A CNN-DRIVEN LOCALLY ADAPTIVE CMOS IMAGE SENSOR

R. Carmona, C. M. Domínguez-Matas, J. Cuadri, F. Jiménez-Garrido and A. Rodríguez-Vázquez

Instituto de Microelectrónica de Sevilla-CNM-CSIC
Campus de la Universidad de Sevilla. Avda. Reina Mercedes s/n,
41012-Sevilla, Spain. E-mail: rcarmona@imse.cnm.es

ABSTRACT

A bioinspired model for mixed-signal array mimics the way in which images are processed in the visual pathway. Focal-plane processing of images permits local adaptation of photoreceptor structures in silicon. Beyond simple resistive grid filtering, nonlinear and anisotropic diffusion can be programmed in this CNN chip. This paper presents the local circuitry for sensors adaptation based on the mixed-signal VLSI parallel processing infrastructure in CMOS.

1. INTRODUCTION

The retina operates on the captured visual stimuli at early stages in the process of vision. Complex spatio-temporal processing encodes visual information into a reduced set of channels [1] to be delivered to the brain by the optic nerve. This model inspires a feasible alternative to conventional digital image processing. We are interested , in particular, in local monitoring and control of the photosensing devices for contrast enhancement. This capability improves the perceived sensation by extracting the reflectance information from the acquired luminance matrix [2]. This task is gracefully implemented in the biological retina. Concurrent sensing and massively parallel processing provides enough computing power to realize these tasks in mixed-signal VLSI. This paper presents a method for locally adapting integrating photosensors exposure time to the illumination conditions that can be implemented in an already functioning CNN chip architecture based on the mammalian retina.

2. BIOINSPIRED CNN PROCESSOR

Visual stimuli trigger the formation of patterns of activation in the retina. These patterns are processed as they advance towards the optic nerve. Contrarily to the spike-like coding of neural information found elsewhere, they are continuous-time analog waves [3]. The biological motivation is the lack of bandwidth offered by the spike-like neural impulses to handle the data contained in the visual stimulus. The captured signals are promediated and the high-gain characteristics of the cones and the bipolar cells are shifted to adapt to light conditions. These operations have a local scope and depend on the recent history of the cells. Once adaptation is achieved, patterns of activity are formed dynamically.

A CNN model has been developed that approximates the observed behaviour of different parts of the retina [4]. This model has been implemented in a 2-layer CNN chip, designed and fabricated in a standard 0.5μm CMOS technology [5]. It contains a central array of $32 \times 32 \times 2$ processing nodes arranged in a 2-layer structure. The evolution of each node, $C(i, j, k)$, in layer k, is described by:

$$\tau_k \frac{dx_{n_{ijk}}}{dt} = -g[x_{ijk}(t)] + b_{00k}u_{ijk} + z_{ijk} + \\ + \sum_{l=-r_1}^{r_1} \sum_{m=-r_1}^{r_1} \sum_{l=-r_1}^{r_1} a_{lmn} f[x_{i+l,j+m,k+n}(t)] \qquad (1)$$

where $f()$ and $g()$ are nonlinear functions of the state variable, $x_{ijk}(t)$. Each layer incorporates intra- and interlayer feedback and feedforward connections, a_{lmn} and b_{lmn}, a bias term z_{ijk}, and its own time constant τ_k. Programming different dynamics in this CNN model is possible by adjusting these parameters. Different reaction-diffusion equations can be mapped into this architecture, resulting in propagative and wave-like phenomena, similar to those found at the biological retina. Fig. 1 shows an effect observed in the outer plexiform layer (OPL) of the retina and programmed in the chip —the detection of spatio-temporal edges followed by de-activation of the patterns—, sampled at different points in the evolution of the network dynamics [5].

Apart from the two different programmable CNN nodes, the basic processing element contains local analog and logic memories, for the storage of intermediate results, and a programmable local logic unit. The basic processing element occupies $188\mu\text{m} \times 186\mu\text{m}$, resulting in a cell density of 29.24cells/mm^2 (not considering the control circuit overhead). The power consumption of the whole chip has been estimated in 300mW. The fastest time constant is designed to be under 100ns. The chip can handle analog data with an

Partially funded by ONR Project N-000140210884, CE Project IST-1999-19007 (DICTAM) and the Spanish MCyT Project TIC1999-0826.

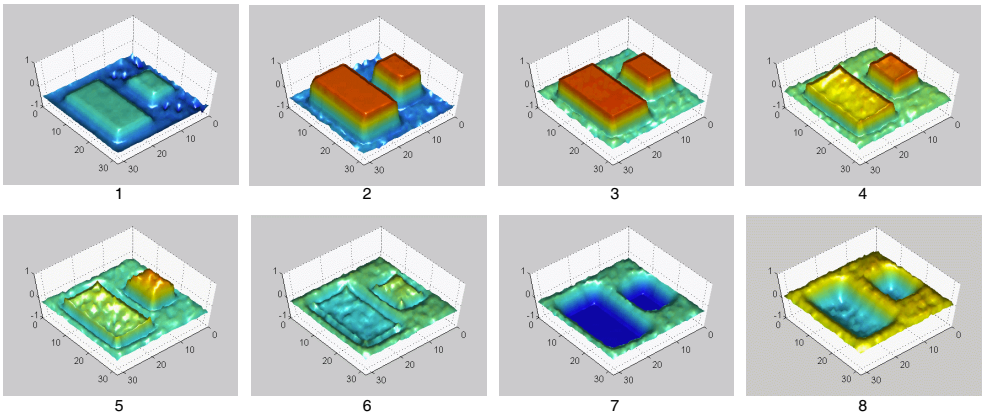

Fig. 1. Spatio-temporal edge detection and de-activation (fast layer).

equivalent resolution of 7.5bits (measured). In future versions we plan to incorporate adaptive control of the sensors based on focal-plane processing feedback.

3. ADAPTIVE OPTO-ELECTRONIC INTERFACE

Capturing light in CMOS technology counts on the separation of photogenerated electron-hole pairs by effect of an electric field, normally provided by a reverse-biased pn junction. Through a linear resistive load, the photocurrent, I_{ph}, generates an instantaneous voltage, proportional to the incident light power. Noise can be filtered out by integrating this current in the diode's parasitic capacitor C_p. Thus, for a given integration time t_{int}, there is a linear relation between the average incident light power density over the sensor, through I_{ph}, and the pixel voltage V_{ph}:

$$V_{ph}(t_{\rm int}) = V_{\rm ref2} - \frac{I_{ph}}{C_p} t_{\rm int} \qquad (2)$$

While CMOS photosensors have a maximum dynamic range of 5 to 7 decades, light intensity on natural scenes can vary over up to 14 decades. Thus, linear sensors usually produce images with over-exposed and under-exposed regions. To accommodate a larger light intensity range within the photodiode dynamic range, voltage compression is required.

The perceptual quality of an image is closely related with the ability of separating the irradiance $E(x,y)$, and the reflectance $\rho(x,y)$ —where most of the relevant information is—, both contained in the luminance signal $I(x,y)$. This task is continuously performed in the retina. As a consequence of Weber's law, the perception gain is inversely proportional to the local average of brightness, $\tilde{I}(x,y)$. If $E(x,y)$ is the main responsible of $\tilde{I}(x,y)$, then [2]:

$$\rho(x,y) \propto \frac{I(x,y)}{\tilde{I}(x,y)} \qquad (3)$$

Based on the intra- and inter-frame correlation found in natural scenes, the local average voltage of a previous frame, $\tilde{V}_{ph}(n-1)$, can be employed to control the integration time of the next frame, $t_{\rm int}(n)$, so as to establish a shorter integration time for strongly illuminated areas and a larger one for the areas lying in the dark. The required local information is provided by the CNN core circuitry in tens of nanoseconds. Fig. 2 depicts the schematic of the local adaptation circuit. Photogenerated currents are integrated in the sense capacitor. $M_{\rm reset}$ works as the electronic

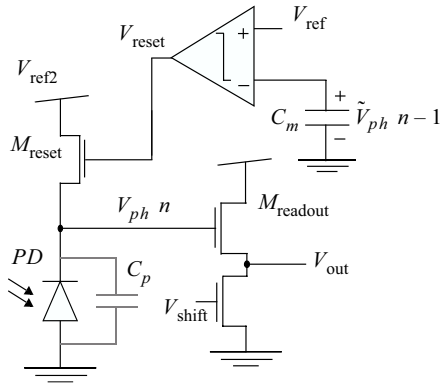

Fig. 2. In-pixel adaptation circuit.

shutter. While the local stored average value, $\tilde{V}_{ph}(n-1)$, remains below a global reference inverted ramp delivered to every pixel, $V_{\rm ref}$, the photosensor voltage is shorted to $V_{\rm ref2}$. When the ramp crosses the value of the stored average, the photodiode starts discharging C_p. The larger $\tilde{V}_{ph}(n-1)$, the sooner the inverse ramp crosses it, and the longer the integration time. Thus, image areas with a stronger average illumination, will have a shorter exposure time, while image areas lying in the dark will have a longer $t_{\rm int}(n)$, following:

$$t_{\rm int}(n) = t_{\rm min} + \frac{t_{\rm max} - t_{\rm min}}{V_{\rm ref2} - V_{\rm min}} \tilde{V}_{ph}(n-1) \qquad (4)$$

where $t_{\max}$ and $t_{\min}$ are the maximum and minimum integration times for each frame. Before the ramp restarts, V_{out} is sampled and $\tilde{V}_{ph}(n)$ is computed by the CNN array.

A high dynamic range image has been built with 11 snapshots of the same scene taken at different integration times. Fig. 4(a) shows a 256-level representation of this HDR image. Information is truncated by the limited DR. If it is re-captured with a simulated pixel array with local adaptation to the average illumination, some information is recovered (Fig. 4(b)). This result can be compared with logarithmic compression (Fig. 4(c)). The average brightness employed is obtained by linear diffusion, but it can be the result of a more involved computation: anisotropic diffusion, diffusion with controlled contours, etc [6].

Stability concerns arise if $\Delta t_{\text{int}} = t_{\max} - t_{\min}$ grows, leading to non-convergence of the series defined by Eq. 4. If the first term, $t_{\min}$ is made dependent on V_{ph}, this series converges for a larger range of Δt_{int}. The global adaptation circuit in Fig. 5 provides an average integration time that is a function of the total image average brightness $\overline{V}_{ph}(n-1)$. Then, if the difference between the maximum and the minimum is fixed to Δt_{int}, Eq. 4 can be rewritten:

$$t_{\text{int}}(n) = \overline{t}_{\text{int}}(n-1) - \frac{\Delta t_{\text{int}}}{2}[1 - \frac{2\tilde{V}_{ph}(n-1)}{V_{\text{ref2}} - V_{\min}}] \quad (5)$$

This is accomplished by a circuit providing a ramp signal whose period is proportional to $\overline{V}_{ph}(n-1)$, i. e. inversely proportional to the history of the average illumination. Fig. 6 shows several frames of an artificially distorted sequence. After a few frames, $\overline{t}_{\text{int}}(n)$ converges and a locally adapted capture takes place. Observing the evolution of $\overline{t}_{\text{int}}(n)$, it dinamically adapts to the average brightness, increasing when dark elementes enter the scene and decreasing if the image gains in brightness. When the image remains still $\overline{t}_{\text{int}}(n)$ converges to an optimum. Fig. 3 displays how tracking a fixed gray level evolves to a stable integration time, and to the same medium gray in all cases.

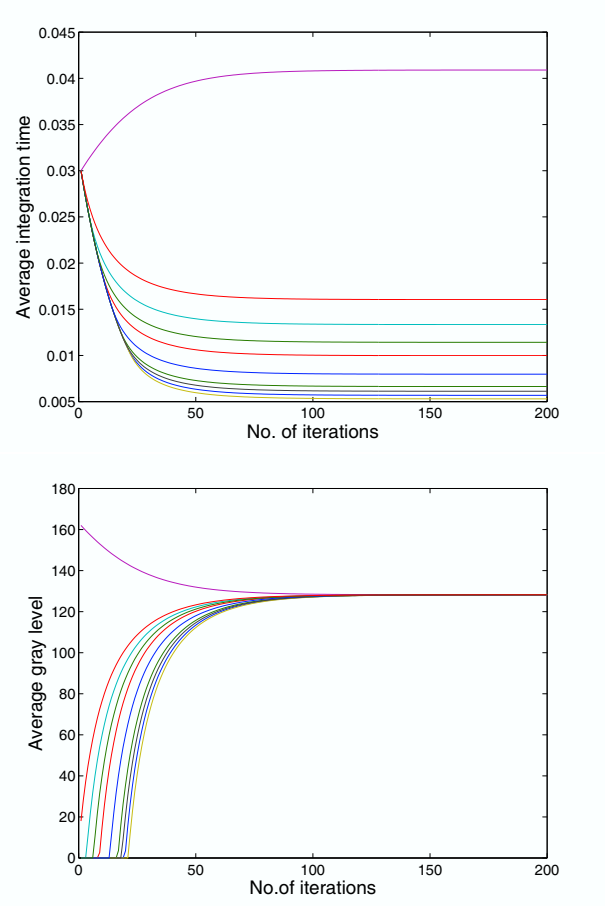

Fig. 3. Simulated capture of a plain gray picture.

4. CONCLUSIONS

A simple but precise model of the real biological retina renders a feasible efficient implementation of an artificial vision device. The results of CNN processing can be employed to locally adapt the photosensors operation. The application of this adaptation method, based on the local average brightness of previous image frames, result in a perceptually enhanced image capture.

5. REFERENCES

[1] B. Roska and F.S. Werblin, "Vertical interactions across ten parallel, stacked representations in the mammalian retina," *Nature*, vol. 410, pp. 583–587, Mar. 2001.

[2] V. Brajovic, "A model for reflectance perception in vision," in *Bioengineered and Bioinspired Systems, Proceedings of SPIE*, May 2003, vol. 5119, pp. 307–315.

[3] F. Werblin, "Synaptic connections, receptive fields and patterns of activity in the tiger salamander retina," *Investigative Ophthalmology and Visual Science*, vol. 32, no. 3, pp. 459–483, Mar. 1991.

[4] F. Werblin; T. Roska and L. O. Chua, "The analogic cellular neural network as a bionic eye," *International Journal of Circuit Theory and Applications*, vol. 23, no. 6, pp. 541–569, Nov. 1995.

[5] R. Carmona et al., "A bio-inspired two-layer mixed-signal flexible programmable chip for early vision," *IEEE Transactions on Neural Networks*, vol. 14, no. 5, pp. 1313–1336, Sept. 2003.

[6] Cs. Rekeczky; T. Roska and A. Ushida, "Cnn-based difference-controlled adaptive nonlinear image filters," *International Journal of Circuit Theory and Applications*, vol. 26, pp. 375–423, July 1998.

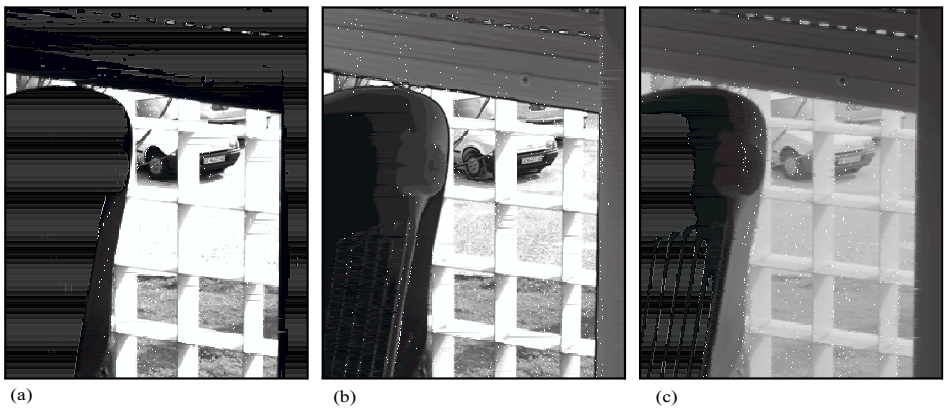

Fig. 4. Simulated capture of a still picture with local adaptation (center).

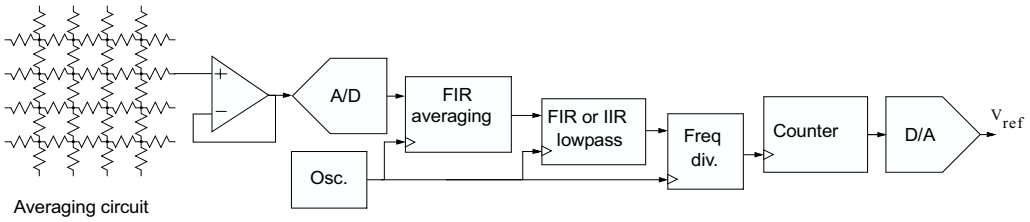

Fig. 5. Conceptual schematic of the global adaptation circuit.

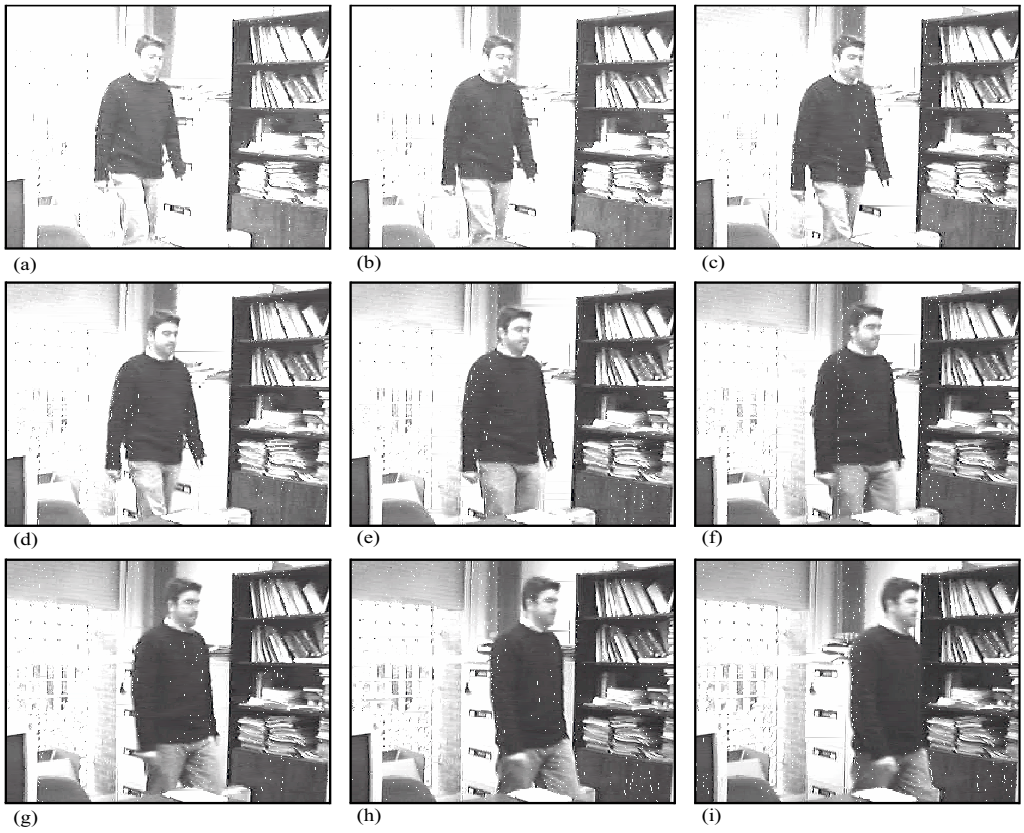

Fig. 6. Simulated adaptation in a sequence taken at 25fps, 1 out of each 3 frames shown.