Doctoral Thesis

Neuromorphic Vision for Robotic Tracking and Navigation

Author(s):
Liu, Hongjie

Publication Date:
2019

Permanent Link:

https://doi.org/10.3929/ethz-b-000385246 →

DISS. ETH NO. 26337

Neuromorphic Vision for Robotic Tracking and Navigation

A thesis submitted to attain the degree of
DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by

HONGJIE LIU
MSc NTU-Singapore EEE
born on February 6th, 1987
citizen of China

accepted on the recommendation of

Prof. Dr. Tobias Delbruck, examiner
Prof. Dr. Benjamin Grewe, co-examiner

2019

Abstract

Commercial frame-based image sensors, i.e. active pixel sensors (APS), read out the accumulated charge in every 4T pixel and converts it to digital representation using column-parallel analog-to-digital converters (ADC). They are good at capturing high-quality images, but ill-suited for many vision applications where dynamic visual content is of higher interest. In contrast, event-based dynamic vision sensors (DVS) employ delta-encoding asynchronous ADCs at the pixel level, which eliminate temporal redundancy and as a consequence lends the benefits of low data rate and fast response, especially when the visual scenes have sparse dynamic activities. The encoded events are transmitted following the Address Event Representation (AER) protocol, and can be processed in software or hardware. With the unique characteristics of sparse and fast readout, DVS may well find applications in various areas such as autonomous vehicles, mobile robotics, surveillance, and industrial monitoring. However, the progress of event-based visual information processing is falling behind. In order to better exploit the advantages of DVS, it is imperative to develop customized event-processing software and hardware systems.

This thesis presents algorithms as well as heterogeneous hardware for processing asynchronous events from DVSs and DAVISs, a type of hybrid sensor that combines the functionality of both APS and DVS. The elaborated topics include detection-based tracking, real-time object avoidance, filtering of background activities, and determination of spatial-temporal correlation of events.

Visual object tracking is a fundamental task in many computer vision problems such as visual analysis, automatic driving and pose estimation. Specifically, in terms of convolution-based object tracking,

since DVS events encode the dynamic information, they can be used to indicate the Region of Interest (ROI) for convolutions. This will reduce the computation comparing to convolving through the whole frame. Using DAVIS, a more robust tracking and navigation system combining both frame-based and event-based approaches is implemented. The detection-based tracking is performed in three steps: ROIs are generated by cluster tracking the event output, likely target locations are detected by classifying foreground and background using convolutional neural network in these ROIs, and finally a particle filter infers the target location from the ROI detections. A few more methods ranging from pure event-based to event-frame based were also introduced.

Real-time object avoidance has been implemented in this thesis using an FPGA setup with a software model that emulates a biological cell that performs early detection of approaching (expanding) dark objects. The input of this hardware-implemented cell model comes from DVS, and an end-to-end event-based processing system is prototyped. The software model was developed in Java and computed with an average processing time of 370 ns/event on an NUC embedded computer. The cell output firing rate for an approaching object depends on the cell parameters that equivalently represent the needed number of input events to reach the firing threshold. For the hardware implementation on a Spartan6 FPGA, the processing time is reduced to 160 ns/event with a 50 MHz clock.

The thesis further presents the chip implementation of filtering of DVS background activities using mixed-signal circuits. The ASIC in 0.18 µm CMOS has a 128×128 array with 20×20 µm^2 cells. Models show it has a power consumption of 1 mW and a latency of 10 ns. It generates a pass flag only for events that are spatiotemporally correlated for post-processing to reduce communication/computation load and improve information relevance. Each filter cell on the chip combines programmable spatial subsampling with a temporal window based on current integration. Power-gating is used to minimize the power consumption by only activating the threshold detection and communication circuits in the cells that receive an input event. This chip targets embedded neuromorphic visual and auditory systems, where low average power consumption and low latency are crucial. It

can also in general be used to detect spatial-temporal correlation of events.

In summary, this thesis presents novel algorithms and hardware implementations based on event-based sensors to build more efficient visual processing systems in combination with accurate deep learning theory.

Keywords: *Neuromorphic Engineering, Dynamic Vision Sensor (DVS), Active Pixel Sensor (APS), Dynamic and Active Vision Sensor (DAVIS), Silicon Retina, Mixed Signal ASIC, FPGA, Approach Cell, Convolutional Neural Network, Robot, Depth, Tracking, Detection, Time to Contact*

Zusammenfassung

Handelsübliche frame-basierte Bildsensoren, d.h. Active Pixel Sensoren (APS), lesen die akkumulierte Ladung in jedem 4T-Pixel aus und wandeln sie mit spaltenparallelen Analog-Digital-Wandlern (ADC) in eine digitale Darstellung um. Sie sind gut in der Aufnahme hochwertiger Bilder, aber ungeeignet für viele Bildverarbeitungsanwendungen, bei denen dynamische und visuelle Inhalte von höherem Interesse sind. Im Gegensatz dazu verwendet der ereignisbasierte Dynamic Vision Sensor (DVS) eine Delta-Codierung von asynchronen ADCs auf Pixelebene, was die zeitliche Redundanz eliminiert und wiederum die Vorteile einer niedrigen Datenrate und schnellen Reaktion bietet, insbesondere wenn die visuellen Aufnahmen geringe dynamische Aktivitäten aufweisen. Die kodierten Ereignisse werden nach dem Address Event Representation (AER) Protokoll übertragen und können in Software oder Hardware verarbeitet werden. Mit den einzigartigen Eigenschaften der spärlichen und schnellen Auslesung kann DVS in verschiedenen Bereichen wie autonome Fahrzeuge, mobile Robotik, Überwachung und industrielle Überwachung eingesetzt werden. Der Fortschritt der ereignisbasierten visuellen Informationsverarbeitung bleibt jedoch in den Kinderschuhen. Um die Vorteile von DVS besser nutzen zu können, ist es unerlässlich, massgeschneiderte Soft- und Hardwaresysteme für die Ereignisverarbeitung zu entwickeln.

Diese Arbeit stellt Algorithmen sowie heterogene Hardware für die Verarbeitung von asynchronen Ereignissen aus DVSs und DAVISs, eine Art hybrider Sensor, der die Funktionalität von APS und

DVS kombiniert. Zu den behandelten Themen gehören detektionsbasiertes Tracking, Tiefenschätzung aus Event-Frames, Echtzeit-Objektvermeidung, Filterung von Hintergrundaktivitäten und Bestimmung der räumlich-zeitlichen Korrelation von Ereignissen.

Visuelle Objektverfolgung ist eine grundlegende Aufgabe bei vielen Bildverarbeitungsproblemen wie visuelle Analyse, automatisches Driving und Pose-Schätzung. Insbesondere im Hinblick auf die faltungsbasierte Objektverfolgung, da DVS-Ereignisse die dynamischen Informationen kodieren, können sie verwendet werden, um die Region of Interest (ROI) für Faltungen anzuzeigen. Dies reduziert die Berechnung im Vergleich zum Zusammenrollen durch den gesamten Rahmen. Mit DAVIS wird ein robusteres Tracking- und Navigationssystem implementiert, das sowohl frame-basierte als auch ereignisbasierte Ansätze kombiniert. Die detektionsbasierte Verfolgung erfolgt in drei Schritten: ROIs werden durch Cluster-Tracking der Ereignisausgabe erzeugt, wahrscheinliche Zielorte werden durch Klassifizierung von Vorder- und Hintergrund unter Verwendung eines neuronalen Faltungsnetzwerks in diesen ROIs erkannt, und schliesslich leitet ein Partikelfilter den Zielort aus den ROI-Erkennungen ab.

Die monokulare Tiefenschätzung mit einem einzigen Ereignisrahmen des DVS wird mit handelsüblichen Deep-Learning-Algorithmen auf der Basis von Faltungsneuronalen Netzen (CNN) untersucht. Dieser Algorithmus kann in selbstfahrenden Autos, Bildbearbeitung und Augmented Reality als preiswerterer alternativer Algorithmus im Vergleich zu Systemen verwendet werden, die auf Lidarsensoren basieren. Die Event-Frames werden auf einer festen Anzahl von Ereignissen integriert, um eine anpassbare Bildrate für unterschiedliche Ereignisraten zu haben. Mit dieser Methode kann die Effizienz des Verfahrens verbessert werden, was die Wahrscheinlichkeit verringert, dass das System bei schnellen und intensiven visuellen Veränderungen Informationen verpasst.

Echtzeit-Objektvermeidung wurde mit einem FPGA-Setup und einem Softwaremodell in dieser These implementiert, das eine biologische Empfindlichkeitszelle emuliert, um eine frühzeitige Erkennung von sich nähernden (expandierenden) dunklen Objekten durchzuführen. Die Eingabe dieses hardwareimplementierten Zellmodells stammt von DVS, und ein End-to-End ereignisbasiertes Verarbeitungssystem wird prototypisch realisiert. Das Softwaremodell wurde

in Java entwickelt und mit einer durchschnittlichen Verarbeitungszeit von 370 ns/Ereignis auf einem NUC Embedded Computer berechnet. Die Zellausgangsfeuerungsrate für ein sich näherndes Objekt hängt von den Zellparametern ab, die äquivalent die erforderliche Anzahl von Eingangsereignissen darstellen, um den Brennschwellenwert zu erreichen. Für die Hardware-Implementierung auf einem Spartan6 FPGA reduziert sich die Verarbeitungszeit auf 160 ns/Ereignis bei einem 50 MHz Takt.

Die These stellt weiterhin die Chip-Implementierung der Filterung von DVS-Hintergrundaktivitäten mit Mixed-Signal-Schaltungen vor. Das ASIC in 0.18 µm CMOS hat ein 128×128 Array mit 20×20 µm^2 Zellen. Es zeigt eine Leistungsaufnahme von 1 mW und eine Latenz von 10 ns. Der berichtete Chip erzeugt ein Pass-Flag nur für Ereignisse, die räumlich und zeitlich für die Nachbearbeitung korreliert sind, um die Kommunikations-/Berechnungslast zu reduzieren und die Informationsrelevanz zu verbessern. Jede Filterzelle auf dem Chip kombiniert programmierbare räumliche Unterabtastung mit einem zeitlichen Fenster, das auf der Stromintegration basiert. Power-Gating wird verwendet, um den Stromverbrauch zu minimieren, indem nur die Schwellenwerterfassungs- und Kommunikationsschaltungen in den Zellen aktiviert werden, die ein Eingangsereignis empfangen. Dieser Chip zielt auf eingebettete neuromorphe visuelle und auditorische Systeme ab, bei denen ein niedriger durchschnittlicher Stromverbrauch und eine geringe Latenzzeit entscheidend sind. Es kann auch im Allgemeinen verwendet werden, um räumlich-zeitliche Zusammenhänge von Ereignissen zu erkennen.

Diese Arbeit präsentiert neue Algorithmen und Implementierungen von Hardware, die auf ereignisbasierten Sensoren basieren, um effizientere visuelle Verarbeitungssysteme in Kombination mit präziser Theorie des tiefen Lernens aufzubauen.

Keywörter: *Neuromorphic Engineering, Dynamic Vision Sensor (DVS), Active Pixel Sensor (APS), Dynamic and Active Vision Sensor (DAVIS), Silicon Retina, FPGA, Approach Cell, Convolutional Neural Network, Robot, Depth, Tracking, Detection, Time to Contact*