# A LOW-POWER VGA FULL-FRAME FEATURE EXTRACTION PROCESSOR

*Dongsuk Jeon, Yejoong Kim, Inhee Lee, Zhengya Zhang, David Blaauw, and Dennis Sylvester*

University of Michigan, Ann Arbor

## ABSTRACT

This paper proposes an energy-efficient VGA full-frame feature extraction processor design. It is based on the SURF algorithm and makes various algorithmic modifications to improve efficiency and reduce hardware overhead while maintaining extraction performance. Low clock frequency and deep parallelism derived from a one-sample-per-cycle matched-throughput architecture provide significantly larger room for voltage scaling and enables full-frame extraction. The proposed design consumes 4.7mW at 400mV and achieves 72% higher energy efficiency than prior work.

***Index Terms***— Feature extraction, Energy-optimal design, SURF

## 1. INTRODUCTION

Various feature extraction algorithms process input image or video to search for interest points based on different characteristics such as local gradient and edges. Each interest point is then described and a multi-dimensional feature vector is generated. Generally feature vectors extracted from an input video or image are compared against vectors already stored in a database and similar feature vectors will be matched.

Robust feature extraction can be used for various applications such as visual navigation, pose estimation, and object recognition. Most feature extraction algorithms incur high computational cost since a full image must go through multiple filters and often >1000 feature vectors can be extracted. Therefore, it is infeasible to implement these algorithms directly on hardware due to high peak performance requirements and large memory space for intermediate data.

However, an embedded system or platform with robust feature extraction enables a variety of useful applications. One example is a MAV (Micro Autonomous Vehicle) with autonomous navigation (Fig. 1). A video camera and embedded processor mounted on a MAV can extract feature vectors from the surrounding landscape. This can be used to determine posture or generate navigation maps by comparing against features extracted from previous frames or stored in a database.

Conventional hardware implementations are highly specialized for certain applications and extract features only from selected ROIs (Region of Interest) to significantly reduce computation. Although this enables energy-efficient feature extraction, a pre-processing algorithm that defines ROIs dominates overall feature extraction performance and must be very accurate. In addition, such approaches are not applicable to applications that require analysis of an entire image such as visual navigation.

In this paper we propose a highly energy-efficient VGA full-frame feature extraction processor based on SURF (Speeded-Up Robust Features). First, we optimize the original SURF algorithm to reduce computation and hardware cost significantly while maintaining extraction performance. Second, we propose a hardware architecture with matched-throughput and low clock frequency that leverages a highly parallelized data path. The proposed design is implemented in 28nm CMOS technology and simulations show it achieves 72% higher energy efficiency than prior work.

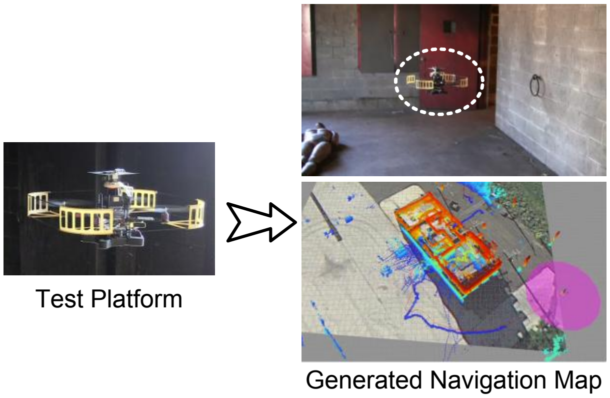

Figure 1. MAV with autonomous navigation system.

## 2. PRIOR WORK

There is active research on ROI-based feature extraction implementations to alleviate high peak performance requirements and enable low-power hardware implementations. Authors in [3] propose an object recognition SoC (System on a Chip) based on the SIFT (Scale-Invariant Feature Transform) algorithm and UVAM (Unified Visual Attention Model), which selects attentive points from input video. It also employs an analog-digital mixed-mode inference engine for initial ROI selection and 4 SIMD (Single Instruction, Multiple Data) and 32 MIMD (Multiple Instruction, Multiple Data) processing elements to perform feature extraction and matching. Reference [5] improves this further by proposing a CAVAM (Context-Aware Visual Attention Model) that considers temporal similarities between successive images as well. In [4], a full-HD wide viewpoint object recognition SoC based on SIFT is proposed. It simplifies the object matching stage by applying a vocabulary tree to characterize an object as a histogram vector. This reduces the otherwise prohibitive amount of memory accesses due to individual feature comparisons. Authors in [6] employ Haar-like features for long-range on-road object detection and use a knowledge-based Kalman object tracking scheme. Generally Haar-like features require significantly less computation than SIFT and this enables low-power operation with reasonable accuracy.

Although defining ROIs significantly reduces the computational workload, some applications require feature

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

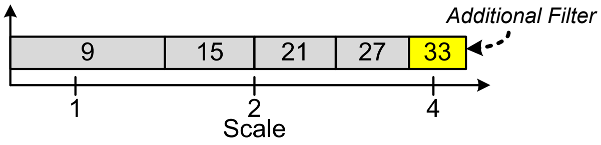

Figure 2. Proposed reduced-scale scheme.

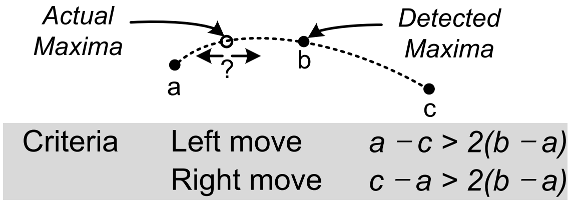

Figure 3. Proposed fast localization scheme.

extraction from the entire image. In addition, the accuracy of the pre-processing step that chooses ROIs directly impacts the performance of entire process and leads to lower overall extraction accuracy than a full-frame approach. Furthermore, a multi-core architecture generally requires a high-throughput communication network among multiple cores and suffers from substantial memory accesses to intermediate data. Also, the varying computational needs of feature extraction makes it necessary to design a system capable of high peak performance. Low-power circuit techniques suitable for variable throughput systems such as DVFS (Dynamic Voltage and Frequency Scaling) inevitably incur overhead for tasks such as voltage regulation.

## 3. PROPOSED APPROACH

To overcome these limitations, we propose an accelerator-based approach to feature extraction with a highly-parallelized architecture. We do not narrow the extraction scope and extract general features from an entire image. Therefore this accelerator is suitable for general applications including visual navigation and pose estimation of MAVs. Also, we co-optimize the hardware architecture and algorithm to maximize energy efficiency while maintaining performance.

### 3.1. Background: SURF algorithm

The SURF algorithm enables low power and low cost implementations in hardware, while providing similar or improved feature extraction performance compared to SIFT [1]. It first calculates a 2-D integrated image from an input image. This simplifies integration-based filter response calculations since integration over a specific regime can be replaced with 2 additions and 2 subtractions of the integrated image. This also significantly reduces the amount of memory accesses. SURF consists of two main steps: detection and description. The detector first applies LoG (Laplacian of Gaussian) box filters with different scales to the input image. It then searches for local maxima in 3×3×3 location-scale space, which represent interest points from the given image, and sends them to the descriptor. The descriptor applies Haar Wavelet Filters to the input image and gathers filter responses around each interest point. Finally it generates scale- and rotation-invariant vectors by orientation assignment and normalization.

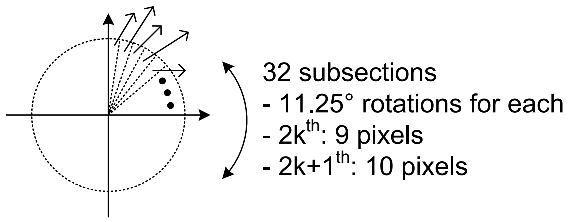

Figure 4. Proposed circular sampling region.

### 3.2. Modified SURF algorithm

Original SURF employs multiple octaves with each octave consisting of 4 different scales, enabling feature extraction from a wide range of scales. However, for the given 640×480 video size the number of detected features in higher octaves is relatively small and those features do not significantly contribute to scale-invariance performance. Therefore, we propose to use a single octave with an additional scale for filter size of 33 to compensate for any performance degradation due to the use of a single octave (Fig. 2).

After interest points are detected in the location-scale space, SURF finds the exact location of maxima by interpolating the determinant of the Hessian matrix as in [8]. Since each point is interpolated in 3×3×3 location-scale space, this requires multiple matrix arithmetic operations that are costly in terms of hardware implementation. To save power and area, we propose a fast localization scheme as shown in Fig. 3. If the actual maxima is expected to be very close to the right (left) of the detected interest point, then we simply choose the right (left) point as the interest point.

In the description stage of the original SURF algorithm, the orientation of each interest point is determined first. Haar wavelet filters are applied to the regions around each point and filter responses are gathered for a sliding orientation window at different angles. The orientation is found to be the angle of the sliding window with the largest summation of all responses in the window. Based on this orientation, the sampling region is then redefined as a shape of a rotated square. This incurs significant overhead since Haar wavelet filter responses must be computed twice; separately for orientation decision and feature description. To remove this overhead, we propose a circular sampling region divided into 32 subsections as shown in Fig. 4. Even if an orientation is not determined prior to filter response calculation, the sampling region is guaranteed to be the same owing to the circular shape. Therefore, we can first apply Haar wavelet filters and determine orientation based on responses gathered from each subsection. Then we have to post-process obtained responses based on the orientation accordingly to achieve rotation-invariance. Responses in each subsection are summed and as a result a single vector is produced for each subsection. The subsection with the largest vector magnitude is then chosen to represent the orientation of that vector. Finally, all vectors are merged together beginning from the largest to form a feature vector, and each vector is rotated based on the orientation found previously. Since the number of sampling points of 2k$^{th}$ subsections is different from 2k+1$^{th}$ subsections, responses are summed over 2 consecutive subsections (window size of 22.5°, moving 11.25° each step).

For better differentiation, we propose outer-product based vector separation. Original SURF gathers absolute values of x- and y- dimensions to double the dimension and enhance differentiation

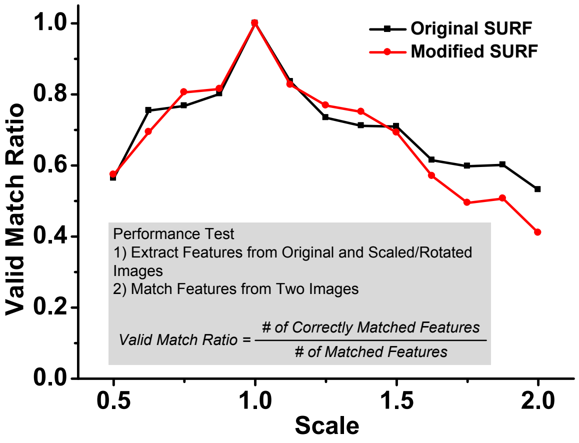

Figure 5. Scale-invariance performance comparison.

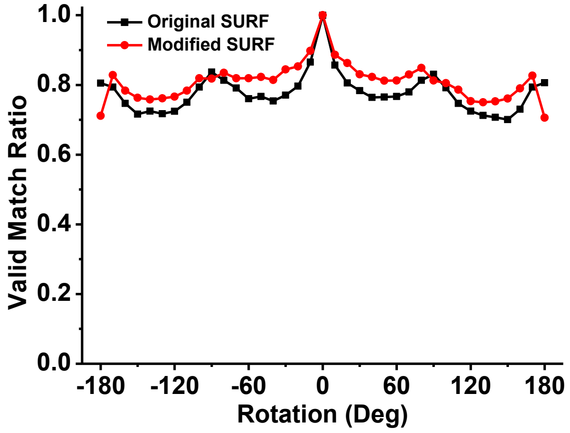

Figure 6. Rotation-invariance performance comparison.

performance. However, this can be achieved only when the orientation is known prior to vector summation over subsections,. This is not possible in modified SURF since only a single vector is stored for each subsection instead of entire filter responses after orientation assignment. To mitigate this issue, we propose to build two different dimensions for each subsection. First, the filter response at each sampling point is categorized based on the sign of x- and y- axis. If the sign of x-axis vector is the same as that of the y-axis, it is added to the first dimension. Otherwise, it goes into the second dimension. After an orientation is determined based on the subsection with largest vector magnitude, each dimension is rotated by the orientation angle. Finally, the sum of these two dimensions becomes the first half dimensions and the sum of absolute values of each dimension becomes the remaining half of the feature vector. The final feature vector has 32 dimensions. This modification reduces the required memory space by 89% for each description processing element.

Fig. 5 and Fig. 6 shows simulated scale- and rotation-invariance performance of original and modified SURF algorithms for images from test vehicle. Results clearly show that modified SURF provides similar performance to the original algorithm.

### 3.3. Optimized hardware architecture

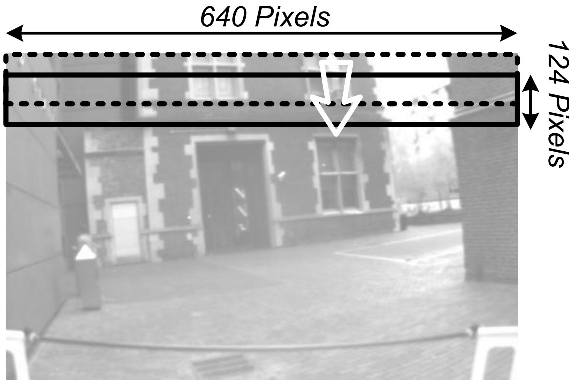

Figure 7. Overlapped processing subsections.

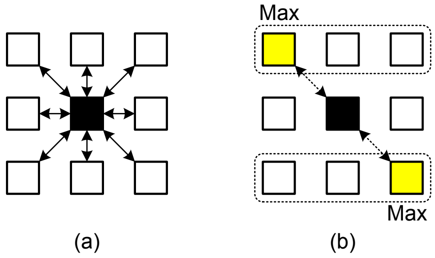

Figure 8. Proposed maxima detection scheme.

Voltage scaling is one of the most promising techniques for low power computation. However, low operating voltages significantly reduce performance and circuits must be carefully designed to meet a given performance requirement. Instead of a variable-throughput multi-core system, we propose to use a throughput-matched one-sample-per-cycle architecture optimized to maximize the benefits of voltage scaling. A highly-pipelined architecture with low clock frequency provides more room for voltage scaling [8]. Since FIFO (First In, First Out) is smaller, faster and more energy-efficient than SRAM for given size in subthreshold regime [2], an accelerator-based approach provides an improved hardware implementation compared to a multi-core architecture.

Since modified SURF has its largest filter size of only 33 pixels and does not require full image storage, an input image is divided into 11 subsections (Fig. 7) to reduce intermediate data storage. Subsections are overlapped by 88 pixels to guarantee feature extraction at borders. Fig. 9 shows the overall architecture. The detector first generates the 2-D integrated image from an original image. Integrator consists of only two adders and one 124-entry FIFO and produces one pixel of the integrated image per cycle in real-time. The integrated image goes through 5 different size filters that represent scales of 9, 15, 21, 27, and 33 pixels, respectively. Box LoG filters are implemented with multiple FIFOs to generate delayed images with different delays. A 3-D local maxima detector searches for local maxima in the 3×3×3 location-scale space. A total of 26 subtractions are used to determine if a given point is larger than all neighboring pixels. However, computation is reduced significantly by reusing previous results. At each cycle, the lower 3 pixels of each scale are processed and the location of the maximum value among them is attached to the lower middle pixel as an additional 2b. Then each cycle target point has to be compared against 8 pixels (maxima of each row) instead of 26 (Fig. 8). This technique reduces the number of comparisons by > 3×. Fast localization is then performed as described in the previous section and the interpolated location is sent to the descriptor.

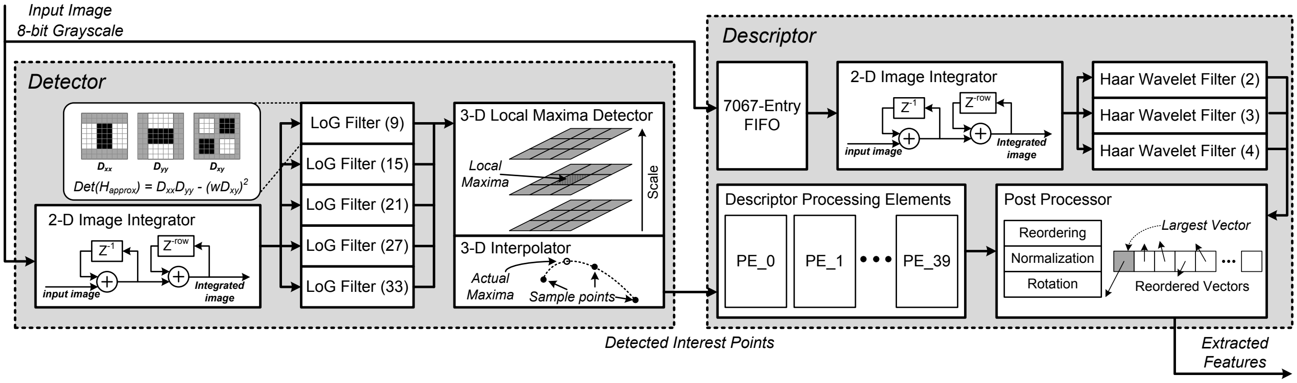

Figure 9. Proposed feature extraction processor architecture.

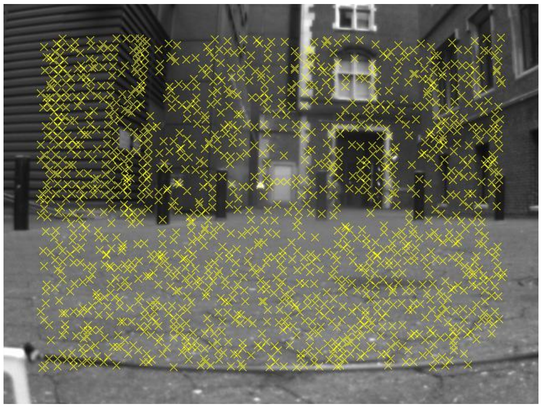

Figure 10. Test image with extracted features.

The descriptor first integrates the raw image to produce a 2-D integrated image. Although the two 2-D integrators in the detector and descriptor are identical, using separate modules saves 56% of FIFO buffer area since the integrated image has more than 2× larger bitwidth. The descriptor has a 7067-entry FIFO to delay the image while the detector processes the image and searches for interest points. It consists of three 113-entry FIFOs and eight 841-entry FIFOs. Three different Haar wavelet filters are applied to the integrated image and filter responses are shared across all processing elements, which greatly reduces communication overhead for multiple processing elements. A controller assigns each interest point to an idle PE (processing elements, 40 PEs in total). PEs that are not in use are power-gated to save active power. Finally, vectors generated from processing elements are post processed. The largest vector is determined from among 16 vectors and an orientation is assigned accordingly. Then the entire feature vector is re-ordered and rotated based on it.

## 4. EXPERIMENTAL RESULTS

The proposed feature extraction processor was implemented and simulated in a commercial 28nm CMOS technology. A standard cell library was re-characterized at the target operating voltage of 400mV. The proposed design was synthesized using Synopsys Design Complier and this re-characterized standard cell library. VCD (Value Change Dump) activity file was also extracted from Verilog simulation with a structural netlist. Total power consumption was measured with PrimeTime simulation that includes parasitic RC elements extracted from post-layout.

Table 1. Characteristic comparison with other fabricated feature extraction processors.

| | **Proposed** | [4] | [5] | [6] |
|---|---|---|---|---|
| Technology | **28nm** | 65nm | 130nm | 40nm |
| Design Target | **Feature Extraction** | Object Recognition | Object Recognition | Object Recognition |
| Base Algorithm | **SURF** | SIFT | SIFT | Haar-like |
| Extraction Scope | **Entire Frame** | ROI only | ROI only | ROI only |
| Input Video | **640×480** | 1920×1080 | 1280×720 | 1280×960 |
| Core Voltage | **400mV** | 1V | 0.7~1.2V | 0.9V |
| Power | **4.7 mW** | 52.5mW | 320mW | 69.3mW |
| Scaled Efficiency | **31.7TOPS/W** | 14.6TOPS/W | 7.8TOPS/W | 18.4TOPS/W |

Simulations show that the proposed design consumes only 4.7mW at 400mV with clock frequency of 27MHz while extracting features from 30 fps VGA video. Average performance is 149GOPS and power efficiency is 31.8 TOPS/W. Table 1 shows characteristic comparison with other fabricated feature extraction processors. The proposed design has significantly lower clock frequency compared to the >100MHz of other works [3-6] and achieves 72% better energy efficiency. Energy efficiency of different technology is scaled as (Scaled Efficiency = Reported Efficiency × Technology$^2$/28nm2 × Voltage/400mV). Figure 10 shows a test image with 1421 features extracted from proposed design. Interest points detected near edges are suppressed for robust feature matching performance.

## 5. CONCLUSIONS

In this paper, we proposed a highly energy-efficient feature extraction processor based on SURF algorithm. First, we modify the original SURF algorithm to maximize energy efficiency and make it suitable for hardware implementation while maintaining performance and accuracy. Second, a matched-throughput parallelized architecture with low clock frequency and operating voltage was proposed. These algorithm-architecture co-optimization techniques enable very low-power full-frame feature extraction and achieve 72% better energy efficiency than prior state-of-art.

## 6. ACKNOWLEDGEMENTS

Funding support by the Army Research Laboratory and test support of Nathan Michael and Shaojie Shen at the University of Pennsylvania are gratefully acknowledged.

## 7. REFERENCES

[1] H. Bay et al., "SURF: Speeded Up Robust Features," *Computer Vision and Image Understanding, Vol. 110, No. 3, pp. 346-359, 2008.*

[2] M. Seok et al., “A 0.27V 30MHz 17.7nJ/transform 1024-pt complex FFT core with super-pipelining,” *IEEE International Solid-State Circuits Conference, 2011.*

[3] S. Lee et al., “A 345mW heterogeneous many-core processor with an intelligent inference engine for robust object recognition,” *IEEE International Solid-State Circuits Conference, 2010.*

[4] Y.-C. Su et al., “A 52mW full HD 160-degree object viewpoint recognition SoC with visual vocabulary processor for wearable vision applications,” *IEEE Symposium on VLSI Circuits, 2011.*

[5] J. Oh et al., “A 320mW 342GOPS real-time moving object recognition processor for HD 720p video streams,” *IEEE International Solid-State Circuits Conference, 2012.*

[6] Y.-M. Tsai et al., “A 69mW 140-meter/60fps and 60-meter/300fps intelligent vision SoC for versatile automotive applications,” *IEEE Symposium on VLSI Circuits, 2012.*

[7] M. Brown and D. Lowe, “Invariant Features from Interest Point Groups,” *British Machine Vision Conference, 2012.*

[8] D. Jeon et al., “A Super-Pipelined Energy Efficient Subthreshold 240MS/s FFT Core in 65nm CMOS,” *IEEE Journal of Solid-State Circuits, Vol. 47, No. 1, pp. 23-34, 2012.*