SMARTCAM FOR REAL-TIME STEREO VISION
Address-event Based Embedded System

Stephan Schraml, Peter Schön and Nenad Milosevic
Austrian Research Centers GmbH, Donau-City-Strasse 1, 1220 Wien, Austria

Keywords: Stereo vision, real-time vision.

Abstract: We present a novel real-time stereo smart camera for sparse disparity (depth) map estimation of moving objects at up to 200 frames/sec. It is based on a 128x128 pixel asynchronous optical transient sensor, using address-event representation (AER) protocol. An address-event based algorithm for stereo depth calculation including calibration, correspondence and reconstruction processing steps is also presented. Due to the on-chip data pre-processing the algorithm can be implemented on a single low-power digital signal processor.

1 INTRODUCTION

Generally speaking, smart cameras are not general-purpose devices since they are specially designed for dedicated applications. The underlying idea of any smart camera system is the specific visual information processing for given application, the goal of which is usually not to provide images of better quality but to understand and describe what is happening in the images for the purpose of better decision-making or as a early-processing step for further back-end processing.

In difference to standard CMOS or CCD imagers a smart camera outputs either the features extracted from the captured images or a high-level description of the scene, which is fed into an automated control system for decision making. On the other side, the aspect of availability of 3D scene information greatly simplifies or is even crucial for many real-world image processing applications.

Smart camera systems are usually subject to many constraints on the design, implementation and production of the device which encapsulates it, such as low power, limited resources, real-time processing and low cost. For such solutions one has to find a reasonable compromise between algorithm complexity and fair approximation which obeys the general assumptions imposed by the selected application. To circumvent these constraints we follow a new paradigm in image processing which make use of vision sensors with focal-plane pre-processing.

In this paper, we present a compact stereo smart camera for sparse depth (disparity) map estimation of moving objects, which uses an optical transient sensor (Lichtsteiner, 2006) as front-end element. This sensor contains an array of autonomous, *self-signaling* pixels which *asynchronously* respond to the temporal changes (transients) of local brightness, and use sparse output representation of image information with minimum amount of redundancy while preserving precise timing information. The transient sensor seems to be a reasonable choice for applications which mainly consist of highly correlated time-variant scenes coupled with large static background regions.

Due to on-chip pre-processing of visual information, coupled with sparse image coding allow us the implementation of (as a general rule) computationally very intensive stereo depth calculation on a single low-power DSP.

This article is organized as follows: Section 2 give a short overview on optical transient sensor and address-event representation (AER) protocol. Section 3 presents the hardware architecture of our embedded system. The implemented stereo algorithm and experimental results in realistic environment are described and shown in Section 4. Finally, Section 5 provides summary, future directions and concludes the paper.

Schraml S., Schön P. and Milosevic N. (2007).
SMARTCAM FOR REAL-TIME STEREO VISION - Address-event Based Embedded System.
In *Proceedings of the Second International Conference on Computer Vision Theory and Applications - IU/MTSV*, pages 466-471

2 TRANSIENT IMAGER AND ADDRESS-EVENT REPRESENTATION

The optical transient sensor (Kramer, 2002) (Lichtsteiner, 2004) is a compact *continuous-time photoreceptor*, whose individual pixels adapt to background illuminance and react to local temporal illuminance changes. Idealized function of each pixel is to compute the rectified derivative of log intensity I

$$\frac{d}{dt}\log I = \frac{dI/dt}{I} \tag{1}$$

The primary goal of logarithmic phototransduction is to compute a "self-normalized" temporal derivative (i.e. contrast transient) which is invariant to absolute illumination. The individual pixels are also polarity-sensitive, i.e. they respond to positive (ON) and negative (OFF) transients at separate channels.

The optical transient imager uses **A**ddress-**E**vent **R**epresentation (AER) output format. This communication protocol is an asynchronous digital multiplexing protocol, previously proposed by (Mahowald, 1992) and (Sivilotti, 1991) in order to model the transmission of neural information in biological systems. Its underlying idea is that the channel bandwidth should be devoted to the transmission of significant signal, i.e. the AER protocol is event-driven since only active pixels transmit their output over the shared bus and the bus stay unused if no changes are detected by the sensor. Several AER designs have been proposed in the literature (Mahowald, 1992)(Mortara, 1998), and the one we used in this paper was developed by (Boahen, 2000). Also, the AER has been already proposed for stereo correspondence calculation of one-dimensional pictures (Häflinger, 2002).

In our case, each time the derivative exceed a given threshold a communication packet (digital pulse) called address-event (AE), is generated and multiplexed onto an arbitrated common binary data bus. The information is encoded in address-event itself, i.e. its address contains the time of origin (time-stamp) t_{ev}, the corresponding array-location of sending pixel *Xtev* and *Ytev*, and the sign of the contrast transient ω_{tev} (ON-positive or OFF-negative).

The signal coming from transient imager can therefore be modelled as a time-series of single address-events, called AE-stream:

$$AE_{stream}(t) = \sum_{t_{ev}} AE(t_{ev}) \tag{2}$$

As the complete temporal information is encoded in time-of-origin t_{ev}, the signal envelope is not significant for further signal processing, and the individual address-event can therefore be modelled as:

$$AE(t_{ev}) = \omega_{t_{ev}} \delta(t - t_{ev}) \delta_{x,x_{t_{ev}}} \delta_{y,y_{t_{ev}}} \tag{3}$$

with ω_{tev} = +1 for positive and ω_{tev} = -1 for negative contrast transients.

Since static scenes produce no signal output, in the frame-representation of sensors field-of-view moving objects are represented as a set of coherent edges, as showed in Figure 4. In order to visualize the AE data, events have been accumulated for a 20 millisecond interval and restored like a video frame. The different grey levels in Figure 4 are proportional to pixel activity per unit time.

Compared with conventional frame-based digital stereo processing, the computation of address-events is obviously much more efficient and requires less memory and computational power for applications where no dense disparity information of moving objects is needed. Moreover it has been proven analytically (Mahowald, 1992), that used with a system that has a sparse activation profile (as transient imager in our case) the address-event communication framework is able to preserve timing information orders of magnitude better than sequential scan.

3 EMBEDDED SYSTEM IMPLEMENTATION

The hardware architecture of our embedded system (shown in Figure 1) consists essentially of following function groups: two optical transient sensors as sensing element, a buffer unit consisting of multiplexer and First-In First-Out (FIFO) memory, and a digital signal processor (DSP) as processing unit.

The transient imager we used here (Lichtsteiner, 2006) has been developed in our group in co-operation with Institute of Neuroinformatics at ETH Zurich, and represents an improved version of imager developed earlier by (Kramer, 2002). It consists of an array of 128x128 pixels, built in a standard 0,35µm CMOS-technology. Each sensor pixel performs at the same time photosensing, signal pre-processing as described in Section 2, and analog-to-digital conversion for interfacing.

The pixel address-event data are read-out through non-greedy Boahen-type arbiter (Boahen, 2000). Our transient imager features wide dynamic range of 120dB of illumination and represents a high-speed real-time device, since the only delay that has to be taken into account is that for arbitration, in order of few hundred nanoseconds (equals to effective framerate > 1000fps). The operating parameters of the imager are tuned by on-board digital-to-analog converters controlled by DSP.

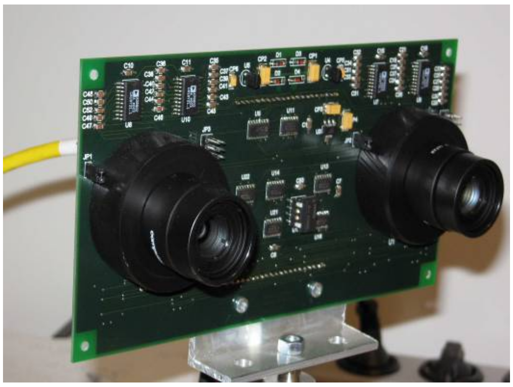

Figure 1: The breadboard of real-time Smart Eye(s) embedded stereo system.

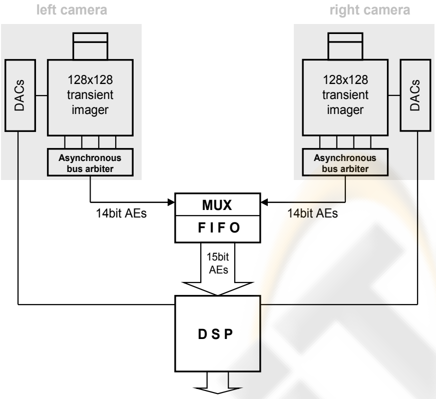

Figure 2: Hardware architecture and signal flow.

The signal flow and a block diagram of our system are plotted in Figure 2: the address-events generated in the transient sensors arrive first at the multiplexer unit which performs Round Robin scheduling. Subsequently they are forwarded to the DSP over a FIFO buffer memory. Every AE received by DSP is labelled by attaching the processor clock ticks with 1ms or less precision as a time stamp. These AE-data are used as input stream for subsequent processing, as described in next section.

All three steps of stereo processing algorithm including rectification, matching and disparity calculation is implemented on Blackfin BF537 DSP from Analog Devices® at 600MHz, with 32MB SDRAM and 4MB flash, on-chip memory.

Our embedded system supports on-board Ethernet function and dissipates in total roughly 5W of electrical power. Obviously it is capable to be used as a compact remote stand-alone application, since it is suitable for self-sufficient battery or solar energy supply and output data can be distributed to any host computer or other IP-client for further high-level image processing tasks.

4 REAL-TIME DEPTH ESTIMATION

Our implemented algorithm for real-time depth estimation consists of three major processing steps, which are common for the most computational stereo problems: (i) camera calibration and rectification, (ii) stereo correspondence calculation and (iii) reconstruction. In the following subsection we give a brief description for each processing step.

4.1 The Implemented Algorithm

In order to achieve a quantitatively accurate measure of depth, first step that has to be performed is the *camera calibration and stereo-rectification* of AE-data. The computation of internal camera parameters (including focal length, principal point, skew coefficient and radial lens distortion), external parameter (rotation matrix and translation vector) and epipolar line alignment was performed using the Camera Calibration Toolbox® (Bouguet, 2005) that employs a nonlinear gradient descent technique.

Due to AER characteristics described in section 2, the camera calibration and rectification of incoming address-events can be done in computationally very efficient manner: based on calibration results a warp-matrix is generated, and the array-addresses *Xtev* and *Ytev* of each AE are corrected instantaneously (on-the-fly) in order to eliminate lens distortion and to have epipolar lines that are aligned with horizontal axis.

In the next step, we performed *stereo correspondence* calculation using standardised area-based algorithm, modified/readapted in order to take advantage of AE-based processing. The matching algorithm is applied successively to pixel event activity accumulated in a time-slot of duration *DT*. Thereby is the event activity of array pixels stored in form of a sorted dynamic list, where the event rate is now encoded in the magnitude of AE_{act}:

$$AE_{act}(kDT;x,y)=\int_{kDT}^{(k+1)DT}AE(t_{ev})dt;\quad k=\mathbf{0},\mathbf{1},\mathbf{2},... \qquad (4)$$

The time-slot duration *DT* (which determines the temporal resolution of our system) is variable parameter and can be chosen regarding to the characteristic time-scale of the scene and settings of temporal imager bias-values. The typical value of *DT* achievable on our system lies between 5ms and 50ms, which is equivalent to the effective frame rate of 200 fps – 20 fps if compared to the conventional frame-based stereo systems

Starting from one active pixel in the left image the algorithm search for the best candidate in the right image by evaluating the similarity metric within the disparity interval between d_{min} and d_{max}. This process is then iterated for the successive AEs lying along the scan line. Our algorithm works by using fixed sized rectangular window (2B+1)·(2B+1) placed around the active pixel of interest.

We employed and tested several types of similarity metrics like Normalized Cross-Correlation, Sum of squared differences and Census-Transform (Zabih, 1994), but the best overall performance was achieved using *Normalized Sum of Absolute Differences* (NSAD). The quantity of the match for a given disparity *d* was therefore evaluated according to NSAD-score:

$$NSAD(d)=\sum_{x=-B}^{+B}\sum_{y=-B}^{+B}\frac{\left|AE_{act}^{L}(x,y)-AE_{act}^{R}(x+d,y)\right|}{\left|AE_{act}^{L}(x,y)\right|+\left|AE_{act}^{R}(x+d,y)\right|} \qquad (5)$$

Finally, the correspondence is computed using the Winner-Take-All principle, by finding the minimum NSAD value for each matched active pixel pair that lie on the same scan line. For the sake of improving match reliability we additionally perform bidirectional consistency check (Di Stefano, 2004) where the roles of two images are reversed, and only those matches that turn out to be coherent after matching left-to-right and right-to-left are considered.

At the end, the conclusive processing step performs *reconstruction*: due to the epipolar constraint the distance *Z* of the scene point from the camera can be calculated by simple triangulation *Z = bf / d*, where *b* is the stereoscopic baseline, *f* is the focal length of the camera lenses and *d* is the estimated disparity. The distance between the cameras *b* is 130mm, and standard camera lenses with focal lengths from 4mm to 8mm have been used.

4.2 Experimental Results

As described in previous section the first processing step is the calibration of two transient imagers and rectification of incoming address-events. Our measurements have shown that the influence of the radial lens distortion for shorter focal lengths is quite dominant and can not be neglected.

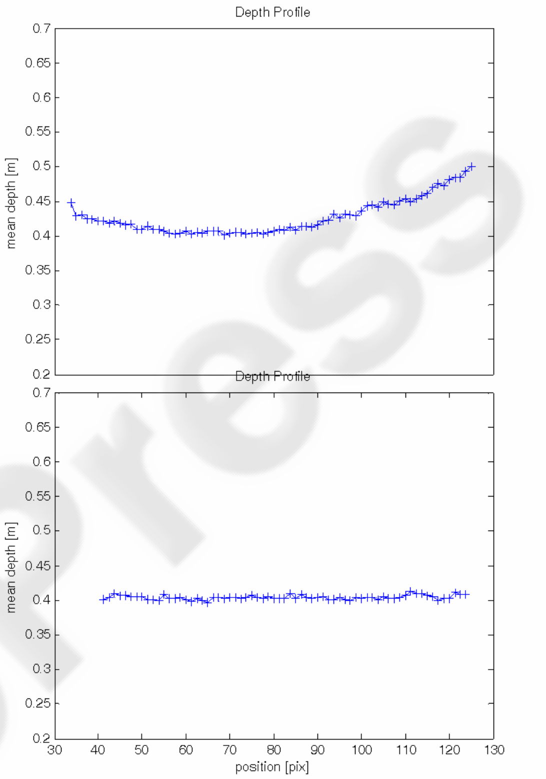

Figure 3: Mean depth of moving target (referred to the left sensor) plotted against horizontal position between 40 and 128 pixel. Each position is defined by a dedicated time slot, as the target swing at equidistance to the sensor with constant velocity.

The experimental result in figure 3 shows the significance of radial lens distortion correction in case of 4mm lens. The figure 3 presents the comparison of the mean depth values of a moving target, as calculated by our stereo algorithm, without (upper figure) and with radial correction (lower figure). As a target we have used a thin pendulum swinging parallel to our system at distance of 40cm. Without lens correction mean depth shows a declination to higher values the farther the object is located from the optical centre (at about 64 pixel).

By applying lens correction mean depth becomes as expected a straight line.

In order to evaluate the efficiency and accuracy of our stereo system under realistic circumstances, we employed our system in indoor environment. The result of a typical surveillance scene is shown in figure 4: upper and middle figures show the event activity profiles of the left and right transient imager respectively, and in the lower figure the corresponding sparse depth map, refered to the left imager, is plotted. The different grey levels encode the event activity and the depth values (given in m) are color-encoded. Darkest red, in depth map, indicates object closest to camera, and darkest blue object farthest from camera.

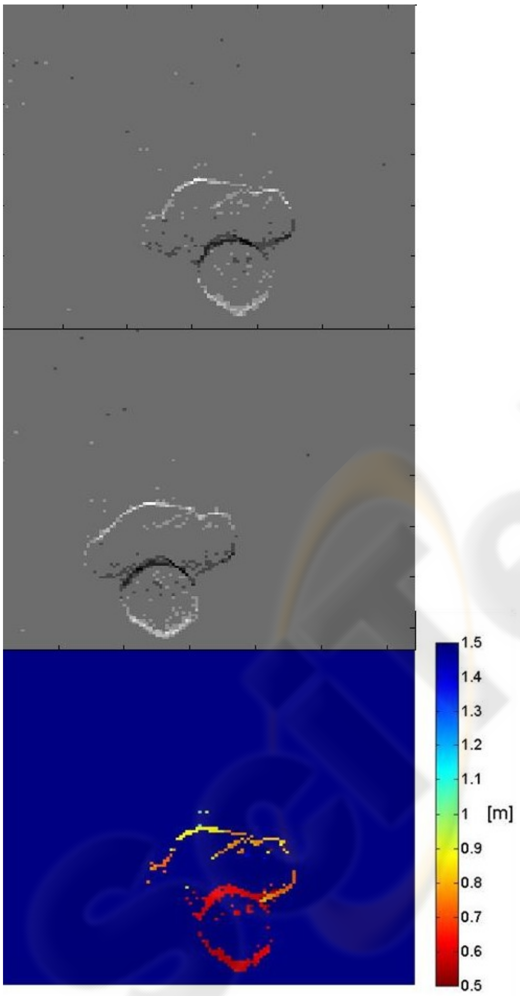

Figure 4: Sparse depth profile of a real surveillance scene (a person moving under the camera mounted on a ceiling) generated from corresponding address-event stereo pair images.

Since we are using real images and because of the characteristic data representation used by transient imager it is very difficult to generate a ground truth data for the purpose of testing our matching algorithm. In our case we evaluate the performance of matching algorithm manually. On average, between 70% and 80% of AEs were matched, and the other events were discarded due to false matching or bidirectional consistency check.

The quantization error of each processing step was not analysed, but instead we have estimated the overall average error of the final depth measurement result. The real world disparity was evaluated by a series of measurement of a rotating disc. Taking equidistant samples we made a data set of which we could estimate the ground truth disparity reasonably accurate and measure the mean root square error (RMSE) of calculated depth values. Figure 5 gives an idea of the overall accuracy of the depth estimation.

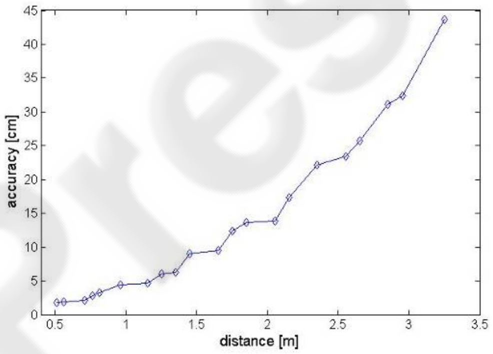

Figure 5: Measured accuracy (RMSE) of our stereo Smartcam using 8mm lenses.

5 CONCLUSIONS AND FUTURE WORK

This paper shows the feasibility of implementing an AE-based stereo algorithm for sparse depth map calculation on a single DSP. The proposed algorithm processes AE-data generated by transient imager and take advantage of its focal pre-processing and efficient asynchronous address-event communication framework.

The complete system is implemented in form of small size, low power and cheap embedded system. Finally, experimental evaluation of our sensor show that is capable to give real-time depth information of moving objects with reasonable accuracy under realistic conditions.

Since our stereo smart camera is a stand-alone, self-contained device that integrates image sensing, stereo depth calculation and communications in one single box, it seems to be suitable for a number of special types of application like:

- Mashine vision in industrial inspection (e.g. 3D-position control; shape, orientation and volume measurement; detection and fast sorting of arbitrary objects on a transportation units)
- Surveillance domain (detecting and counting persons passing through monitored area like doors, portals or terminals; reliable intruders detection due to possibility (Lichtsteiner, 2006) of operating the sensor in night or near-IR range)

ACKNOWLEDGEMENTS

We would like to thank Bernard Kohn und Daniel Bauer for support in hardware implementation. This work is part of VSOC-Project which is funded by "Nationalstiftung für Forschung, Technologie und Entwicklung".

REFERENCES

Boahen, K., 2000. Point-to-Point Connectivity between neuromorphic chips using address events. In *IEEE Trans. on Circuits and Systems II*, 47 (5), 416-433.

Bouguet, J. Y., 2005. *Matlab camera calibration Toolbox*, http://www.vision.caltech.edu/bouguetj/calib_doc.

Di Stefano, L., Marchionni, M., Mattoccia, S., 2004. A fast area-based stereo matching algorithm. In *Image and Vision Computing*, 22, 983-1005.

Häfliger, Ph., Bergh., F., 2002. An integrated circuit computing shift in stereo pictures using time domain spike signals. In *Conference paper NORCHIP 2002*, København.

Kramer, J., 2002. An integrated optical transient sensor. In *IEEE Transactions on Circuits and Systems II,* 49(9), 612-628.

Lichtsteiner, P., Posch, C. and Delbruck, T., 2006. A 128×128 120dB 30mW Asynchronous Vision Sensor that Responds to Relative Intensity Change. In *International Solid State Circuits Conference*, San Francisco, pp. 25-27 .

Lichtsteiner, P., Kramer, J., Delbruck, T., 2004. Improved ON/OFF temporally differentiating address-event imager. In *11th IEEE International Conference on Electronics, Circuits and Systems*, Tel Aviv, Israel.

Moore, R., Lopes, J., 1999. Paper templates. In *TEMPLATE'06, 1st International Conference on Template Production*. INSTICC Press.

Mahowald, M., 1992. *VLSI analogs of neuronal visual processing: a synthesis of form and function*, *Ph.D. dissertation*, California Institute of Technology.

Mortara, A., 1998. A pulsed communication / computation framework for analog VLSI perceptive systems. In *Neuromorphic Systems Engineering*, T. S. Lande, Ed. Norwell, MA: Kluwer Academic, pp. 217-228.

Sivilotti, M., 1991. *Wiring considerations in analog VLSI systems, with application to field-programmable networks", Ph.D. dissertation*, California Institute of Technology.

Zabih, R., Woodfill, J., 1994. Non-parametric Local Transforms for Computing Visual Correspondence. In *Proceedings of 3rd European Conf. on Computer Vision*, pp. 150-158.