A Semi-Global Matching method for large-scale light field images

Xiangsheng Huang
Institute of Automation, Chinese Academy of Sciences, Beijing, China

Ziling Huang
Institute of Automation, Chinese Academy of Sciences, Beijing, China
Ming Lu
Department of Electronic Engineering, TsingHua University, Beijing, China
Pengcheng Ma
Institute of Automation, Chinese Academy of Sciences, Beijing, China
Weili Ding
College of Electrical Engineering, Yanshan University, Qinhuangdao, China

***Abstract*—Semi-Global Matching (SGM) is a robust method in traditional stereo matching. It maintains precise boundary with low computational cost. However, directly applying SGM to light field stereo matching degrades the results greatly due to the sparsity of support points. In this letter, we proposes a novel stereo matching approach for large-scale light field images. We observe that adding weak edges to support points efficiently stabilizes the depth propagation. Based on this observation, we apply a cross detector to obtain support points, and then we propagate the depth of support points to homogeneous region. By solving a semi-global energy minimization problem, the depth information can be well estimated from epipolar plane images. Besides, we introduce a new strategy to deal with occlusion. We iteratively sample the pixels under current disparity hypothesis and the consistency scores are aggregated by a weighted winner-take-all strategy. Our method allows for significant reduce of the disparity search space, the time is halved and the depth is more robust at the occurrence of occlusion. For every pixel, the calculation is based on a single EPI and locally independent. Implementation on GPU shows that our method can achieve state-of-art results with less computational cost.**

***Index Terms*—3D reconstruction; light field;**

Stereo matching plays an important role in many computer vision applications, including 3D reconstruction, image supperresolution [8], synthetic aperture [15]. Traditional stereo methods focus on estimating the pixel correspondence between two or more images. Methods [1], [12], [17] based on local correspondence are typically fast, but these methods require an appropriate choice of window size and cannot be consistently matched in textureless area. Algorithms [7] based on global correspondence overcome afore-mentioned problems by imposing smoothness constraints on the depth image. However, these methods require large computational effort and storage capacity. Besides, these constraints will cause depth image to be over-smoothed and useful edge information to be lost. A compromise strategy [4], [5], [13], [14] is to reduce the ambiguities on correspondence by Semi-Global Matching. This strategy introduces an automatically determined window [5], [14] to preserve local structure. For example, this window can be obtained by over-segmentation [4] or triangulation [13]. The results are more robust with small computational cost.

Since the depth error increases quadratically with the distance, high-resolution images are needed to obtain accurate depth estimation. While traditional stereo matching has been exploited exhaustively, issues on light field gain much more attention recently. For 3D light field, the basic insight to scene reconstruction is first proposed by [2], as shown in Fig.1, images are captured densely along a linear path($C1$,$C2$ and $Cx(x = 3, 4, 5, ...)$ present the focal of different cameras). After obtaining the picture, we pile them up along line s. Slicing these pictures in a fix point v_* ,the EPI images can be obtained. Then every captured scene point corresponds to a linear trace in the epipolar plane image(EPI). The relation between the slope of the trace and the distance to camera can be denoted as

$$d = \frac{fb}{z} \tag{1}$$

where d is the disparity, f is the camera focal length in pixel,

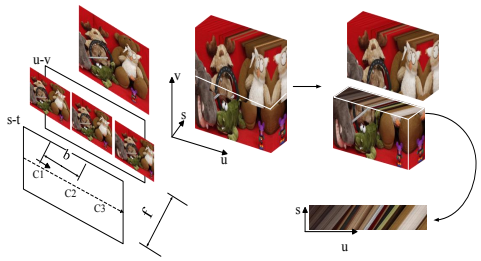

Fig. 1. The EPI-representation of 3D light field proposed by [2]. Every scene point corresponds to a linear trace in the epipolar plane image. The slope of trace is inversely proportional to the disparity.

and b is the metric distance between each adjacent pair of images. By dividing EPI into several tubes, [19] gave a more compact representation of 3D light field. Recently, [11], [18]

used local structure tensor as initial depth, then refined it by solving a total variation optimization problem. However, these global methods require high computational cost. This cost is unacceptable when it comes to high-resolution images. [3] proposed a fine-to-coarse strategy, they estimated depth images at different scales and propagated to full resolution. [16] introduced a bilateral consistency metric on surface camera to filter out occluded pixels, they obtained better reconstruction at the occurrence of occlusion. However, this bilateral metric is constrained to small field of view and cannot applied to real scenes.

In this Letter, we propose a new method to estimate the reconstruction from high resolution 3D light field images. Our method involves support points estimation and depth propagation. In the support points estimation step, we use cross detector to retrieve support points and obtain their depths by photo-consistency constraint. In the depth propagation step, we define a semi-global energy optimization problem similar to [14], also a weighted winner-take-all strategy is introduced to handle occlusion. Unlike [3], our method only estimates depth at full resolution and uses semi-global matching framework during depth propagation. By this framework, we limit the search space and efficiently reduce the computational cost. To the best of our knowledge, our method is the first approach to introduce Semi-Global Matching(SGM) into light field reconstruction.

We denote the light ray as **r**=L(u,v,s), where s is the 1D location and (u,v) represent the direction in image plane. In order to propagate at full resolution, we propose a new edge detector, as illustrated by Fig.2, we observed that there are two kinds of edges in EPI, specifically horizontal edge and vertical edge. Horizontal edge is well defined since it is also horizontal edge in original images. However, vertical edge is ill-defined because the vertical difference in EPI is not equal to difference in original images. So in [3], the author limited his work to horizontal edge. Actually, as shown in [11], [18], depth estimation in EPI is not related to original images. In this paper, we completely transfer the problem to EPI. We name horizontal edge as strong edge and vertical edge as weak edge. When we limit support points to strong edges, the sparsity of strong edges will reduce the consistency of reconstruction. However, introducing weak edges efficiently stabilize the results which allows for the introduction of Semi-Global Matching framework. In this paper, We propose a simple cross check detector, given by

$$S(u,s) = \sum_{(u',s')\in(V(u,s)\cup H(u,s))} \|E(u,s) - E(u',s')\|^2 \quad (2)$$

Where $V(u,s)$ means vertical neighborhood of (u,s) and $H(w,s)$ means horizontal neighborhood of (u,s).We threshold S by ε_e. Here ε_e is set to 0.015, we use this cross detector to check the edge confidence and threshold it to obtain support points.

This extension is simple but very important for semi-global matching. As demonstrated by Fig.3, if only strong edge is used, since the support points are too sparse, the horizontal depth propagation is inconsistent. With little additional cost, we introduce weak edge points, and this strategy allows us to robustly propagate depth into textureless regions.

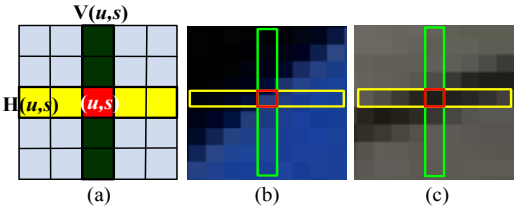

Fig. 2. (a)Illustration of our cross edge detector.(b)strong edge defined by [3].(c)our proposed weak edge

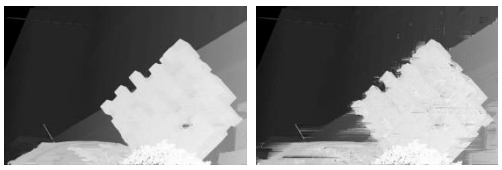

Fig. 3. The effect of adding weak edge. (Left) strong edge only. (Right)combine strong edge with weak edge

The binary mask limits the computation of depth at support points and we equally quantize the disparity between adjacent images into N levels, here N is 256. We obtain the depth by photo-consistency constraint. Before depth estimation, we select a fixed s, namely, we only reconstruct a single image of 3D light field. In this paper, we set s to the center image, denoted by $\hat{s}$. As shown in Fig.4, for every pixel in EPI(u,$\hat{s}$), we assign a hypothetical disparity $\hat{d}\in$[1,N], we can get the pixel set R along the red line w.r.t this disparity $\hat{d}$, R is constructed as follows:

$$R(u,\hat{d}) = \{E(u+(s-\hat{s})\hat{d},s)|s=1,2.....\} \quad (3)$$

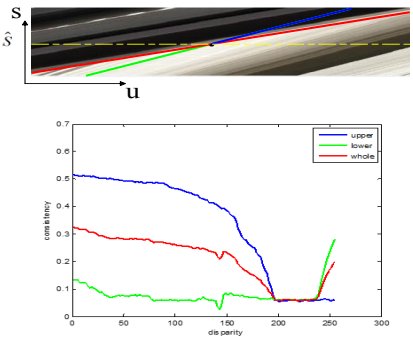

Fig. 4. Illustration of our weighted winner-take-all strategy at occlusion. The consistency in whole trace is small. However, the maximum of upper and lower parts is high. We sample the whole trace several times and use the maximum as the final consistency measurement.

Then we can define a photo consistent score P w.r.t this hypothetical disparity as

$$P_d(u,\hat{s}) = \frac{1}{|R(u,\hat{s})|} \sum_{(u^*,s^*)\in R} \varphi\left(E(u,\hat{s}) - E(u^*,s^*)\right) \tag{4}$$

$$\varphi(s) = 1 - e^{-\frac{s^2}{2\sigma^2}} \tag{5}$$

Here, we set σ = 1.0/255. In our experiments, We observe that this metric function always maintains better distinctiveness as illustrated by Fig.5. The true disparity is always at the local minimum using our metric. In fig.5 we compared our consistency metric with L2-norm(used by [3]), L1-norm at three regions. In the well-defined edge, the three metrics all maintains good distinctiveness. However, in textureless and complicate regions, our metric is better.

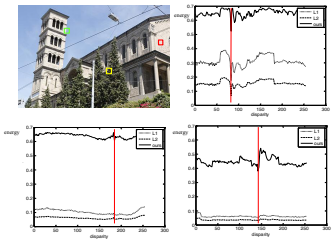

Fig. 5. Comparison of different metric functions under real scene reconstruction. In well-defined edge(top right), all metrics are distinctive. In textureless and complicate regions(second row), our metric stays better result.

Because photo-consistency is reduced at occlusions, the score defined above might fail to get the robust estimation. For example, the red point in Fig.4, occlusion happens when two line cross. The true line at red point is the blue/green line, however, due to occlusion, the consistency is low at this line. But, we observe that the maximum of two parts is high. In order to reliably estimate the slope, we apply a winner-take-all strategy. We randomly select a subset $\mathcal{S}$ from $R(u,\hat{d})$, and measure the consistency at this subset only. Besides, since for adjacent frames, occlusion is more likely to happen, we obtain a weight defined by $\sum_{s\in S}$ 1 - $\mathcal{G}$(s - $\hat{s}$). Here $\mathcal{G}$ is a Gaussian kernel. We weight the photo consistency above to get the final score of current sampling. We repeat this step and select the best score as the final consistency for the hypothetical disparity $\hat{d}$. Unlike [16], we don't differ the unoccluded pixels from the occluded pixels since our reconstruction is not limited to small field of view, the bilateral consistency metric is not suitable here. As shown in Fig.6, by this weighted winner-take-all strategy, the depth at occlusion is consistently estimated.

After we obtained the depths of support points, the depth is horizontally propagated to less detailed region to get the dense reconstruction. For every unknown pixel E(u,$\hat{s}$), we find the nearest two support points E(u_1,$\hat{s}$),E(u_2,$\hat{s}$),and linearly interpolate the expected disparity of E(u,$\hat{s}$). For simplicity, we just denote $\hat{s}$ as s in the following section.

$$\mu(u,s) = \frac{u-u_1}{u_2-u_1} D(u_2,s) + \frac{u_2-u}{u_2-u_1} D(u_1,s) \tag{6}$$

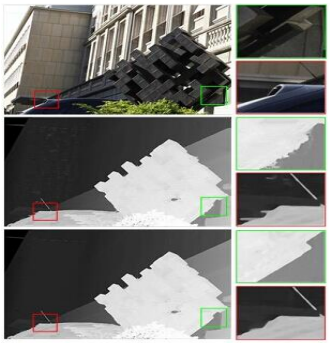

Fig. 6. Illustration of our winner-take-all strategy. As shown in the first row,the depth at occlusion can be consistently estimated with our strategy. The second row is the result of original paper

TABLE I
MAXIMUM DISPARITY

Image Set	Maximum disparity
Bike	12.5
Church	8.5
Statue	7.0
Couch	15.0

Denote the disparity in unknown pixel as d(u,s), we then suppose d(u,s) satisfying a Gaussian distribution with mean μ(u,s) and variation α. Defined by

$$E_p = \begin{cases} \phi + \exp(-\frac{(d-\mu)^2}{2\alpha^2}) & |d-\mu| < 3\alpha \\ 0 & others \end{cases} \tag{7}$$

here ϕ is the positive value to limit the value of prior in case it is too close to zero. As for the likelihood item, Similar to edge points, for a hypothetical d, we first get the set R and then define the likelihood of R for current d using Laplace distribution as

$$E_l = \exp(-\beta \frac{\sum_{E(u^*,s^*)\in R} ||E(u^*,s^*) - E(u,s)||_1}{|R|}) \tag{8}$$

Finally we take the negative logarithm of prior and likelihood to get the energy function for the hypothetical d in unknown point E(u,s) in EPI

$$E(d) = -log(E_p) + -log(E_l) \tag{9}$$

The disparity of E(u,s) can be obtained by minimizing equation(8). Our method is all operated in a single EPI and can be parallelized between EPIs. For every pixel in EPI, the propagation is independent and can be parallelized too. These attributes make our method computationally efficient. By combining EPI analysis with semi-global energy minimization, we get a depth map with precise edge and less noise.

Before evaluating our results, we fixed all the parameters in our experiments. α=5, ϕ=0.02, β=10. Though we equally quantize the disparity into N=256 levels, the maximum disparities vary from different images. as shown in Table 1.

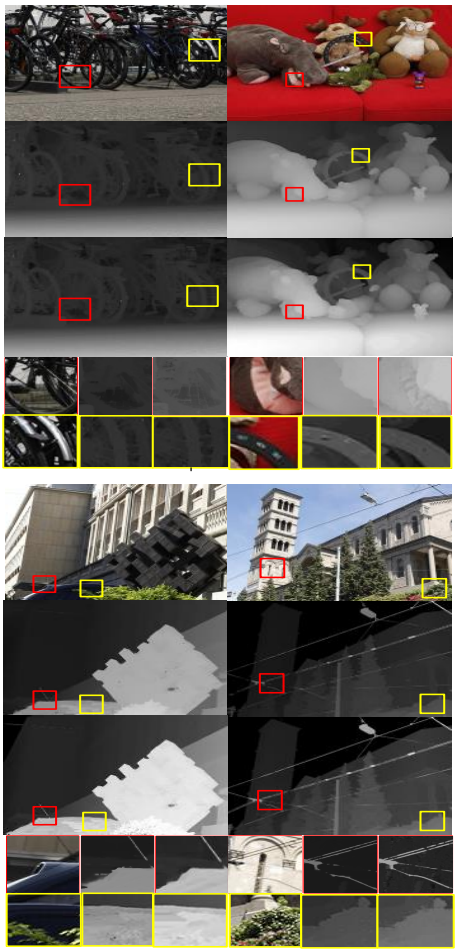

Fig. 7. Complete comparison with [3]'s result. The upper is [3]'s results and the lower is ours. Our results preserve better structure with halved time by introducing semi-global matching framework (red closeup). In complicate regions, the noise is less due to our weighted winner-take-all strategy at occlusion(yellow closeup).

For comparison, we reviewed related work on [3]'s dataset. [10] proposed a high-resolution stereo matching method using local plane sweeps, however this method is only suitable for binocular stereo matching. [9] used a statistical analysis framework to reduce the search space and achieved almost the same speedup as us. However, their method is noisy in textureless regions. Both methods didn't provide a complete comparison on the dataset of [3]. So here we only compare our work with [3]. As shown in Fig.7, with halved computational time, we achieve comparable or even better result(In our experiment,the computational time for orginal paper is 14 minutes to 15 minutes, but for our paper, we only uses 4 minutes to 5 minutes to obtain better results.). From the red closeup of Statue, the colors of car and building are similar, our method preserve better structure here since we only reconstruct the support points at full resolution. Moreover, from the yellow closeup, our propagation strategy is more smooth than the method proposed by [3] in complicate regions due to our weighted winner-take-all strategy at occlusion. To the best of our knowledge, this is the first work to introduce semi-global matching into the reconstruction of large-scale light field images. By semi-global matching, we efficiently reduce the search space of disparity during propagation. Besides, we propose a weighted sampling strategy in order to refine reconstruction at occlusion. Combining these two strategies, we achieve state-of-art results with less computational cost. Our work also has its limitation, As shown in Fig.7, in *Church*, the noise in textureless regions is large due to false estimation at weak edge. However, no filter is used in our method for visual satisfaction. The noise will be reduced if we add bilateral filter to our work, like bilateral median filter used by [3]. Besides, segmentation-based stereo matching methods achieve state-of-art results in binocular stereo matching. In following work, we also plan to combine segmentation with our semi-global matching framework in light field reconstruction.

REFERENCES

[1] Z.Di, L.Jie, and Z.Dongdong, "Robust visual correspondence computation using monogenic curvature phase based mutual information," Opt. Lett. **37,** 10–12 (2012).

[2] R.Bolles, H.Baker, and D.Marimont, "Epipolar-plane image analysis: An approach to determining structure from motion," International Journal of Computer Vision. **1,** 7–55 (1987).

[3] C.Kim, H.Zimmer, Y.Pritch, A.Sorkine-Hornung, and M.Gross, "Scene reconstruction from high spatio-angular resolution light fields," ACM Transactions on Graphics. **73,** 1–12 (2013).

[4] A.Klaus, M.Sormann, and K.Karner, "Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure," ICPR. **3,** 15–18 (2006).

[5] L.Wang, R.Yang, "Global stereo matching leveraged by sparse ground control points," In CVPR pp.3033–3040 (2011).

[6] D.Scharstein, and R.Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," International journal of computer vision. **47,** 7–42 (2002).

[7] V.Kolmogorov, Vladimir and R.Zabih, "Computing visual correspondence with occlusions using graph cuts," International Conference on Computer Vision. **2,** 508–515 (2001).

[8] T.Bishop, S.Zanetti, and P.Favaro, "Light field superresolution," IEEE International Conference on Computational Photography pp.1–9 (2009).

[9] W.Yilin, W.Ke, D.Enrique, Frahm, and M.Jan, "Stereo under Sequential Optimal Sampling: A Statistical Analysis Framework for Search Space Reduction," Computer Vision and Pattern Recognition(CVPR). pp.485–492 (2014).

[10] Sinha, N.Sudipta, D.Scharstein, and R.Szeliski "Efficient High-Resolution Stereo Matching using Local Plane Sweeps," Computer Vision and Pattern Recognition(CVPR). pp.1582–1589 (2014).

[11] S.Wanner, and B.Goldluecke, "Variational Light Field Analysis for Disparity Estimation and Super-Resolution," IEEE Transactions on Pattern Analysis and Machine Intelligence(2014).

[12] L.Jinjun, Z.Hong, J.Kejian, Z.Xiang, and T.Xingmin, "Multiscale stereo analysis based on local-color-phase congruency in the color monogenic signal framework," Opt. Lett. **35,** 2272–2274 (2010).

[13] Hirschmuller, and Heiko, "Stereo processing by semiglobal matching and mutual information," Pattern Analysis and Machine Intelligence, IEEE Transactions on. **30,** 328–341 (2008).

[14] G.Andreas, R.Martin, and U.Raquel, "Efficient Large-Scale Stereo Matching," Asian Conference on Computer Vision(2010).

[15] V.Vaish, B.Wilburn, N.Joshi, and M.Levoy, "Using plane+ parallax for calibrating dense camera arrays," Computer Vision and Pattern Recognition(CVPR), 2004.

[16] C.Can, L.Haiting, Y.Zhan, K.Sing Bing, and Y.Jingyi, "Light Field Stereo Matching Using Bilateral Statistics of Surface Cameras," Computer Vision and Pattern Recognition(CVPR), 2014.

[17] L.Jinjun, Z.Hong, Z.Xiang, and S.Chengying, "Robust stereo image matching using a two-dimensional monogenic wavelet transform," Opt. Lett. **34,** 3514–3516 (2009).

[18] Y. and G. Gbur, "Globally consistent multi-label assignment on the ray space of 4d light fields," Computer Vision and Pattern Recognition (CVPR). pp.3456–3458 (2013).

[19] A.Criminisi, Antonio, S.Kang, R.Swaminathan, R.Szeliski, and P.Anandan,"Extracting layers and analyzing their specular properties using epipolar-plane-image analysis," Computer vision and image understanding. **97,** 51–85 (2005).

[20] H.Xiaoyan, and P.Mordohai, "A quantitative evaluation of confidence measures for stereo vision," Pattern Analysis and Machine Intelligence, IEEE Transactions on. **34,** 2121–2133 (2012).