Challenges in “appearance+shape” loop-closure detection for reliable SfM

Adrien Angeli
Imperial College London
aangeli@doc.ic.ac.uk
http://www.doc.ic.ac.uk/~aangeli

April 12, 2011
Monocular SLAM

- Can we still do SLAM with a single unconstrained camera, flying generally through the world in 3D?

- 30Hz or higher operation required to track agile motion.
- Salient feature patches detected once to serve as long-term visual landmarks.
- Landmarks gradually accumulated and stored indefinitely.

Adrien Angeli
Challenges in “appearance+shape” loop-closure detection for reliable SfM
MonoSLAM

Davison, ICCV 2003; Davison, Molton, Reid, Stasse, PAMI 2007.
General Components of a Scalable SLAM Algorithm

Local Metric  Place Recognition  Global Optimisation
Large Scale Monocular SLAM using Optimisation

Scale Drift-Aware Large Scale Monocular SLAM (Strasdat, Montiel, Davison, Robotics: Science and Systems 2010).

Many of the components of this system are now available in the RobotVision open source software package (openslam.org).
Loop-closure detection

Problem:
- SLAM drifts over time,
- leads to inconsistent global pose/map.

Solution:
- *detect* when the camera is re-visiting a known location,
- compensate for drift (global optimisation).

→ Loop-closure detection is crucial for long-term reliability.
Appearance-based topological place recognition

Classify the most recent image into:

- an existing location of the environment model,
- a new location.

**Bag of words** paradigm for image characterisation:

- unordered set of quantised feature descriptors,
- incremental construction of the dictionary [Filliat, 2007].

**Topological environment model:**

- 1 node per location,
- several close views represent a location.
Discrete Bayes filter [Angeli et al., 2008b], [Angeli et al., 2008a]

Bayesian loop-closure detection

Search for the location $L_{i^*}$ of the environment model that is similar enough to the most recent image $I_t$ to consider that $I_t$ pertains to $L_{i^*}$:

$$L_{i^*} = \arg\max_{i=-1,\ldots,n} p(S_t = i|I_t, M_{t-1})$$  \hspace{1cm} (1)

- $S_t = i \rightarrow "I_t"$ comes from $L_i$, $S_t = -1 \rightarrow$: “new location”.
- $M_{t-1} = \{L_0, \ldots, L_n\}$ is the environment model.

$$p\left(S_t|(z^n)_t, M_{t-1}\right) =$$

$$\eta \left[ \prod_{k=0}^{n} p\left((z_k)_t|S_t, M_{t-1}\right) \right] \sum_{i=-1}^{m} p\left(S_t|S_{t-1} = i, M_{t-1}\right) p\left(S_{t-1} = i|M_{t-1}\right)$$

prediction

likelihood model $\prod_{k=0}^{n} p\left((z_k)_t|S_t, M_{t-1}\right)$

transition model $\sum_{i=-1}^{m} p\left(S_t|S_{t-1} = i, M_{t-1}\right) p\left(S_{t-1} = i|M_{t-1}\right)$  \hspace{1cm} (2)
Likelihood Model and Verification

- Simple voting scheme (inverted index), with \( tf-idf \) weighting [Sivic and Zisserman, 2003].
- Rapid computation, no exhaustive comparison with every location.

Adrien Angeli
Challenges in “appearance+shape” loop-closure detection for reliable SfM
Indoor loop-closure detection results

Start position
Current position
Image viewpoint

No loop-closure detection
Loop-closure detected
Loop-closure rejected

1cm ~ 2.7m

Adrien Angeli
Challenges in “appearance+shape” loop-closure detection for reliable SfM
### Feature spaces and performances

#### Table: Performance with / without colour cue.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>#img</th>
<th>#LC</th>
<th>%TP</th>
<th>#FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indoor SIFT + H</td>
<td>388</td>
<td>217</td>
<td>80</td>
<td>0</td>
</tr>
<tr>
<td>Indoor SIFT</td>
<td>388</td>
<td>217</td>
<td>68</td>
<td>0</td>
</tr>
<tr>
<td>Outdoor SIFT + H</td>
<td>531</td>
<td>301</td>
<td>71</td>
<td>0</td>
</tr>
<tr>
<td>Outdoor SIFT</td>
<td>531</td>
<td>301</td>
<td>70</td>
<td>0</td>
</tr>
</tbody>
</table>

Adding colour information:
- enhances **true positive** rate,
- does not prevent real-time processing.

%TP = rate of correct LCD, #FP = number of incorrect LCD.

#### Table: Performances.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Length</th>
<th>#img</th>
<th>CPU</th>
<th>#SIFT</th>
<th>#H hist.</th>
<th>Frame rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indoor SIFT + H</td>
<td>6m28s</td>
<td>388</td>
<td>2m52s</td>
<td>9201</td>
<td>7284</td>
<td>1Hz</td>
</tr>
<tr>
<td>Indoor SIFT</td>
<td>6m28s</td>
<td>388</td>
<td>1m33s</td>
<td>9201</td>
<td>0</td>
<td>1Hz</td>
</tr>
<tr>
<td>Outdoor SIFT + H</td>
<td>17m42s</td>
<td>531</td>
<td>10m16s</td>
<td>39175</td>
<td>18408</td>
<td>0.5Hz</td>
</tr>
<tr>
<td>Outdoor SIFT</td>
<td>17m42s</td>
<td>531</td>
<td>6m48s</td>
<td>39175</td>
<td>0</td>
<td>0.5Hz</td>
</tr>
</tbody>
</table>

Length = duration of the sequence, CPU = CPU time required to process it.
Further applications

Topo-metric SLAM [Angeli et al., 2009]
Are we fully satisfied with such LCD solution?

Characteristics of the proposed solution:

- reliable and very robust to perceptual aliasing,
- completely incremental solution, fast execution (real-time at 1Hz on a standard laptop).

Limitations:

- responsiveness in situations of high perceptual aliasing could be improved,
- robustness across very different viewpoints can be increased.

→ Appearance information alone is not sufficient, structure information about the scene must be employed in complement.
Improving viewpoint invariance using structure information

- **Problem:** over significantly different viewpoints, feature-based appearance information is not reliable for LCD.
- **Solution:** employ the structure information contained in the map provided by the SLAM system.

**Approach:**

- calculate a structure descriptor + an appearance descriptor for each point feature of the most recent keyframe (KF),
- employ a voting procedure based on structure descriptors only to retrieve most likely-to-match KFs,
- attempt full pairwise KF matching with appearance and structure (*late combination*).
Keyframe retrieval using structure information

Approach (Angeli, Davison 2011 — under reviewing process):

- for each 3D point feature of the KF, approximate the direction and amount of maximum curvature,
- capture the joint distribution of relative curvature amount and orientation over a small neighbourhood around the point of interest.
6DoF Keyframe retrieval results

|     | \( |S \cup A| \) | Perf_{S} |
|-----|----------------|-----------|
| Seq1| 90             | 16%       |
| Seq2| 26             | 73%       |
| Seq3| 43             | 47%       |
| All | 159            | 33%       |

\( \text{Perf}_{S} \) quantifies the number of correct loop-closures that were found using our solution and which could not be detected using a traditional purely appearance-based BoW approach.

Table: Recognition performance.
Are we fully satisfied with such LCD solution?

Improvements over a purely appearance-based approach:

- tighter integration with the SLAM system → re-localisation is trivial,
- enables loop-closure detection over significantly different viewpoints.

Limitations:

- “good” structure not available after a tracking failure → LCD delay,
- can easily fail when the ranges of depths differ a lot from 1 KF to another (e.g., appearance / disappearance of background stuff),
- recall rate at precision $1 < 1$ → there is still room for improvement.

→ Better recognition performance requires a more complete, dense model of the scene.
Live dense surface reconstruction

Newcombe, Lovegrove, Davison 2011 — under reviewing process.
How do we characterise a dense KF?

Keypoints?
- Corners?
- Blobs?

What information should be used?
- Normal map $\rightarrow$ stable across viewpoints.
- Relative orientations between regions of piecewise-constant normal $\rightarrow$ very KF specific, and viewpoint independent.
Dense KF characterisation

Approach:

- find regions of piecewise-constant normal (region growing),
- characterise the appearance and structure of each region using 2 rotation invariant descriptors (extension of Spin Images [Johnson and Hebert, 1999]).
KF matching examples: comparison with SIFT
Conclusion

Reliable SfM / SLAM over the long term requires:

- very accurate local camera tracking,
- loop-closure detection,
- global optimisation.

Robust and reliable loop-closure detection requires:

- a tight LCD / SLAM integration,
- a combination of appearance / structure information,
- that all possible information (dense models) should be used.
Bibliography: I


