Exploring the New Designs of 3D Sensors

Qifeng Chen
Assistant Professors
CSE & ECE
3D Sensors

• The real world is 3D
• 3D sensors play a role in autonomous driving and visual perception
  • Reliable distance information
  • Infer semantic information
Time-of-flight camera
Stereo cameras
LiDAR
Kinect
Depth Sensing Beyond LiDAR Range

Kai Zhang, Jiaxin Xie, Noah Snavely, Qifeng Chen
Depth Sensing Beyond LiDAR Range

Figure 1: Visualization of existing depth-sensing solutions’ maximum range.

A missing piece in long-range depth perception
Motivation

<table>
<thead>
<tr>
<th>Self-driving datasets</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Kitti</td>
<td>80 meters</td>
</tr>
<tr>
<td>Waymo</td>
<td>80 meters</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

Question: can we achieve dense depth sensing beyond LiDAR range with low-cost cameras?

Example application:
 Autonomus trucks driving on highway

60 mph = 96 km/h = 27 m/s
80 meters roughly means 3 seconds

Image sources: velodyne lidar
Problem Setup

**Basic idea:** use two **cameras with telephoto lens** to capture a stereo pair, then reconstruct a dense depth map.

Nikon P1000  
Canon SX70  
Industrial cameras\(^1\)

\(^1\) Industrial cameras are usually much cheaper than consumer ones.
Problem Setup

Important camera setup constraint:

Baseline is restricted to ~2 meters because of typical vehicle size.

What does this mean?

Depth estimation is very sensitive to pose error, especially rotation error.

It’s difficult for hardwares to achieve and maintain this precision.

![Diagram]

Triangulation angle $\theta \approx \frac{b}{z} \approx \frac{2m}{300m} \approx 0.382^\circ$

Estimated depth $\approx z \cdot (1 - \frac{\Delta \phi}{\theta})$

$\Delta \phi$ : rotation error

Relative error in estimated depth
Tentative Solution - SfM

Bas-relief ambiguity in SfM[1]

Big focal length $\rightarrow$ Near-orthographic camera (Weak perspectivity)

Figure 2: Ground-truth (blue) and the reconstructed (red) scene points. The unit for $x$, $y$, $z$ axes is meter.  

Figure 3: Top-down view of ground-truth relative pose (solid) and the recovered one (dashed). $\theta$ is exaggerated for illustration.

Our Approach

- **Pseudo-Rectification**
- **Disparity Estimation**
- **Ambiguity Removal**
- **Offset**

**Raw left view**

**Raw right view**

**Raw back view**

**Pseudo-rectified left view**

**Pseudo-rectified right view**

**Pseudo-rectified back view**

**Ambiguous disparity map**

Estimated ambiguous disparity
Results on synthetic data

<table>
<thead>
<tr>
<th>Method</th>
<th>Failure</th>
<th>&lt;1%</th>
<th>&lt;2%</th>
<th>&lt;3%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>0</td>
<td>45.3%</td>
<td>80.1%</td>
<td>96.9%</td>
</tr>
<tr>
<td>Loop and Zhang [18]</td>
<td>0</td>
<td>1.14%</td>
<td>2.73%</td>
<td>5.99%</td>
</tr>
<tr>
<td>SfM+MVS [19, 20]</td>
<td>15</td>
<td>6.71%</td>
<td>12.7%</td>
<td>19.1%</td>
</tr>
</tbody>
</table>

Table 1: Quantitative results on 40 synthetic scenes for methods in Fig. 7. “Failure” means the number of scenes for which a method fails to output a depth map. The metric is the portion of pixels with relative depth error below certain threshold, i.e., 1%, 2%, 3%, averaged over the successful scenes.

Figure 7: Comparison among different algorithms. For rectification-based methods, the ground-truth depth map has been warped to align with the rectified view. For SfM, we have used the full ground-truth intrinsic matrix.
Results on real-world data

- Ground-truth depth is acquired by the laser rangefinder: only point-wise measurement
- Ground-truth: 302m  Estimated: 300.8m
MFuseNet: Robust Depth Estimation with Learned Multiscopocoic Fusion (ICRA 2020)
Multiscopic Images
Fig. 2: Five images captured using our multiscopic perception system from different viewpoints. The parallax between the center view and any adjacent view is the same.
Fig. 6: The network structure of MFuseNet. For $n$ cost volumes with size $D \times H \times W$, they are processed respectively and then fused to get the final disparity. The feature channels of 3D CNN is 4 such that the size of each cost volume before concatenation is $4 \times D \times H \times W$. 
Fig. 7: Color images and ground-truth disparity maps in the synthetic multiscopic dataset, and the disparity maps obtained by MFuseNet.
Fig. 9: The disparity estimation results of different algorithms for two sets of images, Aloe and Lampshade in the Middlebury 2006 stereo dataset. The first image is the reference RGB image, i.e., the left image for stereo algorithms and the center image for multiscopic algorithms. Two images are used for stereo algorithms, and three images are used for multiscopic algorithms.
Fig. 11: The disparity estimation results of different algorithms for a reflective workpiece.