ACCURATE 3D RECONSTRUCTION FROM CIRCULAR LIGHT FIELD USING CNN-LSTM

Zhengxi Song*, Hao Zhu*, Qi Wu*, Xue Wang*, Hongdong Li† and Qing Wang*

* School of Computer Science, Northwestern Polytechnical University, Xi’an 710072, China 
{songzx,xwang,qwang}@nwpu.edu.cn; zh920307@outlook.com; dongdawuqi@mail.nwpu.edu.cn
† Australian National University, ACT 0200, Australia. {hongdong.li@anu.edu.au}

Abstract

A light field is formed by densely capturing images on a regular sub-aperture grid. Geometry information endowed in the epipolar plane images(EPI) can only lead to a 2.5D reconstruction. In order to obtain a full 360° view of an object, we focus on light fields captured by a circularly moving camera, resulting in circular light fields (or Cir-LFs in short). Compared with traditional EPIs, Circular EPIs(CEPIs) provide unique advantages, such as that corresponding points forming a 3D sinusoid like curve instead of a 2D straight line and geometry information encoded sequentially in multiple adjacent views along the curve. However, current reconstruction methods only focus on the 2D projection of 3D curve, leading to distortions in the reconstructed upper and lower surfaces. We propose to analyze 3D features contained in the 3D CEPI volume and we develop a deep CNN-LSTM network to model the gradient map in the CEPI volume. Additionally, a large scale Cir-LF dataset is constructed for research purpose. Experiments on both synthetic and real scenes demonstrate the effectiveness and generaliability of the proposed method.

Index Terms— Light field, 3D reconstruction, LSTM, Convolutional Neural Networks, Gradients distribution

1. INTRODUCTION

Accurate 3D reconstruction from 2D images is a classic topic in computer vision and graphics communities. The sparsity of image features and the deviation of feature matching hinder the performance of image-based 3D reconstruction[1, 2]. Recently, the light field imaging technology has become increasingly popular in the computational photograph community. Light fields exhibit several unique features for 3D reconstruction, e.g., regular sampling pattern, dense angularly sampling density, and sub-pixel disparity. The motion parallax of corresponding points form a specific trajectory in an image volume, is a constraint by relative motion between a camera and objects/scenes. The Cir-LF, imaging object from 360° with an equal space, brings new insights on the human visual perception.

The Cir-LF forms an image volume (Fig.1, left), which is accumulated with different views, as angular $\theta$. The CEPI is the 2D slice of the volume where $y$ is fixed, and $x$ and $\theta$ vary. Features for 3D reconstruction in traditional light fields are maintaining in Cir-LFs, however, they exhibit encouraging new characteristics. Firstly, corresponding points form a sinusoid like curve, and multiple curves are twining together. Then apart from the eye level CEPI slice, the continuous structures span across multi-slices due to the perspective effect. Single CEPI slice (i.e. green or purple rectangle in Fig.1) suffers from incomplete structures, especially when it is near to lower and upper surfaces. Finally, the gradients of these curves reflect the geometry information. Early methods use a threshold to evaluate the color consistency[3] of curves. Recent methods focus on extracting geometry features, like gradients[4] and Canny edges[5] of each single CEPI slice, which induces distortion errors in upper and lower surface.

To overcome above problems, we introduce a CNN-LSTM neural network to analyze the geometry feature in the Cir-LF. To eliminate ambiguity in the 2D slice, a 3D EPI volume is stacked and fed to the network. Through high-order convolution within a deep learning architecture, we can estimate reliable gradient maps at multiple scales in spatial or...
angular dimensions, or both. Such an approach allows the model to learn representations with geometry information by fully exploiting the 3D EPI volume, enhancing the confidence of gradient map estimation. Finally, these reliable gradient maps can be merged with Cir-LF parameters to achieve an accurate 3D reconstruction. The main contributions of the paper includes, 1) we analyze the 3D feature of EPI curves and the series nature; 2) we design the first learning-based framework for Cir-LF 3D reconstruction; 3) we construct the first large-scale Cir-LF dataset.

2. RELATED WORK

Under two-parallel-planes (TPP) parameterization, a light field is modelled as a 4D function to represent rays in 3D space. Owing to dense sampling, there is a corresponding continuous EPI line in light field for each 3D point. The EPI based analysis has achieved significant advantages such as high efficiency and sub-pixel disparity accuracy, and becomes more and more popular. Deep learning based methods further improves the performance of the EPI analysis by introducing the 2D regularization multi-orientation EPIs. However, TPP light field can only achieve a 2.5D reconstruction.

Feldmann et al.[3] extended the camera motion from a line to a circle, and proposed a depth corrected EPI analysis called Image Cube Trajectory (ICT) analysis. By using color consistency detection, this work tried to extract trajectory of CEPI curves. Yucer et al.[4] analyzed the gradients from a densely sampled 2D circular EPI on an extremely high spatio-angular sampling (10 images per degree). Cserkaszky et al.[9] used a specified ray structure to fit the distorted curves on the CEPI into a standard sinusoid. Vianello et al.[5] focused on the curve instead of the pixels and solve the curve function in the Hough transform space. These two methods have achieved amazing improvements over previous multi-view stereo methods[1] and have opened up novel strategies for light field based 3D reconstruction. However, they have an unavoidable problem. Because the depth changes during the circular light field capture, the trajectory of each point will run through multiple lines, i.e., a 3D trajectory (see Sec.2). Current methods focus on a 2D circular EPI analysis, which is only the 2D projection of a 3D trajectory, thus the ambiguity is introduced. Additionally, when the number of views decreases or the points are close to the camera, the continuous curve in the circular EPI becomes discrete dot lines where previous methods may be failed.

Recently, CNN and LSTM have been placed in one unified framework. CNNs can produce a rich representation of the input image on multi-scale, while LSTMs can modal sequential data.[10]. The Cir-LF shows strong series and periodicity characteristics in the 3D image volume. We are inspired by characteristic of LSTM-CNN networks, and adapt it to model geometry features in 3D CEPI volume. Rather than analyzing single CEPI[5, 4], reliable gradient maps are extracted with multi-scale on both spatial and angular dimensions. Owing to the redundancy in Cir-LFs, extensive experiments demonstrate that the reliable gradient maps are sufficient to a dense and accurate 3D structure estimation.

3. 3D SERIES TRAJECTORY

Suppose that the camera rotates around an axis, a point \( P = (X, Y, Z) \) in 3D space can be expressed by the polar coordinate \((R, \phi, Y)\), where \( R \) is the distance between the \( P \) and the axis and \( \phi \) is the phase offset (see Fig.2(a)). The imaging point in a 3D Cir-LF \( LF(\theta, x, y) \) forms a 3D trajectory which can be described as[5],

\[
\begin{align}
  x(\theta) &= f \cdot \frac{R \sin(\theta + \phi)}{R_{\text{mn}} - R \cos(\theta + \phi)} + x_c \quad (1a) \\
  y(\theta) &= f \cdot \frac{Y}{R_{\text{mn}} - R \cos(\theta + \phi)} + y_c \quad (1b)
\end{align}
\]

where \( f \) is the focal length and \((x_c, y_c)\) is the principle point of the camera. \( R_{\text{mn}} \) is the distance between the camera center.
and the rotational axis (see Fig.2). The series nature is depending on the rotation phase θ so that both the trajectories and their gradients should obey such series nature.

The correspondences between sequential images can be expressed with gradient of the 3D trajectory, and they are promoted to recover the depth. With a given pixel \( p = (\theta_0, x_0) \) in EPI and its 3D position \( (R, \phi) \), the position \( x_1 \) of the corresponding pixel \( p_1 \) in any other view \( \theta_1 \) can be predicted as,

\[
x_1(\theta_1) = \int_{\theta_0}^{\theta_1} x'(\theta)d\theta + x_0(\theta_0),
\]

Different from the TPP light field, the trajectory in Circular Light Field will span across multiple EPIs (Eqn.1b) since the depth changes (Fig.3(a)). Fig.3(b) illustrates several projected trajectories in the y-dimension under different \( Y \). It is noticed that the amplitude increases as the point moves farther away from the center line.

To better understand the phenomenon, the differential function \( y'(\theta) \) in a synthetic circular light field is shown in Fig.4. It can be found that the EPI curve farther away from the central line has a larger \( |y'(\theta)| \). Given a pixel \( p = (\theta_0, x, y_0) \), a larger \( |y'(\theta)| \) indicates that the trajectory will be cut off in the next view \( \theta_0 + 1 \) of the current EPI \( L(\theta, x, y_0) \) and it will appear in another circular EPI \( L(\dot{\theta}, x, y + y'(\theta)) \). In Fig.4(c), two examples of \( y'(\theta) \) in different \( y \) are demonstrated. Because the red line is further away from the central line than the blue one, \( y'(\theta) \) in the red EPI is larger than that of the blue EPI. Additionally, the trajectory in the red EPI is not continuous due to the larger \( y'(\theta) \). To take the influence of \( y'(\theta) \) into the \( x \)-trajectory analysis, our core concern is to extract reliable gradients in the 3D image volume.

4. CNN-LSTM NETWORK FOR CEPI VOLUME

4.1. Network Architecture

With the complete 3D trajectory function, the depth \( Z \) can be computed from the differential \( x'(\theta) \). As a result, the proposed network focuses on learning the differential \( x'(\theta) \) for all pixels. Fig.5 shows the proposed network architecture, with a 3D EPI volume stacked along the \( y \)-axis as the input. The EPI volume is firstly analyzed by a U-shaped CNN-LSTM network to extract the \( x'(\theta) \) of the central EPI. Then the \( x'(\theta) \) and the original EPI volume are concatenated, which are fed to two dense block layers to distinguish the foreground from the background.

The CNN-LSTM network has four levels, where each of these analyzes the EPI at different resolutions. In the top three levels, four convolutional layers are firstly applied to encode the local information. Then two convolutional LSTM layers are cascaded to extract the series features following top-down and left-right directions sequentially. Noting that, given a \( h \times w \times c \) volume \( (h, w \) and \( c \) refer to the height, width and channels, respectively), it is separated as \( h \times 1 \times c \) with length \( w \) and \( 1 \times w \times c \) with length \( h \) series in top-down and left-right LSTM layers, respectively. The kernel sizes of the convolutional LSTM layers are \( 3 \times 1 \) and \( 1 \times 3 \) in left-right and top-down directions respectively. The channel size is 100. (Fig.5). Later, the outputs from top-down LSTM layers are concatenated with the up-convolutional results from a higher level, and be decoded by another four convolutional layers. There are six convolutional layers in the highest 4-th level. Neighbouring levels of the CNN-LSTM are connected by down and up-convolutional layers. The kernel size of all convolutional layers is \( 3 \times 3 \). The channel size for the i-th layer is \( \min(60 \times i, 180) \), \( i = 1, 2, 3, 4 \). After the CNN-LSTM analysis, the \( x'(\theta) \) of the central EPI is extracted. Finally, two standard dense blocks are applied to segment the EPI from foreground and background. Each dense block contains 4 layers.

4.2. Loss Function

Let \( g \) be the ground truth gradients, \( \hat{s} \) and \( s \) is the output and ground truth segmentation, respectively. The loss function is defined as follow:

\[
\mathcal{L} = \mathcal{L}_g(g) + \lambda_1 \mathcal{L}_s(g) + \lambda_2 \mathcal{L}_s(\hat{s}),
\]

where \( \mathcal{L}_g(g) \) and \( \mathcal{L}_s(g) \) are the \( \ell_1 \) loss of gradient predictions under the output foreground mask and the ground truth one, respectively. \( \mathcal{L}_s(\hat{s}) \) is the cross entropy loss between the output segmentation and the ground truth.
Fig. 5. The architecture of the proposed network.

At the training stage, $\lambda_1$ is set as 1 and $\lambda_2 = 3 \times 0.8^{[j/10]}$ changes with the epoch $j$. We use the Adam optimizer. The learning rate is $1e - 4$ initially and decreases 0.99 in each step. All convolutional kernels and bias are initialized using the Xavier method.

4.3. Dataset

To train and evaluate the proposed algorithm, we have rendered 150 Cir-LFs using the POV-Ray, 100 for training and 50 for testing. Each Cir-LF contains 180 views with the resolution of $400 \times 400$. We also provide camera parameters and the depth for each view image. As far as we know, this is the first large scale Cir-LF dataset in the community. The proposed dataset contains different challenging environments such as occlusions, shadows, complex illuminations, reflections and structures with fine details.

5. EXPERIMENTAL RESULTS

To evaluate the effectiveness of our method, experiments on both synthetic and real datasets were performed. In synthetic scenes, we first conduct two experiments to demonstrate the performance of the 3D EPI volume input and the LSTM layers in Fig.5. The results of the proposed algorithm are a set of gradient images for each vertical pixel level. Then they are converted to depth images with each rotation angle $\theta$ of gradient images for each vertical pixel level. Then they are finally merged into 3D point clouds. We also compare our method with two publicly available image-based reconstruction algorithms. The first one is the patch based method Clustering Views for Multi-view Stereo (CMVS) [1]. The second one is the Multi-Views Estimation (MVE) [2].

5.1. Synthetic Scenes

In order to evaluate the effectiveness, we first compared our proposal (3D w. LSTM) with two different network setups, then with both CMVS and MVE. We list quantitative comparisons in Tab.1 with the RMS errors between the ground truth 3D point clouds and the reconstructed point clouds.

5.1.1. 2D EPI vs 3D EPI Volume

Tab.1 also shows the RMS errors of point clouds from the network with 2D and 3D EPI inputs. The network with a 3D EPI input shows great advantages compared with the one with 2D EPI input. This is due to the fact that there is a unavoidable ambiguity in image boundary pixels if only the 2D EPI input is used, as analyzed in the Sec.3.

Table 1. The RMS errors of point clouds of our networks with different configurations compared with two state-of-the-art methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Tellurion</th>
<th>Plant</th>
<th>Hydrant</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMVS[1]</td>
<td>2.266</td>
<td>1.1534</td>
<td>1.230</td>
</tr>
<tr>
<td>MVE[2]</td>
<td>1.251</td>
<td>1.0085</td>
<td>0.525</td>
</tr>
<tr>
<td>Ours (2D w. LSTM)</td>
<td>0.642</td>
<td>0.396</td>
<td>0.415</td>
</tr>
<tr>
<td>Ours (3D w/o. LSTM)</td>
<td>0.655</td>
<td>0.353</td>
<td>0.375</td>
</tr>
<tr>
<td>Ours (3D w. LSTM)</td>
<td><strong>0.592</strong></td>
<td><strong>0.332</strong></td>
<td><strong>0.373</strong></td>
</tr>
</tbody>
</table>

*The terms ‘2D’ and ‘3D’ refer to the 2D and 3D EPI volume inputs respectively.

Fig.6 shows the quantitative and qualitative comparisons between the ground truth gradient map $\hat{z}'(\theta)$ and the output $\hat{z}'(\theta)$. It is noticed that, 2D EPI based analysis achieves similar performance as the results from 3D EPI volume in the center lines (yellow lines in Fig.6). However, as the line goes farther away from the center (red lines in Fig.6), the trajectory becomes less complete and the ambiguity appears such that the performance of 2D EPI based analysis starts to degrade, while the 3D EPI volume based analysis can still provide consistent performance for both the center lines and the boundary lines.
Fig. 7. Qualitative comparisons in gradient maps between the ground truth and the results from networks with and without LSTM layers, respectively.

5.1.2. LSTM Layers

Tab.1 shows the RMSEs between the ground truth 3D point cloud and the reconstructed point clouds from the network with and without LSTM layers. The network with LSTM layers can better capture the constrained features of the distribution of the gradient map $x'(\theta)$ in Sec.3.

Fig.7 gives a good demonstration of this phenomenon. The LSTM layers can model the large gradient areas well. In the green boxes of Fig.7, since $|x'(\theta)|$ is larger than 1 pixel, the trajectory is not continuous and becomes discrete dots. As analyzed in Sec.3, the gradient distribution of a circular EPI is ordered. It can be found that, the reconstructed gradients are over smoothed by other contents in the one without LSTM layers, while the one with LSTM layers can still provide sharp results.

5.1.3. Comparison with State-of-the-arts

In testing synthetic data, visual comparison of the obtained meshes are shown in Fig.8, from left to right are the TellurionData, PlantData and HydrantData, respectively. From top to bottom are input the view, the Groundtruth, the CMVS, the MVE, results from our methods, 2D w. LSTM, 3D w/o. LSTM and 3D w. LSTM, respectively. This scene exhibits similar texture and complex topology (e.g., tree or table). Our method yields the best meshes, which are less noisy than meshes by the MVE and have more details than meshes by the CMVS.

5.2. Real scenes

For the real scene data, we used a NIKON D700 camera with a resolution of $4256 \times 2832$ pixel, a pixel pitch of $\sigma = 8.4 \mu m$ and a focal length $f = 48 mm$ to capture Cir-LFs, as Fig.9. Objects are placed on a high precision rotation stage, and light fields are composed of $N = 180$ images as required. Calibration was performed to remove distortion and determine the correct rotation center. In Teemo and Transformer dataset, our proposal can achieve an accurate and dense result without mesh, as Fig.10 shows. The MVE is better than the CMVS with the increasing number of images acquired, however, it shows higher level of noise in the surface. The stone data is more challenging due to lots of similar details and homogeneous regions. Both the MVE and the CMVS basically fail to recover the object, while our proposal achieves a smooth and robust reconstruction.
6. CONCLUDING REMARKS AND FUTURE WORK

In this paper, we revisit the 3D reconstruction problem under a Cir-LF model, and enable a high fidelity reconstruction on traditionally challenging scenes, such as unfolliaged trees. By raising 3D series trajectory features, we can obtain more convincing correspondences in the 3D image volume which can be describe as gradient maps. Reliable gradient maps are then learned in the 3D EPI volume with CNN-LSTM Neural Networks. Extensive experiments have carried on synthetic and real scene Cir-LFs and the results show that our network can effectively model features in the 3D volume. Finally, a large-scale Cir-LF dataset is constructed for training and testing. Both synthetic and real experiments further verify the robustness and accuracy of our proposal.

Our method is based on the standard circular motion with equal space, which highly relies on the rectified real data. In the future, we will extend our analysis to other camera configurations, such as the hand-held situation. To analyze image trajectories in a unstructured sampling configure, the application will be broadened greatly.

7. REFERENCES