Mobile/Embedded DNN and AI SoCs

Hoi-Jun Yoo
KAIST

Outline

1. **Deep Neural Network Processor**
   – Mobile DNN Applications
   – Basic CNN Architectures

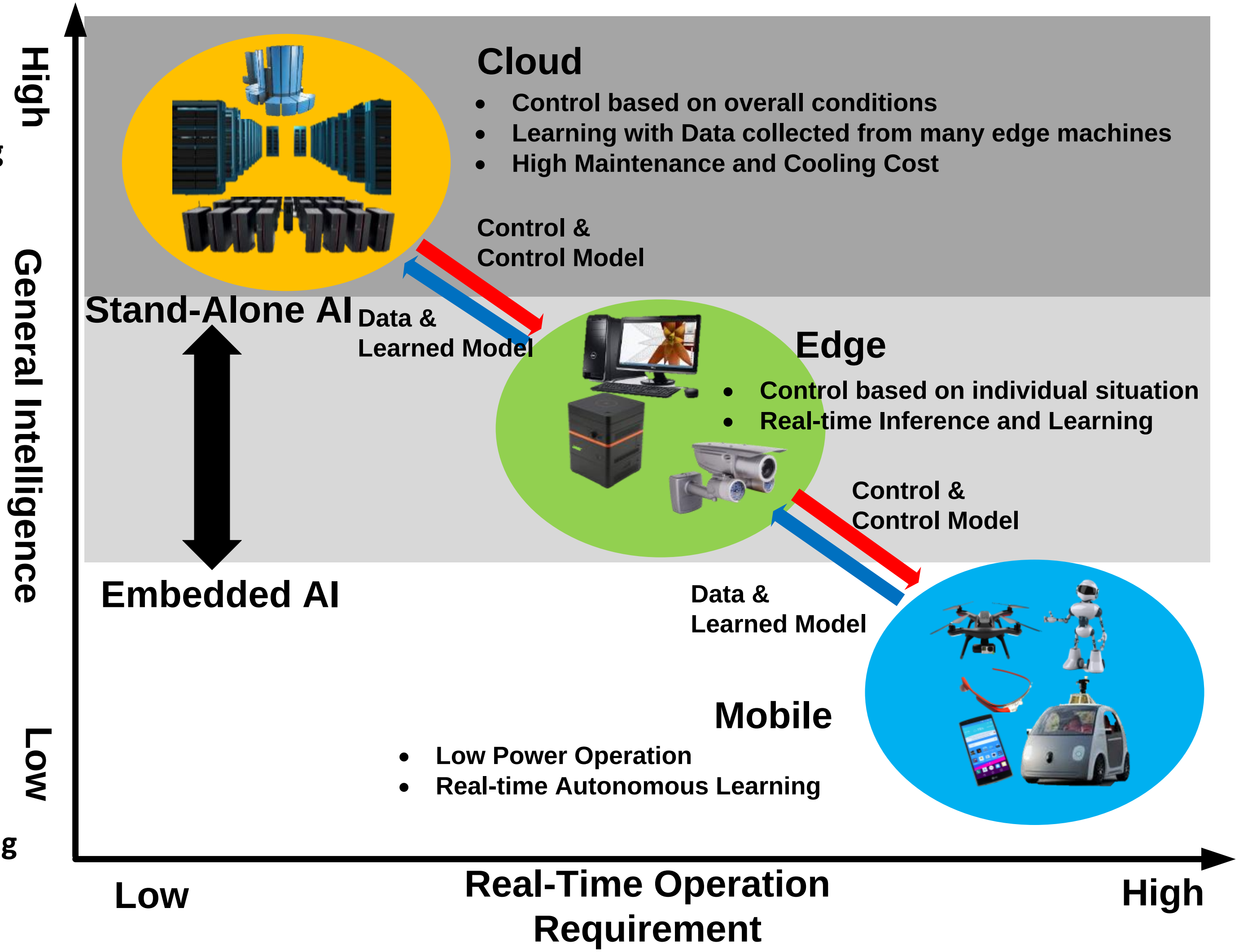
2. **M/E-DNN: Mobile/Embedded Deep Neural Network**
   – Requirements of M/E-DNN
   – M/E-DNN Design & Example

3. **SoC Applications of M/E-DNN**
   – Hybrid CIS, CNNP Processor and DNPU Processor
   – Hybrid Intelligent Systems
   – AR Processor, UI/UX Processor and ADAS Processor

Types of AI SoCs

- General AI
- Scalability
- Global Data Sharing

- Specific AI
- UI / UX
- Limited Data Sharing



High

General Intelligence

Low

Cloud

- Control based on overall conditions
- Learning with Data collected from many edge machines
- High Maintenance and Cooling Cost

Control &
Control Model

Stand-Alone AI

Data &
Learned Model

Edge

- Control based on individual situation
- Real-time Inference and Learning

Control &
Control Model

Embedded AI

Data &
Learned Model

Mobile

- Low Power Operation
- Real-time Autonomous Learning

Low

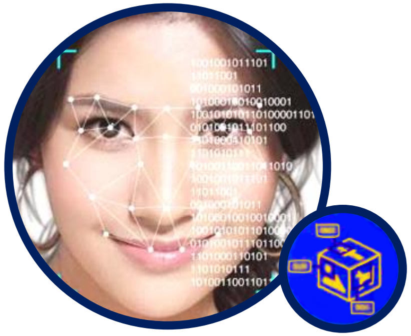
Real-Time Operation Requirement

High

Emerging Mobile Applications

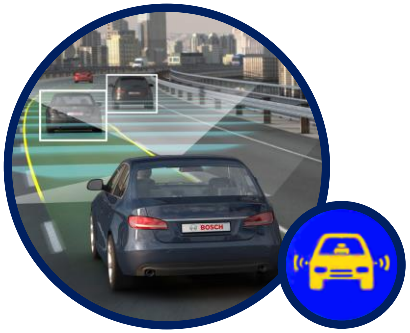
- The Needs for Embedded Vision & AI

Security & Surveillance



Visual Perception & Analytics

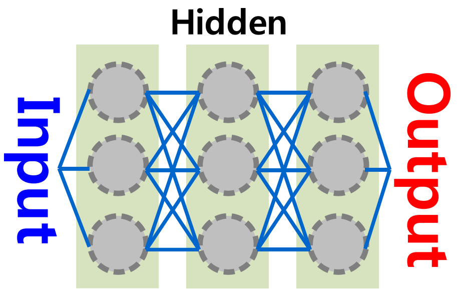
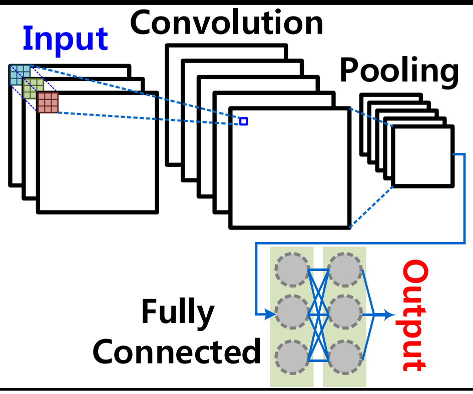
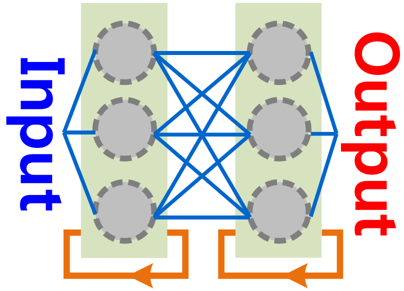
Drones



ADAS & Autonomous Cars

Augmented Reality

Deep Neural Networks

	MLP (Multi-Layer Perceptron)	CNN (Convolutional)	RNN (Recurrent)
		 Input Convolution Pooling Fully Connected Output	
Characteristic	Fully Connected	Convolutional Layer	Sequential Data Feedback Path Internal State
Major Application	Speech Recognition	Vision Processing	Speech Recognition Action Recognition
Number of Layers	3~10 Layers	Max ~100 Layers	3~5 Layers

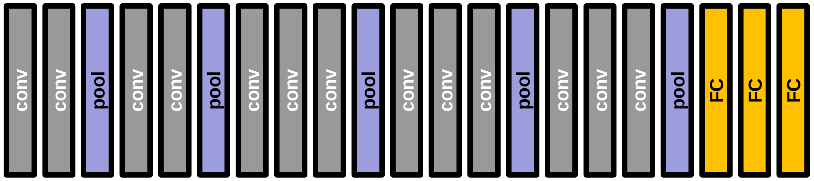
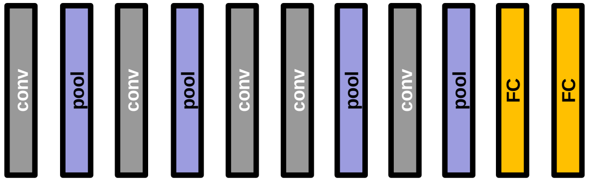
Outline

1. **Deep Neural Network Processor**
   – Mobile DNN Applications
   – Basic CNN Architectures

2. **M/E-DNN: Mobile/Embedded Deep Neural Network**
   – Requirements of M/E-DNN
   – M/E-DNN Design & Example

3. **SoC Applications of M/E-DNN**
   – Hybrid CIS, CNNP Processor and DNPU Processor
   – Hybrid Intelligent Systems
   – AR Processor, UI/UX Processor and ADAS Processor

Mobile/Embedded DNN Requirements

Cloud-based DNN	Mobile/Embedded DNN
 conv conv pool conv conv pool conv conv conv pool conv conv conv pool conv conv conv pool FC FC FC	 conv pool conv pool conv conv pool conv pool FC FC
• Cloud Intelligence • General AI - Very deep network • Communication latency • High Number Precision • High power / cooling cost • Large memory capacity	• On-device Intelligence • Application-specific AI - Specific DNN network • Real-time operation • Reduced Number Precision • Low-power consumption • Efficient memory arch.

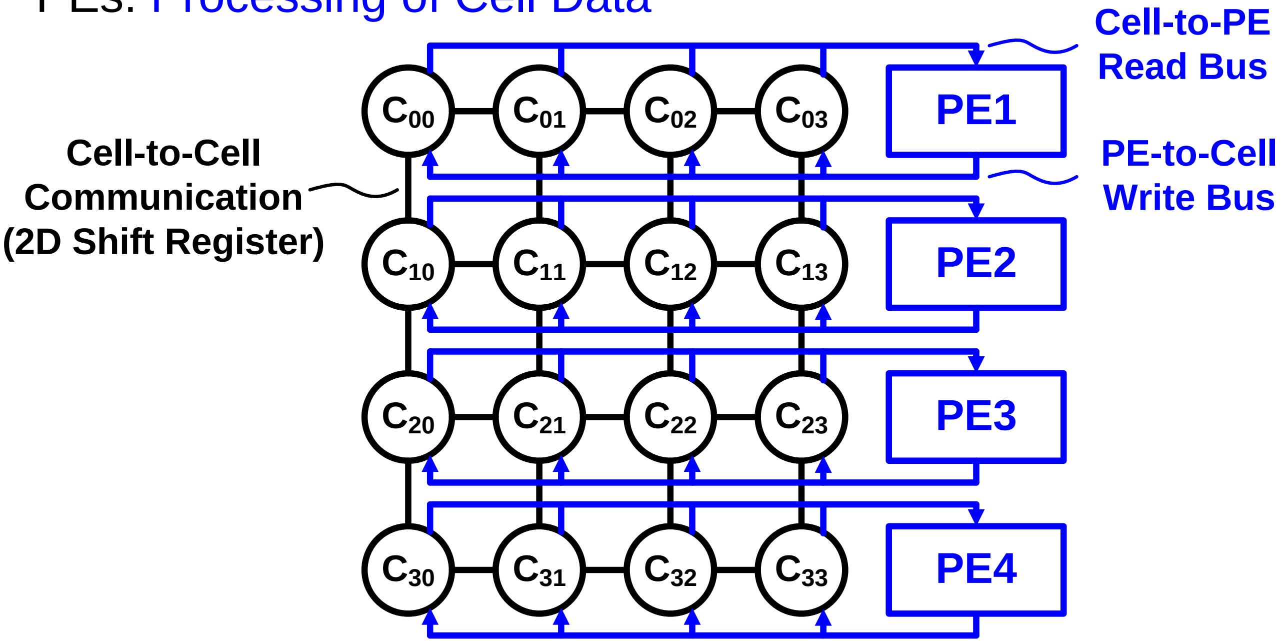
Hoi-Jun Yoo

Mobile/Embedded DNN Architecture

[1] Kwanho Kim et al, ISSCC 2008

[2] Seungjin Lee et al, IEEE TNN 2011

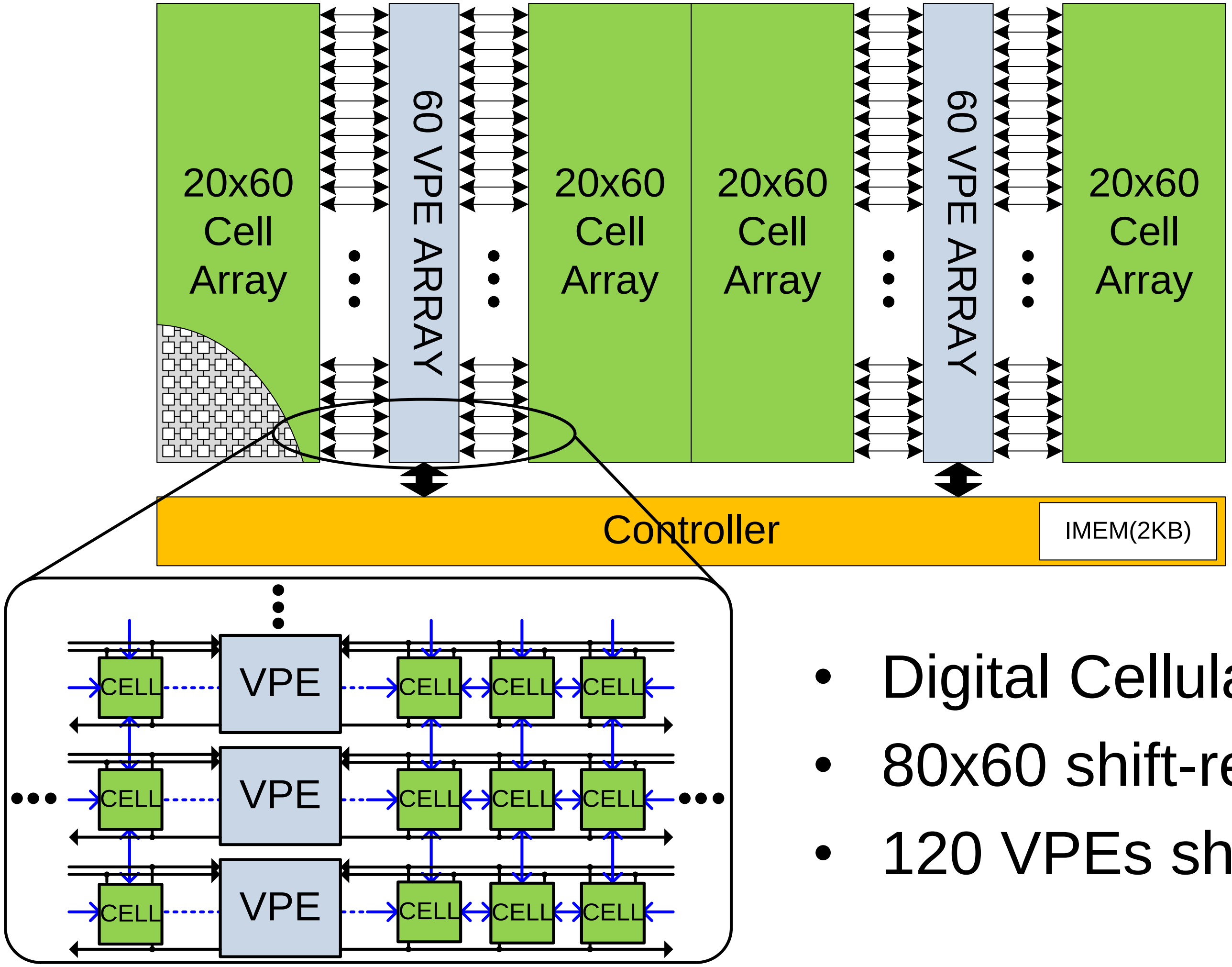
- 2D Cell + Shared PE Digital CNN
  - 2D Cell Array to Remove Emulation Overhead
  - Programmable PE for Flexible Operation
    - Cells: Storage and Communication
    - PEs: Processing of Cell Data



Hoi-Jun Yoo

SRAM based Architecture

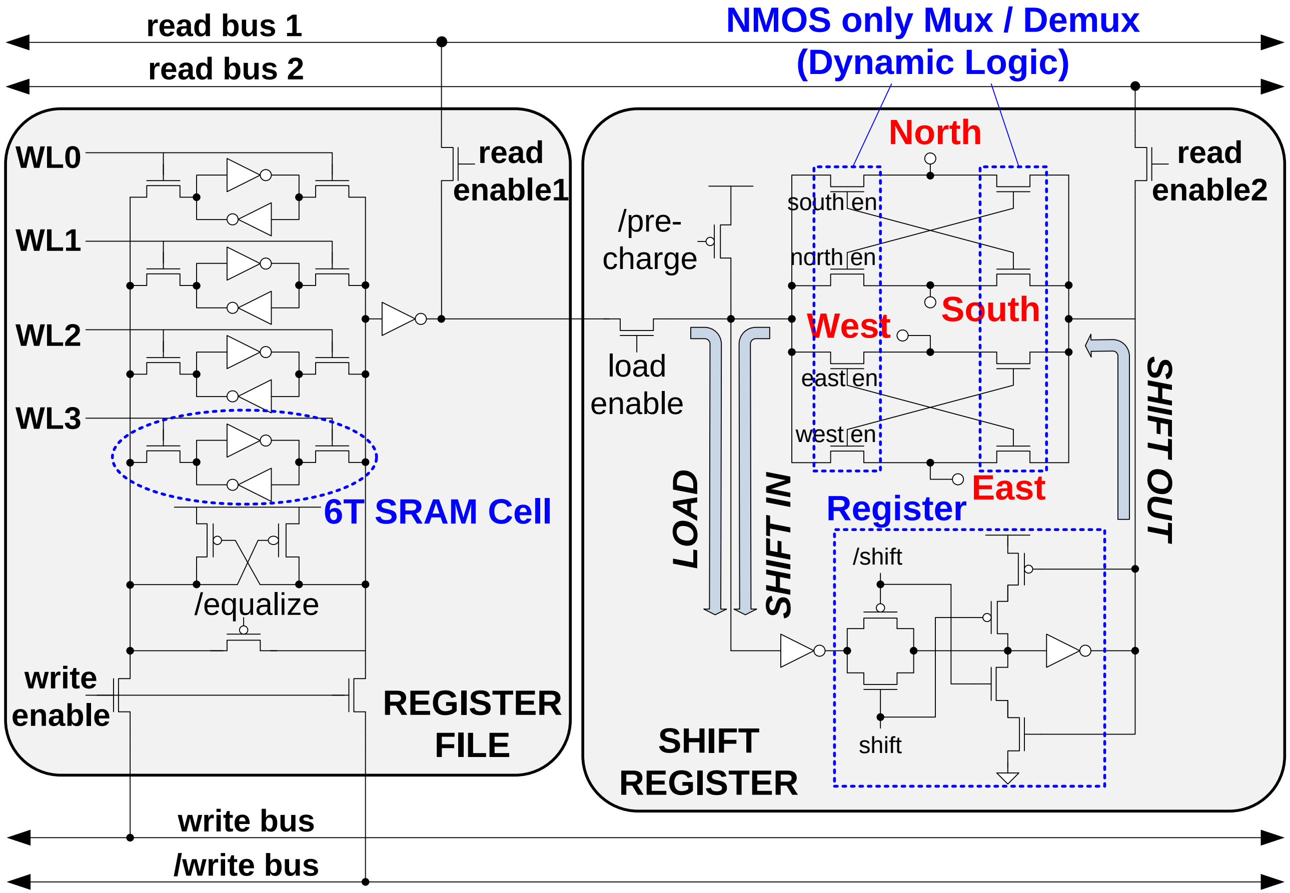
[8] Kwanho Kim et al, ISSCC 2008



- Digital Cellular Neural Net.
- 80x60 shift-register array
- 120 VPEs shared by cells

SRAM Memory Based NN Cell

[8] Kwanho Kim et al, ISSCC 2008



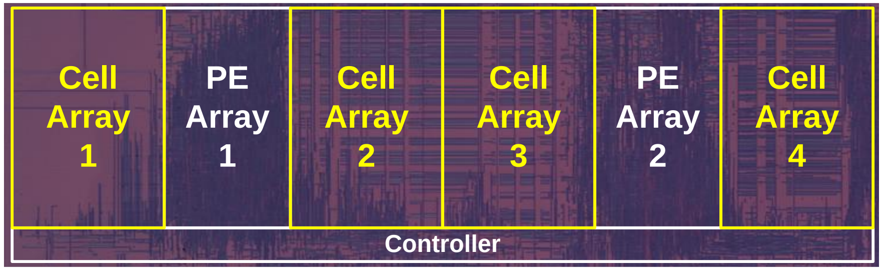
- Dynamic logic-based shift register → Small cell area

Hoi-Jun Yoo

Chip Summary and Applications

[8] Kwanho Kim et al, ISSCC 2008

[10] Seungjin Lee et al, ISSCC 2010



[redacted]

Process Technology	0.13μm 8M CMOS
Area	4.5 mm^2
Number of Cells	4800
Number of PEs	120
Operating Frequency	200 MHz
Peak Performance (Sustained)	24 GOPS (22 GOPS)
Active Power	84 mW
Average Power (15FPS)	6 mW

Outline

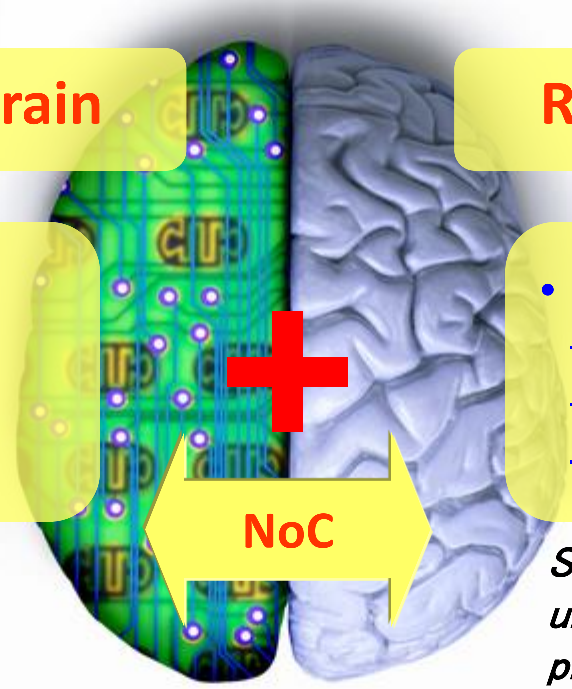
1. **Deep Neural Network Processor**
   - Mobile DNN Applications
   - Basic CNN Architectures

2. **M/E-DNN: Mobile/Embedded Deep Neural Network**
   - Requirements of M/E-DNN
   - M/E-DNN Design & Examples

3. **SoC Applications of M/E-DNN**
   - Hybrid CIS, CNNP Processor and DNPU Processor
   - Hybrid Intelligent Systems
   - AR Processor, UI/UX Processor and ADAS Processor

KAIST Approach: MC Processor + M/E DNN

Heterogeneous Multi-Core SoC for Humanistic Intelligence System



Left brain

- **Hard computing**
  - CPU / DSP
  - Special purpose HW
  - Multiple cores

Fast / accurate computing performance

Right brain

- **Soft computing**
  - Deep Neural Nets
  - Fuzzy logic
  - Learning / evolution

Suitable for solving imprecision, uncertainty, and ambiguous problems

Intelligence computing system:

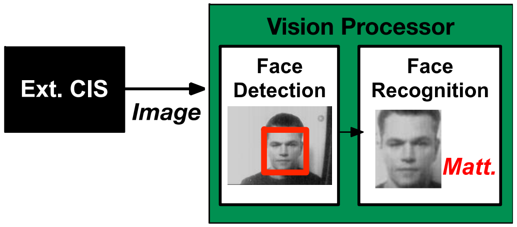
multi-core digital processors + DNN soft-computing hardware (through **Networks-on-Chip**).

Low Power Face Recognition SoC

[1] Kyeongryeol Bong et al, ISSCC 2017

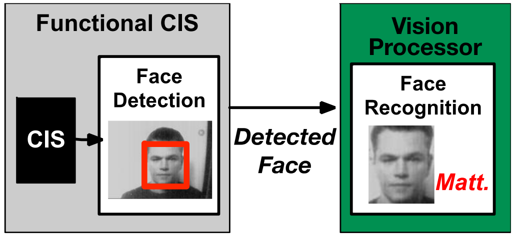
Conventional Face Detector

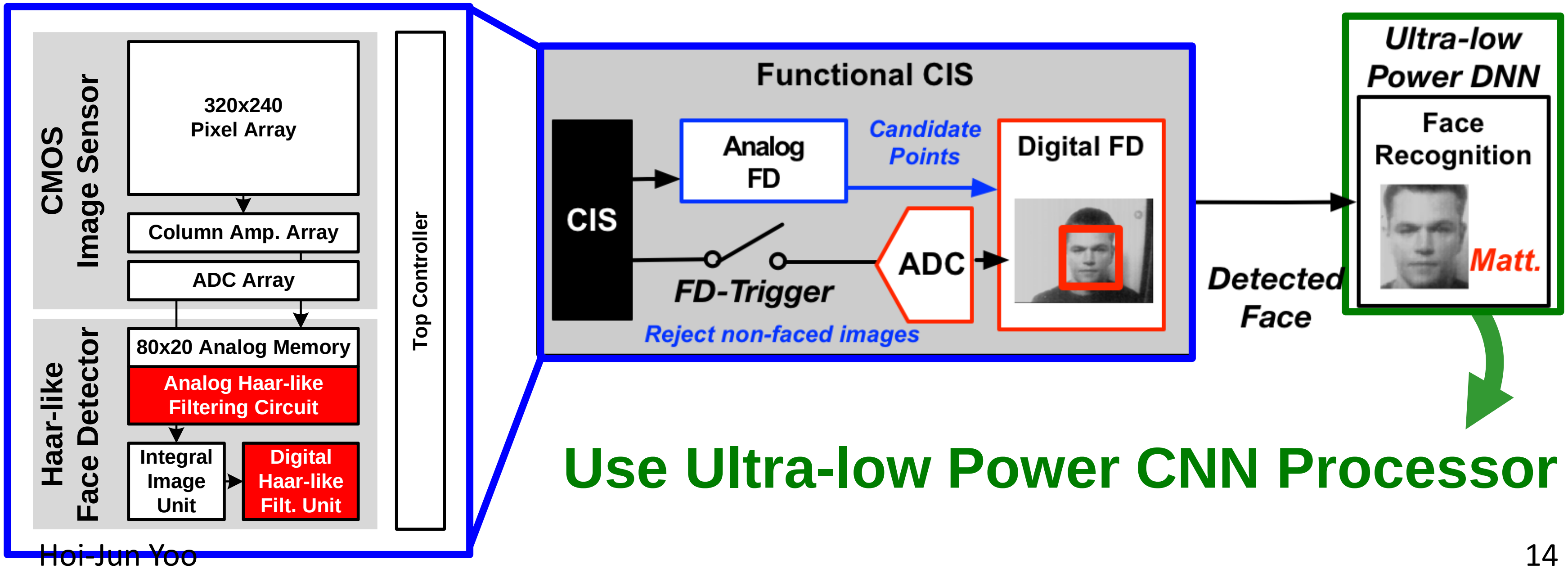
- Face detection by vision processor



Hybrid Face Detector

- Face detection by CMOS image sensor
- Combine analog & digital face detector

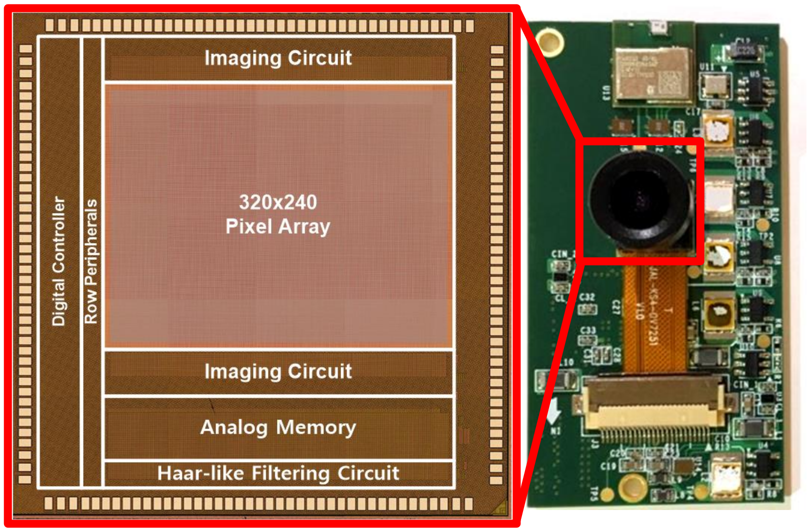




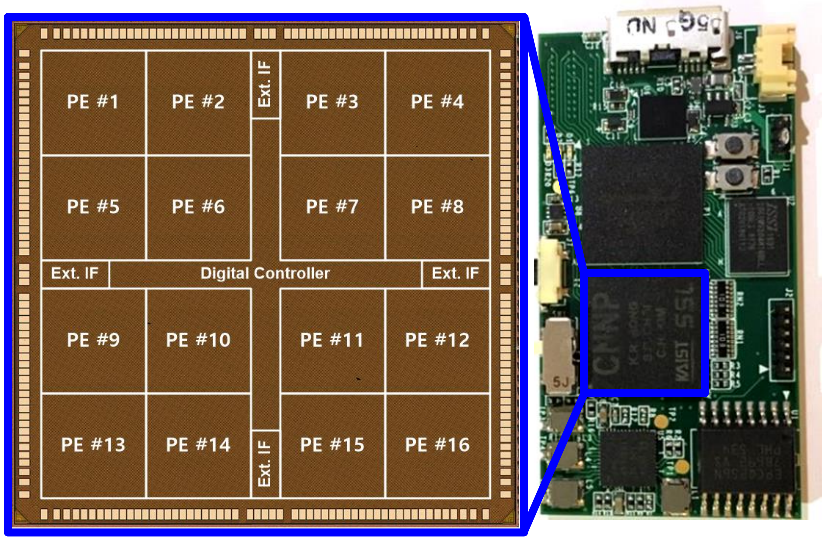
Use Ultra-low Power CNN Processor

Chip & System implementation

[1] Kyeongryeol Bong et al, ISSCC 2017



Always-on face detector (FD)

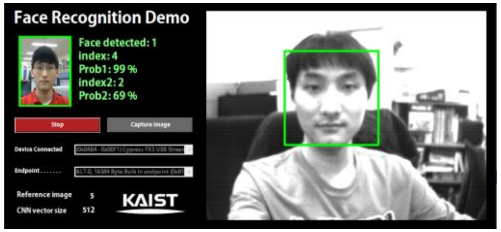


Ultra low-power CNN processor

- **Always-on FD: 3.3mm x 3.36mm CIS in 65nm**
  – 320 x 240 pixel array, Haar-like filtering circuit & analog MEM
- **CNN processor: 4mm x 4mm CNNP in 65nm**
  – 4 x 4 PE array with local T-SRAM

Chip & System implementation

[1] Kyeongryeol Bong et al, ISSCC 2017

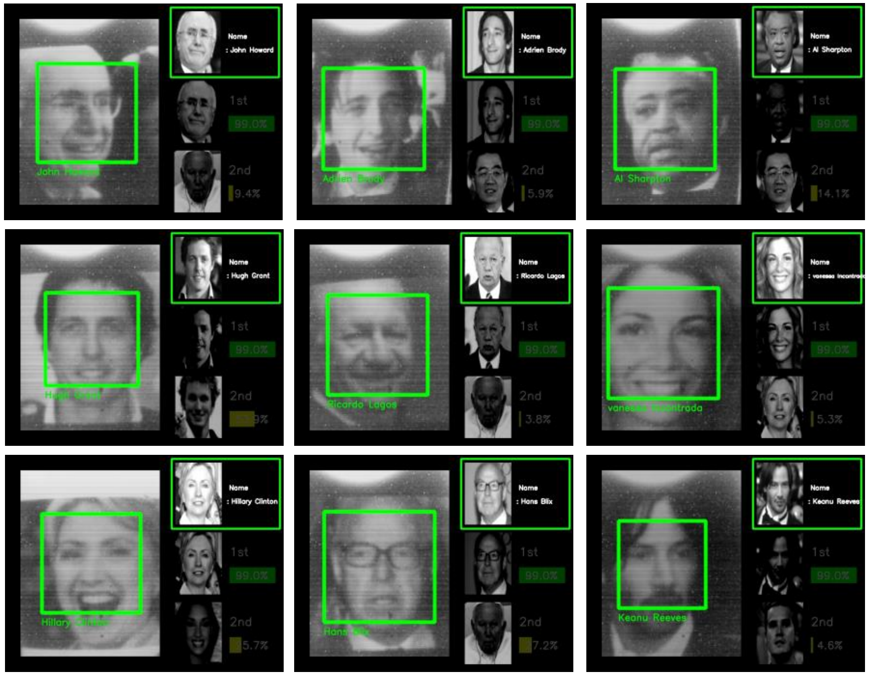

Face Recognition Demo

Face detected: 1
index: 4
Prob1: 99 %
index2: 2
Prob2: 69 %

Device Connected (0x04B4 - 0x00F1) Cypress FX3 USB Stream

Endpoint .... ALT-0, 16384 Byte Bulk in endpoint (0x8

Reference image 5
CNN vector size 512



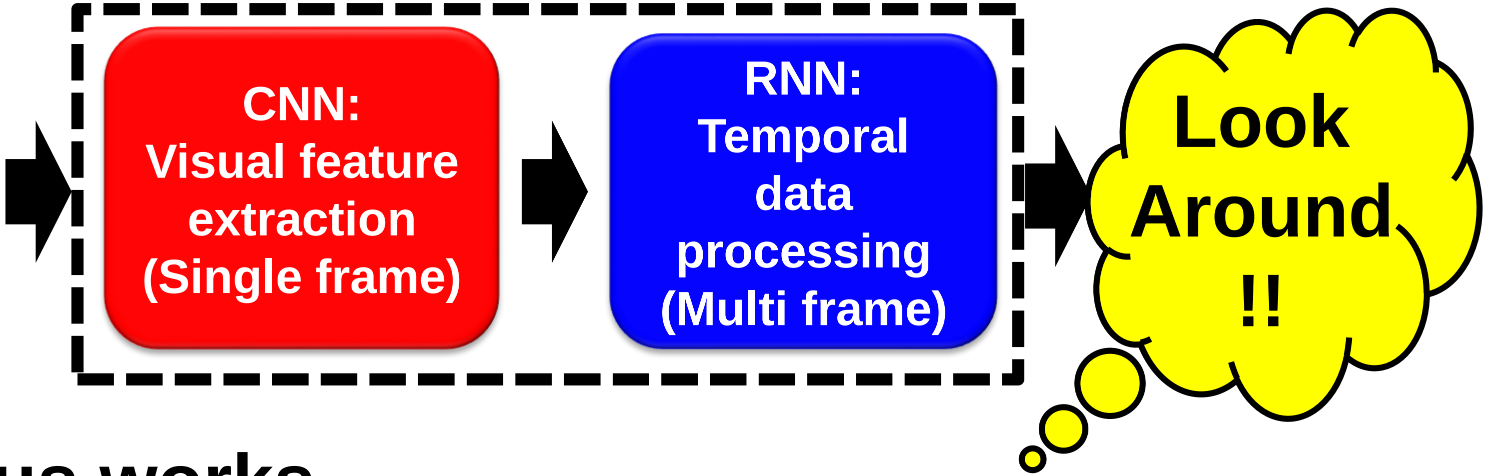
- **The Most Accurate Face Recognition SoC**
  - 97% Achieved using CNN @ LFW dataset
- **0.62mW Face Recognition System w/ imaging**

Hoi-Jun Yoo

CNN + RNN Deep Neural Network

[7] Dongjoo Shin et al, ISSCC 2017

- **CNN: Visual feature extraction and recognition**
  - Face recognition, image classification…
- **RNN: Sequential data recognition and generation**
  - Translation, speech recognition…
- **CNN + RNN: CNN-extracted features → RNN input**



CNN:
Visual feature extraction
(Single frame)

RNN:
Temporal data processing
(Multi frame)

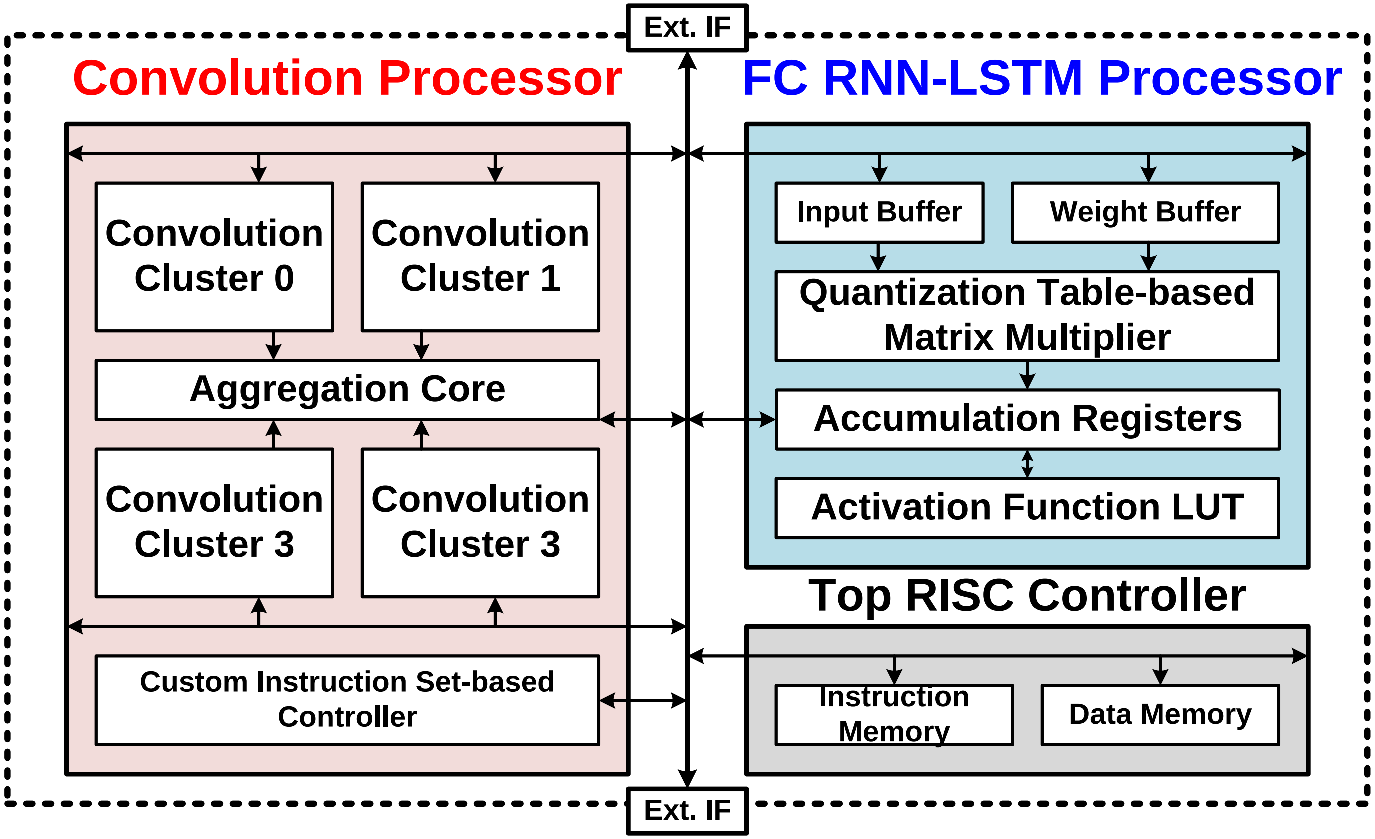
Look Around !!

- **Previous works**
  - **Optimized for convolution layer only: [6], [3]**
  - **Optimized for FC layer and RNN only: [5]**

[3] B. Moons, SOVC 2016
[5] S. Han, ISCA 2016
[6] Y. Chen, ISSCC 2016

Processor for support both CNN & RNN

[7] Dongjoo Shin et al, ISSCC 2017



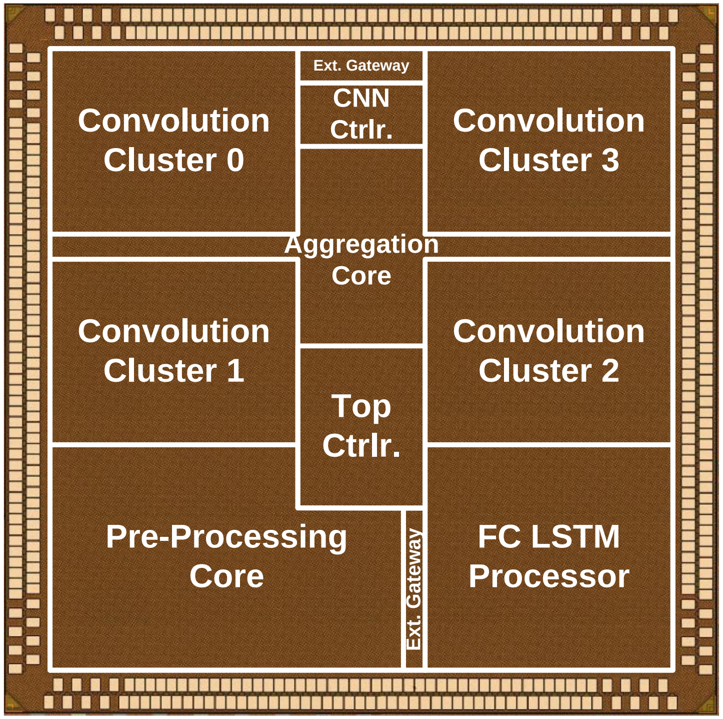
Convolution layer (CNN)

Fully-connected layer (CNN)

Recurrent neural network

DNPU: Chip implementation

[7] Dongjoo Shin et al, ISSCC 2017



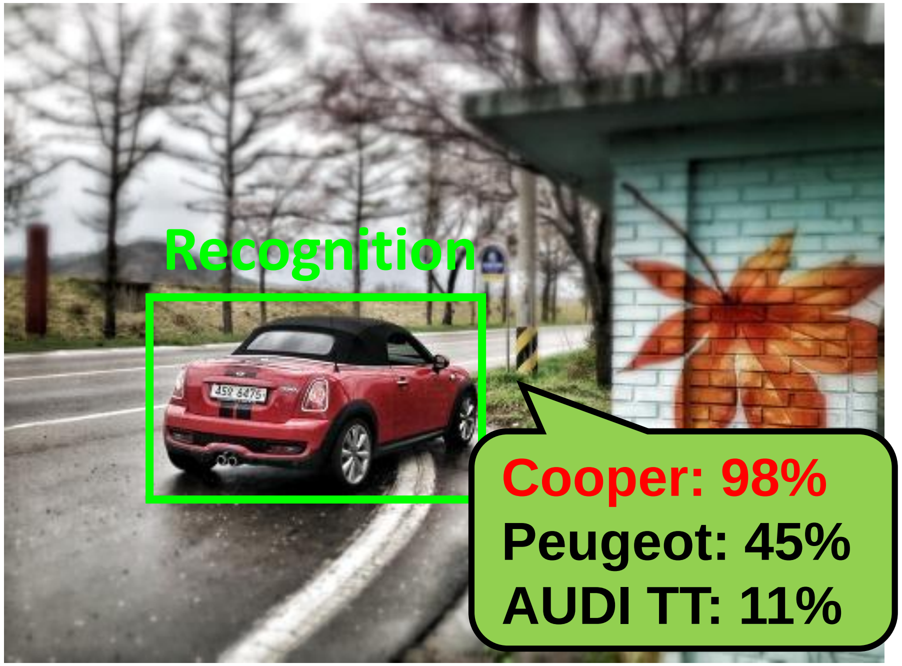
- CNN-RNN processor : 4mm x 4mm in 65nm
- Supply Voltage : 0.77V ~ 1.1V, frequency : 50 ~ 200MHz
- Energy Efficiency : 8.1TOPS/W (50MHz, 0.77V, 4bit word length)
  3.9TOPS/W (200MHz, 1.1V, 16bit word length)

Pet Robot Demonstration Video

[7] Dongjoo Shin et al, ISSCC 2017

Requirements for AR: High-Throughput

5 60148 77156 70
BEEP!
- Mini Cooper
- 1600CC
- $35,000



Marker-based AR

- 2D-barcode / 2D-marker
- Markers on Everything?

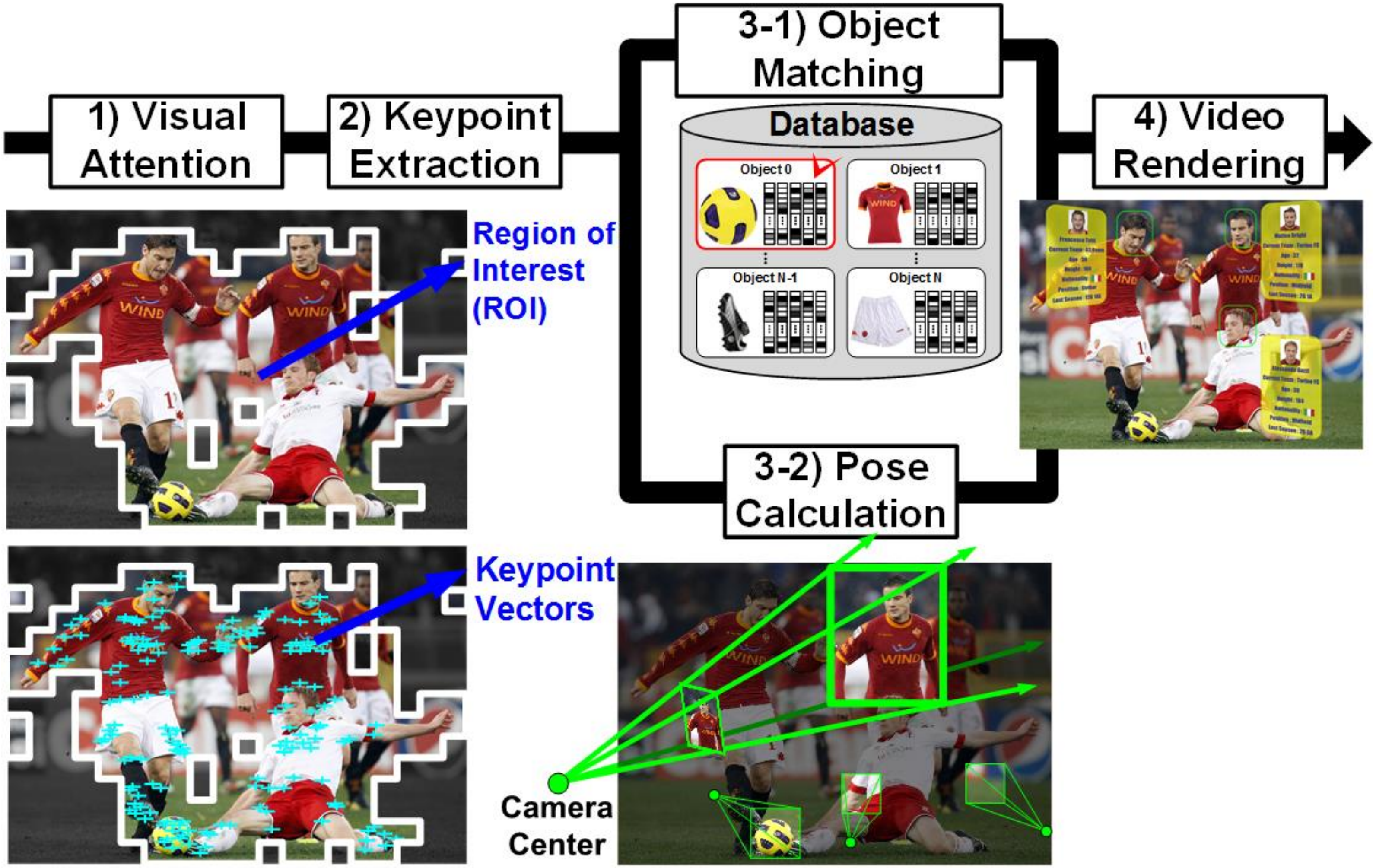
➔ Impossible Solution in Real World

Markerless AR

- Natural Feature Extraction
- >10x Higher Computation

➔ Difficulties in Real-time (30fps) Operation

Markerless AR Process

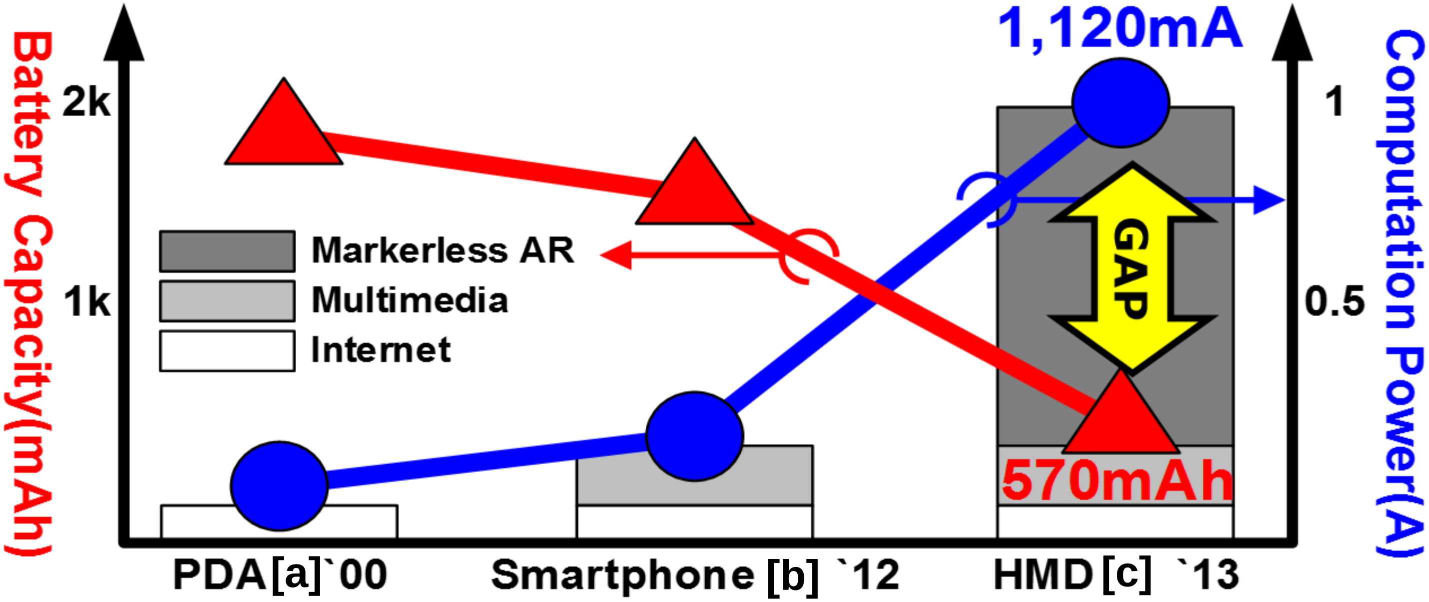


Hoi-Jun Yoo

Requirements for AR in HMD: Energy

Hands-free Head Mounted Display (HMD)

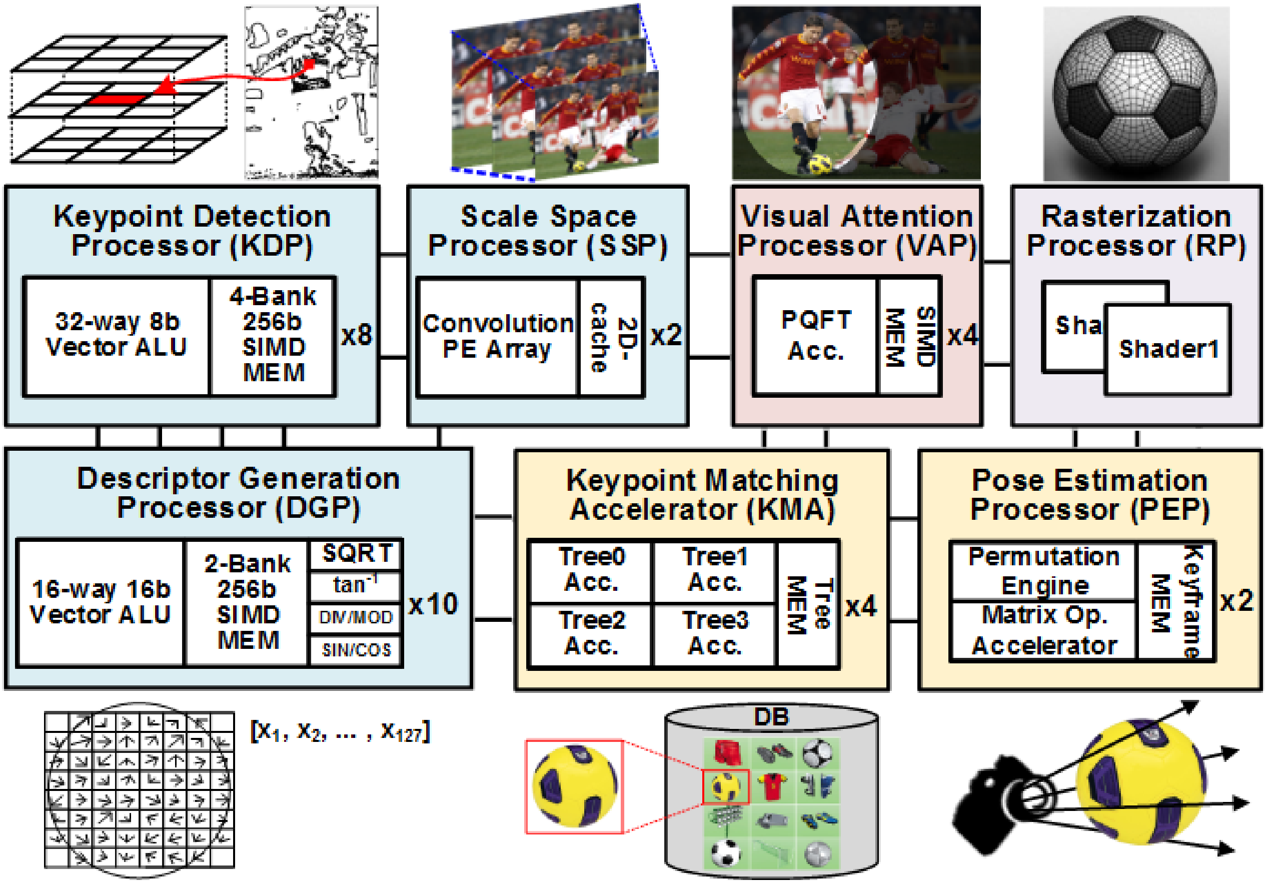
- Smaller Battery Size
- Complicated Markerless AR Functionalities



[a] HP iPAQ [b] Apple iPhone5 [c] Google Project Glass

Hoi-Jun Yoo

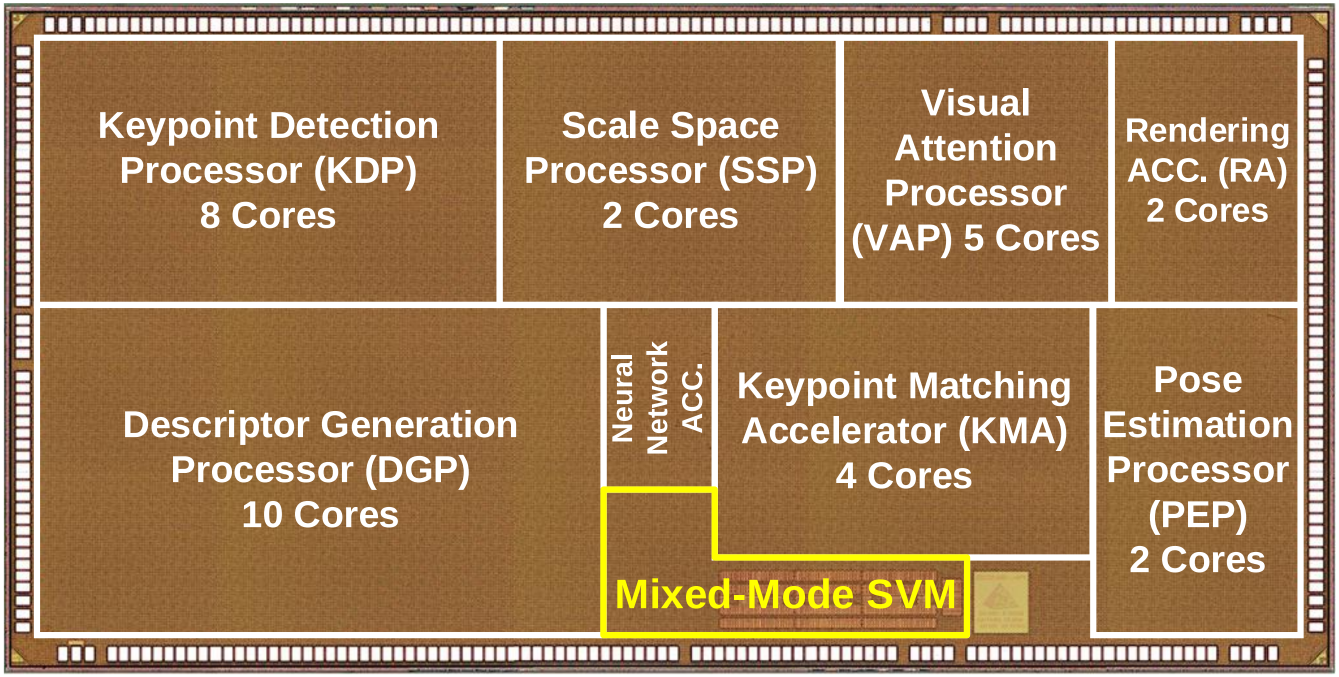
SoC Architecture of K-Glass



[12] Gyeonghoon Kim et al, ISSCC 2014

Hoi-Jun Yoo

BONE-AR: Chip Implementation


Keypoint Detection Processor (KDP) 8 Cores
Scale Space Processor (SSP) 2 Cores
Visual Attention Processor (VAP) 5 Cores
Rendering ACC. (RA) 2 Cores
Descriptor Generation Processor (DGP) 10 Cores
Neural Network ACC.
Keypoint Matching Accelerator (KMA) 4 Cores
Pose Estimation Processor (PEP) 2 Cores
Mixed-Mode SVM

- 4mm x 8mm by 65nm CMOS Technology
- 381mW Average / 778mW Peak Power Consumption
- 1.22 TOPS Peak Performance
- 1.57 TOPS/W Energy Efficiency for 720p Video

[12] Gyeonghoon Kim et al, ISSCC 2014

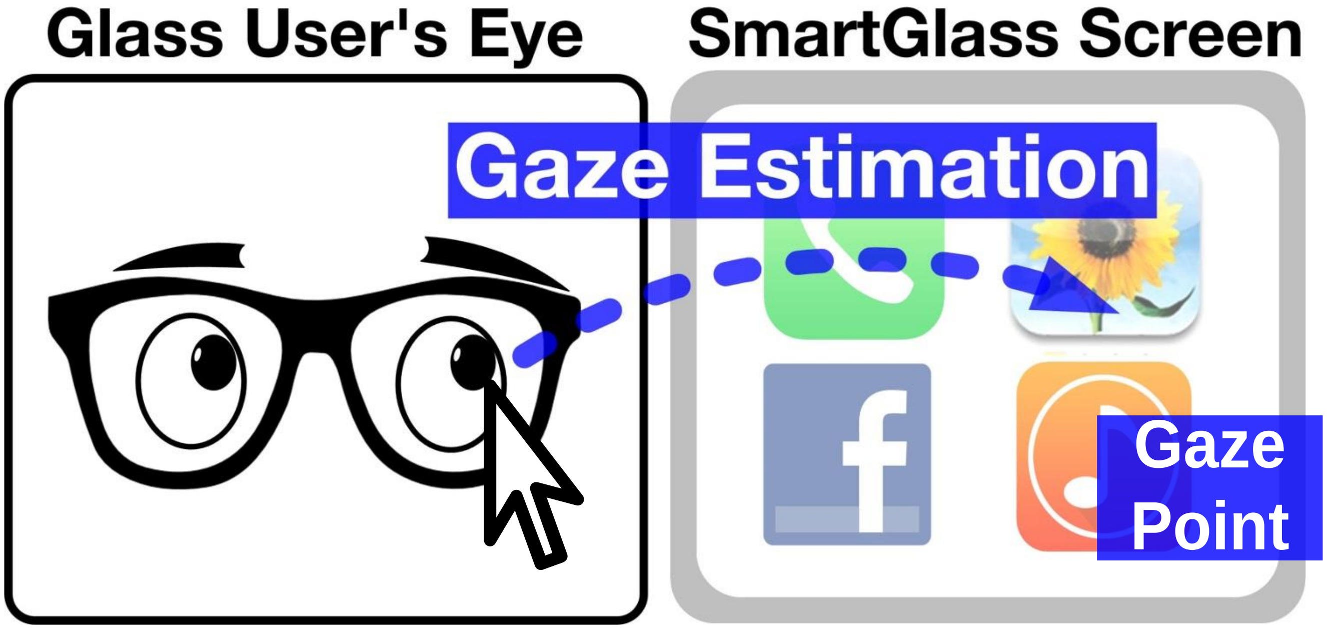
Demo: BONE-AR

K-GLASS

Real-Time Augmented Reality HMD

Gaze User Interface

Step1. “Point” the Cursor by “Gaze”



→ Low-power(<50mW [a]) Always-On Gaze UI

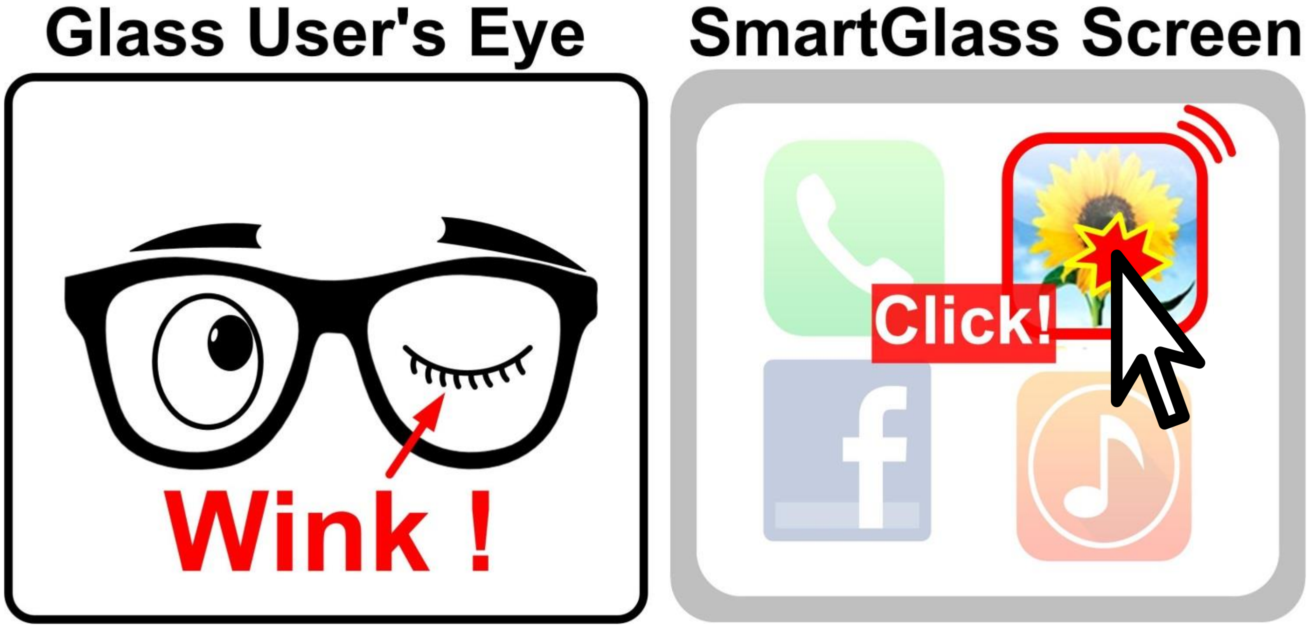
- Previous Work [b] : 100mW

[a] Power Consumption of Wireless Optical Mouse Reported in “Mouse Battery Life Comparison Test Report”, Microsoft Corp., 2004
[b] D. Kim, et al., “A 5000S/s Single-chip Smart Eye-tracking Sensor”, ISSCC 2008

[13] Injoon Hong et al, ISSCC 2015

Gaze User Interface

Step2. “Click” the Target by “Wink”

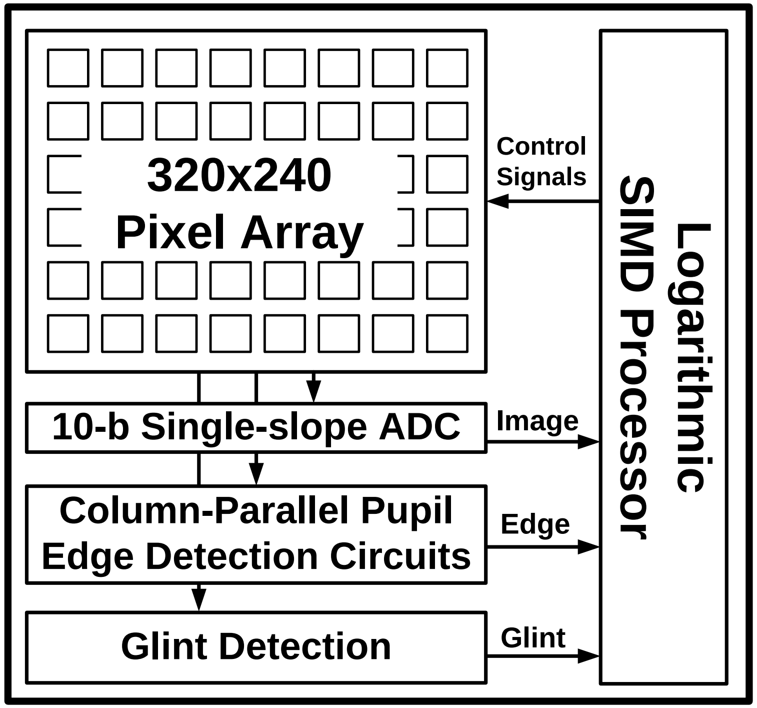


[13] Injoon Hong et al, ISSCC 2015

GIS & BONE-V8: System Architecture

1. 320x240 Gaze Image Sensor (GIS)
2. Multi-core OR Processor (ORP)

Gaze Image Sensor



320x240 Pixel Array

Control Signals

10-b Single-slope ADC

Image

Column-Parallel Pupil Edge Detection Circuits

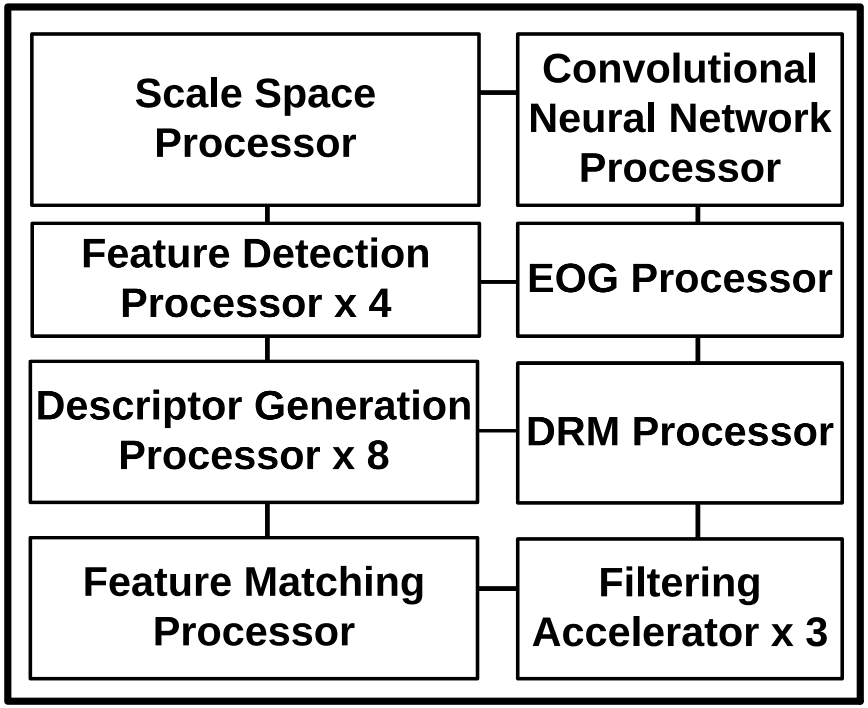
Edge

Glint Detection

Glint

Logarithmic SIMD Processor

Object Recognition Processor



Scale Space Processor

Convolutional Neural Network Processor

Feature Detection Processor x 4

EOG Processor

Descriptor Generation Processor x 8

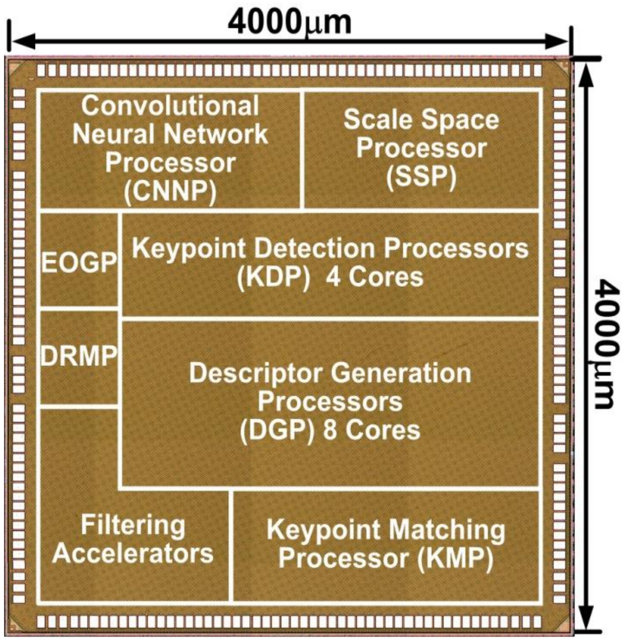
DRM Processor

Feature Matching Processor

Filtering Accelerator x 3

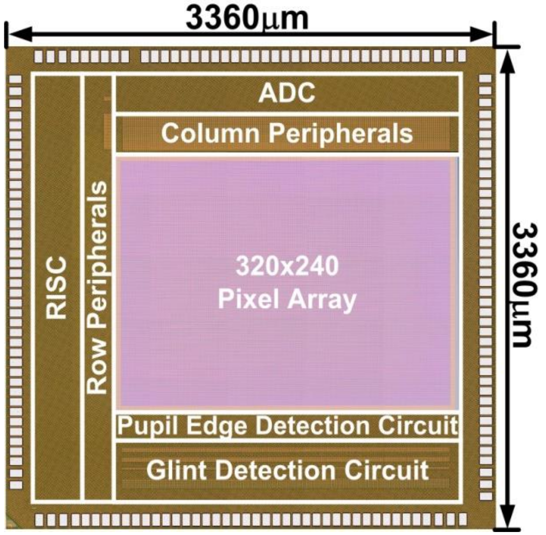
[13] Injoon Hong et al, ISSCC 2015

GIS & BONE-V8: Chip Implementation



4000µm
Convolutional Neural Network Processor (CNNP)
Scale Space Processor (SSP)
EOGP
Keypoint Detection Processors (KDP) 4 Cores
DRMP
Descriptor Generation Processors (DGP) 8 Cores
Filtering Accelerators
Keypoint Matching Processor (KMP)
4000µm

<OR Processor>



<Gaze Image Sensor>

- 4mm x 4mm ORP, 3.36mm x 3.36mm GIS in 65nm
- Real-Time (30fps) OR and Gaze Estimation Performance
- 75mW Average Power (65mW ORP, 10mW GIS)
- 131mW Peak Power (97mW ORP, 34mW GIS)

[13] Injoon Hong et al, ISSCC 2015

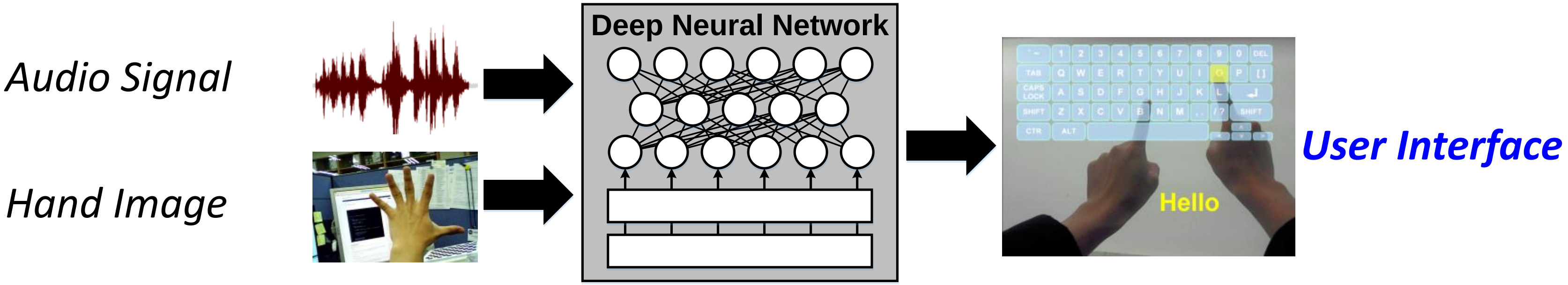
Demo: GIS & BONE-V8

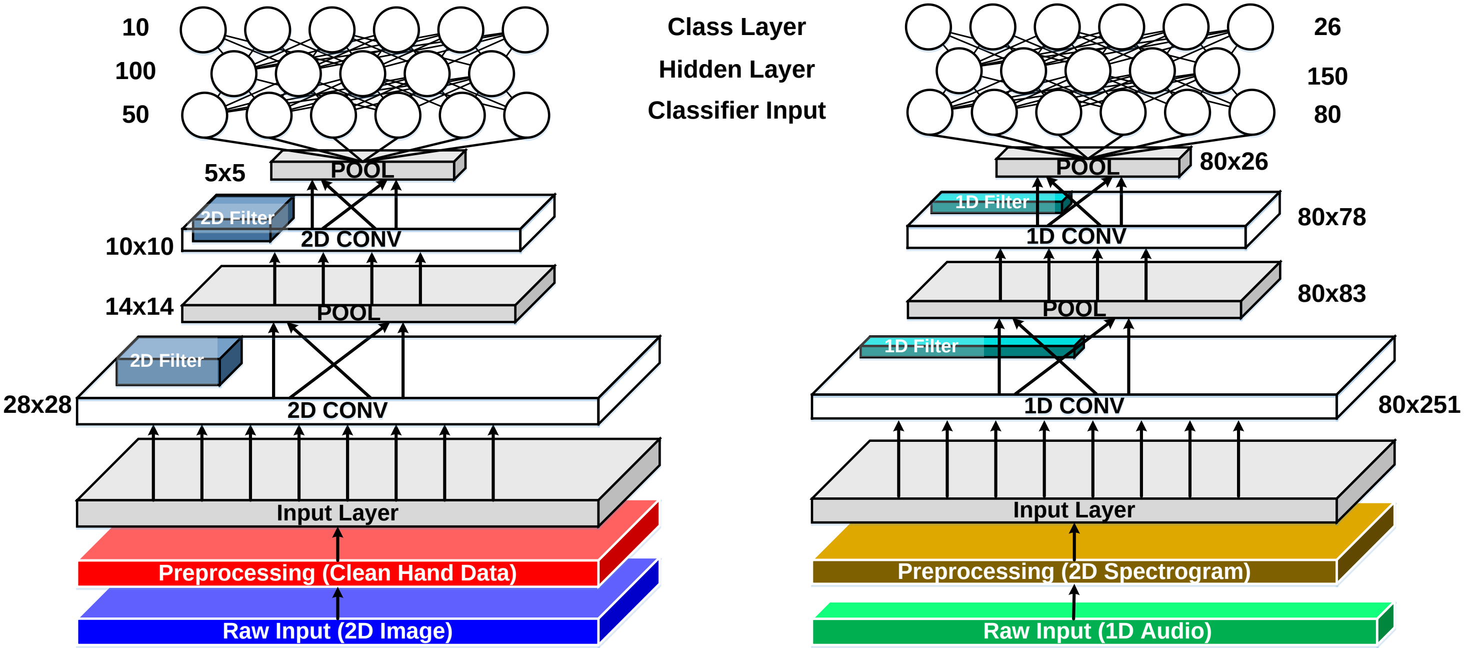
EyeClick

Gaze-Activated Object Recognition HMD

[13] Injoon Hong et al, ISSCC 2015

Multi-Modal UI/UX





<Gesture DNN>

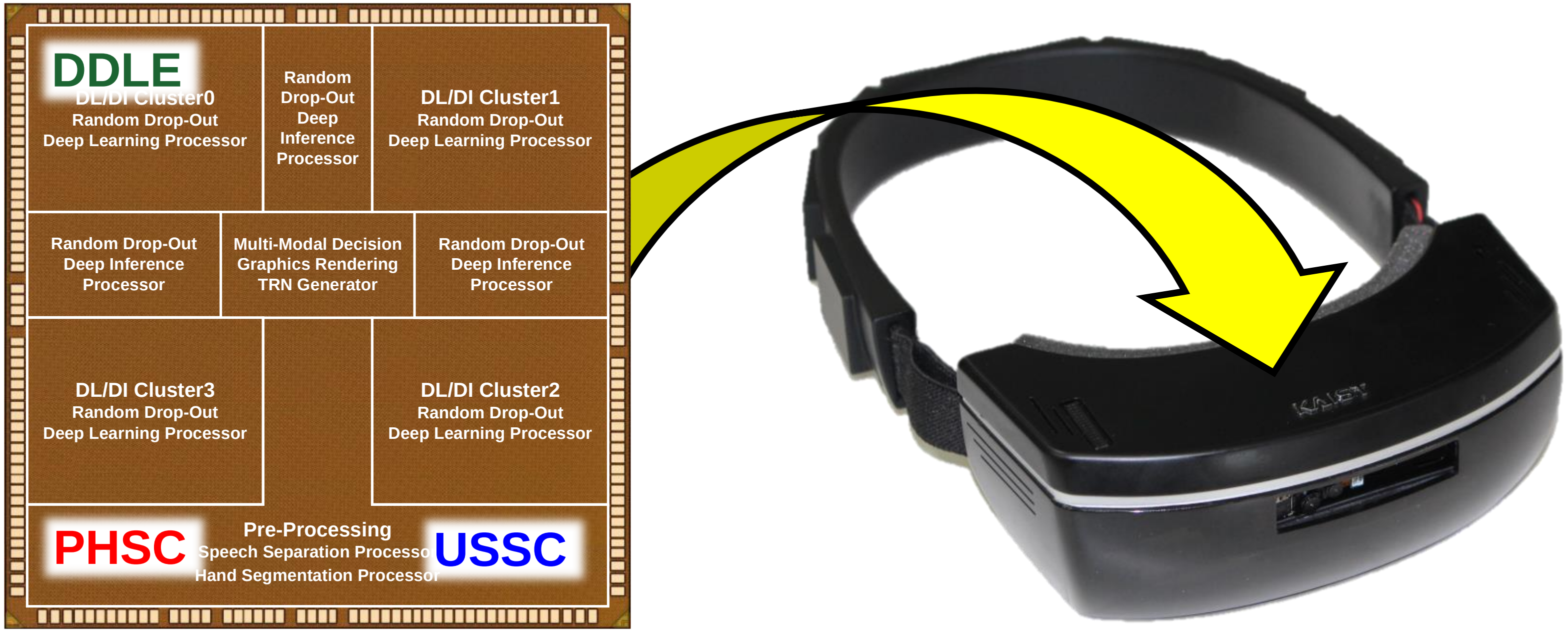
<Speech DNN>

[14] Seongwook Park et al, ISSCC 2016

UI/UX Specialized SoC

UI/UX Processor

K-Glass 3 System



- 4mm x 4mm by 65nm CMOS Technology
- 126mW Deep Learning Processor
- 1.80TOPS/W Smart Glasses Processor

[14] Seongwook Park et al, ISSCC 2016

Hoi-Jun Yoo

K-Glass 3 Demonstration Video

[14] Seongwook Park et al, ISSCC 2016

Summary

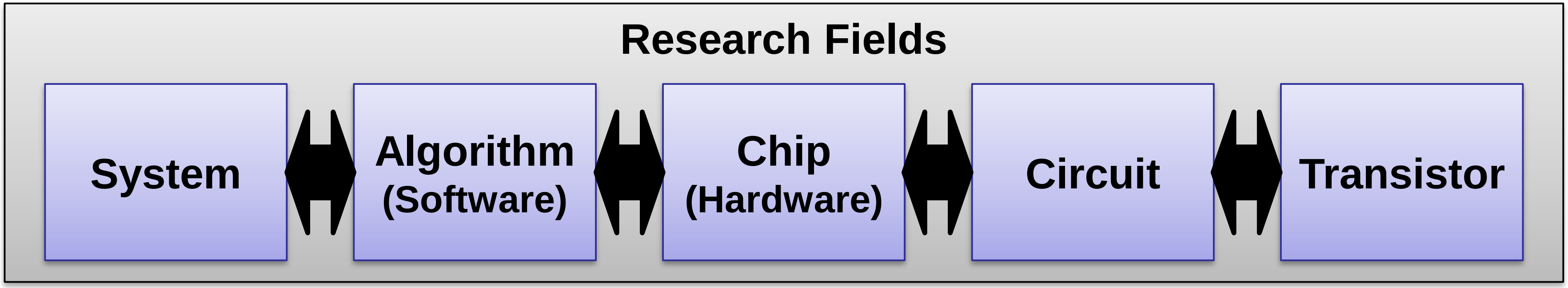
1. **Deep Neural Network Processor**
   - Mobile DNN Applications: AR, Drone, ADAS, Security
   - CNN Hardware Mapping and Architectures
2. **Mobile/Embedded Deep Neural Network**
   - Low-power, Real-time Operations
   - Limited Memory Capacity and Memory Bandwidth
   - Algorithm, Architecture, Circuit Level Optimization
   - Mobile/Embedded-CNNs – SRAM-based Architecture
3. **SoC Applications of Mobile/Embedded-DNN**
   - Hybrid CIS, CNNP Processor for Always-on Face Recognition System
   - DNPU Processor for General Purpose DNN System
   - Hybrid Intelligent Systems
   - AR Processor and UI/UX Processor for Smart Glasses

Collaborative Research

1. **Active Research Exchange Opportunities**
   - Academy-to-Academy
   - Industry-to-Academy

2. **National Funding & Exchange Programs**
   - National organization
   - Equal distribution of ownership policy

3. **Collaboration b/w Different Fields**



References

[1] Kyeongryeol Bong, et al. “A 0.62mW Ultra-Low-Power Convolutional-Neural-Network Face-Recognition Processor and a CIS Integrated with Always-On Haar-Like Face Detector”, ISSCC 2017

[2] Hiroki Nakahara, et al. “A deep convolutional neural network based on nested residue number system”, FBL 2015

[3] Bert Moons, et al. “A 0.3–2.6 TOPS/W precision-scalable processor for real-time large-scale ConvNets”, SOVC 2016

[4] Michael Figurnov, et al. “PerforatedCNNs: Acceleration through Elimination of Redundant Convolutions”, NIPS 2016

[5] Song Han, et al. “Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding”, ICLR 2016

[6] Yu-Hsin Chen, et al. “Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks”, ISSCC 2016

[7] Dongjoo Shin, et al. “DNPU: An 8.1TOPS/W Reconfigurable CNN-RNN Processor for General-Purpose Deep Neural Networks”, ISSCC 2017

[8] Kwanho Kim, et al. “A 125GOPS 583mW Network-on-Chip Based Parallel Processor with Bio-inspired Visual Attention Engine”, ISSCC 2008

[9] Seungjin Lee, et al. “24-GOPS 4.5-mm^2 Digital Cellular Neural Network for Rapid Visual Attention in an Object-Recognition SoC”, IEEE TNN 2011

[10] Seungjin Lee, et al. “A 345mW Heterogeneous Many-Core Processor with an Intelligent Engine for Robust Object Recognition”, ISSCC 2010

[11] Injoon Hong, et al. “A 1.9nJ/Pixel Deep Neural Network Processor for High Speed Visual Attention in a Mobile Vision Recognition SoC” , ASSCC 2015

[12] Gyonghoon Kim, et al. “A 1.22TOPS and 1.52mW/MHz Augmented Reality Multi-Core Processor with Neural Network NoC for HMD Applications”, ISSCC 2014

[13] Injoon Hong, et al. “A 2.71nJ/Pixel 3D-Stacked Gaze-Activated Object Recognition System for Low-power Mobile HMD Applications”, ISSCC 2015

[14] Seongwook Park, et al. “A 126.1mW Real-Time Natural UI/UX Processor with Embedded Deep-Learning Core for Low-Power Smart Glasses”, ISSCC 2016

[15] Seungjin Lee, et al. “The Brain Mimicking Visual Attention Engine: An 80x60 Digital Cellular Neural Network for Rapid Global Feature Extraction” , SOVC 2008

[16] Jinwook Oh, et al. “An Area Efficient Shared Synapse Cellular Neural Network for Low Power Image Processing” , VLSI-DAT 2009

[17] Youchang Kim, et al. “A 4.9mW Neural Network Task Scheduler for Congestion-minimized Network-on-Chip in Multi-core Systems” , ASSCC 2014

[18] Junyoung Park, et al. “A 92mW Real-Time Traffic Sign Recognition System with Robust Light and Dark Adaptation” , ASSCC 2011

[19] Minsu Kim, et al. “A 22.8GOPS 2.83mW Neuro-fuzzy Object Detection Engine for Fast Multi-object Recognition” , SOVC 2009

[20] Seungjin Lee, et al. “A 92mW 76.8GOPS Vector Matching Processor with Parallel Huffman Decoder and Query Re-ordering Buffer for Real-time Object Recognition” , ASSCC 2010

[21] Jinwook Oh, et al. “An Asynchronous Mixed-mode Neuro-Fuzzy Controller for Energy Efficient Machine Intelligence SoC” , ASSCC 2011

[22] Donghyun Kim, et al. “81.6 GOPS Object Recognition Processor Based on a Memory-Centric NoC”, TVLSI 2009

[23] Joo-Young Kim, et al. “A 66fps 38mW Nearest Neighbor Matching Processor with Hierarchical VQ Algorithm for Real-Time Object Recognition” , ASSCC 2008

[24] Minsu Kim, et al. “A 54GOPS 51.8mW Analog-Digital Mixed Mode Neural Perception Engine for Fast Object Detection” , CICC 2009

[25] Jinwook Oh, et al. “A 1.2mW On-Line Learning Mixed Mode Intelligent Inference Engine for Robust Object Recognition” , SOVC 2010

[26] Junyoung Park, et al. “Online Reinforcement Learning NoC for Portable HD Object Recognition Processor” , CICC 2012

[27] Injoon Hong, et al. “A 125,582 vector/s Throughput and 95.1% Accuracy ANN Searching Processor with Neuro-Fuzzy Vision Cache for Real-time Object Recognition” , SOVC 2013

[28] Kyuho Lee, et al. “A Vocabulary Forest-based Object Matching Processor with 2.07M-vec/s Throughput and 13.3nJ/vector Energy in Full-HD Resolution” , SOVC 2014

[29] Jinwook Oh, et al. “A 57mW Embedded Mixed-Mode Neuro-Fuzzy Accelerator for Intelligent Multi-core Processor”, ISSCC 2011

[30] Jinwook Oh, et al. “A 320mW 342GOPS Real-Time Moving Object Recognition Processor for HD 720p Video Streams”, ISSCC 2012

[31] Youchang Kim, et al. “A 0.55V 1.1mW Artificial-Intelligence Processor with PVT Compensation for Micro Robots”, ISSCC 2016

[32] Kyuho J. 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 2016