A 26.5 nJ/px 2.64 Mpx/s CMOS Vision Sensor for Gaussian Pyramid Extraction

M. Suárez*, V.M. Brea*, J. Fernández-Berni†‡, R. Carmona-Galán†, D. Cabello* and A. Rodríguez-Vázquez†‡
*Centro de Investigación en Tecnoloxías da Información (CITIUS)
University of Santiago de Compostela, Santiago de Compostela, Spain
Email: victor.brea@usc.es
†University of Seville, Instituto de Microelectrónica de Sevilla (IMSE-CNM), Seville, Spain
‡CSIC, Instituto de Microelectrónica de Sevilla (IMSE-CNM), Seville, Spain

***Abstract*—This paper introduces a CMOS vision sensor to extract the Gaussian pyramid with an energy cost of 26.5 nJ/px at 2.64 Mpx/s, thus outperforming conventional solutions employing an imager and a separate digital processor. The chip, manufactured in a 0.18 μm CMOS technology, consists of an arrangement of 88 $\times$ 60 processing elements (PEs) which captures images of 176 $\times$ 120 resolution and performs concurrent parallel processing right at pixel level. The Gaussian pyramid is generated by using a switched-capacitor network. Every PE includes four photodiodes, four MiM capacitors, one 8-bit single-slope ADC and one CDS circuit, occupying 44 x 44 μm^2. Suitability of the chip is assessed by using metrics pertaining to visual tracking.**

I. Introduction

CMOS vision sensors differ from imagers in that they are not intended for high quality image reproduction but for fast and power-efficient image analysis [1]. Such challenge calls for embedding parallel processing circuitry close to the sensors and can be addressed by using either per-column or per-pixel processors. This latter case achieves the largest degree of parallelism and hence the maximum speed and power efficiency at the cost of larger pixel pitch and smaller fill factor, unless 3D vertically-integrated technologies are employed. With planar technologies the larger pitch obviously penalizes the quality of imaging. However, this is not a major obstacle for many vision tasks. For instance patients with retinitis pigmentosa are able to see with only a very small fraction of their retina cells still alive. Also, faces can be detected using images captured by QVGA (alternatively, VGA) cameras located at 2.5 m (alternatively, 4 m) from the object. Such low resolution images hence qualify for human machine interface applications and other kind of indoor scenarios where lighting can be controlled and the reduced fill factor is not an insurmountable obstacle. Besides this, analysis show that parallel-processing vision architectures are largely tolerant to individual processor errors, e.g. deviations close to 10% are tolerated in many cases [2].

A major problem faced for vision sensor architects is that there are not standard functional specifications to pursue. This motivates the quest for functions which are suited to a large variety of problems besides being widely employed by computer vision system engineers. This is the case of feature detectors and, particularly, of those based on the Scale Invariant Feature Transform (SIFT). SIFT is used for image retrieval, 3D reconstruction, visual tracking, etc. [3]. The reason why a parallel-processing vision sensor may be advisable for SIFT is that this algorithm requires the calculation of the so-called Gaussian pyramid. It is constructed by applying a Gaussian filter with increasing widths (σ) resulting in different images, also called scales (S), making up an octave (O). This process is repeated O times with S scales each. The origin of a new octave is one half-sized reduction of the former one. The Gaussian pyramid makes SIFT algorithm robust against scale changes, but its calculation takes more than 90% of total time algorithm computation time [4]. The reason for that is the necessity to compute many Gaussian filters - a task which is not the best suited for a serial architecture. It motivates the proposed chip, which employs in-pixel parallel switched-capacitor circuits to calculate the Gaussian pyramid. Owing to the parallel architecture, the chip can be scaled to larger pixel counts and resolution without degrading the computation time and keeping the energy efficiency. Comparison to alternative solutions using conventional architectures shows combined speed-power advantages in the range of two to five orders of magnitude.

II. Chip Design

Fig. 1 shows the chip micrograph with a close-up of the Processing Elements (PEs) with MiM capacitors and photodiodes visible (four of each per PE). The chip occupies 5 $\times$ 5 mm^2, comprising 176 $\times$ 120 photodiodes arranged in 88 $\times$ 60 PEs. In order to shorten routing length and speed up I/O operations the image is read out through two frame buffers outside the PE array. Each PE is connected to two 8-bit registers in the corresponding frame buffer, allowing for reading out pixels outside the chip as they are being A/D converted.

The PE is shown in Fig. 2. Table I outlines its different transistor sizes and the photodiode dimensions. It occupies 44 $\times$ 44 μm^2. The scene is acquired with 4 3T-APS per PE with n-well/p-subs. photodiodes. Every PE contains the local circuitry of an 8-bit single-slope ADC and one CDS circuit. Also, the PE comprises 4 state capacitors C_{pij} with their corresponding switches along the four cardinal directions to configure a double-Euler switched-capacitor network that yields the Gaussian pyramid [5].

The 4 3T-APS structures are biased with only one current source drawing 1 μA. The design of the source follower aims at the largest possible operating range, which is met with low threshold voltage transistors, reaching 1 V of operating range with a gain error spread inferior to 0.4%.

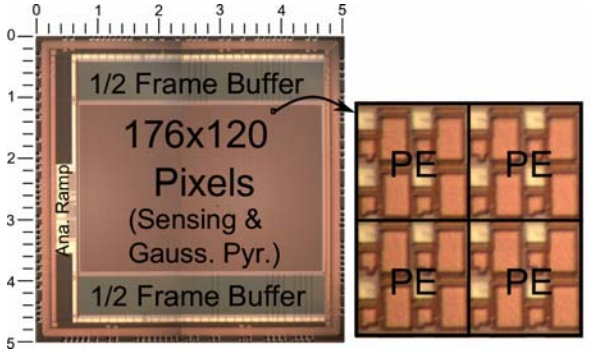

Fig. 1. Chip micrograph with dimensions (in mm) and a close-up of the PEs.

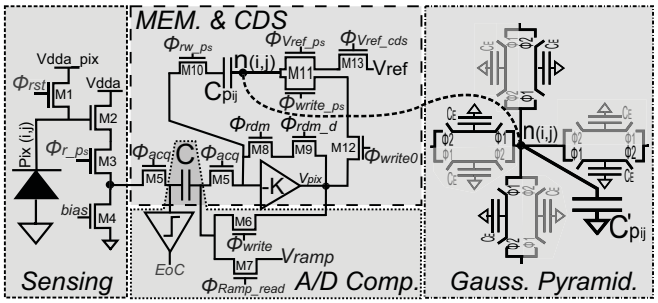

Fig. 2. Processing Element (PE) of the chip.

A time-multiplexed CDS operation reads the image from the four photosites per PE. Fig. 3 helps to understand this operation. Eq. (1) is the output of the CDS stage at every capacitor C_{pij}. Both the state C_{pij} and the capacitor C are sized to the same value; 200 fF, and implemented as MiM structures on Metal5 and Metal6 layers, which in turn, allows for circuitry underneath, saving area. V_{ref} is fixed to 400 mV for the inverter stage $-K$ to fall in the saturation region.

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

The inverter stage $-K$ is a double cascode topology with a high nominal gain of 65 dB to drop linearity errors. Figure 3(b) outlines the transistor sizes of the inverter. Its bias voltages are vbp = 1.2 V, vcp = 0.95 V, and vcn = 0.65 V, giving a bias current I = 1 μA. Additional transistors $enable$ and $enable\_n$ allow for zero static power consumption during standby periods (leakage currents neglected).

As mentioned before, the Gaussian pyramid is provided by a double-Euler switched-capacitor network. A switched-capacitor network minimizes the non-linearity of a conventional RC network, and permits a more accurate control of the σ levels by means of the number of clock cycles (n) of two non-overlapped clock signals (ϕ_1 and ϕ_2 in Fig. 2). The voltage V_{ij} at every state capacitor C_{pij} of the network for a given cycle n is given by Eq. (2), with C_E being the exchange capacitor present in the double-Euler topology (see Fig. 2). In our implementation, we have set C_E = 38.5 fF, while C_{pij} = 330 fF. C_E is implemented with an MOS transistor, while C_{pij} is the combination of an MiM structure with an MOS transistor in parallel.

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

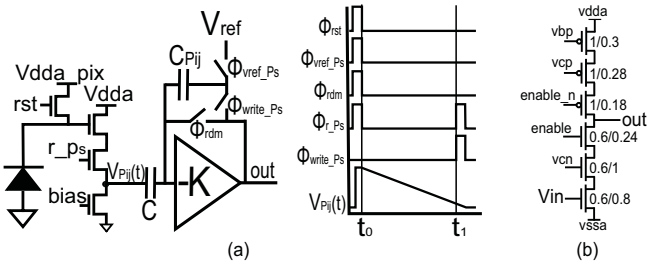

Fig. 3. (a) CDS stage with its time diagram. (b) Amplifier ($-K$) with the transistor dimensions in microns.

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

The starting sigma value is found by comparing Eq. (2) to the equation of the convolution of an image with a Gaussian kernel with non-zero elements along the four cardinal directions. Its value, which in our case is σ_0 = 0.48, along with that of σ as function of n are given by Eq. (3).

$$\sigma_0 = \left(2 \times ln\frac{C_{pij}}{C_E}\right)^{-1/2}, \quad \sigma\,(n) = \sqrt{\frac{2 \cdot n \cdot C_E}{C_{pij}}} \qquad (3)$$

Fig. 4 shows the offset-compensated comparator present in every PE, which, along with global circuitry for biasing, a counter, and the frame buffers mentioned above, makes up an 8-bit single slope ADC. The comparator has two gain stages ($-K$) implemented with the same design as that of the CDS stage (see Fig. 3). Also, the capacitor C used for offset compensation is shared with CDS. Signals $comp\_rst$ and $comp\_rst\_d$ are needed to run bottom sampling, leading to the output of the first inverter given by Eq. (4), with V_Q being the quiescent point of the first inverter, V_{pix} either the signal acquired by the photodiode or a given scale S, and V_{ramp} the ramp of the 8-bit single-slope ADC.

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

The A/D conversion ends with signal EoC at $'0'$. This occurs with the rising edge of the ouput of the first inverter ($inv1$), driving $enable$ of the first inverter to $'0'$ through the feedback loop from the second inverter. The complementary $enable$ of the second inverter is also forced to $'1'$ 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

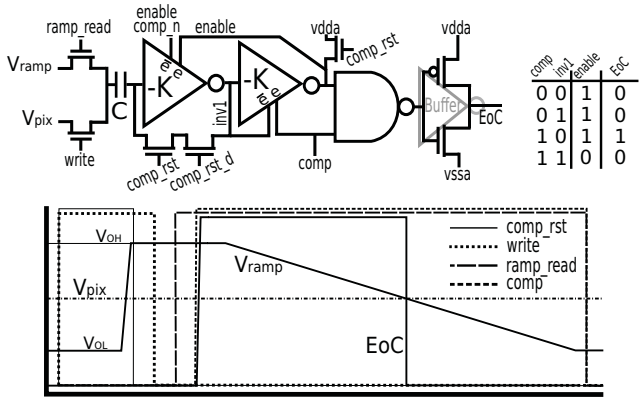

Fig. 4. Offset-compensated comparator for the ADC of the chip.

leads the static power consumption of the two inverters to zero. Finally, the NAND gate with *comp* at '0' drives EoC to '0' too, avoiding writing into the frame buffer before finishing the comparison. The analog ramp of the ADC is provided by an 8-bit current steering DAC located outside the PE array. The measured ADC DNL and INL are below 0.75 and 1.5 LSB, respectively.

Both the input scene and all the scales throughout the Gaussian pyramid are converted to digital. Usually 3 octaves with 6 scales each are used. One A/D conversion per pixel is possible from the second octave on. Nevertheless, for the input scene 4 conversions are required, while 24 conversions are needed for the 6 scales of the first octave. The reason is one ADC per PE (4 pixels), making 40 A/D conversions in total. This number does not change with the PE array size, and it will always be inferior to that of a more conventional ADC per column approach.

III. Experimental Results

Fig. 5 shows several snapshots of the Gaussian pyramid along with the image acquired by the chip. Fig. 6 plots both the expected and the actual on-chip σ as a function of the number of clock cycles n . The upper curve is the experimental σ. The lower curve is the theoretical σ (Eq. (3)). The on-chip σ levels are found by comparing the different on-chip Gaussian-filtered images, known as scales, with the acquired image filtered on a conventional computer within a given range of sigmas $[\sigma_1, \sigma_2]$ around the expected σ value. The minimum RMSE sets the on-chip σ level. Fig. 6 also shows RMSE levels with 255 as full-scale value (FSV). The RMSE slightly changes across octaves, being inferior to 1.2% of FSV. This method accounts for the errors of the on-chip Gaussian pyramid generation and the A/D conversion. The effect of such error levels in terms of an application is addressed in the next section.

The chip consumes 70 mW with scene acquisition and the Gaussian pyramid of 3 octaves with 6 scales each. The Gaussian pyramid is executed in 8 ms (A/D conversions included), with 200 μs per A/D conversion, and 150 ns as the clock cycle for the switched-capacitor network. This leads to 26.5 nJ/px at 2.64 Mpx/s.

Table II puts these numbers in perspective by comparison with more conventional solutions. We have included the power

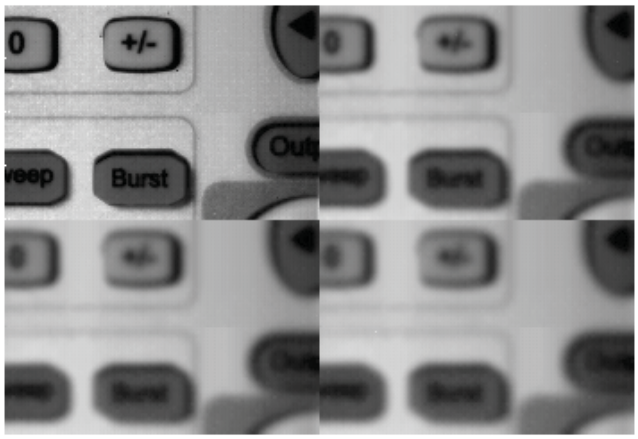

Fig. 5. Image acquisition and different snapshots of the on-chip Gaussian pyramid. The upper left image is the input scene, the rest of the images from left to right and top to down correspond to σ=1,77 (clock cyles n=19), σ=2,17 (n=29), and σ=2,51 (n=39).

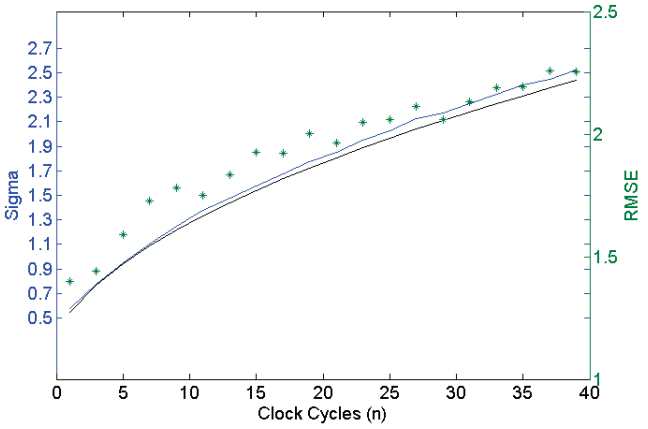

Fig. 6. Expected and actual σ vs. clock cycles (n) plots along with the RMSE values when comparing actual and ideal Gaussian-filtered images.

consumption of conventional CMOS imagers of Omnivision [7] with the image resolution to be tackled by the processor under study. We have not accounted for accesses to external memories, because such costs would also be present if our chip made part of a complete hardware platform for a given application, and because they are very hard to predict even with model memories. The energy cost of our chip outperforms that of an imager and a conventional MPU (even a low-power MPU) in three orders of magnitude with similar or faster processing speed, which leads to a combined speed-power figure of merit from two to five orders of magnitude superior to that of traditional solutions.

IV. Application Assessment

The assessment of the accuracy of the on-chip Gaussian pyramid has been made by incorporating the hardware errors in the interactive tool reported in [3]. This performs visual tracking of six 2D textures on videos with VGA resolution with the SIFT feature detector. The visual tracking metrics always use the so-called homography, defined as the matrix that captures the transformation of the 2D textures from one frame to the next one; e.g. rotation.

TABLE II. COMPARISON OF OUR CHIP WITH CONVENTIONAL SOLUTIONS

HW Solution	Func.	Energy/frame	En./px	Mpx/s
This work 180 nm CMOS	Gauss. Pyr.	176 × 120 resol. 70 mW @ 8 ms 0.56 mJ/frame	26.5 nJ/px	2.64
Ref. [8] OV9655 + Core-i7	Gauss. Pyr.	VGA resol. 90 mW @ 30 fps + 35 W @ 136 ms 4.8 J/frame	15.5 μJ/px	2.26
Ref. [9] OV9655 + Core-2-Duo	Gauss. Pyr.	VGA resolution 90 mW + 35 W @2.1 s 73.7 J/frame	240 μJ/px	0.15
Ref. [10] OV6922 + Qualcomm Snapdragon S4	Gauss. Pyr.	350 × 256 resol. 30 mW + 4 W @ 98.5 ms 0.4 J/frame	4.4 μJ/px	0.91

Repeatability (RP) is the metric that we have calculated to assess the quality of visual tracking with the on-chip Gaussian pyramid. As defined in [3], and formulated in Eq. (5), RP is the set of interest points S_{i-1} and S_{i-2} at frames $i-1$ and $i-2$ such that the geometrical distance between them after applying the corresponding homographies (H_{i-1} and H_{i-2}) from frames $i-1$ and $i-2$ to frame i are below a certain threshold normalized to the total number of interest points S_{i-1} or S_{i-2}. RP gives an estimate of the percentage of interest points whose allocation in successive frames is successfully forecast with the extracted homography.

$$RP = \frac{|(x_a \in S_{i-2}, x_b \in S_{i-1})| \, ||H_{i-2} \cdot x_a - H_{i-1} \cdot x_b|| < \epsilon}{|S_{i-1}|} \tag{5}$$

The RMSE values from the chip calibration have been expressed as local errors at pixel level by finding the standard deviation of the normal distribution which corresponds to the given RMSE level. The normal distribution conveys the variability from chip manufacturing. These errors have been added to every scale of the Gaussian pyramid. Fig. 7 displays RP vs. RMSE for RMSE of 0%, 1%, 2.5% and 5%. Our on-chip RMSE levels are below 1.2% of FSV. RP is the average of the aforementioned six 2D textures throughout all the frames of the corresponding videos with three different image transformations, namely, rotation, zoom and perspective distortion, aiming at the most general scene and motion pattern. The error bars, calculated as the standard deviation across the averaged data, state that the chip error levels do not cause a fatal degradation of RP. In fact, as reported in [3], the temporal distance between consecutive frames has a much bigger impact on RP, so a low computation time is a must, something feasible for the Gaussian pyramid with our proposal. Last, but not least, RP is a percentage, in our case as seen in Fig. 7 always above 0.4, a minimum number of 4 interest points is required to extract a homography H, nevertheless more interest points provide a homography with a higher confidence level. Local noise always increases the number of interest points, so that this is not a concern with analog computation for the Gaussian pyramid.

V. CONCLUSION

This paper presents a proof-of-concept chip for the parallel computation of the Gaussian pyramid. Verifications using visual tracking metrics show that the limited accuracy inherent to this sort of architecture is tolerated due to emergent robustness rendered by the parallelism. Good speed and power efficiency figures are obtained which, owing to the parallelism, are not degraded for larger pixel count. Although the chip employs planar technologies, the principles underlying the proposed architecture remain valid for vertically integrated technologies [6], thus paving the way to the future implementation of large resolution, vertically integrated feature detection vision sensors.

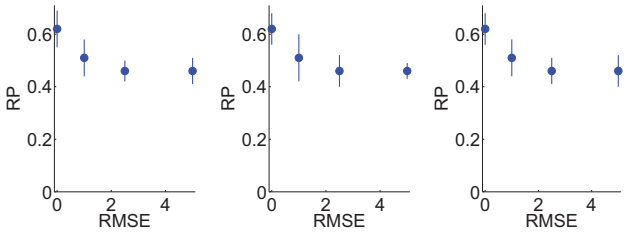

Fig. 7. Repeatability as a function of RMSE for three image transformations, namely, from left to right, rotation, zoom and perspective distortion.

ACKNOWLEDGMENT

This work has been funded by ONR N000141410355, Spanish government projects TEC2009-12686 MICINN, TEC2012-38921-C02 MINECO (European Region Development Fund, ERDF/FEDER), IPT-2011-1625-430000 MINECO, IPC-20111009 CDTI ((ERDF(FEDER)), Junta de Andalucía TIC 2338-2013, Xunta de Galicia EM2013/038, AE CITIUS (CN2012/151, ERDF(FEDER)), and GPC2013/040 ERDF(FEDER). The authors kindly acknowledge the University of California for the software transfer agreement UC Case No. 2014-453 and their PhD student Steffen Gauglitz for his assistance with the software framework to extract metrics for visual tracking.

REFERENCES

[1] A. Rodríguez-Vázquez et al., "A CMOS Vision System On-Chip with Multi-Core, Cellular Sensory-Processing Front-End". Chapter 6 in Cellular Nanoscale Sensory Wave Computers (edited by C. Baatar, W. Porod and T. Roska), Springer 2010.

[2] J. Fernández-Berni et al., "Smart Imaging for Power-Efficient Extraction of Viola-Jones Local Descriptors". Proceedings of SPIE, Vol. 9022, pp. 9022-09, 2014.

[3] S. Gauglitz et al., "Evaluation of Interest Point Detectors and Feature Descriptors for Visual Tracking". Int. J. of Computer Vision, vol. 94, pp. 335-360, 2011.

[4] K. Mizuno et al., "Fast and Low-Memory-Bandwidth Architecture of SIFT Descriptor Generation with Scalability on Speed and Accuracy for VGA Video". FPL 2010, pp. 608-611, 2010.

[5] M. Suárez et al., "Switched-capacitor networks for scale-space generation", ECCTD 2011, pp.190-193, 29-31 Aug. 2011.

[6] 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.

[7] Omnivision. http://www.ovt.com/.

[8] M. Murphy et al., "Image Feature Extraction for Mobile Processors". IEEE IISWC 2009, pp. 138-147, 2009.

[9] Feng-Cheng Huang et al., "High-Performance SIFT Hardware Accelerator for Real-Time Image Feature Extraction". IEEE TCAS-VT, vol. 22, no. 2, pp. 340-351, March 2012.

[10] G. Wang et al., "Workload Analysis and Efficient OpenCL-based Implementation of SIFT Algorithm on a Smartphone". IEEE GlobalSIP 2013, pp. 759-762, 2013.