ENHANCED DEPTH ESTIMATION FOR HAND-HELD LIGHT FIELD CAMERAS

Yanwen Qin, Xin Jin, Senior Member, IEEE, Yanqin Chen, Qionghai Dai, Senior Member, IEEE

Shenzhen Key Lab of Broadband Network and Multimedia,
Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China

ABSTRACT

Conventional depth estimation methods are confined by the occluder and homogenous regions in the scene. In this paper, we propose a new depth estimation and enhancement method. The raw depth is calculated from analyzing the Consistency Metric Range (CMR) in the angular patch. Confident depth map is obtained by analyzing the variation of CMR within a neighborhood around the lowest CMR curves. Confident depth points are propagated to the whole image by global optimization with weighted neighborhood smoothness, gradient and second derivative constraints. Finally, depth is enhanced by using weighted median filter. The experimental results demonstrated the effectiveness of the proposed approach in providing much clearer transitions of texture regions and much smoother homogenous regions after depth propagation.

Index Terms—Light field, depth estimation, depth propagation, consistency metric range, confident depth

1. INTRODUCTION

Nowadays, hand-held light field cameras such as Lytro [1] and Raytrix [2] have been available for commercial and industrial use. The greatest benefits of the hand-held cameras are it can obtain multiple views only at a single shot, and it can refocus [3] locally after capturing the image. Depth information can also be extracted by using this fruitful information.

Recently, several approaches have been proposed in depth estimation for hand-held light field cameras. Tao *et al.* [4] proposed a method of calculating defocus and correspondence, and fuse the two cues together to get a depth map, yet fails in texture-less regions. Jeon *et al.* [6] proposed a new stereo matching based method which introduced phase shift to estimate the sub-pixel shift of sub-aperture images. It is able to estimate the correspondence at sub-pixel accuracy. However, if the distance between two cameras is too large, there will be some correspondence errors because of the limited number of matching blocks. Wanner *et al.* [7] used a structure tensor to compute two slopes in horizontal and vertical directions in Epipolar Plane Image (EPI) of a light field image, and they formulated the problem of estimating depth as a global optimization approach subjected to the epipolar constraint. T. Wang *et al.* [8] developed a method, which treated occlusion explicitly, to enhance the accuracy of depth to some extent. However, the method is based on detecting occlusion boundaries in the image and assumes only one occluder in angular patch. The above methods are not robust to noise and perform worse in texture-less regions.

In this paper, an enhanced depth estimation based on minimizing Consistency Metric Range is proposed. The key to the proposed method is introducing the optimization model of propagating confident depth to unconfident depth points to the scheme. Confidence points are detected by analyzing the variation of CMR within a neighborhood around the lowest CMR curves. Weighted median filter is used to refine depth map to get the final depth map. Experimental results demonstrate that the proposed method can provide a depth map with clearer transition of texture regions and much smoother homogenous regions.

The rest of the paper is organized as follows. The architecture of proposed algorithm is illustrated in Section 2. Section 3 describes the key principle of CMR, confidence analysis of the raw depth and the model designed to propagate the confident depth pixels to unconfident pixels. Experimental results are shown in Section 4. And the conclusions are drawn in Section 5.

2. PROPOSED SYSTEM ARCHITECTURE

The system architecture of proposed algorithm is depicted in Fig.1.

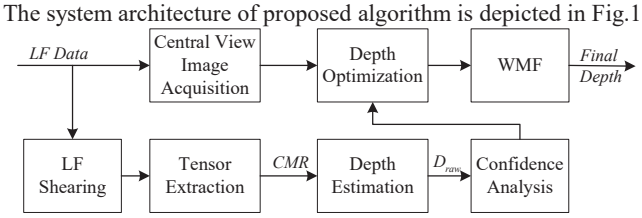


Fig. 1: Framework of the proposed method.

The input is 4D Light Field (LF) data, which are decoded to get the raw input image. Then, light field shearing is performed by [3]

$$L_\alpha(x,y,u,v) = L_F\left(x+u\left(1-\frac{1}{\alpha}\right), y+v\left(1-\frac{1}{\alpha}\right), u, v\right), \quad (1)$$

where L_F is the rectified image; L_α is the refocused 4D light field image at the depth label α ; (x,y) is the spatial coordinates; (u,v) is the angular coordinates. A LF volume is obtained after LF shearing. Each slice of the volume is the image after refocusing which will be used in *Tensor Extraction*. CMR is extracted from the LF volume by *Tensor Extraction*, which varies during refocusing. Initial depth D_{raw} can be generated using CMR, then Confidence Analysis is performed to detect confident depth pixels from D_{raw}.

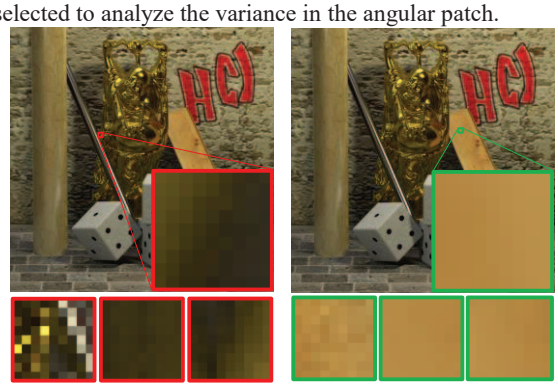
In order to correct the unconfident depth pixels with reasonable depth value, a *Depth Optimization* model considering weighted neighborhood smoothness, gradient and second derivative constraints is used to propagate confident depth to the whole depth map. Finally, weighted median filter (*WMF*) is used to refine the depth.

3. DEPTH ESTIMATION AND OPTIMIZATION

3.1. Angular Patch Analysis

According to [8], the pixels in the angular patch exhibit photo-consistency for Lambertian surfaces if the surfaces are focused, which means the pixels have the same intensity in the angular

patch. In Fig.2, occluded and non-occluded positions in the scene are selected to analyze the variance in the angular patch.



(a) With occlusion (b) Without occlusion

Fig. 2. Angular patch comparison in different positions.

For the red point in Fig.2(a), it will be occluded when changing perspectives. For the angular patch with occlusion, the intensity does not hold photo-consistency. The same angular patch at three different depth labels are shown in the bottom of Fig.2(a) corresponding to depth label α =1, 59, and 100, respectively. We noticed that when α =59, the patch shown in the middle, the variance of pixels' intensity in the angular patch looks much smaller than the other two. In contrast, the green point in Fig. 2 (b) is a patch which will not be occluded during viewpoint changing. It shows photo-consistency by providing the corresponding angular patches with small variance and high similarity among different depth labels, like those shown on the bottom of Fig.2(b) at α =1, 78, and 100. Thus, we hammer at extracting a tensor that is related with the photo-consistency of angular patch at different α .

3.2. Depth from Consistency Metric Range(CMR)

According to above angular patch analysis, a novel method of depth estimation is proposed based on CMR. First, intensity range $R(\boldsymbol{p},\alpha)$ proposed by us before [11] is calculated for every angular patch in the LF volume by:

$$R(\boldsymbol{p},\alpha)=\max_{q\in A(\boldsymbol{p},\alpha)}(I_q)-\min_{q\in A(\boldsymbol{p},\alpha)}(I_q), \tag{2}$$

where $A(\boldsymbol{p},\alpha)$ represents angular patch corresponding to a pixel $\boldsymbol{p}$; $\boldsymbol{p}$=(x,y), in which (x,y) denotes the Cartesian image coordinates of each sub-aperture image; $\boldsymbol{q}$ is a pixel in the angular patch; I is the image after shearing and $I_{\boldsymbol{q}}$ means the intensity value of pixel $\boldsymbol{q}$. Different from that in [11], $R(\boldsymbol{p},\alpha)$ is calculated for each RGB color channel independently. To fuse $R(\boldsymbol{p},\alpha)$ from the three channels, R_{max} and R_{avg} are defined by:

$$R_{max}(\boldsymbol{p},\alpha)=max\left(R_{\boldsymbol{R}}(\boldsymbol{p},\alpha),R_{\boldsymbol{G}}(\boldsymbol{p},\alpha),R_{\boldsymbol{B}}(\boldsymbol{p},\alpha)\right), \tag{3}$$

$$R_{avg}(\boldsymbol{p},\alpha)=\sqrt{\frac{\left(R_{\boldsymbol{R}}^2(\boldsymbol{p},\alpha)+R_{\boldsymbol{G}}^2(\boldsymbol{p},\alpha)+R_{\boldsymbol{B}}^2(\boldsymbol{p},\alpha)\right)}{3}}, \tag{4}$$

where $\{\boldsymbol{R},\boldsymbol{G},\boldsymbol{B}\}$ denotes the 3 color channels. R_{max} performs well in depth estimation if the object has a dominant color channel. While, R_{avg} performs better for the other cases. Then, the proposed tensor CMR is defined as:

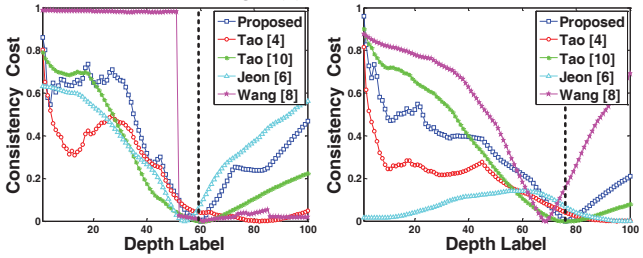
$$C(\boldsymbol{p},\alpha)=\beta R_{max}(\boldsymbol{p},\alpha)+(1-\beta)R_{avg}(\boldsymbol{p},\alpha), \tag{5}$$

where β is a weight and ranges from 0 to 1. CMR can also be interpreted as a cost function of the angular patch. By minimizing $C(\boldsymbol{p},\alpha)$ along the depth label α , we can obtain the raw depth D_{raw} by:

$$D_{raw}(\boldsymbol{p})=\underset{\alpha}{\operatorname{argmin}}\,C(\boldsymbol{p},\alpha), \tag{6}$$

where $D_{raw}(\boldsymbol{p})$ is the depth of the point $\boldsymbol{p}$.

We compared data costs proposed by us and other state-of-the-art approaches for the angular patches in Fig.2 (a) and (b). The data costs varying with the depth label are shown in Fig.3. For a fair comparison, we set the resolution of depth label to be 100. Generally, the depth label at which the data cost reaches the minimum will be selected as the estimated depth label. It is obvious that the proposed tensor will achieve the minimum value together with other methods [4] [6] [8] [10] in the occluded region, as shown in Fig.2(a). While in the non-occluded region, the proposed tensor is comparable with that proposed in [10] and outperforms the others, as shown in Fig. 3(b).

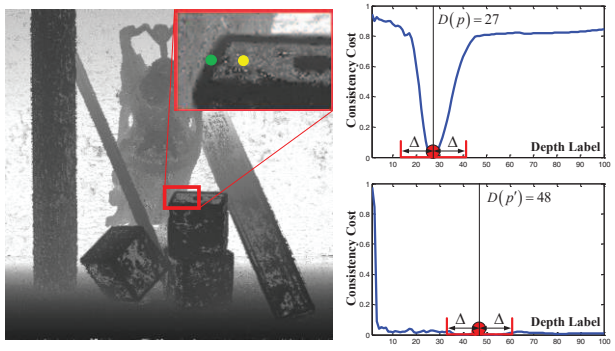


(a) (b)

Fig. 3. CMR cost comparison for the patch in: (a) Fig.2(a); and (b) Fig. 2(b). The depth label at the dashed line corresponds to ground truth.

3.3 Confidence Analysis

After generating an estimated raw depth image D_{raw}, pixels with high confidence are selected by confidence analysis. First, CMR variation with the depth label is analyzed in Fig.4. The green and yellow points are picked at the same depth visually in Fig.4(a). However, their CMR curves look quite different in Fig. 4(b) and the depth labels where they reach the minimum in $D(\boldsymbol{p})$ =27 and $D(\boldsymbol{p}')$ =48. Since generally the CMR curves with drastic change around lowest point will show higher confidence in estimating the depth, we then investigate the variance among a local window around the lowest point.



(a) (b)

Fig. 4. Confidence analysis. (a) Raw depth map. (b) The CMR curves of two points: top one for the green point in (a) and bottom one for the yellow point in (a).

The method to detect unconfident depth pixels is developed as following:

$$\operatorname{var}\left(C(\boldsymbol{r})\right)\Big|_{\boldsymbol{r}\in M(\boldsymbol{p})}\le\tau_{reject}\ ,\ M(\boldsymbol{p})=[D_{raw}(\boldsymbol{p})-\Delta,D_{raw}(\boldsymbol{p})+\Delta] \tag{7}$$

where var(·) is the operation to calculate variance. $M(\boldsymbol{p})$ denotes the neighborhood of the depth label $D_{raw}(\boldsymbol{p})$ along the depth axis as shown in Fig.4. Δ is the size of the neighborhood. If the local variance of $C(\boldsymbol{r})$ is lower than a threshold τ_{reject} , the depth of pixel $\boldsymbol{p}$ is regarded as unconfident depth. So the set of confident depth pixels can be formulated as:

$$\Omega = \left\{ \boldsymbol{p} \middle| \mathrm{var}\left(R(\boldsymbol{r})\right)\Big|_{\boldsymbol{r}\in M(\boldsymbol{p})} > \tau_{reject} \right\}. \tag{8}$$

Using the image in Fig.5 (a) as an instance, its raw depth map is obtained using Eq. (6) and depicted in Fig.5 (b). Note that errors exist in the raw depth map, such as regions on dices. Thus, confidence analysis is needed to detect the unconfident depth points. As shown in Fig.5 (c), red points indicate unconfident depth points when the threshold is set to τ_{reject} =0.04 and Δ =30.

(a) (b) (c)

Fig. 5. Estimated depth map and (a) Color image. (b) Raw depth image. (c) Raw depth map after confidence analysis, red points indicate unconfident points.

3.4 Depth Propagation Model

In order to propagate unconfident depth points with confident depth values to enhance the overall quality of raw depth map, we try to improve the constraint that two neighboring pixels $\boldsymbol{p}$, $\boldsymbol{s}$ should have similar colors if their intensities are similar. Based on the color propagation model in Levin *et.al*. [14]'s work, we minimize the difference between the depth $D(\boldsymbol{p})$ and the weighted average of the depth of pixels belonging to the neighborhood of $\boldsymbol{p}$, the weighted neighborhood smoothness $J_1(D)$ is defined as:

$$J_1(D) = \sum_{\boldsymbol{p}} \left(D(\boldsymbol{p}) - \sum_{\boldsymbol{s}\in N(\boldsymbol{p})} w_{\boldsymbol{ps}} D(\boldsymbol{s}) \right)^2, \tag{9}$$

where $N(\boldsymbol{p})$ is a 3×3 window around $\boldsymbol{p}$; $w_{\boldsymbol{ps}}$ is the weight that sums to one. We use the square difference between two intensities on the central view image I_c to formulate the weighting term as:

$$w_{\boldsymbol{ps}} \propto \exp\left(-\left(I_c(\boldsymbol{p}) - I_c(\boldsymbol{s}) \right)^2 \Big/ 2\sigma_{\boldsymbol{p}}^2 \right), \tag{10}$$

where $\sigma_{\boldsymbol{p}}$ is the variance of intensity in the window $N(\boldsymbol{p})$.

In order to preserve the edges of scenes in final depth map, we minimize the differences of the gradient between I_c and depth D, which is denoted by $J_2(D)$. $J_2(D)$ is formulated as:

$$J_2(D) = \sum_{\boldsymbol{p}} \left(g_D(\boldsymbol{p}) - g_{I_c}(\boldsymbol{p}) \right)^2, \tag{11}$$

where g_D and g_{I_c} represent the gradient of depth map and central view image at pixel $\boldsymbol{p}$, respectively. If the gradient of final depth and color image are similar, J_2 will be small; otherwise J_2 will be large. Furthermore, for the purpose of keeping texture-less regions being smooth enough, we define the second derivative term by:

$$J_3(D) = \sum_{\boldsymbol{p}} \left(\Delta D(\boldsymbol{p}) \right)^2, \tag{12}$$

where ΔD is given by:

$$\Delta D = \left| \partial^2 D / \partial \boldsymbol{p}^2 \right|. \tag{13}$$

Equations (9), (11), (12) should be minimized jointly which leads to a multi-target problem. Finally, the regularized optimization model is defined as:

$$\begin{aligned} D^* &= \arg\min_{D} J_1(D) + \lambda J_2(D) + \eta J_3(D) \\ \boldsymbol{s.t.}\ & D(\Omega) = D_{raw}(\Omega) \end{aligned} \tag{14}$$

where D^* is depth after optimization; Ω is the region with high confident depth map; λ and η control the constraint for edge preserving and flatness of refined depth map, respectively.

Finally, for the reason that WMF performs well both in removing outlier error and preserving edges, weighted median filter (WMF) [12] is used to further refine the depth map D^* to obtain D_{final}.

4. EXPERIMENTAL RESULTS

The performance of the proposed method is evaluated by using the images captured by Lytro1.0 [1]. and images from public synthetic dataset provided by Wanner *et al.* [13]. Depth label Resolution of all methods are tuned to 100 for fair comparison. For our experimental configuration, depth propagation coefficients λ is 10 and η is 0.2. Parameters of other methods are maintained to the default.

Fig.6 shows the processing results for the images captured by Lytro1.0. As shown in the figure, Tao *et al*. [4]'s results look much blurry and fail in homogenous regions especially in the background. Tao *et al*. [10]'s results fail to keep clear boundaries of objects and look blurry as well. Although processing results provided by Jeon *et al*. [6]'s preserve the boundaries of objects, they fail in the presence of occlusion. Wang *et al*. [8]'s results provide the whole depth levels of the scenes, but lose transitions of depth. Obviously, results obtained by the proposed method not only preserve details of boundaries in the scene, but also keep much abundant transitions and high accuracy in homogenous regions. In addition, the proposed method is applied to synthetic scenes [13], as depicted in Fig.7. We can see that processing results shown in Fig.7 demonstrate the same advantages.

Table 1 further measure mean square error (MSE) between the ground truth and the processing results for the above methods on Wanner's datasets [13]. The proposed approach outperforms the others for two scenes and ranks the third for the last. However, average MSE of our method is the lowest among all methods. Combining the results from subjective evaluation of results in Fig.6 and Fig.7, the proposed algorithm can provide more reliable estimation results, richer depth details and lower MSE.

Table 1. MSE Comparison

	Tao *et al.* [4]	Tao *et al.* [10]	Jeon *et al.*[6]	Wang *et al.*[8]	Proposed
Buddha	0.0228	0.0360	0.0184	0.0196	**0.0081**
Buddha2	0.0293	0.0319	0.0157	0.0050	**0.0046**
Mona	0.0341	0.0246	0.0090	**0.0076**	0.0099
Avg.	0.0287	0.0308	0.0144	0.0107	**0.0075**

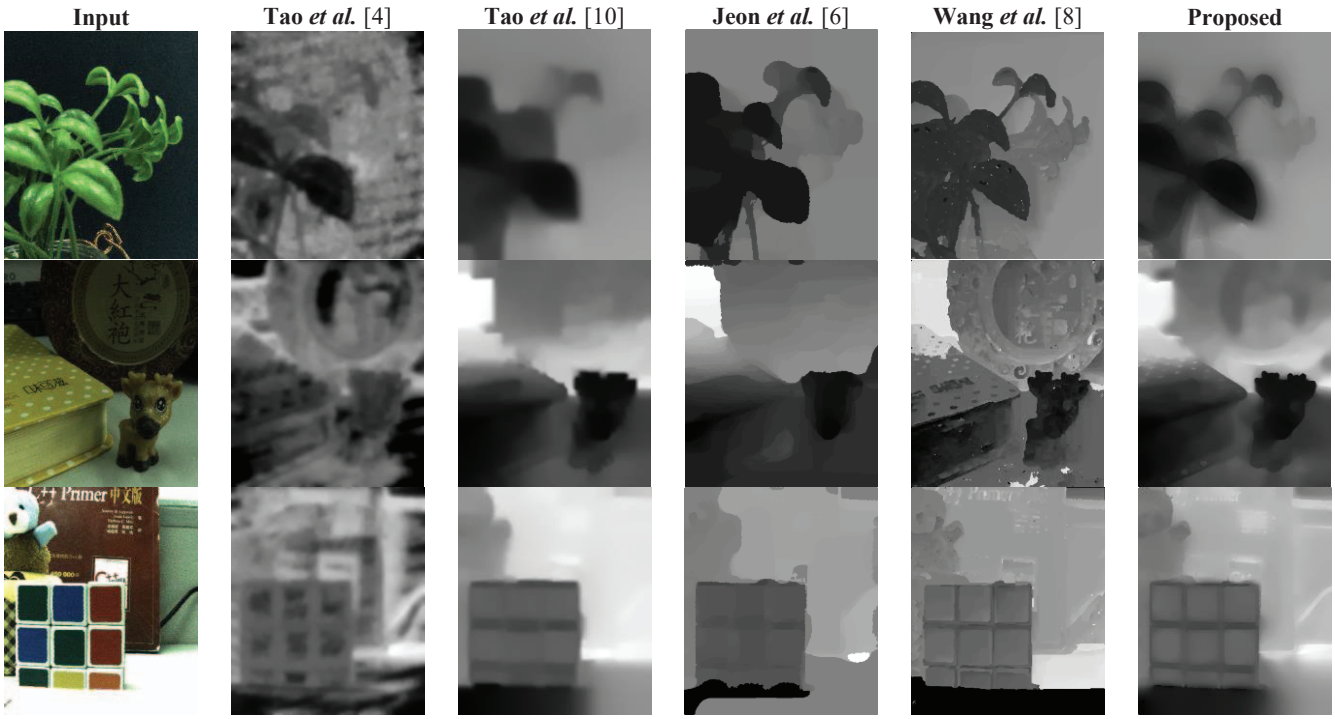


Fig. 6. Processing results comparison of Lytro scenes.

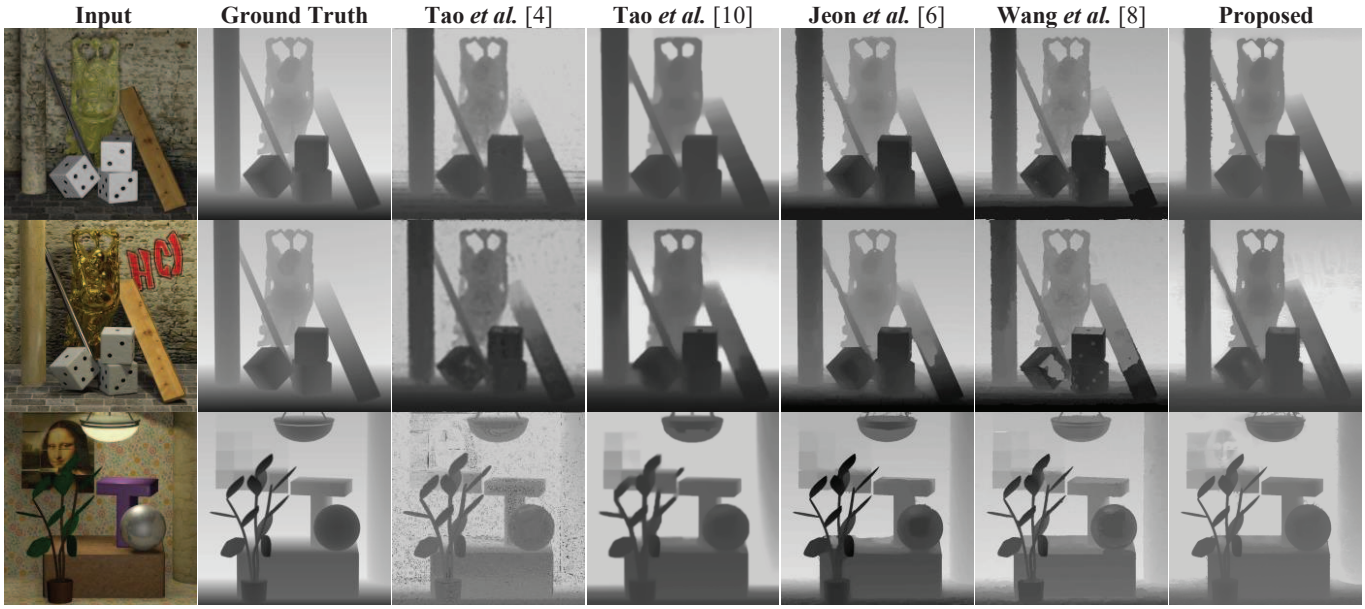


Fig. 7. Processing results comparison of Wanner's datasets [13]. For better visualization, we have tuned the contrast of each result.

5. CONCLUSIONS

Targeting at depth estimation, an enhanced depth estimation method is proposed in this paper, which analyzes the distribution in angular patch. CMR is designed to compute a robustly homogenous raw depth map. Subsequently, confident depth map is obtained by evaluating the variance of CMR curves around a local window. Then, the raw depth map is propagated by global optimization with weighted neighborhood smoothness, gradient and second derivative constraints. Finally, the final depth map is achieved after being refined by WMF. The performance of the proposed method is demonstrated by comparing with existing state-of-the-art algorithms. Experimental results illustrate that the final depth map obtained by using the proposed method can provide much more abundant transitions on boundaries and higher accuracy in texture-less regions for both images captured by Lytro 1.0 and from public synthetic datasets, which will show its potential for wider applications, such as 3D reconstruction etc.

6. ACKNOWLEDGMENT

This work was supported in part by the project of NSFC 61371138 and NSFC-Guangdong Joint Foundation Key Project (U1201255), China.

7. REFERENCES

[1] “Lytro – Home ”, https://illum.lytro.com/

[2] “ Raytrix|3D light field camera technology”, https://www.raytrix.de/

[3] Ng R. Digital light field photography[D]. Stanford university, 2006.

[4] Tao, Michael, et al. “Depth from combining defocus and correspondence using light-field cameras,” Proceedings of the IEEE International Conference on Computer Vision(ICCV), 2013.

[5] “ Middlebury Stereo Evaluation - Version 3” http://vision. middlebury. edu/stereo/eval3/.

[6] Jeon, Hae-Gon, et al. “Accurate depth map estimation from a lenslet light field camera,” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, pp. 1547-1555, 2015.

[7] Wanner, Sven, and Bastian Goldluecke, “Globally consistent depth labeling of 4D light fields,” 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[8] Wang, Ting-Chun, Alexei A. Efros, and Ravi Ramamoorthi. “Occlusion-aware Depth Estimation Using Light-field Cameras,” Proceedings of the IEEE International Conference on Computer Vision(ICCV). 2015.

[9] Williem W, Kyu Park I. “Robust Light Field Depth Estimation for Noisy Scene with Occlusion,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Pp, 4396-4404, 2016.

[10] Tao M W, Su J C, Wang T C, et al. “Depth estimation and specular removal for glossy surfaces using point and line consistency with light-field cameras,” 2015.

[11] Xu Y, Jin X, Dai Q. “Depth estimation by analyzing intensity distribution for light-field cameras,” IEEE International Conference on Image Processing. IEEE, 2015.

[12] Ma Z, He K, Wei Y, et al. “Constant Time Weighted Median Filtering for Stereo Matching and Beyond,” pp.49-56,2013.

[13] Wanner, Sven, Stephan Meister, and Bastian Goldluecke, “Datasets and Benchmarks for Densely Sampled 4D Light Fields,” VMV, 2013.

[14] Levin A, Lischinski D, Weiss Y. “Colorization using optimization,” ACM Transactions on Graphics (TOG). ACM, 2004, 23(3): 689-694.