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

Institut de Robòtica i Informàtica Industrial

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**.

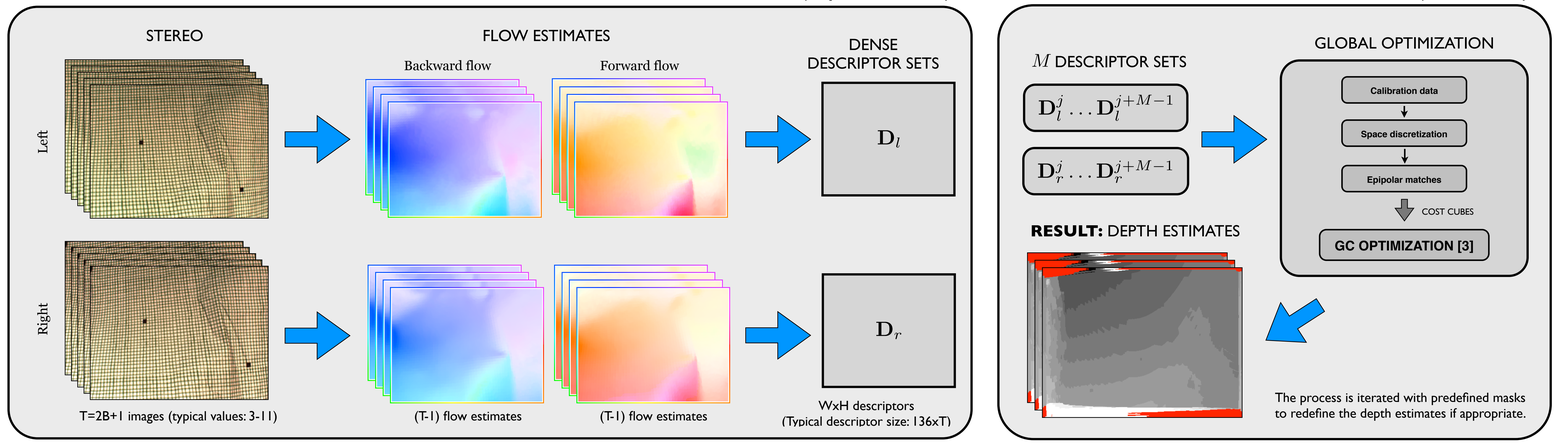
CONTRIBUTIONS

- 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 neighbourhood of a pixel in time. In contrast, most state-of-the-art techniques attempt to describe cubic-shaped volumes of space-time.

METHODOLOGY

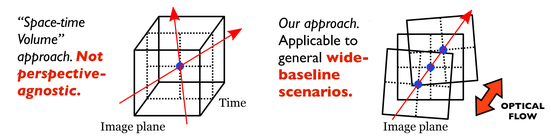
For **descriptor extraction**, we augment a state-of-the-art appearance descriptor with optical flow priors. For **stereo reconstruction** we use a traditional graph-cuts global optimization scheme while enforcing *spatial* and *temporal* consistency.

OVERVIEW



SPATIOTEMPORAL DESCRIPTOR

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



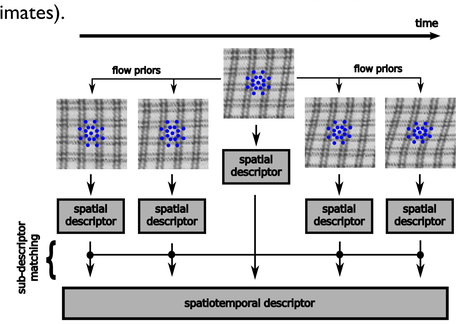
Core idea: capture the **temporal changes around a point.**

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^k:

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:

$$\widetilde{D} = \frac{1}{V} \sum_{l=k-B}^{k+B} v_l \cdot D(\mathbf{D}_1^{\{l\}}(\mathbf{x}), \mathbf{D}_2^{\{l\}}(\mathbf{x})) \qquad (D \text{ as in [1]})$$

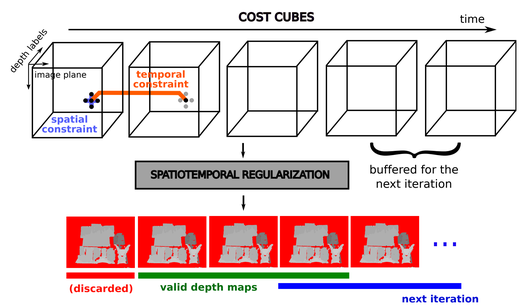
STEREO RECONSTRUCTION

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.

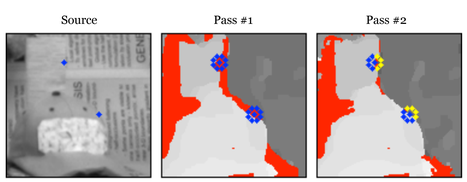
To enforce **spatiotemporal consistency:**

- We perform the optimization over M frames at a time (e.g. 5).
- Every pixel (x, y, t) is linked to its four adjacent neighbours on its frame and to $(x, y, t \pm 1)$.
- The estimates at either end are discarded due to edge artifacts.

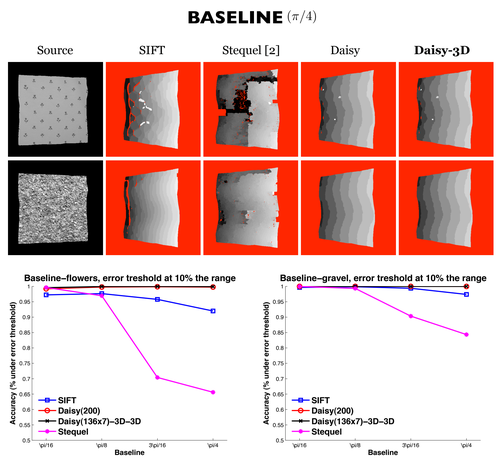


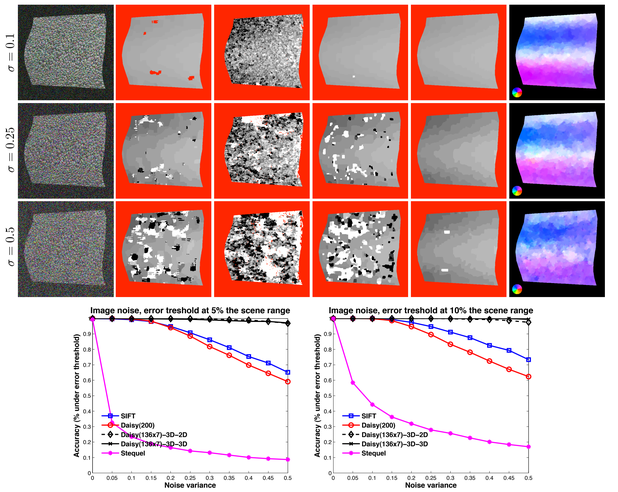
Masks for occlusions

We define a set of **binary masks** over the polar grid as in [1]. The masks are preset (half moons) to enforce spatial coherence, and are used to refine the depth estimates **iterating the stereo process.**

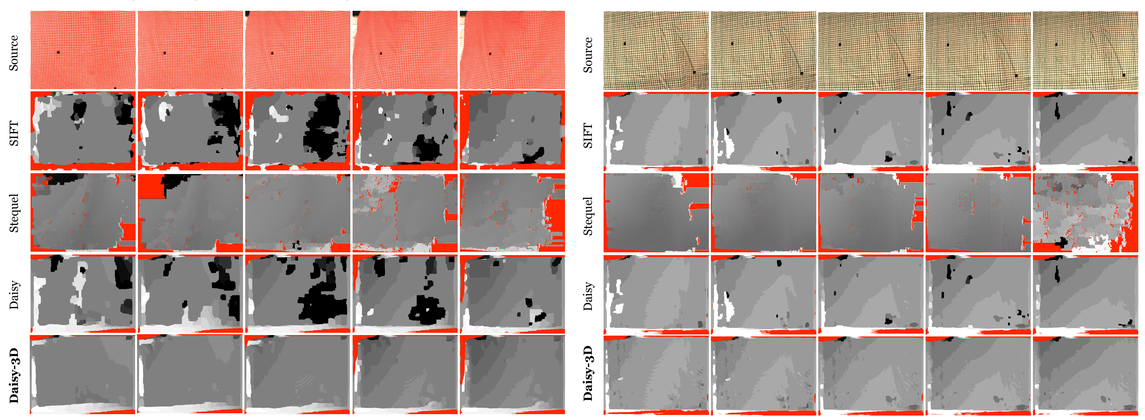


RESULTS - Synthetic data (w/ ground truth)





Real data (w/o ground truth)



References

[1] Tola, E., Lepetit, V., Fua, P.: Daisy: An efficient dense descriptor applied to wide-baseline stereo. T.PAMI (2010)
[2] Sizintsev, M., Wildes, R.: Spatiotemporal stereo via spatiotemporal quadric element (stequel) matching. CVPR (2009)
[3] Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. T.PAMI (2001)