面向增强现实的
单目视觉惯性SLAM算法评测

章国锋

浙江大学CAD&CG国家重点实验室

Augmented Reality

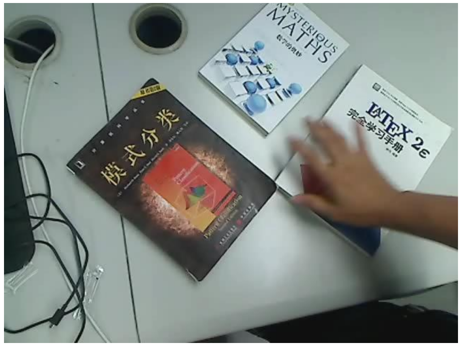
- Integrates digital information or virtual objects with the real environment in real time.
- Presents information more efficiently and intuitively than traditional text, images, and videos

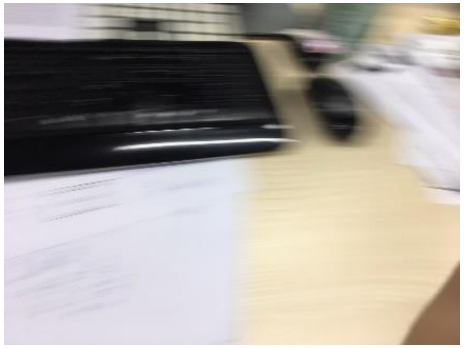
- Wide Applications
  - Education
  - Game
  - Advertising
  - E-commerce
  - Intelligent manufacture
  - Repair assembly
  - Medical

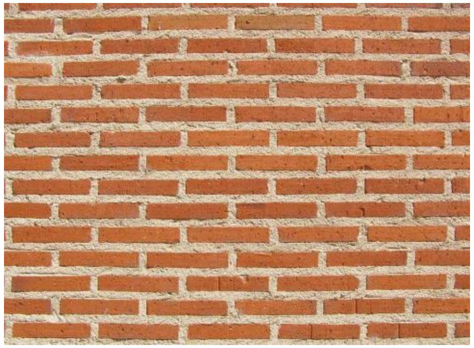
  …

Major Challenges

- Unexpected Situations in Applications
  - A home user may not carefully move the AR device.
  - Real environment may have moving objects, large textureless/repeated regions, and strong occlusions.
- Good User Experiences
  - Accurate and consistent 3D registration.
  - Low frequency of camera lost.
  - Quick recovery from failure status.







Visual-Inertial Dataset

- Typical VIO Dataset (e.g. EuRoC, TUM VI)
  - Synchronized sensors.
  - Global shutter cameras with high quality IMU.
- Mobile Phone Data
  - Sensor synchronization is not so reliable.
  - Rolling shutter camera with low-cost IMU.
- Not for evaluating real AR applications.


EuRoC


TUM VI


Real AR Application

Visual-Inertial Dataset

Comparison of commonly used VISLAM datasets

Dataset	KITTI	EuRoC	TUM VI	ADVIO
Hardware	Car	MAV	Custom Handheld	iPhone 6s
Camera	2×1392×512 10FPS Global Shutter	2×768×480 20FPS Global Shutter	2×1024×1024 20FPS Global Shutter	1×1280×720 60FPS RollingShutter
IMU	OXTS RT 3003 10Hz	ADIS 16488 200Hz	BMI160 200Hz	The IMU of iPhone 6s 100Hz
Ground- truth	OXTS RT 3003 10Hz	VICON/Leica 200Hz	OptiTrack 120Hz (Partially)	Sensor Fusion 100Hz
Environment	Outdoors	Indoors	In-/outdoors	In-/outdoors
Total Distance	39.2 km	0.9 km	20 km	4.5 km
Accuracy	~10 cm	~1 mm	~1 mm	~few dm
Sync	Software	Hardware	Hardware	Software

We need a more appropriate dataset for evaluating SLAM performance in AR applications, along with high accuracy ground-truth.

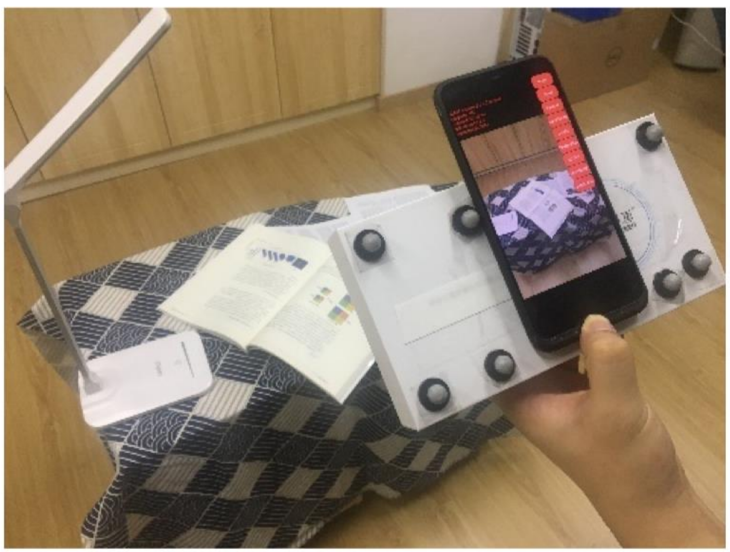
Visual-Inertial Dataset

Comparison of commonly used VISLAM datasets

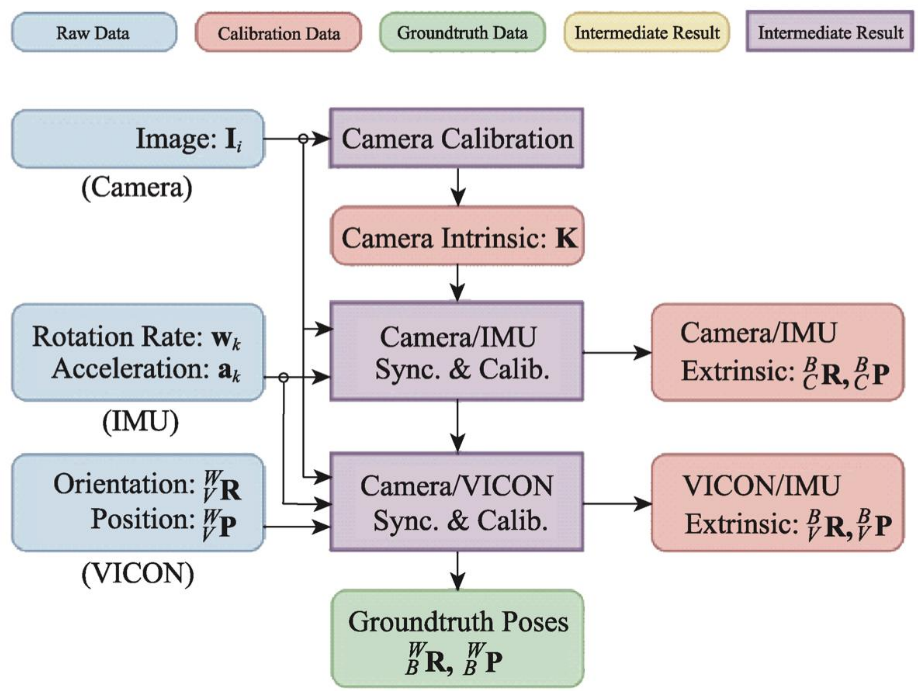
Dataset	KITTI	EuRoC	TUM VI	ADVIO	Ours
Hardware	Car	MAV	Custom Handheld	iPhone 6s	iPhone X/ Xiaomi Mi 8
Camera	2×1392×512 10FPS Global Shutter	2×768×480 20FPS Global Shutter	2×1024×1024 20FPS Global Shutter	1×1280×720 60FPS RollingShutter	1×640×480 30FPS RollingShutter
IMU	OXTS RT 3003 10Hz	ADIS 16488 200Hz	BMI160 200Hz	The IMU of iPhone 6s 100Hz	The IMU of iPhoneX/ The IMU of Xiaomi Mi 8 100Hz/400Hz
Ground- truth	OXTS RT 3003 10Hz	VICON/Leica 200Hz	OptiTrack 120Hz (Partially)	Sensor Fusion 100Hz	VICON 400Hz
Environment	Outdoors	Indoors	In-/outdoors	In-/outdoors	Indoors
Total Distance	39.2 km	0.9 km	20 km	4.5 km	377 m
Accuracy	~10 cm	~1 mm	~1 mm	~few dm	~1 mm
Sync	Software	Hardware	Hardware	Software	Software

Hardware Setup & Data Process

- Two different mobile phones
  - iPhone X (Camera 640x480 30fps, IMU 100Hz)
  - Xiaomi Mi 8 (Camera 640x480 30fps, IMU 400Hz)
- Ground-truth obtained by VICON system at 400Hz

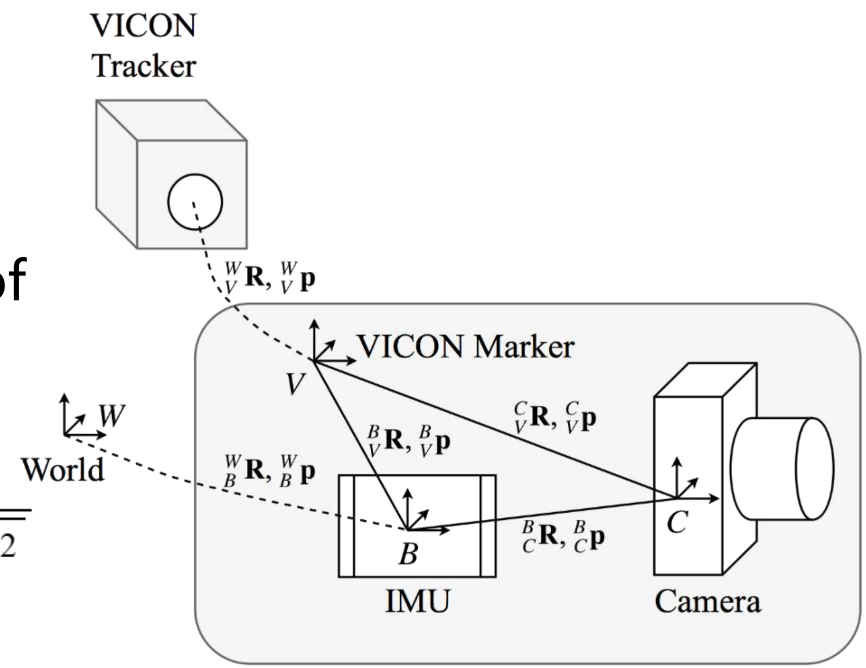


The phone is rigidly attached to a marker object for VICON localization



Device Synchronization and Calibration

- Camera-IMU Synchronization and Calibration
  - MATLAB Toolbox & Kalibr.



- VICON-IMU Synchronization
  - Maximizes the cross-correlation between VICON and IMU angle of rotations.

$$\underset{{}^{B}_{V}t}{\arg\max} \frac{\sum \left\|\theta_V({}_V t)\right\| \left\|\theta_B({}_V t + {}^{B}_{V}t)\right\|}{\sqrt{\sum \left\|\theta_V({}_V t)\right\|^2} \sqrt{\sum \left\|\theta_B({}_V t + {}^{B}_{V}t)\right\|^2}}$$

- VICON-Camera Calibration
  - Aligning the VICON measurements with the camera measurements by Apriltags.

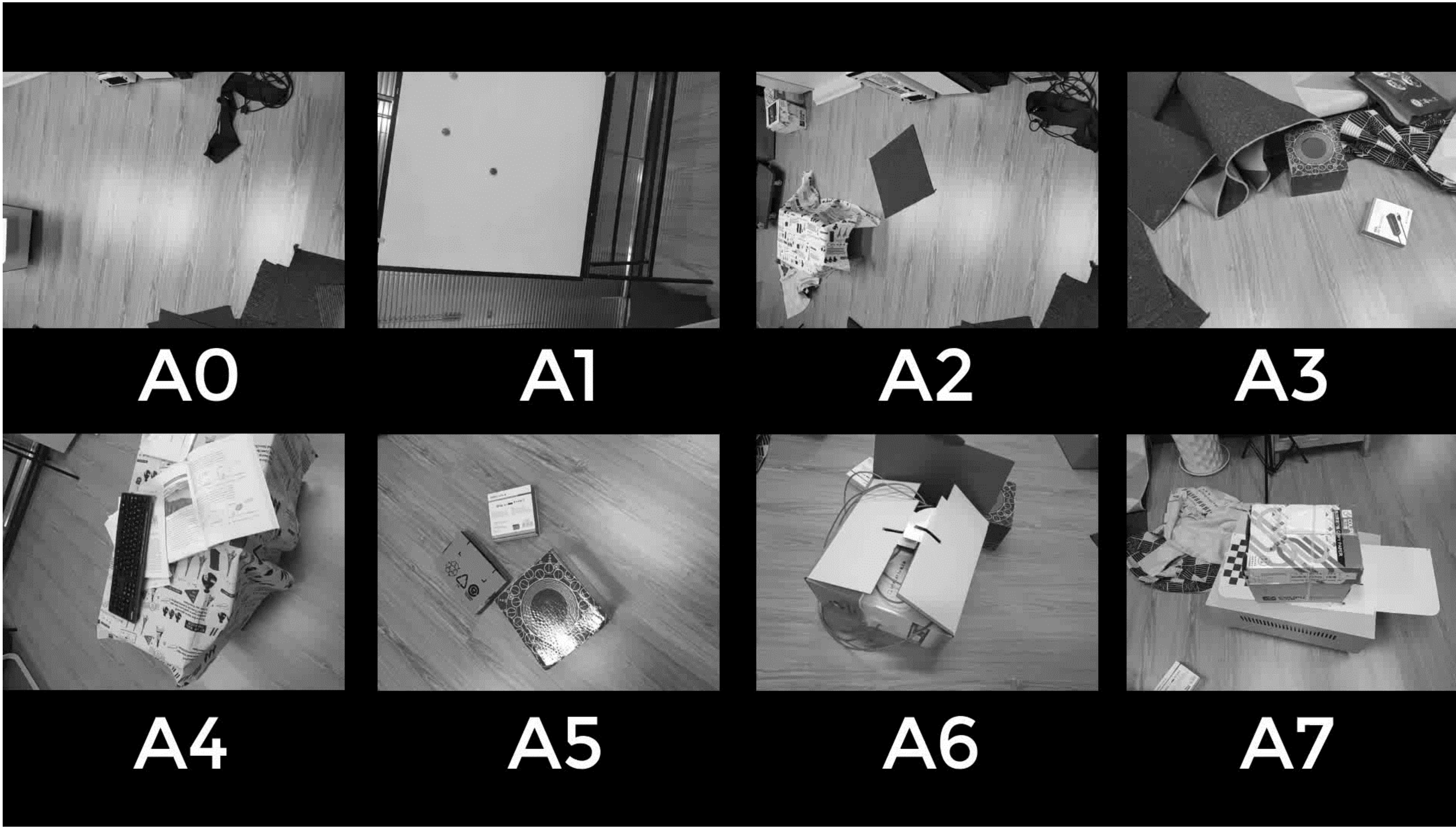
$$\underset{{}^{C}_{V}\mathbf{R},\, {}^{C}_{V}\mathbf{p},\, \{\mathbf{X}_i\}}{\arg\min} \sum_j \sum_i \left\| \pi\left({}^{C}_{V}\mathbf{R}\, {}^{W}_{V}\mathbf{R}^{\top} \left(\mathbf{X}_i - {}^{W}_{V}\mathbf{p}\right) + {}^{C}_{V}\mathbf{p} \right) - \mathbf{x}_{ij} \right\|^2$$

Dataset Motion and Scene Type

- 5 motion types : hold, wave, aiming, inspect, petrol
- 5 scene types : mess, clean, desktop, floor
- 3 segments : static, initialization, main
- B0~B7 are captured for evaluating dedicated criteria

Sequence		Motion	Scene	Description
Xiaomi	A0	inspect+patrol	floor	Walking and looking around the glossy floor.
	A1	inspect+patrol	clean	Walking around some texture-less areas.
	A2	inspect+patrol	mess	Walking around some random objects.
	A3	aiming+inspect	mess+floor	Random objects first, and then glossy floor.
	A4	aiming+inspect	desktop+clean	From a small scene to a texture-less area.
	A5	wave+inspect	desktop+mess	From a small scene to a texture-rich area.
	A6	hold+inspect	desktop	Looking at a small desktop scene.
	A7	inspect+aiming	desktop	Looking at a small desktop scene.
iPhone	B0	rapid-rotation	desktop	Rotating the phone rapidly at some time.
	B1	rapid-translation	desktop	Moving the phone rapidly at some time.
	B2	rapid-shaking	desktop	Shaking the phone violently at some time.
	B3	inspect	moving people	A person walks in and out.
	B4	inspect	covering camera	An object occasionally occluding the camera.
	B5	inspect	desktop	Similar to A6 but with black frames.
	B6	inspect	desktop	Similar to A6 but with black frames.
	B7	inspect	desktop	Similar to A6 but with black frames.

Dataset Preview



Evaluation Criteria

- Tracking Accuracy
- Initialization Quality
- Tracking Robustness
- Relocalization Time

Tracking Accuracy

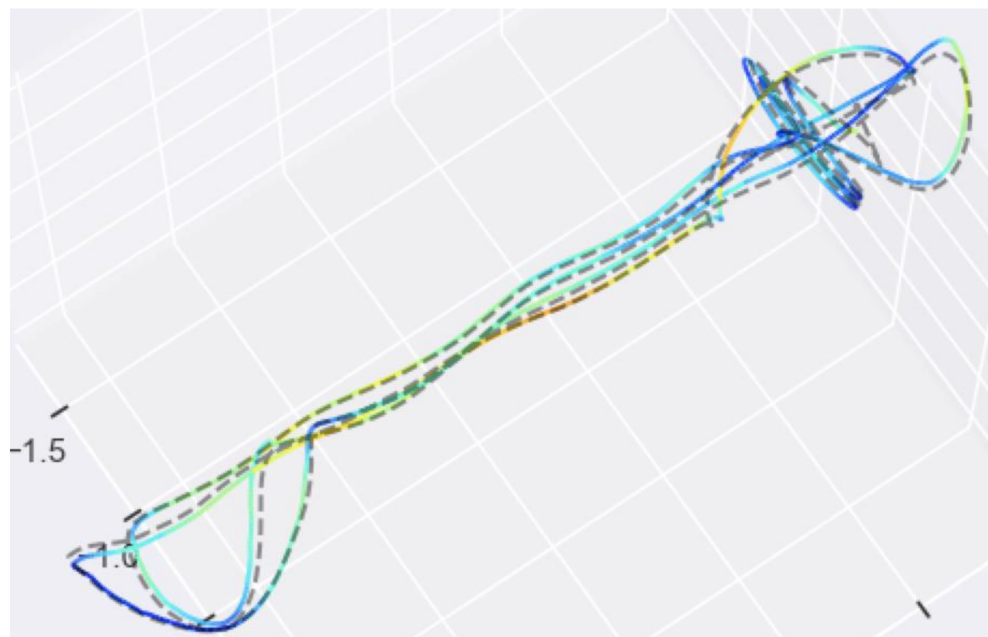
- 4 commonly used criteria:
  - Absolute Positional Error (APE) Relative Positional Error (RPE)

$$\epsilon_{\mathrm{APE}} = \sqrt{\frac{1}{m}\sum_{i=1}^{m} \| \mathbf{p}_{\mathrm{SLAM}}[i] - \mathbf{p}_{\mathrm{GT}}[i] \|^2}$$

$$\epsilon_{\mathrm{ARE}} = \sqrt{\frac{1}{m}\sum_{i=1}^{m} \| \log(\mathbf{R}^{-1}_{\mathrm{SLAM}}[i] \cdot \mathbf{R}_{\mathrm{GT}}[i]) \|^2}$$

  - Absolute Rotational Error (ARE) Relative Rotational Error (RRE)

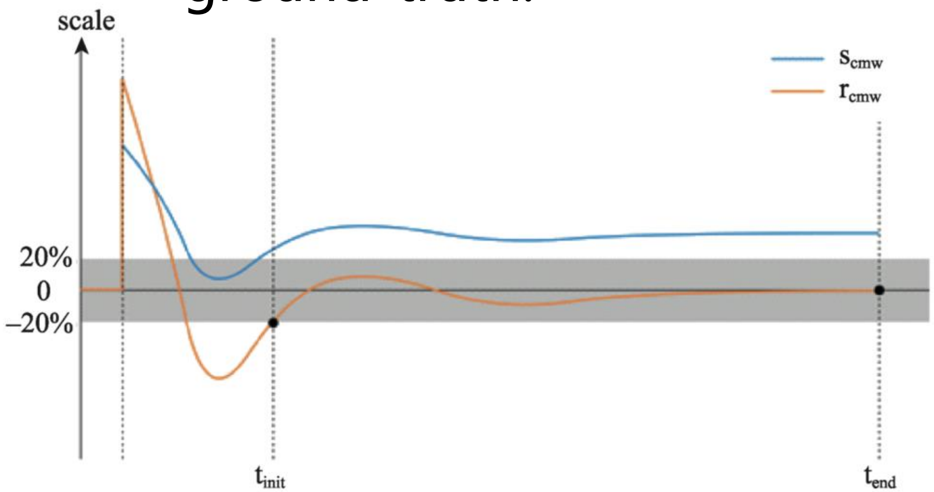
- Completeness
  - the ratio between the number of valid poses and the total number of all poses
  - The poses before the first initialization are not included.



A SLAM Trajectory APE Visualization in Seq. A0

Initialization Quality

- The time t_{init} for scale to converge.
  - Accurate scale is quite important in some AR applications.
  - At the beginning, scale usually fluctuates.
  - We generally insert AR objects after scale converges.
- The quality ϵ_{scale} of converged scale.
  - Key to some applications like AR ruler.
- For VSLAM, true scale is not available.
  - Estimate the global scale by aligning the results with ground-truth.



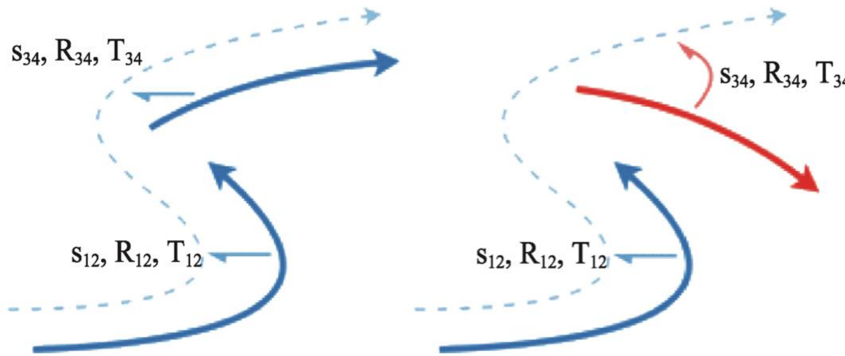
$$\epsilon_{\text{scale}} = \frac{1}{2}\left(\left|\frac{s_{\text{cmw}}(t_{\text{init}})}{s_{\text{g}}} - 1\right| + \left|\frac{s_{\text{g}}}{s_{\text{cmw}}(t_{\text{init}})} - 1\right|\right) \times 100\%$$

$$\epsilon_{\text{init}} = t_{\text{init}}(\epsilon_{\text{scale}} + \beta)^{\alpha}$$

Tracking Robustness

- Relocalization Error
  - The tracking result should be consistent after recovering from lost status.

$$\epsilon_{\mathrm{RL}} = \sum_{i=1}^{n-1} || \log_{\mathrm{Sim}(3)}(\xi_i^{-1}\xi_{i+1}) ||$$



- Lost time: the smaller, the better.
- Smaller tracking error is better.

$$\epsilon_{\mathrm{R}} = (\alpha_{\mathrm{lost}} + \eta_{\mathrm{lost}})(\epsilon_{\mathrm{RL}} + \eta_{\mathrm{APE}}\epsilon_{\mathrm{APE}})$$

Ratio of lost time

APE

Relocalization Time

- Force to enter lost state
  - Manually add black frames.
- Relocalization time measurement
  - VISLAM tends to continue IMU propagation even without sufficient feature matches.
  - Detect relocalization by the jump in trajectory.

$$t_{\mathrm{SLAM}[i]} \equiv \min\left\{t_k > t_{\mathrm{K}[i]} \mid \left\|\mathbf{p}_{\mathrm{SLAM}}[k+1] - \mathbf{p}_{\mathrm{SLAM}}[k]\right\| > \delta\right\}$$

Representative SLAM Systems

- Filtering-based SLAM
  - MonoSLAM : solve camera pose via extended Kalman filter.
  - MSCKF : keep a sliding window of M frames.
  - MSCKF 2.0 : use FEJ to avoid leaking errors.
- Optimization-based SLAM
  - PTAM : use keyframe-based optimization, local tracking and global mapping in two parallel threads.
  - ORB-SLAM2 : use ORB features to improve the system robustness.
  - OKVIS : use sliding-window optimization with both reprojection errors and IMU motion errors.
  - VINS-Mono : use local sliding-window optimization and global pose graph optimization.
- SLAM with Direct Tracking
  - LSD-SLAM, DSO : directly use intensity as measurements and minimize photometric error.

8 Selected VSLAM/VSLAM Systems

- VSLAM
  - PTAM : http://wiki.ros.org/ethzasl_ptam
  - ORB-SLAM2 : https://github.com/raulmur/ORB_SLAM2
  - LSD-SLAM : https://github.com/tum-vision/lsd_slam
  - DSO : https://github.com/JakobEngel/dso
- VISLAM
  - MSCKF : https://github.com/daniilidis-group/msckf_mono
  - OKVIS : https://github.com/ethz-asl/okvis
  - VINS-Mono : https://github.com/HKUST-Aerial-Robotics/VINS-Mono
  - SenseSLAM : http://www.zjucvg.net/senseslam

Experimental Results

- Part of VSLAM Tracking accuracy

APE/RPE (mm)

Sequence	PTAM		ORB-SLAM2		LSD-SLAM		DSO	
A0	**75.442**	6.696	96.777	**5.965**	105.963	11.761	231.860	10.456
A1	113.406	16.344	**95.379**	**10.285**	221.643	23.833	431.929	12.555
A2	**67.099**	6.833	69.486	5.706	310.963	8.156	216.893	**5.337**
A3	**10.913**	4.627	15.310	7.386	199.445	10.872	188.989	**4.294**
A4	21.007	4.773	**10.061**	**2.995**	155.692	10.756	115.477	4.595
A5	40.403	8.926	**29.653**	11.717	249.644	12.302	323.482	**7.978**
A6	19.483	3.051	**12.145**	6.741	49.805	3.018	14.864	**2.561**
A7	13.503	2.462	**5.832**	**1.557**	38.673	2.662	27.142	2.213

Experimental Results

- Part of VSLAM Tracking accuracy

Completeness (%)

Sequence	PTAM	ORB-SLAM2	LSD-SLAM	DSO
A0	**79.386**	65.175	49.513	14.476
A1	60.893	**68.303**	11.511	0.869
A2	**85.348**	79.263	21.804	22.878
A3	71.635	**98.497**	27.112	43.493
A4	95.418	**100.000**	64.283	80.371
A5	87.399	**97.785**	25.033	2.059
A6	97.399	99.786	94.883	**100.000**
A7	**100.000**	**100.000**	98.663	**100.000**

Experimental Results

- Part of VISLAM Tracking accuracy

APE/RPE (mm)

Sequence	MSCKF		OKVIS		VINS-Mono		SenseSLAM	
A0	156.018	7.436	71.677	7.064	160.334	2.798	**58.995**	**2.525**
A1	294.091	14.580	87.73	4.283	253.554	**2.723**	**55.097**	2.876
A2	102.657	10.151	68.381	5.412	102.263	1.976	**36.370**	**1.560**
A3	44.493	3.780	22.949	8.739	29.587	1.278	**17.792**	**0.779**
A4	114.845	8.338	146.89	12.46	37.580	1.042	**15.558**	**0.930**
A5	82.885	8.388	77.924	7.588	40.423	**1.660**	**34.810**	1.954
A6	66.001	6.761	63.895	6.86	80.062	1.404	**20.467**	**0.569**
A7	105.492	4.576	47.465	6.352	25.082	1.138	**10.777**	**0.831**

Experimental Results

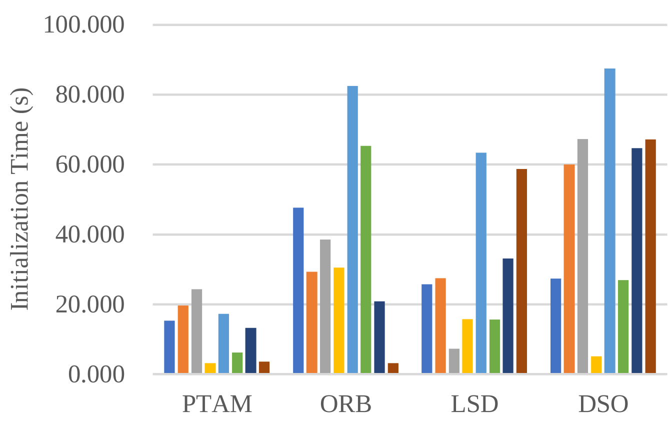
- Part of VISLAM Tracking accuracy

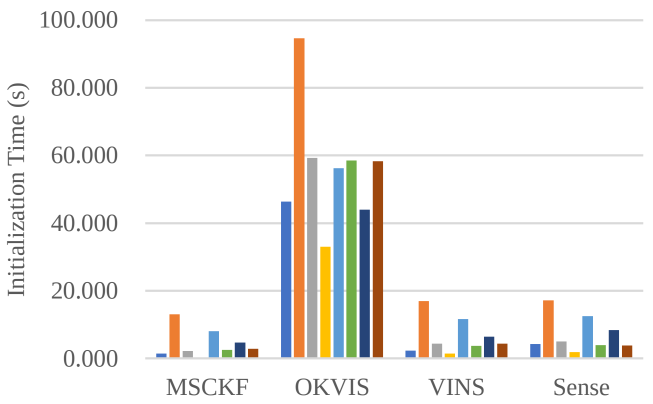
Completeness (%)

Sequence	MSCKF	OKVIS	VINS-Mono	SenseSLAM
A0	40.186	94.255	35.256	**97.317**
A1	1.646	**98.235**	17.902	95.072
A2	61.423	94.959	63.449	**99.707**
A3	97.814	95.972	**100.000**	**100.000**
A4	76.629	97.429	**100.000**	**100.000**
A5	76.738	98.162	**99.866**	99.143
A6	94.128	97.805	81.763	**100.000**
A7	68.341	96.69	**100.000**	**100.000**

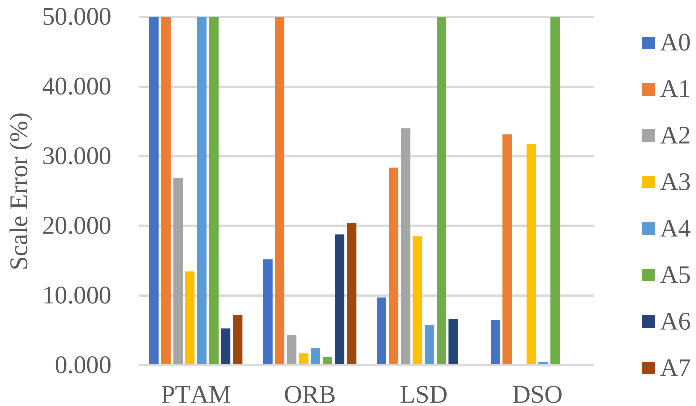
Experimental Results

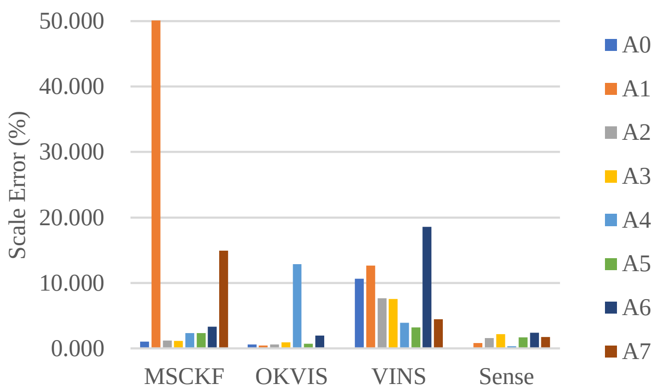
- Initialization Time





- Initialization Scale





Experimental Results

- Initialization Quality $\epsilon_{\text{init}} = t_{\text{init}}(\epsilon_{\text{scale}} + \beta)^{\alpha}$

Sequence	PTAM	ORB SLAM2	LSD SLAM	DSO	MSCKF	OKVIS	VINS Mono	Sense SLAM
A0	41.387	19.172	8.423	**7.460**	**0.211**	5.913	0.783	0.449
A1	22.265	28.141	**14.877**	35.062	49.660	11.487	6.265	**2.324**
A2	12.828	8.913	**4.311**	6.837	**0.331**	7.506	1.300	0.804
A3	**1.193**	5.009	6.960	2.920	**0.035**	4.607	0.441	0.340
A4	16.725	15.324	16.478	**10.450**	1.497	20.964	2.584	**1.456**
A5	14.223	**9.512**	41.941	28.801	**0.463**	7.732	0.763	0.652
A6	**3.322**	9.275	9.158	6.550	**0.991**	7.613	2.857	1.553
A7	**1.027**	1.458	6.176	6.766	1.159	6.265	1.026	**0.650**
Average	14.121	**12.101**	13.541	13.106	6.793	9.011	2.002	**1.029**
Max	41.387	**28.141**	41.941	35.062	49.660	20.964	6.265	**2.324**

Experimental Results

- Tracking Robustness

Sequence	PTAM	ORB SLAM2	LSD SLAM	DSO	MSCKF	OKVIS	VINS Mono	Sense SLAM
B0 (Rapid Rotation)	16.088	3.396	2.068	1.848	---	5.328	16.774	**0.511**
B1 (Rapid Translation)	26.887	7.128	12.739	16.127	---	**5.448**	9.024	7.199
B2 (Rapid Shaking)	36.140	**3.875**	12.476	---	---	24.024	18.062	9.743
B3 (Moving People)	12.779	16.670	22.882	41.294	---	1.636	16.741	**1.089**
B4 (Covering Camera)	20.062	8.265	17.368	---	3.119	13.051	18.619	**1.192**

Experimental Results

- Relocalization Time

Sequence	PTAM	ORB SLAM2	LSD SLAM	VINS-Mono	SenseSLAM
B5 (1s black-out)	1.032	**0.077**	1.082	1.452	0.592
B6 (2s black-out)	**0.366**	0.465	5.413	1.833	1.567
B7 (3s black-out)	0.651	**0.118**	1.834	0.841	0.332
Average	0.683	**0.220**	2.776	1.375	0.830

Disscusion & Conclusion

- Contributions
  - The first public VISLAM benchmark for AR
    - Visual-inertial dataset
    - Evaluation criteria & toolkit

    http://www.zjucvg.net/eval-vislam/dataset/

    https://github.com/zju3dv/eval-vislam
  - Quantitative evaluation for 8 representative systems.
- Future Work
  - Better evaluation on mobile phones.
  - Capture more diverse sequences in a larger outdoor environment.

Jinyu Li, Bangbang Yang, Danpeng Chen, Nan Wang, Guofeng Zhang*, Hujun Bao*. Survey and Evaluation of Monocular Visual-Inertial SLAM Algorithms for Augmented Reality. Journal of Virtual Reality & Intelligent Hardware, published online: http://www.vr-ih.com/vrih/html/EN/10.3724/SP.J.2096-5796.2018.0011/article.html

VRIH Virtual Reality & Intelligent Hardware 虚 拟 现 实 与 智 能 硬 件
科 学 出 版 社 Science Press

虚拟现实与智能硬件（VRIH）

Virtual Reality & Intelligent Hardware

Virtual Reality&Intelligent Hardware 虚拟现实与智能硬件

科学出版社 Science Press

主管单位：中国科学院

主办单位：中国科技出版传媒股份有限公司（科学出版社）
北京航空航天大学

出版单位：北京中科期刊出版有限公司

协办单位：歌尔集团有限公司

海外出版：ScienceDirect (Elsevier)

主要报道学科方向

- 虚拟现实/增强现实
- 低功耗轻量级底层软硬件
- 高性能智能感知
- 高精度运动与姿态控制
- 低功耗广域智能物联
- 端云一体化协同

- ✓ 综述
- ✓ 研究论文
- ✓ 研究快报
- ✓ 案例报道
- ✓ 评述
- ✓ 稿酬从优

Virtual Reality&Intelligent Hardware
虚 拟 现 实 与 智 能 硬 件

科 学 出 版 社
Science Press

- 创刊号（2019年2月已出版）　知名专家撰稿
- 2019年第2—5期　分支学科专刊
- 论文（HTML+PDF）　网刊，SciEngine，ScienceDirect

编辑：祁媛、陈睿超
电话：010-64010640
E-mail：vrih@vip.163.com
网刊：http://www.vr-ih.com

Thank you!