on various pre-processing techniques to convert H3D images into stereo viewpoint images before 3D depth estimation. This chapter presents the proposed pre-processing techniques specifically developed for effective conversion of H3D image data to stereo image data. This chapter also evaluates current H3D pre-processing techniques and compares them against the proposed techniques presented in this chapter. The following make up the first and second block of this research contributions, shown in Figure 3.1 below, highlighted in red.

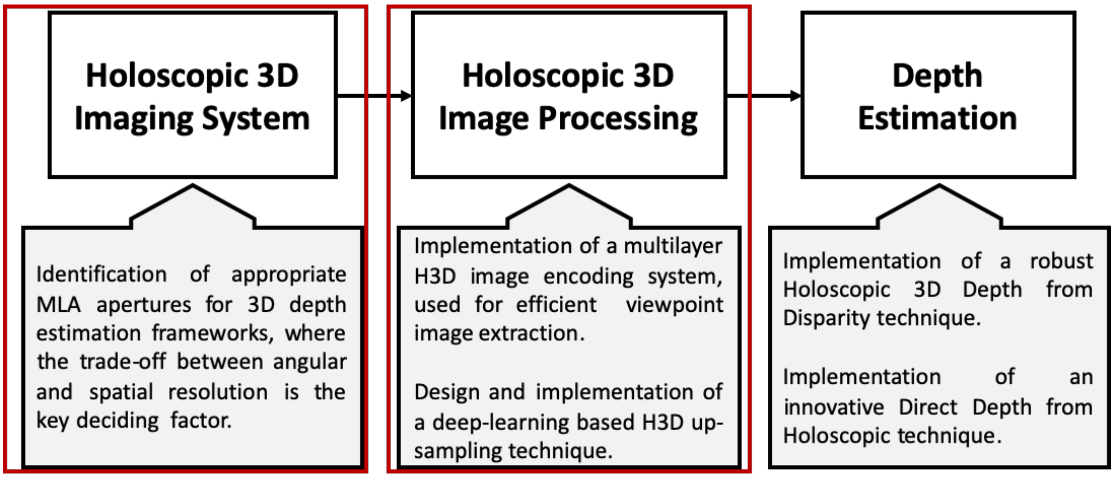

Holoscopic 3D Imaging System
Holoscopic 3D Image Processing
Depth Estimation
Identification of appropriate MLA apertures for 3D depth estimation frameworks, where the trade-off between angular and spatial resolution is the key deciding factor.
Implementation of a multilayer H3D image encoding system, used for efficient viewpoint image extraction.
Design and implementation of a deep-learning based H3D up-sampling technique.
Implementation of a robust Holoscopic 3D Depth from Disparity technique.
Implementation of an innovative Direct Depth from Holoscopic technique.

Figure 3.1 – The research contributions of this chapter, highlighted in red.

The following research contributions come as a result of tackling Holoscopic 3D imaging system-based errors caused by its unique image registration process. These image registration-based errors sometimes have an irreversible negative effect on estimated 3D depth maps. They are also extremely costly to accommodate during 3D depth estimation, should the proposed H3D image pre-processing techniques be ignored.

The Holoscopic 3D image registration principle is based on light diffraction, as shown in Figure 3.2 below.

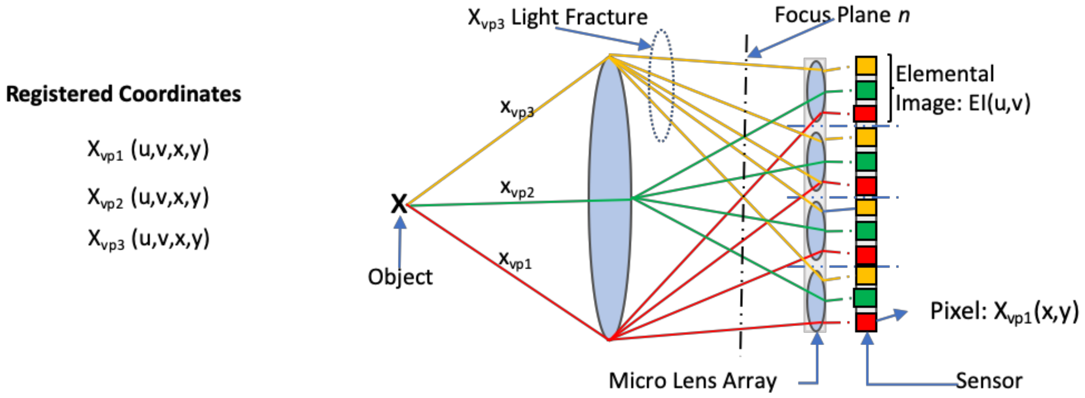


Figure 3.2 – H3D image registration principle, where the viewpoint image (X_{vp1} … $_{vp3}$) of object X is recorded by an H3D imaging sensor. Highlighting different refocusing plane that can be accessed in post-processing.

Figure 3.2 above presents the registration of light rays reflected off "object X". The reflected light is diffracted by the prime lens and then individually recorded by the MLA, resulting in the registration of multiple viewpoint images onto the imaging sensor of the H3D imaging system. Figure 3.2 also presents the focus plane that can be achieved computationally by the interpolation of the diffracted light rays, this is one of the unique abilities of the H3D imaging system and is fully utilized in the proposed H3D depth from disparity framework in Chapter four. The diffraction of light results to the three major drawbacks presented in Figure 3.3 below. These drawbacks inspired the design and development of the following pre-processing techniques presented in this chapter.

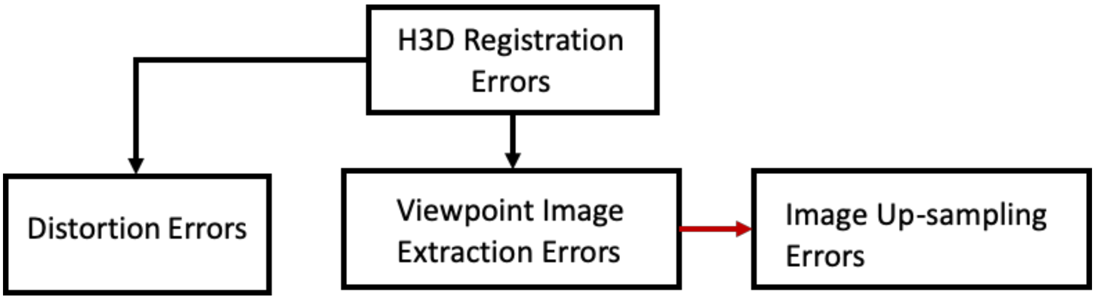


Figure 3.3 – Classification of H3D pixel registration errors that affect the integrity and increases the computational cost of estimating 3D depth information from H3D data.

The three main errors caused by the H3D system registration principle is presented in Figure 3.3 above. The first category, referred to as Distortion errors, is caused by all the lenses of the H3D imaging system. This class of errors cause the diffracted light to be registered at slightly different positions than intended on the H3D image sensor. This error can make the viewpoint extraction process more complex and sometimes impossible to locate all relating pixels needed to reconstruct a viewpoint image. However, existing techniques for reducing the amount of lens distortion errors presented in this chapter is incorporated in the HCA workflow presented in chapter four. The second class of error revolve around the bases of extracting reliable viewpoint images as well. Due to the current 2D image format, the H3D image is saved as, the extraction of viewpoint images is computationally expensive, especially when the complexity of the H3D image data increases. However, the proposed multi-layer H3D image encoding help reduces the computational cost of extracting viewpoint images by **95%.** The multi-layer encoding of the H3D image also aids the innovative Direct depth from Holoscopic (DDH) technique, as the location of EIs is clearly labelled. The third and final class of errors in the "Image up-sampling errors", which is a result of the current H3D imaging systems inability to record high-resolution viewpoint images. By up-sampling the Holoscopic 3D image or viewpoint image used for 3D depth estimation, unwanted artefact or noise is introduced as this can affect the reliability of resulting 3D depth maps. The first and second classes of error can be seen as low-level errors, but the third is a high-level error with leads to the detailed analysis of current interpolation-based techniques to deduce the rate at which noise is introduced into an extracted viewpoint image — the result of the following investigation leading to the proposal of a trade-off between angular and spatial resolution, and the design and suggestion of using learning-based networks to up-sample H3D data.

Before presenting the Holoscopic processing techniques needed for adaptation of 3D images to stereo 2D viewpoint images, the following scene preparation is to be considered and met.

i. Object distance: the object distance is essential as this influences the disparity size between the micro lens array.

ii. Lighting: the light has to be adequate enough for enabling the registration of reliable features. However, some low lighting images can also be recorded as it is easier to improve the overall illumination of medium-light images compared to overexposed images.

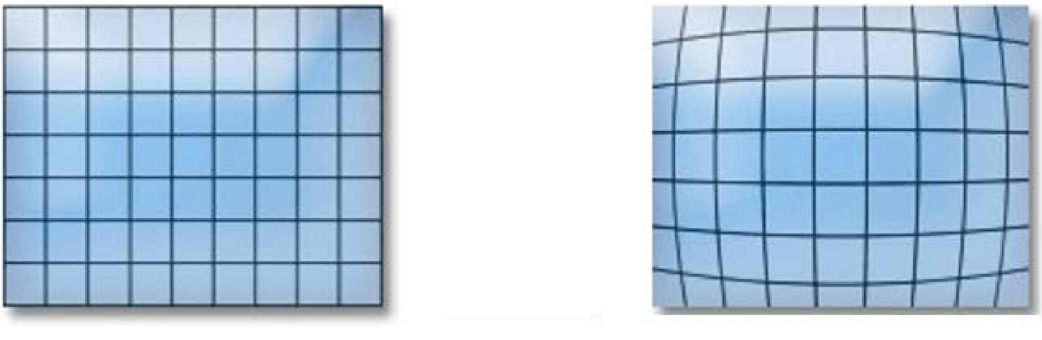
iii. Aperture size: the size of the square aperture also works in controlling the amount of light that is exposed to the H3D sensor, the right amount of light helps prevent ghosting or overlapping of elemental images been recorded.

3.2. Distortion Error Correction

Distortion errors are a result of the diffraction of light that passes through the Holoscopic 3D image camera lenses. The distortion errors associated with the H3D image data are listed below.

I. Barrel distortions:

This error occurs because the image captured is been fitted into a compact state. The error is most noticeable at the edges of the image and decreased as you draw closer to the centre, as shown in Figure 3.4 below. The barrel lens distortion error is easy to notice in an H3D image as one can track the borderlines that separate the array of EIs.



(a) Undistorted view (b) Barrel distortion view

Figure 3.4 – Barrel lens distortion effect often found in images.

II. Dark Moiré effect:

This occurs when a dark borderline is visible around the edges of the acquired H3D image, known as the "dark moiré effect" or vingnetting. This is as a result of the lack of light and the overall micro lens array size being too small to occupy the whole imaging sensor, this highlighted in Figure 3.5 below. When this occurs, the unwanted area of the image is cropped out.

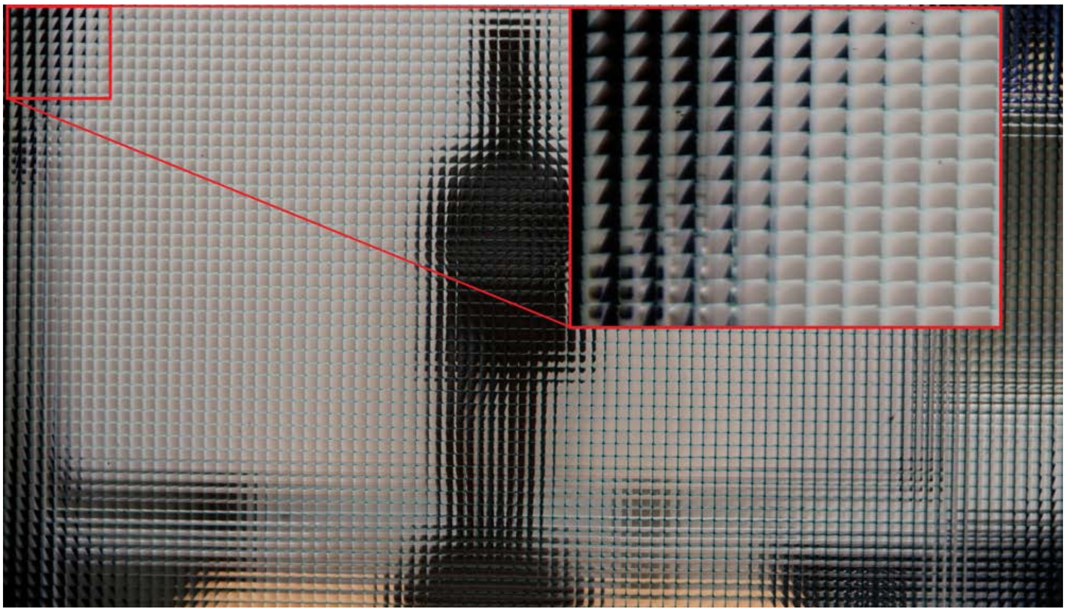


Figure 3.5 – Dark moiré effect usually found at the corner or edges of images.

The mentioned distortion errors above usually result in the mismatch of elemental images as well as the misplacement of viewpoint images. The following distortion can be corrected with two techniques. The first is the use of an advanced photo editing software such as Photoshop. This software has a built-in function that can automatically correct lens distortion, given the appropriate parameter. The second is the use of MATLAB lens correction function with the use of checkerboard images [84][85]. Subsection 3.2.1 below presents the principle behind the two motioned techniques used in the correction of lens-related errors.

3.2.1. H3D Distortion Minimization Techniques

I. Checkerboard technique

The popular checkerboard camera calibration technique [84][85] consists of the analysis of different checkerboard images to estimate the extrinsic parameters of an imaging system, shown in Figure 3.6 below. This technique of camera calibration is mostly used on stereo imaging systems and can be used not only to resolve distortion errors but also used to enforce the Epipolar constraint.

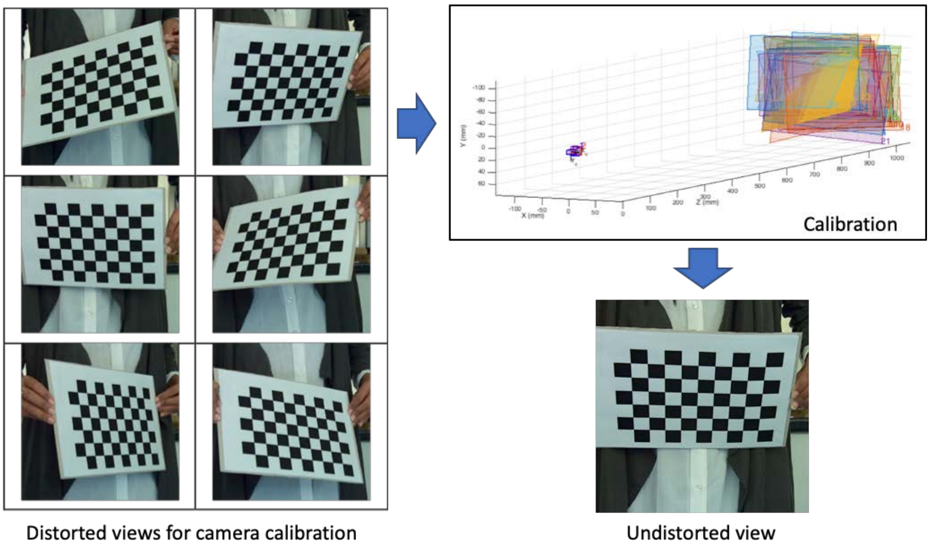


Figure 3.6 – Checkerboard distortion correction technique.[95][96]

II. Grid board technique

The Grid board technique also found in Photoshop corrects lens distortion errors by using a grid formulated based on the micro lens array and lens aperture segmentation grid. This grid is superimposed on the entire H3D image to skew the H3D image in such a way the H3D image grid and the imposed grid align uniformly. The grid-based technique is shown in Figure 3.7 below.

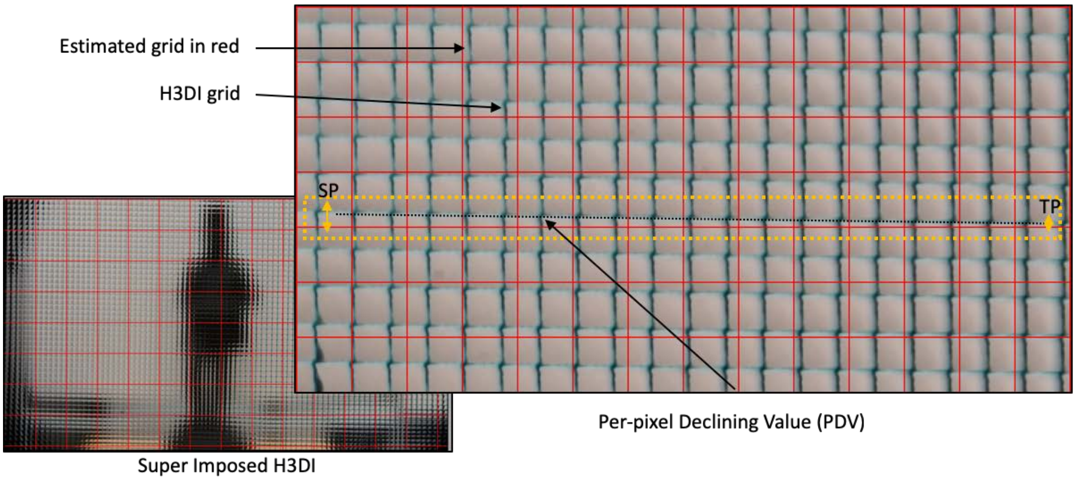
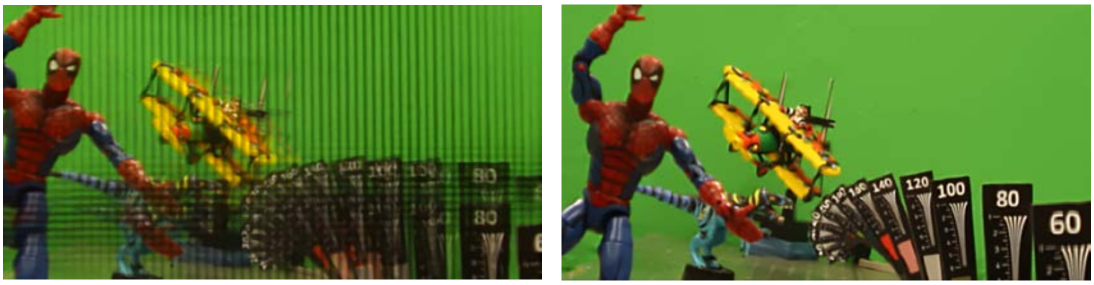


Figure 3.7 – Grid based distortion correction technique.

As mentioned earlier, the H3D image has been corrected of lens distortion by slight skew and transformation, to fix its distortion error. This technique is computationally less expensive and is easily incorporated in the proposed H3D depth estimation frameworks. Secondly, the PDV can also be used (should you skip the lens correction stage) later as a thresh hold to optimize the viewpoint extraction algorithm as it is extremely difficult to totally eliminate the distortions caused by the Holoscopic 3D image system lenses. Viewpoint images and their surrounding neighbours share similar PDV, which can be used to extracts accurate views without the need for high accuracy lens corrections. Effects of the lens distortion error before and after correction is presented in Figure 3.8 below.



(a) Before lens error correction (b) After lens error correction

Figure 3.8 – Viewpoint image extracted from H3D image before distortion correction in (a) and after distortion correction in (b).

3.3. Optimised Viewpoint Extraction Technique

Viewpoint extraction from a Holoscopic 3D image is a tedious process. It consists of the selection of various pixels from multiple EIs to reconstruct a viewpoint image. This is done because a viewpoint ray is usually split into multiple rays and recorded by the micro lens array. Current H3D image can be represented as ***H3DI = [H3DI (x, y)]*** where ***x*** and ***y*** are the horizontal and vertical positions of the H3D image pixels, shown in Figure 3.9 below.

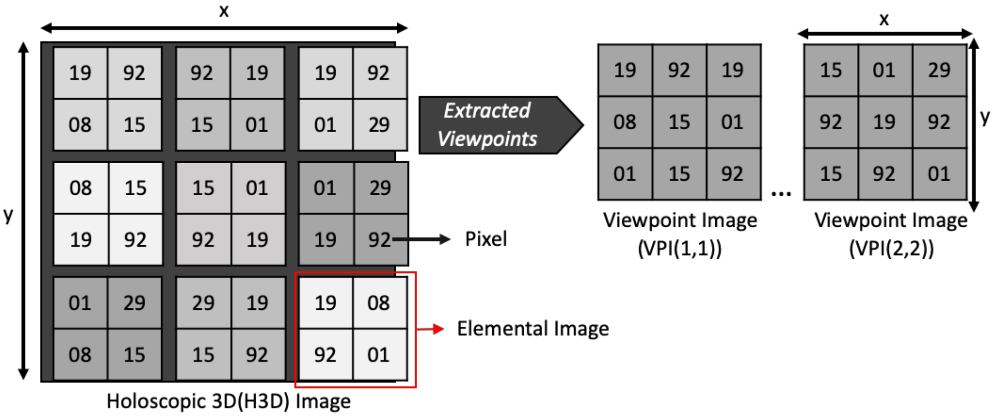


Figure 3.9 – H3D viewpoint extraction process. Producing viewpoint images of size equal to the number of Elemental images (EI).

Extraction of a viewpoint image from an H3D image, requires constant updating of pixel positional values due to the current 2D representation of the Holoscopic 3D image, resulting in more computational cost. The current viewpoint extraction process is defined in Equation (14) below,

$$VI(x, y) = \sum_{k=1}^{K}\sum_{l=1}^{L} H3DI_{k,l}(x + M, y + N) \qquad (14)$$

Where $\boldsymbol{x}$ and $\boldsymbol{y}$ represent the viewpoint image location to be extracted from the Holoscopic 3D image (H3DI), while *M* and *N* are constantly updated variables that ensures the selection of all relating pixel associated with the to be extracted viewpoint image specified *VI(x,y)*. As mentioned earlier, this techniques pixel updating cost compounds to cause high computation cost when the complexity H3D image increases. Specifically, when a large number of MLA is used for H3D data acquisition.

In order to solve the above problem and make the viewpoint extraction process more efficient and scalable, the multi-layer Holoscopic 3D image encoding system is proposed, relabelling the current H3D image to a 5D structure.

3.3.1. Multi-layer H3D Image Encoding for Efficient Viewpoint Extraction

As mentioned above the extraction of viewpoint images from an H3D image is a costly and tedious task, especially when the complexity of the H3D image is greatly increased. To reduce the computational cost that would have been gained if the encoding process is ignored, the conversion of the H3D image to a 5D Metrix is key. This Metrix reflects the actual structure of the H3D image. The extraction of stereo viewpoint image from the encoded H3D image improves computational efficiency due to the labelling all EIs that make up the H3D image, presented in Figure 3.10 below. This, in turn, reduces to overall computation cost for estimating H3D depth from disparity.

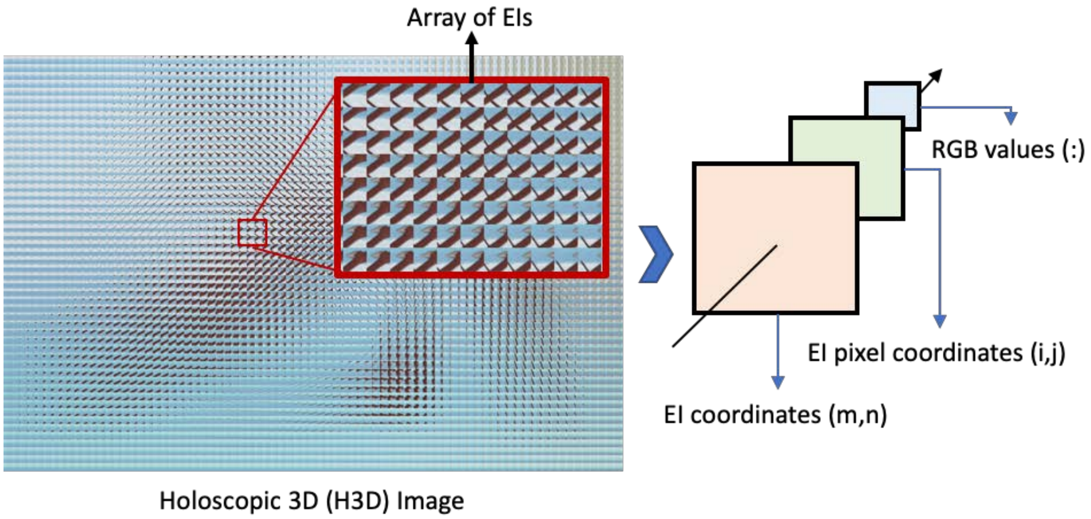


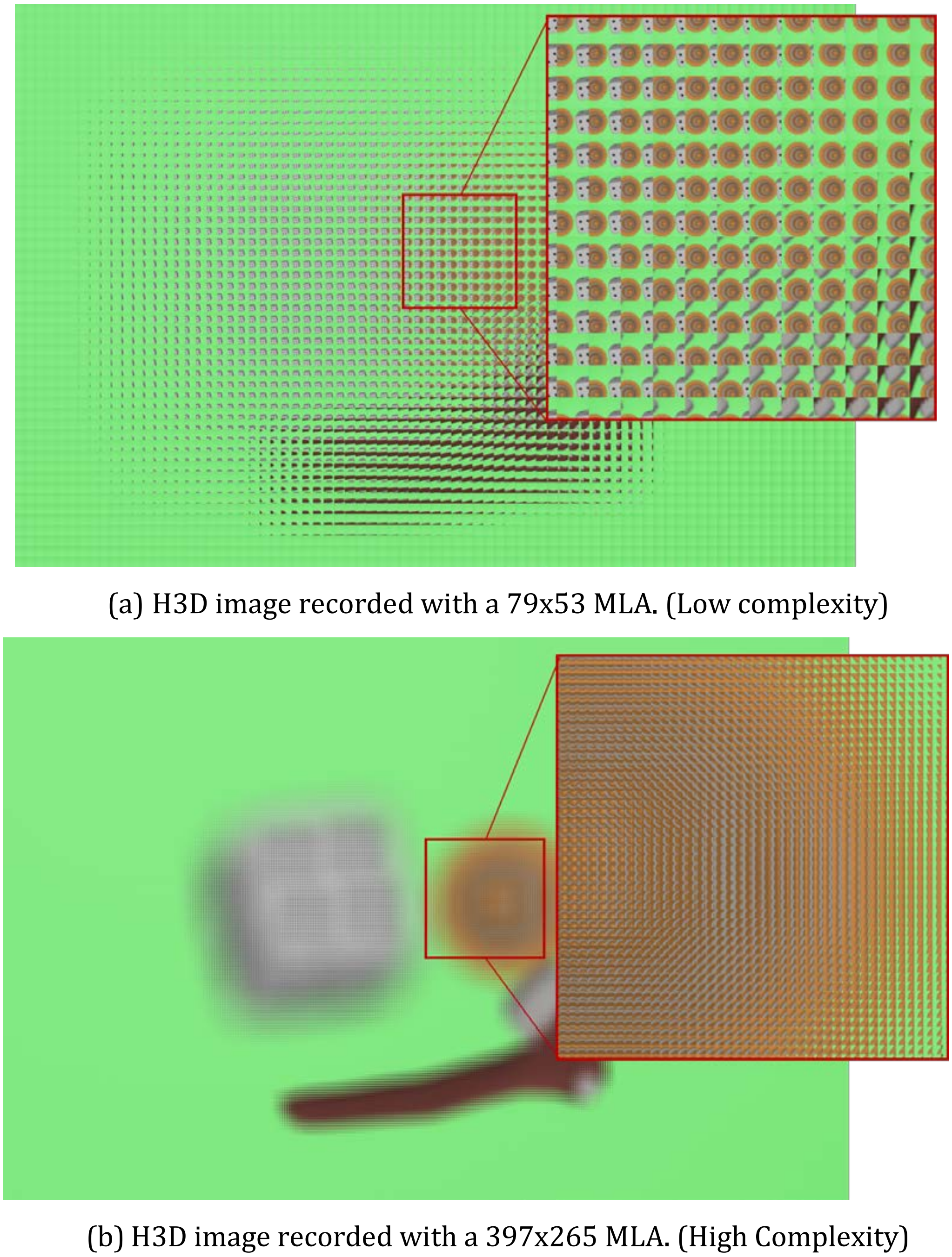
Figure 3.10 - Omnidirectional H3D image and multi-layer structural composition.

Figure 3.10 above reflects the actual structure of an H3D image, represented as $H3DI = [m, n, i, j, :]$, where the horizontal and vertical coordinates of the EI, viewpoint images and their RGB are all labelled. This multi-layer encoding system is more convenient for image processing purposes. It reduces the computational cost needed in the execution of H3D depth estimation formulations presented in chapter four and five. Equation (15) below is formulated to extract viewpoint images from this multi-layer H3D image format, as pixels at same location in each EI is selected to make up a viewpoint image.

$$VI(x, y) = \sum_{k=1}^{K} \sum_{l=1}^{L} EI_{k,l}(x, y) \tag{15}$$

Where *VI (x, y)* is the coordinate of the viewpoint image to be extracted from the array of elemental images ($EI_{k,\,l}$).

The cost of extracting VI using the technique presented in Equation (14) in comparison to the optimised technique represented in Equation (15) is presented below in Figure 3.12. However, Synthetic H3D images are used in testing the efficiency of both viewpoint extraction techniques presented in Equations (14) and (15). The SH3DIs are computationally generated in the Persistence of Vision Ray Tracer software, referred to as POV-RAY, which is written in C++ and developed by the POV-Team [86]. The use of POV-RAY serves as a cost-effective means of generating H3D data with different MLA specifications.



(a) H3D image recorded with a 79x53 MLA. (Low complexity)

(b) H3D image recorded with a 397x265 MLA. (High Complexity)

Figure 3.11 – Synthetic Holoscopic 3D image data set used for evaluating the optimized viewpoint extraction technique against the previous viewpoint extraction technique.

Figure 3.11 above presents the low and high complexity H3D data set used to test the scalability of the current and new viewpoint extraction techniques mentioned earlier.

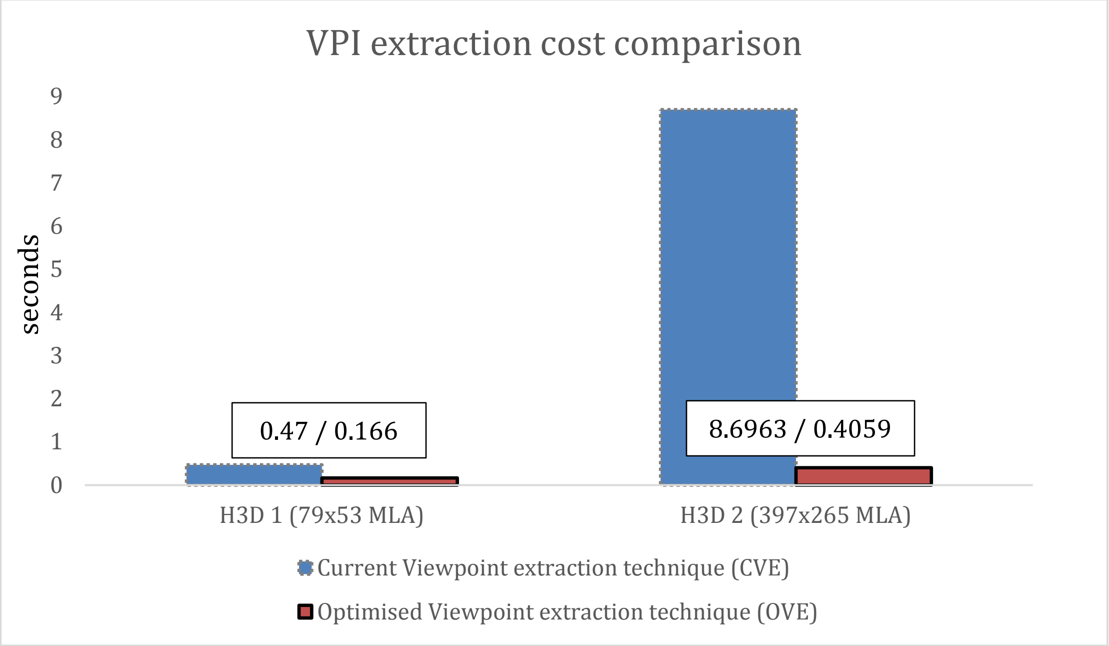


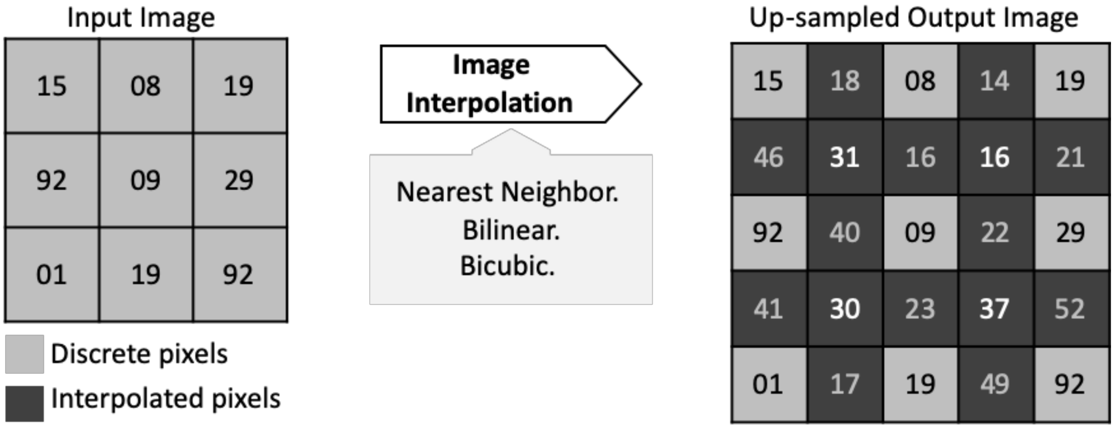
Figure 3.12 – Result indicating how well the optimised viewpoint extraction technique copes with complex H3D data compared to previous viewpoint extraction techniques.

Figure 3.12 indicates how well the new optimised viewpoint extraction technique presented Equation (15) scales when presented with complex H3D data, unlike the current viewpoint extraction technique represented Equation (14). However, Equation (14) and Equations (15) above presents the extraction of viewpoint images without any form of image up-sampling or interpolation.

To extract reliable viewpoint images with high spatial resolution for H3D depth estimation. Current up-sampling techniques are evaluated using industrial standard image quality assessment matrix PSNR and SSIM. The results from the evaluating led to the identification of the optimal trade-off between angular and spatial information and the proposal of a deep-learning-based H3D image up-sampling technique.

3.4. Evaluation of Current State-of-the-art Interpolation Techniques

Interpolation is the process of up-sampling or down-sampling an image from one resolution to the other without losing essential information by applying approximating continues functions on discrete samples [87].



15	08	19
92	09	29
01	19	92

15	18	08	14	19
46	31	16	16	21
92	40	09	22	29
41	30	23	37	52
01	17	19	49	92

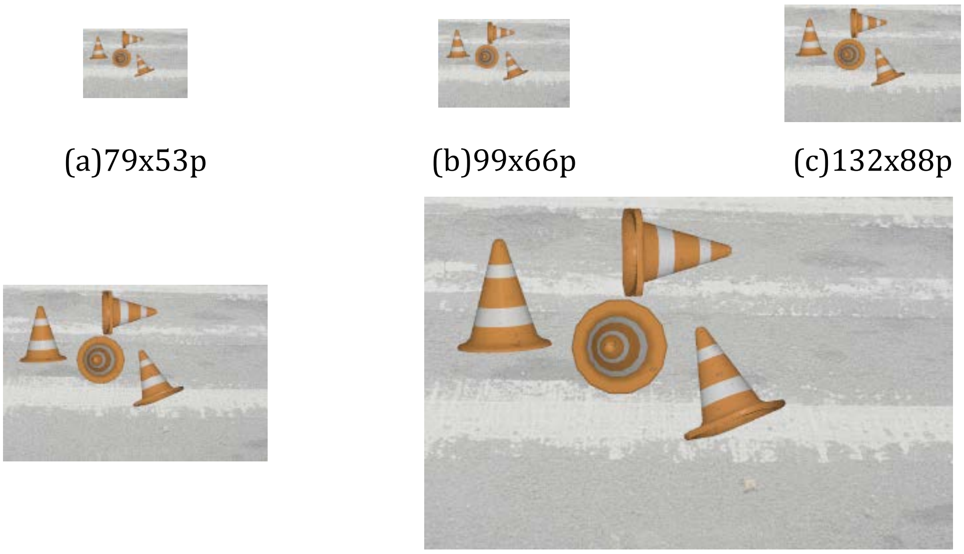
Figure 3.13 – Interpolation based image up-sampling principle.

Figure 3.13 above presents the basic concept behind image interpolation (up-sampling), where discrete pixels are interpolated to make up newly introduced pixels.

As the low-resolution problem is a significant limitation that affects the efficient estimation of depth information from an H3D image, the consideration of the optimal trade-off between angular and spatial resolution helps improve VI quality by increasing the default H3D image size and reducing the scaling factor. This section presents the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [77] evaluation result of various interpolation techniques used in up-sampling the H3D viewpoint images. This evaluation Metrix require an identical target image to evaluate the up-sampled viewpoint image. The targeted viewpoint resolution used for this viewpoint image quality assessment ranges from Standardbred definition (SD) to high definition (HD) as most 3D depth estimation data resolution are within that range [9][88][89]. The results of this evaluation help provide an estimate of the amount of noise or unwanted artefact been introduced in images in correlation to their scaling factor. Leading to the suggestion and consideration of the MLA size when recording H3D data for 3D depth estimation purposes.

The default viewpoint image is always the same as the microlens array size, and the total number of viewpoint images recorded by the Holoscopic 3D imaging system is the same as the size of the EI. It is well known in Stereovision and

Computer vision societies that 3D depth estimation techniques work better with high-resolution image data [90][23]. The reason behind this is due to the higher chance of extracting more reliable features for stereo correspondence. The inability to extract high-resolution viewpoint images is a significant drawback of the Holoscopic 3D imaging system, leading to the use of various interpolation technique to up-sample viewpoint images. However, interpolation techniques always introduce unwanted artefacts in the resulting VI, leading to the evaluation of state-of-the-art interpolation techniques in this section, such as the Bicubic, Bilinear and Nearest Neighbour techniques. POV-Ray light tracer is used in generating the various H3D image data of different MLA parameters used in evaluating the interpolation techniques mention. Figure 3.14 below presents the rendered SH3D image data.



(a)79x53p (b)99x66p (c)132x88p

(d)198x132p (e)397x265p

Figure 3.14 - SH3DIs used in evaluating state-of-the-art interpolation technques.

Subsections 3.4.1 to 3.4.3 presents the results of the image quality assessments of the various SH3D image data set ranging from 79x53 to 397x265 MLA sizes. A brief introduction of the techniques used for evaluation is presented.

3.4.1. H3D Image: Nearest Neighbour Interpolation

The nearest neighbour interpolation technique is the most basic and does not require much computational power. This technique does not interpolate pixels based on surrounding pixels but recreates duplicates of pixels closest to the linear position at which the intended interpolated pixel is to be placed. One of this technique drawback is that it introduces unwanted artefacts that cause up-sampled images to look pixelated. Table 3.1 below presents the *PSNR* and *SSIM* image quality assessment results of default Holoscopic viewpoint images up-sampled by the Nearest neighbour interpolation technique.

Table 3.1

Nearest neighbour PSNR assessment results.

VI resolution (Pixels)	**480x300p PSNR**	**720x480p PSNR**	**1920x1380p PSNR**
79x53	*26.6674*	*26.3288*	*26.0340*
99x66	*27.5794*	*27.1218*	*26.8026*
132x88	*28.6127*	*28.1307*	*27.7222*
198x132	*30.5036*	*29.5963*	*29.0477*
397x265	***33.5005***	***31.5856***	***30.7088***

Nearest neighbour SSIM assessment results.

VI resolution (Pixels)	**480x300p PSNR**	**720x480p PSNR**	**1920x1380p PSNR**
79x53	*0.6765*	*0.6970*	*0.8329*
99x66	*0.6990*	*0.7098*	*0.8349*
132x88	*0.7371*	*0.7357*	*0.8380*
198x132	*0.8116*	*0.7916*	*0.8504*
397x265	***0.9109***	***0.8674***	***0.8717***

The default viewpoint resolution is upscaled to the 480p, 720p and 1920p resolution using the Nearest neighbour interpolation technique. The quality assessment of the resulting image is assessed with PSNR and SSIM matrix.

3.4.2. H3D Image: Bilinear Interpolation

Bilinear interpolation averages the linear and neighbouring horizontal pixels to acquire the value of the desired interpolated pixel, providing better results compared to the Nearest neighbour interpolation technique. However, it is computationally more demanding. Depending on the scaling factor, the centre positions between the discrete pixel samples is interpolate first, and this process is repeated until the suggested positions for the interpolated pixels have neighbouring pixels at all directions. Table 3.2 below presents the *PSNR* and *SSIM* image quality assessment score of default Holoscopic viewpoint images up-sampled by the Bilinear interpolation technique.

Table 3.2

Bilinear PSNR assessment results.

VI resolution (Pixels)	**480x300p PSNR**	**720x480p PSNR**	**1920x1380p PSNR**
79x53	*27.6321*	*27.2607*	*27.7467*
99x66	*28.5997*	*28.1701*	*27.7467*
132x88	*29.8006*	*29.2245*	*28.7004*
198x132	*31.9171*	*31.0312*	*30.2765*
397x265	***37.7689***	***34.9186***	***33.2698***

Bilinear SSIM assessment results.

VI resolution (Pixels)	**480x300p PSNR**	**720x480p PSNR**	**1920x1380p PSNR**
79x53	*0.7221*	*0.7400*	*0.8653*
99x66	*0.7412*	*0.7527*	*0.8653*
132x88	*0.7729*	*0.7739*	*0.8732*
198x132	*0.8353*	*0.8211*	*0.8894*
397x265	***0.95279***	***0.9234***	***0.9323***

The default viewpoint resolution is upscaled to the 480p, 720p and 1920p resolution using the Bilinear interpolation technique. The quality assessment of the resulting image is assessed with PSNR and SSIM matrix.

3.4.3. H3D Image: Bicubic Interpolation

Bicubic interpolation is the most complex compared to the Bilinear and Nearest neighbour interpolation techniques. The Bicubic considers the weighted values of sixteen neighbouring pixels to predict the value of the desired interpolated pixel [91]. This technique is computationally more expensive than the Nearest neighbour and Bilinear technique. Table 3.3 below presents the *PSNR* and *SSIM* quality score of default Holoscopic viewpoint images up-sampled by the Bicubic interpolation technique.

Table 3.3

Bicubic PSNR assessment results.

VI resolution (Pixels)	480x300p PSNR	720x480p PSNR	1920x1380p PSNR
79x53	*27.9513*	*27.5726*	*27.2040*
99x66	*28.8830*	*28.4307*	*27.9944*
132x88	*30.0841*	*29.5295*	*28.9905*
198x132	*32.2609*	*31.4944*	*30.7356*
397x265	***38.8198***	***36.8267***	***35.0139***

Bicubic SSIM assessment results.

VI resolution (Pixels)	480x300p PSNR	720x480p PSNR	1920x1380p PSNR
79x53	*0.7188*	*0.7373*	*0.8591*
99x66	*0.7355*	*0.7472*	*0.8633*
132x88	*0.7682*	*0.7699*	*0.8711*
198x132	*0.8360*	*0.8254*	*0.8892*
397x265	***0.9705***	***0.9564***	***0.9530***

The default viewpoint resolution is upscaled to the 480p, 720p and 1920p resolution with the Bicubic interpolation technique. The quality assessment of the resulting image is assessed with PSNR and SSIM matrix.

3.4.4. Trade-off Consideration

Based on the results presented in Table 3.1 to Table 3.3, it is clear that the bigger the sampling factor, the greater the amount of unwanted artefact is introduced in any extracted viewpoint image. This, in turn, affects the quality of 3D depth maps because the unwanted artefacts make it more difficult to extract reliable features. Therefore, the use for smaller MLA sizes will reduce the scaling factor resulting in more quality viewpoint images. However, this comes at the cost of the angular information within the image. Furthermore, since the H3D depth from disparity does not need more than two viewpoint images to extract 3D depth information, the trade-off has less impact on the resulting 3D depth map.

The trade-off further investigated in chapter four, where the H3D depth from disparity framework presented, where a disparity test range is conducted. The graphs are shown below in Figure 3.15, and Figure 3.16 visually presents how the MLA sizes affect the image quality of the up-sampled default Holoscopic viewpoint images during H3D adaptation when using current up-sampling techniques.

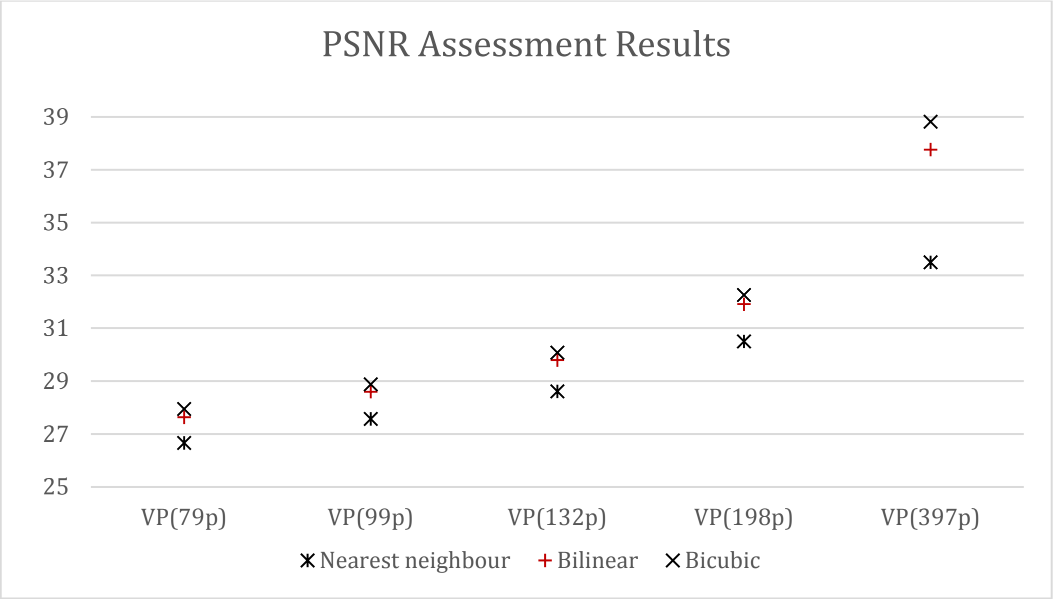


Figure 3.15 - Interpolated viewpoint image quality assessment results of PSNR. The default Holoscopic viewpoint image ranging from 79x53pixels to 397x265pixels(p), all upscaled to 480p.

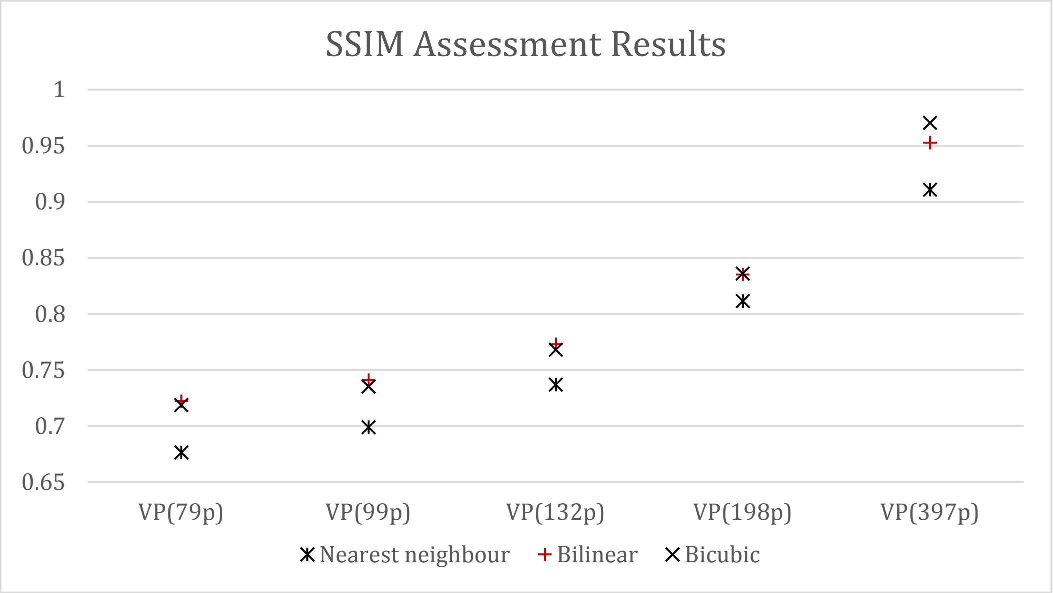


Figure 3.16 - Interpolated viewpoint image quality assessment results of SSIM. The default Holoscopic Viewpoint image ranging from 79x53pixels to 397x265pixels(p) all upscaled to 480p.

The H3D data used in this evaluation ranges from 79p to 397p, clearly shown in Figure 3.15 and 3.16 above. The results closer to 1 and 40 indicates high-quality up-sampling, which consist of a smaller scaling factor.

The next subsection presents the second proposed solution for addressing the low-resolution problem of the H3D imaging system.

3.4.5. Holoscopic 3D Deep-learning Configuration

This section presents the second solution for the low-resolution problem associated with the H3D imaging system. A Single Image Super-Resolution (SISR) technique that can be applied either on the entire Holoscopic 3D image or the extracted viewpoint images. The results of these deep-learning-based techniques are promising and could lead to future works where the full incorporation of this technique into the Holoscopic Content Adaptation framework developed.

Based on the following low-resolution problem that affects the 3D depth matching process, H3D training and remodelling layer is fitted to a CNN. The training layer extracts EIs from the H3D image and uses it as training data, while the remodelling layer reconstructs the up-sampled elemental images into a single super-resolution H3D image. The learning-based up-sampling technique requires reference data for each training data set. The training data is a mixture of down-sampled segments of 2D images and EIs, using the default data mix as reference data. The residual difference is then learned by the network to produce HD elemental images that are reconstructed into an SRH3DI. The proposed deep-learning configuration for SH3D up-sampling is presented in Figure 3.22 below.

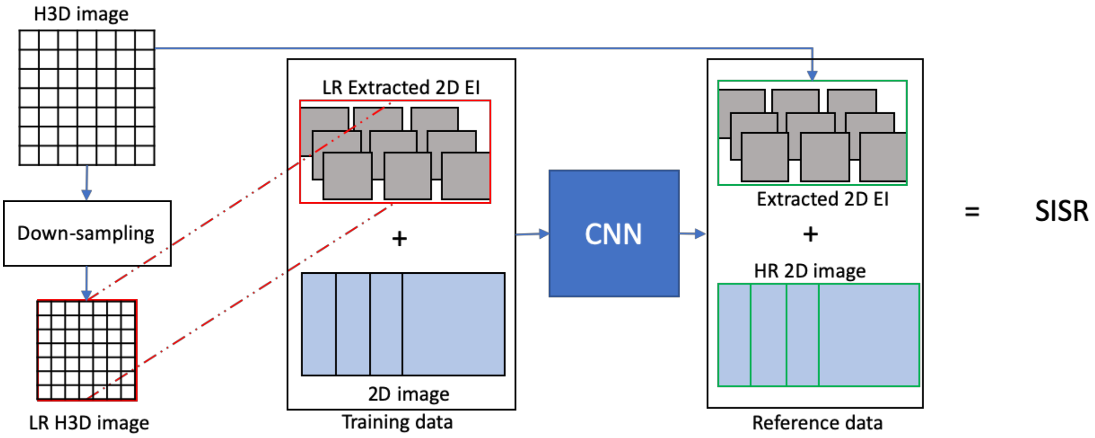


Figure 3.17 – Architecture of the proposed deep-learning network for single H3DI up-sampling.

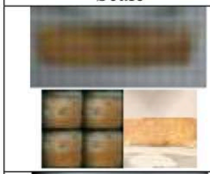
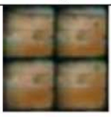
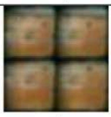
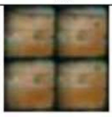
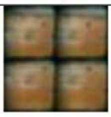
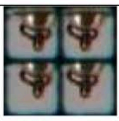
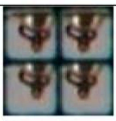
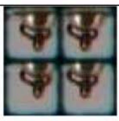
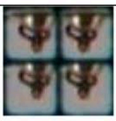
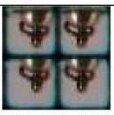
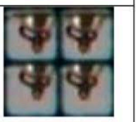
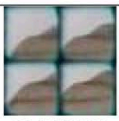
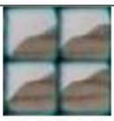
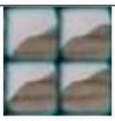
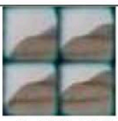
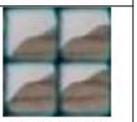
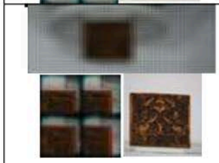
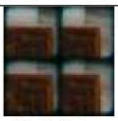
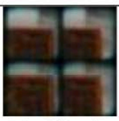
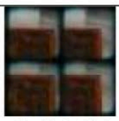
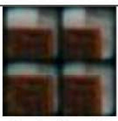
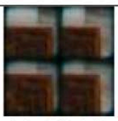
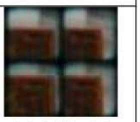
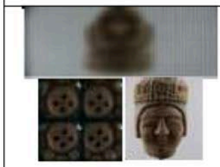
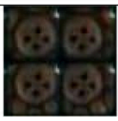
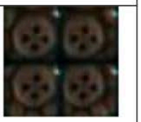
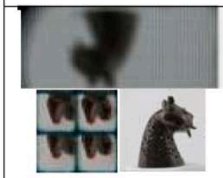
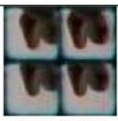
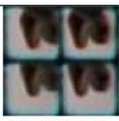
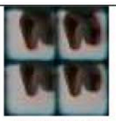
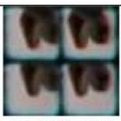
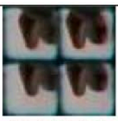
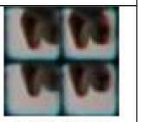
The proposed learning-based network uses additional 2D images segments, down-sampled to match the size of the corresponding EI data due to lack of H3D database. The results presented in the next section proves the lack of H3D training data does not affect the quality of the up-sampled H3D image as previously expected.

I) Results

Table 3.4 below presents the results obtained from the proposed configuration presented in Figure 3.22 above. The image quality is assets using PSNR and SSIM, where the feature and structural qualities are accessed and evaluated against their respective reference images. The images are scaled from two to four times its original size and values on the bottom left are PSNR results while values on the right are SSIM evaluation result.

Table 3.4

Quality comparison of pretrained SR models on a selection of H3D data, using PSNR and SSIM.

	Bicubic		SRCNN		VDSR	
Scale	×2	×4	×2	×4	×2	×4
	 38.87/0.891	 34.11/0.792	 39.12/0.901	 35.39/0.803	 40.09/0.917	 35.19/0.830
	 39.58/0.911	 35.10/0.817	 40.31/0.823	 35.67/0.793	 41.71/0.897	 36.21/0.819
	 39.15/0.893	 35.27/0.841	 40.12/0.857	 37.12/0.806	 40.02/0.936	 37.63/0.871
	 38.89/0.875	 33.73/0.748	 39.11/0.900	 35.12/0.734	 39.59/0.843	 36.66/0.793
	 39.01/0.918	 34.52/0.852	 40.03/0.910	 35.43/0.817	 40.19/0.918	 35.73/0.858
	 39.21/0.881 PSNR / SSIM	 34.09/0.858 PSNR / SSIM	 40.14/0.895 PSNR / SSIM	 34.57/0.785 PSNR / SSIM	 40.36/0.920 PSNR / SSIM	 36.01/0.829 PSNR / SSIM

The results presented in Table 3.4 above are promising; this serves as encouragement for the incorporation of deep learning-based techniques in H3D viewpoint extraction. However, the proposed deep-learning technique is not yet optimized to extract viewpoint images at different focal planes. Therefore, the patch-based viewpoint extraction technique, derived from the Bicubic and Bilinear, is used for extracting viewpoint images for the proposed H3D 3D depth estimation technique presented in chapter four.

3.5. Summary

This chapter presents the various Holoscopic pre-processing techniques that can be used to aid the extraction of reliable viewpoint images for H3D depth estimation. The motivation behind the design and development of the following H3D pre-processing techniques mentioned is to counterattack the significant limitations associated with the H3D recording principle.

The most obvious limitation is the H3D inability to reproduce HD viewpoint images. However, the process of extracting the viewpoint images itself is a complicated task. Current viewpoint extraction techniques cannot efficiently scale whenever handed the task of extracting viewpoint images from complex H3D data. This chapter presents a more effective viewpoint extraction technique that is a result of the multi-layer H3D image encoding system, where the EI images are stacked and clearly labelled. As pixel interpolation is also an integral step in viewpoint extraction, evaluation of current state-of-the-art single image interpolation techniques is presented in the above chapter. The results of the evaluation led to the deduction of the optimal trade-off between angular and spatial information for H3D depth from disparity techniques. The chapter also presents a deep-learning-based technique that uses H3D and 2D images as training data, resulting in up-sampling of high-resolution EIs that are reconstructed to produce super-resolution H3D images.

The techniques are used in the next chapter to create a robust Holoscopic Content Adaptation framework that is capable of extracting viewpoint image from low-resolution H3D data and scales comfortably when tasked with the extraction of viewpoint images from more complex H3D data. The following HCA framework consists mainly of three stages, namely; ii) Distortion Error Correction, ii) Multi-layer H3D Encoding, and iii) Viewpoint Image Extrapolation.

CHAPTER 4: HOLOSCOPIC 3D DEPTH ESTIMATION FROM DISPARITY

This chapter presents the proposed framework for Holoscopic 3D depth from disparity. The framework consists of the most compatible 3D depth techniques with H3D data and a specially developed H3D content adaptation technique. The chapter layout is as follows: 4.1 Introduction, 4.2 Holoscopic Content Adaptation, 4.3 Similarity Measure and Depth Estimation, 4.4 Results and Evaluation, 4.5 Smart 3D Depth Mapping Design, and 4.6 Summary.

4.1. Introduction

3D depth estimation has a wide application area such as Space exploration, AR/VR, Robotics, Autonomous navigation, and Biomedical science. 3D depth from disparity is a technique that estimates 3D depth information by calculating the difference in distance between corresponding pixels of a stereo viewpoint image. The technique presented in this chapter can be classified as a semi-global technique, as it does not take into account the matching costs of the entire image when estimating 3D depth. The resulting 3D depth information is stored in a 2D depth map array that contains the ***z*** values of each pixel from the reference viewpoint image. The following Holoscopic 3D depth from disparity (H3DDD) framework consists of three main stages, namely i) Holoscopic Content Adaptation (HCA), ii) Feature Matching and iii) Disparity Estimation and Optimisation. However, Feature Matching, Disparity Estimation and Optimisation often work hand in hand and is discussed under similarity measure, and depth estimation presented Figure 4.1 below.

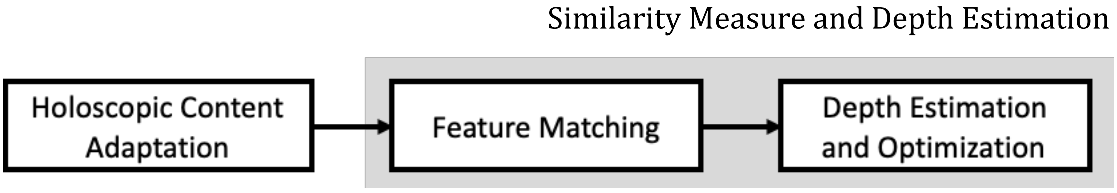

Similarity Measure and Depth Estimation
Holoscopic Content Adaptation
Feature Matching
Depth Estimation and Optimization

Figure 4.1 -The Holoscopic 3D depth from Disparity (H3DDD) Framework.

The Holoscopic Content Adaptation (HCA) stage of the proposed H3DDD framework presented in Figure 4.1 above converts H3D image to stereo images for successful estimation of 3D depth. The HCA stage is crucial as current 3D depth estimation frameworks [92][93][80] find it tedious to estimate reliable 3D depth information directly from a Holoscopic 3D image. In cases where the successful extraction of stereo viewpoint is accomplished, there is no indication of feature quality control. This leads to the use of inconsistent 3D depth optimisation techniques, making current 3D depth frameworks lack stability. However, the H3DDD framework presented in this chapter takes into consideration MLA specifications to improve the quality of viewpoint images extracted. Therefore, improving the overall chances of estimating reliable 3D depth maps. The proposed framework also takes advantage of the Holoscopic 3D system ability to extract viewpoint images of different focal planes after capture, where current frameworks do not. This results in the extraction of semi segmented viewpoint images that do not only result in high feature matching results but also results in the estimation of 3D depth information from stereo viewpoint images with smaller baselines.

This chapter also presents a detail evaluation of the proposed framework and the Holoscopic 3D dataset used during evaluation. Testing the frameworks disparity range, capability and scalability capabilities. The chapter then concludes with the design of a smart 3D depth calibration framework for Holoscopic 3D depth from Disparity techniques, aimed to reduce the computational cost of the technique.

The following is part of the third block of this research's contribution, shown in Figure 4.2 below, highlighted in red and emphasised in green.

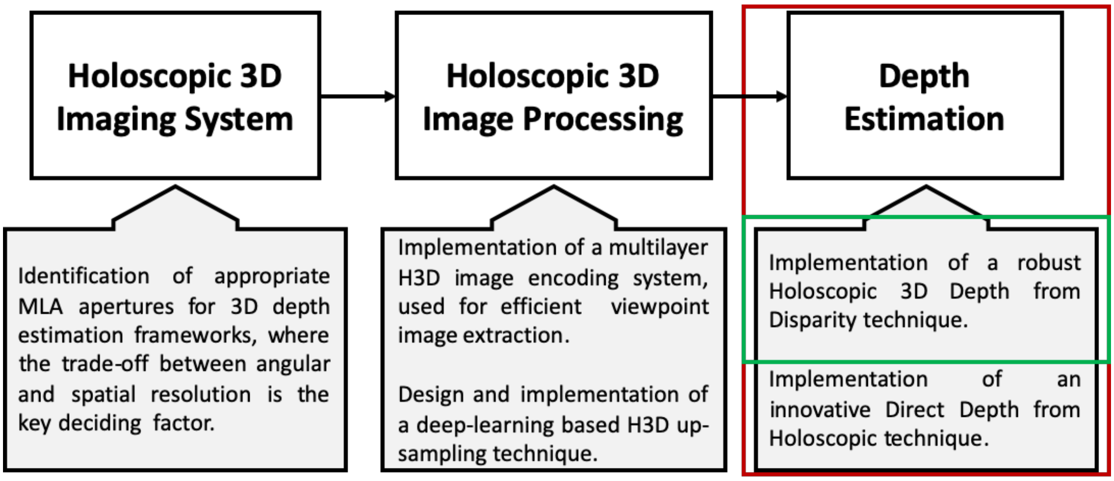

Holoscopic 3D Imaging System
Holoscopic 3D Image Processing
Depth Estimation
Identification of appropriate MLA apertures for 3D depth estimation frameworks, where the trade-off between angular and spatial resolution is the key deciding factor.
Implementation of a multilayer H3D image encoding system, used for efficient viewpoint image extraction.
Design and implementation of a deep-learning based H3D up-sampling technique.
Implementation of a robust Holoscopic 3D Depth from Disparity technique.
Implementation of an innovative Direct Depth from Holoscopic technique.

Figure 4.2 – The research contribution of this chapter highlighted in red and emphasised in green.

4.2. Holoscopic Content Adaptation

The Holoscopic content adaptation (HCA) stage focuses on converting Holoscopic 3D images into suitable stereo image data that can be used in 3D depth estimation. The workflow of the HCA process is presented in Figure 4.3 below, where the workflow consists of three significant steps.

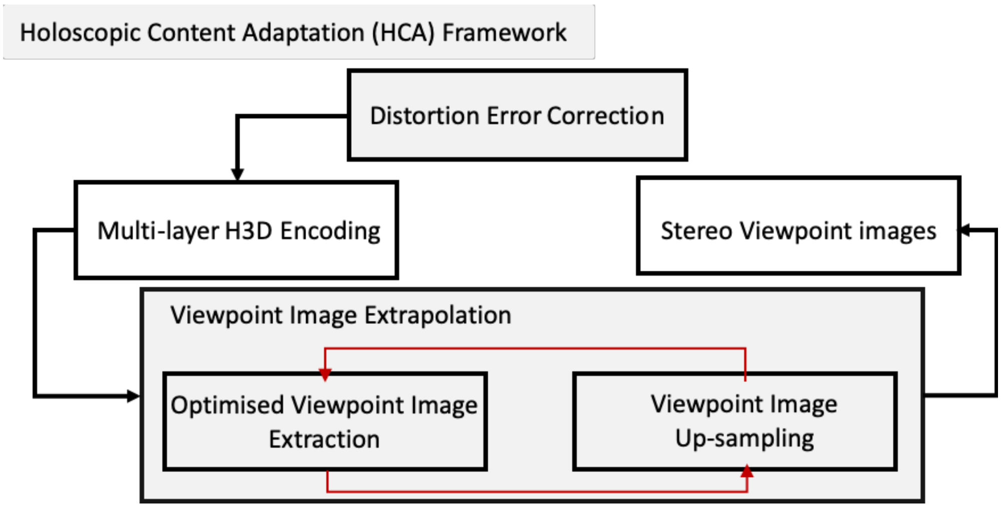

Holoscopic Content Adaptation (HCA) Framework
Distortion Error Correction
Multi-layer H3D Encoding
Stereo Viewpoint images
Viewpoint Image Extrapolation
Optimised Viewpoint Image Extraction
Viewpoint Image Up-sampling

Figure 4.3 – The HCA process found in the H3DDD framework.

HCA process presented in Figure 4.3 above consist of the three following steps; i) Distortion Error Correction, ii) Multi-layer H3D Encoding, and iii) Viewpoint Image Extrapolation.

Distortion Error Correction as the name implies is the process of reducing or eliminating all barrel and dark moiré errors. Distortion errors result in inconsistent registration of viewpoint image pixels across the entire H3D image. Without the correction of these distortion errors, 3D depth estimation is near impossible as the process of extracting reliable viewpoint images is rendered extremely complex. Chapter three presents the detail explanation of the Grid board technique used in the lens correction process of all H3D data used in this research.

Secondly, the encoding of the Holoscopic 3D image is conducted to reduce the computational cost acquired during viewpoint extrapolation. The multi-layer H3D formatting process converts the Holoscopic 3D image into a 5D matrix, labelling all the Elemental images and its respective pixels. Chapter three-section three contains the full details of the mentioned multi-layer H3D encoding process.

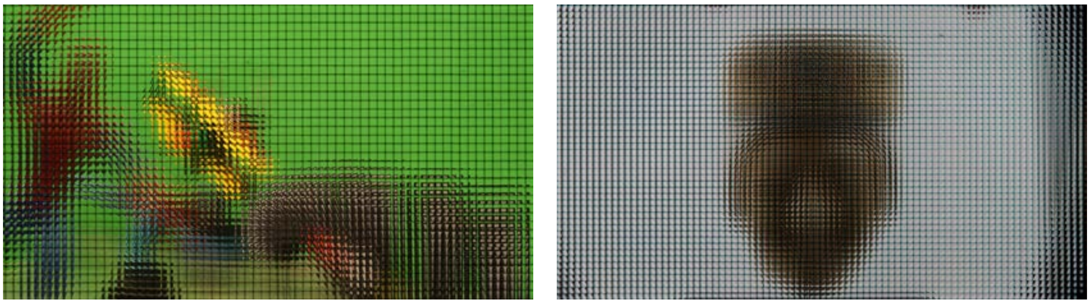
Lastly, the viewpoint extrapolation process, where stereo viewpoint images with an adequate spatial resolution (320p and above) are extracted from a single H3D image. Since the resolution of viewpoint images is directly proportional to the number MLAs used during capture, viewpoint images have very low spatial information. As a result, the extraction and matching of reliable features become extremely complex. The viewpoint extrapolation stage up-samples viewpoint images with a patch-based technique. This patch-based technique acquires the neighbouring viewpoint pixels around the pixels of the intended viewpoint image to be extracted. The extracted window of pixels is then interpolated and smoothed over with a Gaussian blur. It is wise to note that the patch-based uses as little image filter as possible as over smoothing of the viewpoint images, can confuse the 3D depth estimation algorithm, especially at the matching cost stage. This is because of the lack of sharp difference between pixels will make it difficult to extract even the easiest edge features within an image.

The next subsection presents the H3D data used in the evaluation and explanation of the proposed H3DDD framework.

4.2.1. H3D Data Set and Extracted Viewpoint Images

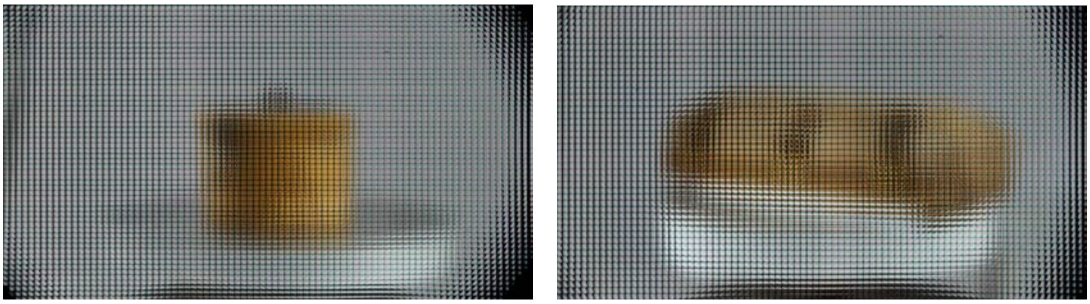
The subsection presents all H3D images and their extraction parameters used in this chapter for evaluating the proposed H3D depth from disparity framework.

The H3D data presented in Figure 4.4 below is captured with the Brunel H3D image system, sensor size of *35x24mm* and a dot pixel pitch of *0.00451mm*. The Spiderman data is captured in the Brunel University lab under normal lighting conditions. However, the Bighead, Vase and Flax are H3D image recordings of some of Qatar cultural asset, only available at their national museum.



(a) Spiderman: (64x34 MLA) 5160×2743p

(b) Bighead: (85x45 MLA) 7916×4412p



(c) Vase: (85x56 MLA) 7952×5232p (d) Flax: (85x47 MLA) 7936×4408p

Figure 4.4 - Holoscopic 3D dataset, captured in Brunel lab and Qatar national museum. Highlighting their MLA specifications and image size.

Figure 4.4 above presents the Holoscopic 3D data set used, highlighting MLA size and image resolution. The maximum number of viewpoint images that can be extracted from the following data set is the same as the size of their respective elemental image. However, for this H3D depth estimation framework, the disparity is calculated only from a single direction. Therefore, the maximum number of viewpoint images available for H3D depth estimation is the same as the size of the EI in one axis (x or y). This value also serves as the maximum disparity size of the Holoscopic 3D image in theory. However, depending on the patch size, only a limited amount of viewpoint images can be extracted, meaning only a limited amount of disparity can be assessed regardless of the maximum disparity.

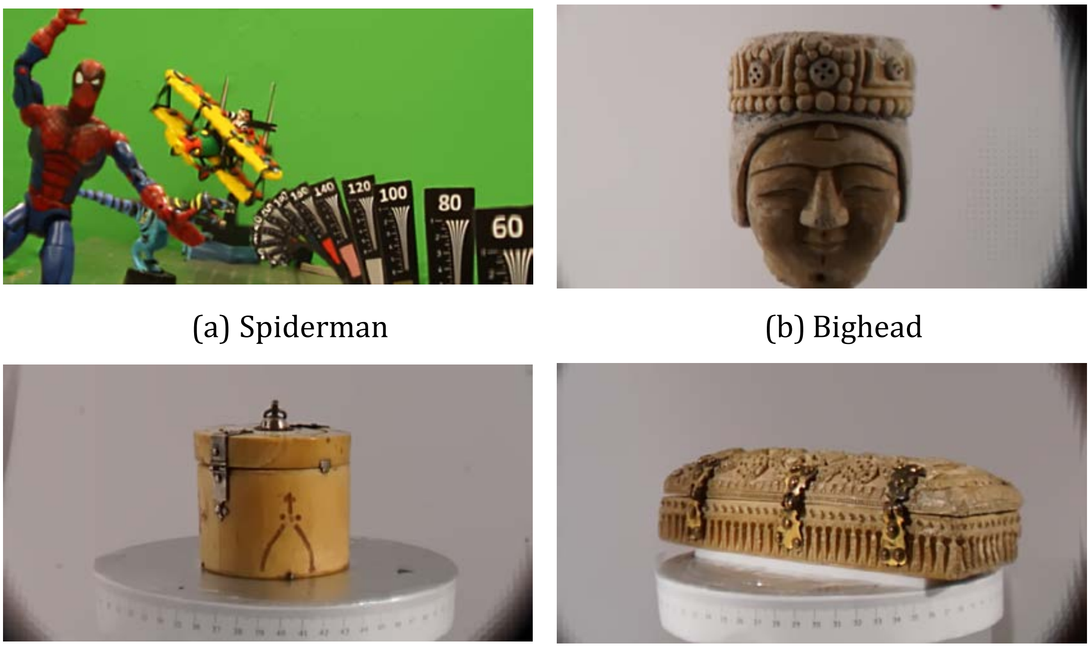
Table 4.1 below presents the starting variables of all the extracted viewpoint images, using the proposed Holoscopic Content Adaptation framework mentioned earlier.

Table 4.1

Holoscopic viewpoint extraction variables for all reference viewpoint images used in estimation of H3D depth through disparity.

	Number of VI	**VI x-axis location**	**VI y-axis location**	**Patch size *(win)***	**x-shift *(xs)***	**y-shift *(ys)***
Spiderman	82	16	29	13	-0.6	-0.7
Bighead	93	16	29	13	0.3	0.9
Vase	92	10	40	13	1.7	1.2
Flax	92	20	40	13	1.7	1

Samples of the viewpoint images extracted with the following parameters above are presented in Figure 4.5 below, where stereo views of the following image samples will be used as our prime stereo input data.



(a) Spiderman

(b) Bighead

(c) Vase

(d) Flax

Figure 4.5 - Holoscopic viewpoint image samples extracted from the Holoscopic 3D data set presented in Figure 4.4 above.

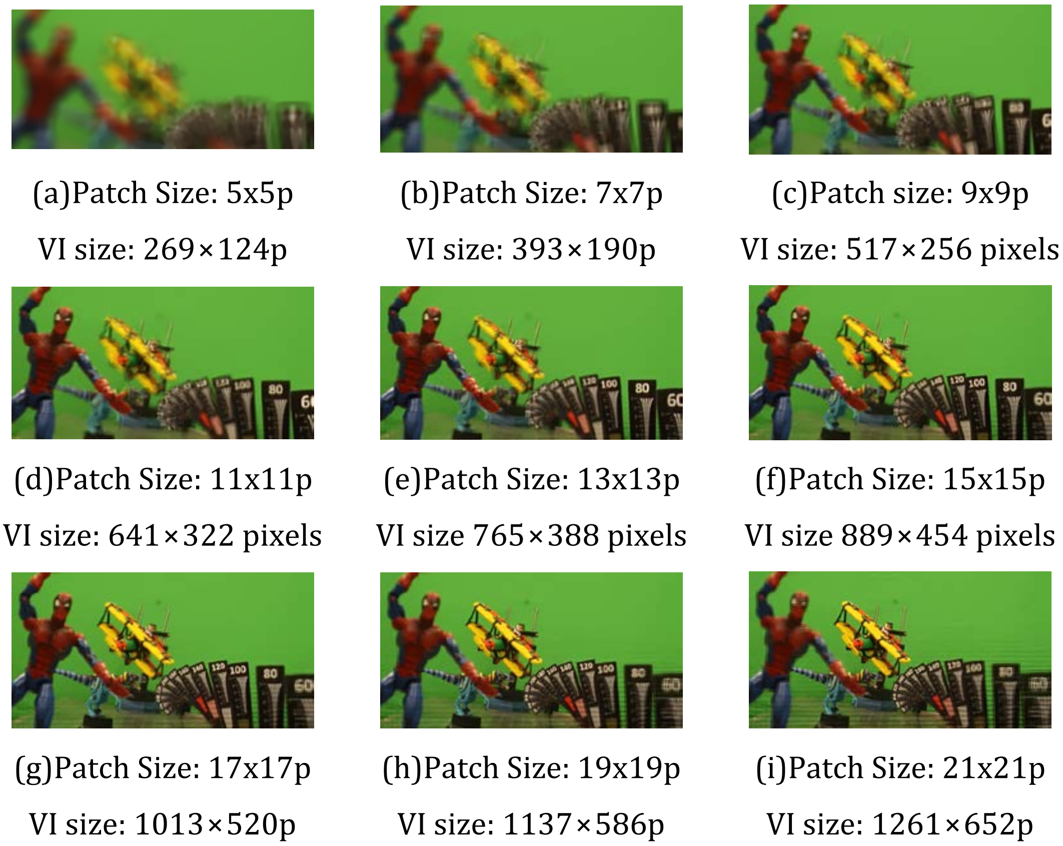
Table 4.2 below presents the average computational cost in seconds taken to extract the following viewpoint images, patch size ranging from 9p to 21p.

Table 4.2

The computational cost taken in seconds to extract viewpoint image from patch sizes 9p to 21p.

Window size (pixels)	Time taken in seconds (secs)
9x9	5.2047 secs
13x13	9.3444 secs
15x15	12.2895 secs
17x17	14.9178 secs
19x19	18.2112 secs
21x21	21.8869 secs

Figure 4.6 below presents the VIs extracted with varying patch sizes, detailing the VI resolution in pixels (p).



(a)Patch Size: 5x5p
VI size: 269×124p

(b)Patch Size: 7x7p
VI size: 393×190p

(c)Patch size: 9x9p
VI size: 517×256 pixels

(d)Patch Size: 11x11p
VI size: 641×322 pixels

(e)Patch Size: 13x13p
VI size 765×388 pixels

(f)Patch Size: 15x15p
VI size 889×454 pixels

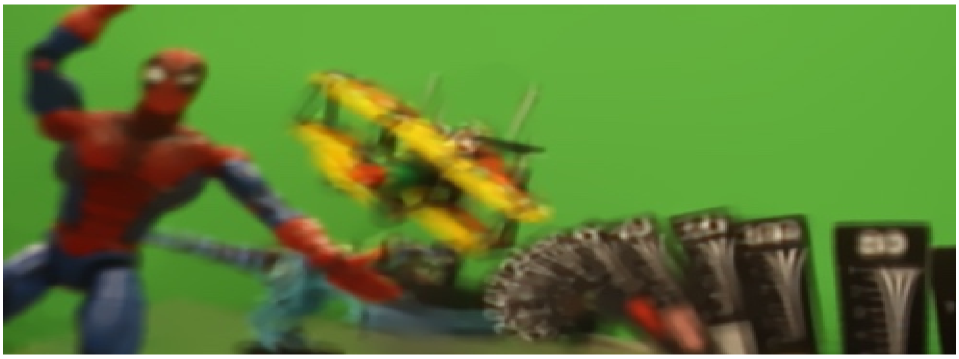
(g)Patch Size: 17x17p
VI size: 1013×520p

(h)Patch Size: 19x19p
VI size: 1137×586p

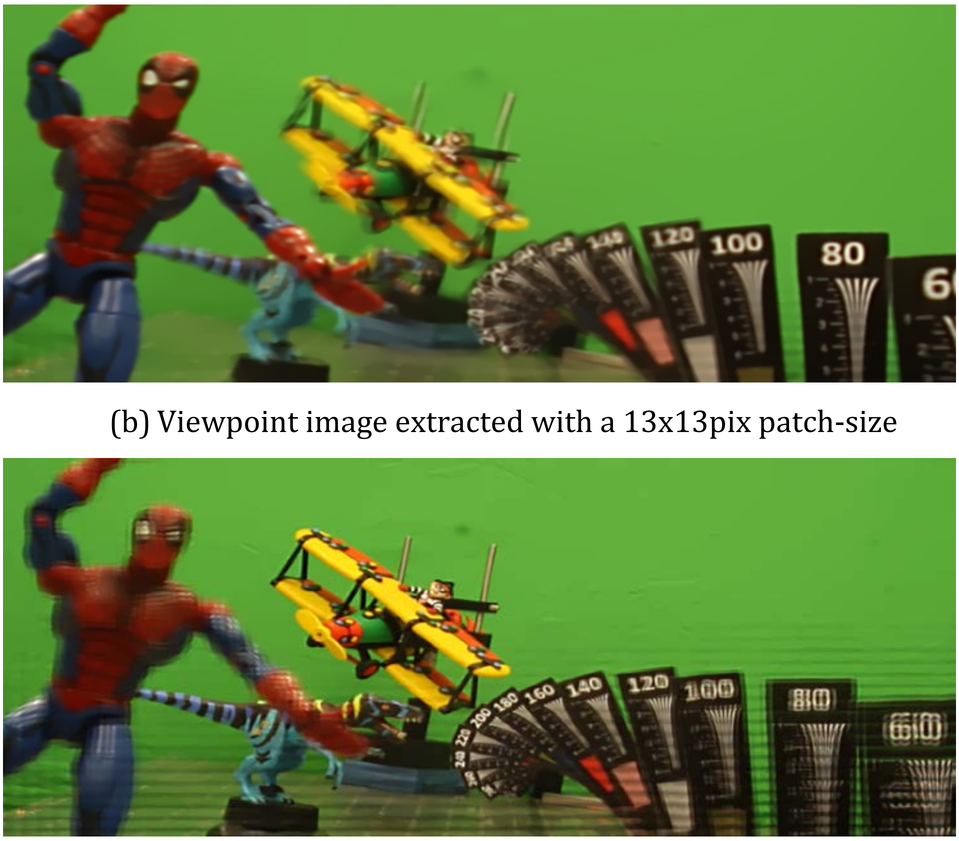
(i)Patch Size: 21x21p
VI size: 1261×652p

Figure 4.6 – Viewpoint image samples extracted with various patch sizes ranging from 5x5pixels to 21x21pixels(p).

Based on the viewpoint images presented in Figure 4.6 above, it is evident that the patch-size variable of the Holoscopic content adaptation process has an effect on the quality and size of the viewpoint image. However, the patch-size variable also has an effect on the focusing plane of the extracted viewpoint images. This is a unique ability of the Holoscopic 3D imaging system, and it is exploited to aid the reliable estimation of 3D depth maps. Figure 4.7 below presents a more evident effect of the patch-size variable on the extracted viewpoint images.



(a) Viewpoint image extracted with a 7x7pix patch-size



(b) Viewpoint image extracted with a 13x13pix patch-size

(c) Viewpoint image extracted with 21x21pix patch-size

Figure 4.7 – Viewpoint images highlighting the effects various patch-size effect on the H3D image focusing planes.

Figure 4.7 above presents three viewpoint images from the table of viewpoint images in Figure 4.6. Figure 4.7a presents a viewpoint image from a 7x7 patch-size that results in an out of focus low-resolution viewpoint image. However, as the patch-size increases, the focal range move to where the foreground objects are in focus and the background becomes out of focus and shown in Figure 5.4b above. Figure 5.4c, on the other hand, presents an inverse focus, where the background features are in full focus while the foreground is slightly out of focus. Figures 5.4b and Figure 5.4c presents the idea data set and this slight segmentation of features assist in reliable feature matching and 3D depth estimation. Test results proving this is presented later on the chapter in section 4.4.

The next section presents the detail explanation of the similarity measure and depth estimation stage of the Holoscopic 3D depth form disparity framework.

4.3. Similarity Measure and Depth Estimation

Matching cost in 3D depth estimation determines the similarity measure of pixels by either the use of global or local search. However, the matching technique used in the H3D depth from disparity framework presented in this chapter is a local technique, where the mean square error (MSE) between pixels values of the window patch between stereo viewpoint images are compared. This is done to find the best match between pixels of the stereo viewpoint images, followed by the calculation of 3D depth through disparity. Based on the Holoscopic 3D imaging system principles mentioned earlier in chapter two, it is obvious that the lack of disparity between the viewpoint images is a drawback that can result in the inaccurate 3D depth map. This small baseline problem can render the use of block matching techniques obsolete when estimating 3D depth information from stereo viewpoint images. The suggested solution for the following drawback is the use of a dynamic subpixel block matching technique, and the use of semi segmented viewpoint images, resulting in the estimation of reliable 3D depth maps. However, due to the computational cost that occurs during dynamic subpixel matching, the use of pyramiding and telescopic search to guide the block matching is introduced to reduce the severity of this problem [94][95]. A step by step explanation of the feature extraction and disparity estimation process is presented in pseudo code in Algorithm (1) below, following by an explanation of its key functions in sections 4.3.1 and 4.3.2.

Algorithm 1: Dynamic Depth Extraction Technique from Holoscopic 3D Image

```
Input:  H3D: Left view and right view
        Iter: number of iterations
        finf: false infinity
for i = 1 to Iter do
    Disparity cost = finf
    set min/max row bonds for image block
    for j = 1 to Iter do
        compute disparity bonds
        Compute as save all matching cost
        for d = min disparity to max disparity
            Disparity cost = sum of absolute difference
            end for
        end for
        Process scanline disparity cost with dynamic programming
        for k = 1 to Iter do
            cfinf = (disparity cost size – i +1) * finf
            Construct matrix for finding optimal move for each column
            Record optimal routes
        end for
        for l = 1 to Iter do
            Recover optimal route
        end for
    end for
Output: Depth map
```

4.3.1. Concatenation of Viewpoint Images

The extracted H3D viewpoint pair used is first converted into grayscale before the commence of the single-channel matching process. Although the use of coloured images sometimes results in better estimates, it is not efficient. The composite of the stereo viewpoint image is presented in Figure 4.8 below.

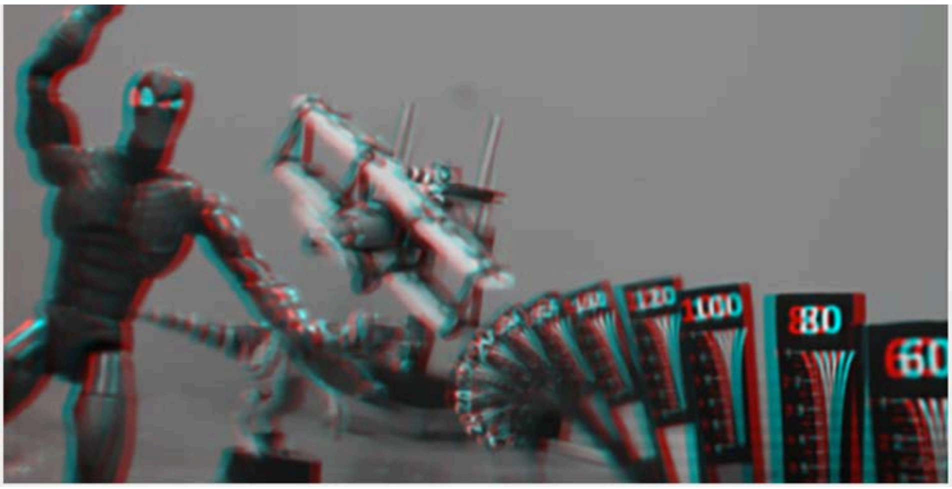


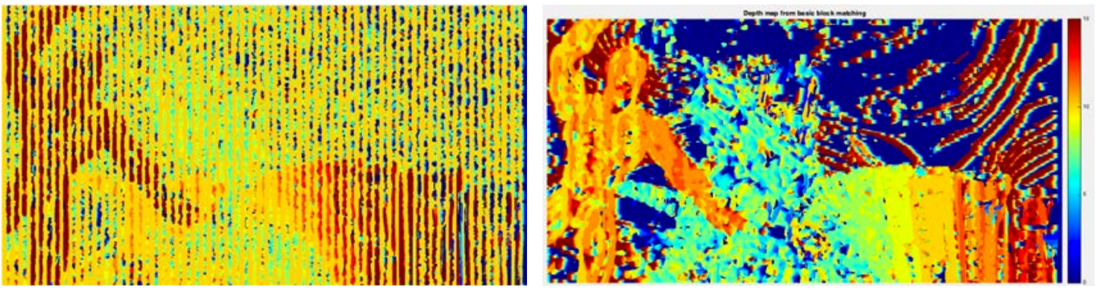
Figure 4.8 – Concatenated viewpoint image, presenting the left and right images in cyan and red.

This form of representation is popularly referred to as an anaglyph stereo representation, where one viewpoint image is tinted in blue or cyan, and the other is tinted in red. This help shows clearly if the disparity is present within the extracted viewpoint images, followed by the use of a concatenate function that links the stereo images together. This viewpoint concatenation is straight forward when the stereo image dataset is extracted from Holoscopic 3D image as there is no need for calibration. H3D imaging image rectification is not needed. This is as a result of Holoscopic 3D imaging system able to record multiple viewpoints of a scene in a single snapshot where stereo or Multiview imaging system fall short.

4.3.2. Block Matching with Dynamic Subpixel Accuracy

Once the stereo viewpoint images are extracted and concatenated, feature matching is the next step that follows. Area-based or feature-based matching techniques can be used to find the corresponding pixels, by searching along the horizontal scanline of the stereo viewpoint images. For this 3D depth estimation framework, the use of area-based matching with dynamic subpixel accuracy is applied. This is because of the baseline problem that occurs in Holoscopic 3D data set. The dynamic subpixel matching technique takes into account the cost of corresponding neighbouring pixels and refines an initial 3D depth map that is estimation without such consideration.

In cases where the baseline between viewpoint images are small, 3D depth map estimated from such viewpoint images have a high amount of noise, rendering the 3D depth map unreliable as presented in Figure 4.9a. The noise initially introduced during viewpoint extrapolation stage also contributes to this problem; that is why the 3D depth map in Figure 4.9b still have some noise.



(a) Without subpixel accuracy (b) With subpixel accuracy

Figure 4.9 – Direct comparison of feature matching result without subpixel accuracy and with subpixel accuracy.

Figure 4.9 above presents a direct comparison of the 3D depth map estimated with and without subpixel accuracy, indicating a definite improvement. The 3D depth map presented in Figure 5.9b above is extracted with the suggested feature matching technique. However, the use of this technique for estimating 3D depth maps is computationally demanding and depending on the window size used for matching this could be even more expensive. Apart from the following problems mentioned, traces of noise are still noticeable in the estimated 3D depth map, although not enough to render the 3D depth map unreliable. This leads to the use of dynamic matching technique implementation as a pyramid based structure [96] [97] to help tackle the problem at hand. Figure 4.10 below presents the final depth map after pyramid optimisation.

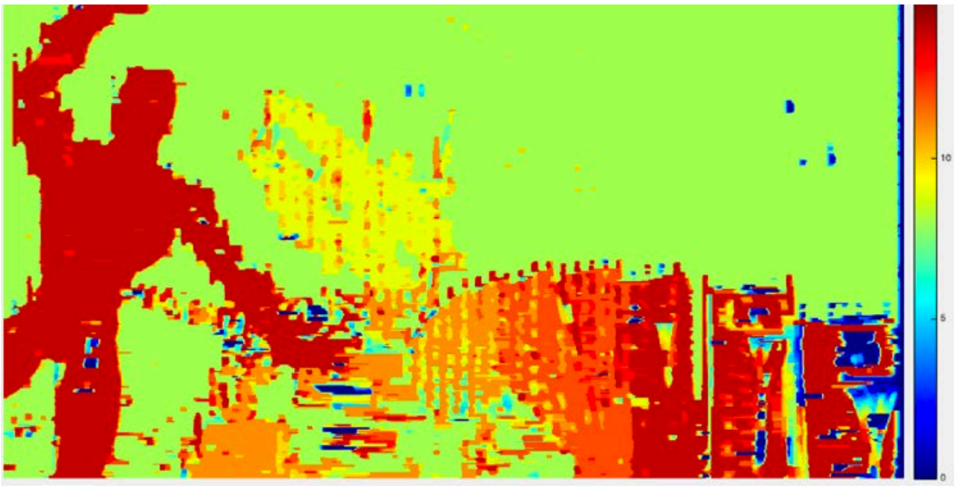


Figure 4.10 – Estimated 3D depth map from a single Holoscopic 3D image with dynamic pyramid optimisation.

The pyramid structure is a type of multi-scale signal representation, and in terms of 3D depth estimation, corresponding viewpoint images are subjected to repeated smoothing and subsampling. During subsampling, Gaussian filters are used to reduce the subsampling error caused by the pixel by pixel approach.

Section 4.4 presents the detail evaluation of the above framework, presenting the computationally cost acquired by this approach. This limitation led to the design of a smart 3D depth map mapping between H3D depth maps, presented in section 4.5, to possibly help reduce the computation burden of the H3D depth technique.

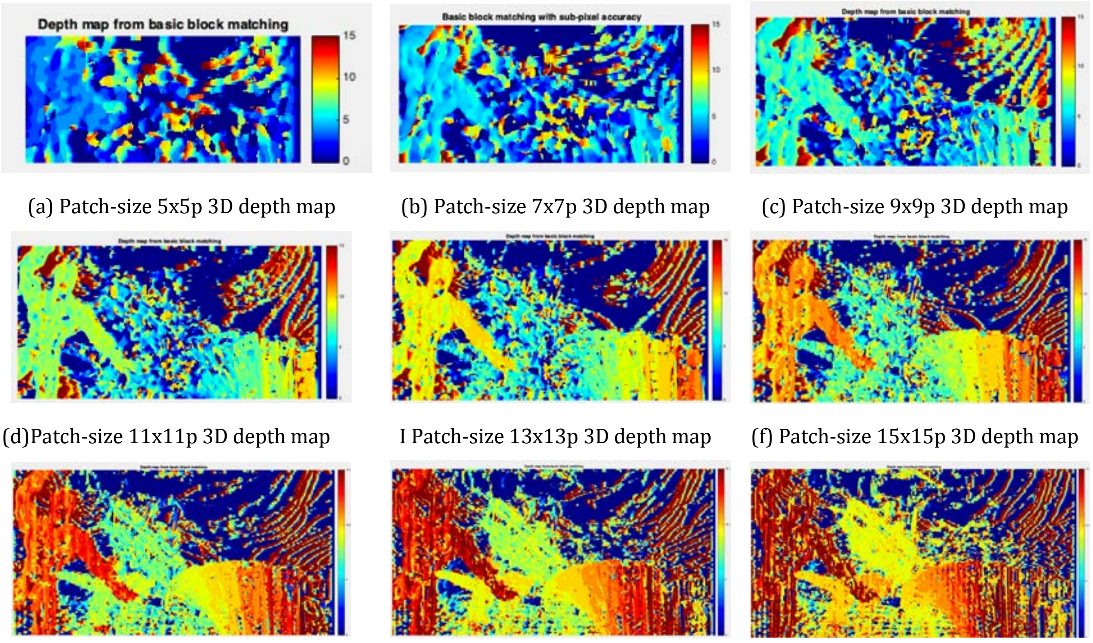
4.4. Results and Evaluation

This section presents all 3D depth results estimated by the H3DDD technique presented in this chapter. The image resolution range, matching window size range and disparity test range is conducted to test the robustness the H3DDD technique. Concluding with the direct comparison of this technique against current state-of-the-art techniques.

The H3DDD depth is implemented on an Intel(R) Core (TM) i7-4790 CPU @ 3.60 GHz CPU.

4.4.1. Image Resolution Test

The 3D depth maps in this section are estimated from viewpoint images with different patch sizes. This test emphasises how the lack of spatial information within the extracted viewpoint image can affect the resulting 3D depth map. Figure 4.6 in section 4.2 presents the VIs extracted from different patch sizes, detailing their VI resolution in pixels (p).



(a) Patch-size 5x5p 3D depth map (b) Patch-size 7x7p 3D depth map (c) Patch-size 9x9p 3D depth map

(d)Patch-size 11x11p 3D depth map I Patch-size 13x13p 3D depth map (f) Patch-size 15x15p 3D depth map

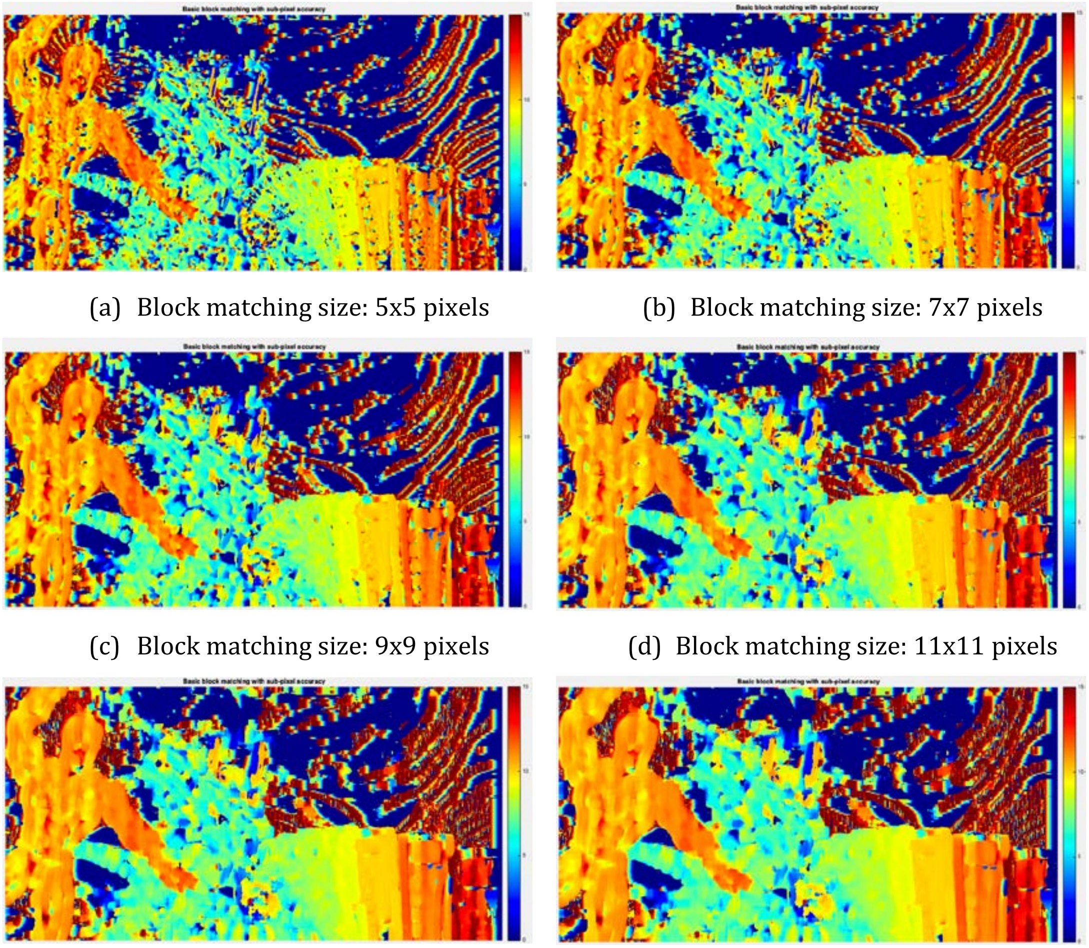
(g) Patch-size 17x17p 3D depth map (h)Patch-size 19x19p 3D depth map (i) Patch-size 21x21p 3D depth map

Figure 4.11– 3D depth map estimation of stereo viewpoint images extracted with patch sizes ranging from 5x5pixels to 21x21pixels.

Results from Figure 4.11 above clearly indicates that viewpoint image with higher resolution results to better estimation of 3D depth maps. However, this comes at computation cost during 3D depth estimation. Therefore, the use of viewpoint images with just enough spatial information is advice. Also, as mentioned earlier, the ability to refocus at different planes of the viewpoint image begin to take effect as the patch size increases. The following also contributes to the improvement of the estimated 3D depth maps presented above.

4.4.2. Match Window Size Test

The 3D depth maps in this section are estimated from stereo viewpoint images with different block matching sizes, ranging from 5x5 pixels to 21x21 pixels, shown in Figure 4.12 below.



(a) Block matching size: 5x5 pixels

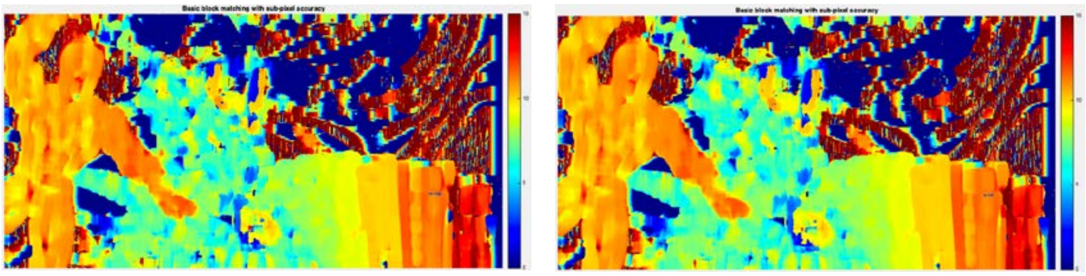
(b) Block matching size: 7x7 pixels

(c) Block matching size: 9x9 pixels

(d) Block matching size: 11x11 pixels

(e) Block matching size: 13x13 pixels

(f) Block matching size: 15x15 pixels



(g) Block matching size: 17x17 pixels

(h) Block matching size: 21x21 pixels

Figure 4.12 – 3D depth map estimation with block matching windows ranging from 5x5p to 21x21p.

As presented in Figure 4.12 above, 3D depth maps estimated with small matching block sizes of 5x5 to 9x9 pixels have patchy looking 3D depth maps, especially within prominent features. However, as the block matching sizes increase, this patchy effect is gradually reduced at a minimal increase in computational cost. However, 3D depth maps estimated from small patch-sizes can also be optimized to match depth maps estimated with a larger matching window. Figure 4.13 presents the time taken measured in seconds (secs), to estimate each 3D depth map from the window sizes ranging from 5x5pixels to 21x21pixels.

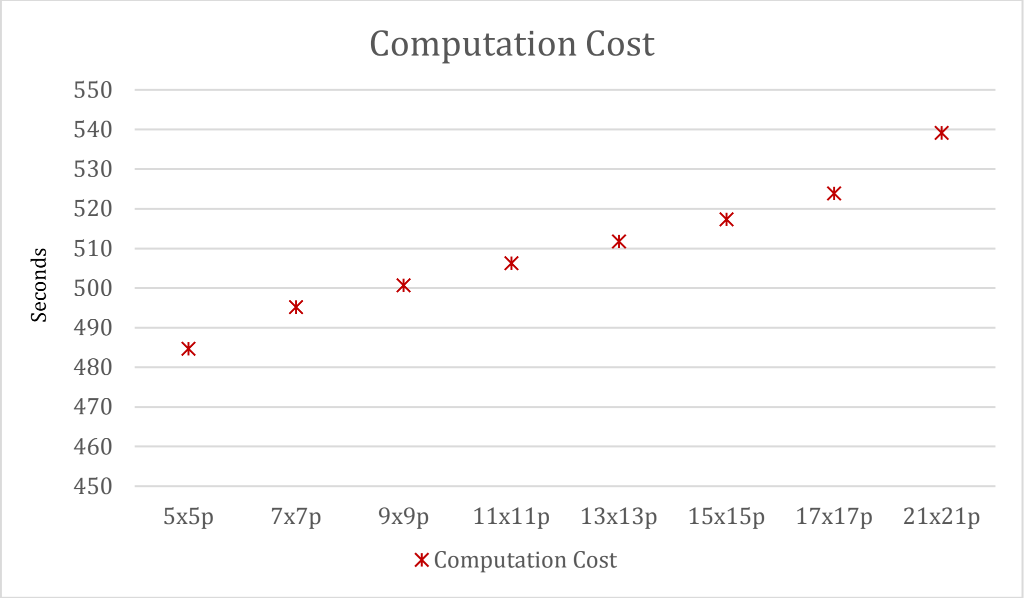
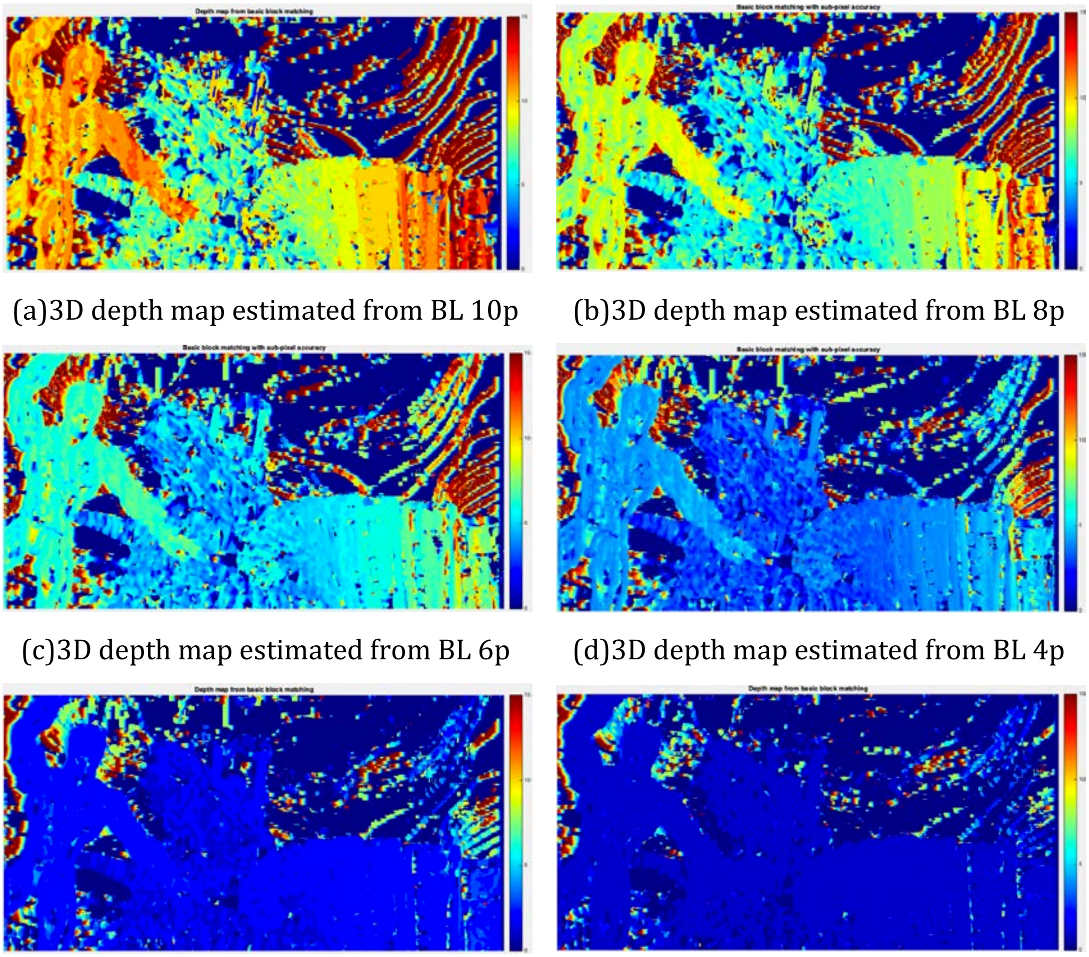


Figure 4.13 – The computation cost as the matching window size increases, measured in seconds.

4.4.3. Disparity Range Test

The 3D depth maps in this section are estimated from stereo viewpoint images of varying disparity sizes. This helps in deducing the best trade-off between angular and spatial information. Figure 4.14 below presents the 3D depth maps estimated with different disparity sizes ranging from 1 to 10 pixels between the stereo viewpoint images.



(a)3D depth map estimated from BL 10p (b)3D depth map estimated from BL 8p

(c)3D depth map estimated from BL 6p (d)3D depth map estimated from BL 4p

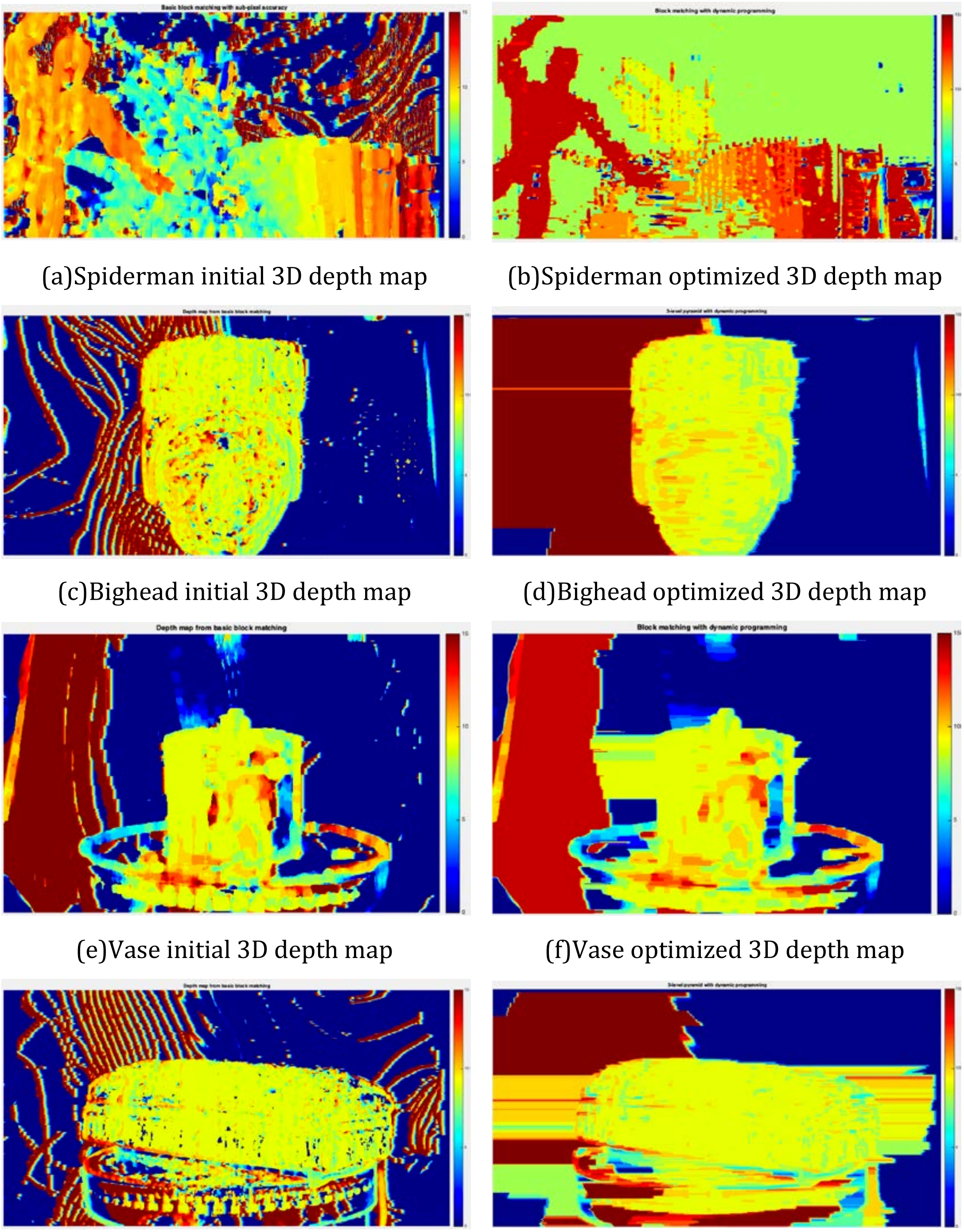
(e)3D depth map estimated from BL 2p (f)3D depth map estimated from BL 1p

Figure 4.14 – 3D depth map estimated from stereo viewpoint images with a baseline (BL) ranging from 1pixels to 10pixels(p).

The following 3D depth maps present in Figure 4.14 above begins to present a precise segmentation of features when the baseline between viewpoint images in about four pixels and above. This is only possible due to the semi segmentation of the extracted viewpoint images. This information can lead to the use of smaller microlens arrays to increase the default resolution of viewpoint images without interpolation. Emphasising on the fact that MLAs specifications should be considered when recording H3D data for H3D depth from disparity.

4.4.4. Multi Data 3D Depth Maps and Point Cloud Reconstruction

The 3D depth maps in this section are estimated from stereo viewpoint images of all the H3D data set presented in Figure 4.4 above. Figure 4.15 shows the before and after optimisation effect.



(a)Spiderman initial 3D depth map

(b)Spiderman optimized 3D depth map

(c)Bighead initial 3D depth map

(d)Bighead optimized 3D depth map

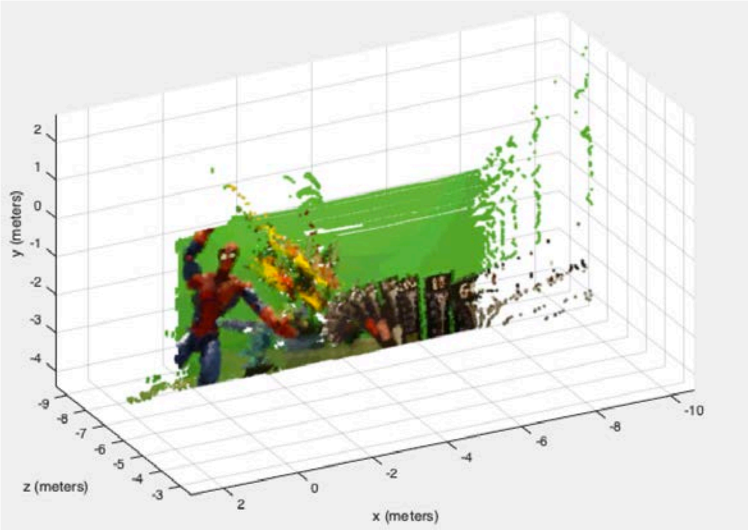
(e)Vase initial 3D depth map

(f)Vase optimized 3D depth map

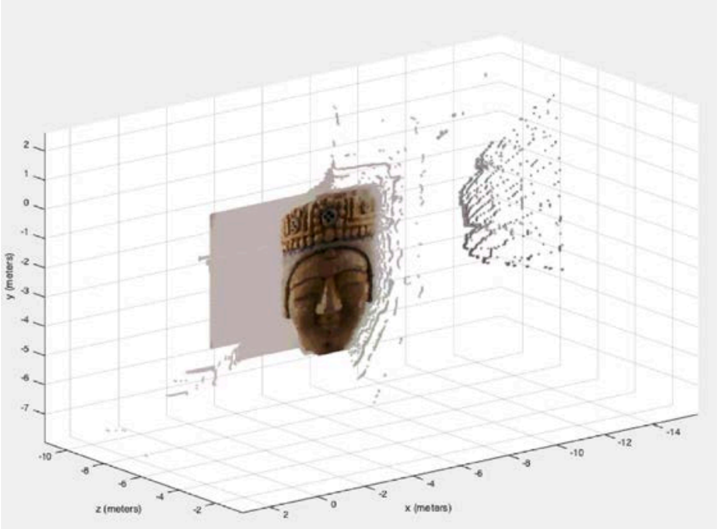
(g)Flax initial 3D depth map (h)Flax optimized 3D depth map

Figure 4.15– 3D depth maps of H3D data set, with a side by side presentation of 3D depth maps before and after depth map optimization.

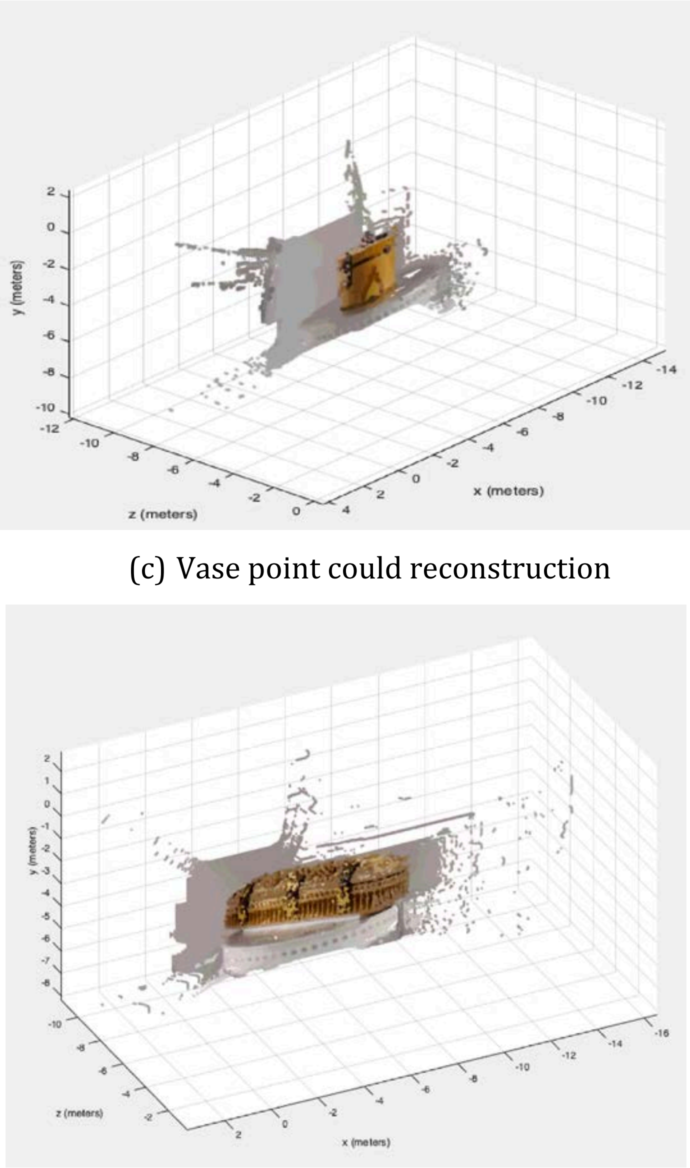
The 3D depth maps estimated using the proposed framework shows how robust the H3D depth from disparity technique is in handling varying Holoscopic 3D data of varying specifications and feature properties. The point cloud representation of the following 3D depth maps in Figure 4.15 above is presented in Figure 4.16 below.



(a) Spiderman point could reconstruction



(b) Bighead point could reconstruction



(c) Vase point could reconstruction

(d) Flax point could reconstruction

Figure 4.16 – Point cloud representation of the above 3D depth maps in Figure 4.15.

4.4.5. Direct Comparison

This section presents the direct comparison of the proposed H3D depth estimation and state-of-the-art depth estimation technique developed by Szeliski et al., which can be downloaded at the Middlebury website[3] [9]. This section also presents the complexity assessment based on the stereo correspondence constraints presented in chapter two. The complexity evaluation matrix uses a binary ranking system to present the amount of constraint used in the stereo depth technique, averaging the score against the total number of constraints. The

[3] vision.middlebury.edu/stereo/

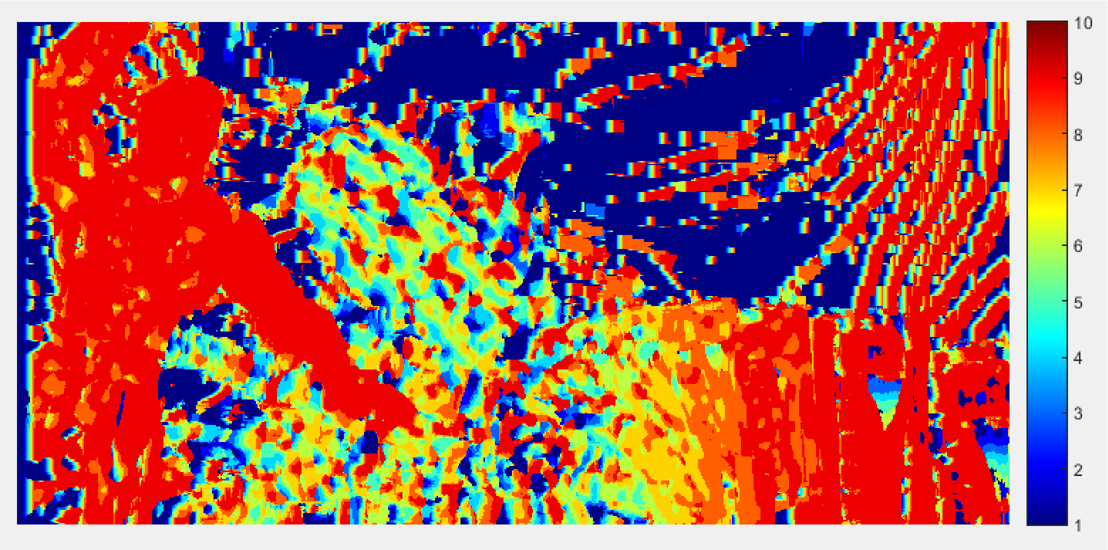
computational cost is then taken in time (seconds) concluding the two-step evaluation process. Table 4.3 below presents the stereo constraints and their descriptions.

Table 4.3

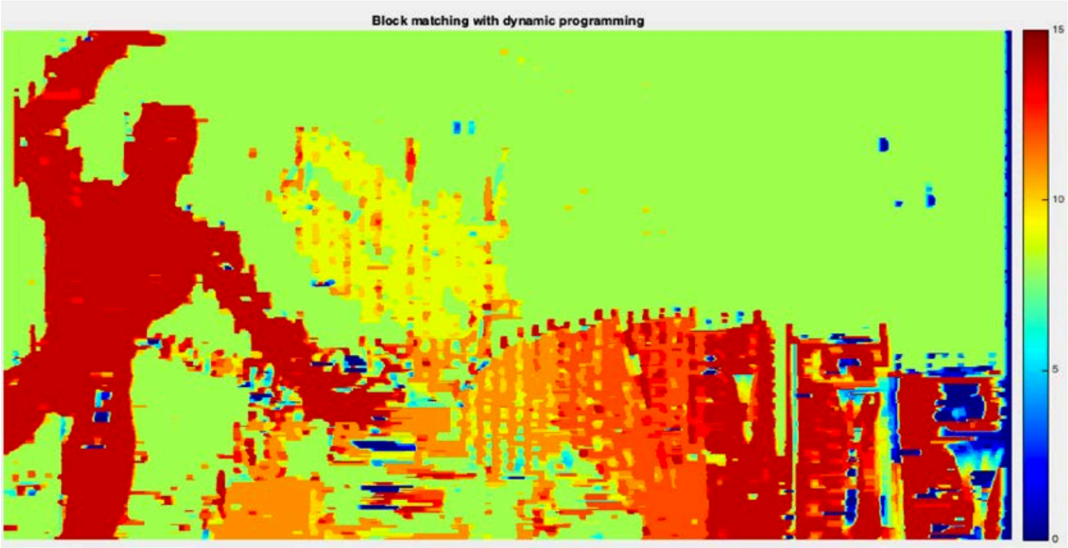
The stereo depth constrains, presenting their identification code and brief description.

ID	Stereo Constrains	Description
C1	Similarity Constraint	Enforces pixel matching between stereo viewpoint images.
C2	Epipolar Constraint	Enforces stereo image rectification.
C3	Uniqueness Constraint	Enforces that each pixel has a one unique match or occlusion.
C4	Positional Constraint	Enforces matching of pixels at irregular positions, the cross-eye depth problem.
C5	Disparity Constraint	Enforces smoothness between pixels of similar disparity.
C6	Structural Constraint	Enforces structural shape of depth map to be the same as the reference viewpoint image.

Figure 4.17 below presents the 3D depth map estimated from the proposed H3D depth from disparity framework presented in this chapter, compared directly to depth map estimated from state-of-the-art stereo 3D depth estimation technique.



(a) Szeliski 3D depth result with our data set



(b) Our 3D depth result with our data set

Figure 4.17– Direct comparisons of the H3DDD frameworks and current state-of-the-art framework using the Spiderman H3D image.

The following comparison indicates that the Holoscopic content adaptation of the proposed 3D depth from disparity framework can be used on other 3D depth from disparity frameworks. However, since the other frameworks where designed to estimate 3D depth information from Stereo/Multiview imaging stereo data that have different properties from H3D images. Their frameworks fail to match the accuracy of the proposed framework in this chapter. The complexity of the following 3D depth from disparity frameworks are presented in Table 4.4 below, where the number closest to one (1) entails that the 3D depth estimation framework is more complex.

Table 4.4

Complexity evaluation result of the H3DDD and Szeliski et.al. depth frameworks.

3D Depth frameworks	*Stereo constraints*						
	C1	**C2**	**C3**	**C4**	**C5**	**C6**	**Avg.**
3D Depth Szeliski et al.	1	1	1	0	1	0	0.6
H3D Depth from Disparity	1	1	0	0	1	0	**0.5**

The computational cost of the following 3D depth estimation techniques presented in Table 4.4 above is shown in Figure 4.18 below.

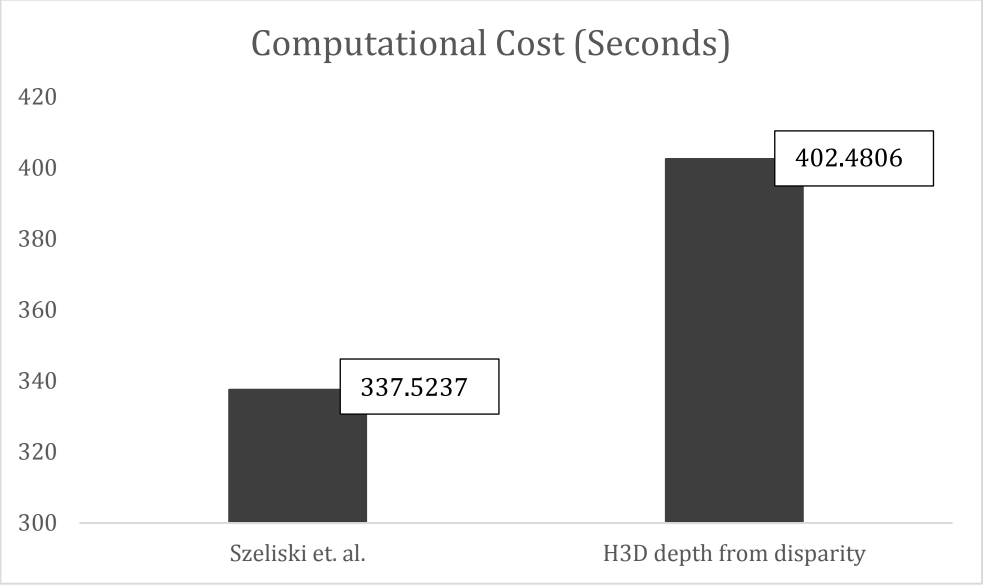


Figure 4.18 – Computational cost of Szeliski et.al. and H3DDD depth techniques measured in seconds.

Although the presented framework in this chapter is very robust, this comes at a computational cost, as shown in Figure 4.18 above, this results in the use of this technique in only near-real-time applications. To further improve the computational efficiency of the framework, the proposal of a smart 3D depth mapping framework is presented next.

4.5. Smart 3D Depth Mapping Design

The smart 3D depth mapping design presented in this section is proposed to reduce the computational workload taken to estimated 3D depth maps from high-resolution images. The design takes advantage of the HCA ability to extract multiple viewpoint images of different image sizes. It is well established that the higher the image resolution, the greater the chance of extracting reliable 3D depth maps. The smart 3D depth mapping takes advantage of this by estimating various 3D depth maps of low-resolution images and high-resolution images of the same viewpoint image. The low-resolution 3D depth map is then up-sampled and mapped against the high-resolution 3D depth map, training a deep-learning network to understand the residual difference between the two 3D depth maps. This design workflow is presented below in Figure 4.19.

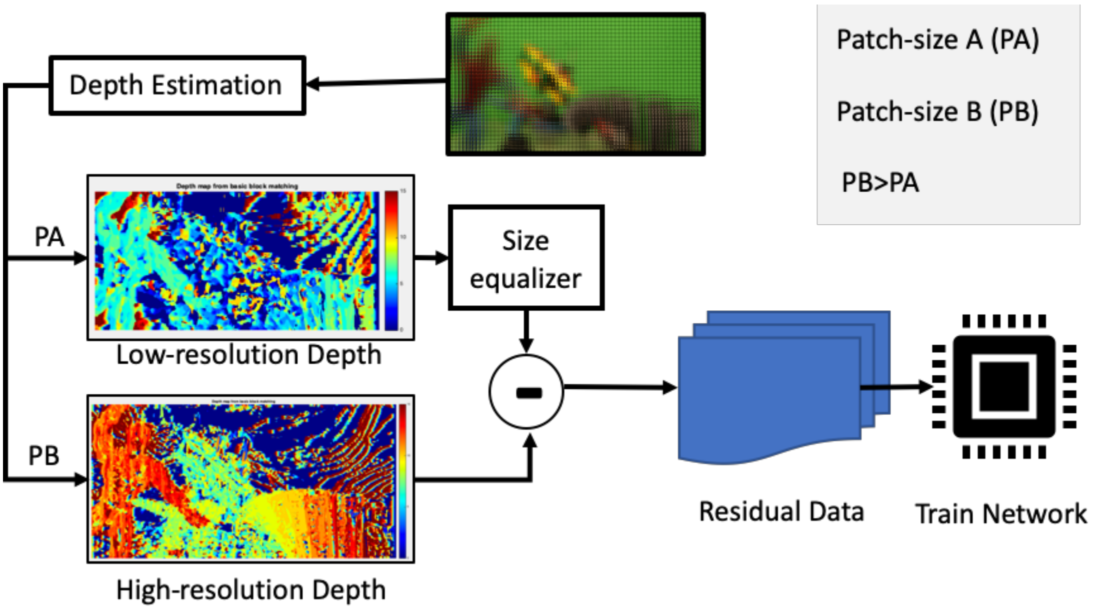


Figure 4.19 – Architecture of the Smart 3D depth mapping technique for H3D images.

The design above will result in the estimation of 3D depth maps with low-resolution viewpoint images after the intimal mapping of 3D depth information between low- and high-resolution 3D depth maps. The subsequent 3D depth estimations will take lesser computational effort as the initial 3D depth map is estimated from low-resolution then optimized into high-quality 3D depth maps. The smart 3D depth mapping will also consider the suitable parameters when handling a large data set, making the presented framework even more robust. The full implementation of this will be done in future works.

4.6. Summary

This chapter presents a standard and robust 3D depth estimation framework for Holoscopic 3D data. The proposed technique takes advantage of various H3D image properties when extracting viewpoint images for the 3D depth from disparity framework. This includes the ability to record multiple viewpoint images and the ability to refocus at any focal plane after capture. As the low-resolution problem that is commonly associated with the H3D data is resolved with smart viewpoint extraction technique that is capable of handling complex H3D data. The Holoscopic content adaptation process can also be used to extract viewpoint images for other stereo 3D depth estimation frameworks. However, current 3D depth from disparity frameworks is designed against the image properties of standard 2D data. The 3D depth calculation and optimization section of the proposed framework in this chapter employs 3D depth techniques that are most suitable for H3D data. This includes the use of subpixel and pyramid dynamic programming to match the corresponding pixel and optimize initial estimates of 3D depth maps. The documented depth map results range from image resolution test to disparity range test. The image resolution test clearly indicates that the higher the resolution of extracted viewpoint images, the better the 3D depth maps.

Matching window experiment is conducted as well, however, does not have a huge enough impact on the 3D depth map estimated, therefore opting for the window size with the least computational burden been considered. The reason as to why the size of the matching window does not have an impact is due to the fact that initial 3D depth maps estimates from small window sizes can still be optimized to produce a reliable 3D depth map.

Lastly, the disparity range test indicates that 3D depth information can be extracted from stereo viewpoint images with a baseline as small as 10pixels (0.41mm). This proves and elaborates on the suggestions that Holoscopic 3D data captured with smaller MLAs might be most suitable for 3D depth from disparity frameworks.

The chapter also presents a smart 3D depth mapping design that aims to reduce the overall computational time taken to estimation 3D depth information the intimal mapping of low and high-resolution 3D depth maps.

CHAPTER 5: INNOVATIVE DIRECT 3D DEPTH ESTIMATION FROM H3D IMAGE

This chapter presents the innovative direct 3D depth from Holoscopic 3D image framework. Due to the inability to accurately predict or quantify the amount of unwanted noise introduced in viewpoint images and the loss of angular information, this approach extracts 3D depth information directly from a single H3D image by estimating disparity at the EI level. The chapter layout is as follows: 5.1 Introduction, 5.2 Holoscopic 3D Data, 5.3 Feature Extraction with Census Transform, 5.4 Similarity Measure and 3D Depth Estimation, 5.5 3D Depth Optimization with Segmentation, 5.6 Results and Evaluation and 5.7 Summary.

5.1. Introduction

This chapter presents an innovative 3D depth estimation technique specifically designed to estimate 3D depth information directly from a Holoscopic 3D image. The technique referred to as Direct 3D depth from Holoscopic (DDH) can estimate 3D depth information from a single H3D image Omni directionally. The motivation behind the development of this depth technique is to by-pass the content adaptation stage that often introduces unwanted artefacts into extracted viewpoint images and results in loss of angular information. In doing so, the quality of extracted 3D depth information is preserved. The unique feature in this 3D depth estimation technique is the implementation of a similarity measure that estimates 3D depth information by calculating the disparity between Elemental images. Depending on the Elemental image position of the start viewpoint image pixel (VIP) and position in which the corresponding VIP is located, the disparity is calculated by their squared differences. The workflow of the following 3D depth estimation technique is presented in Figure 5.1 below.

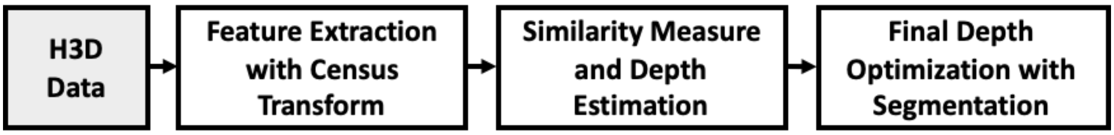

H3D Data
Feature Extraction with Census Transform
Similarity Measure and Depth Estimation
Final Depth Optimization with Segmentation

Figure 5.1 – The innovative Direct 3D Depth from Holoscopic framework.

The Direct 3D depth from Holoscopic technique workflow presented in Figure 5.1 above consists of four steps: (i) H3D pre-processing, (ii) Feature extraction with census transform, (iii) Similarity measure and 3D depth estimation and, (iv) Final 3D depth optimisation with segmentation. The following is discussed in detail in their respective sections, before concluding with the detailed testing and evaluation of the DDH technique. Due to the DDH technique novelty, an indirect comparison to existing stereo techniques is conducted, where similar viewpoint images estimated by the DDH technique is extracted, up-sampled and estimated by current state-of-the-art stereo matching techniques. A complexity and cost evaluation are also presented, where the DDH technique outperforms current state-of-the-art techniques. The pseudo code of the DDH technique presented in Algorithm (2) below is as follows:

Algorithm 2: Direct Depth from Holoscopic 3D Image Algorithm

```
Input:  H3D: single Holoscopic 3D image
        Iter: number of iterations
        **for** i = 1 to Iter **do**
            best_match = initialised disparity diff
            **for** j = 1 to Iter **do**
                ref_p = reference pixel position
                ref_patch = block of pixels including reference pixel
                Apply gaussian and census transform to ref_patch
                **for** k = 1 to Iter **do**
                    tar_p = target pixel position
                    tar_patch = block of pixels including target pixel
                    Apply gaussian and census transform to tar_patch
                    find best match for ref_patch
                  **If** curr_match < best_match
                    best_match = curr_match
                    Z = depth value
                  **end if**
                **end for**
            Depth map (i,j) = Z
            **end for**
        **end for**
Output: Depth map
```

The following is part of the third block and final contribution of this research, shown in Figure 5.2 below, highlighted in red and emphasised in green.

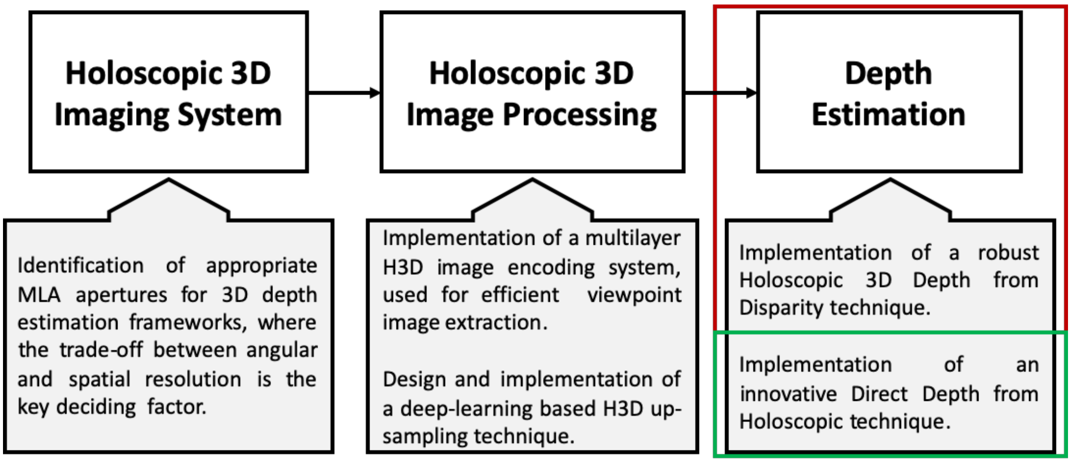

Holoscopic 3D Imaging System
Holoscopic 3D Image Processing
Depth Estimation
Identification of appropriate MLA apertures for 3D depth estimation frameworks, where the trade-off between angular and spatial resolution is the key deciding factor.
Implementation of a multilayer H3D image encoding system, used for efficient viewpoint image extraction.
Design and implementation of a deep-learning based H3D up-sampling technique.
Implementation of a robust Holoscopic 3D Depth from Disparity technique.
Implementation of an innovative Direct Depth from Holoscopic technique.

Figure 5.2 – The research contribution of this chapter highlighted in red and emphasised in green.

5.2. Holoscopic 3D Data

The synthetic Holoscopic 3D image used in this chapter mimics the real registration process of Brunel University Holoscopic 3D imaging system principles. However, the distortions that occur when recording such imaging in the real world do not occur in the virtual environment, leading to the generation of error-free synthetic Holoscopic 3D (SH3D) images. The SH3D images used in this chapter as the prime H3D dataset is computationally generated in The Persistence of Vision Ray Tracer software referred to as POV-RAY, written in C++ and developed by the POV-Team [86]. However, due to the lack of maturity of the software, the rendering of highly textured objects is challenging. On the other hand, its ability to render error free-viewpoint images of different EI sizes is the main reason why the software is employed in this chapter.

The synthetic Holoscopic 3D data set used in this chapter is rendered in accordance with the principles of an H3D imaging system with a square Omni-directional MLA. The virtual H3D system has a sensor size of 35.9x24mm and dot pixel pitch size of 0.00451mm. The specifications of the synthetic H3D data set is presented in Table 5.1, while the data sample is presented in Figure 5.3 below.

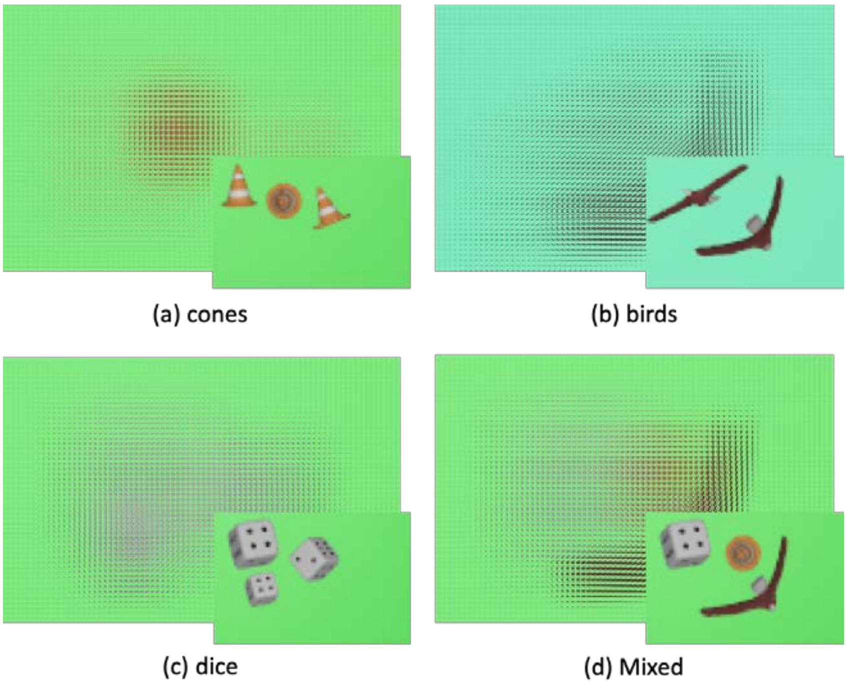


Figure 5.3 – Samples of SH3D dataset used in this chapter and with their respective viewpoint images, all rendered from POV-RAY.

Table 5.1

Full specifications of the Synthetic Holoscopic 3D image data rendered from POV-RAY.

EI Size (pixel)	VI SIZE (pixel)	MLA Size
100x100	79x53	79x53
80x80	99x66	99x66
60x60	132x88	132x88
40x40	198x132	198x132
20x20	397x265	397x265

Table 5.1 above presents a clear correlation between MLA and viewpoint image resolutions. Table 5.1 also presents the Synthetic Holoscopic 3D image data set properties used in the following subsection, testing the robustness of DDH technique to be presented.

The following section discusses the feature extraction process of the DDH 3D depth estimation technique.

5.3. Feature Extraction with Census Transform

The feature extraction process employed by the DDH technique can be seen as an area-based matching technique. This technique, like all other area-based matching technique, employs the use of block matching windows to extract feature within an image with a pixel by pixel movement. The workflow of the feature extraction process presented in Figure 5.4 below.

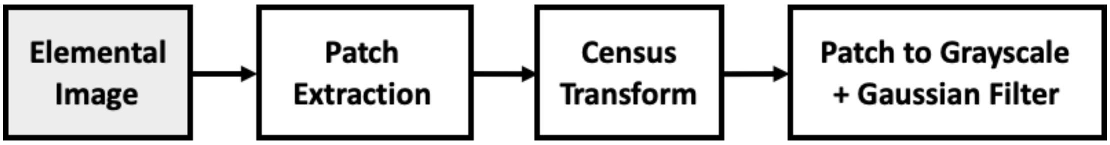


Figure 5.4 – Feature extraction process with Census Transform.

Figure 5.4 above presents the detailed workflow of the feature extraction technique used in the DDH 3D depth estimation process. The technique extracts a widow patch around the viewpoint image pixel (VIP) in which 3D depth is to be currently estimated. The extracted window-patch then converted to grayscale to before the application of the CT. The Census transform applied on the window reduces the intensity component of the extracted window-parch into binary intensity values depending on the value of the reference pixel. This is done to prevent the extracted features form been affected by global radiometric differences, such as global illumination differences. The Gaussian filter is then applied to help create a contrast between pixels and at the same time, group pixels of the same properties. As the DDH is a 3D depth technique that estimates 3D information on a pixel by pixel bases, the next section discusses the similarity measure and 3D depth estimation process of the following technique.

5.4. Similarity Measure and 3D Depth Estimation

Current stereo algorithms estimate 3D depth through disparity or triangulation, given a set of stereo images, a feature is selected from a reference image and a scan is executed on the target image to find its corresponding match. The squared difference of the match is stored as the disparity value of the reference pixel. The DDH technique, on the other hand, estimates the disparity difference of a reference pixel in relation to the position of the EI origin position and corresponding EI match location. As discussed earlier in section 5.3 above, once the feature of the start VIP is extracted, the summation of the window is executed before a unidirectional scan through aligning a set of Elemental images is conducted to find the best match. This innovative similarity measure can be defined as,

$$EI_{(u,v)}(x,y,d) = \sum_{EI_{(u,v)}(x,y,)\in W}^{EI_{(U,V)}} \left(\left(EI_{(u,v)}(x,y) - EI_{(u-d,v)}(x,y) \right)^{2} \right) \qquad (16)$$

Equation (16) above presents the DDH similarity measure and disparity estimation, where $EI_{(u,v)}$, is the starting position in which the reference pixel candidate is located, defined as $EI_{(u,v)}(x,y)$, and the disparity d, is calculated from squared difference between the $EI_{(u,v)}$ and $EI_{(u-d,v)}$, where the best match of the reference VIP is located.

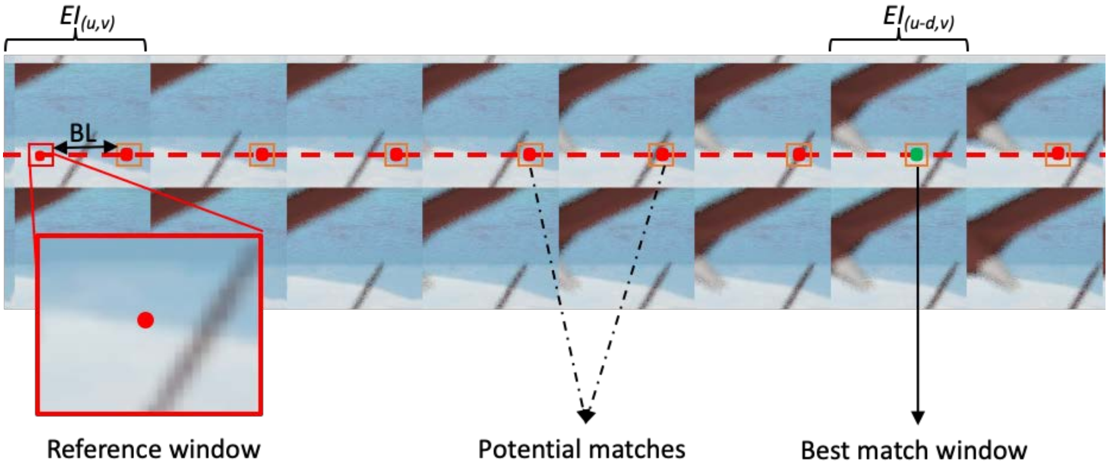


Figure 5.5 - Principle of the similarity measure for the Direct 3D Depth from Holoscopic 3D image, whereas BL is the distance between viewpoint pixels.

However, to improve the quality of the estimated results, a threshold is applied, reducing the number valid of Elemental images a feature point can be searched in depending on the reference VIP's Elemental image location. Figure 5.5 above further gives a graphical explanation of the principle behind the innovative DDH similarity measure and initial 3D depth estimation process. The following technique presented in this chapter currently searches for matches along the x-axis of the Holoscopic 3D image. However, the following matching technique has the ability to estimated 3D depth inform from both axes, making use of the vertical as well as the horizontal 3D depth information recorded by the Omni-directional MLA of the H3D imaging system. In order to successfully estimate the disparity of a viewpoint image, all VIP that combines to make up a specific reference viewpoint is matched against all VIP that combines to make the targeted viewpoint. This is shown in Figure 5.6 below.

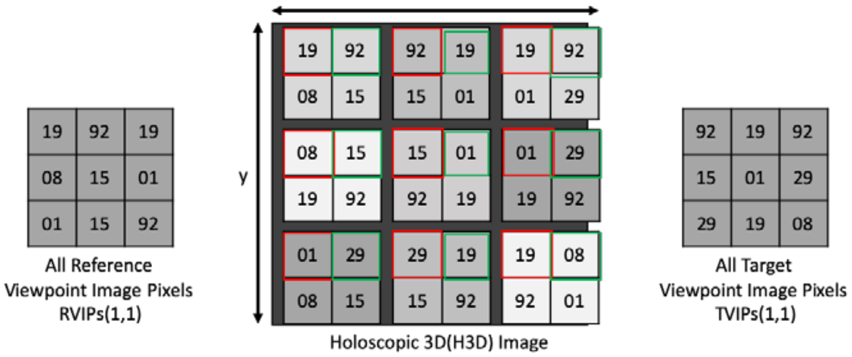


Figure 5.6 – H3D image indicating all reference and target viewpoint image pixels.

Viewpoint image pixels (VIP) are pixels that combine to make up a specific viewpoint image. They are distributed across the Holoscopic 3D image in uniform distance to each other on both x and y-axis, provided there are no lens distortion errors. The following results to all viewpoint pixels having the same neighbouring pixels, therefore the use of an area-based matching technique to estimate their individual disparities. The area-based matching technique focuses on the RVIPs and its neighbours as well as the TVIPs and its neighbours, ignoring any pixel that does not belong to those set of pixels. The window size has to be within the size of the Elemental image and depending how big it is, the computational cost could increase accordingly.

The second advantage the DDH technique have over 3D depth estimation techniques is its ability easily handle edges without the need for image padding done by stereo 3D depth techniques. Image padding is usually done by stereo based techniques because the stereo image data needed for 3D depth estimation do not have neighbouring pixels surrounding the pixels that make up the edge or boundary of the images. This often results in more computational effort to estimate the 3D depth information of image edges, mainly when the data is acquired with an out of bound error. As for the DDH technique, due to the H3D imaging systems unique ability to record multiple viewpoints of any giving scene, by merely choosing a reference viewpoint image with adequate neighbouring pixels (see Figure 6.5) surrounding it, rectifies this problem showing another massive potential of the H3D image and DDH technique.

Although this chapter only presents how the Census Transform (CT) and the sum of squared difference (SSD) are adapted in finding the best match and initial disparity map, other area-based techniques can also be incorporated in the DDH technique.

The next section presents the 3D depth optimization technique used to refine the 3D depth map extract from this stage.

5.5. 3D Depth Optimization with Segmentation

The 3D depth optimization stage of the DDH technique uses segmentation to refine the initial 3D depth map. In order to achieve this, a smoothness term is applied to the reference viewpoint image, labelling each pixel into feature segments, resulting in an energy minimization problem. As presented in Figure 5.7 below, the segmentation principle is elucidated with a directed weighted graph A = (K, U), consisting of nodes U and a set of directed edges K that connects them. In 3D depth estimation, the nodes can be referred to as pixels or reliable image features and depending on the number of objects recorded by an imaging system, extra nodes or terminals are created for labelling, in which nodes are grouped. As for Graph A(K,U) that consist of two main features, they are ladled as the "source" $\boldsymbol{s} \in \boldsymbol{U}$ and the "sink" $\boldsymbol{t} \in \boldsymbol{U}$ nodes.

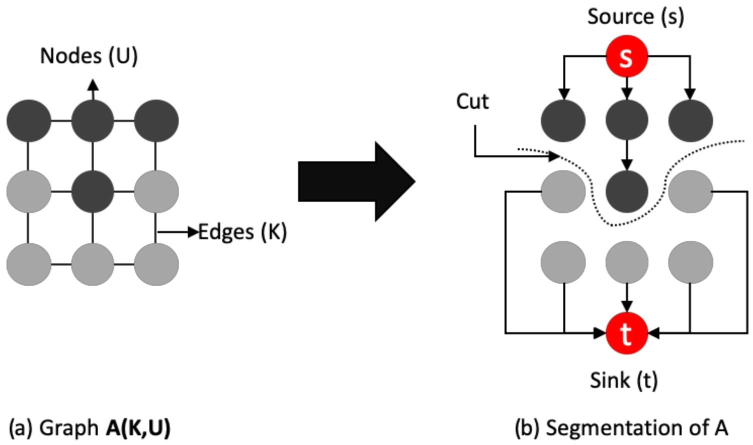
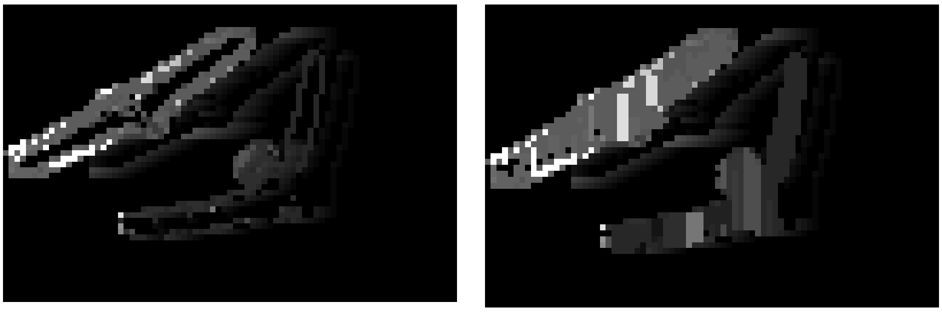


Figure 5.7 – Segmentation of Graph A (K, U).

However, the segmentation technique elucidated in Figure 5.7 above, is used in correcting only disparity discontinuities that occur inside 3D depth segments, as shown in Figure 5.8 below. The binary technique takes the maximum value around the discrete pixel within the enclosed feature to optimize the disparity depth map.



(a)Initial 3D depth estimate (b) Optimized 3D depth result

Figure 5.8 – 3D depth map before and after the result is optimized.

Other 3D depth optimization techniques like the graph cut formulations presented in chapter three can also be used in place of the segmentation technique presented above to improve initially estimated 3D depth maps.

All test results gotten from this innovative DDH technique is presented in the next section.

5.6. Results and Evaluation

This section is subdivided into three subsections where all 3D depth map estimated by the DDH technique are presented. The first section presents a disparity test where viewpoint images with baselines ranging from 7 to 22pixels are documented. Secondly, the EI size test is conducted where the correlation between the window patch and the EI size is examined. The final experiment is an indirect comparison, complexity test and computational efficiency test between this research innovative DDH technique and state-of-the-art stereo techniques from the Middlebury website[4] and Hae-Gon et al. [9] [98].

The DDH is implemented on an Intel© Core © i7-4790 CPU @ 3.60 GHz CPU.

5.6.1. Disparity Test

The 3D depth maps estimated from the synthetic Holoscopic data presented in Figure 5.3 above is presented in this section. The 3D depth maps are estimated from VIPs with disparity or baseline sizes varying from 7 to 82 pixels. The

[4] vision.middlebury.edu/stereo/

Synthetic Holoscopic data set has an Elemental image resolution of 100x100 pixels, with a default viewpoint image resolution of 79x53 pixels. The window patch size used for this experiment is a fixed 15x15pixels. The following parameters are presented in Table 5.2 below.

Table 5.2

SH3D image parameters and disparity range values used for the disparity test.

MLA size (Pixels)	VIP size (Pixels)	EI Size (Pixels)	Disparity Range (Pixels)	Patch Size (Pixels)
79x53	79x53	100x100	7 – 82	15x15

The 3D depth maps estimated form the following parameters are as follows.

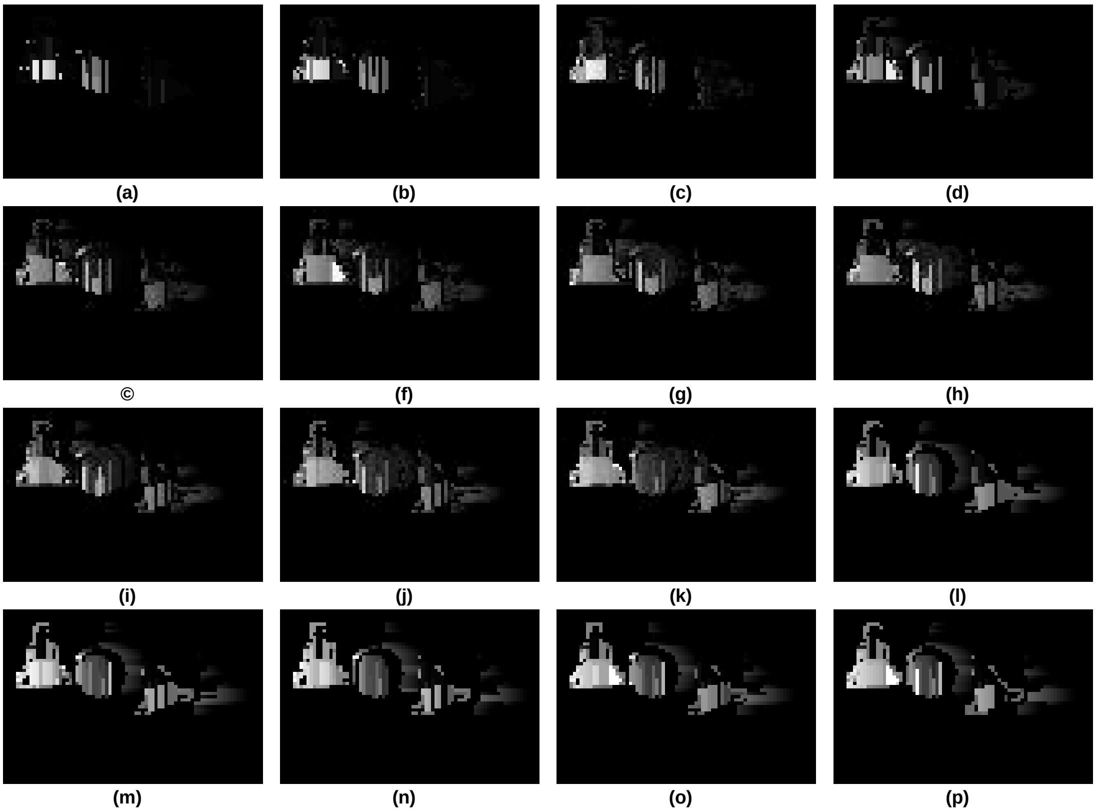


Figure 5.9 – 3D depth maps estimated from the SH3D data set labelled “cones”. All 3D depth maps have a disparity difference of 5pixels between each result.

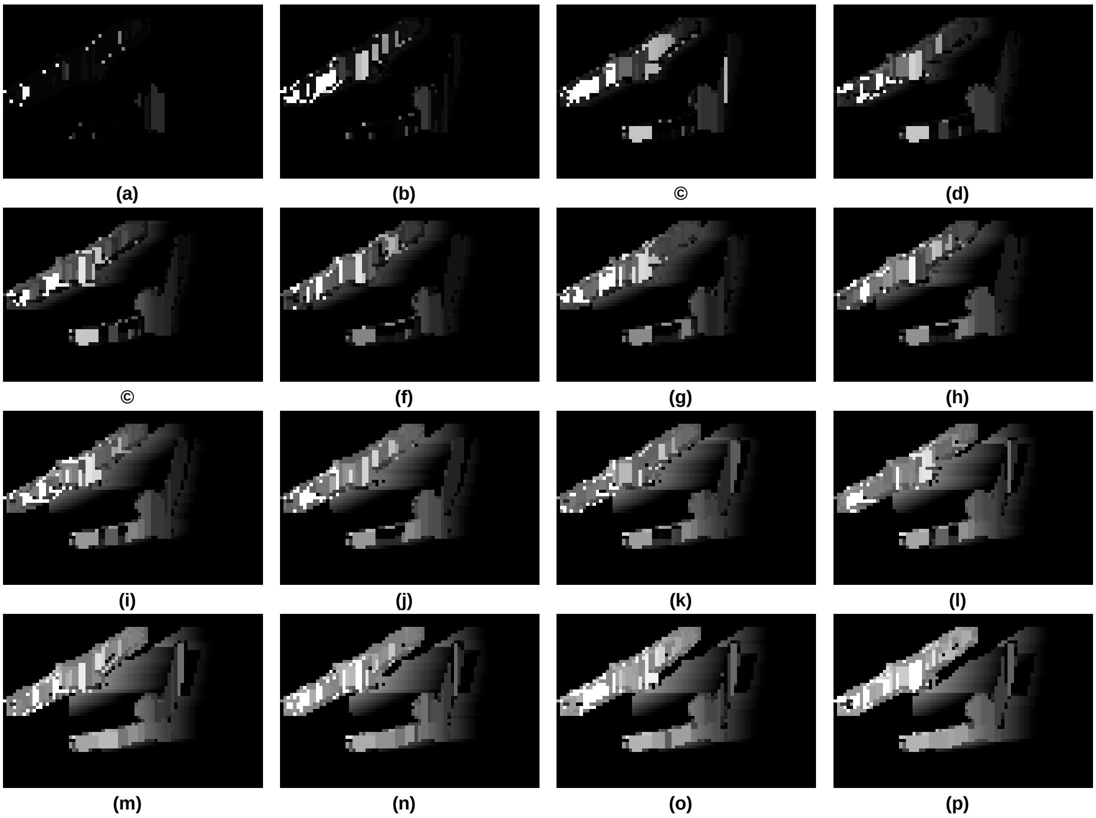


Figure 5.10 – 3D depth maps estimated form the SH3D data set labelled “birds”. All 3D depth maps have a disparity difference of 5pixles between each result.

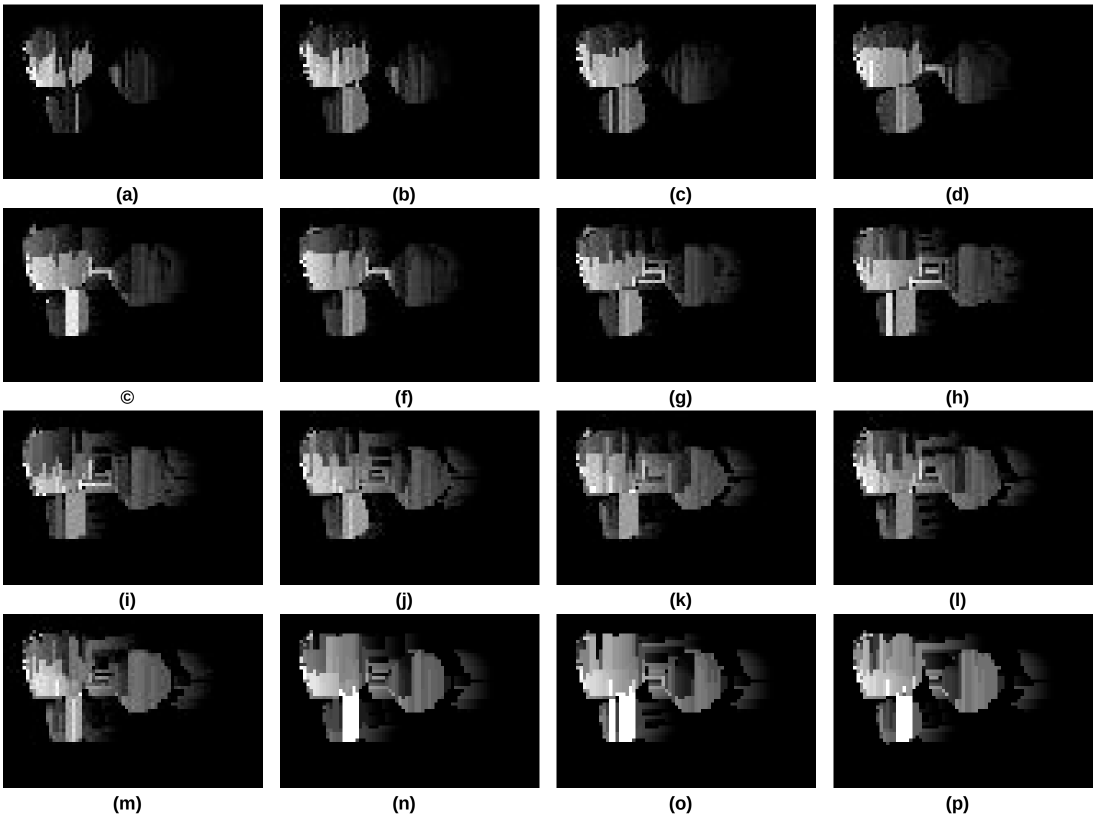


Figure 5.11 – 3D depth maps estimated from the SH3D data set labelled “dice”. All 3D depth maps have a disparity difference of 5pixels between each result.

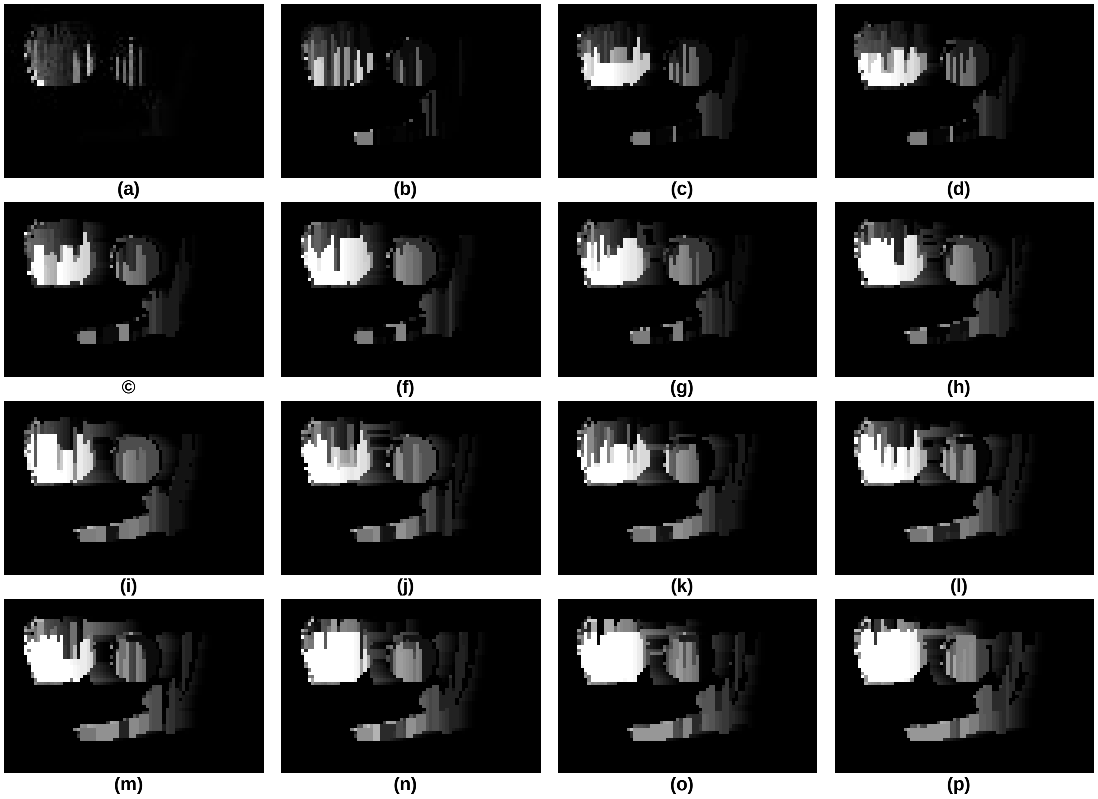


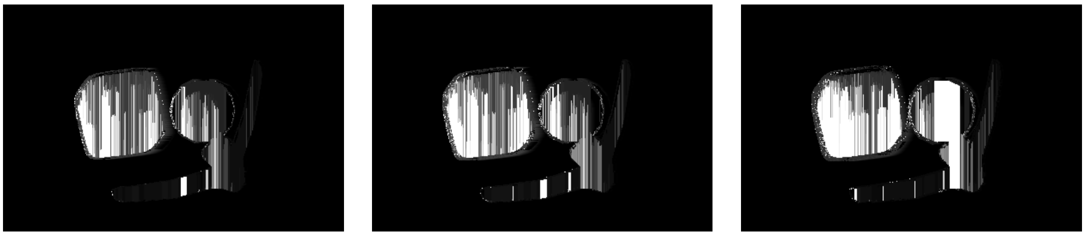
Figure 5.12 – 3D depth maps estimated from the SH3D data set labelled “mixed”. All 3D depth maps have a disparity difference of 5pixels between each result.

The 3D depth results presented in section 5.6.1 indicates the more significant the disparity, the better the 3D depth map. However, the disparity size is restricted by the number of horizontal pixels within the Elemental image and the window size. This leads to the next analysis of finding the best trade-off between Elemental image size and window size.

Secondly, the data set used for this experiment lack adequate texture within them, Figures 5.9 to Figure 5.11 register less accurate disparity values compared to the Figure 5.12 test results as it is rendered with objects of different shapes and textures, making it easier for image features to be matched accurately. Therefore, the Mixed Holoscopic data set is to be used for experiments presented in section 5.6.2 and 5.6.3.

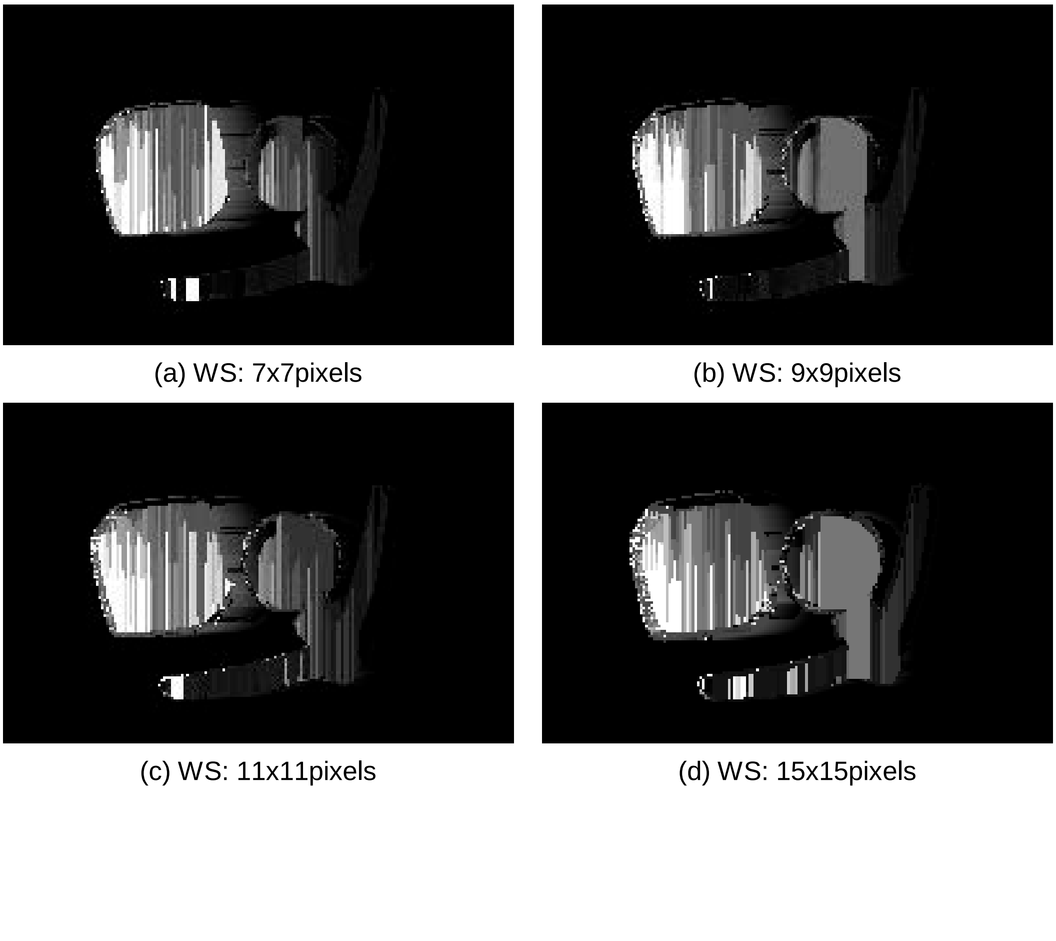
5.6.2. Elemental Image Size vs Block Matching Size Test

In this subsection, the 3D depth results gotten from different Elemental image sizes and viewpoint image resolutions, ranging from 79x53 pixels to 397x265 pixels are presented in Figure 5.13 to Figure 5.16. The results presented below indicates that regardless of the disparity difference or size of the Elemental image, it is still possible to extract 3D depth information from any Holoscopic 3D image. However, since the current DDH technique only extracts the 3D depth information of a viewpoint image from one direction, the maximum disparity size is always limited to the horizontal number pixels. Since the larger the disparity between the viewpoint images the better the 3D depth results, it is clear that H3D images that produce low-resolution viewpoint images will most likely produce better 3D depth results due to their more significant disparity range.



(a) WS: 7x7pixels (b) WS: 9x9pixels (c) WS: 11x11pixels

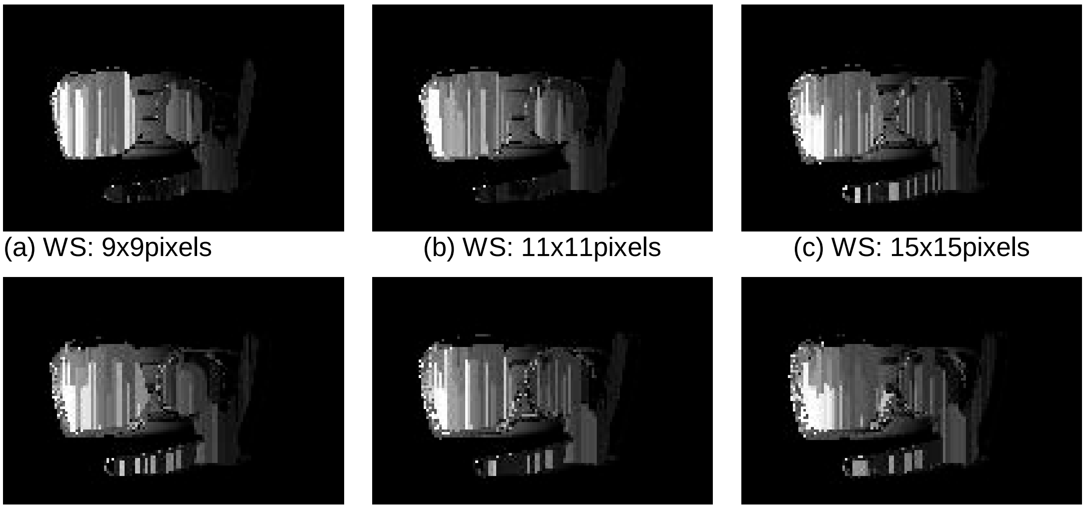
Figure 5.13 – Elemental image size vs matching window test results, estimated from SH3D data set labelled "mixed". The Elemental image size of the SH3D data is 20x20pixels and the viewpoint image size is 397x265pixels.



(a) WS: 7x7pixels (b) WS: 9x9pixels

(c) WS: 11x11pixels (d) WS: 15x15pixels

Figure 5.14 – Elemental image size vs matching window test results, estimated from SH3D data set labelled "mixed". The Elemental image size of the SH3D data is 40x40pixels and the viewpoint image size is 198x132pixels.

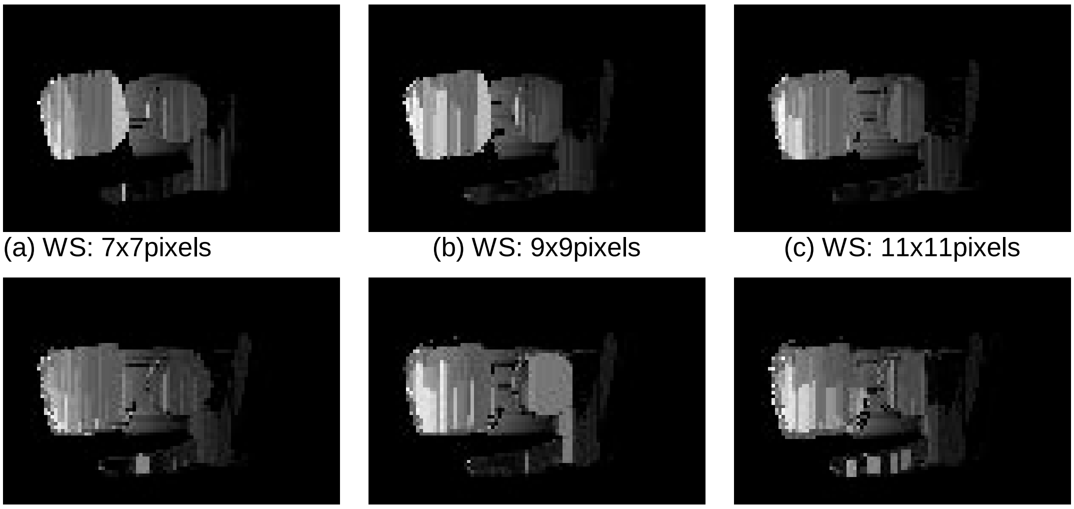


(a) WS: 9x9pixels (b) WS: 11x11pixels (c) WS: 15x15pixels

(a) WS: 17x17pixels (b) WS: 19x19pixels (c) WS: 21x21pixels

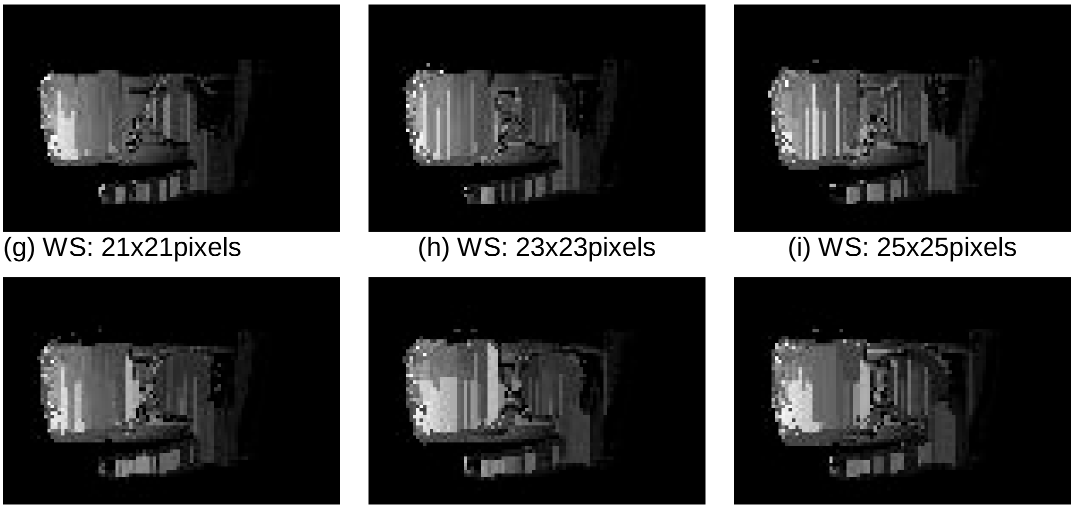
Figure 5.15 – Elemental image size vs matching window test results, estimated from SH3D data set labelled "mixed". The Elemental image size of the SH3D data is 60x60pixels and the viewpoint image size is 132x88pixels.

Secondly, another observation is that, as the window size increases the maximum disparity between the viewpoint images decreases in order to accommodate the window size growth, making it challenging to optimise 3D depth results when there is a low number of pixels in the Elemental images.



(a) WS: 7x7pixels (b) WS: 9x9pixels (c) WS: 11x11pixels

(d) WS: 15x15pixels © WS: 17x17pixels (f) WS: 19x19pixels



(g) WS: 21x21pixels (h) WS: 23x23pixels (i) WS: 25x25pixels

(j) WS: 27x27pixels (k) WS: 29x29pixels (l) WS: 31x31pixels

Figure 5.16 – Elemental image size vs matching window test results, estimated from SH3D data set labelled “mixed”. The Elemental image size of the SH3D data is 80x80pixels and the viewpoint image size is 99x66pixels.

This effect occurred mostly in Elemental images with an image resolution of 20x20 pixels to 40x40 pixels, and the computation cost needed to computed 3D depth information from H3D image increases since the smaller the Elemental image size, the higher the overall number Elemental images the algorithm has to search in order to find the best match.

Elemental image sizes ranging from 60x60 pixels to 100x100 pixels begin to show a better distinction within the disparity maps, however, finding the best trade-off between the window size and disparity is limited as choosing the biggest window does not always result in better 3D depth maps.

5.6.3. Comparison of Proposed Technique

Since there is no H3D depth estimation technique, like the DDH presented in this chapter, the technique is evaluated by comparing the results gotten to other state-of-the-art stereo and 3D depth estimation techniques. The 3D depth technique developed by Schavstein et al. can be downloaded from the Middlebury website [9] and the second is developed Hae-gon et al. [98]. The Middlebury 3D depth technique is implemented to estimate 3D depth information from stereo images recorded by a 2D imaging system. This means the 3D depth algorithm is not implemented to accommodate the H3D system small baseline problem. The second algorithm is implemented by Hae-gon et al., on like the previous algorithm, it is implemented to accommodate the small baseline problem that comes with H3D images. However, they estimate 3D information from an array of viewpoint images which is subjected to some form of viewpoint image interpolation. The results of these comparisons are presented in Figure 5.18 below; however, the viewpoint image sample of the interpolated samples used in presented in Figure 5.17.

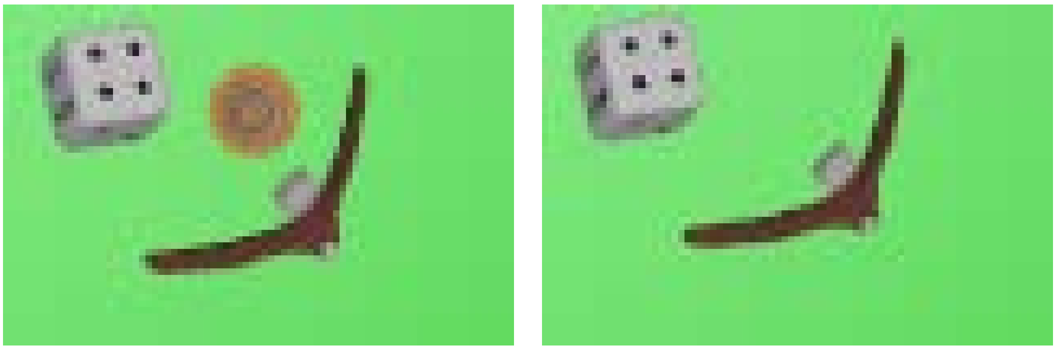
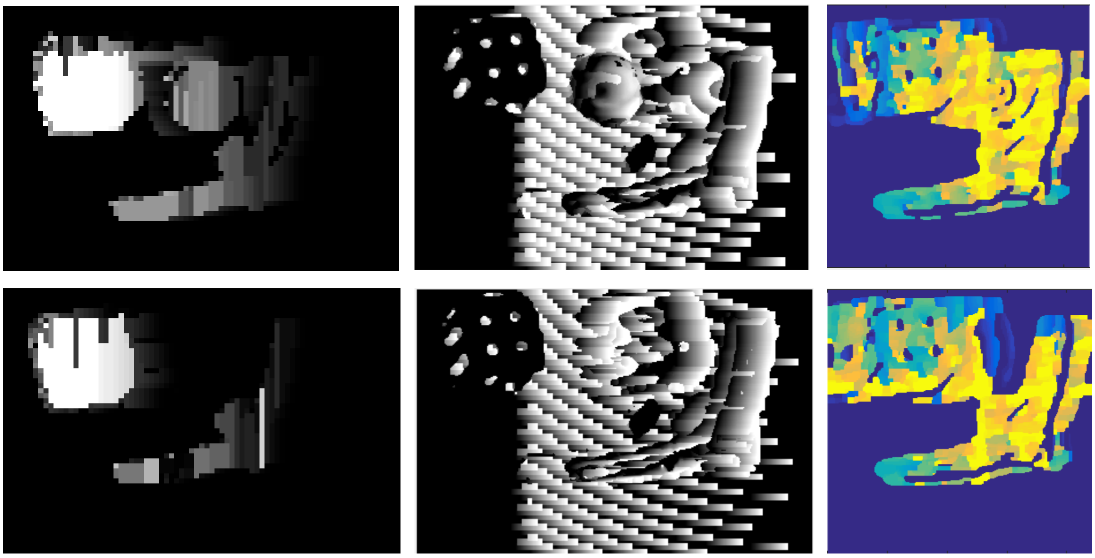


Figure 5.17 – Holoscopic viewpoint image samples used for the direct comparison, up-sampled from a viewpoint image of default size 79x53p, excluding the DDH which is extracted directly .



(a) DDH 3D depth result

(b) Schavstein et al. 3D depth results

(c)Hae-Gon et al. 3D depth results

Figure 5.18 – The innovative DDH 3D depth map compared to other state-of-the-art 3D depth algorithms.

The complexity of the following 3D depth from disparity frameworks are presented in Table 5.3 below, where the number closest to one (1) entails that the 3D depth estimation framework is more complex.

Table 5.3

Complexity evaluation results of the DDH, Szeliski et al. and Hae-gon et al. depth frameworks.

3D Depth frameworks	*Stereo constraints*						
	C1	**C2**	**C3**	**C4**	**C5**	**C6**	**Avg.**
DDH	1	1	0	0	1	0	**0.5**
3D Depth Schavstein et al.	1	1	1	0	1	0	0.6
3D Depth Hae-gon et al.	1	1	1	1	1	0	0.8

The computational cost of the following 3D depth estimation techniques presented in Table 5.3 above is shown in Figure 5.19 below.

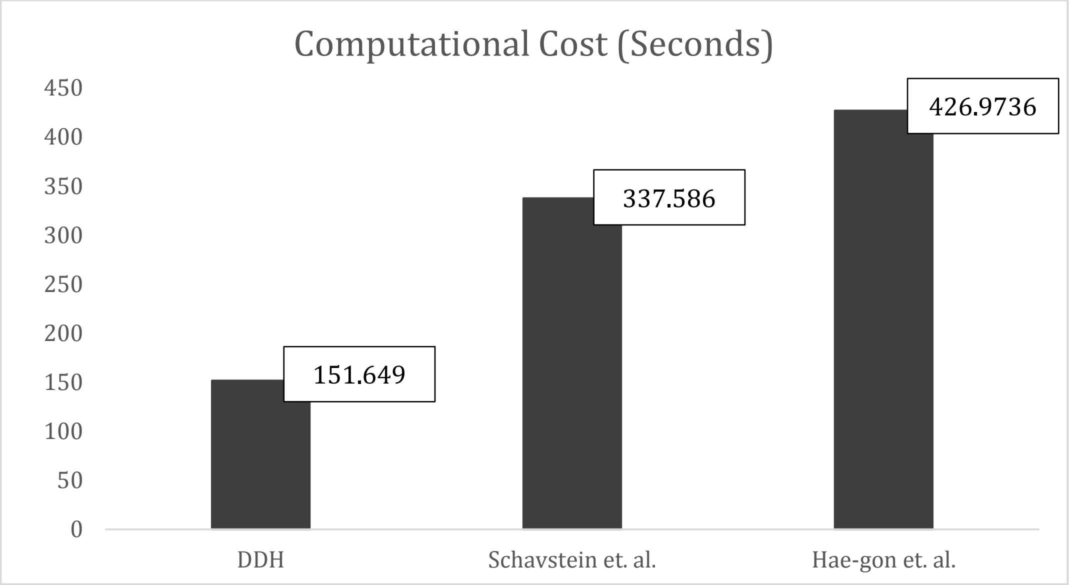


Figure 5.19 – The computational cost of the DDH, Szeliski et al. and Hae-gon et al. measured in seconds.

Figure 5.19 presents the computational cost of the DDH technique presented in this chapter compared to other state-of-the-art techniques. The DDH technique is the most efficient of the three and also extracts higher quality depth maps. The DDH technique can, therefore, be optimised and used in Realtime applications, where all angular information can be assessed, Making this the only true H3D depth estimation technique.

5.7. Summary

The following chapter presents the innovative Direct 3D depth from Holoscopic technique, implemented to estimate 3D depth information directly from a single H3D image without the need of the adaptation of H3D images into stereo viewpoint images. The DDH technique has the ability to estimated 3D depth information of a viewpoint image in both directions of the H3D image. However, only the estimation of 3D depth information from the x-axis is documented in this chapter. The DDH does not require the padding of viewpoint images as the use of VIPs with adequate neighbouring pixels makes it is possible to estimate 3D depth information around image borders accurately.

As the current H3D imaging systems lack MLA of different sizes, a synthetic H3D image database is rendered in POV ray to help deduce the core principle of the DDH technique. On like H3D images that are exposed to lens distortion errors, the synthetic H3D image data is free from any form of distortion, making the H3D data rectified. However, the synthetic data lack detail texturing that can be seen in real H3D data. The data set labelled "mixed" served as the prime data set as it had the most similar properties to the real H3D image data.

The DDH technique is made up of the following three stages; (i) Feature extraction with Census transform, (ii) Similarity measure and 3D depth estimation and (iii) 3D depth optimisation with segmentation. Feature extraction technique in the depth framework can be grouped as an area-based feature extraction technique. The Similarity measure and 3D depth estimation are implemented to traverse along both axis EI of the H3D image, resulting in the estimation of 3D depth at the EI level.

The DDH technique presented in this chapter is the first of its kind as current 3D depth estimation frameworks prefer to convert their 3D image into 2D images, in the process losing a lot of 3D depth information. Therefore, an indirect comparison of this technique and current state-of-the-art stereo matching 3D depth frameworks is conducted by extracting the VIP used for 3D depth estimation by the DDH to stereo viewpoint image. The following viewpoint image is used as a data set for the stereo matching techniques to be compared. Other tests, such as the disparity range test and Elemental image size VS matching window size, were conducted to see the limits of the DDH technique. The main conclusion has been the DDH technique works better with Holoscopic 3D images of bigger and fewer MLAs.

CHAPTER 6: CONCLUSIONS AND FURTHER WORKS

In this chapter, the main findings and contribution in regard to the research topic "Robust 3D depth estimation from an H3D image" is summarised, and general conclusion based on the research contributions presented in this thesis documented. The chapter concludes with the suggestions for further research topics, taking into consideration this thesis limitations and strengths.

6.1. Conclusions

Holoscopic 3D imaging is a new research area. Therefore, there is no standard platform to process H3D data captured by the imaging system. This limits the application areas of which the H3D imaging systems can be used. In place of this system, stereo imaging systems and active systems are used as current industrial systems for recording 3D depth information of a scene, but the following system is costly in comparison to H3D imaging systems. However, due to the limitations associated with the H3D imaging system, industries continue to rely on the current inefficient way of recording and estimating 3D depth information. This gap led to the research "Robust 3D depth estimation from a single H3D image".

The incorporation of the research finding, and contributions presented in this thesis leads to industrial usage of the H3D imaging system, making use of the 3D depth qualities along with other attributes of the H3D imaging system in different application areas. The main research goal is subdivided into aims and objectives that constitute the comprehensive qualitative research on 3D depth sensors and estimation techniques, and the development of scalable H3D depth estimation and H3D pre-processing techniques.

(1) This study led to the design and implementation of multilayer H3D encoding technique that enables efficient viewpoint image extraction from complex H3D data, reducing the computational cost by ***95%***.

(2) This study led to the finding that suggests the consideration of the trade-off between angular and spatial information, resulting in higher quality viewpoint images. The MLA has a direct correlation to the scaling factor of the viewpoint images and the disparity range when adapting H3D data for 3D depth estimation. Therefore, the considerations of MLA aperture sizes when recording H3D for 3D depth is highly recommended. Taking into account of the MLA size can increase the spatial resolution or improve the angular information depending on the H3D 3D depth estimation framework or application area.

(3) This study led to the design and implementation of deep-learning-based single image upscaling technique for H3D images, where the use of EIs and low-resolution 2D data is used as training data. The results found outperformed current H3D upscaling techniques, proving that 2D data can be used as training data for H3D image resolution problems.

Based on the following contributions above, (4) the implementation of a robust H3D 3D depth from disparity framework is implemented. The robust framework has an H3D content adaptation technique capable of extracting semi segmented viewpoint images, making use of the H3D imaging system unique ability to refocus at any point after capture. This contribution then led to the proposal of a 3D depth mapping technique to reduce the computation cost to make the 3D depth estimation technique usable in near real-time applications.

Although the following H3D 3D depth from disparity technique performs better than other depth estimation techniques, the following limitations below still affect the quality of H3D data and their resulting 3D depth maps;

i. The introduction of unwanted artefacts that affect the quality of 3D depth maps.
ii. The loss of valuable angular information when adapting 3D content into 2D content.

The following limitations above led to (5) the implementation of a 3D depth technique that can extract 3D depth information directly from a single H3D image. The innovative Direct 3D depth from Holoscopic (DDH) technique results in 100% preservation of the quality and integrity of the angular information recorded by the H3D imaging system. The innovative technique has a unique similarity measure and 3D depth estimation technique that calculates the disparity of a VIP by the sum of the squared difference between the RVIP Elemental image position and TVIP Elemental image location. The DDH technique also can estimate disparity from both directions of an omnidirectional H3D image.

Based on the research findings for MLA consideration and multilayer H3D image encoding, the development of the second generation H3D imaging system should be considered. The new H3D imaging system having the inbuild version of the proposed multilayer H3D encoding program will make the H3D imaging an industrial tool.

The detail examination and implementation of various pre-processing solutions that are aimed at addressing the current drawbacks of the H3D imaging system should lead to the development of a low-cost H3D image pre-processing platform. The following would allow the average to explore the advantages of the H3D imaging system with can lead to the replacement of multi-lens systems that are found on smartphones like "iPhone 11" and "Samsung Galaxy X". The H3D image pre-processing findings have already been used in the CEPROHA project where the digitization of cultural assets was achieved.

Based on the H3D 3D depth estimation frameworks presented in this thesis, the implementation of cost-effective applications that are heavily dependent on 3D depth information is possible. These applications include autonomous navigation, 3D digitization, and other various depth-related applications, further expanding the use of H3D imaging systems in communities as there is no doubt that it will eventually replace conventional 2D imaging systems in the near future.

6.2. Further Work

The development of an automated technique for extracting viewpoint images could be a sensible research area. This research area will lead to the redesign of the current H3D imaging system with perfect MLA calibration to the sensor. Followed by a deep-learning-based viewpoint extraction technique that can automatically correct lens distortion errors and extract the viewpoint image with minimal input from the user.

The implementation of a 3D depth mapping technique that can be used to map low-resolution 3D depth maps to their respective high respective resolution counterparts. In doing so, when estimating 3D depth information form video H3D data or multi-frame H3D data, the estimation and optimization of low-resolution 3D depth to high-resolution 3D depth maps will significantly reduce the computational expensive.

Further research into 3D scene reconstruction from H3D video data is also research that benefits from this research.

Further research into the extension of the direct 3D depth from Holoscopic 3D image technique to accommodate real H3D images is expected. This is a possibility because synthetic H3D images where used to prove and standardize the DDH core 3D depth estimation principles. The synthetic H3D used had no lens-related errors, where the real H3D data will at least have lens distortion error which will undoubtedly lead to extension and optimization of the current DDH framework to accommodate more image restrictions.

A. Appendix

This appendix provides additional details about the Holoscopic 3D computer graphics data set in this thesis. The CG H3D data is used in chapter 3 and chapter 5 for image sampling evaluation and direct Holoscopic depth estimation.

POV ray configuration of Holoscopic 3D image rendering

Parameter	*Value*	*Description*
MLA size	86x48	Includes MLA affected by vignetting.
Valid MLA size	84x47	Only MLA not affected by vignetting.
Dot pixel pitch	0.00451mm	Size of single pixel on sensor.
Censor Length	35.9x24mm	Width and length of sensor.
Max resolution	7952x5304 pixels (8K)	Max. H3D image size.
Max Obj. in scene	3	

Scene properties	No shadow	Do not render shadows of scene.
Scene gamma	2.2	
Focal length range	1:0.5:11	
Depth range	Focal length*EI length	

CG Holoscopic 3D image data set properties

Square MLA size (mm)	*EI size / No. VPIs (pixels)*	*Default VPI size (pixels)*
0.451^2	100^2	79x53
0.4059^2	90^2	88x58
0.3608^2	80^2	99x66
0.3151^2	70^2	113x75
0.2706^2	60^2	132x88
0.2255^2	50^2	159x106
0.1804^2	40^2	198x132
0.1353^2	30^2	265x178
0.0902^2	20^2	397x265
0.0451^2	10^2	795x533

B. Appendix

This appendix provides details about the patch-based interpolation [99] technique used in chapter 4 for viewpoint extrapolation.

A Holoscopic 3D image is made up of a set of Elemental images and the size of each Elemental image equals the number of viewpoint images recorded in a single snapshot. Furthermore, the total number of micro lens array (MLA) equals the default size of each viewpoint image recorded. This resulting to one of the major setbacks of the Holoscopic imaging system, its inability to record high resolution that results to difficulty in feature matching. However, details of the patch-based image sampling technique [99] discussed in chapter 4, is utilized to avoid mismatching problems associated with the low-resolution set back. This defined in Equation B.1 below as:

$$SRVI = \sum_{k=1}^{N}\sum_{l=1}^{N} K_{n+\delta_x(1+i),i,\ \ m+\delta_y(1+j),j} \qquad (B.1)$$

Where the up-sampled list of reference images LRVPIs (K) and the resulting super resolution Viewpoint image is donated as SRVI. The coordinates of the SRVI is donated as n and m while the Vis index numbering range from 1 to N and represented by k and l. The omnidirectional sub-pixel shift parameters are δ_x and δ_y respectively. The shift parameter value δ is defined as:

$$\delta = \frac{pixel\ size}{N} \qquad (B.2)$$

However, due to over or underfitting of the reconstruction of SRVI, the resulting image do not have well defined features [100], resulting in the smoothing and sharping of the shift-integrated LRVPIs with Gaussian blurring low-pass filter kernel and image sharping techniques [101].

C. Appendix

This appendix provides details about the Holoscopic 3D image data set used in conjunction with 2D image data sets for viewpoint image up-sampling with deep learning in this thesis.

As mentioned in chapter two section three, deep-learning based networks require a large data set in order to accurately create a regression model. In terms of image up-sampling, the network is used to deduce the difference between up-sampled viewpoint images and their reference image. In doing so, the successful and accurate up-sampling of the Holoscopic viewpoint images is achieved. The data collection guide for testing and training learning-based networks is listed below.

i. Acquisition of raw benchmark Holoscopic 3D data of size ranging from 2k to 8k in pixels.
ii. Raw Holoscopic 3D data is subdivided as two classes, training and test data.

iii. The training data of size ranging from 2k to 8k is then subsampled and broken down into Elemental or sub aperture images.

iv. The raw Holoscopic data is also broken down into Elemental images and used as reference data of the previously subsampled EIs.

v. As this data is insufficient, 2D image data is used to supplement the training data.

vi. Once the network creates a regression model that can up-sample the training data with high accuracy.

vii. The up-sampled Elemental image is then reconstructed into a single Holoscopic image.

viii. The results evaluated with PSNR and SSIM, state-of-the-art image evaluation metric.

REFERENCE

[1] R. M. Anderson, A. S. Souza and L. M. G. Goncalves, *Current Advancements in Stereo Vision, Chapter 5: Depth Estimation - An Introduction*. Intech, 2012.

[2] S. A. El Mesloul Nasri, A. H. Sadka, N. Doghmane, and K. Khelil, "Group of pictures effects on proposed multiview video coding scheme," *2017 40th Int. Conf. Telecommun. Signal Process. TSP 2017*, vol. 2017-Janua, pp. 548–554, 2017.

[3] K. E. Ozden, K. Cornelis, L. Van Eycken, and L. Van Gool, "Reconstructing 3D trajectories of independently moving objects using generic constraints," *Comput. Vis. Image Underst.*, 2004.

[4] H. Pan, T. Guan, Y. Luo, L. Duan, Y. Tian, L. Yi, Y. Zhoa, J. Yu, "Dense 3D reconstruction combining depth and RGB information," *Neurocomputing*, vol. 175, pp. 644–651, 2016.

[5] F. E. Ives, "Parallax stereogram and process of making same," US patent 725,567, Apr. 14, 1902.

[6] G. Lippmann, "La Photographies Integrale," *Comtes Rendus, Academie des Sciences*, vol. 146, pp. 446–451, 1908.

[7] G. Lippmann, "Epreuves reversibles donnant la sensation du relief," *J. Phys.(Paris), Vol. 7, pp. 821-825,* 1908.

[8] A. Bogusz, "Holoscopy and holoscopic principles," J. Opt. 20(6), 281–284, 1989.

[9] D. Scharstein, R. Szeliski, and R. Zabih, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," in *Proceedings - IEEE Workshop on Stereo and Multi-Baseline Vision, SMBV 2001*, 2001.

[10] M. A. Lozano and F. Escolano, "Graph matching and clustering using kernel attributes," *Neurocomputing 113:177-194*, 2013.

[11] J. Tanida, T. Kumagai, K. Yamada, S. Miyatake, and K. Ishida, "Thin observation module by bound optics (TOMBO) : concept and experimenal verification" Applied Optics 40, 11 (April) 1806-1813, 2001.

[12] J. M. Benjamin, "The Laser Cane," *Bull. Prosthet. Res.*, pp. 443–450, 1974.

[13] J. Jung, J. Y. Lee, Y. Jeong, and I. S. Kweon, "Time-of-Flight Sensor Calibration for a Color and Depth Camera Pair," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 7, pp. 1501-1503. Jul. 2015.

[14] F. Devernay and P. Beardsley, "*Stereoscopic Cinema*," in *Image and Geometry Processing for 3D Cinematograpy*, Remi Ronfard and Gabriel Taubin, Eds. Springer-Verlag, 2010.

[15] C. Wheatstone, "*Contributions to the physiology of vision - Part the second. On some remarkable, and hitherto unobserved, phenomena of binocular vision,*" Philosophical transaction of the Royal Society 142, 1-17, 1852.

[16] H. Urey, K. V Chellappan, E. Erden, and P. Surman, "State of the Art in Stereoscopic and Autostereoscopic Displays," *Proc. IEEE*, vol. 99, no. 4, pp. 540–555, 2011.

[17] C. Zhang and T. Chen, "*Multi-View Imaging : Capturing and Rendering Interactive Enviroments*," Computer Vision for Interactive and Intelligent Enviroment, pp.51-67, Nov 2005.

[18] P. T. Boufounos, "Depth sensing using active coherent illumination," *in Proc. IEEE Int. Conf. Acoust. Speech and Signal Processing*, March 25-30, 2012.

[19] E. A. Bernal, L. K. Mestha, and E. Shilla, "Non contact monitoring of

respiratory function via depth sensing," *IEEE-EMBS Int. Conference on Biomedical and Health Informatics*, pp. 101–104, 2014.

[20] T. Y. Kuo and Y. C. Lo, "Depth estimation from a monocular view of the outdoors," *IEEE Trans. Consumer Electron.*, vol 57, no. 2, pp.817-822, May 2011.

[21] J. Lin, X. Ji, W. Xu, and Q. Dai, "Absolute depth estimation from a single defocused image," *IEEE Trans. Image Process.*, vol 22, no. 11, pp. 4535-4550, Nov. 2013.

[22] D. Scharstein, R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondance algorithms," IJCV 47(1-3), 7-42, 2002.

[23] R. Szeliski, R. Zabih, D. Scharstein "A Comparative Study of Energy Minimization Methods for Markov Random Fields with Smoothness-Based Priors," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 6, pp. 1068-1080, June 2008.

[24] G. Medioni and R. Nevatia, "Segment-based stereo matching.," *Comput. Vision, Graph. Image Processing*, vol. 31, no. 1, pp. 2–18, 1985.

[25] Y. S. Heo, K. M. Lee, and S. U. Lee, "Joint depth map and color consistency estimation for stereo images with different illuminations and cameras," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1094-1106, May 2013.

[26] N. Kwak and J.-W. Lee, "Feature extraction based on subspace methods for regression problems," *Neurocomputing* 73 (10), 1740-1751, 2010.

[27] X. Luan, F. Yu, H. Zhou, X. Li, D. Song, and B. Wu, "Illumination-robust area-based stereo matching with improved census transform," *Proc. 2012 Int. Conf. Meas. Inf. Control. MIC 2012*, vol. 1, no. Mic, pp. 194–197, 2012.

[28] M. B. Hisham, S. N. Yaakob, R. A. A. Raof, A. B. A. Nazren, and N. M. W. Embedded, "Template Matching using Sum of Squared Difference and Normalized Cross Correlation," *2015 IEEE Student Conf. Res. Dev. (SCOReD). IEEE*, pp. 100–104, 2015.

[29] S. Birchfield and C. Tomasi, "A pixel dissimilarity measure that is insensitive to image sampling," *IEEE Trans. Pattern Anal. Mach. Intell.*, 1998.

[30] G. S. Cox, "Template Matching and Measures of Match in Image Processing," Review paper, Dept. of E.E., University of Cape Town, 1995.

[31] S. Wei and S. Lai, "Fast Template Matching Based on Normalized Cross Correlation With Adaptive Multilevel Winner Update," IEEE Trans. Image Processing, vol. 17, no. 11, pp. 2227–2235, Nov. 2008.

[32] T. Tomioka, K. Mishiba, Y. Oyamada, and K. Kondo, "Depth map estimation using census transform for light field cameras," *IEICE Trans. Inf. Syst.*, vol. E100D, no. 11, pp. 2711–2720, 2017.

[33] S. Guo, P. Xu, and Y. Zheng, "Semi-global matching based disparity estimate using fast Census transform," *Proc. - 2016 9th Int. Congr. Image Signal Process. Biomed. Eng. Informatics, CISP-BMEI 2016*, no. 61401398, pp. 548–552, 2017.

[34] C. Schmid, A. Zisserman, and R. Mohr, "Integrating Geometric and Photometric Information for Image Retrieval," In International Workshop on Shape, Contour and Grouping in Computer Vision, pp. 217–233, 1998.

[35] L. Szumilas, H. Wildenauer, and A. Hanbury, "Invariant Shape Matching for Detection of Semi-local Image Structures," in Lecture Notes in Computer Science vol. 5627, pp. 551–562, 2009.

[36] Y. Xia, A. Tung, S. Ho, and Y. Ji, “A novel wavelet stereo matching method to improve DEM accuracy generated from SPOT stereo image pairs,” *IGARSS 2001. Scanning Present Resolv. Futur. Proceedings. IEEE 2001 Int. Geosci. Remote Sens. Symp. (Cat. No.01CH37217)*, vol. 7, no. C, pp. 3277–3279 vol.7, 2001.

[37] G. Pajares, H. Manuel, D. Cruz, and H. A. Lo, “Relaxation labeling in stereo image matching,” Pattern Recog., vol. 33, no. 1, pp. 53-68, Jan. 2000.

[38] Y. Jia, Y. Xu, W. Liu, C. Yang, and Y. Zhu, “A Miniature Stereo Vision Machine for Real-Time Dense Depth mapping,” Proceedings of Third International Conference of Computer Vision Systems 2626, pp. 268–277, 2003.

[39] S. P. Balan, P. Leya, and E. Sunny, “Survey on Feature Extraction Techniques in Image Processing,” International Journal for Research in Applied Science and Engineering Technology (IJRASET) vol. 6, no. 3, pp. 217–222, 2018.

[40] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” IJCV vol. 60, no. 2, pp. 91–110, 2004.

[41] H. Bay, T. Tuytelaars, and L. Van Gool, “SURF : Speeded Up Robust Features.” in Proc. Eur. Conf. Computer Vision, Graz, Austria, pp. 404-417, May 2006.

[42] N. Dalal, B. Triggs, and D. Europe, “Histograms of Oriented Gradients for Human Detection,” Proc. IEEE Conference Computer Vision and Pattern Recognition, 2005.

[43] S. Paris, F. X. Sillion, and L. Quan, “A surface reconstruction method using global graph cut optimization,” in *International Journal of Computer Vision*, 2006.

[44] P. Kamencay and M. Breznan, “A Stereo Depth Recovery Method Using Belief Propagation.” In Proceedings of 21st International Conference Radioelektronika. Brno (Czech Republic), 2011, p. 383-386. 2011.

[45] T. Collins, “Graph Cut Matching In Computer Vision,” *Computing*, no. February, pp. 1–10, 2004.

[46] M. Bleyer and M. Gelautz, “Graph-cut-based stereo matching using image segmentation with symmetrical treatment of occlusions,” *Signal Process. Image Commun.*, vol. 22, no. 2, pp. 127–143, 2007.

[47] S. Roy and I. J. Cox, “A maximum flow formulation of the n-camera stereo correspondence problem,” *IEEE Comput. Vis.*, pp. 492–499, 1998.

[48] Y. Boykov, O. Veksler, and R. Zabih, “Fast approximate energy minimization via graph cuts,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, 2001.

[49] V. Kolmogorov and R. Zabih “Computing Visual Correspondence with Occlusions using via Graph Cuts,” Proc. Int'l Conf. Computer Vision, vol 2, pp. 508–515, 2001.

[50] V. Kolmogorov and R. Zabih, “Multi-camera Scene Reconstruction via Graph Cuts.” Proc. Seventh European Conf. Computer Vision, vol 3, pp. 82-96, May 2002.

[51] V. Kolmogorov and R. Zabih, “What Energy Functions Can Be Minimized via Graph Cuts ?,” IEEE Trans. Pattern Analysis and Machine Intelligence vol. 26, no. 2, pp. 147–159, Feb. 2004.

[52] V. Kolmogorov, R. Zabih, and S. Gortler, “Generalized Multi-camera Scene Reconstruction Using Graph Cuts,” EMMCVPR, pp. 501–516, 2003.

[53] Y. Boykov, O. Velsler and R. Zabih, “Markov random fields with efficient approximations.” Proc. IEEE Conf. Computer Vision and Pattern

Recognition, pp. 648-665, 1998.

[54] C. L. Zitnick and S. B. Kang, "Stereo for image-based rendering using image over-segmentation," *Int. J. Computer Vision*, vol. 75, pp. 49-65, Oct. 2007.

[55] J. Sun, N.-N. Zheng, and H.-Y. Shum, "Stereo Matching Using Belief Propagation." in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 7, pp. 787-800, July 2003.

[56] A. P. Pentland, "New Sense for Depth of Field," *IEEE Trans. Pattern Analysis and Machchine Intelligence*, vol. PAMI-9, no. 4, pp. 523–531, 1987.

[57] M. Subbarao, "Parallel Depth Recovery by Changing Camera Parameters," *Proc. IEEE Conf. Vis. Pattern Recognit.*, pp. 149–155, 1988.

[58] P. Favaro, S. Soatto, M. Burger, and S. J. Osher, "Shape from defocus via diffusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2008.

[59] C. Li, S. Su, Y. Matsushita, K. Zhou, and S. Lin, "Bayesian depth-from-defocus with shading constraints," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013.

[60] S. W. Bailey, J. I. Echevarria, B. Bodenheimer, and D. Gutierrez, "Fast depth from defocus from focal stacks," *Vis. Comput.*, vol. 31, no. 12, pp. 1697-1708, Dec. 2015.

[61] C. Zhou, S. Lin, and S. K. Nayar, "Coded aperture pairs for depth from defocus" *in Proceedings of ICCV*, pp. 325-332, 2009.

[62] A. Saxena, S. H. Chung, and A. Y. Ng, "3-D depth reconstruction from a single still image," *Int. J. Comput. Vis.*, vol. 76, pp. 53-69, 2008.

[63] S. Choi, D. Min, B. Ham, Y. Kim, C. Oh, and K. Sohn, "Depth Analogy: Data-Driven Approach for Single Image Depth Estimation Using Gradient Samples," *IEEE Trans. IMAGE Process.*, vol. 24, no. 12, 2015.

[64] W. Yang, X. Zhang, Y. Tian, W. Wang, J. Xue, and Q. Liao, "Deep Learning for Single Image Super-Resolution : A Brief Review," in IEEE Transactions on Multimedia, vol. 21, no. 12, pp. 1–17, Dec. 2019.

[65] C. Dong, C. C. Loy, and K. He, "Image Super-Resolution Using Deep Convolutional Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, 2016.

[66] Y. B. Goodfellow, Ian, Y. Bengio, A. Courville, *Deep Learning*. Cambridge: MIT press, 2016.

[67] C. C. Chang, "Adaptive multiple sets of CSS features for hand posture recognition," *Neurocomputing*, vol. 69, no. 16-18, pp. 2017-2025, Oct. 2006.

[68] M. Hosseini Kamal, B. Heshmat, R. Raskar, P. Vandergheynst, and G. Wetzstein, "Tensor low-rank and sparse light field photography," *Comput. Vis. Image Underst.*, vol. 145, pp. 172-181, 2016.

[69] L. Deng, "A tutorial survey of architectures , algorithms, and applications for deep learning" APSIPA Transaction Signal Inf. Process, vol. 3, no. May, 2014.

[70] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436-444, 2015.

[71] D. Glasner, S. Bagon, and M. Irani, "Super-Resolution from a Single Image," *2009 IEEE 12th Int. Conf. Comput. Vis.*, no. Iccv, pp. 349–356, 2009.

[72] J. Kim, J. K. Lee, and K. M. Lee, "Accurate Image Super-Resolution Using Very Deep Convolutional Networks," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1646-1654, 2016.

[73] Y. Bengio, P. Simard, P. Frasconi, and S. Member, "Learning Long-Term Dependencies with Gradient Descent is Difficult," in IEEE Transactions on

Neural Networks, vol. 5, no. 2, pp. 157-166, March 1994.

[74] W. Shi, J. Caballero, F. Huszar, J. Totz,.. “Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network,” *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1874–1883, 2016.

[75] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, “Residual Dense Network for Image Super-Resolution,” *Comput. Vis. Pattern Recognit.*, pp. 2472-2481, 2018.

[76] Y. Fan, Z. Wang, and X. Wang, “Wide Activation for Efficient and Accurate Image Super-Resolution,” *Comput. Vis. Pattern Recognit.*, 2018.

[77] A. Horé and D. Ziou, “Image quality metrics: PSNR vs. SSIM,” *Proc. - Int. Conf. Pattern Recognit.*, pp. 2366–2369, 2010.

[78] R. Kreis, “Issues of spectral quality in clinical 1 H-magnetic resonance spectroscopy and a gallery of artifacts,”,NMR in Biomedecine, vol. 17, no. 6, pp. 361–381, 2004.

[79] I, Avcibas, B. Sankur, and K. Sayood, “Statistical evaluation of image quality measures,” J. Electron. Imag., vol. 11, no. 2, pp. 206–223, Apr. 2002.

[80] M. Bleyer and M. Gelautz, “A layered stereo matching algorithm using image segmentation and global visibility constraints,” *ISPRS J. Photogramm. Remote Sens.*, vol. 59, no. 3, pp. 128–150, 2005.

[81] P. Slav and M. Cadik, “Evaluation of Two Principal Approaches to Objective Image Quality Assessment.” Proceedings. Eighth International Conference on Information Visualisation, pp. 513-518, 2004.

[82] T. B. Nguyen and D. Ziou, “Contextual and Non-Contextual Performance Evaluation of Edge Detectors” Pattern Recognition Letteres, vol. 21, no. 9, pp. 805–816, 2000.

[83] Z. Wang, A. C. Bovik, H. R. Sheikh, S. Member, E. P. Simoncelli, and S. Member, “Image Quality Assessment : From Error Visibility to Structural Similarity,” in IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, April 2004.

[84] A. Fetić and D. Jurić, “The procedure of a camera calibration using Camera Calibration Toolbox for MATLAB,”2012 Proceedings of 35th International Convention MIPRO, pp. 1752–1757, 2012.

[85] A. Bushnevskiy , L. Sorgi and B.Rosenhahn, "*Multimode camera calibaration*" *2016 IEEE Int. Conf. Image Process. (ICIP)*, Phoenix, AZ, pp. 1165–1169, 2016.

[86] H. Shang, F. Zhao and H. Zhao, “The analysis of errors for field experiment based on POV-Ray,” IEEE International Geoscience and Remote Sensing Symposium, Munich, pp. 4805-4808, 2012.

[87] A. Prajapati, S. Naik, and S. Mehta, “Evaluation of Different Image Interpolation Algorithms,” *Int. J. Comput. Appl.*, vol. 58, no. 12, pp. 6–12, 2012.

[88] H. Hirschmuller and D. Scharstein, “Evaluation of Cost Functions for Stereo Matching,” *Conf. Comput. Vis. Pattern Recognit.*, MN, pp. 1–8, 2007.

[89] D. Scharstein, “High-resolution stereo datasets with subpixel-accurate ground truth,” *in Proceedings* German Conference of Patternn Recognition. (GCPR) vol. 8753, no. 1, pp. 31–42, 2014.

[90] D. Wan and J. Zhou, “Stereo vision using two PTZ cameras,” *Comput. Vis. Image Underst.*, vol 112, no. 2, pp. 184-194, 2008.

[91] P. R. Rajarapollu and V. R. Mankar, “Bicubic Interpolation Algorithm

Implementation for Image Appearance Enhancement," International Journal of Computer Science and Technology (IJCST), vol. 8, no. 2, pp. 23–26, 2017.

[92] N. Shao, H. L. Le, and L. Z. Zhang, "Stereo vision robot obstacle detection based on the SIFT," Second WRI Global Congress on Intelligent Systems, Wuhan, pp. 274–277, 2010.

[93] S. Ramalingam, S. K. Lodha, and P. Sturm, "A generic structure-from-motion framework," *Comput. Vis. Image Underst.*, vol. 103, no. 3, pp. 218-228, 2006.

[94] P. Thevenaz, U. E. Ruttimann, M. Unser, and S. Member, "A Pyramid Approach to Subpixel Registration Based on Intensity," IEEE Transactions on Image Processing, vol. 7, no. 1, pp. 27–41, 1998.

[95] A. Koschan, V. Rodehorst, and K. Spiller, "Color Stereo Vision Using Hierarchical Block Matching and Active Color Illumination," Proceedings of 13th International Conference on Pattern Recognition,vol. 1, pp. 835–839, 1996.

[96] J. Ha and H. Jeong, "Pyramid Structure for DP-based Stereo Matching Algorithm Department of Electrical Engineering Department of Electrical Engineering," *2011 6th Int. Conf. Comput. Sci. Converg. Inf. Technol.*, pp. 245–248, 2011.

[97] F. Solina, W. G. Kropatsch, R. Klette, A. Koschan, and V. Rodehorst, "Dense depth maps by active color illumination and image pyramids," *Adv. Comput. Vision. Adv. Comput. Sci.*, pp. 137–148, 1997.

[98] H. G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y. Tai and I. Kweon "Accurate depth map estimation from a lenslet light field camera," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, MA, pp. 1547-1555, 2015.

[99] H. Ur and D. Gross, "Improved resolution from Sub-pixel Shifted Pictures," *CVGIP: Graphical Models and Image Processing*, Vol. 54. No. 2, pp. 181-186, 1992.

[100] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Computer Vision*, Vol. 60, No. 2, pp. 91-110, 2004.

[101] H. Greenspan, C. Anderson, and S. Akber, "Image Enhancement by Nonlinear Extrapolation in Frequency Space," *IEEE Transactions on Image Processing*, Vol. 9, No. 6, 2000.