Low Power and Low Latency Line Streaming Image Processing for 3D User Interface on Smart Devices

Sungpill Choi, Hoi-jun Yoo
School of Electrical Engineering, KAIST

Abstract A low-power 3D hand-gesture tracking (HGT) processor is proposed for real-time user interaction in smart devices. It is optimized for 3-level (Functional, processing, and ALU): 1) Region of hands are extracted in low resolution and it reduces overall redundant 3D HGT computation; 2) Line streaming CNN processing improves pipeline utilization and reduces overall 3D HGT latency; 3) Shifter-based stereo aggregation processing elements reduce power and area. The proposed chip facilitates real-time 3D HGI with 33.5mW and maximally 10.8mm error.

Keywords: 3D hand gesture tracking, low power, low latency

1. Introduction

In these days, smart devices such as head mounted displays (HMDs) which supports virtual reality (VR) or mixed reality (MR) are widely used. Its number is over 12.4M and it is expected to increase exponentially for next several years [1]. Unlike previous mobile devices such as smart phones or tablets, VR/MR devices require novel user interface (UI) instead of conventional 2D-based UIs which are mouse or touch interface. Thus, to provide better VR/MR user experiences (UX), 3D hand-gesture tracking (HGT) has drawn active attention for a future UI. However, because HGT requires large computations and memory accesses compared with conventional UIs, it is impossible to realize real-time latency and low power on mobile application processors (APs) or mobile GPUs.

This paper presents a chip for real-time latency (33.3ms) HGT with 9.02mW power consumption which is ~1% of nominal mobile APs. To reduce power consumption, we utilizes stereovision systems for 3D depth acquisition instead of time of flight sensors (ToFs) which consumes large power (~225mW) [2]. Also, we optimize additional computations from stereovision processing to reduce overall system power consumption. In this paper, we will introduce 3-level hardware optimization methods for real-time 3D HGT.

2. Overall system

Fig. 1 illustrates overall 3D HGT processing. It composed of 4 stages which are stereo calibration, hand detection, depth estimation, and hand tracking. Boundaries of each stages are determined whether processing should be processed sequential. Due to non-decomposed processing, latency optimization is performed in each stages.

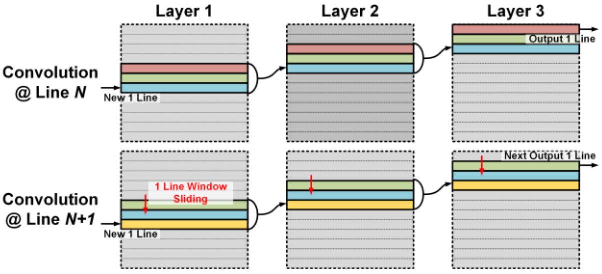


Fig. 2. Line streaming CNN processing.

A. Function-level optimization

In 3D HGT, hands usually occupies maximally 54% of overall image area when hands locates at 15cm distance which is a nominal scenario(15cm~45cm). Thus, most of regions are background in input images and our system set region of interest (ROI) to reduce this redundancy. It is generated by 4 layer CNN with 4 channels for each channels and 80X60 low-resolution images.

B. Processing-level optimization

To acquire accurate depth data such as ToF, CNN-based stereovision[3] is essential. Even though we reduce computation by selective processing, because CNN feature extraction and feature matching requires large computations, 3D HGT cannot meet real-time latency. In conventional CNN hardware, CNN is processed layer-by-layer to maximize data reusability[4-6]. But this causes utilization degradation because matching

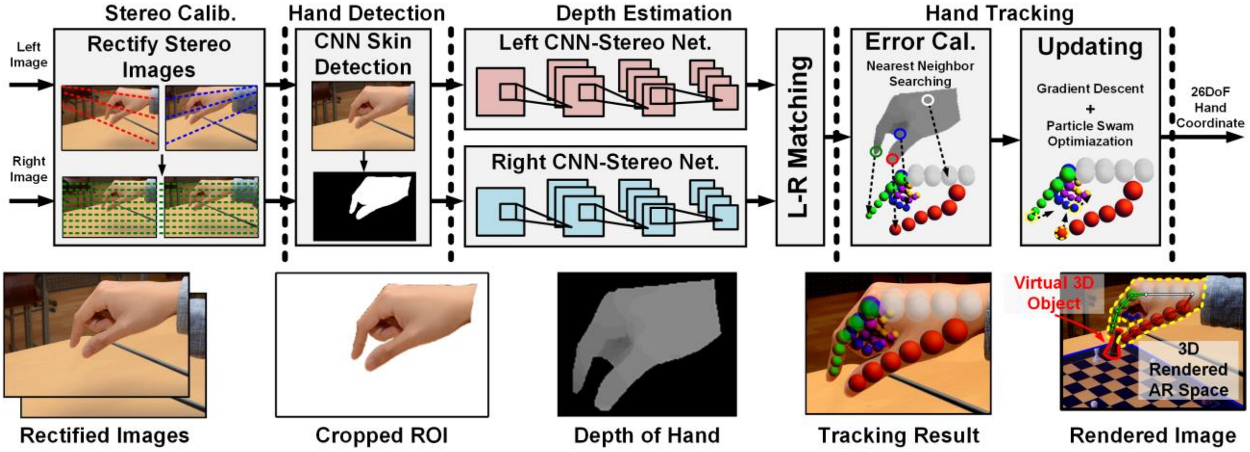
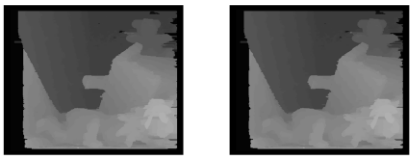


Fig. 1. Flow of 3D hand-gesture tracking & visualization.



Error Region	Adaptive Support-Weight*	Shifter Based Aggregation	Diff.
All	10.6%	10.7%	+ 0.1%
Non-Occ	7.2%	7.1%	- 0.1%
Disc	16.5%	16.7%	+ 0.2%

Fig. 3. Results of shift-based matching

hardware should wait CNN processing. To resolve it, as shown in fig. 2, the proposed hardware performs CNN line-by-line and matching operation can be hidden behind line CNN processing. As a results, we reduce 30.6% overall latency.

In addition, our hand detection works as line streaming processing and we adopt line streaming calibration algorithm in [7]. Thus, line streming hand detection processing also can be hidden behind stereo calibration and it reduces additional 3.4% latency.

C. ALU-level optimization

In 3D HGT, stereo feature matching operation requires massive floating-point computation (172Gflops). Unlike fixed-point CNN processing, it causes large power consumption and large area. To reduce power and area, we propose shift-based matching algorithm. In stereo matching, because it utilizes a winner-takes-all(WTA) method, we can convert expnent operations in [8] to shift operation. As a result, we reduce 92.2% of power and 96.4% of area with neglible error compare with [8] as shown in fig. 3.

3. Implementation results

The proposed 3D HGT processor is fabricated in 65nm 1P8M CMOS technology and it occupies 4X4mm^2 with 781.5KB SRAM. Energy efficiency of line streaming CNN core is 1.80Tops/W and it processes 3D HGT with 33.3ms latency. I consumes 9.02mW for one hand 3D HGT and 31.20mW for two hand 3D HGT. Optimal range of the system is from 20cm to 40cm.

Fig. 5 shows an evaluation system of proposed chip. The evaluation system is connected to host computer with USB 3.0 I/F. We setup both of stereo cameras and ToF sensor to evaluate hand tracking accuracy. Fig. 6 shows evaluation results of our test system. We tested it from 15cm to 50cm and calculated tracking errors at fingertips, hand nodes compared with ToF sensor data. Average error is 4.3mm and maximum error is 10.6mm in test environments.

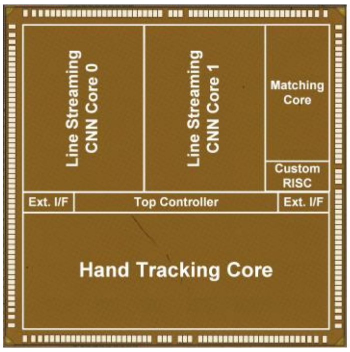
4. Conclusion

A 9.02mW 3D hand gesture tracking (HGT) processor is proposed for 33.3ms latency user interaction in smart devices. As a result, it realizes always-on 3D UI for AR/VR devices

References

[1] Statistics of AR/VR devices. Available: https://www.statista.com/statistics/539713/worldwide-virtual-and-augmented-reality-hardware-shipments/

[2] C. S. Bamji et al., "IMpixel 65nm BSI 320MHz demodulated TOF Image sensor with 3μm global shutter pixels and analog binning," ISSCC, 2018, pp. 94-96, 2018.



Tehnology	65nm 1P8M CMOS	
Die Area	4mm x 4mm (16mm^2)	
On-chip Memory	CNN-Stereo Engine	490.5 KB
	ICP-PSO Engine	291 KB
Supply Voltage	0.7V ~ 1.2V	
Clock Freqency	10MHz ~ 100MHz	
1-Hand Real-time Scenario (0.85V/50MHz)		
Power	CNN-Stereo Engine	3.13mW
	ICP-PSO Engine	5.89mW
	Total	9.02mW
2-Hands Real-time Scenario (1.1V/90MHz)		
Power	CNN-Stereo Engine	10.84mW
	ICP-PSO Engine	20.36mW
	Total	31.20mW
Latency	30ms (33.3fps)	
HGR Range	20cm ~ 40cm	
Hand Tracking Error	Max	10.6mm
	Average	4.3mm

Fig. 4. Chip photography

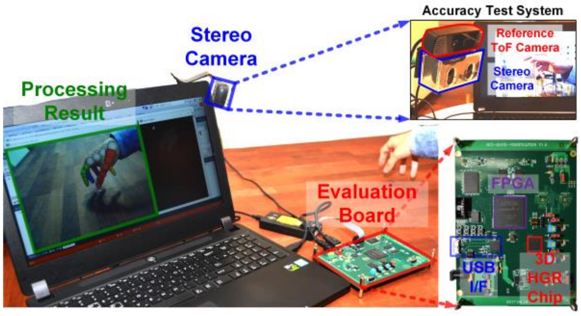


Fig. 5. Evaluation system

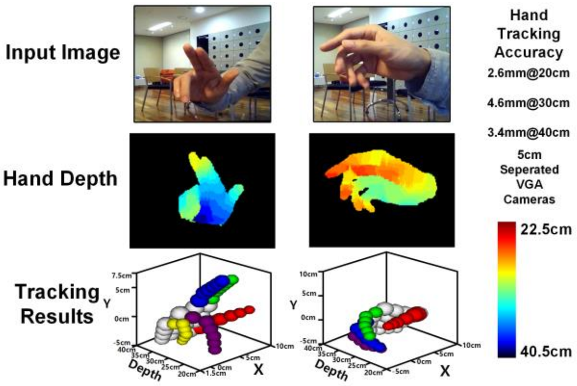


Fig. 6. 3D Hand gesture tracking results

[3] W. Luo, et al., "Efficient Deep Learning for Stereo Matching," CVPR, pp. 5695-5703, 2016.

[4] K. Bong, et al., "14.6 A 0.62mW ultra-low-power convolutional-neural-network face-recognition processor and a CIS integrated with always-on haar-like face detector," ISSCC, pp. 248-249,.

[5] Y. H. Chen, et al., "14.5 Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," ISSCC, pp. 262-263,.

[6] D. Shin, et al., "14.2 DNPU: An 8.1TOPS/W reconfigurable CNN-RNN processor for general purpose deep neural networks," ISSCC, pp. 240-241, 2017.

[7] H. Tong, Nonlinear Time Series, J. B. Elsner, ed: “A Dynamical System Approach,” Oxford University Press, Oxford (1990).

[8] C. Kim, et al., "A CMOS Image Sensor-Based Stereo Matching Accelerator With Focal-Plane Sparse Rectification and Analog Census Transform," TCAS-I, vol. 63, no. 12, pp. 2180-2188, Dec. 2016.

[9] Kuk-Jin Yoon, et al., "Locally adaptive support-weight approach for visual correspondence search," CVPR, pp.924-931, 2005.