A 0.62mW Ultra-Low-Power Convolutional-Neural-Network Face-Recognition Processor and a CIS Integrated with Always-On Haar-Like Face Detector

*Kyeongryeol Bong*, S. Choi, C. Kim, S. Kang, Y. Kim, and Hoi-Jun Yoo

Semiconductor System Laboratory
School of EE, KAIST
Outline

- Introduction
- Proposed Face Recognition System
- Low-power Key Features
  1. Analog-Digital Hybrid Haar-like Face Detector (HHFD)
  2. Separable Filter approximation for convolutional layers (SF-CONV)
  3. Transpose-read SRAM (T-SRAM)
- Implementation Results
- Conclusion
DNN Researches

- **Cloud DNN**
  - GPU
  - 5.75 TOPS
  - ~250 W
  - NVIDIA Titan X

- **Edge DNN**
  - FPGA
  - ~1.2 TOPS
  - ~40 W
  - Arria 10 GX1150

- **Mobile DNN**
  - SoC
  - 0.1-0.3 TOPS
  - 90-400 mW

- **GPU**
  - 5.75 TOPS
  - ~250 W
  - NervanaSys-32

- **SoC**
  - 0.1-0.3 TOPS
  - 90-400 mW

© 2017 IEEE
International Solid-State Circuits Conference
More Smart Devices Per Person

- Always-on Face Recognition (AoFR) for “Unlock” (User Authentication)
  - Non-intrusive way for multiple devices

**<Smart Devices Per Person>**

<table>
<thead>
<tr>
<th>Year</th>
<th>Devices</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003</td>
<td>0.08</td>
</tr>
<tr>
<td>2010</td>
<td>1.84</td>
</tr>
<tr>
<td>2015</td>
<td>3.47</td>
</tr>
<tr>
<td>2020</td>
<td>6.58</td>
</tr>
</tbody>
</table>

**<User Authentication w/ Face>**
New UI for IoT

- Put Smart Face Recognition as New UI

Intelligence of Things

- Things can recognize users
- More interactive with users

Requirements

- Ultra-low power consumption
- Interactive intelligence
Low-power Face Recognition System

- Conventional FR System [1]

- Functional CIS + FR System

Transfer & Process only detected face images
Low-power Face Recognition System

- Proposed Hybrid Face Detector
  - Analog FD
    * Reject non-faced images instantly w/o initial processing
  - Digital FD
    * Detect face location w/ cascaded classifiers
  → Using Analog FD as Trigger for low-power FD
Ultra-low Power DNN

Algorithm
- Reduced Number Precision
- Tensor Decomposition

Architecture
- PE Array Optimization
- Distributed Memory Architecture

Circuit
- Customized SRAM
- Near-threshold Logic
Distributed Memory Architecture

Centered Global Memory Architecture
- SRAM as a global buffer
- Memory bandwidth
- Data reuse distance
- Architecture Scalability

Distributed Local Memory Architecture
- No global buffer
- Memory bandwidth
- Data reuse distance
- Architecture Scalability

This Work
Contribution of This Work

1. 24μW Always-on 320x240 CIS w/ FD
   1) Analog-digital hybrid Haar-like face detector (Low-power)

2. 0.6mW CNN Processor for FR
   1) Separable filter appr. for conv. layers (Low-power)
   2) Transpose-read SRAM (Low-power)
   3) Voltage & frequency scaling (Low-power)

0.62mW Low-power Always-on Face Recognition System
Overall System Architecture

1. 320x240 CIS with Always-on FD
2. CNN Processor (CNNP) for FR
Haar-like Filter
- Intensity summation over rectangles (black and white)

Cascade Classifier
- >10 Classifier Stages
- >1000 Haar-like filters

Check $I_{white} - I_{black} > I_{TH}$
Analog & Digital Haar-like Filtering

- **Digital Haar-like Filtering Unit (DHFU)**
  - Integral Image Gen.  Block sum becomes 3 ADD/SUB

- **Analog Haar-like Filtering Circuit (AHFC)**
  - Analog MEM: keep voltage intensities read from CIS pixels
  - Haar-like filt. circuit: charge-mode sum using analog MEM

\[ I_{int}(x,y) = \sum_{i=0}^{x} \sum_{j=0}^{y} I(i,j) \]

\[
\begin{align*}
D - C - B + A &= D - C - B + A \\
F - E - D + C &= F - E - D + C
\end{align*}
\]
Analog Haar-like Filtering Circuit

- 80x20 Analog Memory
  - Keep voltage intensities read out from pixels
  - 1 storage cap.
  - 1 unity gain buffer
  - 1 output cap. & several switches

- Analog Haar-like Filtering Circuit
  - Charge-mode summation
    - Black rectangle $\rightarrow C_{\text{black}}$
    - White rectangle $\rightarrow C_{\text{white}}$
Analog vs. Digital Haar-like Filtering

- **Digital Haar-like Filtering Unit (DHFU)**
  - Large initial processing (integral image gen.)
  - Simple operations for filtering

  ➔ Efficient for *long-lived* windows

- **Analog Haar-like Filtering Circuit (AHFC)**
  - Static current in analog MEM & filtering circuit
  - Direct processing (no initial processing)

  ➔ Efficient for *early-rejected* windows
Proposed Analog-Digital Hybrid FD

- **AHFC for 1st Stage**
  - Static current during only 1 stage

- **DHFU for Remaining Stages**
  - Integral image gen. for only passed windows
Energy-efficiency in Hybrid FD

- > 60% windows are rejected at 1st Stage
  - 60% workload reduction for integral image gen.
- With Hybrid FD, overall energy consumption is reduced by 40%

![Graph showing energy consumption comparison between Digital Processing and Hybrid Processing](image)
Separable Filter Approximation

Conv. with $2D \ d \times \ d$ Filter

Conv. with $1D \ d \times 1\ X$-Filter
+ Conv. with $1D \ 1 \times \ d\ Y$-Filter

$1-D\ X$-filter ($dx1$)

$1-D\ Y$-filter ($dx1$)

Workload Reduction w/ SF-CONV

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2D</td>
<td>60x60</td>
<td>5x5</td>
<td>48</td>
<td>192</td>
</tr>
<tr>
<td>#MAC</td>
<td></td>
<td></td>
<td></td>
<td>723M</td>
</tr>
</tbody>
</table>

<Conv. w/o Approximation>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Vertical (Ver.)</td>
<td>60x60</td>
<td>5x1</td>
<td>48</td>
<td>n:48</td>
</tr>
<tr>
<td>Horizontal (Hor.)</td>
<td>56x60</td>
<td>1x5</td>
<td>n:48</td>
<td>192</td>
</tr>
<tr>
<td>#MAC</td>
<td></td>
<td></td>
<td></td>
<td>183M</td>
</tr>
</tbody>
</table>

<Conv. w/ Separable Filter>
Workload Reduction w/ SF-CONV

- **2x ~ 3x** Workload (# of MAC) Reduction
  - Enables low-power CNN processing

- **< 1%** Accuracy Degradation

<table>
<thead>
<tr>
<th>Layer</th>
<th>Normalized # of MAC</th>
<th>Classification Accuracy*</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Appr.</td>
<td>1</td>
<td>97.6%</td>
</tr>
<tr>
<td>Conv2</td>
<td>0.57</td>
<td>97.4%</td>
</tr>
<tr>
<td>Conv2 &amp; 3</td>
<td>0.34</td>
<td>96.8%</td>
</tr>
</tbody>
</table>

*Tested with LFW (Labeled Face in Wild) dataset
Memory Access in SF-CONV

- Inefficient Memory Access for Image Fetching
  - Inefficient Vertical image filtering
  - 4.7x increased activity factor ($\propto P_{\text{Dyn}}$)

![Diagram showing memory access for horizontal and vertical filtering](image)

**<Hor. Filtering>**

**<Ver. Filtering>**

© 2017 IEEE International Solid-State Circuits Conference
Transpose-Read SRAM (T-SRAM)

- **V-WD & V-SA:** *transposed read/write*
  - Vertical wordline-driver & Sense amplifier
  - Output vertical 1-D vector

- **H-WD & H-SA:** *conventional read/write*
  - Output horizontal 1-D vector
SF-CONV w/ T-SRAM

Input Image

1-D Vert. Filtered Image

Output Image

1-D Vert. Filtering

1-D Hor. Filtering

Input Image in T-SRAM

1-D vert. filtered Image

Output Image

(H-SA) V-SA

1-D Vert. Filter

H-SA

V-WD

1-D Vert. Read

(Transposed Read)

V-WD

1-D Hor. Read

(Conv. Read)
Transpose-Read SRAM

- Two Read Modes
  - Conventional Read & Transpose Read

- 7T SRAM Cell
  - Single-ended read using **decoupled MOS**
  - **Bitline & wordline share** between conventional read & transpose read
Transpose-Read SRAM

- **Two Read Modes**
  - Conventional Read & Transpose Read

- **7T SRAM Cell**
  - Single-ended read using **decoupled MOS**
  - **Bitline & wordline share** between conventional read & transpose read
Memory Access in SF-CONV w/ T-SRAM

- Efficient Memory Access for Vertical Filtering Using Transpose Read
  - 4.2x decreased activity factor ($\propto P_{\text{Dyn}}$)

![Diagram](memory_access.png)
- **FD:** 3.3mm x 3.36mm CIS in 65nm
  - 320 x 240 pixel array, Haar-like filtering circuit & analog MEM
- **FR:** 4mm x 4mm CNNP in 65nm
  - 4 x 4 PE array with local T-SRAM
2x Energy Efficiency Improvement
- 211mW @ 100MHz, 0.8V
- 5.3mW @ 5MHz, 0.46V
System Implementation

Functional CIS

FD Layer

ULP-DNN Processor

FR Layer

FPGA
Experiment Results

- Only < 1% accuracy loss from SF-CONV
- 97% achieved using CNN @ LFW dataset
### Performance Comparison

<table>
<thead>
<tr>
<th></th>
<th>SOVC '15</th>
<th>This Work</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Technology</strong></td>
<td>TSMC 40nm</td>
<td>Samsung 65nm</td>
</tr>
<tr>
<td><strong>Algorithm</strong></td>
<td>FD: Haar-like Cascade Classifier</td>
<td>FD: Haar-like Cascade Classifier</td>
</tr>
<tr>
<td></td>
<td>FR: PCA + SVM</td>
<td>FR: CNN</td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td>81% @ 32-class in LFW</td>
<td>97% @ LFW</td>
</tr>
<tr>
<td><strong>Resolution</strong></td>
<td>HD</td>
<td>QVGA</td>
</tr>
<tr>
<td><strong>Framerate</strong></td>
<td>5.5fps</td>
<td>1fps</td>
</tr>
<tr>
<td><strong>Power</strong></td>
<td>23mW @ 600mV, 100MHz (w/o imaging)</td>
<td>0.62mW @ 460mV, 5MHz (w/ imaging)</td>
</tr>
</tbody>
</table>

- **The Most Accurate Face Recognition SoC**
  - x1.20 accuracy improvement with CNN (97%)

- **Ultra-Low-Power Consumption**
  - 0.62mW power consumption w/ imaging
Conclusion

- An Ultra-Low-Power Face Recognition System with CIS and CNN
- Ultra-Low-Power Features:
  1. Analog-Digital Hybrid Haar-like Face Detector
  2. Separable Filter Approximation for Conv. Layers
  3. Transpose-Read SRAM
  4. NTV Implementation with Voltage & Freq. Scaling

0.62mW Face Recognition System is Realized for *Always-on* and *IoT User Interface* Application