ASPEN: Acceleration of Visual-Inertial Odometry for Extended Reality on an FPGA

Kathleen Feng

Stanford University

31 August 2023

Extended Reality

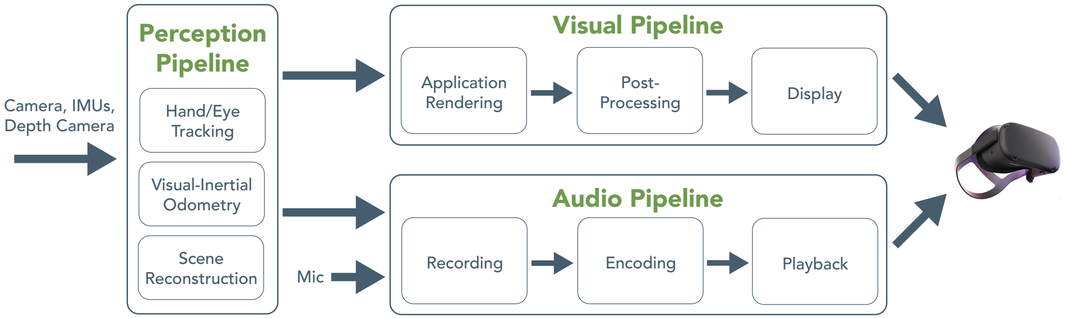
Augmented Reality (AR): overlay of digital information on real world, supplements reality

Virtual Reality (VR): full immersion into computer-generated environment, opaque displays and sensory input

Mixed Reality (MR): manipulation of both physical and virtual items, interaction with real world and virtual environment

Extended Reality (XR): umbrella term for AR+VR+MR

Extended Reality Pipeline



Constraints for an Ideal System

Three axes to evaluate XR system:

Constraints for an Ideal System

Three axes to evaluate XR system:

1. Performance:
    - Motion-to-photon (MTP) latency: under 20 ms (VR), 7 ms (AR)
    - Throughput: ideally 100 gigapixels/s

Constraints for an Ideal System

Three axes to evaluate XR system:

1. Performance:
    - Motion-to-photon (MTP) latency: under 20 ms (VR), 7 ms (AR)
    - Throughput: ideally 100 gigapixels/s
2. Power:
    - Power budget: under 1-2 W (VR), 0.1-0.2 W (AR)

Constraints for an Ideal System

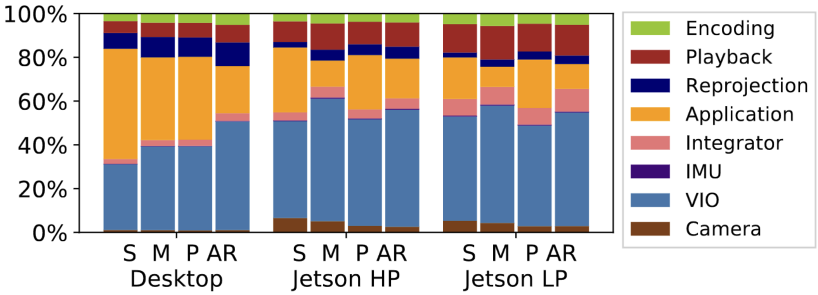
Three axes to evaluate XR system:

1. Performance:
   - Motion-to-photon (MTP) latency: under 20 ms (VR), 7 ms (AR)
   - Throughput: ideally 100 gigapixels/s
2. Power:
   - Power budget: under 1-2 W (VR), 0.1-0.2 W (AR)
3. Quality:
   - Accuracy in subtasks: visual-inertial odometry, eye gaze estimation
   - User experience: interaction, responsiveness,...

Current State of Commerical XR Systems

	Ideal	Varjo VR-3	Quest 2	HoloLens 2	Quest Pro
Usage	–	VR	VR	MR	MR
Resolution (MPixels)	400	15.7	7.0	4.4	6.9
Refresh Rate (Hz)	240	90	90	120	90
MTP latency (ms)	$<$ 20 (VR) $<$ 7 (AR)	$<$ 20	N/A	$<$ 9	N/A
Power (W)	$<$ 1-2 (VR) $<$ 0.1-0.2 (AR)	N/A	N/A	$>$ 7	N/A
Mass (g)	100-200	944	503	566	722

CPU Processing Time Breakdown

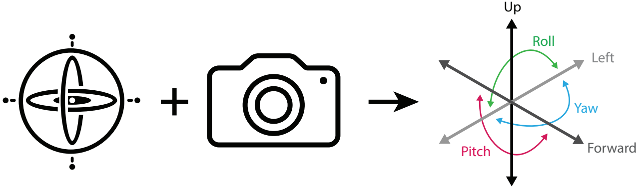


Source: ILLIXR [Huzaifa 2021]

VIO: Visual-Inertial Odometry

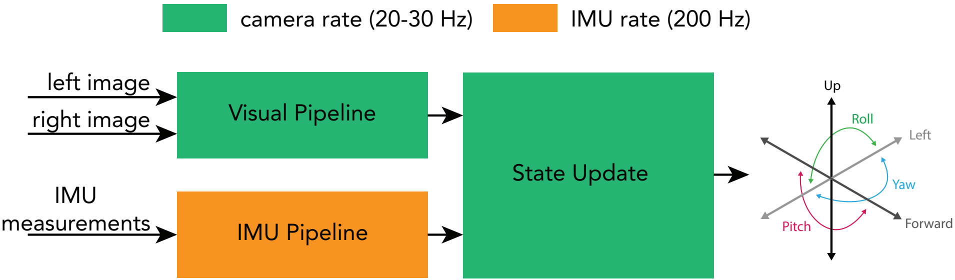
Calculates 3D user position from sensors

- IMUs, cameras $\Rightarrow (x, y, z, \theta, \phi, \psi)$
- Most dominating subtask, represents $\sim$ 40% of XR workload



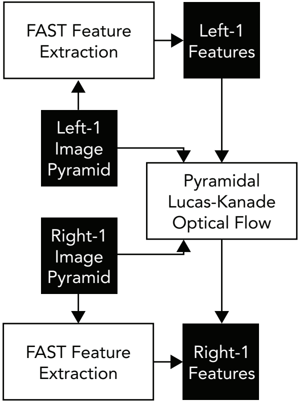
Using OpenVINS as gold model [https://docs.openvins.com/index.html]

VIO: Visual-Inertial Odometry



Visual Pipeline

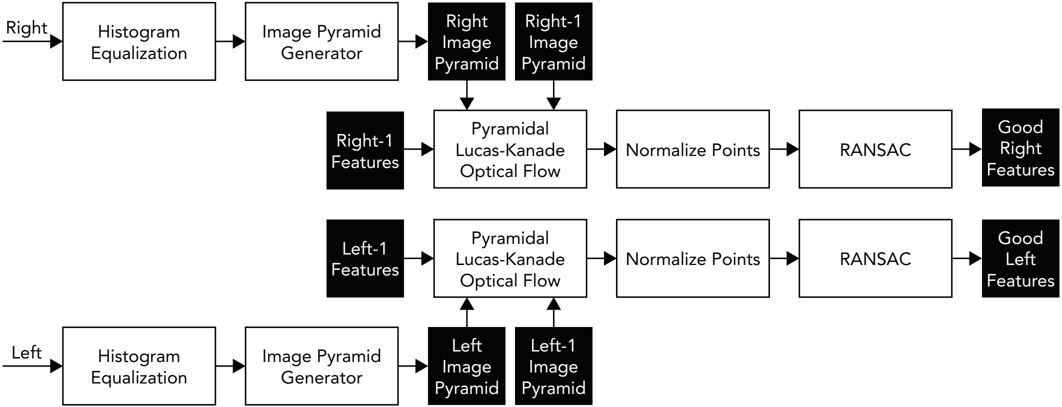
1. Extract features (if needed) from previous frame
2. Track features between previous and current frame



(1)

Visual Pipeline

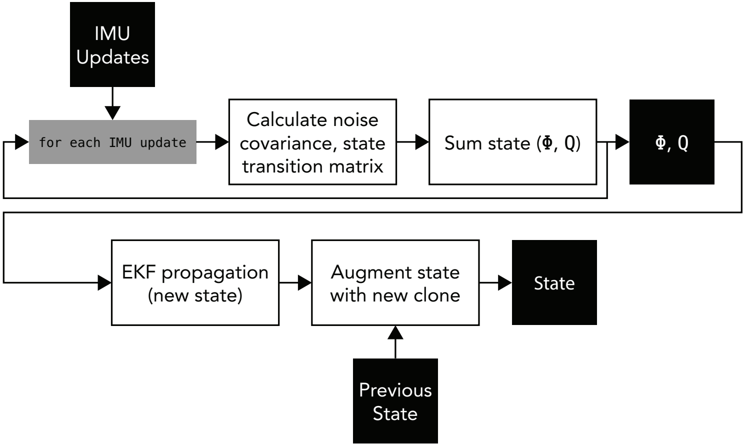
1. Extract features (if needed) from previous frame
2. Track features between previous and current frame



(2)

IMU Pipeline

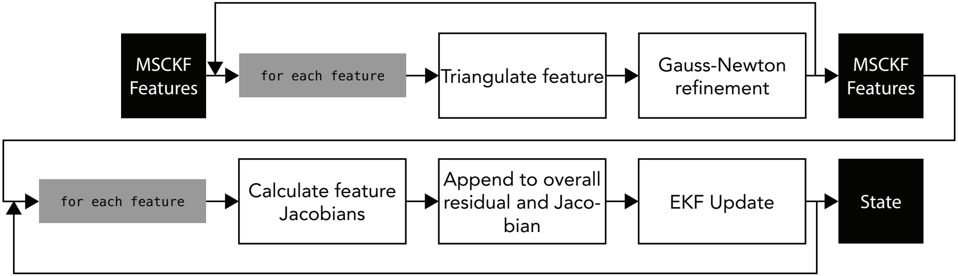
Provides fast state updates between camera frames



State Update

Features two update paths for two sets of features

1. MSCKF features: newer, freshly extracted features
2. SLAM features: features that have appeared in several consecutive frames

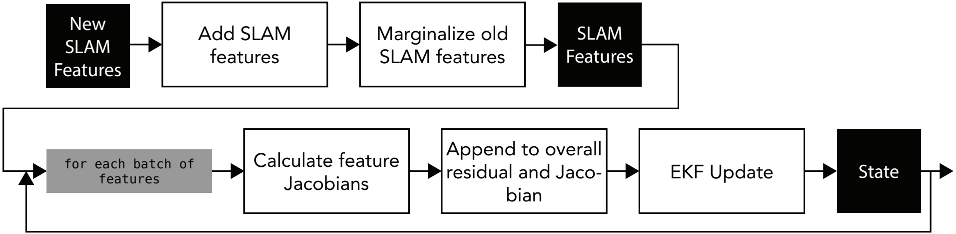


(1)

State Update

Features two update paths for two sets of features

1. MSCKF features: newer, freshly extracted features
2. SLAM features: features that have appeared in several consecutive frames

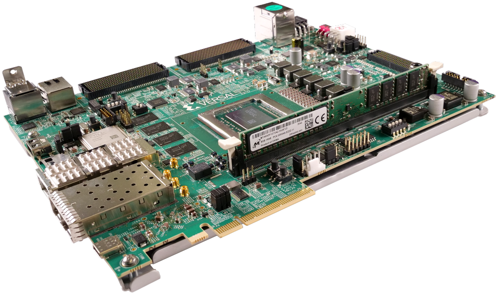


(2)

FPGA Acceleration

Currently prototyping accelerator on Xilinx Versal VCK190 FPGA

- Using Vitis HLS to implement vision and IMU pipelines on reconfigurable fabric
- On-board application processor performing all other computation
- In the process of testing end-to-end application on FPGA



Visual Pipeline Optimizations

- Switched almost everything to integer
- Absolute trajectory error: difference between the estimated trajectory and groundtruth after it has been aligned

$$e_{ATE} = \sqrt{\frac{1}{K}\sum_{k=1}^{K} ||\mathrm{x}_{k,i} \boxminus \hat{\mathrm{x}}^{+}_{k,i}||_2^2}$$

	Orientation Error (Θ)	Position Error (m)
TUM VI Room 1 (orig)	1.451	0.056
TUM VI Room 1 (int)	1.854	0.054
TUM VI Room 2 (orig)	1.448	0.071
TUM VI Room 2 (int)	1.422	0.074
TUM VI Room 3 (orig)	1.354	0.061
TUM VI Room 3 (int)	1.428	0.076

FPGA Results

For a 512×512 input image

	Vision	IMU
Clock Period	12 ns	8 ns
BRAM	243 (12%)	16 (∼0%)
DSP	1559 (79%)	756 (38%)
FF	230272 (12%)	64032 (3%)
LUT	553169 (61%)	99706 (11%)
URAM	21 (4%)	0 (0%)

Visual Pipeline Preliminary Latency per Kernel

For a 512×512 input image

- ✓ Functionally correct
- × Optimization in progress

	Clock (ns)	Cycles	Latency (ms)
Histogram Equalization	5	771630	3.86
FAST corner detection	12	1202358	14.4
Pyramid Generation + Optical Flow	12	10587926	127
Undistortion	5	43249	0.216
RANSAC	5	2915580	14.6
Estimated Total			~290

IMU Pipeline Preliminary Latency

Batching 11 IMU measurements together

- ✓ Functionally correct
- ✓ Optimized

	Clock (ns)	Cycles	Latency (ms)
IMU pipeline	5.3	1285844	6.82

Future Plans

1. Continue to improve on FPGA prototype
2. Develop and tapeout SoC for accelerating visual-inertial odometry and rest of perception pipeline in November
3. Evaluate performance/power/quality benefit of accelerator in context of entire XR application pipeline
4. Research further into the other pipelines (visual and audio) and integrate with the perception pipeline accelerator