Scalable Structure from Motion for Densely Sampled Videos Supplemental Material

B. Resch[1,2] H. P. A. Lensch[2,3] O. Wang[1] M. Pollefeys[3] A. Sorkine-Hornung[1]
[1]Disney Research Zurich [2]Tübingen University [3]ETH Zurich

In this supplemental material, we present some more information about our test datasets. Additionally, we show results from reconstructions of datasets not used for evaluation in the paper: Stanford Light Fields Archive datasets, 5k video and a cooperative data acquisition scenario where the same room was captured with different cameras.

1. Paper dataset information

We used the following datasets in the paper for which we show some example images and illustrations of the reconstructions, as well as individual reconstruction timings (Figure 1).

SHIP The SHIP dataset consists of a video sequence with 4353 frames, recorded with a DSLR camera in 1080p. The camera was mounted on a slowly moving crane to simulate high framerates: the camera translation from frame to frame is about 1/4 of the other test scenes in the paper, so it corresponds to the same camera velocity at about 240 fps. The trajectory starts and stops at the same position which allows camera drift to be measured. Refer to Figure 2 for some example images and a 3D visualization of the reconstruction.

OFFICE The OFFICE dataset consists of a video sequence with 1333 frames, recorded with a GoPro Hero 3 camera in 1080p 60 fps wide angle mode. The trajectory starts and stops at the same position which allows camera drift to be measured. Refer to Figure 3 for some example images and a 3D visualization of the reconstruction.

BENCH The BENCH dataset consists of a video sequence with 1055 frames, recorded with a GoPro Hero 3 camera in 1080p 60 fps wide angle mode. The trajectory starts and stops at the same position which allows camera drift to be measured. Refer to Figure 4 for some example images and a 3D visualization of the reconstruction.

STAIRWAY The STAIRWAY dataset consists of a video sequence with 876 frames, recorded with a GoPro Hero 3 camera in 1080p 60 fps wide angle mode. The trajectory starts and stops at the same position which allows camera drift to be measured. Refer to Figure 5 for some example images and a 3D visualization of the reconstruction.

OFFICE 2 The OFFICE 2 dataset consists of a video sequence with 1180 frames, recorded with a GoPro Hero 3 camera in 1080p 60 fps wide angle mode. The trajectory starts and stops at the same position which allows camera drift to be measured. Refer to Figure 6 for some example images and a 3D visualization of the reconstruction.

SWINGING The SWINGING dataset consists of a video sequence with 1257 frames, recorded with a GoPro Hero 3 camera in 1080p 60 fps wide angle mode. The trajectory starts and stops at the same position which allows camera drift to be measured. Refer to Figure 7 for some example images and a 3D visualization of the reconstruction.

2 LOOPS The 2 LOOPS dataset consists of a video sequence with 1715 frames, recorded with a GoPro Hero 3 camera in 1080p 60 fps wide angle mode. The trajectory starts and stops at the same position which allows camera drift to be measured. The camera circles an object twice which allows to compare our algorithm with and without global anchor constraints between the two loops. Refer to Figure 8 for some example images and a 3D visualization of the reconstruction.

2. Pseudocode

For easier reimplementation, we give the pseudo code of our method in Algorithm 1.

3. Stanford lightfields

We reconstructed some of the Stanford Light Field Archive datasets by treating the individual 10 megapixel raw images in zigzag order as video stream. Each dataset contains an array of 17x17=289 camera poses. For timings refer to Figure 9. An exemplary reconstruction visualization of the lego truck scene can be found in Figure 10.

Algorithm 1 High level pseudo code of the proposed method. This algorithm covers the main function "reconstruct" and a selected subset its called functions.

function RECONSTRUCT($images$, $costMatrix$)
  $windows \leftarrow \emptyset$ ▷ *windows* only contains keyframes
  for all $imageSeq \subseteq images$ **do**
    $tracks \leftarrow$ TRACKING($imageSeq$)
    $startPoses \leftarrow$ INITWINDOWBA($imageSeq, tracks$) ▷ See paper Section 3.2.1.
    $windows \leftarrow windows \cup$ WINDOWBA($startPoses, imageSeq, tracks$)
  end for
  $windows \leftarrow windows \cup$ ANCHORHANDLING($windows, costMatrix$)
  $constraints \leftarrow$ EXTRACTCONSTRAINTS($windows$) ▷ *constraints* only contains information about keyframes
  $scene \leftarrow$ INITPOSESLINEAR($constraints$) ▷ *scene* only contains keyframes poses here
  $scene \leftarrow$ OPTIMIZE($scene$)
  $scene \leftarrow$ INTERPOLATENONKEYFRAMEPOSES($scene$)
  $scene \leftarrow$ OPTIMIZE($scene$)
end function

function TRACKING($images$) ▷ See paper Section 3.1.
  $allTracks \leftarrow \emptyset$
  $activeTracks \leftarrow \emptyset$
  for $image \in images$ **do**
    for all $track \in activeTracks$ **do**
      $feature \leftarrow$ FINDFEATURE($image, track.previousImagePattern$)
      $feature \leftarrow$ REFINEFEATURE($feature, image, track.firstImagePattern$)
      if $feature \neq$ **null then**
        EXTENDTRACK($track, feature$)
      else
        $activeTracks \leftarrow activeTracks \setminus \{track\}$
      end if
    end for
    $newTracks \leftarrow$ FINDNEWKEYPOINTS($image$)
    $newTracks \leftarrow$ FILTERTOOCLOSE($newTracks, activeTracks$)
    $activeTracks \leftarrow activeTracks \cup newTracks$
    $allTracks \leftarrow allTracks \cup newTracks$
  end for
  return $allTracks$
end function

function WINDOWBA($startPoses, images, tracks$) ▷ See paper Section 3.2.4.
  $currentWindow \leftarrow$ CREATESCENE($startPoses$)
  $windows \leftarrow \{currentWindow\}$
  while last image not processed **do**
    $currentWindow \leftarrow$ REMOVEUNCONFIDENTPOSES($currentWindow$)
    $baPoses \leftarrow$ SELECTBASUBSET($currentWindow$)
    $lastConfidentPose \leftarrow$ **null**
    while true do
      $candidate \leftarrow$ PICKIMAGELININCREASINGOFFSET($images$)
      if ISCONFIDENT($baPoses, candidate$) **then** ▷ Includes optimizing for the candidate's pose against baPoses
        $lastConfidentPose \leftarrow candidate$
      else
        break
      end if
    end while

```
        currentWindow ← ADDPOSE(currentWindow, lastConfidentPose)
        windows ← windows ∪ {currentWindow}
        constraints = EXTRACTCONSTRAINTS(windows)
        constraints = FILTERBYPOSES(currentWindow.cameraPoses)
        currentWindow ← INITPOSESLINEAR(constraints)
    end while
    return windows
end function

function ANCHORHANDLING(windows, costMatrix)                    ▷ See paper Section 3.3.
    anchorWindows ← ∅
    varianceMatrix ← ESTIMATEVARIANCES(windows)
    samplingMatrix ← (1 − costMatrix) ∘ varianceMatrix
    while |anchorWindows| < maxGlobalAnchors do
        cameraPair ← IMPORTANCESAMPLEDPAIR(samplingMatrix)
        window ← BESTWINDOWCONTAINING(windows, cameraPair.first)
        anchorWindow ← ADDPOSE(window, cameraPair.second)
        anchorWindow ← OPTIMIZE(anchorWindow)
        if ISSTABLE(anchorWindow, cameraPair.second) then
            anchorWindows ← anchorWindows ∪ {anchorWindow}
        end if
    end while
    return anchorWindows
end function

function OPTIMIZE(scene)
    scene.points ← FINDPOINTS(scene.cameraPoses)
    scene.points ← SUBSAMPLEPOINTS(scene.points)                ▷ See paper Section 3.2.2.
    scene ← BA(scene)
    while config says so do
        scene.points ← FINDPOINTS(scene.cameraPoses)
        scene.points ← FILTERBADPOINTS(scene.points)
        scene.points ← SUBSAMPLEPOINTS(scene.points)            ▷ See paper Section 3.2.2.
        scene ← BA(scene)
    end while
    return scene
end function
```

4. 5k high resolution video

We reconstructed a scene from very high resolution video with 5120x2700 pixels and 901 frames. Timings and a visualization of the reconstruction can be found in Figure 11.

Note that the major bottleneck in this reconstruction was the subsampling of points whose acceleration is left for future work.

5. Large, cooperative reconstruction

We reconstructed an office room from video footage with 14254 frames that consists of several video sequences recorded with three different cameras. We used a GoPro Hero3, a budged consumer Canon S100 photo camera and a Sony camcorder and recorded all video sequences in FullHD with frame rates between 25 and 60 Hz.

Example images, a visualization of the reconstructed scene and camera tracks as well as timing information can be found in Figure 12.

Refer to Figure 13 for an illustration of the effect of different framerates on the camera tracks.

5.1. Video

Please refer also to the supplied video for a more three-dimensional impression of the reconstruction.

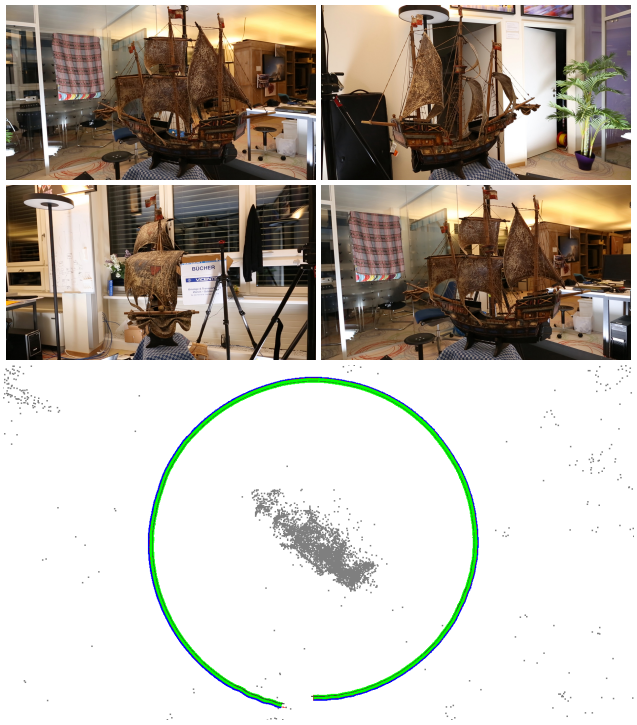

Figure 2. Ship scene: Input images and visualization - topdown view. Note that the loop is not closed because this scene was reconstructed without global anchor constraints to measure camera drift.

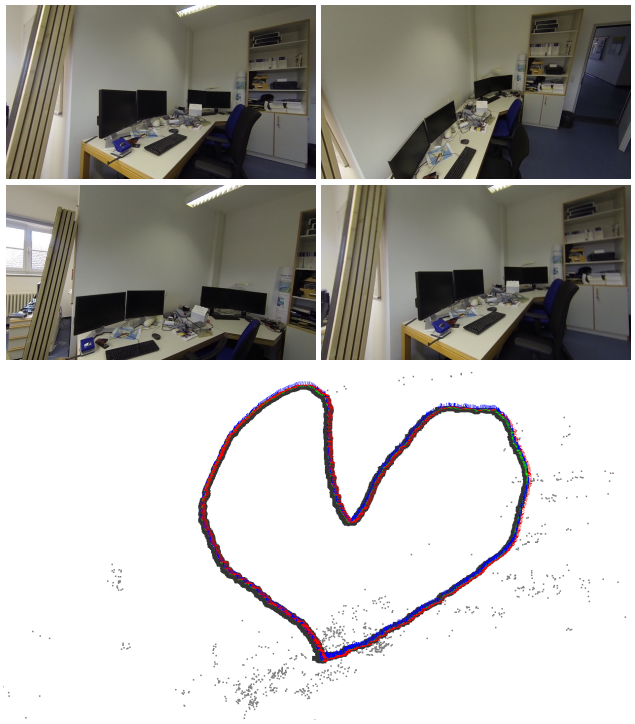

Figure 3. Office scene: Input images and visualization - viewing in camera direction.

Figure 4. Bench scene: Input images and visualization - frontal view.

Figure 5. Stairway scene: Input images and visualization - topdown view.

Figure 6. Office 2 scene: Input images and visualization - topdown view. Note that the loop is not closed because this scene was reconstructed without global anchor constraints to measure camera drift.

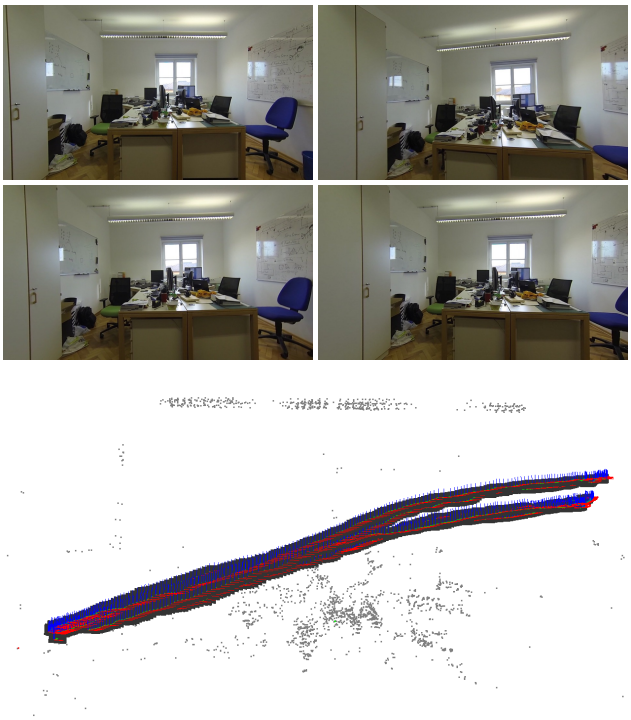

Figure 7. Swinging scene: Input images and visualization - viewing in camera direction.

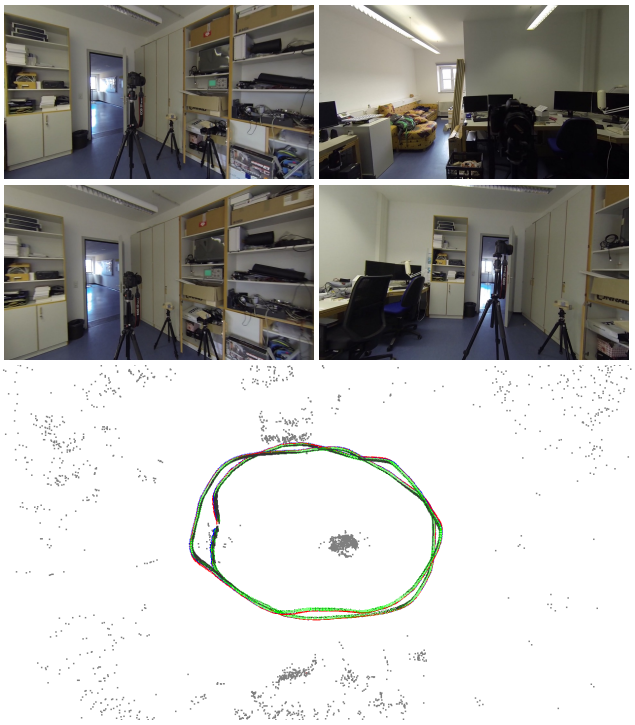

Figure 8. 2 loops scene: Input images and visualization - topdown view.

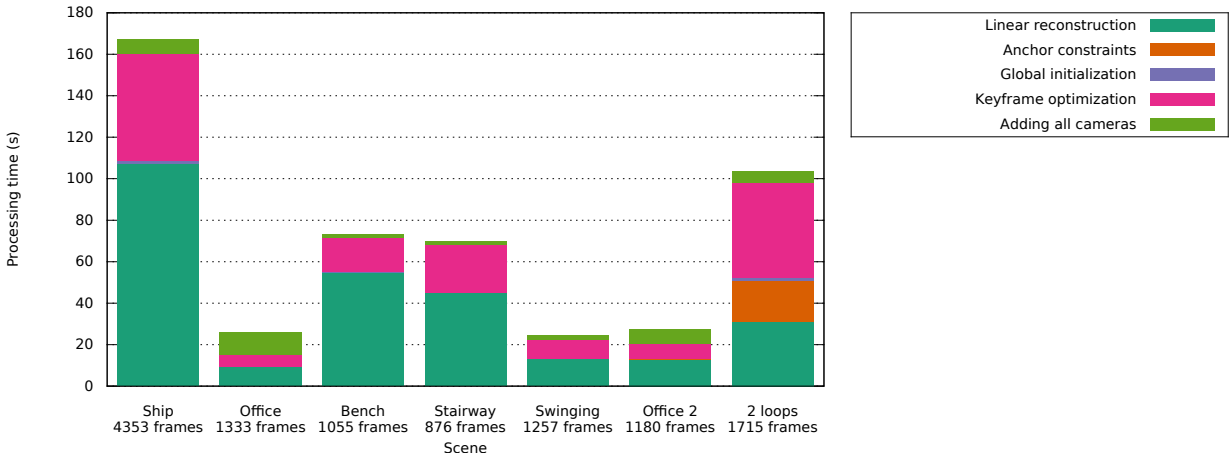

Figure 1. Reconstruction timings for the test scenes used in the paper. Reconstruction does not depend directly on the number of frames but on how much scene information was captured with a specific camera movement. Redundant or unnecessary information is skipped by our frame selection and point subsampling. Note that most scenes were reconstructed without anchor constraints to be able to evaluate the camera drift.

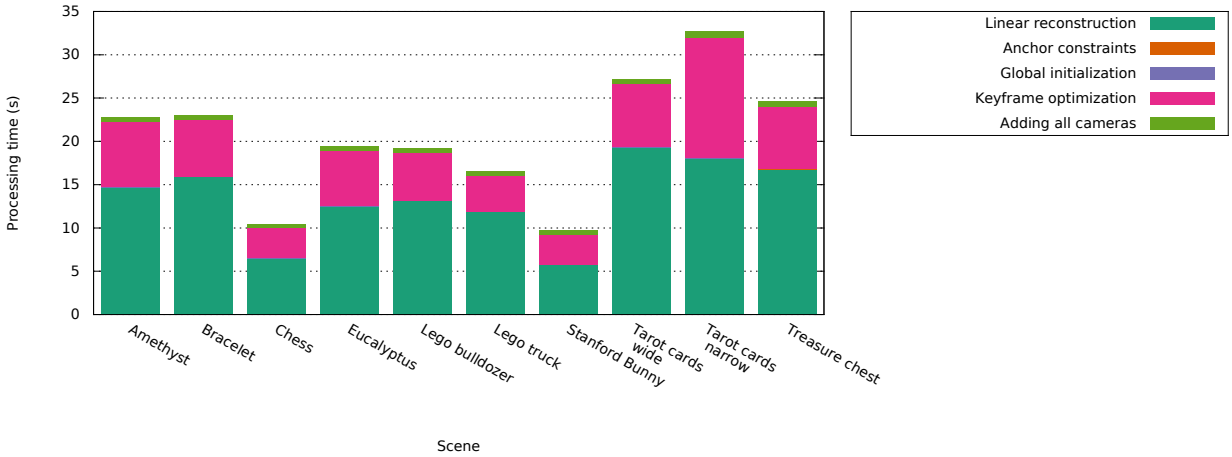

Figure 9. Timings for the stanford dataset reconstruction. Each scene contains 289 10MPixel images.

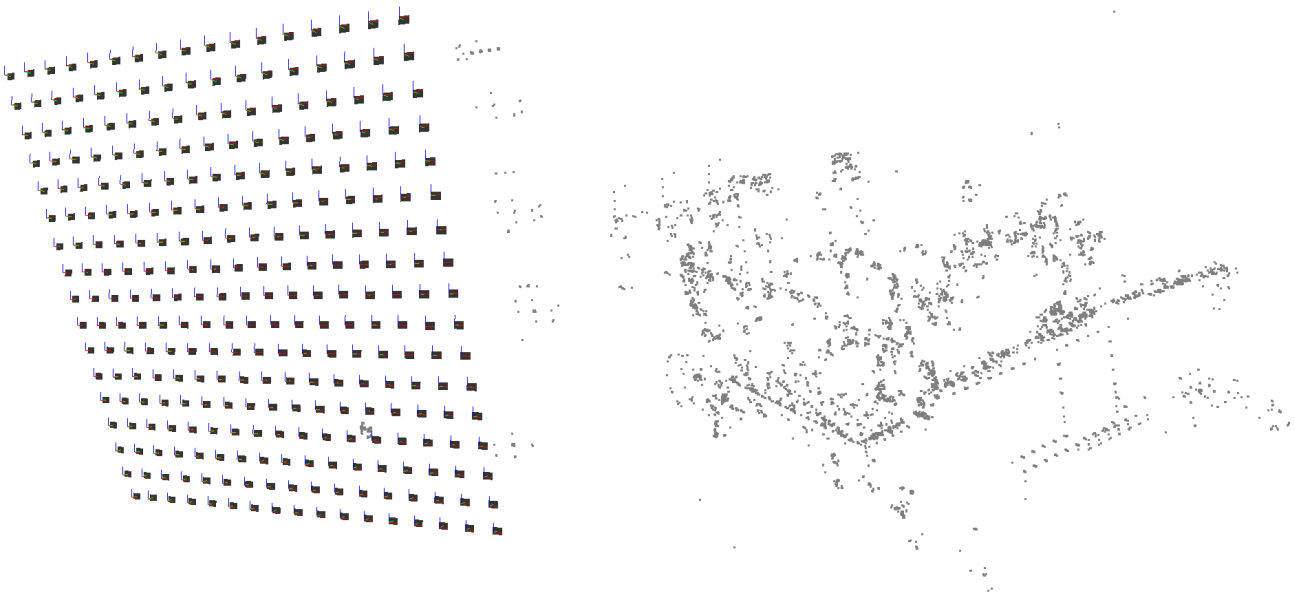

Figure 10. Reconstruction of the Stanford Light Field Archieve lego bulldozer scene.

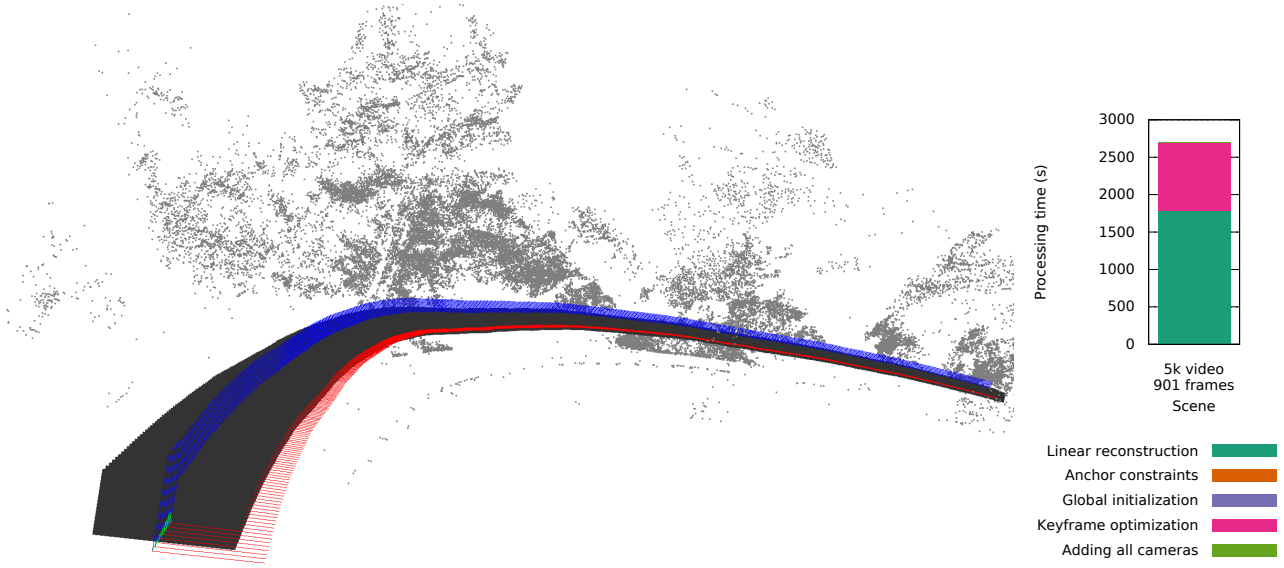

Figure 11. Reconstruction visualization and timings and reconstruction from a very high resolution (5k) video. Most time is spend on point subsampling during window BA and the final BA.

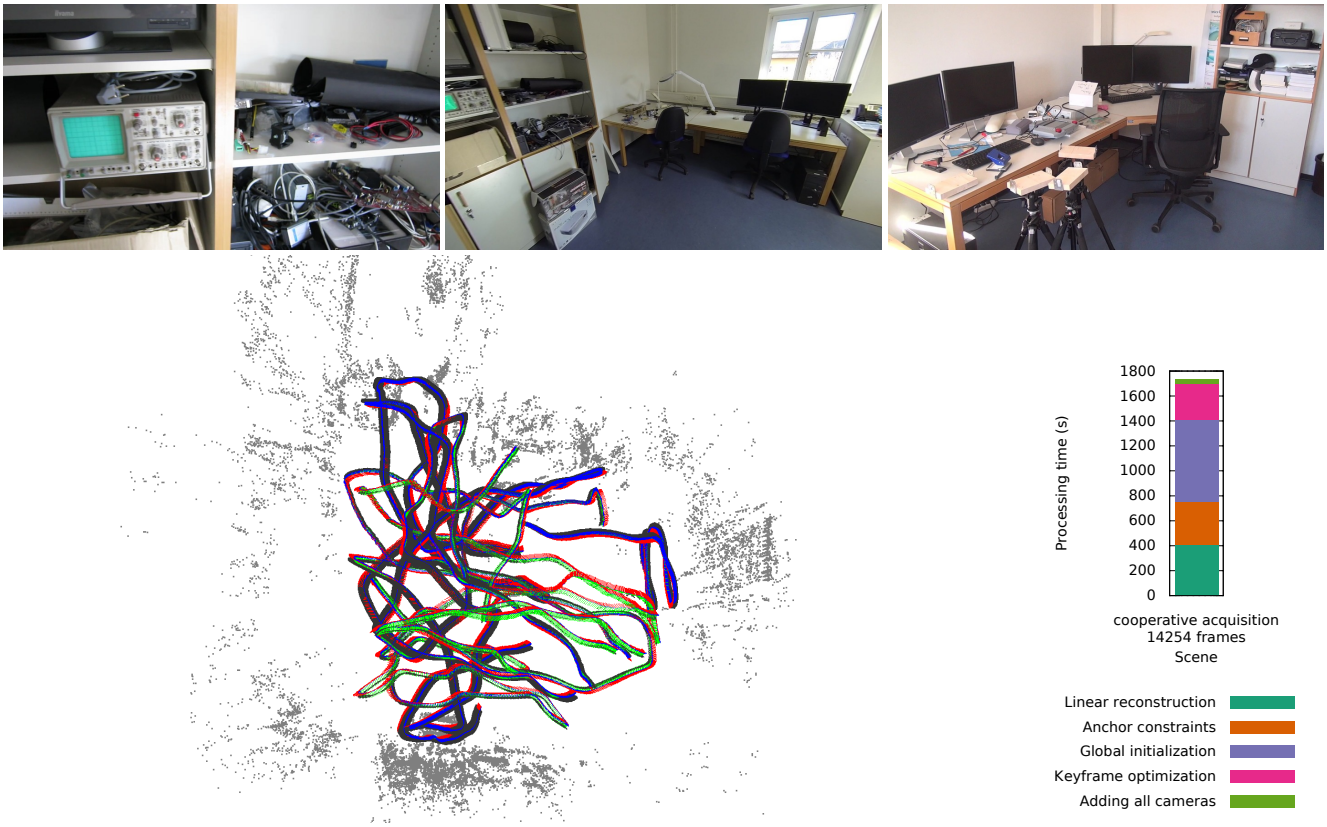

Figure 12. Example images, reconstruction visualization and timings from cooperatively captured video sequence set containing 14254 frames from three different cameras. Examples images from left to right: Canon S100, GoPro Hero3, Sony camcorder. All reconstructed camera trajectories are linked into a common global context by the anchor constraints.

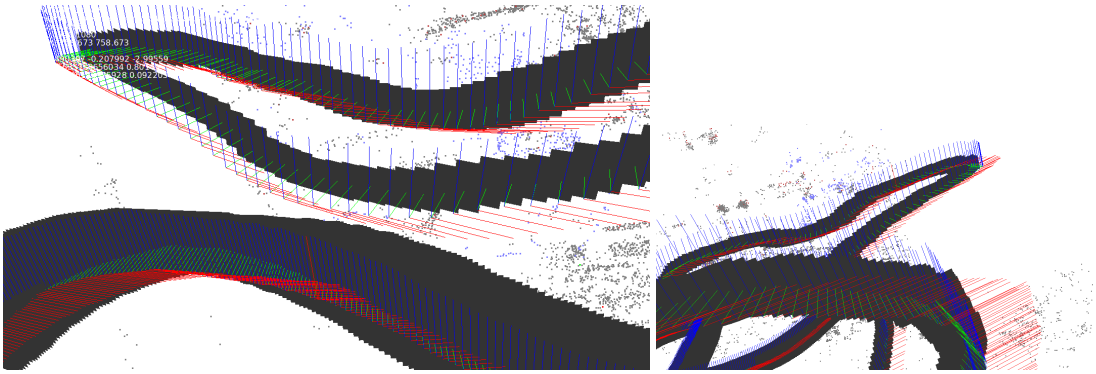

Figure 13. Effect of different framerates on camera trajectories. Left: Canon S100 (25 fps) vs. Gopro Hero 3 (60 fps). Right: Canon S100 (25 fps) vs. Sony camcorder (50 fps).