Computer Vision and Image Understanding 145 (2016) 148–159

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Robust depth estimation for light field via spinning parallelogram operator

Shuo Zhang[a], Hao Sheng[a,*], Chao Li[a,b], Jun Zhang[c], Zhang Xiong[a]

[a] State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing, PR China
[b] Shenzhen Key Laboratory of Data Vitalization, Research Institute in Shenzhen, Beihang University, Shenzhen, PR China
[c] Department of Electrical Engineering and Computer Science, University of Wisconsin-Milwaukee, Milwaukee 53201, USA

ARTICLE INFO

Article history:
Received 15 January 2015
Accepted 22 December 2015

Keywords:
Light field
Epipolar plane image
Depth estimation
Spinning parallelogram operator

ABSTRACT

Removing the influence of occlusion on the depth estimation for light field images has always been a difficult problem, especially for highly noisy and aliased images captured by plenoptic cameras. In this paper, a spinning parallelogram operator (SPO) is integrated into a depth estimation framework to solve these problems. Utilizing the regions divided by the operator in an Epipolar Plane Image (EPI), the lines that indicate depth information are located by maximizing the distribution distances of the regions. Unlike traditional multi-view stereo matching methods, the distance measure is able to keep the correct depth information even if they are occluded or noisy. We further choose the relative reliable information among the rich structures in the light field to reduce the influences of occlusion and ambiguity. The discrete labeling problem is then solved by a filter-based algorithm to fast approximate the optimal solution. The major advantage of the proposed method is that it is insensitive to occlusion, noise, and aliasing, and has no requirement for depth range and angular resolution. It therefore can be used in various light field images, especially in plenoptic camera images. Experimental results demonstrate that the proposed method outperforms state-of-the-art depth estimation methods on light field images, including both real world images and synthetic images, especially near occlusion boundaries.

1. Introduction

The space of light rays in a scene that intersect a plane from different directions may be interpreted as a 4D light field [1]. It captures not only the accumulated intensity in the scene, but also the direction of each light ray. Hence, the light field can reveal the structure of the scene and is adopted in a wide range of applications, *e.g.* light field rendering [2], super-resolution [3–5], digital refocusing [6] and 3D reconstruction [7].

However, it used to be difficult to manufacture a device to acquire the full light field until the concept of "plenoptic camera" [8] was proposed a few years ago. Using an array of micro-lenses, a single camera is capable of capturing multiple views of the scene simultaneously. More and more this type of specially designed cameras, *e.g.* Lytro [6] and Raytrix [9], have appeared in order to meet the increasing demands in various light field applications.

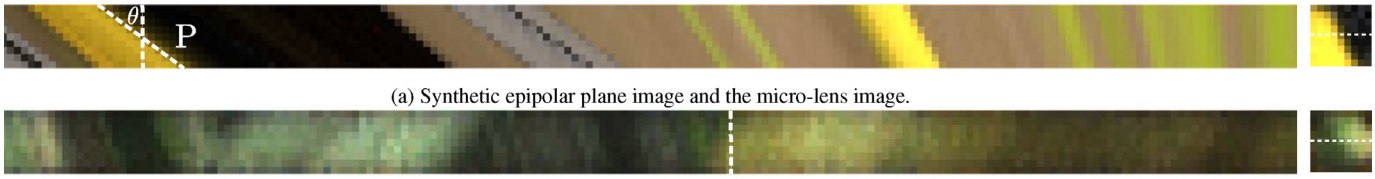
As a critical step in light field image processing, much attention has been paid to efficient and robust algorithms for depth estimation from light fields. On the one hand, light field images can be processed for multiple images from different perspectives of the scene, namely sub-aperture images in light field. Therefore, lots of methods based on multi-view stereo matching [10–12] have been proposed. On the other hand, because of the continuous angular space, some methods based on estimating the slopes of the lines in the Epipolar Plane Image (EPI) [13,14] have been developed. However, these methods have some problems with occluded and noisy regions, especially in plenoptic light field images.

Occlusion has been a tough problem for depth estimation [11,15] both in multi-view and light field images. Occlusion occurs when points near the camera occlude points far away from the camera. As a result, the points far away from the camera are only visible in some sub-aperture images and are occluded by front points in other sub-aperture images. In such case, we cannot rely on finding the points' correct positions in every sub-aperture images to estimate the depth like traditional stereo matching techniques.

Since a plenoptic camera samples the light field and provides angular as well as spatial information on the distribution of light rays in space, the spatial and angular resolutions of the captured images are limited by the hardware. Consequently, images captured by these cameras may be highly aliased due to the sparse

* Corresponding author. fax: +86 010 82338199.
E-mail address: shenghao@buaa.edu.cn (H. Sheng).

http://dx.doi.org/10.1016/j.cviu.2015.12.007
1077-3142/



(a) Synthetic epipolar plane image and the micro-lens image.

(b) Lytro epipolar plane image and the micro-lens image.

Fig. 1. The epipolar plane image (EPI) and the micro-lens image. Compared with Synthetic image, EPIs from plenoptic camera are filled with many noises and aliasing, and have vertical luminance changes due to the circular micro-lens used.

sampling [10] in space. The limited angular resolution, a large depth range and different kinds of noise in the real environment, along with the occlusion problems, reconstruction errors increase in plenoptic camera images. Overall, depth from the plenoptic cameras is difficult to estimate due to the above problems for most methods (some examples can be found in Fig. 1).

In this paper, we propose a novel spinning parallelogram operator (SPO) to locate lines and calculate their orientations in an EPI for local depth estimation. The proposed method measures the slopes by maximizing distribution distances between two parts of the parallelogram window to extract depth information. The spinning parallelogram operator has been demonstrated to be insensitive to occlusion, noise, spatial aliasing, limited angular resolution and EPI specific.

The proposed method also defines a confidence matrix to achieve the maximum utilization of information in horizontal and vertical slices, and calculates the depth values for individual views. In addition, a filter-based method [16], which is less computationally demanding but achieves similar performance as energy-based methods, is introduced to obtain a high-quality global depth map. Based on the confidence metric which is used to gauge reliability, we can obtain the refined depth map.

Experimental results show that our method is effective for both real camera images and synthetic light field images. For plenoptic camera images, including Lytro and Raytrix images, which are highly noisy and aliased, our method outperforms state-of-the-art ones for light field depth estimation and shows the structure of the object more clearly. For the other types of images, which are more similar to multi-view images, our method achieves better performances near the occlusion boundaries and ambiguous regions. Regions with strong specular highlights or with no texture is a typical failure case and further preprocessing or postprocessing is needed.

The main contributions of our work are

1. We propose a spinning parallelogram operator (SPO) for depth estimation on the EPI, which can be applied to both synthetic and real light field camera images. This method is insensitive to occlusion, noise, spatial aliasing, angular resolution, and large depth range.
2. We introduce a confidence metric to integrate information from the rich structure of the light field, and to further reduce the influence of occlusion and ambiguities.
3. We adapt the spinning parallelogram operator to different kinds of datasets and achieve the final refined depth maps.

2. Related work

Due to the different characters of the datasets, the works for depth information in light field images can be divided into two parts. For synthetic light field images, which has less noise and does not have cross-talk artifacts, research has been designed towards solving problems such as the computational speed and occlusion. For real camera images, research has been focused on how to modify and use the imperfect information to recover depth.

To obtain depth information from synthetic light field images, a lot of efforts have been made, among which multi-view stereo matching methods [11,12] and structure tensor based [13,14] methods are notable. Taking the same pixel underneath each micro-lens, we can get sub-aperture images from different angles (*i.e.* multi-view images), so that the traditional stereo matching using multi-view images is available. Yu et al. [12] developed a novel line-assisted graph-cut algorithm that encoded 3D line constraints into light field stereo matching. Chen et al. [11] recently introduced a bilateral consistency metric for light field stereo matching to handle significant occlusion. However, these stereo correspondence methods usually have matching ambiguities in noisy and aliased regions. Moreover, the occlusion problem can only be solved in images with a large baseline and high angular resolution, which means that it is not applicable for real camera light field images.

Another representative work is from Wanner et al. [13,14], who develop a structure tensor based approach to measure each pixel's direction in the EPI, and simultaneously take global visibility constraints into account to make consistent depth map and segmentation approachable. These methods do not try to match pixels in different positions, hence the processing speed has increased. However, because the structure tensor relies on high angular resolution to guarantee that the disparities between neighboring views are less than around two pixels, the method cannot be used in light field images with sparsely sampled view point.

For real camera images, images are highly noisy and aliased along with low angular resolution. Bishop and Favaro [17] used an anti-aliasing filter to avoid cross-talk image artifacts before depth estimation. They formulated a novel photo consistency constraint designed for a plenoptic camera and performed matching directly on the raw data. Tao et al. [18] proposed to combine defocus with correspondence to estimate depth. Sabater et al. [19] developed a demultiplexing method that can acquire image information directly from the undemosaiced raw data and then estimated disparities from the incomplete views. In addition, some works [20–22] focused on calibration and rectification for plenoptic cameras have been proposed to extract sub-aperture images accurately. Jeon et al. [23] designed an algorithm to challenge the real world examples with sub-pixel accuracy. However, these methods have not proposed an effective method for the occlusion problem so that they cannot be extended to more complex real scenes.

Unlike all the above-mentioned work, we propose a novel spinning parallelogram operator (SPO) for depth estimation in different kinds of light field images. Local depth information is extracted by maximizing the differences between the separated region in a parallelogram window. Rules for choosing directions for individual images are then proposed to exploit the redundancy information in the light field. Then a filter-based method [16], which can be parallelized on GPU, is introduced to cover textureless regions and makes the algorithm efficient. Compared to the above methods, the proposed method is able to handle all the light field images and is not limited by the depth range and angular resolution.

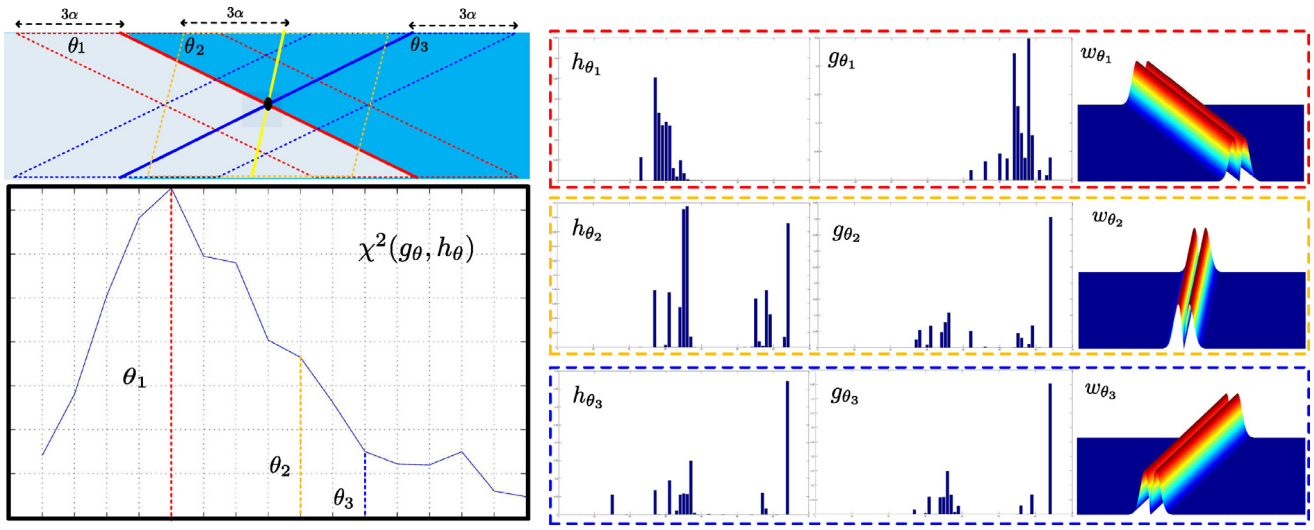


Fig. 2. The spinning parallelogram operator (SPO) separates the EPI into two regions $g(i)$ and $h(i)$ according to different angles θ. The slope of the operator indicates the depth information of the reference point, and can be figured out by maximizing difference $\chi^2(g_\theta, h_\theta)$ between the two distributions of pixel values on each side of the operator. The weighting function $w(i, j)$ is determined by the distance from the point to the center line.

3. Line assignments based on spinning parallelogram operator

In this paper, a 2D plane is used to parametrize the 4D light field, where light through Point P intersects the main lens plane at (u, v) and intersects the micro lens plane at (x, y). The light field is then expressed as $L(x, y, u, v)$.

For light field $L(x, y, u, v)$, if the coordinate (y^*, v^*) is fixed and (x, u) is changed, as shown in Eq. (1), we can get a 2D image $I_{y,v}(x, u)$ called an epipolar plane image (EPI), the same as $I_{x,u}(y, v)$. It can be viewed as a 2D slice of a constant angular image stack in the light field, which reflects the change of the same point in images from different angles.

$$I_{y,v}(x, u) = L(x, y^*, u, v^*). \tag{1}$$

A point in the real world is imaged at different positions in sub-aperture images according to different imaging positions. As shown in Fig. 1, the coordinate x of point P will change if we vary coordinate u, which means the direction of the line in the EPI is able to reflect the depth information. Specifically if we use θ to define the direction of the corresponding line, the actual depth Z can be described according to [13]:

$$Z = f\frac{\Delta u}{\Delta x} = \frac{f}{tan\theta}, \tag{2}$$

where f is the distance between the parallel planes. Angle θ is adopted to reflect the rate of change between Δu and Δx, showing the shift of a scene point moving between the views. The depth estimation is then essentially equivalent to the slope estimation of lines in the epipolar plane images. In this paper, we calculate the disparity θ instead of depth Z.

Limited by hardware, light field images from plenoptic cameras, *e.g.* Lytro, are noisy along with cross-talk artifacts and low angular resolution, as shown in Fig. 1. In addition, the circular microlens used cause luminance changes on the edges of the EPI. These features lead to high matching uncertainties in general methods. Moreover, occlusion and ambiguity regions, as well as the large range of depth, also affect the accuracy of previous methods. In this section, we introduce a new spinning parallelogram operator (SPO) to deal with such situations in EPIs.

3.1. Spinning parallelogram operator (SPO)

We consider how to estimate the local direction of a line at a specific point in an EPI. In EPI based work [14,24], the Canny edge operator or the structure tensor is used to obtain local disparity estimates, which can be regard as computing the weighted mean of each side of the line and then finding the distance between the two means. However, this assumption does not hold well in EPIs because the sides of a line may include other features, like occlusion points and noises. Following the traditional contour detection work [25–27], we assume that the two distributions of pixel values on either side of a line in an EPI are different.

The original compass operator [25] created a circle split into two semicircles with an edge as the diameter. The possible boundary and its orientation at each image pixel is then detected by finding the largest distance of the separated two sides among the different orientations. As the distributions used in the compass operator is able to better describe the two sides of the boundaries than traditional edge detection techniques, the compass operator has been widely applied especially in natural image boundary detection [26] and image segmentation [27].

Unlike natural image, the EPI has “boundaries” which is the imaging pixels of the same point in different views. For a Lambertian scene, the EPI is supposed to contain a set of linear structures, which contains the disparity information of the points. Therefore, the points around a specific line can be used to figure out the orientation of the line. Instead of creating a circle in compass operator, we create a parallelogram operator whose orientation is equal to the possible slope θ, as shown in Fig. 2. As we can imagine that the correct matching points, as well as the line, will separate the parallelogram into two different parts, thus the differences between the two parts can be utilized to predict the orientation of the line.

We define the center point of the parallelogram as the reference point, and then the orientation of the center line that passes through the center point is the disparity to be estimated. The center lines with different orientations, divide the window into two parts of the same size. The correct line, indicating the disparity information, can be figured out by finding the maximum distance between the distributions of pixel values on either side of the lines.

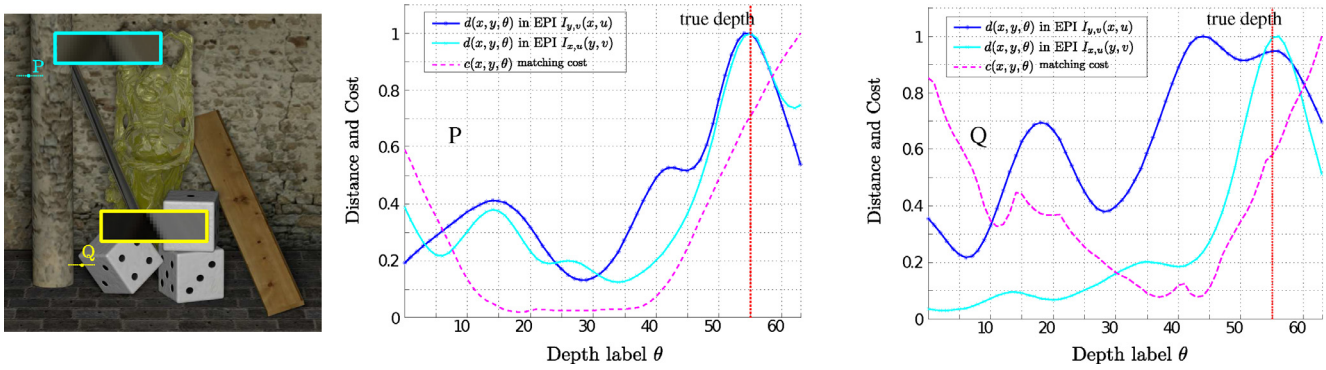


Fig. 3. When the occlusion occurs, the traditional stereo matching algorithm (we use the squared intensity difference as the matching term) produces a large matching cost at the right depth, and estimates the wrong depth at the minimum cost. By contrast, the proposed method picks up the maximum distance in all the possible depth labels. The distances keep at a relatively high value, even if heavy occlusion exists.

First, we define the size of parallelogram operator using a weighted function. The weights in the window are used to measure each pixel's contribution to the distance. We create the weighting function $w(i, j)$ by using the derivative of a Gaussian, as mentioned in Fig. 2. Specifically, for the reference point $P(x_r, u_r)$ in the EPI $I_{y,v}(x, u)$, the corresponding pixels in the defined window are weighted by:

$$w_\theta(i, j) = c \cdot d_\theta(i, j) \cdot e^{\frac{-d_\theta^2(i,j)}{2\alpha^2}}, \tag{3}$$

where $d_\theta(i, j) = i - (x_r + (j - u_r) \cdot tan\theta)$ is the distance between the pixel and the matching point P in the same row, and $(x_r + (j - u_r) \cdot tan\theta)$ is the hypothetical matching line's position in the row j. α is a scale parameter, which is determined by the complexity of the data. c is a normalizing constant. The height of the window is defined as the number of sup-aperture in one angle, *i.e.* angular resolution, and width is 3α.

Notice that we set the weights of the point according to its horizontal distance from the hypothetical matching line so that the point that has the same distance from the line will contribute the same amount of information to the distance measurement. This design separates the window based on the pixel's distance to the hypothetical line. When dealing with large disparities, which contains stair-step structures due to the sparse sampling, the weighted window shows the similar stair-step structure to estimate the slope. Compared with the other slope estimation methods, which need the densely sampling of the large slope, the SPO can be applied immediately.

The weight function approaches zero as we move the location of the pixels parallelly outward, and achieves the maximum value on the close sides of the line, as shown in the last column of Fig. 2. The function finally forms a parallelogram, and the two parallel edges are rotated around the reference point as the angular θ changes. The center line, parallel to the two borders, separates the parallelogram into two regions. Comparing the distribution distance between the two regions, the slope of the true line can be found easily.

3.2. Distance measure

It is a common practice to use distances such as L_2, χ^2 or Earth Mover's distance(EMD) to measure the difference between the distributions of pixel colors. As analysed in [26], we choose to use difference χ^2 of the color histograms:

$$\chi^2(g_\theta, h_\theta) = \sum_i \frac{(g_\theta(i) - h_\theta(i))^2}{g_\theta(i) + h_\theta(i)}. \tag{4}$$

where $g_\theta(i)$ and $h_\theta(i)$ are the histograms of the separated parts. If the χ^2 distance is large, it indicates that the two sides of the parallelogram are different, which means a straight edge exists at the hypothetical matching line. The χ^2 distance can reflect the difference adequately, whose computation is also efficient. Notice that we flip the negative side of the function for better comparison. Actually, we add one positive side to the other negative side to calculate the χ^2 using convolution for a simple and efficient calculation.

3.3. Profile of the spinning parallelogram operator

Traditional stereo matching algorithms estimate depth by searching for the minimum matching cost for the related points. Here we compare the SPO with the pixel-based stereo matching term, *i.e.* squared intensity difference using all aperture images. Specifically, when a point is occluded by another object in some sub-aperture images, the matching point will include the occlusion, which leads to a large matching error at the right depth, as shown in Fig. 3. This means that the useful information in the stereo matching (the low matching cost of the unoccluded point) has been lost. In [11], Chen et al. set a metric to weight the possibilities of the occlusion to exclude the occlusion. However, they need dense angular sampling to guarantee reliability, and tend to be more ambiguous during the matching process.

By contrast, the proposed method picks up the maximum distance between the two regions. It still keeps the correct distance information even if there are ambiguities and occlusion. The reason is that the occlusion will generate a junction in the EPI, which means the right line is intersected by another occlusion line. The distance between the histograms remains at a local maximum value in both orientations. For example, in Fig. 3, Point P and Q is occluded in some views, which causes errors using simple square intensity matching costs in multi-view stereo. However, the distance in the proposed method always remains at a relatively high value at the accurate depth. Theoretically, because the number of pixels of the parallelogram we used in the distance computation is larger than just using the matching points themselves, the insensitivity to occlusion and noises is better.

Additionally, for a EPI $I_{y,v}(x, u)$, as all the points in the EPI have been calculated for all depth labels, the $d_{y,v}(x, u)$ can be used at all points. This means, after computing on all EPIs in a light field images, the algorithm is able to get the depth images of all views.

4. Depth estimation via spinning parallelogram operator

We define $d_{y,v}(x, u, \theta)$ as the histogram distance measured by spinning parallelogram operator on the EPI $I_{y,v}(x, u)$. A large difference between the two halves indicates a discontinuity in the EPI. Hence the local depth estimation can be calculated by taking the maximum response to orientations and a measure of boundary

strength at each pixel is yielded:

$$\Theta_{y,v}(x,u) = arg\max_{\theta} d_{y,v}(x,u,\theta). \quad (5)$$

After calculating $d_{y,v}(x,u,\theta)$ in the horizontal slice image , we can get the corresponding depth map for the reference view $\Theta_{u,v}(x,y)$ by fixing u, v and combining the results. In the same way, $d_{x,u}(y,v,\theta)$ in vertical slice image $I_{x,u}(y,v)$ is calculated. In this section, we develop a confidence based method to interpret the information and utilize a filter-based algorithm to better approximate the true depth.

4.1. Depth estimation on individual view

In this section, we integrate the estimations $d_{x,u}(y,v,\theta)$ and $d_{y,v}(x,u,\theta)$ into a consistent single disparity map. As we have shown, unlike traditional multi-view stereo matching methods, the sampling in angular resolution of light field images is continuous and the angle is diverse. Therefore, a point may be distinct in horizontal slice and ambiguous in vertical slice. Also, it can be occluded in some slices and unoccluded in the other slices. Here we try to extract the credible information among $d_{x,u}(y,v,\theta)$ and $d_{y,v}(x,u,\theta)$ on the basis of different scene structures.

We first define the corresponding confidence c to measure for the reliability of the depth estimation. Unreliable depth estimates, such as the regions with occlusion and ambiguities, always have average scores among all the possible disparities. Specially, for occlusion regions, the distances get high values at both the true disparity and the disparity of the occlusion. For ambiguous regions, the distributions remain the same so that the distances remain at a low value. Accordingly, the confidence c is defined as the difference between the maximum score and the average score:

$$c = exp\left(-\frac{\bar{d}/d_{max}}{2\sigma^2}\right), \quad (6)$$

where $d_{max} = \max_{d} d(\theta)$ and $\bar{d} = \sum_{\theta} d(\theta)$.

The confidence measure c is meaningful as the ambiguous homogeneous regions would have a low $\bar{d}$ and d_{max} and make the confidence c lower. Similarly, the occlusion regions would have a high score in both $\bar{d}$ and d_{max}, where c is low. The example is shown in Fig. 4, where a simple summation cannot achieve the correct information but the weighted sum is able to get the maximum distance at the correct depth. The term σ controls the sensitivity of the confidence and is set as 0.26 in the experiment.

In this paper, we choose different slices according to the borders directions to achieve more accurate results. Specially, we combine the depth scores in accordance with the confidence to integrate the estimations $d_{x,u}(y,v,\theta)$ and $d_{y,v}(x,u,\theta)$, so that the more reliable information can be chosen to reduce the influence of occlusion and ambiguous regions. After obtaining the confidence of each EPI slice, we first normalize the confidences of the same point to keep the relative relation with the other points. Then the weighted summation of $d_{x,u}(y,v,\theta)$ and $d_{y,v}(x,u,\theta)$ is set as:

$$\begin{aligned} d_{u,v}(x,y,\theta) = {} & c_{y,v^*}(x,u^*)d_{y,v^*}(x,u^*,\theta) \\ & + c_{x,u^*}(y,v^*)d_{x,v^*}(y,v^*,\theta). \end{aligned} \quad (7)$$

where the confidence of occlusion regions and ambiguous regions are low. Compared with the general summation, which may introduce the incorrect information from the $d_{y,v^*}(x,u^*,\theta)$ and $d_{x,v^*}(y,v^*,\theta)$, the weighted summation will choose the more credible information for further process, as shown in Fig. 4.

4.2. Depth integration

After obtaining the local depth estimations for the individual points, we further take into account the constraint between the points. In particular, we propagate the correct information with higher confidence to the similar regions with less texture. In order to obtain a global depth map, the most popular method formulated the problem as an energy model and solve it by global approaches such as graph cut [28] or belief propagation [15]. However, both methods are computationally expensive especially with a large number of views and disparity labels. In this section we show that the filter-based method can be used to approximate the global results for synthetic images. For real light field images, further optimizations are needed to obtain a much smoother depth map.

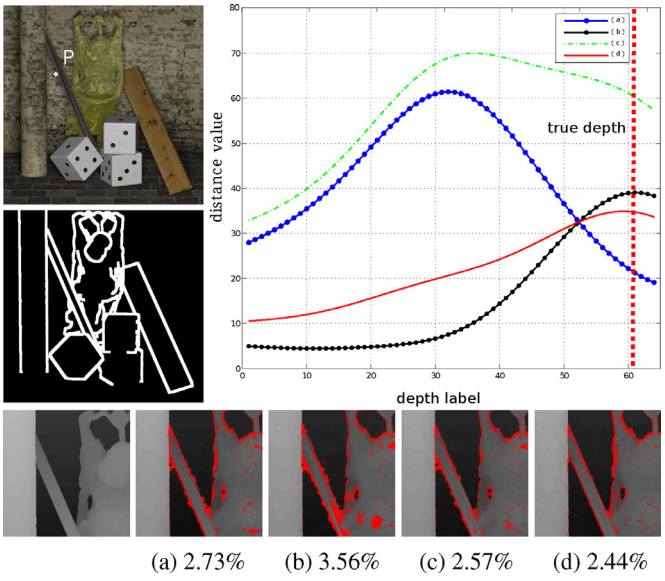
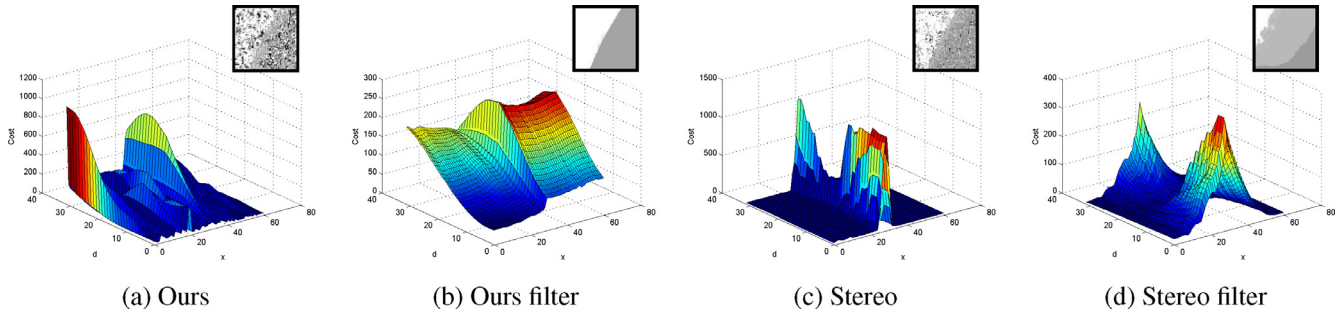


Fig. 4. The image "Buddha" and the occlusion regions that used to calculate the error rate (Occ). The line chart shows the distances calculated by SPO of Point P, where (a) and (b) is $d_{x,u^*}(y,v^*,\theta)$ and $d_{y,v^*}(x,u^*,\theta)$ respectively, (c) is the simple summation, and (d) is the proposed weighted summation. We can find that the summation based on the proposed confidence achieves the maximum at the correct depth label. Results from each step of the proposed method using 5 sub-aperture images after filtering are shown in the bottom with error rates.

The calculated distance volumes at each possible depth label are filtered using guided filter [29], which can perform as an edge-preserving smoothing operator. The weight in the filter is large when the pixel has approximate color and is small otherwise. The guided filter has been widely used in labeling problems [16] for fast filtering of cost volumes. However, as the values are averaged if pixels are in similar color, the occlusion regions, especially the ambiguously occluded ones, will propagate wrong information to other regions with the same color. Specifically, the matching costs of the occlusion regions calculated by traditional stereo matching methods can be high at the correct depth label. In this case, the large costs will propagate to the similar regions and spread the mistakes widely, as shown in Fig. 5. This property makes filter-based method inadequate for the multi-view matching stereo, unless the occlusion are excluded strictly, as [11] does.

By contrast, the proposed depth estimation method depends on the maximum response of the distance. When occlusion occurs, the difference still remains a relatively large value at the right label. The propagation process further averages the differences, and decreases the influence of the occlusion. Unlike stereo matching, which needs extra occlusion estimation, the proposed method is able to remove the occlusion effects automatically during the process.

For synthetic images, the disparities are smooth enough after guided filtering. However, for Lytro images with lots of noises and aliasing, the disparity map is still too noisy. Here, we recommend



(a) Ours (b) Ours filter (c) Stereo (d) Stereo filter

Fig. 5. The performance of guided filter with different local data term settings along the depth discontinuity region. Notice that we choose the maximum distance for the proposed SPO data term and minimum cost for the stereo matching. The corresponding depth map is shown in the top right corner of the figure. This figure shows that the proposed data term is much insensitive to occlusion than the stereo data term, and can be used for filter-based method immediately to achieve the global results.

further using disparity optimization and enhancement method like Jeon et al. [23]. The initial disparity after guided filter is then optimized using weighted median filter [30], graph cuts [31], and finally the iterative refinement [32].

After filtering, a few regions still cannot be propagated perfectly. This happens frequently in regions with devoid of any texture. In order to recover these regions, we create a global confidence measurement as Eq. (6) to determine if a pixel can be reliably labeled or not. If the confidence is sufficiently high, the maximum response of the depth score after filtering is assigned as the depth value. Otherwise, the pixel is marked as unreliable. The marked pixel will be filled according to adjacent pixels based on the same color coherence.

5. Experiment

In this paper, the parameters used in the experiment are first analyzed. Then, the performance of the proposed method on synthetic light field images with available ground truth [33] is evaluated. This kind of images has low noise and does not have cross-talk artifacts, and is more like multi-view images. Our results are compared with the multi-view stereo-based method [31] (MVGC), the bilateral consistency metric designed for removing occlusion [11] (BCM) and tensor structure based method [13] (GCLD), which have shown good performance on the ideal light field images recently. Experiments of images with different angular resolution and different kinds of noise are also performed.

As a quality measurement, we use the square disparity error (MSE), *i.e.* the percentage of depth value below a relative error based on the ground truth as in [33,34]. Specifically, the overall error rate (All) is calculated as $e = \frac{1}{N}\sum(|d(s) - d_T(s)| > \delta_d)$, where s is the pixel we need to evaluate, d and d_T is the corresponding calculated depth and the ground truth respectively. Moreover, the error rate (Occ), which is only measured on the occlusion pixels, is used to observe how well the method is able to handle the occlusion problem clearly. The occlusion regions are calculated based on the ground truth, we extract regions where depth values change sharply as in [11,15]. The example images are shown in Fig. 4.

We also perform depth reconstruction on plenoptic light field images [18,23] of real scenes, which are highly noisy and aliased. Moreover the depth estimations of these images are only reliable on a small number of pixels (less than 7 views per direction). For a better comparison, we compare the disparity after using further optimizations and enhancements like [23]. The proposed method is mainly compared with Tao et al. [18] (CDC), who combined the defocus and correspondence, and Jeon et al. [23] (PSSD), who used the phase shift based sub-pixel displacement. Both methods are designed for plenoptic images and have shown good performance.

For a real especially a complex scenario, the occlusion problem should be taken into adequately account. The high quality of depth map provides credible information for further matting or segmentation. In order to reserve the original structure, the initial matching costs or the matching distances need to keep as much useful information as possible and eliminate occlusions at the same time. In the experiments, we perform detailed comparisons on the depth discontinuous regions in various data sets.

5.1. Optimal parameter selection

In this subsection, the parameter α used in Eq. (3) is investigated. The scale α determines the number of pixels used to calculate the distance, as well as the width of the window $w_\theta(i, j)$. Specifically, pixels within $3 \cdot \alpha$ of the matching point are added to the histogram. As the number of the pixels increases, the disparities are more accurate in unoccluded regions especially in textureless regions, but occlusions are more likely to influence the distance calculation. Optimal parameters are then found by testing a number of different parameters on different data sets. In general, it can be noted that for the synthetic images with many textures, the scale α of 0.8 is always reasonable. Results computed using horizontal EPIs of 5 views with different α are depicted in Fig. 6. Compared the depth map labelled '1' with '4' in Fig. 6, we can find that as the α increases, the ability for occlusion exclusion decreases due to the number of the occluded pixels added. By contrast, the disparity is smoother with a large α especially in textureless regions.

We also evaluate the number of bins, which is used to compute the distances of the color histograms. As we can see in Fig. 6, the precision increases as the number of bins increases and reaches an acceptable value at 64 bins. The histogram separates the pixel into different bins. As the number of bins increases, the similar pixels are easier to be distinguished, as shown in depth map labelled '2' with '3' in Fig. 6. For the synthetic images, the small texture can be detected using more than 64 bins. However, limited by the computational complexity, we cannot increase the number of bins without restrictions. Furthermore, for the images with noise, the pixels coming from the same point in the scene may have different intensities. In this case, the number of bins is supposed to be small in order not to introduce incorrect discrimination to increase the distance. The parameter for noisy images are analyzed in detail in Section 5.4.

5.2. Synthetic images

In this section, we first show detailed visual and quantitative results from each step of the method in Fig. 4. The depth images calculated from two different slices are merged into a single depth map. We choose the weight of each slice based on its confidence. As shown in the line chart, both the distance of point P calculated from $d_{y,v^*}(x, u^*, \theta)$ and the simple summation of the two slices does not keep the maximum value at the correct depth label.

Table 1
The error rate of the estimated depth compared with ground truth (%).

Scene	GCLD [13]		BCM [11]		MVGC [31]		Ours	
	Overall	Occlusion	Overall	Occlusion	Overall	Occlusion	Overall	Occlusion
Cube	**0.92**	13.29	1.11	**9.72**	1.53	12.57	0.98	9.96
Buddha	2.41	15.01	1.72	8.34	3.17	13.64	**1.50**	**7.99**
Horse	15.36	23.02	5.26	9.33	8.37	15.16	**2.40**	**7.39**
Papillon	19.24	24.44	12.86	12.83	15.75	20.83	**5.02**	**7.50**
Mona	12.52	20.44	10.00	16.48	13.17	18.59	**6.53**	**10.52**

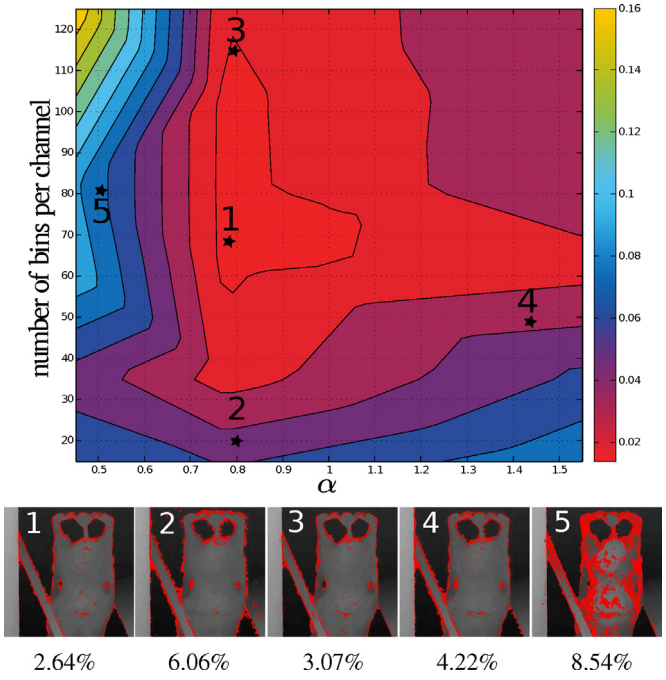


Fig. 6. The evaluation of different parameters on image "Buddha" with 5 sub-aperture images depending on $I_{x,u}(y,v)$ after guided filtering. The error pixels are labels red for distinct comparison. We can observe that the errors rate decreases as the number of bins increases. The scale used in Eq. (3) determines the width of the window and reaches the minimum error rate when $\alpha = 0.8$. The visual disparity images with different combinations of parameters are also shown.

However, it is verified that the confidence correctly measures the reliability of each distance and achieves lower overall error rate in the merged depth map.

Then the comparisons with advanced methods on synthetic light field images with 9 × 9 views are performed. We separate the color channels in 64 bins and set α equal to 0.8, then we discretize the depth map into 64 steps. Table 1 summarizes the accuracies of the proposed method compared with the state-of-the-art techniques for the whole images (All) and occlusion regions (Occ).

We can observe that compared with GCDL [13], our results are better especially in depth discontinuous regions and acquire almost equal error rate overall. Based on the similar theory, GCDL detects the angles of lines in EPIs using structure tensors. Without the quantization process, the depth map is smooth and efficient. The large-scale error is concentrated around depth discontinuities. By way of comparison, the distributions used in SPO is more robust in occlusions. BCM [11] achieves good results when the scene is complex because they use all the views for occlusion estimation before stereo matching. However, they pick up only a part of views with fewer distinctions to compute the stereo cost, so that the error increases when the images have lots of ambiguous regions as shown in images "Horses" and "Papillon", in Fig. 7. On the contrary, the proposed SPO uses the surrounding regions to approach the accurate depth and is able to handle both the occlusion and ambiguity problems. Fig. 8 shows some close-up views for a more clear comparison. For some scenes with complex or trivial structures, our results show comparable results near the occlusion boundaries and acquire a higher accuracy.

5.3. Angular resolution

In this section, we test the proposed method on light field images "StillLife" with different angular resolutions. The numerical results, compared with the state-of-art methods are illustrated in Table 2. It should be noted that method BCM [11], designed for heavy occlusion images, is not effective if they use a small set of views (*e.g.*, 5 ∼ 10). The reason is that the robust metric used to detect the possible occlusion in stereo matching needs densely sampled views to measure the occlusion reliably. The structure tensor based method [13] (GCLD) requires nine views in each angular direction of the light field to acquire high accuracy. What's

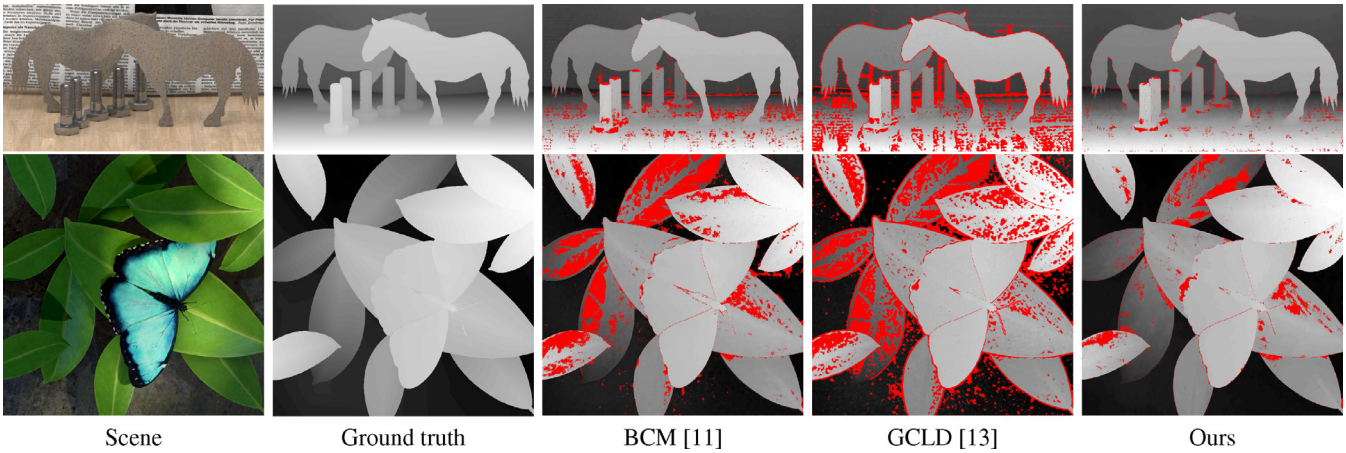


Fig. 7. Depth estimation from synthetic images "Horse" and "Papillon".

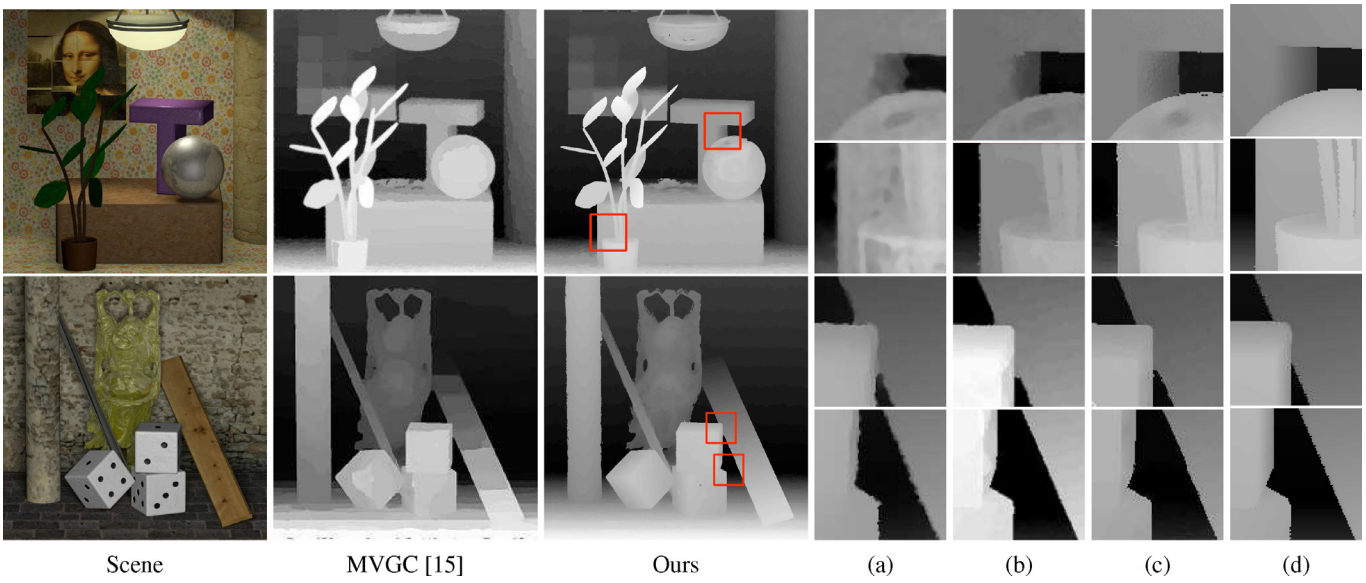


Fig. 8. (a) GCLD [13] (b) BCM [11] (c) Ours (d) Ground Truth. Depth estimation from synthetic images "Buddha" and "Mona ".

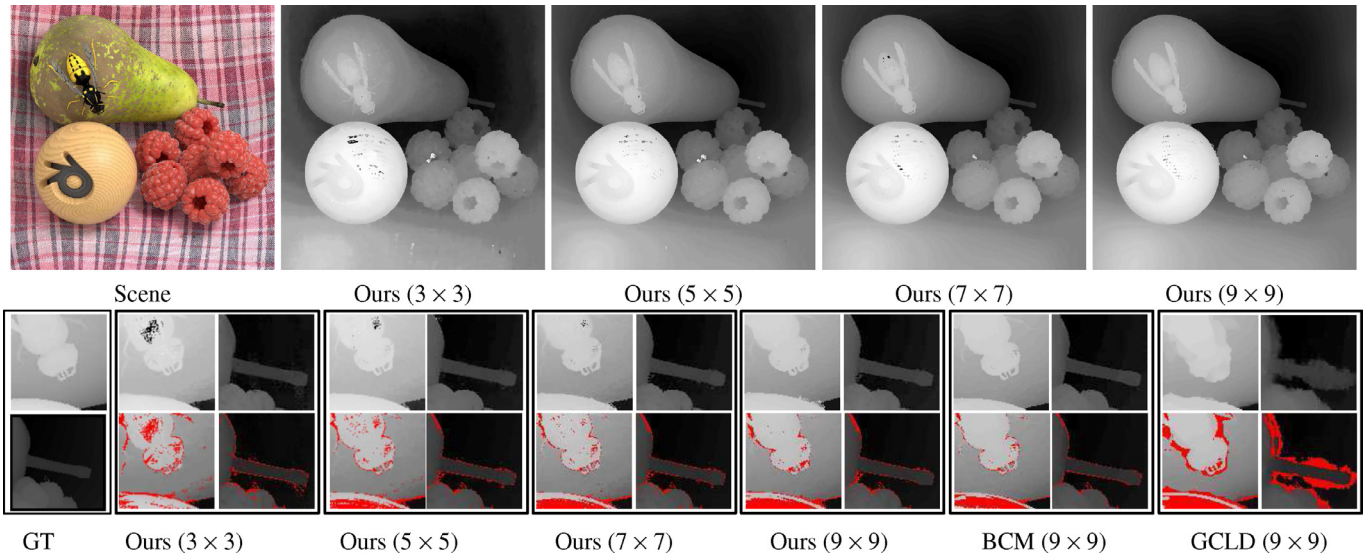


Fig. 9. Depth estimation from synthetic image "stillLife" using different angular resolutions (3–9 views in different directions). The proposed SPO is able to estimate disparities correctly with small angular resolution (less than 7 views) and acquires clear disparity edges in occluded regions.

Table 2
The error rate of the estimated depth compared with ground truth (%).

Angular resolution	GCLD [13]		BCM [11]		Ours	
	All	Occ	All	Occ	All	Occ
9 × 9	4.30	8.35	**1.51**	6.53	1.61	**5.81**
7 × 7	5.64	12.72	2.88	7.83	**1.77**	**5.88**
5 × 5	8.31	19.39	3.07	9.05	**1.69**	**6.41**
3 × 3	–	–	5.59	13.21	**3.18**	**9.08**

more, it requires that the sampling of view points must be sufficiently dense (less than two pixels between neighboring views) because the structure tensors are computed on a 3 × 3 stencil.

In comparison, the number of views in our method is not restricted. The parameter is set as analyzed in Section 5.1. The visual results, shown in Fig. 9, verify that we can acquire the depth map with clear edges even with 3 views in each directions. This quality is quite important for real light field images captured by plenoptic cameras, which have fewer available views for depth estimation because of the hardware limitations. We also verify that we can handle regions with high frequency patterns, like the background of the image "StillLife".

We also show the breaking structure in an EPI when we use only 5 sub-aperture images to estimate large disparities in "StillLife" image. As shown in Fig. 10, the disparities between neighboring views are more than 3 pixels. The corresponding weight shows the similar breaking structure, because we explicitly weight the pixel to two regions according to the distance between the pixel and the matching point P in the same row, as Eq. (3). As a result, we can measure the slope correctly and handle a larger depth range.

5.4. Image with artifacts

In order to verify the robustness of the proposed method, we add different artifacts to the synthetic light field images to

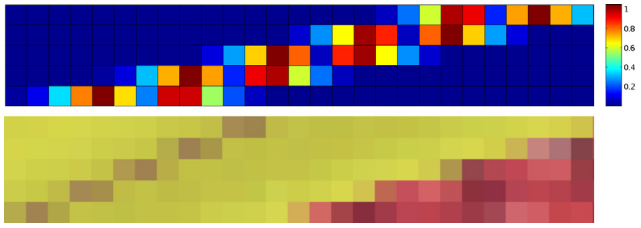


Fig. 10. The spinning parallelogram operator (SPO) on the EPI with large disparities. We can notice that the disparities between neighboring views are more than 3 pixels. The corresponding weight in SPO shows the breaking structure, which is able to fit the large slope and achieves correct depth maps.

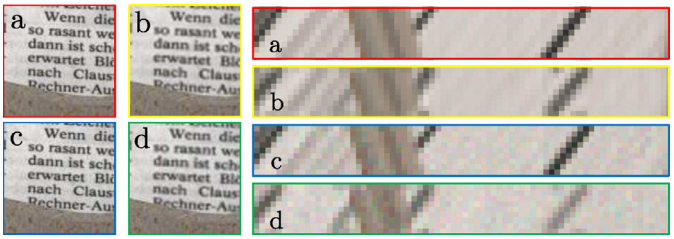


Fig. 11. Light field image "Horse" with different artifacts. (a) original image (b) image with aliasing (c) image with Gaussian noises (d) image with both aliasing and Gaussian noises.

imitate the real environment. To simulate the aliasing in the real light field images, we first reduce the resolution to its 1/2 size and then resize it to its original size using linear interpolation. We implement the resizing operation on the raw image because the images captured by light field camera are arranged in such form. This operation also adds some vignettings, which can emulate the effects of the micro-lenses. The image is then added Gaussian noise with zero mean and a variance equal to 2% of the image dynamic range as in [35]. Example noisy images are shown in Fig. 11.

We first evaluate the parameter, *i.e.* the number of bins, on the noisy images. As we mentioned in Section 5.1, the error rate decreases as the number of bins increase. However, when the noise is added, the error rate increases as more bins are used to calculate the distance, especially for ambiguous regions. The reason is that if we separate the regions into small range bins, the noisy pixels are more likely to be assigned to the wrong bin. Consequently, the noisy pixels affect the distance measurement and make it hard for the distances to achieve the maximum value at the right depth. According to the above analysis, we separate the regions into 16 ∼ 48 bins in this experiment. The quantitative results are illustrated in Table. 3.

Table 3
The error rate of the estimated depth compared with ground truth (%).

Bins	Original		Gaussian		Artifacts		Both	
	All	Occ	All	Occ	All	Occ	All	Occ
16	3.88	12.76	6.424	18.89	5.23	17.30	**10.11**	23.54
32	2.79	8.10	**6.01**	**16.19**	3.56	11.68	11.08	21.50
48	2.42	7.75	6.92	19.63	**3.37**	**10.60**	12.60	**21.19**
64	**2.40**	**7.39**	12.69	21.94	3.71	10.80	22.27	25.39

The qualitative and quantitative results are compared with BCM [11] and PSSD [23] as shown in Fig. 12. The proposed method is able to handle such noise and aliasing, and achieves the acceptable error rate, which means the SPO is robust to this kind of noise. If the noises expand, we can expect further denoising, optimization and refinement, such as TV-L^1 denoising and graph cuts [31], as we do with Lytro images in the next section.

5.5. Lytro images

In this section, the performance of our method on Lytro images is evaluated. As shown in Fig. 13, the sub-aperture images obtained from the Lytro camera are highly noisy and aliased. Therefore the depth image after guided filtering (GF) is still noisy, which means the depth is not adequately propagated. For better showing the effectiveness of the proposed SPO, we implement the further optimization (FO) as Jeon et al. [23]. We show the example images calculated from each step of our method in Fig. 13. The scale α is set to 1 ∼ 3 for the images with more textureless regions and the number of bins is set to 36 ∼ 48 due to the noisy environment.

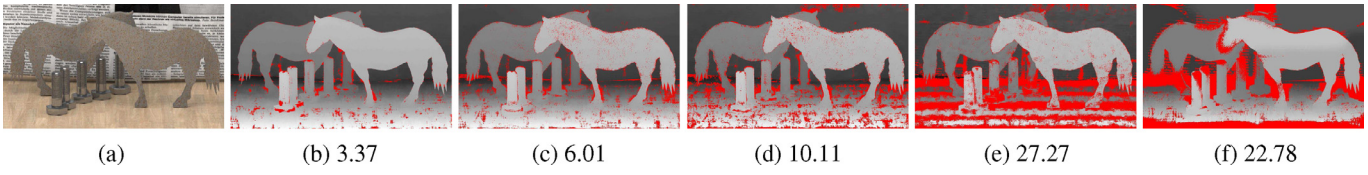


Fig. 12. Depth estimation on image "Horse" with different artifacts along with the corresponding error rates. The error pixels are labels red for distinct comparison. (a) Shows the image with both aliasing and Gaussian noises. (b) And (c) Shows our results on image with aliasing or Gaussian noises, respectively. (d–f) Shows the depth estimation on image with both aliasing and Gaussian. (d) Ours (e) BCM (f) PSSD.

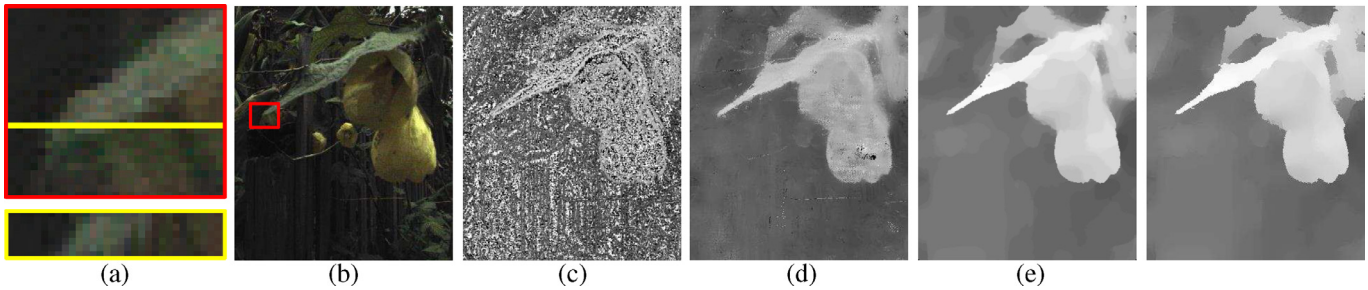


Fig. 13. Depth estimation for Lytro images at each step of our method. (a) The sub-aperture image. (b) The depth map based on the initial SPO distance described in Section 4.1. (c) The depth map after guided filter refinement in Section 4.2. (d) The depth map after the multi-label optimization. (e) The depth map after iterative refinement. The processes in (d) and (e) can be found in [23].

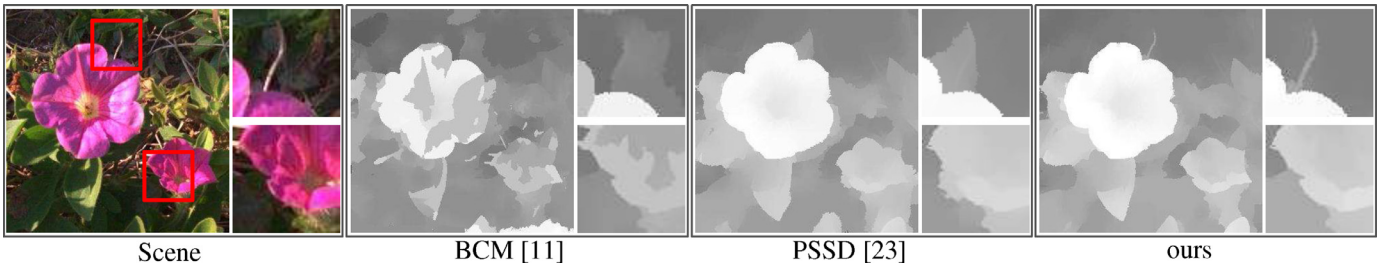


Fig. 14. Zoom-up disparity images of the distortion-corrected Lytro images. The proposed method is able to better retain the structure of the scene than PSSD, CDC and BCM.

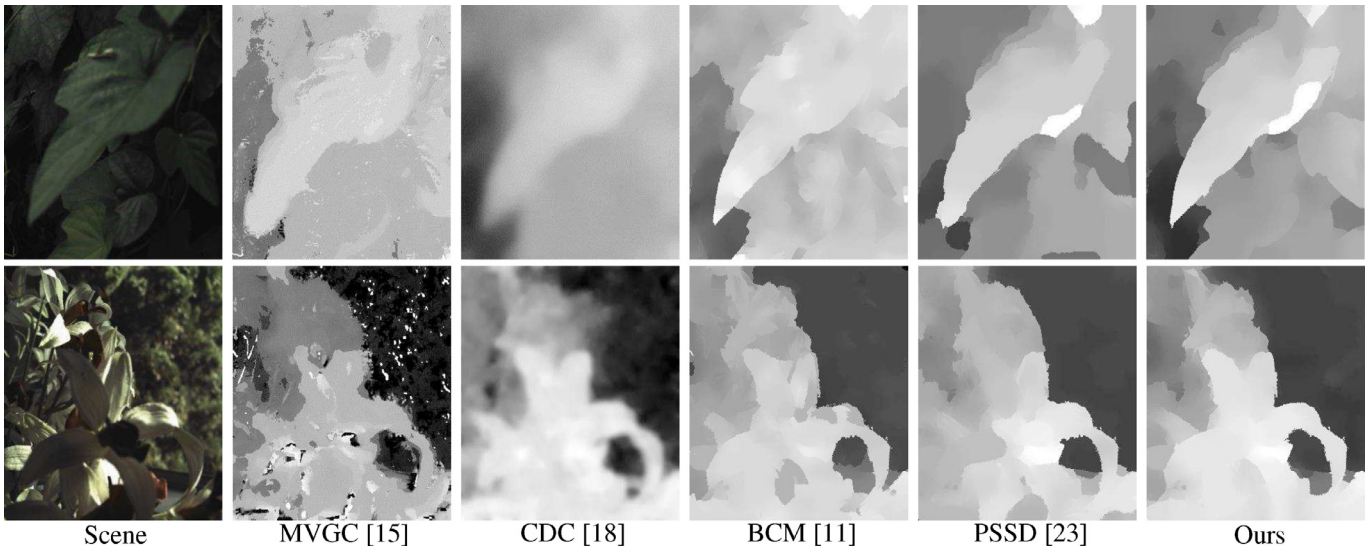


Fig. 15. Depth estimation for Lytro images using different methods.

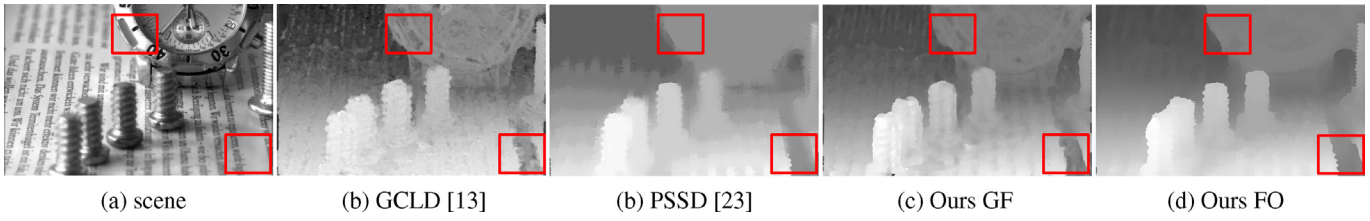


Fig. 16. Depth estimation on raytrix image "Watch" with specular and textureless regions. The error pixels are labels red for distinct comparison. (c) Ours after guided filtering. (d) Ours after further optimization.

Fig. 14 shows the light field images after distortion estimation and correction (Jeon et al. [23]). We compare the results with Jeon et al. [23] (PSSD) who developed a phase shift based sub-pixel displacement to calculate the cost volume. We also try the stereo matching based method [11], which is designed for heavy occlusion images on the light field images. We find that the bilateral consistency metric they used is invalid due to noises and aliasing. Even though the depth edges are clear, the depth is not smooth on flat regions. Another relevant study is from Tao et al. [18] (CDC), who used both stereo and defocus cues to better approximate the true depth maps of Lytro images. Using the distortion-corrected Lytro images and the same depth optimization method as PSSD, we can observe that the proposed method is able to better retain the structures of the scene (Fig. 14) than PSSD and CDC.

More comparative results are illustrated in Fig. 15. Clearly, our depth map has clear edge features and reflects the structure of the object in the scene. For algorithms designed for synthetic images, such as MVGC and GCLD, the estimated depth and confidence is influenced by noises and aliasing so that structure information is lost during the calculation. Moreover, a small number of views here also influences the depth estimate and confidence measures.

5.6. Raytrix images

In this section, we show experimental results with light field images captured by a Raytrix camera [9]. The image "Watch" shows some regions with highlight specular and textureless. Comparisons with GCLD [13] and PSSD [23] are presented in Fig. 16. We can find that the depth estimation is stable except on points with highlight specular or without texture, as shown in the red boxes in Fig. 16. The specular intensity changes as the view changes and does not form a line in the EPI, so that the traditional stereo matching and EPI based methods cannot detect the disparity. As a result, for scenes which are not Lambertian, we need to estimate and remove the specular components as [36] first. For textureless regions, we rely on the further integration and optimization to find the true disparity.

Table 4
The run time analysis on Lytro images.

Method	Local estimation				Optimazation	
	Ours	PSSD	CDC	GCDL	GF	FO
Time	65*s*	231*s*	121s	<1s	63s	75s

5.7. Run time analysis

In order to reduce the complex of the method, we process each histogram bin separately as [27]. Let I denote the EPI, and let $I^b(x, y)$ be 1 if $I(x, y)$ falls in histogram bin b and 0 otherwise. Then the parallelogram window is treated as a convolution kernel and implemented on the histogram images. In particular, estimating the depth on an EPI with B bins, N channels and D depth label takes NBD convolution operations. Then the distance summation between the histogram images is computed.

The run time for local depth estimation of the proposed method and the other state-of-art techniques is illustrated in Table. 4. The variables are selected for Lytro images with 64 depth label and 3 channels. The number of bins is set as 48 in our method. All the algorithm are implemented in Matlab using a computer equipped with Intel i7 3.60 GHz CPU and 8 GB RAM. Notably, the entire local cue computations and the guided filtering in our method can be easily parallelized using GPUs and we can refer to [27] and [29] for further analyses.

6. Conclusion

Taking into account the special structure of the light field data, we propose a novel local depth estimation method based on a spinning parallelogram operator. The proposed algorithm uses the EPI to find the corresponding line and sets local and global confidence to deal with occlusions. Then we have designed a rule to choose the information for individual views during the depth scores calculation, which can further handle occlusions. The problem has then been formulated as a discrete labeling problem and solved using a filter-based algorithm. Compared with the state-of-the-art stereo matching methods, tensor structure based methods and methods designed for Lytro images, the proposed method is more robust to noise, artifacts, occlusion. Experimental results show that our method performs excellently in Lytro images, especially near occlusion boundaries. Moreover, the algorithm is less affected by the limitation of depth range and angular resolution. Future work shall be devoted to incorporating different sets in different cameras for further improvement in depth estimation process, and in extending this method to various applications.

Acknowledgments

This study was partially supported by the National Natural Science Foundation of China (No. 61472019), the National High Technology Research and Development Program of China (No. 2013AA01A603) and the National Science & Technology Pillar Program (No.2015BAF14B01). Supported by the Programme of Introducing Talents of Discipline to Universities and the Open Fund of the State Key Laboratory of Software Development Environment under grant #SKLSDE-2015KF-01

References

[1] M. Levoy, P. Hanrahan, Light field rendering, in: ACM Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, 1996, pp. 31–42.
[2] L. McMillan, G. Bishop, Plenoptic modeling: an image-based rendering system, in: ACM Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques, 1995, pp. 39–46.
[3] T.E. Bishop, S. Zanetti, P. Favaro, Light field superresolution, in: Computational Photography (ICCP), 2009 IEEE International Conference on, IEEE, 2009, pp. 1–9.
[4] T.E. Bishop, P. Favaro, The light field camera: Extended depth of field, aliasing, and superresolution, IEEE Trans. Patt. Anal. Machine Intell. 34 (5) (2012) 972–986.
[5] S. Wanner, B. Goldluecke, Spatial and angular variational super-resolution of 4d light fields, in: Computer Vision–ECCV 2012, Springer, 2012, pp. 608–621.
[6] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, P. Hanrahan, Light field photography with a hand-held plenoptic camera, Comput. Sci. Technical Report CSTR 2 (11) (2005).
[7] M. Levoy, Light fields and computational imaging, IEEE Comput. 39 (8) (2006) 46–55.
[8] E.H. Adelson, J.Y. Wang, Single lens stereo with a plenoptic camera, IEEE Trans. Patt. Anal. Mach. Intell. 14 (2) (1992) 99–106.
[9] C. Perwa, L. Wietzke, The next generation of photography, http://www.raytrix.de.
[10] T.E. Bishop, P. Favaro, Plenoptic depth estimation from multiple aliased views, in: Proceedings of IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops), IEEE, 2009, pp. 1622–1629.
[11] C. Chen, H. Lin, Z. Yu, S.B. Kang, J. Yu, Light field stereo matching using bilateral statistics of surface cameras, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.
[12] Z. Yu, X. Guo, H. Ling, A. Lumsdaine, J. Yu, Line assisted light field triangulation and stereo matching, in: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2792–2799.
[13] S. Wanner, B. Goldluecke, Globally consistent depth labeling of 4d light fields, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 41–48.
[14] S. Wanner, B. Goldluecke, Variational light field analysis for disparity estimation and super-resolution, Trans. Patt. Anal. Mach. Intell. IEEE 36 (3) (2014) 606–619.
[15] J. Sun, N.-N. Zheng, H.-Y. Shum, Stereo matching using belief propagation, IEEE Trans. Patt. Anal. Mach. Intell. 25 (7) (2003) 787–800.
[16] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, M. Gelautz, Fast cost-volume filtering for visual correspondence and beyond, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 3017–3024.
[17] T.E. Bishop, P. Favaro, Full-resolution depth map estimation from an aliased plenoptic light field, in: Proceedings of Asian Conference of Computer Vision (ACCV) 2010, Springer, 2011, pp. 186–200.
[18] M.W. Tao, S. Hadap, J. Malik, R. Ramamoorthi, Depth from combining defocus and correspondence using light-field cameras, in: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2013, pp. 673–680.
[19] N. Sabater, V. Drazic, M. Seifi, G. Sandri, P. Perez, Light-field demultiplexing and disparity estimation, technical report: hal-00925652 (2014).
[20] D.G. Dansereau, O. Pizarro, S.B. Williams, Decoding, calibration and rectification for lenselet-based plenoptic cameras, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, IEEE, 2013, pp. 1027–1034.
[21] D. Cho, M. Lee, S. Kim, Y.-W. Tai, Modeling the calibration pipeline of the lytro camera for high quality light-field image reconstruction, in: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2013, IEEE, 2013, pp. 3280–3287.
[22] Y. Bok, H.-G. Jeon, I.S. Kweon, Geometric calibration of micro-lens-based light-field cameras using line features, in: Computer Vision–ECCV 2014, Springer, 2014, pp. 47–61.
[23] H.-G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y.-W. Tai, I.S. Kweon, Accurate depth map estimation from a lenslet light field camera, in: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2015, IEEE, 2015, pp. 1547–1555.
[24] A. Criminisi, S.B. Kang, R. Swaminathan, R. Szeliski, P. Anandan, Extracting layers and analyzing their specular properties using epipolar-plane-image analysis, Comput. Vision Image Understanding 97 (1) (2005) 51–85.
[25] M.A. Ruzon, C. Tomasi, Color edge detection with the compass operator, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2, 1999.
[26] D.R. Martin, C.C. Fowlkes, J. Malik, Learning to detect natural image boundaries using local brightness, color, and texture cues, IEEE Trans. Patt. Anal. Mach. Intell. 26 (5) (2004) 530–549.
[27] P. Arbelaez, M. Maire, C. Fowlkes, J. Malik, Contour detection and hierarchical image segmentation, IEEE Trans. Patt. Anal. Mach. Intell. 33 (5) (2011) 898–916.
[28] Y. Boykov, O. Veksler, R. Zabih, Fast approximate energy minimization via graph cuts, IEEE Trans. Patt. Anal. Mach. Intell. 23 (11) (2001) 1222–1239.
[29] K. He, J. Sun, X. Tang, Guided image filtering, in: European Conference on Computer Vision (ECCV), Springer, 2010, pp. 1–14.
[30] Z. Ma, K. He, Y. Wei, J. Sun, E. Wu, Constant time weighted median filtering for stereo matching and beyond, in: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2013, IEEE, 2013, pp. 49–56.
[31] V. Kolmogorov, R. Zabih, Multi-camera scene reconstruction via graph cuts, in: European Conference on Computer Vision (ECCV), Springer, 2002, pp. 82–96.
[32] Q. Yang, R. Yang, J. Davis, D. Nistér, Spatial-depth super resolution for range images, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07., IEEE, 2007, pp. 1–8.

[33] S. Wanner, S. Meister, B. Goldluecke, Datasets and benchmarks for densely sampled 4d light fields, in: Vision, Modeling & Visualization, The Eurographics Association, 2013, pp. 225–226.
[34] D. Scharstein, R. Szeliski, A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, Int. J. Comput. Vision 47 (1-3) (2002) 7–42.
[35] S. Heber, R. Ranftl, T. Pock, Variational shape from light field, in: Energy Minimization Methods in Computer Vision and Pattern Recognition, Springer, 2013, pp. 66–79.
[36] M.W. Tao, T.-C. Wang, J. Malik, R. Ramamoorthi, Depth estimation for glossy surfaces with light-field cameras, in: Proceedings of European Conference on Computer Vision Workshops on Light Fields Computer Vision (ECCV Workshops), Springer, 2014.