Bio-inspired stereo vision system with silicon retina imagers

Jürgen Kogler, Christoph Sulzbachner and Wilfried Kubinger

AIT Austrian Institute of Technology GmbH
A-1220 Vienna, Austria
{juergen.kogler,christoph.sulzbachner,wilfried.kubinger}@ait.ac.at

Abstract. This paper presents a silicon retina-based stereo vision system, which is used for a pre-crash warning application for side impacts. We use silicon retina imagers for this task, because the advantages of the camera, derived from the human vision system, are high temporal resolution up to $1ms$ and the handling of various lighting conditions with a dynamic range of $\sim 120dB$. A silicon retina delivers asynchronous data which are called *address events* (AE). Different stereo matching algorithms are available, but these algorithms normally work with full frame images. In this paper we evaluate how the AE data from the silicon retina sensors must be adapted to work with full-frame area-based and feature-based stereo matching algorithms.

1 Introduction

Advanced Driver Assistance Systems (ADAS) currently available on the market perform a specific function like lane departure warning (LDW), collision warning, or high beam assist. ADAS are entering only slowly into the market because cost-effective solutions are still missing, which would allow extensive market penetration and an increase in number of sensors and supported safety functions.

For example BMW offers for the latest series 5 and 6 a LDW system as optional equipment which costs ∼950$. This price is for vehicles in the higher price segment acceptable, but not for low price and economy vehicles. Recent studies (2005 [12]) have been made to evaluate customer desirability and willingness to pay for active and passive safety systems in passenger cars. The result is that an acceptable price is below the current market prediction, so manufactures need to find cheaper solutions to increase the customer acceptance.

In the EU-funded project ADOSE[1] we use a *Silicon Retina Sensor* for reducing the costs. That kind of sensor overcomes limitations of classical vision systems with high temporal resolution, allowing to react to fast motion in the visual field, on-sensor pre-processing to significantly reduce both memory requirements and processing power, and high dynamic range for dealing with difficult lighting situations encountered in real-world traffic situations. Efficient pre-processing of visual information on the focal plane of the silicon retina vision chip allows cost

[1] www.adose-eu.org

effective computation of scene depth using a single low-cost, low-power *Digital Signal Processor* (DSP).

The silicon retina is specifically tailored to serve as a pre-crash sensor for side impacts (e.g., for the pre-ignition and preparation of a side airbag). In this paper we describe the principle of this sensor technology and how we use the specific silicon retina data in stereo matching algorithms.

2 Bio-inspired silicon retina imagers

The silicon retina imager is derived from the human vision system and is represented by an analog chip which delivers intensity changes as output. Fukushima etal [2] describe in their work an early implementation of an artificial retina. The first retina imager on silicon basis is described in the work of Mead and Mahowald [7], which have also established the term *Silicon Retina*. The work from Litzenberger etal [5] describes a vehicle counting system using the same silicon retina sensor described in the work from Lichtsteiner etal [6], which is developed at the AIT[2]/ETH[3] and also used for the described stereo vision system in this paper.

The silicon retina delivers, for each pixel that has exceeded a defined intensity change threshold, the coordinates of the pixel, a timestamp and the polarity which signals a rising intensity (ON-event) or a falling intensity (OFF-event). The description of the exact data structure from the silicon retina is described in section 2.1. For the setting of the threshold, which defines when an intensity change should trigger an AE, 12 different bias voltages are available in the silicon retina sensor. Each pixel of the silicon retina is connected via an analog circuit to its neighbors which are necessary for the intensity measurements. Based on these additional circuits on the sensor area, the density of the pixels is not as high as on conventional monochrome/color sensors, which results for our sensor in a resolution of 128×128 pixels with a pixel pitch of $40\mu m$.

Due to the low resolution and the asynchronous transmission of AEs from pixels where an intensity change has been occurred, a temporal resolution up to $1ms$ is reached. In figure 1 on the right side the speed of a silicon retina imager compared to a monochrome camera (Basler A601f) is shown. The top image pair on the right shows a running LED pattern with a frequency of $45Hz$. Both camera types recognize the LED movement. The frequency of the LED pattern in the bottom right image pair is $450Hz$. The silicon retina can capture the LED hopping sequence, but the monochrome camera can not capture the fast moving pattern and therefore, more than one LED is visible in a single image. A further benefit of the silicon retina is the high dynamic range up to $120dB$ for various lighting conditions, which is demonstrated in figure 1 on the left side. The top left image pair shows a moving hand in an average illuminated room with an illumination of $\sim 1000\frac{lm}{m^2}$. In both images the hand is clearly visible. In the bottom left image pair a moving hand is captured from both camera types

[2] Austrian Institute of Technology GmbH
[3] Eidgenössische Technische Hochschule Zürich

too, but in a room with an illumination of $\sim 5 \frac{lm}{m^2}$. Here, only the silicon retina sensor recognizes the hand.

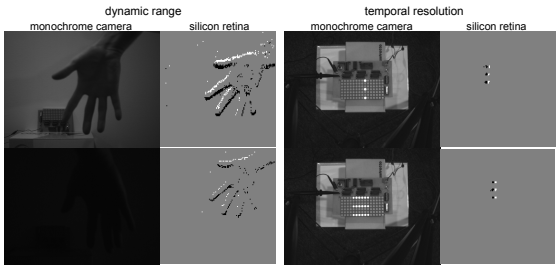

Fig. 1. *Left*: The top pair shows a hand moved under office illumination conditions ($\sim 1000 \frac{lm}{m^2}$) and the lower pair on the left side shows the same scene with an illumination of $\sim 5 \frac{lm}{m^2}$. *Right*: The LED running speed in the top image pair is $45 Hz$ and in the lower image pair on the right $450 Hz$.

2.1 Address-Event data format

The silicon retina is free running and sends only data if the intensity changes generate AEs. These AEs can happen anytime and therefore the silicon retina sensor adds a timestamp, represented by a 32 bit value, to the AEs (location and polarity) and forwards the AEs to the processing unit. The location of the event is addressed by its coordinates (x,y). Both values (x,y) are mapped to a 7 bit representation in the data format. The polarization p of an event is described by one bit. A high bit denotes an OFF-event and low bit an ON-event.

Table 1 shows a comparison between a monochrome sensor and a silicon retina imager with the same resolution for a typical application. Both imagers have different types of data representation and therefore the calculation of the transfer rate is carried out, which makes a direct comparison of both sensors possible. For future purpose the AE data structure will be improved so that the bits/AE decreases and the transfer performance will increase. The data amount of a silicon retina imager with an average address event rate of 50000 AE/s is at the moment ~ 2.3 times lower than a monochrome sensor with $60 fps$.

2.2 Address-Event converter

Before the AE data can be used with full frame image processing algorithms, the data structure is changed into a frame format. For this reason an address event to frame converter has been implemented.

The silicon retina sensor delivers permanently ON- and OFF-events, which are marked with a timestamp t_{ev}. The frame converter collects the address events over a defined time period $\Delta t = [t_{start} : t_{end}]$ and inserts these events into a

Table 1. Data rate of a monochrome sensor and a silicon retina imager

	128×128 monochrome sensor	128×128 silicon retina
Transfered Data	983.040[a] pix/s	50.000[b] AE/s
Data Size	8[c] bit/pix	64[d] bit/AE
Transfer rate	7.5 MBit/s	∼3.2 MBit/s

[a] at 60 fps
[b] average address events (AE) per second (measured with fast movements in front of the camera $\Longrightarrow$ distance $\sim 1m$)
[c] 256 grayscale values
[d] 32 bit timestamp, 15 bit data and 17 bit reserved

frame. After the time period a frame is closed and the generation of the next frame starts. The definition of an event frame is

$$AE_{frame} = \int_{t_{start}}^{t_{end}} AE_{xy}(t_{ev}) dt_{ev} \tag{1}$$

Different algorithm approaches need a different frame format. The silicon retina stereo camera system in this paper is evaluated with two algorithms derived from two different categories. The first algorithm is an area-based approach, which works with the comparison of frame windows. The second algorithm is a feature-based variant which matches identified features. Both categories need differently constructed frames from the converter. Due to this reason, the converter offers configurations to fulfil these requirements. Figure 2 shows on the left side the output frame of the converter with the collected ON- and OFF-events. The resolution of the timestamp mechanism of the silicon retina is $1ms$, but for

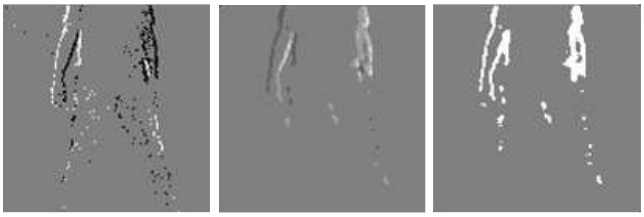

Fig. 2. Different results of AE to frame converter

the algorithm used in this paper a Δt of the $10ms$ and $20ms$ is used. The Δt is changed for different conditions which produce a different number of events.

The image in the middle of figure 2 shows a frame built for an area-based matching algorithm. For this reason each event received in the defined time

period is interpreted as a gray value, with

$$AE_{frame} = \int_{t_{start}}^{t_{end}} graystep(AE_{xy}(t_{ev}))dt_{ev}. \tag{2}$$

The background of the frame is initialized with 128 (based on a 8 bit grayscale model) and each ON-event adds a gray value and an OFF-event subtracts one. In 3 the function for generating a gray value frame is shown. The 8 bit grayscale model limits the additions and subtractions of the $\Delta_{grayvalue}$ and saturates if an overflow occurs.

$$graystep(AE_{xy}(t_{ev})) = \begin{cases} AE_{frame_{xy}} + \Delta_{grayvalue} & AE_{xy}(t_{ev}) = ON_{event} \\ AE_{frame_{xy}} - \Delta_{grayvalue} & AE_{xy}(t_{ev}) = OFF_{event} \end{cases} \tag{3}$$

The right image in figure 2 shows a frame built for a feature-based image processing algorithm. Multiple received events within the defined time period are overwritten in this case of frame building. Equation 4 shows the frame building and the used simplify function is illustrated in (5).

$$AE_{frame} = \int_{t_{start}}^{t_{end}} simplify(AE_{xy}(t_{ev}), conv_{on})dt_{ev} \tag{4}$$

The simplify function gets a second parameter ($conv_{on}$) to decide the event variant (only ON or OFF). This frame is prepared for different kind of feature-based algorithms and also for algorithms based on segmentation.

$$simplify(AE_{xy}(t_{ev}), conv_{on}) = \begin{cases} ON_{ev} & AE_{xy}(t_{ev}) = ON_{ev} \wedge conv_{on} = 1 \\ ON_{ev} & AE_{xy}(t_{ev}) = ON_{ev} \wedge conv_{on} = 0 \\ ON_{ev} & AE_{xy}(t_{ev}) = OFF_{ev} \wedge conv_{on} = 1 \\ 0 & AE_{xy}(t_{ev}) = OFF_{ev} \wedge conv_{on} = 0 \end{cases} \tag{5}$$

Both specialized generated frames (middle and right in figure 2) can optionally filtered with a median filter to reduce noise and small objects. With this settings every Δt a new frame from the left and right address event stream is generated. These frames are now handled as images for the stereo matching algorithms described in the next section.

3 Stereo matching

The main task of this stereo vision sensor is the extraction of depth information from the viewed scenery for the application mentioned in section 1. It is a challenging task to handle the asynchronous incoming AEs for the stereo matching process. Hess [4] used a global disparity filter in his work to find a main disparity of the received events. In a second approach he worked with a general disparity which considers each incoming event separately, but this is a time consuming task and needs a new kind of a stereo matching implementation. In our work we evaluate the opportunity to use standard stereo vision algorithms for AE data from silicon retina imagers. In section 3.1 the suitability of an area-based algorithm is analyzed and a feature-based approach is described in section 3.2.

3.1 Area-based approach for AE stereo matching

For the evaluation of the area-based stereo matching of AE images a simple correlation method is used. In the work from Scharstein and Szeliski [8] many different area-based approaches are compared and evaluated, and for the silicon retina stereo matching a *Sum of Absolute Differences* (SAD) algorithm is used. Before the silicon retina output is processed by the SAD algorithm the data stream is converted into a grayvalue frame (Figure 2 in the middle). A block matching, only with ON- and OFF- events, would produce a lot of similar costs and a lot of mismatches may appear. The grayscale images have more than two values and therefore, the statistical significance of the block is larger.

Derived from the application scenario the distance of a closer coming object must be estimated. Therefore, the distance measurement does not have to be exact and the search space is restricted to one horizontal scanline without a prior rectification step. For each pixel the disparity is calculated and after that the average disparity is calculated which represents the main disparity of the whole object. Results of the algorithm are shown in section 4.2.

3.2 Feature-based approach for AE stereo matching

For feature-based stereo matching, features must be extracted from the image. Shi and Tomasi [10] give more detail about features in their work. For the evaluation of the feature-based stereo matching with silicon retina cameras, a segment center matching approach is chosen. Tang etal [11] describe in their work an approach for matching feature points. An assumption, derived from the application scenario is, that an object comes closer to the sensor and the distance of the object must be estimated. That means no occlusions with other objects and exact distance measurements of each pixel respectively the closer coming object.

Additional processing is required for the extraction of the segment centers, but usually the stereo matching process is less costly. For the segment extraction a morphological *erosion* followed by a *dilation* is applied [3]. After that the *flood fill* labeling [1] function is used, which labels connected areas (segments). A pixel-by-pixel matching is not possible and therefore, it must be defined how the whole segment shall be matched. In a first step the features are ordered downwards according to their area pixel count. This method is only useful, if the found segments in the left and right image are nearly the same. As representative point of the segment the center is chosen. The center of the corresponding segment in the left and right frame can differ. Due to this reason the confidence of the found centers are checked. This mechanism checks the differences of center points, if they are too large, the center points are ignored for the matching. If the center points lie within the predefined tolerances, the disparity is calculated which stands for the disparity of the whole object. Results of the algorithm are shown in section 4.3.

4 Experimental results and discussion

This section presents results of the stereo matching on AE frames. In a first step the sensor setup and configuration used for the tests are described.

4.1 Stereo sensor setup

The stereo system consists of two silicon retina imagers and are mounted on a $0.45m$ baseline. Both cameras are connected to an Ethernet switch, which joins both address event data streams packaged in UDP packets and sends it to the further processing unit. The left and the right camera must have the same knowledge of time, therefore we have realized a master-slave synchronization concept where the master camera sends the timestamp to the slave. In a first step for the offline processing and algorithm evaluation a PC-based system is used.

The silicon retina cameras of the stereo vision system are also equipped with lenses which must be focussed before they can be used. Due to the fact that an output is only delivered if intensity changes are recognized in front of the sensor, a stimuli is necessary which generates an continuous sensor output and can used for the adjustment of the camera lenses. Therefore, we are using blinking lights to get focused camera (depth of field is infinity). The input AE frames for the algorithms are shown in figure 3.

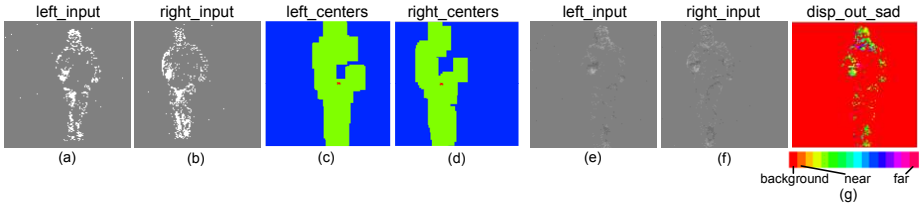

Fig. 3. (a,b): Input pair for the feature-based algorithm. (c,d): Segment centers as disparity representatives. (e,f): Input pair for the area-based algorithm. (g): Disparity output of the SAD.

4.2 Results of the area-based approach

The algorithm parameter of the SAD is the correlation window size. We tested the algorithm with an object at three different distances ($2m$, $4m$, $6m$) and different settings of the address event converter.

In figure 4 the results of the SAD algorithm processing AE frames are given. On the x-axis the different converter settings at three different distances are shown. The first number represents the object distance in meters, the second

value describes the time period for collecting address events and the last value represents the graystep for the accumulation function described in section 2.2. For each distance all four converter settings with four different SAD correlation window sizes are tested. The output on the y-axis is the average relative error of the distance estimation based on 500 image pairs.

The results in figure 4 show that the average relative disparity error increases with the distance of the object. In near distances the results are influenced by the correlation window size, especially there is a significant difference between the usage of a 3×3 window and a 9×9 window. In the distance of $4m$ and $6m$ the results with a timestamp collection time Δt of $20ms$ are better. The third parameter of the generated input AE frame is the grayscale step size which has no influence at any distance. Generally we reach with the SAD stereo matching approach used for AE frames in the main operating distance of $4m$ an minimal error of 8%. That is equivalent to an estimated distance range of $3.68m$-$4.32m$. In figure 3 (e,f,g) an example of an input stereo pair for the area-based algorithm and the SAD disparity output are shown.

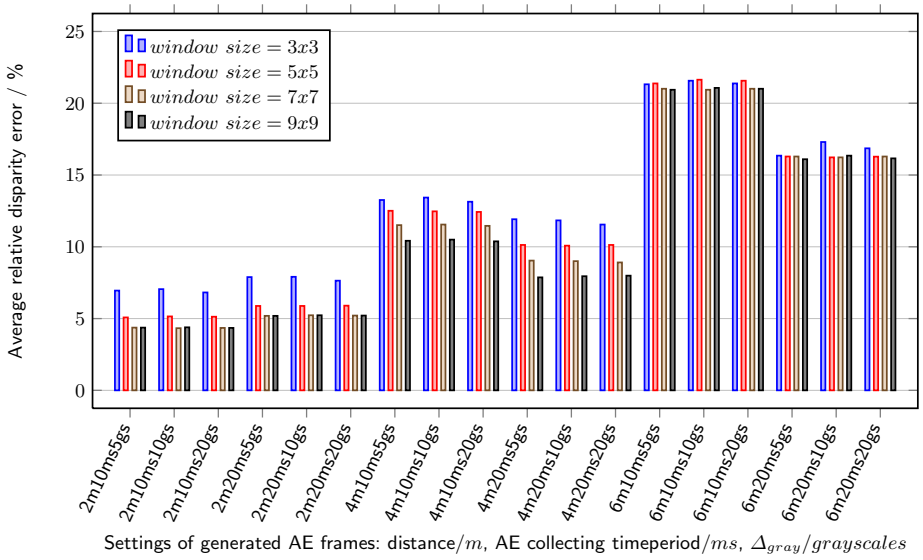

Fig. 4. Results of the area-based stereo matching algorithm on address event frames.

4.3 Results of the feature-based approach

This section shows results of the feature-based stereo matching on AE frames. The algorithm parameter of the feature center matching is the morphological erosion and dilation function at the beginning of the algorithm.

In figure 5 the results of the feature-based algorithm processing AE frames are given. For the center matching only the collecting time period Δt of the address events is varied, which is shown with the second value from the descriptors on the x-axis. All converter settings with three different morphological erosion and dilation settings are tested. The structuring element is always a square. The results on the y-axis shows the average relative disparity error of the feature center matching at three different distance with two different address converter settings and with three different morphological function combinations. The results are based on 500 image pair samples.

The results in figure 5 show that the average relative disparity error depends on the sizes of the structuring elements. At all distances the morphological combination *erosion=3* and *dilation=5* produces the best results. The timestamp collection time Δt has only a significant influence at the distance of $6m$. In the main operating distance of $4m$ the minimal error is 17%. That is equivalent to an estimated distance range of $3.32m$-$4.68m$. In figure 3 (a,b,c,d) an example of an input stereo pair for the feature-based algorithm and the segment centers are shown.

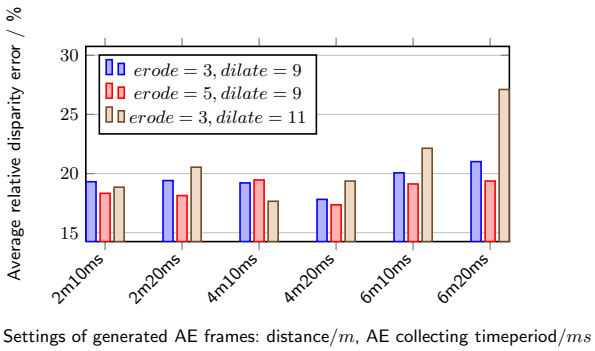

Fig. 5. Results of the feature-based stereo matching algorithm on address event frames.

5 Future work

As presented in this paper, several algorithms which are suited for monochrome or color stereo matching can be used for silicon retina image processing, too. The algorithms have to be adapted and the address event representation of the information has to be converted to conventional data structures representing images (AE frames). This conversion is a time-consuming process and therefore, the advantage of the asynchronous data delivery and high temporal resolution of the silicon retina sensor is not used very efficiently. Additionally, the results showed us, that the feature-based approach produces errors, which are too high for the estimation of the distance. The area-based approach is better and the

estimated distances are precise enough for the usage in the distance estimation of approaching objects.

In the next step we will analyze how we can process the delivered data from the silicon retina camera in a more efficient way. For this reason we want to design an algorithm which can handle the asynchronous data and processes each incoming event without any frame generation strategies. Additionally we would like to implement a suitable calibration and rectification step for the silicon retina stereo vision system, which makes a more exact distance estimation possible. The next version of the algorithm shall run on an embedded platform based on a TMS320C64x+ DSP core from *Texas Instruments* for the evaluation of real time capabilities.

6 Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement n° ICT-216049 (ADOSE).

References

1. Burger, W., Burge, M.J.: Digital Image Processing An Algorithmic Introduction using JAVA. Springer-Science/Business Media LLC, First Edition, (2008)
2. Fukushima, K., Yamaguchi, Y., Yasuda, M. and Nagata, S.: An Electronic Model of the Retina. Proceedings of the IEEE (Volume 58 / Issue 12), 1950–1951, (1970)
3. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Prentice Hall/Pearson Education International, Second Edition, (2002)
4. Hess, P.: Low-Level Stereo Matching using Event-based Silicon Retinas. Semesterarbeit am Institut für Neuroinformatik, ETH Zürich, (2006)
5. Litzenberger, M., Kohn, B., Gritsch, G., Donath, N., Poscha, C., Belbachir, N.A., Garn, H.: Vehicle Counting with an Embedded Traffic Data System using an Optical Transient Sensor. Proceedings of the IEEE Intelligent Transportation Systems Conference (ITSC'07), Seattle/USA, (2007)
6. Lichtsteiner, P., Posch, C. and Delbruck T.: A 128128 120dB 30mW Asynchronous Vision Sensor that Responds to Relative Intensity Change. IEEE International Solid-State Circuits Conference (ISSCC'06), San Francisco/USA, (2006).
7. Mead, C., Mahowald, M.: A silicon model of early visual processing. Neural Networks Journal, Vol 1(1), 91–97, (1988)
8. Scharstein, D., Szeliski, R.: A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. International Journal of Computer Vision, Vol 47(1-3), 7–42, (2002)
9. Schreer, O.: Stereoanalyse und Bildanalyse. Springer-Verlag, (2005)
10. Shi, J., Tomasi, C.: Good Features to Track. Proceedings of the IEEE Computer Vision and Pattern Recognition Conference (CVPR'94), Seattle/USA, (1994)
11. Tang, B., AitBoudaoud, D., Matuszewski, B.J., Shark, L.K.: An Efficient Feature Based Matching Algorithm for Stereo Images. Proceedings of the IEEE Geometric Modeling and Imaging Conference (GMAI'06), London/UK, (2006)
12. Frost & Sullivan, European Markets for Advanced Driver Assistance Systems, B844-18, (2006)