Event-based Stereo Matching Approaches for Frameless Address Event Stereo Data

Jürgen Kogler, Martin Humenberger and Christoph Sulzbachner

AIT Austrian Institute of Technology GmbH
Donau-City-Strasse 1, 1220 Vienna, Austria
{juergen.kogler.fl|martin.humenberger|christoph.sulzbachner}@ait.ac.at

Abstract. In this paper we present different approaches of 3D stereo matching for bio-inspired image sensors. In contrast to conventional digital cameras, this image sensor, called *Silicon Retina*, delivers asynchronous events instead of synchronous intensity or color images. The events represent either an increase (on-event) or a decrease (off-event) of a pixel's intensity. The sensor can provide events with a time resolution of up to $1ms$ and it operates in a dynamic range of up to 120dB. In this work we use two silicon retina cameras as a stereo sensor setup for 3D reconstruction of the observed scene, as already known from conventional cameras. The polarity, the timestamp, and a history of the events are used for stereo matching. Due to the different information content and data type of the events, in comparison to conventional pixels, standard stereo matching approaches cannot directly be used. Thus, we developed an area-based, an event-image-based, and a time-based approach and evaluated the results achieving promising results for stereo matching based on events.

1 Introduction

Different industry, home, or automotive applications use 3D information of the observed scene for reliable operation. Such applications include, e.g., driver assistance, home care, or industrial production. Laser range finders (LIDAR, light detection and ranging), time-of-flight (TOF) cameras, and ultrasonic sound sensors are commonly used for this purpose. All of these sensors are embedded, which means that a processing unit is integrated on-board which processes the raw sensor data and produces the output information directly on the device. Thus, no additional processing power is needed for depth measurement. This is advantageous in different applications where the processing power is limited but the processing effort is high. Additional benefits of embedded systems are the low power consumption and the small form factor. An alternative approach for real 3D sensing, which means not only depth measurement, is stereo vision. Classic stereo vision uses two conventional digital cameras [1], which are mounted side-by-side, separated by the baseline to capture the same scene from two different view points. The exact geometry between the cameras and the correspondences between the images are used to reconstruct the 3D data of the observed scene. A

different approach of stereo vision with digital cameras is using two silicon retina sensors and will be discussed in this work. The silicon retina is an optical sensor which has a time resolution of $1ms$ and a dynamic range of 120dB. Both are advantageous for applications in uncontrolled environments with varying lighting conditions and where fast movements occur. In contrast to conventional image sensors, such as CCD or CMOS, the silicon retina generates so called address events instead of intensity values. Due to the fact that events represent intensity changes in the observed scene, only moving objects or objects with changing color are recognized by the sensor. The data transmission is asynchronous, which means that events are transmitted only once after they occur without a fixed frame rate. With the maximum possible data rate, the system can be arbitrarily dimensioned, e.g., in terms of processing power.

The remainder of the paper is organized as follows. First, section 2 gives an overview about the related work of stereo vision algorithms for conventional cameras and silicon retina sensors. Section 3 describes the silicon retina sensor and its basic characteristics. Then, section 4 presents the three introduced stereo matching approaches for the silicon retina. Finally, section 5 shows the results and gives an outlook about our future research.

2 Related Work

Classic stereo matching algorithms can be subdivided into area- and feature-based approaches. Area-based algorithms use each pixel for correspondence analysis, independent of the scene content. An overview and comparison of different, mostly area-based, techniques is presented in the work of Scharstein and Szeliski [2], Brown *et al.* [3], and Banks *et al.* [4]. Feature-based approaches calculate the correspondences of certain features in the images. Such features are introduced in the work of Shi and Tomasi [5]. In the work of Tang [6] an example of feature-based matching is shown where extracted feature points are connected to chains and further used for the matching step. An interesting point is to analyze if such algorithms are suitable for silicon retina stereo vision. Schraml *et al.* [7] have evaluated several area-based costs functions for grayscale images produced with a silicon retina stereo system (aggregation of events over time), including *Normalized Cross-Correlation* (NCC), *Normalized Sum of Absolute Differences* (NSAD), *Sum of Squared Differences* (SSD) and *Census-Transform*. More detailed information about the grayscale image generation with silicon retina data will be given in section 4.3. The results show that the best method and, thus, chosen for further investigations, is the NSAD approach where the depth measurement has an average error of $\sim$10% in 3m distance. In our previous work [8] we compared an area-based, feature-based, and a new time-based approach, using the characteristics of the silicon retina, with each other. The outcome of the evaluation was that the area-based algorithm achieved nearly the same results as Schraml. The best result of the feature-based algorithm had an average depth error of $\sim$18%. For the newly introduced time-based algorithm we achieved first promising results but the comparison with the other approaches was not possible

because we had to use different test data sets. Both papers proof that existing area-based matching approaches can be used for silicon retina stereo vision as long as images are created. Furthermore, the new time-based approach delivers the most promising results because it directly uses the elementary characteristics of the sensor, event polarity and time, without preprocessing. Therefore, the next step is to compare these algorithms directly in a way to find approaches which strongly exploit all the sensor characteristics.

3 Silicon Retina Stereo Sensor

In 1988, Mead and Mahowald [9] developed an electronic silicon model which reproduced the basic steps of human visual processing. One year later, Mahowald and Mead [10] implemented the first retina sensor based on silicon and established the name *Silicon Retina*.

As a reminder, the silicon retina generates events instead of intensity values. The polarity of the produced events can either be on, for an increase of the intensity, or off, for a decrease. An event encodes the pixel location on chip, the time (with a time resolution of $1ms$) when the event occurred, and the polarity (on or off). The polarity and the time information will further be used to deploy new techniques for correspondence analysis. The data format is called *Address-Event-Representation* (AER) protocol which was introduced by Sivilotti [11] and Mahowald [12] in order to model the transmission of neural information within biological systems. It has to be mentioned that the description of the silicon retina is restricted to the functional behavior due to the algorithmic content of this work. Technical details can be found in the work of Lichtsteiner *et al.* [13, 14]. The new generation of this sensor with a higher resolution in time and space as well as a higher dynamic range is described in the work of Posch *et al.* [15].

4 Silicon Retina Stereo Matching

In this paper we introduce three different stereo matching algorithms for silicon retina data. The first matching algorithm, described in section 4.3, is an area-based SAD (sum of absolute differences) approach which is derived from conventional stereo matching with grayscale or color images. To do this, grayscale images have to be generated out of the fired events. In section 4.4 an event-image-based algorithm is presented which uses images, generated out of aggregated events in a different way, for stereo matching. The third approach presented in section 4.5 is time-based and the correspondence search is directly applied onto the events. In the first and the second approach a pre-processing step is needed to collect the events over time to generate the grayscale or event image. The third method exploits the additional time information and does not need any frame creation. For stereo camera geometry setup, calibration, and rectification we modified and used existing methods, as well as have been modified for the silicon retina sensors.

4.1 Framework

Figure 1 shows the basic work flow which have all three approaches in common. Step 1 is the data acquisition where timed address events are generated by the sensor, then rectified and undistorted with our software. The three stereo matching approaches need three different types of data representation which is explained in detail in the next section. Basically, the mathematical background of silicon retina calibration and rectification is the same as of conventional camera calibration and can be found in [16]. The key difference is the data acquisition because static calibration pattern cannot be used here. To overcome this, we use a flashing checkerboard pattern to generate events. Step 2 is the stereo matching itself and, thus, the main part of the work flow. It consists of the matching costs calculation and optimization. Due to the previous rectification, the search is carried out along an horizontal line within the disparity range. The generated costs represent the probability of correct matching for each possible candidate. Step 3 is the disparity optimization which is a minimum or maximum search through the matching costs of all matching candidates for each pixel up to now. The result is a disparity map, ready for further processing such as 3D reconstruction.

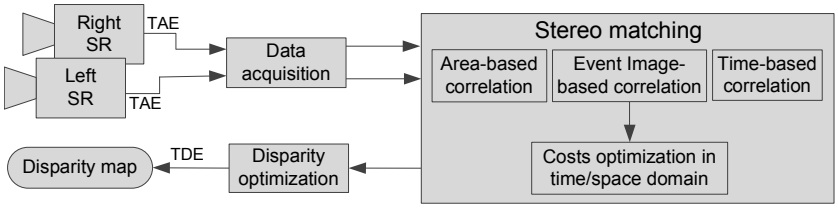

Fig. 1. The work flow of the proposed algorithms

4.2 Data Representation

Before the stereo matching approaches are explained in detail, we want to introduce the used data representation steps, illustrated in figure 2. First, timed-address events fire over a certain time period. They can be described with

$$\mathrm{AE}(u,v,t) = \begin{cases} +1 & I(u,v,t) - I(u,v,t-\Delta t) > \Delta I_{on} \\ -1 & I(u,v,t) - I(u,v,t-\Delta t) < \Delta I_{off} \\ 0 & \text{background - no activity} \end{cases} \tag{1}$$
$$\forall u \in [0,\ldots,H_{res}-1] \wedge \forall v \in [0,\ldots,V_{res}-1],$$

where $I(u,v,t)$ is the intensity of the pixel at position u,v (H_{res} and V_{res} represent the horizontal and vertical resolution of the sensor) and time t. The time is the absolute time from the sensor and the timestamp resolution is $\Delta t = 1ms$. If

the difference between the current intensity value at the time t and the previous intensity value at the time $\Delta t = 1ms$ exceeds a positive threshold ΔI_{on} then an on-event occurred. The difference can also exceed a negative threshold ΔI_{off}, which fires an off-event. Second, if needed, the events are collected and further transformed to proper image formats. In this work, we used event images and grayscale images. The following sections introduce the three different matching approaches where each is based on one of the data representation types. Each approach benefits from space and time data in a different way.

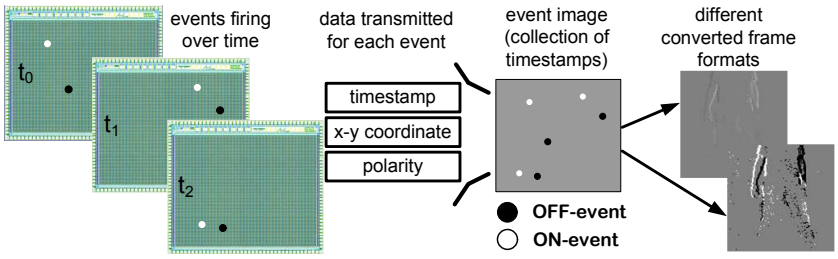

Fig. 2. Data representation types: From left to right: Single events, event collection, transformed grayscale image and event-image

4.3 Area-based Approach

The first approach is dedicated to conventional stereo matching. This means that the events are collected over a certain period of time and a grayscale image, we call it address event frame ($\mathrm{AEF}_{t,L/R}$ at time t for the left and right sensor), is generated out of them with

$$\mathrm{AEF}_{t,L/R}(u,v) = 128 + \sum_{i=0}^{h_{max}} \mathrm{AE}_{L/R}(u,v,t-i\Delta t), \tag{2}$$

where h_{max} is the time history within the events which are considered for the generation of the AEF. The optimum duration of the time history strongly depends on the movement and the speed of the movement in the observed scene. The longer the history, the more details of movements get lost. This address event frame has to be calculated for the left and the right silicon retina camera for each time step. As can be seen in (2), the intensity values of the resulting grayscale image depend on the number of on- and- off events at a position (u, v) during the time period $n\Delta t$ (history). This grayscale image can then further be used for stereo matching algorithms known from conventional stereo vision. Figure 2 shows an example of such a frame generation. The top right image is the resulting grayscale image. Due to the less information such an image contains,

we decided to use a sum of absolute difference (SAD) metric for costs calculation with

$$\text{DSI-AB}_t(u,v,d) = \sum_i^n \sum_j^m |\text{AEF}_{t,R}(u+i,v+j) - \text{AEF}_{t,L}(u-d+i,v+j)|, \tag{3}$$

where n and m define the block for an additional costs aggregation. The calculated costs are stored in the so-called disparity space image (DSI), which is a three-dimensional data structure of the size disparities $\times$ width $\times$ height at time t and is further used to search for the best matching candidate.

4.4 Event-image-based Approach

The second approach generates frames out of the events as well, with the difference that no grayscale image is built. Here, the events firing during the time history are collected only, therefore each occurred event is stored in the address event image ($\text{AEI}_{t,l/r}$ at time t for the left and right sensor) which is built with

$$\begin{gathered}\text{AEI}_{t,L/R}(u,v) = \text{AE}_{L/R}(u,v,t-n\Delta t) \\ \forall n \in [0,\ldots,h_{max}] \ \wedge \text{AE}_{L/R}(u,v,t-n\Delta t) \neq 0.\end{gathered} \tag{4}$$

If a new event fired at the same position, the old event will be overwritten. As well as the area-based grayscale image generation, the event-image-based approach collects events during the time period $n\Delta t$. The main difference is that the events are then directly used without grayscale conversion and the last occurred event within the time at a coordinate is valid. That is the reason why we call the collection-frame, address event image.

For the correspondence search vectors are generated, which have a tri-state logic. The generation of the tri-state vector image AEI-Tri is done with

$$\begin{gathered}\text{AEI-Tri}_{t,L/R}(u,v) = \otimes(\text{AEI}_{t,L/R}(u+i,v+j)) \\ \forall \text{AEI}_{t,L/R}(u,v) \neq 0 \ \wedge i \in [-m,\ldots,m] \ \wedge j \in [-n,\ldots,n]\end{gathered} \tag{5}$$

where the tri-state function $\otimes$ concatenates the neighborhood $m \times n$ of a pixel $\text{AEI}_{t,L/R}(u,v)$ to a vector.

After all vectors for an event image are generated the costs generation takes place. For the tri-state logic a function ξ is used to compare the values of the vectors between left and right. The DSI is calculated with

$$\text{DSI-EB}_t(u,v,d) = \xi(\text{AEI-Tri}_{t,L}(u,v), \text{AEI-Tri}_{t,R}(u-d,v))) \tag{6}$$

where

$$\xi(p_1,p_2) = \sum_i^{m \times n} \begin{cases} 1, p_1(i) <> p_2(i) \\ 0, \text{otherwise} \end{cases} . \tag{7}$$

Figure 3 illustrates with an example how the matching with the tri-state logic works.

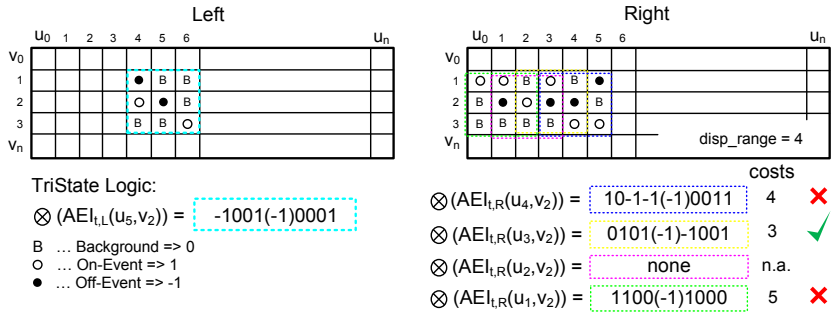

Fig. 3. Event-image-based matching: The neighborhood of the matching candidate is encoded in a bitvector (tri-state logic) which is used for calculating the matching costs

4.5 Time-based Approach

The usage of the former mentioned area-based and event-image-based stereo algorithms has shown a reduction of the advantages of the asynchronous data interface and the high temporal resolution of the silicon retina cameras. Due to this fact, a frame-less and therefore purely event-based stereo matching approach is presented here. It fully exploits the characteristics of the silicon retina technology. The proposed algorithm uses the time difference between events as the primary matching costs. Both silicon retina cameras are perfectly synchronized which means that correct matching events have exactly the same timestamp and polarity. Obviously, there are a few problems known from conventional stereo matching. For example if a series of events fire at the same time (with exactly the same timestamp) when movement is detected in the observed area in front of the camera. Even if the contrast between fore- and background is high enough so that, e.g. for a moving object, at one vertical border only on-events and at the other only off-events fire, there suppose to be more than one matching candidate. This is similar to conventional stereo matching when more than one matching candidate with the same matching costs exists. To overcome this problem, as well as in the approaches mentioned above, a neighborhood of the actual pixel will be taken into consideration. For silicon retina stereo vision, this neighborhood has two dimensions, space, as mentioned above, and time, as will be explained here. The time information for each event gets completely lost in the previous approaches. Here, the costs calculation is first done with

$$\text{DSI-TB}_t(u,v,d) = \begin{cases} \text{wf}(\tau(\text{AE}_L(u,v,t)) - \tau(\text{AE}_R(u-d,v,t_{last}))), \\ \qquad\qquad \text{AE}_L(u,v,t) = \text{AE}_R(u-d,v,t_{last}) \\ 0, \text{no matching candidate} \end{cases} \tag{8}$$

with

$$\tau(\text{AE}(u,v,t)) = t, \tag{9}$$

where the time difference of each possible matching candidate within the disparity range is analyzed. The event time t_{last} of the right side describes the last valid time of the event with respect to the maximum time history.

The weighting function is defined as

$$\mathrm{wf}(\Delta t_{LR}) = \begin{cases} h_{max}\Delta t - \Delta t_{LR}, & \text{invers linear} == \text{true} \\ h_{max}\Delta t \cdot e^{-\frac{\Delta t_{LR}}{a}}, & \text{gaussian} == \text{true} \end{cases}, \tag{10}$$

where it has to be distinguished between inverse linear and Gaussian.

Figure 4 shows an example which clarifies (8) and (9). In the first step, the matching of the events is carried out where for each generated event a corresponding event on the opposite side is searched within the disparity search range. For this search, all events of the current timestamp, as well as events from the past are used. This means also previous events of a certain time history $n\Delta t$ and the same polarity are considered during the correspondence search. In this example, the off-event on the left side is searched in the right side. As can be seen, assuming the time to be a sufficient metric for costs calculation, the best matching candidate is the off-event with disparity 3 and timestamp 10. If this event would not be generated, the best match would be the event with disparity 6 and timestamp 5.

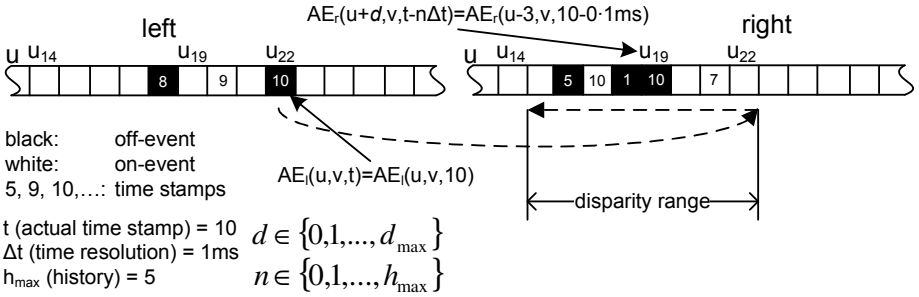

Fig. 4. Time-based matching: The time difference between event are weighted and used as matching costs

In this case this event cannot be correct because it fired at a previous time than the event we are searching for. Thus, a weighting function is used to adapt the costs value (10) in a way that events near the actual event (in time) are more likely to be correct than events more in the past. After that, an aggregation step uses a defined neighborhood to include local information in the matching. If the aggregation step is not done because of any reason, the weighting function is unnecessary.

5 Experimental Results

For the evaluation of the different event-based stereo matching approaches a silicon retina stereo camera setup is used. The system consists of two silicon retina sensors with a baseline of $250mm$, a resolution of 128×128 pixels, a pixel pitch of $40\mu m$, and $12mm$ lenses. The stereo matching algorithm generates a disparity map which is further used for distance calculation of the events. Therefore, a moving object which causes intensity changes and, thus, creates events is placed in front of the stereo sensor system at known distances. In detail, it is a rotating disk with a black and white pattern mounted at distances between $1.5m$ and $5.0m$. All distances are measured with a point laser to have an exact ground truth value. Figure 5 shows the monochrome image (a) of the rotating disk, (b)-(d) show the rotating disk with collected events for time history of $1ms$, $10ms$ and $100ms$.

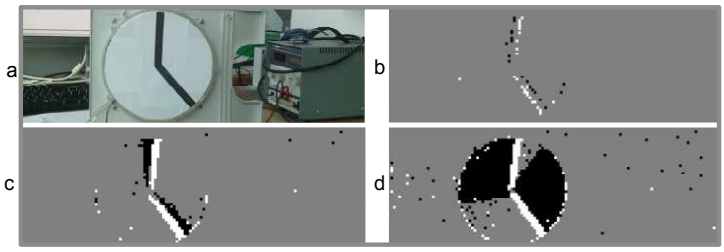

Fig. 5. Rotating disk, (a) monochrome image, silicon retina event-image with a time history of (b) $1ms$, (c) $10ms$ and (d) $100ms$.

The rotating disk is placed in $0.5m$ intervals between $1.5m$ and $5.0m$ in front of the cameras. Before the stereo matching a calibration step is applied. In the first test the calculated disparities from the stereo matching algorithms are compared with the measured ground truth distances of the rotating disk.

Table 1 lists the results of the stereo matching using the three introduced stereo matching approaches. For evaluation the average disparity of all pixels within a time history of $5ms$ is used. The table shows the average disparity, the calculated distance, and the average distance error between ground truth distance and processed distance. Additionally, an average error over all distances summarizes the results for each algorithm approach.

The results show an average distance error of 4.91% if the area-based approach is used followed by the event-image-based approach which achieves an average error of 4.93%. The time-based stereo matching algorithm achieves the best result with an average error of 3.48%.

In the second test we have determined the detection rate of all three stereo matching approaches. For this test the disparity maps in the distances $1.5m$, $2.5m$ and $5.0m$ with all stereo matching algorithms are calculated. Figure 6 shows the average detection rate calculated for a period of $2.5s$.

Table 1. Calculated distances and average distance errors of all three different stereo matching approaches

distances	Area-based			Event-image-based			Time-based		
	disp (px)	z (m)	err. (%)	disp (px)	z (m)	err. (%)	disp (px)	z (m)	err. (%)
1.5m	45.7	1.63	8.67	44.4	1.68	12.0	48.0	1.55	3.33
2.0m	35.7	2.09	4.50	35.3	2.11	5.50	36.6	2.04	2.00
2.5m	28.4	2.63	5.20	28.4	2.63	5.20	29.0	2.57	2.80
3.0m	23.6	3.16	5.33	23.8	3.13	4.33	24.0	3.11	3.67
3.5m	20.1	3.71	6.00	20.4	3.66	4.57	20.5	3.64	4.00
4.0m	17.7	4.21	5.25	18.0	4.14	3.50	17.9	4.17	4.25
4.5m	16.1	4.63	2.89	16.0	4.66	3.56	15.9	4.69	4.22
5.0m	14.7	5.07	1.40	14.8	5.04	0.81	14.4	5.18	3.60
average			4.91			4.93			3.48

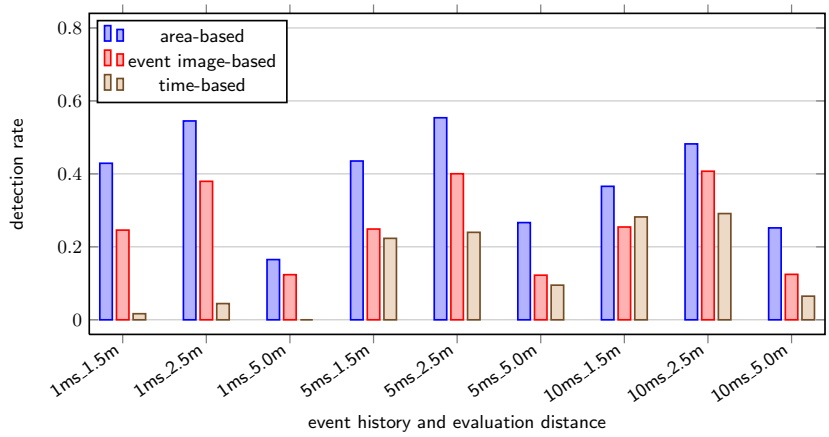

Fig. 6. Average detection rate of all matching algorithms within a time period of $2.5s$

A disparity value within a range of -1 and +1 of the ground truth value is set as correct. Additionally, the detection rate is determined with three different time histories ($1ms$, $5ms$ and $10ms$). Figure 6 shows that the area-based algorithm has the best average detection rate up to ∼55% for each time history and at the distance of $2.5m$.

Figure 7 shows the best detection rate measured for one timestamp within an analysis time of $2.5s$. These results represent the best possible outcome of the stereo matching algorithms.

Figure 7 depicts that the detection rate achieved in a distance of $5.0m$ is up to 100%. This can be led back to the fact that in a distance of $5.0m$ the rotating disk produces only a few events and it can easily happen that these few events are perfectly matched.

In figure 6 and 7 the area-based approach generally achieves the better results. The area-based approach has the best average detection rates at the middle

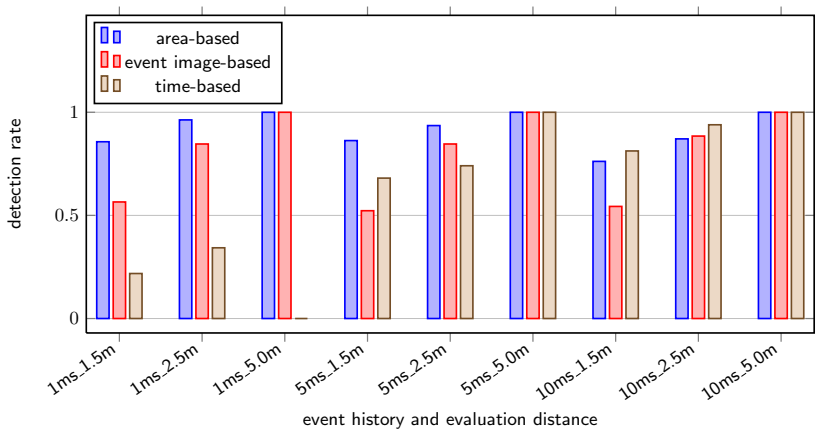

Fig. 7. Best detection rate measured for one timestamp within a time period of 2.5s

distance and lower time histories for the image generation. Taken the best detection rate under consideration, the area-based method delivers improved results with a marginal influence of the time history. The time-based algorithm depicts a bad detection rate with low time histories but it significantly increases for a rising time history. The result of the event-image-based approach lies between the detection rate of the other matching algorithms.

Summarizing all results the time-based approach has the best accuracy in consideration of the absolute distance estimation. In contrast, the area-based approach has a better detection rate. This leads to the conclusion that an algorithm combination of both approaches could result in a better stereo matching algorithm for silicon retina stereo camera systems.

6 Conclusion and Future Work

In this paper we presented the development and evaluation of different stereo matching algorithms for a silicon retina stereo vision system. The silicon retina technology differs in comparison to conventional image sensors and therefore existing algorithms can only be used with proper adaptations. Furthermore we showed that silicon retina data can be directly used for stereo matching as well. All introduced algorithm approaches are tested with real world data which delivered promising results. In a next step we will refine the algorithms and combine them in different ways to benefit from the strength of each single approach.

Acknowledgment

This work is supported by the AAL JP project Grant CARE "aal-2008-1-078". The authors would like to thank all CARE participants working on the success of the project.

References

1. Belbachir, A.: Smart Cameras. Springer New York Dordrecht Heidelberg London (2010)
2. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision **47** (2002) 7–42
3. Brown, M.Z., Burschka, D., Hager, G.D.: Advances in computational stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence **25** (2003) 993–1008
4. Banks, J., Bennamoun, M., Corke, P.: Non-parametric techniques for fast and robust stereo matching. In: Proceedings of the IEEE Region 10th Annual Conference on Speech and Image Technologies for Computing and Telecommunications, Brisbane/Australia (1997) 365–368
5. Shi, J., Tomasi, C.: Good features to track. In: Proceedings of the IEEE Computer Vision and Pattern Recognition Conference, Seattle/USA (1994) 593–600
6. Tang, B., AitBoudaoud, D., Matuszewski, B., Shark, L.: An efficient feature based matching algorithm for stereo images. In: roceedings of the IEEE Geometric Modeling and Imaging Conference, London/UK (2006) 195–202
7. Schraml, S., Schön, P., Milosevic, N.: Smartcam for real-time stereo vision - address-event based embedded system. In: In: Ranchordas, Alpesh, Arajo, Helder and Vitri, Jordi (eds.) VISAPP 2007 - Proceedings of the Second International Conference on Computer Vision Theory and Applications. Volume 2., Barcelona/Spain (2007) 466–471
8. Kogler, J., Sulzbachner, C., Humenberger, M., Eibensteiner, F.: Address-event based stereo vision with bio-inspired silicon retina imagers. In: Advances in Theory and Applications of Stereo Vision edited by Asim Bhatti, InTech Open Books (2011)
9. Mead, C., Mahowald, M.: A silicon model of early visual processing. Neural Networks Journal **1** (1988) 91–97
10. Mahowald, M., Mead, C.: Silicon retina. Analog VLSI and Neural Systems (1989) 257–278
11. Sivilotti, M.: Wiring consideration in analog vlsi systems with application to field programmable networks. Phd-thesis, California Institute of Technology (1991)
12. Mahowald, M.: VLSI analogs of neuronal visual processing: a synthesis of form and function. Phd-thesis, California Institute of Technology (1992)
13. Lichtsteiner, P., Posch, C., Delbruck, T.: A 128×128 120db 30mw asynchronous vision sensor that responds to relative intensity change. In: Proceedings of the IEEE International Solid-State Circuits Conference, SanFrancisco/USA (2006)
14. Lichtsteiner, P., Posch, C., Delbruck, T.: A 128×128 120 db 15 μs latency asynchronous temporal constrast vision sensor. IEEE journal of solid-state circuits **43** (2008)
15. C. Posch and D. Matolin and R. Wohlgenannt: A QVGA 143 dB Dynamic Range Frame-Free PWM Image Sensor With Lossless Pixel-Level Video Compression and Time-Domain CDS. IEEE Journal of Solid-State Circuits **46** (2011) 259–275
16. Zhang, Z.: A flexible new technique for camera calibration. Technical Report MSRTR9-71, Microsoft Research (2002)