Are we ready for Autonomous Driving?
The KITTI Vision Benchmark Suite

Philip Lenz¹  Andreas Geiger²  Christoph Stiller¹  Raquel Urtasun³

¹Karlsruhe Institute of Technology  ²Max-Planck-Institute IS  ³Toyota Technological Institute at Chicago

Philip Lenz (KIT)  Andreas Geiger (MPI-IS)  Christoph Stiller (KIT)  Raquel Urtasun (TTI-C)
The Challenge of Autonomous Cars

State of the art
- Localization, path planning, obstacle avoidance
- Heavy use of 3D laser scanner and detailed maps

Problems for computer vision
- Stereo, optical flow, localization
- Object detection, recognition and tracking
- Semantic segmentation, 3D scene understanding
The Challenge of Autonomous Cars

State of the art
- Localization, path planning, obstacle avoidance
- Heavy use of 3D laser scanner and detailed maps

Problems for computer vision
- Stereo, optical flow, localization
- Object detection, recognition and tracking
- Semantic segmentation, 3D scene understanding

P. Lenz: The KITTI Vision Benchmark Suite – www.mrt.kit.edu
The Challenge of Autonomous Cars

State of the art

- Localization, path planning, obstacle avoidance
- Heavy use of 3D laser scanner and detailed maps

Problems for computer vision

- Stereo, optical flow, localization
- Object detection, recognition and tracking
- Semantic segmentation, 3D scene understanding
The Challenge of Autonomous Cars

State of the art

- Localization, path planning, obstacle avoidance
- Heavy use of 3D laser scanner and detailed maps

Problems for computer vision

- Stereo, optical flow, localization
- Object detection, recognition and tracking
- Semantic segmentation, 3D scene understanding
The Challenge of Autonomous Cars

State of the art

- Localization, path planning, obstacle avoidance
- Heavy use of 3D laser scanner and detailed maps

Problems for computer vision

- Stereo, optical flow, localization
- Object detection, recognition and tracking
- Semantic segmentation, 3D scene understanding
The Challenge of Autonomous Cars

State of the art
- Localization, path planning, obstacle avoidance
- Heavy use of 3D laser scanner and detailed maps

Problems for computer vision
- Stereo, optical flow, localization
- Object detection, recognition and tracking
- Semantic segmentation, 3D scene understanding
The Challenge of Autonomous Cars

State of the art

- Localization, path planning, obstacle avoidance
- Heavy use of 3D laser scanner and detailed maps

Problems for computer vision

- Stereo, optical flow, localization
- Object detection, recognition and tracking
- Semantic segmentation, 3D scene understanding
The Challenge of Autonomous Cars

State of the art
- Localization, path planning, obstacle avoidance
- Heavy use of 3D laser scanner and detailed maps

Problems for computer vision
- Stereo, optical flow, localization
- Object detection, recognition and tracking
- Semantic segmentation, 3D scene understanding
The Challenge of Autonomous Cars

State of the art
- Localization, path planning, obstacle avoidance
- Heavy use of 3D laser scanner and detailed maps

Problems for computer vision
- Stereo, optical flow, localization
- Object detection, recognition and tracking
- Semantic segmentation, 3D scene understanding
The KITTI Vision Benchmark Suite

- Two stereo rigs (1392 × 512 px, 54 cm base, 90° opening)
- Velodyne laser scanner, GPS+IMU localization
- 6 hours of recordings, 10 frames per second
The KITTI Vision Benchmark Suite

- Two stereo rigs (1392 × 512 px, 54 cm base, 90° opening)
- Velodyne laser scanner, GPS+IMU localization
- 6 hours of recordings, 10 frames per second
The KITTI Vision Benchmark Suite

- Two stereo rigs (1392 × 512 px, 54 cm base, 90° opening)
- Velodyne laser scanner, GPS+IMU localization
- 6 hours of recordings, 10 frames per second
Sensor Calibration Challenges

- Camera ↔ camera calibration
- Velodyne ↔ camera registration
- GPS ↔ Velodyne registration

Geiger et al., ICRA 2012
ICP + Hand-eye calibration

P. Lenz: The KITTI Vision Benchmark Suite – www.mrt.kit.edu
Sensor Calibration Challenges

360° Velodyne Laserscanner

Stereo Camera Rig

Camera ↔ camera calibration

Velodyne ↔ camera registration

GPS ↔ Velodyne registration

Geiger et al., ICRA 2012

ICP + Hand-eye calibration

P. Lenz: The KITTI Vision Benchmark Suite – www.mrt.kit.edu
Sensor Calibration Challenges

360° Velodyne Laserscanner

- Camera ↔ camera calibration
- Velodyne ↔ camera registration
- GPS ↔ Velodyne registration

\[ T_{GPS} \quad T_{C} \quad T_{Velodyne} \]

\{ Geiger et al., ICRA 2012 \}
\{ ICP + Hand-eye calibration \}

P. Lenz: The KITTI Vision Benchmark Suite – www.mrt.kit.edu
Data Annotation Challenges

- **3D object labels:** 22 Annotators
- **Occlusion labels:** Mechanical Turk
Data Statistics

Number of Labels

- Car: 150k
- Van: 50k
- Truck: 0
- Pedestrian (sitting): 100k
- Cyclist: 150k
- Tram: 0
- Misc: 0

P. Lenz: The KITTI Vision Benchmark Suite – www.mrt.kit.edu
Data Statistics

Number of Labels

- Car: 150k
- Van: 50k
- Truck: 0
- Pedestrian (sitting): 100k
- Cyclist: 150k
- Tram: 0
- Misc: 0

P. Lenz: The KITTI Vision Benchmark Suite – www.mrt.kit.edu
Data Statistics

- Fully Visible
- Partly Occluded
- Largely Occluded
Novel Challenges

Middlebury Stereo Evaluation – Version 2

Average errors: 2 – 3% (non-occluded regions)
Novel Challenges

Middlebury Stereo Evaluation – Version 2

Average errors: 2 – 3% (non-occluded regions)

<table>
<thead>
<tr>
<th>Error Threshold = 1</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm</td>
<td>Avg.</td>
<td>Tsukuba ground truth</td>
<td>Venus ground truth</td>
<td>Teddy ground truth</td>
</tr>
<tr>
<td>CoopRegion [41]</td>
<td>8.8</td>
<td>0.87</td>
<td>1.16</td>
<td>4.16</td>
</tr>
<tr>
<td>AdaptingBP [17]</td>
<td>9.0</td>
<td>1.11</td>
<td>1.37</td>
<td>5.79</td>
</tr>
<tr>
<td>ADCensus [94]</td>
<td>7.3</td>
<td>1.07</td>
<td>1.48</td>
<td>5.73</td>
</tr>
<tr>
<td>SurfaceStereo [79]</td>
<td>18.2</td>
<td>1.28</td>
<td>1.65</td>
<td>6.78</td>
</tr>
<tr>
<td>GC+SegmBorder [57]</td>
<td>27.1</td>
<td>1.47</td>
<td>1.82</td>
<td>7.86</td>
</tr>
<tr>
<td>WarpMat [55]</td>
<td>20.8</td>
<td>1.16</td>
<td>1.35</td>
<td>6.04</td>
</tr>
<tr>
<td>RDP [102]</td>
<td>12.5</td>
<td>0.97</td>
<td>1.39</td>
<td>5.00</td>
</tr>
<tr>
<td>RVbased [116]</td>
<td>11.6</td>
<td>0.95</td>
<td>1.42</td>
<td>4.98</td>
</tr>
<tr>
<td>OutlierConf [42]</td>
<td>12.9</td>
<td>0.88</td>
<td>1.43</td>
<td>4.74</td>
</tr>
</tbody>
</table>
Novel Challenges

Fast guided cost-volume filtering (Rhemann et al., CVPR 2011)

Middlebury, Errors: 2.7%

Error threshold: 1 px (Middlebury) / 3 px (KITTI)
Novel Challenges

Fast guided cost-volume filtering (Rhemann et al., CVPR 2011)

Middlebury, Errors: 2.7%

Kitti, Errors: 46.3%

- Error threshold: 1 px (Middlebury) / 3 px (Kitti)
Novel Challenges

So what is the difference?

Middlebury

- Laboratory
  - Lambertian
  - Rich in texture
  - Medium-size label set
  - Largely fronto-parallel

KITTI

- Moving vehicle
  - Specularities
  - Sensor saturation
  - Large label set
  - Strong slants
Novel Challenges

So what is the difference?

Middlebury
- Laboratory
- Lambertian
- Rich in texture
- Medium-size label set
- Largely fronto-parallel

KITTI
- Moving vehicle
- Specularities
- Sensor saturation
- Large label set
- Strong slants
Novel Challenges

So what is the difference?

**Middlebury**
- Laboratory
- Lambertian
- Rich in texture
- Medium-size label set
- Largely fronto-parallel

**KITTI**
- Moving vehicle
- Specularities
- Sensor saturation
- Large label set
- Strong slants
Novel Challenges

So what is the difference?

Middlebury
- Laboratory
- Lambertian
- Rich in texture
- Medium-size label set
- Largely fronto-parallel

KITTI
- Moving vehicle
- Specularities
- Sensor saturation
- Large label set
- Strong slants

P. Lenz: The KITTI Vision Benchmark Suite – www.mrt.kit.edu
Novel Challenges

So what is the difference?

<table>
<thead>
<tr>
<th>Middlebury</th>
<th>KITTI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laboratory</td>
<td>Moving vehicle</td>
</tr>
<tr>
<td>Lambertian</td>
<td>Specularities</td>
</tr>
<tr>
<td>Rich in texture</td>
<td>Sensor saturation</td>
</tr>
<tr>
<td>Medium-size label set</td>
<td>Large label set</td>
</tr>
<tr>
<td>Largely fronto-parallel</td>
<td>Strong slants</td>
</tr>
</tbody>
</table>
### Stereo Evaluation

**200 training images / 200 test images**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Method</th>
<th>Setting</th>
<th>Code</th>
<th>Out-Noc</th>
<th>Out-All</th>
<th>Avg-Noc</th>
<th>Avg-All</th>
<th>Density</th>
<th>Runtime</th>
<th>Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PCBP-SS</td>
<td>☑️</td>
<td>☑️</td>
<td>3.49 %</td>
<td>4.79 %</td>
<td>0.8 px</td>
<td>1.0 px</td>
<td>100.00 %</td>
<td>5 min</td>
<td>4 cores @ 2.5 Ghz (Matlab + C/C++)</td>
</tr>
<tr>
<td>2</td>
<td>StereoSLIC</td>
<td>☑️</td>
<td>☑️</td>
<td>3.99 %</td>
<td>5.17 %</td>
<td>0.9 px</td>
<td>1.0 px</td>
<td>99.89 %</td>
<td>2.3 s</td>
<td>1 core @ 3.0 Ghz (C/C++)</td>
</tr>
<tr>
<td>3</td>
<td>PR-ST+G</td>
<td>☑️</td>
<td>☑️</td>
<td>4.09 %</td>
<td>4.95 %</td>
<td>0.9 px</td>
<td>1.0 px</td>
<td>100.00 %</td>
<td>200 s</td>
<td>4 cores @ 3.0 Ghz (Matlab + C/C++)</td>
</tr>
<tr>
<td>4</td>
<td>PCBP</td>
<td>☑️</td>
<td>☑️</td>
<td>4.13 %</td>
<td>5.45 %</td>
<td>0.9 px</td>
<td>1.2 px</td>
<td>100.00 %</td>
<td>5 min</td>
<td>4 cores @ 2.5 Ghz (Matlab + C/C++)</td>
</tr>
<tr>
<td>5</td>
<td>PR-Sceneflow</td>
<td>☑️</td>
<td>☑️</td>
<td>4.46 %</td>
<td>5.32 %</td>
<td>1.0 px</td>
<td>1.1 px</td>
<td>100.00 %</td>
<td>150 sec</td>
<td>4 core @ 3.0 Ghz (Matlab - C/C++)</td>
</tr>
<tr>
<td>6</td>
<td>WSGM</td>
<td>☑️</td>
<td>☑️</td>
<td>5.03 %</td>
<td>6.24 %</td>
<td>1.3 px</td>
<td>1.6 px</td>
<td>97.03 %</td>
<td>6s</td>
<td>1 core @ 3.5 Ghz (C/C++)</td>
</tr>
<tr>
<td>7</td>
<td>ATGV</td>
<td>☑️</td>
<td>☑️</td>
<td>5.05 %</td>
<td>5.91 %</td>
<td>1.0 px</td>
<td>1.6 px</td>
<td>100.00 %</td>
<td>6 min</td>
<td>&gt;8 cores @ 3.0 Ghz (Matlab + C/C++)</td>
</tr>
<tr>
<td>8</td>
<td>iSGM</td>
<td>☑️</td>
<td>☑️</td>
<td>5.16 %</td>
<td>7.19 %</td>
<td>1.2 px</td>
<td>2.1 px</td>
<td>94.70 %</td>
<td>8 s</td>
<td>2 cores @ 2.5 Ghz (C/C++)</td>
</tr>
<tr>
<td>9</td>
<td>OCRG-SGBM2</td>
<td>☑️</td>
<td>☑️</td>
<td>5.42 %</td>
<td>6.54 %</td>
<td>1.0 px</td>
<td>1.2 px</td>
<td>100.00 %</td>
<td>2 s</td>
<td>1 core @ 2.5 Ghz (C/C++)</td>
</tr>
<tr>
<td>10</td>
<td>AABM</td>
<td>☑️</td>
<td>☑️</td>
<td>5.50 %</td>
<td>5.60 %</td>
<td>1.1 px</td>
<td>1.3 px</td>
<td>100.00 %</td>
<td>0.43 s</td>
<td>1 core @ 3.0 Ghz (C/C++)</td>
</tr>
</tbody>
</table>


# Stereo Evaluation

## 200 training images / 200 test images

### Stereo Evaluation

<table>
<thead>
<tr>
<th>Rank</th>
<th>Method</th>
<th>Setting</th>
<th>Code</th>
<th>Out-Noc</th>
<th>Out-All</th>
<th>Avg-Noc</th>
<th>Avg-All</th>
<th>Density</th>
<th>Runtime</th>
<th>Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PCBP-SS</td>
<td>3.49 %</td>
<td>4.79 %</td>
<td>0.8 px</td>
<td>1.0 px</td>
<td>100.00 %</td>
<td>5 min</td>
<td>4 cores @ 2.5 Ghz (Matlab + C/C++)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>StereoSLIC</td>
<td>3.99 %</td>
<td>5.17 %</td>
<td>0.9 px</td>
<td>1.0 px</td>
<td>99.89 %</td>
<td>2.3 s</td>
<td>1 core @ 3.0 Ghz (C/C++)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>PR-St+e</td>
<td>4.09 %</td>
<td>4.95 %</td>
<td>0.9 px</td>
<td>1.0 px</td>
<td>100.00 %</td>
<td>200 s</td>
<td>4 cores @ 3.0 Ghz (Matlab + C/C++)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>PCBP</td>
<td>4.13 %</td>
<td>5.45 %</td>
<td>0.9 px</td>
<td>1.2 px</td>
<td>100.00 %</td>
<td>5 min</td>
<td>4 cores @ 2.5 Ghz (Matlab + C/C++)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>PR-Sceneflow</td>
<td>4.46 %</td>
<td>5.32 %</td>
<td>1.0 px</td>
<td>1.1 px</td>
<td>100.00 %</td>
<td>150 sec</td>
<td>4 core @ 3.0 Ghz (Matlab - C/C++)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Anonymous submission:

<table>
<thead>
<tr>
<th>Rank</th>
<th>Method</th>
<th>Setting</th>
<th>Code</th>
<th>Out-Noc</th>
<th>Out-All</th>
<th>Avg-Noc</th>
<th>Avg-All</th>
<th>Density</th>
<th>Runtime</th>
<th>Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>wSGM</td>
<td>5.03 %</td>
<td>6.24 %</td>
<td>1.3 px</td>
<td>1.6 px</td>
<td>97.03 %</td>
<td>6s</td>
<td>1 core @ 3.5 Ghz (C/C++)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>ATGV</td>
<td>5.05 %</td>
<td>5.91 %</td>
<td>1.0 px</td>
<td>1.6 px</td>
<td>100.00 %</td>
<td>6 min</td>
<td>&gt;8 cores @ 3.0 Ghz (Matlab + C/C++)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>iSGM</td>
<td>5.16 %</td>
<td>7.19 %</td>
<td>1.2 px</td>
<td>2.1 px</td>
<td>94.70 %</td>
<td>8 s</td>
<td>2 cores @ 2.5 Ghz (C/C++)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>OCV-SGBM2</td>
<td>5.42 %</td>
<td>6.54 %</td>
<td>1.0 px</td>
<td>1.2 px</td>
<td>100.00 %</td>
<td>2 s</td>
<td>1 core @ 2.5 Ghz (C/C++)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>AABM</td>
<td>5.50 %</td>
<td>5.60 %</td>
<td>1.1 px</td>
<td>1.3 px</td>
<td>100.00 %</td>
<td>0.43 s</td>
<td>1 core @ 3.0 Ghz (C/C++)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


Stereo Evaluation

Particle Convex Belief Propagation (PCBP): **Best Results**

- Errors: 0.5%
- Natural scenes, lots of texture, no objects
- A couple of wrong pixels at poles
Stereo Evaluation

Particle Convex Belief Propagation (PCBP): Worst Results

Errors: 19.5%

- Inner city scenes, lots of objects
- Textureless surfaces, sensor saturation, reflections

Errors: 21.1%
Optical Flow Evaluation

Second Order Total Generalized Variation: **Best Results**

Errors: 0.5%

City scenes with slow motion (intersections)
Small flow vectors (< 30 px)
Optical Flow Evaluation

Second Order Total Generalized Variation: **Worst Results**

- Errors: 56.5%
- Errors: 58.8%

**Difficult lighting conditions, highway driving**

**Large flow vectors (> 150 px)**
22 sequences – 40 kilometers
11 training sequences / 11 test sequences

<table>
<thead>
<tr>
<th>Rank</th>
<th>Method</th>
<th>Submitted</th>
<th>Translation</th>
<th>Rotation</th>
<th>Runtime</th>
<th>Environment</th>
<th>Compare</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GT VO3pt</td>
<td>8 Jun. 2012</td>
<td>2.21 %</td>
<td>0.0117 [deg/m]</td>
<td>1.26 s</td>
<td>1 core @ 2.5 Ghz (C/C++)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>VIS02-S</td>
<td>12 Mar. 2012</td>
<td>2.28 %</td>
<td>0.0154 [deg/m]</td>
<td>0.05 s</td>
<td>1 core @ 2.5 Ghz (C/C++)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>VO3pt</td>
<td>15 Mar. 2012</td>
<td>2.93 %</td>
<td>0.0116 [deg/m]</td>
<td>0.56 s</td>
<td>1 core @ 2.0 Ghz (C/C++)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>VO3ptLB3</td>
<td>14 Mar. 2012</td>
<td>3.17 %</td>
<td>0.0180 [deg/m]</td>
<td>0.57 s</td>
<td>1 core @ 2.0 Ghz (C/C++)</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>VOFS</td>
<td>15 Mar. 2012</td>
<td>4.21 %</td>
<td>0.0158 [deg/m]</td>
<td>0.51 s</td>
<td>1 core @ 2.0 Ghz (C/C++)</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>VOFSLB3</td>
<td>15 Mar. 2012</td>
<td>4.35 %</td>
<td>0.0189 [deg/m]</td>
<td>0.52 s</td>
<td>1 core @ 2.0 Ghz (C/C++)</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>VIS02-M</td>
<td>4 Apr. 2012</td>
<td>13.79 %</td>
<td>0.0372 [deg/m]</td>
<td>0.1 s</td>
<td>1 core @ 2.5 Ghz (C/C++)</td>
<td></td>
</tr>
</tbody>
</table>

Andreas Geiger, Julius Ziegler and Christoph Stiller. *StereoScan: Dense 3d Reconstruction in Real-time*. IEEE Intelligent Vehicles Symposium 2011.


Challenges: Visual Odometry

- Ground truth from GPS+IMU
- **Metric**: For all frame combinations \((i, j)\):
Challenges: Visual Odometry

- Ground truth from GPS+IMU
- **Metric:** For all frame combinations \((i, j)\):

\[
\hat{T}_{i \rightarrow j}
\]
Challenges: Visual Odometry

- Ground truth from **GPS+IMU**
- **Metric:** For all frame combinations \((i, j)\):

\[
\mathbf{T}_{i \rightarrow j}
\]

![Graph showing translation and rotation errors for different methods with varying speeds.](image)

- **Translation Error [%]**
- **Rotation Error [deg/m]**

<table>
<thead>
<tr>
<th>Method</th>
<th>Translation Error [%]</th>
<th>Rotation Error [deg/m]</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIS2-S</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VO3ptLBA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VO3pt</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VOFSLBA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VOFS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VIS2-M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTvO3pt</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

P. Lenz: The KITTI Vision Benchmark Suite – www.mrt.kit.edu
Object Detection and Orientation

Image Selection for Training and Test

- Sequences are added exclusively to training or test set.
- Maximize no. of non-occluded objects.
- Images with many objects.
- Diversity in object orientation → Entropy Maximization.

\[ X \leftarrow X \cup \text{argmax}_x \left[ \alpha \cdot \text{noc}(x) + \frac{1}{C} \sum_{c=1}^{C} H_c(X \cup x) \right] \]

- \( X \): current set.
- \( x \): image from the whole dataset.
- \( \text{noc}(x) \): no. of non-occluded objects in an image.
- \( C \): no. of object classes.
- \( H_c \): entropy of class \( c \) with respect to the orientation.
Image Selection for Training and Test

- Sequences are added exclusively to training or test set
- Maximize no. of non-occluded objects
- Images with many objects
- Diversity in object orientation $\rightarrow$ Entropy Maximization

$$
\mathcal{X} \leftarrow \mathcal{X} \cup \arg\max_x \left[ \alpha \cdot noc(x) + \frac{1}{C} \sum_{c=1}^{C} H_c (\mathcal{X} \cup x) \right]
$$

- $\mathcal{X}$: current set
- $x$: image from the whole dataset
- $noc(x)$: no. of non-occluded objects in an image
- $C$: no. of object classes
- $H_c$: entropy of class $c$ with respect to the orientation
Object Detection and Orientation

Image Selection for Training and Test

- Sequences are added exclusively to training or testset
- Maximize no. of non-occluded objects
- Images with many objects
- Diversity in object orientation $\rightarrow$ Entropy Maximization

\[
\mathcal{X} \leftarrow \mathcal{X} \cup \text{argmax}_x \left[ \alpha \cdot noc(x) + \frac{1}{C} \sum_{c=1}^{C} H_c(\mathcal{X} \cup x) \right]
\]

- $\mathcal{X}$: current set
- $x$: image from the whole dataset
- $noc(x)$: no. of non-occluded objects in an image
- $C$: no. of object classes
- $H_c$: entropy of class $c$ with respect to the orientation
## Object Detection Evaluation

### Evaluation Criteria

<table>
<thead>
<tr>
<th>difficulty</th>
<th>height</th>
<th>occlusion</th>
<th>truncation</th>
</tr>
</thead>
<tbody>
<tr>
<td>easy</td>
<td>&gt; 40px</td>
<td>fully visible</td>
<td>&lt; 0.15</td>
</tr>
<tr>
<td>moderate</td>
<td>&gt; 25px</td>
<td>partly occluded</td>
<td>&lt; 0.30</td>
</tr>
<tr>
<td>hard</td>
<td>&gt; 25px</td>
<td>mostly occluded</td>
<td>&lt; 0.50</td>
</tr>
</tbody>
</table>

### Overlap $o$ Criterion per Class

<table>
<thead>
<tr>
<th>class</th>
<th>overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>car</td>
<td>&gt; 0.7</td>
</tr>
<tr>
<td>pedestrian</td>
<td>&gt; 0.5</td>
</tr>
<tr>
<td>cyclist</td>
<td>&gt; 0.5</td>
</tr>
</tbody>
</table>

\[
o = \frac{\text{GT} \cap \text{DET}}{\text{GT} \cup \text{DET}}
\]

Object Detector used as baseline for the benchmark

- Discriminatively Trained Deformable Part Models v4 [1]

Object Detection Evaluation

Evaluation Criteria

<table>
<thead>
<tr>
<th>difficulty</th>
<th>height</th>
<th>occlusion</th>
<th>truncation</th>
</tr>
</thead>
<tbody>
<tr>
<td>easy</td>
<td>&gt; 40px</td>
<td>fully visible</td>
<td>&lt; 0.15</td>
</tr>
<tr>
<td>moderate</td>
<td>&gt; 25px</td>
<td>partly occluded</td>
<td>&lt; 0.30</td>
</tr>
<tr>
<td>hard</td>
<td>&gt; 25px</td>
<td>mostly occluded</td>
<td>&lt; 0.50</td>
</tr>
</tbody>
</table>

Overlap \( o \) Criterion per Class

\[
o = \frac{GT \cap DET}{GT \cup DET}
\]

<table>
<thead>
<tr>
<th>class</th>
<th>overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>car</td>
<td>&gt; 0.7</td>
</tr>
<tr>
<td>pedestrian</td>
<td>&gt; 0.5</td>
</tr>
<tr>
<td>cyclist</td>
<td>&gt; 0.5</td>
</tr>
</tbody>
</table>

Object Detector used as baseline for the benchmark

Discriminatively Trained Deformable Part Models v4 [1]

Object Detection Evaluation

Evaluation Criteria

<table>
<thead>
<tr>
<th>difficulty</th>
<th>height</th>
<th>occlusion</th>
<th>truncation</th>
</tr>
</thead>
<tbody>
<tr>
<td>easy</td>
<td>&gt; 40px</td>
<td>fully visible</td>
<td>&lt; 0.15</td>
</tr>
<tr>
<td>moderate</td>
<td>&gt; 25px</td>
<td>partly occluded</td>
<td>&lt; 0.30</td>
</tr>
<tr>
<td>hard</td>
<td>&gt; 25px</td>
<td>mostly occluded</td>
<td>&lt; 0.50</td>
</tr>
</tbody>
</table>

Overlap $o$ Criterion per Class

<table>
<thead>
<tr>
<th>class</th>
<th>overlap</th>
<th>$o = \frac{GT \cap DET}{GT \cup DET}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>car</td>
<td>&gt; 0.7</td>
<td></td>
</tr>
<tr>
<td>pedestrian</td>
<td>&gt; 0.5</td>
<td></td>
</tr>
<tr>
<td>cyclist</td>
<td>&gt; 0.5</td>
<td></td>
</tr>
</tbody>
</table>

Object Detector used as baseline for the benchmark

- Discriminatively Trained Deformable Part Models v4 [1]

Object Orientation Evaluation

Evaluation Metric: Orientation

Average Orientation Similarity (AOS)

$$AOS = \frac{1}{11} \sum_{r \in \{0, 0.1, ..., 1\}} \max_{\tilde{r}: \tilde{r} \geq r} s(\tilde{r})$$

Normalized Cosine Similarity $s(r)$

$$s(r) = \frac{1}{|D(r)|} \sum_{i \in D(r)} \frac{1 + \cos \Delta_{\theta}^{(i)}}{2} \delta_i$$

- $D(r)$: set of all object detections at recall rate $r$
- $\Delta_{\theta}^{(i)}$: difference in angle between estimated and ground truth orientation of detection $i$
- $\delta_i = 1$ if detection $i$ has been assigned to a ground truth bounding box
- $\delta_i = 0$ if no ground truth has been assigned
Object Orientation Evaluation

Evaluation Metric: Orientation

Average Orientation Similarity (AOS)

\[
AOS = \frac{1}{11} \sum_{r \in \{0, 0.1, \ldots, 1\}} \max_{\tilde{r} : \tilde{r} \geq r} s(\tilde{r})
\]

Normalized Cosine Similarity \( s(r) \)

\[
s(r) = \frac{1}{|D(r)|} \sum_{i \in D(r)} \frac{1 + \cos \Delta^{(i)}_{\theta}}{2} \delta_i
\]

- \( D(r) \): set of all object detections at recall rate \( r \)
- \( \Delta^{(i)}_{\theta} \): difference in angle between estimated and ground truth orientation of detection \( i \)
- \( \delta_i = 1 \) if detection \( i \) has been assigned to a ground truth bounding box
- \( \delta_i = 0 \) if no ground truth has been assigned
Object Detection Evaluation

![Graph showing Orientation Similarity vs Recall for different difficulty levels of cars.]

- Easy (red line)
- Moderate (green dashed line)
- Hard (blue dotted line)

P. Lenz: The KITTI Vision Benchmark Suite – www.mrt.kit.edu
Object Detection Evaluation

Pedestrian

- Easy
- Moderate
- Hard

Precision
Recall
Object Detection Evaluation

Pedestrian

Orientation Similarity

Recall

Easy
Moderate
Hard
Which cars are hard to see?

Bad illumination conditions
Which cars are hard to see?

Bad illumination conditions
Which cars are hard to see?

Occlusions
Which cars are hard to see?

Occlusions
Which pedestrians are hard to see?

Children
Which pedestrians are hard to see?

Carried Items
Conclusion

Where are we now?

- Realistic dataset with 3D ground truth
  - Stereo
  - Optical flow
  - SLAM
  - Object detection / orientation estimation
  - Object Tracking
    - “Recognition meets Reconstruction Challenge”
- Complement existing benchmarks / reduce overfitting
- Submit your results: www.cvlibs.net/datasets/kitti
Conclusion

Where are we now?

- Realistic dataset with 3D ground truth
  - Stereo
  - Optical flow
  - SLAM
  - Object detection / orientation estimation
  - Object Tracking
    “Recognition meets Reconstruction Challenge”

- Complement existing benchmarks / reduce overfitting
- Submit your results: www.cvlibs.net/datasets/kitti
Conclusion

Where are we now?

- Realistic dataset with 3D ground truth
  - Stereo
  - Optical flow
  - SLAM
  - Object detection / orientation estimation
  - Object Tracking
    “Recognition meets Reconstruction Challenge”
- Complement existing benchmarks / reduce overfitting

Submit your results: www.cvlibs.net/datasets/kitti
Conclusion

Where are we now?

- Realistic dataset with 3D ground truth
  - Stereo
  - Optical flow
  - SLAM
  - Object detection / orientation estimation
  - Object Tracking
    “Recognition meets Reconstruction Challenge”
- Complement existing benchmarks / reduce overfitting
- Submit your results: www.cvlibs.net/datasets/kitti
Conclusion

But KITTI is much more ...

- 3D object tracking
- Loop closure (SLAM)
- Structure-from-Motion
- Semantic segmentation (class labels)
- 3D scene understanding (layout and objects)
- Use of maps

Lenz et al., IV 2011
Conclusion

But KITTI is much more ...

- 3D object tracking
- Loop closure (SLAM)
- Structure-from-Motion
- Semantic segmentation (class labels)
- 3D scene understanding (layout and objects)
- Use of maps

Williams et al., RAS 2009
Conclusion

But KITTI is much more ... 
- 3D object tracking
- Loop closure (SLAM)
- Structure-from-Motion
- Semantic segmentation (class labels)
- 3D scene understanding (layout and objects)
- Use of maps

Agarwal et al., ICCV 2009
Conclusion

But KITTI is much more ...

- 3D object tracking
- Loop closure (SLAM)
- Structure-from-Motion
- Semantic segmentation (class labels)
- 3D scene understanding (layout and objects)
- Use of maps

Wojek et al., ECCV 2008
Conclusion

But KITTI is much more ...

- 3D object tracking
- Loop closure (SLAM)
- Structure-from-Motion
- Semantic segmentation (class labels)
- 3D scene understanding (layout and objects)
- Use of maps

Geiger et al., CVPR 2011 and NIPS 2011
Conclusion

But KITTI is much more ...

- 3D object tracking
- Loop closure (SLAM)
- Structure-from-Motion
- Semantic segmentation (class labels)
- 3D scene understanding (layout and objects)
- Use of maps

OpenStreetMap – The free Wiki World Map
## Related Datasets and Benchmarks

<table>
<thead>
<tr>
<th>Stereo / Optical Flow</th>
<th>setting</th>
<th>#images</th>
<th>ground truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>EISATS</td>
<td>synthetic</td>
<td>500</td>
<td>dense</td>
</tr>
<tr>
<td>Middlebury</td>
<td>laboratory</td>
<td>40</td>
<td>dense</td>
</tr>
<tr>
<td>Make3D Stereo</td>
<td>real</td>
<td>260</td>
<td>0.5 %</td>
</tr>
<tr>
<td>Ladicky</td>
<td>real</td>
<td>70</td>
<td>manual</td>
</tr>
<tr>
<td><strong>KITTI</strong></td>
<td>real</td>
<td>400</td>
<td>50 %</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SLAM</th>
<th>setting</th>
<th>length</th>
<th>metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>TUM RGB-D</td>
<td>indoor</td>
<td>0.4 km</td>
<td>✓</td>
</tr>
<tr>
<td>New College</td>
<td>outdoor</td>
<td>2.2 km</td>
<td></td>
</tr>
<tr>
<td>Malaga 2009</td>
<td>outdoor</td>
<td>6.4 km</td>
<td></td>
</tr>
<tr>
<td>Ford Campus</td>
<td>outdoor</td>
<td>5.1 km</td>
<td></td>
</tr>
<tr>
<td><strong>KITTI</strong></td>
<td>outdoor</td>
<td>39.2 km</td>
<td>✓</td>
</tr>
</tbody>
</table>
## Related Datasets and Benchmarks

<table>
<thead>
<tr>
<th>Object Detection</th>
<th>#cat.</th>
<th>#labels/cat.</th>
<th>occlusion</th>
<th>3D</th>
<th>orientation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caltech 101</td>
<td>101</td>
<td>40-800</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIT StreetScenes</td>
<td>9</td>
<td>3k</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LabelMe</td>
<td>4000</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ETHZ Pedestrian</td>
<td>1</td>
<td>12k</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PASCAL 2011</td>
<td>20</td>
<td>1k</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Daimler</td>
<td>1</td>
<td>56k</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caltech Pedestrian</td>
<td>1</td>
<td>350k</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>COIL-100</td>
<td>100</td>
<td>72</td>
<td>✓</td>
<td></td>
<td>discrete</td>
</tr>
<tr>
<td>EPFL Multi-View</td>
<td>20</td>
<td>90</td>
<td>✓</td>
<td></td>
<td>discrete</td>
</tr>
<tr>
<td>Caltech 3D</td>
<td>100</td>
<td>144</td>
<td>✓</td>
<td>✓</td>
<td>discrete</td>
</tr>
<tr>
<td><strong>KITTl</strong></td>
<td><strong>3</strong></td>
<td><strong>1k - 40k</strong></td>
<td>✓</td>
<td>✓</td>
<td><strong>continuous</strong></td>
</tr>
</tbody>
</table>