In-Pixel Generation of Gaussian Pyramid Images by Block Reusing in 3D-CMOS

M. Suárez, V.M. Brea, D. Cabello
Centro de Investigación en Tecnologías de la Información
(CITIUS)
University of Santiago de Compostela
Santiago de Compostela, Spain
Email:manuel.suarez.cambre@usc.es

R. Carmona-Galán
A. Rodríguez-Vázquez
Instituto de Microelectrónica de Sevilla (IMSE-CNM)
CSIC-Universidad de Sevilla, Spain

Abstract— **This paper introduces an architecture of a switched-capacitor network for Gaussian pyramid generation. Gaussian pyramids are used in modern scale- and rotation-invariant feature detectors or in visual attention. Our switched-capacitor architecture is conceived within the framework of a CMOS-3D-based vision system. As such, it is also used during the acquisition phase to perform analog storage and Correlated Double Sampling (CDS). The paper addresses mismatch, and switching errors like feedthrough and charge injection. The paper also gives an estimate of the area occupied by each pixel on the 130nm CMOS-3D technology by Tezzaron. The validity of our proposal is assessed through object detection in a scale- and rotation-invariant feature detector.**

I. Introduction

Conventional image processing architectures operate frame-by-frame: frames are first captured, then codified in digital domain and finally processed. This approach benefits from the enormous computational power of digital processors in scaled-down technologies, but it is neither the most efficient one in terms of *processing speed* (time lag from inputs to actions) nor in terms of *energy consumption* [1]. The rationale for such sub-optimum efficiency is that frames do involve huge amount of data (number of pixels and number of bits employed per pixel) many of which do not carry useful information but all of which must be processed [2].

Among the different ways to increase architectural efficiency, this paper addresses the conception of new sensors capable to extract and deliver image *features* in addition to capturing and delivering image frames. This approach has a twofold rationale. On the one hand, it is fully compatible with the methodology followed by vision system architects, who are accustomed to using features when realizing scene interpretation and attentional [3]. On the other hand, it employs close-to-the-sensors concurrent processing to extract features, thereby yielding very large speed and energy efficiency [4].

The sensor architecture in this paper is specifically conceived for the SIFT (Scale Invariant Feature Transform) algorithm; a scale- and rotation-invariant feature detector algorithm customarily employed for object detection and classification, image retrieval, image registration and tracking [5]. A key ingredient of this algorithm is the extraction of *Gaussian pyramids*, which comprise a set of images of different resolutions called *octaves*. Every octave is the result of a $1/4$ downscaling of the previous octave. In turn, every octave is made up of a series of images called *scales*. Every scale is the result of performing a Gaussian filtering with given width (σ-level) on a previous scale. The main challenge for the extraction of Gaussian pyramids is to implement programmable Gaussian filters in accurate and controllable manner.

This paper addresses the challenge of implementing Gaussian filters and extracting Gaussian pyramids by using a 3D integration technology. We specifically employ the 130nm CMOS-3D technology from Tezzaron [6] which in its current version consists of two vertically-interconnected tiers tied to a 1Gb DRAM standard macro. Although Gaussian pyramids can be extracted by using programmable vision chips realized in conventional planar technologies [7][8], using these planar technologies largely penalizes the pixel pitch and hence the image quality. This drawback is overcome with 3D technologies owing to the vertical distribution of sensing and processing resources across the vertical layers.

Although the paper provides a global view of the architecture, emphasis is placed on the description of the first tier and more specifically on the implementation of the diffusion grid used for Gaussian filtering. We propose an implementation based on switched-capacitor networks [9] as this method provides the emulation of an inherently linear diffusion network, as compared to networks of nonlinear resistors [8]. Also, this approach enables in-pixel processing elements to be multiplexed in time to operate into different data and reused for concurrent implementation of CDS (Correlated Double Sampling) and ADC (Analog-to-Digital Conversion).

This paper starts with the description of the context for an elementary focal-plane processing cell, which is later described. The different functionalities exhibited by time-multiplexing and re-use of the signal processing blocks are then explained. Finally, with the help of a behavioral model, the influence of mismatch and switching errors in the final system performance is assessed.

II. 3D System Architecture

SIFT can be split in two phases: 1) where the so-called keypoints are extracted, and 2) where every keypoint is repre-

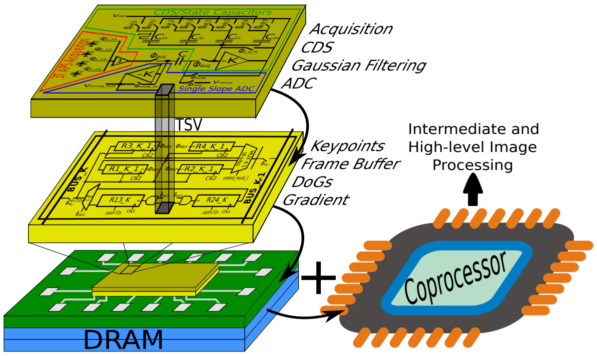

Fig. 1. CMOS-3D-based vision system.

sented with a descriptor vector. In the first phase, all the pixels of the image are involved, rendering a high volume of data. In the second phase, only the keypoints are involved (estimated as 1% of of the pixels of the image [5]).

The application and the resolution of the image set the amount of scales and octaves. Usually three to four octaves with six scales each are needed. The keypoints are located across the Gaussian pyramid on the Difference-of-Gaussians (DoG) in a given octave. The DoGs are calculated by subtracting two successive σ levels. The keypoints are the extrema among three successive DoG images calculated within a 3×3 neighborhood (27 neighbors).

Fig. 1 displays the complete system architecture where the hardware for Gaussian pyramid is embedded. It comprises a CMOS-3D stack with two tiers on top of a DRAM, and a coprocessor. The top tier in the stack includes signal acquisition, Gaussian pyramid generation and its digitization. The second tier contains a digital buffer, the DRAM memory controller and image subtraction used for gradient calculation, difference of Gaussians and keypoint location. The off-chip coprocessor is meant to perform intermediate- and high-level image processing as well as all the communication protocols. Since future updates of the CMOS-3D Tezzaron technology contemplate additional tiers, the architecture is conceived for modularity and scalability.

III. Pixel Architecture

Fig. 2 depicts the architecture of the elementary processor in the top tier of the CMOS-3D stack. This cell is responsible for offset-corrected image capture, pixel binning and diffusion in cooperation with the neighboring cells, and contributes to fully-parallel single-ramp A/D conversion. The sensors will be implemented as conventional 3T-Active Pixel Sensors (APS). The area constraints force us to assign 4 photodiodes per elementary processor. Correlated Double Sampling (CDS) and in-pixel ADC cannot be run in parallel for the four pixels due to hardware-sharing within every cell. The CDS is carried out with the circuit enclosed in green in Fig. 2. We also outline the data path for one of the pixels with a green dashed line. ϕ_{r_si}, ϕ_{vref_si}, ϕ_{acq},$\phi_{r\_d\_m}$ and ϕ_{write_si} are the signals involved in CDS. This is a very well-known circuit [10]. At the end of this stage, the pixel value is stored in what we named as state capacitors, C_{si} (node n_i) in Fig. 2, which act as analog

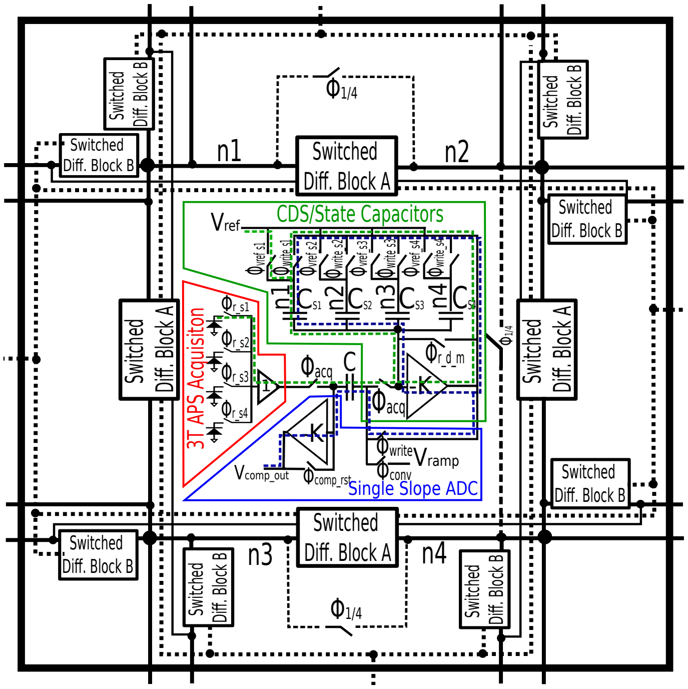

Fig. 2. Schematic of a cell in the top tier.

memories not only during the acquisition phase, but also, as we will see in Section IV, for the Gaussian pyramid generation.

The architecture performs single-slope ADC conversion. This circuitry is distributed in the two tiers. The comparator, encircled as ADC in Fig. 2, lies in the top tier. The counter and the registers are allocated in the bottom tier. This way, only one Through-Silicon-Via (TSV) is needed for communication between the top and the bottom tiers. The signals ϕ_{write}, ϕ_{conv} and $\phi_{r\_d\_m}$ control this operation. The capacitance C is used for both CDS and ADC.

Finally, the Gaussian pyramid generation is carried out by the state capacitors and the switches and exchange capacitors displayed on Fig. 2. Signals ϕ_{Diff} are used for this operation. This is addressed in Section IV.

The cell sketched in Fig. 2 contains 4 APS, 2 inverters of gain $-K$, 4 state capacitors, an additional capacitor for offset-cancellation in the comparator used for ADC, plus 16 exchange capacitors, and around 70 switches. We can give an area estimate per pixel if we account for: 1) an area of $5\mu m \times 5\mu m$ per photodiode, 2) state capacitors $C_s = 100fF$, which have a density of $1fF/\mu m^2$ for the $130nm$ CMOS-3D Tezzaron technology, 3) exchange capacitors $C_E = 10fF$, 4) double-cascode inverters to enhance a high enough gain ($K > 60dB$) to reduce errors in closed-loop configurations, amounting to $50\mu m^2$ if we take the implementation presented in [11] as a reference (FDSOI $150nm$), 5) $2\mu m^2$ per switch, and 6) a TSV of $5\mu m^2$. All in all yields around $250\mu m^2$ per pixel, which we overestimate up to $300\mu m^2$, accounting for the routing. These numbers would lead to $23mm^2$ for an image of QVGA resolution. The ratio $C_s/C_E = 100fF/10fF$, as will be seen in Section IV, will be employed for the graph of Fig. 5, which permits a σ from 0.45, enough for SIFT-based applications.

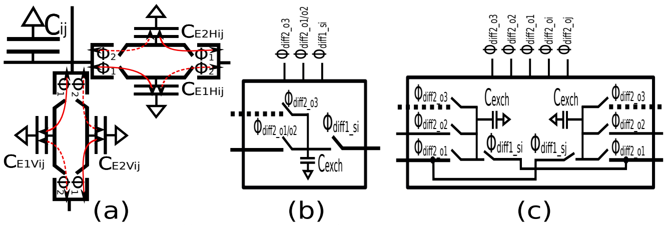

Fig. 3. Different schemes of switched-capacitor networks.

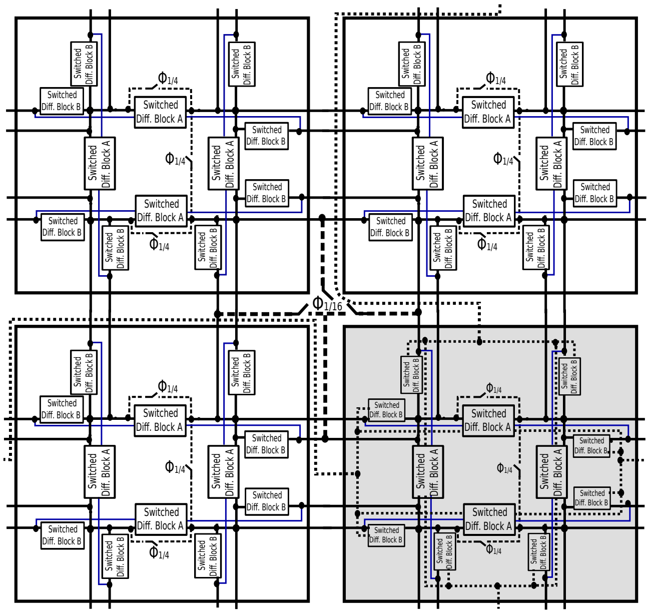

Fig. 4. Schematics of the diffusion network for a grid of 16 x 16 pixels of our system.

IV. Gaussian Pyramid Generation

The system shown in Fig. 1 provides three octaves with six scales each. These numbers suffice for SIFT [5].

A. Nominal Analysis

The Gaussian pyramid is generated with a parallel Forward-Euler switched-capacitor network. Fig. 3.a sketches the interaction of a given state capacitor C_{ij} with two neighbors. Signals ϕ_1 and ϕ_2 are non-overlapping clock signals. A 4-connected 2D network is completed with two more interactions along the rest of cardinal directions (not shown in Fig. 3.a). In such a network the relation between σ-levels and number of clock cycles n, $\sigma = \sigma(n)$ is given by Eq. (1) [9]. As seen, the C/C_E ratio determines the σ values.

$$\sigma(n) = \sqrt{\frac{2nC_E}{4C_E + C}} \qquad (1)$$

In SIFT the σ-levels cannot change from octave to octave. Nevertheless, building the Gaussian pyramid leads to the reconfiguration of the grid topology. This might change the σ-levels across octaves. Fig. 2, Fig. 3 and Fig. 4 illustrate how the grid topology of our network changes when moving across octaves. In Fig. 4, the circuits enclosed in a square are the cells of Fig. 2, made up of 4 pixels and their corresponding diffusion blocks to interact with their neighborhood. In a 4-connected network every pixel is connected to the nearest four neighbors along the four cardinal directions. That's why there are four diffusion blocks connected to the state capacitors C_{si}, labeled as nodes n_i in Fig. 2. Going from the first to the second octave means to downscale the image from M x N to $1/2M$ x $1/2N$ resolution. This is performed by joining the four pixels (state capacitors) with signal $\phi_{1/4}$ on (Fig. 2 and Fig. 4). Now, every cell is only one pixel resulting from an average of the 4 pixels within a cell. Nevertheless, when doing this, the state capacitor of a cell becomes $4C_s$, while we only have two exchanging or switched-diffusion blocks along every cardinal direction, i.e. $2C_E$ (those marked as Switched Diff. Block B in Fig. 4), which are shown at schematic level in Fig. 3.b. Keeping the C/C_E ratio, and thus the σ-levels, forces to $4C_E$ instead of $2C_E$. This is achieved by shorting the exchange capacitors of the exchanging blocks of the four pixels within a cell (blue lines in Fig. 4), those labeled Switched Diff. Blocks A in Fig. 4, to the surrounding cells. This is performed by signals ϕ_{diff2_o2} in Fig. 3.c. Thick solid lines show the connections among cells in Fig. 4.

A similar process occurs with the leap from the second to the third octave. This involves one more set of switches, those controlled by signal $\phi_{1/16}$ (shown in Fig. 4). When $\phi_{1/16}$ is on, the values stored in four cells (16 pixels) are averaged in only one pixel, performing the downscaling of the original image from M x N to $1/4M$ x $1/4N$ resolution. Now we have a macrocell of 16 x 16 pixels. Fig. 4 displays the cell used for the scale-generation within the third octave. The connectivity among pixels is also shown in Fig. 4 as a dotted-line.

B. Error Analysis

1) Mismatch: The main source of mismatch is the spread in the capacitance values (C and C_E). Such a spread makes that every pixel have a different Gaussian kernel. This causes two effects. On the one hand, the relation $\sigma = \sigma(n)$ might change. On the other hand, some pixels will have a greater smooth effect than others. This will give different figures of merit as $recall$ and $precision$ in object detection (Section IV.D).

Fig. 5 shows the effect of mismatch on the relation σ-levels-number of clock cycles (n). As in [9] the σ-levels are found by comparing the images from a convolution of Gaussian kernels with well-defined σ-levels to images produced by our switched-capacitor network with mismatch errors through a vehavioural model in MATLAB. The algorithm searches for the σ that provides the least RMSE. Simulations of 50 random normal distributions with a standard deviation of $6\sigma = \sqrt{C}$ were run. As seen, the effect of mismatch on the relationship σ with cycles amount $(\sigma(n))$ is barely noticeable. This is because the σ-levels are extracted from the whole image, accounting for a global effect. On average the local variations of C/C_E are almost cancelled out over the whole image.

2) Switching Errors: Feedthrough and charge injection are the main errors coming from the switches.

In our system (depicted in Fig. 3.a) the feedthrough is caused by the coupling of the clock signal (either ϕ_1, or ϕ_2) through the overlapping capacitances between the switches and the state capacitor (C_{ij}). In this case, a falling edge of ϕ_1 is followed by a rising edge of ϕ_2. As a first order, we

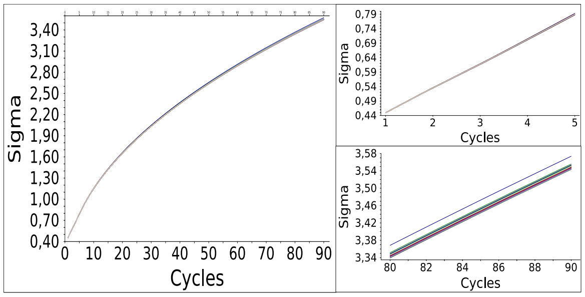

Fig. 5. Mismatch effect on σ with filtering cycles amount n.

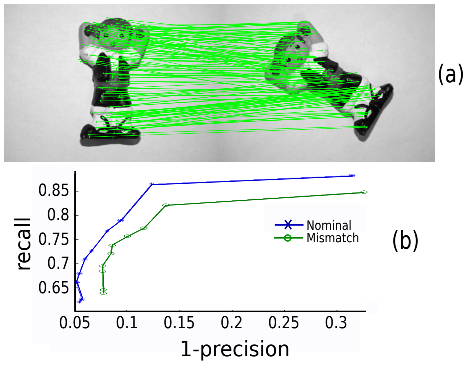

Fig. 6. (a) SIFT for object detection. (b) recall vs. 1-precision plot for object detection on the images (a).

consider that the succession of both edges leads to a zero error. Nevertheless, in practice there will be a residual error caused by the mismatch between the overlapping capacitances of switches ϕ_1 and ϕ_2.

A similar process takes place for charge injection. Turning off ϕ_1 is immediately followed by turning on ϕ_2. The charge injected by the first switch into the state capacitor C_{ij} will be collected by the second one under the assumption of equal geometries in both switches. This would lead again to a zero error. In practice, mismatch will yield a non-zero error.

C. SIFT-based Assessment: Object Detection

This section presents an example of object detection performed by our switched-capacitor network subject to mismatch errors. In a first order we assume that both charge injection and feedthrough errors are canceled due to the switching mechanism of the network as it was discussed in Section IV.B. Fig. 6 (a) displays an example of an object with a rotated version of 45^o. Fig. 6 (b) plots the $recall$ vs. $1 - precision$ graph. The precision is defined as $p = tp/(tp + fp)$, and the recall as $r = tp/(tp + fn)$, with tp being the number of true positives, fp the number of false positives and fn the number of false negatives. $tp + fp$ is the number of matches, shown as overlapping points in Fig. 6 (a). The number of matches is calculated by comparing the descriptor vectors of two keypoints. If the difference of modules of such vectors is below (above) a certain threshold (th), the corresponding pair of keypoints is a match. A match becomes tp when it also complies with the location condition; otherwise it is a false positive. The location condition can be checked easily in a known transformation as that of Fig. 6. It is also possible to calculate fn. Fig. 6 (b) was obtained from changing th for the original image rotated 45^o. As it can be seen, the shape of the curve with the switched-capacitor networks subject to mismatch resembles that of the nominal one, but with worse performance. This is due to the local effect of mismatch, which causes every pixel to have a different kernel from the nominal Gaussian kernel. In this case a random normal distribution with a standard deviation $6\sigma = \sqrt{C}$ were run. The application dictates whether or not the mismatch is detrimental.

V. Conclusions

This paper addresses the architecture of a switched-capacitor network for Gaussian pyramid generation. Gaussian pyramids are widely used in image processing tasks as the SIFT algorithm or in visual attention. The switched-capacitor network discussed in this work is embedded in the top tier of a CMOS-3D stack with two tiers. Hardware-sharing among different functions is implemented in order to reduce area occupancy per pixel. In our case, the state capacitors of the network are used for analog storage and CDS calculation. The paper has shown the suitability of our architecture to succeed against mismatch, and function errors like charge injection and feedthrough. We have estimated an area of $300\mu m^2$ per pixel ($1200\mu m^2$ by cell) on the $130nm$ CMOS-3D technology from Tezzaron. The next step will be the full-custom design of the network.

VI. Acknowledgment

This work has been funded by Xunta de Galicia (Spain) and MICINN (Spain) through projects 10PXIB206037PR and TEC2009-12686, and by ONR through Project N000141110312.

References

[1] W. Zhang Q. Fu, N.-J. Wu ,*"A Programmable Vision Chip Based on Multiple Levels of Parallel Processors,"* IEEE JSSC, vol.46, no.9, pp.2132-2147, Sept. 2011.

[2] N. Balakrishnan et al.*"A New Image Representation Algorithm Inspired by Image Submodality Models, Redundancy Reduction, and Learning in Biological Vision,"* IEEE Trans. PAMI, vol.27, no.9, pp.1367-1378, Sept. 2005.

[3] L. Itti et al.,*"A Model of Saliency-Based Visual Attention for Rapid Scene Analysis"* , IEEE Trans. PAMI, vol. 20, no. 11, pp. 1254-1259, Nov. 1998.

[4] J. E. Eklund, C. Svensson, A. Astrom,*"VLSI implementation of a focal plane image processor-a realization of the near-sensor image processing concept,"* IEEE Trans. VLSI, vol.4, no.3, pp.322-335, Sept. 1996.

[5] D.G. Lowe,*"Distinctive Image Features from Scale-Invariant Keypoints"* , Int. J. Comput. Vis., vol. 60, no. 2, pp. 91-110, 2004.

[6] http://www.tezzaron.com.

[7] A. Rodríguez-Vázquez et al.,*"The Eye-RIS CMOS Vision System"*, in H. Casier et al. (eds.), Analog Circuit Design. Springer, 2008.

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

[9] M. Suárez et al.,*"Switched-Capacitor Networks for Scale-Space Generation"* , 20th European Conference on Circuit Theory and Design, pp. 189-192, Linkping, Sweden, August 29-31, 2011.

[10] Yu M Chi et al.,*"CMOS Camera With In-Pixel Temporal Change Detection and ADC"* , IEEE Journal of Solid-State Circuits, vol. 42, NO. 10, pp. 2187-2196, October 2007.

[11] A. Rodríguez-Vázquez et al.,*"A 3D Chip Architecture for Optical Sensing and Concurrent Processing"* , in F. Berghmans, A. G. Mignani, C. A. van Hoof (Eds.): Optical Sensing and Detection, Proceedings of SPIE, vol. 7726, pp. 772613-1-772613-12, April 12-15, 2010.