Light Field Stereo Matching Using Bilateral Statistics of Surface Cameras

Can Chen[1] Haiting Lin[1] Zhan Yu[1] Sing Bing Kang[2] Jingyi Yu[1]
[1]University of Delaware [2]Microsoft Research
canchen,haiting,yshmzhan,jingyiyu@udel.edu sbkang@microsoft.com

Abstract

In this paper, we introduce a bilateral consistency metric on the surface camera (SCam) [26] for light field stereo matching to handle significant occlusions. The concept of SCam is used to model angular radiance distribution with respect to a 3D point. Our bilateral consistency metric is used to indicate the probability of occlusions by analyzing the SCams. We further show how to distinguish between on-surface and free space, textured and non-textured, and Lambertian and specular through bilateral SCam analysis. To speed up the matching process, we apply the edge-preserving guided filter [14] on the consistency-disparity curves. Experimental results show that our technique outperforms both the state-of-the-art and the recent light field stereo matching methods, especially near occlusion boundaries.

1. Introduction

Stereo matching is a long standing problem in computer vision. Traditional binocular stereo suffers from ambiguity as a result of partial occlusions: it is not possible to establish correspondences on points observed in one view but occluded in the other. Solutions such as graph-cut rely on the smoothness term to “fill in” these holes. Occlusion ambiguity can be potentially addressed by using multiple views separated by different baselines. For example, an optimal joint estimate can be achieved by conducting pairwise binocular stereo matching on rectified images and then fusing all estimates onto a common 3D model [5]. For more general multi-view setups, a volumetric representation of the scene can be recovered using space carving [11].

Much has been done on handling occlusions for multi-view stereo matching. The graph-cut framework [10] uses an occlusion term for checking if the depth assignment violates the visibility constraint. [25] adds an additional second order smoothness terms and solves the graph-cut problem using Quadratic Pseudo-Boolean Optimization (QPBO) [15]. Bleyer et al. [3] impose soft segmentation and minimum description length to improve occlusion estimation. Heavy occlusion, however, remains difficult to address even with a large number of views. Fig. 7 shows plants with dense foliage that would be a challenge for any stereo algorithm; indeed, results from the multi-view stereo algorithm of [10] contain significant errors at occlusion boundaries.

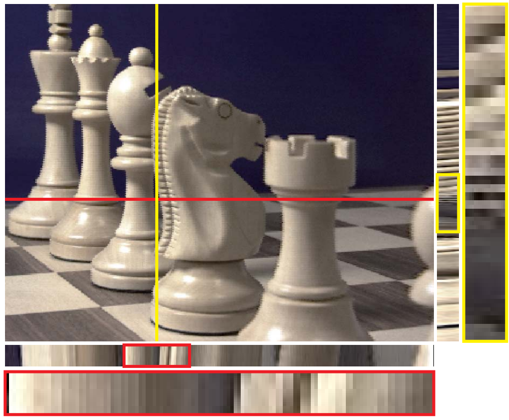


Figure 1. Epipolar Image (EPI) under occlusions. The chess scene captured by Lytro exhibits complex occlusions. Closeup views of the EPIs (of the red and yellow scanlines) show that near occlusion boundaries, the EPIs do not form clear structures for reliable direction/depth estimation.

There are commercially available light field cameras such as the Lytro and Raytrix cameras which are capable of capturing a few hundred of views in a single shot. For example, using an 11MP sensor and 0.1 million microlenses, the Lytro camera can acquire 100 views of the scene. These advances have renewed interest in stereo matching. Heber *et al.* [6] propose a matching cost based on Active Wavefront Sampling, which dose not explicitly model occlusion. Wanner and Goldlücke [18, 19, 22] apply structure tensors to estimate the feature pixels’ directions in 2D Epipolar Image (EPI) and use them in stereo matching and object segmentation. Yu *et al.* [27] encode the constraints of 3D lines to further improve the reconstruction quality. However, both techniques are vulnerable to heavy occlusion: the directional field become too random to estimate (Fig. 1) and

3D lines are partitioned into small, incoherent segments.

In this paper, we present a new light field stereo matching algorithm that is capable of handling heavy occlusion. Our approach builds upon the analysis of the angular statistics of the light field: for every 3D point, we trace all rays passing through it back to the camera array. We then construct an angular sampling image called the surface camera or SCam [26]. We propose a new bilateral metric to measure the probability of occlusions of the SCam. We further show how to distinguish between on-surface and free space, textured and non-textured, and Lambertian and specular through global and local SCam consistency analysis.

Specifically, for a pixel in the reference view, we compute its bilateral metric at every possible depth. We show that the metric reaches a local minimum at the ground truth depth; this is a unique property to light field stereo. To reliably handle ambiguous textureless surfaces, we introduce an additional set of local and global confidence metrics to gauge the reliability in depth estimation. Finally, we develop an efficient filtering-based light field stereo matching technique that can be parallelized on the GPU for efficient processing. Experimental results show that our technique significantly outperforms both the state-of-the-art and the recent light field stereo matching methods.

2. Angular Light Field Statistics

To represent a light field, it is common practice to use the two-plane parametrization (2PP) as shown in Fig. 2. The camera plane st is at $z = 0$ and the image plane uv is at $z = 1$. A ray r not parallel to the parametrization planes in the space is uniquely specified as a 4D vector $r = [s, t, u, v]$, where $[s, t, 0]$ is its intersection with the st plane and $[u, v, 1]$ with the uv plane. In practice, $[s, t]$ can be viewed as the coordinates of the camera and $[u, v]$ as those of the sensor.

We first describe our notations. Our analysis is conducted with respect to a reference $R(s_r, t_r)$, *e.g.*, the central camera of the light field. Let p denote a pixel at location (u, v) in R. We use p_d to represent the 3D point (u, v, d), *i.e.*, d is the depth along ray $[s_r, t_r, u, v]$. For every 3D point p_d, we back project it to every camera (s, t) in the light field. This is equivalent to gathering, at each st camera, a ray (pixel) $[u(s, t), v(s, t)]$ that passes through p_d and then organizing them as an st image. This model has been previously referred to as the Surface Camera (SCam) [26] or the Surface Light Field [24]. We denote the SCam at p_d as $A_{p_d}(s, t)$.

2.1. SCam Analysis

The characteristics of an SCam image A_{p_d} depends on the 3D location of p_d, scene structure, and scene reflectance/texture properties. In our analysis, we assume the field-of-view (FOV) of the SCam is small (we use the Lytro camera, whose Scam has about 10 degree FOV).

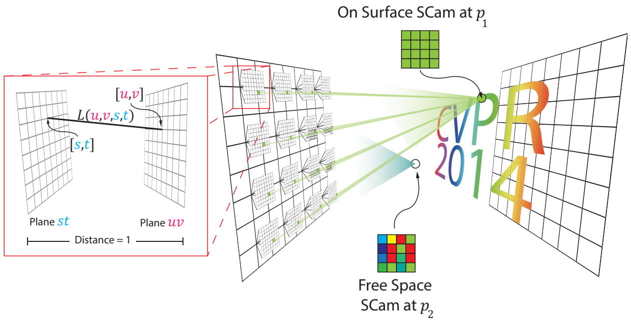


Figure 2. A light field is parameterized using the *st* camera plane and the *uv* image plane. A Surface Camera at P is formed by tracing all rays passing through P back to all st cameras to form an *st* image. [26].

Consider the simplest case of an unoccluded point on a Lambertian surface. Rays from different views should have identical radiances and its SCam should be an image of constant color image, *e.g.*, the SCam of green circled point at its true depth in Fig. 3. If p_d lies on a specular surface, the SCam will then exhibit view dependencies. However, since the field-of-view is small, the variation will be small and its SCam will appear smooth, *e.g.*, the red circled point at its true depth in Fig. 3. This is the on-surface case.

If p_d is free space, then its SCam will appear precisely as a pinhole image at p_d, *i.e.*, it will collect rays emitting from different points from nearby surfaces. If the nearby surfaces to p_d are textureless, its SCam will appear similar to the on-surface cases, exhibiting depth ambiguity commonly observed in stereo matching. However, if the nearby surface is highly textured, then the SCam will resemble the texture pattern, *e.g.*, the brown or green circled point at a wrong depth in Fig. 3. This is the free space case.

Next, let us consider the occlusion case. Recall that if p_d lies on some surface that is occluded from some views in the light field, part of its SCam should appear as on-surface and part free-space. The SCam images in this case will exhibit abrupt color/texture changes. If the scene is heavily occluded, the SCam will exhibit complex textures and colors, *e.g.*, the blue circled point at its ground true depth in Fig. 3.

2.2. Bilateral Consistency Metric (BCM)

Our analysis further reveals that we can potentially analyze the content of the SCams to distinguish their occlusion profiles. Let $\Omega^{v*}_{p_d}$ be the set of rays (pixels) collected by the SCam that reach p_d without occlusion and $\Omega^{o*}_{p_d}$ be those that are blocked from reaching p_d due to occlusions.

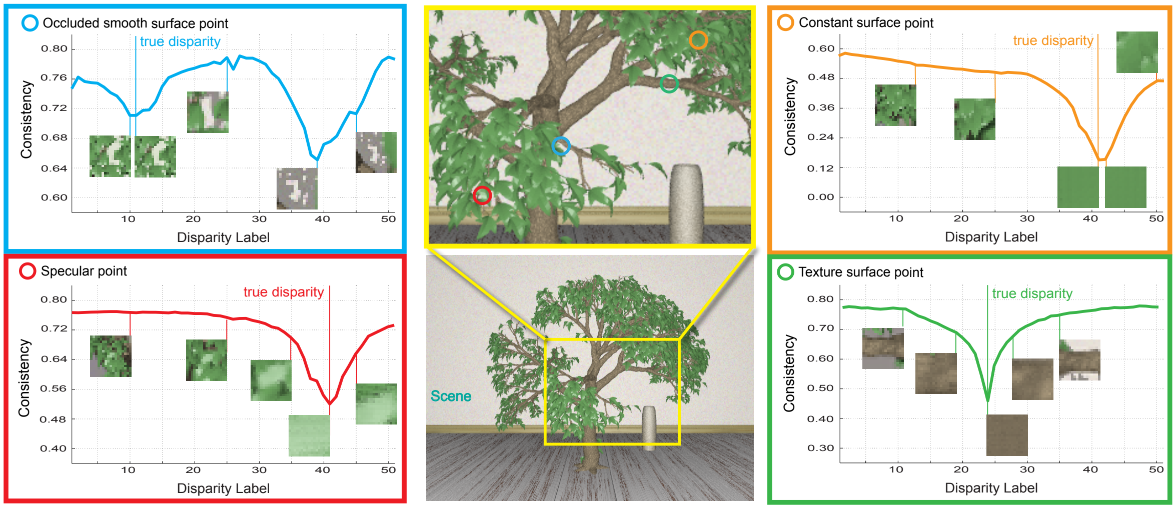


Figure 3. The consistency-depth (C-D) curves of different points in a plant scene. Consistency is measured using the bilateral consistency metric and we map depth to disparity for clarity. The C-D curves have different profiles for different points. Notice that the ground truth disparities correspond to the local/global minimum. The SCams at the sampled disparities are also plotted on the curve.

We define the following consistency measurement:

$$c^*_{p_d} = \frac{1}{|\Omega^{v*}_{p_d}|} \sum_{(s,t)\in\Omega^{v*}_{p_d}} \rho(A_{p_d}(s,t) - A_{p_d}(s_r,t_r)), \qquad (1)$$

where $|\Omega|$ is the size of the set Ω and $\rho(x) = 1 - e^{-\frac{x^2}{2\sigma^2}}$ is a distance function, where $\sigma(= \frac{1}{255}$ in our experiments) controls the sensitivity of the function to large distances. The reason for using this distance function instead of L2 norm is explained in Section 2.3.

For the on-surface case, we have $\rho(A_{p_d}(s,t) - A_{p_d}(s_r,t_r)) = 0, (s,t) \in \Omega^{v*}_{p_d}$ and therefore $c^*_p = 0$. Even if the surface is non-Lambertion, $c^*_{p_d}$ would still close to zero due to the narrow SCam field-of-view assumption. However, if p_d is free space or occluded, $c^*_{p_d}$ would be larger.

The simplest stereo matching algorithm would be to locate p_d that minimizes $c^*_{p_d}$. Unfortunately, there are two significant issues. First, the set $\Omega^{v*}_{p_d}$ is unknown before scene geometry is recovered. Second, if the surface is textureless, the off-surface SCam will appear as if it is on-surface, causing ambiguity. Previous multi-view algorithms [8, 10, 23] implicitly incorporates this visibility constraint into their optimization framework, *e.g.*, as consistency priors.

To handle these two issues, we introduce a new bilateral consistency metric (BCM) $P_{A_{p_d}}$ to estimate the probability of each pixel in SCam A_{p_d} belonging to $\Omega^{v*}_{p_d}$. Our metric resembles bilateral filters on color and spatial [17] for estimating how close each pixel in the SCam $A_{p_d}(s,t)$ is to the reference pixel $A_{p_d}(s_r,t_r)$:

$$\begin{aligned} P_{A_{p_d}}(s,t) &= e^{-\frac{d_c^2}{2\sigma_c^2} - \frac{d_s^2}{2\sigma_s^2}}, \\ d_c &= A_{p_d}(s,t) - A_{p_d}(s_r,t_r), \\ d_s &= (s,t) - (s_r,t_r), \end{aligned} \qquad (2)$$

where σ_c and σ_s are color and spatial variances (set as $\frac{3}{255}$ and $\frac{1}{4}$ in our experiments) respectively.

If we assume that the size of $\Omega^{v*}_{p_d}$ is at least N_v, we can sort all pixels in A_{p_d} using their BCM and approximate $\Omega^{v*}_{p_d}$ as:

$$\Omega^{v}_{p_d} = \{(s,t) | P_{A_{p_d}}(s,t) \geq \min(P_{Thresh}, P^{N_v}_{A_{p_d}})\} \qquad (3)$$

where $P_{Thresh}(= 0.5)$ is a predefined threshold and $P^{N_v}_{A_p}$ is the N_v-th highest BCM among all pixels in the SCam. (N_v is set as the half of the total number of views.) Finally, we can use the estimated visibility set to compute the consistency measure c as:

$$c_{p_d} = \frac{1}{|\Omega^{v}_{p_d}|} \sum_{(s,t)\in\Omega^{v}_{p_d}} \rho(A_{p_d}(s,t) - A_{p_d}(s_r,t_r)). \qquad (4)$$

2.3. Consistency-Depth Profile

Given a pixel p in the reference view, we can compute the consistency measure c_{p_d} for every hypothetical depth d. Therefore, we can form a c_{p_d}- d curve or the C-D curve. Let us study the profile of the C-D curves. For the on-surface case, if the surface is highly textured, the consistency measure should reach global minimum at the ground truth depth. If the surface is textureless, then we should

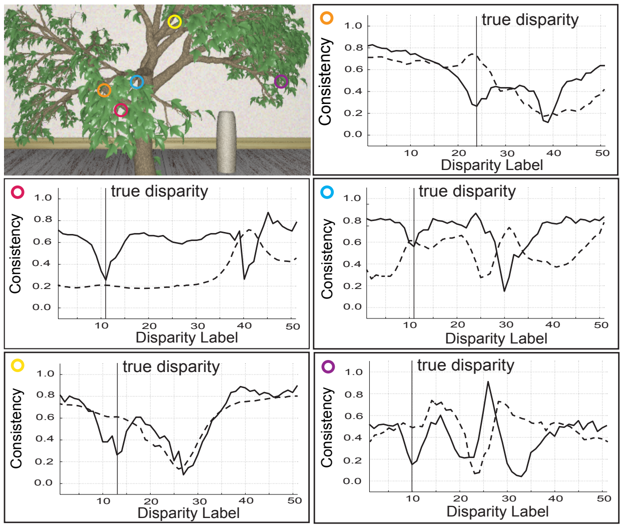


Figure 4. C-D curves for partially occluded scene points. The ground truth disparity always correspond to the local minimum on the curve when using our bilateral consistency metric (solid curves). However, it does not hold when using the traditional L2 consistency metric.

observe a trough (wide flat range of minima) due to ambiguity. For the occlusion case, if the occluding surface has a different color from the ground truth point, since our BCM already excludes the pixels from the occluders, the consistency measure should still be the global minimum.

However, if the occluder has a similar color of the reference pixel, BCM cannot reliably prohibit its corresponding pixels (denoted as set $\Omega^{err}_{p_d}$) from being included in $\Omega^{v}_{p_d}$. As a result, c_{p_d} may not be the global minimum. However, it will always correspond to a local minimum, as shown in Fig. 4. In particular, small depth changes in either direction $\pm\delta$ will produce values $c_{p_{d\pm\delta}}$ that are larger than c_{p_d}. Conceptually, this is because the incoherence for the pixels in $\Omega^{v*}_{p_d}$ will increase, whereas the incoherence for the pixels in $\Omega^{err}_{p_d}(=\Omega^{v}_{p_d}\setminus\Omega^{v*}_{p_d})$ will remain approximately the same due to the insensitiveness of ρ to large distances. The complete proof can be found in the Appendix A.

In summary, for each pixel p, its ground truth depth should always correspond to some minimum on its C-D curve. It is important to note that the local minimum property is unique for our C-D curve based on robust distance function ρ. The property does not hold when using the classical L2 norm to measuring the consistency. Fig. 4 compares our C-D curve and L2 norm curve v_{p_d}:

$$v_{p_d}=\frac{1}{|\Omega^{v}_{p_d}|}\sum_{(s,t)\in\Omega^{v}_{p_d}}(A_{p_d}(s,t)-A_{p_d}(s_r,t_r))^2. \qquad (5)$$

Fig. 4 shows that in the presence of occlusion, the correct depth is not always at a minimum of v_{p_d}.

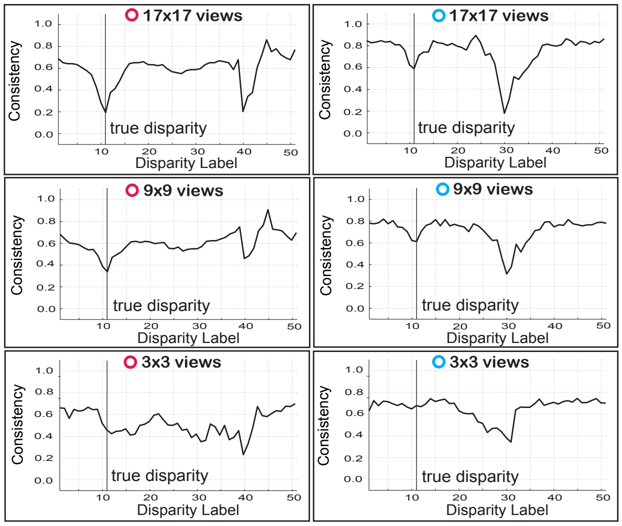


Figure 5. The C-D becomes more oscillating when the number of view decreases. The ground truth depths of the red and blue points of Fig. 4 no longer correspond to the local minimums with 3 × 3 views.

Note that our robust metric for stereo matching is applicable only for light fields with large numbers of densely sampled views. It will not be effective if we use a small set of views (*e.g.*, $5\sim 10$) as in traditional multi-view stereo matching. This is because $\Omega^{v}_{p_d}$ will be too small to reliably measure. Fig. 5 shows the C-D curves computed using fewer SCams for the same scene. When the number of views is fewer than 5×5, there is a noticeable degradation of the local minimum property.

3. Light Field Stereo Matching

Our new light field stereo matching algorithm is based on the minimum property of the C-D curve. The most direct implementation would be to use the consistency measure c_{p_d} in place of the data term in classical graph-cut [4] or belief propagation [16] algorithms. However, both algorithms are computationally expensive, especially with a large number of views and disparity labels. We instead developed a more efficient technique based on recent filtering-based stereo matching methods [13, 14] that can be parallelized on the GPU.

We first apply the edge-preserving guided filter [14] to smooth the initial consistency measure c_{p_d}. To assign depth/disparity to every pixel p, we refine c_{p_d} using a local confidence measure for resolving the ambiguity caused by a valley of minima on the C-D curve. We then measure the profile of p's C-D curve (*e.g.*, the number of local minima) to determine an overall confidence. If the confidence is sufficiently high, we assign p's depth using the global minimum on the C-D curve. Otherwise, we mark the pixel's depth as “unknown”. Once we process all pixels, we apply depth propagation similar to [12] to fill in the holes associated with the “unknown” pixels based on the color coherence. The complete algorithm is shown in Algorithm 1.

Algorithm 1 Light Field Stereo Matching

```
Require: Light Field Input $I_{(s,t)}$, Reference View $(s_r, t_r)$
  procedure DEPTH ESTIMATION
    for all $p$ do
      for all $d \in [d_{min}, d_{max}]$ do
        Compute $c_{p_d}$ based on Eq. 5;
        Apply edge preserving filter on $c_{p_d}$;
        Compute local confidence $f^l_{p_d}$ using Eq. 7;
        Compute $\tilde{c}_{p_d}$ using Eq. 8;
      end for
      Compute the reliability $f^g$ and mark $p$ as reliable/unreliable.
      If $p$ is reliable, assign the global min of $\tilde{c}_{p_d}$ to $p$.
      Otherwise, assign $p$ as unknown.
    end for
    Fill in “unknown” pixels using propagation.
  end procedure
```

Local Confidence Measurement. If the region is textureless, matching becomes ambiguous, since the ground truth c_{p_d} lies within a valley of global minima on the C-D curve. In the MRF-based solutions, the issue is addressed via the smoothness prior, *i.e.*, by forcing the pixel's depth to be similar to its reliable neighbors. Several confidence measures have been previous proposed [7], *e.g.*, based on the existence of local minima and/or flat regions of low cost on each individual cost profile.

We propose a new local confidence measurement that makes use of the C-D curve and spatial coherence. Instead of examining the profile along dimension d only, our confidence measurement incorporates nearby pixels; we compute the stability of c_{p_d} under perturbation. Let A'_{p_d} be a perturbed version of SCam image A_{p_d}:

$$A'_{p_d}(s,t) = \frac{1}{|\Omega'|} \cdot \sum_{(du,dv)\in\Omega'} I_{(s,t)}(u_{p_d}+du, v_{p_d}+dv)) \quad (6)$$

where $\Omega' = \{(-1,0),(1,0),(0,-1),(0,1)\}$, $(du,dv) \in \Omega'$, $I_{(s,t)}$ is the image of st-th view and $A_{p_d}(s,t) = I_{(s,t)}(u_{p_d}, v_{p_d})$.

We define the local confidence $f^l_{p_d}$ as:

$$f^l_{p_d} = 1 - e^{-(c_{p_d} - c'_{p_d})^2/(2\sigma_l^2)}, \quad (7)$$

where c'_{p_d} is the consistency measure of A'_{p_d} and the term σ_l controls the scale of the confidence and is same to all pixels. The local confidence measure penalizes the depth estimated at textureless region while giving more confidence for those depth estimated from textured region. We use the local confidence measure to compute the new consistency measure $\tilde{c}_{p_d}$ as:

$$\tilde{c}_{p_d} = 1 - (1 - c_{p_d}) \cdot f^l_{p_d}. \quad (8)$$

Reliability Measure. Next, we assign depths to each pixel based on their consistency $\tilde{c}_{p_d}$. Specifically, for each pixel p, we select the depth d that corresponds to the smallest $\tilde{c}_{p_d}$ for all possible d. Recall that the C-D curve can have multiple local minima adding to the local confidence measurements. Therefore, the assignment using the global minimum is not reliable. This happens frequently in the cases where occluders have similar colors to that of the occluded smooth surface.

Instead, we compute a global confidence measurement to determine if a pixel can be reliably labeled or not. Let $\{\tilde{c}_{p_{d_1}}, \tilde{c}_{p_{d_2}}, ..., \tilde{c}_{p_{d_n}}\}$ represent a set of local minima on the C-D curve arranged in ascending order, where n is the number of local minima whose $\tilde{c}_{p_d}$ is less than a predefined threshold. The global confidence for a depth estimation is defined as:

$$f^g = (\tilde{c}_{p_{d_1}} - \tilde{c}_{p_{d_2}})/(\tilde{c}_{p_{d_1}} - \tilde{c}_{p_{d_n}}). \quad (9)$$

The global confidence is computing the relative gap between the global and the second smallest local minimum. The smaller the gap, the less reliable of the depth estimation using the global minimum. For pixels deemed unreliable, we resort to depth propagation [12].

4. Experiments

We first evaluate our algorithm on the Synthetic Light Field datasets used in [18, 20, 21]. We compare our technique with the recent method of globally consistent depth labeling (GCDL) [18] and line-assisted graph-cut (LAGC) [27] using the results or source code from the respective authors. We also implemented the more classical multi-view graph-cut (MVGC) [10]. Finally, in Fig. 8, we compare with the published results of the fast light field stereo (FLFS) [9] whose source code is not yet available.

Fig. 6 compares results generated by different methods on three sets of light fields (Maria, cube and still life) with the ground truth disparity maps. All scenes exhibit complex occlusions. For clarity, we further show the closeup views of the error maps. Our technique outperforms GCDL, MVGC, and LAGC near the occlusion edges in all three examples and our depth maps better preserve the contour of the occlusion boundaries. MVGC produces much noisier results whereas GCDL and LAGC incur boundary fattening.

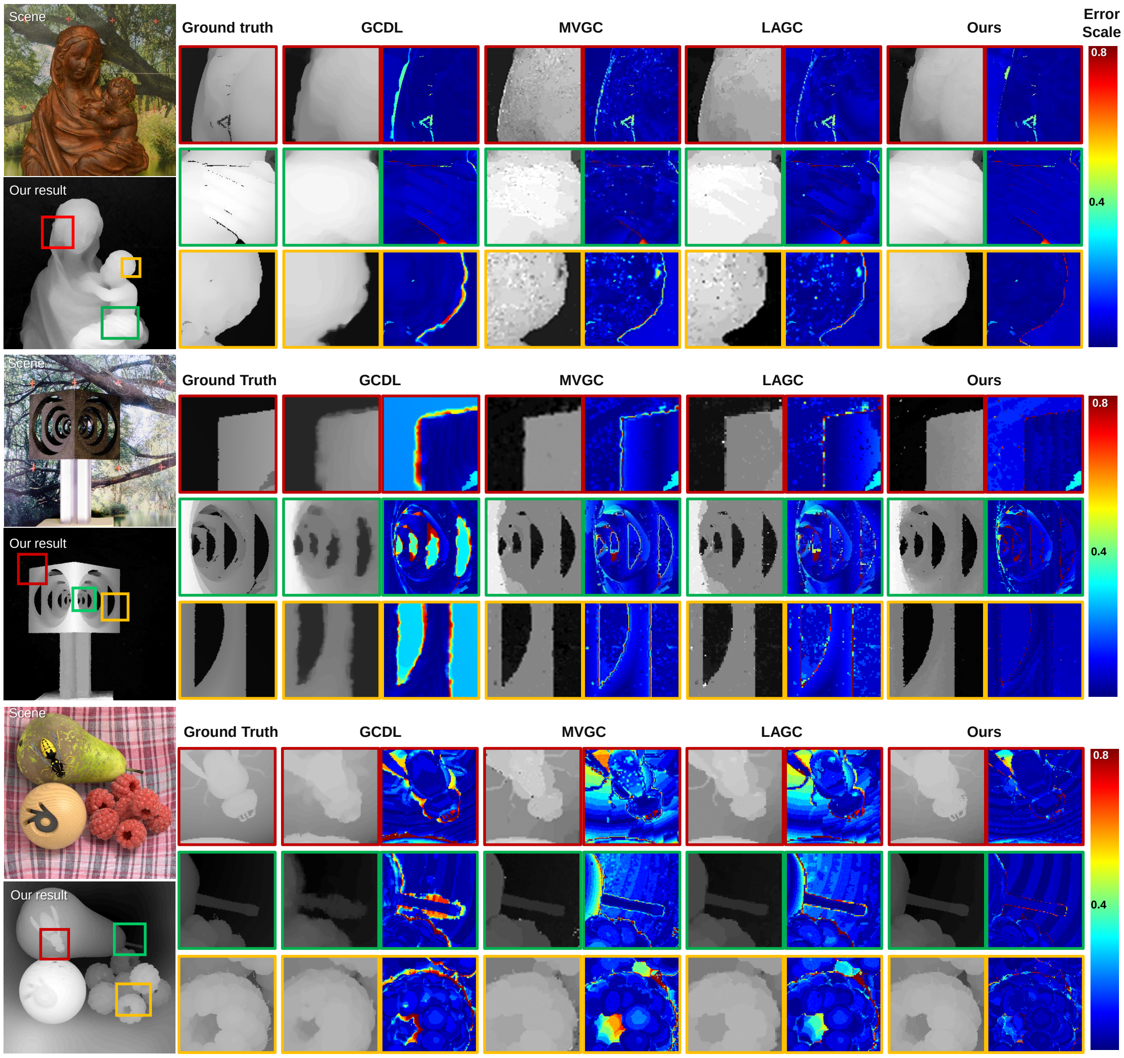


Figure 6. Reconstruction errors using our approach, GCDL [18], MVGC [10], and LAGC [27] on the Maria scene.

Table 1. Mean square errors of disparity recovery computed for all pixels and occlusion pixels.

Scene	GCLD		MVGC		LAGC		Our result	
	overall	occlusion	overall	occlusion	overall	occlusion	overall	occlusion
maria	**0.001074**	0.016240	0.002281	0.015064	0.002548	0.016888	0.001273	**0.013989**
cube	**0.008675**	0.054887	0.019067	0.034319	0.018621	0.044417	0.010940	**0.032793**
still life	0.033672	0.202133	0.066728	0.134544	0.041361	0.183859	**0.012868**	**0.066527**
Tree	N/A	N/A	0.088057	0.088366	0.073049	0.073479	**0.072461**	**0.072523**

Our technique is particularly suitable for handling scenes with heavy occlusions. Fig. 7 shows two synthetic light fields of trees rendered using the POV-Ray ray-tracer [2] at a resolution of 600 × 470 × 17 × 17 and 480 × 640 × 17 × 17. The scenes both have disparity ranges of 0 to 13 pixels. Notice that the structure tensor based GCDL cannot handle such a large disparity range. Heavy occlusion is also a challenge to classical MVGC. Notice that parts of the scenes such as leaves lack texture while other parts have complex textures. Our result is able to accurately recover most leaves and at the same time maintain smooth disparity transitions of the ground, the background wall, and the trunk of the trees. In contrast, MVGC produces strong visual artifacts by merging individual leaves into groups.

Table 1 summarizes the accuracy of our technique and these state-of-the-art techniques for all pixels and for occlusion pixels. Specifically, we compute the mean square errors of the disparity maps as in [21]. The overall errors using GCDL, LAGC, and our technique are comparably low when occlusions are less severe. However,

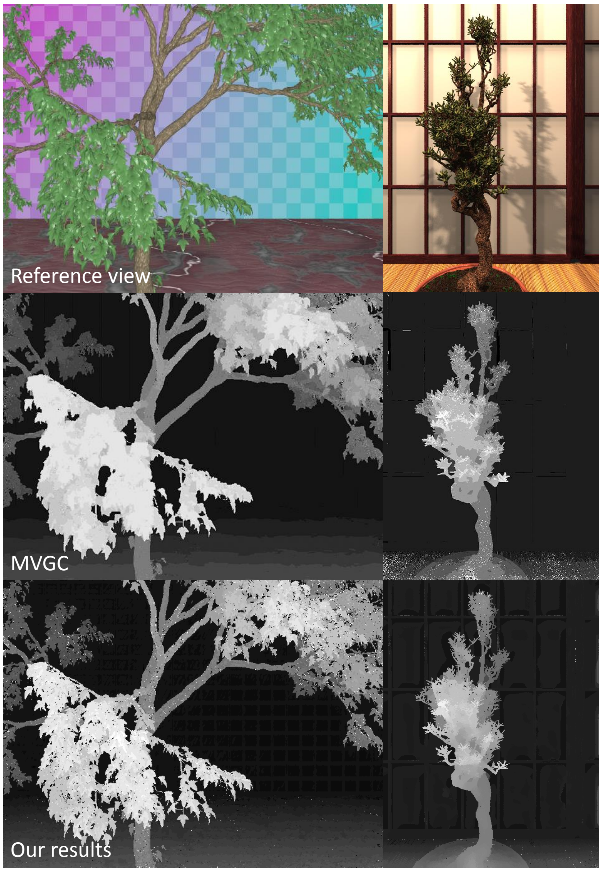


Figure 7. Stereo matching results on heavily foliaged plants using MVGC and our technique.

for heavily-occluded scenes (the bottom rows) such as the plants (Fig. 7), our technique performs better, especially near the occlusion boundaries. Recall that GCDL can only handle small disparity range (generally $[-4, 4]$) since it relies on subpixel accuracy for tensor estimation. We did not include its results on the plants models. Our technique does not have disparity range requirements and can handle these large disparity ranges.

Next, we compare our techniques with GCDL [18] and FLFS [9] on the Stanford Gantry data set [1]. The light field contains 17×17 views at a resolution of 1280×960 of a Lego gantry crane model. The disparity range is between -4 to 4 pixels and we discretize it into 80 steps. Fig. 8 shows the comparisons. All three methods produce reasonable results. However, near the occlusion boundaries such as the hoist rope from the crane and the contours of the headlights and windows, our solution much better preserves the shape of the contours.

Fig. 9 shows the effectiveness of guided filtering. We compute the disparity maps on the Buddha scene and the Mona Lisa scene, both composed of 9×9 views at a resolution of 768×768 with disparity range $[-1.2, 1.5]$. The Buddha scene consists of large textureless regions such as the wood plank and the dices. Without applying the guided filter, the disparities of the dots on the dices are incorrectly estimated. This is because the ground truth disparity lies at a trough on its C-D curve whose confidences are equally low. Using the guided filter, our technique is able to bias more towards the boundary of the dot whose confidence is much higher due to textures and therefore forces the interior of the dots to choose the correct disparity. The process is analogous to adding the smoothness prior to the graph-cut framework to resolve ambiguity.

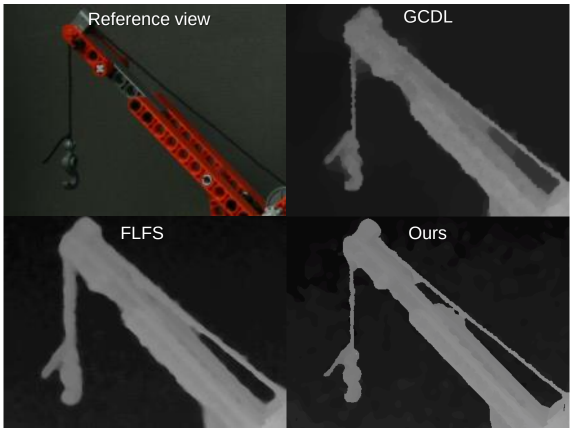


Figure 8. Comparison using our method, GCDL [18] and FLFS [9] on the Stanford Gantry data.

5. Discussion and Future Work

We have presented a new light field stereo matching algorithm by modeling the angular light field statistics using the SCam. To characterize SCam, we have developed a new bilateral consistency metric for measuring the reliability of the SCam. The analysis reveals the importance of the SCam structure in multiview stereo. In particular, we have shown that the consistency-depth curve has the property that the ground truth depth always corresponds to a minimum on the curve. This is a unique property under the dense view assumption and the bilateral metric. Finally, we have developed a filter-based stereo matching technique and demonstrated that it outperforms the state-of-the-art solutions especially near the occlusion boundaries.

The key component of our approach is the bilateral consistency metric. This metric, however, is biased towards to the reference view as it uses the color of the reference pixel as the mean of the bilateral filter. In contrast, the traditional multi-view stereo such as MVGC uses the mean of all pixels in the SCam as the reference and compute the L2 difference. Therefore, MVGC is expected to be more robust when the input images are noisy. The problem can be partially com-

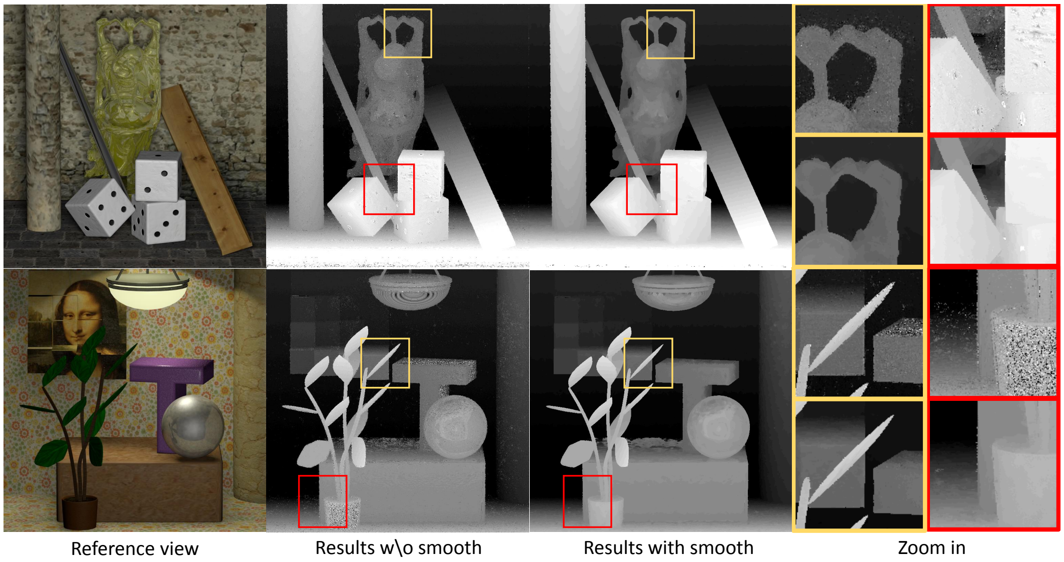


Figure 9. Results of our method with and without guided filtering.

pensated by first denoising the input images. Alternatively, we can incorporate the noise model into the consistency metric, which is part of our immediate future work. Finally, we plan to build a database of light field data with small to large disparity ranges and use it as a benchmark to compare existing light field stereo algorithms.

Acknowledgement

This project was partially supported by the National Science Foundation under grants IIS-CAREER-0845268 and IIS-1218156.

References

[1] The (new) stanford light field archive. http://lightfield.stanford.edu/lfs.html. 7

[2] The persistence of vision raytracer. http://www.povray.org/. 6

[3] M. Bleyer, C. Rother, and P. Kohli. Surface stereo with soft segmentation. In *CVPR*, 2010. 1

[4] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. In *ICCV*, 1999. 4

[5] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In *ACM SIGGRAPH*, 1996. 1

[6] S. Heber, R. Ranftl, and T. Pock. Variational Shape from Light Field. In *International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition*, 2013. 1

[7] X. Hu and P. Mordohai. A quantitative evaluation of confidence measures for stereo vision. *IEEE TPAMI*, 2012. 5

[8] S. B. Kang, R. Szeliski, and J. Chai. Handling occlusions in dense multi-view stereo. In *CVPR*, 2001. 3

[9] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross. Scene reconstruction from high spatio-angular resolution light fields. In *ACM SIGGRAPH*, 2013. 5, 7

[10] V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph cuts. In *in European Conference on Computer Vision*, pages 82–96, 2002. 1, 3, 5, 6

[11] K. N. Kutulakos and S. M. Seitz. A theory of shape by space carving. *International Journal of Computer Vision*, 38(3):199C218, 2000. 1

[12] A. Levin, D. Lischinski, and Y. Weiss. Colorization using optimization. In *ACM SIGGRAPH*, 2004. 5

[13] Z. Ma, K. He, Y. Wei, J. Sun, and E. Wu. Constant time weighted median filtering for stereo matching and beyond. In *ICCV*, 2013. 4

[14] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz. Fast cost-volume filtering for visual correspondence and beyond. In *CVPR*, 2011. 1, 4

[15] C. Rother, V. Kolmogorov, V. Lempitsky, and M. Szummer. Optimizing binary mrfs via extended roof duality. In *CVPR*, 2007. 1

[16] J. Sun, N.-N. Zheng, and H.-Y. Shum. Stereo matching using belief propagation. *IEEE TPAMI*, 2003. 4

[17] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *ICCV*, 1998. 3

[18] S. Wanner and B. Goldluecke. Globally consistent depth labeling of 4d light fields. In *CVPR*, 2012. 1, 5, 6, 7

[19] S. Wanner and B. Goldluecke. Spatial and angular variational superresolution of 4d light fields. In *ECCV*, 2012. 1

[20] S. Wanner and B. Goldluecke. Variational light field analysis for disparity estimation and super-resolution. *IEEE TPAMI*, 2013. 5

[21] S. Wanner, S. Meister, and B. Goldluecke. Datasets and benchmarks for densely sampled 4d light fields. In *Vision, Modelling and Visualization (VMV)*, 2013. 5, 6

[22] S. Wanner, C. Straehle, and B. Goldluecke. Globally consistent multi-label assignment on the ray space of 4d light fields. In *CVPR*, 2013. 1

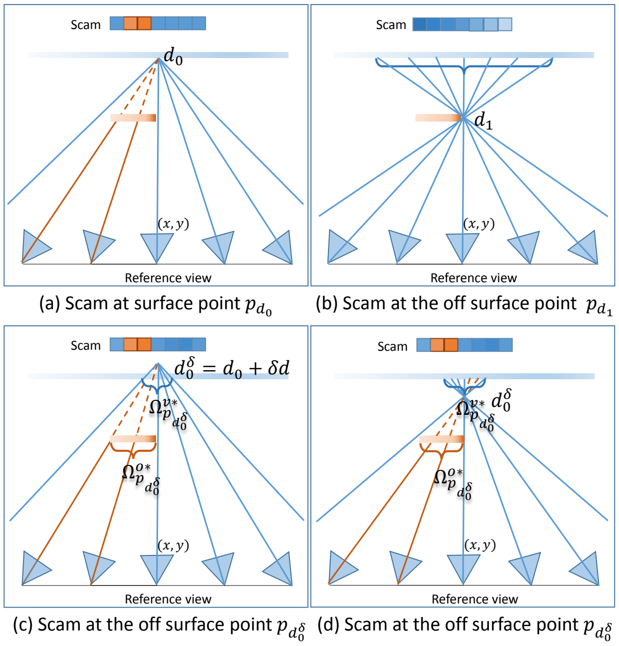
[23] Y. WEI and L. QUAN. Asymmetrical occlusion handling using graph cut for multi-view stereo. In *CVPR*, 2005. 3

[24] D. Wood, D. Azuma, W. Aldinger, B. Curless, T. Duchamp, D. Salesin, and W. Stuetzle. Surface light fields for 3d photography. In *ACM SIGGRAPH*, 2000. 2

[25] O. J. Woodford, P. H. S. Torr, I. D. Reid, and A. W. Fitzgibbon. Global stereo reconstruction under second order smoothness priors. In *CVPR*, 2008. 1

[26] J. Yu, L. McMillan, and S. Gortler. Surface camera (scam) light field rendering. *International Journal of Image and Graphics (IJIG)*, 4, 2004. 1, 2

[27] Z. Yu, X. Guo, H. Lin, A. Lumsdaine, and J. Yu. Line-assisted light field triangulation and stereo matching. In *ICCV*, 2013. 1, 5, 6


Scam
d_0
(x, y)
Reference view
Scam
d_1
(x, y)
Reference view
Scam
$d_0^\delta = d_0 + \delta d$
$\Omega^{v*}_{p_{d_0^\delta}}$
$\Omega^{o*}_{p_{d_0^\delta}}$
(x, y)
Reference view
Scam
d_0^δ
$\Omega^{v*}_{p_{d_0^\delta}}$
$\Omega^{o*}_{p_{d_0^\delta}}$
(x, y)
Reference view

(a) Scam at surface point p_{d_0}

(b) Scam at the off surface point p_{d_1}

(c) Scam at the off surface point $p_{d_0^\delta}$

(d) Scam at the off surface point $p_{d_0^\delta}$

Figure 10. Smooth surface with occlusions.

A. Local Minimums of C-D Curve

Fig. 10(a) shows an illustration of a smooth surface occluded by a surface with a similar color. We assume that the occluder is at depth d_1 but has color close to the surface point p_{d_0}, *i.e.*, the corresponding pixels are identified as in $\Omega^v_{p_{d_0}}$. The occluder is also assumed to have relatively larger color distances to the reference color compared to the neighbor surface points next to p_{d_0}. Because of the larger distances, this can result in $c_{p_{d_0}} > c_{p_{d_1}}$, where p_{d_1} is at the depth of the occluder as shown in Fig. 10(b).

This implies that the ground truth depth may not correspond to the global minimum on the C-D curve. Nevertheless, we show that they tend to correspond to the local minimum. We approximate the derivative of the cost function at d_0 with respect to d as:

$$\frac{\partial c_{p_d}}{\partial d}|_{d_0} = \lim_{\delta d \to 0} \frac{1}{\delta d} \cdot \\ \left(\frac{1}{|\Omega^v_{p_{d_0^\delta}}|} \sum_{(s,t)\in\Omega^v_{p_{d_0^\delta}}} \rho(A_{p_{d_0^\delta}}(s,t) - A_{p_{d_0^\delta}}(s_r,t_r)) \right. \\ \left. - \frac{1}{|\Omega^v_{p_{d_0}}|} \sum_{(s,t)\in\Omega^v_{p_{d_0}}} \rho(A_{p_{d_0}}(s,t) - A_{p_{d_0}}(s_r,t_r))\right) \tag{10}$$

where $p_{d_0^\delta}$ is at $d_0^\delta = d_0 + \delta d$ and $A_{p_{d_0}}(s_r, t_r) = A_{p_{d_0^\delta}}(s_r, t_r)$. Assume δd is small enough that $\Omega^v_{p_{d_0^\delta}} = \Omega^v_{p_{d_0}} = \Omega^{o*}_{p_{d_0}} \bigcup \Omega^{v*}_{p_{d_0}}$ and the misidentification remains the same as shown in the Fig. 10(c) and (d).

Because $\rho(\cdot)$ is insensitive to large distances, we have

$$\sum_{(s,t)\in\Omega^{o*}_{p_{d_0^\delta}}} \rho(A_{p_{d_0^\delta}}(s,t) - A_{p_{d_0^\delta}}(s_r,t_r)) \approx \\ \sum_{(s,t)\in\Omega^{o*}_{p_{d_0}}} \rho(A_{p_{d_0}}(s,t) - A_{p_{d_0}}(s_r,t_r)). \tag{11}$$

Therefore, Eq. 10 reduces to:

$$\frac{\partial c_{p_d}}{\partial d}|_{d_0} \approx \lim_{\delta d \to 0} \frac{1}{\delta d} \frac{1}{|\Omega^{v*}_{p_{d_0}}|} \cdot \\ \sum_{(s,t)\in\Omega^{v*}_{p_{d_0}}} \rho(A_{p_{d_0^\delta}}(s,t) - A_{p_{d_0^\delta}}(s_r,t_r)) \tag{12}$$

Since $\rho(\cdot) \geq 0$, the sign of the partial derivative $\frac{\partial c_{p_d}}{\partial d}|_{d_0}$ coincides with the sign of δd (*i.e.*, one side positive and the other side negative). This concludes that $\frac{\partial c_{p_d}}{\partial d}|_{d_0} = 0$ by the intermediate value theorem. Hence the ground truth depth correspond to the *local* minimum of our C-D curve. In the cases where Eq. 11 is not satisfied, the local minimum property may not hold. However it is rare.