Learning Temporal Features with CNNs for Monocular Visual Ego Motion Estimation

Michael Weber, Christoph Rist and J. Marius Zöllner
Data Flow

**Input:**
Monocular color images
Consolidate consecutive frames to one data point

**Label:**
Transformation
Rotation as Tait-Bryan-Rotation

**Output:**
6-dim Vector
Translation + Rotation

- Yaw
- Pitch
- Roll

![Diagram showing data flow and transformation labels]
Extracting Temporal Information within CNNs

- Design principles for extracting temporal information

Karpathy et al - Large-scale video classification with convolutional neural networks
Extracting Temporal Information within CNNs

- Design principles for extracting temporal information

(a) Late Fusion
(b) Early Fusion
(c) Slow Fusion

Karpathy et al - Large-scale video classification with convolutional neural networks
Other Visual Odometry CNN Work

• Not generalizing to unseen data
Early Fusion Architecture

- Two Input Frames
- Spatial Dropout

![Diagram of Early Fusion Architecture](image-url)
Slow Fusion Architecture

- 3D Convolution Layers
- 2D Pooling
- Spatial Dropout
3D Convolution Layers

Tran et al. – Learning Spatiotemporal Features with 3D Convolutional Networks
Evaluation Data – KITTI Odometry Dataset

- 11 Sequences with given Ground Truth for Training/Validation purposes
- 11 Sequences for Testing (KITTI Benchmark)
- Inner city road / rural road / highway / campus
- Rectified RGB-stereo images, ~10 Hz
- Image Resolution approx. 1240x370
Data Augmentation

- Probabilistic input modification
- Temporal Flipping: Virtually reversing
- Synthetic zero movement
Training Details

- **Loss:**
  \[ L_i = ||\Delta \hat{t}_i - \Delta t_i||^2 + \beta ||\Delta \hat{r}_i - \Delta r_i||^2 \]
  
  \[ \text{Balance factor between translational and rotational error} \]

- **Optimizer:** SGD/Adam, Mini-Batch Training
- **Training on 10 sequences, Validation on 1 sequence**
- **44,000 data points**
- **Slow Fusion Network trained for 350 epochs**
- **Early Fusion Network trained for 1000 epochs**
Results

<table>
<thead>
<tr>
<th>CNN architecture</th>
<th>Inference time</th>
<th>Per step</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slow Fusion Network</td>
<td>23.40 ms</td>
<td>2.62 ms</td>
</tr>
<tr>
<td>Early Fusion Network</td>
<td>6.37 ms</td>
<td>6.37 ms</td>
</tr>
</tbody>
</table>
Results – KITTI Testsequence 15

(c) Sequence 15: Paths
Thank you for your attention! Questions?