CMOS-3D Smart Imager Architectures for Feature Detection

M. Suárez, V.M. Brea, J. Fernández-Berni, R. Carmona-Galán, G. Liñán, D. Cabello and A. Rodríguez-Vázquez *Fellow, IEEE,*

***Abstract*—This paper reports a multi-layered smart image sensor architecture for feature extraction based on detection of interest points. The architecture is conceived for 3D IC technologies consisting of two layers (tiers) plus memory. The top tier includes sensing and processing circuitry aimed to perform Gaussian filtering and generate Gaussian pyramids in fully concurrent way. The circuitry in this tier operates in mixed-signal domain. It embeds in-pixel Correlated Double Sampling (CDS), a switched-capacitor network for Gaussian pyramid generation, analog memories and a comparator for in-pixel ADC (Analog to Digital Conversion). This tier can be further split into two for improved resolution; one containing the sensors and another containing a capacitor per sensor plus the mixed-signal processing circuitry. Regarding the bottom tier, it embeds digital circuitry entitled for the calculation of Harris, Hessian and DoG detectors. The overall system can hence be configured by the user to detect interest points by using the algorithm out of these three better suited to practical applications. The paper describes the different kind of algorithms featured and the circuitry employed at top and bottom tiers. The Gaussian pyramid is implemented with a Switched-Capacitor (SC) network in less than 50 μs, outperforming more conventional solutions.**

***Index Terms*—Smart CMOS Imagers, Chip architectures, Feature image extraction, Gaussian pyramid.**

I. Introduction

CMOS technologies enable to embed readout, analog-to-digital conversion, control, processing and communication circuitry along with high-quality photo-sensors. These embedding capabilities are instrumental for the realization of high-performance smart imagers and compact vision systems [1]. Besides the incorporation of on-chip smartness, CMOS Image Sensors (CIS) art is evolving towards: enhanced photo-sensing quality [2], [3]; increased spatial resolution, frame rate and data rate [4]; ability to detect low-light and multi-spectral signals [5], [6]; and enlarged flexibility and configurability [7], among others. All these developments are opening new application domains which were not feasible for CCDs [8]. CISs are also replacing CCDs at applications which used to be CCD niches [9]. All in all, the outcome is that CISs currently dominate the market of area imagers, with more than 90% of the total share [10].

The most important asset of CISs is the incorporation of intelligence on-chip [1]. Different levels of intelligence can be contemplated. The lowest involves basically readout, control and error correction, and is the only one yet exploited by industry [11], [12]. Higher intelligence levels, as required to analyzing, extracting and interpreting the information contained into images have been explored for years at academia [13] - [25], but with scarce industrial impact [11]. From now on we will refer to CISs with high-level intelligence attributes as CVSs, where CVS stands for CMOS Vision Sensor.

A common strategy to design CVSs consists of using a sensing front-end composed of an array of smart pixels. As a difference to active pixels, which embed the minimum circuitry required for photo-sensing and pixel addressing, smart pixels include also memories and processors [14]. Smart pixels are usually arranged into Single Instruction Multiple Data (SIMD) architectures [27], which perform parallel processing by making sensory data to interact following uniform laws. SIMD sensory front-ends extract image features (such as salient points, edges, etc.) through the parallel implementation of early vision tasks [28] using either hardwired, software-controlled or mixed architectures. The extracted features can then be used, instead of full frames, as inputs for subsequent image analysis. This significantly reduces the computational load of any ulterior processing, what in turn increases the efficiency in the realization of vision tasks. Conventional solutions, on the contrary, simply consists of a sensory front-end that delivers full frames to a digital processor [29].

CVS-SIMD chips to realize early vision tasks at thousands fps (frames per second) rate and with very low power consumption have been reported elsewhere [16] - [18]. The industrial art includes also vision-systems that combine early processing (to extract features) and post processing into a single chip [11], [15]. However, a drawback of these sensors, and the ultimate reason why they are not yet widely employed, is the rather large pixel pitch and reduced fill factor of their smart pixels. These features make CVS-SIMDs to have limited sensitivity and modest spatial resolution, thereby constraining their usage to applications with limited field-of-view and active illumination. This paper describes a CVS-SIMD architecture which overcomes these drawbacks through the usage of 3-D integration technologies [30] and the subsequent improvement of the form factors and footprints of the functional structures embedded at the pixels and at the whole chip. These improvements are achieved owing to the vertical distribution of sensing

Manuscript received June 30, 2012. This work has been funded by Xunta de Galicia (SPAIN) through project 10PXIB206037PR, MICINN (SPAIN) through projects TEC2009-12686, IPT-2011-1625-430000, and by ONR through Project N000141110312.

M. Suárez, V.M. Brea and D. Cabello are with the Centro de Investigación en Tecnologías de la Información (CITIUS), University of Santiago de Compostela, SPAIN. Contact: manuel.suarez.cambre@usc.es, victor.brea@usc.es.

R. Carmona-Galán and G. Liñán are with Instituto de Microelectrónica de Sevilla (IMSE-CNM), CSIC, SPAIN. Contact: rcarmona@imse-cnm.csic.es.

J. Fernández-Berni and A. Rodríguez-Vázquez are with the University of Seville, SPAIN, and with Instituto de Microelectrónica de Sevilla (IMSE-CNM), CSIC, SPAIN. Contact: angel@imse-cnm.csic.es.

and processing resources across the vertical layers of a 3-D structure.

The architecture described in this paper tackles the detection of interest points for scale-invariant and rotation-invariant feature detectors. Interest points are customarily employed by vision system architects for object detection and classification, image retrieval, image registration and tracking [31]. Out from the different algorithms, our architecture can be configured by the user to implement three widely accepted ones, namely: i) the classical Harris-corner detector [32]; ii) the Scale Invariant Feature Detector (SIFT) [33]; and iii) the Hessian operator, which is the one used in the Speed-Up Robust Features (SURF) [34]. A basic rationale for configurability is the lack of uniqueness of detector algorithms, as [35] demonstrates for volume reconstruction applications. Our proposal faces this drawback by allowing users to select the specific detector which is better suited to each application. Configurability also enables the accuracy-speed trade-off to be tuned. To the best of our knowledge, the architecture reported in this paper is the first one implementing different interest point detectors into a single, dedicated system.

A key ingredient of the proposed architecture is the extraction of Gaussian pyramids, which comprise a set of images of different resolutions called octaves. Every octave is the result of down-sampling the previous octave by a factor of 2. 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. We particularly employ Switched-Capacitor (SC) networks, based on our previous proposal reported in [36], as this method enables to emulate inherently linear diffusion networks, which is advantageous versus using nonlinear resistors [16], [37]. Besides this, other relevant feature of the herein reported architecture arises from the possibility of in-pixel processing elements to be multiplexed in time to operate on different data and to be used for concurrent implementation of CDS (Correlated Double Sampling) and ADC (Analog-to-Digital Conversion).

This paper is not intended to reporting a specific silicon prototype, but a novel sensory-processing architecture whose physical implementation relies on CMOS standard structures; i.e., novelty relies on the way circuit structures are arranged into the architecture. Silicon implementation themselves are similar to other previously demonstrated through state-of-the-art silicon by the authors and other groups. We first briefly describe the functional target and overview previous hardware implementations of feature detectors, on the one hand, and of Gaussian filters, on the other. Then, the proposed architecture is addressed, including a comparison with state-of-the-art custom chips performing feature detection. Finally, the conclusions are outlined.

II. Feature Detectors: Algorithms, Hardware Implementations and General 3-D Architectures

Feature detectors are widely used for computer vision applications such as object detection and classification, image retrieval, 3D reconstruction or tracking, among others [38]. Ideal feature detectors should be invariant against image changes caused by translations, rotations, scale modifications, partial occlusions and affine transformations; and should not be impacted by neither added noise nor changes of the illumination conditions. Their outcome should be the same despite these changes, although repeatability is usually traded for computer efficiency in practical implementations.

Feature detector implementation involves seeking for the matching of a set of local features between two images and comprises three main stages, namely: 1) interest points (also called keypoints) detection; 2) descriptor vector generation; and 3) matching. There are different techniques for each of these steps; and many ways of combining these techniques towards feature detection. For instance, [39] overviews up to 20 different combinations of interest point detectors, descriptor vectors and matching techniques for visual tracking. The three algorithms chosen for implementation into our architecture, namely the classical Harris-corner detector, the Scale Invariant Feature Detector (SIFT) and the Speed-Up Robust Features (SURF), are widely used and cover an ample application range. Furthermore, the three detectors addressed in our architecture share the common feature of using convolution type operators with Gaussian kernels, thus being very well suited for re-configuration.

A. Detection of Interest Points at the Different Algorithms

The classical Harris corner detector finds keypoints through the so-called first moment or autocorrelation matrix of Eq.(1), where I_x and I_y denote the first derivatives along the x and y directions in the image plane, and w_k is the neighborhood around a point (x_k, y_k) in the image plane used for the calculation of the derivatives [32]. Usually, previous to obtaining the autocorrelation matrix, the image is filtered with a Gaussian kernel, which enhances robustness with noisy images. A pixel is an interest point if it is a maximum of the autocorrelation function. The algorithm is completed with a non-maximum suppression based on the trace and the determinant of the autocorrelation matrix, both formulated in Eq.(2) and Eq.(3). The variable R defined in Eq.(4) states whether or not a pixel is a corner. If R is positive, the pixel is regarded as a corner. If R is negative the pixel is sorted out as an edge, and if R is zero or small, the pixel is part of a flat or homogeneous region. k is a constant that can be modified in order to enhance the stability of the interest points. This tuning will depend on the application. The Harris corner detector is invariant to rotations, but it yields very limited invariance to scale changes. Its simplicity and relatively low computation time when compared to other point detectors are its main assets.

$$A(x,y) = \begin{pmatrix} B & C \\ C & D \end{pmatrix} = \begin{pmatrix} \sum\limits_{w_k} I_x^2 & \sum\limits_{w_k} I_{xy}^2 \\ \sum\limits_{w_k} I_{xy}^2 & \sum\limits_{w_k} I_y^2 \end{pmatrix} \quad (1)$$

$$Tr(A) = B + D \quad (2)$$

$$Det(A) = BD - C^2 \tag{3}$$

$$R = Det(A) - k \cdot Tr^2(A) \tag{4}$$

The Harris corner detector is provided with scale invariance by incorporating scale space to the image representation. This is known as Harris-Laplace [40]. The keypoints are found as the local extrema over the so-called Laplacian-of-Gaussian (LoG) defined in Eq.(5), where σ_n is the corresponding scale, and L_{xx} and L_{yy} denote the second derivatives along the x and y directions. As in the classical Harris algorithm, a non-maximum suppression can be applied.

$$|LoG(x_k, y_k, \sigma_n)| = {\sigma_n}^2 |L_{xx} + L_{yy}| \tag{5}$$

Let us now consider SURF; it employs the so-called Hessian operator [34]. The Hessian operator is a matrix whose entries are the second derivatives around a pixel. These second derivatives are smoothed by Gaussian filters with the aim of reducing noise levels. The interest points are calculated as the extrema of the Hessian matrix. The same non-maximum suppression method as in the classical Harris corner detector is used. Also, Hessian-Laplace is possible.

The Difference-of-Gaussians (DoG) is the method used by SIFT for interest point detection [33]. SIFT relies on the scale space. In this method the keypoints are found as the local extrema among three successive DoGs. A DoG is obtained as the difference between two consecutive scales. The scales are generated through the Gaussian pyramid. Usually, a pyramid consisting of 3 octaves with 5 scales each suffice for many applications. Although the interest point detection in SIFT is based on DoGs, local maxima directly among the scales is also suggested by D. Lowe [33].

B. Overview of Feature Detectors Hardware Art

Only dedicated hardware, using either FPGAs or custom chips, is addressed here. Also, since most interest point detectors rely on the generation of the Gaussian pyramid and the application of a differentiation operator to the generated scales, only SIFT realizations are covered; the conclusions drawn for SIFT apply to other algorithms. Out from the SIFT implementations addressed here, [43] - [45] correspond to FPGA implementations, while [46] and [47] correspond to custom chips. None of these implementations includes the sensing devices; in all cases images are provided by separate, external sensors.

In [43] the image is split into regions of interest and a pipeline flow is realized at each region, thus increasing parallelism. Frame rates of 56fps for VGA images are obtained by using a pyramid with 18 modules (3 octaves with 6 scales each). Based on the observation that the generation of the Gaussian pyramid consumes up to 92% of the total resources, separable Gaussian kernels are employed for increased computation efficiency.

The work in [44] modifies the original SIFT algorithm to reduce area and computation time, while keeping good feature matching. It detects features at 30fps on QVGA images. Two different strategies are adopted for efficient generation of the pyramid, namely: i) using one Gaussian filter module per scale and octave – 18 modules in total; ii) separating the Gaussian kernel into horizontal and vertical sub-kernels.

Reference [45] achieves video frame rate processing for VGA images by using a scheme similar to that in [11]; namely, by first splitting the image into what they call segments, and then applying pipelining within each segment.

In reference [46], parallelism is achieved by using several SIMD units for the Gaussian pyramid generation. Also, VGA images are split into 300 regions of 32×32 each, and the SIFT algorithm is completed only in those regions where features are detected. This approach yields 30fps for VGA images.

Reference [47] also employs 18 modules for Gaussian pyramid generation. A distinctive feature of the adopted approach is that feature matching is performed once every 30 frames. In the rest of frames, object level matching is realized. These strategies feature real time operation with full HD images.

The most important message conveyed by the description above is that the generation of Gaussian pyramids represents the Gordiano knot for interest point detector implementation. Different architectural strategies are employed to cope with the computational load of Gaussian pyramid generation, such as: reducing the number of octaves and scales, using several Gaussian kernels, separating these kernels into horizontal and vertical sub-kernels, etc. Ultimately, all these strategies are meant to increase parallelism by evolving from a single-core processing architecture to a multiple-core one. Under these premises the usage of CVS-SIMD chips with dedicated smart-pixels seems to be pertinent for two reasons: i) they yield the maximum possible degree of parallelism by using a processing core per pixel; ii) they combine the sensing and processing operations into a single system, thus precluding the usage of external sensors.

The potential advantages of CVS-SIMDs can be assessed by comparing performance levels of the CVS-SIMD chip in [16], to those of the digital chip in [46]. The former extracts the Gaussian pyramid by using a resistive grid with a pixel per resistive node. The chip permits a fine control of the σ-level of the Gaussian kernel by means of a sampling mechanism. The complete Gaussian pyramid is generated in less than 10 μs (without accounting for the downloading of Gaussian pyramid images) for images with QCIF resolution. Regarding the former chip, [46], it performs feature detection (Gaussian pyramid and extrema location) in 180 μs over a 32×32 image. Then, for a VGA image, by splitting it into 300 regions and using 4 units for Gaussian filtering and extrema location, 13.5 ms are required to complete the detection of interest points. These two chips do not perform the same function; however, since Gaussian pyramid generation takes around 90% of the resources for interest point detection, the three-orders-of-magnitude difference among them motivates further exploration of CVSSIMD solutions. Potential advantages of these architectural solutions with distributed pre-processing versus conventional digital architectures are further discussed in [16] and references thereof.

Our approach for Gaussian pyramid generation is conceptually equivalent to that in [16]; in both cases, resistive grids

are employed to obtain natural solutions of the heat-diffusion equations and hence implement Gaussian filters. Leaving aside the fact that [16] does not address 3-D integration, differences arise in the approach used to realize the resistive grid. The solution reported in [16] employs continuous-time circuits with the resistors implemented through MOS-transistors. This has two major drawbacks: on the one hand, it raises non-linearity problems; on the other hand, it requires high precision circuits to control the time constant and hence set the sigma-level for Gaussian filtering. On the contrary, we herein employ discrete-time resistive grids with resistors emulated by SC circuits – based on our previous proposal in [36]. Thus, sigma-levels are controlled in very simple manner, by setting the number of clock cycles, and the operation is inherently linear.

C. General CMOS-3D Architectures for Feature Detectors

Algorithms previously described for interest point detection involve different hierarchical steps each with different image representations, different amount of data and different abstraction levels. Resorting to 3D technologies enables emulating such a hierarchical, conceptual structure and, particularly, mapping every function of a feature detector onto corresponding physical tiers – see Fig. 1(a). This would lead to the highest possible performance, as the hardware of every tier would suit the function to be implemented. In this figure, all low-level image processing functions but extrema location are at pixel level and would be implemented in the analog domain. Extrema location is more conveniently implemented in the digital domain, especially if the non-maximum suppression shown in Eq.(2), Eq.(3) and Eq.(4) is executed. Still, a pre-selection of possible extrema location can be incorporated to the analog plane, and thus realized in parallel, for better tradeoff between processing speed and power consumption. As Fig. 1(a) shows, the circuits for A/D conversion would be distributed among the tier for extrema location and the tier for the composition of the Harris, Hessian and DoG images. The large area for the circuits for A/D conversion (ideally in-pixel A/D conversion to keep the highest possible parallelism) and for the digital circuits found in the tier for extrema location would decrease parallelism in both tiers. The rest of the tiers would work in the digital domain. At the intermediate-level of image processing, the feature description tier would provide certain degree of parallelism as the descriptor vector of every keypoint can be calculated independently of any other interest point. At this level, only a percentage of the $M \times N$ pixels of the image are processed (typically 1% of the $M \times N$ pixels [33]). At the highest level of abstraction, feature matching would imply more irregular and complex memory accesses. In this case, task parallelism would replace data parallelism.

Fig. 1(b) shows a more realistic architecture. This approach would be possible on CMOS-3D technologies like that of MIT Lincoln Laboratories, as that reported in reference [49], where the sensors were bump-bonded to an FDSOI CMOS stack of three tiers. It is important to emphasize that in Fig. 1(b) the sensing plane is separated from the rest of tiers not to degrade the fill-factor. The processing continues being next to the sensor thanks to the TSVs between tiers and a photosensor-per-processor assignment is still possible. In this architecture,

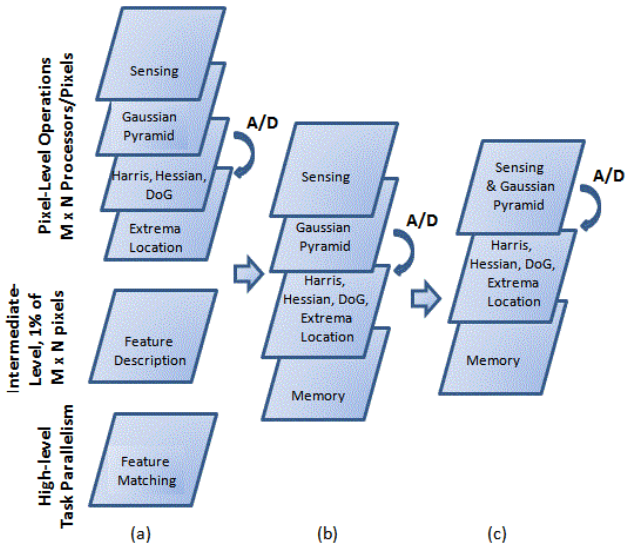

Fig. 1. Different partitions of the functionality of feature detectors across different tiers on CMOS-3D technology: (a) ideal situation; one tier for every function, (b) a more realistic situation, with the sensors in a dedicated tier not to degrade fill-factor, and (c) another example of CMOS-3D technology with the sensing and the Gaussian pyramid construction in the same tier.

the A/D conversion would be allocated in the second and third tiers, rendering again the extrema location in the digital domain. The bottom tier is a memory to store scales and extrema or keypoints. Every keypoint is stored with its σ-level and its (x, y) coordinates in the image plane. It should be noted that this architecture would perform interest point detection. Feature description and matching would be implemented on a companion chip.

If only two tiers with an eventual memory block are available, the distribution of pixels and smart processing in two different tiers is not feasible. This approach is shown in Fig. 1(c), where sensing and Gaussian pyramid lie in the same tier. We leave the Harris, Hessian and DoG calculations, as well as extrema location for the bottom tier, while the eventual memory block would be used to store scales and extrema locations. Circuits for in-pixel A/D conversion are distributed among the top and the bottom tiers. Sensing and Gaussian pyramid generation with one pixel-per-processor assignment allows exploiting the inherent parallelism of low-level image processing. The degree of parallelism is reduced in the bottom tier. Finally, and as in the architecture of Fig. 1(b), the implementation of Fig. 1(c) is oriented to interest point detection; the difference is that now the limited number of tiers causes sensing and Gaussian pyramid generation to lie in the same tier, degrading the pixel pitch and fill factor.

III. CMOS-3D STACK FOR INTEREST POINT DETECTORS

Fig. 2 shows an architecture for interest point detection where sensors lie in a dedicated tier as in reference [49], so that different active pixels are possible. The last tier in this architecture is a DRAM memory block which is actually the implementation choice in [50]. Unfortunately, this block cannot be used as a frame buffer due to the fully-parallel

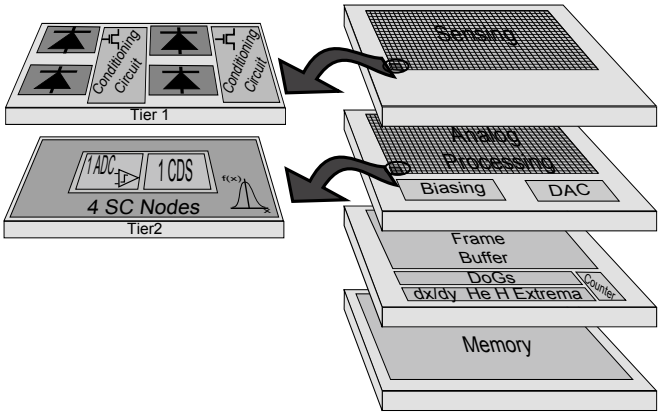

Fig. 2. Functionality distribution across tiers on a 3D-IC technology.

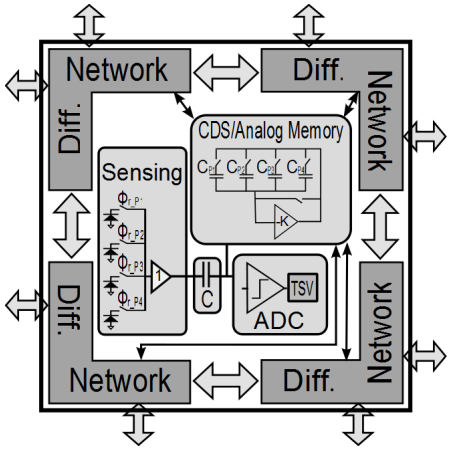

Fig. 3. Simplified schematic of the cell or processor element in the top tier of the CMOS-3D stack.

asynchronous access required by the top tier. For prototyping purposes, the sensors (tier 1 in Fig. 2) and the pre-processing (tier 2 in Fig. 2) are in the same tier - the one described below. The processing circuitry in this tier is devoted to the construction of the Gaussian pyramid. In the herein reported prototype each cell in this tier has 4 SC diffusion links, one CDS and part of an ADC, with the latter being distributed between two tiers. As a consequence, for an image of $M \times N$ pixels or photodiodes we need $1/4 \times M \times N$ cells. The fact that there is only one CDS and A/D converter per 4 photodiodes and 4 SC nodes causes the acquisition to be in 4 successive cycles, requiring additional switches to complete the conditioning circuits for sensing. Also, the analog to digital conversion of either the input image or a given scale takes 4 consecutive cycles. Note, however, that we still have a pixel per SC network node arrangement, keeping the parallelism in the Gaussian filtering.

The bottom tier contains the registers of the single slope A/D converter, one 8-bit register per pixel, and the circuits for the local extrema calculation of Harris, Hessian and DoG detectors. Finally, there is only one TSV per cell connecting the two tiers of processing. This TSV drives the enable signal of the registers of the bottom tier.

A. Top-tier Design

A simplified schematic of the cell in the top tier is shown in Fig. 3. Hardware re-using is a must in order to keep the cell with a reasonable area. The sensing is performed with the classical 3T Active Pixel Sensor (APS). The four photodiodes share a unity gain stage implemented with a source follower. The block labeled Analog Memory and CDS contains the four state capacitors (C_{Pi}) of the SC network, one per photosensor, which are also employed as analog memories to store the input image through the four cycles during the acquisition. The capacitor C in conjunction with the inverting stage K and the corresponding capacitor C_{Pi} perform in-pixel CDS. Four cycles are needed to complete the CDS acquisition of the 4 photodiodes. An offset-compensated comparator made up of the inverter within the ADC block and the capacitor C is the circuitry of the in-pixel 8-bits single slope ADC laid in the top tier. The ADC is completed with the registers in the bottom tier. The output of the comparator is the enable signal for such registers; hence only one TSV per cell is required to interconnect tiers. It should be noted that the top tier provides either the input image after running CDS, or a given diffusion or Gaussian filtered image with a particular sigma. In both cases, the four pixels per cell are stored in the state capacitors (C_{Pi}), and four cycles of conversion take place to transfer the whole image between tiers.

1) Nominal Analysis: Fig. 4 displays the schematic of the CDS circuit with its corresponding time diagram for a given photodiode or state capacitor (C_{Pi}). Signal V_{ref} is an analog reference common to the four state capacitors. The output of this circuit is given by Eq.(6), with $V_{Pi}(t_0)$ and $V_{Pi}(t_1)$ being the readings from the photodiode/pixel P_i at the beginning and end of the exposition time.

$$V(C_{Pi}) = V_{ref} + \frac{C}{C_{Pi}}(V_{Pi}(t_0) - V_{Pi}(t1)) \qquad (6)$$

Fig. 5 shows the complete scheme of the 8-bit single-slope A/D converter with its timing diagram. The signal V_{ramp} is the global ramp for all the in-pixel ADCs. The result of the comparator is given by Eq.(7). If the gain stage K is large enough, a small $(V_{ramp} - V_{Pi})$ value around V_Q, with V_Q being the quiescent point of the comparator, will lead the output of the comparator to a logic state, enabling/disabling the writing of the registers allocated in the bottom tier.

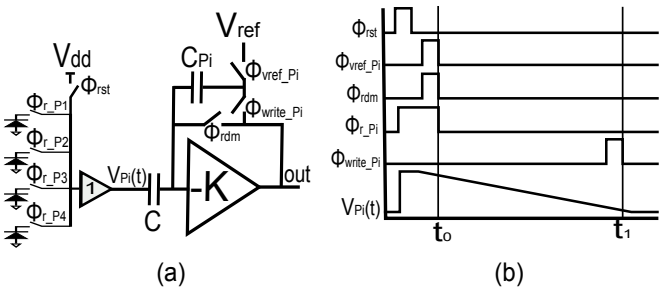

Fig. 4. Schematic and time diagram of the CDS circuit in the top-tier of the CMOS-3D stack.

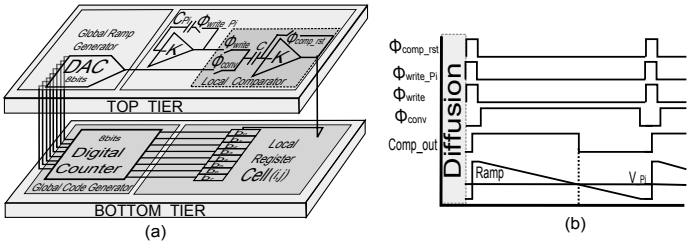

Fig. 5. Schematic and time diagram of the ADC distributed across the two tiers of the CMOS-3D stack.

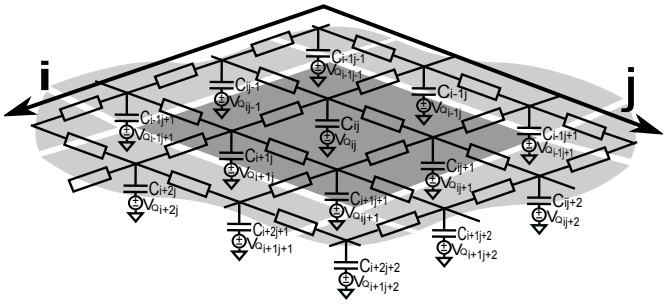

Fig. 6. Schematic of a switched-capacitor network.

$$V_{out} = -K(V_{ramp} - V_{Pi}) + V_Q \tag{7}$$

The Gaussian filtering, also called diffusion, is performed with a SC network. Fig. 6 and Fig. 7 sketch the SC network used for Gaussian pyramid construction. In Fig. 6 every state capacitor is identified as $C_{Pi,j}$ (C_{Pi} in Fig. 3), with i, j being the coordinates in the grid (image), and $V_{Pi,j}$ the voltage stored at capacitor $C_{Pi,j}$. The diffusion or exchange of charge takes place with the neighbors located along the four cardinal directions with two non-overlapping clock cycles (signals ϕ_1 and ϕ_2 in Fig.7(a) by means of the bi-linear SC network displayed on Fig. 7(a) [36]. Fig. 7(d) plots a 4×4 grid, in which every square represents a cell of 4 state capacitors. The state capacitors of every cell (4 pixels) are connected with the switched diffusion block A (Fig. 7(a)). The inter-cell connections among neighboring cells are implemented with the switched diffusion block B (7(d)). It should be noted that along every cardinal direction for inter-cell connectivity there are two switched diffusion blocks of type B between two nodes, making up the bi-linear SC implementation.

On the other hand, the σ level is defined by the number of cycles of ϕ_1 and ϕ_2 and the relation between the state capacitor $C_{Pi,j}$ and the exchange capacitors C_E. If we denote by n the number of cycles, the voltage at a state capacitor $C_{Pi,j}$ at cycle n is given by Eq.(8).

$$V_{ij}(n) = V_{ij}(n-1) + [V_{i-1j}(n-1) + V_{i+1j}(n-1) + \\ + V_{ij-1}(n-1) + V_{ij+1}(n-1) - 4V_{ij}(n-1)] \frac{\frac{C_E}{C}}{1+4\frac{C_E}{C}} \tag{8}$$

A similar equation can be found for the case of the Gaussian

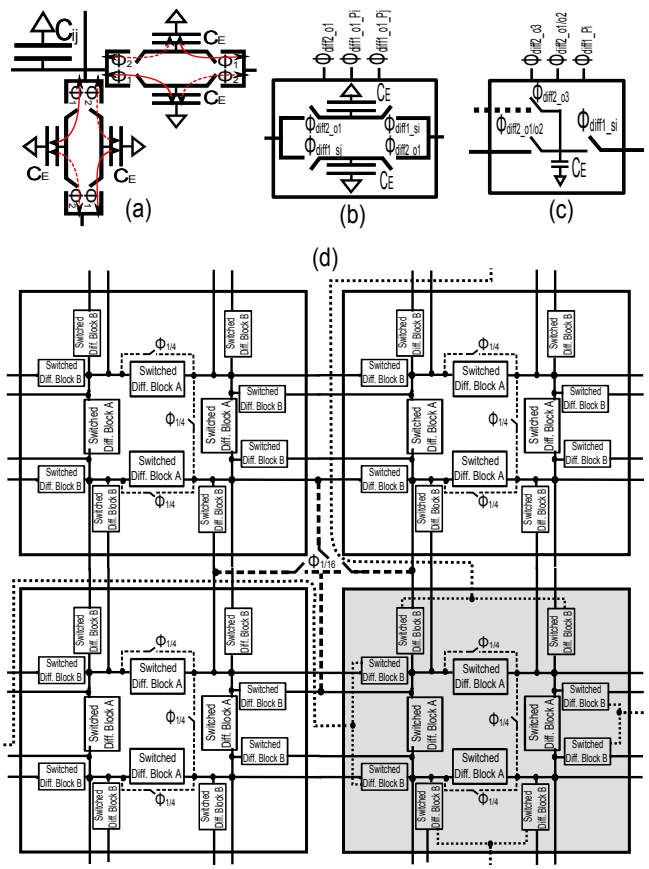

Fig. 7. Schematic of the: (a) Diffusion network node, (b) SC resistor (Block A), (c) half SC resistor (block B) and (d) 4×4 Network.

kernel, as listed in Eq.(9).

$$V_{ij}(n) = V_{ij}(n-1) + [V_{i-1j}(n-1) + V_{i+1j}(n-1) + \\ + V_{ij-1}(n-1) + V_{ij+1}(n-1) - 4V_{ij}(n-1)] \frac{e^{-\frac{1}{2\sigma^2}}}{1+4e^{-\frac{1}{2\sigma^2}}} \tag{9}$$

By looking at equations (7) and (8), it is easy to identify the σ achieved per each individual cycle [36], as indicated in Eq.(10).

$$\sigma_0 = \left(2 \times ln\frac{C_{i,j}}{C_E}\right)^{-1/2} \tag{10}$$

The application of two successive Gaussian filters with σ_0 is equivalent to a Gaussian kernel with a certain σ level. This allows our SC network, which has a level of filtering σ_0 fixed by the $C_{i,j}/C_E$ ratio, to approach up any Gaussian kernel by recursive filtering of Gaussian kernels with σ_0. The dependence of σ with the number of cycles n is given by Eq.(11), explained in [36].

$$\sigma = \sqrt{\frac{2nC_E}{4C_E + C_{i,j}}} \tag{11}$$

The sequence of operations required for the transition between octaves is shown in the time diagram of Fig. 8. Signals ϕ_{diff_si} are used for initialization purposes at the beginning of every scale [36]. Signals ϕ_{diff_oi} with the ending

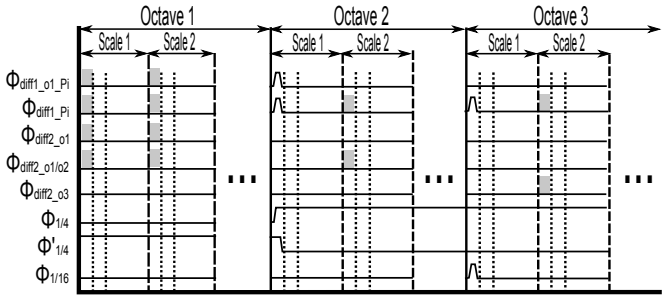

Fig. 8. Control signals for the generation of octaves and scales.

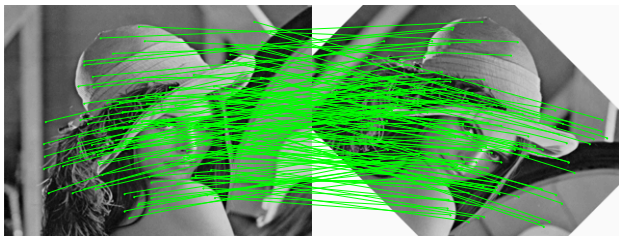

Fig. 9. Examples of matches between interest points of two images.

oi meaning the octave where they are employed. For instance, signal ϕ_{diff_o1} drives the switches that control the charge exchange among nodes of the SC network for the first octave, and signal $\phi_{diff\_o1/o2}$ does the same for the first two octaves. Signals $\phi_{1/4}$ and $\phi_{1/16}$ perform the $1/2$ downscaling between octaves, meaning $1/4$ of the pixel count, by merging the value of four state capacitors. Signal $\phi_{1/4}$ performs the binning of the four state capacitors within a cell, while signal $\phi_{1/16}$ does the same for inter-cell connectivity. It should be taken into account that when going from the first to the second octave the merging of four state capacitors changes the capacitance from $C_{Pi,j}$ to $4\times C_{Pi,j}$, while C_E changes to $2\times C_E$, modifying the ratio $C_{i,j}/C_E$, and thus the σ level. This is fixed by decreasing the new state capacitor from $4\times$ to $2 \times C_{i,j}$ with the switch $\phi'_{1/4}$ (not shown in the figures). The transition from the second to the third octave proceeds without any further modification. On this occasion, only one of the four cells would interact with its neighbors. Fig. 7 shades in gray color the cell that would interact with its neighbors in the third octave. The cell connections are indicated with dotted lines in Fig. 7.

2) Error Analyses: In this work the design parameters are set by the loss of true matches of interest points when comparing two objects with the SIFT algorithm. Fig. 9 shows an example of matching between interest points for an object with a known rotation. The true matches are obtained by the comparison of the descriptor vectors associated with every interest point, and by spatial matching. In the case of a known transformation, the spatial matching is easy to check by comparing the locations of the interest points of the query image (left hand-side of Fig. 9) after applying the known transformation with those of the image in the database (right hand-side of Fig. 9). The spatial distance between a pair of interest points should be inferior to a certain threshold. When a pair of keypoints complies with the descriptor vectors and the spatial criteria, they are sorted out as a true match. If the transformation is unknown, statistical methods like RANdom SAmple Consensus (RANSAC) are applied to provide the transformation [51].

The number of interest points and true matches should be large enough not to compromise the task (e.g. object detection). Unfortunately, a hardware implementation leads to an avoidable loss of true matches when compared to a pure software solution due to circuit non-idealities (e.g. finite gain) and parameter deviations as mismatch and global variations. In this work we have included the effect of hardware deviations following a normal distribution for the following parameters of an image with $M \times N$ resolution:

1) the gain of the source follower of every cell (a normal distribution of 1/4 $M \times N$ values)
2) the gain K of the inverter and the capacitor C employed for the CDS (a normal distribution of 1/4 $M \times N$ values)
3) the capacitors that make up the diffusion network ($C_{Pi,j}$ and C_E), (a normal distribution of $M \times N$ values)

The variation in the gain of the source follower K_{SF} for a given pixel i, j is gathered in Eq.(12).

$$V_{out_{i,j}} = \left(K_{SF_{i,j}} + \Delta K_{SF_{i,j}}\right) V_{in_{i,j}} \tag{12}$$

Eq. (13) conveys the variation of the capacitors C and $C_{Pi,j}$ in the output voltage of the CDS stage. C and $C_{Pi,j}$ are designed nominally identical. The dependence of C with the pixel is indicated with the sub index CDS_i,j in Eq. (13). Eq. (13) also shows the effect of a finite gain K, and its dependence with the pixel location i, j. Eq. (6) is the output voltage of the CDS stage without variations.

$$\begin{aligned} V_{out_{i,j}} = & \frac{K_{i,j}C_{CDS_{i,j}}}{C_P i_{i,j}+(K_{i,j}+1)C_{CDS_{i,j}}} V_{ref} + \\ & + \left[1 - \frac{K_{i,j}C_{CDS_{i,j}}}{C_{Pi,j}+(K_{i,j}+1)C_{CDS_{i,j}}}\right] V_Q \\ & + \frac{K_{i,j}C_{CDS_{i,j}}}{C_{Pi_{i,j}}+(K_{i,j}+1)C_{CDS_{i,j}}} \left[V_{ph}(t_0) - V_{ph}(t_1)\right] \end{aligned} \tag{13}$$

Eq. (14) captures the effects of variations of the state and exchange capacitors, $C_{Pi,j}$, and C_E, in the diffusion or Gaussian filtering. Eq. (8) represents the diffusion without variations.

$$\begin{aligned} V_{i,j}(n) = & \frac{V_{i,j}(n-1)C_{Pi,j}+V_{i,j+1}(n-1)C_{Ei,j+1}}{C_{Pi,j}+C_{Ei,j+1}+C_{Ei+1,j}+C_{Ei,j-11}+C_{Ei-1,j}} + \\ & \frac{V_{i+1,j}(n-1)C_{Ei+1,j}+V_{i,j-1}(n-1)C_{Ei,j-1}+V_{i-1j}(n-1)C_{Ei-1,j}}{C_{Pi,j}+C_{Ei,j+1}+C_{Ei+1,j}+C_{Ei-1,j}+C_{Ei-1,j}} \end{aligned} \tag{14}$$

Fig. 10 shows the effect of hardware variations in loss of true matches. These curves were taken for VGA images. The nominal gain of the inverter stage for CDS was fixed at K=2000, the nominal gain of the source follower at K_{SF}=0.9, and the nominal capacitances were C=150 fF, $C_{Pi,j}$=150 fF, and C_E=20 fF. These values are reasonable in a hardware implementation. The x axis represents the value of 6σ in the distribution of the parameters in the array. For the simulations, the implementation of the SIFT algorithm found in [52] was modified including the variations in K, K_{SF} and the capacitors C_{CDS} (labeled C_{nom} in Fig. 10), the state capacitors

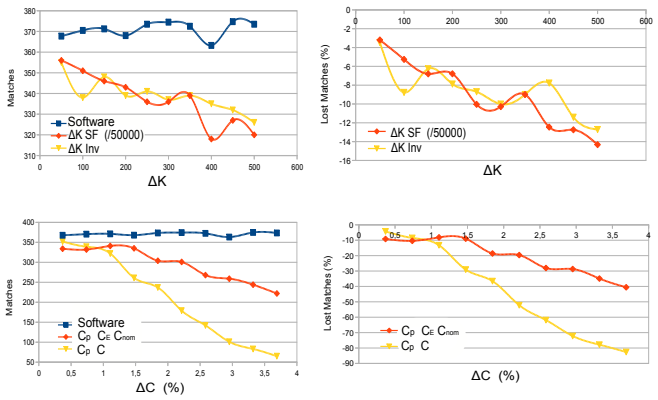

Fig. 10. Effect of variations in the gain of the source follower K_{SF}, the gain of the CDS stage K, and the capacitors used for the CDS stage C, and the capacitors of the SC network; the state C_P and exchange C_E capacitors.

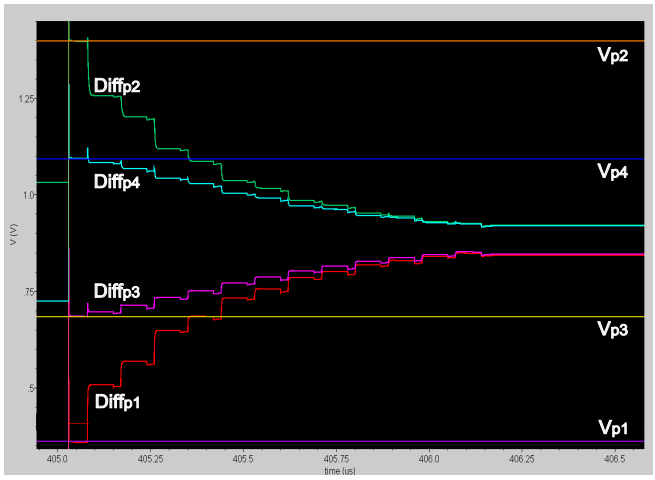

Fig. 11. Evolution of the voltage at the four state capacitors of a cell after 10 cycles of diffusion in the SC network.

TABLE I
ERRORS IN THE EVOLUTION OF THE FOUR CAPACITORS OF A CELL IN THE SC NETWORK IMPLEMENTED.

Diff	Vp1 E_rr (%)	Vp2 E_rr (%)	Vp3 E_rr (%)	Vp4 E_rr (%)
1	-0,3179	0,2912	0,4836	0,2254
2	0,5124	0,3382	0,5307	0,3633
3	-0,132	0,3338	0,147	0,2423
4	0,2742	0,469	0,4136	0,4386
5	-0,1112	0,3643	0,0728	0,2995
6	0,2283	0,5218	0,3405	0,4855
7	-0,0649	0,3601	0,0456	0,3196
8	0,2354	0,5287	0,3092	0,5077
9	-0,0173	0,339	0,0468	0,3192
10	0,2588	0,5201	0,3039	0,5106

$C_{Pi,j}$ (identified as C_P) and the exchange capacitors of the diffusion network C_E.

The top two sub-figures in Fig. 10 display the effect of variations in K, K_{SF}. The same variation in both parameters was used. In absolute terms the variation in K_{SF} is smaller than that of K, as K_{SF} is much smaller than K. In this case, variations in K and K_{SF} of $\Delta K = 6\sigma = 300$ for $K = 2000$, and $\Delta K_{SF} = 6\sigma = 7mV/V$ for $K_{SF} = 0.9$ produce an acceptable loss of around 10% of true matches. These variation levels are feasible in a circuit realization. We can also observe a smaller variation in the number of true matches in the software version caused by the statistical nature of the RANSAC method used for matching.

The bottom two sub-figures in Fig. 10 collect the effect of variations in C. We have adopted the same level of variation in all the capacitances. Two curves are plotted: 1) with variations in C_P and C_E, and 2) with variations in C_{CDS} and C_P. The variations in C_{CDS} and C_P yield a larger loss percentage of true matches. This is reasonable as the C_{CDS}/C_P ratio defines the gain of the CDS stage (see Eq.6). In this case, variations in C_{CDS} and C_P above 1%-2% would lead to a significant loss of true matches.

The cell of the top tier (see Fig. 3) has been implemented with the CMOS UMC 0.18 μm technology. The gain of the CDS and the ADC blocks is realized with a double cascode inverter. The state capacitors of the diffusion network C_{Pi} have been designed with an MIM structure with metal layers M5 and M6, and a nominal value $C_{Pij} = 150fF$, and the presence of a transistor with the drain and source terminals shorted. This enhances the capacitance without a significant degradation in area. This, in turn, helps decrease dynamic effects like charge injection and feedthrough.

Fig. 11 shows the evolution of the voltage at the four state capacitors of a cell after 10 cycles of diffusion in the SC network, with every diffusion cycle taking 90 ns (this time can easily be shortened). As we can see, the SC network settles down to a stationary state in less than 5 μs. Less than 50 μs would suffice for a complete Gaussian pyramid of 3 octaves with 6 scales each. These data are in line with those published in [16], outperforming conventional solutions like those found in [46] and [47] with a speed-up factor, exclusively considering Gaussian pyramid generation, close to 1000×.

Table I lists the voltage errors of our SC realization expressed as % at every node C_{Pi} when compared to the expected values of an error-free SC network. These errors should be added to the circuit deviations (mismatch and global variations) collected in Fig. 10. The cell has been designed to keep dynamic errors and circuit deviations from their nominal values at low enough levels as not to cause further losses of true matches. This, despite the analog processing, would permit to provide an overall output close to that of a pure software solution for the realization of a feature detector.

3) Top tier Performance Metrics: Fig. 12 shows the layout of a top-tier cell. This layout is used to extract metrics on area consumption, power dissipation and processing speed — used in Section III.C for comparison purposes. The photodiodes labeled $PD1$ - $PD4$ occupy an area of 8 μm × 8 μm. They are made up of an n-type diffusion over a p-type substrate to avoid the area overhead of an n-type well. The state capacitors (C_{Pi}) are laid down with metal layers M4 and M5, thus allowing to place circuitry underneath. The circuitry for charge exchange with the neighbors along the north direction is below the state capacitor between the photodiodes $PD1$ and $PD2$. The circuits for connecting neighbors across the south direction are

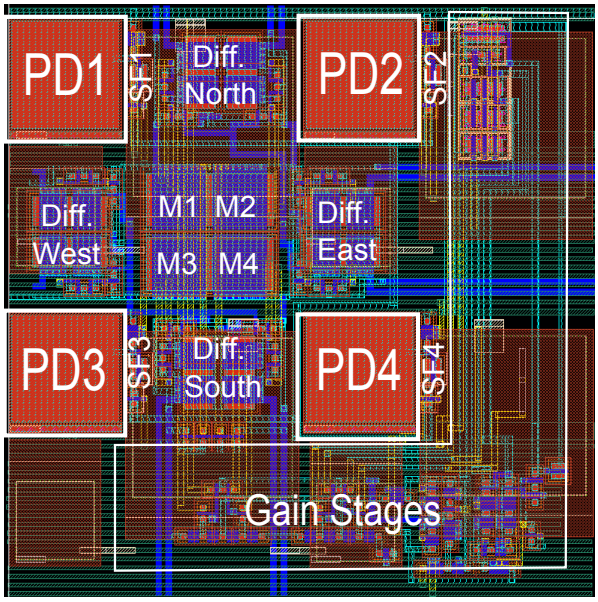

Fig. 12. Layout of a cell in the top tier.

below the capacitor between the photodiodes $PD3$ and $PD4$: the connections with the western and eastern neighbors are below the capacitors between $PD1$ and $PD3$, and $PD2$ and $PD4$, respectively. Switches for configurability are distributed below the capacitor between $PD3$ and $PD4$ and the rest of the cell. The gain stages are also labeled in Fig. 12. The area occupied by a cell is 43×43 μm^2 which results into 475μm^2 per pixel and amounts up to 36.5mm^2 for a QVGA sensor.

Three stages are identified regarding power consumption and processing speed:

1) the acquisition of the scene by the photodiodes
2) the storage of the acquired image into the state capacitors CPi and the diffusion of Gaussian filtering in the SC network
3) the in-pixel A/D conversion

Regarding acquisition phase the following has to be considered: i) the photodiodes and ii) the source follower. In a worst case scenario the voltage at the photodiodes changes from 0 to V_dd (1.8V). In the CMOS UMC 0.18μm technology the intrinsic capacitance amounts to 54fF for a 8×8 μm^2 photodiode. The source follower consumes power during the reset and during the sampling photodiode phases, which last for 1.5μs and 10μs respectively. The bias current for the source follower is set to 1μA. This leads to a total charge variation during the acquisition of ΔQ_{acq}=46.4pC per cell (four photodiodes) — 11.6pC per pixel.

The power dissipated by the storage of the acquired scene in the state capacitors and the diffusion or Gaussian filtering in the SC network can be estimated taking into account that: i) the gain stage used as buffer (see Fig. 3) is biased at 1μA; ii) the gain stage consumes power in voltage follower mode throughout the sampling of the acquired image by the source followers and their reset, T_{sf} = 11.5 μs, and during the reading of the values stored at the state capacitors C_{Pi}, namely, the Gaussian pyramid; and iii) 18 scales are generated, with T_{diff} = 2.5μs per scale. The total charge variation caused by storage

TABLE II
FIGURES OF MERIT OF THE CIRCUITS IN THE TOP TIER

Area	Processing Speed Gauss. Pyramid	Power Dissipation
36.5 mm^2	50 μs w/o ADC 3.74 ms w ADC	75 mW w A/D and im. acq.

of values at the state capacitors and the Gaussian filtering, ΔQ_{mem}, is given in Eq. (15), amounting to 110.5pC per pixel. Notice that 36 readings with T_{vf} = 1.5μs per reading cycle are needed for the subsequent A/D conversion of every scale required for DoGs; 24 for the first octave, as there are 4 readings per cell, and 6 readings per octave for the second and third octaves.

$$\Delta Q_{mem} = 1\mu A \times (T_{sf} + T_{vf} \times 36 + T_{diff} \times 18) = 110.5pC \quad (15)$$

Finally, the charge variation due to the in-pixel A/D conversion has been estimated through electrical simulations of a complete ramp (single-slope), amounting to approximately 110pC per pixel, which, with 36 conversions, yields a total charge variation due to the A/D of ΔQ_{AD} = 3960pC per pixel. All in all, the charge variation per pixel results in ΔQ_{px} = 4082.1pC.

The computation time is estimated through Eq. (16), assuming: i) an integration (exposure) time T_{int} = 120μs; ii) a sampling and reset time through the source followers T_{sf} = 11.5 μs; iii) a time for every reading through the gain stage in voltage follower mode T_{vf} = 1.5 μs; iv) a time for scale T_{diff} = 2.5μs; and v) a time for comparison estimated through electrical simulations for the in-pixel ADC, T_{AD} = 100μs per scale at most. This leads to a time per frame T_{fr} = 4.27ms; 3.74ms without the acquisition time.

$$T_{fr} = (T_{sf} + T_{int}) \times 4 + (T_{vf} + T_{diff} + T_{AD}) \times 36 = 4.27ms \quad (16)$$

The above data lead to current per pixel $I_{px} \approx 1\mu$A, which amounts to 1.8μW per pixel, leading to 150mW for a QVGA array as the worst case. We have estimated an average consumption of 75mW for the whole array.

Table II includes previously calculated metrics. It is seen that the A/D conversion of every scale penalizes the processing speed. This can be partially alleviated through the distribution of photodiodes and circuitry in two different tiers. If this was case, every photodiode would be assigned to one cell, cutting the number of A/D conversions from 24 down to 6 for the first octave, i.e. from 3.74ms to 1.8ms. Note that no other A/D conversion is required, and that in a more conventional approach with a digital chip like those of references [46] and [47], the image is taken from an external camera board; camera board and chip combined would have a similar bottleneck due to A/D conversion.

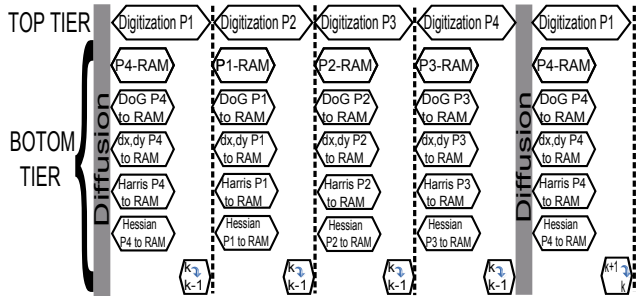

Fig. 13. Sequence of operations in the CMOS-3D stack.

B. Bottom-tier Design

As it was said before, the A/D converter is shared among the two tiers, with the comparators in the top tier, and a set of $M \times N$ 8-bit registers in the bottom tier for an image of $M \times N$ resolution. Such registers store the scales of the Gaussian pyramid or the input image. The registers are arranged in a mesh of 1/4 $M \times N$ banks. Every bank contains 6 8-bit registers; 4 of them for the scale $k-1$, and two for the current scale k. This permits to process the image at scale $k-1$ while the image at scale k is being read from the top tier as the A/D conversion is being performed. The pixels of scale $k-1$ stored in every bank register are named as: $P1$, $P2$, $P3$ and $P4$, corresponding with image coordinates: (i, j), $(i, j+1)$, $(i+1, j)$ and $(i+1, j+1)$. The four pixels $P1 - P4$ are digitized in series; hence, four conversion cycles are needed for the whole image in the first octave. The 1/4 downscaling for the second and third octaves, however, permits to perform the analog to digital conversion of the whole image in just one cycle.

Fig. 13 shows the operations executed in the CMOS-3D stack for the first octave. The top tier provides the diffusion or scale k. After this, several operations are run in parallel along the four cycles needed to perform the A/D conversion of the whole $M \times N$ image for the first octave. The operations running during the analog to digital conversion of pixel $P1$ at scale k are:

1) writing of pixel $P4$ of scale $k-1$ into the memory
2) the difference of Gaussian between scales $k-1$ and $k-2$, namely DoG(k-1)
3) horizontal, dx, and vertical derivates, dy, which are used not only for Harris, but also for subsequent stages of the SIFT algorithm
4) Harris and Hessian detection over scales

The results of the above operations are sorted out in groups of 128 bits (16 words of 8 bits each) and transfer in burst mode to the DRAM memory. Fig. 14 depicts the architecture of the circuit located in the bottom tier. The images from the buffer array are read row by row in groups of 20 registers in order to provide the first and second derivatives along rows and columns (as we will see below, the two nearest neighbors along a row are required to calculate the second derivative). For every row i, the 20 columns of pixels Pi within a row are selected through the multiplexers seen in Fig. 14. Four multiplexers are needed for this task. Two of them are shared by the first and the second octaves for scales k and $k-1$.

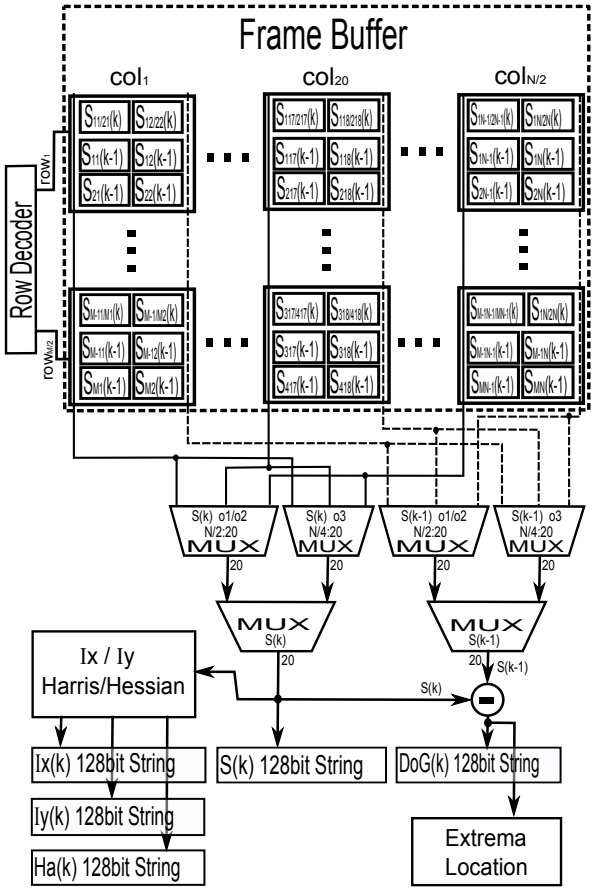

Fig. 14. Architecture of the digital circuit located in the bottom tier of the CMOS-3D stack.

Both scales are required in the DoG calculation. The two other multiplexers are employed for the third octave. It should be noted that for the first and the second octave, the multiplexers can be shared, as we always access all the registers along a row. In the case of the first octave we need 4 reading cycles to transfer pixels $P1 - P4$. In the second octave and beyond, we do such a transfer in only one cycle due to the 1/4 downscaling between octaves. The lowest frequency needed for reading and performing all these operations is set by the first octave, being 10 Mhz in order to read all the pixels Pi in less than 100 μs, which is the time for the A/D conversion.

Fig. 15 shows the schematic of a bank register in the bottom tier. Every bank comprises 6 8-bit registers, and as said before, there is only one TSV per register bank connecting the two tiers. This is a 1-bit signal driving two AND gates with ϕ_{conv13} and ϕ_{conv24} as inputs, yielding the enable signals for the top two registers, R_13_K and R_24_K. The top two registers store the pixels of scale k. The four bottom registers keep pixels $P1 - P4$ for scale $k-1$. Scales k and $k-1$ are available on the corresponding buses of every set of registers for DoG calculation. The sequence of operations to achieve every scale in the first octave is as follows. Pixel $P1$ is digitized into register R_13_K with ϕ_{conv13} on. Subsequently, pixel $P2$ is digitized and stored in register R_24_K, following a similar process with signal ϕ_{conv24} on. During this phase the DoG for all pixels $P1$ of scale $k-1$ are calculated and

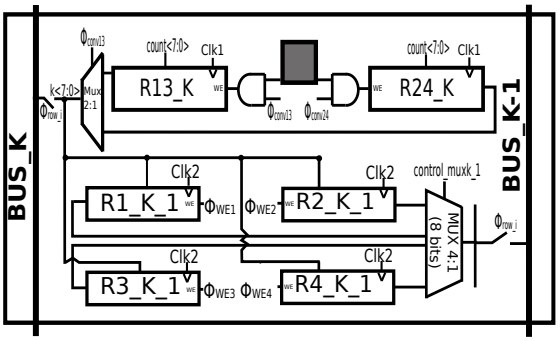

Fig. 15. A register bank of the bottom tier. R_13_K and R_24_K let the conversion of scale k while scale $k-1$ is being read. The other registers store the pixel values of scale $k-1$.

written into the DRAM. After reading pixels $P1$ the content of register R_13_K is transferred into register R_1_K by means of signal ϕ_{WE1} on. Later on, pixels $P3$ are digitized in register R_13_K while the pixels $P2$ are read, and the process continues up to pixels $P4$, completing the first octave.

The pixel arrangement of four pixels within a cell along with their serial analog to digital conversion causes not to have all the pixels available at one cycle of digitization. For instance, during the digitization of pixel $P2$ at scale k, only the pixels $P1$ of the scale k are available in the register banks. Besides, the reading of the register banks for further processing is carried out in a row by row order. As a result, only the derivatives along certain directions can be provided. It is possible, for instance, to calculate the derivative along the x direction (dx) of pixel $P2$ during the cycle of analog to digital conversion of $P2$, by accessing the adjacent pixels $P1$ along a row. It would not be possible, however, to provide the derivatives along the vertical direction, because pixels $P3$ are not available yet.

The solution adopted in our architecture to overcome this issue is to calculate the first derivatives or gradient along a different set of axis which has been rotated 45 degrees with respect to the conventional coordinates. The gradient is now calculated with the next set of equations:

$$d_x^{'}(i,j) = I(i+1,j+1) - I(i-1,j-1) \tag{17}$$

$$d_y^{'}(i,j) = I(i+1,j-1) - I(i-1,j+1) \tag{18}$$

This process is illustrated in Fig. 16, which shows a set of register banks in the bottom tier. In this case, only the pixels $P1$, marked in gray, are available. The pixels where the derivatives are calculated are marked with an X. It is possible to provide the derivatives along the new x and y axes for pixels $P2$ as indicated in Fig. 16. This process is being done at the same time as the reading of the register banks.

The strategy we followed for the second derivatives, needed for the Hessian matrix, is different. An approach to the second derivative is made by recreating the neighbor located one pixel apart along the horizontal and vertical directions. This is done by interpolating the pixels located two pixels apart from the one under study. Thus, in this procedure, the neighbor at $(i+1,j)$ is generated as $I(i+1,j) = [I(i+2,j) + I(i,j)]/2$, yielding the following set of equations:

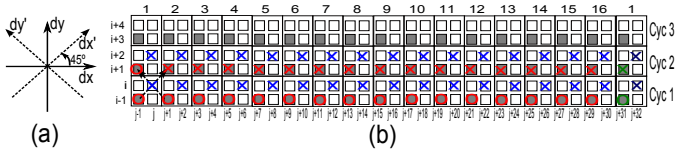

Fig. 16. Arrangement for gradient calculation.

$$I_{xx}(i,j) = [I(i,j+n)) + I(i,j-n) - 2*I(i,j)]/n \tag{19}$$

$$I_{yy}(i,j) = [I(i+n,j) + I(i-n,j)) - 2*I(i,j)]/n \tag{20}$$

$$I_{xy}(i,j) = [I(i+n,j+n) + I(i-n,j-n) - \\ -I(i-n,j+n]/n \tag{21}$$

where $n = 2$ for the first octave, and $n = 1$ for the next octaves, given that at the second and third octaves every pixel has the right neighbors along horizontal and vertical directions to perform the Hessian matrix [53].

The frequency specifications in our system are set by the operations run in parallel during the A/D conversion of a given pixel (see Fig. 9), which was estimated at 100 μs. Two types of operations are run in parallel, namely, data calculation and memory-writing.

In our design, the DoGs, dx and dy are provided in groups of 16 pixels, so for a given pixel P_i, the total number of clock cycles is given by: $M \times N/(16 \times 4)$. In a VGA image, this renders a minimum clock frequency of 10 MHz. Our circuit has been synthesized on a Virtex-6 from Xilinx, reaching 375 MHz. This frequency would lead to less than 12 μs for the DoGs, dx and dy in a VGA image, leaving still 88 μs for their memory storage.

C. Comparison with State-of-the-Art Chips

The circuits in the top tier can be compared with state-of-the-art custom chip performing similar operations, like Gaussian filtering, Gaussian pyramid generation, feature detection, feature matching, SIFT, etc. The implementation based on FPGA reported in [45] is not included in the comparison because no data concerning power consumption and area occupation is provided. Three different figures of merit have been considered: i) the time it takes to realize the corresponding operation, what gives an idea of its speed; ii) the power per pixel and fps, what is a measure of the energy efficiency; and iii) fps per mm^2 scaled by a normalized resolution –actually 1920×1080-pixels, the resolution of the chip in [47]–, what gives an idea of the area efficiency of the implementation in connection with the processing speed.

It is important to mention that the functions implemented by the different chips are relatively diverse, i. e. in some of them only Gaussian filtering is performed [16], in others the DoG is available [54], other chips have the complete feature extraction [46] [47]. Some chips contain photosensors [54] [7] [55] [16], and the subsequent ADC, and some others need to be fed with an already digitized image [46] [47]. Therefore, we are comparing quite heterogeneous systems.

TABLE III
COMPARISON WITH SIMILAR STATE-OF-THE-ART CUSTOM CHIPS

Chip (technology) and functionality	Time per task (μs)	Power/ (fps·px) (nJ/px)	(fps·px)/ (Area·HD) (s·mm^2)$^-$1
Ref. [54] (0.8μm CMOS) Gaussian/DoG filter + local extrema detection	42	300	0.006
Ref. [7] (0.35μm CMOS) Multiresolution imaging + motion detection	1040	76	0.019
Ref. [55] (0.35μm CMOS) Edge filtering based on convolution masks	1910	130	0.014
Ref. [16] (0.35μm CMOS) Gaussian filtering by focal-plane diffusion	1.5 w/o ADC	400	0.009
Ref. [46] (0.13μm CMOS) Feature detection and matching QVGA	5400 (only feat. det.)	37 (SIFT)	0.09 (SIFT)
Ref. [47] (65nm CMOS) Feature detection on full-HD	110.9 (only 1 feat.)	0.8 (SIFT)	4.7 (SIFT)
This work (0.18μm CMOS) Gaussian pyramid QVGA	50 (pyram. gen.)	3.6 w ADC 0.05 w/o ADC	0.27 w ADC 20.3 w/o ADC

Concerning the first column of Table III, it can be seen that our chip is very fast in computing the Gaussian pyramid. It is clear that the difference between the time it takes to provide the complete Gaussian pyramid for a QVGA image (50μs) is very competitive. Then A/D conversion requires some additional 3.74ms. DoG and extrema detection implemented in a Virtex-6 FPGA from Xilinx take 12μs for a VGA image. Concerning the times reported in [47] it is not clear how many keypoints they are considering, therefore, their operation time can be considerably longer.

In the second column of Table III, the energy efficiency of the system is shown. The proposed chip is the most efficient if the fact that the chip in [47] does not contain photosensors nor A/D converters are considered.

In the third column, performances distribute along three orders of magnitude, being the number of pixels per frame the major factor in order to distinguish the best performer. This supports the idea of vertically integrating functionality in order to maintain a large pixel count.

IV. CONCLUSION

There are many applications where enhanced image resolution, the basic challenge for consumer applications, must be complemented with other features such as speed and smartness. For instance, sensors intended for surveillance applications should be capable to analyzing complex spatial-temporal scenes and combining high-quality image recording of significant events with high-speed decision making. Just to mention another example, scientific applications call for the smart selection of salient points and region-of-interests and for the ultra-high-speed downloading of the so-selected areas. Also, machine vision sensors (as those employed for inspection) require image content analysis and decision making to be completed with the largest possible throughput. These features call for the embedding of processing circuitry within the sensor chip. Conventional architectures for such embedding consist of a sensor array, a readout section and a data conversion section followed by digital processing block. Analysis of the power budget in this conventional architecture shows that most of the power is used by the digital processing section owing to the necessity to handle huge amount of data. For increased efficiency, alternative architectures consisting of distributed, multi-core processor arrays to enable progressive processing, and hence reduction of the data as they proceed through the processing chain, are worth considering. These multi-core architectures largely benefit from the possibility of arranging the required functions into vertically interconnected tiers, as it is actually enabled by 3D integration technologies. This paper shows that complex interest point detection algorithms can be mapped onto multi-layered architectures suitable for 3D implementation which report significant speed advantages as compared to conventional solutions. Such speed advantage becomes more evident at the lowest level of processing, especially in the generation of the Gaussian pyramid, a key issue for interest point detectors might take up to 90% of the computation time of a modern scale- and rotation- invariant feature detector as SIFT. The paper shows that using a SC network provides a 1000x speed enhancement in the generation of Gaussian pyramids when compared to conventional solutions.

REFERENCES

[1] J. Ohta, Smart CMOS Image Sensors And Applications. CRC Press, 2008.
[2] H. Abe, "Device technologies for high quality and smaller pixel in CCD and CMOS image sensors,", *IEEE International Electron Devices Meeting, Technical Digest 2004*,pp. 989- 992, 13-15 Dec. 2004.
[3] http://www.chipworks.com/. "A Survey of Recent Image Sensor Pixel Structures", May 2012.
[4] I. Takayanagi and J. Nakamura, "High-Resolution CMOS Video Image Sensors". *Proceedings of the IEEE, to appear*.
[5] R. Xu et al., "A 1500 fps Highly Sensitive 256 $\times$ 256 CMOS Imaging Sensor With In-Pixel Calibration". *IEEE J. Solid-State Circuits*, vol. 47, no. 6, pp. 1408-1418, June 2012.
[6] D. Hertel et al., "A Low-Cost VIS-NIR True Color Night Vision Based on a Wide Dynamic Range CMOS Imager". *2009 IEEE Intelligent Vehicles Symposium*, pp. 273-278, 2009.
[7] J. Choi et al., "A Spatial-Temporal Multiresolution CMOS Image Sensor with Adaptive Frame Rates for Tracking the Moving Objects in Region-of-Interest and Suppressing Motion Blur". *IEEE J. Solid-State Circuits*, vol. 42, no. 12, pp. 2978-2989, Dec. 2007.
[8] Tower Jazz Semiconductors. http://www.jazzsemi.com/.
[9] Sony Corporation. "The Advantage of the CMOS Sensor", April 2011.
[10] Yole Development. http://www.yole.fr/.
[11] AnaFocus Ltd. http://www.anafocus.com/.
[12] Omnivision. http://www.ovt.com/.
[13] J.E. Eklund et al., "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.
[14] A. Rodríguez-Vázquez et al., "ACE16k: the third generation of mixed-signal SIMD-CNN ACE chips toward VSoCs". *IEEE Trans. on Circuits and Systems I*, vol. 51, no. 5, pp. 851-863, May 2004.
[15] 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.

[16] 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.
[17] W. Zhang et al., "A Programmable Vision Chip Based on Multiple Levels of Parallel Processors". *IEEE J. Solid-State Circuits*, vol.46, no.9, pp.2132-2147, Sept. 2011.
[18] J. Fernández-Berni et al., *Low-Power Smart Imagers for Vision-Enabled Sensor Networks*. Springer, May 2012.
[19] J. Poikonen et al., "MIPA4k: A 64x64 cell mixed-mode image processor array". *IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1927-1930, 2009.
[20] A. Lopich and P. Dudek, "A SIMD Cellular Processor Array Vision Chip With Asynchronous Processing Capabilities". *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 58, no. 10, pp. 2420-2431, 2011.
[21] W. Miao et al., "A programmable SIMD vision chip for real-time vision applications". *IEEE Journal of Solid-State Circuits*, vol. 43, no. 6, pp. 1470-1479, 2008.
[22] W. Zhang et al, "A Programmable Vision Chip Based on Multiple Levels of Parallel Processors". *IEEE Journal of Solid-State Circuits*, vol. 46, no. 9, pp. 2132-2147, 2011.
[23] A. A. Abbo et al, "Xetal-II: A 107 GOPS, 600 mW Massively Parallel Processor for Video Scene Analysis". *IEEE J. Solid-State Circuits*, vol. 43, no. 1, pp. 192-201, 2008.
[24] T. Komuro et al, "A dynamically reconfigurable SIMD processor for a vision chip". *IEEE Journal of Solid-State Circuits*, Vol. 39, No. 1, pp. 265-268, 2004.
[25] A. Dupret et al, "A DSP-like analogue processing unit for smart image sensors". *Int. J. of Circuit Theory and Applications*, vol. 30, pp. 595-609, 2002.
[26] S. Lee et al, "24-GOPS 4.5-mm2 Digital Cellular Neural Network for Rapid Visual Attention in an Object-Recognition SoC". *IEEE Transactions on Neural Networks*, vol. 22, no. 1, pp. 64-73, 2011
[27] Alireza Moini, Vision Chips. Springer, 2000.
[28] C. Tomasi, Early Vision. *Encyclopedia of Cognitive Sciences*. Nature Pub. Group. McMillan 2002.
[29] 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.
[30] P. Garrou et al., "Handbook of 3D Integration: Technology and Applications of 3D Integrated Circuits", Wiley-VCH, 2008.
[31] 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.
[32] C. Harris and M. Stephens, "A Combined Corner and Edge Detection". *4th Alvey Vision Conference*, pp. 147-151, 1988.
[33] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints". *International Journal of Computer Vision*, vol. 60(2): 91-110, 2004.
[34] H. Bay et al., "Speed-Up Robust Features (SURF)". *Computer Vision and Image Understanding (CVIU)*, vol. 110, no. 3, pp. 346-359, 2008.
[35] H. Aanaes et al., "Interesting Interest Points: A Comparative Study of Interest Point Performance on a Unique Data Set". *Int. J. of Computer Vision*, vol. 97, pp.18-35, 2012.
[36] M. Suárez et al., "Switched-Capacitor Networks for Scale-Space Generation". *20th European Conference on Circuit Theory and Design*, pp.189-192, Linkoping, Sweden, August 29-31, 2011.
[37] J. Fernández-Berni et al., " "Multi-Resolution Low-Power Gaussian Filtering by Reconfigurable Focal-Plane Binning". *Proc. SPIE 8068, 806806*, 2011.
[38] T. Tuytelaars and K. Mikolajczyk, "Local Invariant Feature Detectors: A Survey", *Foundations and Trends in Computer Graphics and Vision*, vol. 3, no. 3, Now Publishers Inc., 2008.
[39] 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.
[40] K. Mikolajczyk and C. Schmid, "Scale & Affine Invariant Interest Point Detectors". *Int. J. of Computer Vision*, vol. 60(1), pp. 63-86, 2004.
[41] J. Matas et al., "Robust Wide-Baseline Stereo from Maximally Stable Extremal Regions". *Image and Vision Computing*, vol. 22(10), pp. 761-767, 2004.
[42] M. Trajkovic and M. Hedley, "Fast Corner Detection". *Image and Vision Computing*, vol. 16(2), pp. 75-87, 1998.
[43] K. Mizuno et al., "Fast and Low-Memory-Bandwidth Architecture of SIFT Descriptor Generation with Scalability on Speed and Accuracy for VGA Video". *International Conference on Field Programmable and Applications 2010 (FPL 2010)*, pp. 608-611, 2010.
[44] V. Bonato et al., "A Parallel Hardware Architecture for Scale and Rotation Invariant Feature Detector". *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 18, no. 12, pp. 1703-1712, Dec. 2008.
[45] F.C. Huan et al., "High-Performance SIFT Hardware Accelerator for Real-Time Image Feature Extraction". *IEEE Tran. on Circuits and Systems for Video Technology*, vol. 22, no. 3, pp. 340-351, March 2012.
[46] S. Lee et al., "A 345 mW Heterogeneous Many-Core Processor With an Intelligent Inference Engine for Robust Object Recognition". *IEEE J. of Solid-State Circuits*, vol. 46, no. 1, pp. 42-51, January 2011.
[47] 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.
[48] 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.
[49] A. Rodríguez-Vázquez et al.,"A 3D Chip Architecture For Optical Sensing and Concurrent Processing", *SPIE Photonics Europe, Optical Sensing and Detection.*, 12-15 April 2010. Proceedings of SPIE Volume: 7726-39. DOI: 10.1117/12.855027.
[50] Tezzaron Semiconductors. http://www.tezzaron.com/.
[51] M.A. Fischler and R.C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography". *Comm. of the ACM*, vol. 24, pp 381-395, 1981.
[52] A. Vedaldi. http://www.vlfeat.org/ vedaldi/code/sift.html.
[53] B. Han et al., "Optimized Algorithm for Gaussian Second Order Derivative Filters in Fast Hessian Detector," *2010 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT)*,vol.4, pp.623-627, 9-11 July 2010.
[54] Y. Ni and J. Guan, "A 256x256 pixel smart CMOS image sensor for line-based stereo vision applications". *IEEE J. of Solid-State Circuits*, vol. 35, no. 7, pp. 1055-1061, July 2000.
[55] N. Takahashi, K. Fujita T. Shibata, "A Pixel-Parallel Self-Similitude Processing for Multiple-Resolution Edge-Filtering Analog Image Sensors". *IEEE Trans. on CAS-I*, vol. 562, no. 11, pp. 2384-2392, Nov. 2009.

PLACE PHOTO HERE

M. Suárez is currently a PhD Student at Centro de Investigación en Tecnologías de la Información (CITIUS), University of Santiago de Compostela, SPAIN. His main research interests lie in the field of the implementation of feature detectors on focal-plane chips with CMOS-3D technologies.

PLACE PHOTO HERE

V.M.Brea received his PhD in Physics in 2003. Currently he is an Associate Professor at Centro de Investigación en Tecnologías de la Información (CITIUS), University of Santiago de Compostela, SPAIN. His main researh interests lie in the design of efficient architectures and CMOS solutions for computer vision, especially in early vision. Dr. Brea has been the Program Chair of 11th International Workshop on Cellular Neural Networks and Their Applications, CNNA 2008, held in Santiago de Compostela. He is also co-recepient of the Best Paper Award of the European Conference on Circuit Theory and Design in 2003.

PLACE PHOTO HERE

J. Fernández-Berni received the B.Eng. degree in electronics and telecommunication in September 2004 from the University of Seville, Spain. He then spent four months at the Center for Astrobiology (CAB) in Madrid, Spain, granted by the Spanish National Research Council (CSIC). From January 2005 to September 2006, he was working in the telecommunication industry. In October 2006, he joined the Institute of Microelectronics of Seville (IMSE-CNM-CSIC) as a graduate student, receiving the M.Sc. degree in microelectronics in December 2008, and his PhD in 2011. He was visiting the Computer and Automation Research Institute of the Hungarian Academy of Science in Budapest for a term, where he worked in vision system integration. His main areas of interest are mixed-signal design and VLSI implementation of low-power focal-plane processing arrays with applications to vision-enabled wireless sensor networks.

PLACE PHOTO HERE

R. Carmona-Galán (M'04) graduated in Physics and received the Ph.D. in microelectronics from the University of Seville, Spain, in 1993 and 2002, respectively. He started research activities in analog and mixed-signal integrated circuit design as an undergraduate student at the National Center for Microelectronics (CSIC) in Seville. From 1994 to 1996 he was funded by Iberdrola S.A. From July 1996 to June 1998, he worked as a Research Assistant at Prof. Chuas laboratory in the EECS Department of the University of California, Berkeley. From 1999 to 2005 he was an Assistant Professor of the Department of Electronics of the University of Seville. He taught circuit analysis and synthesis at the School of Engineering. Since 2005, he has been a Tenured Scientist at the Institute of Microelectronics of Seville (IMSE-CNM-CSIC). His main research focus has been on VLSI implementation of concurrent sensor/processor arrays for real-time image processing and vision. He also held a PostDoc at the University of Notre Dame, Indiana (2006 - 2007), where he worked in interfaces for CMOS-compatible nanostructures for multispectral light sensing. He has collaborated with start-up companies in Seville (AnaFocus) and Berkeley (Eutecus). He has designed several vision chips implementing different focalplane operators for early vision processing. His current research interests lie in the design of low-power smart image sensors and 3-D integrated circuits for autonomous vision systems. He has authored more than 60 papers in refereed journals and conferences and several book chapters.

Dr. Carmona-Galán received a Best Paper Award in 1999 from the International Journal of Circuit Theory and Applications. He is a co-recipient of an award of the ACET in 2002 and a Certificate of Teaching Excellence from the University of Seville.

PLACE PHOTO HERE

G. Liñán received the Licenciado and Doctor (Ph.D.) degrees in physics, with specialty in electronics, from the University of Seville, Seville, Spain, in 1996 and 2002, respectively. In 1995, he was a Graduate Student of the Spanish Ministry of Education at the Institute of Microelectronics of Seville (IMSE), National Microelectronics Center (CNM), Spanish Microelectronics Center (CSIC), Seville, where he also received a doctoral grant from 1997 to 1999, which was funded by the Andalusian Government, and where he is currently with the Department of Analog and Mixed-Signal Circuit Design. From February 2000 to June 2004, he was an Assistant Professor with the Department of Electronics and Electromagnetism, School of Engineering, University of Seville and at the Faculty of Physics. Since June 2004, he has been a Tenured Scientist of the Spanish Council of Research. His main areas of interest include the design and VLSI implementation of massively parallel analog-/mixed-signal image processors.

Dr. Linán received the 1999 Best Paper Award and the 2002 Best Paper Award from the International Journal of Circuit Theory and Applications.He is also a corecipient of the Most Original Project Award and the 2002 Salv i Campillo Award, which were conceded by the Catalonian Association of Telecommunication Engineers.

PLACE PHOTO HERE

D. Cabello (M'96) received the B.Sc. and Ph.D. degrees in Physics from the University of Granada, Granada, Spain, and the University of Santiago de Compostela, Santiago de Compostela, Spain, in 1978 and 1984, respectively. Currently, he is a Professor of Electronics at Centro de Investigación en Tecnologías de la Información (CITIUS), University of Santiago de Compostela, SPAIN. He has been the Dean in the Faculty of Physics between 1997 and 2002, and the Head of the Department of Electronics and Computer Science between 2002 and 2006, both in the University of Santiago de Compostela. His main researh interests lie in the design of efficient architectures and CMOS solutions for computer vision, especially in early vision. Professor Cabello has been General Co-Chair of 11th International Workshop on Cellular Neural Networks and Their Applications, CNNA 2008, held in Santiago de Compostela. He is also co-recepient of the Best Paper Award of the European Conference on Circuit Theory and Design in 2003.

PLACE PHOTO HERE

Ángel Rodríguez-Vázquez (F'96) is currently a Full Professor of electronics with the University of Seville, Seville, Spain and is appointed for research at the Institute of Microelectronics of Seville, Centro Nacional de Microelectrnica, Consejo Superior de Investigaciones CientficasUniversity of Seville. He has authored eight books; approximately 50 chapters in edited books, including original tutorials on chaotic integrated circuits, design of data converters, and design of chips for vision; and some 500 articles in peer-reviewed specialized publications. His research work is widely quoted, and he has an h-index of 35. His current research interests are in the areas of imagers and vision systems using 3-D integration technologies and of ultra low-power medical electronic devices. Prof. Rodríguez-Vázquez has served and is currently serving as an Editor, an Associate Editor, and a Guest Editor for different IEEE and non-IEEE journals. He is in the committee of many international journals and conferences and has chaired different international IEEE and Society of Photo-Optical Instrumentation Engineers conferences. He has received a number of international awards for his researchwork (IEEE GuilleminCauer Best PaperAward, two Best Paper awards from Wileys International Journal of Circuit Theory and Applications, IEEE European Conference on Circuit Theory and Design Best Paper Award, and IEEE International Symposium on Circuits and Systems Best Demo-Paper Award).