Segment-Based Depth Estimation in Light Field Using Graph Cut

Wenjie Shao[1*], Hao Sheng[1,2], Chao Li[1,2]

[1]State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing 100191, P.R.China
[2]Shenzhen Key Laboratory of Data Vitalization, Research Institute in Shenzhen, Beihang University, Shenzhen 518057, P.R.China

Abstract. In this paper, we present a depth-extracting method on the scenes of 4D light fields. The method is based on image segmentation and epipolar plane images. We extract disparity map and reliability map from the original image of 4D light fields. Then this information is applied to image segmentation, in which a large number of planes are produced, so that the disparity map which is consist of pixels can be transferred to the disparity map which is consist of planes. In the resulting optimization problem, graph-cut technique is used to assign a corresponding disparity plane to each segment. Our method is tested on a number of synthetic and real-world examples captured with a light field camera, and compared to ground truth where available. Furthermore, an approach to optimize the method to reduce the running time is also proposed.

Keywords: Light fields; Depth Estimation; Graph Cuts; Imgae Segmentation;

1 Introduction

The 4D light field has been established as a promising paradigm to describe the visual appearance of a scene. Compared to a traditional 2D image, it contains not only the accumulated information at each image point, but also separate information for each ray direction. With such additional information, a wide range of applications in a light field have been developed. They achieved some functions which can't be applied in ordinary camera. Digital zooming, also called refocusing, is an important use[7]. It can invert a defocused image to highly accurate focused image and reduce the difficulty of auto focusing. Super-resolution reconstruction is another active research in 4D light field.

More importantly, reconstruction from 4D light fields requires getting an accurate disparity map, which is still a challenging in computer vision. There are many mature methods in depth estimation, such as stereo matching[16]. They can be divided into two major classes, local algorithms and global algorithms. Both of them have some disadvantages. Local algorithms only uses the pixel which is in a finite neighboring window so that it has low accuracy. Global algorithms make explicit smoothness assumptions of the disparity map and they

require a lot of time in the minimization techniques[9]. As is commonly known, the real scene structure could be approximated by a set of planes in disparity space. After image segmentation, every region could be fitted to a plane formula using pixel locations and local depth estimates. However, a common scene could be divided into thousands of regions after Mean shift algorithm and it will cost a lot of time to give every region a fitting plane, and more importantly, a large number of the planes are inaccurate[13]. In our method, several measures are used to choose the accurate planes. For example, a precise plane must be big enough which means it has a lot of points and the points with high reliability must occupy most of all points in this plane[2]. After plane fitting, the problem is converted into assigning labels to planes which can be easily formulated as an energy minimization problem in the segment domain.

To solve the energy minimization problem, graph cuts technique and Markov random fields is used. Li Hong and George Chen proposed a graph cuts method in stereo matching based on separated pixels[6]. In general, the number of segments is much less than pixels, which leads to a simple graph structure and fast computation. But because of our over-segmentation, the runtime tends to exceeded our expectation. We apply an optimism method to reduce the number of regions. We combine those regions whose plane equation is similar to a large region. According to our experiments, this approach will reduce the runtime by thirty percent .

This paper proposes a region-based global algorithm to obtain the disparity map in 4D light fields. It makes full use of image information and produced much better results than stereo matching[4]. More importantly, it greatly reduces the running time which is a big problem for global algorithms, especially in 4D light fields. It combines the traditional image segmentation with light fields.

The rest of the paper is organized as follows: First we introduce the related work on image segmentation in Section 2. Then we present how to obtain disparity map in the 3D light field (Section 3) and how to give every region produced by Meanshift a plane label (Section 4). In Section 5 we apply graph cuts to solve the energy minimization problem which assign the corresponding disparity plane to each segment. We provide experimental results in Section 6. Finally, we conclude our paper and discuss related advantages and disadvantages of our approach in Section 7.

2 Related work

The concept of light fields mainly came from computer graphics. One of the first approaches using *EPIs* to analyze the scene geometry was by Bolled et al.[10]. To reconstruct the 3D structure, they detect edges, peaks and troughs with a subsequent line fitting in the *EPI*. Another approach is presented by Criminisi et al.[2], who use an iterative extraction procedure for collections of EPI-lines of the same depth, which they call an *EPI*-tube. They also proposed procedure to remove specular highlights from already extracted *EPI*-tubes.

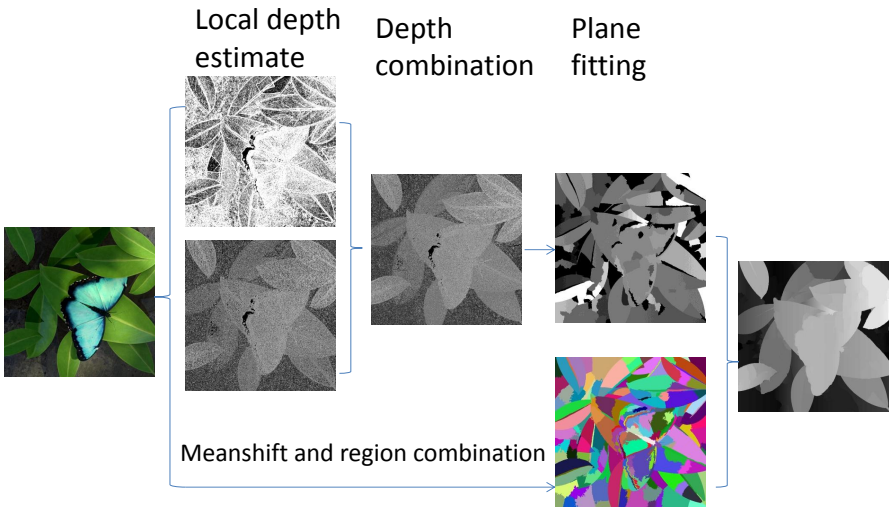

Fig. 1. The whole process of our method, from the original picture to the final depth image.We extract disparity map and reliability map from the original image of 4D light fields. Then we apply these information on image segmentation, in which we get accurate segments in a large number of planes, so that we can get transfer the disparity map which is consist of pixels to the disparity map which is consist of planes.

Sven Wanner and Bastian Golduecke introduced a novel local data term for depth estimation[13], which is tailored to the structure of light field data. They use the coherence of the structure tensor as the reliability measure and the direction of the local level lines to obtain the depth estimated.

Image segmentation is an active research area, especially in stereo matching. Andreas Klaus, Mario Sormann and Konrad Karner proposed segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure[3]. They apply Mean Shift color segmentation to image segmentation which is first used in Comaniciu and Meer. The main advantage of the mean-shift approach is the fact that edge information is incorporated as well.

Christoph Straehle and Sven Wanner gave an assignment of globally consistent multi-labels on the 4D light fields[15]. They provided an optimal data structure for label optimization by implicitly providing scene geometry information. It is thus possible to consistently optimize label assignments over all views simultaneously.

In Matousek et al.[6], a cost function is formulated to minimize a weighted path length between points in the first and the last row of an EPI, which prefers constant intensity in a small neighborhood of each EPI-line. However, their method only works in the absence of occlusions.Berentetal.[8] deal with the si-

multaneous segmentation of EPI-tubes by a region competition method using active contours, imposing geometric properties to enforce correct occlusion ordering.

Multi view object co-segmentation is similar to light field image segmentation. Adrsh Kowdle and Sudipta N.Sinha proposed an algorithm formulated using an energy minimization framework that combines stereo and appearance cues, where for each surface; an appearance model is learnt using an unsupervised approach[1]. Dorit S.Hochbaum proposed an efficient algorithm for Co-segmentation on object which is in two images with arbitrary background[14]. To solve the difficult optimization problem of Markov Random Field (MRF), they designed a new algorithm which can solve for optimal in polynomial time using a maximum flow procedure on an appropriately constructed graph[11].

3 Local depth estimate in 3D light field

As mentioned before, *EPI*, which can be viewed as 2D slices of constant angular and spatial directions could be obtained through the Lumigraph from 3D light field. There are several ways to represent light fields. In this paper, we adopt the two-plane parametrization. It can be treated as a map

$$L : \Omega \times \Pi \to \mathbb{R}, \qquad (x, y, s, t) \mapsto L(x, y, s, t) \tag{1}$$

It can be viewed as an assignment of an intensity value to the ray $(R_{x,y,s,t})$ passing through $(x, y) \in \Omega$ and $(s, t) \in \Pi$. Fix a horizontal line of constant (y) in the image plane and a constant camera coordinate (t), and restrict the light field to an (x,s) slice Σ_{y^*,t^*}. The *EPI* could be described as.

$$\begin{aligned} S_{y^*,t_*} : \Sigma_{y^*,t^*} \to \mathbb{R}, \\ (x, s) \mapsto S_{y^*,t^*}(x, s) \coloneqq L(x, y*, s, t^*) \end{aligned} \tag{2}$$

The next step is to get the depth of every pixel. In the *EPI*, a pixel is a slant. According to its depth in the scene, the slant slope is different. The farther the object, more inclined the line is. There is a very simple formula to calculate its slope.

$$l = -f\frac{\Delta s}{\Delta x} \tag{3}$$

However there are several problems in this formula. The first is some of the lines which are corresponding to a fixed pixel whose texture is not clear enough is very hard to be distinguished from many lines. So the depth could be wrong. Another problem is the *EPI* will contain some lines which are not long enough, which means there are some obstructions blocking the pixel. It could be solved by only using part of the line to calculate despite that it will reduce the accuracy.

Fortunately there is a measure of accuracy. Using the coherence of the structure tensor to calculate reliability is advisable. It can effectively reflect the accuracy of the depth estimate of every pixel.

$$r_{y^*,t^*} \coloneqq \frac{(J_{xx} - J_{yy})^2 + 4{J_{xy}}^2}{(J_{yy} + J_{xx})^2} \tag{4}$$

Using the local depth estimates for all the *EPIs* in horizontal and vertical directions, respectively, different view points could produce different disparity maps and reliability maps. For example, for the original picture from a light field camera which produces 9×9 photos, using the local depth estimates we can get 9×9 *EPIs*, it is necessary to choose how to combine such information into one disparity map. Paper[5] used an algorithm which is similar to stereo matching. It is very complex and for some scenes which have little texture, its performance is not good enough. So our method only apply a very simple method. It gives up some information and only uses two *EPIs*. Obviously both of them are obtained from the same view point. A simple method is to combine them with a comparison of whose reliability is larger.

Fig. 2. Combination of two disparity maps in different directions. The left is the horizontal disparity map. The middle is vertical disparity map. The right is the result.

As indicated in Fig. 2, the horizontal *EPI* is sensitive to horizontal edges because the corresponding reliability is bigger in horizontal direction. The vertical is similar. After combining them, both directions seem clear.

4 Plane segmentation and fitting

In the above steps, texture is used to get rough depth information. The next step is to combine these pixels into several regions. Color is a good feature for segmentation. Our method uses Mean Shift to split the original picture, which is corresponding to the center image in light field, into small regions. Because of the view point of the center image is the same as the disparity map and the reliability map, the deviation could be reduced from parallax maximally. In this part, an over-segmented image is preferred and these tiny regions will be combined at last. It may cost extra time for the merging step. But if some region boundaries are not boundaries in the segmented image, nothing can be done to correct this error. The mean-shift analysis approach is essentially defined as a gradient ascent search for maxima in a density function defined over a high dimensional feature space. To get the over-segmentation image, the spatial radiation and color radiation of Mean Shift need to be a small value, especially the color radiation, when the image is almost a same color.

After plane segmentation, thousands of, maybe more, regions are produced. The next step is to specify a fitting plane to each region with the disparity and

reliability from last part. Every plane can be describe as

$$ax + by + cz + 1 = 0 \tag{5}$$

Where x and y are the coordinates of a pixel, z is the disparity of the pixel. A plane is been calculated for each region using least square. It is easy to realize and efficient to solve the plane fitting problem. If the disparity of each pixel is not all accurate, it will affect the result of least square remarkably. Fortunately there is reliability to exclude those impurities with threshold. Note that the large number of regions will cost a very long time to calculate the fitting plane of each region and in fact there is no need to divide a scene into thousands of planes. We use a method to eliminate the regions with small reliability and only calculate those high reliable planes. The method comprises several thresholds:

$$\begin{cases} |N_1| > T_1 & N_1 = \{p | r_p > r_c \& p \in R\} \\ |N_2| > T_2 & N_2 = \{p | p \in R\} \\ \frac{|N_1|}{|N_2|} > T_3 & \\ |N_3| > T_4 & N_3 = \{p | p \ is \ on \ edge \ \& p \in R\} \end{cases} \tag{6}$$

1) $|N_1|$ is the number of reliable pixel is larger than T1, where T1 is a fixed amount; 2)$|N_2|$ is the number of all pixels is larger than T2, where T2 is a fixed amount; 3) the percentage of reliable pixels in all pixels is larger than T3, where T3 is a fixed amount; 4)$|N_3|$ is the reliable pixels which is on the edge is larger than T4, where T4 is a fixed amount. The above T1, T2, T3, T4 need to be set before plane fitting to insure that the number of plane, after elimination, is at a proper size. In some cases, the number of regions generated by Mean shift is not so many, maybe a hundred, then we can only use the principle 1) to assure the reliability of plane.

After this step, there are several planes. These planes are all extracted from the scene. It means the scene could be rebuilt by these planes. This will be described in the next part.

5 Disparity plane assignment

In this section, a final global optimum for the disparity plane assignment is searched. In general, a global optimization function could be divided into two parts, the local cost and the global cost. It is necessary to find a solution to make the sum of both costs to be the minimum.

5.1 Plane assignment

In our method, the local costs could be treated as the deviation when we assign a plane to a region. It means that the smaller the difference between the disparity which is calculated with the plane formula and the original disparity, the smaller the local cost is. On the other hand, the global cost is behalf of the deviation between adjacent regions. When two planes are assigned respectively

to two adjacent regions, it is preferred to assign similar planes to similar regions. Similar regions in here mean both regions have similar disparity in average and in boundary. Similar planes mean both planes have similar parameters, it can be formulated as

$$S = (\frac{a_1}{a_2} - \frac{a_2}{a_1})^2 + (\frac{b_1}{b_2} - \frac{b_2}{b_1})^2 + (\frac{c_1}{c_2} - \frac{c_2}{c_1})^2 \tag{7}$$

where a, b, c are the parameters of plane formula. S is the similarity of both planes. In the above steps, a lot of regions which is divided from the image and several planes come out. Next is to assign a corresponding plane to every region. Therefore, the problem is formulated as an energy minimization problem. The energy for an assignment is given by:

$$E = E_{DATA} + E_{SMOOTH} \tag{8}$$

where the local cost could be calculated as

$$E_{DATA} = \sum |d_p - d_o| e^{1-\frac{s}{n}} \tag{9}$$

and the global cost could be calculated as

$$E_{SMOOTH} = \sum^{Pairs} (\frac{1}{n} \sum d_p^1 - \frac{1}{m} \sum d_p^2) \tag{10}$$

where d_p is the disparity of the pixel when set the region with a fixed plane, d_o is the original disparity of the pixel, s is the number of fitting pixels of a fixed plane, which means the pixel is near from the plane, n is the number of pixels in the region. *Pairs* is the amount of adjacent region. Before these variables are calculated, we eliminate those unreliable pixels for each region. The same process is carried out in plane fitting. This is necessary in our method because it is sensitive for the small errors. So making full use of reliability map to reduce the deviation is important. Note that there is a little difference in eliminating process. In plane fitting, a threshold is set to determine reliability of each pixel and delete the pixels with small reliability. However, in this part, we statistic amount of pixels for each disparity in one region and choose the major, depends on the sum of pixels to choose the top three or four, to calculate the cost. This is because the reliability is calculated pixel by pixel and represents changes in small area. In the optimizing process, it is necessary to take the whole region into account. The pixels whose amount is in major to represent the region are chosen; and in our experiment it provided better results than simply using a threshold.

5.2 Region Merging

Several algorithms based on graph cuts have been developed recently to efficiently solve the problem of energy minimization. Here we use the Markov random

Fig. 3. Threshold in region merging. Regions in red rectangle need to be combined. The left is small region in a huge region. The middle is thin and long region between two regions. The right is small region surrounded by several regions.

vector field algorithm in graph cuts to solve this problem. Its time-requirement is in proportion to the nodes and segments number. To make this process faster, a region merging process is applied before energy minimization, as the term suggests, those tiny regions are combined with low confidence into a larger, more reliable region. Our method uses fifteen thresholds, some of which are like the threshold in plane fitting, to find the adjacent regions that could be combined. As shown in Fig.3: 1) if a region is small enough, and only has one neighbor which means this region is surrounded by a large region, it could be combine with the surrounding region; 2) if a region is thin and long and has only two neighbors which in general means it is an edge of over-segmentation, the similarity is calculated respectively between the region and its two neighbors, then we choose the similar one to combine into; 3) if a region is surrounded by several regions whose number of pixels is much larger than itself and these regions has similar disparity, assuming all of these regions comprise a big region. Furthermore, if several similar regions could be connected to a closed area, it is possible to assume it a bigger region. To use this principle, its necessary to strict the definition of similarity .

6 Experiments

In this section, we discuss our experiments for evaluating the performance of our method qualitatively and relative to earlier approaches. Besides, we precede our experimentation to assume the running time. The dataset and groundtruth are from HCI light field archive, which have 9×9 images at resolution 768×768 (picture horse is 1024×576 resolution) per picture.

With our experiments we show that our methods provide better results in some part of image with comparable computational effort. In the first step, we compared the consistency between our result and groundtruth. In the second step, we compared the running time and accuracy between our method and HCI's local method. Then, the influence of the amount of merging terms on run time and average error is measured. Finally, we compared the running time between method without region merging and methods with merging.

Consistency with groundtruth. Fig 4 shows the result image of our method. One could see the boundary of our result is smooth and in some large

plane we could find clearly that there is gradual change in gray level, which means depth estimate changes smooth in those planes. It is more real in a 3D scene. Besides, in the textureless region, such as the light in MonasRoom, our results has good performance. It means our global method provide cracking result in depth recovery. For comparing local depth estimate and global estimate, we could find the result almost eliminate all noise produced by occlusion or textureless. But when some regions, such as the background which is surrounded by leaves in Papillon, is occluded in all pictures, it will get wrong depth label in global method. That's because it's local cost is large for all plane label, the smooth cost must be set small for minimization, the region will be given a label similar with surrounding. To solve this problem, the local depth estimate must be improved. This is what we will do in future.

Comparison with other. Table 1 shows the quality and running time between our method and other global methods in plane segmentation. To highlight the various, the experiments are implemented on the same platform, which is on an ATI Radeon HD 4300 hosted on two Intel Core E7500 CPUs.

While testing the running time, we run these method three times. Note that the process only use one CPU core. HCI method means the local depth estimate method in Paper [12]. The global method of them is very time consuming hence is not on the same order of magnitude with the local method, even running on GPU. So we did not show its running time in Table 1. It can be seen that although there are many inputs in our method and global method is used for final smoothing, we use the same time, some of them even to be short, to process our experiment. It shows that our method is efficient and can deal with large 4D light fields quickly.

images	Ave-running time(s)				Ave-error(%)			
	local depth	global	global without merging	HCI	local depth	global without merging	global	HCI
Papillon	5.6	28.7	243.2	26.0	12.4	4.8	4.8	5.6
MonasRoom	10.4	31.4	291.4	29.8	25.1	1.8	3.1	2.1
Horse	3.7	38.8	223.4	47.6	16.1	1.7	3.8	1.8
Buddha	4.7	34.2	173.9	40.1	10.6	2.0	4.3	1.9

Table 1. Comparison of our method and HCI method. In these experiments, HCI global method is applied on single view. Our method is applied on two views. All of these method are implemented on an ATI Radeon HD 4300 hosted on two Intel Core E7500 CPUs.

While testing accuracy, we static the number of pixels, which has more than two error gray level from groundtruth. Before the comparison, gray order reset is processed to keep the result and groundtruth in the same depth level. Obviously the error reduces a lot after global optimization. While compared with HCI local depth estimation, our method has better accuracy and is less time consuming.

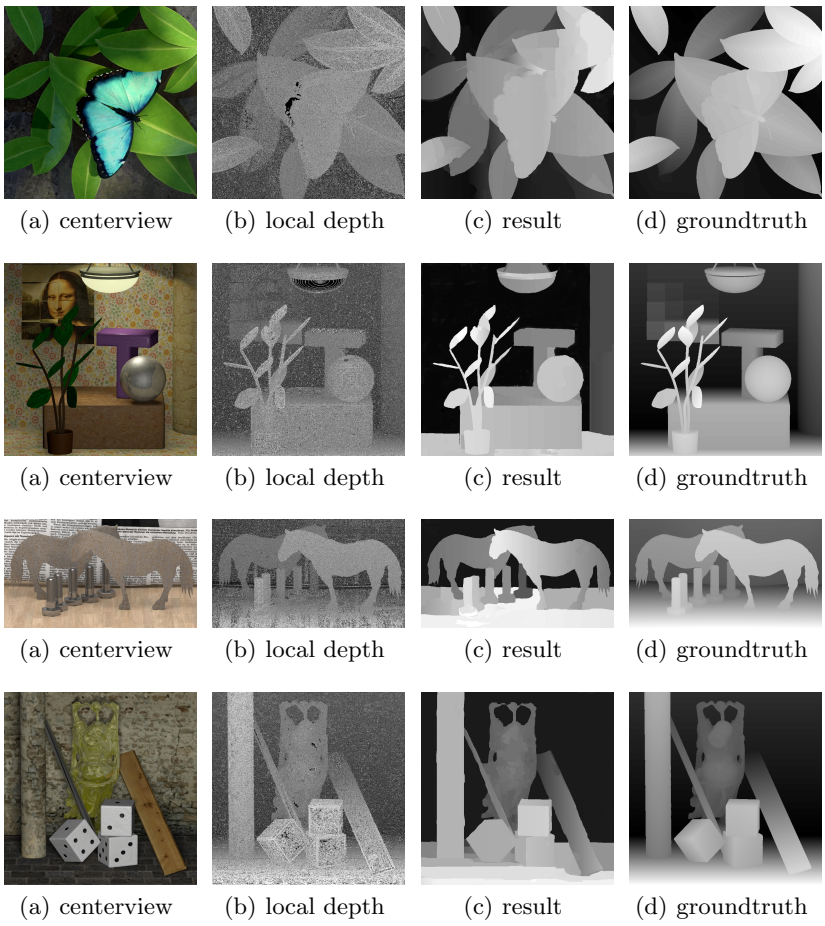

Fig. 4. Experiment result. (a) is the center view image in 4D light filed. (b) is local depth estimate after merging. (c) is the result of our method. (d) is ground truth

Influence of merging terms. In this part, we tried to change the number of merging terms and find out how the terms influence run time and average error. Fig.5 shows the change curve of picture MonasRoom. As the chart indicates that the run time and average error decrease while the terms' amount increasing, but both of them will increase after the trough. Note that the trough differs from each other and in some cases it will emerge twice. It all depends on the selection of thresholds.

Comparison with methods without region merging. We test our global method without region merging. As Table 1 shows, it costs decuple times more than the method which is proceeded after region merging. Note that in some cases, too many terms for merging will increase the run time. Because every merging procedure needs to proceed separately to reduce the influence between

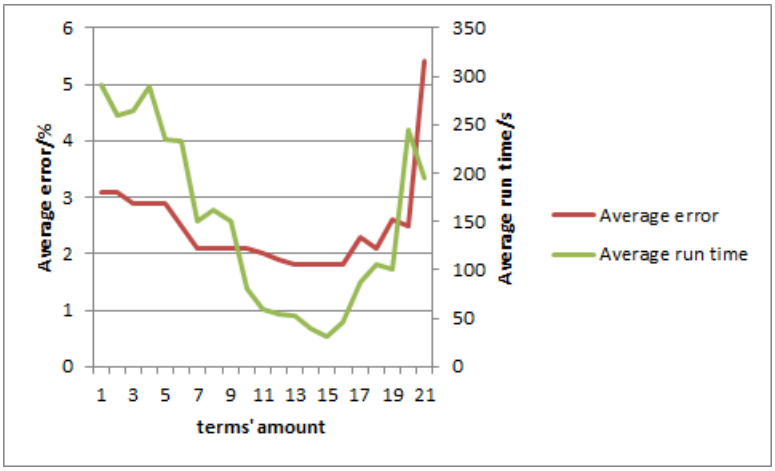

Fig. 5. The influence of merging terms on run time and average error.

each other. Terms increasing will lead to more recursion. In our experiments, fifteen terms produced best results and required the shortest run time.

7 Conclusion

We demonstrated a depth-extracting method on the scene of 4D light fields. It has good performance in processing 4D light field images, especially those that do not contain obvious texture. We applied a highly efficient method of stereo matching on a dense 4D light field and changed some means to make the method fitting the circumstance well. Owing to the large amount information provided by the 4D light field, the deviation could be smaller than the traditional stereo match method. Compared with those existing global algorithm, our method provide a less time consuming and highly efficient way to distract the depth.

Apart from those advantages, the current version of our algorithm will not be able to handle the situation comprised with too many small objects rich in texture. It will cost a very long time and reach a bad result eventually. In future research, we plane to use more information in the 4D light fields to obtain a high efficient method in depth-extracting.

Acknowledgement. This study was partially supported by the National Natural Science Foundation of China (No. 61370122), the National High Technology Research and Development Program of China (No. 2013AA01A603) and the National Aerospace Science Foundation of China (No.2013ZC51). 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-2015ZX-21.

References

1. A.Chambolle, T.Pock: A first-order primal-dual algorithm for convex problems with applications to imaging. J.Math.Imaging Vis 40(1), 120–145 (2011)
2. A.Criminisi, S.Kang, R.Swaminathan, R.Szeliski, P.Anandan: Extracting layers and analyzing their specular properties using epipolar-plane-image analysis. Computer vision and image understanding 97(1), 51–85 (2005)
3. B.Goldluecke, S.Wanner: The variational structure of disparity and regularization of 4d light fields. In: International Conference on Computer Vision and Pattern Recognition (2013)
4. E.Strekalovskiy, D.Cremers: Generalized ordering constraints for multilabel optimization. In: International Conference on Computer Vision (2011)
5. He, K, Sun, J, Tang, X: Single image haze removal using dark channel prior. In: Computer Vision and Pattern Recognition (2009)
6. J.Berent, P.Dragotti: Segmentation of epipolar-plane image volumes with occlusion and disocclusion competition. In: IEEE 8th Workshop on Multimedia Signal Processing. pp. 182–185 (2006)
7. M.Bleyer, C.Rother, P.Kohli, D.Scharstein, S.Sinha: Object stereo - joint stereo matching and object segmentation. In: Computer Vision and Pattern Recognition (2011)
8. M.Matousek, T.Werner, , V.Hlavac: Accurate correspondences from epipolar plane images. Computer Vision Winter Workshop pp. 181–189 (2001)
9. N.Campbell, G.Vogiatzis, C.Hernandez, , R.Cipolla: Automatic object segmentation from calibrated images. In: Conference for Visual Media Production (2011)
10. R.Bolles, H.Baker, D.Marimont: Epipolar-plane image analysis: An approach to determining structure from motion. International Journal of Computer Vision 1(1), 7–55 (1987)
11. S.Vicente, V.Kolmogorov, C.Rother: Joint optimization of segmentation and appearance models. International Conference on Computer Vision and Pattern Recognition (2009)
12. S.Wanner, B.Goldluecke: Globally consistent depth labeling of 4d light fields. In: Computer Vision and Pattern Recognition. pp. 41–48 (2012)
13. S.Wanner, J.Fehr, B.Jahne: Generating epi representations of 4d light fields with a single lens focused plenoptic camera. Advances in Visual Computing pp. 90–101 (2011)
14. T.Bishop, P.Favaro: Full-resolution depth map estimation from an aliased plenoptic light field. In: Asian Conference of Computer Vision (2011)
15. T.Pock, D.Cremers, H.Bischof, A.Chambolle: Global solutions of variational models with convex regularization. SIAM Journal on Imaging Sciences (2010)
16. W.Lee, W.Wontack, E.Boyer: Silhouette segmentation in multiple views. In: Transactions on Pattern Analysis and Machine Intelligence (2010)