18.1 A 2.71nJ/Pixel 3D-Stacked Gaze-Activated Object-Recognition System for Low-Power Mobile HMD Applications

Injoon Hong, Kyeongryeol Bong, Dongjoo Shin, Seongwook Park, Kyuho Lee, Youchang Kim, Hoi-Jun Yoo

KAIST, Daejeon, Korea

Smart eyeglasses or head-mounted displays (HMDs) have been gaining traction as next-generation mainstream wearable devices. However, previous HMD systems [1] have had limited application, primarily due to their lacking a smart user interface (UI) and user experience (UX). Since HMD systems have a small compact wearable platform, their UI requires new modalities, rather than a computer mouse or a 2D touch panel. Recent speech-recognition-based UIs require voice input to reveal the user's intention to not only HMD users but also others, which raises privacy concerns in a public space. In addition, prior works [2-3] attempted to support object recognition (OR) or augmented reality (AR) in smart eyeglasses, but consumed considerable power, >381mW, resulting in <6 hours operation time with a 2100mWh battery.

In this paper, we propose a low-power OR system with an intention-concealed gaze UI for HMD applications, which can be used all day long with battery power. Fig. 18.1.1 shows the overall system concept of the 3D stacked-gaze-activated OR system on the HMD platform. The gaze-image sensor (GIS) stacked on the OR processor (ORP) detects the HMD user's pupil center location and as the user moves his gaze on the display screen, he can point the cursor at an object on the screen. Then, the electrooculography (EOG) sensor monitors eyelid movement, such as a wink, to 'click' the cursor for the target object selection. After the user selects a region-of-interest (ROI) within the image from the front camera with the gaze 'pointing' and the wink 'click', the ORP performs OR only for the selected ROI, leading to significant power reduction. There are 3 key features for the low-power gaze-activated OR system: 1) the GIS with focal-plane pupil-edge detection circuits for low-power gaze estimation, 2) a reconfigurable convolutional neural-network processor (CNNP) with a tile-parallel architecture to reduce external memory accesses, 3) a descriptor-generation processor (DGP) with trigonometric function approximation based on a look-up table (LUT) for low-power processing.

Figure 18.1.2 depicts the overall system architecture of the gaze-activated OR system, stacking the GIS chip and the ORP chip vertically. The ORP has a total of 20 IPs connected by a 2D NoC. The EOG processor (EOGP) is a programmable RISC to detect the wink command from EOG signals, and the DRM processor (DRMP) controls DVFS based on the real-time workload estimation. The CNNP reduces the gaze-selected 640×360 ROI into ~100 32×32 ROI tiles with sparse-feature learning by the CNN. From the ROI tiles selected by the CNNP, the scale-space processor (SSP), feature-detection processor (FDP) and descriptor-generation processor (DGP) operate in a tile-level task pipeline to extract SIFT feature vectors. The feature-matching processor (FMP) adopts a low-power vocabulary forest processor architecture [2] and makes the final recognition decision. The GIS in Fig. 18.1.2 has 320×240 pixel arrays, a 10b ADC, and mixed-mode pupil-edge-detection circuits (PEDC) for low-power gaze estimation.

Figure 18.1.3 shows the key circuit diagrams of the PEDC for the new GIS-friendly pupil-edge-detection algorithm: XY pupil detection (XY-PD). The XY-PD compares intensity values with 4 neighboring pixels along the X/Y axis of the sensor, which is more power efficient than the previous 8-pixel weighted-sum operation [4]. For the XY-PD, the PEDC converts photo voltages V1 and V2 into currents I1 and I2 by source followers M1 and M2, respectively. Then, I1 is compared with $I2+I_{TH}$ and its digital voltage output is latched. A PEDC is shared with all pixels connected to a column for column-parallel processing. Since column-parallel processing requires fewer PEDCs than pixel-level processing, and the XY-PD requires only 4 pixel comparisons rather than the 8-pixel weighted-sum operation, this work consumes 2.9× less power, even with 16× better resolution than the previous work [4]. In addition, the RISC in the GIS reduces sub-pixel pupil detection error by fitting an optimum ellipse [5] on the boundary of the dark pupil with the data obtained from the PEDC. As a result, the proposed GIS obtains ~0.45° gaze-estimation accuracy, which is equivalent to only 17 pixels on a virtual 100-inch screen of 1920×1080 resolution at 10ft distance.

The proposed CNNP of Fig. 18.1.4 employs a sparse-feature-based CNN – a promising neural-network architecture for deep-learning applications [6]. It can reduce the database size to 8kB from the 1MB associated with previous work [3] obviating the need for massive external memory accesses. In addition, the CNNP has a tile-parallel architecture with a multi-layer-perceptrons (MLP) classifier for fast ROI tile selection. The same reconfigurable 200-MAC (multiply-accumulate) array performs the tile-parallel convolutions in the CNN mode with the offline learned sparse features, and then the MAC operates in MLP mode to make the final decision for the tile selection. It supports 8-tile parallel convolution operations with 8-way MAC arrays in CNN mode, and 200 parallel MAC operations with 200-way MAC arrays in MLP mode. As a result, it takes 3ms to process a 640×360 ROI rescaled in size to 160×90, and consumes 18mW peak power, which is 1.8× smaller than a previous hierarchical convolution accelerator [3].

Figure 18.1.5 shows the LUT-based low-power DGP architecture to calculate the trigonometric functions with low-power consumption. Since the rotation-invariant feature vector can be obtained with an 11.25° orientation binning operation, rather than with precise arctan ALU operations, we use a LUT to remove the power-hungry arctan ALU [2]. The orientation LUT in Fig. 18.1.5 determines the bin index among 32 orientation bins containing arctan results. The proposed trigonometric LUT reduces power by 1.5×, with <0.4% matching error, which is negligible considering the 11.25° accuracy requirement.

Figure 18.1.6 gives measurement results for the CNNP with the tile-parallel CNN for ROI selection. The C1 result shows the intermediate image result extracted from the convolution operation on 50 sparse features in the C1 layer of Fig. 18.1.4. The C3 result shows the intermediate result of the convolution operation at the C3 layer with 250 sparse features. The MLP accepts the C3 results to select the final ~100 ROI tiles among 225 tiles, achieving 56% workload reduction. It is noticeable that even the C3 results look very similar to the final selected ROIs. The DRMP chooses one of 4 different DVFS modes based on a workload estimation, shown in Fig. 18.1.6. There are 3 dynamic power domains in the ORP: the SSP and FDP are in domain1, the DGP in domain 2, and the FMP in domain 3. Because the workload of domain 1 is proportional to the number of ROI tiles, the DRMP assigns {V_{DD1}:0.7V f_1:20MHz} to the SSP and FDP for a scene with 30 ROI tiles, and {V_{DD1}:1.2V f_1:100MHz} to the SSP and FDP for a scene with 215 ROI tiles. In addition, since the workloads of domains 2 and 3 are proportional to the number of features, the DRMP applies {V_{DD2}:0.7V f_2:10MHz} to the DGP and {V_{DD3}:0.7V f_3:20MHz} to the FMP for a tile with 2 feature vectors and {V_{DD2}:0.7V f_2:50MHz} to DGP and {V_{DD3}:0.7V f_3:100MHz} to the FMP for a tile with 30 feature vectors. Consequently, ORP power is reduced by 30% compared to the ORP without the DRMP.

The proposed 3.36×3.36mm^2 GIS and 4×4mm^2 ORP are fabricated in 65nm CMOS technology integrating 3.68M NAND2 equivalent gate count and 400kB SRAM as shown in Fig. 18.1.7. The GIS is 3D stacked on the ORP by wire bonding. The peak system performance is 151.3GOPS and its peak power consumption is 131mW. Since the size of the ROI is reduced by the GIS and the CNNP, the OR system consumes 75mW average power with 30fps throughput in 720p resolution with the help of the LUT-based DGP and the DRMP-controlled DVFS. As a result, compared to the previous ORP [2], the proposed system consumes 3.4× lower power and 2.71nJ/pixel energy consumption. In conclusion, the low-power OR system with 3D stacking of GIS and ORP is successfully realized for a gaze-based smart intuitive UI suitable for HMD systems.

References:

[1] Google Glass. Retrieved from: https://www.google.com/glass/start/

[2] G. Kim, *et al.*, "A 1.22TOPS and 1.52mW/Mhz Augmented Reality Multi-Core Processor with Neural Network NoC for HMD Applications," *ISSCC Dig. Tech. Papers*, pp. 182-183, Feb. 2014.

[3] J. Park, *et al.*, "A 646 GOPS/W Multi-classifier Many-core Processor with Cortex-like Architecture for Super-Resolution Recognition," *ISSCC Dig. Tech. Papers*, pp. 168-169, Feb. 2013.

[4] D. Kim, *et al.*, "A 5000S/s Single-chip Smart Eye-tracking Sensor," *ISSCC Dig. Tech. Papers*, pp. 46-47, Feb. 2008.

[5] D. Li, *et al.*, "Starburst: A Hybrid Algorithm for Video-based Eye Tracking Combining Feature-based and Model-based Approaches," *IEEE Computer Vision and Pattern Recognition*, pp. 79, 2005.

[6] K. Jarret, "What is the Best Multi-Stage Architecture for Object Recognition?" *IEEE International Conference on Computer Vision*, pp. 2146-2153, 2009.

 978-1-4799-6224-2/15/$31.00

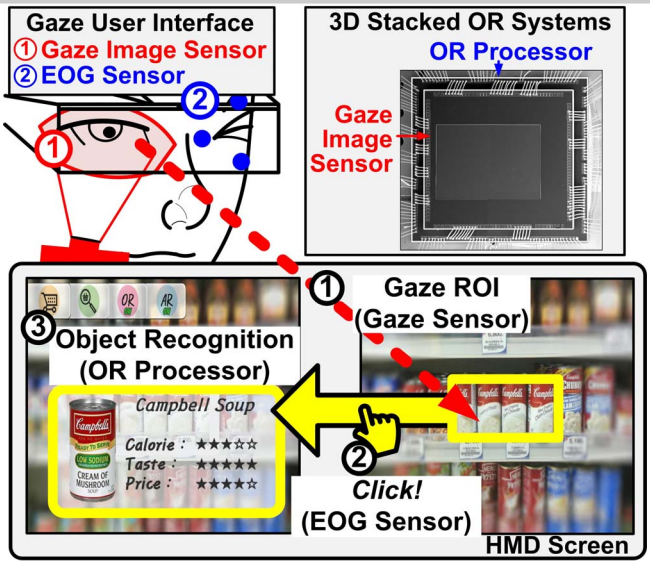


Figure 18.1.1: A gaze-activated object-recognition system.

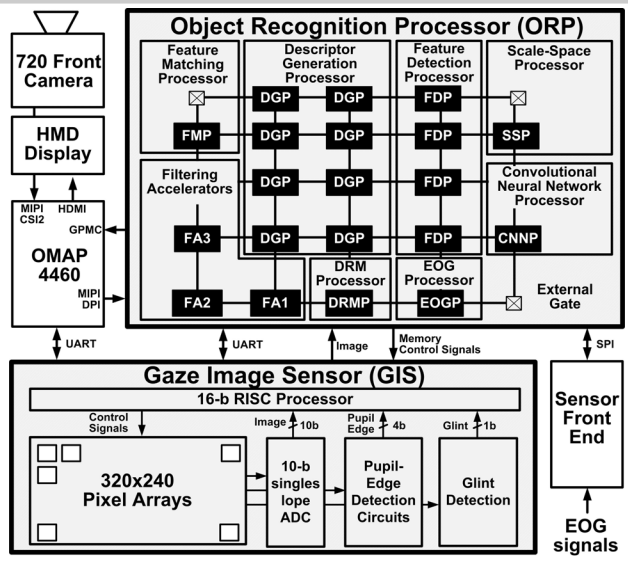
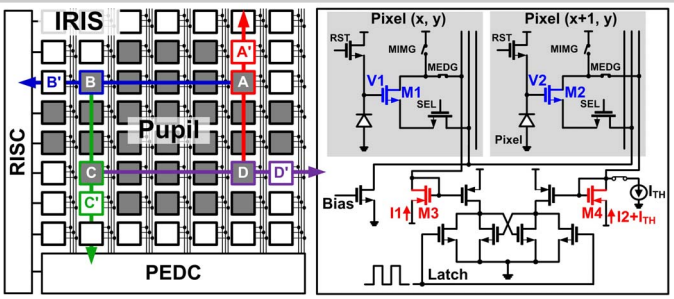
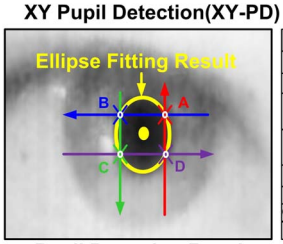


Figure 18.1.2: System architecture of proposed OR system.



XY Pupil Detection(XY-PD) | Pupil Edge Detection Circuit(PEDC)



Pupil Detection Result

	This Work	ISSCC 2008[4]
Functionality	PD +Gaze Estimation	PD
Resolution	320x240	80x60
PD Algorithm	XY-PD + Ellipse Fitting	Shrinking & Expanding
Processing Architecture	Column-Parallel + Digital RISC	In-Pixel
Fill Factor	60%	20%
PD Error	<1 pixel	√2 pixel
Gaze Accuracy	0.5°	-
Peak Power	34mW	100mW

Performance Comparison

Figure 18.1.3: A mixed-mode PEDC in gaze image sensor.

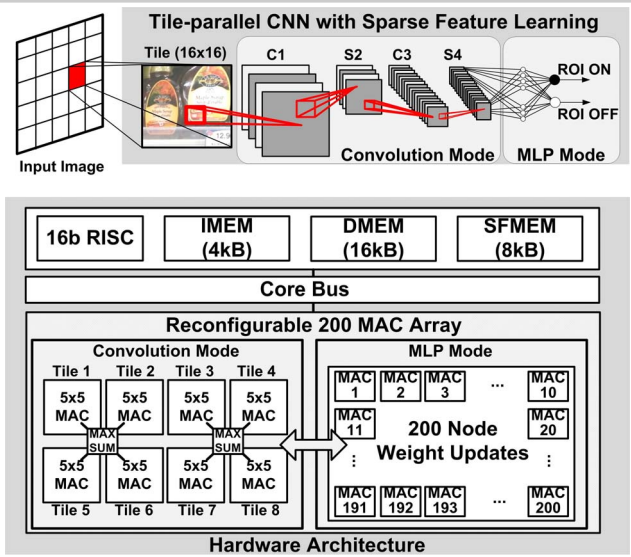
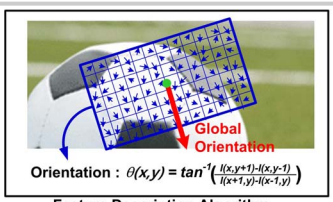


Figure 18.1.4: A reconfigurable tile-parallel CNNP architecture.



Feature Description Algorithm

LUT Orientation Assignment

dx sign	dy sign	dy/dx	θ(x,y) bin
+	+	0≤ dy/dx <51	0≤ θ <11.25°
-	+	0≤ dy/dx <51	168.75° ≤ θ <180°
...		...	...
-	-	dy/dx >1287	258.75°≤ θ <270°

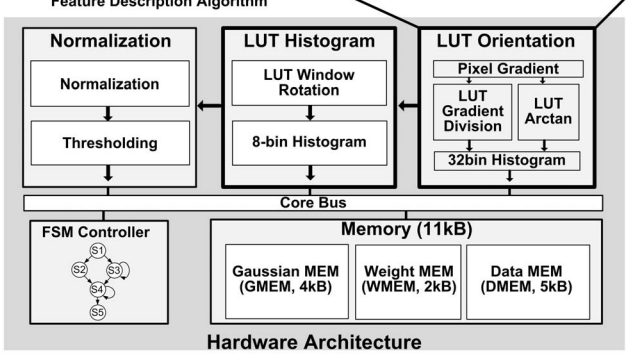


Figure 18.1.5: A LUT-based DGP architecture.

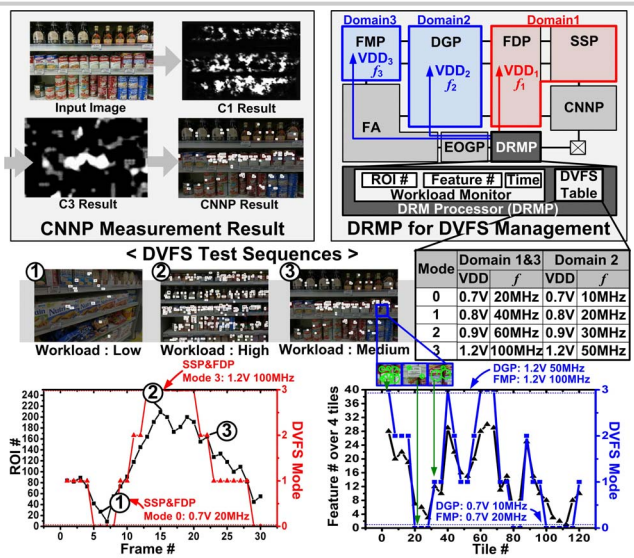
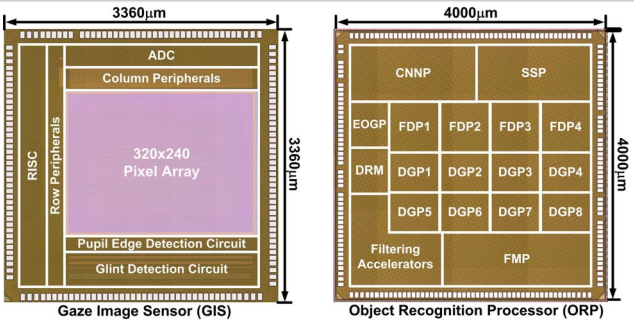


Figure 18.1.6: Measurement results of the DRMP for DVFS control.



Technology	65nm 1P8M Logic CMOS		
Chip Size	GIS	3360x3360μm²	
	ORP	4000x4000μm²	
Frequency	GIS	50MHz	
	ORP	Nominal	50MHz: DGP 100MHz: FDP/SSP/FMP 200Mhz: CNNP/DRMP EOGP/FA/NoC
		DVFS	10 – 100 MHz (36FO4)
Voltage	GIS	2.5V(Pixel) / 1.2V(Others)	
	ORP	Nominal	1.2V
		DVFS	0.7 – 1.2V
Gate / SRAM	3.68M / 400kB		
Peak Performance	GIS	0.3 GOPS	
	ORP	SSP	37.4 GOPS
		4xFDP	50.4 GOPS
		8xDGP	9.2 GOPS
		CNNP	42.0 GOPS
		Others	12 GOPS
	Total		151.3 GOPS

	ISSCC'13 [2]	ISSCC'14 [1]	This Work
Functionality	SIFT-based OR	SIFT-based OR[1)]	SIFT-based OR
Resolution	720p	720p	720p
User Interface	X	X	Gaze UI
Process	130nm	65nm	65nm
Throughput	30fps	30fps	30fps
Performance	271.4 GOPS	1103 GOPS[1)]	151 GOPS
Peak Power	420mW	669mW[1)]	97mW(ORP) + 34mW (GIS)
Average Power	260mW	328mW[1)]	65mW(OR) + 10mW (GIS)
Energy Consumption	9.4nJ/pixel	11.8nJ/pixel	2.71nJ/pixel

1) Comparison with Object Recognition Performance in ISSCC'14

Figure 18.1.7: Chip photograph and performance summary.

978-1-4799-6224-2/15/$31.00