Depth Estimation with a Plenoptic Camera

Steven P. Carpenter
Auburn University, Auburn, AL, 36849

The plenoptic camera is a tool capable of recording significantly more data concerning a particular image than a conventional camera. This additional data is analyzed to produce several views of a particular phenomenon as though taken at different locations with a single exposure. The separate views are then compared using correlation algorithms on features within a scene to find a disparity map, which may be translated into a depth map using the appropriate inverse relation. This depth map is useful in analytical research with applications to structural vibrations and 3-D flow visualization, providing a means to determine the distance from the camera to a given feature and its fluctuation.

Nomenclature

\begin{align*}
C &= \text{Confidence value} \\
c &= \text{Correlation coefficient} \\
D &= \text{Distance to a plane conjugate to sensor plane} \\
d &= \text{Depth} \\
F &= \text{Focal length of lens} \\
f &= \text{Distance between lens and sensor plane} \\
h &= \text{Disparity of features} \\
I &= \text{Light intensity} \\
Im &= \text{Image} \\
P &= \text{Pixel} \\
r &= \text{Displacement of aperture} \\
u &= \text{Subpixel location in the } x \text{ direction} \\
v &= \text{Subpixel location in the } y \text{ direction} \\
x &= \text{Macropixel location in the } x \text{ direction} \\
y &= \text{Macropixel location in the } y \text{ direction}
\end{align*}

I. Introduction

The plenoptic camera is an imaging tool that records distinctive qualities of incoming light that a conventional camera cannot, allowing for simple experimental apparatuses for essential types of analysis. The camera is composed of a large main lens which focuses light onto an array of smaller lenses which project the light onto a sensor plane. The second layer of lenses provides a means to separate incoming light rays according to their incident angle, adding two additional dimensions to the camera output. In a conventional camera, this information is lost when the light intensity is summed at the sensor plane. The additional data allow for:

1) Computation of images from several perspectives with a single capture
2) Refocusing of an image during analysis
3) Depth estimation

A. Use for Depth Map Generation with a Plenoptic Camera

Innovation in the scientific and engineering community relies heavily on the analysis of empirically collected data. New theories must be thoroughly tested to be accepted, and old theories must continue to produce accurate predictions of results. The testing of old and new theories alike is done by experiment; in order to perform experiments, one must have the necessary tools.

Capturing an image of an experiment is common practice for the purposes of analysis and documentation, and therefore cameras are common experimental tools. Several simple and complex imaging techniques exist in the field

---

1Senior, Department of Aerospace Engineering, 371 W Glenn Ave, Apt 73, Auburn, AL, 36830

American Institute of Aeronautics and Astronautics
of aerospace engineering, such as shadowgraph and schlieren photography, and particle image velocimetry. These techniques are excellent for depicting density gradients in a fluid and for finding the instantaneous velocity field of a flow.

The plenoptic camera generates a representation of the lightfield emitted by the captured scene. By analyzing the lightfield, it is possible to generate separate views of a given scene as though they were taken by several cameras at different locations at the same instant in time. Further analysis of the data, such as that performed on stereo images, allows for the generation of a depth map of the scene. The depth map provides an additional useful means for the researcher to analyze three dimensional (3-D) flow and vibrations in 3-D structures. Prior to the invention of the plenoptic camera, similar results could be obtained from stereo images, in which two or more images taken at the same time are compared to find the disparity for a given feature between the two scenes. Disparity is inversely related to depth.

B. History of the Plenoptic Camera

The history of the plenoptic camera begins with the practice of integral photography, with the first production of 3-D images made by Lippman in 1908.1 His method consisted of capturing an image upon a plate using an array of lenses. Then, by placing a light source behind the plate, a viewer could see an image that appeared to change with viewing position. This technique was later improved by Ives, who added a larger lens in front of the array that could focus the image in front of, behind, or on the array of lenses. This greatly increased the depth of field of the produced image.2

The progress made by Lippmann, Ives, and many others, was used by Adelson and Wang to create the first single camera capable of generating a fully 3-D image, using a similar array of lenses and larger focus lens, as seen in Figure 1.3 This camera offered a vast improvement over other set-ups that produced similar results, such as binocular or trinocular stereo imaging. These techniques, which require two or three cameras respectively, are less effective, adding to the bulk and cost of the equipment and requiring extensive calibration.

II. Extracting Images from the Lightfield

In order to produce a depth map with a plenoptic camera, the camera output image must first be parsed into separate images.

A. Initial Data

The image generated by a plenoptic camera is shown in Figure 2. As seen, the image is divided into macropixels, which represent the microlens array, and subpixels, which represent the sensors laying behind one microlens. In order to generate separate images, it is useful to divide the initial 2-D image into an array representing the 4-D lightfield with the form \( I(x,y,u,v) \). To do this, the centers of the macropixels must be determined. This may be
accomplished by experiment, reducing the main lens aperture until the incoming light illuminates only the center of each microlens. The resulting dot matrix displayed in the generated image provides the location of the macropixel centers. This location is constant in all images generated by the camera for a particular configuration. The average distance, in pixels, between the centers is considered to be equivalent to twice the radius of each macropixel.

With the centers and radius of each macropixel determined, the array is created by sectioning the image into a number of squares equal to the number of macropixels. These squares are centered on and circumscribe each macropixel. This requires interpolation, as the coordinate location of each macropixel center is usually not a set of integers.

B. Determining Sub-Images

In each microlens, the angle of refraction for a given light ray depends upon its angle of incidence. If the lenses in the array are characteristically similar, each lens will refract incoming light from a certain location to the same location in its associated macropixel. Because the light sensors have a finite, non-zero size, each subpixel represents the small range of light rays that are integrated by its associated sensor.

Due to this correspondence, a conventional image may be constructed by taking a subpixel from a given location within each macropixel. Using the previously defined matrix $I$, an image can therefore be determined. For each $x$ and $y$ in $I$, while holding $u$ and $v$ constant, $I_m(x,y)$ is given by $I(x,y,u,v)$. Two images determined by this method are shown in Figure 3. Because subpixels were chosen with only a horizontal displacement, the resultant feature disparity is purely horizontal. It is therefore assumed a priori that pure horizontal displacement in viewing position results in horizontal disparities, and vertical displacements correspond to vertical disparities.

III. Depth From Image Pairs

When a viewer takes in a scene from different positions, the image that the viewer sees changes. For the general case of stereo images, objects that are near to the viewer displace more quickly than objects that are further away. This indicates an inverse relation between the depth and disparity of features. In the plenoptic camera, objects that are very near to the camera displace to the left as the viewing window moves to the right, objects that are at the focus plane of the camera do not displace, and objects that are further away displace to the right. This is visible in Figure 3: the left domino is close to the camera, and displaces to the left, the right domino is far from the camera, and displaces to the right. Adelson and Wang use the method of least triangles to determine the relation between depth and disparity:

$$\frac{1}{d} = \frac{h}{r} \left( \frac{1}{F} - \frac{1}{D} \right) + \frac{1}{D} \tag{1}$$

The depth, $d$, to a feature is therefore dependent upon several known, constant quantities and a variable disparity, $h$. As this is a one-to-one mapping over the useful set of possible values, the disparity map alone gives a good visual representation of depth without the application of the relation. In order to determine the depth of a certain feature in a pair of images, it is therefore necessary to find the distance that the feature displaces from one image to the other. Several techniques exist to perform this analysis. One of the most common is that which is performed on binocular stereo images. This technique is treated at length in Ref. 5.

A. Stereo Image Method

The traditional binocular stereo method for depth estimation requires two images taken from separate locations with cameras that are on the same plane with respect to the object plane and which have the same orientation. This requires exact calibration during set up, which is generally inflexible. The images produced by the plenoptic camera...
are always located on the same plane, and all sensors have the same orientation with respect to the object. The images produced by a plenoptic camera are therefore well suited to this method.

A cross correlation algorithm is used to estimate feature disparity between images. In an image $I_m$, a viewing window containing pixels $P$ and centered at $I_m(x,y)$ is selected to represent the feature at that location. That window is then compared to windows at a range of locations in the next image around $I_m(x,y)$. Because the images were derived from subpixels with purely horizontal displacement, the search range need only check for horizontal disparity, greatly reducing the time cost of the algorithm. Each comparison produces a correlation coefficient according to the Equation 2.6

$$C = \frac{\sum_{P} (I_m(x,y) - I_m(x+h,y)) (I_m(x,y) - I_m(x+h,y))}{\sqrt{\sum_{P} (I_m(x,y) - I_m(x,y))^2 + \sum_{P} (I_m(x,y) - I_m(x,y))^2}}$$ (2)

The disparity $h$ at $I_m(x,y)$ is then found by determining the value of $h$ that minimizes $c$. Additional accuracy is achieved by fitting a parabola to the data surrounding this value of $h$, and using the location of its minimum for the true value of $h$. This process is iterated for all locations in $I_m$ to generate a disparity map of the image. An example disparity map produced with this method is shown in Figure 4. In this figure, note that the domino on the left is generally black, indicating negative disparity and consequently an object near to the camera. The middle domino is at the focus plane and has a negligible disparity. The white domino is further away from the camera with a positive disparity.

C. Plenoptic Stereo Method

The traditional stereo method was developed in order to produce appreciable results for a depth map, but was limited by experimenter’s ability to capture several images. While capturing more than two images is possible, the requirement for additional equipment and calibration is highly prohibitive. While methods for trinocular stereo exist, the plenoptic camera generates as many images as it has subpixels in a given macropixel.

This method performs the regular stereo method for several image pairs. Holding $I_m$ constant, the viewing windows are compared to several images generated by other subpixels. In Figure 5, the three vertical dominos are in the same position as in Figure 4, and there is an additional horizontal domino, the front of which lays on the focus plane. To produce Figure 5, $I_m$ was compared to four other images taken from subpixels with pure horizontal displacements from the original. The resulting disparities were then assigned weights corresponding to their correlation and summed. The visual noise using this method at object level is significantly less than that present when using the traditional stereo method. The disparity map for the horizontal domino demonstrates the capability of the plenoptic stereo method to process objects that have variable depth values.

Figure 4: Binocular Stereo Disparity Map. a) Ground truth, b)Disparity Map. The black to white scale indicates negative to positive disparity. The middle domino lays at the focus plane, with approximately no disparity.

Figure 5: Plenoptic Stereo Disparity Map. a) Ground truth, b)Disparity Map. The black to white scale indicates negative to positive disparity. The middle domino lays at the focus plane, with approximately no disparity.
D. Least Squares Gradient

This method, used by Adelson and Wang\(^3,8\), uses gradients in a given image and in successive images to calculate the disparity for a feature at location \(\text{Im}_1(x,y)\). For this method, let \(\text{Im}_2\), \(\text{Im}_3\), \(\text{Im}_4\), and \(\text{Im}_5\) stand for the images generated by the subpixels to the left, right, top, and bottom of the original subpixel. A viewing window consisting of pixels \(P\) is selected for all locations in \(\text{Im}_1\). The feature disparity at any location in \(\text{Im}_1\) is given by Equation 3.

\[
h(x, y) = \frac{\sum p(x, y)}{\sum (x^2 + y^2)}
\]

In this, \(I_x\) and \(I_y\) stand for the gradient in \(\text{Im}_1\) at \((x,y)\) in the x and y directions respectively. \(I_{vx}\) is the gradient in horizontal view shift, given by finding the slope in \(I\) from \(\text{Im}_2(x,y)\) and \(\text{Im}_3(x,y)\). Similarly, \(I_{vy}\) is the gradient in vertical view shift, given by finding the slope in \(I\) from \(\text{Im}_4(x,y)\) to \(\text{Im}_5(x,y)\). Figure 6 demonstrates the resultant disparity map from this method. This method is excellent for determining which regions are not useful by assigning confidence measures, but the resulting depth map lacks the clarity of the plenoptic stereo method.

\[\text{Figure 6: Least Square Gradient Disparity Map. a) Ground truth, b) Disparity Map. Regions of low confidence are displayed as black. There is very little discrimination between domino disparity levels.}\]

IV. Problems and Future Work

There are inherent limitations in the algorithms used to generate disparity maps, and also in the plenoptic camera

A. Correspondence Problem

In order to calculate disparity, each algorithm used a viewing window in a particular image that was compared to similar windows in other images. This functions well in scenes without occlusions. However, when one object lays behind another in the viewing window, part of the original matching window will no longer be visible in the comparison images. When using the stereo method, this can lead to minimized correlation at incorrect disparities around the occlusion boundary. This can be seen in Figure 7, in which the right domino is occluded by the central domino, and the depth of the area at the occlusion boundary is incorrect. When using the gradient method, this will produce a low measure of confidence. If the image has occluded objects, the effect of the correspondence problem is minimized by using only images from subpixels that are adjacent to one another, limiting the amount of the object that is hidden from view.\(^9\)

\[\text{Figure 7: Correspondence Problem. When part of the image is hidden, the correlation fails to recognize that the object it is searching for does not exist. This confuses the results at that location.}\]

B. Low Texture Areas

Areas with very few features have little data with which the cross correlation can discriminate. In this case, the area surrounding a location will contain several very good matches, and therefore the minimized correlation coefficient can be chosen at the incorrect disparity. For the gradient method, this results in a low confidence measure for the location. This problem is shown in the above figures in the noise of the background of each image.\(^{10}\)

C. Blurring

In the process of generating images from subpixels, it was assumed that the each pixel only received light from a single microlens. At the edges of the macropixel, however, this is not necessarily the case. Due to the initial step of
averaging to find the location and radius of each macropixel, and interpolating to find the intensity at each subpixel location, the subpixels at the boundaries of each macropixel may contain data from more than one microlens. This problem is avoided in this study by reducing the radius considered by one pixel.  

D. Resolution

Plenoptic cameras require that the sensors be divided among a number of microlenses. Because of this, the spatial resolution in each dimension must be limited compared to that available to a conventional camera with a similar number of pixels. In effect, the spatial resolution in each dimension is reduced by a factor of twice the radius of a macropixel. High fidelity depth maps require many subpixels, however increasing the number of subpixels reduces the total image resolution.

E. Future Work

The current aim for improvement to the algorithm is to reduce the background noise in the result. This will be accomplished by implementing a bandwidth filter to the input data, reducing the impact of regions with very little or no texture. After this, the algorithm will be expanded to handle stereo matching in two dimensions, with the ultimate aim of using all input data in the final weighted disparity map.

The scope of the project is to provide a tool capable of generating a high fidelity depth map of any scene. Such an instrument would find use in the aerospace industry in 3-D visualization and in measuring structural deformation. If needed, the plenoptic camera could be used to capture video, as it essentially the same mechanically as a conventional camera, and the depth information could be used to produce a highly accurate model of a fluctuating flow or wing in that unsteady flow.

V. Conclusion

The plenoptic camera is a tool that has an abundance of applications in research. Depth estimation is one particularly useful capability that the camera possesses. The plenoptic camera performs this function with less need for equipment calibration and also performs the task with better results than the traditional method of stereo images, the most known alternative. In order to estimate depth with this camera, the camera output may be interpreted as a set of sub-images. To produce an accurate depth map, one needs to calculate the disparity between features in image pairs. In the future, the algorithm described will be improved by dealing with background noise and by improving the correlation algorithm that generates the disparity map. The completed algorithm may be used to visualize 3-D flow and structural vibration.

References