DISS. ETH No. 16898

An AER Temporal Contrast Vision Sensor

A dissertation submitted to the

SWISS FEDERAL INSTITUTE OF TECHNOLOGY ZURICH

for the degree of

DOCTOR OF NATURAL SCIENCES

presented by

PATRICK LICHTSTEINER

Dipl. Phys. ETH

born 01.03.1976

citizen of
Sempach (Lucerne) and Schötz (Lucerne)

accepted on the recommendation of

Prof. Dr. Rodney Douglas, examiner
Dr. Tobias Delbrück, coexaminer
Prof. Dr. Bernabé Linares-Barranco, coexaminer

2006

Summary

Electronic eyes became very abundant in our environment in the last few years. Every railway station, airport, bank, super-market and large public building is equipped with surveillance cameras. But surveillance is just one of many application fields for electronic vision devices. Artificial vision is also used in industrial manufacturing, in safety systems in industrial environments, for visual quality control and failure investigation, for visual stock control, for barcode reading, to control automated guided vehicles etc. In these applications, human vision is replaced by an electronic camera paired with sophisticated computer vision software running on the computer to which the camera is attached. The complete system, comprising one or more cameras, computers and sometimes also actuators (i. e. robots), is called a machine vision system.

Although machine vision systems are widely and successfully used, they are limited to perform narrowly defined tasks under highly controlled and repetitive circumstances. No machine vision system comes even close to the capabilities of human vision in terms of image comprehension, tolerance to lighting variations, image degradation, scene variability etc. Machine vision systems do not *see* in the same way that human beings are able to and the principle differences between machine vision systems and biological vision systems start right at the front end electronic cameras that are not equivalent to human eyes.

Almost all machine vision systems rely on standard electronic imagers that produce image at a regular frequency. Frame-based architectures carry hidden costs because they are based on a stroboscopic series of snapshots taken at a constant rate. All pixels are sampled repetitively even if they have nothing novel to say. The pixel bandwidth is limited by the identical global sampling rate, the frame rate. The dynamic range of frame based cameras is typically limited by the finite pixel capacity for integrated photo charge as well as by the identical pixel gain and integration time. In unsupervised environments with natural lighting, the disadvantages of limited dynamic range and bandwidth can become the limiting factor of the whole machine vision system.

Biological visual systems have no notion of a frame. The retinal photoreceptors work in continuous time lacking any global reset or readout signal. They are embedded in a neuronal network that does significant focal plane computation such as gain control and redundancy reduction. The retinal outputs are massively parallel and data-driven. In vertebrates, the decisions on when to quantize are made by the ganglion cells that project to the brain. A quantization is immediately communicated from the retina to the brain by the axon of the ganglion cell in the form of a digital pulse (this principle is often called *integrate and fire*). The information about the spatial origin of this pulse is given by its unique communication channel, and information about the intensity or

intensity differences is coded in the relative timing of a sequence of such pulses. Eyes are optimized for the extraction of relevant information from a visual scene and in this respect outperform any existing camera by far.

These considerations underlie the belief that emulating this asynchronous data driven biological architecture will lead to capturing the power of biological vision into electronic devices. The continuing increase in integration density in silicon integrated circuits means that it is becoming increasingly feasible to develop vision sensor devices with complex pixels that still have acceptable pixel size and fill factor.

In this thesis we will describe the development of a vision sensor that imitates some specific retinal functionality; the sensitivity to local relative temporal changes in illumination. It loosely models the transient pathway in the retina by emulating three key principles of biological vision: its event-based output, its representation of relative luminance change, and its representation of sign into separate positive and negative channels. Our pixel combines an active continuous-time logarithmic photo sensor with a well-matched self-timed switched-capacitor amplifier. Each pixel continuously monitors its photo current for changes and responds with an ON or OFF event that represents a fractional increase or decrease in intensity that exceeds a tunable threshold. Events are immediately communicated asynchronously off chip on a self-timed bus using *Address Event Representation* (AER).

The latest vision sensor that we designed (chapter 6) contains a 128×128 array of pixels that encode scene reflectance change by outputting asynchronous address-events. These events are triggered by local relative intensity changes that exceed a global threshold. These events typically signify object movement and have sub-millisecond timing precision. An individual pixel bandwidth of above $3kHz$ under $1klux$ scene illumination and an array mismatch of 2.1% in relative intensity is achieved. The chip is built in a $0.35\mu m$ 4M 2P process yielding 40×40 μm^2 pixels with 9.4% fill-factor. The measured dynamic range is $120dB$ and chip power consumption is $23mW$. The response latency varies slowly with illumination and is $15\mu s$ at $> 1klux$ pixel illumination. This combination of characteristics provides unique capabilities for low-latency dynamic vision under wide illumination range with low computational requirements. The sensor has already been used successfully for different experimental applications.

The thesis is structured into 7 chapters. Chapter 1 is an introduction that encompasses a short descriptions of electronic cameras and recent development trends in imager design, relates some common knowledge about eyes, familiarizes the reader with neuromorphic engineering and AER communication and finishes in a discussion of the state of the art of comparable sensors and the justification for this work.

The description of stepwise advances in the development of our sensor covers two different pixel designs described in chapters 2 and 3. The first of these designs existed before we started this work and was done by Jörg Kramer.

In chapter 2 we will describe Jörg Kramer's pixel design, his transient imager and our own implementation of his design. The conclusions we draw after the characterization results of our first transient imager acquaint the reader with the rationale to switch to a new pixel design.

This novel pixel design is described in chapter 3. The description of this pixel design and a discussion of its three different implementations with thorough characterization results are given in chapters 4 to 6. The thesis is concluded by

chapter 7.

The work and achievements presented in this thesis stem from teamwork by me and my supervisor Tobi Delbrück. We had some help from other people, specifically Christoph Posch who joined our design, implementation and characterization efforts somewhere between chapter 4 and 6. The rest of the people who contributed to this work will be named in the acknowledgments or directly in the text at the appropriate point. Please note, that I use 'we' throughout the thesis instead of specifically naming the persons meant. This 'we' includes Tobi Delbrück and myself and in chapter 4 to 6 also Christoph Posch.

Zusammenfassung

Elektronische Augen beobachten seit Jahrzehnten unsere Umwelt. Jeder Bahnhof, jeder Flughafen und auch jede Bank ist mit Überwachungskameras ausgerüstet. Video-Überwachung ist aber nur eine von vielen Anwendungen für elektronische Kameras. Künstliches Sehen wird heutzutage auch in der industriellen Herstellung, in Sicherheitssystemen in Industrieumgebungen, für die visuelle Qualitätskontrolle und Fehlerprüfung sowie zur Strichcode Erkennung, für das Führen automatischer Fahrzeuge, etc. verwendet. In diesen Applikationen wird das menschliche Sehen durch eine elektronische Kamera im Verbund mit hochentwickelten computergestützen Bildanalyseverfahren ersetzt. Ein solches künstliches Sehsystem beinhaltet eine oder mehrere Kameras, Computer und manchmal auch Aktoren (z.B. Roboter) und wird maschinelles Bildverarbeitungssystem (machine vision system) genannt.

Obwohl solche maschinelle Bildverarbeitungssysteme weit verbreitet sind und erfolgreich eingesetzt werden, sind sie in ihrem Anwendungsbereich eingeschränkt auf eng definierte und repetitive Aufgaben und eine rigide angpasste Umgebung. Kein maschinelles Bildverarbeitungssystem kommt auch nur annähernd an die Fähigkeiten menschlichen Sehens heran in Bezug auf Bildverständnis, Toleranz für verschiedene Beleuchtungsverhältnisse, Stabilität bei variablen visuellen Gegebenheiten etc. Maschinelle Bildverarbeitungssysteme 'sehen' nicht in der gleichen Art, wie es Menschen können. Die Unterschiede zwischen den künstlichen und biologischen Systemen beginnen schon beim visuellen Sensor, denn elektronische Kameras unterscheiden sich grundsätzlich vom menschlichen Auge.

Nahezu alle maschinellen Bildverarbeitungssysteme verlassen sich auf konventionelle elektronische Kameras, welche in konstanter Abfolge Bilder, so genannte Frames (eng. für Rahmen), generieren. Framebasierte Sensorarchitekturen bringen versteckte Nachteile mit sich, denn sie sind auf stroboskopisch aufgenommene Serien von Bildern, die in einer regelmässigen Frequenz erzeugt werden, basiert. Alle Pixel werden repetitiv ausgelesen, auch wenn sich zwischenzeitlich nichts verändert hat. Somit werden in einem erheblichen Mass redundante Daten erzeugt. Die Pixelbandbreite ist limitiert durch die identische globale Auslesefrequenz, die so genannte Frame-Rate. Der Bereich der verwertbaren Lichtintensitäten von konventionellen Kameras ist typischerweise limitiert durch die beschränkte Kapazität der Pixel die elektrisch photoinduzierte Ladung zu speichern, wie auch durch die einheitliche Integrationszeit, welche die Pixel aufwenden, um photoinduzierte Ladung zu akkumulieren. In unbeaufsichtigten Umfeldern mit natürlichen Lichtverhältnissen können die aufgeführten Nachteile sehr stark ins Gewicht fallen.

Biologische Sehsysteme funktionieren ohne Frames. Die Photorezepto-

ren in der Retina arbeiten zeitkontinuierlich. Sie kommen ohne globale Rückstellungsmechanismen und Auslesesignale aus. Sie sind in ein Netz von Neuronen eingebettet. Die entscheidende Verarbeitung des visuellen Eingangs, wie zum Beispiel Verstärkungsregelung und Redundanzverminderung, wird durch das neuronale Netz direkt in der Brennebene der Retina vollführt. Der Retina-Output findet parallel und ereignisgetrieben statt. Bei Wirbeltieren wird die Entscheidung, zu welchem Zeitpunkt eine Quantisierung vorgenommen wird, durch die Ganglionzellen, die ins Hirn projizieren, getroffen. Eine Quantisierung, sprich das Erreichen eines bestimmten Schwellwertes, wird unverzüglich mit einem digitalen Puls, welcher durch das Axon der Ganglionzelle transportiert wird, von der Retina ins Hirn kommuniziert. Die Information der örtlichen Herkunft des Pulses ist durch das Axon, das den Puls transportiert, eindeutig gegeben und Informationen über die Intensität oder Intensitätsveränderungen sind durch die relativen zeitlichen Abstände der Pulse auf einem Axon gegeben. Augen sind optimiert, um effizient Informationen aus visuellen Gegebenheiten zu beziehen und übertreffen in dieser Hinsicht alle technischen visuellen Sensoren bei weitem.

Diese Überlegungen liegen der Überzeugung zu Grunde, dass eine Nachahmung der ereignisgetriebenen asynchronen Architektur von biologischen Systemen dazu führt, dass die Kraft des biologischen Sehens in elektronischen Schaltungen eingefangen werden kann. Durch die kontinuierliche Verkleinerung von integrierten elektronischen Schaltungen wird es zunehmend möglich, kompliziertere Pixels auszugestalten, ohne dass die Grösse dieser Pixel ein akzeptables Auflösungsvermögen eines visuellen Sensors verhindert.

In der vorliegenden Arbeit beschreiben wir die Entwicklung eines visuellen Sensors, der eine spezifische Funktionalität einer Retina nachempfindet; nämlich die Sensitivität auf lokale zeitliche Veränderungen der Lichteinstrahlung. Der visuelle Sensor, der ausschliesslich auf relative Belichtungsveränderungen reagiert, modelliert lose den Anfang des magnozellulären Pfades von biologischen Sehsystemen, indem er dem biologischen Vorbild folgend, drei Schlüsselprinzipien nachahmt: Die ereignisgetriebene Ausgabe, die Repräsentation von relativen Belichtungsveränderungen und die Darstellung des Vorzeichens in separaten positiven und negativen Kanälen. Unser Pixel kombiniert einen aktiven kontinuierlich operierenden logarithmischen Photosensor mit einer gut abstimmbaren Kapazitätsumschalter Schaltung, in welcher die Umschaltung der Kapazitäten durch das Pixel selbst zeitlich gesteuert wird. Jedes Pixel überwacht kontinuierlich seinen Photostrom auf Veränderungen. Es antwortet mit ON und OFF Ereignissen auf relative Erhöhungen oder Erniedrigungen der Belichtungsintensität, welche einen wählbaren Schwellwert übersteigen. Unter Zuhilfenahme des so genannten *Address Event Representation* (AER) Kommunikationsprotokolls werden Ereignisse sogleich asynchron aus dem Chip kommuniziert.

Der neueste Sehsensor, den wir konstruiert haben (Kapitel 6), beeinhaltet ein quadratisches Feld von 128 × 128 Pixeln, die auf relative Belichtungsveränderungen mit der asynchronen Ausgabe ihrer Adresse (ihre Position in X und Y Koordinaten) reagieren. Typischerweise bedeutet das Ereignis einer Adressausgabe, dass sich ein Objekt im Blickfeld des Sensors bewegt hat. Die zeitliche Präzision solcher Ereignisse ist im sub-Milisekunden Bereich. Mit dem Sensor wird eine individuelle Pixel-Bandbreite von $> 3kHz$ bei einer Beleuchtung der visuellen Szene von $< 1kLux$ erreicht. Der Pixel Versatz beträgt 2.1% in relativer Lichtintensität. Der Chip ist in einem $0.35\mu m$, 4 Metal, 2 Poly Pro-

zess gebaut. Die Pixelgrösse von $40 \times 40\ \mu m^2$ enthält einen Füllfaktor von 9.4%. Der gemessene dynamische Bereich beträgt $120dB$ und der Leistungsverbrauch des Chips beläuft sich auf $23mW$. Die Reaktionslatenzzeit variiert langsam mit der Belichtungsstärke und entspricht $15\mu s$ bei $> 1kLux$ Pixel Beleuchtung. Diese Kombination von Charakteristiken bietet einzigartige Möglichkeiten für Sehsysteme, die in unkontrollierten Lichtverhältnissen arbeiten und auf eine kurze Reaktionszeit angewiesen sind.

Das vorliegende Dokument ist in 7 Kapitel gegliedert. Kapitel 1 enthält eine Einleitung, eine kurze Beschreibung von elektronischen Kameras und zeigt die neuesten Entwicklungen im Kamerabereich auf. Es berichtet über Augen, macht den Leser mit dem neuromorphen Ingenieurwesen und AER Kommunikation bekannt und endet in einer Begründung der Bedeutsamkeit dieser Arbeit.

Die Beschreibung der schrittweisen Entwicklung unseres Sensors beinhaltet zwei verschiedene Pixelausgestaltungen, welche in den Kapiteln 2 und 3 vorgestellt werden. Die erste Pixelvariante wurde von Jörg Kramer entwickelt und existierte bereits vor Beginn dieser Arbeit.

In Kapitel 2 werden wir Jörg Kramers Pixelaufbau, seinen Sensor und unsere eigene Implementation seines Pixels beschreiben. Das Endergebnis unserer Charakterisierungsresultate wird dem Leser die Begründung für unseren Entschluss, ein neues Pixel zu entwickeln, nahe bringen.

Das neue Pixel beschreiben wir ausführlich in Kapitel 3. Die Schilderung der Implementierung des neuen Pixels in drei verschiedene Chips und die Charakterisierung dieser Chips füllen die Kapitel 4 bis 6. Mit abschliessenden Erklärungen in Kapitel 7 wird dieses Dokument beendet.

Die Arbeit und die Ergebnisse die hier präsentiert werden, entstanden in enger Zusammenarbeit zwischen mir und dem Leiter meines Doktorats, Tobi Delbrück. Wir hatten Hilfe von anderen Personen, im speziellen von Christoph Posch, der unsere Entwicklungs- und Charakterisierungsanstrengungen unterstützt hat. Christoph Poschs Beteiligung beginnt in Kapitel 4. Weitere Personen, die zum Gelingen dieser Arbeit beitrugen, werden in den Danksagungen oder an entsprechender Stelle im Dokument genannt. In diesem Dokument verwende ich die Wir Form und beziehe mich dabei auf Tobi Delbrück und mich, sowie in den Kapiteln 4 bis 6 auch auf Christoph Posch.