Light Field Assisted Stereo Matching using Depth from Focus and Image-Guided Cost-Volume Filtering

Jędrzej Kowalczuk, Eric T. Psota, and Lance C. Pérez
Department of Electrical Engineering, University of Nebraska, Lincoln, NE, U.S.A.

Abstract—Light field photography advances upon current digital imaging technology by making it possible to adjust focus after capturing a photograph. This capability is enabled by an array of microlenses mounted above the image sensor, allowing the camera to simultaneously capture both light intensity and approximate angle of incidence. The ability to adjust focus after capture makes light field photography well-suited for computer vision techniques that aim to determine the depth of objects in a scene, such as depth from focus/defocus. Another commonly known method for extracting depth from images is stereo matching, which seeks to obtain a disparity map linking corresponding objects that appear in both images. This disparity map, along with the known geometry of the cameras can be used to compute the depth of objects in the scene. Whereas depth from focus/defocus using traditional cameras requires multiple sequential image captures, light field cameras capture these images simultaneously, making them more suitable for use in conjunction with stereo matching. A method is presented that combines both light field photography and stereo matching to achieve an enhanced disparity map. This joint approach takes advantage of the unique strengths of both methods to produce a result that is more accurate than either method produces in isolation.

Keywords: Stereo matching, stereo correspondence, light field photography, depth from focus.

1. Introduction

Over the last three decades, computer vision researchers have developed a variety of methods for computing the three-dimensional structure of a scene. These methods can generally be divided into active and passive methods. Active methods typically require light to be projected onto the scene, and include laser scanning [1], structured light [2], [3], and infrared range sensing [4]. While these methods produce accurate depth maps of the scene, they require additional hardware and are often restricted to small-scale indoor environments. In contrast, passive methods compute scene depths without requiring the projection of light, thus making them suitable for a broad range of applications. Commonly known passive methods include stereo matching [5], structure from motion [6], and depth from focus/defocus [7], [8].

The goal of stereo matching is to find correspondences between objects that appear in two images of the scene. By finding positional offsets, i.e., disparities, relating objects that appear in the two images, it is possible to use the known geometry of the stereo camera configuration to compute the depth and scale of these objects. Unfortunately, the accuracy of stereo matching suffers from several inherent difficulties. These difficulties include low-texture areas, depth discontinuities, occlusions, slanted surfaces, repeating patterns, reflections, and specularities. Global stereo matching methods, such as graph-cuts and image segmentation [9], have significantly reduced the detrimental effects of low-texture surfaces, often by assuming they are part of one large, flat surface. The introduction of adaptive support weights for stereo matching cost aggregation [10] addresses the challenge posed by depth discontinuities, and has been adopted by many recently proposed stereo matching algorithms [11], [12]. Occlusions, on the other hand, require algorithms to effectively “guess” the disparity values, since occluded surfaces are only visible in one of the two images.

Techniques that extract depth using focal plane characteristics can be divided into two sub-categories: depth from focus and depth from defocus [8]. A depth from focus (DFF) system adjusts the camera’s focus settings, or the distance of the focal plane, until the point of interest comes into focus. Given that there is a one-to-one relationship between the depth of an object point and the depth of the focal plane that puts that point in focus, it is possible to use depth from focus to obtain the depth of every point in the scene. In contrast, depth from defocus (DFD) operates by comparing a reference image obtained with a small aperture to images obtained with a large aperture. Depth is then approximated by determining the amount of blur (defocus) that must be applied to the reference image to reproduce portions of images with large aperture. The amount of blur, or the size of the blur kernel, corresponds to a particular depth value and larger apertures generally produce more precise depth measurements.

An advantage of depth from focus/defocus is that these methods do not suffer from many of the difficulties associated with stereo matching. For example, a repeating pattern will not confuse DFF or DFD, as they operate locally and do not require a scan-line search. Until recently, DFF and DFD were largely limited to post-processing of images of static scenes, where a set of focus-adjusted images is obtained in sequence. With the availability of light field cameras produced by Lytro (Lytro Inc., Mountain View, CA), it has become possible to obtain a large set of images with different focus settings simultaneously [13]. Light field cameras effectively capture both light intensity and angle of incidence on the same image sensor using an array of microlenses.

In this paper, a method is proposed that uses a stereo configuration of light field cameras. The method begins by
capturing a pair of light field photographs, performing depth estimation using DFF, and constructing a pair of images, such that every point in the images is in focus. These images are then processed using stereo matching with image-guided filtering, and the reliability of stereo matching is assessed using a back-and-forth consistency check. Finally, DFF and stereo matching costs are combined using a weighted summation, and the Winner-Takes-All decision criteria is used to find minimum-cost matches and assign disparities. It is shown that this approach improves upon the accuracy of the disparity map produced using either stereo matching or DFF in isolation.

2. Background

Stereo matching is formulated as the process of computing a disparity map that defines correspondences between pixels in a pair of stereo images. Stereo matching methods are commonly built on top of a Winner-Takes-All (WTA) framework, which operates on a set of precomputed per-pixel matching costs. For a particular pixel of interest in the reference image the WTA approach selects the candidate match characterized by the minimum matching cost. Typically, these costs are aggregated to form a stereo matching cost volume \( C_{SM}(x, y, d) \) holds the per-pixel dissimilarity metric, e.g., the sum of absolute color differences or the sum of squared color differences, evaluated between pixel \( p \) at location \((x, y)\) in the reference image and pixel \( \bar{p} \) at location \((x - d, y)\) in the target image. The disparities associated with minimum-cost matches are recovered by finding disparity arguments that minimize \( C_{SM}(x, y, d) \). For a pixel \( p \) in the reference image, the disparity is found using the decision criteria given by

\[
d_p = \arg\min_d C_{SM}(x, y, d) . \tag{1}
\]

To enforce local consistency of matches and reduce the noise present in the resulting disparity maps, window-based cost aggregation techniques have been introduced into the WTA framework. Among these methods, adaptive support-weight cost aggregation presented by Yoon and Kweon [14] has been shown to produce highly accurate disparity maps and aggregate matching cost over a larger window size, preventing the application of larger filter kernels.

The weights of the guided filter in a support window \( \Omega_p \) are given by

\[
w(p, q) = \frac{1}{|\Omega_p|} \sum_{q \in \Omega_p} \left( \frac{I(p) - \mu_{\Omega_p}}{\sigma_{\Omega_p}^2 + \epsilon} \right)^2 , \tag{4}
\]

where \(|\Omega_p|\) is the number of pixels in the support window \( \Omega_p \), \( I \) is the guidance image (here, a grayscale version of the reference image), \( \mu_{\Omega_p} \) and \( \sigma_{\Omega_p}^2 \) are the mean and the variance of the guidance image in the support window, and \( \epsilon \) is a regularization parameter used to control the strength of smoothing. Once the weights are computed, the cost volume is updated as

\[
C_{SM}'(x, y, d) = \sum_{q \in \Omega_p} w(p, q) C_{SM}(x, y, d) , \tag{5}
\]

and matches are selected according to (1).

2.1 Stereo Matching using Cost-Volume Filtering

In fact, the adaptive support-weight cost aggregation is equivalent to smoothing the initial cost volume (layer by layer) using a bilateral filter. The edge-preserving property of the bilateral filter captures object boundaries in the disparity maps, however, the complexity of filtering grows exponentially with increasing window size, preventing the application of larger filter kernels.

To address this, Rheemann et al. [19], proposed cost-volume filtering using a guided image filter [20] that closely mimics the behavior of the bilateral filter, yet can be computed in constant time, regardless of the window size. Apart from fast operation, the ability to aggregate matching cost over a larger window results in exceptionally accurate disparity maps, making their approach the best-performing local stereo matching algorithm among those listed on the Middlebury stereo benchmark [5], [2], approaching the accuracy of algorithms that perform global energy minimization.

2.2 Light Field Photography

In February of 2012, Lytro released the first consumer-level light field camera. The primary advantage for photography enthusiasts is the ability to adjust focus after capturing the photo. Figure 1 illustrates a simplified model of the Lytro camera. Much like ordinary cameras, which capture the intensity of light on an image sensor, the Lytro camera incorporates a lens assembly that allows for zooming and focusing. However, unlike ordinary cameras, the Lytro camera uses a fixed aperture of \( f/2.0 \) and light passes through an array of microlenses that allow the camera to separate light captured from a variety of
different angles before making contact with the image sensor. In effect, the Lytro camera captures both the light’s intensity and its angle of incidence. By post processing this data it is possible to mimic movement of the image sensor in relation to the lens assembly, thus reproducing the effects of a standard focus adjustment.

![Diagram of image capture using Lytro’s light field camera.](image)

This technology lends itself well to the depth from focus and depth from defocus research areas which, prior to the introduction of light field cameras, required the acquisition of multiple images of a static environment captured using different focus settings. After capturing a photograph with Lytro’s light field camera, the user can extract multiple images of the scene with different focus settings. Hence, to extract depth information at a particular point in the scene, it is possible to scan through the image set in search of the one that provides the highest level of sharpness, and then interpolate the depth from that image.

### 3. Light Field Assisted Stereo Matching

Both stereo matching and depth from focus provide the capability to produce depth maps of the observed scene. While image blur is a key to the successful operation of DFF, it is undesirable for stereo matching. Thus, the first stage of the proposed method involves creating all-in-focus images from the photographs captured by the light field cameras. Figure 2 shows three of the twelve images of a test scene obtained with both the left and right light field cameras. In the top row of Figure 2, the background is in focus, and the second and third rows correspond to focal planes moving toward the camera.

To create an all-in-focus image $I_{\text{Sharp}}$, the sharpness of each pixel is evaluated for each of the twelve images $I_1, \ldots, I_{12}$ obtained from the light field photograph. Sharpness in an image can be quantified by integrating the amount of high frequency content surrounding a pixel. To create a high frequency image, a low-pass filtered image is first generated using $I_{\text{LPF},k} = I_k \ast H_1$, where $H_1$ is a $5 \times 5$ Gaussian filter with standard deviation $\sigma = 2$. Then, the low-pass filtered image is subtracted from the original image to create a high-pass filtered image $I_{\text{HPF},k} = |I_k - I_{\text{LPF},k}|$. The amount of high frequency content surrounding a pixel is aggregated using a second stage of low-pass filtering, and the resulting cost for evaluating the high frequency content in image $k$ is given by

$$C_{\text{HPF},k} = I_{\text{HPF},k} \ast H_2,$$

where $H_2$ is an $11 \times 11$ Gaussian filter with standard deviation $\sigma = 5$. Once the cost are computed for each of the 12 images, the sharpest image for each pixel is determined using

$$I_{\text{Sharp}}(x, y) = I_k(x, y) \mid \arg \max_{j=1,\ldots,12} C_{\text{HPF},j}(x, y) = k,$$

for all pixel locations $(x, y)$ within the image.
Fig. 3: A grayscale map of indices of the sharpest images calculated for each pixel (top row) and all-in-focus images obtained by combining sharp pixels from the image sets together (bottom row).

Figure 3 provides a grayscale visualization of the sharpest images chosen for each pixel in the images, and the resulting image $I_{\text{sharp}}$ created using both the left and right light field captures. This pair of all-in-focus images is now suitable for processing using stereo matching.

### 3.1 Depth-from-Focus Cost Volume

The cost volume $C_{\text{HPF}}$ computed in Equation (6) can be used to approximate the disparity between the left and right images after assigning a unique disparity to each of the images extracted from the light field photograph. However, the range of stereo matching disparity values is much larger than 12 between the $1024 \times 1024$ images given in Figure 3. In fact, the range of disparities between these images is approximately 20-100. Thus, it is necessary to interpolate the cost volume $C_{\text{HPF}}$ to achieve the same level of precision as stereo matching for this image set, resulting in a high-precision cost volume that will be denoted $C_{\text{DFF}}$. Figure 4(c) shows a set of high-frequency costs evaluated for the 12 images extracted from the light field capture, labeled as “DFF:Measurements”. An illustration of the interpolated cost is also given in Figure 4(c), producing a minimum cost disparity of approximately $d = 41$.

### 3.2 Cost Volume Merging and Joint Disparity Selection

Using the image-guided cost-volume filtering approach given in [19], the cost of matching all-in-focus images can be evaluated over the entire range of disparity hypotheses. Figure 4 illustrates the matching cost between the reference image patch in Figure 4(a) and its corresponding search space given in Figure 4(b). Note that, while the correct disparity is $d = 44$ for this particular pixel, the cost of choosing $d = 81$ is lower, and would be chosen using a pure image-guided stereo matching algorithm. Also given in Figure 4(c) is the sampled cost volume $C_{\text{HPF}}$ and the interpolated cost volume $C_{\text{DFF}}$. While pure stereo matching fails to select the correct disparity due to the repeating pattern in the target search space, the interpolated cost $C_{\text{DFF}}$ clearly favors the disparity $d = 44$ over $d = 81$. It is worth noting that the precision of stereo
matching is much higher than that achieved using depth from focus, thus creating a tradeoff between the two methods. A joint cost volume $C_{DFF+SM}$ is created by merging together the cost volume $C_{DFF}$ and the stereo matching cost volume $C_{SM}$ using a weighted summation. Figure 4(d) shows the joint cost equal to $C_{DFF+SM} = 0.05 \times C_{DFF} + 20 \times C_{SM}$. In this example, the stereo matching ambiguity can be correctly resolved using WTA after incorporating information from depth from focus.

Stereo matching errors can often be detected using a back-and-forth consistency check operation. If the match chosen for pixel $p$ in the reference image is $\bar{p}$, yet the match chosen for pixel $\bar{p}$ in the target image is not $p$, the disparity assigned to pixel $p$ is deemed inconsistent. The most common cause of inconsistencies in a well-calibrated stereo image pair is the presence of occlusions, where an area in the scene is only visible in only one of the images. Low-contrast areas and specularities can also cause inconsistent matching. In order to take advantage of consistency checking for error detection, the cost volume summation can be modified to favor the DFF matching costs for pixel locations where stereo matching is inconsistent.

4. Results

The stereo matching results given in Figures 5(a) and 5(b) were generated using the image-guided cost-volume filtering approach described in Section 2.1, with a support window $\Omega_p$ of size $33 \times 33$. Figure 5(a) shows the disparity map obtained for the left image using the Winner-Takes-All match selection criteria, and Figure 5(b) highlights inconsistent disparities where the mapping between pixels is not bijective. Figure 5(c) shows the corresponding disparity map extracted from the depth-from-focus cost volume $C_{DFF}$.

Figure 5(d) shows a disparity map produced using the proposed joint method incorporating depth from focus and image-guided stereo matching. The joint matching cost volume at every pixel is computed using

$$C_{DFF+SM} = \begin{cases} 
0.05 \times C_{DFF} + 20 \times C_{SM} & \text{if } p \text{ is consistent} \\
0.075 \times C_{DFF} + 20 \times C_{SM} & \text{otherwise}
\end{cases},$$

where DFF matching costs are favored in cases where stereo matching produces inconsistent disparities.

Conventional stereo matching fails in the two regions highlighted within Figure 5(a), among which the region marked with a green border illustrates errors caused by repeating patterns and the region marked with a blue border illustrates errors caused by occlusions. Within these two regions, the benefits of applying the joint method are apparent, in that the joint method successfully resolves the disparities associated with the repeating metallic mesh and correctly handles the occlusions around the pens and markers.

5. Conclusion

A method is given for using stereo images obtained via light field photography to enhance the accuracy of the disparity map obtained using conventional stereo matching. The precision of stereo matching and the accuracy of depth from focus are combined by computing a weighted sum of their respective
cost volumes over the entire disparity space. It is shown that many of the common difficulties associated with stereo matching, including repeated patterns and occlusions, can be resolved by using interpolated depth-from-focus information obtained from images extracted from the light field photograph. While light field photography is still in its infancy, the results given in this paper demonstrate the potential gains that can be achieved using this technology.

Applying depth from focus to light field photographs produces reliable depth for objects close to the camera, however, for objects that are far away from the camera the reliability quickly diminishes. This limitation, imposed by the size of the aperture is similar to the limitations of stereo matching caused by the distance between the two cameras. Therefore, future research and applications involving the integration of light field photography and stereo matching should consider this tradeoff in order to design passive systems that capture accurate depth information over a wide range.

References