FAST RESPONSE AGGREGATION FOR DEPTH ESTIMATION USING LIGHT FIELD CAMERA

Cao Yang[⋆] *Kai Kang*[⋆] *Jing Zhang*[⋆] *Zengfu Wang*[⋆‡]

[⋆] Department of Automation, University of Science and Technology of China
[‡]Hefei Institute of Intelligent Machines, Chinese Academy of Sciences
forrest@ustc.edu.cn

ABSTRACT

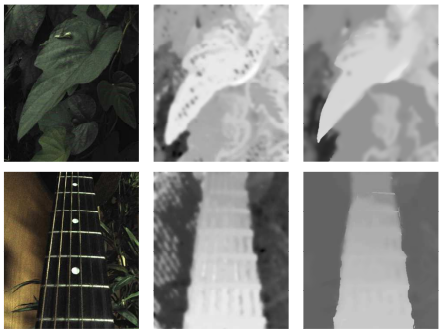
Light field cameras have been recently shown to be very effective in applications such as multifocusing and 3D reconstruction. These cameras can provide depth cues from both defocus and correspondence in a single snapshot. In this paper, we present a fast response aggregation framework for depth estimation by jointly using defocus and correspondence cues. Different from existing approaches, we perform a fast gradient preserving filtering in a label domain, instead of in a depth domain, to efficiently compute a dense depth map. The proposed approach comprises of three steps: 1) constructing defocus and correspondence response volumes, 2) adaptively smoothing the two volumes and performing Winner-Takes-All label selections, and 3) post-processing by using nonlocal image guided averaging. With such a compact framework, currently best depth estimation results can be achieved. This compact framework is suitable for various applications such as object segmentation and surface reconstruction.

Index Terms— depth estimation, light-field, response aggregation

1. INTRODUCTION

The concept of light-field cameras has been proposed by Adelson and Wang [1] over twenty years ago. Ng et al. [2] are the first to introduce the prototype of micro-lenses array light-field camera, which provides a multiple-view of a scene in a single snapshot, recording the distribution of light rays in space. As explained in [2], we can refocus images by shearing the epipolar image (EPI) extracted from light-field data to achieve many multiple-views focused at different depths. Therefore, we can calculate depths from defocus and correspondence cues simultaneously.

How to accurately estimate depth from defocus and correspondence cues has been extensively studied. Schechner et al. [3] and Vaish et al. [4] have discussed the strengths and weaknesses of each cue. Generally speaking, the defocus cues are good at repeating textures and noisy regions, while correspondence cues have better performance in bright points and features. Most existing work on depth estimation only exploit one cue or another, since it is difficult to acquire and combine both cues in the same framework.



(a) (b) (c)

Fig. 1. Depth estimates from light-field image. (a) the central view, (b) result of [5], (c) result of the proposed method.

With the emergence of light-field camera, it becomes easy to exploit defocus and correspondence cues jointly for depth estimation. Recently, Tao et al. [5] combine the two cues to estimate dense depth map. They present the defocus and correspondence measures for different shear angles of EPI, and adopt Markov Random Fields (MRF) [6] to fuse both cues according to confidence measures introduced by Hirschmuller [7]. However, in regions where both cues show low confidence, their depth estimates degrade significantly. In addition, due to the lack of edge-preserving property in postprocessing, their results tend to be blurry at edges.

In this paper, we propose a fast response aggregation framework for depth estimation by jointly using defocus and correspondence cues. This framework is shown in Fig. 2. We first construct defocus and correspondence response volumes, and apply an image guided filtering on them. Then, the Winner-Takes-All strategy is used to obtain the initial depth estimates. Finally, we utilize nonlocal image guided averaging (NLGA) [8] to preserve sharp depth edges in post-processing. This proposed approach can produce high-quality depth estimates with sharp edges, as shown in Fig. 1, which enable applications such as object segmentation and surface

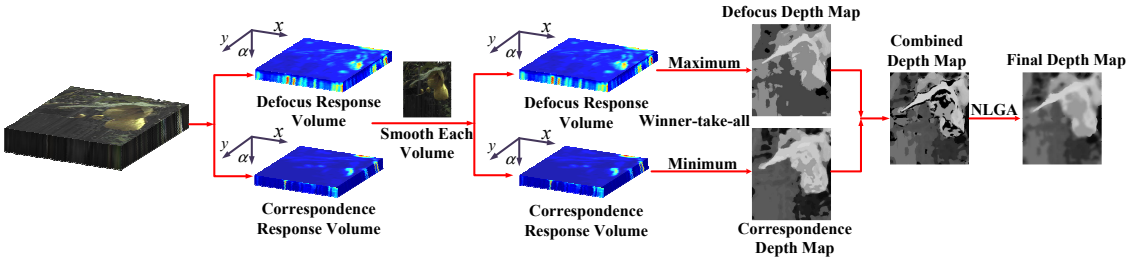


Fig. 2. The pipeline of the proposed algorithm. We first construct defocus and correspondence response volumes according to the sheared EPI. Then, we smooth the volumes by image guided filter, and use Winner-Takes-All strategy to obtain the initial depth maps. Next we combine defocus depth map with correspondence one. Finally we use NLGA to refine the combined depth map.

reconstruction.

2. ALGORITHM

The pipeline of the proposed algorithm is shown in Fig. 2. Similar to [5], we first construct the defocus and correspondence response volumes by shearing EPI. Different from their method, we first filter each slice of response volumes under the guidance of central view image. Then, the depth estimates from the above response volumes are determined in a winner-take-all fashion, respectively. Next, we average the estimates for pixels passing the validity-check. We adopt weighted median filtering method to fill the invalid pixels. At the end, we utilize NLGA to refine the outputs.

2.1. Response Volume Construction

A light-field image after decoding are presented in the left of Fig. 3. It consists of serval sub-images, that are captured in a single snapshot but with different view points. Fig. 3 also explains the concept of EPI. We first stack all images along a line of view points (denoted by red rectangle block), then cut through the stack (denoted by yellow line), forming a cut plane (denoted by yellow rectangle block) called an epipolar plane image (EPI). The rich structure (denoted by color lines within the yellow rectangle block) emerges in the EPI.

For simplicity, we explain the proposed algorithm on 2D EPI shown in Fig. 3 (Note that we implement our algorithm on 4D EPI). Ng et al. have demonstrated how to shear the EPI to achieve refocus in [2].

$$L_\alpha(x,\mu) = L_0(x + s(1-\frac{1}{\alpha}), s), \tag{1}$$

where L_0 denotes the input EPI, L_α denotes the shearing angular value, x represents the spatial domain and s represents the view domain. $\bar{L}_\alpha$ is the average of sheared EPI across the view dimension s, which can be interpreted as the refocused image under shearing value α. The refocus processing is presented in Fig. 3.

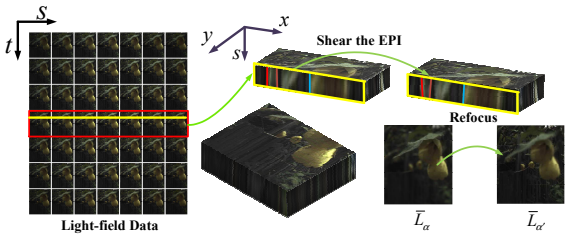


Fig. 3. A light-field image after decoding. It consists of serval sub-images presented in the left side, that are captured in a snapshot but with different view points. A cut (denoted by yellow line) through the stack of sub-images along one horizontal dimension (denoted by red rectangle block), forming a cut plane (denoted by yellow rectangle block) called an epipolar plane image (EPI). The rich structure (denoted by color lines within the yellow rectangle block) emerges in the EPI. Shearing EPI means to refocus image shown in the right bottom.

In [5], Tao et al. have presented two effective measures to describe defocus and correspondence responses. They treat the spatial variance of EPI integrated across the view dimension as the defocus measure $D_\alpha(x)$.

$$D_\alpha(x) = \frac{1}{|W_D|}\sum_{x'\in W_D} \left|\Delta_x \bar{L}_\alpha(x')\right|, \tag{2}$$

where W_D is the window size around the current pixel, and $\Delta_x(x)$ is the Laplacian operator. Besides, they regard the view variance as the correspondence measure $C_\alpha(x)$.

$$C_\alpha(x) = \sqrt{\frac{1}{N_u}\sum_{u'}\left(L_\alpha(x,u') - \bar{L}_\alpha(x)\right)^2} \tag{3}$$

Also, they average $C_\alpha(x)$ by W_c size window for greater robustness. According to Eqn. 2 and Eqn. 3, for each pixel in the image, we can measure defocus and correspondence responses for each shearing angular value α. Then we can construct

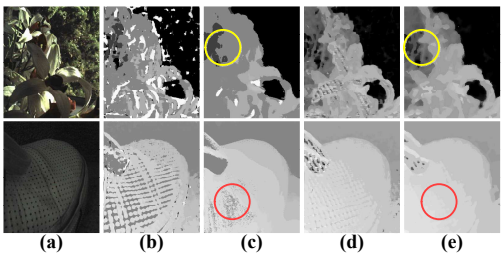


Fig. 4. The depth maps before and after smoothing response volumes. For the depth map, the lighter pixels are regarded as closer point in scene to the camera and darker as father. (a) is center view, (b) and (c) are the defocus depth maps before and after smoothing respectively, (d) and (e) are the correspondence depth maps before and after smoothing respectively.

the response volume that is a 3D array which stores the responses for a certain shearing value α at pixel x. Noticed that the sheared value α is discrete.

2.2. Fast Response Volume Filtering

The optimal shearing angular value α implies the true depth information. For defocus cues, the optimal $\alpha_d^*(x)$ can be found by locating the largest response, and for correspondence cues, $\alpha_c^*(x)$ can be found by locating the smallest response.

$$\alpha_d^*(x) = \arg\max_{\alpha} D_\alpha(x), \tag{4a}$$

$$\alpha_c^*(x) = \arg\min_{\alpha} C_\alpha(x), \tag{4b}$$

The above problem is a typical multi-label problem. Hosni et al. proposed a simple framework to achieve high-quality solutions for general multi-label problems [9]. In their work, the resultant label is effectively smoothed by a very fast edge preserving filter [10], where the label transitions are aligned with color edges of the input image. Similar to their method, we propose to smooth the response volumes and then find the optimal angle in a winner-take-all fashion. For each slice of the response volume, we smooth it under the guidance of center view image.

$$V'_{i,\alpha} = \sum_j W_{i,j}(I) V_{j,\alpha}, \tag{5}$$

where V represent the response volume, i and j are pixel indices. The filter weights $W_{i,j}$ is calculated from the guidance image I as follows.

$$W_{i,j} = \frac{1}{|\Omega|^2} \sum_{k:(i,j)\in\Omega_k} \left[1 + (I_i - \mu_k)^T (\Sigma_k + \varepsilon U)^{-1} (I_j - \mu_k)\right], \tag{6}$$

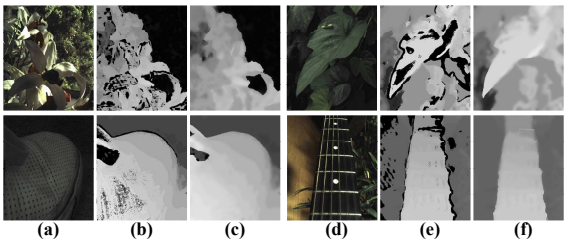


Fig. 5. (a) and (d) are the central view, (b) and (e) are the combined depth map where black pixel refers to the invalid value which need to be filled by reliable neighbor value, (c) and (f) are the results refined by weighted median filter and NLGA.

Here, μ_k and Σ_k are the mean vector and covariance of I in a squared window Ω_k centered at pixel k. U denotes the identity matrix and is a smoothness parameter. $|\Omega|$ is the number of pixels in N_k. We recommend referring to [10] for further details.

The comparison of the depth maps before and after smoothing is shown in the Fig. 4. It can be easily seen that most of the outliers have been eliminated.

2.3. Combination and Refinement

After the processing steps as outlined above, we obtain two depth estimates from two different cues. However, they may be inconsistent in some regions, as shown in regions enclosed by the color circles in Fig. 4. To overcome this problem, we use a validity-check technique to assign binary validity value for each pixel. For each pixel, if the difference between two depth estimates is smaller than a threshold, the validity value is assigned to be 1, otherwise 0. For pixels passing the validity-check, we compute the average of the two estimates as the initial depth value. Fig. 5(b) and Fig. 5(e) show the initial depth map after validity-check. For the invalid pixels, we adopt weighted median filtering method to fill them [9].

Then, we use NLGA filter to refine the filling result to smooth the remaining residual artifacts. NLGA is an edge-preserving filtering method developed recently and is able to exploit the nonlocal self-similarity of the guidance image. The explicit form of the filter kernel weight can be formulated as:

$$W_{i,j} = \sum_{k:(i,j)\in\Omega_k^n} w_{ik} w_{kj} \times \left[1 + \left(I_k - \overline{I_w(\Omega_k^n)}\right)^T (\Sigma_{k,w} + \varepsilon U)^{-1} \left(I_j - \overline{I_w(\Omega_k^n)}\right)\right], \tag{7}$$

where Ω_k^n represents the nonlocal neighborhood of pixel i, w_{ij} is the nonlocal weight. $\overline{I_w(\Omega_k^n)}$ and $\Sigma_{k,w}$ are the weighted mean vector and covariance matrix in Ω_k^n. The refinement results are shown in Fig. 5(f).

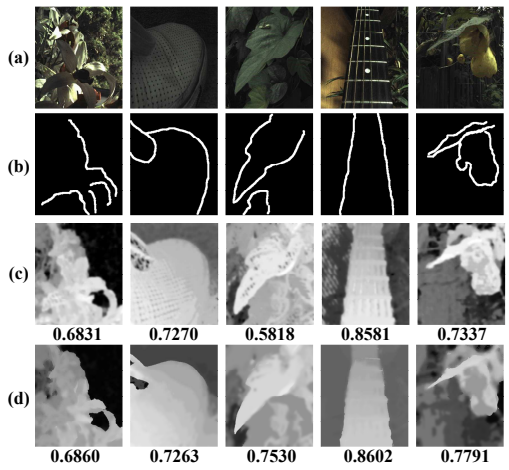


Fig. 6. The comparison between our results and Tao et al.'s. (a) is the central view, (b) is the occlusion boundaries marked by user manually, (c) presents Tao et al.'s results, (d) presents our results. The value under each depth map is AUC value.

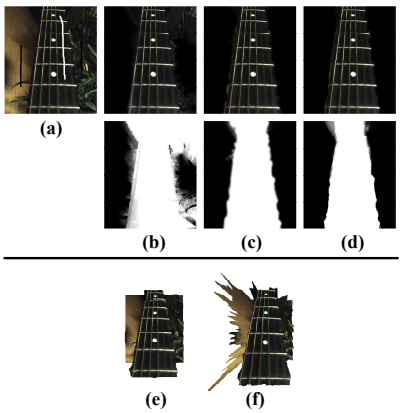


Fig. 7. Two applications. (a) the center view, (b) object segmentation by matting method [11], (c) object segmentation by using Tao et al.'s depth map (d) object segmentation result by using the proposed method's depth map, (e) surface construction by using our depth map, (f)surface construction by using Tao et al.'s depth map. The second row presents each segmentation's matte that the white regions indicate the foreground object.

3. RESULTS AND ANALYSIS

In this section, we explore the potential of the proposed algorithm based on light-field images released by [5]. The comparison between our results and that of Tao's are shown in Fig. 6. In our implementation, we adopt the same parameters as those reported in [5], when we construct the response volumes.

For quantitative comparison, we evaluate the performance of depth map in terms of occlusion boundary detection by calculating the area under the receiver operating characteristic (ROC) curve, i.e. the AUC value. Here, we regard the occlusion boundary detection as a classifier and the gradient of depth maps as the classification results. Given a certain threshold, the points above the threshold are classified as positive and the others as negative. For any threshold, true and false positive rate are obtained. By varying the threshold at some intervals within the maximum range, the ROC curve can be plotted and the area under the curve (AUC) can be calculated. The depth map with larger AUC has superior quality.

According to AUC evaluation, the proposed scheme outperforms Tao et al.'s in terms of occlusion boundary detection. As seen from Fig. 6, the surface of the foreground leaf, and the background of chord are consistent. Visually, the edge of leaf in calabash is sharper. For the shoes example, though the AUC measures are close, our results have better visual appearance, since the proposed scheme removed the unexpected textures.

We also show that the proposed scheme produces high-quality depth maps that can be used for object segmentation and surface construction.

Object Segmentation. With a simple stroke, we can extract objects with more accurate boundaries by using depth map, as shown in Fig. 7. Apparently, the depth map we obtained is helpful to reduce the ambiguity that comes from color space, and produce shaper boundaries compared with Tao's.

Surface Construction. We can remap the pixels into 3D space to achieve surface reconstruction according to the obtained depth map, as shown in Fig. 7.

4. CONCLUSIONS

In this paper, we propose a fast response aggregation framework for depth estimation by jointly using defocus and correspondence cues. High quality depth maps can be achieved by filtering each cue's response volume. The validity-check processing is able to preserve the consistent estimates and exclude others. The pixels fail to pass validity-check are filled via a weighted median filtering method. High-quality depth estimates with sharp edges can be obtained by using NLGA filtering method to refine the filling results. Finally, the visual inspection and quantitative assessment confirm the effectiveness of the proposed method. The depth maps obtained from this compact framework is able to lead to improved performance in two applications.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for very helpful comments and suggestions. This work was supported by the National Natural Science Foundation of China (No. 61472380).

5. REFERENCES

[1] E. Adelson and J. Wang, "Single lens stereo with a plenoptic camera," *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 14, pp. 99–106, 1992.

[2] R. Ng, M. Levoy, M. Bredif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," Tech. Rep., Stanford University, 2005.

[3] Yoav Y. Schechner and Nahum Kiryati, "Depth from defocus vs. stereo: How different really are they?," *International Journal of Computer Vision*, vol. 39, pp. 141–162, 2000.

[4] Vaibhav Vaish, Richard Szeliski, C. L. Zitnick, and Sing Bing Kang, "Reconstructing occluded surfaces using synthetic apertures: Stereo, focus and robust measures," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006.

[5] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[6] Allison Janoch, Sergey Karayev, Yangqing Jia, Jonathan T. Barron, Mario Fritz, Kate Saenko, and Trevor Darrell, "A category-level 3-d object dataset: Putting the kinect to work," in *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2011.

[7] Heiko Hirschmuller, Peter R. Innocent, and Jon Garibaldi, "Real-time correlation-based stereo vision with reduced border errors," *International Journal of Computer Vision*, vol. 47, pp. 229–246, 2002.

[8] Jing Zhang, Yang Cao, and Zengfu Wang, "A new image filtering method: Nonlocal image guided averaging," in *IEEE International Conference on Acoustic, Speech and Signal Processing*, 2014.

[9] A. Hosni, C. Rhemann, M. Bleyer, C. Rother, and M. Gelautz., "Fast cost-volume filtering for visual correspondence and beyond," *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 35, pp. 504–511, 2013.

[10] K. He, J. Sun, and X. Tang, "Guided image filtering," in *European Conference on Computer Vision (ECCV)*, 2010.

[11] A. Levin, A. Rav-Acha, and D. Lischinski, "Spectral matting," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.