Integration of Deep Optical Flow in Visual-Inertial Odometry

Semester Thesis

Jingkun Feng
Advisor: Mariia Gladkova

January 31st, 2022
Outline

- Introduction and Motivation
- Preliminaries
  - Optical Flow
  - Basalt VIO
- Integration and Outlier Removal
- Evaluation
- Discussion
- Summary
Introduction and Motivation

- Before, handcrafted optical flow
- Recently, deep optical flow with rise of deep learning
- Inspired by DF-VO from Zhang et al [2]

**Aim** to explore probability of leveraging deep optical flow to improve the **accuracy** and **robustness** of a state-of-the-art VIO system.
Preliminaries

Optical Flow

- A displacement vector describes apparent motion of the same pixel in consecutive frames.

![Optical Flow Diagram](image)

Fig 1. Optical flow for a single pixel. Constant intensity is assumed: \( I(x_1, y_1, t_1) = I(x_1, y_1, t_1) = I(x_1, y_1, t_1) \)

- Useful for feature tracking

- Assumptions:
  - Brightness constancy
  - Constant motion in a local neighborhood (Lucas-Kanade method [5])
  - Spatially smooth motion (Horn-Schunck method [6])

- Sparse or dense vector field

![Sparse Optical Flow](image)  ![Color Coded Dense Optical Flow](image)

Fig 2. Sparse optical flow  Fig 3. Color coded dense optical flow
Preliminaries

Basalt VIO [1]

- Consists of **visual-inertial odometry** and visual-inertial mapping
- Algorithm framework of Basalt VIO

![Fig 4. Basalt VIO framework](image)

- Patch-based KLT for tracking
  - Locally-scaled sum of squared differences (LSSD)
  - Coarse-to-fine optimization using pyramidal approaches
Preliminaries
Basalt VIO

- Locally-scaled sum of squared differences (LSSD)
  - Patch $\Omega$
  - Desired transformation $T \in SE(2)$ between two matching patches in adjacent images
  - Average intensity of all pixels in the patch $\bar{I}$
  - Residual $r$ of an increment $\xi$

$$r_i(\xi) = \frac{I_{t+1}(Tx_i)}{I_{t+1}} - \frac{I_t(x_i)}{I_t}$$

- Minimize LSSD over patches to obtain $T$

$$\argmin_{T \in SE(2)} \sum_{x_i \in \Omega} (r_i(\xi))^2$$

- Coarse-to-fine optimization using pyramidal approaches
  - Achieve robustness to large displacements in the image
  - The pyramid level is fixed
  $\rightarrow$ only robust to large displacements in certain degree
Integration and Outlier Removal

Integration

- Extract **FAST** keypoints
  - Split the image into regular cells
  - Extract and track the *keypoint with strongest response* in each cell
  - Resample if no keypoint remains in the cell
Integration and Outlier Removal

Integration

- Extract **FAST** keypoints
  - Split the image into regular cells
  - Extract and track the keypoint with strongest response in each cell
  - Resample if no keypoint remains in the cell

- Deep optical flow for temporal feature tracking
  - Predict forward optical flow using **Recurrent All-Pairs Field Transforms (RAFT)** # [3]
  - Use deep optical flow as prior to warp patches
  - **Refine by minimizing LSSD**

- Pyramidal KLT for stereo matching

# The model we used is the pretrained model released in the official repo of RAFT.
Integration and Outlier Removal

Outlier Removal

1. Forward-backward flow inconsistency
   - To remove outliers in temporal feature tracking

2. Epipolar constraint
   - To remove outliers in stereo matching
Integration and Outlier Removal

Outlier Removal

Forward-backward flow inconsistency

- Predict backward optical flow
- Track points from the current frame to the target frame and back
- Calculate distance between initial position and position after the second tracking
- Large distance denotes high inconsistency → to remove
Integration and Outlier Removal

Outlier Removal

Epipolar constraint

- Check epipolar geometry of correspondences on stereo images
- Calibration $\rightarrow$ Fundamental matrix $F$
- $x'Fx = 0$

- Remove points on the right frame if constraint is violated
- Keep points on the left frame

Fig 7. Epipolar geometry
Evaluation Dataset

1. KITTI Odometry [4]
   - 11 stereo sequences of various driving scenarios with ground-truth
   - Due to storage limitation, long sequences (02, 05, 08) are excluded
   - Grayscale and color images
   - No IMU data

2. EuRoC MAV [9]
   - 11 sequences of different difficulties with accurate motion ground-truth
   - Collected on-board a drone (6 DoF)
   - Grayscale images
   - IMU measurements
Evaluation

Evaluation Metrics

1. Root mean squared absolute trajectory error: $ATE$
2. Relative pose error: translational $RPE_{tran}$ and rotational $RPE_{rot}$
3. Average translational and rotational error: $t_{err}$ and $r_{err}$

Notation:

- Estimated camera pose: $Q \in SE(3)$
- Ground-truth camera pose: $P \in SE(3)$
- Translation and rotation part of a rigid body transformation $T$: $trans(T)$, $rot(T)$
Evaluation
Evaluation Metrics – Root Mean Squared Absolute Trajectory Error (ATE)

- Evaluate global consistency
- Align the estimated and the ground-truth trajectory with a transformation matrix $S$ (Horn method [])
- Absolute trajectory error matrix at time step $i$
  \[ E_i := Q_i^{-1}SP_i \]
- Compute the root mean squared error over all time indices
  \[
  ATE := \sqrt{\frac{1}{m} \sum_{i=1}^{m} ||\text{trans}(E_i)||^2}
  \]
Evaluation
Relative Pose Error ($RPE_{rot}$, $RPE_{tran}$)

- Evaluate local consistency
- Relative pose error matrix $F_{i:\Delta} := (Q_i^{-1}Q_{i+\Delta})^{-1} (P_i^{-1}P_{i+\Delta})$
- Translational part

$$RPE_{trans} := \sqrt{\frac{1}{m} \sum_{i=1}^{m} \|\text{trans}(F_i)\|^2} \quad \text{for} \quad i = 1, \ldots, n$$

- Rotational part

$$RPE_{rot} := \frac{1}{m} \sum_{i=1}^{m} \angle F_i \quad \text{for} \quad i = 1, \ldots, n \quad \text{where} \quad \angle F_i := \arccos\left(\frac{\text{tr}(\text{rot}(F_i)) - 1}{2}\right)$$
Evaluation
Average translational and rotational error

- Specific metric adopted to evaluation on KITTI Odometry
- Measures errors as function of the trajectory length
Evaluation Results

- On KITTI Odometry

<table>
<thead>
<tr>
<th>Method</th>
<th>Metric</th>
<th>01</th>
<th>03</th>
<th>04</th>
<th>05</th>
<th>06</th>
<th>07</th>
<th>09</th>
<th>10</th>
<th>Avg. excl. 01</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>$t_{err}$</td>
<td>4.5239</td>
<td>0.9962</td>
<td>1.1921</td>
<td>1.7646</td>
<td>1.0605</td>
<td>0.8625</td>
<td>1.0590</td>
<td>0.5892</td>
<td>0.9320</td>
</tr>
<tr>
<td></td>
<td>$r_{err}$</td>
<td>0.1713</td>
<td>0.2293</td>
<td>0.1922</td>
<td>0.2276</td>
<td>0.2313</td>
<td>0.4851</td>
<td>0.1937</td>
<td>0.2652</td>
<td>0.2606</td>
</tr>
<tr>
<td></td>
<td>ATE</td>
<td>30.7334</td>
<td>1.3648</td>
<td>1.2690</td>
<td>2.7245</td>
<td>2.5591</td>
<td>1.5547</td>
<td>4.3127</td>
<td>0.9834</td>
<td>2.1098</td>
</tr>
<tr>
<td></td>
<td>RPE$_{tran}$</td>
<td>0.6737</td>
<td>0.0143</td>
<td>0.0267</td>
<td>0.0136</td>
<td>0.0183</td>
<td>0.0113</td>
<td>0.0213</td>
<td>0.0139</td>
<td>0.0171</td>
</tr>
<tr>
<td></td>
<td>RPE$_{rot}$</td>
<td>0.0469</td>
<td>0.0328</td>
<td>0.0237</td>
<td>0.0309</td>
<td>0.0239</td>
<td>0.0281</td>
<td>0.0332</td>
<td>0.0383</td>
<td>0.0301</td>
</tr>
<tr>
<td>Ours</td>
<td>$t_{err}$</td>
<td>1.7562</td>
<td>0.9033</td>
<td>0.9665</td>
<td>0.6996</td>
<td>0.9144</td>
<td>X</td>
<td>0.9602</td>
<td>0.6122</td>
<td>0.8427</td>
</tr>
<tr>
<td></td>
<td>$r_{err}$</td>
<td>0.1258</td>
<td>0.2144</td>
<td>0.2342</td>
<td>0.2262</td>
<td>0.2432</td>
<td>X</td>
<td>0.1819</td>
<td>0.2459</td>
<td>0.2243</td>
</tr>
<tr>
<td></td>
<td>ATE</td>
<td>5.1679</td>
<td>1.0309</td>
<td>1.0170</td>
<td>2.2426</td>
<td>2.4629</td>
<td>X</td>
<td>3.7208</td>
<td>0.9023</td>
<td>1.8961</td>
</tr>
<tr>
<td></td>
<td>RPE$_{tran}$</td>
<td>0.2653</td>
<td>0.0133</td>
<td>0.0240</td>
<td>0.0118</td>
<td>0.0146</td>
<td>X</td>
<td>0.0188</td>
<td>0.0131</td>
<td>0.0159</td>
</tr>
<tr>
<td></td>
<td>RPE$_{rot}$</td>
<td>0.0324</td>
<td>0.0325</td>
<td>0.0226</td>
<td>0.0303</td>
<td>0.0228</td>
<td>X</td>
<td>0.0322</td>
<td>0.0380</td>
<td>0.0297</td>
</tr>
</tbody>
</table>

Table 1: Evaluation results on KITTI Odometry (Seq. 01, 03-07, 09, 10).

- On EuRoC MAV

<table>
<thead>
<tr>
<th>Method</th>
<th>Metric</th>
<th>MH_01</th>
<th>MH_02</th>
<th>MH_03</th>
<th>MH_04</th>
<th>MH_05</th>
<th>V1_01</th>
<th>V1_02</th>
<th>V1_03</th>
<th>V2_01</th>
<th>V2_02</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>ATE</td>
<td>0.09081</td>
<td>0.05387</td>
<td>0.08488</td>
<td>0.10852</td>
<td>0.12732</td>
<td>0.04284</td>
<td>0.05636</td>
<td>0.07201</td>
<td>0.05636</td>
<td>0.06414</td>
<td>0.07650</td>
</tr>
<tr>
<td></td>
<td>RPE$_{tran}$</td>
<td>0.00138</td>
<td>0.00180</td>
<td>0.00374</td>
<td>0.00509</td>
<td>0.00370</td>
<td>0.00229</td>
<td>0.00295</td>
<td>0.00508</td>
<td>0.00118</td>
<td>0.00988</td>
<td>0.00371</td>
</tr>
<tr>
<td></td>
<td>RPE$_{rot}$</td>
<td>0.00040</td>
<td>0.00043</td>
<td>0.00055</td>
<td>0.00069</td>
<td>0.00054</td>
<td>0.00068</td>
<td>0.00086</td>
<td>0.00107</td>
<td>0.00067</td>
<td>0.00098</td>
<td>0.00069</td>
</tr>
<tr>
<td>Ours</td>
<td>ATE</td>
<td>0.08618</td>
<td>0.05395</td>
<td>0.07096</td>
<td>0.10008</td>
<td>0.10767</td>
<td>0.04322</td>
<td>0.04114</td>
<td>0.04876</td>
<td>0.03777</td>
<td>0.03974</td>
<td>0.06295</td>
</tr>
<tr>
<td></td>
<td>RPE$_{tran}$</td>
<td>0.00036</td>
<td>0.00138</td>
<td>0.00353</td>
<td>0.00485</td>
<td>0.00357</td>
<td>0.00228</td>
<td>0.00265</td>
<td>0.00348</td>
<td>0.00110</td>
<td>0.00298</td>
<td>0.00272</td>
</tr>
<tr>
<td></td>
<td>RPE$_{rot}$</td>
<td>0.00038</td>
<td>0.00041</td>
<td>0.00054</td>
<td>0.00066</td>
<td>0.00051</td>
<td>0.00067</td>
<td>0.00082</td>
<td>0.00104</td>
<td>0.00065</td>
<td>0.00091</td>
<td>0.00066</td>
</tr>
</tbody>
</table>

Table 2: Evaluation results on EuRoC MAV (V2_03 is excluded).

- Outperforms the original in terms of global and local accuracy
- However, our system fails at a single frame on KITTI 07.
Discussion

Failure Case

Fig 8. Failure case. Time step above is t and below is t+1.
Discussion

Ablation Study

1. Optical flow inference: Grayscale vs. color images

2. With or without refinement using LSSD

3. … (for other studies please refer to the paper)
Discussion

Ablation Study – Grayscale vs. RGB

- Color images are more informative than grayscale images
- Most existing datasets (e.g., Flyingthings [9] and Sintel [8]) contain merely color images.
- Currently proposed deep-learning-based methods mainly train on color images.

- Evaluated on KITTI Odometry
Discussion

Ablation Study – Grayscale vs. RGB

- Using color images for optical flow inference can boost performance in pose estimation.

<table>
<thead>
<tr>
<th>Method</th>
<th>Metric</th>
<th>03</th>
<th>04</th>
<th>05</th>
<th>06</th>
<th>09</th>
<th>10</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grayscale</td>
<td>$t_{err}$</td>
<td>0.9033</td>
<td>0.9665</td>
<td>0.6996</td>
<td>0.9144</td>
<td>0.9602</td>
<td>0.6122</td>
<td>0.8427</td>
</tr>
<tr>
<td></td>
<td>$r_{err}$</td>
<td>0.2144</td>
<td>0.2342</td>
<td>0.2262</td>
<td>0.2432</td>
<td>0.1819</td>
<td>0.2459</td>
<td>0.2243</td>
</tr>
<tr>
<td></td>
<td>ATE</td>
<td>1.0309</td>
<td>1.0170</td>
<td>2.2426</td>
<td>2.4629</td>
<td>3.7208</td>
<td>0.9023</td>
<td>1.8961</td>
</tr>
<tr>
<td></td>
<td>$RPE_{tran}$</td>
<td>0.0133</td>
<td>0.0240</td>
<td>0.0118</td>
<td>0.0146</td>
<td>0.0188</td>
<td>0.0131</td>
<td>0.0159</td>
</tr>
<tr>
<td></td>
<td>$RPE_{rot}$</td>
<td>0.0325</td>
<td>0.0226</td>
<td>0.0303</td>
<td>0.0228</td>
<td>0.0322</td>
<td>0.0380</td>
<td>0.0297</td>
</tr>
<tr>
<td>RGB</td>
<td>$t_{err}$</td>
<td>0.8827</td>
<td>1.0082</td>
<td>0.6946</td>
<td>0.9170</td>
<td>0.9644</td>
<td>0.5622</td>
<td>0.8382</td>
</tr>
<tr>
<td></td>
<td>$r_{err}$</td>
<td>0.2249</td>
<td>0.2282</td>
<td>0.2234</td>
<td>0.2420</td>
<td>0.1849</td>
<td>0.2278</td>
<td>0.2219</td>
</tr>
<tr>
<td></td>
<td>ATE</td>
<td>1.0049</td>
<td>1.0668</td>
<td>2.1745</td>
<td>2.3706</td>
<td>3.7324</td>
<td>0.8626</td>
<td>1.8686</td>
</tr>
<tr>
<td></td>
<td>$RPE_{tran}$</td>
<td>0.0134</td>
<td>0.0241</td>
<td>0.0118</td>
<td>0.0147</td>
<td>0.0188</td>
<td>0.0130</td>
<td>0.0159</td>
</tr>
<tr>
<td></td>
<td>$RPE_{rot}$</td>
<td>0.0324</td>
<td>0.0228</td>
<td>0.0303</td>
<td>0.0228</td>
<td>0.0322</td>
<td>0.0380</td>
<td>0.0297</td>
</tr>
</tbody>
</table>

Table 3 Evaluation results of ablation study about the image format used for inference on KITTI Odometry (Seq.03, 04, 05, 06, 09, 10).

- But
  - the improvement is not significant, about 1% in average ATE.
  - only some of the datasets provide RGB images.
Discussion

Ablation Study – Refinement

- In general, refinement helps achieve more accurate trajectory estimation.
- System with refined optical flow has obvious larger drift in KITTI 03 and 06.

<table>
<thead>
<tr>
<th>Method</th>
<th>Metric</th>
<th>03</th>
<th>04</th>
<th>05</th>
<th>06</th>
<th>09</th>
<th>10</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Refined</td>
<td>t_{err}</td>
<td>0.9033</td>
<td>0.9665</td>
<td>0.6996</td>
<td>0.9144</td>
<td>0.9602</td>
<td>0.6122</td>
<td>0.8427</td>
</tr>
<tr>
<td></td>
<td>r_{err}</td>
<td>0.2144</td>
<td>0.2342</td>
<td>0.2262</td>
<td>0.2432</td>
<td>0.1819</td>
<td>0.2459</td>
<td>0.2243</td>
</tr>
<tr>
<td></td>
<td>ATE</td>
<td>1.0309</td>
<td>1.0170</td>
<td>2.2426</td>
<td>2.4629</td>
<td>3.7208</td>
<td>0.9023</td>
<td>1.8961</td>
</tr>
<tr>
<td></td>
<td>RPE_{tran}</td>
<td>0.0133</td>
<td>0.0240</td>
<td>0.0118</td>
<td>0.0146</td>
<td>0.0188</td>
<td>0.0131</td>
<td>0.0159</td>
</tr>
<tr>
<td></td>
<td>RPE_{rot}</td>
<td>0.0325</td>
<td>0.0226</td>
<td>0.0303</td>
<td>0.0228</td>
<td>0.0322</td>
<td>0.0380</td>
<td>0.0297</td>
</tr>
<tr>
<td>Not refined</td>
<td>t_{err}</td>
<td>0.6714</td>
<td>1.0386</td>
<td>0.8153</td>
<td>1.0155</td>
<td>1.0376</td>
<td>0.6348</td>
<td>0.8689</td>
</tr>
<tr>
<td></td>
<td>r_{err}</td>
<td>0.2595</td>
<td>0.4916</td>
<td>0.2730</td>
<td>0.3138</td>
<td>0.2466</td>
<td>0.3638</td>
<td>0.3247</td>
</tr>
<tr>
<td></td>
<td>ATE</td>
<td>0.6747</td>
<td>0.9074</td>
<td>3.3607</td>
<td>2.0447</td>
<td>4.4613</td>
<td>1.1232</td>
<td>2.0953</td>
</tr>
<tr>
<td></td>
<td>RPE_{tran}</td>
<td>0.0147</td>
<td>0.0385</td>
<td>0.0154</td>
<td>0.0225</td>
<td>0.0242</td>
<td>0.0165</td>
<td>0.0220</td>
</tr>
<tr>
<td></td>
<td>RPE_{rot}</td>
<td>0.0333</td>
<td>0.0287</td>
<td>0.0327</td>
<td>0.0279</td>
<td>0.0352</td>
<td>0.0410</td>
<td>0.0331</td>
</tr>
</tbody>
</table>

Table 4 Evaluation results of ablation study about refinement of the deep optical flow on KITTI Odometry (Seq.03, 04, 05, 06, 09, 10).

<table>
<thead>
<tr>
<th>Method</th>
<th>Metric</th>
<th>MH_01</th>
<th>MH_02</th>
<th>MH_03</th>
<th>MH_04</th>
<th>MH_05</th>
<th>V1_01</th>
<th>V1_02</th>
<th>V1_03</th>
<th>V2_01</th>
<th>V2_02</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Refined</td>
<td>ATE</td>
<td>0.0862</td>
<td>0.0540</td>
<td>0.0710</td>
<td>0.1001</td>
<td>0.1077</td>
<td>0.0432</td>
<td>0.0411</td>
<td>0.0488</td>
<td>0.0378</td>
<td>0.0397</td>
<td>0.0629</td>
</tr>
<tr>
<td></td>
<td>RPE_{tran}</td>
<td>0.0014</td>
<td>0.0014</td>
<td>0.0035</td>
<td>0.0048</td>
<td>0.0036</td>
<td>0.0023</td>
<td>0.0026</td>
<td>0.0035</td>
<td>0.0011</td>
<td>0.0030</td>
<td>0.0027</td>
</tr>
<tr>
<td></td>
<td>RPE_{rot}</td>
<td>0.0004</td>
<td>0.0004</td>
<td>0.0005</td>
<td>0.0007</td>
<td>0.0005</td>
<td>0.0007</td>
<td>0.0008</td>
<td>0.0010</td>
<td>0.0007</td>
<td>0.0009</td>
<td>0.0006</td>
</tr>
<tr>
<td>Not refined</td>
<td>ATE</td>
<td>0.2238</td>
<td>0.1707</td>
<td>0.1429</td>
<td>0.4098</td>
<td>0.3747</td>
<td>0.0543</td>
<td>0.0549</td>
<td>0.0508</td>
<td>0.0397</td>
<td>0.0530</td>
<td>0.1574</td>
</tr>
<tr>
<td></td>
<td>RPE_{tran}</td>
<td>0.0022</td>
<td>0.0027</td>
<td>0.0044</td>
<td>0.0082</td>
<td>0.0063</td>
<td>0.0024</td>
<td>0.0030</td>
<td>0.0035</td>
<td>0.0014</td>
<td>0.0026</td>
<td>0.0038</td>
</tr>
<tr>
<td></td>
<td>RPE_{rot}</td>
<td>0.0005</td>
<td>0.0005</td>
<td>0.0006</td>
<td>0.0008</td>
<td>0.0006</td>
<td>0.0007</td>
<td>0.0009</td>
<td>0.0011</td>
<td>0.0007</td>
<td>0.0009</td>
<td>0.0007</td>
</tr>
</tbody>
</table>

Table 5 Evaluation results of ablation study about refinement of the deep optical flow on EuRoC MAV.
Discussion

Ablation Study – Refinement

- In general, refinement helps achieve more accurate trajectory estimation
- System with refined optical flow has obviously larger drift on KITTI 03 and 06
Discussion
Timing and Efficiency

- **Not efficient**
  - A huge part of available information is not in use.
    - About 300 pixels out of \((370 \times 1226)\) pixels

- **Not real-time capable**
  - Original Basalt VIO is around 4 times faster than real-time
    - Frame rate of EuRoC is 30 fps (0.03s per frame)
    - About 7.5 ms per frame on EuRoC
  - However, Optical flow inference is very "time consuming".
    - 0.4 s per frame on EuRoC using RAFT
Summary

- We extended the Basalt VIO by integrating deep optical flow
  - replace the pyramid KLT tracker in BASALT VIO with refined deep optical flow
  - remove outliers using forward-backward flow inconsistency and epipolar constraint

- According to the evaluation, our system outperforms the original Basalt VIO w.r.t accuracy of trajectory estimation.

- However, our integration has drawbacks
  - less robust to dynamic objects
  - inefficient in terms of the usage of available information
  - not real time capable
Thank you!
Reference


Reference


Appendix
Qualitative Evaluation Results – KITTI Odometry

Qualitative evaluation results on KITTI Odometry Seq. 01, 03-06, 09, 10
Appendix
Qualitative Evaluation Results – EuRoC MAV

Qualitative evaluation results on EuRoC MAV (V2_03 is excluded)