Stereo Matching Based Estimation of Depth Map from Stereo Image Pair

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Technology

in

Computer Science and Applications

Submitted By

Khushboo Jain
(Roll No. 601403014)

Under the Supervision of:

Dr. Husanbir Singh Pannu
Lecturer

Ms. Vineeta Bassi
Assistant Professor

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR UNIVERSITY
PATIALA-147004
JUNE, 2016

Certificate

I hereby certify that the work which is being presented in the thesis entitled, **"Stereo Matching based Estimation of Depth Map from Stereo Image Pair"**, in partial fulfillment of the requirements for the award of degree of Master of Technology in Computer Science and Applications submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of *Dr. Husanbir Singh Pannu*, *Ms. Vineeta Bassi* and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Khushboo Jain

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

Dr. H.S.Pannu
Lecturer
CSED

Ms.Vineeta Bassi
Assistant Professor
CSED

Countersigned by:

Dr. Maninder Singh
Head
Computer Science and Engineering Dept.
Thapar University
Patiala

Dr. S.S.Bhatia
Dean (Academic Affairs)
Thapar University
Patiala

Acknowledgment

First and foremost, I would like to thank my supervisors **Dr. Husanbir Singh Pannu** and **Ms. Vineeta Bassi** for their guidance and encouragement through out my work. They have set a high academic and professional standard for me to follow. I would also like to thank Mr. Dilbag Singh, for all his patience, time and scholarly insight provided in this work. I am grateful to Mr. Gaurav Vijayvargiya from IIT Bombay who brought this topic to my knowledge.

I would like to thank **Dr. S.S.Bhatia**, Dean of Academic Affairs, for giving provisions of the entire required infrastructure such as computer labs, library facilities, immensely useful for learners to equip themselves with the latest in the field.

I take this opportunity to express my appreciations towards the **Dr. Maninder Singh**, Head CSED, for his kind help and cooperation, and other faculty members for their constructive suggestions through out the research. I am also thankful to all my respected teachers in the Department and all my friends, for their direct or indirect help, inspiration and motivation.

I want to express my greatest gratitude to my dear parents for their endless love, constant support. Last but not least, I would like to thank **Lord Ganesha** for everything.

Khushboo Jain

Abstract

Stereo vision is a technique of depth perception, in which the information about the depth is inferred from the two (or more) images captured from different perspectives of a scene. These images is known as stereo image pair. Practical applications where stereo vision technology plays a role may include autonomous vehicle guidance, aerial photogrammetry, robotics vision, object tracking and industrial automation. Many automated vision systems could benefit substantially from depth maps.

A depth map is a grayscale image that contains depth information for each pixel in an image. Traditionally depth maps were extracted using stereo camera approach. Depth estimation from stereo involves finding disparities along the same scanline, also called as epipolar lines. Such a search process typically requires a prior adjustment of the images known as rectification step to ensure that epipolar lines are well aligned. Still, the approaches for estimation of depth map suffer from either limited reliability and robustness when tested on stereo image pair or large time of computation. A novel approach for depth estimation that integrates the image filtering into a stereo matching framework is introduced. Experiment helps verifying the sustenance of high quality in depth maps, while reducing the average percent of bad pixels to 3.58%.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Background

Using a camera for capturing and storing visual data has been in use for long. However, it was the use of visual data to accomplish complete or partial automation that gave birth to machine vision. Many researchers have been, and still are, exploring the domain of extracting useful information from images and videos. Understanding a scene structure has been a tremendous problem in machine vision. Given a few views of a scene, research is in progress to build complete systems that can estimate the structure of the scene.

One of the crucial steps in understanding a scene is to estimate a depth map. A depth map is a map which stores for each pixel a value that quantifies the distance of the object from the camera. Stereo vision technique is used for depth perception, in which the information about the depth can be inferred with the help of two (or more) images captured from different perspectives of the scene. A more detailed explanation is provided in the further section of this chapter. However, the human ability to perform stereo vision has been observed by scientists since the 19th century. The human vision system routinely and with apparent ease perceives the depth of surroundings and 3-D spatial relationships of objects. However, till date computerized solutions for stereo vision computation fall far short of the human vision ability, especially when it comes to natural or real time images. Algorithmic solutions made for artificial stereo visionary are therefore highly optimized for a particular application area. Applications for stereo vision include, but are not limited to aerial photogrammetry, autonomous vehicle guidance, robotic vision, collision detection, complete navigation system for blind, industrial automation and many more.

The fundamental topic of the thesis work has been depth estimation from stereo image pair, which basically is about estimating the depth to different objects present in a scene of camera view. As the name directly suggests, depth estimation, is the method to estimate the exact distance between two components in the scene. Thus, the very naive way to compute this distance measure is to

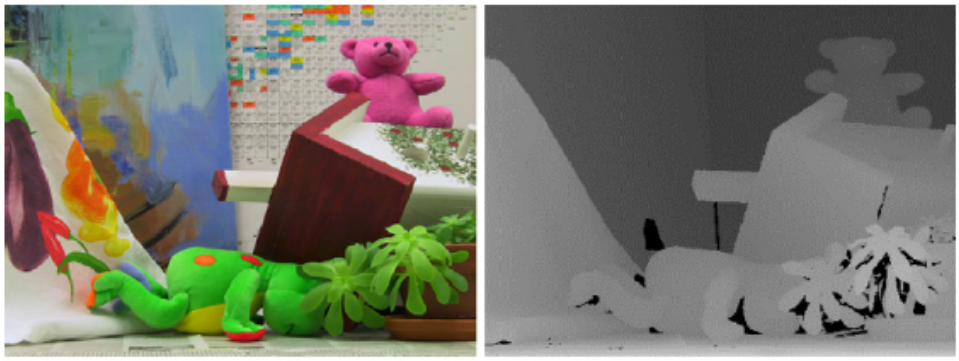


Figure 1.1: Tsukuba 2D camera image and depth map belonging to the image. Pixels go darker as the depth grows.

simply use a ruler and get the distance information, however, in the field of computer vision same thing can be done in a more intelligent way using stereo vision method. Commonly, a camera pair is used to take the images simultaneously and then compute the depth of the objects which lie apart from the cameras location by taking into consideration the minute differences between the two images of the same scene captured by the two different cameras individually. The main kernel function which is used in stereo vision is commonly known as stereo matching which calculates the differences on the pixel level lies within the two images and thereafter generates the depth map or disparity information for each pixel in the image. The figure 1.1 shows the well known standard stereo image Tsukuba and its corresponding depth map. As the depth from the camera increases in scene, it can be noted that pixels in the become more darker. The details of fundamentals behind the stereo vision and the process of estimating the depth information is briefed in the further sections of this chapter.

1.1 Fundamentals behind Stereo Vision

Since last three decades, Stereo vision has been into the computer vision prominent research areas. Therefore there are various concepts and terms which are assumed to be trivial in those research areas. Hereby this section provides a brief background.

1.1.1 Acquiring Depth Information

Stereo vision is one of a number of techniques used to discern the three dimensional information of a scene. Other techniques which can be used instead of, or complementary to stereo vision, include:

1. Active range finding methods which works by applying controlled energy beam to the objects in the scene and detecting the reflected energy. These techniques, although well suited to static scenes and underground situations, are too slow for real-time applications.

2. Structured lighting approaches which include striped, grid and patterned lighting. These tend to be useful in hard real time domains likewise industrial automation and robotics vision.

3. Monocular image based techniques which use information from a single image, and include depth from focus, shape from shading, depth from occlusion cues and depth from texture gradient. Depth and object orientation are typically inferred from statistical assumptions.

Nowadays, 2D image depth knowledge may be estimated in various ways. A method based on stereo matching is at present used to generate depth maps. Hence, the focus of the thesis work is on how to use stereo correspondences method for estimating depth from a image pair. The main principle of stereo matching is basically defined as searching matching pixels in a stereo pair images taken from two different cameras which lie at slightly horizontally separated locations. The technique is well explained in detail in the further section.

1.1.2 Stereoscopic Image Pair

A stereo image pair has two images of the same scene, known as left view image and the right view image and which are captured from slight horizontally separated location point. A scene is depicted in figure 1.2 where two pictures taken from two horizontally separated camera positions, giving left view image and right view image. The objects in the scene which lie close to the camera location will appear more towards the right in the left view image and more towards the left in the right view image. Faraway objects, such as the sun and the cloud, will be located at approximately the same position in both images. The effect of taking images likewise can be clearly seen in figure 1.2. Nearer objects are in the left view image are slightly towards the right while in the right view image slightly towards the left. Whereas, the objects which are far away will be depicted at almost the same position in both the images and that make up the stereo pair of the scene. The fundamental fact is that the horizontal displacement of objects in left and right

Figure 1.2: Stereoscopic images, left view image and right view image of the same scene.

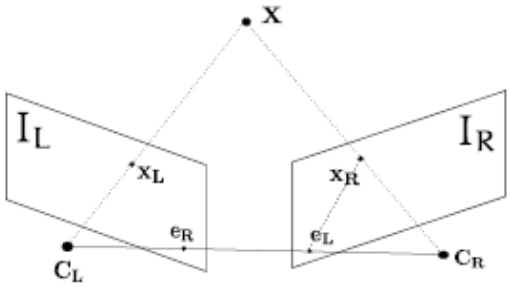


Figure 1.3: I_L and I_R are two views of the same scene.

view images directly depends upon the value of distance between the object and capturing camera point location. The main aim of stereo matching method is to search the corresponding matching pixels in left and right view images.

1.1.3 Stereo Case Analog to Human Eye

In humans, the stereo vision case works on the principle of binocular vision. The figure 1.3 describes the analogy to human way of perception of world. Similarly in the stereo case, let there be two pinhole cameras situated at C_L and C_R which acts like left viewing eye and right viewing eye respectively. Consider two views, I_L and I_R, of the same scene shown in the figure 1.3 .

Consider a point X in 3D world system, the image formed of point in left image plane as x_L and in right image plane be x_R. Here x_L and x_R are known as corresponding points. As observed, 3D space point X can be estimated as the

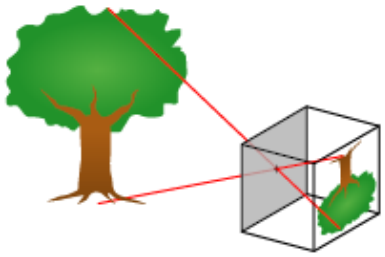


Figure 1.4: The camera pinhole model

intersection of the two lines i.e by joining x_L to the center point of camera C_L, of image plane view I_L and x_R to the center point of camera C_L, of image plane view I_R. This whole process is commonly known as triangulation. The image formed by the camera C_L in image plane IR is e_L and the image formed by the camera C_R in image plane I_L is e_R. The point e_L and e_R are known as epipolar points. They are also known as epipoles. Epipolar Line is the line joining camera center C_L and world system point X in image plane I_R. Any line which passes through the epipole is called as epipolar lines. The plane joining both the camera center C_R, C_L and point X is known as epipolar plane. Thus the core problem of finding the point x_R given the point x_L reduces to a one dimension search problem along some epipolar line. Thus finally the difference in the positions of these points is called disparity, d_p.

1.1.4 Pinhole Camera Model

There are several camera models, but the simplest of them is the pinhole model. If we consider a camera without any lens to bend the light, we end up with a small "hole" where the light is allowed to pass through, and is captured on a plane further back in the device, illustrated in figure 1.4. The plane at the back of the device where the image is projected, is known to be the focal plane or image plane. The point letting in the light is called the camera center, and can be seen as the point where we place the aperture, which lets light in. In practice when constructing cameras, it is natural to put the image plane in the back of the device, behind the camera center. For the purpose of the theoretical model, we can move the image plane in front of center of projection, and still pretend that the same rays of light hits the plane, as if it were in the back. The effect of this is that the image is not flipped horizontally, nor vertically, anymore.

A pinhole camera model is used vary widely in projection geometry to model the projection of 3D points in space to image points. It describes the relationship between the 3D point coordinates and projection point on the image plane of pin-

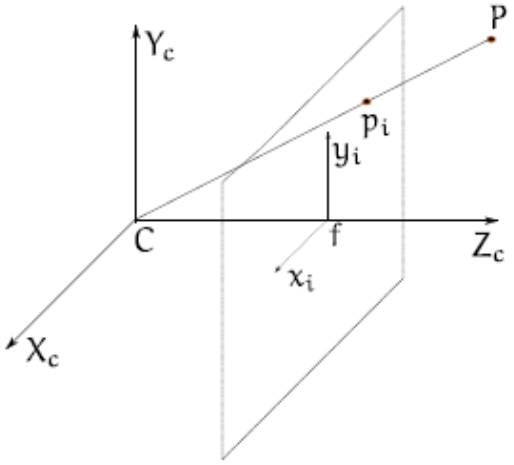


Figure 1.5: Pinhole camera model with camera center at C and projecting point P on the image plane as p_i

hole camera. The mathematics of the pinhole camera model is described with the help of the figure 1.5. The orthogonal 3D coordinate system with its origin at C. This is also where camera center point is situated. Three axes of coordinate system are labeled as X_c , Y_c and Z_c. Axis Z_c is the viewing direction and is called as the principal axis. The image plane where the 3D world is projected is parallel to X_c, Y_c. It is located at distance from the camera center C. Assume a point P in the world coordinate system and the point P_i representing the projection of point P on the image plane. The x_i and y_i axes represents 2D coordinate system in image plane and are parallel to X and Y, respectively. This demonstrates a very basic model for a camera. However, for ideal cases, it is the best robust model. The projection of point P from world coordinate system as point p_i in the image plane can be obtained with the help of standard transformation matrix which is also known a the camera matrix or projection matrix. Let P be written as $[x, y, z, 1]$ and p_i be written as $[i, j, 1]$ in homogeneous coordinate system. Thus the camera matrix can be written as:

$$T_p = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{f} & 0 \end{bmatrix}$$

$$\begin{bmatrix} i \\ j \end{bmatrix} = \frac{f}{z} \begin{bmatrix} x \\ y \end{bmatrix}$$

The above equations provide the relationship between point P and the projected point p_i where f stands for the camera focal length.

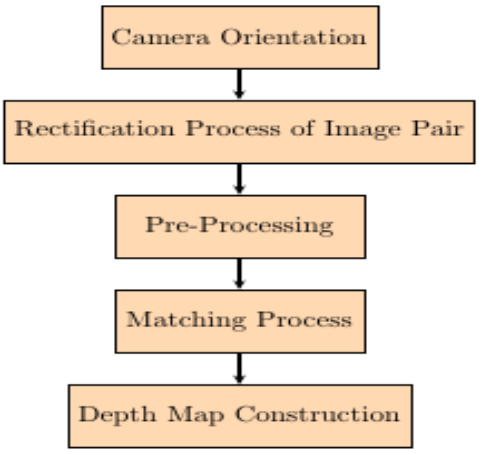


Figure 1.6: Generalized stereo vision paradigm

1.1.5 Stereo Algorithm Paradigm

The figure 1.6 shows the general steps of stereo vision paradigm. These steps involves:

1. Camera Orientation: Camera orientation basically comprised of two main components which are interior and exterior parameters. Interior orientation of camera refers to just the internal geometric configuration of camera. It mainly includes parameters like the focal length, principal point location, and characteristics of distortion of lens system. While the exterior orientation features includes the location of camera during the time image was captured. It comprises of the location of the optical center and also optical axis direction.

2. Rectification Process: Epipolar geometry helps a lot in improving the speed of matching methods just by applying constraints in the search for a corresponding match to linear search in one dimension. The process of rectification transforms the images in such a way that the epipolar lines become co-incident with horizontal scan lines, and this further reduces the complexity of implementation of matching process. The camera orientation information is needed while performing image rectification process.

3. Preprocessing: The rest of the preprocessing of the images happens prior matching process. To name a few examples of preprocessing technique which may be needed like transformation, noise reduction, hole filling.

4. Matching Process: It simply comprises of locating corresponding matching points in the stereo image pair. A depth map can be estimated using the disparity values obtained from with the help of matching process and camera orientation parameters.

5. Depth Map Construction: The constructed depth map is often known as 2D representation rather than a full 3D representation, since depth information can only be discerned for visible surfaces in the scene. No information can be extracted about the extent and shape of objects behind the visible surfaces.

1.1.6 Common Stereo Pair Dataset

Although there is varied collection of the dataset avaliable on the middlebury stereo homepage, few stereo pair are extensively used nowadays for benchmarking. The figure 1.7 shows the left view of the these stereo image pair commonly used and they are known as Teddy, Tsukuba, Cones, Venus.

1.2 Disparity Estimation

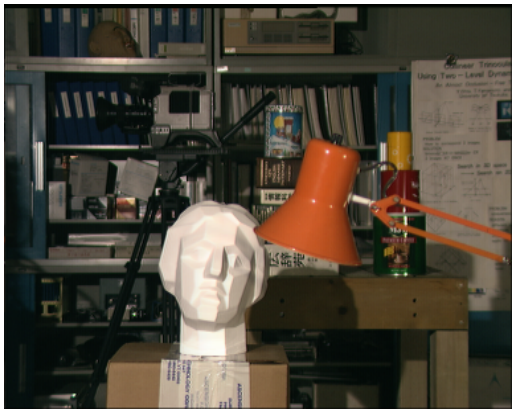
From the earliest findings into this topic, it is well known that humans perceive depth based on the differences in appearance between the left and right eye. Left image is bit shifted from right image and this shift is called disparity. Disparity gives us the depth from the camera and relates inversely proportional to the distance from the viewer's location.

1.2.1 Depth Map

Image comprises of pixels and is arranged as rows and columns making x, y as two cardinal axes of 2D image coordinate system. The figure 1.8 is an example of 10×5 image. The rows are called as scanlines. The columns which are along x axis and rows which are along y axis is known as width as well as height of the image respectively. So here 10 is the width and 5 is the height.

When a scene is captured by a camera, 3D world coordinates are projected onto 2D plane. With the loss of one dimension, the objects in the scene with coordinates (x_w, y_w, z_w) are projected as camera coordinates (x_c, y_c). The dimension of the depth map is same as the image dimension $x \times y$. Each element in the depth map contains the depth value of the corresponding pixel in the image of the scene. Image may also contain an extra coordinate representing the depth value additional to the color values. It is worth noting that the depth maps represents

(a) Teddy Image



(b) Tsukuba Image

(c) Cones Image

(d) Venus Image

Figure 1.7: Common stereo vision image pair datasets.

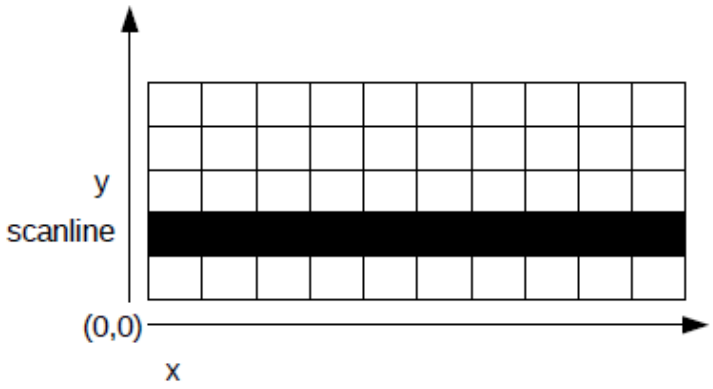


Figure 1.8: Raster digital image.

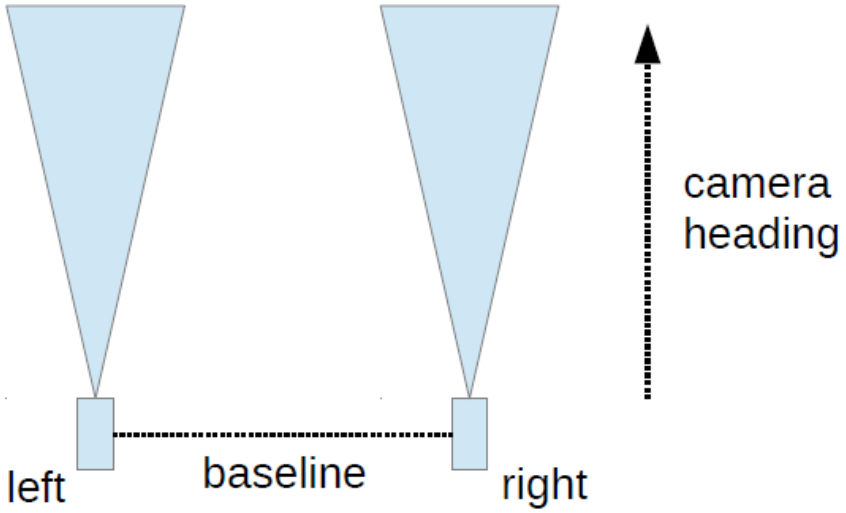


Figure 1.9: A stereo camera pair.

the distance between the point in the scene and the camera. Depth map is a gray scale image which comprises of the gray pixel value with a range of [0-255] . In this, the value 0 of the pixel denotes that 3D pixel is at the most distant place in the scene while the value 255 of gray pixel denotes that 3D pixel is situated at the most near place in the scene. Every pixel in the depth map defines the position of the pixel along Z-axis.

1.2.2 Disparity

Stereo vision is a concept of capturing two images of same scene with the help of two cameras situated at two different viewpoints and calculating depth of the pixels using this information. The distance between the centers of two images is called as baseline. It is approximated that distance between two cameras is same as baseline. It is to be noted that cameras must have same orientation and baseline must be orthogonal to camera locations. It is referred to the two cameras as the left and right camera, as illustrated in figure 1.9.

In other cases, need of scaling, rotation or other transformations may be required. This is done to get the horizontal epipolar lines. The underlying principle of the stereo matching is to search for the corresponding matching pixels in the stereo image pair. For every pixel in the left image, the search is conducted to find the corresponding pixel in the right image. The difference of distance of corresponding pixels is called as disparity. The search of the matching pixel is complex

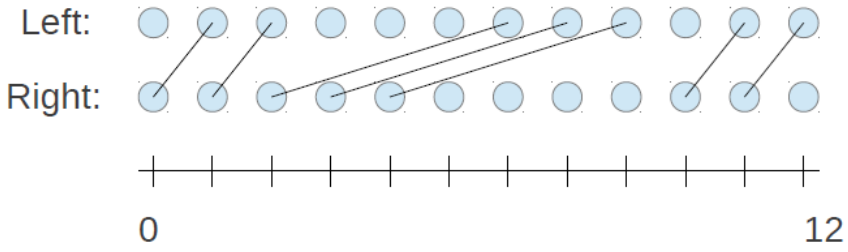


Figure 1.10: Disparity from left to right image with blue dots representing pixels and lines representing the corresponding match found.

and time consuming task. After image rectification, epipolar lines are horizontal and thus the matching pixel in the right image lies either on somewhere to the left of the pixel or on the pixel position. Thus the search is restricted to same scanline or towards the left. If the pixel coordinates in the left image is (x_1, y) and in the right image is (x_2, y) then relation between x_1, x_2 will $x_1 \geq x_2$ and disparity values is given as $\delta = x_1 - x_2$

In figure 1.10 a row of pixels of stereo image pair represented with blue dots is shown. As noted, the set $S = (1,0), (2,1), (6,2), (7,3), (8,4), (10,9), (11,10)$ of matching pixels can be formed. Their corresponding pixel difference or the disparity values can be represented by set $D = 1, 1, 4, 4, 4, 1, 1$. Thus, it is clear that scene consists of three objects and the middle one is closer to the camera compared to the side ones as the disparity value of middle one is 4 while other two object have disparity as 1.

1.2.3 Epipolar Lines

It is known so far that depth estimation algorithms based on stereo vision, need images of the same scene, taken by two cameras situated at varied positions. Further, the stereo correspondence algorithms desire that a feature point detected at pixel $p_1 = (x_1, y_1)$ in the first image, should be found at $p_2 = (x_2, y_2)$ in the second image, to limit the search space for matching pixels along the epipolar lines. In other words, the pixels in both images can vary along the x-axis, but they must be on the same image scanline. Achieving this effect by perfect real-world camera alignment is extremely difficult, and software correction is used instead. To do this we need to make use of epipolar geometry, which is the intrinsic projective geometry between two views. What kind of objects are in the scene, and how far away the objects are, does not change the epipolar geometry, it is only dependent on the intrinsics, and extrinsics of the cameras.

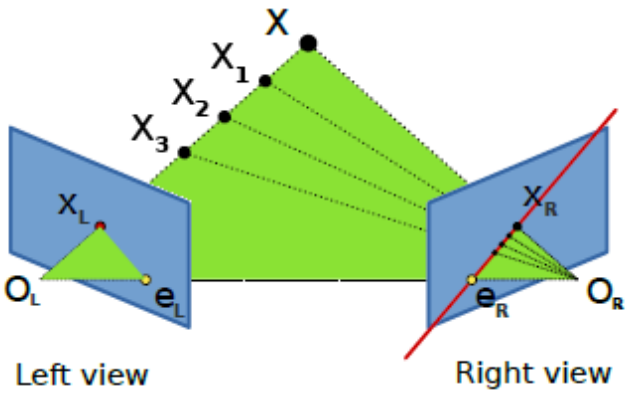


Figure 1.11: The epipolar lines $e_R - x_R$ and $e_L - x_L$.

Figure 1.11, shows a point X observed by two cameras, with center of projection O_L and O_R. X_L is the point where the ray from X hits the left image plane, and similarly with X_R for the right view. Notice that the points X, O_L and O_R together defines a plane, which is called the epipolar plane. The intersection of the image planes and the epipolar plane describes a line in the two image planes, which is called the epipolar lines. Thus, the epipolar lines change if any of our three points defining the epiplane change. Points that lie in the plane will be visible in the epipolar lines. Though, the epipolar lines do not need to be parallel to any of the cardinal axes in the image planes. With the help of binocular stereo vision concepts, the position of a point lying in space can be estimated by just finding the point of intersection of two lines that passes through the center of projection and the projection of point in image.

1.2.4 Image Rectification

The objective behind rectification of images is to transform the epipolar lines of two camera images in such a way that they become aligned horizontally. It is necessary to first perform image rectification so as to significantly lesson the complexity of pixel matching. This can be obtained by using linear transformations which involves rotation, translation and skewing the camera images. In these types of transformations, intrinsic parameters of camera along with the information about mutual camera positioning and orientation of camera are also required. This concept is explained with the help of the figure 1.12 in which camera images before and after rectification is shown. Before rectification pixel matching approach was 2D search problem. After rectification, pixel matching approach can be done through one dimensional linear search.

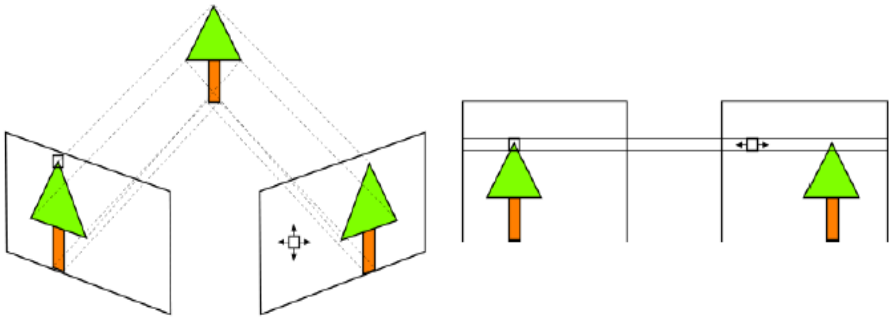


Figure 1.12: Camera images before and after rectification.

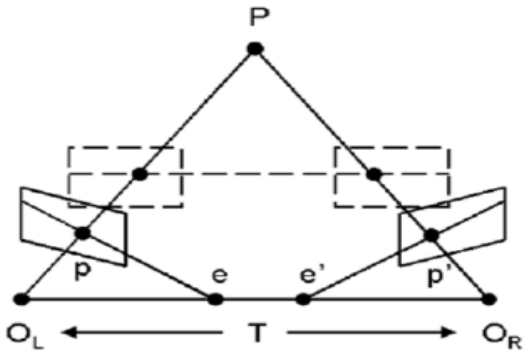


Figure 1.13: Original camera image planes represented with solid line borders while rectified image planes represented with dashed line borders.

The principle behind rectification of image is well illustrated with the help of figure 1.13 as the original camera image planes represented with solid line borders and the rectified image planes represented with dashed line borders. Let us consider the point P in the world coordinate system and its projection onto the both image plane be p and p' on left and right image plane respectively. Epipolar lines which are drawn along the points p and p' intersect the baseline T at e and e' which are known as epipoles.

1.2.5 Relation between Disparity and Depth

The output obtained as a result of stereo matching algorithm technique is disparity values of the pixels not the depth. The relation between the disparity and the depth is that disparity value of the pixel is inversely proportional to its depth from camera. Hence to obtain the depth value of the pixel from its corresponding disparity is done using triangulation.

With the help of a stereo image pair which is rectified, triangulation approach can be used to calculate the depth. The principle is illustrated in figure 1.14 for an arbitrarily located 3D point P. Hence, as observed all of following relations are independent of y ,hold:

$$d = x_L - x_R = f\left(\frac{x_p + l}{Z_p} - \frac{x_p - l}{Z_p}\right) = \frac{2fl}{Z_p} \tag{1.1}$$

$$Z_p = \frac{2fl}{d} \tag{1.2}$$

Here, in figure 1.14 the disparity value is calculated with the difference of $d = x_L - x_R$, where x_L represents the x-coordinate of the projection x_p upon the left camera image plane Im_L and x_R represents x-coordinate of the projection upon the right image plane Im_R. This directly means that once disparity value is estimated, depth map can be generated, given few camera parameters such as focal length f and baseline distance as $T = 2l$.

1.2.6 Ground Truth Disparity Map

In figure 1.15, disparity map has been displayed which shows true disparities for corresponding pixels of the image. The disparity map with ”true” values of disparity is known as ground truth. It acts as a base in evaluation of different approaches used for calculate disparity values of the scene.

There are many ways of calculating the ground truth disparity maps. One of the straight-forward method for creating the ground truth disparity map is with the help of range cameras. Other method based on the piecewise planar technique. It

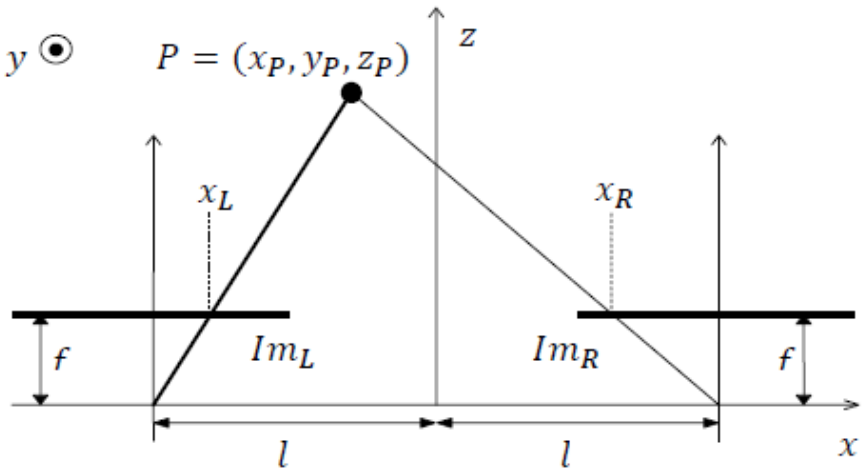


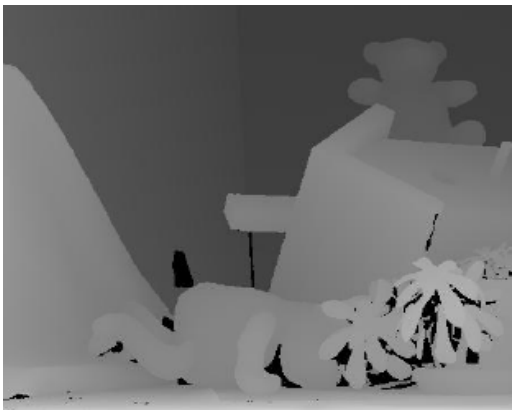
Figure 1.14: Disparity calculation phenomenon.

is calculated by decomposing scenes in piecewise planar surfaces only and these are labeled manually. Thus, the corresponding points in the surfaces are recognized in the stereo images of high resolution and disparity is modelled with plane fitting.

This ultimately produces ground truth disparity maps with for downsampled images. In order to produce better challenging stereo image pair along with their ground truth data, researcher in the field used a method which is based on structured light. In the figure 1.15, ground truth disparity maps belonging to the image pairs shown in section 1.1.6 in figure 1.7 are displayed (note that the ground truth disparity maps are scaled to use a larger range of the available intensity interval).

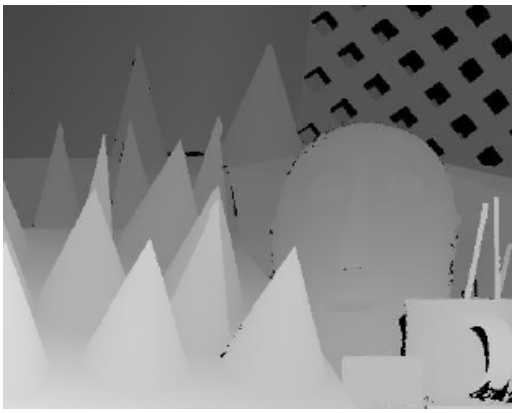
1.2.7 Applications

Depth maps have several applications like vision for robots, complete navigation system for blind, tracking objects and a lot more as discussed below. One of the application of depth maps is collision detection. There are many collision detection systems for cars on the marked already, mostly based on active systems such as laser or radar. By using stereo vision, it is possible to track multiple objects at the same time, and more importantly to determine where they are located and how much space the objects occupy. A pair of cameras could be mounted in front or at the back of a car, and the depth maps could be used to determine when an object is closer than a threshold, in which case it could alarm the driver. Depth maps can also be used in robotics, where a robot could have a stereo camera pair integrated to function as eyes. Besides capturing visual information from the



(a) Teddy ground truth, created from a technique based on structured light

(b) Tsukuba ground truth, consisting of fronto-parallel planar surfaces



(c) Cones ground truth data, created from a technique based on structured light

(d) Venus ground truth, consisting of planar surfaces

Figure 1.15: Ground truth data with reference to datasets mentioned in section 1.1.6

scene, the robot can use the camera pair to generate depth maps, and use the depth-maps to navigate in its environment. The information can be used for avoiding crashing into objects, simply determine which object is currently the closest and prefer an interaction with this object.

Few more examples of such applications include autonomous vehicles and robotic devices and toys. More specific applications include LHD (load,haul,dump) vehicle navigation, automated roof bolting, and autonomous machines for locating and breaking rocks so that they fit through a grid.

So as to conclude, the application of estimating the depth map using stereo matching are not limited to one area and can be widely used for many industrial purposes.

1.3 Stereo Matching Constraints

Apart from the epipolar constraint used in image rectification, there are several other constraints that may be exploited in stereo correspondence. The constraints are sometimes implemented and used explicitly, but are normally used implicitly. Some of the most commonly referred constraints will be brought up here.

1.3.1 Pixel Intensity Similarity

The similarity constraint is very fundamental and states that pixel intensities belonging to the same point of an object in the left and right images should be almost equal for corresponding pixels. Thus, their color and intensity value must not vary significantly between different viewpoints.

1.3.2 Uniqueness

The uniqueness constraint is also essential and means that for each pixel in left input image there must be only one matching corresponding pixel in the right image. Even though this may seem natural, it will be seen further that this constraint is not trivial as it does not hold at all times.

1.3.3 Smoothness and Continuity

A constraint that it referred to very often in modern stereo correspondence research is the smoothness constraint, sometimes also called the continuity constraint. This constraint postulates that the depth of physical surfaces in the real world is a smoothly varying property, except for at object boundaries where discontinuities may be present.

Figure 1.16: Occlusion in the Tsukuba image pair.

1.4 Fundamental Problems in Stereo Correspondence

There are a lot of challenges when solving stereo correspondence problems and some of the most important problems encountered in stereo matching will be described here.

1.4.1 Occlusion

One problem which needs to be dealt with is occlusion. Occlusion occurs when a point in the 3D space in front of the cameras gets depicted in one of the images and is blocked from being depicted in the other one. Refer figure 1.16 for an illustration of the problem. Here, yellow areas in the left image are occluded from the right camera and areas of the same color in the right image are occluded from the left camera.

For the observant reader, it may be obvious that the earlier mentioned uniqueness constraint does not hold for occluded areas. This is used in stereo algorithms, and will be described in section 4.3.4 .

There is no general solution to the occlusion problem and as occlusion may lead to spurious matches, one goal in stereo correspondence research is to somehow minimize the impact of occlusion on the final result.

In figure 1.17 the points P_v are visible to both cameras and disparity estimation is therefore possible for these points. However, the points P_o are occluded for the right camera, and no matching pixels will be found.

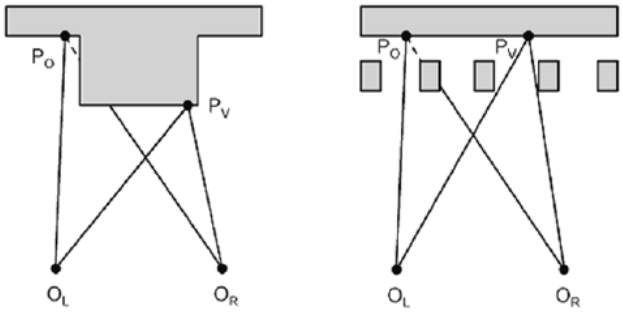


Figure 1.17: Occlusion phenomenon.

1.4.2 Uniform Texture and Lack of Texture

Another challenge is to handle surfaces with uniform texture and surfaces that lack texture. When surfaces with these properties are encountered in a stereo image pair, it immediately becomes complicated to decide which pixels in the left image corresponds to which pixel in right image. This, in turn, leads to an elevated risk of spurious matches – with resulting mismatch errors in the calculated disparity map.

1.4.3 Sensor Noise and Bias

Camera sensor noise is yet another issue for stereo matching in that two matching pixels that should have the same intensities may not. This problem is especially apparent in poorly textured image regions or in images taken under poor lighting conditions, thus having low signal-to-noise ratio. If two different cameras are used to capture the stereo images there might also be a difference in gain or a bias in the intensity values of the matching pixels in left and right image. Depending on which matching technique being used, the images may have to be preprocessed to correct this issue in order to prevent mismatching.

1.5 Thesis Organization

The further thesis document is organized in the following order. The chapter 2 focuses on to theoretical information about the techniques that are commonly counted under modern stereo correspondence algorithms category. A survey of modern stereo correspondence algorithms is presented in this chapter.

In chapter 3, the detailed description of stereo correspondence problem is mentioned. The challenges that come along in stereo correspondence are briefed. The

survey conducted in chapter 2 has served to influence the contents of chapter 4. Detailed description of stereo algorithm and proposed work is defined in this chapter.

An implementation description of a proposed stereo algorithm along with the generated results follows in chapter 5 and lastly the conclusion, contribution and future scope are presented in chapter 6.

Chapter 2

Literature Review

This chapter says about the gradual evolution of the stereo matching techniques. The description is divided in two subsections. First section briefs about the database collection of stereo vision history and evaluation process of the all existing and recently formed stereo matching algorithm. Second section is specific about the broad classification of stereo algorithms identified during the literature survey. There has been a lot of research done in area of computer vision techniques, especially in the field of stereo vision. A detailed overview of prominent methods and their comparison with respect to performance can be obtained from surveys [1].

2.1 Online Repository and Evaluation Process Review

A systematic taxonomy [2] provides insight roadmap of evolution of dense stereo correspondence algorithms. This gives a knowledge pool of current stereo matching methods and their essential components. An up-to-date homepage [3] which displays a comparison table on basis of performance among various proposed algorithms comes along this taxonomy. Till today, experimental results from over 80 stereo algorithms are been published on this page. There also provided links to research papers presenting these algorithms. The homepage [3] also contains collection of several stereo image pairs dataset [4] together with ground truth dataset. This is a wide platform which gives way to test and evaluate algorithm with the already existing ones. After done with the experimentation part there is a way to submit the final results in order to take part of this performance comparison table as shown in figure 2.1.

Algorithm	Avg. Rank	Tsukuba ground truth nonocc	Tsukuba ground truth all	Tsukuba ground truth disc	Venus ground truth nonocc	Venus ground truth all	Venus ground truth disc	Teddy ground truth nonocc	Teddy ground truth all	Teddy ground truth disc	Cones ground truth nonocc	Cones ground truth all	Cones ground truth disc	Average percent of bad pixels (explanation)
AdaptingBP [17]	4.5	1.11 9	1.37 5	5.79 11	**0.10** 1	0.21 3	1.44 2	4.22 3	7.06 4	11.8 4	2.48 2	7.92 7	7.32 3	4.23
CoopRegion [41]	4.5	**0.87** 1	**1.16** 1	**4.61** 1	0.11 2	0.21 2	1.54 4	5.16 10	8.31 7	13.0 7	2.79 6	7.18 4	8.01 9	4.41
DoubleBP [35]	5.9	0.88 3	1.29 2	4.76 3	0.13 5	0.45 12	1.87 8	3.53 2	8.30 6	**9.63** 1	2.90 7	8.78 16	7.79 6	4.19
OutlierConf [42]	7.0	0.88 2	1.43 7	4.74 2	0.18 11	0.26 7	2.40 14	5.01 6	9.12 10	12.8 6	2.78 5	8.57 12	6.99 2	4.60
SubPixDoubleBP [30]	9.2	1.24 16	1.76 16	5.98 12	0.12 4	0.46 13	1.74 7	**3.45** 1	8.38 8	10.0 2	2.93 9	8.73 15	7.91 8	4.39
WarpMat [55]	11.3	1.16 10	1.35 4	6.04 13	0.18 12	0.24 5	2.44 15	5.02 7	9.30 11	13.0 9	3.49 17	8.47 11	9.01 22	4.98
Undr+OvrSeg [48]	14.8	1.89 37	2.22 33	7.22 29	0.11 3	0.22 4	**1.34** 1	6.51 16	9.98 12	16.4 19	2.92 8	8.00 8	7.90 7	5.39
GC+SegmBorder [57]	15.6	1.47 28	1.82 18	7.86 34	0.19 13	0.31 8	2.44 15	4.25 4	**5.55** 1	10.9 3	4.99 45	**5.78** 1	8.66 17	4.52
AdaptOvrSegBP [33]	16.7	1.69 31	2.04 28	5.64 9	0.14 6	**0.20** 1	1.47 3	7.04 27	11.1 15	16.4 21	3.60 21	8.96 19	8.84 19	5.59
GeoSup [64]	17.8	1.45 27	1.83 20	7.71 33	0.14 7	0.26 6	1.90 9	6.88 24	13.2 29	16.1 16	2.94 10	8.89 18	8.32 14	5.80

Figure 2.1: The top ten dense stereo matching algorithms listed in the evaluation table on the Middlebury Stereo homepage.

2.2 Modern Stereo Vision Ideas and Review

The purpose this section is to provide an insight of few modern existing stereo matching approaches and techniques. The research taxonomy on the stereo correspondence proposed by Scharstein and Szeliski [2] broadly categorizes these algorithms in two classes as local methods and global methods. In past few years, there has been a lot of improvement done and various algorithms are implemented which can be briefly described below:

2.3 Local Methods

Local methods[5; 6; 7; 8; 9; 10; 11] are widely referred as window based methods, where the estimation of disparity at a some given point just depends on their intensity values bounded within a definite size window. In local stereo matching algorithm, firstly matching cost is calculated locally within a fixed sized window and then disparity refinement is achieved. Limited size window is used to preserve rich texture and edge information. The drawback of small size window is it produces noisy disparity map. Whereas large size window produces smoothness in the disparity map but edge information is not preserved. They work by considering implicit smoothness assumptions by simply aggregating support[12; 13; 5].

Table 2.1: Brief overview of all local algorithm techniques.

Methods	***Advantages***	***Limitations***	***References***
SAD, SSD, NCC	Fast speed, high efficiency	At the end of the treatment effect is not good texture area	[14]
TSAD, TSSD	Reduces noise, making the cost of matching more robust	Low reliable	[9] [10]
Combined stereo matching	Match combine several images to enhance the robustness of low texture regions to solve illumination intensity sensitive issue	Time consuming	[13]
Bilateral filter	Simple and widely used	Slow, can not meet the real-time requirements	[15] [16]
AD	Calculation method and layered window is linear to improve the computing speed	Accuracy is not high at the end of the texture region	[17]
Adaptive Weight	Better deal with similar color without interfering in a layer between the pixel depth	Calculation is very slow	[18; 19; 9]

2.3.1 Local Edge preserving Smoothing Filter

Preserving edges simply implies ability to detect the discontinuity of the object in the scene and to clearly represent the boundary in the depth map. With the help of edge information preserving and smoothing method, the details of the input image will be smoothed while preserving the edges. All the edge information preserving approaches can be broadly divided into two categories. One of them is local filter based algorithms such as median filter, tree filter [20], bilateral filter [15] and its iterative version [16] , and other is Guided Image Filter (GIF) [21; 22] and Weighted Guided Image Filter(WGIF) [23] which are reviewed in section 2.4.1.

Bilateral filtering [15] handles images just by combining the range filter concept with domain filter concept to preserve edge information. Although, it is widely used weighted average filter due to its simplicity, but still suffer from gradient reversal artifacts near edges.

2.3.2 Matching Cost Function

Matching cost is estimated based on similarity measures. To name the earliest of them are absolute difference value metric (AD)[5] and the square difference [6] measure (SD). They have high computation speed, better efficiency however they have less immunity to noise and are sensitive to light intensity resulting in noisy disparity map. In order to reduce this noise, common approaches based on local window are sum of square difference (SSD) and sum of absolute difference (SAD) [24], but edge information is still not preserved in this. Normalized cross correlation (NCC) [14] is more robust and has same complexity as SSD and SAD. The mathematical details are discussed in chapter 4.

2.3.3 Adaptive Weight and Modified

Recently, algorithm follow cost aggregation method by setting weights of pixels within aggregation windows based on their color intensity similarity and their distance to the center pixel as proposed in [18][23]. The main principle behind the adaptive support weight [19] is to calculate weights individually to every pixel lying within the fixed window. Although it brings smoothness, but complexity is increased and results in edge fattening effect. In order accelerate, various filters are introduced like bilateral filter [15; 16], trilateral filter [25], tree filter [20] and guided image filter [21]. This helps in preserving the edge distinctiveness.

2.4 Global Methods

The optimization objective function is framed as combination of a fidelity data term and a regularization smoothness term. The detailed explanation about both is given in section 4.2.2. With variations of fidelity terms or variations of regularization terms, various methods are proposed and thus results are established. The number of iterations are used to solve all these problems and this makes global optimization based methods very time consuming.

Traditional global approaches are based on the concept of regularization [26; 27] and Markov random fields (MRF)[28; 29; 30; 31]. Recently, various techniques are introduced based on global optimization, which works by minimizing a global energy function while generating the disparity map. Global algorithms are embedded with optimization technique just after the phase of disparity computation thereafter skipping the aggregation step which is compensated by the global smoothness constraints. After the global energy has been obtained, various methods can be used to obtain minimized energy framework as done in [32] [33]. Widespread global stereo correspondence methods which include dynamic programming [34] , belief propagation [35], block matching [17] and graph cuts approach [36]. Dynamic programming aproach is basically used [13] reduce the occlusion. Although, improvements are also made with the help of connected and undirected graph and its constraints are obtained by the relevance among pixels in graph. however, belief propagation method which simply works with help of passing messages around the image is used in [35][29], and by minimization of the matching cost function optimization can be achieved [12].

2.4.1 Global Edge Preserving Filters

With refrence to section, 2.3.1, to overcome the drwbacks of existing filters, Guided Image Filter (GIF) [22] was introduced. The idea behind this is implementing a linear transform for representing the pixel value in a fixed size window. It is different from other approaches as, GIF computes the resultant image by considering the configuration of a reference image into consideration. Nevertheless, this filter cannot handle the image well when comes of edge information and also there are some halos (gaps) in the disparity map. These halos lower the visual quality of disparity map.

In [37], a extension of GIF, weighted guided image filter (WGIF) was introduced to reduce halo artifacts. Also a factor which handles edge information was proposed. This reduces the halo artifacts but still, it fails preserve edge information well in few cases as it implements process of image filtering and edge-preserving combined.

Table 2.2: Brief overview of all global algorithm techniques.

Methods	***Advantages***	***Limitations***	***References***
GIF	Cutting edge treatment is better	Slower operation	[22][38][21]
Dynamic programming	Fast and can reach almost near real-time	Lower accuracy than other global optimization algorithms	[34]
Belief propagation	Fast	Iterations more difficult convergence algorithm in poor	[39][35][36]
GCP	Non-textured area boundaries and occlusion effect have some improvement	Non-autonomous	[40]

2.5 Comparison in Local and Global Methods

In comparison to local algorithms, global methods usually posses better robustness and can easily estimate much more reliable disparity maps. Due to the fact that global method for generating disparity maps exhaustively takes other pixel value into consideration which results in optimal results. Consequently, they effect the smoothness in rather adverse way. Global algorithms makes disparity smooth everywhere and this results to poor results. Edge quality and the occlusion pints are extremely affected thereby effecting the final disparity maps. At present, there are few algorithms which are edges and occlusion specific. Till now, there are only few algorithms which are specific to edge information awareness or occlussion problem[41].

Chapter 3

Problem Statement

The chapter states the goal, objective and motivation behind the thesis work. Later sections contains the problem statement and gaps in the literature.

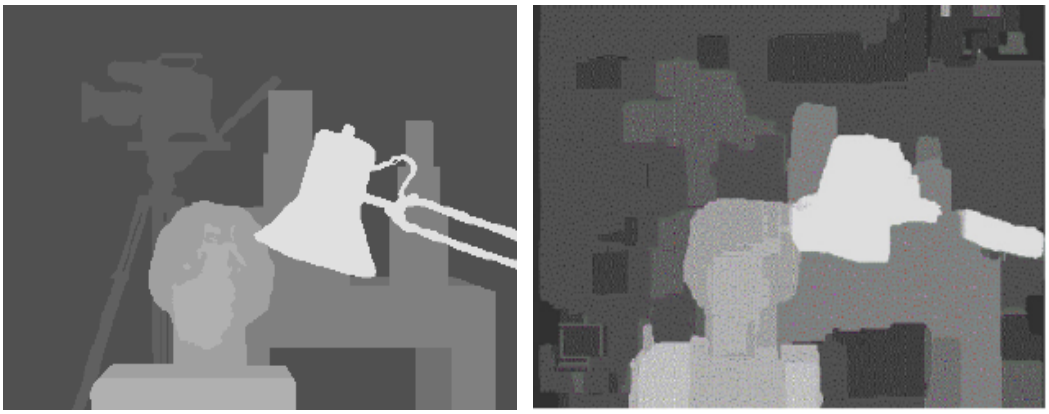
3.1 Goal and Objective of Stereo Vision

The thesis work goal stands for construction of the depth map of a scene around using two stereo images which are captured at the same time, each of them acquired from a different angle in space around. These images are simply be captured either using multiple cameras or by just one moving camera around the around. The two camera system is commonly known as binocular vision.

The main objective of this thesis work has always been to look closer both vertically and horizontally in the topic and to investigate existing stereo matching algorithms in order to get a direction within the field. Moreover, algorithmic components which are found interesting have been considered for experimental trials. Various stereo correspondences algorithms has also been looked into and researched to some extent.

3.2 Motivation

The application domains like automatic navigation, robot vision or 3D scene reconstruction where multiple views are available anyways and demand depth-maps motivated the research. Moreover, such pipelines demand a more reliable performance, which acted as a motivation for exploring the parallelization. The declining costs of stereo imaging has pushed the urge of better algorithmic way to find the depth map and hence motivated to explore advantages it provides. There are several depth map estimation algorithms that perform well. But most approaches



(a) Tsukuba standard disparity map (b) Tsukubar disparity map with halos

Figure 3.1: Halos in disparity map

fail to provide the clear depth map and suffer from artifacts and halos. Such limitations motivated the course of this work to develop a complete solution. Hence a strong motivation compelled to develop a useful extension to the existing depth estimation approaches, in particular to gracefully handle halo, computations and better edge preserving.

3.3 Problem Statement

Estimating dense depth maps from multiple views is a well known problem area in image processing. Several methods with different trade-offs in computational complexity and estimation quality exist. Most methods require rectified stereo images as input to efficiently estimate the disparity map between two views. More than two views are rarely used since the homography constraint between stereo images is considered sufficient.

Problems that is being tackled by this work is removal of halos to improve the visual quality and obtaining resulting depth map which represents the image well near the edges. As illustrated in figure 3.1, halos can be easily seen around the lamp and other areas. Robust depth map estimation can be performed as a solution to a global dense correspondence problem. Such an problem in general is time consuming. Therefore this thesis explores the problem of speeding up of such a process.

Three important criteria for assessing stereo matching algorithms are speed, reliability and accuracy. In this context, reliability refers to the ability of matching algorithms to return the correct match, and not some erroneous match. Accuracy, on the other hand, refers to the precision to which depth information can be

computed. For automation applications, speed and reliability are of foremost importance. Accuracy is of secondary importance for this application, as it is more important to discern reliable information about the presence and overall extent of objects, rather than exact depth information. Furthermore, for this application, a dense depth map is required. The thesis focuses in particular on analyzing the reliability of matching algorithms suitable for the application domain.

3.4 Gaps Analysis

The research questions that guided the thesis work are summarized here:

1. Can the traditional edge preserving approach be replaced with an approach using Gradient Guided Image Filter for better edge detection?

2. Can the halos be minimized to better present the depth map?

3. Can average percent of bad pixels be reduced?

Chapter 4

Proposed Work

The process of estimating 3D depth model out of two or more images by searching for corresponding matching pixels in the images and thereafter estimating a 3D depth for per pixel using their 2D positions is known as Stereo Matching. This chapter addresses the question that how one can build a depth map and to highlight common fundamental concepts of stereo correspondence algorithm of current scenario.

4.1 Types of Correspondence

This section states brief discussion about constructing sparse and dense depth map correspondence that attempts to allot depths measures to each pixel value in the input reference image.

4.1.1 Sparse Correspondence

Earlier stereo matching methods used feature based technique. To elaborate the process, firstly extraction of set of potential matching image pixel locations is done, using techniques like edge detectors or interest operators, and then thereafter search for corresponding matching pixel locations is carried out in other image of pair using patch-based metric technique. This became limitation to carry out sparse correspondence as due to lack of availability of computational resources, also due to less satisfying results produced. In few applications, it was required to match scenes which have very different illumination potentially and edges are most suitable features. Later these 3D sparse reconstructions can be interpolated with surface fitting methods.

4.1.2 Dense Correspondence

Sparse matching methods are now occasionally used in real time, as of now most of the stereo matching methods nowadays focuses on the dense correspondence technique, as it is more demanding for applications like image rendering or 3D modeling. However, the problem becomes more challenging as compared to sparse correspondence methods, because inferring out depth values from texture less areas needs quite certain amount of detailed technical work. Based on the detailed observation it can be concluded that stereo matching algorithms usually performs following steps in common:

1. Matching cost computation
2. Cost aggregation;
3. Disparity computation and optimization; and
4. Disparity refinement.

4.2 Classification of Dense Correspondence

In stereo matching research usually algorithms are broadly classified into two categories local methods or global methods. Although, there exist few methods which lie in between these two classes. The classification done on the basis of approach used while assigning disparities as explained further in detail.

4.2.1 Local Methods

The local algorithms [6; 7; 8] which are window based employs the disparity estimation for a point and depends only on the neighboring pixels intensity values which lie within that finite square window, generally these methods makes implicit smoothness simply by aggregating the cost. Local methods [5; 9; 10], as the name suggests, only use the information which is located in a near neighborhood around the pixels which are compared. The general approach for local method is to assign the disparity value that will minimize the cost function for each pixel coordinate individually for guidance image. This process is termed as Winner-take-all (WTA) technique of optimization. A simple illustration of the method is given in figure 4.1. Consider, in Tsukuba image pair the pixel on the left edge of the red lamp. In this example, a 9×9 fixed size square window centered around the compared pixels with the cost function as C_{SAD} and the range of disparity search as $0 - 20$ pixels. There is rise of the cost when the window enters the white

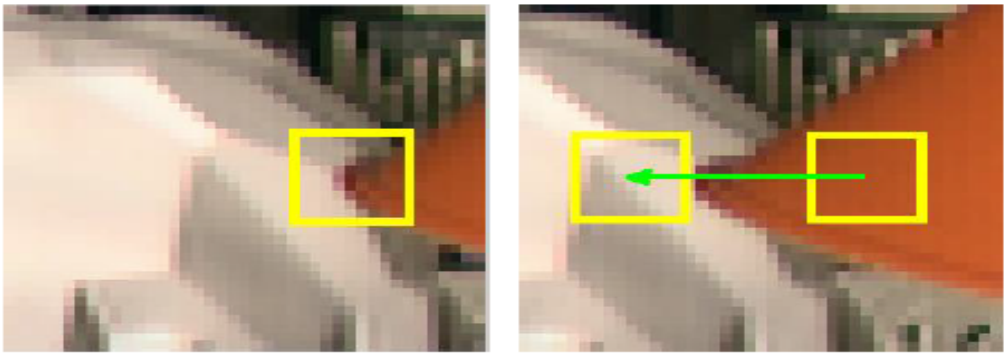


Figure 4.1: Local matching in the Tsukuba image pair where disparities are calculated for pixels along the green line.

area of the sculpture behind the lamp. Few algorithms may cleanly decompose into the following three steps as follows:

1. Matching cost: It is estimated by squared difference of the pixel intensity values with known disparity.

2. Cost aggregation: It is known as summing up the total matching cost within that fixed size square window with constant disparity.

3. Winner-Take-All technique: It is used for selecting the minimal or winning aggregated value of disparity at each pixel.

The advantages of using local methods is that they run fast and are not complex, thus are perfect for real time based applications. Whereas the pitfalls are poor occlusion handling and having high noise sensitivity.

4.2.2 Global Methods

In many applications, it is been proved that it is better to completely formulate the desired goals of the transformation with the help of some optimization criterion first and thereafter infering the best matching solution according to this criterion. Regularization methods [27] formulates a global energy function which encapsulates the all the expected characteristics solution and thereafter finding a optimum energy solution with minimum value. However, it must be noted thhat this energy term is comes from statistical mechanics field, in which the labeling problems are usually solved to obtain the spin of the particles [42] .

Regularization

The regularization concept [26] was first introduced by statisticians while fitting models to data which is heavily constrained the solution space. Consider an example of finding a smooth surface that approximately passes through the data points set. Such a problem is known as ill posed as there can be many different possible surfaces can be there which fit this data. As to quantify, a norm on the solution space must be defined which is also known as smoothness term. Along with the smoothness term, another term is also required i.e. data term. To obtain a total global energy function that can be optimized, the two energy terms are simply added together. Using this in stereo matching, the energy function comprises of two values data correspondence term and smoothness term:

$$E(d) = E_{data}(d) + \lambda . E_{smooth}(d) \tag{4.1}$$

In the above equation, d stands for $d(x, y)$ and it means it is the disparity map for guidance image whereas for adjusting the smoothness λ parameter is used. The term $E_{data}(d)$ usually stands for the sum of all costs for a given disparity map, based on cost of matching.

$$E_{data}(d) = \sum_{(x,y) \in Im} C(x, y, d(x, y)) \tag{4.2}$$

When global optimization technique is used cost aggregation is mostly skipped. The smoothness term $\lambda . E_{smooth}(d)$ depends on differences between the disparity value and intensity value of neighborhood pixels, as stated below:

$$\begin{aligned} E_{smooth}(d) = \sum_{(x,y) \in Im} & \rho(d(x, y) - d(x + 1, y), I(x, y) - I(x + 1, y)) \\ & + \rho(d(x, y) - d(x, y + 1), I(x, y) - I(x, y + 1)) \\ & = \sum_{(i,j) \in N} V_{ij}(d_i, d_j) \end{aligned} \tag{4.3}$$

where N defines the set of all four-connecting neighborhood pixels and one neighborhood pixel pair is included only once. On the basis of observations, it is assumed that depth discontinuities generally coincides with color edges or intensity edges in data of real world, the basic idea is to monotonically increase ρ with difference in disparity to compensate a discontinuous result with penalty, and be able to reduce this value of penalty for disparity discontinuity situated at the edges, at the same time. Such a behavior can be achieved by making ρ a product of two functions which can be written as:

$$\rho(d,i) = \rho_d(d).\rho_i(i) \tag{4.4}$$

In the above equation, ρ_d increases with the value of $|d|$ and decreases with the value of $|i|$. The function value ρ_d is usually based on the truncated difference magnitude of linear disparity stated as:

$$\rho_d(d) = min(|d|, \alpha) \tag{4.5}$$

The reason behind truncating this penalty is to avoid over smoothness in the result, especially around the boundaries of the object where the presence of large disparity jumps is the most. Avoiding truncation of the penalty value in such areas may lead to various smaller jumps. The value of parameter α is generally fixed to some of disparity search range. The value of intensity function ρ_i is typically based on image intensities of the near-by adjacent pixels. The function may looks as below:

$$\rho_i(i) = \begin{cases} c, & \text{when } |i| < T \\ 1, & otherwise \end{cases} \tag{4.6}$$

where c is a constant term having value of less than and is also a threshold with which it is determined what to be considered a intensity edge in the guidance image. As expected, the penalty value for the disparity jumps will thus be reduced at intensity edges of the image.

Although computational complexity for global methods is much higher than local methods, still benefits lies in better occlusion handling, uniform texture of the object and adjustments of smoothness of the result. Apart from these global methods are much more tolerant to the noise than local methods.

4.3 Stereo Matching Algorithm

Broadly speaking, proposed thesis work comprises of, a global dense stereo matching algorithm based on improved guided image filter i.e. gradient guided image filter [43] and Markov random fields (MRF)[44]. Edges information in the disparity maps are more accurate and it is due to high uniqueness of the gradient guided image filter. Thereafter, smoothness of the disparity map is maintained with the help of image segmentation technique known as mean shift [45]. It search for the sections having same color, and then smoothing is done. These techniques are explained in the section below:

4.3.1 Preserving Edges based on Guided Filter

Guided Image Filter (GIF) [22] is used to preserve the edges in the stereo matching. In GIF, there are two image one G is the guidance image and X is the filtered

image. Both may be identical. There is fixed sized window $\omega_{r'}(p)$ with radius r' and centered around pixel p. Linear transformation is done on image G in square window $\omega_{r'}(p')$ to generate Z as output image.

$$Z(p) = a_{p'}G(p) + b_{p'}, \forall p \in \omega_{r'}(p'), \tag{4.7}$$

Here $a_{p'}$ and $b_{p'}$ are the fixed value constants in square window. Optimum value of $a_{p'}$ and $b_{p'}$ is estimated by minimizing the cost function $E(a_{p'}b_{p'})$ and can be stated as:

$$E = \sum_{p \in \omega_{r'}(p')} [(a_{p'}G(p) + b_{p'} - X(p))^2 + \lambda.a_{p'}^2], \tag{4.8}$$

here λ is regularization parameter. It is used to compensate the large value of $a_{p'}$ by penalty. The following expression presents the optimum value of $a_{p'}$ and $b_{p'}$:

$$a_{p'} = \frac{\mu G.X, r'(p') - \mu G, r'(p')\mu X, r'(p')}{\sigma^2_{G,r'}(p') + \lambda} \tag{4.9}$$

$$b_{p'} = \mu X, r'(p') - a_{p'}\mu G, r'(p'), \tag{4.10}$$

where "." operator refers to product of two matrices (element wise), $\mu G.X, r'(p'), \mu G, r'(p')$ and $\mu X, r'(p')$ stands for the mean values of $G.X$, where G and X lies in window $\omega_{r'}(p')$ respectively.

Another extension of GIF is [37], weighted guided image filter (WGIF). The purpose behind the extension was to lesson down the halos obtained in GIF. A factor was introduced to become aware of edges. Although, it worked well in doing the same, but fall short in preserving edges well in some datasets.

4.3.2 Proposed Gradient Guided Filter

In both GIF and WGIF constraints are applied to get obtain the pixel values and to smooth the pixel values. There is no other specified constraints for the edge information in both. Thus, both fall short in preserving the edge information well. It is because both of them consider the process of filtering of image and process of edge preservation together. It is scientifically known that gradients are considered to be the integral as the way human beings perceive images of the world. The human cortical cells could be related to the gradients in an image. Therefore, it is expected to implement a new filtering approach comprising of the explicit constraints which can treat the edge information in order to make sure that the gradient values of the both input and output images is same.

Thus, gradient domain based guided image filter is proposed in this section. The introduced filter [43] not only consists of edge information preserving constraint and also successfully take care of the edge information better when compared to both the GIF and WGIF. In linear model 4.7 it can be noted that $\nabla Z(p) =$

$ap'\nabla G(p)$. Thus, smoothness term of Z in $\omega'_r(p')$ is the value of $a_{p'}$. When the value becomes 1 of $a_{p'}$, that indicates that the edges information is well preserved and this happens when the pixel p' is at an edge. Otherwise, if the pixel p' lies in some flat space, then the expected value of $a_{p'}$ becomes 0 so as to smooth out this flat region. Thus, considering this, a new cost function can be stated as:

$$E = \sum_{p \in \omega_{r'}(p')} [(a_{p'} G(p) + b_{p'} - X(p))^2 + \frac{\lambda}{\Gamma_G(p')} . (a_{p'} - \gamma_{p'})^2], \tag{4.11}$$

where $\gamma_{p'}$ tends to attain a value 1 if the pixel p' is on the edge and tends to attain value 0 if pixel lies in some smooth region. It can be deduced as, the value of $a_{p'}$ also attains 1 if the pixel p' is lies on edge and attains value 0 same pixel lies in some smooth region. This makes the proposed filter strong as it becomes less sensitive to the value of the λ selection. Consequently, edge information is better preserved with the help of proposed filter. The optimum values of $a_{p'}$ and $b_{p'}$ are computed as following:

$$a_{p'} = \frac{\mu G.X, r'(p') - \mu G, r'(p') \mu X, r'(p') + \frac{\lambda}{\Gamma_G(p')} \gamma_{p'}}{\sigma^2_{G,r'}(p') + \frac{\lambda}{\Gamma_G(p')}} \tag{4.12}$$

$$b_{p'} = \mu X, r'(p') - a_{p'} \mu G, r'(p'), \tag{4.13}$$

The final value of $Z(p)$ is given as follows:

$$\hat{Z}(p) = \bar{a}_{p'} G(p) + \bar{b}_{p'}, \tag{4.14}$$

where $\bar{a}_{p'}$ and $\bar{b}_{p'}$ are the mean of $a_{p'}$ and $b_{p'}$ in the window respectively.

Therefore, gradient domain based guided image filter is proposed explicit edge aware constraint. The filter consists of optimization and the cost function is comprises of data fidelity term and regularization term. This regularization term is the one which includes edge preserving constraint. It differs from other such terms in both GIF [21] and WGIF [23].

4.3.3 Matching Cost Computation

Markov random field (MRF)[44] provides a base for modeling contextual limitations in visual techniques and its interpretation. It plays a vital role in developing vision based algorithms optimally. It specifies how to assign a characteristic value to corresponding pixels. Conventional MRF theory can be used to generate the exact characteristic value for every pixel.

In stereo vision problems, the main purpose is finding out the disparity for every pixel. The image pair, left image I_l and right image I_r. I_l is the reference

or guidance image. As per the MRF theory, the set of disparity can be written as $L_d = l_1, l_2, l_3, ..., l_M,$. The further sections provides details about the pixel matching and common measures of cost function evaluations.

Pixel matching and cost function estimation

The main aim of the pixel matching [8] [13] [12] is to find similar matching pixel while searching linearly in the guidance image. This match is judged on the basis of similarity measure of the pixel. However much more common terminology is dissimilarity measure or also known as matching cost. Dissimilarity measures is indirectly proportional to the matching pixel value i.e. it increases with the decrease in similarity value of the two pixels.

A common way of defining matching cost is with the help of a $C(x, y, d)$ function in terms of guidance image coordinates and disparity value. The cost function results in the dissimilarity value for pixel or a coordinate (x, y, d). The disparity space comprises of image pixel coordinates and the value of disparity search range. Usually, range value is specified before the experiment manually and it also varies with the characteristics of input image.

Common measures for cost aggregation

Since last few decades, there has been a recent development in dissimilarity measures. There are few common measures which are used for pixel-by-pixel comparison and those are Absolute Intensity Difference (AD), Squared Intensity Difference (SD) and Absolute Gradient Difference (GRAD). The extension of these techniques is also widely used. It is based on doing comparison with square window regions which is centered about the reference and the search pixels. Thus the extended version of these measures is absolute difference as Sum of Absolute Intensity Differences (SAD), Squared Absolute Difference as Sum of Squared Intensity Differences (SSD) and Absolute Gradient Difference as Sum of Absolute Gradient Differences (SGRAD). Total summing up the costs over all window regions is termed as cost aggregation. It is mainly done to reduce the high noise sensitivity of pixel-by-pixel comparison.

However, this region based technique may be responsible for low detailed outputs. Therefore, choosing a aggregation window size helps maintaining the trade off between noise sensitivity and loss of detail.

The varies cost functions for dissimilarity measures discussed above can mathematically be written as:

$$C_{AD}(x, y, d) = |I_L(x, y) - I_R(x - d, y)| \tag{4.15}$$

$$C_{SD}(x,y,d) = |I_L(x,y) - I_R(x-d,y)|^2 \tag{4.16}$$

$$C_{GRAD}(x,y,d) = |\nabla_x I_L(x,y) - \nabla_x I_R(x-d,y)| + |\nabla_y I_L(x,y) - \nabla_y I_R(x-d,y)| \tag{4.17}$$

$$C_{SAD}(x,y,d) = \sum_{(u,v)\in W(x,y)} |I_L(u,v) - I_R(u-d,v)| \tag{4.18}$$

$$C_{SSD}(x,y,d) = \sum_{(u,v)\in W(x,y)} |I_L(u,v) - I_R(u-d,v)|^2 \tag{4.19}$$

$$\begin{aligned} C_{SGRAD}(x,y,d) = & \sum_{(u,v)\in W(x,y)} |\nabla_x I_L(u,v) - \nabla_x I_R(u-d,v)| \\ & + \sum_{(u,v)\in W(x,y)} |\nabla_y I_L(u,v) - \nabla_y I_R(u-d,v)| \end{aligned} \tag{4.20}$$

Here, I_L and I_R refers to the intensity of pixels of the left and right image, respectively. $W(x,y)$ is window having position (x,y), and ∇_x and ∇_y are gradient operators.

There exist other cost aggregation technique and dissimilarity measures methods beside those as mentioned above which include normalized measures, like normalized cross correlation (NCC) and its extension normalized sum of squared differences (NSSD). Other than these, few numerous local transforms such as census transform, rank transform and fourier transform may also be applied on regions around the pixels. The requirement of normalization may be reduced with the help of image pre-processing however, local transforms very rarely seem to have any solid advantage over common gradient and difference based measures.

Adaptive support-weights in cost aggregation

A way to aggregate cost, which has recently started to show up in algorithms, is to set weights of pixels within aggregation windows on the basis of both color intensity similarity and measured distance to the center pixel. One of the approaches to do this is based on the bilateral filtering technique[15; 16] which was introduced into computer vision in the late nineties. Various techniques have been used in segmentation for the purpose of cost aggregation. The method of cost aggregation presented in [10] uses the larger size pixel square window, and then weighting pixels separately based on whether these pixels which belong to similar segment as of center pixel or does not belong to same segment. Thus the pixels which are

located in similar segment as that of the center pixel then they are weighted with value of 1 , and those pixels which lie outside the segment are weighted with the δ small value.

4.3.4 Handling of Occlusion and Outliers

It is of great advantage that occlusion is dealt within stereo algorithm implicitly. Moreover, there are various algorithms implemented to deal the occlusion explicitly. One of these algorithms is right-to-left and left-to-right cross checking. It is popular because it is simple and much effective in comparison to other approaches to remove occlusion. In order to increase reliability of the algorithm, consider cost function and estimate the difference between winning disparity cost value and second lowest cost value of disparity. As mentioned earlier, truncating cost may greatly help improving the result value at disparity discontinuity where occlusion is noticeable problem.

Cross Checking between Left and Right Disparity Estimates

Cross-checking is carried out between left and right input image pair by calculating two different disparity maps, i.e. one map for each view of input image pair. Then performing the warping of disparity values to corresponding expected location in other view of the image pair. Thus, reliability of disparity map can be estimated by simply comparing the disparities values of original computed maps for each view with disparity value present in warped version which originated from the other view of the stereo pair. Pixels fulfilling the equalities below are consistent and therefore not considered to be occluded:

$$d_{LEFT}(x, y) = d_{RIGHT}(x - d_{LEFT}(x, y), y) \tag{4.21}$$

$$d_{RIGHT}(x, y) = d_{LEFT}(x - d_{RIGHT}(x, y), y) \tag{4.22}$$

In the 4.21 and 4.22 relations, $d_{LEFT}(x, y)$ and $d_{RIGHT}(x, y)$ represents disparity maps for the left and right views of the scene. Pixels with disagreeing disparity values in the above comparisons are considered to be unreliable. The disagreement may very well be explained by occlusion, but this method also handles mismatches resulting from other reasons such as noise or lack of texture. Unreliable pixels can be excluded from contributing to the final result, which may improve the quality.

Truncating Cost

When cost aggregation is done, the final result may be some what affected by the outlier pixels in the image. If the final aggregated cost of the correct disparity

that would been chosen in the Winner-take-all optimization technique, if noise free images were there then it is influenced by outliers, then some another value for disparity may be selected. This problem often came across in discontinuous disparity area in other words where both foreground as well as background pixels are present at the same time in aggregation window. The cost truncation may help in enhancing the global optimization. The effects of truncating cost function helps in reducing outliers influences. The truncation function can be written as:

$$C(x, y, d) = min(C(x, y, d), \alpha) \tag{4.23}$$

The threshold value is fixed with the outcome of reliable disparity estimations which in turn have been verified by cross checking. The value, although, is fixed slightly above what is actual cost value of reliable disparities estimation. However, cost truncation function is being implemented just to manually fix the threshold value. The parameter, as known, is dependent on the characteristics of input images, as one of them are noise level. This shows that either for every image pair the parameter has to be always fixed manually, or rather fix some high value, which in turn lowers improvement of result.

4.3.5 Image Segmentation for Smoothness

Since last decade, Image segmentation technique has marked a successful contribution in stereo correspondence algorithms. The aim of segmentation of image technique is to decrease total number of colors of input guidance image and thereafter grouping neighborhood pixels of same color value together so as to form bounded closed segments. The mean-shift based color segmentation technique has been widely used in stereo algorithms nowadays and has proven to be better way of segmenting image into regions with boundaries that usually coincide with object borders.

Mean-shift Image Segmentation

The mean shift segmentation technique is mode-fitting and non-parametric approach based technique. It works using mode which is the local maximum of the probability density function and being non parametric it makes no pre-assumptions on the statistical properties of the data. This technique can also be used for feature space analysis of an image to find the color space belonging to an image and segment image into clusters. The data contained in image is modeled as sample with help of multivariate distribution and with unknown (PDF) probability density function and thus finding its modes.

Considering statistically, if given the number of data samples, kernel density based estimation is a fine way to determine unknown probability density function. Underlying phenomenon can be understood with the help of univariate kernel based density estimator can be written as:

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} K(x - x_i) \tag{4.24}$$

Similarly, a Gaussian kernel can be expressed as following:

$$K(x) = \frac{1}{\sqrt{2\pi}h} exp\left(-\frac{x^2}{2h^2}\right) \tag{4.25}$$

where h denotes the parameter which controls bandwidth of kernel.

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) \tag{4.26}$$

For the d dimensional set of data points, [32] authors uses simplified multivariate based kernel density estimator and that can be written as:

$$K_h(x) = \frac{c}{h^d} k\left(\frac{||x||^2}{h^2}\right) \tag{4.27}$$

Now as the estimator is being defined, the goal now shifted onto searching for the local maximum of density. This can be implemented without even calculating the density value itself, and by estimating its gradient value as follows:

(4.28)

With the mathematical calculation done, the mean shift final formulations can be done as follows:

$$m(x) = \frac{\sum_{i=1}^{n} x_i G(x - x_i)}{\sum_{i=1}^{n} G(x - x_i)} - x \tag{4.29}$$

Therefore, mean-shift vector represents weighted mean of the data point set estimated in terms of x. Clearly, it always directs towards the direction of gradient, in other words towards the local maximum of density. For the sensible use of mean-shift image segmentation technique, it is required to include spatial terms in the kernel function.

Setting Smoothness Terms based on Segmentation Results

With the help of segmentation information it will be simpler to make disparity discontinuities boundaries coincide with the segment border in global optimization. The energy function of this type $E(d) = E_{data}(d) + \lambda.E_{smooth}(d)$,smoothness term can be stated as following:

$$E_{smooth}(d) = \sum_{(i,j)\in N} V_{ij}(d_i, d_j) \tag{4.30}$$

$$V_{ij}(d_i, d_j) = \begin{cases} \lambda.\rho_d(d_i - d_j), & \text{for } s_i = s_j \\ \gamma.\lambda.\rho_d(d_i - d_j), & \text{for } s_i != s_j \end{cases} \tag{4.31}$$

This clearly shows that when the neighborhood pixels i, j belongs to one segment of image such that $s_i = s_j$, then the parameter for smoothness is equal to λ . Whereas, when the neighborhood pixels are situated in different segments of image, then cost is narrowed down with the help of multiplication of λ parameter with some constant known as γ where $\gamma \in (0, 1)$, thus enhancing the disparity jumps which occurs at the segment boundaries.

Chapter 5

Experimental Results and Implementation

This chapter presents the results obtained after implementing the proposed stereo algorithm. To serve this purpose, the existing algorithms have been considered as the source of quantitative comparison. Through out the thesis work, it has always been desirable to implement the algorithm using techniques which are of less computational complexity.

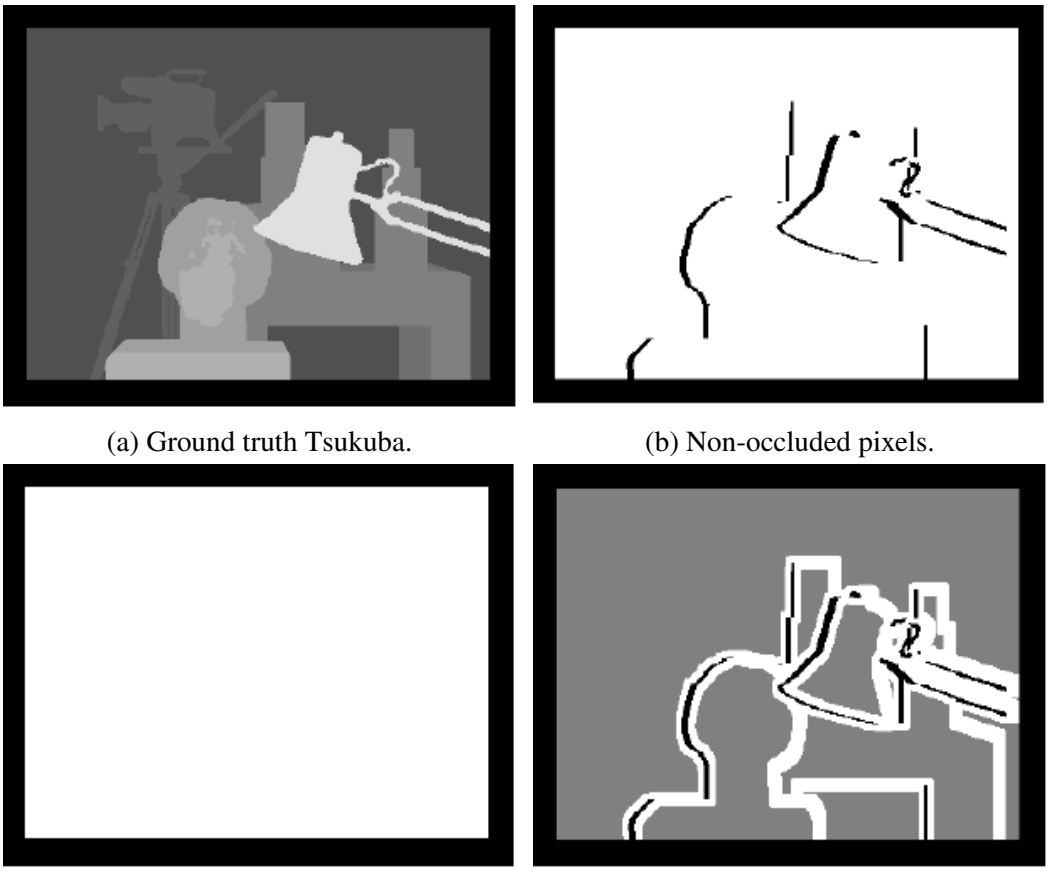
5.1 Error Evaluation Using Ground Truth Data

The authors behind the taxonomy [2] and the Middlebury Stereo homepage [3] have developed an objective performance measure that compares pixel subsets of disparity maps coming from different algorithms against the same pixel subsets in ground truth data.

The subsets are non-occluded pixels (nonocc), all pixels (all) and disparity discontinuity pixels (disc). They are determined by performing warping of ground truth disparity maps or by selecting pixels near discontinuities where the change in disparity is larger than a specified threshold value. In figure 5.1, the subsets as computed for the Tsukuba stereo images are shown. Here the error is measured only in white areas. The quantitative measure that has been used to evaluate performance of algorithms is the percentage of bad pixels in subsets of estimated disparity maps, which is defined as:

$$B_S = \frac{1}{N_S} \sum_{(x,y)\in S} |d_{estimate}(x,y) - d_{groundtruth}(x,y)| > \delta \qquad (5.1)$$

Here, S denotes the desired subset of pixels. N_S is the number of pixels in S and δ is the error tolerance. Throughout this work, a tolerance of 1.0 has been



(a) Ground truth Tsukuba. (b) Non-occluded pixels.

(c) All pixels. (d) Pixels of discontinuous regions.

Figure 5.1: Errors are evaluated in white areas.

used as this corresponds to the value used in [2] and several other studies.

5.2 Gradient Guided Filter

The gradient guided filter has been implemented using a MATLAB version10 using computer vision toolbox. The section 4.3.2 describes the following energy function:

$$E = \sum_{p \in \omega_{r'}(p')} \left[(a_{p'} G(p) + b_{p'} - X(p))^2 + \frac{\lambda}{\Gamma_G(p')} . (a_{p'} - \gamma_{p'})^2 \right], \tag{5.2}$$

where $\gamma_{p'}$ is described in section 4.3.2. The optimal values of $a_{p'}$ and $b_{p'}$ are computed as following:

$$a_{p'} = \frac{\mu G.X, r'(p') - \mu G, r'(p')\mu X, r'(p') + \frac{\lambda}{\Gamma_G(p')}\gamma_{p'}}{\sigma^2_{G,r'}(p') + \frac{\lambda}{\Gamma_G(p')}} \tag{5.3}$$

$$b_{p'} = \mu X, r'(p') - a_{p'}\mu G, r'(p'), \tag{5.4}$$

The ultimate $Z(p)$ is given as follows:

$$\hat{Z}(p) = \bar{a}_{p'}G(p) + \bar{b}_{p'}, \tag{5.5}$$

where $\bar{a}_{p'}$ and $\bar{b}_{p'}$ are the mean of $a_{p'}$ and $b_{p'}$ in the window respectively.

GGIF has four parameters that control the output; the guidance image I, filtering input image p, local window radius and regularization parameter. This function returns a filtered image.

5.3 Pixel Matching

Pixel have been matched by implementing the following two differences, absolute intensity differences and absolute gradient differences. Further to implement cross checking later, following cost functions are computed and result is stored for both left and right view of the image:

$$C_{ADL}(x, y, d) = |I_L(x, y) - I_R(x - d, y)| \tag{5.6}$$

$$C_{ADR}(x, y, d) = |I_L(x + d, y) - I_R(x, y)| \tag{5.7}$$

$$C_{GRADL}(x, y, d) = |\nabla_x I_L(x, y) - \nabla_x I_R(x-d, y)| + |\nabla_y I_L(x, y) - \nabla_y I_R(x-d, y)| \tag{5.8}$$

$$C_{GRADR}(x, y, d) = |\nabla_x I_L(x+d, y) - \nabla_x I_R(x, y)| + |\nabla_y I_L(x+d, y) - \nabla_y I_R(x, y)| \tag{5.9}$$

When horizontal pixel coordinates point outside the available image coordinate range, they are set to the smallest or largest available coordinate value. The BT dissimilarity measure was also implemented, and may optionally replace the absolute values in the above equations when desired.

5.4 Winner-Take-All Optimization

The WTA technique explained is implemented as following:

5.4.1 Cost aggregation

First, cost aggregation is performed, if desired. Referring to the section 4.3.3 aggregation is implemented as an averaging operation using a square window, with manually specified window size. To speed things up, the aggregation is done by performing 2D convolution on each disparity layer of the pixel matching cost function:

$$C_{SAD}(x, y, d) = C_{AD}(x, y, d) * W(x, y) \tag{5.10}$$

$$C_{SGRAD}(x, y, d) = C_{GRAD}(x, y, d) * W(x, y) \tag{5.11}$$

Here, $W(x, y)$ is a window containing (x, y). And $C_{SAD}(x, y, d)$, $C_{AD}(x, y, d)$, $C_{SGRAD}(x, y, d)$ and $C_{GRAD}(x, y, d)$ is defined in section 4.3.3.

5.4.2 WTA Optimization and Cross Checking

A weighting between SAD and SGRAD cost is calculated from the amount of consistency in a left to right and right to left cross checking, explained in section 4.3.4. The WTA-optimization and subsequent cross-checking is performed for a number of discrete weights $\omega_i \in [0, 1]$, which are predetermined:

$$C_L^i(x, y, d) = (1 - \omega).C_{SADL}(x, y, d) + \omega_i.C_{SGRADL}(x, y, d) \tag{5.12}$$

$$C_R^i(x, y, d) = (1 - \omega).C_{SADR}(x, y, d) + \omega_i.C_{SGRADR}(x, y, d) \tag{5.13}$$

$$d_L^i(x, y) = argmin_d C_L^i(x, y, d) \tag{5.14}$$

$$d_R^i(x, y) = argmin_d C_R^i(x, y, d) \tag{5.15}$$

The disparity estimates $d_L^i(x, y)$ and $d_R^i(x, y)$ created for every weight are warped as described in Section 3.5.1, and the weight ω_i giving the largest number of valid cross matches is considered to be optimal. Also, a bad pixel map $b(x, y)$ is computed. This map contains information on which pixels that failed the mutual left and right cross checking test. This information is used later to exclude bad disparity estimates from influencing the final result, and also for cost truncation.

5.4.3 Cost Truncation

To limit the influence of outliers in later steps of the algorithm, truncation of the cost is performed as explained in section 4.3.4. The implemented cost truncation makes use of the cross checking information and stores the WTA cost for reliable matches in a list. The list is then histogram analyzed, and the frequency of cost values belonging to the reliable matches is computed. A cost truncation threshold is determined by summation of the number of occurrences for each of the unique reliable cost values, starting at the lowest cost. When the number of summed cost value occurrences corresponds to a certain percentage of all reliable matches, the summation is terminated. The threshold is set to the cost value just above the highest cost value whose number of occurrences took part in the summation.

5.5 Image Segmentation

The mean-shift segmentation has been implemented using a MATLAB version10 using computer vision toolbox. The mean-shift segmentation algorithm works as described in section 4.3.5, and has one parameters that control the output; the bandwidth b_w. This function returns a segmented version of the input image and a label map, where each pixel has been assigned an integer label number corresponding to a segment of clustered pixels.

5.6 Experimental Results

The proposed algorithm has been analyzed with the Middlebury stereo testbed results. The four standard image dataset Teddy, Tsukuba, Cones and Venus are used for testing the algorithms and their reults are shown in figure 5.2, 5.3, 5.4, 5.5 respectively. Results include the disparity map obtained without filtering process and disparity map obtained after filtering process. The disparity maps generated by newly proposed algorithm for all the four standard dataset differ from ground truth disparity maps in minute details. It can also be noted that proposed algorithm does perform well in preserving the edge information and smoothness of the object as compared to disparity maps obtained with GIF filtering.

Throughout the experiments, the same error measures as those on the Middlebury Stereo homepage [3] will be computed and used. These are the percentages of bad pixels in non-occluded areas, all areas and discontinuous areas. For clarity, these errors will be abbreviated $nonocc, all$ and $disc$. Here, nonocc refer to non occluded areas, disc refer to edges and all refer to the whole area.

In table 5.1, 5.2, 5.3 and 5.4 respectively represents proposed algorithm comparison with stereo matching based on GIF and Middlebury Testbed results re-

spectively for standard images Teddy, Tsukuba, Cones and Venus. The table 5.5 presents the average percent of bad pixels for all three algorithms.

.

Table 5.1: Error results comparing for Teddy dataset.

Algorithm	***Non-occluded***	***All***	***Discontinuous***
Proposed	3.54	6.93	9.91
GIF	4.84	7.06	13.8
Testbed	4.08	5.98	11.4

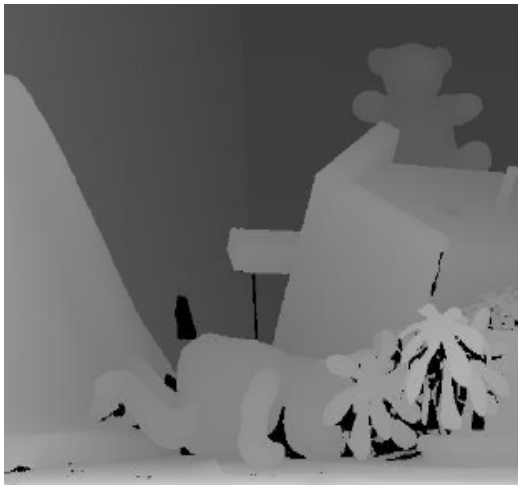
Table 5.2: Error results comparing for Tsukuba dataset.

Algorithm	***Non-occluded***	***All***	***Discontinuous***
Proposed	0.87	1.16	4.66
GIF	1.01	1.42	6.23
Testbed	0.93	1.37	5.05

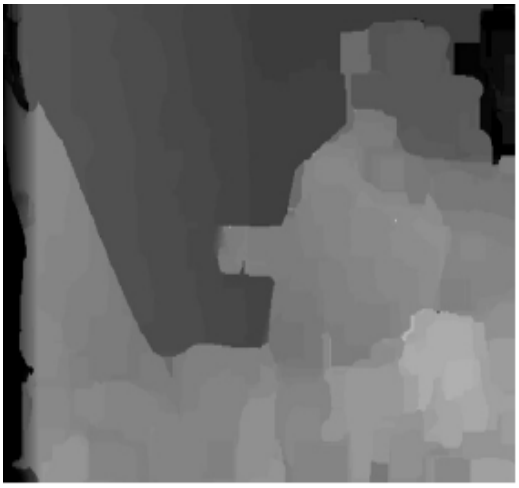
Table 5.3: Error results comparing for Cones dataset.

Algorithm	***Non-occluded***	***All***	***Discontinuous***
Proposed	1.76	6.70	5.99
GIF	2.28	6.68	6.95
Testbed	2.14	6.97	6.27

(a) Teddy dataset.



(b) Teddy standard disparity map.



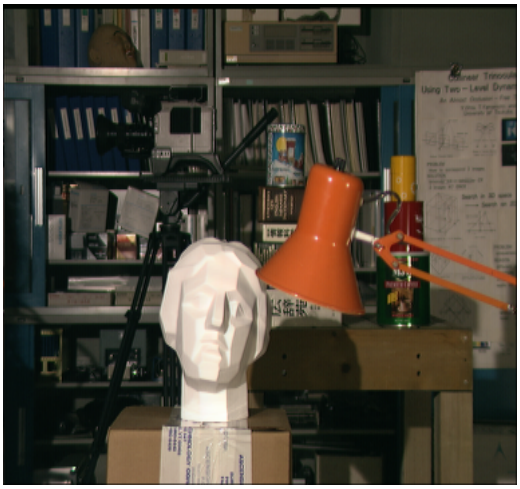
(c) Disparity map GIF filter.

(d) Teddy GGIF filtered disparity map.

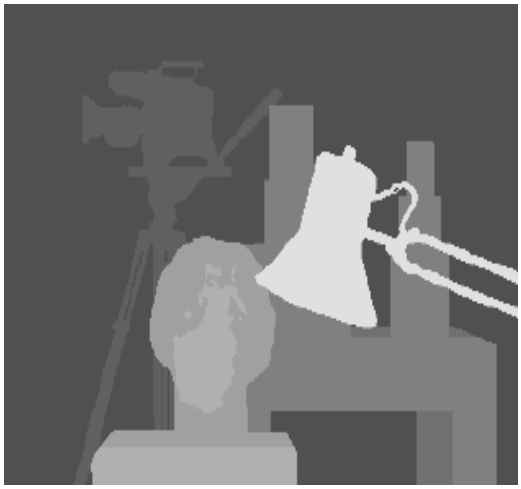
Figure 5.2: Proposed algorithm disparity map for Teddy dataset.

Table 5.4: Error results comparing for Venus dataset.

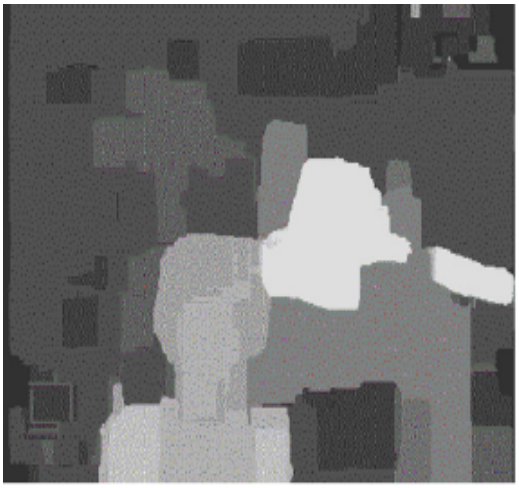
Algorithm	***Non-occluded***	***All***	***Discontinuous***
Proposed	0.08	0.20	1.12
GIF	0.09	0.22	1.15
Testbed	0.07	0.17	1.04



(a) Tsukuba dataset.



(b) Tsukuba standard disparity map.



(c) Disparity map with GIF filter.

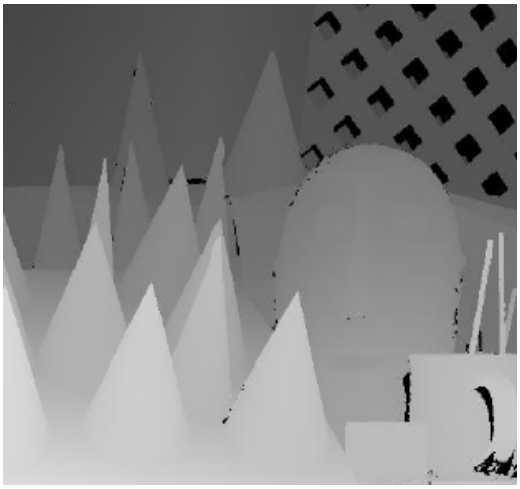
(d) Tsukuba GGIF filtered disparity map.

Figure 5.3: Proposed algorithm disparity map for Tsukuba dataset.

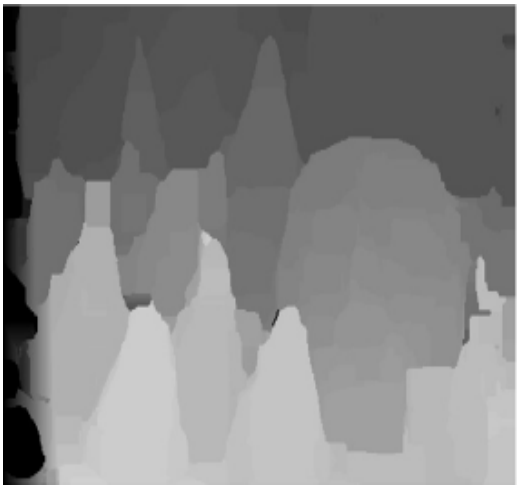
Table 5.5: Average percent of bad pixels.

Algorithm	***Average Error%***
Proposed	3.58
GIF	4.31
Testbed	3.79

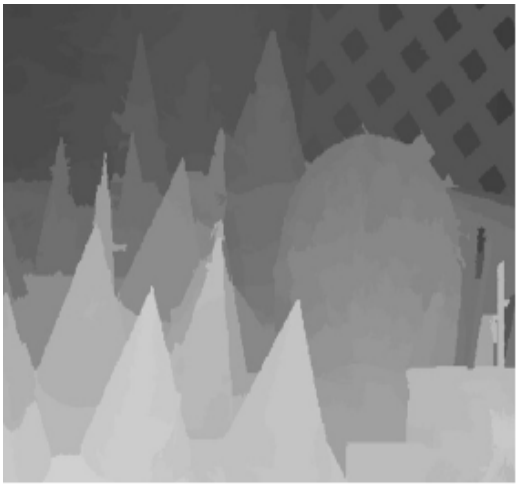
(a) Cones dataset.



(b) Cones standard disparity map

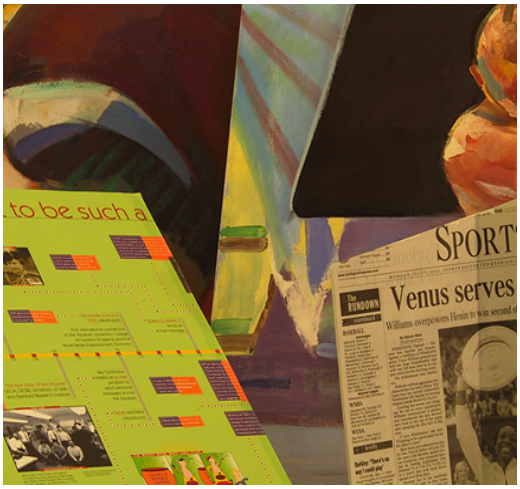


(c) Disparity map with GIF filter.

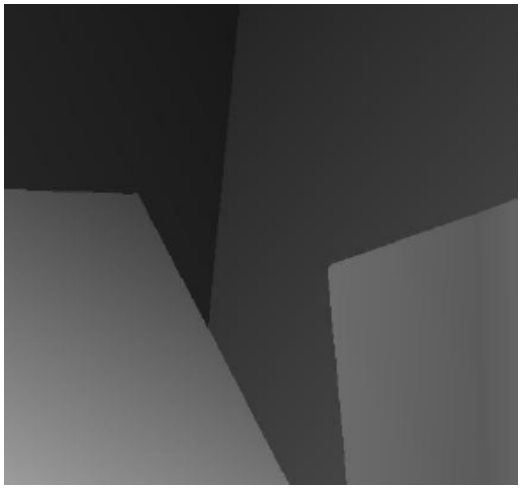


(d) Cones GGIF filtered disparity map.

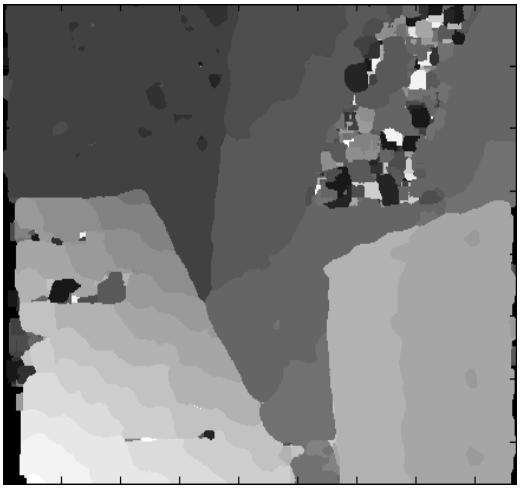
Figure 5.4: Proposed algorithm disparity map for Cones dataset.



(a) Venus dataset.



(b) Venus standard disparity map.



(c) Disparity map with GIF filter.

(d) Venus GGIF filtered disparity map.

Figure 5.5: Proposed algorithm disparity map for Venus dataset.

Chapter 6

Conclusion and Future Scope

This chapter concludes the document with a listing of important conclusions, contributions and the scope for further extension.

6.1 Conclusion and Contribution

Some important conclusions drawn through the course of the research work are presented as following:

1. The area of depth map estimation continues to provide several challenges to researchers over the years. There are many situations where even the best proposed techniques fail to estimate a good depth map.

2. A complete algorithmic solution is developed that is capable of performing matching embedded with better filtering technique.

3. Algorithm has proved successful on several different type of datasets which are used for experimentation and the results are presented for four standard dataset.

4. Experiment helps verifying the sustenance of high quality in depth maps, while reducing the average percent of bad pixels to 3.58%.

5. Improved version of Guided Image Filter as Gradient Guided Image Filter was implemented successfully which in turn improved the visual quality of depth map and retained edge information better.

6. Experiments shows that halos are minimized in the depth maps and are also visually visible in the results.

6.2 Future Scope

Since a novel approach is proposed in the thesis, it can be refined to perfection in several directions. A few goals that could spark further research based on the work presented can be listed as following :

1. Methods that are robust to noise and that can handle real world stereo images also, while still maintaining reasonable computational time, need to be found.

2. Reducing occlusion problems and to increase quality/speed would be interesting to look further into. Some proposed ideas were found, but it seems as if not much research has been made within this field. Also, several of the proposed ideas seemed to be slow by their outlines.

3. Segmentation of image algorithm can be improved with better termination condition and less comptutaion time so as to have a better object detection.

References

[1] X. Hu and P. Mordohai, "A quantitative evaluation of confidence measures for stereo vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 2121–2133, Nov 2012.

[2] R. Szeliski, *Computer Vision: Algorithms and Applications*. New York, NY, USA: Springer-Verlag New York, Inc., 1st ed., 2010.

[3] "Middlebury Homepage." `vision.middlebury.edu/stereo/`. Accessed: 2010-09-30.

[4] "Middlebury Dataset." `http://vision.middlebury.edu/stereo/data/`. Accessed: 2010-09-30.

[5] C. C. Pham and J. W. Jeon, "Domain transformation-based efficient cost aggregation for local stereo matching," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, pp. 1119–1130, July 2013.

[6] V. D. Nguyen, D. D. Nguyen, S. J. Lee, and J. W. Jeon, "Local density encoding for robust stereo matching," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, pp. 2049–2062, Dec 2014.

[7] T. Mouats, N. Aouf, and M. A. Richardson, "A novel image representation via local frequency analysis for illumination invariant stereo matching," *IEEE Transactions on Image Processing*, vol. 24, pp. 2685–2700, Sept 2015.

[8] J. Jiao, R. Wang, W. Wang, S. Dong, Z. Wang, and W. Gao, "Local stereo matching with improved matching cost and disparity refinement," *IEEE MultiMedia*, vol. 21, pp. 16–27, Oct 2014.

[9] L. Xu, O. C. Au, W. Sun, L. Fang, F. Zou, and J. Li, "Stereo matching with optimal local adaptive radiometric compensation," *IEEE Signal Processing Letters*, vol. 22, pp. 131–135, Feb 2015.

[10] K. J. Yoon, "Stereo matching based on nonlinear diffusion with disparity-dependent support weights," *IET Computer Vision*, vol. 6, pp. 306–313, July 2012.

[11] A. Hosni, M. Bleyer, C. Rhemann, M. Gelautz, and C. Rother, "Real-time local stereo matching using guided image filtering," in *2011 IEEE International Conference on Multimedia and Expo*, pp. 1–6, July 2011.

[12] D. Min, J. Lu, and M. N. Do, "Joint histogram-based cost aggregation for stereo matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 2539–2545, Oct 2013.

[13] G. A. Kordelas, D. S. Alexiadis, P. Daras, and E. Izquierdo, "Content-based guided image filtering, weighted semi-global optimization, and efficient disparity refinement for fast and accurate disparity estimation," *IEEE Transactions on Multimedia*, vol. 18, pp. 155–170, Feb 2016.

[14] Y. S. Heo, K. M. Lee, and S. U. Lee, "Robust stereo matching using adaptive normalized cross-correlation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 4, pp. 807–822, 2011.

[15] L. Fu, G. Peng, and W. Song, "Histogram-based cost aggregation strategy with joint bilateral filtering for stereo matching," *IET Computer Vision*, vol. 10, no. 3, pp. 173–181, 2016.

[16] Q. Yang, "Hardware-efficient bilateral filtering for stereo matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, pp. 1026–1032, May 2014.

[17] A. K. Jain and T. Q. Nguyen, "Discriminability limits in spatio-temporal stereo block matching," *IEEE Transactions on Image Processing*, vol. 23, pp. 2328–2342, May 2014.

[18] J. Kowalczuk, E. T. Psota, and L. C. Perez, "Real-time stereo matching on cuda using an iterative refinement method for adaptive support-weight correspondences," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, pp. 94–104, Jan 2013.

[19] A. Hosni, M. Bleyer, and M. Gelautz, "Secrets of adaptive support weight techniques for local stereo matching," *Computer Vision and Image Understanding*, vol. 117, no. 6, pp. 620 – 632, 2013.

[20] Q. Yang, "Stereo matching using tree filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, pp. 834–846, April 2015.

[21] C. Ttofis, C. Kyrkou, and T. Theocharides, "A low-cost real-time embedded stereo vision system for accurate disparity estimation based on guided image filtering," *IEEE Transactions on Computers*, vol. PP, no. 99, pp. 1–1, 2015.

[22] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 1397–1409, June 2013.

[23] G. S. Hong, M. S. Koo, A. Saha, and B. G. Kim, "Efficient local stereo matching technique using weighted guided image filtering (wgif)," in *2016 IEEE International Conference on Consumer Electronics (ICCE)*, pp. 484–485, Jan 2016.

[24] Y. Zhou and C. Hou, "Stereo matching based on guided filter and segmentation," *Optik-International Journal for Light and Electron Optics*, vol. 126, no. 9, pp. 1052–1056, 2015.

[25] D. Chen, M. Ardabilian, and L. Chen, "A fast trilateral filter-based adaptive support weight method for stereo matching," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, pp. 730–743, May 2015.

[26] T. Poggio, V. Torre, and C. Koch, "Computational vision and regularization theory," *Nature*, vol. 317, no. 6035, pp. 314–319, 1985.

[27] A. Geiger, M. Roser, and R. Urtasun, "Efficient large-scale stereo matching," in *Computer Vision–ACCV 2010*, pp. 25–38, Springer, 2010.

[28] K.-L. Tang, C.-K. Tang, and T.-T. Wong, "A markov random field approach for dense photometric stereo," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, pp. 1197 vol. 2–, June 2005.

[29] E. P. Kim, J. Choi, N. R. Shanbhag, and R. A. Rutenbar, "Error resilient and energy efficient mrf message-passing-based stereo matching," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 24, pp. 897–908, March 2016.

[30] J. Choi and R. A. Rutenbar, "Fpga acceleration of markov random field trw-s inference for stereo matching," in *Formal Methods and Models for Codesign (MEMOCODE), 2013 Eleventh IEEE/ACM International [2]Conference on*, pp. 139–142, Oct 2013.

[31] J. Choi and R. A. Rutenbar, "Video-rate stereo matching using markov random field trw-s inference on a hybrid cpu x002b;fpga computing platform,"

IEEE Transactions on Circuits and Systems for Video Technology, vol. 26, pp. 385–398, Feb 2016.

[32] A. Arranz, A. Sanchez, and M. Alvar, "Multiresolution energy minimisation framework for stereo matching," *IET Computer Vision*, vol. 6, pp. 425–434, Sept 2012.

[33] M. G. Mozerov and J. van de Weijer, "Accurate stereo matching by two-step energy minimization," *IEEE Transactions on Image Processing*, vol. 24, pp. 1153–1163, March 2015.

[34] J. Cai, "Integration of optical flow and dynamic programming for stereo matching," *IET Image Processing*, vol. 6, pp. 205–212, April 2012.

[35] Y. C. Lai, C. C. Cheng, C. K. Liang, and L. G. Chen, "Efficient message reduction algorithm for stereo matching using belief propagation," in *2010 IEEE International Conference on Image Processing*, pp. 2977–2980, Sept 2010.

[36] J. H. Kim, J. W. Kwon, and Y. H. Ko, "Multi-baseline based texture adaptive belief propagation stereo matching technique for dense depth-map acquisition," in *Electronics, Information and Communications (ICEIC), 2014 International Conference on*, pp. 1–2, Jan 2014.

[37] Z. Li, J. Zheng, Z. Zhu, W. Yao, and S. Wu, "Weighted guided image filtering," *IEEE Transactions on Image Processing*, vol. 24, pp. 120–129, Jan 2015.

[38] Q. Yang, D. Li, L. Wang, and M. Zhang, "Full-image guided filtering for fast stereo matching," *IEEE Signal Processing Letters*, vol. 20, pp. 237–240, March 2013.

[39] J. Huang, "Stereo matching based on segmented b-spline surface fitting and accelerated region belief propagation," *IET Computer Vision*, vol. 9, no. 4, pp. 456–466, 2015.

[40] C. Shi, G. Wang, X. Yin, X. Pei, B. He, and X. Lin, "High-accuracy stereo matching based on adaptive ground control points," *IEEE Transactions on Image Processing*, vol. 24, pp. 1412–1423, April 2015.

[41] S. Huq, A. Koschan, and M. Abidi, "Occlusion filling in stereo: Theory and experiments," *Computer Vision and Image Understanding*, vol. 117, no. 6, pp. 688 – 704, 2013.

[42] O. Veksler, *Efficient Graph-based Energy Minimization Methods in Computer Vision*. PhD thesis, Ithaca, NY, USA, 1999. AAI9939932.

[43] F. Kou, W. Chen, C. Wen, and Z. Li, "Gradient domain guided image filtering," *IEEE Transactions on Image Processing*, vol. 24, pp. 4528–4539, Nov 2015.

[44] S. Z. Li, *Markov random field modeling in image analysis*. Springer Science & Business Media, 2009.

[45] "Shai Bagon(2010)." `http://www.wisdom.weizmann.ac.il/~bagon/matlab.html`. Shai Bagon's MATLAB code homepage. [Online].

Publications

1. K. Jain, H. S. Pannu and V. Bassi, "Complete Survey of 3D Model Generation," in *IEEE International Conference On Innovations In Information, Embedded And Communication Systems*, vol. 4, pp. 2013-2015, March 2016.

2. K. Jain, H. S. Pannu and D. S. Gill, "Stereo Matching Based Depth Map Estimation from Stereo Image Pair" in *Advanced Robotics Journal of Taylor and Francis*, July 2016.

3. K. Jain, H. S. Pannu and D. S. Gill, "3D Surface Reconstruction using Support Vector Regression and Tuning Automation" in *Computer Applications in Engineering Education Journal of Wiley Online Library*, July 2016.

Video Presentation and Plagiarism Report

The video presentation on the thesis work can be found on following link:
`https://www.youtube.com/channel/UCWuRpdHimavFpehexJVHH4g`

The generated plagiarism report is attached.

StereoVision

ORIGINALITY REPORT

9%	4%	8%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Kou, Fei, Weihai Chen, Changyun Wen, and Zhengguo Li. "Gradient Domain Guided Image Filtering", IEEE Transactions on Image Processing, 2015. Publication	1%
2	Submitted to iGroup Student Paper	1%
3	Zhou, Yuan, and Chunping Hou. "Stereo matching based on guided filter and segmentation", Optik - International Journal for Light and Electron Optics, 2015. Publication	1%
4	Stephen Se. "Passive 3D Imaging", 3D Imaging Analysis and Applications, 2012 Publication	<1%
5	ira.lib.polyu.edu.hk Internet Source	<1%
6	Submitted to Middle East Technical University Student Paper	<1%
7	www.ijcte.org Internet Source	<1%