Light Field Depth from Multi-scale Particle Filtering

Jie Chen∗†, Lap-Pui Chau∗† and He Li†

∗School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
†ST Engineering - NTU Corporate Lab, Nanyang Technological University, Singapore
‡Singapore Technologies Dynamics Pte Ltd
Email: Chen.Jie@ntu.edu.sg, ELPChau@ntu.edu.sg, Lihe@stengg.com

Abstract—Rich information could be extracted from the high dimensional light field (LF) data, and one of the most fundamental output is scene depth. State-of-the-art depth calculation methods produce noisy calculations especially over texture-less regions. Based on Super-pixel segmentation, we propose to incorporate multi-level disparity information into a Bayesian Particle Filtering framework. Each pixels’ individual as well as regional information are involved to give Maximum A Posteriori (MAP) predictions based on our proposed statistical model. The method can produce equivalent or better scene depth interpolation results than some of the state-of-the art methods, with possible potential in image processing applications such as scene alignment and stabilization.

Index Terms—Light Field, Bayesian, Particle Filter, Depth Interpolation

I. INTRODUCTION

The light field is a vector function that describes the amount of light propagating in every direction through every point in space [1]. Conventional cameras use a converging lens to record a 2D projection of the high dimensional LF data. This inevitably loses a lot of information, especially each light ray’s directional information. The light field cameras are designed to resolve the integration process and capture the extra directional information, which is useful in many applications such as refocusing, depth inference [2] [3], and 3D reconstruction [4] [5] etc.

The challenge of depth inference from light field is similar to that of traditional stereo vision, however instead of a pair of input images, the light field extends the disparity space to a continuous or multiple discrete ones. Bishop et al. [6] implemented an iterative algorithm to estimate the scene depth by iteratively searching and filtering among multiple aliased views for a best correspondence match.

As the dimension at the aperture plane are usually densely sampled, this enables the formation of an epipolar-plane image (EPI) based on which the correspondence problem can be greatly relieved, as the corresponding scene points will linearly line up as tilting straight lines in the EPI. By finding their slopes, the depth map can be calculated. Wanner et al. used a structure tensor based on local gradients to estimate the direction of lines on EPI [2]. The tensor is easy and highly efficient to implement. Kim et al. [4] proposed a scoring mechanism for all the hypothetical disparities for each scene point, and choose the highest scored disparity as their estimation. This method proves to be more accurate than the previous however slightly more time and memory consuming.

Either based on correspondence matching or EPI based slope calculation, the output depth map estimation is expected to be noisy due to several reasons such as the lack of matching clue for homogeneous regions, view-dependant effects such as occlusion or specularity. Some sort of filtering or optimization is always need for a refined output, see examples such as [7]–[9]. Notably, Wanner et al. [2] used an energy minimization approach combined with state-of-the-art convex optimization algorithms; they produced high quality depth maps, however at the cost of very high computational complexity. A “fine-to-coarse” approach was implemented in Kim’s recent work [4], where disparity estimations from lower resolution EPIs were used to interpolate the missing data in higher resolution EPIs due to lack of estimation confidence. Kim’s method produces state-of-the-art quality depth map outputs, with efficient implementation (in an order of 1 minute for a 2M pixel image). Che et al. used an adaptive guided filtering algorithm for depth interpolation [10], this method executes efficiently however it has problem in regions where confident predictions are sparse.

Based on Super-pixel segmentation, we propose to incorporate multi-level disparity information into a Bayesian Particle Filtering framework. Each pixels’ individual as well as regional information are involved to give Maximum A Posteriori (MAP) predictions based on our proposed statistical model. The method can produce equivalent or better scene depth interpolation results than some of the state-of-the art methods, with possible potential in image processing applications such as scene alignment and stabilization.

II. PROPOSED ALGORITHM

A discretized light field data usually comprises of an array of images that were virtually taken from a closely positioned multi-camera array which we denote as \( I_F(v) \), where \( v = 1, 2, ..., V \) denotes the view index, with total LF view number \( V \). The linear relationships of the LF viewing angles determines the linear configuration of disparity values between each views. For a \( 5 \times 5 \) discretized light field, the horizontal and vertical disparity ratios between each off-center views with respect to the center view position are shown in matrices \( M_H \) and \( M_V \) in Eqn. (1) respectively, where each matrix element
$M_H(v)$, $M_V(v)$ represents the relative position of a LF view.

\[
M_H = \begin{bmatrix}
-2 & -2 & -2 & -2 & -2 \\
-1 & -1 & -1 & -1 & -1 \\
0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 \\
2 & 2 & 2 & 2 & 2 \\
\end{bmatrix},
\]

\[
M_V = \begin{bmatrix}
-2 & -1 & 0 & 1 & 2 \\
-2 & -1 & 0 & 1 & 2 \\
-2 & -1 & 0 & 1 & 2 \\
-2 & -1 & 0 & 1 & 2 \\
-2 & -1 & 0 & 1 & 2 \\
\end{bmatrix}. \quad (1)
\]

The disparity value of a certain scene point can be expressed as $S_d$ [11]:

\[
S_d(dp, v) = dp \times (M_H(v) + M_V(v)), \quad v = 1, 2, \ldots, V. \quad (2)
\]

where $dp$ is the unit disparity between adjacent views, which is a parameter only related to the scene point’s distance to the camera once the camera’s settings are fixed.

According to Eqn. (2), if given a certain scene point $x_0$ with unit disparity value $dp'$, if we shift all the views $v$ respectively along the vector $S_d(dp', v)$, then the scene point $x_0$ will be perfectly aligned across all views if the Lambertion assumption is met., i.e.,

\[
\sum_{v=1}^{V} ||F(I_F(x_0, v), S_d(dp', v)) - L_F(x_0, v_c)||_2^2 = 0, \quad (3)
\]

where $F(i, j)$ is a translation and interpolation operator that translates $i$ along the vector $j$, $v_c$ is the index for central view. Bicubic interpolation will be used when translation falls on sub-pixel positions.

Based on Eqn. (3), a certain scene point’s disparity value could be estimated. For a sequence of disparity values that cover the desirable resolution and range of the LF: $DP = [dp_1, dp_2, \ldots, dp_K]$, the following matching error could be calculated for each element in $DP$:

\[
E_1(x_0, dp) = \sum_{v=1}^{V} ||F(I_F(x_0, v), S_d(dp, v)) - I_F(x_0, v_c)||_2^2. \quad (4)
\]

The value $E_1(x_0, dp)$ could be understood as the amount of “mismatch” at $x_0$ for a certain disparity $dp$. The subscript $1$ denotes that this value is calculated on a single pixel basis. The disparity for pixel $x_0$ could be estimated as the one $dp$ that produces the smallest matching error:

\[
D_1(x_0) = \arg \min_{dp} E_1(x_0, dp) \quad (5)
\]

Fig. 2(a) gives a visual demonstration of disparity estimations based on Eqn. (5). From the image, we can see disparity estimations are more accurate in regions with sharp edges or clear textures; in regions with small intensity variations, the estimation tend to be noisy and incorrect. We propose to alleviate such problem by considering a larger region instead of a single pixel in Eqn. (5). A region can provide more information for matching and produce less noisy outputs.

This consideration leads us to the method of superpixel (SP) segmentation.

A. Multiscale Superpixel Segmentation and Multiscale Disparity Estimation

The concept of “superpixel” has been widely used in various computation vision applications such as image segmentation [12], object tracking [13], etc. It groups pixels into perceptually meaningful atomic regions to replace the rigid structure of the pixel grid. We implement the SLIC superpixel segmentation algorithm [14] in our work. SLIC is an iterative regional pixel clustering algorithm that cluster each pixel to an initiated center grid according to the their respective normalized spatial and color distances.

In this paper we propose to build a hierarchy of segmentation with different superpixel (SP) sizes. Fig. 1(b) to (f) gives an illustration of such multiscale segmentation on the LF image "truck" in Fig. 1(a).

Based on the multiscale segmentation, a superpixel’s mismatch error can be calculated as:

\[
E_l(x_0, dp) = \frac{1}{N_{l,k}} \sum_{x_0, \hat{x} \in \Omega_{l,k}} E_1(\hat{x}, dp). \quad (6)
\]

where $E_l$ is the average matching error for all pixels that belong to the superpixel $\Omega_{l,k}$ at a certain given disparity $dp$. $l$ and $k$ denotes the superpixel scale and index respectively.

The disparity value that produces the smallest average matching errors for all the pixels, will be assigned as the superpixel $\Omega_{l,k}$’s disparity:

\[
D_l(x_0) = \arg \min_{dp} E_l(x_0, dp), \quad (7)
\]

Fig. 2 from (b) to (f) gives an illustration of the calculated disparity values for each SP in different scales for the LF image "truck" in Fig. 1(a).

Disparity calculated from SP segmentation generalizes all mismatch errors in the region. This greatly benefits smooth and non-textual areas that would tend to produce false estimation on a single pixel basis.
and contours of target objects becomes irregular. However, the edges superpixel), The prediction becomes more stable especially from (b) to (f), as the scales goes higher (more pixels per as compared to individual pixels. As can be seen in Fig. 7 better disparity prediction over smooth and non-textual regions

dimension

\(l\) is set to 10 in our experiments. Pixel numbers of different SP scales are set as \([1, 20, 60, 100, 150, 200, 300, 400, 800, 1000]\), respectively.

1) The Scale Observation Likelihood: In this subsection, we aim to design a statistic model for the scale observation likelihood \(p(y|l)\), i.e., given a pixels observation \(y\) (e.g., pixel intensity, neighbourhood variation, and disparity predictions from all superpixel scales), we aim to calculate the possibility that the disparity calculated from superpixel scale \(l\) is "most correct".

The mismatch error for each SP in Eqn. 6 could be used as the metric for how likely the given scale \(l\) is the correct one. In order to preserve the relative ratios of each error and convert them into normalized distributions, the following calculation will be implemented:

\[
p(y|l) = \mathcal{N}_l[(\max_{i} E_i + \min_{i} E_i) - E_l] \tag{9}
\]

where \(\mathcal{N}_l\) denotes the normalization operation along the dimension \(l\) for each given observation. Fig.3 gives an illustration of 50 sample data with their respective mismatch error in (a) and normalized observation likelihood in (b).

2) The Texture Intensity Prior \(p_e(l)\): Superpixels give a better disparity prediction over smooth and non-textual regions as compared to individual pixels. As can be seen in Fig. 7 from (b) to (f), as the scales goes higher (more pixels per superpixel), The prediction becomes more stable especially for large and smooth background regions. However, the edges and contours of target objects becomes irregular.

We propose a Bayesian multiscale framework for the prediction of LF scene points’ disparity value. The framework is based on the disparity values calculated from different SP scales. These pre-calculated atoms could be understood as basic units in a typical particle filter framework \[15\], where the output is usually an linear combination of the particles.

\[1\] For pixels in smooth and low contextual intensity variations, larger SP scales gives a much better disparity

\[2\] The Texture Intensity Prior

\[3\] The Texture Intensity Prior

\[4\] For pixels in smooth and low contextual intensity variations, larger SP scales gives a much better disparity

\[5\] For pixels in smooth and low contextual intensity variations, larger SP scales gives a much better disparity
prediction.

2) As the contextual variation goes lower, smaller SPs will be more and more likely to give the correct prediction.

3) The Cross-Scale Prior $p_s(l)$: On top of the contextual gradient intensity prior, a cross-scale prior is also implemented. It can be observed from Fig. 2 that larger SP scales give a more stable disparity prediction and smaller SP scales are much more noisy. The following prior is devised to regulate the variation of smaller SP scales:

$$p_s(x_l) = \frac{1}{\sum_l} e^{-\frac{(D_l - D_{l+1})^2}{\sigma^2}}$$  \hspace{1cm} (11)

An illustration of the prior function is shown in Fig. 6. As can be seen, the prior will encourage the choice of SP levels when their neighboring larger scale produces similar disparity prediction. The tolerance (variance) gets larger for pixels with larger gradient intensities, as sudden disparity changes tend to happen over the pixels that show intense edges.

III. EXPERIMENTS

A. Evaluation of the Statistic Model

We evaluate the imposed scale prior in this subsection. We reconstruct the benchmark LF data Mona and Truck with and without the cross-scale prior, and their respective reconstructions are shown in Fig. 7. As can be seen from the figure, without the proposed prior, LF regions with large non-textual regions tend to give wrong predictions, and very noisy predictions also appear from smaller SP scales when they are involved into the final output.

B. Quantitative Evaluation of the Proposed Algorithm

We compare disparity outputs from our proposed Bayesian inference framework with some of the state-of-the-art methods on the benchmark LF images Buddha, Maria, Medieval, and Monas given ground truth disparity. Data from competing methods are imported from [2]. It can be seen from Table I, the proposed method can produce equal or better disparity reconstruction than some of the state-of-the-art methods.
TABLE I: Detailed evaluation of disparity reconstruction errors. Data from competing methods are imported from [2]. The values in the table show the mean squared error in pixels times 100, i.e. a value of 0.81 means that the mean squared error in pixels is 0.0081.

<table>
<thead>
<tr>
<th>Lightfield</th>
<th>EPI_L</th>
<th>EPI_S</th>
<th>EPI_C</th>
<th>EPI_G</th>
<th>ST_AV_L</th>
<th>ST_AV_S</th>
<th>ST_AV_C</th>
<th>ST_AV_G</th>
<th>ST_CH_L</th>
<th>ST_CH_S</th>
<th>ST_CH_G</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buddha</td>
<td>0.81</td>
<td>0.57</td>
<td>0.55</td>
<td>0.62</td>
<td>1.20</td>
<td>0.78</td>
<td>0.90</td>
<td>1.01</td>
<td>0.67</td>
<td>0.80</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>Maria</td>
<td>0.19</td>
<td>0.10</td>
<td>0.10</td>
<td>0.11</td>
<td>0.34</td>
<td>0.11</td>
<td>0.11</td>
<td>0.51</td>
<td>0.11</td>
<td>0.11</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>Medieval</td>
<td>1.69</td>
<td>1.15</td>
<td>1.10</td>
<td>1.24</td>
<td>7.22</td>
<td>0.91</td>
<td>0.76</td>
<td>12.14</td>
<td>1.08</td>
<td>0.79</td>
<td>1.30</td>
<td></td>
</tr>
<tr>
<td>Monas</td>
<td>1.15</td>
<td>0.90</td>
<td>0.82</td>
<td>0.93</td>
<td>2.25</td>
<td>1.05</td>
<td>0.79</td>
<td>2.28</td>
<td>1.02</td>
<td>0.81</td>
<td>1.01</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 8: Visual Comparison of disparity reconstruction for Mona. (a) is the output from (EPI_L), and (b) is from ours.

C. Qualitative Evaluation of the Proposed Algorithm

We compare the disparity reconstruction calculated in our proposed method with state-of-the-art methods (“EPI_L” and “EPI_C” in Table I) on the benchmark LF images Mona, Watch, Maria, and Medieval, and the details are shown in Fig. 8, 9, 10, and 11, respectively. As can be generalized, our proposed method can produce more stable and correct predictions especially in large and texture-less regions. For other textured parts, our method can produce equivalent or better outputs.

IV. CONCLUSION

In this work, we focus on the calculation and filtering of scene depth based on light field data. State-of-the-art methods produce noisy calculations especially over texture-less regions. Based on Super-pixel segmentation, we propose to incorporate multi-level disparity information into a Bayesian Particle Filtering framework. Each pixels’ individual as well as regional information are involved to give Maximum A Posteriori (MAP) predictions based on our proposed statistical model. The method can produce equal or better scene depth interpolation results than some of the state-of-the-art methods, with possible potential in image processing applications such as scene alignment and stabilization.

ACKNOWLEDGMENT

The research was partially supported by the ST Engineering-NTU Corporate Lab through the NRF corporate lab@university scheme.

REFERENCES

Fig. 11: Visual Comparison of disparity reconstruction for Medieval. (a) is the output from (EPI L), and (b) is from ours.


