Spatiotemporal Descriptor for Wide-Baseline Stereo
Reconstruction of Non-Rigid and Ambiguous scenes

E. Trulls, A. Sanfeliu, F. Moreno–Noguer
Institut de Robòtica i Informàtica Industrial, CSIC-UPC, 08028 Barcelona, Spain

PROBLEM
Computing dense depth maps from stereo video sequences under complex ambiguities such as highly repetitive patterns, noisy images and non-rigid deformations, on wide baseline setups with occlusions.

CONTRIBUTIONS
• A dense stereo reconstruction algorithm that can handle very challenging situations on wide baseline scenarios.
• A new approach to spatiotemporal stereo, warping appearance descriptors to capture the evolution of the neigbourhood of a pixel in time. In contrast, most state-of-the-art techniques attempt to describe cubic-shaped volumes of space-time.

METHODOLOGY
For descriptor extraction, we augment a state-of-the-art appearance descriptor with optical flow priors. For stereo reconstruction we use a traditional graph-cuts global optimization scheme while enforcing spatial and temporal consistency.

SPATIOTEMPORAL DESCRIPTOR
Many spatiotemporal strategies treat space-time as a volume, which does not translate to the general case of wide-baseline set-ups.

Core ideas: capture the temporal changes around a point.

We base our descriptor on Daisy, a SIFT-like descriptor (histograms of gradient orientations). Daisy is:
• Designed for dense computation.
• Computed over a discrete, adaptable grid.

To build the spatiotemporal descriptor for frame $F^t$:
1) We estimate backward and forward flow vectors for every consecutive pair over $T$ frames.
2) The grid is warped (a) translating each point, and (b) reorienting the descriptor orientation relative to the center.
3) The $3D'$ descriptor is assembled concatenating the $2D'$ descriptors, validating them against the original frame to discard bad matches (occlusions, lighting changes, bad flow estimates).

The distance between spatiotemporal descriptors is defined as the average distance between valid sub-descriptor pairs:

$$d = \frac{1}{p \cdot (p+1) / 2} \cdot D(D_1^{u1}(x), D_1^{u2}(x))$$

($D$ as in [1])

STEREO RECONSTRUCTION
For stereo reconstruction, we:
• Use a pair or calibrated monocular cameras.
• Discretize 3D space into a given number of depth bins.
• Compute the distance between every possible match restricted to epipolar geometry, and store the best match for every depth bin.
• Cast the results into a graph-cuts [3] global optimization algorithm with a truncated linear model for the smoothness cost.

To enforce spatiotemporal consistency:
• We perform the optimization over $3M$ frames at a time (e.g. 3).
• Every pixel $(x, y, t)$ is linked to its four adjacent neighbours in its frame and to $(x, y, t-1)$.
• The estimates at either end are discarded due to edge artefacts.

RESULTS - Synthetic data (w/ ground truth)

Table: Accuracy (% under error threshold)

<table>
<thead>
<tr>
<th>Noise variance</th>
<th>0.05</th>
<th>0.10</th>
<th>0.15</th>
<th>0.20</th>
<th>0.25</th>
<th>0.30</th>
<th>0.35</th>
<th>0.40</th>
<th>0.45</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daisy(136x7)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Daisy(200)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Baselines:
- Baseline 1 (0.05)
- Baseline 2 (0.10)
- Baseline 3 (0.15)
- Baseline 4 (0.20)
- Baseline 5 (0.25)
- Baseline 6 (0.30)
- Baseline 7 (0.35)
- Baseline 8 (0.40)
- Baseline 9 (0.45)

References