Focal-Plane CMOS Wavelet Feature Extraction for Real-Time Pattern Recognition

Ashkan Olyaei and Roman Genov

Intelligent Sensory Microsystems Laboratory, University of Toronto
10 King's College Road, Toronto, ON M5S 3G4 Canada

ABSTRACT

Kernel-based pattern recognition paradigms such as support vector machines (SVM) require computationally intensive feature extraction methods for high-performance real-time object detection in video. The CMOS sensory parallel processor architecture presented here computes delta-sigma ($\Delta\Sigma$)-modulated Haar wavelet transform on the focal plane in real time. The active pixel array is integrated with a bank of column-parallel first-order incremental oversampling analog-to-digital converters (ADCs). Each ADC performs distributed spatial focal-plane sampling and concurrent weighted average quantization. The architecture is benchmarked in SVM face detection on the MIT CBCL data set. At 90% detection rate, first-level Haar wavelet feature extraction yields a 7.9% reduction in the number of false positives when compared to classification with no feature extraction. The architecture yields 1.4 GMACS simulated computational throughput at SVGA imager resolution at 8-bit output depth.

Keywords: Image sensor, focal-plane processing, pattern recognition, Haar wavelet, mixed-signal VLSI

1. INTRODUCTION

Real-time pattern recognition in video faces two essential challenges. One is the large degree of disparity between objects within a class due to variations in images such as scale, orientation, and illumination variance. The other is the computational complexity of high-performance classification algorithms, such as support vector machines (SVM). For the latter, the Kerneltron, a massively parallel 1 GMACS/mW VLSI array processor for kernel-based pattern recognition has been recently developed by us.[1] Feature extraction methods resolve the first challenge. Invariant features of objects within a class are extracted, improving the generalization ability of the classifier, and often reducing the requirements on the size of the training set and memory.

Haar wavelet transform is a popular feature extraction method. It yields information about spatial gradient of image intensity, not just its absolute value. Extracting Haar wavelet features from streaming video is very computationally intensive, requiring vast computing resources. CMOS integrated imaging technologies allow to perform low-cost, low-power signal processing directly on the focal plane, eliminating the need for an external processor.[2] The intrinsic parallelism of such computing yields throughput and energy efficiency well above those of modern digital processors, allowing to extract complex features in real time.

Various active pixel spatial filtering techniques suitable for on-focal-plane analog VLSI implementation of Haar wavelet transform have been developed. Switched capacitor implementations use charge sharing to compute average sum and difference but require matched capacitors and have limited scalability.[3,4] Current-mode implementations,[5] including those with weighted averaging,[6] use zero-latency current-mode addition but deploy multiple matched current mirrors at the expense of increased pixel area. Current integration and gain-stage voltage summation[7] utilized in variable resolution imaging do not allow for weighted averaging and require additional column-parallel amplifiers. All of the aforementioned architectures perform computing in analog VLSI domain and require an extra analog-to-digital converter (ADC) to provide the digital output.

We present a mixed-signal VLSI architecture of a digital CMOS imager and real-time Haar wavelet feature extractor with focal-plane computing being an intrinsic part of the quantization process. Our approach combines weighted spatial averaging and oversampling quantization in a single $\Delta\Sigma$-modulated analog-to-digital conversion cycle. The approach yields

Further author information: (Send correspondence to R.G.)
A.O.: E-mail: olyaei@eecg.toronto.edu
R.G.: E-mail: roman@eecg.toronto.edu

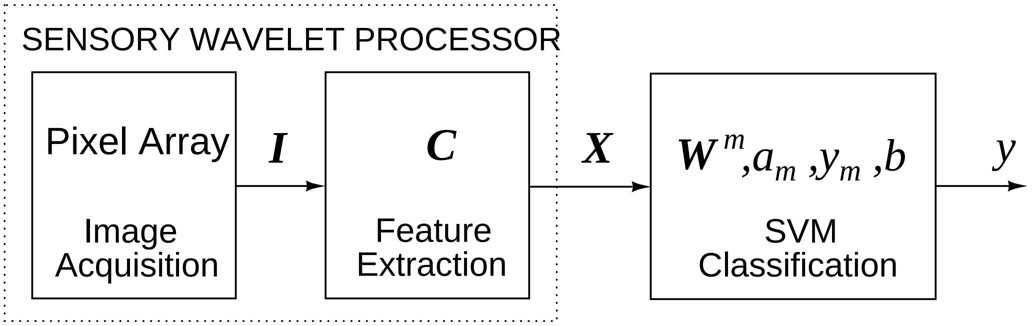

Figure 1. Functional block diagram for a sensory pattern recognition system. The sensory wavelet processor consists of the imager pixel array and the feature extractor.

no overhead in time and negligent overhead in area compared to a baseline raw image readout and oversampling quantization architecture. The architecture yields 1.4 GMACS throughput with SVGA imager resolution at 8-bit output depth. The rest of this paper is organized as follows. Section 2 gives an overview of a real-time visual pattern recognition microsystem combining Haar wavelet feature extraction and support vector machine (SVM) data classification. Section 3 presents the architecture and circuits of the sensory Haar wavelet processor. Section 4 presents SVM classification performance results validating functionality of the sensory processor architecture.

2. REAL-TIME VISUAL PATTERN RECOGNITION MICROSYSTEMS

The block diagram of the real-time visual pattern recognition microsystem combining a sensory feature extractor and a data classifier is shown in Fig. 1. The next sections present an overview of the feature extraction and data classification functional blocks.

2.1. Haar Wavelet Feature Extraction

To enhance the performance of a classifier, set of features generalizing across variations in visual objects is extracted from the input. Both training and classification are then performed on these extracted features. Haar wavelet transform enhances classification performance of high-performance kernel-based learning paradigms such as support vector machines (SVM),[8, 9] and is simple enough to be implemented in a mixed-signal VLSI technology.

By extracting horizontal, vertical and diagonal edges, Haar wavelets register the relationship between intensities among neighboring pixels in different orientations and hence form a "ratio template".[8] The ratio template is independent of illumination conditions. It truly captures the ratio between various features of an object, which within a class exhibit greater correlation than the absolute intensity values. Ratio templates effectively reduce the variability of objects within a class. As a result, a more precise object description is generated at a cost of lower spatial resolution. This enables the classifier to detect complex objects such as faces or pedestrians at higher detection rates with lower false positive.

The one-dimensional Haar wavelet is composed of the scaling function ϕ(t), or the father wavelet, and the wavelet prototype function ψ(t) also known as the mother wavelet:

$$\phi(t) = \begin{cases} 1 & \text{if -1} \leq \text{t} \leq 1 \\ 0 & \text{otherwise,} \end{cases}$$

$$\psi(t) = \begin{cases} 1 & \text{if } 0 < \text{t} \leq 1 \\ -1 & \text{if -1} \leq \text{t} \leq 0 \\ 0 & \text{otherwise.} \end{cases}$$

The combination of scaling and wavelet prototype functions in a two-dimensional space yields the following scalar, horizontal, vertical, and diagonal two-dimensional Haar wavelet functions:

$$\begin{aligned}\phi(x,y) &= \frac{1}{4}\phi(x)\phi(y),\\ \psi^H(x,y) &= \frac{1}{4}\psi(x)\phi(y),\\ \psi^V(x,y) &= \frac{1}{4}\phi(x)\psi(y),\\ \psi^D(x,y) &= \frac{1}{4}\psi(x)\psi(y),\end{aligned}$$

where the $\frac{1}{4}$ coefficient is applied so that the maximum value of the wavelet transform stays within the image intensity range. The equivalent spatial kernels of the scalar and wavelet functions for the first-level Haar wavelet are:

$$\begin{aligned}\mathbf{\Phi_1} &= \frac{1}{4}\begin{pmatrix} +1 & +1 \\ +1 & +1 \end{pmatrix},\\ \mathbf{\Psi_1^H} &= \frac{1}{4}\begin{pmatrix} +1 & -1 \\ +1 & -1 \end{pmatrix},\\ \mathbf{\Psi_1^V} &= \frac{1}{4}\begin{pmatrix} +1 & +1 \\ -1 & -1 \end{pmatrix},\\ \mathbf{\Psi_1^D} &= \frac{1}{4}\begin{pmatrix} +1 & -1 \\ -1 & +1 \end{pmatrix}.\end{aligned}$$

The transformation of the image fragment **I** into the feature **X** is of the linear form:

$$X_{ij} = \sum_{h=1}^{K}\sum_{v=1}^{K} C_{hv}\, I_{xy}, \tag{1}$$

$$x = h + (i-1)K, \quad i = 1, 2, \ldots, \frac{H}{K}, \tag{2}$$

$$y = v + (j-1)K, \quad j = 1, 2, \ldots, \frac{V}{K}, \tag{3}$$

$$K = 2^l, \quad l = 1, \ldots, L, \tag{4}$$

where $C_{hv} \in \{-1, +1\}$ are the Haar block matrix coefficients comprising a square spatial kernel; *H* and *V* are the image fragment horizontal and vertical sizes, multiples of the kernel size K; *h* and *v* are the horizontal and vertical Haar kernel indices; i and j are the indices of the Haar transformed image; L is the number of levels of Haar transform, and

$$\mathbf{C} \in \{\mathbf{\Phi_1}, \mathbf{\Psi_1^H}, \mathbf{\Psi_1^V}, \mathbf{\Psi_1^D}\}. \tag{5}$$

L-level Haar wavelet features for $L > 1$ can be obtained by repetitive use of level-one transform, or directly using $3L + 1$ spatially averaging Haar wavelet coefficient kernels of size $K = 2^l$, with $l = 1, \ldots, L$.

2.2. Support Vector Machine Classification

Support vector machine (SVM) is a kernel-based learning paradigm for data classification and function approximation rooted in Vapnik's statistical learning theory.[10] SVMs outperform many of the existing pattern recognition methods in complex tasks, even with sparse data. In the simple case of two-class separable data classification, support vector machine maps the wavelet transformed input data into a higher dimensional space where the two classes are linearly separable and classifies the data by constructing an optimal hyperplane that separates two classes. The unique separating hyperplane is determined by maximizing its distance to the closest points of the training set called support vectors. Support vectors typically comprise a small subset of the training set. Hence, only a subset of the training set is needed to construct the separating hyperplane.

In its basic form, an SVM classifies a two-dimensional feature input $\mathbf{X}$ into class $y \in \{-1, +1\}$ based on the support vectors $\mathbf{W}^m$ and their corresponding classes y_m as

$$y = \mathrm{sign}(\sum_{m=1}^{M} a_m \, y_m \, K(\mathbf{W}^m, \mathbf{X}) - b), \tag{6}$$

where $K(\cdot, \cdot)$ is a symmetric positive-definite kernel function chosen subject to fairly mild constraints.[10] The coefficients a_m and b are calculated in training by solving a linearly constrained quadratic programming (QP) problem, $a_{\,m}$ being zero for all the training set points with the exception of support vectors.[10]

Kernel function $K(\cdot, \cdot)$ maps the input space into a higher dimensional space to realize a non-linear separating hyperplane for data classification. Inner-product based kernels perform well in classification and are simple to implement in hardware:

$$K(\mathbf{W}^m, \mathbf{X}) = f(\sum_{n=1}^{3L+1} \sum_{i=1}^{\frac{H}{K_n}} \sum_{j=1}^{\frac{V}{K_n}} W_{ij}^{m,n} X_{ij}^{n}), \tag{7}$$

where $f(\cdot)$ is a monotonically non-decreasing scalar function subject to the Mercer condition on $K(\cdot, \cdot)$, such as polynomial or sigmoid.[10]

The most computationally intensive part of the SVM decision rule (6) is inner-product calculation of the wavelet-transformed input and the support vectors, which is equivalent to vector-matrix multiplication (VMM). While an energy-efficient, compact, and real-time implementation of VMM was reported by us in[1], in what follows we present a high-throughput mixed-signal VLSI architecture of Haar feature extraction on the focal plane, as the necessary pre-processing step.

3. SENSORY VLSI WAVELET PROCESSOR

Image acquisition and wavelet feature extraction are both performed on the focal plane of the sensory processor. Its mixed-signal VLSI architecture is depicted in Fig. 2.

3.1. Analog Image Acquisition

Analog image acquisition is performed by the active pixel array, the row control and the correlated double sampling (CDS) units. The active pixel comprises a resetable n^+-diffusion–p-substrate photodiode and a selectable source follower with shared column-parallel current source biased with *IbiasCol* current as shown in Fig. 3. The row control unit generates the digital signals *Reset* and *RowSelect* controlling the integration and readout phases. Fixed pattern noise (FPN) due to pixel-to-pixel transistor mismatch is one of the dominant noises in CMOS imagers. Switched-capacitor CDS unit suppresses FPN by subtracting the reset pixel output from the sensed pixel output.

3.2. On-Focal-Plane Haar Wavelet Feature Extraction

As detailed in Section 2.1, L-level Haar wavelet transform requires correlating an image with $3L + 1$ square spatial kernels of sizes $K = 2^l$, $l = 1, \ldots, L$. For $L = 1$, the Haar kernels are of the form (5). The presented mixed-signal VLSI architecture computes multiple kernel-image correlations in parallel within a single oversampling quantization cycle, with negligent space overhead and no time overhead as compared to a standard digital imager employing column-parallel oversampling quantizers.

The active pixel array is integrated with a bank of column-parallel oversampling analog-to-digital converters (ADCs). Each ADC is configured to sample K adjacent pixels in a column within a single quantization cycle. Uniform pixel sampling yields a digital representation of a column-wise spatial average of the acquired image segment. Correlation of a K-pixel long fragment of an image column with a signed unity-amplitude one-dimensional kernel of size K is computed by combining the uniform pixel sampling and pixel output sign inversion. A double sampling sign unit circuit determines the sign of each pixel output by selecting a switched-capacitor sampling sequence order. Sign values are presented bit-serially

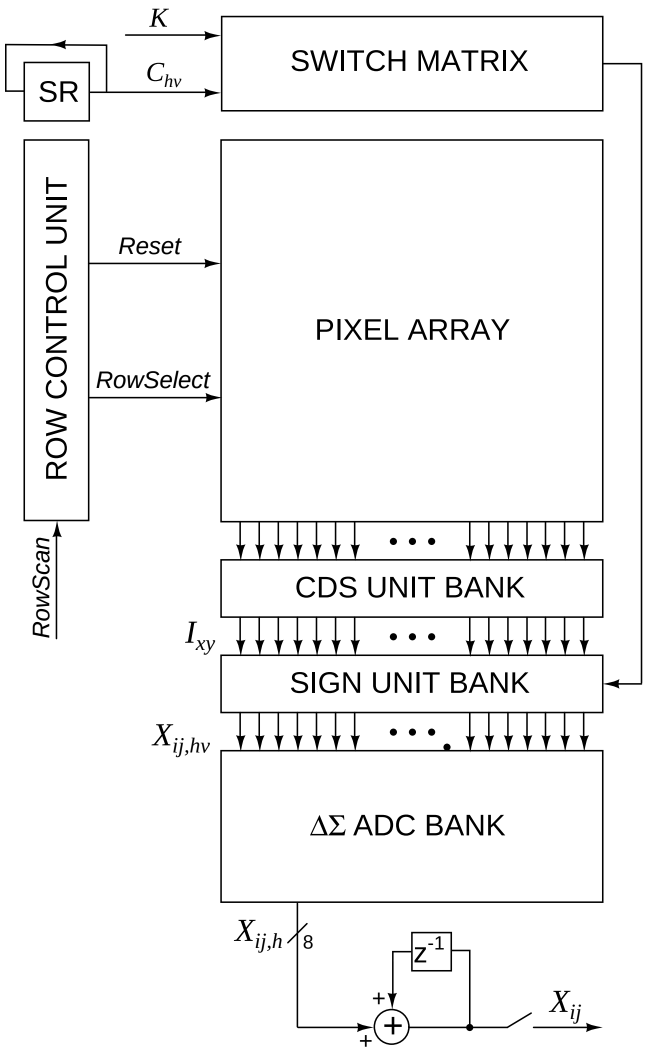

Figure 2. Sensory wavelet processor top-level architecture.

from a ring shift register, *SR*, with a sequence period of K values, synchronously with image read-out clock *RowScan*, as shown in Fig. 2. This amounts to computing one-dimensional column-parallel spatial Haar wavelet transform:

$$X_{ij,h} = \sum_{v=1}^{K} X_{ij,hv} = \sum_{v=1}^{K} C_{hv}\, I_{xy}, \tag{8}$$

where the notation is consistent with that of (2)–(4).

The oversampling quantizer is implemented as a first-order incremental $\Delta\Sigma$-modulated ADC. The implementation is compact as the decimating filter is a simple digital counter. The block diagram of the ADC is depicted in Fig. 4.

The switch matrix routes the K different time-dependent *Sign* signals to H/K groups of adjacent column-parallel sign unit circuits. A simple digital delay and adder loop performs spatial accumulation over K adjacent ADC outputs as they are read out:

$$X_{ij} = \sum_{h=1}^{K} X_{ij,h} = \sum_{h=1}^{K}\sum_{v=1}^{K} C_{hv}\, I_{xy}. \tag{9}$$

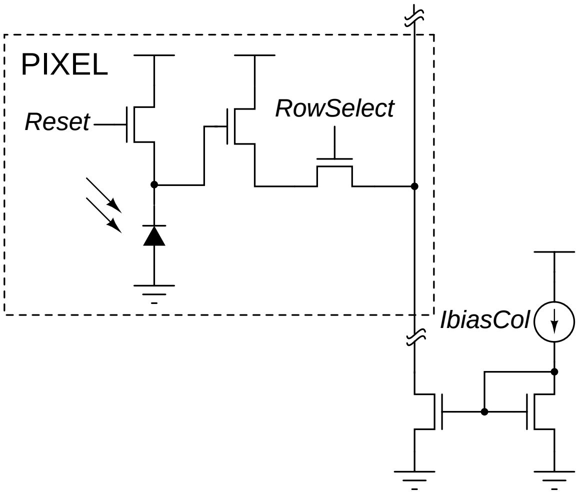

Figure 3. Photodiode active pixel circuit.

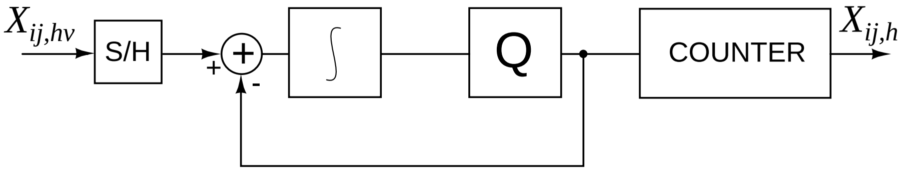

Figure 4. Block diagram of a first-order $\Delta\Sigma$ incremental A/D converter with comparator, sample-and-hold circuit, integrator and counter.

This mixed-signal VLSI computation realizes a two-dimensional Haar wavelet transform in (1) of up to L levels. The area overhead of sign unit circuits, switch matrix and digital accumulator scales linearly with the imager size and is negligent. As computing is interleaved with quantization no extra computational time is required above that of raw image oversampling quantization in a baseline digital imager.

4. RESULTS

We have validated the functionality of the sensory wavelet processor with support vector machine (SVM) data classification in a face detection task. Two cases are compared: no feature extraction and first-level Haar transform computed on the described architecture.

The SVM classifier with second degree polynomial kernel was trained on 6977 images (2429 faces and 4548 non-faces) and tested on 24045 images (472 faces and 23573 non-faces) from the MIT CBCL data set. Images are 16×16 pixels in size, 8-bit grayscale and histogram-equalized.

The solid curve in Fig. 5 illustrates the receiver operating characteristic (ROC) of the SVM classifier with no feature extraction, while the x-marked ROC curve corresponds to the SVM classifier with one-level Haar wavelet feature extraction. At 90% detection rate, the false positive rate decreases by 7.9%, *i.e.,* over 1800 non-faces are no longer misclassified. Additionally, Haar wavelet feature extraction decreases the number of support vectors from 410 to 322, reducing the memory requirements and computational complexity of the classifier by 21%.

The photodiode active pixel is simulated to have the worst case integration time of 10 msec. Assuming no pipelining, at 30 frames per second frame rate, additional 23 msec are available for Haar wavelet image processing. Based on a conservative quantizer sampling rate of 40 ksps, the computational throughput is estimated to be 1.4 GMACS for SVGA image resolution. The estimate is given for the three-level Haar wavelet transform computed on the presented architecture employing $\Delta\Sigma$-modulated algorithmic ADC.[11]

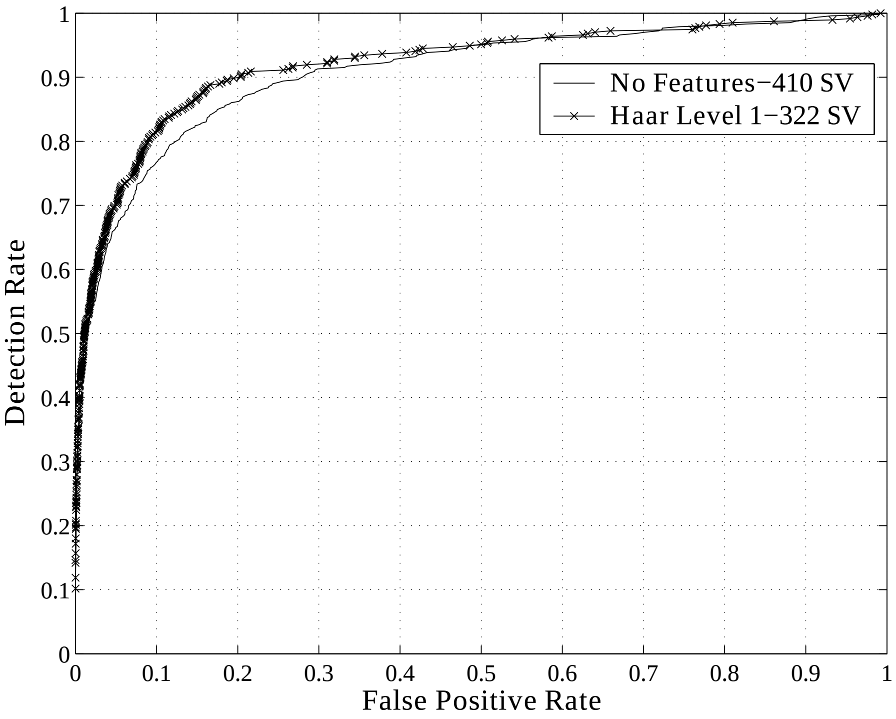

Figure 5. Receiver operating characteristic (ROC) curves for SVM polynomial classification with and without Haar wavelet feature extraction. The training and testing data are MIT CBCL face data sets.

5. CONCLUSION

We present a mixed-signal VLSI architecture of a digital CMOS imager computing two-dimensional Haar wavelet transform on the focal plane in real time. The approach combines weighted spatial averaging and oversampling quantization in a single $\Delta\Sigma$-modulated analog-to-digital conversion cycle, making focal-plane computing an intrinsic part of the quantization process. The approach yields no overhead in time and negligent overhead in area compared to a baseline digital imager employing oversampling quantization. Haar wavelet feature extraction architecture enhances classification performance in pattern recognition, yielding 1.4 GMACS simulated computational throughput at SVGA imager resolution at 8-bit output depth.

ACKNOWLEDGMENTS

This work was supported by Natural Sciences and Engineering Research Council (NSERC), Communications and Information Technology Ontario (CITO), and Canada Foundation for Innovation (CFI).

REFERENCES

1. R. Genov, and G. Cauwenberghs, “Kerneltron: support vector ”machine” in silicon,” *IEEE T. Neural Networks,* vol. **14** (5), 2003.
2. E.R. Fossum, “CMOS image sensor: Electronic camera-on-a-chip,” *IEEE Trans. Electron Devices*, vol. **44**, pp 1689-1698, 1997.
3. S.E. Kemeny, R. Panicacci, B. Pain, L. Matthies, and E.R. Fossum, “Multiresolution image sensor,” *IEEE Trans. Circuits and Systems for Video Technology,* vol. **7** (4), pp 575-583, 1997.
4. Q. Luo, and J.G. Harris, “A novel integration of on-sensor wavelet compression for a CMOS imager,” IEEE Int. Symp. on Circuits and Systems (ISCAS’02), Scottsdale, AZ, May 26-29, 2002. ISCAS’2002.
5. A. Fish, and O. Yadid-Pecht, “CMOS current/voltage mode winner-take-all circuit with spatial filtering,” IEEE Int. Symp. on Circuits and Systems (ISCAS’01), vol. **3**, pp 636-639, 2001.
6. V. Gruev, and R. Etienne-Cummings, “Implementation of steerable spatiotemporal image filters on the focal plane, ” *IEEE T. Circuits and Systems II,* vol. **49** (4), 2002.

7. Z. Zhou, B. Pain, and E.R. Fossum, “Frame-transfer CMOS active pixel sensor with pixel binning,” *IEEE Trans. Electron Devices*, vol. **44** (10), pp 1764-1768, 1997.
8. M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio, “Pedestrian Detection Using Wavelet Templates,” in *Computer Vision and Pattern Recognition*, pp 193-199, 1997.
9. C. Papageorgiou, M. Oren, and T. Poggio, “A General Framework for Object Detection,” in *Proceedings of International Conference on Computer Vision*, 1998.
10. V. Vapnik, *The Nature of Statistical Learning Theory,* 2nd ed., Springer-Verlag, 1999.
11. G. Mulliken, F. Adil, G. Cauwenberghs, and R. Genov, “Delta-Sigma Algorithmic Analog-to-Digital Conversion,” IEEE Int. Symp. on Circuits and Systems (ISCAS’02), Scottsdale, AZ, May 26-29, 2002. ISCAS’2002.