Real-time Image-based 6-DOF Localization in Large-Scale Environments

Hyon Lim[1] Sudipta N. Sinha[2] Michael F. Cohen[2] Matthew Uyttendaele[2]

[1] Seoul National University [2] Microsoft Research Redmond

SEOUL NATIONAL UNIVERSITY

Microsoft Research

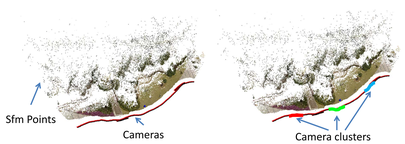
Contribution

We estimate precise 6-dof camera pose from monocular video given a SfM scene reconstruction computed offline. Our key contributions are:

- A new framework for accurate video-rate localization that consistently maintains low latency.
- An efficient 2d-3d direct feature matching approach interleaved with 2d keypoint tracking.

Our method scales to large scenes (with 100K+ 3D pts), is faster than keyframe recognition-based approaches [3], is more robust than SLAM methods and practical for onboard localization for MAVs.

Offline Map Construction

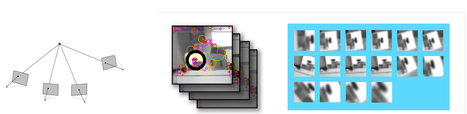


Structure from Motion

- Images → 3D points + calibrated cameras

Camera clustering

- Clustered based on visibility of SfM 3D points.
- Overlapping clusters used for coarse localization.



Multi-scale feature extraction

- Harris corners extracted at multiple scales.
- T2-8a-2r6s-32d DAISY descriptors [2], (32 dim) extracted, needed for 2d-3d feature matching.
- A bag of DAISY descriptors extracted from different images is associated with each 3D point.
- Global kd-tree index for all DAISY descriptors.
- Descriptors tagged with cluster + 3D point-id.
- Avoids online scale-invariant keypoint detection.

2D-3D Feature Matching

Coarse place recognition

- DAISY matching → cluster voting.
- Matched descriptors from images not in selected cluster, pruned → less outliers during RANSAC.

Global matching

- k-ANN query for a set of DAISY descriptors $\{q_i\}$.
- Priority search on single kd-tree (d = 32).
- $\forall q_i$, k descriptors $\{d_{ij}\}$ retrieved; sorted by distance $\{s_{ij}\}$ from q_i, d_{i0} is 1-NN.
- Matching strength for each d_{ij}, $m_{ij} = s_{i0}/s_{ij}$.
- All descriptors vote for clusters; votes weighted by strengths m_{ij}. Top-scoring cluster(s) selected.
- Prune $\{d_{ij}\}$ based on selected cluster(s).
- $\forall d_{ij}$, unique 3d point matches computed; Each point has a matching strength = $\sum m_{ij}$ for all retrieved descriptors associated with it.
- Ratio test performed for matching 3d points; One-to-one 2d-3d matches found for some q_i.
- Unlike common practice, 1-to-many matches also retained and handled during RANSAC.

Guided matching

- Incrementally update the selected *location cluster*.
- Exploit known 2d-3d matches (and pose) to accelerate priority search on kd-tree.
- Out-of-scope descriptors pruned during backtracking step in kdtree query. Lot faster than pruning afterwards.
- 1-to-many matches geometrically disambiguated using camera pose estimate in previous frame.

Keypoint Tracking

- Harris corners extracted (at a single scale).
- Binary 256-bit BRIEF [1] descriptors used.
- Tracking by re-detection and local matching.
- Hamming distances computation + ratio test.
- Tracks updated after 2d-3d pose estimation.

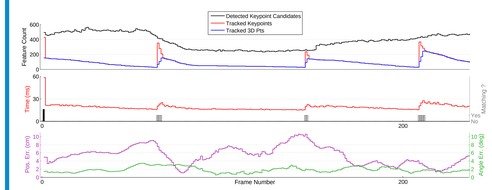
Online localization

Algorithm

$P \leftarrow$ COMPUTE-POSE (f, T, M)

KALMAN-FILTER-PREDICT ( )
$P \leftarrow \varnothing$
$K \leftarrow$ EXTRACT-KEYPOINTS (f)
$\eta \leftarrow$ TRACK-2D (f, K, T)
if $\eta < \kappa_1$ **then**
  ADD-GOOD-FEATURES (f, K, T)
end if
$C_1 \leftarrow$ FETCH-2D-3D-MATCHES-FROM-TABLE (T)
if $|C_1| > \kappa_2$ **then**
  $P \leftarrow$ ESTIMATE-POSE (C_1)
  if MATCHES-PENDING (T) **then**
    GUIDED-MATCHING (f, T, P, M)
  end if
else
  GLOBAL-MATCHING (f, T, M)
end if
$C_2 \leftarrow$ FETCH-2D-3D-MATCHES-FROM-TABLE (T)
if $|C_2| > \kappa_2 \wedge C_1 \neq C_2$ **then**
  $P \leftarrow$ ESTIMATE-POSE (C_2)
end if
KALMAN-FILTER-UPDATE (P)
return P

Real-time Statistics



Advantages

- Place recognition prunes outliers efficiently.
- BRIEF tracking efficiently propagates 2d-3d matches.
- Direct 2d-3d matching faster than keyframe matching.
- DAISY + kd-tree query computation minimized.
- Descriptor computation spread over multiple frames.

References

[1] M. Calonder, V. Lepetit, C. Strecha and P. Fua, "BRIEF: Binary Robust Independent Elementary Features", *ECCV*, 2010.

[2] S. Winder, G. Hua, M. Brown, "Picking the best DAISY", *CVPR*, 2009.

[3] Z. Dong, G. Zhang, J. Jia and H. Bao, "Keyframe-Based Real-Time Camera Tracking", *ICCV* 2009.

Results

Datasets

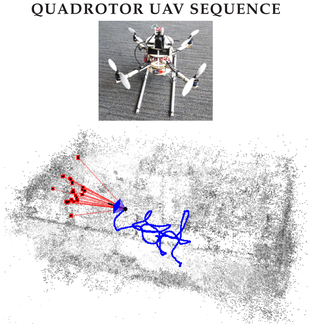
Name	size (m.)	Cams.	Pts.	D	L	Mem.
LAB	8 × 5	2111	76,560	1,019,253	450	124 MB
HALL	30 × 12	2749	88,248	1,377,785	253	111 MB
OUTDOOR1	from [11]	1448	120,313	1,241,045	188	117 MB
OUTDOOR2	from [11]	1011	26,484	1,282,227	126	107 MB

Test Sequences

Map	Sequence	F	F_{loc}	T (ms.)	M	T_M (ms.)
LAB	WALK1	237	237 (100%)	19 ± 4	10 (4%)	27 ± 11
LAB	WALK2	3793	3790 (99.9%)	18 ± 3	210 (6%)	23 ± 6
LAB	FLIGHT1	1000	1000 (100%)	17 ± 2	34 (3%)	22 ± 5
LAB	FLIGHT2	1210	1204 (99.5%)	17 ± 3	47 (4%)	23 ± 7
HALL	WALK1	475	475 (100%)	17 ± 3	27 (6%)	20 ± 7
HALL	WALK2	713	712 (100%)	17 ± 2	30 (4%)	20 ± 4
HALL	WALK3	540	540 (100%)	16 ± 1	33 (6%)	18 ± 3
HALL	WALK4	201	201 (100%)	16 ± 8	4 (2%)	24 ± 7
OUTDOOR1		1033	1033 (100%)	21 ± 4	169 (16%)	25 ± 6
OUTDOOR2		605	603 (99.6%)	27 ± 10	115 (19%)	40 ± 15

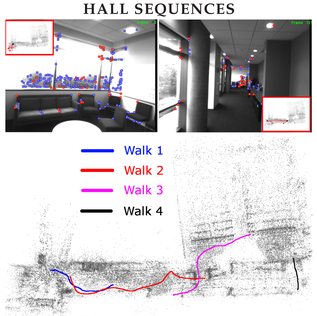
- Over 30Hz on 2.66Ghz Core2 Duo laptop.
- Runs at 12Hz on FitPC (1.6GHz Atom, 2GB RAM).
- 5× faster than [3] (single core) on OUTDOOR datasets.
- Position and orientation accuracy was within 5.1 cm. and 1.7 degrees in LAB scene.

Quadrotor UAV sequence



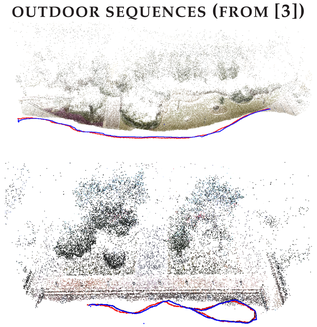
Runs at 12Hz (single core) on the compact FitPC computer onboard quadrotor MAV. 99% frames localized in the LAB scene.

Hall sequences



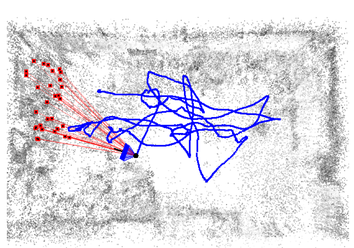
Larger HALL scene with ambiguous features, a narrow corridor and doorways between rooms. All frames were accurately localized.

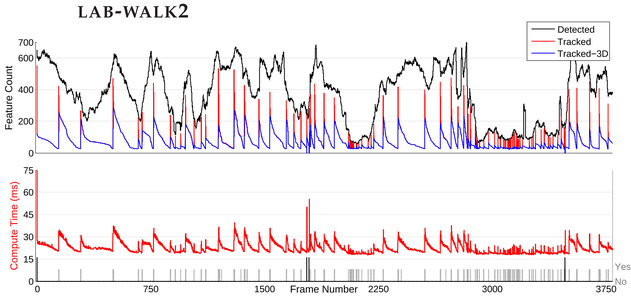
Outdoor sequences (from [3])



Camera trajectory from our method (in blue) coincides well with the trajectory from offline SfM on the same images (in red).

LAB-WALK2





- Camera consistently localized in 2+ minute sequence despite lower tracking efficiency in less textured areas.
- Twice tracking was lost, requiring relocalization.
- Plot shows all guided/global matching invocations.