Low-power focal-plane dynamic texture segmentation based on programmable image binning and diffusion hardware

Jorge Fernández-Berni and Ricardo Carmona-Galán

Institute of Microelectronics of Seville (IMSE-CNM)
Consejo Superior de Investigaciones Científicas y Universidad de Sevilla
Parque Tecnológico de la Cartuja. Calle Américo Vespucio s/n, 41092, Seville, Spain

ABSTRACT

Stand-alone applications of vision are severely constrained by their limited power budget. This is one of the main reasons why vision has not yet been widely incorporated into wireless sensor networks. For them, image processing should be suscribed to the sensor node in order to reduce network traffic and its associated power consumption. In this scenario, operating the conventional acquisition-digitization-processing chain is unfeasible under tight power limitations. A bio-inspired scheme can be followed to meet the timing requirements while maintaining a low power consumption. In our approach, part of the low-level image processing is conveyed to the focal-plane thus speeding up system operation. Moreover, if a moderate accuracy is permissible, signal processing is realized in the analog domain, resulting in a highly efficient implementation. In this paper we propose a circuit to realize dynamic texture segmentation based on focal-plane spatial bandpass filtering of image subdivisions. By the appropriate binning, we introduce some constrains into the spatial extent of the targeted texture. By running time-controlled linear diffusion within each bin, a specific band of spatial frequencies can be highlighted. Measuring the average energy of the components in that band at each image bin the presence of a targeted texture can be detected and quantified. The resulting low-resolution representation of the scene can be then employed to track the texture along an image flow. An application specific chip, based on this analysis, is being developed for natural spaces monitoring by means of a network of low-power vision systems.

Keywords: Wireless sensor networks, bio-inspired processing architecture, dynamic textures, VLSI implementation, diffusion

1. INTRODUCTION

Wireless Sensor Networks (WSN)[1] represent a clear example of the advances reached in communications and electronic. In these networks, tiny autonomous sensors are capable of capturing information from their surroundings, processing this information and communicating the results if necessary. Data obtained by sensors normally comes from scalar measurements like temperature, pressure or humidity. This permits meeting strict power budgets while employing a conventional serial processing scheme. Complex computations are not usually necessary and the amount of information to be either processed in-situ or broadcasted is small. But recently, a step further has been proposed: wireless multimedia sensor networks.[2] This implies the incorporation of vision capabilities into the network nodes. At this point a crucial problem arises: image processing means a great deal of computation to be realized over a quite significant amount of raw data. In these conditions, pure digital processing schemes could only achieve real-time performance at the expense of a very high power consumption, especially high when compared to the limited resources of the sensor nodes.

In this paper we present an approach based on a bio-inspired processing scheme which addresses this problem. By carefully designing focal-plane analog hardware performing low-level tasks[3] over the images it is possible to obtain a reduced representation of such images at very low energy cost. A digital processor would carry out medium- and high-level tasks over this reduced representation in order to achieve the final result. This processing

Further author information:
Jorge Fernández-Berni: E-mail: berni@imse.cnm.es, Telephone: +34 954 46 66 66
Ricardo Carmona-Galán: E-mail: rcarmona@imse.cnm.es, Telephone: +34 954 46 66 66

Bioengineered and Bioinspired Systems IV, edited by Ángel B. Rodríguez-Vázquez, Ricardo A. Carmona-Galán,Gustavo Liñán-Cembrano, Proc. of SPIE, Vol. 7365, 736502
 · doi: 10.1117/12.822590

chain is similar to that of biological vision sensors, where the retina plays the focal-plane hardware role, pre-processing the scene just captured. The outcome is a still retinotopic but simplified —in terms of the number of data— and elaborated —in terms of the nature of the data— version of this scene which is sent to the visual cortex for further understanding.[4] It has been demonstrated that a physical implementation based on this approach[5] can reach a high performance with less cost and power than their digital counterparts. More specifically, we apply the previous approach to the segmentation of dynamic textures. A temporal or dynamic texture (DT) is a spatially-repetitive time-varying visual pattern whose temporal variation presents certain stationarity.[6] An additional feature of a DT is its indeterminate spatial and temporal extent. Smoke colums, wave patterns, swaying trees or a flock of birds are some examples of dynamic textures. The interest of implementing these detection capabilities in WSNs comes from two facts. First, one of the most suitable applications for WSNs is environmental monitoring.[1] Second, DTs are very common in natural scenes.[7] In order to carry out segmentation of DTs, we make use of binning and filtering at the focal-plane. This processing can be realized by an arbitrarily gated resistive grid. The resistors of this grid are implemented by MOS transistors biased in the ohmic region. The control of their gate voltage permits to set the size of the bins according to the scale to be surveyed for detecting the targeted textures. Within every bin, the pixel values are diffused during a certain —controlled— period of time. This extracts relevant information at the spatial frequency bands associated to the frequency signature of the DT. Finally, once the diffusion is stopped at the required time instant, the total energy of every bin is calculated. A simplified representation of the original scene is built, by means of which temporal evolution of the aimed textures can be tracked.

2. BIO-INSPIRED PROCESSING SCHEME

Existing research on DT recognition makes use of complex computations based normally on the optic flow in order to extract motion features.[8] This allows the classification of sequences containing different categories of textures. However, optic flow computation implies heavy processing load or, in other words, high energy consumption. We propose a different approach based on a bio-inspired processing architecture (Fig. 1). By means of focal-plane processing —concurrent and therefore fully-parallel connected to the sensor— a scene composed of $M \times N$ pixels is transformed into a reduced representation composed of $m \times n$ values (with $M \gg m$ and $N \gg n$). It is this reduced representation which is then converted to the digital domain in order to be post-processed. This alleviates the intrinsic bottleneck associated to the serialization as well as the processing load of the digital processor.

At this point the key element is the focal-plane processing that permits texture segmentation. The spatial repeatability of dynamic textures is exploited in the form of a distinctive frequency signature for each texture. Thus we propose to divide the scene into portions, or bins, of size $W \times H$ pixels. Then, we will represent each bin by a single value correlated to the chance of containing textures close to the targeted ones. The original amount of information will be reduced $W \times H$ times, resulting in a total number of bins:

$$m \times n = \frac{M \times N}{W \times H} \tag{1}$$

The size of the bins will be determined not only by the spatial frequency signature but also by the scale in which we look for it. It must also allow to track the texture across the scene. As an example, in Fig. 2 a binning process has been applied to a scene containing a flock of birds. The value of each bin is calculated as:

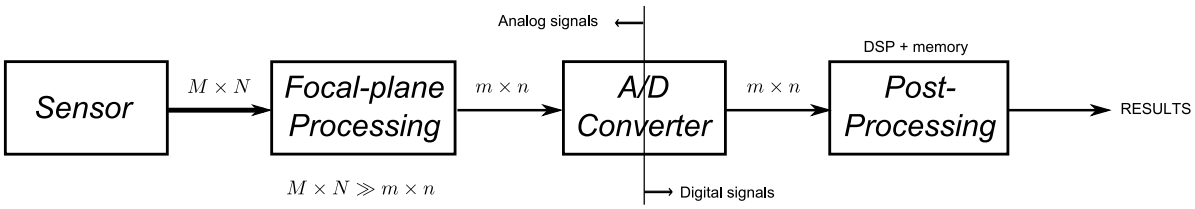

Figure 1. Processing scheme proposed

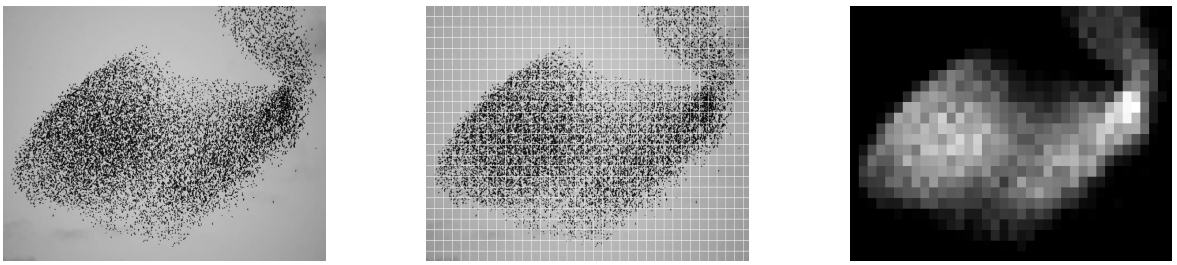

Figure 2. Binning and filtering applied to a scene containing a flock of birds

$$B_{kl} = \frac{\sum_{\forall \mathbf{k} > \mathbf{0}} E_{kl}(\mathbf{k})}{\sum_{\forall \mathbf{k}} E_{kl}(\mathbf{k})} \tag{2}$$

where $\mathbf{k}^T = (u, v)^T$ represents the wave number and $E_{kl}(\mathbf{k})$ the energy associated to the frequency component $\mathbf{k}$ within the bin B_{kl}. As can be seen, even the density of birds within the flock can be observed without performing a pixel-level analysis.

The next step is to design power-efficient analog hardware capable of realizing both pixel binning and spatial filtering. This hardware must extract information about any selected band of frequencies, defined within each bin. This will permit searching for specific frequency signatures by region. A family of spatial filters extensively employed in literature because of their suitable properties to develop a multiscale representation of the image is the Gaussian lowpass filters. They are closely related to diffusion and a feasible implementation with MOS-based resistances, which in turns will allow to control the size of the bins, will be proposed.

3. SPATIAL FILTERING WITH LINEAR RESISTIVE GRIDS

The impulsive response of a Gaussian filter is given by:[9]

$$h(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{3}$$

where σ is the standard deviation of a Gaussian distribution centered at the origin. It expresses the spreading of the original image information towards the neighbouring pixels. In the Fourier space, this function is represented by:

$$H(k_x, k_y) = e^{-2\pi^2\sigma^2(k_x^2+k_y^2)} \tag{4}$$

Let us consider now the diffusion equation:

$$\frac{\partial V}{\partial t}(x, y, t) = D\nabla^2 V(x, y, t) \tag{5}$$

whose spatial Fourier transform is:

$$\frac{\partial \hat{V}}{\partial t}(k_x, k_y, t) = -4\pi^2 D(k_x^2 + k_y^2)\hat{V}(k_x, k_y, t) \tag{6}$$

and solving:

$$\hat{V}(k_x, k_y, t) = \hat{V}(k_x, k_y, 0)e^{-4\pi^2 Dt(k_x^2+k_y^2)} \tag{7}$$

Therefore, diffusion during a period of time denoted by t is equivalent to the application of a Gaussian filter whose spread is controlled by this t through:

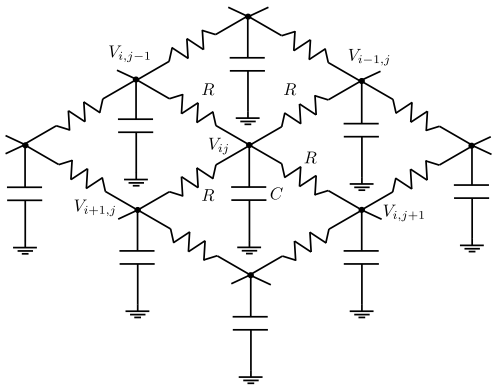

Figure 3. Resistive grid with time-controlled smoothing factor

$$\sigma = \sqrt{2Dt} \tag{8}$$

thus, the longer the diffusion time the larger the spread of the original image information towards its vicinity. If we substract the result of diffusing the original function, $V(x, y, 0)$, during two time periods t_1 and t_2, the outcome is equivalent to apply a DoG bandpass filter.[10] Therefore, time-controlled diffusion permits to evaluate the magnitude of different spatial frequency components. In order to implement this diffusion, we are going to make use of a resistive grid. Resistor networks have been a useful tool in many branches of science and technology since long ago.[11]

Consider the resistive grid of Fig. 3, where the smoothing factor is determined by the amount of time that the initial node voltages are allowed to diffuse, thus differing from static implementations of resistive grid filtering.[12] If we permit the network to evolve from the initial state, the equation satisfied at each node is:

$$\tau \frac{dV_{ij}}{dt} = -4V_{ij} + V_{i+1,j} + V_{i-1,j} + V_{i,j+1} + V_{i,j-1} \tag{9}$$

where $\tau = RC$, and the indexes i and j are employed to denote the position of the pixel in the now discrete grid in which diffusion takes place. Applying the DFT to this equation we obtain:

$$\tau \frac{d\hat{V}_{uv}}{dt} = -4\hat{V}_{uv} + e^{\frac{2\pi iu}{M}}\hat{V}_{uv} + e^{\frac{-2\pi iu}{M}}\hat{V}_{uv} + e^{\frac{2\pi iv}{N}}\hat{V}_{uv} + e^{\frac{-2\pi iv}{N}}\hat{V}_{uv} \tag{10}$$

where we have considered that the set of pixels being processed is of size $M \times N$. Eq.(10) can be rewritten as:

$$\tau \frac{d\hat{V}_{uv}}{dt} = [-4 + 2cos(\frac{2\pi u}{M}) + 2cos(\frac{2\pi v}{N})]\hat{V}_{uv} \tag{11}$$

and now solving in the time domain we obtain:

$$\hat{V}_{uv}(t) = \hat{V}_{uv}(0)e^{\frac{2t}{\tau}[cos(\frac{2\pi u}{M})+cos(\frac{2\pi v}{N})-2]} \tag{12}$$

where $\hat{V}_{uv}(0)$ represents the DFT of the image defined by the initial voltages at the capacitors and $\hat{V}_{uv}(t)$ is the DFT of the image defined by those same node voltages after a certain time interval t since the network started to evolve. A transfer function can be defined as follows:

$$\hat{H}_{uv}(t) = \frac{\hat{V}_{uv}(t)}{\hat{V}_{uv}(0)} = e^{\frac{2t}{\tau}[cos(\frac{2\pi u}{M})+cos(\frac{2\pi v}{N})-2]} \tag{13}$$

what reflects an approximation of a Gaussian spatial filtering where t determines the cut-off frequency. Let us represent the transfer function defined by Eq. (13) over the corresponding discrete Fourier plane for different

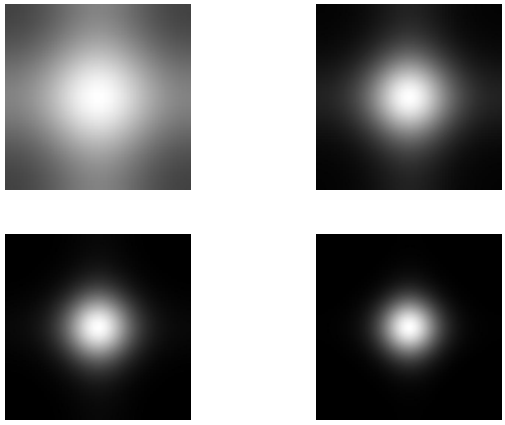

Figure 4. Transfer function of the resistive grid for increasing values of t

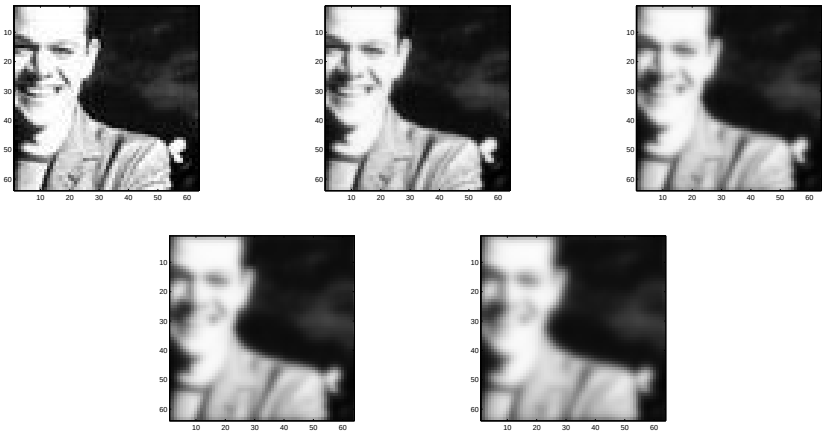

Figure 5. Original image and outcome of applying to it the transfer functions in Fig. 4, respectively

values of t. For the sake of simplicity, we are going to consider square images. The result is depicted in Fig. 4. Take into account that the center of the planes corresponds to the wave number $(u, v)^T = (0, 0)^T$ where the transfer function always equals its maximum value for any value of t, that is, $\hat{H}_{00}(t) = 1$. The outcome of applying the transfer functions in Fig. 4 is shown in Fig. 5. It can be seen that the operation realized by the grid is ideally an isotropic lowpass filtering whose bandwidth is determined by t. The longer the interval t the narrower the bandwidth. A certain distortion appears as higher frequencies are to be filtered. It is due to the 4-connected interconnection pattern among neighbor pixels which constrains the filtering of the highest spatial frequency to only four spatial directions.[13] The smaller the frequencies considered the more the possible spatial directions involved in the filtering and therefore the more its isotropic shape.

One of the most interesting aspects of the transfer function just described is that information about any band of spatial frequencies with approximately the same norm can be easily extracted from the difference between two filtered images. In Fig. 6 the normalized difference between the adjacent transfer functions in Fig. 4 are depicted. Note that the bandwidth of the resulting bandpass filters depends on the time intervals considered. The ideally isotropic shape of the filtering performed by the proposed resistive grid has an enormous importance when it comes to segment dynamic textures. It implies that any spatial frequency signature defined can be detected whatever its spatial orientation be.

Finally, there is an additional issue to be remarked about the physics behind the circuit realizing the filtering

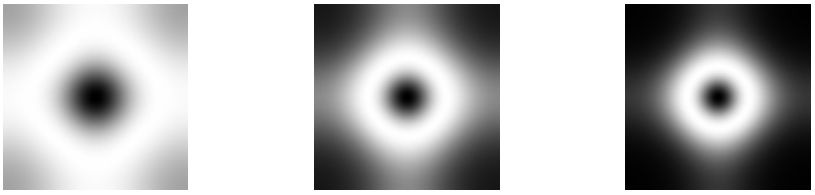

Figure 6. Normalized difference between the adjacent transfer functions in Fig. 4

just explained. It is composed of capacitors connected by resistors where a simple charge diffusion process which does not require further energy insertion is carried out. That is to say, the signal processing within the circuit is massively parallel and ultra power-efficient.

4. PHYSICAL IMPLEMENTATION OF IMAGE BINNING AND FILTERING

We have demonstrated that a simple resistive grid meets two essential features for the proposed processing scheme: power-efficient operation and flexibility for extracting information about different bands of spatial frequencies. To achieve a VLSI implementation of programmable image binning based on such a circuit we propose the hardware structure depicted in Fig. 7. It consists of a $M \times N$ grid where the value of each pixel is stored in a capacitor. Each capacitor is 4-connected to the neighboring capacitors by means of MOS transistors. Both the capacitors and the transistors are nominally identical throughout the grid. The gate voltage of each transistor is controlled by the corresponding row or column selection signal. When selected, i. e. the control signal is high, the MOS are biased in the ohmic region, behaving as resistors connecting two nodes. If the control signal is low the MOS transistors are off, establishing the boundary of a bin. Thus the particular distribution of 0's and 1's in the set of row and column selection signals determines the size and amount of bins in which the image plane is divided. For instance, bins with a size of 2×2 pixels would be established by the distribution in Fig. 7. Finally, once the image plane division is defined, a charge diffusion process like that of the resistive grid described in the previous section is performed within the bin whenever the corresponding control signals remain high. By switching to low these signals, the diffusion is stopped.

As a measure of the accuracy and robustness reached by a grid implemented by MOS transistors with respect to an ideal resistive grid, we have implemented in HSPICE an NMOS grid. The models belong to a standard 0.35μm CMOS process. The main parameters employed are nominally $W = 1\mu$m, $L = 4\mu$m, $\mu_0 = 3.41 \cdot 10^{-2}\text{m}^2/\text{Vs}$ and $t_{ox} = 7$nm. Independent deviations following gaussian distributions with $\sigma = 10\%$ of μ_0 and t_{ox} from their nominal values are introduced in a 64×64 grid. Initializing the grid with the original image in Fig. 5 and permitting the network to evolve, the RMSE of the voltages at the capacitors with respect to an ideal resistive grid is always less than 0.75%, that is, an accuracy better than 7 bits. To visualize this accuracy, we have represented in Fig. 8 the output images for an ideal resistive grid and for the NMOS grid implemented at the time instant in which the RMSE in the NMOS grid reaches its maximum. As can be seen, the outcome is perceptually equivalent. We have also represented their difference normalized by a maximum observed error between pixels of 1.88%.

5. COMPUTING THE ENERGY OF THE BINS

Notwithstanding the above mentioned, the hardware proposed in Fig. 7 does not still match the focal-plane processing defined in Fig. 1. It is necessary to represent every bin by only one value once the difussion has been stopped. This is the way a reduced representation of the scene is built. Let $V_{ij}(t)$ be the voltages at the $M \times N$ grid at time instant t. That is, the raw image after diffusion for t seconds. For a discrete-time signal, the total energy is defined as the sum of the squared amplitude of the samples. In the case of a discrete-space signal, i. e. the image of size $M \times N$ pixels, the energy of the complete image at any given time instant is given by the sum:

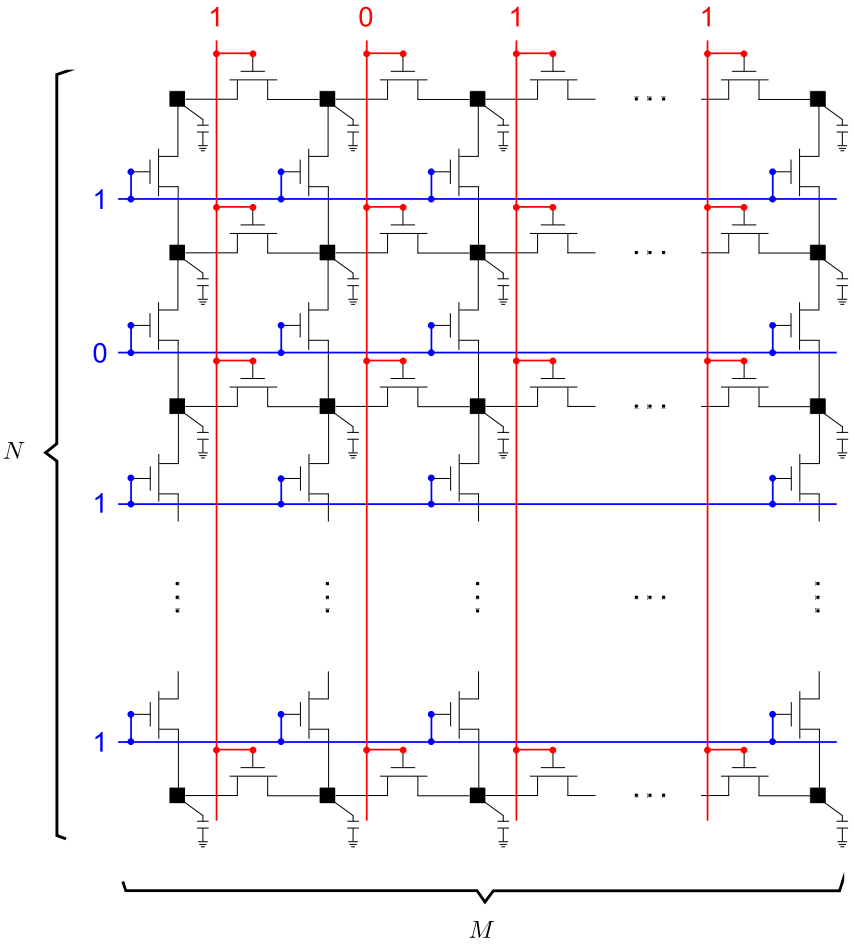

Figure 7. Hardware structure for programmable image binning and filtering

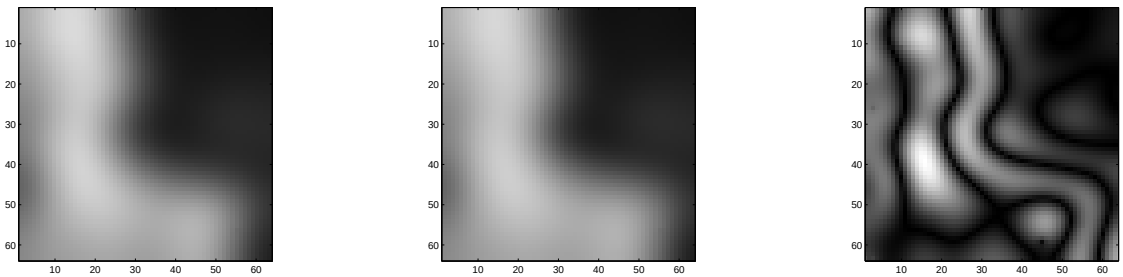

Figure 8. Outcome of an ideal resistive grid and a grid composed of NMOS, respectively, at the time instant in which the RMSE in the NMOS grid reaches its maximum value with respect to the ideal resistive grid. Their difference normalized by a maximum observed error between pixels of 1.88% is also represented

$$E(t) = \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} |V_{ij}(t)|^2 \tag{14}$$

what matches the value obtained by adding up the energies corresponding to the different components of the spatial DFT of the image:

$$E(t) = \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} |\hat{V}_{uv}(t)|^2 \tag{15}$$

Eqs. (14) and (15) mean that the energy of the image accounts for the filtering undergone during the diffusion. The total charge in the whole capacitor array is conserved, but, naturally, the system evolves towards the less energetic configuration. Thus the energy at each time instant is a measure of the evolution of the diffusion process. The longer t the less $E(t)$. The energy lost between two consecutive points in time during the difussion corresponds to that of the spatial frequencies filtered. Notice that changing the reference level for the amplitude of the pixels does not have an effect beyond the dc component of the image spectrum. A constant value added to every pixel does not eliminate nor modify any of the spatial frequency components already present, apart from that at the origin of the Fourier space.

In order to analyze the presence of different spatial frequency components within a particular bin of the image, we would need to measure the energy of the bin pixels once filtered. Remember that for analyzing a particular band of frequencies we will substract two lowpass filtered versions of the image. In this way, only the components of the targeted frequency band will remain. This will allow to track changes at that band without pixel-level analysis. The hardware employed to calculate the energy of the bins at the pixel-level is very simple. It consists of a MOS grid like that of Fig. 7 linked pixel to pixel with the MOS grid performing diffusion. For this new grid, all the capacitors must be pre-charged to a reference voltage V_{REF}. The link between both grids is realized by means of the circuit depicted in Fig. 9 where C_P represents the capacitor storing the pixel value and C_E the corresponding capacitor pre-charged to V_{REF}. Once the diffusion has been stopped after a certain time interval t, the switch S_E in all the pixels is switched ON during a fixed period of time T_E. It leads to a final voltage at C_E:

$$V_{E_{ij}} = V_{REF} - \frac{T_E}{C_E}\beta[V_{ij}(t) - V_{th}]^2 \tag{16}$$

We are assuming that all the transistors M_E, operating in saturation, are nominally identical. Deviations occur from pixel to pixel due to mismatch in the threshold voltage (V_{th}), the transconductance parameter (β), and the body-effect constant (γ, not in this equation). Being area dependent effects, transistors M_E are tailored to control the resulting error in the computation. Also, mobility degradation contributes to the deviation from the behaviour depicted in Eq. (16). The useful signal range will be limited by this. When S_E is switched back to OFF in all the pixels after the period of time T_E, charge redistribution takes place in the grid, with the same binning scheme as the diffusion. It results in averaging the pixel energy within each bin:

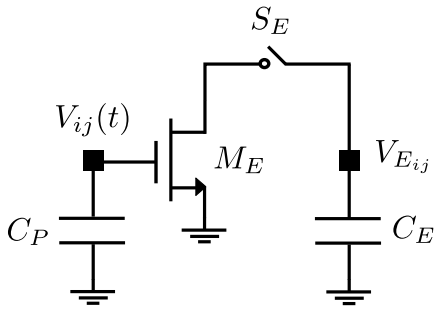

Figure 9. Circuit linking pixel to pixel the two MOS grids

$$\bar{V}_{E_{kl}} = V_{REF} - \frac{\beta T_E}{WHC_E} \sum_{i=kW}^{kW+W-1} \sum_{j=lH}^{lH+H-1} [V_{ij}(t) - V_{th}]^2 \quad (17)$$

where indexes k and l identify the bin. This voltage is proportional to the total energy of the pixels of that bin t seconds after the diffusion started. As referred before, the offset introduced by V_{th} does not affect any spatial frequency other than the dc component. Finally, in order to achieve a reduced representation of the scene, only one pixel out of every bin needs to be read as all the capacitors within the bin will be at the same voltage $\bar{V}_{E_{kl}}$.

6. CONCLUSIONS

A new generic approach to segment dynamic textures has been presented. It is based on a bio-inspired processing architecture suitable for real-time vision applications where the power requirements demand very low energy consumption. At the circuit level, we make intensive use of simple resistive grids. It permits the design of massively parallel and ultra power-efficient analog hardware flexible enough to achieve programmable image binning and filtering and to extract information about different frequency bands at the focal plane. Finally, a reduced representation of the scene is obtained by computing the total energy of the pixels within the bins established. This alleviates the workload of the digital post-processing in order to reduce as much as possible the power consumption while meeting the timing requirements of real-time vision applications

ACKNOWLEDGMENTS

This work is partially funded by the Andalusian regional government (Junta de Andalucía-CICE) through project 2006-TIC-2352 and by the Spanish Ministry of Science (MICINN) through project TEC 2006-15722.

REFERENCES

[1] Akyildiz, I., Su, W., Sankarasubramaniam, Y., and Cayirci, E., "A survey on sensor networks," *IEEE Communications Magazine* **40**(8), 102–114 (2002).

[2] Akyildiz, I., Melodia, T., and Chowdhury, K., "A survey on wireless multimedia sensor networks," *Computer Networks* **51**(4), 921–960 (2007).

[3] Pirsch, P. and Stolberg, H., "VLSI implementations of image and video multimedia processing systems," *IEEE Transactions on Circuits and Systems for Video Technology* **8**(7), 878–891 (1998).

[4] Roska, B. and Werblin, F., "Vertical interactions across ten parallel, stacked representations in the mammalian retina," *Nature* **410**, 583–587 (2001).

[5] Liñán Cembrano, G., Rodríguez-Vázquez, A., Carmona-Galán, R., Jímenez-Garrido, F., S., E., and Domínguez-Castro, R., "A 1000 fps at 128x128 vision processor with 8-bit digitized i/o," *IEEE Journal of Solid-State Circuits* **39**(7), 1044–1055 (2004).

[6] Nelson, R. and Polana, R., "Qualitative recognition of motion using temporal texture," *CVGIP: Image Understanding* **56**(1), 78–89 (1992).

[7] Péteri, R., Huskies, M., and Fazekas, S., "Dyntex: A comprehensive database of dynamic textures." http://www.cwi.nl/projects/dyntex/.

[8] Chetverikov, D. and Péteri, R., "A brief survey of dynamic texture description and recognition," in [*International Conference on Computer Recognition Systems (CORES'05)*], 17–26 (2005).

[9] Jahne, B., Hauβecker, H., and Geiβler, P., [*Handbook of Computer Vision and Applications*], vol. 2, ch. 4, Academic Press (1999).

[10] Wilson, H. and Giese, S., "Threshold visibility of frequency gradient patterns," *Vision Research* **17**(10), 1177–1190 (1977).

[11] Liebmann, G., "Solution of partial differential equations with a resistance network analogue," *British Journal of Applied Physics* **1**(4), 92–103 (1950).

[12] Shi, B. and Chua, L., "Resistive grid image filtering: input/output analysis via the CNN framework," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications* **39**, 531–548 (Jul 1992).

[13] Rau, R. and McClellan, J., "Efficient approximation of gaussian filters," *IEEE Transactions on Signal Processing* **45**, 468–471 (Feb 1997).