Visual sensor with resolution enhancement by mechanical vibrations

Oliver Landolt* and Ania Mitros
California Institute of Technology, M/S 139-74, Pasadena CA 91125
landolt@klab.caltech.edu, ania@klab.caltech.edu
http://www.klab.caltech.edu/~ania/research/

Abstract

The resolution limit of visual sensors due to finite pixel spacing can be overcome by applying continuous low-amplitude vibrations to the image – or taking advantage of existing vibrations in the environment. Thereby, spatial intensity gradients turn into temporal intensity fluctuations which can be detected and processed by every pixel independently from the others. This approach enhances resolution and virtually eliminates fixed-pattern noise. A visual sensing microsystem taking advantage of this principle is described. It incorporates a custom analog integrated circuit implementing an array of 32 by 32 pixels with local temporal signal processing. Another key component is a resonant mechanical device producing low-amplitude image scanning movements powered by environmental vibrations.

1 Introduction

Image sensors can be classified in two broad categories on the basis of their purpose. Cameras are meant to acquire images for replication at another place or time for the benefit of human observers. Visual sensors are meant to extract information about a visual scene for purposes such as robot navigation. In the second category, it is preferable to incorporate visual data processing as early as possible in the signal flow to reduce the cost of transmitting and processing the tremendous amount of redundant raw image data delivered by an array of photoreceptors. Consistently with this requirement, a number of visual sensing integrated circuits incorporating some amount of processing within each pixel have been described in the literature, many of which are inspired to some degree by biological neural structures [1][2]. Adding substantial local processing into every pixel unavoidably leads to a steep increase in silicon area compared to the area devoted to photodetection, hence a reduction of the total number of pixels which can be integrated on a chip. The resulting loss in spatial sampling rate is a handicap of existing visual sensors with focal plane processing, in comparison to traditional approaches of machine vision combining a camera – with a fill factor close to 100% – with external processing hardware. Another issue plaguing visual sensors is fixed-pattern noise caused by random spatial fluctuations of device parameters within a pixel array. The level of pixel mismatch is frequently such that only strongly contrasted edges can be detected reliably, whereas dimmer image features are lost in fixed-pattern noise. Signal processing techniques capable of overcoming this problem tend to introduce undesirable side effects, such as temporal sampling or hardware overhead for offset storage.

In this paper, we introduce a new principle for the acquisition of visual information, which extends the effective resolution of a pixel array far beyond the limit imposed by pixel spacing. It is also inherently insensitive to fixed-pattern noise. Instead of measuring the distribution of light intensity at fixed locations, we propose to apply continuous small-amplitude oscillatory movements to the imaging system. As a result of such movements, spatial varia-

*New affiliation: Agilent Technologies, PO Box 10350, Palo Alto CA 94303, oliver_landolt@agilent.com

tions of light intensity in the image turn into temporal fluctuations of light intensity at every photoreceptor. For instance, sweeping a photoreceptor over a thin spatial feature – such as a cable in an outdoor scene – can produce a detectable impulse of photocurrent, even if this feature is much thinner than pixel spacing. The effective spatial resolution of the sensor is limited by the focusing optics and pixel temporal bandwidth. Knowing the pattern of movements applied to the system, local spatial features can be retrieved from the temporal waveform detected by each photoreceptor.

In the field of image sensing for restitution purposes, a number of devices have been proposed, which apply subpixel shifts to an image by optical means in order to enhance the intrinsic resolution of a camera. This procedure differs from our scheme in that a camera delivers frames of raw image data at discrete locations, while our sensor exploits temporal waveforms of light intensity resulting from continuous movements to extract spatial image features. A truly continuous two-pixel scanning visual system implemented with discrete components has been described [3]. More recently, the effect of scanning was verified using two linear arrays of p-i-n photodiodes on a millimeter scale mobile robot, and a 1-D microlens array was fabricated for the same purpose [4]. In addition, an elegant implementation of a 2-D pixel array exploiting vibrations has been described [5] and implemented in CMOS [6], which incorporates local feature processing based on correlation between the photoreceptor signal and a template waveform. A limitation of this scheme is that only one type of feature can be detected at a time. In this paper, we propose an implementation of a vibrating 2-D visual sensor which encodes temporal signal features in the timing of digital pulses transmitted off-chip in real time. Spiking patterns can be processed by external hardware to detect a possibly large number of features in parallel.

2 Resolution enhancement by scanning

Let us first consider the case of a 1-D image $I(x)$ of an unchanging visual scene projected onto the surface of a visual sensor. If this image is shifting at a velocity v over the sensor as a consequence of mechanical vibrations occurring at some point in the optical path, a single photoreceptor will detect a light intensity $I_{pix}(t) = I(x_0 + v \cdot t)$, where x_0 depends on the location of the photoreceptor on the sensor. The spatial distribution of light intensity within the image is transformed into a temporal signal. Assuming a constant scanning velocity v, the spectrum of the temporal signal is related to the spatial spectrum of the image by linear scaling of the frequency axis:

$$f_T = v \cdot f_S \tag{1}$$

where f_T designates temporal frequency whereas f_S designates spatial frequency in the image plane. If the photoreceptor has a temporal bandwidth of f_{Tmax}, the spatial cutoff frequency for a scanning pixel will be $f_{Smax} = f_{Tmax}/v$. The spatial bandwidth of a non-scanning image sensor is entirely dependent on the spacing Δx of its photoreceptors and equals $1/(2\Delta x)$. Thus, scanning can improve the spatial resolution provided that

$$\frac{f_{Tmax}}{v} > \frac{1}{2\Delta x} \tag{2}$$

In the case of a 2-D image subject to mechanical vibrations along both axes, each photoreceptor acquires visual information along a curvilinear scanning path determined by image movements. Continuous image data is collected along the scanning path with a resolution determined by the same analysis as in the 1-D case. One dimension of the image features to be detected must be larger than the gap between nearest segments of the curvilinear scanning path. These gaps are smaller or equal to the scanning amplitude, which is intended to be on the order of pixel spacing.

In the implementation described in the present paper, the visual sensor is designed to operate at a scanning frequency of 300Hz with a photoreceptor bandwidth of at least 10KHz. In the first version of the

image sensor, photoreceptor spacing is $68.5\mu m$. Using these parameters and the equations introduced earlier in this section, under the hypothesis of circular scanning with a diameter of Δx, it can be shown that the effective spatial resolution in the image plane along the scanning path is about $6.5\mu m$. In terms of viewing angle, the resolution would be about 0.08^{o} for a focal length of $4.5mm$. This resolution is close to the diffraction limit of our focusing optics.

3 Pixel-level signal processing

The continuous temporal waveforms delivered by the photoreceptors carry high-resolution visual information, but this information is not readily available in a suitable format for machine vision applications. Besides, it would be impractical to send out of the chip all continuous waveforms as provided by the photoreceptors. For these reasons, signal processing must be performed locally in every pixel for the purpose of detecting key features in the temporal waveform, and encoding them in a format compatible with off-chip communication and subsequent processing. In the first stage of signal processing (Fig. 1), the current delivered by the photoreceptor is applied to a logarithmic current-to-voltage converter. The same visual scene under different illumination levels would produce images differing only by a scaling factor in intensity. After logarithmic transform, the temporal waveforms produced by scanning would differ only in their DC component, which is ignored by subsequent processing stages. Therefore, the logarithmic operator contributes to make visual data delivered by the sensor independent of illumination level [1]. The signal resulting from logarithmic compression is differentiated with respect to time and half-wave rectified, whereby both the positive and the negative fraction are retained separately. Current signals delivered at both outputs of the rectifier are sent to independent non-leaky integrate-and-fire circuits, where the charge is accumulated over time until the resulting voltage reaches a threshold. At this point, the integrate-and-fire block emits a brief pulse ("spike"), resets its integrator and resumes operation. Spikes from the positive and negative integrate-and-fire blocks are the final output signals delivered by the pixel. They are sent off-chip by means of a so-called address-event communication bus tailored specifically to this application [7]. This communication scheme consists of briefly flashing the address of the firing pixel on the output bus whenever a spike occurs somewhere in the array. Spikes are brief enough that collisions due to simultaneous firing of multiple pixels are rare at an overall data rate of several million spikes per second [7]. The processing scheme described in the next section is tolerant to limited spike loss due to collisions.

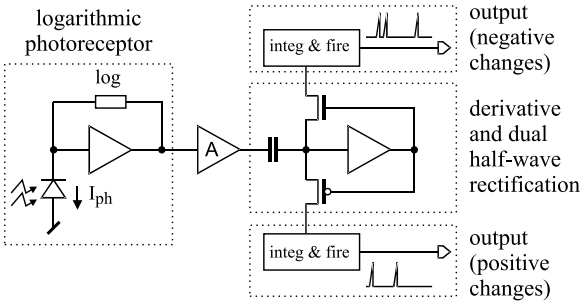

Figure 1: Functional block schematic of a single pixel.

4 Building image feature maps

The spike train emitted by every single pixel can be used in combination with the trajectory of the scanning device to identify spatial features present in the small image area scanned by this pixel. For conceptual simplicity, we first describe the case of a 1-D scanning motion. As the photoreceptor sweeps over an image pattern, charge is added to the integrators at a rate proportional to the image gradient. Thereby, if the same image segment is scanned multiple times, the firing threshold will be reached more frequently at locations of steeper gradients – although not always at the exact same locations in general. We have observed empirically that the probability of spiking at a specific point of an image is proportional to the intensity gradient at this location. Therefore, the image gradient is revealed by the histogram of average spiking frequency versus position. Such a

histogram can be computed by sampling the position of the scanning device every time a spike occurs, and incrementing the bin corresponding to this instant position. In this histogram, high peaks reveal sharp edges, whereas fainter gradients would result in lower spiking frequencies. Additional visual information can be retrieved by distinguishing spikes emitted by the positive or the negative integrator of the pixel. It is also useful to take the direction of scanning into account – instead of just position – in order to retrieve the sign of the gradient accurately.

This feature extraction approach can be extended to the case of 2-D scanning. The general idea is again to divide the region scanned by a pixel into smaller fields and count the number of spikes occurring in each field over a preset integration time. However, in the 2-D case, these fields can be given arbitrary shapes to detect spatial image features such as oriented line segments, edges or junctions. For instance, in order to detect and locate oriented edges, the visual field of a pixel can be divided into an array of parallel stripes. If an edge lies within a stripe, this particular field will receive a large number of spikes, whereas other fields will receive very few. As a matter of fact, each stripe constitutes the *receptive field* of a spatial feature detector. Collectively, the set of spiking frequencies received by all fields represent a spatial *feature map*, somewhat analogous to the maps found in the visual cortex of animals. The number and the nature of features is determined by the mapping between histogram bins and the instant position of the scanning device at the time a spike is emitted. Instead of just using the position, it can be useful to consider also the direction in which the scanning device is moving, because a gradient is directional by definition. If the location of an image feature does not matter as much as the nature of this feature, the direction of scanning can be used alone. It should be noted that a local feature map can be built for each pixel individually without ever combining information across different pixels. This property eliminates fixed-pattern noise problems which typically affect the computation of gradients or high-pass filtering in conventional approaches. It should also be noted that sensitivity and signal-to-noise ratio can be traded for bandwidth by adjusting the integration time in the computation of the histograms. The above signal processing principles apply to scanning paths of any shape. If a periodic scanning pattern is imposed, the feature type can be thought of as being encoded in the phase relationship of succeeding spikes. For example, for a repeated circular scan over a single feature, the shape of the spike histogram determines the feature type while the phase indicates its orientation.

Given that spikes are transmitted sequentially by the image sensing front-end described in Section 3 with a latency of only a few tens of nanoseconds, histograms can be computed by extremely simple and efficient hardware consisting mostly of an adder and a memory bank (Fig. 2). Each pixel has its own section of memory, the size of which depends on the number of different features which are detected. The position and/or direction of movement of the scanning device, which are measured by external sensors, are converted into the address of a memory cell by a look-up table which determines the shapes of the receptive fields. This table is shared by all pixels because the scanning movements are identical. Whenever a spike occurs, the content of the currently selected memory cell in the section belonging to the firing pixel is incremented by the adder. If overlapping receptive fields exist, then multiple adders operating in parallel on independent memory banks can be used. Alternatively, several memory cells can be successively incremented in response to a single incoming spike. The introduction of a first-in first-out (FIFO) buffer between the visual sensor and the external processing hardware as depicted in Figure 2 relaxes latency specifications of the adder and memory blocks. After a given integration time, the memory bank contains a spatial feature map which can be read out and used by a higher level visual processor. The simplicity of the hardware required to build features maps is a decisive advantage of the proposed visual sensor. With traditional approaches, spatial feature maps typically require convolutions between high-resolution raw image data and a number of different kernels corresponding to all features of interest. This operation is very computationally intensive and requires powerful digital signal processors to perform in real time.

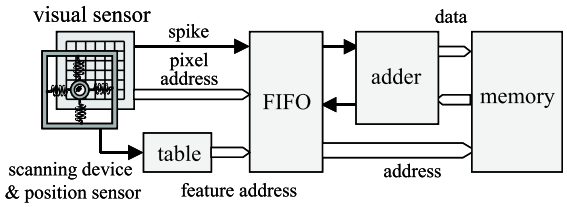

Figure 2: Block diagram of the envisioned hardware implementation of a feature map extractor using the proposed scanning visual sensor.

5 Integrated Circuit

5.1 Overview

We have designed a custom VLSI chip incorporating a 32 by 32 array of pixels implementing an analog signal processing chain as described in Section 3, together with a digital communication bus to transmit visual information outside the chip. The chip has been manufactured in a $0.6\mu m$, double poly, triple metal CMOS process. The chip is designed for a supply voltage of 3V. The pixel array occupies 2.2mm by 2.2mm and the entire chip area is about $10mm^2$. A single pixel circuit has an area of $68.5\mu m$ by $68.5\mu m$. The area of the photodiode is $10\mu m$ by $10\mu m$, which results in a fill factor of 2.1%. The remainder of the area is occupied by signal processing circuitry. Design details of this chip have been published in [7].

5.1.1 Measured pixel output

Each pixel is expected to fire spikes at a rate proportional to the derivative of its input signal (Section 4). To verify this property, we have recorded the timing of spikes generated by an individual test pixel in response to a sine wave current input. The spikes are histogrammed against the phase of the sine wave (Fig. 3). As expected, the envelope of the histogram of the spikes corresponds to the derivative of the input. Since the signal is half-wave rectified and each portion is fed to a separate integrate-and-fire circuit, only half the derivative waveform is portrayed. A small number of spikes occurs during the wrong phase of the signal. These spikes are due to intrinsic shot noise in the photodiode and subsequent transistors.

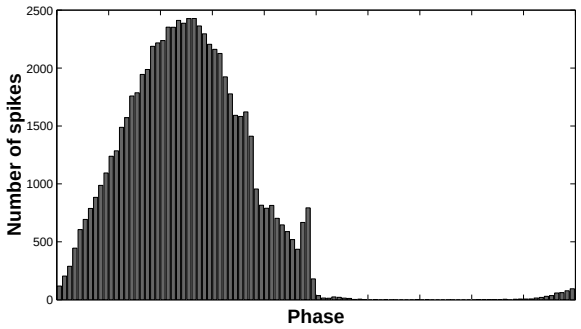

Figure 3: Histogram of spike output of the test pixel to multiple cycles of a sine wave input.

5.1.2 Photoreceptor noise

Intrinsic noise within the photoreceptor circuit limits the contrast range which the visual sensor is able to detect. Noise-induced fluctuations at the photoreceptor level are passed down the signal processing chain and ultimately turn into stochastic spike firing at some baseline rate. Signal-induced spikes superimpose to this baseline. At sufficient image contrast, the noise-induced firing rate should be much lower than the actual signal. Unfortunately, in the first version of the chip, we measured an unacceptably high baseline rate of noise-induced spikes, effectively preventing the acquisition of natural images in normal lighting conditions.

This flaw has been addressed in a redesign incorporating several techniques aiming at reducing the impact of intrinsic noise sources. The pixel bandwidth is limited as much as allowed by resolution enhancement constraints (10KHz). This filtering of the high-frequency noise is accomplished by a second-order filter built into the amplifiers following the photoreceptor. In addition, noise-induced spiking is reduced by introducing an amplitude threshold before rectification. Since this threshold also suppresses very low amplitude signals, it is carefully chosen to preserve a useful range of detectable image contrast (above 3%). Results from the second version of the chip are not yet available as of this writing.

5.1.3 Power consumption

Under a supply voltage of 3V, the chip consumes between $22\mu A$ and $26\mu A$ of DC current depending on illumination conditions. In addition, it consumes an amount of dynamic power roughly proportional to the data rate at the digital output bus. At a rate of 1M spikes/second, the dynamic consumption is about 1mA.

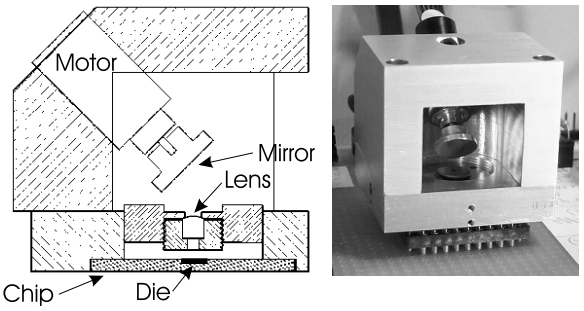

Figure 4: Drawing and photograph of mechanical device producing circular scanning.

6 Mechanical Design

The purpose of the mechanical component of our visual sensing microsystem is to keep the image focused onto the chip in steady small-amplitude motion. We have pursued two different approaches in parallel. The first device produces a well-controlled circular scanning path under the action of a motor. It consists of a tilted mirror spinning in front of the focusing lens (Fig. 4). The mirror is mounted on the shaft of a DC motor tilted at an angle of about 45° with respect to the optical axis of the lens. The mirror is not exactly perpendicular to the motor shaft, but tilted by a small angle ϵ of about 0.5°. Rotation of the motor causes the reflective surface to wobble, thereby causing the image to travel a circular path with a radius of 2ϵ in viewing angle. A position encoder on the motor axis indicates the orientation of the mirror at all times. The signal from this encoder serves as a reference for the interpretation of spiking patterns delivered by the visual sensor (Section 4). This device was easy to build and provides accurate control over the scanning path. Therefore, it is most appropriate for laboratory experiments.

For applications where space and power consumption are an issue, we have designed an alternative device where scanning is caused by displacements of the lens by mechanical vibrations. In this device, the lens is mounted on springs allowing lateral X-Y displacements but maintaining constant spacing between the lens and the chip. If the system is mounted onto a vibrating platform such as a vehicle driving on a rough surface, the mechanical energy available in the vicinity of the resonance frequency of the lens/spring system will cause scanning movements. The amplitude must be on the order of pixel spacing on the chip, e.g. a few tens of microns. The shape of the scanning path will depend on the relative magnitudes and phases of vibrations applied to the X and the Y axes, and on the relative resonance frequencies along these axes. In conditions where damping is low, the scanning path is a Lissajou figure, or elliptical if the resonance frequencies match. Conceivably, the relative resonance frequencies could be tuned to minimize gaps in the curvilinear scanning path (Section 2). As the scanning path will vary over time depending on environmental vibratory conditions, it is necessary to continuously monitor the position of the lens for interpreting the spiking patterns as described in Section 4. The lens position is monitored by capacitive measurements between the lens socket and surrounding fixed electrodes. This measurement has been implemented by off-the-shelf analog components. A prototype scanning device operating on the principle described herein has been manufactured (Fig. 5). Measurements show that the springs have a radially symmetric spring constant of about 12KN/m. The maximum resonance frequency is 645Hz; it can be reduced by adding mass to the lens socket. A mechanical power of about 75μW is required to sustain oscillation with a peak-to-peak amplitude equal to pixel spacing. In applications where this amount of power is not guaranteed to be available in the environment, the system can be mounted on piezoelectric actuators.

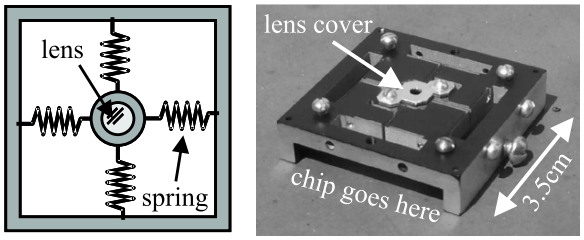

Figure 5: Left: Schematic top view of a mechanical device producing scanning powered by environmental vibrations. Right: Photograph of the actual device.

7 Conclusion

A new approach to visual sensing for machine vision purposes has been described, which relies on mechanical vibrations in the optical path to turn image features into temporal signals. These signals are processed locally in every pixel and encoded in the timing of digital pulse streams suitable for subsequent processing. A scheme is proposed for building spatial feature maps from these pulse streams using simple digital hardware. Compared to other visual sensors with focal-plane processing, this approach achieves much higher effective resolution despite a modest number of pixels and a low fill factor, due to the properties of continuous scanning. Compared to visual sensing approaches based on conventional high resolution cameras, our sensor offers a considerable reduction in subsequent data processing power requirements, because substantial preprocessing is carried out on-chip, and because the format of the output data lends itself well to real-time image feature extraction. One of the proposed mechanical implementations of the scanning function targets small robotics applications with extremely stringent power consumption specifications. In light of the long communication latency, the efficacy of a robot on a planetary mission could be greatly enhanced by autonomous navigation capability facilitated by image preprocessing with simple hardware. Vibrations of useful magnitude are likely to be plentiful on a rover driving on rough terrain such as the surface of a remote planet. Aircraft such as helicopters are also likely platforms to exploit this visual sensing principle.

8 Acknowledgements

This work was funded by the Office of Naval Research, DARPA, and the Center for Neuromorphic Systems Engineering as part of the National Science Foundation Engineering Research Center Program.

References

[1] C. Mead. *Analog VLSI and Neural Systems*. Addison Wesley, 1989.

[2] T.S. Lande. *Neuromorphic Systems Engineering - Neural Networks in Silicon*. Kluwer Academic Publishers, Dordrecht, 1998.

[3] S. Viollet and N. Franceschini. Visual servo system based on a biologically-inspired scanning sensor. In *Sensor Fusion and Decentralized Control in Robotic Systems II*, volume 3839, pages 144–155, Bellingham, 1999. SPIE.

[4] K. Hoshino, F. Mura, and I. Shimoyama. Design and performance of a micro-sized biomorphic compound eye with a scanning retina. *Journal of Microelectromechanical Systems*, 9(1):32–37, 2000.

[5] A. Kimachi, R. Imaizumi, and S. Ando. Intelligent image sensor with a vibratory mirror mimicking involuntary eye movement. In *Technical Digest of the 16th Sensor Symposium*, pages 171–176, 1998.

[6] S. Ando and A. Kimachi. Time-domain correlation image sensor: First cmos realization of demodulator pixels array. In *Proc. 1999 IEEE Workshop on Charge-Coupled Devices and Advanced Image Sensors*, pages 33–36, Karuizawa, Japan, 1999.

[7] O. Landolt, A. Mitros, and Koch C. Visual sensor with resolution enhancement by mechanical vibrations. In *Proc. 2001 Conf. Advanced Research in VLSI*, pages 249–264, Salt Lake City, Utah, 2001.