3.7 A 1920×1080 30fps 2.3TOPS/W Stereo-Depth Processor for Robust Autonomous Navigation

Ziyun Li, Qing Dong, Mehdi Saligane, Benjamin Kempke, Shijia Yang, Zhengya Zhang, Ronald Dreslinski, Dennis Sylvester, David Blaauw, Hun Seok Kim

University of Michigan, Ann Arbor, MI

Precise depth estimation is a key kernel function to realizing autonomous navigation on micro-aerial vehicles (MAVs). The state-of-the-art semi-global matching (SGM) algorithm has become favored for its high accuracy. In particular, it effectively handles low texture regions due to its global optimization of the disparity between a left and right image over the entire frame. However, SGM involves massively parallel computation (~2TOP/s) and extremely high bandwidth memory access (38.6Tb/s) for 30fps HD resolution. This leads to ~20s runtime for an HD image pair on a 3GHz CPU [1] requiring ~386MB memory and >35W power consumption. Together, these factors place it well outside the realm of MAVs. Prior ASIC implementations have used either simpler local methods [2] or aggressively truncated global algorithms [3] that produce a depth map with significantly inferior quality or limited disparity range (32 or 64 pixels) and therefore fail to support standard automotive scene benchmarks [2-5]. In addition, due to the high memory requirement of SGM, prior methods [3-4] have used external DRAM to store intermediate computation, significantly reducing performance and efficiency.

This paper presents a stereo vision processor that fully implements the SGM algorithm on a single chip. The design uses a new image-scanning stride to enable a deeply pipelined implementation with ultra-wide (1612b) custom SRAM for 1.64Tb/s on-chip access bandwidth. Our design is the first ASIC to report performance under the industrial standard KITTI benchmark that renders realistic automobile scenes. The proposed design supports 512-level depth resolution on full HD (1920×1080) resolution with real-time 30fps, consuming 836mW from a 0.75V supply in 40nm CMOS. We also integrate the stereo chip with a quadcopter and demonstrate its operation in real-time flight.

Figure 3.7.1 (top right) visualizes the output difference between the local sum of absolute difference (SAD) algorithm and SGM, clearly illustrating the higher quality of SGM. To make a single-chip SGM implementation feasible, i.e. remove the need for external DRAM, we first observe that inter-pixel correlation diminishes when pixel pairs are more than 50 pixels apart. Hence, the proposed design processes the input image in units of 50×50 overlapping pixel blocks. Adjacent blocks are overlapped by 8 pixels to allow cost aggregation across block boundaries. This technique reduces the memory requirement for storing intermediate aggregation results by 95.4%. Fig. 3.7.1 shows a side-by-side comparison of this block-based SGM and the original SGM, which are almost identical. Fig. 3.7.1 also presents quantitative results evaluated on 194 KITTI test cases showing only 0.5% accuracy degradation.

As shown in Fig. 3.7.2, the processor streams left and right image blocks into two on-chip interleaved image buffers (30Kb each). It then performs a 7×7 census transformation on each pixel using its surrounding pixels and compares each census-transformed pixel on the left image with census-transformed pixels on the right image at 128 different disparity locations. This produces 128 Hamming distances (6b each) for each pixel that represent the 'local' matching cost for the 128 disparities. The processor then aggregates the local matching costs (separately for each disparity) along 8 paths over the 50×50 block. By searching the sum of the aggregated cost for each disparity for the minimum value, the processor obtains the coarse (integer) SGM depth output. It then refines this depth precision by performing a quadratic fitting on three aggregated costs around the minimum using a look-up table to provide sub-pixel depth accuracy.

Conventionally, SGM is implemented with a forward and a backward raster scan, with each scan performing aggregation along 4 paths (total of 8 paths). However, following this conventional raster scan order results in a data dependency where the previous pixel must complete its computation before the current pixel can be aggregated (Fig. 3.7.3, left). This dependency dominates the critical path, limiting the clock frequency and voltage scalability for low power operation. We therefore propose a dependency-resolving scan in which pixel processing proceeds diagonally (Fig. 3.7.3, right). When a pixel (F) is fetched into the pipeline, the aggregated costs of all previous pixels (light gray and dark gray) are already computed and stored in high-bandwidth custom SRAMs. This mechanism enables aggressive pipelining, yielding a 3× performance gain. As shown in the block diagram in Fig. 3.7.4 (top), our design leverages parallelism in cost aggregation by running 4 paths in parallel on 4 aggregation units, with each aggregation unit containing 128 processing elements and 512 selection units, resulting in 1.882TOP/s.

Figure 3.7.4 (bottom) shows the proposed architecture of the customized compact high-bandwidth SRAM. In the proposed design, the row buffers are read and written simultaneously at 170MHz, and all 128 previous aggregated costs are accessed in a single cycle. This approach achieves the required memory bandwidth of 1.64Tb/s for the 3 row buffers accessed in parallel. This bandwidth would incur large chip area and power overhead if realized with compiled SRAMs. To provide an efficient area/power solution, we use a custom high-bandwidth SRAM that leverages the design's highly parallelized structure in which each bank has only 50 words with a single word size of 403b (Fig. 3.7.4). All four banks in one SRAM are read and written concurrently, realizing a 1612b dual port access. To reduce leakage power in the 40nm technology, the custom 8T memory bitcell uses HVT transistors. Unlike conventional 8T cells, the read transistor stack is flipped such that the read transistor is not connected to RBL, reducing coupling between RWL and the short, low capacitance RBL. Skewed inverters are used in place of conventional sense amplifiers (1612 per SRAM), reducing sense amp overhead by 2.8×. Overall, each 80Kb SRAM consumes 6mW with 548.1Gb/s bandwidth.

The vision processor is fabricated in 40nm GP CMOS. Fig. 3.7.5 shows the measurement setup and real-time demonstration platform mounted on a quadcopter. Real-time image streams captured by the stereo camera are rectified, block-partitioned by a Samsung Exynos-5422 processor on the ODRIOD-XU4 board, and then transmitted to the stereo processor through a USB3.0 interface. The processed real-time depth and confidence maps provide feedback to the Exynos processor through another USB3.0 channel. At 0.9V nominal voltage, the real-time VGA (HD) frame processing latency of the stereo processor is 4.1ms (26ms). In a KITTI automobile scene (Fig. 3.7.5, bottom) and a quadcopter scene 'on-the-fly' captured by our demonstration platform (Fig. 3.7.5, middle), large (>100 pixels) disparity frequently occurs, and the proposed processor is able to generate an accurate depth map over the entire image due to its 512 levels of resolution. Fig. 3.7.6 shows the voltage and frequency scaling of the chip and provides comparison with prior work. The proposed processor achieves 7% outlier accuracy on KITTI and an 8× improvement in disparity range compared with [2-5]. Note that [2-5] all lack a standard benchmark evaluation because of their limited depth range. Our system consumes 836mW to process 30fps full HD images at 0.0262nJ 'normalized energy' (an FoM proposed in [4] and defined in Fig. 3.7.6 top, left), marking a 5.8× improvement over listed prior work. Power reduces to 55mW for VGA images at 30fps, yielding 0.0117nJ normalized energy. Fig. 3.7.7 shows the die photo and a performance summary.

Acknowledgements:
We thank TSMC University Shuttle Program for chip fabrication.

References:

[1] H. Hirschmuller, et al., "Accurate and Efficient Stereo Processing by Semi-Global Matching and Mutual," *Computer Vision and Pattern Recognition*, pp. 807-814, 2005.

[2] M. Hariyama, et al., "VLSI Processor For Reliable Stereo Matching Based on Window-Parallel Logic-In-Memory Architecture," *IEEE Symp. VLSI Circuits*, pp. 166-169, 2004.

[3] K. Lee, et al., "A 502GOPS and 0.984mW Dual-Mode ADAS SoC with RNN-FIS Engine for Intention Prediction in Automotive Black-Box System," *ISSCC*, pp. 256-257, 2016.

[4] H-H. Chen, et al., "A 1920×1080 30fps 611mW Five-View Depth-Estimation Processor for Light-Field Applications," *ISSCC*, pp. 422-423, 2015.

[5] J. Park, et al., "A 30fps Stereo Matching Processor Based on Belief Propagation with Disparity-Parallel PE Array Architecture," *ISCAS*, pp. 453-454, 2010.

 978-1-5090-3758-2/17/$31.00

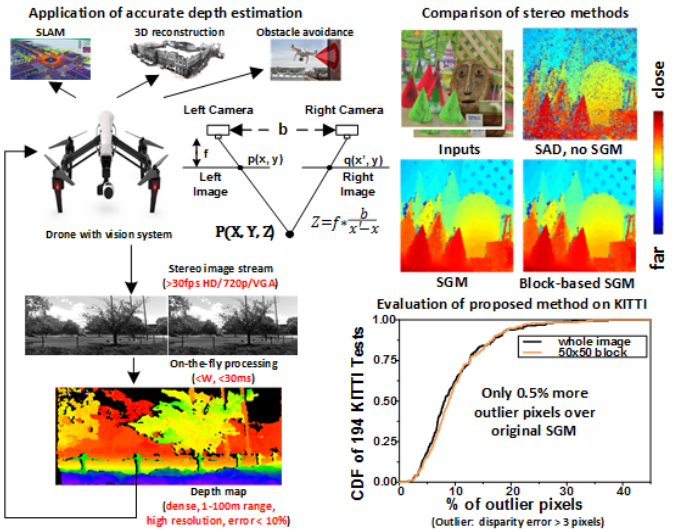


Figure 3.7.1: Depth estimation on MAVs and associated requirements. Comparison of local, SGM and proposed block-based SGM.

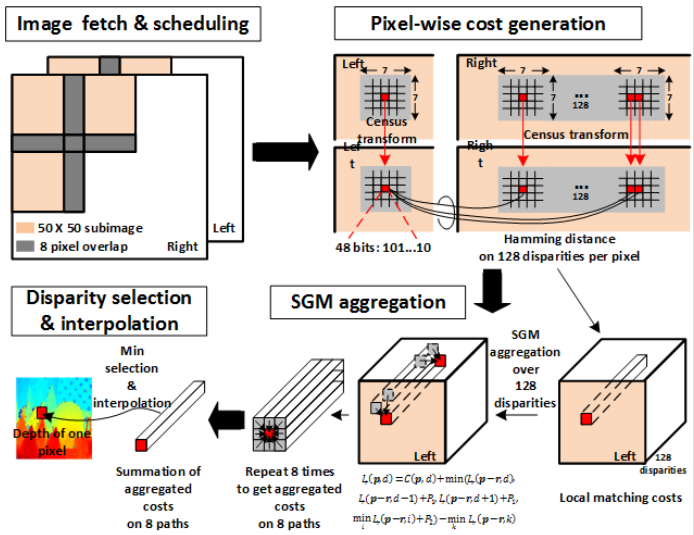


Figure 3.7.2: Proposed block-based SGM processing procedure.

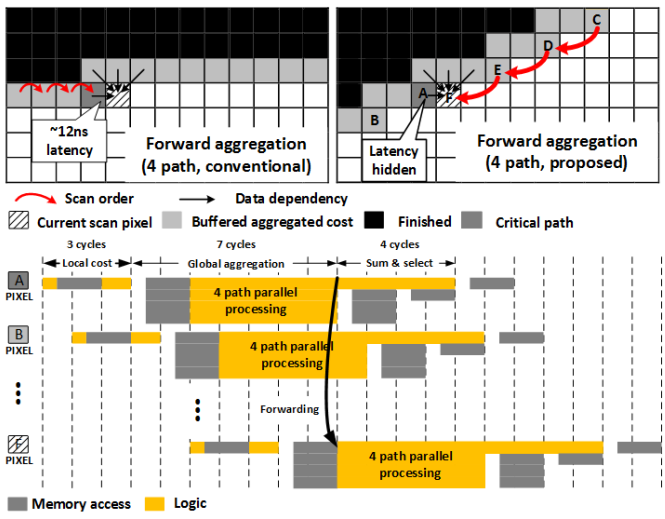


Figure 3.7.3: Proposed dependency-resolving scan and corresponding pipelining and forwarding scheme.

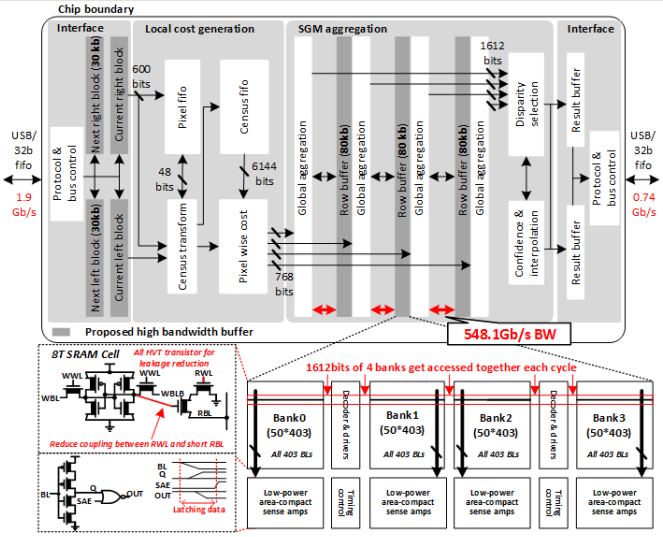


Figure 3.7.4: Chip block diagram and schematic of proposed compact high bandwidth SRAM.

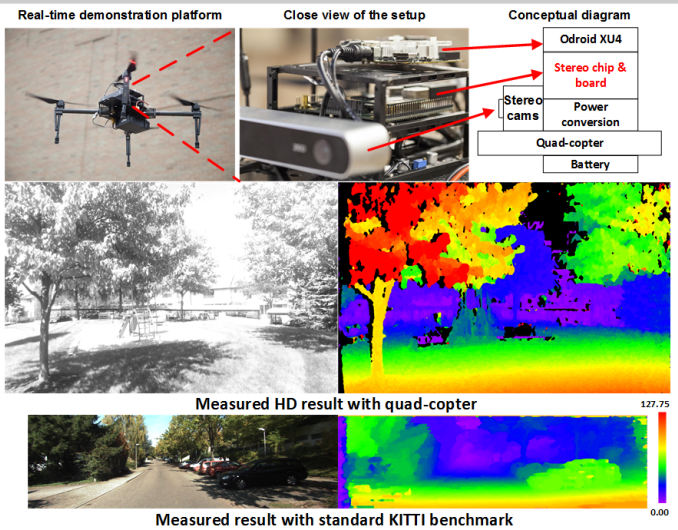
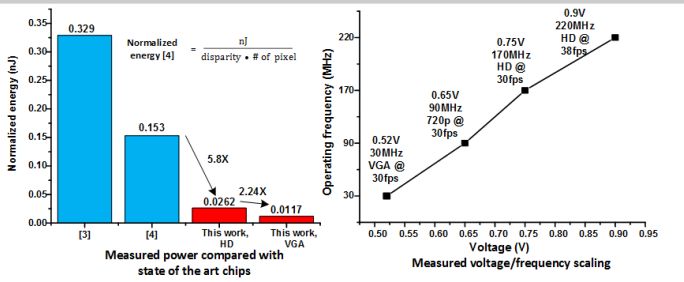


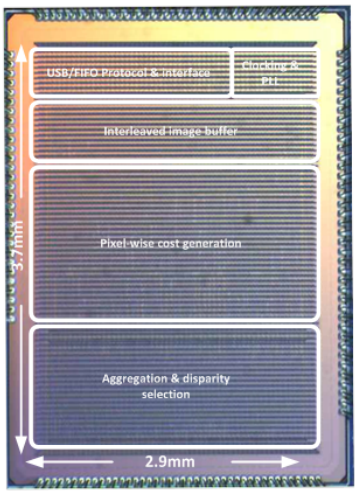
Figure 3.7.5: Real-time demonstration setup and visualization of chip measurements.



	VLSI 2004[2]	ISSCC 2015[4]	ISSCC 2016[3]	This work
Method	SAD	5-View BP	Truncated SGM	SGM
Technology	180nm	40nm	65nm	40nm
Chip area	25mm^2	22mm^2	16mm^2	10.8mm^2
On-chip memory	N.A	352Kb	3946.9Kb	1064Kb
Frequency	125MHz	215MHz	250MHz	170MHz
Throughput & image size	320 X 240 30fps	1920 X 1080 30fps	1280 X 720 30fps	1920 X 1080 30fps
Depth level	N.A	64 (32 + 1 bit fractional)	64	512 (128 + 2 bit fractional)
Accuracy (Outlier % on KITTI)	Not reported due to limited depth range			7%
Operating voltage	1.8 V	0.9 V	1.2V	0.75V @ 30 fps HD 0.52V @ 30 fps VGA
Power	N.A	1019mW (includes DRAM power)	582mW (excludes DRAM power)	836mW @ 30 fps HD 55 mW @ 30 fps VGA
Normalized energy*	N.A	0.153nJ	0.329nJ	0.0262nJ @ 30 fps HD 0.0117nJ @ 30 fps VGA

* Normalized energy [4] - equation shown in Fig 6 (top left)

Figure 3.7.6: Chip measurements and comparison with recent prior works.



	This work
Algorithm	8 Path SGM
Technology	40nm
Core Area	10.8mm^2
On-chip memory	1064Kb
Frequency	170MHz
Image size & throughput	1920 X 1080 30 fps
Depth	512 (128 + 2 bit fractional)
Benchmark evaluation	7% outlier @ KITTI
Operating voltage	0.75V @ 30 fps HD 0.52V @ 30 fps VGA
Power	836mW @ 30 fps HD 55 mW @ 30 fps VGA
Normalized energy	0.0262nJ @ 30 fps HD 0.0117nJ @ 30 fps VGA

Figure 3.7.7: Die photo and summary of performance.

978-1-5090-3758-2/17/$31.00