Extended Abstracts of the 2016 International Conference on Solid State Devices and Materials, Tsukuba, 2016, pp457-458

J-5-04

# Real-Time Haar-like Feature Extraction Coprocessor with Pixel-Based Pipelined Hardware Architecture for Flexible Low-Power Object Detection and Recognition

Aiwen Luo*, Fengwei An#, Yuki Fujita*, Xiangyu Zhang#, Lei Chen+, Hans Jürgen Mattausch*+

*Graduate School of Advanced Sciences of Matter, +HiSIM Research Center, #Graduate School of Engineering
Hiroshima University, 1-3-1 Kagamiyama, Higashi-hiroshima city, 739-8530, Japan
E-mail: aiwen-luo, anfengwei, fujita-yuuki, zhangxiangyu, chen, hjm@hiroshima-u.ac.jp

**Abstract**

**Image feature extraction is an important task for image or video analysis in machine vision. We report a coprocessor for Haar-like feature extraction with pixel-based pipelined hardware architecture. Instead of the conventional initial pixel-storage in large frame-buffer memories, the coprocessor synchronizes with image-sensor working frequency and immediately processes every input pixel for online feature construction. A prototype chip is fabricated in 180nm CMOS technology and requires only 12 KB on-chip memory for extracting the 1680-dimensional Haar-like SURF feature descriptor. During 30 fps VGA video input at 12.5 MHz frequency, it consumes only 4.78 mW power at 1.8 V supply voltage.**

## 1. Introduction

Feature-vector-based object recognition is widely used for many innovative applications, such as robotic and wearable vision or interactive games. Here a feature vector represents a reduced set of characteristic values extracted from an input image. Recognition applies such a reduced representation instead of the complete initial set of pixels. The existing feature descriptors, including Scale Invariant Feature Transform (SIFT) [1], and its successor Speeded Up Robust Features (SURF) [2], have demonstrated their robustness in many experiments and applications. Haar-wavelet responses are calculated for SURF-descriptor feature extraction. Unfortunately, traditional software implementation is not suitable for such complex feature extraction in efficiency-critical mobile applications, due to high resource and power consumption [3].

In this paper, we present a real-time cell-based Haar-like feature extractor with pixel-based pipelined hardware architecture in 0.18μm CMOS technology, having high flexibility for different size images, very high energy efficiency, and strongly reduced requirements for on-chip memory.

## 2. Cell-Based Haar-like Feature Extractor

We define an 8×8 pixel array as a cell which is divided into 4×4 pixel sub-cells. For each sub-cell, Haar-wavelet responses $D_x$ and $D_y$ in horizontal and vertical direction are calculated according to (1) and (2), respectively.

$$D_x = \sum_{x \in left\_sub-cell} p(x) - \sum_{x \in right\_sub-cell} p(x) \quad (1)$$

$$D_y = \sum_{y \in up\_sub-cell} p(y) - \sum_{y \in down\_sub-cell} p(y) \quad (2)$$

Here, p(x) and p(y) represent sub-cell-pixel values. Absolute values $|D_x|$ and $|D_y|$ are also determined to capture the polarity of intensity changes. Finally, the sub-cell results are added up to form a 4-dimensional Haar-like descriptor $\boldsymbol{v} = \{\Sigma D_x, \Sigma D_y, \Sigma |D_x|, \Sigma |D_y|\}$ for each cell.

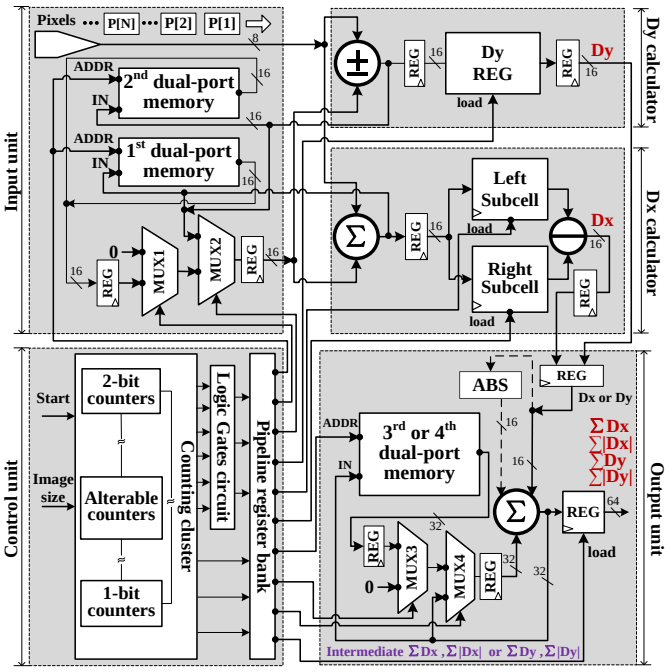


Fig. 1 Pixel-based pipeline (PBP) architecture for extracting Haar-like descriptor vectors of flexible-size images.

This paper reports the pixel-based pipelined architecture and its circuit implementation for such cell-based Haar-like features extraction, which can describe the target objects among complex backgrounds. Obtainable processing speed is only limited by the pixel frequency of the image sensor. For a search window (SW) of 64×128 pixels, a 1680-dimensional Haar-like feature vector is obtained, which contains the target-object information.

## 3. Pixel-Based Pipelined Architecture

The developed pixel-based pipeline architecture, shown in Fig. 1, has scalable input-image size and can capture target objects among complex backgrounds in real time. Processing speed is only limited by the pixel-transfer frequency from the image sensor. Alterable counters are included for enabling image-size flexibility. $D_x$ and $D_y$ of each sub-cell are calculated by the circuitry in the upper right of the Fig. 1. For this purpose, the pixel values of left and right half in one sub-cell are added up and stored temporarily in the 1st dual-port memory of the $D_x$ calculator. The final accumulation results for left and right part of each sub-cell are then loaded into the corresponding registers (i.e. Left_Subcell and Right_Subcell)

for calculation of the $D_x$ response. On the other hand, the $D_y$ calculator has a simpler structure than the $D_x$ calculator, due to line-wise input of the pixels from the image sensor. The upper half of each sub-cell is summed up and then the lower half is subtracted by an adder/subtractor circuit. The partial results of $D_y$ for each sub-cell are stored in the corresponding allocated space of the 2nd dual-port memory until $D_y$ completion and then $D_y$ is loaded immediately into register ($D_y$ REG). Once the last pixel of a sub-cell has been processed, the $D_x$ or $D_y$ result will be transferred to the output unit for calculating the feature vector $\boldsymbol{v}$ of the corresponding cell.

For an image with $n \times m$ pixels, only $n/2$ intermediate sub-cell summation results for $D_x$ in the 1st dual-port memory and $n/4$ intermediate sub-cell summation results for $D_y$ need to be stored in 2nd dual-port memory, respectively. The 3rd and 4th dual-port memories have to store just $n/8$ intermediate cell summation results of $\Sigma D_x$, $\Sigma |D_x|$ and $\Sigma D_y$, $\Sigma |D_y|$, respectively. In fact, the height of the input image is unlimited since the storage space of these four memories is reused for each cell row in the image. Once processing for a cell row is completed, the storage space can afterwards be overwritten by the data for the next cell row.

Rather than buffering entire images for integral image computing, the proposed architecture can therefore extract Haar-like descriptors in synchronization with image-sensor working frequency $\boldsymbol{f}$. The total image-processing time becomes consequently $(n \times m + DPR)/f$ seconds, where $DPR$ represents the clock-cycle delay of the pipeline registers. Furthermore, each cell-feature vector $\boldsymbol{v}$ is extracted only once and is then reutilized, according to the cell position in each related SW, for parallel feature-vector extraction of all these related SWs, without complex iteration or re-computation. This results in high processing speed, small memory usage and low power dissipation at the same time.

## 4. Measurement Results

The proposed coprocessor for Haar-like descriptors is prototyped in 180 nm CMOS technology with 1.76 mm$^2$ (1.82 mm×0.97 mm) core area, as shown in Fig. 2. The power consumption of the chip is 4.78 mW at 1.8 V supply voltage and 12.5MHz frequency, which is sufficient for 30 fps VGA-image processing from general image sensors. Fig. 3 demonstrates possible further power-consumption reduction to 1.57 mW at 1 V supply voltage and 12.5 MHz clock frequency. Correct chip operation up to 200 MHz is verified, indicating

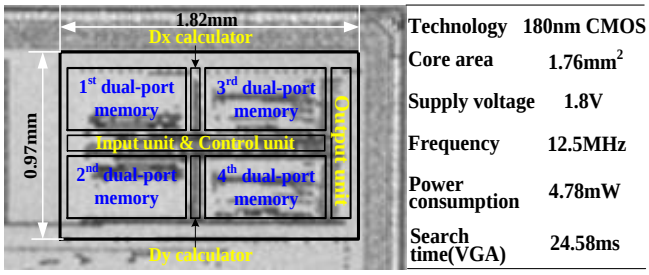


Fig. 2 Microphotograph and parameters of the fabricated chip in 180nm CMOS technology for cell-based Haar-like feature extraction with 1.76mm$^2$ (1.82mm×0.97mm) core area.

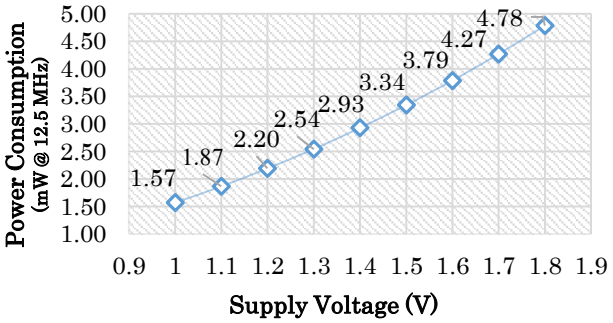


Fig. 3 Power consumption at 12.5 MHz working frequency with different supply voltage.

Table I Performance Comparison to previous work

| | This work | TCSVT[1] |
|---|---|---|
| Technology | 0.18 µm | 0.18 µm |
| Image size | VGA | VGA |
| Function | Haar-like hardware | SIFT hardware |
| Frequency | 100 MHz | 100 MHz |
| Power | 36.25 mW | - |
| Search windows | 851 | 151 |
| Processing time | 3.072ms | 8.397ms |
| Memory usage | 96 Kb | 5.729 Mb |

real-time processing capability beyond VGA size images.

Comparison to previous work [1] is shown in Table 1, verifying that the proposed architecture can extract the 1680-dimensional Haar-like feature vectors of all 851 SWs in a VGA frame with a factor 10 shorter processing time of 3.072ms (approximately 31.57ms would be consumed for 851 SWs in [1]) and a factor 60 smaller memory size of only 96 Kb. In addition, the image-size flexibility of the proposed architecture can be exploited for real-time processing of higher resolution images, e.g. XGA or Full HD, at 120 MHz operating frequency.

In conclusion, the proposed hardware architecture for extracting feature vectors of Haar-like descriptors has high image-size flexibility, very fast real-time processing speed, high energy efficiency, and low memory requirements, making it suitable for a large variety of different mobile applications.


## Acknowledgements

This research was supported by grant 25420332 from the Ministry of Science and Education, Japan. The VLSI-chip was fabricated through the chip fabrication program of VDEC, the University of Tokyo, in collaboration with, Rohm, Synopsys, and Cadence.


## References


[1] F.C Huang, S.Y Huang, J.W. Ker, Y.C. Chen. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 22, No. 3, 2012, pp.340–351.

[2] H. Bay, T. Tuytelaars, and L. V. Gool. SURF, European Conference on Computer Vision, (2006), pp.404–417.

[3] Joo-Young Kim, Minsu Kim, et al, A 201.4 GOPS 496mW Real-Time Multi-Object Recognition Processor with Bio-Inspired Neural Perception Engine, IEEE Journal of Solid-State Circuits. Vol.45, pp 32 - 45, Jan. 2010.