A 176×120 Pixel CMOS Vision Chip for Gaussian Filtering with Massivelly Parallel CDS and A/D-Conversion

Manuel Suárez
Víctor M. Brea
Diego Cabello
Centro de Investigación en Tecnoloxías da Información (CITIUS)
University of Santiago de Compostela
Santiago de Compostela, Spain
Email: manuel.suarez.cambre@usc.es

Jorge Fernández-Berni
Ricardo Carmona-Galán
Ángel Rodríguez-Vázquez
Centro Microelectrónica de Sevilla CNM-IMSE
University of Seville
Seville, Spain

***Abstract*—This paper conveys a proof-of-concept chip for Gaussian pyramid generation for image feature detectors. Gaussian filtering and image resizing are performed with a switched-capacitor (SC) network. The chip is conceived as the mapping of a CMOS-3D architecture for feature detectors onto a conventional technology, with some functionality removed, and the corresponding area overhead with respect to a CMOS-3D architecture, but preserving masivelly parallel Correlated Double Sampling (CDS) and A/D conversion. The chip has been fabricated on a die of 5×5 mm^2 with 0.18 μm CMOS technology, achieving an array of 176×120 sensing elements (pixels). The pixels are arranged in Processing Elements (PEs). Every PE comprises four photodiodes, four SC nodes, one CDS circuit, and local circuitry for one ADC. Every PE occupies an area of 44×44 μm^2. The chip senses an image and computes the Gaussian pyramid with an average power consumption lower than 75 nW/pixel at 30 frames/s.**

I. Introduction

The advances in the computational power of the last years allow doing tasks as traffic or citizen control, surveillance, robot guidance or augmented reality in reasonable computing time. Feature detectors have become common place in these applications. The Scale Invariant Feature Detector (SIFT) [1] is a state-of-the-art feature detector. As any other feature detector, SIFT comprises low- and intermediate-level processing stages. The first stage is the generation of a Gaussian pyramid or scale-space. The scale-space generation is defined as the sucessive Gaussian filtered versions of the incoming image. The incoming image is filtered with rising up widths (σ in the Gaussian filter), also called scales (S) to provide one set (octave) of filtered images. This process is done O times to get O octaves with S scales each. The origin of a new octave is one half-sized reduction of a previous octave.

Low-level image tasks like convolution-type operations as Gaussian pyramids are better suited for a Single Instruction Multiple Data (SIMD) architecture with a Processing Element (PE) per pixel. The Gaussian filtering is naturally performed by both a resistive-capacitor (RC) grid, or a switched-capacitor (SC) network [2], [3]. These solutions outperform some other paradigms like those based on cellular non-linear networks [4].

Intermediate-level stages work on a reduced set of pixels. In SIFT, this stage works on the extrema, obtained from the Gaussian pyramid. The extrema amount to the 1% of the pixels in the image [1]. In this case another kind of parallelism arises, and the digital domain emerges as a better solution to perform more complex functions like feature description.

This paper addresses a proof-of-concept chip conceived as the mapping of a CMOS-3D architecture for feature detectors onto a conventional CMOS technology. The result is a chip with some functionality removed, and area overhead with respect to a CMOS-3D architecture. The chip, however preserves the massive parallelism for CDS and A/D conversion, with an assignment of 4 sensing elements to 4 nodes of an SC network for Gaussian pyramid, and one CDS stage, and the local circuitry for A/D conversion.

II. Vision Chips for Feature Detectors on CMOS-3D Architectures

In the realm of feature detectors, CMOS-3D technology comes out as a possible solution to embed the whole algorithm onto a single die thanks to the distribution of functionality across different tiers interconnected with the so-called Through-Silicon-Vias (TSV) [5]. The spread of functionalities as shown in Fig. 1(a) might permit to optimize functions across different image processing levels, or within a given level itself without degrading fill-factor or foodprint. For instance, if several tiers were available, a specific tier could be left for sensing in order to enhance certain characteristics like dynamic range or spectral response. Also, a second layer with PEs that include CDS or in-pixel ADC, a third layer to store a digital image in a frame buffer with circuitry for extrema detection, or even a fourth tier for higher-level processing could be incorporated. Such a CMOS-3D architecture would provide more parallelism than that of a conventional CMOS imager, which usually counts on per-column CDS and ADC circuits.

III. Vision Chip for Feature Detectors on Conventional CMOS Technology

The chip addressed in this paper conveys an array of 176 x 120 pixels in a conventional 2D CMOS UMC 0.18 μm technology. The layout of the chip is displayed on Fig. 1(b). The chip is manufactured on a 5×5 mm^2 die. Every PE in the

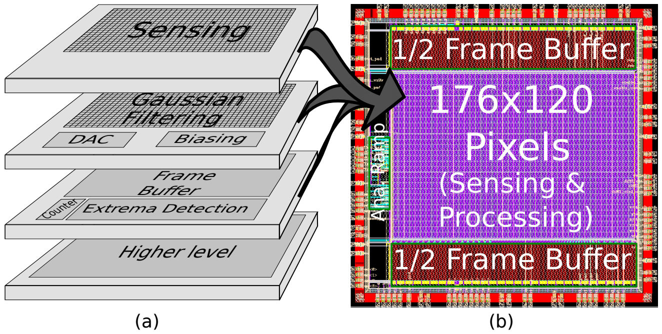

Fig. 1. Several functions of a feature detector on a CMOS-3D stack (a). Gaussian filtering with massively parallel ADC and CDS mapped onto conventional CMOS technology. Extrema location has not been mapped (b).

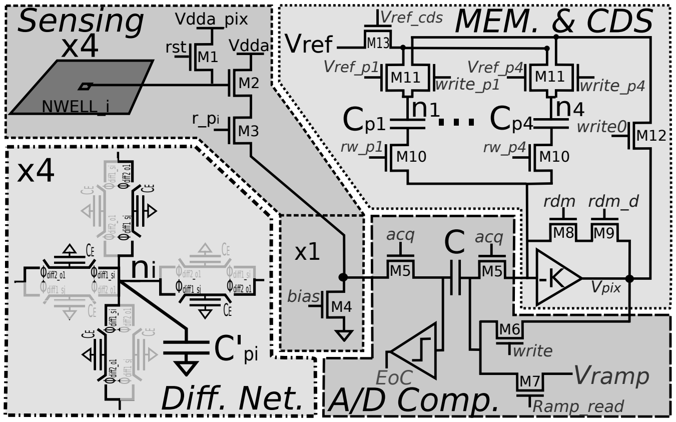

Fig. 2. Scheme of the Processing Element (PE) of the chip.

array comprises 4 photodiodes, 4 nodes of an SC network for Gaussian filtering, one CDS circuit and one comparator, the latter being part of an 8-bit single-slope ADC. Furthermore, the area constraint obliges to reuse circuitry between Gaussian filtering, CDS and A/D conversion. The registers that complete the ADCs are laid down outside the array, and labeled as 1/2 frame buffer in Fig. 1(b). The analog ramp for the ADCs along with additional biasing circuits are placed outside the array too. This floorplan resembles an approach with two tiers in a CMOS-3D technology, with the upper tier for sensing and Gaussian filtering, and the bottom tier for the frame buffer of the ADC, both connected with TSVs. The key difference between the chip addressed in this paper and a chip over a CMOS-3D stack lies in the routing overhead between the PE array and the frame buffer, as in CMOS-3D technology a TSV per PE would connect the frame buffer with the comparator to complete the ADC, lowering the fan-out.

A. Processing Element (PE)

The PE's architecture comprises four main blocks: i) 4 *3T-APS* (3 Transistors - Active Pixel Sensor) structure, ii) 4 capacitors C_{pi} with an inverter as gain stage to work as *memories* and realize the CDS, iii) a *switched-diffusion network* for Gaussian filtering with communication with the neighbors along the four cardinal directions, and iv) a capacitor C for CDS that is reused together with a comparator for *A/D conversion*. Fig. 2 shows a PE circuit, whilst the sizes of transistors and photosensors are listed in Table I. Every PE occupies an area of $44 \times 44 \mu m^2$. Concerning functionality, the main functions performed by the PE are: i) image acquisition and CDS, ii) Gaussian Filtering, and iii) A/D conversion.

TABLE I. PE TRANSISTOR SIZES (IN MICRONS).

	Width	Length		Width	Length
Photodiode	7.4	6.7	M1	0.24	1
M2	1.6	0.3	M3	0.24	0.6
M4	0.6	0.8	M5	0.24	1.4
M6	0.24	0.8	M7	0.24	1
M8	0.24	0.3	M9	0.24	0.8
M10	0.24	0.2	M11	0.24	0.8
M12	0.24	0.1	M13	0.24	0.4

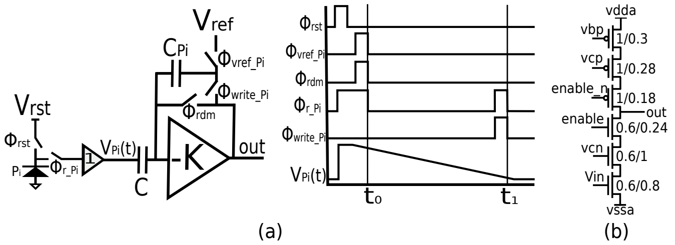

Fig. 3. (a) CDS stage with its time diagram. (b) Amplifier ($-K$) with the transistor dimensions in microns.

1) Image acquisition and CDS: The image acquisition in a PE is done by 4 photodiodes with 4 source followers and 4 selecting transistors, but only one current source shared by the 4 photodiodes biased at $1\mu A$ (Fig. 2). The sensing element is an n-well diode to enhace the spectral response at longer wavelengths. The source follower is designed to achieve the largest operating range through the use of low threshold voltage transistors and hard reset with a voltage supply $vdda\_pix$ higher than the power supply $vdda$. The source follower provides a gain spread less than 0.4% with an operation range of 1V, and an average power consumption of 2.5 nW per PE at 30 frames/s (13.2 μW for the whole array).

The analog memories highlighted by the upper polygon in Fig. 2 are built with 4 state capacitors $C_{pi} = 200fF$, and one shared inverter with gain $-K$. The capacitors are MIMcaps on Metal5 and Metal6 layers. These memories have different functions along the image processing path: CDS, image storage and Gaussian filtering. This stage works together with the 3T-APS structures and a capacitor $C = 200fF$ to remove the Fixed Pattern Noise (FPN) and other low frequency noise components of the sensed signal through the CDS circuit shown in Fig.3. Eq. (1) is the output of the CDS. V_{ref} is fixed to 400 mV to ensure the saturation region of the inverter.

$$V(C_{Pi}) = V_{ref} + \frac{C}{C_{Pi}}[V_{Pi}(t_0) - V_{Pi}(t_1)] \qquad (1)$$

The inverter has been designed with a double cascode configuration in order to achieve a high nominal gain of 65 dB, required for low linearity errors. The schematic of the inverter with its transistor sizes in microns is displayed on Fig. 3(b). The bias voltages for the cascode inverter are $vbp = 1.2V$, $vcp = 0.95V$ and $vcn = 0.65V$ respectively, providing a bias current $I = 1\mu A$. Additional transistors *enable* and *enable_n* permit to cut down to zero the static power consumption during standby periods (leakage currents neglected). As seen below, the gain stages in the comparator for A/D conversion use the same design. The double-cascode inverter has an average

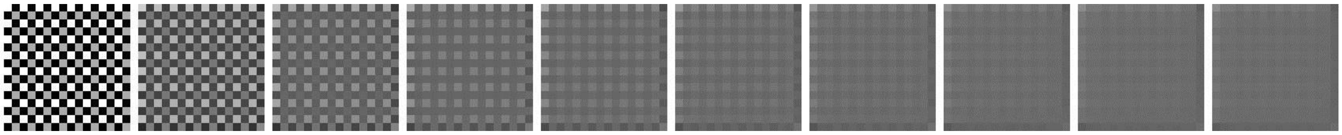

Fig. 4. Sensed image and 9 successive Gaussian filtering cycles from simulated results of 16×16 pixels.

power consumption of 15.4 nW at 30 frames/s due to the fact that they are off during the 85% of the computing time.

2) Gaussian Filtering: The Gaussian filtering and the octaves generation are implemented by a switched-capacitor network [3]. The switched-capacitor network minimizes the non-linearity of a conventional RC network implemented by MOS transistors [2],[3]. On the other hand, they permit a more accurate control of the Gaussian width σ by the number of switching cycles. Eq. (2) represents the behavior of one node of the network in a given diffusion cycle n. An extra capacitor $C'_{pi} = 130$ fF (see Fig. 2) implemented as an MOS capacitor has been added in parallel with the MIMcaps in order to reduce switching and leakage errors, turning that the capacitance associated with an SC node ij in the array to $C_{ij} = C_{p_{ij}} + C'_{p_{ij}}$. The exchange capacitor C_E, also implemented with an MOS transistor, has a value of 28.5 fF.

$$V_{ij}(n) = V_{ij}(n-1) + [V_{i-1j}(n-1) + V_{i+1j}(n-1) + \\ + V_{ij-1}(n-1) + V_{ij+1}(n-1) - 4V_{ij}(n-1)]\frac{\frac{C_E}{C_{ij}}}{1+4\frac{C_E}{C_{ij}}} \quad (2)$$

From Eq. (2), the σ_0 width per iteration is given by Eq. (3).

$$\sigma_0 = \left(2 \times ln\frac{C_{ij}}{C_E}\right)^{-1/2} \quad (3)$$

The diffusion network is performed by a double Euler configuration, highlighted in the bottom-left hand-side of Fig. 2 [6]. The Gaussian width per Gaussian filtering or diffusion cycle is $\sigma_0 = 0.48$. Fig. 4 illustrates 9 diffusion cycles for a 16×16 image from simulated PEs with an asymptotic $RMSE = 0.6\ LSB$.

3) A/D Conversion: As mentioned before, the in-PE single-slope ADC is distributed within and outside the pixel array. The comparator is located within every PE, while the frame buffer is outside the pixel array. Global circuitry for the generation of the single-slope or analog ramp along with biasing circuitry are also needed to complete the ADC. The function of the ADC is to digitize either the input image or the scales to perform extrema detection.

The comparator is shown in Fig. 5. It is an offset-compensated topology with two gain stages $(-K)$ implemented with a double cascode structure with the same biasing and transistor sizes as those of the amplifier used for CDS. The capacitor C used for offset compensation is shared with the CDS stage (see Fig. 3). Signals $comp\_rst$ and $comp\_rst\_d$ are used to apply the bottom sampling technique to cancel offset, leading to the output of the first inverter given by:

$$inv1 = V_Q + K(V_{pix} - V_{ramp}) \quad (4)$$

with V_Q being the quiescent point of the first inverter, and V_{pix} and V_{ramp} being the signal acquired by the photodiode or a given scale S, and the ramp of the 8-bit single-slope ADC, respectively.

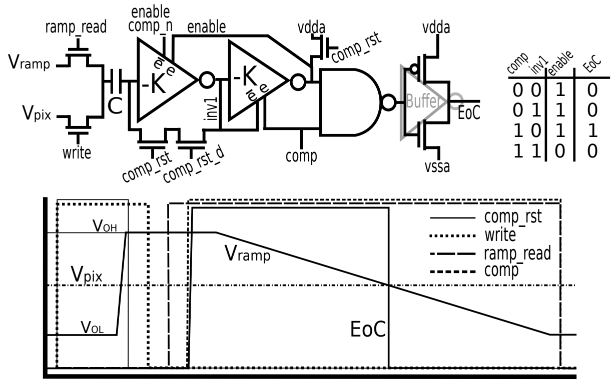

Fig. 5. Offset-conpensated comparator used in the proof-of-concept chip with feedback loop for gain enhancement and power save.

The A/D-conversion finishes when signal EoC goes down to the $'0'$ logic state. This happens with the transition of the ouput of the first inverter $(inv1)$ from $'0'$ to $'1'$, leading the $enable$ input of the first inverter to $'0'$ through the feedback loop from the second inverter. The complementary $enable$ input of the second inverter is also tied to logic state $'1'$ through $inv1$. The feedback loop reinforces the logic states of both inverters after the zero crossing between V_{pix} and V_{ramp}, and it also cuts down to zero the static power consumption of the two inverters. Finally, the NAND gate with the signal $comp$ to $'0'$ produces a signal EoC tied to zero, avoiding writing into the frame buffer while the comparison is not taking place. The comparator is the most expensive stage in power consumption per PE, with an average value of 265 nW/PE at 30 frames/s.

The frame buffer stores the scales or the input image generated in the pixel array. The signal EoC provided by the comparator at every pixel finishes the reading of the registers, storing the 8-bit word that unleashes the conversion. The 8-bit word is generated by an 8-bit counter implemented with a D-type flip-flops.

The frame buffer is laid down in the top and bottom sides of the die, in such a way that the upper 60 pixels drive the top frame buffer, while the remaining 60 lower pixels are read out by the bottom frame buffer. This floorplan minimizes routing. Also, as seen in Fig. 6, every 1/2 frame buffer (see Fig. 2) comprises 352 x 15 8-bit registers. In turn, every 1/2 frame buffer is split into two regions of 176 x 15 registers each. As there is only one comparator or ADC per 4 pixels or photodiodes, 4 bursts are needed to complete the reading of the array. The split of the 1/2 frame buffers into two regions allows for reading 176 x 15 pixels into the registers at the same time that 176 x 15 pixels are read out of the chip.

The ADC is complete with an 8-bit Digital-to-Analog

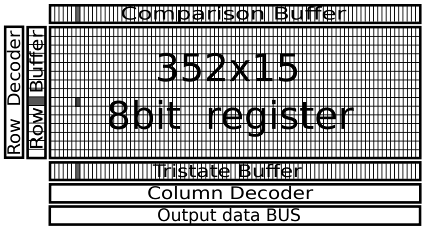

Fig. 6. Frame buffer or registers for the 8-bit single slope ADC.

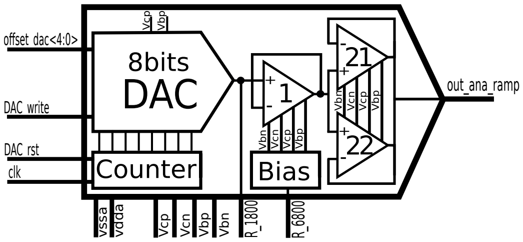

Fig. 7. Analog ramp generator made up of a thermometric current steering DAC with its buffers to drive the 176 × 120 pixel array.

Converter (DAC) and its corresponding buffers to provide the single-slope ramp to the comparator. Fig. 7 shows the schematic of the analog ramp generator. This is implemented with a current steering thermometer DAC driven by a counter, with the current sources implemented with cascode stages biased at 2 μA. The currents are transformed into voltages in an external resistor of R=1.8 $K\Omega$, providing a voltage of 3.6 mV per LSB. The DAC includes a circuit to calibrate offset with 5 bits of resolution, while the gain errors can be minimized by tuning the external resistor. At simulation level the non-linearity errors were $INL = \pm 0.19$ LSB, and $DNL = \pm 0.006$ LSB. The ramp generated by the DAC is buffered to the pixel array with two stages of folded cascode OTAs, labeled '1', and '21' and '22' in Fig. 7, and biased at 50 μA, and 600 μA respectively.

B. Chip Comparison

Although there are SIFT implementations over programmable hardware like FPGAs or GPUs, we focus on the most recent ASICs to compare with the chip described in this paper. Table II summarizes the most relevant performance metrics. It should be noted that the chips in [7] and [8] are implemented in the digital domain, and that both use additional strategies to drop power consumption and computation time. For instance, the chip in [8] utilizes a visual attention algorithm to run SIFT only on the image regions with relevant information. It splits an HD 720p image into 3600 tiles of 16 × 16 pixels each, usually processing less than 1/3 of the pixels in the image, which enables video frame rate. The data in the first column of Table II, however, refers to a case where all the pixels of an HD 720p are computed, to set a fairer comparison with our approach. The available data for the chip of reference [7] refers to the detection of only one feature. It is not clear how long it would take to detect all the features of an image. In both chips there would be an additional step of image acquisition and A/D conversion. A/D conversion in our chip clearly worsens the performance metrics, still, and taking into account that our data are simulated results, the data collected in Table II show that the expected performance is comparable to that of the state-of-the-art custom chips for SIFT.

TABLE II. COMPARISON WITH CUSTOM CHIPS RUNNING SIFT

Chip (technology) and functionality	Time per task (μs)	Power/(fps·px) (nJ/px)	(fps·px)/(Area·FullHD) $(\text{s}\cdot\text{mm}^2)^-1$
Ref. [7] (65nm CMOS) Feature detection on full-HD	110.9 (only 1 feat.)	0.8 (SIFT)	4.7 (SIFT)
Ref. [8] (0.13μm CMOS) Object Recognition on HD 720p	720 (feature detec.)	10.5 (SIFT)	0.416 (SIFT)
This work (0.18μm CMOS) Gaussian pyramid (176x120)	50 (pyram. gen.) (3 octaves 6 scales)	3.6 with ADC 0.05 without ADC	0.27 with ADC 20.3 without ADC

IV. Conclusion

This paper has addressed a visual chip on 0.18 μm CMOS technology with an array of 176 × 120 pixels for running Gaussian pyramid on an SC network with an assignment of 4 photodiodes per 4 SC nodes, one CDS and one ADC. The chip is the mapping of a CMOS-3D architecture for feature detection with reduced functionality, and area overhead due to the lack of TSVs. The data collected from this implementation, however, would make it easier a future design on CMOS-3D technology. The performance metrics from the extracted layouts are comparable to those of state-of-the-art chips that include SIFT for object detection and recognition. If available, experimental results will be shown at the paper presentation.

Acknowledgment

This work has been funded by Xunta de Galicia (Spain) through project 10PXIB206037PR, MICINN (Spain) through TEC2009-12686, MINECO (Spain) through projects IPT-2011-1625-430000, TEC2012-38921-C02 (co-funded by the European Region Development Fund, ERDF/FEDER), AE CITIUS (EDRF/FEDER, under the project CN2012/151 of the Galician Ministry of Education, Spain), and by ONR through Project N000141110312.

References

[1] D. Lowe (2004)., "Distinctive Image Features from Scale-Invariant Keypoints". International Journal of Computer Vision 60 (2): 91.

[2] J. Fernández-Berni et al., "FLIP-Q: A QCIF Resolution Focal-Plane Array for Low-Power Image Processing". *IEEE J. of Solid-State Circuits*, vol. 46, No. 3, pp. 669-680, March 2011.

[3] M. Suárez et al., "CMOS-3D Smart Imager Architectures for Feature Detection", IEEE Journal on Emerging and Selected Topics in Circuits and Systems, vol.2, no.4, pp.723-736, Dec. 2012.

[4] J. Fernández-Berni et al., "On the Implementation of Linear Diffusion in Transconductance-based Cellular Nonlinear Networks". Int. J. of Circuit Theory and Applications, 37(4),pp. 543-567, May 2009.

[5] R.S. Patti, "Three-Dimensional Integrated Circuits and the Future of System-on-Chip Design". *Proceedings of IEEE*, vol. 94, no. 6, pp. 1214-1224, June 2006.

[6] M. Suárez et al., "Switched-capacitor networks for scale-space generation", 2011 20th European Conference on Circuit Theory and Design (ECCTD 2011), pp.190-193, 29-31 Aug. 2011.

[7] Y.C. Su et al., "A 52 mW Full HD 160-Degree Object Viewpoint Recognition SoC With Visual Vocabulary Processor for Wearable Vision Applications". *IEEE J. of Solid-State Circuits*, vol. 47, no. 4, pp. 797-809, April 2012.

[8] Jinwook Oh et al. "A 320 mW 342 GOPS Real-Time Dynamic Object Recognition Processor for HD 720p Video Streams", IEEE Journal of Solid-State Circuits, vol.48, no.1, pp.33-45, Jan. 2013.