Image Depth Estimation Using Stereo Vision

Mrinall Umasudhan

April 25, 2022

Candidate ID: 943289432.

Contents

§1 Introduction

One of the most explored problems in the field of computer vision is the process of accurately estimating the real-world depth of a pixel within a two dimensional image. The inference of three dimensional information is done by using multiple two dimensional views of a scene, the process being deemed the name stereo vision.

§1.1 Applications

A common counter-argument to the practicality of stereo vision algorithms are the presence of other sensors that do not make use of visual data such as ultrasonic or time of flight distance sensors. While these sensors are not not impacted by factors that would be detrimental to the accuracy of stereo vision algorithms such as the lack of adequate lighting, "stereo vision has the advantage that it achieves the 3-D acquisition without energy emission or moving parts" (https://research.csiro.au/qi/stereo-vision/). Moreover, whereas traditional distance sensors focus on a singular point in space, stereo vision algorithms are only limited by the camera's field of view, making the depth analysis large area far more simple and cost effective. Finally, stereo vision algorithms are able to easily work in conjunction with other computer vision techniques such as machine learning based object detection models when compared to the previous depth estimation approaches as it already tracks depth on the same image plane that a object detection model may be implemented on. These factors allow for a far greater analysis of the various shapes and angles in an image leading to its usage in various fields.

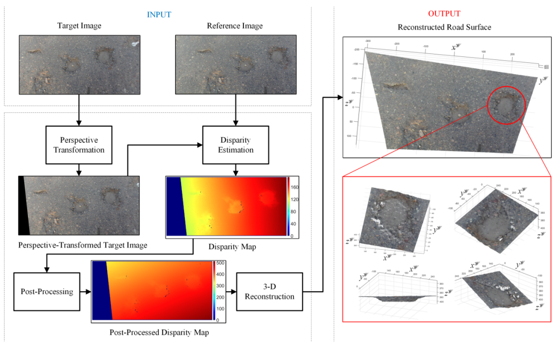

Figure 1: Stereo Vision For Road Deformity Detection

A common application of stereo vision algorithms is in the quality control process of industrial factories. Factories must analyze each finished product for deformities in order to maintain a standard of quality in their products. However, many factories output a high volume of product every day meaning that the human analysis of such product would be far too expensive and inefficient when considering the large amounts of workers needed to manually inspect each products as well as the time it takes for the inspection of a product. The installation of multiple distance sensors in order to analyze each square inch of a product would also be far too expensive. However, because a factory is in a controlled environment with uniform lighting and the object is one of known shape, the usage of a stereo vision algorithm would be ideal for the situation as stereo cameras are able to analyze objects with their large field of view and can easily detect deformities as the object being analyzed is of known geometry, meaning the algorithm can compare each depth point of the current object to the depth of a model product, reporting any deformities both accurately and efficiently. and can easily detect deformities as the object being analyzed is of known geometry, meaning the algorithm can compare each depth point of the current object to the depth of a model product, reporting any deformities both accurately and efficiently.

A paper done by the University of Bristol explored deformity analysis preformed by stereo algorithms further by acquiring three dimensional road data for autonomous cars further displaying the capabilities of stereo algorithms. The process which this algorithm follows can be seen in the figure above.

§1.2 Overview and Purpose

The essence of many successful stereo vision algorithms can be summarized in three key steps:

1. Triangulation: The process of assigning depth values to each pixel in the image using multiple two dimensional views of a scene and the specific parameters from camera hardware, difference in location of the cameras used, and the disparity in the pixels from each view of the scene.

2. Calibration: The process of correcting image distortion caused by the spherical geometry of the camera lens and reifying the two dimensional views of the scene such that the objects in study are on the same plane.

3. Pixel Correspondence: In order to apply the triangulation process the algorithm must be able to match a pixel from one view of the scene to another view taken from a separate camera also known as the disparity value of this pixel.

In modern research, the most studied step of the algorithm is the process of pixel Correspondence, better known as stereo matching. As of now there are many optimization techniques being applied to stereo matching algorithms in order to increase their efficiency and accuracy. Firstly, this paper explains the math and logic behind each portion of a successful stereo vision system while providing implementations. finding value in optimization techniques when compared to more standard stereo matching approaches. In order to analyze and implement a sound stereo vision algorithm as well as a optimized matching algorithm scholarly sources were regarding stereo vision and optimization techniques for pixel correspondence. After implementing a standard matching algorithm as well as another using the optimization technique known as dynamic programming, I found that there was a significant increase in both accuracy and efficiency in the depth estimates provided from the algorithm.

§2 Triangulation

The core of every stereo vision algorithm is to find the depth of a pixel using multiple two dimensional views of the scene, more formally this process is known as the backwards projection of a camera from image coordinates into three dimensional world coordinates. In order to derive the formulas for the backwards projection of a camera, the formulas by which a camera uses in order to map world coordinates into image coordinates; this process can be explained as the forwards projection model.

§2.1 Forward Projection Model

Formally defined, the forward projection model "describes the mathematical relationship between the coordinates of a point in three-dimensional space and its projection onto the image plane of an ideal **pinhole camera**, where the camera aperture is described as a point and no lenses are used to focus light" (**wikipedia**). The usage of a pinhole camera allows for the elimination of lense distortion when mapping to the image plane, simplifying the formulas significantly.

> **Remark 2.1.** The majority of cameras used in stereo vision algorithms, including those used in this paper, use lenses contradicting the pinhole camera model. However, because researchers calibrate their camera's in order to remove distortion from the images returned, the pinhole camera model can still be applied.

The forward projection model of converting a 3D camera point into a 2D pixel coordinate is defined using the formula below:

Theorem 2.2 (Forward Projection Equation)

$$(u, v) = (f_x \cdot \frac{x_c}{z_c} + o_x, f_y \cdot \frac{y_c}{z_c} + o_y)$$

(u, v)	two dimensional pixel coordinates	f_x	focal length on x-axis
x_c	x position on scene coordinate frame	f_y	focal length on y-axis
y_c	y position in scene coordinate frame	o_x	image center on x-axis
z_c	depth of point in scene coordinate frame	o_y	image center on y-axis

Whereas the other paramteres of the equation are self-explanatory, the focal length (f) requires further explanation. The focal length is "the distance between the lens and the image sensor when the subject is in focus" (Nikon Website). As such, using this information the forward projection equations essentially show how a ray from the camera to the scene is mapped to an image.

§2.2 Derivation of Backwards Projection Model

It is clear that deriving the depth of a pixel from manipulating the forward projection equations is impossible given the inquality in depth measurements when using the x and y pixels. Therefore it is evident that additional information is needed in order to infer depth. This is where the usage of multiple viewpoints of a scene is needed.

> **Remark 2.3.** Although many stereo vision systems use more than two viewpoints of a scene, in order to simplify the implementation process a simple (binocular) stereo system will be used.

As mentioned prior the forward projection equations essentially represent the camera as projecting a ray from the image into a scene point. Using this fact, an additional camera which is calibrated to be on the same plane as the original camera may be used to project another ray from the corresponding image point onto the scene. By finding the intersection of these two rays the depth of a pixel in the scene may be found.

Using the depth measurement of a pixel, it is also possible to derive (x, y) coordinates of a scene from the image coordinate frame, giving the full scene coordinate frame:

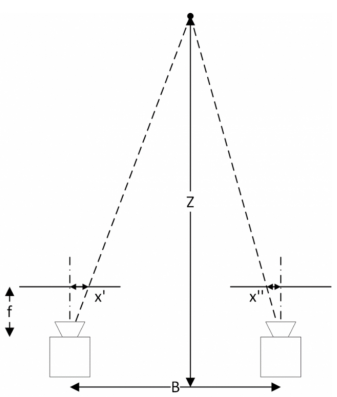

Figure 2: Simple stereo system

Theorem 2.4 (Backward Projection Equations)

$$z = \frac{b \cdot f_x}{(u_r - u_l)} \tag{1}$$
$$x = \frac{z}{f_x} \cdot (u_l - o_x) \tag{2}$$
$$y = \frac{z}{f_y} \cdot (v_l - o_y) \tag{3}$$

(u_l, v_l)	pixel coordinates on left camera	f_x	focal length on x-axis
(u_r, v_r)	pixel coordinates on right camera	f_y	focal length on y-axis
x	x position on scene coordinate frame	o_x	image center on x-axis
y	y position in scene coordinate frame	o_y	image center on y-axis
z	depth of point in scene coordinate frame	b	baseline distance

The two rays projected from the camera, along with the calibrated basline measurement (distance between left and right cameras) form a triangle, allowing for the derivations of the formulas shown above. However, in order to attain the parameters in these formulas that are not immediatly present in the image such as focal length, and the baseline as well as correcting for lense distortion camera calibration is required. Moreover, in order to make use of the ray projected from the additional viewpoint,

§3 Camera Calibration

§3.1 Intrinsic matrix

Remark 3.1. Notes

1. External Parameters: Position and Orientation of the camera with respect to the world coordinate frame...
2. Internal Parameters: How the camera maps world points into the image coordinate frame.
3. For the case of this program becuase of camera model was constructed to match an ideal model, we only need to make use of two variables the focal length and the distnce between the two cameras d.
4. Use this as a good guide for the process: https://medium.com/swlh/i-see-you-computer-vision-fundamentals-64cc662d0b05. Essentianlly use an mage of known geometry and rearange the original projection eqations to find both the fx anf fy
5. In many modern cameras, the lenses are contructed in such a manner that pixel quality is maximized, however the shape of these lense can lead to singificant distortion in the image taken when compared to the true real world position.
6. There consists of two types of distortion **radial and tangential**
7. In the paper you can describe the correction equation as well as what they algo:ref but in the actual Implementation you would have to use the built in function since a custom distorition function is way to complex , have a custom intrinsic and extrinsic matrix computation tool though.

§3.2 Extrinsic Parameters

§3.3 Lens Distortion

§3.3.1 Radial Distortion

§3.3.2 Tangential Distortion

§4 Stereo Rectification

§4.1 Window Based SSD Disparity Estimation

§4.2 Adaptible Window Optimizations

Remark 4.1. Notes

1. Now that the two cameras have been calibrated such that they are on a identical image plane wiht one another you can now find the disparity between pixels of the two images in order to find depth using the triangulation model.
2. One way this may be done is through a window based method, where we search for the object selected in one image the seocnd by creating a window and linearly searching for identical pixels on the second image. This method may be done efficiently due to camera calibration as the pixel range is on the same scan line.
3. We can explore other methods of stereo rectification if we have time as the computation is pretty simple. You would have the use a minimmum squared difference algorithm on the average intensity of the window.

§4.3 Dynamic Programming Based Disparity Estimation

§5 Implementation

§5.1 Hardware

§5.2 Program

§6 Testing

§7 Conclusion

§8 Works Cited

§9 Appendix