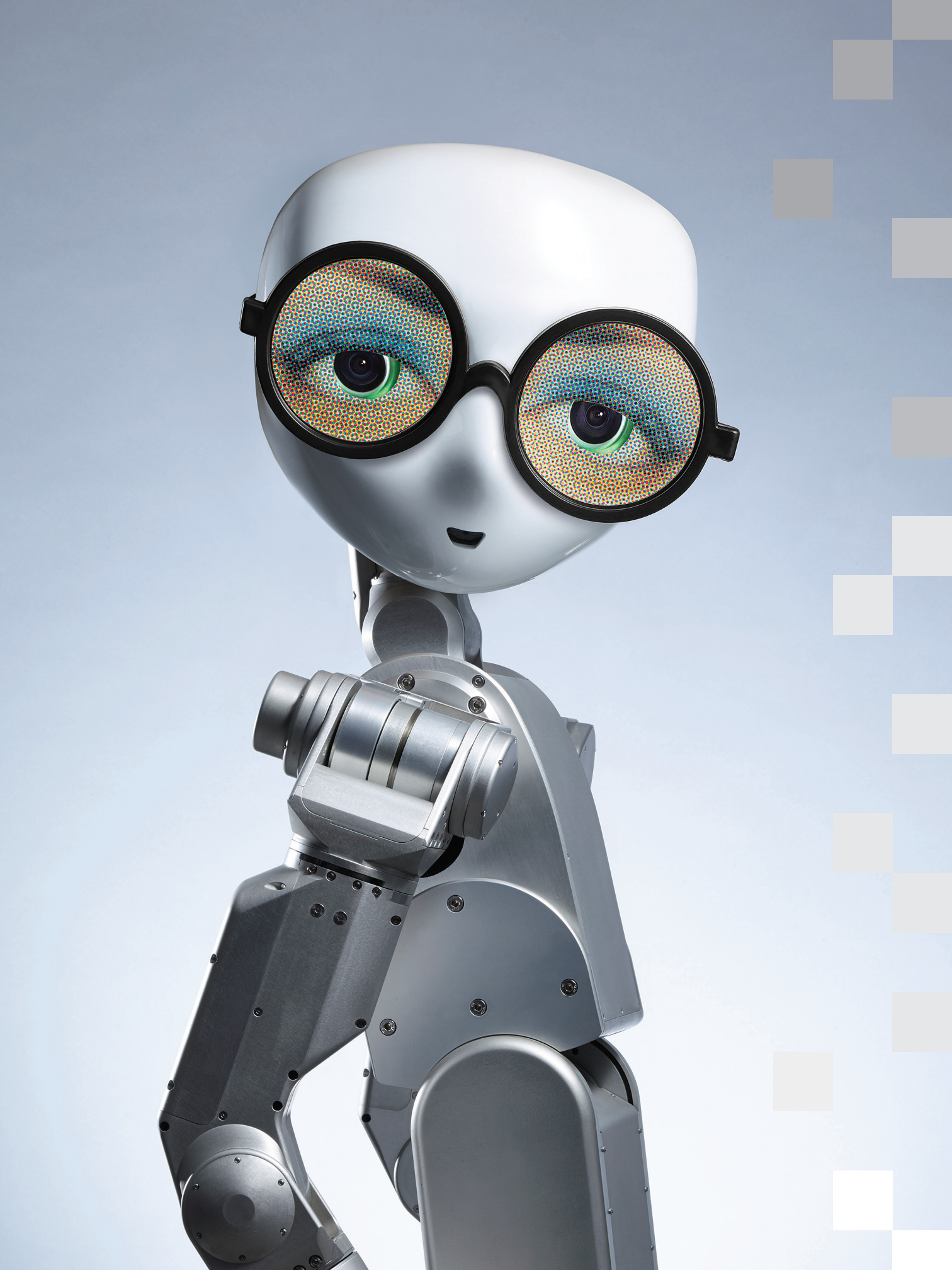


Giving Machines Humanlike

Vision similar to our own would let devices capture images more efficiently

by CHRISTOPH POSCH, RYAD BENOSMAN & RALPH ETIENNE-CUMMINGS

Photo-illustration by **THE VOORHES**

When Eadweard Muybridge set up his cameras at Leland Stanford's Palo Alto horse farm in 1878, he could scarcely have imagined the revolution he was about to spark. Muybridge rigged a dozen or more separate cameras using trip wires so that they triggered in a rapid-fire sequence that would record one of Stanford's thoroughbreds at speed. The photographic results ended a debate among racing enthusiasts, establishing that a galloping horse briefly has all four legs off the ground—although it happens so fast it's impossible for anyone to see. More important, Muybridge soon figured out how to replay copies of the images he took of animal gaits in a way that made his subjects appear to move.

Generations of film and video cameras, including today's best imaging systems, can trace their lineage back to Muybridge's boxy cameras. Of course, modern equipment uses solid-state detectors instead of glass plates, and the number of frames that can be taken each second is vastly greater. But the basic strategy is identical: You capture a sequence of still images, which when played back rapidly gives the viewer the illusion of motion.

If the images are to be analyzed by a computer rather than viewed, there's no need to worry about whether the illusion is a good one, but you might still need to record lots of frames each second to track the action properly.

Actually, even with a high frame rate, your equipment may not be up to the task: Whatever you are trying to analyze could be changing too quickly. What then do you do? Many engineers would answer that question by looking for ways to boost the video frame rate using electronics with higher throughput. We argue that you'd be better off reconsidering the whole problem and designing your video equipment so it works less like Muybridge's cameras and instead functions more like his eyes.

The general strategy of creating electronic signal-processing systems inspired by biological ones is called neuromorphic engineering. For decades, this endeavor has been an exercise in pure research, but over the past 10 years or so, we and other investigators have been pursuing this approach to build practical vision systems. To understand how an artificial eye of the kind we've been investigating can outperform even a high-speed video camera, let us first disabuse you of the idea that the way modern video gear operates is sensible.

Imagine for a moment that you're trying to analyze something that happens really fast, say, a pitcher throwing a baseball. If you try to use a conventional video camera, which records at something like 30 or perhaps even 60 frames per second, you'll miss most of the movement of the pitcher's arm as he whips the ball toward the plate. Perhaps some frames will catch his arm in different positions. But you'll capture relatively little information of interest, along with much redundant imagery of the pitcher's mound, the infield turf, and other unchanging parts of the background. That is, the scene you record will be under- and oversampled at the same time!

There's no way to avoid that problem given that all parts of the image sensor in your camera share a common timing source. While this weakness won't be a problem for a casual viewer, if you wanted a computer to analyze nuances of the pitcher's arm motion, your data will be woefully inadequate. In some cases, sophisticated postprocessing might let you derive the results you wanted. But this brute-force approach would fail you in environments with limited power, bandwidth, and computing resources such as on mobile devices, multicopter drones, or other kinds of small robots.

The machine-vision community has been stuck with this basic problem for decades. But the situation may soon be changing for the better as we and other researchers develop equipment that samples different parts of the scene at different rates, mimicking how the eye works. With such gear, those parts of the scene that contain fast motions are sampled rapidly, while slow-changing portions are sampled at lower rates, going all the way down to zero if nothing changes.

OUT THE GAIT: A sequence of frames, like those of Eadweard Muybridge's animal-locomotion studies (such as the one shown above), undersamples parts of the scene—like the horse's swiftly moving legs here—while oversampling the parts that remain static.

Getting video cameras to work this way is tricky, because you don't know beforehand which parts of the scene will change and how rapidly they will do so. But as we describe below, the human eye and brain deal with this problem all the time. And the rewards of copying how they work would be enormous. Not only would it make fast-changing subjects—explosions, insects in flight, shattering glass—more amenable to analysis, it would also allow the video cameras on smartphones and other battery-operated devices to record ordinary motions using much less power.

Engineers often liken the eye to a video camera. There are some similarities to be sure, but in truth the eye is a much more complicated creation. In particular, people's retinas don't just turn light into electrical signals: They process the output of the eye's photoreceptor cells in sophisticated ways, capturing the stuff of interest—spatial and temporal changes—and sending that information to the brain in an amazingly efficient manner.

Knowing how well this approach works for eyes, we and others are studying machine-vision systems in which each pixel adjusts its own sampling in response to changes in the amount of incident light it receives. What's needed to implement this scheme is electronic circuitry that can track the amplitudes of each pixel continuously and record changes of only those pixels that shift in light level by some very small prescribed amount.

This approach is called level-crossing sampling. In the past, some people have explored using it for audio signals—for example, to cut down on the amount of data you'd have to record with the usual constant-rate sampling. And academic researchers have been building electronic analogues of the retina in silicon for research purposes since the late 1980s. But only in the past decade have engineers attempted to apply level-crossing sampling to the practical real-time acquisition of images.

Inspired by the biology of the eye and brain, we began developing imagers containing arrays of independently operating pixel sensors in the early 2000s. In our more recent cameras, each pixel is attached to a level-crossing detector and a separate exposure-measurement circuit. For each individual pixel, the electronics detect when the amplitude of that pixel's signal reaches a previously established threshold above or below the last-recorded signal level, at which point the new level is then recorded. In this way every pixel optimizes its own sampling depending on the changes in the light it takes in.

With this arrangement, if the amount of light reaching a given pixel changes quickly, that pixel is sampled frequently. If nothing changes, the pixel stops acquiring what would just prove to be redundant information and goes idle until things start to happen again in its tiny field of view. The electronic circuitry associated with that pixel outputs a new measurement just as soon as a change is detected, and it also keeps track of the position in the sensor array of the pixel experiencing that change. These outputs, or "events," are encoded according to a protocol called Address Event Representation, which came out of Carver Mead's lab at Caltech in the early 1990s. The train of events such a vision sensor outputs thus resembles the train of spikes you see when you measure signals traveling along a nerve.

CORBIS

The key is that the visual information is not acquired or recorded as the usual series of complete frames separated by milliseconds. Rather, it's generated at a much higher rate—but only from parts of the image where there are new readings. As a result, just the information that is relevant is acquired, transmitted, stored, and eventually processed by machine-vision algorithms.

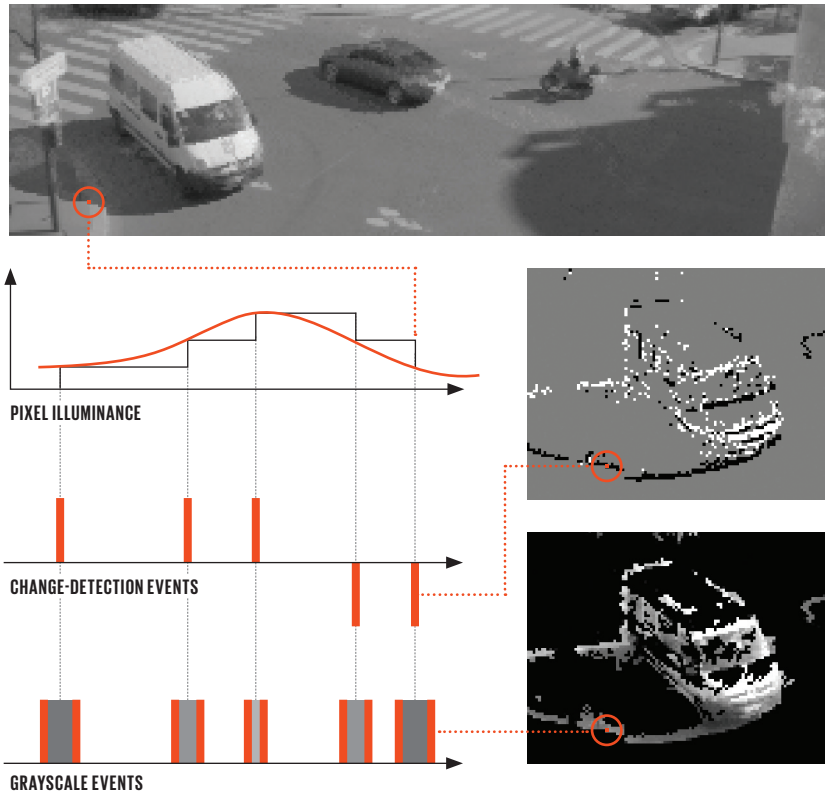
We designed the level-crossing and recording circuits in our camera to react with blazing speed. With our equipment, data acquisition and readout times of a few tens of nanoseconds are possible in brightly lit scenes. For standard room-light levels, acquisition and readout require a few tens of microseconds. These rates are beyond all but the most sophisticated high-speed video cameras available today, cameras costing hundreds of thousands of dollars. And even if you could afford such a camera, it would deluge you with mostly worthless information. Sampling different pixels at different rates, on the other hand, reduces not just equipment cost but also power consumption, transmission bandwidth, and memory requirements—advantages that extend well beyond the acquisition stage. But you'll squander those benefits if all you do is reconstruct a series of ordinary video frames from the data so that you can apply conventional image-processing algorithms.

To fully unlock the potential of eyelike vision sensors, you need to abandon the whole notion of a video frame. That can be a little hard to get your head around, but as soon as you do that, you become liberated, and the subsequent processing you do to the data can resolve things that you could otherwise easily miss—including the detailed arm motions of our hypothetical baseball pitcher.

To do this, though, you'll have to rethink how you process the data, and you'll probably have to write new code instead of using a standard video-analysis library. But the mathematical formulations appropriate for this new kind of video camera are simple and elegant, and they yield some very efficient algorithms. Indeed, in applying such algorithms to the output of our autosampling vision sensors, we were able to show that certain real-time vision tasks could be run at a rate of tens to even hundreds of kilohertz, whereas conventional frame-based video-analysis techniques applied to the same situation topped out at a painfully slow 60 hertz.

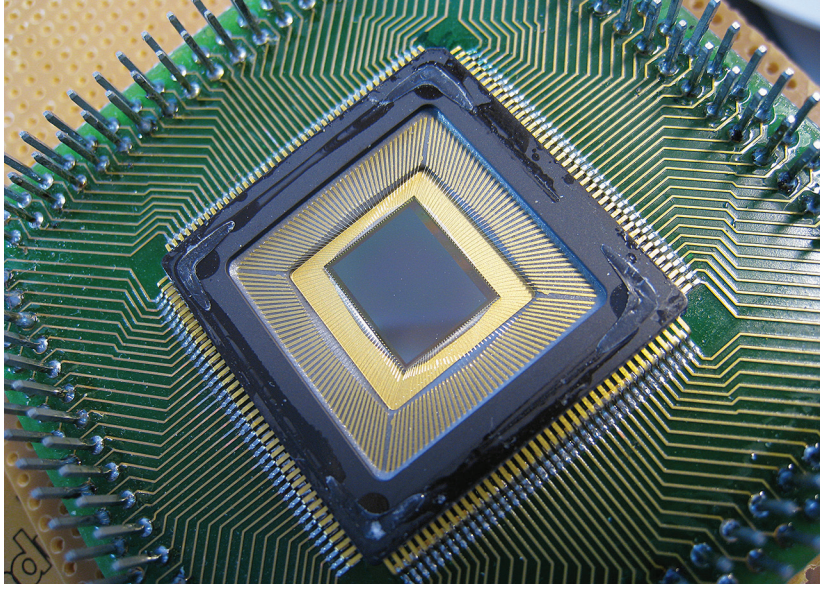
Another advantage of analyzing the nearly con-tinuous data streams from our eyelike sensors instead of a series of conventional video frames is that we can make good use of signal timing, just as biological neurons do. This is perhaps best explained with a specific example.

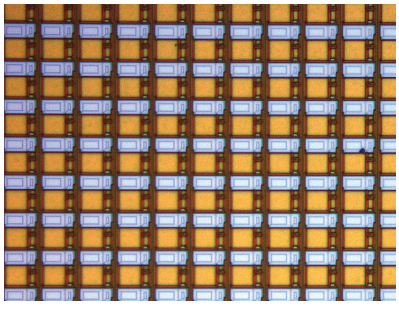
Event-Based Video at a Crossroads

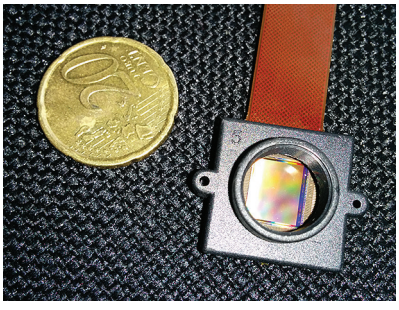


AN EVENT-BASED vision sensor takes in a traffic intersection [top]. The sensor outputs a measurement only when the light striking a pixel shifts by a preset amount. When the light level crosses that threshold, circuitry attached to the pixel produces a "change-detection event" and triggers an illuminance measurement, encoded by the interval between two more events. The results for the entire pixel array can be viewed for a given time slice, but only changing pixels contribute to these images: the positive and negative level-crossing events [white and black, respectively, middle right] and associated illuminance measurements [nonblack areas, bottom right].

CHRONOCAM (3)







EYES OF SILICON: The author's asynchronous time-based image sensor is attached to a test fixture [top]. A close-up of a portion of the sensor's pixel array reveals its many photodiodes [blue areas, left]. A packaged vision sensor [right] serves as the input device for Pixium Vision's artificial retina.

Suppose you wanted to design a mobile robot that uses a machine-vision system to navigate its environment. Clearly, having a 3-D map of the things around it would be helpful. So you'd no doubt outfit the robot with two somewhat separated cameras so that it had stereo vision. That much is simple enough. But now you have to program its robotic brain to analyze the data it receives from its cameras and turn that into a representation of 3-D space.

If both cameras record something distinct—let's say it's a person stepping in front of the robot—it's easy enough to work out how far away the person is. But suppose two different people enter the robot's field of view at the same time. Or six people. Working out which one is which in the two camera views now gets more challenging. And without being able to ascertain identities for certain, the robot will not be able to determine the 3-D position of each one of these human obstacles.

With vision sensors of the type we've been studying, such matching operations become simpler: You just need to look for coincidences in the readings from the two cameras. If pixels from separate cameras register changes at the very same instant, they are almost certainly observing the same event. Applying some standard geometrical tests to the observed coincidences can further nail down the match.

Tobi Delbrück and others at ETH Zurich demonstrated the power of this approach in 2007 by building a small-scale robotic soccer goalie using an eyelike sensor that was broadly similar to ours. It had a reaction time under 3 milliseconds. (Peter Schmeichel, eat your heart out.) Were you to try to achieve that speed using a conventional video camera, you'd need to find one that could record some hundreds of frames per second, and the computational burden would be enormous. But with Delbrück's neuromorphic Dynamic Vision Sensor, the computer running his soccer goalie was loping along at a mere 4 percent CPU load.

Compared with standard video techniques, neuromorphic vision sensors offer increased speed, greater dynamic range, and savings in computational cost. As a result, demanding machine-vision tasks—such as mapping the environment in 3-D, tracking multiple objects, or responding quickly to perceived actions—can run at kilohertz rates on cheap battery-powered hardware. So this kind of equipment would allow for "always-on" visual input on smart mobile devices, which is currently impossible because of the amount of power such computationally intense tasks consume.

Another natural application of neuromorphic vision sensors is in electronic retinal implants for restoring sight to those whose vision has been lost to disease. Indeed, two of us (Posch and Benosman) helped to found Pixium Vision, a French company that has developed a neuromorphic retinal implant, which is now undergoing clinical trials. Unlike competing implants under development, which are frame based, Pixium's products use event-based sampling to provide patients with visual stimulation. Right now, these implants are able to give patients only a general ability to perceive light and shapes. But the technology should improve swiftly over the next few years and perhaps one day will be able to offer people who have lost their natural vision the ability to recognize faces—all thanks to artificial retinas inspired by real ones.

You can expect eyelike vision sensors to evolve from the pioneering designs available today into forms that eventually play a big role in medical technology, robotics, and more. Indeed, it wouldn't surprise us if they proved just as seminal as Muybridge's wooden cameras. ■

POST YOUR COMMENTS at http://spectrum.ieee.org/machinevision1215

CHRONOCAM (3)