Noise Robust Depth from Focus using a Ring Difference Filter

Jaeheung Surh, Hae-Gon Jeon, Yunwon Park, Sunghoon Im, Hyowon Ha, In So Kweon
Robotics and Computer Vision Lab., KAIST
{jhsurh, hgjeon, ywpark, shim, hwha, iskweon}@rcv.kaist.ac.kr

Abstract

Depth from focus (DfF) is a method of estimating depth of a scene by using the information acquired through the change of the focus of a camera. Within the framework of DfF, the focus measure (FM) forms the foundation on which the accuracy of the output is determined. With the result from the FM, the role of a DfF pipeline is to determine and recalculate unreliable measurements while enhancing those that are reliable. In this paper, we propose a new FM that more accurately and robustly measures focus, which we call the "ring difference filter" (RDF). FMs can usually be categorized as confident local methods or noise robust non-local methods. RDF's unique ring-and-disk structure allows it to have the advantageous sides of both local and non-local FMs. We then describe an efficient pipeline that utilizes the properties that the RDF brings. Our method is able to reproduce results that are on par with or even better than those of the state-of-the-art, while spending less time in computation.

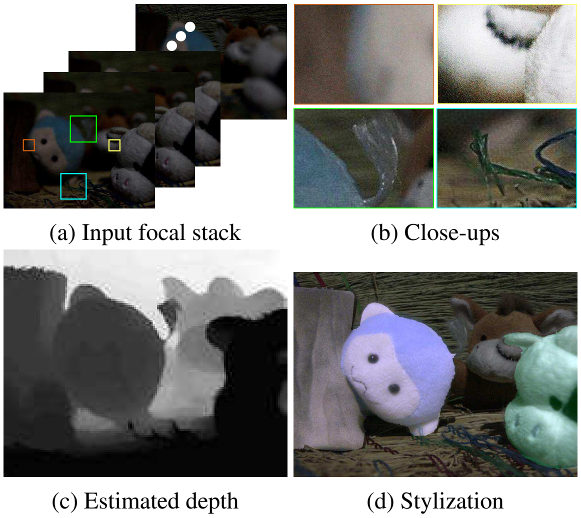

(a) Input focal stack
(b) Close-ups
(c) Estimated depth
(d) Stylization

Figure 1. Given a series of images for a scene with different focus settings, or the focal stack (a), we estimate an accurate depth map (c). Note the high noise levels and fine structures in the close-up (b) of the focal stack. These results can then be used for image processing applications, like stylization (d).

1. Introduction

As the computing horsepower of hand-held devices grows, interests into the capture of depth information for these electronics have been increasing. The accurate computation of scene depth forms the base for a wide variety of highly sought-after applications, like synthetic focus, 3D parallax, and augmented or virtual reality (AR/VR). In order to address these needs, many works of research have pursued different approaches for acquiring a scene's depth, while minimizing the overhead involved in such tasks. However, following the current interests in AR/VR on mobile devices, we chose to focus on methods that may be viable for that platform.

One method that achieves this is through the use of active sensing devices. Numerous products, such as Google's Project Tango [1], Occpital's Structure Sensor [5], or Microsoft's Kinect [4], and research [21, 14] have utilized this method for depth measurements. This approach revolves around the projection of light with a known pattern to determine its deformation as it returns to the system. The depth is then calculated with the assumption that the pattern's deformation is dependent only on the structure of the scene. However, such a system often requires expensive or specialized equipment to work, depends on the accuracy of its calibration, and fails outside of environments with controlled lighting conditions. Another group of approaches for depth acquisition is the use of light-field [32, 3, 42, 20] or stereo systems [2, 6]. These setups utilize multiple cameras or complex systems that simulate the capture of multiple cameras to estimate depth by finding correspondences in the pixels of images with a certain amount of baseline. Despite the abundance of works that follow these types of procedures, they are often heavily dependent on the accurate retrieval of correspondences and are often burdensome to set up and/or calibrate. Structure from small motion [50, 19, 15] is another avenue that tries to obtain the depth information of a scene. It consists of a single camera taking multiple images from various locations to determine

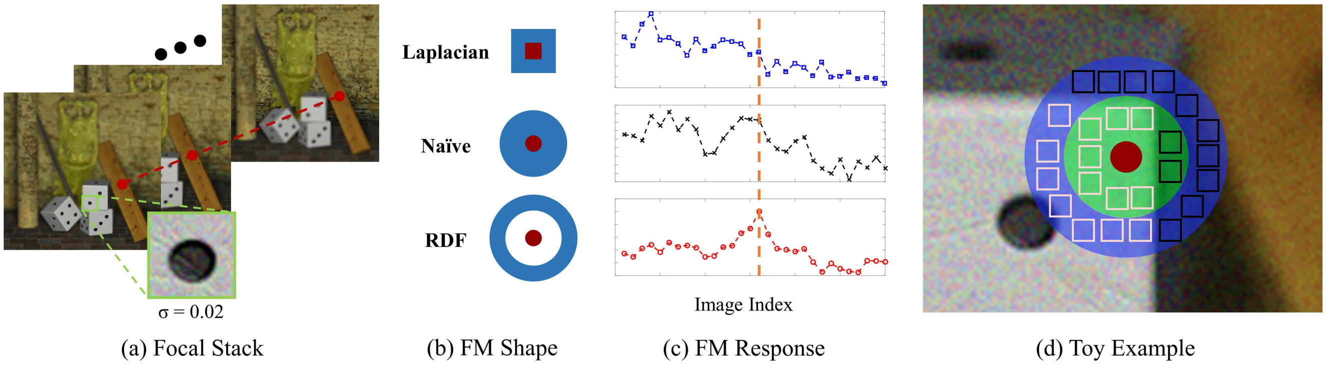


Figure 2. Comparison of different focus measures for depth from focus. (a) Input noisy focal stack (noise $\sigma = 0.02$). (b) Shape of focus measures from top to bottom: Laplacian, naïve, and RDF. (c) Focus measure response results of the three measures at the location marked by the red dots and dashed line in (a). Note that the correct image index is shown by the vertical dashed orange line. (d) Toy example of the difference between the naïve approach and RDF. White boxes are patches that are similar to the POI and black are dissimilar.

the location of corresponding points within each image. The caveat with this approach is that it cannot account for the focus changes that auto-focus on a mobile phone inevitably presents.

These shortcomings, when it comes to the problem of calculating the depth from a mobile phone camera, leads us to examine depth from focus (DfF) as an alternative. DfF takes in a focal stack, Fig.1(a), to output a depth map, Fig.1(c), simply by using the focus change. In this paper, we propose a new measure that more robustly and accurately determines the focus within the DfF framework and a pipeline through which the depth of a scene can be retrieved efficiently. Our major contribution lies in the proposal of a new focus measure (FM) that is more robust to noise than previously benchmarked measures. We call the filter that computes this measure the ring difference filter (RDF). To demonstrate the robustness of our FM, we qualitatively and quantitatively compare our results to that of the top-ranked FMs examined in [31], using a synthetic light-field dataset. To demonstrate our pipeline's effectiveness, we evaluate the proposed FM through a set of real-world evaluations and a comparison with a state-of-the-art approach [40]. Some applications, like synthetic focus and stylization (Fig.1(d)), of our work are shown.

2. Related Works

The focus measure (FM) is the measure of how in focus a pixel is on a given image. The extent to which a pixel is focused is described by the inner workings of the camera lens setup modeled by the thin-lens model. For any given point in the scene, the lens focuses the light on the focus plane. The scene point is then presented as a focused pixel on the image if the imaging sensor is located on the focus plane during the capture. The FM determines how close the focus plane and the imaging sensor are, and the depth can be recovered by using this measurement.

A variety of FMs have been proposed and form the foundation for a wide range of image processing and computer vision tasks including, but not limited to, depth from focus (DfF) [40], edge detection [24], and autofocus [29]. We referred to [31] for the different groups of FMs that are most commonly used. Gradient-based measures [7, 11, 9, 25, 13] leverage the values obtained from the first derivative of the image with the assumption that focused images contain sharper edges. With the same assumption, Laplacian-based measures (LAP) [8, 41, 28] utilize the second derivative of the image, instead. Wavelet-based measures (WAV) [18, 46, 47] use the frequency and spatial domain information provided by the discrete wavelet transform of the image, while discrete cosine transform-based measures [23, 37, 22] solely exploit the frequency domain information provided by the discrete wavelet transform of the image. Some employ statistical measures to extrapolate the degree of focus. These measures are named statistics-based measures [34, 39, 12, 49, 30]. Then, there are the ones that do not belong to any of the other group, which is miscellaneous measures [26, 16, 27, 38].

In this paper, we propose the use of a filter called the "ring difference filter" (RDF) as the focus measure. The filter maintains high robustness and confidence by incorporating both local and non-local characteristics. This is done by utilizing information in a relatively large window of neighboring pixels and strategically placing a gap space to ignore certain regions of the window.

3. Ring Difference Filter

We propose an FM that is both robust to noise and confident in its measurement, which we call the ring difference filter (RDF). The structure of RDF can be seen in the last row of Fig.2(b). It is a unique combination consisting of a ring and a disk. The disk, marked in red, focuses on the pixel of interest (POI) and the ring, marked in blue, surrounds the disk. For all FMs depicted in the figure, the red regions are negative weights that add up to -1, while the blue are positive and add up to 1. We call the pixels within the disk area as the region of interest (ROI), those within the ring as the ring pixels, and those between the disk and the ring as the gap pixels. Formally, given that $\mathbf{x_0}$ is the position of the POI,

$$\mathcal{RDF} = \begin{cases} -\frac{1}{\pi r_1^2} & |\mathbf{x_0} - \mathbf{x}| \leq r_1 \\ \frac{1}{\pi(r_3^2 - r_2^2)} & r_2 < |\mathbf{x_0} - \mathbf{x}| \leq r_3 \ , \\ 0 & otherwise \end{cases} \quad (1)$$

where $\mathbf{x}$ is the pixel position, r_1 is the ROI radius, and r_2 and r_3 are the inner and outer radii of the ring, respectively. RDF measures the focus of the POI by finding the difference between the average value of the ROI and the average value of the ring pixels, while ignoring the gap pixels.

3.1. RDF Structure Rationale

In order to explain RDF as being a noise robust focus measure, we must first examine the classic Laplacian filter, illustrated in the first row of Fig.2(b). The filter tries to estimate the discrete second derivative of the image at the POI. This is achieved by comparing the value at the POI to the 8 pixels around it. However, if the image were to contain noise, like in Fig.2(a)[1], the error due to this noise would be greater than if more pixels are compared, like with the other two filters in Fig.2(b). By sampling more pixels, the effects of noise is distributed amongst a larger set of sampled points, suppressing its effects.

However, naïvely increasing the sampled points to within a given radius, like the second filter in Fig.2(b), is not ideal. It can be safely assumed that pixels nearby the POI contain similar values and those that are farther away are more likely to be different. Similar pixel values from nearby regions are redundant and only decreases the confidence of the measurement since the gradient is maximized when the pixel is in focus.

To counteract the noise within the image while maintaining confidence, we propose a gap space as shown in the last filter in Fig.2(b). The toy example in Fig.2(d) explains how this is true. It stands to reason that when looking nearby the POI, which is marked in red, those pixels will resemble the POI. This is why over half the area for comparison that the naïve filter looks at, marked in green, consists of the same information, even though the POI is at the edge. By placing a gap and looking further away, almost 3/4 the area that RDF looks at, marked in blue, consists of dissimilar pixels, allowing RDF to more confidently determine edges than the naïve approach does. The confidence of RDF's results and how it synergizes with refinement methods will be dealt with in more detail in Sec.3.2.

	Winner Margin (scale: $\times 10^{-3}$)			
Noise level σ	0	0.005	0.01	0.02
Laplacian	0.137	0.002	-0.228	-0.212
Naive	0.227	0.424	0.552	0.674
RDF	**0.321**	**0.483**	**0.584**	**0.697**

	Curvature (scale: $\times 10^{-5}$)			
Noise level σ	0	0.005	0.01	0.02
Laplacian	0.881	-0.287	-2.354	-2.848
Naive	1.880	4.372	5.998	7.518
RDF	**2.599**	**4.779**	**6.355**	**7.999**

Table 1. Confidence measure of each focus measure.

These robust and confident characteristics are shown in the filter response results of the mentioned filters. An example of the filter response for the marked location of a noisy input focal stack is shown in Fig.2(c). The focal stack consists of 30 images of different focus and the location marked in red (Fig.2(a)) is in focus at the 16th image. As shown in the figure, the Laplacian and naïve filter are unable to discern the true label due to their lack of robustness and confidence, respectively. RDF, however, outputs a more robust and confident output, shown by its correct prediction and sharpness in its results, respectively.

3.2. RDF Confidence Analysis

In this section, we numerically verify the confidence of RDF and the advantage it brings. We used the 7 light-field datasets of resolution 768×768 or higher, provided by [44]. We quantized the depth map into 30 equally spaced depth labels and generate a set of depth-dependently blurred images for each label. We then applied different levels of signal dependent Gaussian noise to simulate realistic camera noise.

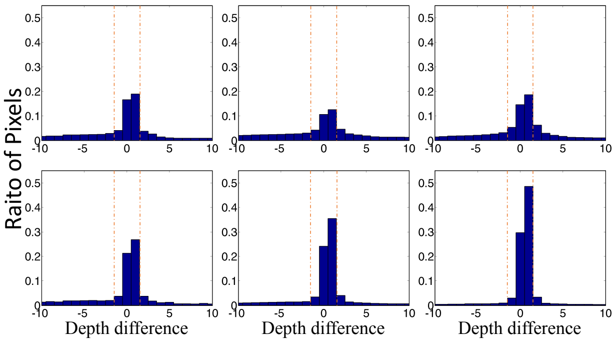
As a demonstration of the confidence of RDF, we employed a confidence measure (CM) proposed in [35]. The CM, named Winner Margin (WM) is defined as follows:

$$WM = \frac{c_1 - c_{2m}}{\sum_l c(l)}, \quad (2)$$

where c_1 and c_{2m} are the maximum response value and the second local maximum response value of the FM, respectively, l is the image index, and $c(l)$ is the FM response at image index l. To put it in simpler terms, the measure determines how much the global maximum peak "wins" against the second maximum peak.

To show how high the confidence is when the FM is correct against how low it is when incorrect, we subtract the mean confidence of the incorrect pixels from the mean confidence of the correct pixels. The results of this CM

[1] To simulate realistic camera noise, we applied a signal-dependent Gaussian noise with a standard deviation of σ to the clean images as done in [36] for all synthetic datasets in this paper. Unless mentioned otherwise, noise is set to $\sigma = 0.02$.



(a) Laplacian (b) Naive (c) RDF

Figure 3. Depth difference histogram. (Top) Histogram of depth label error by initial focus measure response. (Bottom) Histogram of depth label error after using aggregation on the initial response. The vertical lines denote the margin of correctness.

for different FM responses with datasets that have varying noise levels, σ, are shown in Table.1. Each FM generates 30 responses for each of the 4276224 pixel locations of the datasets. As one can see, RDF is able to maintain high confidence for correct labels and low confidence for incorrect labels, compared to other FMs.

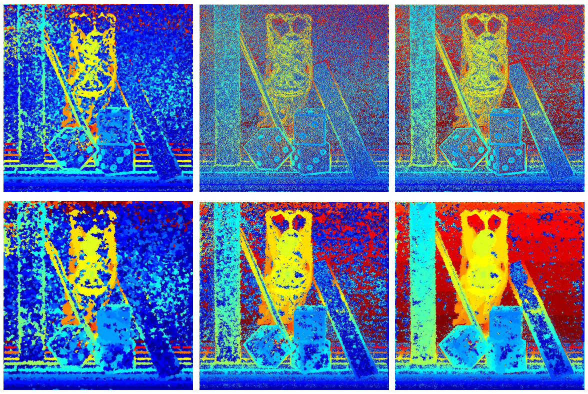
Another measure that we use to determine confidence is by looking at the curvature [10] (CUR) near the maximum response value. The measure is defined as follows:

$$CUR = 2c(l_{max}) - c(l_{max} - 1) - c(l_{max} + 1), \quad (3)$$

where l_{max} denotes the image index that outputs the maximum response. This CM essentially determines the degree to which the global maximum is a peak (high confidence) or a valley (low confidence). Like WM, we subtract the mean confidence of the incorrect pixels from the mean confidence of the correct pixels. As the results show in Table.1, RDF outperforms the other FMs in its curvature, demonstrating high confidence.

The high confidence of RDF synergizes well with the cost aggregation step [43] that comes after the initial response measurement as a refinement. Cost aggregation is a cost volume refinement method that is widely used in stereo matching [48, 33]. It works by propagating the weighted response at one location to nearby locations. The weight is determined by the color difference between the pixels. Cost aggregation synergizes with FMs well as they are able to propagate correct labels from the edges into homogeneous regions. However, for the aggregated result to be optimal, the confidence for correct labels must be high and vice versa for incorrect labels.

In Fig.3, we compared the error of the estimations against the ground truth. Before aggregation, the Laplacian filter has the most number of estimations that show no error. However, its measurements are not confident, which we can see from the high number of other labels that are also estimated. The naive approach fares better, but it is unable to be as confident in its measurements as RDF is. The impact of their confidence is shown after the FM responses are aggregated. The Laplacian filter encounters a degradation in performance, while RDF shows a vast improvement compared to the others.



(a) Laplacian (b) Naive (c) RDF

Figure 4. Depth estimation results. (Top) Depth map estimated by initial focus measure response. (Bottom) Depth map results after using tree aggregation on the initial response.

An example of this phenomenon is shown in Fig.4, using the tree-based cost aggregation proposed in [48]. As shown in this figure, due to the high confidence for correct labels of RDF, cost aggregation is able to more optimally propagate the correct labels to homogeneous regions. The other FMs are unable to do so to the degree that RDF does.

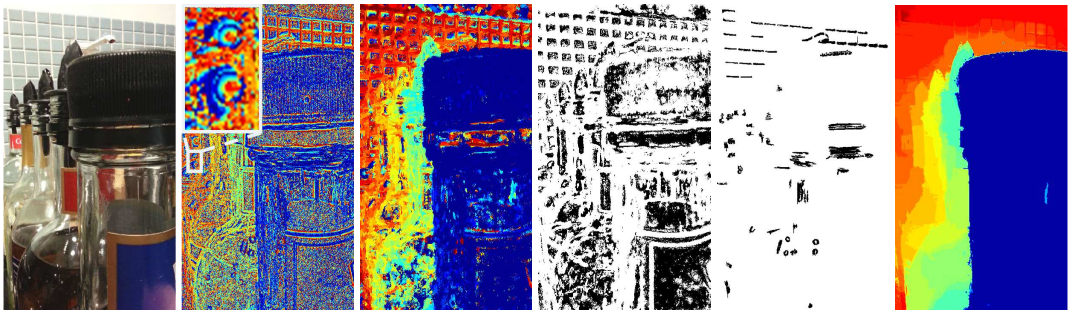
4. Proposed Pipeline

In this section, we describe the pipeline that utilizes our proposed FM. The algorithm can be broken down into three parts, the alignment in Sec.4.1, initial depth calculation in Sec.4.2, and depth refinement in Sec.4.3.

4.1. Image Alignment

The purpose of this alignment step is to compensate for the image appearance change due to the magnification from the focus change and the slight translations from the user's hand shaking during capture. As the result of the alignment process, the focal stack should resemble that from a static and telecentric camera.

We opted for a global homography-based alignment method to ensure that the original shapes of the bokeh in the images are preserved and that the algorithm runs more efficiently. We found that, on average, mobile phone cameras take anywhere between half a second and a third of a second to capture the span of its entire focus settings. The amount of local parallax encountered within this time frame under a typical capturing scenario is negligible and can be ignored without noticeable performance issues, allowing us



(a) All-in-focus (b) Initial (c) Aggregated (d) MAD mask (e) Bokeh mask (f) Final depth

Figure 5. Estimated depth maps and masks at different step of our pipeline.

to choose homography as the core of our alignment method.

There are two additional policies that we employed to maximize efficiency of the alignment. First, the alignment is applied on the images after rescaling them to half the size. The reduction in size helps suppress the extent of defocus blur present in the images, meaning that the images being aligned encounters less ambiguity. This allows the algorithm to become more robust to blur ambiguities that lead to misalignment while also reducing the computation needed for the final alignment. The second way we maintain efficiency is by initializing the frame that needs to be aligned by warping it with the previous homography warp. This is done with the assumption that the zoom and translation in the focal stack are temporally continuous. Therefore, it stands to reason that the image transformation of the image should follow that of the previous one with a small additional warp. By initializing each successive alignment this way, our method is able to avoid calculating the homography between two vastly different images, leading to faster convergence.

4.2. Initial Depth Acquisition

Once the focal stack has been properly aligned, the initial depth of the scene can be obtained by using the filter outlined in Sec. 3. This is done by applying the filter to each frame of the aligned focal stack to construct a cost volume, namely,

$$\mathcal{C}\left(x,y,l\right) = \hat{I}_l(x,y) * RDF(x,y), \tag{4}$$

where x and y denote the pixel location on the image, $\hat{I}$ denotes the aligned image, l denotes the image index within the focal stack, and $*$ denotes a 2D convolution. The initial depth, $\mathcal{D}_{init}$, can be seen by executing the following operation:

$$\mathcal{D}_{init}\left(x,y\right) = \underset{l}{\text{argmax}}\ \mathcal{C}\left(x,y,l\right). \tag{5}$$

However, as one can see from Fig.5(b), the number of correctly labeled depth pixels are very sparse and the amount of noise present in homogeneous regions are quite high. To remedy this, we employ the cost aggregation method proposed in [48] on the cost volume to reduce the noise and propagate dominant focus measure responses in the cost volume. The advantages resulting from this procedure is twofold, as shown in Fig.5(a) and (c). The obvious outcome of the two is the improved initial depth map. The other is the improved rendering of the all-in-focus image. The all-in-focus image can be obtained by stitching pixels from the images corresponding to the labeled depth:

$$\bar{I}\left(x,y\right) = \hat{I}_d\left(x,y\right), \tag{6}$$

$$d = \mathcal{D}_a\left(x,y\right), \tag{7}$$

where $\bar{I}$ and $\mathcal{D}_a$ denote the all-in-focus image and the depth acquired from the aggregated cost volume, respectively. This improved all-in-focus will be used in the next section.

4.3. Unreliable Depth Rejection and Recovery

Even with the cost aggregation, the resulting depth map remains erroneous in many regions. However, a trend can be found in the map. One is that the depth labeling at almost all the edges present in the image are correct and dominant. Another is that most of the erroneous labels are located in regions where texture is lacking. Finally, the locations with bokeh, the reflective regions with bright light that saturate the imaging sensor, exhibit a rainbow-like halo (white box in Fig.5(b)) that contaminates the depth results. These locations with unreliable outcomes must be resolved for an acceptable output. This can be done through the combined use of two different binary masks that segment reliable points and those that are not.

The first of these binary masks is obtained with a statistical measure of the cost volume. The measure employed is based on the Median of Absolute Deviation (MAD) [17], a robust measure of statistical dispersion. Our implementation uses the median-normalized MAD to extract the corresponding binary mask, defined as follows:

$$\mathcal{B}_{MAD}\left(x,y\right) = \begin{cases} 1, & \mathcal{C}_{MAD} > T_{MAD} \\ 0, & \text{otherwise} \end{cases}, \text{where} \tag{8}$$

$$\mathcal{C}_{MAD} = \frac{\operatorname*{med}_{l} \left| \mathcal{C}_a(x,y,l) - \operatorname*{med}_{k} \mathcal{C}_a(x,y,k) \right|}{\operatorname*{med}_{l} \mathcal{C}_a(x,y,l)}, \tag{9}$$

where $\mathcal{B}_{MAD}$, $\mathcal{C}_{MAD}$, and T_{MAD} are the MAD-based binary mask, cost value, and threshold, respectively, and med is the median operator. The binary mask that results from this operation is shown in Fig.5(d). Since the focus measure outputs a high response on an edge in focus and gradually moves towards zero as it becomes out of focus, the statistical dispersion should be high. In homogeneous regions, however, the focus measurements should remain stable and low, resulting in a low dispersion. As for noise, the focus measure would show a sudden peak. Although the usual dispersion measures, like variance, will count these noise as statistically significant, MAD is able to reject these values by focusing on the median, rather than the mean. The resulting binary mask acquired by the measure is one that rejects depth from homogeneous or noisy regions and keeps the edges, handling the first two trends that result in errors.

The second binary mask is obtained using the image intensity changes to handle bokeh. Due to the nature of the bokeh, the saturated regions' boundaries are detected as edges that constantly expand or contract. To detect and locate these regions, the amount of intensity change that occurs in that location must be measured as the algorithm sweeps through the focal stack. The measure we propose is the difference between the maximum and minimum gray-level intensities, described by the following operation:

$$\mathcal{B}_{bokeh}(x,y) = \begin{cases} 1, & \Delta_l \mathcal{G}_l(x,y) < T_{bokeh} \\ 0, & \text{otherwise} \end{cases}, \text{where} \tag{10}$$

$$\Delta_l \mathcal{G}_l(x,y) = \max_{l} \mathcal{G}_l(x,y) - \min_{l} \mathcal{G}_l(x,y) \tag{11}$$

where $\mathcal{B}_{bokeh}$ is the bokeh binary mask, $\mathcal{G}_l$ is the gray-level image of the l-th frame in the focal stack, and T_{bokeh} is the bokeh measure threshold. The dispersion of the intensity may be a viable measure to detect bokeh, similar to what was done for the other binary mask. However, the median of the intensities is not suitable as it disregards a large portion of the data as irrelevant. Since the level of noise encountered in the image intensities are relatively small compared to that of a saturated pixel, measuring the degree of fluctuation is actually more accurate. To this end, the binary mask for bokeh is obtained as Fig.5(e).

Once the two masks are obtained, we reject the unreliable depth labels in the aggregated depth map, $\mathcal{D}_a$, through an element-wise multiplication:

$$\mathcal{D}_{\mathcal{B}}(x,y) = \mathcal{D}_a(x,y) \cdot \mathcal{B}_{MAD}(x,y) \cdot \mathcal{B}_{bokeh}(x,y), \tag{12}$$

where $\mathcal{D}_{\mathcal{B}}$ is the depth map with unreliable labels rejected. To reconstruct the depth in the empty spaces left by the rejection procedure, we employ the tree-based propagation [48] again on $\mathcal{D}_{\mathcal{B}}$, using the all-in-focus image, $\bar{I}_a$, generated by the aggregated depth map as the algorithm's guidance. The resulting output is the final depth of this pipeline (Fig.5(f)).

5. Experimental Results

In order to demonstrate the validity of RDF and our pipeline, we conducted a set of experiments through different sets of synthetic and real-world datasets. To test our proposed focus measure, we benchmarked our filter with the top-ranked focus measures in [31] on some synthetic light-field datasets, in Sec.5.1, and real-world datasets, in Sec.5.2. In Sec.5.3, we qualitatively compared our results from those shown in [40] on their dataset.

For our threshold values, we set $\mathcal{B}_{MAD} = 0.1$ and $\mathcal{B}_{bokeh} = 0.15$. The radius parameters are $r_1 = 1$, $r_2 = 3$, and $r_3 = 5$. On average, for a focal stack of 25 frames with a resolution of 640×360, our algorithm takes 6.7s on an Intel Core i7 3.60 GHz CPU. Of that time, alignment takes 3.27s, bokeh and MAD takes 0.34s, RDF takes 0.38s, aggregation takes 1.54s, and propagation takes 1.56s.

5.1. RDF Robustness Evaluation

In order to test our filter's robustness quantitatively, we tested our filter and the three top-ranked focus measures from [31] on the synthetic light-field dataset provided by [44]. The focal stacks were generated using the same setup as in Sec.3.2. By measuring the errors in the results, we can see how noise robust each measure is.

The three FMs from [31] are the sum of wavelet coefficients (WAV) [47], modified Laplacian (LAP) [28], and the eigenvalues-based focus measure (EIG) [45]. The results of this test across datasets are shown in Table.2. We can see that WAV and LAP both give acceptable results when noise is not an issue, but as noise increases, these measures degrade rapidly. The results of EIG shows an immunity towards noise, as it stays relatively stable throughout, but the accuracy of the results are lacking compared to the other measures. The proposed method is able to acquire accurate depth and is robust to noise. Visualizations of the datasets will be in the supplementary materials.

The focus measurement times for LAP, WAV, EIG, and RDF are 1.04s, 2.08s, 582.87s, and 0.38s, respectively.

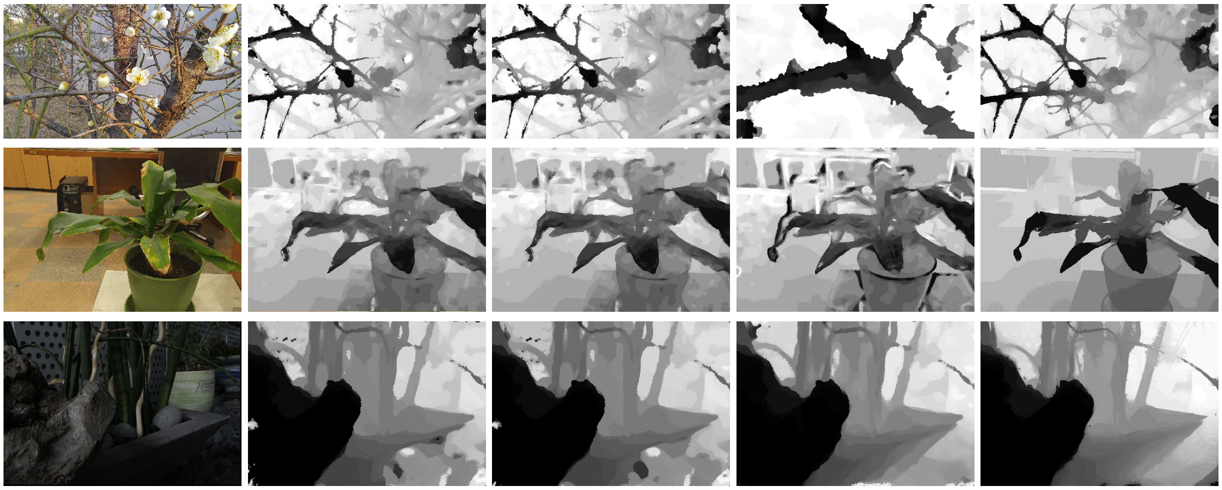
5.2. Real-World Results

The results of the depth estimation using different focus measures, including our own, are shown in Fig.6. One of them, the first row of Fig.6, was shot outside while the wind was blowing the flowers. Another, the middle row of Fig.6, was shot indoors in a low-light environment. The last row of Fig.6, was captured using 122 shots to demonstrate the

	Root mean square error								Bad pixel ratio (%)							
	σ =0				σ =0.005				σ =0				σ =0.005			
	WAV	LAP	EIG	RDF	WAV	LAP	EIG	RDF	WAV	LAP	EIG	RDF	WAV	LAP	EIG	RDF
Buddha	2.796	2.328	3.075	1.039	3.774	3.174	3.200	1.057	17.17	14.68	22.57	10.17	29.31	23.18	22.79	10.55
Buddha2	1.779	1.384	2.409	1.029	1.979	1.719	2.438	1.083	11.31	10.05	16.35	8.19	12.19	10.83	16.42	8.47
Horses	1.020	0.907	1.287	0.535	1.201	1.202	1.421	0.548	9.87	9.45	14.74	5.02	10.36	10.13	14.83	5.16
Medieval	0.916	0.812	1.474	1.284	1.081	0.942	1.489	1.359	5.80	5.34	12.09	5.42	6.21	5.76	12.05	5.59
Mona	3.188	3.076	4.557	1.884	3.775	3.388	4.605	1.535	15.89	14.16	23.78	11.79	21.32	21.01	24.01	12.12
Papillon	4.264	3.926	5.813	2.202	5.997	5.408	5.935	2.309	10.39	9.13	19.69	4.91	23.29	21.88	20.10	5.38
StillLife	1.327	1.178	2.359	0.687	1.418	1.265	2.385	0.703	11.66	11.08	17.78	10.82	12.27	11.61	17.81	10.93
Average	**2.184**	**1.944**	**2.996**	**1.237**	**2.746**	**2.442**	**3.067**	**1.228**	**11.72**	**10.55**	**18.14**	**8.04**	**16.42**	**14.91**	**18.28**	**8.31**

	σ =0.01				σ =0.02				σ =0.01				σ =0.02			
	WAV	LAP	EIG	RDF	WAV	LAP	EIG	RDF	WAV	LAP	EIG	RDF	WAV	LAP	EIG	RDF
Buddha	6.484	5.375	3.258	1.104	7.932	7.660	3.394	1.263	57.33	49.46	23.84	11.78	72.24	73.43	26.60	16.58
Buddha2	2.349	1.979	2.462	1.209	4.890	4.458	2.515	1.560	15.42	14.15	16.54	9.02	44.26	41.72	17.75	11.02
Horses	1.598	1.760	1.539	0.572	4.168	5.055	1.707	0.639	13.46	14.29	15.07	5.63	29.59	31.75	16.53	7.19
Medieval	1.301	1.189	1.489	1.439	2.199	2.543	1.516	1.561	7.25	7.08	12.21	5.95	16.92	19.64	12.75	7.22
Mona	5.097	4.782	4.676	1.520	8.052	8.105	4.826	1.987	40.49	40.77	24.65	13.15	76.06	78.42	30.03	18.79
Papillon	8.137	7.905	5.938	2.584	9.493	9.558	6.333	3.342	51.88	53.59	22.23	7.02	73.59	78.06	38.35	24.37
StillLife	1.621	1.492	2.380	0.716	3.105	3.094	2.431	0.809	14.08	14.11	17.93	11.18	24.09	26.28	18.71	12.30
Average	**3.798**	**3.497**	**3.106**	**1.306**	**5.691**	**5.782**	**3.246**	**1.594**	**28.56**	**27.64**	**18.92**	**9.10**	**48.11**	**49.90**	**22.96**	**13.92**

Table 2. Quantitative evaluations. The root mean square error and the percentage of pixels labeled incorrectly in each light-field dataset using each focus measure with different noise levels. (red = best, green = second best)



(a) All-in-focus (b) WAV (c) LAP (d) EIG (e) RDF

Figure 6. Depth map comparison on real-world datasets. First two rows were captured on an LG V10 smartphone and the last row was taken using a Canon 60D DSLR camera.

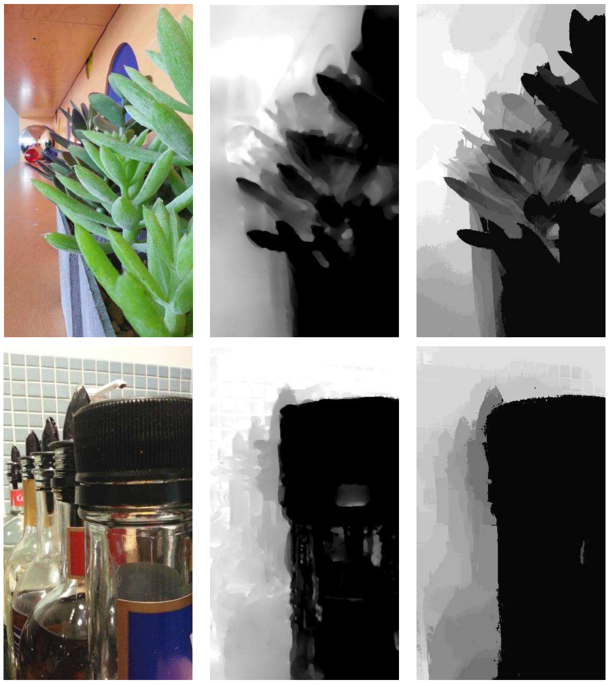
depth resolution of our proposed method in low-light conditions. These results show that the alignment using homography is enough to handle the kinds of motion generated by the waving branches, that the RDF is able to find edges within a low-light setting, and that RDF can differentiate even the shallowest of depths.

The artifacts present in the first set is most likely due to the response asymmetry of the responses for different filters. What this means is that the edge values are skewed for certain colors and, therefore, over-weigh other edge responses. In these cases, the response from one edge will overwrite the response from another because of the measure's sensitivity towards signals that are other than edges. The noisiness of the second and third sets is most likely due to the noise generated by the gain of the camera. In low light situations, the camera automatically increases the gain to become more sensitive to light, which also makes it sensitive to noise. This noise can be mitigated for RDF, which is why the results for the specific filter looks more refined.

5.3. Comparison with the State-of-the-Art

To demonstrate the practicality of our method, we compare our results to the state-of-the-art smartphone-based DfF [40]. As shown in Fig.7, our method shows finer details in the first row, due to the use of a confident focus measure in conjunction with an image-based propagation method. Also, our mask-based rejection method handles bokeh, in the second row of Fig.7, without a problem. We note that the computation time for the state-of-the-art is considerably larger, at 20 minutes with the same setup with only 25 images, as opposed to our 6.7s.

Suwajanakorn *et al.* [40] explain that the optical flow



(a) All-in-focus (b) Suwajanakorn *et al.* (c) Ours

Figure 7. Comparison between the results of our proposed method and that of Suwajanakorn *et al.* [40]. The results for the cited work is directly copied from their paper.

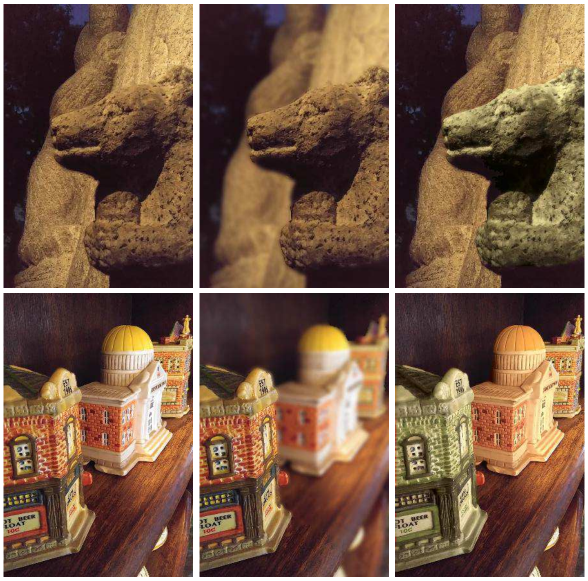
used in their method to align the images take around 8 minutes, as opposed to the homography align that we use, which only takes 8s. Another major bottleneck for their method is the refinement of the depth map, which takes around 3 minutes, while we utilize refinement methods that are well-known in the stereo matching community. These bottlenecks could have been avoided if the initial depth that they acquired were closer to the ground truth. Our refinement allows the results from the proposed pipeline to be used as a initial point that starts closer to the solution, for the convex optimization that [40] proposes for metric calibration.

5.4. Depth-Aware Image Processing Applications

An accurate depth map can be used for various consumer applications. Digital refocusing and image stylization are two such examples.

Digital refocusing is one of the most popular depth-aware image processing technique in which the in-focus region of the photo is changed and the defocus blur is enhanced [2, 6]. An accurate depth map is essential in creating a realistic blur to apply on the photo. In Fig.8(b), we synthetically blur the all-in-focus image to simulate a photo with a shallow depth of field.

Image stylization is another popular application in which the look of the image is changed. Using an accurate depth map, the pixels within a depth range can be modified to generate an aesthetically pleasing photo, like in Fig.8(c). These application examples show that our depth maps are sufficient to work well even under low-light conditions.



(a) All-in-focus (b) Refocusing (c) Stylization

Figure 8. Digital refocusing and stylization examples.

6. Conclusion

This paper proposes a new focus measure that more robustly and accurately measures the degree of focus. The key concept behind the measure is that by inserting a gap and looking at pixels that are farther away from the POI, the filter can encounter more of the informative pixel values. The pipeline that can be used as a result of such an improved measure is more computationally efficient and sufficient to give results better than that of the state-of-the-art.

However, there are some limitations to this method. The aggregation and propagation methods all depend on color dissimilarity for differing depth. This is not absolutely true in the real-world and there are corner cases in which this fails. Handling these corner cases effectively and efficiently will be part of our future research avenues.

Another direction for further research may be the handling of the quantization effect. Since DfF works by sampling discrete focus settings to acquire the focal stack, the output depth is inevitably quantized. Further research may be done to leverage other photometric cues, like shading or semantic information.

Acknowledgement. This work was supported by the Technology Innovation Program (No. 2017-10069072) funded By the Ministry of Trade, Industry & Energy (MOTIE, Korea). Hae-Gon Jeon and Sunghoon Im were partially supported by Global Ph.D. Fellowship Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2015034617, NRF-2016907531).

References

[1] Google inc., project tango. https://www.google.com/atap/project-tango/. 1

[2] HTC One (m8). http://www.htc.com/us/smartphones/htc-one-m8/. 1, 8

[3] The lytro camera. http://www.lytro.com/. 1

[4] Microsoft inc., kinect 2. https://www.microsoft.com/en-us/download/details.aspx?id=44561/. 1

[5] Occipital inc., structure sensor. http://structure.io/. 1

[6] Venue 8 7000 series. http://www.dell.com/en-us/shop/productdetails/dell-venue-8-7840-tablet/. 1, 8

[7] M. B. Ahmad and T. S. Choi. Application of three dimensional shape from image focus in lcd/tft displays manufacturing. *IEEE Transactions on Consumer Electronics*, 53(1):1–4, 2007. 2

[8] Y. An, G. Kang, I.-J. Kim, H.-S. Chung, and J. Park. Shape from focus through laplacian using 3d window. In *International Conference on Future Generation Communication and Networking*, volume 2, pages 46–50. IEEE, 2008. 2

[9] N. N. K. Chern, P. A. Neow, and M. H. Ang Jr. Practical issues in pixel-based autofocusing for machine vision. In *IEEE International Conference on Robotics and Automation*, volume 3, pages 2791–2796. IEEE, 2001. 2

[10] G. Egnal, M. Mintz, and R. Wildes. A stereo confidence metric using single view imagery with comparison to five alternative approaches. *Image and Vision Computing*, 22(12):943–957, 2004. 4

[11] A. M. Eskicioglu and P. S. Fisher. Image quality measures and their performance. *IEEE Transactions on Communications*, 43(12):2959–2965, 1995. 2

[12] L. Firestone, K. Cook, K. Culp, N. Talsania, and K. Preston. Comparison of autofocus methods for automated microscopy. *Cytometry*, 12(3):195–206, 1991. 2

[13] J.-M. Geusebroek, F. Cornelissen, A. W. Smeulders, and H. Geerts. Robust autofocusing in microscopy. *Cytometry*, 39(1):1–9, 2000. 2

[14] M. Gupta, Q. Yin, and S. Nayar. Structured Light in Sunlight. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, Dec 2013. 1

[15] H. Ha, S. Im, J. Park, H.-G. Jeon, and I. S. Kweon. High-quality depth from uncalibrated small motion clip. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016. 1

[16] F. S. Helmli and S. Scherer. Adaptive shape from focus with an error estimation in light microscopy. In *International Symposium on Image and Signal Processing and Analysis*, pages 188–193. IEEE, 2001. 2

[17] D. C. Hoaglin, F. Mosteller, and J. W. Tukey. *Understanding robust and exploratory data analysis*, volume 3. Wiley New York, 1983. 5

[18] J.-T. Huang, C.-H. Shen, S.-M. Phoong, and H. Chen. Robust measure of image focus in the wavelet domain. In *International Symposium on Intelligent Signal Processing and Communication Systems*, pages 157–160. IEEE, 2005. 2

[19] S. Im, H. Ha, G. Choe, H.-G. Jeon, K. Joo, and I. S. Kweon. High quality structure from small motion for rolling shutter cameras. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 837–845. IEEE, 2015. 1

[20] H.-G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y.-W. Tai, and I. S. Kweon. Accurate depth map estimation from a lenslet light field camera. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[21] K. Jo, M. Gupta, and S. K. Nayar. Spedo: 6 dof ego-motion sensor using speckle defocus imaging. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 4319–4327, 2015. 1

[22] S.-Y. Lee, Y. Kumar, J.-M. Cho, S.-W. Lee, and S.-W. Kim. Enhanced autofocus algorithm using robust focus measure and fuzzy reasoning. *IEEE Transactions on Circuits and Systems for Video Technology,*, 18(9):1237–1246, 2008. 2

[23] S.-Y. Lee, J.-T. Yoo, Y. Kumar, and S.-W. Kim. Reduced energy-ratio measure for robust autofocusing in digital camera. *IEEE Signal Processing Letters*, 16(2):133–136, 2009. 2

[24] R. Maini and H. Aggarwal. Study and comparison of various image edge detection techniques. *International journal of image processing*, 3(1):1–11, 2009. 2

[25] A. S. Malik and T.-S. Choi. A novel algorithm for estimation of depth map using image focus for 3d shape recovery in the presence of noise. *Pattern Recognition*, 41(7):2200–2225, 2008. 2

[26] R. Minhas, A. A. Mohammed, Q. J. Wu, and M. A. Sid-Ahmed. 3d shape from focus and depth map computation using steerable filters. In *Image Analysis and Recognition*, pages 573–583. Springer, 2009. 2

[27] H. Nanda and R. Cutler. Practical calibrations for a real-time digital omnidirectional camera. *CVPR Technical Sketch*, 20, 2001. 2

[28] S. K. Nayar and Y. Nakagawa. Shape from focus. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 16(8):824–831, 1994. 2, 6

[29] O. Osibote, R. Dendere, S. Krishnan, and T. Douglas. Automated focusing in bright-field microscopy for tuberculosis detection. *Journal of microscopy*, 240(2):155–163, 2010. 2

[30] J. L. Pech-Pacheco, G. Cristóbal, J. Chamorro-Martinez, and J. Fernández-Valdivia. Diatom autofocusing in brightfield microscopy: a comparative study. In *International Conference on Pattern Recognition*, volume 3, pages 314–317. IEEE, 2000. 2

[31] S. Pertuz, D. Puig, and M. A. Garcia. Analysis of focus measure operators for shape-from-focus. *Pattern Recognition*, 46(5):1415–1432, 2013. 2, 6

[32] Raytrix. 3d light field camera technology. http://www.raytrix.de/. 1

[33] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz. Fast cost-volume filtering for visual correspondence and beyond. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 4

[34] A. Santos, C. Ortiz de Solórzano, J. J. Vaquero, J. Pena, N. Malpica, and F. Del Pozo. Evaluation of autofocus functions in molecular cytogenetic analysis. *Journal of microscopy*, 188(3):264–272, 1997. 2

[35] D. Scharstein and R. Szeliski. Stereo matching with nonlinear diffusion. *International Journal of Computer Vision (IJCV)*, 28:155–174, 1998. 3

[36] Y. Schechner, S. Nayar, and P. Belhumeur. Multiplexing for optimal lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 29(8):1339–1354, 2007. 3

[37] C.-H. Shen and H. H. Chen. 1 robust focus measure for low-contrast images. 2006. 2

[38] M. V. Shirvaikar. An optimal measure for camera focus and exposure. In *Proceedings of the Thirty-Sixth Southeastern Symposium on System Theory*, pages 472–475. IEEE, 2004. 2

[39] Y. Sun, S. Duthaler, and B. J. Nelson. Autofocusing in computer microscopy: selecting the optimal focus algorithm. *Microscopy research and technique*, 65(3):139–149, 2004. 2

[40] S. Suwajanakorn, C. Hernandez, and S. M. Seitz. Depth from focus with your mobile phone. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3497–3506, 2015. 2, 6, 7, 8

[41] A. Thelen, S. Frey, S. Hirsch, and P. Hering. Improvements in shape-from-focus for holographic reconstructions with regard to focus operators, neighborhood-size, and height value interpolation. *IEEE Transactions on Image Processing (TIP)*, 18(1):151–157, 2009. 2

[42] K. Venkataraman, D. Lelescu, J. Duparré, A. McMahon, G. Molina, P. Chatterjee, R. Mullis, and S. Nayar. Picam: an ultra-thin high performance monolithic camera array. *ACM Transactions on Graphics (TOG)*, 32(6):166, 2013. 1

[43] L. Wang and R. Yang. Global stereo matching leveraged by sparse ground control points. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 4

[44] S. Wanner, S. Meister, and B. Goldluecke. Datasets and benchmarks for densely sampled 4d light fields. In *Vision, Modelling and Visualization*, pages 225–226. Citeseer, 2013. 3, 6

[45] C.-Y. Wee and R. Paramesran. Image sharpness measure using eigenvalues. In *International Conference on Signal Processing*, pages 840–843. IEEE, 2008. 6

[46] H. Xie, W. Rong, and L. Sun. Wavelet-based focus measure and 3-d surface reconstruction method for microscopy images. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 229–234. IEEE, 2006. 2

[47] G. Yang and B. J. Nelson. Wavelet-based autofocusing and unsupervised segmentation of microscopic images. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 3, pages 2143–2148. IEEE, 2003. 2, 6

[48] Q. Yang. A non-local cost aggregation method for stereo matching. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 4, 5, 6

[49] P.-T. Yap and P. Raveendran. Image focus measure based on chebyshev moments. In *IEE Proceedings Vision, Image and Signal Processing*, volume 151, pages 128–136. IET, 2004. 2

[50] F. Yu and D. Gallup. 3d reconstruction from accidental motion. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014. 1