sensors

MDPI

Article

Construction of All-in-Focus Images Assisted by Depth Sensing

Hang Liu [1], Hengyu Li [1,*], Jun Luo [1], Shaorong Xie [1] and Yu Sun [1,2]

[1] School of Mechatronic Engineering and Automation, Shanghai University, No. 99 Shangda Road BaoShan District, Shanghai 200444, China; liuhang@shu.edu.cn (H.L.); luojun@shu.edu.cn (J.L.); srxie@shu.edu.cn (S.X.); sun@mie.utoronto.ca (Y.S.)

[2] Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON M5S 3G8, Canada

* Correspondence: lihengyu@shu.edu.cn

Received: 13 February 2019; Accepted: 15 March 2019; Published: 22 March 2019

Abstract: Multi-focus image fusion is a technique for obtaining an all-in-focus image in which all objects are in focus to extend the limited depth of field (DoF) of an imaging system. Different from traditional RGB-based methods, this paper presents a new multi-focus image fusion method assisted by depth sensing. In this work, a depth sensor is used together with a colour camera to capture images of a scene. A graph-based segmentation algorithm is used to segment the depth map from the depth sensor, and the segmented regions are used to guide a focus algorithm to locate in-focus image blocks from among multi-focus source images to construct the reference all-in-focus image. Five test scenes and six evaluation metrics were used to compare the proposed method and representative state-of-the-art algorithms. Experimental results quantitatively demonstrate that this method outperforms existing methods in both speed and quality (in terms of comprehensive fusion metrics). The generated images can potentially be used as reference all-in-focus images.

Keywords: all-in-focus; image fusion; depth sensing

1. Introduction

The depth of field (DoF) of an imaging system is limited. With a fixed focus setting, only objects in a particular depth range appear focused in the captured source image, whereas objects in other depth ranges are defocused and blurred. An all-in-focus image in which all objects are in focus has many applications, such as digital photography [1], medical imaging [2], and microscopic imaging [3,4]. A number of all-in-focus imaging methods have been proposed, which can be grouped into two categories: point spread function (PSF)-based methods and RGB-based multi-focus image fusion methods.

The PSF-based methods obtain an all-in-focus image by estimating the PSF of the imaging system and restoring an all-in-focus image based on the estimated PSF. A partially-focused image can be modelled as an all-in-focus image convolved with a PSF. Deconvolution methods first estimate the PSF and then deconvolve with this PSF to restore an all-in-focus image. The PSF of a partially-focused image is non-uniform because the farther an object is from the DoF of an imaging system, the larger is the extent of blurriness of the object in an image. One type of deconvolution method directly estimates the non-uniform PSF of an imaging system using specially-designed cameras [5] or a camera with a specially-designed lattice-focal lens [6]. Instead of estimating the non-uniform PSF, the other type of deconvolution method first constructs an image with uniform blur and then estimates a uniform PSF. The image with uniform blur can be obtained by scanning the focus positions [4,7] or moving the lens or image detector [8] during a single detector exposure. The wave-front coding technique is another

approach to obtain a uniform blur image by adding a suitable phase mask to the aperture plane and making the optical transfer function of the imaging system defocus invariant [9–12]. The deconvolution methods enable single-shot extended DoF imaging. However, deconvolution ringing artefacts can appear in the resulting image, and high frequencies can be captured with lower fidelity [8].

In RGB-based multi-focus image fusion methods, in-focus image blocks are distinguished from among multiple multi-focus source images that are captured using different focus settings, to construct an all-in-focus image. Existing multi-focus image fusion algorithms include multi-scale transform [13,14], feature space transform [15,16], spatial domain methods [2,17–20], pulse coupled neural networks [21,22], and deep convolutional neural networks [23].

In multi-focus image fusion, one challenge is to obtain a reference all-in-focus image, which better reflects the ground truth, to which other methods are directly compared. Due to the lack of reference images, a number of metrics were defined for indirectly comparing the performance across multi-focus image fusion methods. As discussed in [24], various metrics, such as information theory-based metrics, image feature-based metrics, image structural similarity-based metrics, and human perception-based metrics [25], were developed because they all represent different aspects of the quality of an all-in-focus image.

In order to obtain a reference all-in-focus image, if the distances between all objects and the camera are known, the in-focus image blocks can be directly determined by choosing those objects whose distances are within the DoF of the camera. This is enabled by the advent and rapid advances of depth sensors (e.g., Microsoft Kinect and ZED stereo camera), which provide a convenient approach for accurately determining the distances of objects in a scene.

Actually, depth maps from depth sensors and colour images from traditional cameras are complementary to each other. Depth maps provide depth information of objects, which have been integrated with colour images to improve the performance of object tracking [26], the resolution of colour images [27], the detection of perspective-invariant features [28], etc. Compared with colour images, the resolution of depth maps from consumer depth sensors is lower with much noise. Thus, colour images have also been used to improve the resolution of depth maps and reduce the noise [29–31].

In this paper, our idea is to use the depth information from a depth map to assist the fusion of multiple multi-focus source images to construct an all-in-focus image. To our knowledge, this is the first work to use a depth map to help solve the all-in-focus imaging problem. Instead of distinguishing in-focus image blocks from among multi-focus source images, a graph-based depth map segmentation algorithm is proposed to directly obtain in-focus image block regions by segmenting the depth map. The distances of objects in each segmented in-focus image block region are confined to be within the DoF of the camera such that all objects in the region appear focused in a multi-focus source image. These regions are used to guide the focus algorithm to locate an in-focus image for each region from among multi-focus source images to construct an all-in-focus image. Experimental results quantitatively demonstrate that this method outperforms existing methods in both speed and quality (in terms of fusion metrics); thus, the generated images can potentially be used as reference all-in-focus images. The proposed method is not dependent on a specific depth sensor and can be implemented with structured light-based depth sensors (e.g., Microsoft Kinect v1), time of flight-based depth sensors (e.g., Microsoft Kinect v2), stereo cameras (e.g., ZED stereo camera), and laser scanners.

2. Multi-Focus Image Fusion System

In Figure 1, the image detector is at a distance of v from a lens with focal length of f. A scene point M, at a distance of u from the lens, is imaged in focus at m. If the lens moves forward with a distance of p from the lens, then M is imaged as a blurred circle centred around m', while the scene point N at a distance of u' $(u' < u)$ from the lens is imaged in focus at n. In optics, if the distance between m and m' is less than the radius of the circle of confusion (CoC) in the image plane, all the scene points between M and N appear acceptably sharp in the image. This indicates that by changing the distance between

the lens and image detector while capturing images, objects at different distance ranges appear focused in order in the captured multi-focus source images. DoF can be divided into back DoF (denoted by b_DoF in this work) and front DoF (denoted by f_DoF in this work), which indicate the depth range of objects after and before the precisely in-focus scene point that can appear acceptably sharp in an image.

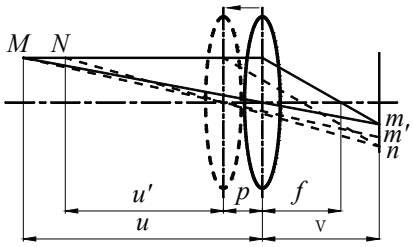


Figure 1. A scene point M at a distance of u from the lens is imaged in focus by an image detector at a distance of vfrom the lens with a focal length f. If the lens moves forward with a distance of p, M is imaged as a blurred circle around m', while the near scene point N at a distance of u' from the lens is imaged in focus at n.

Figure 2a shows our multi-focus image fusion system, which consists of a focus-tunable Pentax K01 colour camera with an 18–55-mm lens and a Kinect depth sensor. The diameter of the CoC of the colour camera δ is 0.019 mm; the aperture value F was set to 4.0; and the focal length f was set to 24 mm. The flowchart of the proposed multi-focus image fusion method is shown in Figure 2b. In this method, the depth map and multi-focus source images of an unknown scene are captured using the Kinect depth sensor and Pentax colour camera, respectively. Then, the depth map is segmented into multiple in-focus image block regions, and the objects in each region are within a DoF and all appear focused. These segmented in-focus image block regions are used to guide the focus algorithm to locate an in-focus image from among multi-focus source images for each region. Finally, the all-in-focus image is constructed by combining the in-focus images of all segmented regions.

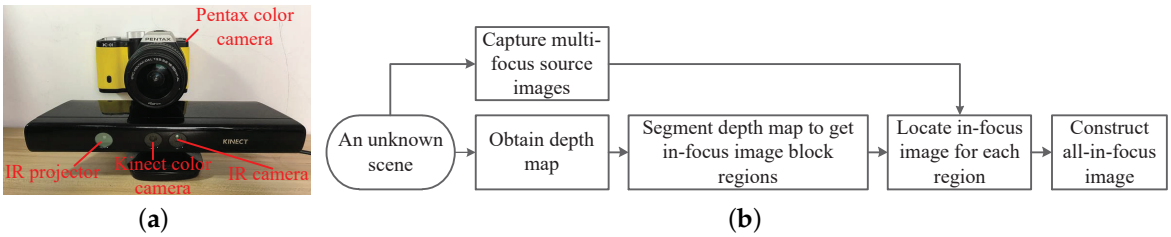


Figure 2. (**a**) Setup used in this work for evaluating the proposed multi-focus image fusion method. (**b**) Flowchart of the method.

3. Detailed Methods

Figure 3 uses an example to illustrate the main steps and intermediate results when the proposed multi-focus image fusion method is applied to construct an all-in-focus image of a scene. Firstly, the depth map from the Kinect depth sensor is preprocessed to align with the colour image captured with the colour camera, based on a stereo calibration method, and to recover the missing depth values. A graph-based image segmentation algorithm is then used to segment the preprocessed depth map into regions. A focus algorithm is used to locate an in-focus image for each region from among multi-focus source images to construct an all-in-focus image.

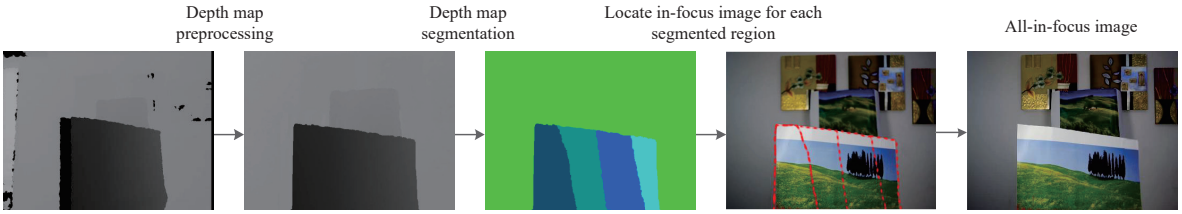


Figure 3. Main steps and intermediate results when applying the proposed multi-focus image fusion method to construct an all-in-focus image of a real scene.

3.1. Depth Map Preprocessing

3.1.1. Align Depth Map with Colour Image

Microsoft Kinect contains a depth sensor and an RGB camera that provides both depth and colour streams with a resolution of 640 × 480 at 30 Hz. The depth sensor consists of an infrared (IR) projector combined with an IR camera. The IR projector projects a set of IR dots, and the IR camera observes each dot and matches it with a dot in the known projector pattern to obtain a depth map. The operating range of the present Kinect depth sensor is between 0.5 m and 5.0 m [32].

Due to the different spatial positions and intrinsic parameters of the IR camera of the Kinect depth sensor and of the Pentax colour camera, the depth map is not aligned with the colour image. To align the depth map with the colour image, the depth map is first mapped to 3D points in the IR camera's coordinate system using the intrinsic parameters of the IR camera. Then, these 3D points are transformed to the Pentax colour camera's coordinate system using extrinsic parameters that relate the IR camera's coordinate system and the colour camera's coordinate system. Finally, the transformed 3D points are mapped to the colour image coordinate system using the intrinsic parameters of the colour camera.

Let (u_0, v_0) denote the coordinates of the principal point of the IR camera, f_x and f_y denote the scale factors in the image u and v axes of the IR camera, and u_0, v_0, f_x, and f_y be the intrinsic parameters of the IR camera. Let $[u, v, Z]$ represent a pixel in the depth map, Z represent the depth value in $[u, v]$, and $[X, Y, Z]^T$ represent the mapped 3D point of $[u, v]$ in the IR camera coordinate system. According to the pinhole camera model, the values of X and Y can be calculated according to:

$$\begin{aligned} X &= (u - u_0)Z/f_x, \\ Y &= (v - v_0)Z/f_y. \end{aligned} \tag{1}$$

Let R and T represent the rotation and translation that relate the coordinate system of the IR camera of the Kinect depth sensor and the colour camera's coordinate system. R and T are the extrinsic parameters. R is a 3×3 matrix, and T is a 3×1 matrix. The relationship between the transformed 3D point $[X', Y', Z']^T$ in the colour camera's coordinate system and $[X, Y, Z]^T$ can be expressed as:

$$\left[X', Y', Z'\right]^T = R[X, Y, Z]^T + T. \tag{2}$$

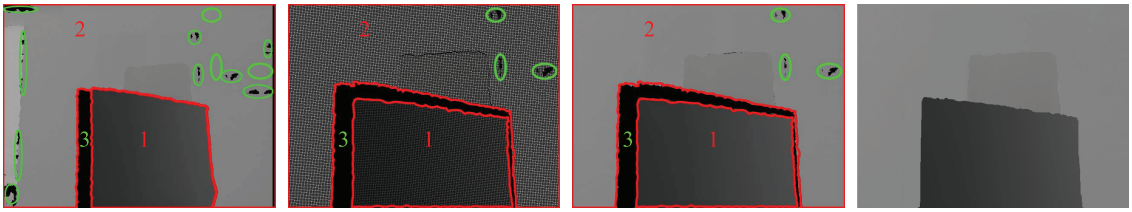
Let (u_0', v_0') denote the coordinates of the principal point of the colour camera and $f_x{}'$ and $f_y{}'$ denote the scale factors in the image u' and v' axes of the colour camera. After mapping $[X', Y', Z']^T$ to the colour image coordinate system, the aligned depth point $[u', v', Z']$ can be obtained, where u' and v' are calculated according to:

$$\begin{aligned} u' &= \tfrac{X'}{Z'} f_x{}' + u_0{}' \\ v' &= \tfrac{Y'}{Z'} f_y{}' + v_0{}'. \end{aligned} \tag{3}$$

The intrinsic parameters of the IR camera of the Kinect depth sensor and the colour camera and their extrinsic parameters are determined using a stereo camera calibration method. In the example shown in Figure 4a, there are many pixels in Regions 1 and 2 that have a value of zero because the aligned depth Regions 1 and 2 are larger than their corresponding Regions 1 and 2 in Figure 4a,

and these pixels do not obtain depth values from Figure 4a. A dilation operation is used to recover the depth value of these pixels. A 3×3 rectangular structuring element is used to dilate the source depth map to determine the shape of a pixel's neighbourhood over which the maximum is taken, according to:

$$dst(x,y) = \max_{(x',y'):element(x',y')\neq 0} src(x+x',y+y'). \tag{4}$$



(**a**) (**b**) (**c**) (**d**)

Figure 4. Depth map preprocessing: (**a**) raw depth map from the Kinect depth sensor, (**b**) aligned depth map, (**c**) aligned depth map after dilation, and (**d**) aligned depth map after hole filling. Black holes are labelled with green-coloured ellipses; the largest hole is labelled with "3" in green colour; the front object region and background region are labelled with "1" and "2", respectively.

3.1.2. Depth Map Hole Filling

From the aligned depth map (Figure 4c), there still exist a number of black holes that are labelled with green-coloured ellipses, and the largest hole labelled with "3" in green colour. These holes are caused by the structured light that the IR projector of the Kinect depth sensor emits, which was reflected in multiple directions, encountered transparent objects, and scattered from object surfaces [33]. To avoid incorrect segmentation, these depth holes must be filled.

The task is to use valid depth values around depth holes to fill the depth holes. Vijayanagar et al. [34] proposed a multi-resolution anisotropic diffusion (AD) method, which uses the colour image to diffuse the depth map and requires this process to be iterated many times in the multi-resolutions of the colour image for each resolution. Differently, as discussed in the next sub-section on depth map segmentation, the depth value of a filled hole only needs to be within the DoF at its neighbouring valid depth value. Therefore, the AD method is applied more efficiently in our work. (1) The AD filter is only applied to the depth map of the original size. (2) The conduction coefficients are only computed from the depth map. (3) Only one iteration of AD is applied because after one iteration, the differences between the depth value of the recovered pixel and its neighbours become less than the DoF at the recovered depth value, and thus, incorrect segmentation is avoided.

For an image I, the discrete form of the anisotropic diffusion equation, according to [35], is:

$$\begin{aligned} I^{t+1}_{(i,j)} = I^{t}_{(i,j)} + \lambda(C_N \cdot d_N + C_S \cdot d_S \\ + C_W \cdot d_W + C_E \cdot d_E)^{t}_{(i,j)}, \end{aligned} \tag{5}$$

where $0 \leq \lambda \leq 0.25$ for the equation to be stable, t indicates the current iteration, d represents the depth value difference between the pixel $I_{(i,j)}$ and one of its four neighbours, and the subscripts N, S, E, and W denote the neighbouring pixels to the north, south, east, and west. The conduction coefficient C is:

$$C = g(d) = e^{(-(d/K)^2)}, \tag{6}$$

where K is the standard deviation.

To recover the depth value of $I_{(i,j)}$, since the IR projector is located on the right side of Kinect and the IR camera is on the left side, the main depth holes (Region 3 in Figure 4c) are always to the left of an object, and we replace $I_{(i,j)}$ with $I_{(i-2,j)}$ to fill the depth holes. Thus, (5) is rewritten as:

$$\begin{aligned} I_{(i,j)} = I_{(i-2,j)} + \lambda(C_N \cdot d_N + C_S \cdot d_S \\ + C_W \cdot d_W + C_E \cdot d_E)_{(i-2,j)}, \end{aligned} \tag{7}$$

where:

$$\begin{aligned} d_N &= I_{(i-3,j)} - I_{(i-2,j)}, \\ d_S &= I_{(i-1,j)} - I_{(i-2,j)}, \\ d_W &= I_{(i-2,j-1)} - I_{(i-2,j)}, \\ d_E &= I_{(i-2,j+1)} - I_{(i-2,j)}. \end{aligned} \tag{8}$$

The aligned depth map after hole filling is shown in Figure 4d.

3.2. Graph-Based Depth Map Segmentation

After preprocessing the depth map, the depth map is segmented into distinct image block regions. Each segmented region must satisfy the DoF rule, as described below, to ensure all objects in this region appear in focus. In Figure 5, scene point L is at a distance of u_l from the lens, M is at a distance of u, and S is at a distance of u_s. The three points are imaged as l at a distance of v_l, m at a distance of v, and s at a distance of v_s. Among the three scene points, only M is imaged in perfect focus at the image detector; L and S are imaged as a blurred circle with diameter δ centred around m. The DoF consists of two parts, the back DoF (b_DoF) and front DoF (f_DoF), and their values at a distance of u can be derived as:

$$b\_DoF(u) = u_l - u = F\delta u^2/(f^2 - F\delta u) \tag{9}$$

$$f\_DoF(u) = u - u_s = F\delta u^2/(f^2 + F\delta u) \tag{10}$$

where $F = f/d$ is the aperture value. Let Min and Max represent the minimum and maximum depth values in a segmented region, respectively, and let $Diff$ represent the difference between Min and Max (i.e., $Diff$ = Max - Min). Let $b\_DoF(Min)$ represent the back DoF when the camera is in focus at Min, $f\_DoF(Max)$ represent the front DoF when the camera is in focus at Max, and $MaxDoF$ represent the larger value between $b\_DoF(Min)$ and $f\_DoF(Max)$. To ensure all objects in a segmented region all appear focused, the DoF rule requires that $Diff$ be smaller than $MaxDoF$ (i.e., $Diff < MaxDoF$).

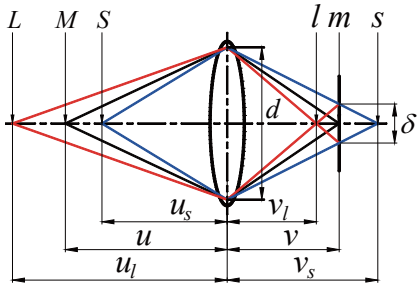
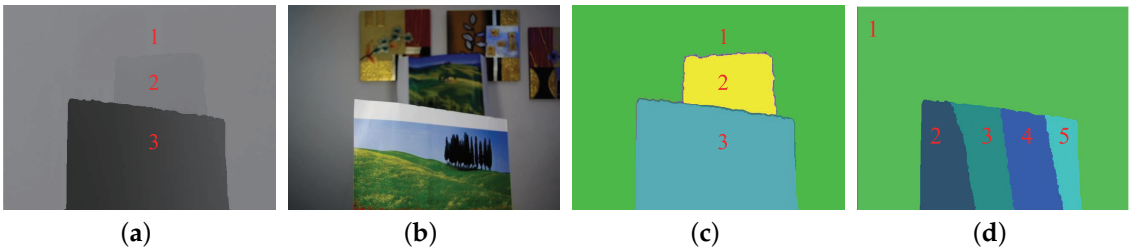


Figure 5. Diagram for calculating the depth of field (DoF).

In graph theory-based segmentation algorithms, a graph with vertices, image pixels, and edges corresponding to pairs of neighbouring vertices is established. Each edge has a weight initialized by the difference between the values of pixels on each side of the edge. In existing graph theory-based segmentation algorithms, blocks of pixels with low variability tend to be segmented into a single region. For an object with a wide depth range, the entire object crosses multiple DoFs and cannot appear focused in one focus setting. In this case, standard graph-based segmentation algorithms would incorrectly segment the entire object into a single region. For objects within a specific DoF of

the camera, but with different depth values, the standard graph-based segmentation algorithms may unnecessarily segment these objects into different regions.

Figure 6a is the depth map of a real scene with its corresponding colour image shown in Figure 6b). We first applied the classic graph-based segmentation algorithm (Felz algorithm) [36], which segmented the depth map into three regions (Figure 6c). Table 1 summarizes the values of Min, Max, $Diff$, and $MaxDoF$ of each segmented region. For Region 3, $Diff$ is larger than $MaxDoF$, indicating that all the objects in Region 3 cannot appear focused in one focus setting. Since the depth values in Region 3 change gradually from 832 mm–1360 mm, they were incorrectly segmented into a single region. For Regions 1 and 2, when the camera was set to focus at the minimum depth value in Region 2 (2417 mm), and b_DoF was 1132 mm, which is larger than the difference (698 mm) between the minimum depth value in Region 2 (2417 mm) and the maximum depth value in Region 1 (3115 mm), indicating that the objects in Regions 1 and 2 can appear focused in one focus setting. In summary, with the Felz algorithm, Regions 1 and 2 in Figure 6c were unnecessarily segmented into two regions, and Region 3 in Figure 6c was incorrectly regarded as a single region.



(**a**) (**b**) (**c**) (**d**)

Figure 6. Depth map segmentation using the standard graph-based segmentation algorithm and our modified graph-based segmentation algorithm. (**a**) Raw depth map of a real scene. (**b**) Colour image of the scene. (**c**) Segmentation result by using the standard graph-based Felz algorithm. (**d**) Segmentation result using our modified algorithm.

Table 1. Depth values (in mm) of segmented regions in Figure 6c.

Region	Min	Max	$Diff$	$MaxDoF$	$Diff < MaxDoF$?
1	2722	3115	393	1526	Yes
2	2417	2639	222	1132	Yes
3	832	1360	528	207	No

In our depth map segmentation, a graph-based representation of the depth map is first established, in which pixels are nodes and edge weights measure the dissimilarity between nodes (e.g., depth differences). Given two components, C_1 and C_2, let min and max represent the minimum and maximum depth values among all the depth pixels within C_1 and C_2, $diff$ equal max minus min, and $b\_DoF(min)$ and $f\_DoF(max)$ represent the back DoF and front DoF when the camera is set to focus at min and max, respectively. To ensure that the final segmented regions can all appear focused in one focus setting of the camera, we then impose the rule of DoF, i.e., only if $diff$ is less than the larger value of $b\_DoF(min)$ and $f\_DoF(max)$ can the two components be merged.

The segmentation result using the proposed graph-based depth map segmentation algorithm is shown in Figure 6d. The Min, Max, $Diff$, and $MaxDoF$ values of each segmented region are shown in Table 2. It can be seen that in every region, $Diff$ is less than $MaxDoF$, indicating that all objects within each region can appear focused in one focus setting.

Table 2. Depth values (in mm) of segmented regions in Figure 6d.

Region	Min	Max	$Diff$	$MaxDoF$	$Diff < MaxDoF$?
1	2463	3140	677	1186	Yes
2	855	962	107	109	Yes
3	950	1085	135	136	Yes
4	1088	1269	181	182	Yes
5	1273	1412	139	257	Yes

3.3. Construct All-in-Focus Image

Based on the segmented regions on the depth map, a focus algorithm is guided to locate an in-focus image for each region from among the multi-focus images captured at different focus settings. In our work, the focus algorithm of the normalized variance (NV) is used due to its best overall performance in terms of accuracy, number of false maxima, and noise level [37]. Consider a grey image I of size $M \times N$, where M equals the number of rows and N is the number of columns. NV is computed according to:

$$NV = \frac{1}{M \times N \times \mu} \sum_{M} \sum_{N} \left(I(x,y) - \mu \right)^2, \tag{11}$$

where μ is the mean grey value of image I and $I(x,y)$ is the grey value of the pixel at position (x,y) of image I.

4. Experiments

4.1. Evaluation Metrics

Seven representative fusion methods were selected for comprehensive comparisons with our proposed method. These five methods are discrete wavelet transform (DWT) [38], nonsubsampled contourlet transform (NSCT) [39], image matting (IM) [18], guided filtering (GF) [17], spatial frequency-motivated pulse coupled neural networks in the nonsubsampled contourlet transform domain (NSCT-PCNN) [22], dense SIFT (DSIFT) [24], and the deep convolutional neural network (DCNN) [23]. DWT and NSCT are multi-scale transform methods; IM and GF are spatial methods; NSCT-PCNN is a PCNN-based and multi-scale transform method; DSIFT is a feature space method; and DCNN is a deep learning method. The source codes of these algorithms were obtained online (see the Supplementary Material).

In image fusion applications, there is a lack of a reference image or a fused image as the ground truth for comparing different algorithms. As reported in [25], fusion metrics are categorized into four groups: (1) information theory-based metrics, (2) image feature-based metrics, (3) image structural similarity-based metrics, and (4) human perception-inspired fusion metrics. In the experiments, six fusion metrics covering all four categories were chosen, including normalized mutual information Q_{MI} [40], nonlinear correlation information entropy Q_{NCIE} [41], gradient-based fusion metric Q_G [42], phase congruency-based fusion metric Q_P [25], Yang's fusion metric Q_Y [43], and the Chen–Blum metric Q_{CB} [44]. Q_{MI} and Q_{NCIE} are information theory-based metrics; Q_G and Q_P are image feature-based metrics; Q_Y is an image structural similarity-based metric; and Q_{CB} is a human perception-based metric. These six fusion metrics were implemented using the image fusion evaluation toolbox at https://github.com/zhengliu6699. For all six metrics, a larger value indicates a better fusion result.

4.2. Source Images

Multi-focus source images from five different scenes were captured and used in this study (Figure S1 in the Supplementary Material). Figure 7 shows the source images of one of the scenes. Figure 7a is the depth map of the scene. The depth map segmentation resulted in only two regions: the front region and the background region, as shown in Figure 7b. Thus, the focus algorithm was

guided to locate the two multi-focus source images (Figure 7c,d), which were then used to construct an all-in-focus image. The scenes tested in this work were intentionally set to have only two regions, and there were only two multi-focus source images because the online image fusion evaluation toolbox (https://github.com/zhengliu6699) was designed to evaluate the fusion performance of two source images. In addition, all the source codes of different multi-focus image fusion methods were also designed to fuse two images. The source images of other scenes are provided in the Supplementary Material and can be downloaded from the author's GitHub website (https://github.com/robotVisionHang).

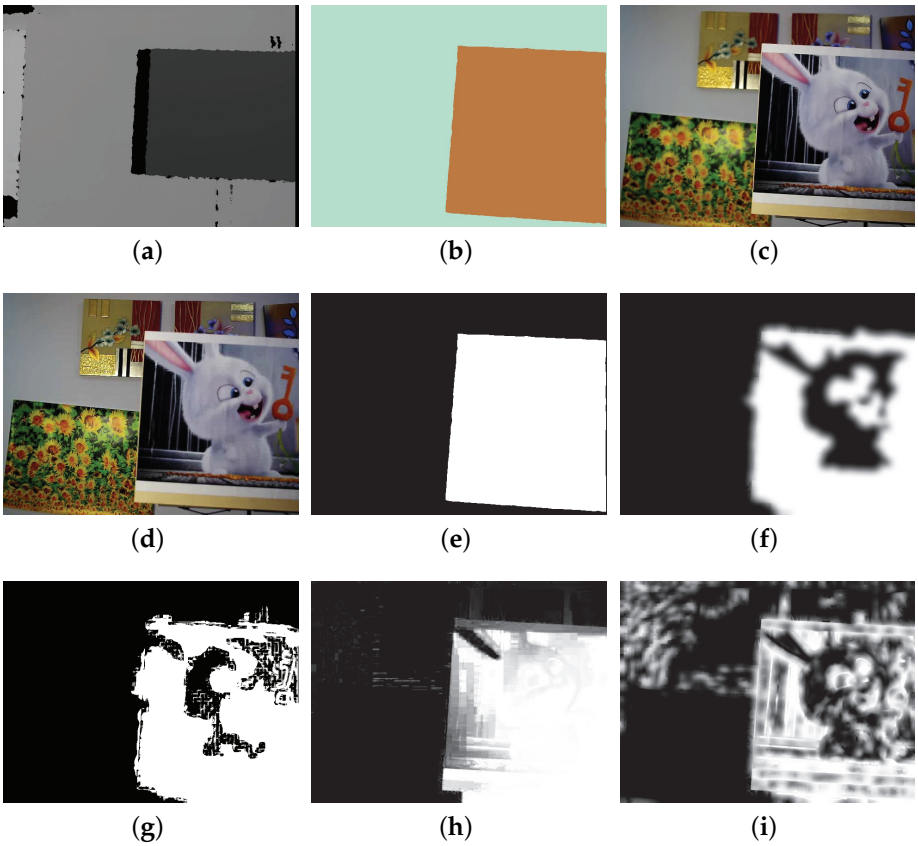


Figure 7. Images of one test scene (**a**–**d**) and weight maps of different multi-focus image fusion methods (**e**–**i**). This is the fifth scene in Figure S1 (Supplementary Material). (**a**) Depth map of the scene. (**b**) Segmentation result of the depth map. (**c**) Multi-focus source image with the front object in best focus. (**d**) Multi-focus source image with the background objects in best focus. Weight maps generated by (**e**) our proposed method, (**f**) DCNN, (**g**) dense SIFT (DSIFT), (**h**) image matting (IM), and (**i**) guided filtering (GF). These weight maps are also shown in the fifth group of weight maps in Figure S2 (Supplementary Material).

4.3. Comparison Results

The assessment metric values of the all-in-focus images constructed using our proposed method and other multi-focus image fusion algorithms for different scenes are summarized in Table 3. For each metric, the numbers in parentheses denote the score of each of the seven methods. The highest score was seven, and the lowest score was one. The higher the score, the better the method.

Table 4 shows the number of times of each method receiving a score, the total score of each method, and the overall ranking of the eight methods. Among the eight methods, our proposed method received a score of eight for the highest number of times and had the highest overall ranking. The results also reveal that our proposed method, DCNN, DSIFT, and IM outperformed GF, NSCT-PCNN, DWT,

and NSCT, and GF performed better than other multi-scale transform methods (NSCT-PCNN, DWT, and NSCT).

The core process of state-of-the-art RGB-based multi-focus image fusion methods (e.g., DCNN, DSIFT, GF, and IM) is to compute a weight map by comparing the relative clearness level of multi-focus source images based on the deep convolutional neural network, dense SIFT feature, guided filter, and image matting, respectively. In our proposed method, the weight map was generated through segmenting the depth map. Take A and B as two multi-focus source images, and W is the weight map. A fused image, F, is constructed according to:

$$F = (1.0 - W) * A + W * B, \tag{12}$$

where $*$ is an operation of pixel-wise multiplication. The range of values for W is 0.0–1.0. In a position (i, j) within W, a value of 0.0 means the fusion method judges that A is definitely clearer than B, and a value of 1.0 means B is definitely clearer than A in (i, j). If the fusion method is uncertain about whether A is definitely clearer than B, it assigns a value between 0.0 and 1.0 to represent the clearness level of A compared with B. A value less than 0.5 indicates that A is considered to be probably clearer than B; a value of 0.5 indicates that A and B are considered to be equally clear; and a value higher than 0.5 indicates that B is considered to be probably clearer than A.

Table 3. Quantitative assessments of the proposed all-in-focus imaging method and other existing multi-focus image fusion methods. Parentheses denote the scores of a method when compared with the other six methods. The higher the score, the better the method. Eight is the highest score, and one is the lowest score. NSCT-PCNN, pulse coupled neural networks in the nonsubsampled contourlet transform domain.

Scenes	Metrics	Methods							
		DWT	NSCT	IM	GF	NSCT-PCNN	DSIFT	DCNN	Ours
1	Q_{MI}	1.1478(2)	1.0451(1)	1.3869(5)	1.3402(4)	1.3372(3)	1.4235(8)	1.3903(6)	1.4201(7)
	Q_{NCIE}	0.8463(2)	0.8408(1)	0.8629(4)	0.8597(3)	0.8646(6)	0.8681(8)	0.8635(5)	0.8653(7)
	Q_G	0.6694(3)	0.4408(1)	0.6998(5)	0.6946(4)	0.6421(2)	0.7079(6)	0.7094(7)	0.7153(8)
	Q_P	0.8344(2)	0.7255(1)	0.9129(8)	0.9023(4)	0.8516(3)	0.9049(5)	0.9112(7)	0.9099(6)
	Q_Y	0.8992(2)	0.7262(1)	0.9548(5)	0.9412(4)	0.9275(3)	0.9710(6)	0.9721(7)	0.9766(8)
	Q_{CB}	0.7372(2)	0.6935(1)	0.7688(5)	0.7634(4)	0.7977(8)	0.7575(3)	0.7708(6)	0.7742(7)
2	Q_{MI}	0.9504(2)	0.8125(1)	1.2323(7)	1.1674(4)	1.0457(3)	1.2308(6)	1.2250(5)	1.2504(8)
	Q_{NCIE}	0.8308(2)	0.8250(1)	0.8480(7)	0.8426(4)	0.8374(3)	0.8468(6)	0.8465(5)	0.8489(8)
	Q_G	0.6387(3)	0.3889(1)	0.6855(6)	0.6747(4)	0.5777(2)	0.6834(5)	0.6879(7)	0.6954(8)
	Q_P	0.8273(3)	0.6922(1)	0.9159(5)	0.9175(6)	0.8269(2)	0.9141(4)	0.9206(8)	0.9191(7)
	Q_Y	0.9012(3)	0.6908(1)	0.9655(6)	0.9431(4)	0.8976(2)	0.9627(5)	0.9716(7)	0.9832(8)
	Q_{CB}	0.7231(2)	0.6681(1)	0.7856(6)	0.7627(3)	0.7744(4)	0.7832(5)	0.7887(7)	0.7977(8)
3	Q_{MI}	0.9101(2)	0.8422(1)	1.1820(5)	1.1500(4)	1.0052(3)	1.2015(7)	1.1927(6)	1.2089(8)
	Q_{NCIE}	0.8284(2)	0.8255(1)	0.8437(5)	0.8414(4)	0.8344(3)	0.8448(7)	0.8442(6)	0.8454(8)
	Q_G	0.6608(3)	0.4649(1)	0.7039(5)	0.6998(4)	0.5672(2)	0.7079(6)	0.7099(7)	0.7143(8)
	Q_P	0.8266(3)	0.7660(1)	0.9070(5)	0.9115(7)	0.8053(2)	0.9112(6)	0.9127(8)	0.9033(4)
	Q_Y	0.9151(3)	0.7796(1)	0.9742(5)	0.9602(4)	0.8834(2)	0.9759(6)	0.97997	0.9825(8)
	Q_{CB}	0.7059(2)	0.6699(1)	0.7816(5)	0.7681(4)	0.7169(3)	0.7903(6)	0.7949(7)	0.7954(8)
4	Q_{MI}	0.8384(2)	0.7653(1)	1.1384(5)	1.0978(4)	0.9426(3)	1.1727(7)	1.1520(6)	1.1828(8)
	Q_{NCIE}	0.8249(2)	0.8220(1)	0.8408(5)	0.8382(4)	0.8310(3)	0.8430(7)	0.8415(6)	0.8439(8)
	Q_G	0.6269(3)	0.4355(1)	0.6738(5)	0.6642(4)	0.5434(2)	0.6786(6)	0.6822(7)	0.6886(8)
	Q_P	0.7967(3)	0.7586(2)	0.8972(4)	0.9039(6)	0.7443(1)	0.9020(5)	0.9048(7)	0.9067(8)
	Q_Y	0.9047(3)	0.7491(1)	0.9692(5)	0.9500(4)	0.8729(2)	0.9777(6)	0.9837(7)	0.9890(8)
	Q_{CB}	0.6908(2)	0.6486(1)	0.7713(5)	0.7527(4)	0.7075(3)	0.7828(6)	0.7852(8)	0.7834(7)
5 (Figure 7)	Q_{MI}	0.9352(2)	0.8659(1)	1.1746(5)	1.1420(4)	0.9868(3)	1.2248(7)	1.1968(6)	1.2311(8)
	Q_{NCIE}	0.8305(2)	0.8276(1)	0.8444(5)	0.8435(4)	0.8335(3)	0.8481(7)	0.8465(6)	0.8482(8)
	Q_G	0.6432(3)	0.4472(1)	0.6720(5)	0.6594(4)	0.5506(2)	0.6751(6)	0.6753(7)	0.6885(8)
	Q_P	0.8381(3)	0.7649(1)	0.9011(7)	0.8953(4)	0.7858(2)	0.8973(5)	0.8984(6)	0.9214(8)
	Q_Y	0.9016(3)	0.7483(1)	0.9628(5)	0.9419(4)	0.8702(2)	0.9698(6)	0.9769(7)	0.9802(8)
	Q_{CB}	0.7117(2)	0.6785(1)	0.7860(5)	0.7607(4)	0.7186(3)	0.7966(7)	0.7964(6)	0.8014(8)

Table 4. Scores and rankings of the methods.

Number of Times / Scores / Methods	8	7	6	5	4	3	2	1	Total Scores	Ranking
Ours	23	5	1	0	1	0	0	0	229	1
DCNN	3	14	10	3	0	0	0	0	197	2
DSIFT	2	7	13	6	1	1	0	0	180	3
IM	1	3	3	21	2	0	0	0	160	4
GF	0	1	2	0	25	2	0	0	125	5
NSCT-PCNN	1	0	1	0	1	14	12	1	85	6
DWT	0	0	0	0	0	0	13	17	73	7
NSCT	0	0	0	0	0	0	1	29	31	8

The better performance of our proposed method compared to the other multi-focus image fusion methods can be understood by examining the weight maps they generated. For GF, the weight map for the detail layer was used to reconstruct the base layer and the detail layer of the fused image due to its more detailed reflection of the level of sharpness compared with the weight map for the base layer. Interestingly, the fused image reconstructed only with a detail layer (vs. with both base layer and detail layer [17]) generally obtained a higher score (see Table S1 in the Supplementary Material).

The values in the weight map of DSIFT can take on 0.0, 0.5, or 1.0, and for DCNN, IM and GF, the values ranged from 0.0–1.0. In our proposed method, the weight map was generated through the segmented regions. For a scene with only two segmented regions, the values in the weight map within a segmented region were all zeros, since the pixels of one multi-focus source image within this region were considered in best focus. Similarly, the values in the weight map within the other segmented region were all ones.

To fuse the multi-focus source images shown in Figure 7c,d, the weight maps generated by our proposed method, DCNN, DSIFT, IM, and GF are shown in Figure 7. The weight maps of other test scenes can be found in the Supplementary Material. This scene only contains two regions, the front region and the background region. During image capturing, the distance from the front region and the background region was set to be sufficiently large to ensure that when one region is in focus, the other region is defocused. Figure 7e shows that the white front region and black background region are completely separated; the weight values in the front region are all ones, and the weight values in background region are all zeros, accurately reflecting the sharpness level of this scene. However, in Figure 7f–i, it can be seen that none of the DCNN, DSIFT, IM, and GF methods were able to generate a weight map as clean as the weight map generated by our proposed method because they rely on the colour information of the multi-focus source images for computing weight maps, which is susceptible to lighting, noise, and the texture of objects. Differently, our proposed method circumvents these limitations by making use of the depth map to directly determine weight maps.

The time consumption for constructing an all-in-focus image using our proposed method and other multi-focus image fusion algorithms was also quantified and compared. The sizes of the multi-focus source images and depth maps were 640 × 480. Tests were conducted on a computer with a 4-GHz CPU and 32 GB of RAM. The time consumption of our proposed method reported in Table 5 includes preprocessing the depth map, segmenting the depth map, and selecting in-focus images from multi-focus source images to construct an all-in-focus image. Our method took 33 ms on average to construct an all-in-focus image, among which preprocessing the depth holes cost 5 ms, segmenting the depth map cost 27.5 ms, and selecting in-focus images to construct the all-in-focus image cost 0.5 ms. The significantly lower time consumption of our method, compared to the RGB-based methods (see Table 5), is due to the low computational complexity stemming from the assistance of the depth map. Note that in practice, there are usually more than two multi-focus source images to be used to construct an all-in-focus image of a scene, and in accordance, the time consumption of other multi-focus image

fusion methods increases linearly. Differently, for the proposed all-in-focus imaging method, since the time cost of preprocessing and segmenting the depth map is linear to the size of the depth map [36], as long as the size of the depth map from the depth sensor is fixed, the time cost of preprocessing and segmenting the depth map stays constant. Although the time cost of selecting in-focus images is linear to the number of multi-focus source images, due to its low computational complexity, the time consumption of our proposed method does not increase significantly when the number of multiple multi-focus source images becomes higher.

The proposed multi-focus image fusion method is highly dependent on the depth map from the depth sensor. Presently, the range of the Kinect depth sensor is limited to 0.5 m–5 m. However, the proposed method is not dependent on a specific depth sensor. For instance, the ZED stereo camera has a significantly larger operating range (0.5 m–20 m) and can obtain depth maps with a size up to 4416 × 1242 at 15 fps. Figure 8 shows the use of the ZED stereo camera for obtaining the depth map of more complex nature scenes.

Table 5. Running time (seconds) of the proposed method and existing multi-focus image fusion algorithms for the five test scenes.

Scenes	Methods							
	DWT	NSCT	IM	GF	NSCT-PCNN	DSIFT	DCNN	Ours
1	0.2054	35.7285	3.2084	0.3351	243.2443	8.8385	132.9873	0.030
2	0.2031	35.5960	3.1097	0.3491	243.8029	11.4488	131.7024	0.035
3	0.2061	35.7128	2.9816	0.3473	243.4221	7.6047	131.6626	0.033
4	0.2039	35.7426	2.9719	0.3457	243.8831	7.3378	127.3014	0.032
5 (Figure 7)	0.2050	35.7939	2.9131	0.3452	243.1754	9.4629	132.2269	0.035
Average	0.2047	35.7148	3.0369	0.3445	243.5056	8.9385	131.1761	0.033

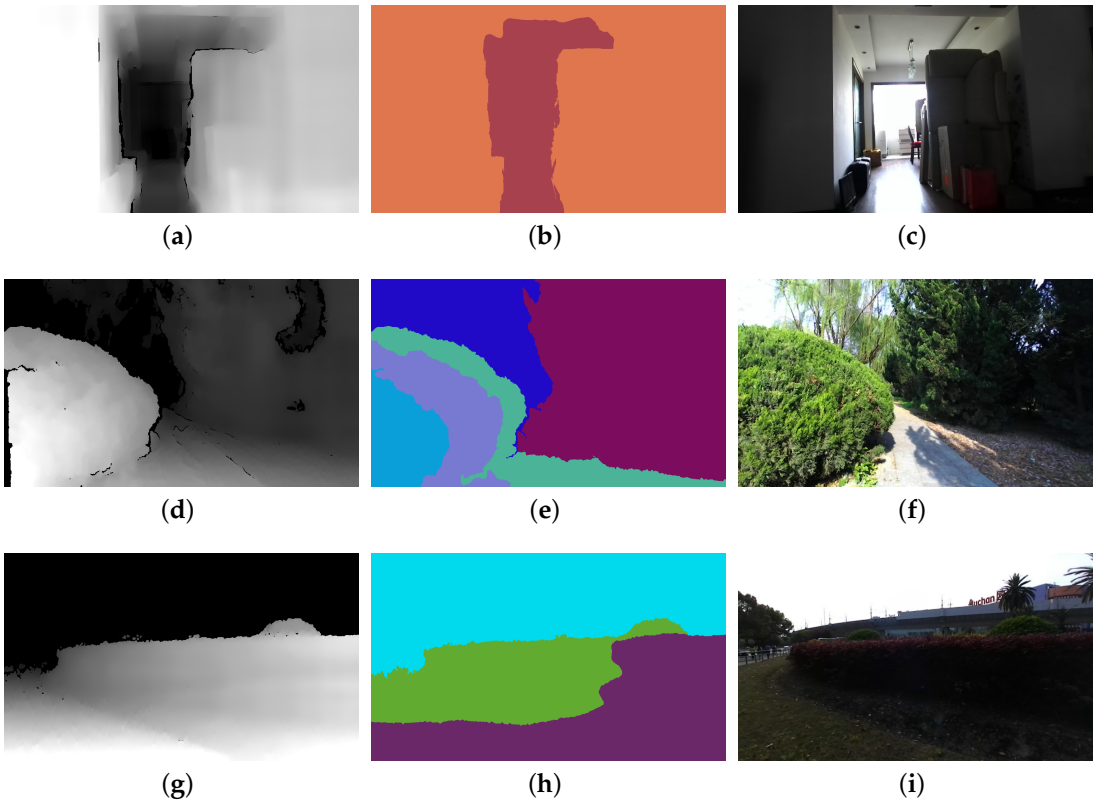


Figure 8. (**a**,**d**,**g**) Depth maps obtained via the use of a ZEDstereo camera. (**b**,**e**,**h**) In-focus image block regions determined by segmenting depth maps. (**c**,**f**,**i**) Corresponding all-in-focus colour images.

5. Conclusions

This paper reported an efficient multi-focus image fusion method assisted by depth sensing. The depth map from a depth sensor was segmented with a modified graph-based segmentation algorithm. The segmented regions were used to guide a focus algorithm to locate an in-focus image for each region from among multi-focus images. The all-in-focus image was constructed by combining the in-focus images selected in each segmented region. The experimental results demonstrated the advantages of the proposed method by comparing the method with other algorithms in terms of six fusion metrics and time consumption. The proposed method enables the construction of an all-in-focus image within 33 ms and provides a practical approach for constructing high-quality all-in-focus images that can potentially be used as reference images.

Supplementary Materials: Supplementary materials are available online at http://www.mdpi.com/1424-8220/19/6/1409/s1. Table S1: The quantitative assessments of the GF method reconstructs a fused image using two different base and detail layers (GF_DIFF) and the same detail layer (GF_SAME), Figure S1: Multi-focus source images captured at five different scenes, the four images in each row belong to a same scene. In each row, the first image is the depth map of the scene, the second image is the segmentation result of the depth map, the front object is in best focus in the third image and the background objects is in best focus in the last image, (a)–(d) belong to the first scene, (e)–(h) belong to the second scene, (i)–(l) belong to the third scene, (m)–(p) belong to the fourth scene, (q)–(t) belong to the fifth scene, Figure S2: Weight maps generated by different methods when fusing different groups of multi-focus source images in Fig. S1. In each row, the weight map from left to right is generated by the proposed method, DCNN, DSIFT, IM and GF respectively. (a)–(e) belong to the first scene, (f)–(j) belong to the second scene, (k)–(o) belong to the third scene, (p)–(t) belong to the fourth scene, (u)–(y) belong to the fifth scene.

Author Contributions: Conceptualization and methodology, H.L. (Hang Liu); software, H.L. (Hang Liu); supervision, H.L. (Hengyu Li) and Y.S.; project administration and funding acquisition, H.L. (Hengyu Li), J.L., S.X. and Y.S.

Funding: This research was funded by the National Natural Science Foundation of China (Grant Numbers 61525305 and 61625304), the Shanghai Natural Science Foundation (Grant Numbers 17ZR1409700 and 18ZR1415300), the basic research project of Shanghai Municipal Science and Technology Commission (Grant Number 16JC1400900), and the Shandong Provincial Natural Science Foundation (Grant Number ZR2017BF021).

Acknowledgments: The authors would like to thank the authors of [25] for having shared their image fusion evaluation toolbox.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Li, S.; Kang, X.; Fang, L.; Hu, J.; Yin, H. Pixel-level image fusion: A survey of the state of the art. *Inf. Fusion* **2017**, *33*, 100–112. [CrossRef]
2. Motta, D.; De Matos, L.; De Souza, A.C.; Marcato, R.; Paiva, A.; De Carvalho, L.A.V. All-in-focus imaging technique used to improve 3D retinal fundus image reconstruction. In Proceedings of the 30th Annual ACM Symposium on Applied Computing, Salamanca, Spain, 13–17 April 2015; pp. 26–31.
3. Nguyen, C.N.; Ohara, K.; Avci, E.; Takubo, T.; Mae, Y.; Arai, T. Real-time precise 3D measurement of micro transparent objects using All-In-Focus imaging system. *J. Micro-Bio Robot.* **2012**, *7*, 21–31. [CrossRef]
4. Liu, S.; Hua, H. Extended depth-of-field microscopic imaging with a variable focus microscope objective. *Opt. Express* **2011**, *19*, 353–362. [CrossRef] [PubMed]
5. Bishop, T.E.; Favaro, P. The Light Field Camera: Extended Depth of Field, Aliasing, and Superresolution. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 972–986. [CrossRef] [PubMed]
6. Gario, P. 4D Frequency Analysis of Computational Cameras for Depth of Field Extension. *ACM Trans. Graph.* **2009**, *28*, 341–352.
7. Iwai, D.; Mihara, S.; Sato, K. Extended Depth-of-Field Projector by Fast Focal Sweep Projection. *IEEE Trans. Vis. Comput. Graph.* **2015**, *21*, 462–470. [CrossRef] [PubMed]
8. Kuthirummal, S.; Nagahara, H.; Zhou, C.; Nayar, S.K. Flexible Depth of Field Photography. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 58–71. [CrossRef]
9. Dowski, E.R.; Cathey, W.T. Extended depth of field through wave-front coding. *Appl. Opt.* **1995**, *34*, 1859–1866. [CrossRef]

10. Yang, Q.; Liu, L.; Sun, J. Optimized phase pupil masks for extended depth of field. *Opt. Commun.* **2007**, *272*, 56–66. [CrossRef]
11. Zhao, H.; Li, Y. Optimized sinusoidal phase mask to extend the depth of field of an incoherent imaging system. *Opt. Lett.* **2008**, *33*, 1171–1173. [CrossRef]
12. Cossairt, O.; Zhou, C.; Nayar, S. Diffusion coded photography for extended depth of field. *ACM Trans. Graph.* **2010**, *29*, 31. [CrossRef]
13. Zhao, H.; Shang, Z.; Tang, Y.Y.; Fang, B. Multi-focus image fusion based on the neighbor distance. *Pattern Recognit.* **2013**, *46*, 1002–1011. [CrossRef]
14. Li, S.; Yang, B.; Hu, J. Performance comparison of different multi-resolution transforms for image fusion. *Inf. Fusion* **2011**, *12*, 74–84. [CrossRef]
15. Wan, T.; Zhu, C.; Qin, Z. Multifocus image fusion based on robust principal component analysis. *Pattern Recognit. Lett.* **2013**, *34*, 1001–1008. [CrossRef]
16. Liang, J.; He, Y.; Liu, D.; Zeng, X. Image fusion using higher order singular value decomposition. *Image Process. IEEE Trans.* **2012**, *21*, 2898–2909. [CrossRef] [PubMed]
17. Li, S.; Kang, X.; Hu, J. Image fusion with guided filtering. *IEEE Trans. Image Process.* **2013**, *22*, 2864–2875. [PubMed]
18. Li, S.; Kang, X.; Hu, J.; Yang, B. Image matting for fusion of multi-focus images in dynamic scenes. *Inf. Fusion* **2013**, *14*, 147–162. [CrossRef]
19. Mckee, G.T. Everywhere-in-focus image fusion using controlablle cameras. *Proc. SPIE Int. Soc. Opt. Eng.* **1996**, *2905*, 227–234.
20. Ohba, K.; Ortega, J.C.P.; Tanie, K.; Tsuji, M.; Yamada, S. Microscopic vision system with all-in-focus and depth images. *Mach. Vis. Appl.* **2003**, *15*, 55–62. [CrossRef]
21. Huang, W.; Jing, Z. Multi-focus image fusion using pulse coupled neural network. *Pattern Recognit. Lett.* **2007**, *28*, 1123–1132. [CrossRef]
22. Xiao-Bo, Q.U.; Yan, J.W.; Xiao, H.Z.; Zhu, Z.Q. Image Fusion Algorithm Based on Spatial Frequency-Motivated Pulse Coupled Neural Networks in Nonsubsampled Contourlet Transform Domain. *Acta Autom. Sin.* **2008**, *34*, 1508–1514.
23. Liu, Y.; Chen, X.; Peng, H.; Wang, Z. Multi-focus image fusion with a deep convolutional neural network. *Inf. Fusion* **2017**, *36*, 191–207. [CrossRef]
24. Liu, Y.; Liu, S.; Wang, Z. Multi-focus image fusion with dense SIFT. *Inf. Fusion* **2015**, *23*, 139–155. [CrossRef]
25. Liu, Z.; Blasch, E.; Xue, Z.; Zhao, J.; Laganière, R.; Wu, W. Objective Assessment of Multiresolution Image Fusion Algorithms for Context Enhancement in Night Vision: A Comparative Study. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 94. [CrossRef]
26. Liu, Y.; Jing, X.Y.; Nie, J.; Gao, H.; Liu, J.; Jiang, G.P. Context-Aware Three-Dimensional Mean-Shift With Occlusion Handling for Robust Object Tracking in RGB-D Videos. *IEEE Trans. Multimed.* **2019**, *21*, 664–677. [CrossRef]
27. Zhao, L.; Bai, H.; Liang, J.; Zeng, B.; Wang, A.; Zhao, Y. Simultaneous color-depth super-resolution with conditional generative adversarial networks. *Pattern Recognit.* **2019**, *88*, 356–369. [CrossRef]
28. Yu, Q.; Liang, J.; Xiao, J.; Lu, H.; Zheng, Z. A Novel perspective invariant feature transform for RGB-D images. *Comput. Vis. Image Underst.* **2018**, *167*, 109–120. [CrossRef]
29. Eichhardt, I.; Chetverikov, D.; Jankó, Z. Image-guided ToF depth upsampling: A survey. *Mach. Vis. Appl.* **2017**, *28*, 267–282. [CrossRef]
30. Yang, Q.; Yang, R.; Davis, J.; Nistér, D. Spatial-Depth Super Resolution for Range Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007.
31. Zhang, S.; Wang, C.; Chan, S.C. A New High Resolution Depth Map Estimation System Using Stereo Vision and Kinect Depth Sensing. *J. Signal Process. Syst.* **2015**, *79*, 19–31. [CrossRef]
32. Khoshelham, K.; Elberink, S.O. Accuracy and resolution of kinect depth data for indoor mapping applications. *Sensors* **2012**, *12*, 1437–1454. [CrossRef]
33. Han, J.; Shao, L.; Xu, D.; Shotton, J. Enhanced computer vision with microsoft kinect sensor: A review. *IEEE Trans. Cybern.* **2013**, *43*, 1318–1334. [PubMed]
34. Vijayanagar, K.R.; Loghman, M.; Kim, J. Real-Time Refinement of Kinect Depth Maps Using Multi-Resolution Anisotropic Diffusion. *Mob. Netw. Appl.* **2014**, *19*, 414–425. [CrossRef]

35. Perona, P. Scale-space and edge detection using anisotropic diffusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **1990**, *12*, 629–639. [CrossRef]
36. Felzenszwalb, P.F.; Huttenlocher, D.P. Efficient Graph-Based Image Segmentation. *Int. J. Comput. Vis.* **2004**, *59*, 167–181. [CrossRef]
37. Sun, Y.; Duthaler, S.; Nelson, B.J. Autofocusing in computer microscopy: selecting the optimal focus algorithm. *Microsc. Res. Tech.* **2004**, *65*, 139. [CrossRef]
38. Li, H.; Manjunath, B.S.; Mitra, S.K. Multisensor Image Fusion Using the Wavelet Transform. In Proceedings of the 1994 International Conference on Image Processing, ICIP-94, Austin, TX, USA, 13–16 November 2002; pp. 235–245.
39. Zhang, Q.; Guo, B.L. Multifocus image fusion using the nonsubsampled contourlet transform. *Signal Process.* **2009**, *89*, 1334–1346. [CrossRef]
40. Hossny, M.; Nahavandi, S.; Creighton, D. Comments on 'Information measure for performance of image fusion'. *Electron. Lett.* **2008**, *44*, 1066–1067. [CrossRef]
41. Wang, Q.; Shen, Y.; Zhang, J.Q. A nonlinear correlation measure for multivariable data set. *Phys. Nonlinear Phenom.* **2005**, *200*, 287–295. [CrossRef]
42. Xydeas, C.S.; Petrovic, V. Objective image fusion performance measure. *Mil. Tech. Cour.* **2000**, *36*, 308–309. [CrossRef]
43. Yang, C.; Zhang, J.Q.; Wang, X.R.; Liu, X. A novel similarity based quality metric for image fusion. *Inf. Fusion* **2008**, *9*, 156–160. [CrossRef]
44. Chen, Y.; Blum, R.S. A New Automated Quality Assessment Algorithm for Image Fusion. *Image Vis. Comput.* **2009**, *27*, 1421–1432. [CrossRef]